Editorial

AN EXCELLENT news that I am very happy to share with our readers is the inclusion of our journal in the ISI Web of Science, as part of its SciELO collection. This opens new international horizons to our journal, both in terms of readership, which, with the great visibility of the journal through the most authoritative scientific information platform in the world will now be more numerous and more international, and in terms of the excellence of the published materials, since the journal is now more than even attractive for international authors. No doubt, in the coming years we will witness growth of quality and reputation of our journal.

This achievement is the result of a great work of the new editorial team, which has renewed the image of the journal starting from 2008 and is constantly working on its improvement. I would like to congratulate our associate editors, reviewers, and most importantly, our authors and readers with this new achievement and thank them for their constant effort devoted to the journal. Very special thanks go to Antonio Sánchez Pereyra and his team from the Latin American Bibliography department of the UNAM and SciELO Mexico for their constant support and great help in our editorial work.

This issue of the journal Polibits includes ten papers by authors from nine different countries: Argentina, Brazil, Chile, Colombia, Ecuador, Mexico, Switzerland, USA, and Vietnam. The papers included in this issue are devoted to such topics as computer vision, machine learning, bioinformatics, natural language processing, computer-assisted learning, databases, computer security, logic, optimization, and scheduling.

C. L. Sabharwal and **J. L. Leopold** from **USA** in their paper "A Completeness of Metrics for Topological Relations in 3D Qualitative Spatial Reasoning" propose novel metrics for qualitative spatial reasoning. Reasoning about locations of objects, such as whether they touch each other or intersect, to which direction one object is located with respect to another, how far the two objects are from each other, etc., has important application in a number of areas, ranging from robotics, computer vision, and geoprocessing (think Google Maps) to natural language generation, when it comes to a natural language description of a scene or, say, driving directions (think again Google Maps). While previous works concentrated on object on a plane, the authors present their metrics for the objects in the three-dimensional space.

R. Z. Nunes Marques et al. from **Brazil** in their paper "EMiner: A Tool for Selecting Classification Algorithms and Optimal Parameters" describe their system that helps in selecting machine learning algorithms and their parameters. While there is a wide range of machine learning techniques,

selecting the right one for a particular task remains an art very difficult to master; in fact this is the very reason for the existence of so many different techniques for more or less the same task: there is no one-size-fits-all methods. What is more, many of these techniques have parameters that should be set by the user—while users often have little clue of how to choose them. A system capable of automatically make such decision will be of great help for the users of machine learning algorithms.

F. Rinaldi et al. from **Switzerland** and **Mexico** in their paper "An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database" explain how natural language processing technology can be used for creation biological knowledge bases. The amount of published literature on biological processes and properties of living organisms by far exceeds any human ability for information processing, and keeps growing. To be able to orient in this ocean of information, the researchers are in need of structured, ordered, well-organized knowledge bases semi-automatically learnt from this huge body of published literature. One of such efforts is RegulonDB, a large biomedical knowledge base. The authors' current effort is devoted to semi-automatic enrichment of this knowledge base.

C. Peñuela et al. from **Colombia** in their paper "Warnings and Recommendation System for an E-Learning Platform" introduce a system for timely alerting students of possible learning problems and signs of poor academic performance and recommending them learning material with which they can improve their learning trajectory. The system relies on a set of existing user profiles and on the observation of the online behavior of the given student to detect deviation from the expected trajectory and early signs of failure. It is especially valuable that the student is not only capable of alerting the student but also of recommending necessary measure to improve the alarming situation.

N. Rodriguez and **L. Barba** from **Chile** and **Ecuador** in their paper "Bi-variate Wavelet Autoregressive Model for Multi-step-ahead Forecasting of Fish Catches" consider a task of time series forecasting; specifically, forecasting of fishery outcome. They apply Fourier spectrum analysis to the time series in question to separate high-frequency component, responsible for short-term predictions, from the low-frequency component that can be used to predict the long-term average behavior of the signal. While it is common to just discard the high-frequency component, the authors build a bi-variant model that uses both component for more accurate forecasting.

T. K. Dang and **C. N. Ngo** from **Vietnam** in their paper "Location Privacy-Aware Nearest-Neighbor Query with

Complex Cloaked Regions" address the problem of protecting users' privacy in the world of proliferation of ubiquitous Internet-connected location-aware portable devices. A common technique related to this problem is to handle the location of the user in communication with external services not as a point but as a region that is still meaningful for the service (such as "in such city" or "in such supermarket") but does not pose too high security risk. While previous works consider mainly rectangular regions, the authors deal with complexly-shaped regions: a city or a supermarket are not rectangles.

R. Flores-Carapia et al. from **Mexico** in their paper "Cipher Image Damage: An Application of Filters" present a technique for restoring encrypted images after the data have been damaged, which can be, for example, the result of an attack on the computer system, of hardware failure, or of communication problems. The paper combines image processing technique with cryptographic techniques. The images considered by the authors range from photos of human faces to dense small-print text. The authors also propose statistical and information-theoretic tests to measure the quality of the image restoration process.

E. Zurek et al. from **Colombia** in their paper "An Implementation of Propositional Logic Resolution Applying a Novel Specific Algebra" propose a novel approach to the problem of satisfiability of a logical formula, which is one of fundamental problems of automatic reasoning. For this, the authors develop a logical algebra that allows them to use the resolution-refutation technique to refute or prove satisfiability of the formula.

R. de Abreu Batista and **A. L. Cetertich Bazzan** from **Brazil** in their paper "Identification of Central Points in Road Networks using Betweenness Centrality Combined with Traffic Demand" address a practical problem of traffic control in a large city. They solve an important task of identifying central points in the road network, which are highly susceptible to bottleneck and jam problems. They apply their technique to analysis of traffic networks of two Brazilian cities, Porto Alegre and Sioux Falls. Their research has implication for urban design, optimization of traffic-related services, and control of traffic in large cities, thus reducing the costs, time spent by people in the traffic jams, as well as risks of traffic accidents.

V. Yannibelli and **A. Amandi** from **Argentina** in their paper "Project Scheduling: A Memetic Algorithm with Diversity-Adaptive Components that Optimizes the Effectiveness of Human Resources" introduce a technique that relies on a memetic algorithm for optimization of the use of human resources. Human resources is one of the most important, and most costly, components of the functioning of companies and organizations; thus, the optimization of their use can result in large cost savings for a company. While a large number of techniques have been proposed for this purpose, the authors' results significantly outperform the current state of the art on all six test problems considered in this paper.

This issue of the journal will be useful to researchers, students, and practitioners working in the corresponding areas, as well as to general public interested in advances in computer science, artificial intelligence, and computer engineering.

Dr. Alexander Gelbukh

Instituto Politécnico Nacional, Mexico City, Mexico
Editor-in-Chief

A Completeness of Metrics for Topological Relations in 3D Qualitative Spatial Reasoning

Chaman L. Sabharwal and Jennifer L. Leopold

Abstract—For qualitative spatial reasoning, there are various dimensions of objects. A considerable amount of effort has been devoted to 2D representation and analysis of spatial relations. Here we present an exposition for 3D objects. There are three types of binary relations between pairs of objects: topological connectivity, cardinal directions, and distance relations. The combinations of these relations can provide additional useful knowledge. The spatial databases include data and the spatial relations to facilitate end-user spatial querying, it also is important to associate natural language with these relations. Some work has been done in this regard for line-region and region-region topological relations in 2D, and very recent work has initiated the association between natural language, topology, and metrics for 3D objects. However, prior efforts have lacked rigorous analysis, expressive power, and completeness of the associated metrics. Herein we present a detailed study of new metrics required to bridge the gap between topological connectivity and size information for integrating reasoning in spatial databases. The complete set of metrics that we present should be useful for a variety of applications dealing with 3D objects including regions with vague boundaries.

Keywords—Region connection calculus, metrics, spatial reasoning, qualitative reasoning.

I. Introduction

S PATIAL relation theory is the foundation for using spatial databases effectively. In qualitative spatial reasoning, there are various types of object dimensions. A considerable amount of effort has been devoted to 2D. Here we present exposition for 3D objects. There are three types of binary relations between pairs of objects. The combinations of these relations can provide additional useful knowledge required for some applications. The spatial databases include data and the spatial relations to facilitate end-user spatial querying and retrieving and analyzing spatial knowledge quickly. Qualitative spatial reasoning is intrinsically useful even when spatial information is imprecise or incomplete. The reasons are: (1) precise information may not be available or may not even be required, (2) detailed parameters may not be necessary before proceeding to decision making, and (3) complex decisions sometimes must be made in a relatively short period of time.

However, qualitative reasoning can result in ambiguous solutions due to incomplete or imprecise quantitative information. In RCC8 [1], [2], the regions have a well-defined interior, boundary, and exterior. The RCC8 relations are bivalent with true and false crisp values. Mathematically defined and computer drawn objects are crisp and well defined, whereas hand-drawn regions tend to have a vague boundary [3]. When regions are vague, the relations between regions can be vague also. That results in the possible values for relations being true, false, or even 'maybe.' RCC8 assumes that regions are crisp; hence the relations are crisp. While topology is sufficient to determine the spatial connectivity relations, it lacks the capability to determine the degree (or extent) of connectivity or separation of such relations.

For example, in Fig. 1, for two objects A and B, the RCC8 disjoint relation, DC(A,B), evaluates to true, yet it does not provide any information about the degree of separation; we do not know how close the two objects are — are they almost touching or are they far apart? The usefulness of metrics lies in providing such additional information, which can be useful in some applications.
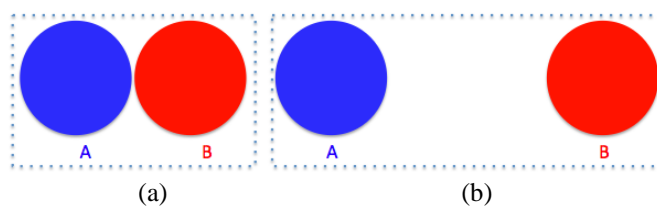
Fig. 1. Two examples of disjoint objects. RCC8 determines that A and B are disjoint, yet it does not tell if they are almost touching or far apart. That is, it does not quantify the degree of disjointness.

For another example, in Fig. 2, for two objects A and B, the RCC8 proper overlap relation, PO(A,B), evaluates to true, yet it does not provide any information about the degree of connectivity; we do not know how much is the overlap—are the objects barely overlapping or are they are almost equal?
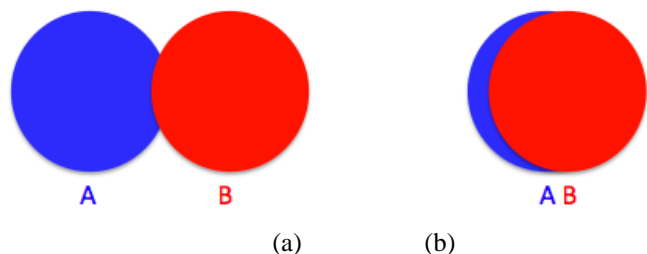

(a)   (b)

Fig. 2. Two examples of proper overlap. RCC8 determines that there is an overlap between A and B, but it does not quantify the proper overlap. In (a) the objects are barely overlapping, and in (b) they are almost equal.

Similarly in Fig. 2, a metric can determine which object is larger or if they are equal independent of location. Metrics are quantitative, whereas topology is qualitative; both together can supplement each other in terms of spatial knowledge. The metric refinements provide for quality of connectivity of each relation. The goal of this exposition is to bridge the gap between topology and size via metrics.

The paper is organized as follows. Section II provides a brief mathematical background relevant to subsequent discussions in the paper. Section III explains the motivation for metrics. Section IV discusses the development of our metrics, as well as the association between size and topology. Section V explains the association between connectivity, size, and metrics. Section VI gives the conclusion and future directions.

II. BACKGROUND

A. Spatial Relations in General

Historically, there are two approaches to topological region connection calculus, one is based on first order logic [1], and the second is based on the 9-intersection model [2]. Both of these approaches assume that regions are in 2D and the regions are crisp, and that relation membership values are true and false only. For qualitative distances, metrics were used in 1D to differentiate relative terms of proximity like *very close*, *close*, *far*, and *very far* [4]. To refine natural language ontology and topological relationships, metrics were introduced for line-region and region-region connectivity in 2D [5]. These approaches lack expressing the strength of relation, and the combination of the connectivity and size information. Recently more attention has been directed to these issues for vague regions in 2D [6] and for natural language ontology in 3D [7]. However, prior work has been deficient in rigorous analysis, expressive

power, and completeness of the metrics. The complete set of metrics presented herein differs from the previous approaches in its completeness and enhanced expressiveness.

B. Mathematical Preliminaries

R^3 denotes the three-dimensional space endowed with a distance metric. Here the mathematical notions of *subset, proper subset, equal sets, empty set ($\emptyset$), union, intersection, universal complement,* and *relative complement* are the same as those typically defined in set theory. The notions of *neighborhood, open set, closed set, limit point, boundary, interior, exterior,* and *closure* of sets are as in point-set topology. The interior, boundary, and exterior of any region are disjoint, and their union is the universe.

A set is *connected* if it cannot be represented as the union of disjoint open sets. For any non-empty bounded set A, we use symbols A^c, A^i, A^b, and A^e to represent the universal complement, interior (Int(A)), boundary (Bnd(A)), and exterior (Ext(A)) of a set A, respectively. Two regions A and B are equal if $A^i == B^i$, $A^b == B^b$, and $A^e == B^e$ are true. For our discussion, we assume that every region A is a non-empty, bounded, regular closed, connected set without holes; specifically, A^b is a closed curve in 2D, and a closed surface in 3D. A spatial region A is closed if it contains the interior and boundary, and is denoted by $\overline{A}$. Thus the regions are regular closed sets, meaning closure of the interior of a region is itself, $A = \overline{A^i}$. For spatial regions, we use *weak* connectivity: two regions A and B are connected if $\overline{A} \cap B \neq \emptyset$.

C. Metric Spaces and Spatial Metrics

Topologically a metric m on a metric space satisfies three properties: (1) $m(A, A) = 0$, *identity*; (2) $m(A, B) = m(B, A)$, *symmetry*; and (3) $m(A, B) \leq m(A, C) + m(C, B)$, *triangle inequality*. Furthermore a metric is translation invariant if $m(A,B) = m(A + t, B + t)$ where A + t is the translation of object A by t. Most of the time, the metric represents the Euclidean distance between a pair of objects. For spatial objects a qualitative metric does not necessarily follow this rule.

We will define metrics for topological relations that capture the semantic knowledge about the relation between a pair of objects. These metrics overcome the limitations of topological spatial relations. For example, to measure the part of A split by B (i.e., the part common to A and B), the metric is defined by $m(A, B) = \text{volume}(A \cap B)/\text{volume}(A)$. As such, for this metric m, (1) $m(A, A) = 1$, not zero, *anti-identity*; (2) $m(A, B)$ is not necessarily equal to $m(B, A)$, *anti-symmetric*; and (3) m does not satisfy the triangle inequality, *anti-triangle inequality*. For our purpose this metric provides very useful information to determine the quality of topological connectivity relations (see Section V).

D. Region Connection Calculus Spatial Relations

Much of the foundational research on qualitative spatial reasoning concerns a region connection calculus (RCC) that describes 2D regions (i.e., topological space) by their possible relations to each other. RCC8 can be formalized by using first order logic [1] or using the 9-intersection model [2]. Conceptually, for any two regions, there are three possibilities: (1) *One object is outside the other*; this results in the RCC8 relation DC (disconnected) or EC (externally connected). (2) *One object overlaps the other across boundaries*; this corresponds to the RCC8 relation PO (proper overlap). (3) *One object is inside the other*; this results in topological relation EQ (equal) or PP (proper part). To make the relations jointly exhaustive and pairwise distinct (JEPD), there is a converse relation denoted by PPc (proper part *converse*), PPc(A,B) ≡ PP(B,A). For completeness, RCC8 decomposes proper part into two relations: TPP (tangential proper part) and NTPP (non-tangential Proper part). Similarly for PPc, RCC8 defines TPPc and NTPPc. The RCC8 relations are pictorially described in Fig. 3.
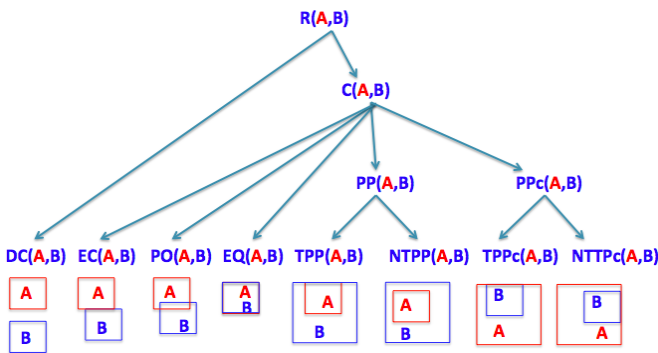

Fig. 3. RCC8 Relations in 2D.

Each of the RCC8 relations can be uniquely described by using a 9-Intersection framework. This is a comprehensive way to look at the relation between two regions. Table 1 depicts the 9-Intersection matrix between two regions A and B, where Int represents the region's interior, Bnd denotes the boundary, and Ext represents the exterior. The predicate IntInt(A,B) is a binary relation that represents the intersection between the interiors of region A and region B; the value of this function is either true (non-empty) or false (empty) for that intersection. Similarly, there are other predicates for the intersection of A's interior, exterior, or boundary with those of B.

For two non-empty bounded regions, A and B, the intersection of their exteriors is always non-empty; it adds nothing to the discrimination and knowledge discovery about regions. In our prior work [7], we have consistently replaced the 9-Intersection 3×3 matrix with the 8-Intersection to define the spatial relations. The values of the 8-Intersection for the RCC8 relations are given below in Table 2.

TABLE 1.
9-INTERSECTION 3×3 MATRIX AND REDUCED 4-INTERSECTION 2×2 MATRIX (SHADED) FOR CALCULATING RCC8 RELATIONS

	Interior	Boundary	Exterior
Interior	Int(A)∩Int(B)	Int(A)∩Bnd(B)	Int(A)∩Ext(B)
Boundary	Bnd(A)∩Int(B)	Bnd(A)∩Bnd(B)	Bnd(A)∩Ext(B)
Exterior	Ext(A)∩Int(B)	Ext(A)∩Bnd(B)	Ext(A)∩Ext(B)

From careful analysis of Table 2, we see that the IntInt and BndBnd columns have the most useful information in the sense that they are sufficient to partition the RCC8 relations into three classes: {DC, EC}, {NTTP, NTTPc}, and {PO, EQ, TPP, TPPc}.

TABLE 2.
THE VALUES OF THE 8-INTERSECTION VECTORS AND 4-INTERSECTION VECTORS (SHADED) THAT ARE REQUIRED TO DISTINGUISH RCC8 RELATIONS

	IntInt	BndBnd	IntBnd	BndInt	IntExt	BndExt	ExtInt	ExtBnd
DC	F	F	F	F	T	T	T	T
EC	F	T	F	F	T	T	T	T
NTPP	T	F	F	T	F	F	T	T
NTPPc	T	F	T	F	T	T	F	F
EQ	T	T	F	F	F	F	F	F
TPP	T	T	F	T	F	F	T	T
TPPc	T	T	T	F	T	T	F	F
PO	T	T	T	T	T	T	T	T

Further analysis of Table 2 indicates that only 4-intersections are sufficient for classification of topological relations [11]. This table can be interpreted and formulated in terms of rules for system integration. These rules are displayed for visualization in the form of a decision tree in Fig. 4.
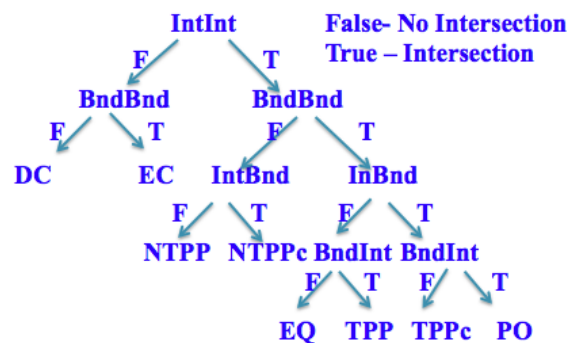

Fig. 4. Tree for classifying the topological relations, (T/F) represents whether or not the objects intersect.

Originally, region connection calculus was designed for 2D [1], [2]; later it was extended to 3D [8], [6]. In [5], metrics were used for associating line-region and region-region connectivity in 2D to natural language. The metrics were adapted from [5] for qualitative study of the dependency

between metrics and topological relations, and between metrics and natural-language terms; conclusions then were drawn for association between several natural-language terms and the topological connectivity RCC8 terms [7]. However, the 2D metrics were adopted and adapted to 3D objects without any regard for viability or completeness. Herein we introduce new metrics and describe a complete set of metrics to explore the degree of association between them in terms of strength of connectivity and relative size information.

III. Motivation for Metrics

In qualitative spatial reasoning, there are three distinct properties for reasoning about spatial objects: connection, dimension, and direction. Reasoning over *combinations* of these properties can provide additional useful knowledge. The prior efforts [5] have lacked rigorous analysis, expressive power, and completeness of the associated metrics. Revision of the metrics is required before we can begin to bridge the gap between topological connectivity and size information for automated spatial reasoning.

We start with following example for motivation to study the degree (or extent) of spatial relations. This example centers around one metric and one pair of objects; see Fig. 5 for concept illustration. Consider the volume of interior of an object A split by the volume of interior of an object B; let this be denoted by metric, IVsIV(A,B). The definition of metric considers one object, the first parameter, as the reference object and the other object, the second parameter, as the target object. The target object is the object that performs the splitting action on the reference object. This metric calculates how much of reference object A is part of the target object B. Since sizes of objects can vary in units of measurement, it is more reasonable to compare qualitative or relative sizes for objects. Recall from section II.B that A^i represents the interior of A. We define the relative (i.e., normalized) part of A common in B by the equation,

$$IVsIV(A,B) = \frac{volume(A^i \subsetneq B^i)}{volume(A^i)}$$

With this metric, let us see in what ways, the connectivity and size information are useful in spatial reasoning.

(1) *RCC8 topological relation:* Suppose that for objects A and B in Fig. 5, we have IVsIV(B,A) = 1. This implies B is a proper part of A, PP(B,A), which is an RCC8 qualitative connectivity relation. Without the metric, in general, this relation is computed by using the 9-intersection model involving various pairwise intersections before arriving at this conclusion [2], [10]. The metric provides this information much more quickly and efficiently.

(2) *Size relations*: In Fig. 5, suppose IVsIV(A,B) = 0.1, which implies that 10% of A is part of B and part of B is 10% of A. From step (1), IVsIV(B,A) = 1, which means all of B is a part of A. With both the metrics, we conclude that B is equal

to 10% of A, and that B is part of A. Therefore, B is much smaller than A for the size relation (i.e., A is much bigger than B). In general, if IVsIV(A,B) < IVsIV(B,A), then A is larger than B in size (i.e., or B is smaller than A in size). Thus the metric is a useful tool for qualitative size comparison of pairs of objects.
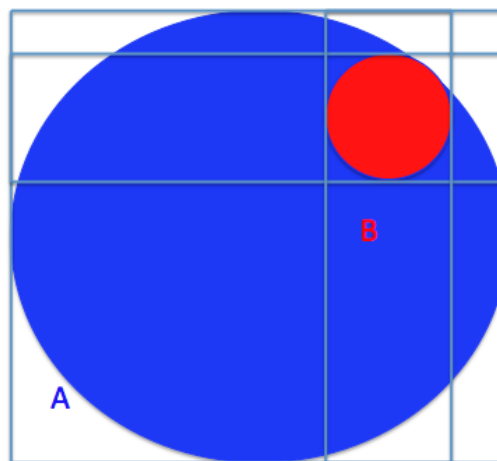

Fig. 5. Object B is a proper part of A, B is much smaller than A in size, and B is in the northeast relative to A.

(3) *Cardinal direction relations:* We will concentrate on steps (1) and (2) in this paper. The grid is generated by grid lines for A and B, where the minimum-bounding rectangle is composed of horizontal and vertical gridlines. The detailed discussion of directions metrics is beyond the scope of this exposition; the reader may consult [9], [10]. The direction metric in [9], [10] determines that B is in the northeast of part of A. With this directional knowledge, it means that in addition to B being a tangential proper part of A, TPP(B,A), tangency is in the northeast direction.

Thus we see that B is a proper part of A, and B is much smaller than A, i.e. B is simply 10% of A. Moreover, B is a tangential proper of A and is located in the northeast part of A.

For an example of the need and usefulness of the metrics, see Section V. Later we will discuss how these metrics measure the degree of connectivity (as shown at the end in Fig. 12), and strengthen the topological classification tree, see Fig. 4.

IV. Introduction to Metrics

With relative movement of objects, their spatial relation change over time. Spatial relations can be used to detect such changes. Shift, erosion and dilation are the most common changes that take place in geographical regions with those changes occurring in a continuous pattern. Quantitative metrics are defined to determine the extent of temporal

connectivity of the topological relations between pairs of objects in 3D. The metrics are normalized so that the metric values are constrained to [0, 1]. The metrics also facilitate determining the topological relations between objects. As seen in Fig. 5, a metric can be used to derive the qualitative size of the overlap. The overlap relation, PO (A,B), is symmetric, but the overlap metric IVsIV (A,B) is anti-symmetric. The metric values are also sensitive to the location of the objects in addition to topological connectivity, see Fig. 2 and 3.

For the purposes of precisely defining the metrics herein, we will need two additional topological concepts in addition to the traditional interior, boundary, and exterior parts of an object (or region). The classical crisp boundary of an object A is denoted by A^b, see Fig 6(a); for fuzzy regions, the boundary interior neighborhood (Bin) is denoted by A^{bi}, see. Fig. 6(b), and the boundary exterior neighborhood (Bex) is denoted by A^{be}, see. Fig. 6(c). We give the complete details of these concepts in Section IV.B; an application can selectively use the kind of boundary information available. The exterior and interior boundary neighborhoods even may be combined into one fuzzy/thick boundary which is denoted by A^{bt}, defined as $A^{bt} \equiv A^{bi} \cup A^{be}$, see Fig. 7.
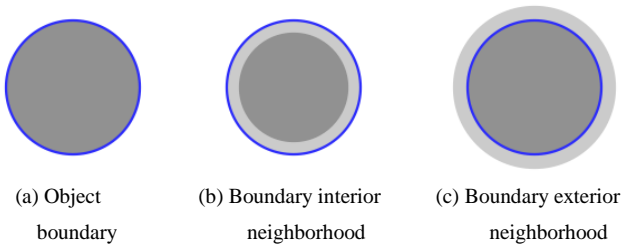

Fig. 6. (a) A 3D object, (b) the interior neighborhood of the boundary of the object, and (c) the exterior neighborhood of the boundary of the object.

Based on these five region parameters, the 9-Intersection table expands to a 25-Intersection table; see Table 3. For 9-intersection, there are $2^9 = 512$ possible combinations out of which only eight are physically realizable; see. Fig. 3. Similarly out of 2^{25} possible combinations derivable from the five region parameters, only a few are physically possible. The possible relations using metrics are as crisp as for bivalent 9-intersection values, see Section V.

A. Volume Considerations

For 3D regions, the volume of a region is a positive quantity, as is the volume enclosed by a cube or a sphere. The classical crisp boundary of a 3D object is 2D, and the volume of a 2D region in a plane or space is zero. Topological relations are predicates that represent the existence of a relation between two objects; metrics measure the strength of the relation or degree of connectivity.

Recall from Section III, the metric IVsIV(A,B) can be used to determine the extent of the overlap $A \cap B$ relative to A,

whereas the metric IVsIV(B,A) determines the extent of overlap $A \cap B$ relative to B. For ease and consistency, the metrics are always normalized with respect to the first parameter, of the metric function. The metric IVsIV(A,B) is *anti-symmetric*. It represents the amount of overlap relative to first argument of the metric.
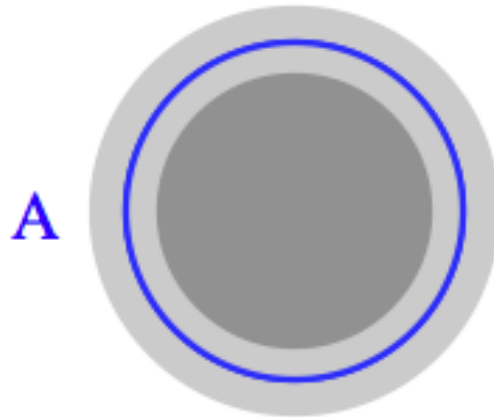

Fig. 7. The space relative to object A is partitioned into 5 parts: interior (dark), A^i; interior-neighborhood (light dark inside the boundary), A^{bi}; boundary, A^b; exterior neighborhood (light dark outside the boundary), A^{be}; and exterior (all outside), A^e

For practical applications, the first parameter is never the exterior volume of an object, because the exterior of a bounded object is unbounded with infinite volume. Also it is observed that since volume(A) = volume($A \cap B^i$) + volume($A \cap B^e$), then IVsIV(A,B^e) = 1- IVsIV(A,B^i).

B. Boundary Considerations

The boundary neighborhood is the region within some small positive radius of the boundary. This is useful for regions with a vague boundary. There are two types of neighborhoods, the boundary interior neighborhood, A^{bi}, and the boundary exterior neighborhood, A^{be}; see Fig. 6. By combining the two, we can create a thick boundary for vague regions; see Fig. 7.

Several metrics are designed for cases where the boundary is vague; these are discussed in Sections IV.F.1 and IV.F.2. To compensate for an accurate crisp boundary, an application-dependent small neighborhood is used to account for the thickness of the boundary. For the 3D object shown in Fig. 6(a), let the boundary interior neighborhood of A^b of some radius r>0, be denoted by A^{bi} or $N_{Ir}(A^b)$, i.e., $A^{bi} \equiv N_{Ir}(A^b)$ (Fig. 6(b)), and let the boundary exterior neighborhood of A^b of some radius r>0, be denoted by A^{be} or $N_{Er}(A^b)$, i.e., $A^{be} \equiv N_{Er}(A^b)$; see Fig. 6(c). The smaller the value of r, the less the ambiguity in the object boundary. Fig. 7 depicts how these terms apply to space partitioning. We denote the qualitative interior neighborhood by $\Delta_I A$ and exterior neighborhood by

$\Delta_E A$ without specific reference to r, as $\Delta_I A \equiv A^{bi}$ and $\Delta_E A \equiv A^{be}$ in the equations that follow in this paper.

Many times in geographical information system (GIS) applications the region's exact boundary is not available. Thus the problem in spatial domains becomes that of how to identify and represent these objects. In such analyses, the external connectedness would be resolved by using a metric that considers the boundary exterior neighborhoods, BexsBex, and examining whether the value BexsBex (A,B) < min (r₁,r₂) (instead of BsB (A,B)=0, which only considers the crisp boundaries) where the objects have boundary exterior r₁- and r₂-neighborhoods for thick boundaries of objects. For definitions of these metrics see Sections IV.F.1 and IV.F.2.

In fact, some applications may need only one r-neighborhood (the combination of r₁-interior and r₂-exterior neighborhood along a vague boundary), while others may need two separate neighborhoods as in [5]. The value of r = min(r₁,r₂) is specified by the application. For some applications (e.g. numerical calculations) it is approximately one percent of the sum of the radii of two spheres. Intuitively, r accounts for the minimum thickness of the boundary for the vague region. In other applications, in order for the metrics to be useful, the radial distance r is chosen to be equal to distance of boundary of object A from the boundary of the object B, i.e. $r_{ab} = dist(A^b, B^b) = min\{dist(x,y): x \in A^b$ and $y \in B^b\}$.

C. Intersections in General

All the metrics and topological relations involve intersections (see Table 3) between a pair of objects. An intersection between a pair of objects may be interior to interior (i.e., 3D), or boundary to boundary (neighborhood), which may be turn out to be 2D, or 1D or even 0D. Metrics measure the quantitative values for topological relations. The intersection of 3D objects may remain 3D, as in the case of PO(A,B). If the intersection such as $A^i \cap B^i$ exists, then we can calculate the volume of the 3D intersection $A^i \cap B^i$, which is practical. But if the boundary is 2D, the volume of the boundary is zero, which does not provide any useful information. The intersection between two 3D objects may also be 3D, 2D, 1D, or even 0D, see Fig. 8. Since intersection is a significant component of topological relations, we can extract useful information from intersections of lower dimensional components also. We can calculate the area of a 2D object (e.g., $A \cap B^b$ may be a 2D surface), and surface area can provide essential information for relations EC(A,B), TPP(A,B), and TPPc(A,B). For example, if two cubes touch face to face, they intersect in a surface; the volume of intersection will be zero, but surface area will be positive, which can still provide a measure of how close the objects are to each other. So we will need metrics that accommodate 2D surface area also. Sometimes intersection is a curve or a line segment, in which case we can analyze the strength of the relation from the length of the segment. Consequently, we

also need metrics that handle the length of edge intersection. For a single point intersection (degenerate line segment), the volume of a point is zero, as are the area and length of a single point, see Fig. 8.
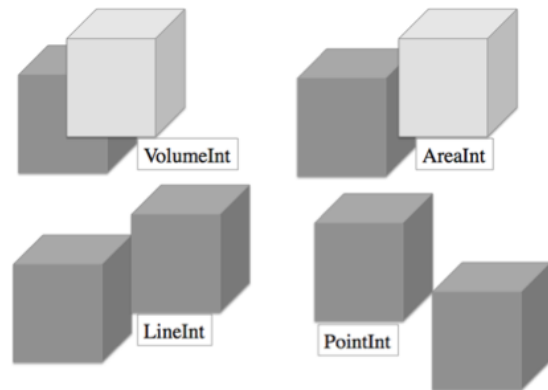

Fig. 8. The intersection two 3D objects can be a point 0D, a line segment 1D, surface area 2D, or volume 3D.

D. Space Partitioning

Usually, each object divides the 3D space into three parts: interior, boundary and exterior. In reality, it is five parts, see Fig. 7. The interior and exterior of the object are 3D parts of space, and the boundary of the object is 2D. The intersection between two 3D objects can be 3D, or a 2D surface, or a 1D curve, or a line segment, or even 0D (i.e., a point). In many geographical applications, regions may not have a well-defined boundary. For example, the shoreline boundary of a lake is not fixed. If the lake is surrounded with a road, the road can serve as the boundary for practical purposes. We need to compensate for the blur in the boundary. Consequently we utilize two additional topological regions: Boundary inner neighborhood (Bin) and Boundary exterior neighborhood (Bex). They can be used to measure how close the objects are from boundary to boundary. The thick boundary becomes a 3D object rather than a 2D object, so the volume calculation for boundary becomes meaningful. For non-intersecting objects, it can be used to account for the distance between them, and for the tangential proper part relation between objects A and B, NTPP(A, B), it can measure how close the inner object A is from the object inner boundary B^{bi}. Thus the terms Boundary interior neighborhood (Bin) and Boundary exterior neighborhood (Bex) for an object A account for the fuzziness, $A^{bt} \equiv A^{bi} \cup A^{be}$, in the boundary description or the thickness of the boundary; see Fig. 6 and Fig. 7.

E. 25-Intersection

To keep full generality available to the end-user, an object space can be defined in terms of five parts: interior, boundary, exterior, boundary interior neighborhood, and boundary exterior neighborhood. As descriptive as we can be for

symbols to be close to natural language: we use Int(A) for A^i the interior of A, Ext(A) for A^e the exterior of A, Bnd(A) for A^b the boundary of A, Bin(A) for A^{bi} the boundary interior neighborhood A, and Bex(A) for A^{be} the boundary exterior neighborhood of the boundary of A. This will lead to a 25-intersection table where the boundary can be a crisp boundary A^b, or a thick boundary $A^{bt} \equiv A^{bi} \cup A^{be}$; see Table 3 for all 25 combinations of intersections.

TABLE 3
25-INTERSECTION TABLE.

	Int	Bnd	Ext	Bin	Bex
Int	$A^i \cap B^i$	$A^i \cap B^b$	$A^i \cap B^e$	$A^i \cap B^{bi}$	$A^i \cap B^{be}$
Bnd	$A^b \cap B^i$	$A^b \cap B^b$	$A^b \cap B^e$	$A^b \cap B^{bi}$	$A^b \cap B^{be}$
Ext	$A^e \cap B^i$	$A^e \cap B^b$	$A^e \cap B^e$	$A^e \cap B^{bi}$	$A^e \cap B^{be}$
Bin	$A^{bi} \cap B^i$	$A^{bi} \cap B^b$	$A^{bi} \cap B^e$	$A^{bi} \cap B^{bi}$	$A^{bi} \cap B^{be}$
Bex	$A^{be} \cap B^i$	$A^{be} \cap B^b$	$A^{be} \cap B^e$	$A^{be} \cap B^{bi}$	$A^{be} \cap B^{be}$

Now Bnd(A) represents the crisp boundary of A, if any, whereas Bin(A) and Bex(A) account for the crisp representations of the vague boundary. There are 2^{25} possible 25-intersection vectors in all. However, all the vectors are not physically realizable. For example, all entries in any row in Table 3 cannot be true simultaneously, and all entries in any column in Table 3 cannot be true simultaneously. Another use of the metrics is to see, for the proper part relation between A and B, PP(A,B), how far the inner object A is from the inner boundary neighborhood of the enclosing object, B^{bi}. A commonly used predicate for determining *weak* connectivity between crisp regions is boundary-boundary intersection, $A^b \cap B^b$. We must be mindful that space now is portioned into five parts instead of three parts. It is clear that A^i, A^e are open sets, and A^b is a closed set. For spatial reasoning, when A^{bi} and A^{be} are used, they are semi-open, semi-closed sets — open towards A^b and closed towards the inside of A^{bi} and the outside of A^{be}.

F. Developing Spatial Metrics

Here we complete the development of the remaining metrics; an application may selectively use the metrics applicable to the problem at hand. If IVsIV(A,B)=0 and BexsIV(A,B)=0, it means that A,B are at least r = d_{ab} (distance between the exterior neighborhood of A and the boundary of B). Conventionally, a 4-intersection [10] (BndBnd, IntBnd, BndInt, IntInt) is sufficient to represent crisp 3D data. Some applications may represent Bex and Bin separately [5], while fuzzy logic applications may need to combine Bex and Bin into Bnd [6]. For all 25 intersections (see Table 3) the metrics are defined by normalizing the intersections. There are 25 possible pairwise intersections to be considered in the metrics. For one pair of objects, consider the eight distinct versions {(A,B), (A,B^e), (A^e,B), (A^e,B^e),

(B,A), (B,A^e), (B^e,A), (B^e,A^e)} as input arguments for which a metric value may be computed. That is, the domain for each metric consists of eight distinct pairs corresponding to each input pair of objects A and B. Since metrics are normalized, some metrics may not be realizable; for example, IVsIV cannot be defined for the combinations {(A^e,B), (A^e,B^e), (B^e,A), (B^e,A^e)} because the corresponding metrics involve infinity. In fact, five of the metrics are impossible (not realizable); see Table 4. Here we will identify the possible (reasonable) 20 metrics.

TABLE 4
Complete list of metrics corresponding to 25 intersections in table 1. 20 metrics are viable and 5 metrics are not possible.

Possible	Impossible
IVsIV, IVsEV	EVsIV, EVsEV
BinsIV, BinsEV, IVsBin	EVsBin
BexsIV, BexsEV, IVsBex	EVsBex
BinsBin, BinsBex,	
BexsBin, BexBex	
BsIV, BsEV, IVsB	EVsB
BsBin,BsBex,	
BinsB, BexsB	
BsB	

Since the metrics are anti-symmetric, the converse metrics can be defined by switching arguments A and B (e.g., the converse of IVsIV(A,B) is IVsIV(B,A)). To make the list of metrics exhaustive, we can append suffix c to the name to indicate the converse metric when needed. Table 2 lists directly possible and impossible metrics, which are further developed in detail in Sections IV.F.1 and IV.F.2.

Next we will define 20 viable metrics and show their connection with the RCC8 topological relations and size relations on 3D objects only. First we look at the two metrics together: IVsIV(A,B) and IVsEV(A,B) which measure how much space one object shares with the other object. We have already defined interior volume split by interior volume, IVsIV(A,B), earlier in the motivation discussion, Section III.

F.1. Volume Metrics: segmentation of interior of reference object by the interior and exterior of target object

Recall, interior volume splitting (IVsIV) measures the scaled (normalized) part of reference object that is split by the interior of the target object. It calculates how much of A is part of B. The crisp boundary of a 3D object is 2D. Here boundary does not matter, as the volume of the boundary is zero. Exterior volume splitting (IVsEV) describes the proportion of reference object's interior that is split by the other target object's exterior. The exterior volume splitting metric (IVsEV) is defined by

$$IVsEV(A, B) = \frac{volume(A^i \subsetneqq B^e)}{volume(A^i)}$$

It measures how much A is away from B. Again, boundary does not matter. Observe that volume(A) = volume(A∩B) + volume(A∩B^e), and hence IVsEV(A,B) = 1 - IVsIV(A,B). The metric value is between 0 and 1, inclusive. If the metric value IVsIV(A,B) = 0, the objects are disjoint or externally connected. If the metric value IVsIV(A,B) > 0, then this value indicates two things. First, A^i∩B^i ≠ ∅. Usually, the truth value of A^i∩B^i is established by considering the intersection of the boundaries of two objects (extensive computation takes place because the objects are represented with boundary information only). Here if the metric value IVsIV(A,B) > 0, we can quickly determine the truth value of A^i∩B^i. Secondly, the actual value of the metric IVsIV(A,B) measures what relative portion of object A is common with object B; the larger the value of the metric, the larger the commonality and conversely. Let

$$x = \frac{volume(A^i \subsetneqq B^i)}{volume(A^i)} * 100 \quad y = \frac{volume(B^i \subsetneqq A^i)}{volume(B^i)} * 100$$

This can directly answer queries such as object A has x percent in common with B, whereas object B has y percent in common with A, provided A and B intersect. If x=y=0, then the objects are either externally connected or disjoint, but this metric alone does not tell how far apart they are. In order to determine that, we simply compute the distance between the boundaries to differentiate between DC and EC. The metric does embody knowledge about which object is larger. If x<y, then object A is bigger than object B.

F.2. Boundary Metrics: segmentation of thick boundary of reference object by the interior and thick boundary of target object

Recall, for the 3D object shown in Fig. 6(a), A^{be} is the boundary exterior neighborhood of A^b with some radius (Fig. 6(b)), and A^{bi} is the boundary interior neighborhood of A^b with some radius, see Fig. 6(c). The value of the radius is application-dependent. We use the qualitative interior and exterior neighborhood without specific reference to r, as $\Delta_I A \equiv A^{bi}$ and $\Delta_E A \equiv A^{be}$ in the following equations.

Considering the interior neighborhood of the reference object, we define the closeness to interior volume (BinsIV) as follows:

$$BinsIV(A, B) = \frac{volume(D_I A \cap B^i)}{volume(D_I A)}$$

This metric contributes to the overall degree of relations of PO, EQ, TPP, and TPPc.

Similarly, we can consider the exterior neighborhood of an object, and can define a metric for exterior volume closeness (*BexsIV*) by replacing $\Delta_I A$ by $\Delta_E A$.

$$BexsIV(A, B) = \frac{volume(D_E A \cap B^i)}{volume(D_E A)}$$

This metric is a measure of how much of the exterior neighborhood of A^b is aligned with the interior of B. This metric is useful for analyzing the degree of relations of PO, EQ, TPP, and TPPc.

Similarly the metrics for the exterior of B are defined for completeness as follows:

BeinsEV(A,B) becomes

$$BinsEV(A, B) = \frac{volume(D_I A \cap B^e)}{volume(D_I A)}$$

BexsEV(A,B) is defined by replacing $\Delta_I A$ by $\Delta_E A$.

$$BexsEV(A, B) = \frac{volume(D_E A \cap B^e)}{volume(D_E A)}$$

Boundary-boundary intersection is an integral predicate for distinguishing RCC8 relations.

Similarly, for quantitative metrics, it can be important to consider how much of the inside and outside of the boundary neighborhood of reference object is shared with the boundary neighborhood of the other object.

BinsBin(A,B) is designed to measure how much of the Interior Neighborhood of A is split by the Interior Neighborhood of B. This metric is useful for fuzzy regions with fuzzy interior boundary.

$$BinsBin(A, B) = \frac{volume(D_I A \cap D_I B)}{volume(D_I A)}$$

BexsBin(A,B) is designed to measure how much of the Exterior Neighborhood of A is split by the Interior Neighborhood of B.

$$BexsBin(A, B) = \frac{volume(D_E A \cap D_I B)}{volume(D_E A)}$$

This metric is useful when the region is vague around both sides of the boundary.

BinsBex(A,B) is defined by replacing $\Delta_I A$ by $\Delta_E A$ and is designed to measure how much of the interior neighborhood of A is split by the exterior neighborhood of B, It is useful to analyze topological relations DC and EC.

$$BinsBex(A, B) = \frac{volume(D_I A \cap D_E B)}{volume(D_I A)}$$

BexsBex(A,B) is designed to measure how much of the exterior neighborhood of A is split by the exterior neighborhood of B. This metric is useful for fuzzy regions, if BexsBex(A,B) = 0 then we can narrow down the candidates of possible relations between A and B to DC, NTPP, and NTPPc.

$$BexsBex(A, B) = \frac{volume(D_E A \cap D_E B)}{volume(D_E A)}$$

F.3. Boundary Metrics: segmentation of crisp boundary of reference object by the interior and the boundary of target object

We define several splitting metrics to specifically examine the proportion of the boundary of the reference object that is split by the volume, boundary neighborhoods, and boundary of the target object; we denote these metrics accordingly for boundary splitting. It should be noted that there are five versions of the equations for this metric. First, the boundary may be the thick boundary composite neighborhood (interior and exterior), in which case it is a volume, see Fig. 8. If the boundary is a simple boundary, it is a 2D area. Therefore, for numerator calculations, we will be calculating $A^b \cap B$ as either a volume or an area. It also is possible that $A^b \cap B$ is an edge (a curve or a line segment). For example, for two cubes, a cube edge may intersect the face of the cube as a line segment or an edge of another cube in a line segment, or even as a single point (i.e., a degenerate line segment). If $A^b \cap B$ is an edge, we calculate edge length. For the denominator, $volume(A^b)$ and $area(A^b)$ are self-evident depending on whether we have a thick or simple boundary. However, $length(A^b)$ calls for an explanation. In the numerator, when $length(A^b \cap B)$ is applicable, then this intersection is part of an edge in A^b; $length(A^b)$ is computed as the length of the enclosing edge. These metrics are defined and described below. The converses of the metrics can be derived similarly.

BsIV(A,B) measures the Boundary of A split by the Interior Volume of B.

$$BsIV(A, B) = \frac{volume(A^b \subsetneq B^i)}{volume(A^b)} \ or \ \frac{area(A^b \subsetneq B^i)}{area(A^b)} \ or \ \frac{length(A^b \subsetneq B^i)}{length(A^b)}$$

BsEV(A,B) is defined by replacing B^i by B^e and measures the Boundary of A split by the Exterior Volume of B.

$$BsEV(A, B) = \frac{volume(A^b \subsetneq B^e)}{volume(A^b)} \ or \ \frac{area(A^b \subsetneq B^e)}{area(A^b)} \ or \ \frac{length(A^b \subsetneq B^e)}{length(A^b)}$$

BsBin(A,B) is defined by replacing B^i by $\Delta_I(B)$ and measures the Boundary of A split by the Interior Neighborhood of B.

$$BsBin(A, B) = \frac{volume(A^b \cap D_I B)}{volume(A^b)} \ or \ \frac{area(A^b \cap D_I B)}{area(A^b)} \ or \ \frac{length(A^b \cap D_I B)}{length(A^b)}$$

BsBex(A,B) is defined by replacing B^i by $\Delta_E B$ and measures the Boundary of A split by the Exterior Neighborhood of B.

$$BsBex(A, B) = \frac{volume(A^b \cap D_E B)}{volume(A^b)} \ or \ \frac{area(A^b \cap D_E B)}{area(A^b)} \ or \ \frac{length(A^b \cap D_E B)}{length(A^b)}$$

BsB(A,B) is defined by replacing B^i by B^b and measures the Boundary of A split by the Boundary of B.

$$BsB(A, B) = \frac{volume(A^b \subsetneq B^b)}{volume(A^b)} \ or \ \frac{area(A^b \subsetneq B^b)}{area(A^b)} \ or \ \frac{length(A^b \subsetneq B^b)}{length(A^b)}$$

This metric is again directly applicable to computing $A^b \cap B^b$ which is used to distinguish many of the RCC8 relations. This subsequently allows us to narrow down the candidates of possible relations between A and B to DC, NTPP, and NTPPc.

For crisp regions, we have an interior, boundary, and exterior. For vague regions, we have boundary interior and exterior neighborhoods. The smaller the radius for boundary neighborhoods, the smaller the ambiguity in the object boundary. For consistency, we can combine the interior and exterior neighborhoods into one, which we call a thick boundary. For a thick boundary, the object has three disjoint crisp parts: the interior, the thick boundary, and the exterior. Now we can reason with these parts similar to how we use crisp regions for determining the spatial relations.

F.4. Anatomy of Metrics

The following discussion is applicable when there is positive distance between the boundaries, $A^b \cap B^b = \varnothing$, irrespective of $A^i \cap B^i \neq \varnothing$, see Fig. 9, or $A^i \cap B^i = \varnothing$, see Fig. 10. In Fig. 7, we see that the space relative to an object A can be partitioned into 5 parts: interior (dark), interior-neighborhood (light dark inside the boundary), boundary, exterior neighborhood (light dark), and exterior (all outside). In Fig 9, B shares only A^i. In Fig 10, B shares only A^e. In Fig 11, B shares all five parts, A^i, A^{bi}, A^b, A^{be}, and A^e.

Fig. 9. The space relative to object A is partitioned into 5 parts: interior (dark), interior-neighborhood (light dark inside the boundary), boundary, exterior neighborhood (light dark), and exterior (all outside). A and B have disjoint boundaries, B is inside A, but away from the boundary by r = d_{ab}.
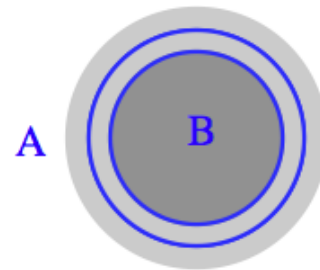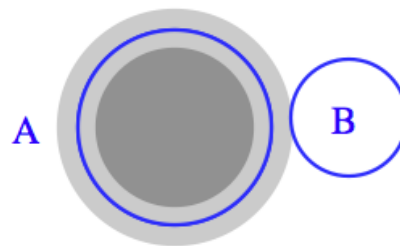
Fig. 10. The space relative to object A is partitioned into 5 parts: interior (dark), interior-neighborhood (light dark inside the boundary), boundary, exterior neighborhood (light dark), and exterior (all outside). A and B have disjoint boundaries, B is outside A ,but away from the boundary by r = d$_{ab}$.
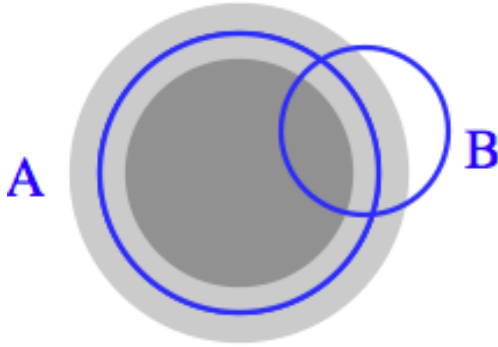
quality of the topological relations. Finally the metric based tree is generated that captures the topological and metrical relations in a single hybrid tree.

Fig. 11. An example where B shares A^i, $\Delta_I A \equiv A^{bi}$, A^b, $\Delta_E A \equiv A^{be}$, and A^e .

Fig. 12. Tree for the metrics required for classification and the contribution (0/+) of the respective metrics to the overall quality of classification.

V. Spatial Relations: Topological and Metrical

If the regions are crisp, we can use the 9-intersection model for determining connectivity relations for 2D connectivity knowledge [2],[10], and for relative size information we use the 3D metrics from Section IV. The relative size of the objects and boundary is obtained by using volume metrics IVsIV, IVsB, and boundary-related BsB metrics. Metrics measure the degree of connectivity; for example, for the proper overlap relation PO(A,B), IVsIV metric helps to determine the relative extent of overlap of each object. In Section IV we discussed which metrics are specific to each of the connectivity relations. If one or both regions have vague boundaries, we can use metrics to create a thick boundary, A^{bt} $\equiv$ A$^{bi}\cup$A^{be}, by using the interior and exterior neighborhoods. Again we have, crisp interior A^i, exterior A^e, and thick boundary A^{bt}.

By using the 9-Intersection model on A^i, A^e, and A^{bt}, we can derive the connectivity, degree of connectivity, and relative size information for vague regions. Other applications such as natural language and topological association [5] can use appropriate combinations of these topological parts. Fig. 5 provides a visual summary of: (1) what intersections are required to classify each topological relation, and (2) the contribution (T / F) indication if the intersection is non-empty or empty. This tree can be used to classify crisp topological relations. Fig. 12 provides a visual summary of: (1) what metrics are required to classify each topological relation, and (2) the contribution (0 / +) each metric has with regards to the overall quality of the relation. This tree can be used to classify crisp relations. Similarly, a tree could be generated for vague regions with appropriate metrics from the set of 20 metrics.

The general frame work for combining topology and metrics given in Fig. 12 can be explained diagrammatically in Fig. 13 as follows. First of all topological relations are determined. Then metrics are developed to strengthen the

Fig. 13. The structure of topological relations system and metrics relations leading to hybrid system.

VI. Conclusions and Future Directions

Herein we presented an exhaustive set of metrics for use with both crisp and vague regions, and showed how each metric is linked to RCC8 relations for 3D objects. Our metrics are systematically defined and are more expressive (consistent with natural language) than previously published efforts. Further, we showed the association between our metrics and the topology and size of objects.

This work should be useful for a variety of applications dealing with automated spatial reasoning in 3D. In the future, we plan to use these metrics to associate natural language terminology with 3D region connection calculus including occlusion considerations. Also we will explore the applications of these metrics between heterogeneous dimension objects, O$_m \in$R^m and O$_n \in$R^n for m, n $\in\{1,2,3\}$.

REFERENCES

[1] D. A. Randell, Z. Cui, and A. G. Cohn, "A Spatial Logic Based on Regions and Connection." *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, Cambridge, MA, pp. 156–176, 1992.

[2] M. J. Egenhofer and J. Herring, "Categorizing binary topological relations between regions, lines, and points in geographic databases." *NCGIA Technical Reports* 91-7, 1991.

[3] S. Schockaert, M. De Cock, and E. E. Kerre, "Spatial reasoning in a fuzzy region connection calculus," *Artificial Intelligence*, 173(2), pp. 258–298, 2009.

[4] D. Hernandez, E. C. Clementini, and P. Di Felice. "Qualitative Distances," *2nd Int. Conf. on Spatial Information Theory*, LNCS 988, Springer-Verlag, pp. 45–57, 1995.

[5] R. Shariff, M. J. Egenhofer, and D. M. Mark, "Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms." *International Journal of Geographical Information Science*, 12, pp. 215–246, 1998.

[6] M. Schneider, "Vague topological predicates for crisp regions through metric refinements," P. F. Fisher (Eds.) *Developments in Spatial Data Handling, 11th International Symposium on Spatial Data Handling*, 1

[7] J. Leopold, C. Sabharwal, and K. Ward, "Spatial Relations Between 3D Objects: The Association Between Natural Language, Topology, and Metrics". *Proceedings of the Twentieth International Conference on Distributed Multimedia Systems Pittsburgh DMS 2014*, August 27–29, pp. 241–249, 2014.

[8] J. Albath, J. L. Leopold, C. L. Sabharwal, and A. M. Maglia, " RCC-3D: Qualitative Spatial Reasoning in 3D," *Proceedings of the 23rd International Conference on Computer Applications in Industry and Engineering*, pp. 74–79, 2010.

[9] C. L. Sabharwal and J. L. Leopold. "Cardinal Direction Relations in Qualitative Spatial Reasoning," *International Journal of Computer Science & Information Technology (IJCSIT)* 6(1), pp. 1–13, 2014. DOI 10.5121/ijcsit.2014.6101

[10] C. Sabharwal and J. Leopold, "Reducing 9-Intersection to 4-Intersection for Identifying Relations in Region Connection Calculus," *Proceedings of the 24th International Conference on Computer Applications in Industry and Engineering (CAINE 2011)*, pp. 118–123, 2011.

[11] C. Sabharwal and J. Leopold, "Modeling Cardinal Direction Relations in 3D for Qualitative Spatial Reasoning," *Proceedings of Mining Intelligence and Knowledge Exploration, MIKE 2014,* in *Lecture Notes in Artificial Intelligence* 8891, pp. 199–214, 2014.

Springer, pp. 149–162, 2005.

EMiner: A Tool for Selecting Classification Algorithms and Optimal Parameters

Rayrone Zirtany Nunes Marques, Luciano Reis Coutinho, Tiago Bonini Borchartt, Samyr Béliche Vale, and Francisco José da Silva e Silva

Abstract—In this paper, Genetic Algorithm (GA) is used to search for combinations of learning algorithms and associated parameters with maximum accuracy. An important feature of the approach is that the GA initial population is formed by using parameter values gathered from ExpDB (a public database of data mining experiments). The proposed approach was implemented in a tool called EMiner, built on top of a grid based software infrastructure for developing collaborative applications in medicine and healthcare domains (ECADeG project). Experiments on 16 datasets from the UCI repository were performed. The results obtained have shown that the strategy of combining the data from ExpDB via GA is effective in finding classification models with good accuracy.

Index Terms—Data mining, medicine and healthcare, algorithm selection, parameter optimization, genetic algorithms.

I. INTRODUCTION

MEDICINE and healthcare are important application areas for data mining technology [1], [2], [3], [4]. Given its nature—viz. to extract information and knowledge from large amounts of raw data—data mining has much to contribute in helping physicians and other researches to analyse medical databases for improving their understanding of diseases, diagnostic skills, treatment procedures, etc. In this sense, one of the most popular and effective data mining techniques has been the use of machine learning algorithms to infer classification models from training datasets [5].

For instance, consider a dataset $\mathcal{D}$ of patient records $\{r_1, \ldots, r_m\}$. Each record r_i being a tuple $(x_1, \ldots, x_n, y)$ of attributes describing the medical condition of a patient. The attributes x_i are data items detailing symptoms, disease factors, and/or results of clinical tests collected from the patient. The attribute y summarizes the patient medical condition in terms of some disease type. In this setting, a completely correct classification model consists in a function $f_{\mathcal{D}}$ that maps the attributes x_i to y such that $f_{\mathcal{D}}(x_1, \ldots, x_n) = y$ iff $(x_1, \ldots, x_n, y) \in \mathcal{D}$. In practice—mainly due to uncertainty, or incomplete and inconsistent data—a completely correct classification model $f_{\mathcal{D}}$ is an impossible idealization. Instead, what is feasible is an approximate classification model $\hat{f}_{\mathcal{D}}$ that tries to minimize the classification error

against the current available records and the records added to $\mathcal{D}$ in the future. The approximate classification model $\hat{f}_{\mathcal{D}}$ then represents the knowledge extracted from $\mathcal{D}$ relating antecedents to consequents, and can be used researchers and physicians to explain, diagnose and treat diseases.

The research in machine learning algorithms has produced many approaches for creating good classifiers [5]. On the one hand, such an abundance provides the researcher with proven techniques to mine accurate and general classification models out of raw data. On the other hand, the researcher is faced with the problem of choosing among tenths, or even hundreds, of sophisticated mining algorithms, each one having a set of (hyper)parameters that need to be adjust for the proper functioning of the algorithm [6], [7]. Considering that the medical researcher is not necessarily an expert in data mining, the right choice of algorithm and associated parameter values becomes a challenging task prone to error, or leading to sub-optimal selection based on intuitive appeal and/or reputation. As noted in [6], this suggests an important challenge for machine learning: given a dataset $\mathcal{D}$, how to automatically and simultaneously choose a learning algorithm and set its (hyper)parameters to optimize accuracy and generality; a challenge that can be named *Combined Algorithm Selection and Hyperparameter* optimization problem (or CASH problem).

In this paper, we present an approach to the CASH problem originally developed in the context of ECADeG Project[1] [8]. We put forward a methodology and a tool called EMiner. Our methodology starts with a given dataset $\mathcal{D}$ and a predefined set of available classification algorithms. In general, this set of algorithms is arbitrary. The aim is to select the best classifier, in terms of accuracy/generality, obtainable from the available algorithms with adequate parameter settings. For this, we iterate over all classification algorithms applying a Genetic Algorithm (GA) [9]. This GA is used to evolve a population of candidate classifiers, each candidate classifier consisting in the current algorithm instantiated with a possibly different set of parameter values. Traditionally, the initial population of GA is random. In our case, we propose to form the initial population by using parameter values retrieved from a public database that collects the results of data mining experiments, called ExpDB [10]. With this, i.e. by combining GA with

The authors are with the Universidade Federal do Maranhão, Programa de Pós-Graduação em Ciência da Computação, Av. dos Portugueses 1966, Bacanga, São Luís, MA, Brazil (e-mail: rayronezirtany@gmail.com, {lrc, tiagobonini, samyr, fssilva}@deinf.ufma.br).

[1]*Enabling Collaborative Applications for Desktop Grids* is a project whose aim is to provide a grid based software infrastructure to support the development of collaborative applications in medicine and healthcare domains.

parameters values coming from the experimental data available in ExpDB, we show that we can automatically reach at good classifiers outperforming default parameters settings, with less experiments than is necessary if we conduct a non-guided search over the state of possible algorithms and associated parameters.

The rest of the paper is organized as follows. Section 2 discuss related work. In Section 3 the technical concepts and specific learning algorithms used in this work are presented. Section 4 describes the proposed methodology and the EMiner tool. In Section 5 we present some of the experiments we have conducted to assess and validate our approach to the CASH problem. Finally, in Section 6 we draw our final conclusions and comment on future work.

II. RELATED WORK

In this section we comment on related work sorted in two main themes: approaches to CASH optimization problem, and approaches employing GA to optimize parameter values.

Dealing with the CASH optimization problem, we found [6] and [7]. Thornton et al. [6] use recent innovations in Bayesian optimization to find the best parameter values for a classification algorithm. They formally define the problem of algorithm selection and parameter optimization in classification problems, naming this type of problem by the acronym CASH (*Combined Algorithm Selection Hyperparameter*) optimization problem. They demonstrate that the learning algorithm itself can be considered a parameter and that Bayesian optimization obtains high quality results in a relatively reasonable time. Finally, to fully automate the approach, they built a tool called Auto-Weka, extending/adapting the traditional tool Weka.

Leite et al. [7] propose a new technique for selecting classification algorithms called *active testing*. This technique selects the most useful algorithm using cross-validation testing on a tournament where, in each round of selection and test, it is chosen the algorithm that outperformed the algorithms that won the previous round. The most promising algorithm is then chosen based on the tournament history over similar datasets. To evaluate the approach, 292 combinations of algorithms-parameters are used to analyze 76 datasets. The results showed that active testing quickly come up with a combination algorithm-parameter whose performance is close to optimal.

Among the approaches employing GA to optmize parameter values, we found the work by Samadzadegan et al. [11] using GA to select an optimal kernel and respective parameters in the learning of *Support Vector Machines* (SVM). The results showed that the proposed method outperformed in terms of accuracy when compared to traditional *Grid Search* (brute force). Another work is [12] that also use GA to optimize SVM kernel and parameters. The author explains that the traditional combination causes a problem of premature convergence that limits the accuracy of the SVM obtained. Then he suggests new genetic operators (called IO-GA) specially designed to

optimize the SVM kernel/parameters. The results showed an increase in classification accuracy over the traditional operators. Several other work in the literature address the issue of parameter optimization for classification algorithms by using GA [13], [14], [2], [4]. In general, they do not take into account previous experiments to initially guide their search for better results, and are limited to the parameter optimization of a single classification algorithm.

III. CONCEPTUAL BACKGROUND

In this section we further detail the CASH optimization problem, the specific classification algorithms we have used in our experiments, the notion of genetic algorithms (GA) and the specific GA we have applied in our approach. We end the section with a description of the ExpDB project.

A. CASH Optimization Problem

To better characterize the CASH optimization problem, let us first consider the restricted problem of *parameter optimization* for a fixed machine learning algorithm $\mathcal{A}$ when this is applied to infer a classification model $\hat{f}_\mathcal{D}$ from a given dataset $\mathcal{D}$. Suppose that $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the list of (hyper)parameters of $\mathcal{A}$, where each λ_i is a variable defined over some domain $\{\alpha_i\}$ of values (discrete or continuous). Given $\mathcal{D} = \{(x_1, \ldots, x_n, y)\}$, the problem of optimizing the parameters of $\mathcal{A}$ consists in finding assignments $\lambda_i := \alpha_i^*$, $1 \le i \le n$, leading to maximum accuracy/generality of $\mathcal{A}$ over $\mathcal{D}$. Thus, denoting the set of assignments $\{\lambda_i := \alpha_i\}_{1 \le i \le n}$ by α and the measure of accuracy/generality of $\mathcal{A}$ (with parameter values α) over $\mathcal{D}$ by $\mathcal{M}_{ag}[\mathcal{A}(\alpha), \mathcal{D}]$, the parameter optimization problem can be formulated compactly as the search for $\alpha^* \in \mathbf{argmax}_\alpha \mathcal{M}_{ag}[\mathcal{A}(\alpha), \mathcal{D}]$.

Traditionally a method of measuring accuracy/generality of a given combination of algorithm and parameter values is to perform a cross-validation test with k-folds [15]. In this method, the data set $\mathcal{D}$ is randomly divided into k mutually exclusive subsets $B_1, B_2, \ldots, B_k$ of approximately equal size called *folds*. Then, the classification algorithm $\mathcal{A}$ with appropriate parameters values α is executed k times; each time $t \in \{1, 2, ..., k\}$ the algorithm $\mathcal{A}(\alpha)$ is trained using $\mathcal{D} - B_t$ and tested against B_t. At the end, the measure of accuracy/generality is obtained as follows

$$\mathcal{M}_{ag}[\mathcal{A}(\alpha), \mathcal{D}] = \frac{1}{k} \sum_{t=1}^{k} \frac{|B_t^c|}{|B_t|} \times 100, \qquad (1)$$

where $|B_t|$ is total number of records in B_t, and $|B_t^c|$ is number of records correctly classified by the model $\hat{f}_{\mathcal{D}-B_t}$ obtained at time t; a record is correctly classified when $\hat{f}_{\mathcal{D}-B_t}(x_1, \ldots, x_n) = y$ implies $(x_1, \ldots, x_n, y) \in B_t$.

In this work we use the method of k-folds cross-validation to estimate the accuracy/generality of the algorithms in our experiments. We adopt $k = 10$ because it is a value that provides good estimates as established by several research papers [15], [7], [6], [10].

Finally, from this formulation of the restricted problem of parameter optimization, the more general CASH optimization problem consists in the search for both algorithm $\mathcal{A}^*$ and respective parameter values α^* that lead to maximum accuracy/generality in $\mathcal{D}$; i.e., find $\mathcal{A}^*(\alpha^*) \in$ **argmax** $\mathcal{M}_{ag}[\mathcal{A}(\alpha), \mathcal{D}]$.
$$\underset{\mathcal{A}(\alpha)}{}$$

B. Classification Algorithms

In medical application of data mining two crucial feature are the *transparency* and the *interpretability* of the models obtained [1]. By this we mean the possibility of human analysis and validation of the models generated by a given learning technique. This is crucial because the physician or researcher using the models in general wants to be able to explain and justify its decisions when being guided by some classification model.

Taking into account these desirable features, the classification algorithms we have used in our work are divided into two categories: algorithms based on *decision trees induction* and algorithms based on *classification rule induction*. The choice of these categories is justified by the transparency and interpretability of the classifiers obtained with them, once both categories are based on human comprehensible representations, which facilitates their analysis and validation. Among the algorithms for decision trees induction we have used J48 [16], RandomTree [3] and REPTree [5]. For classification rule induction JRip [17], PART [5], and Ridor [5]. In Table I we summarize the (hyper)parameters of used classification algorithms.

C. Genetic Algorithms

Genetic Algorithms (GA) are optimization algorithms based on principles of natural selection and genetics [9]. The main idea is that in a population, the individuals with more favorable genetic traits are more likely to survive, reproduce and give birth to an increasingly fit offspring, while less fit individuals tend to become extinct. In GA, each individual in the population (in the form of chromosomes) represents a candidate solution for a given problem. To find better individuals, the GA uses a random search strategy, favoring points of high fitness; i.e., points at which the function to be minimized or maximized assumes relatively low or high values. The search process is iterative, where each iteration is called a generation. In each generation, the mechanisms of selection and reproduction are applied to individuals of the population. Through selection, it is determined which individuals are apt to reproduce, generating a certain number of offspring. The probability of an individual being selected is proportional to its fitness; i.e., the higher fitness of an individual the greater chance of it being selected.

Depending on the problem, the chromosomes can be further divided in logical sections called genes. In this paper, chromosomes are used to represent the set of parameters

TABLE I
CLASSIFICATION ALGORITHMS AND RESPECTIVE (HYPER)PARAMETERS.

Algorithm	Parameter	Type	Def. Value
J48	Use binary splits	boolean	false
	Confidence threshold for pruning	double	0.25
	Cleanup after the tree has been built	boolean	true
	Subtree raising to be performed	boolean	true
	Minimum number of instances per leaf	integer	2
	Number of folds for reduced error pruning	integer	3
	Use reduced error pruning	boolean	false
	Random number seed for reduced-error pruning	integer	1
	Use Laplace correction	boolean	false
	Unpruned tree	boolean	false
R. Tree	Minimum number of instances for leaf	double	1.0
	The number of attributes considered for a split	integer	0
	The random seed to use	integer	1
REPTree	Maximum tree depth	integer	-1
	Minimum class variance proportion of train	double	0.001
	Minimum number of instances	integer	2
	Switch off pruning	boolean	false
	Number of folds for reduced error pruning	integer	3
	Seed for random data shuffling	integer	1
JRip	Switch off checking error rate $>= 0.5$	boolean	true
	Min weights of instances within a split	double	2.0
	Switch off pruning	boolean	false
	Number of folds for reduced error pruning	integer	3
	Number of optimization runs	integer	2
	Seed for random data shuffling	integer	1
PART	Use binary splits on nominal attributes	boolean	false
	Confidence factor for rule pruning	double	0.25
	Generate unpruned decision list	boolean	false
	Minimum number of instances per leaf rule	integer	2
	Number of folds for reduced error pruning	integer	3
	Use reduced error pruning	boolean	false
	Random number seed for reduced-error pruning	integer	1
Ridor	Minimal weights of instances within a split	double	2.0
	Number of folds for reduced error pruning	integer	3
	Number of randomization shuffles	integer	10
	Use error rate of all data	boolean	false
	Use majority class as default class in each step	boolean	false

values peculiar to a given classification algorithm. Thus, each gene determines the value of a specific parameter can take. Once we are using 6 different classification algorithms, each one with different parameters, we have defined 6 types of chromosomes in order to meet the particularities of each one of them. For instance, for the J48 algorithm, which has 10 parameters in total, we have defined a chromosome divided in 10 genes, each gene with a given number of bits appropriate to the parameter types.

The evaluation of the individuals is done by means of a fitness function. This is intended to measure how apt is a given individual (solution) from the current population. The aptitude of an individual influences in its survival probability. In this work, the fitness function used is the mean accuracy/generality of the classifier (algorithm + parameter values) after the process of k-fold cross-validation given by Eq. (1).

Via crossover, the information contained in the genes of two existing individuals are combined in order to create new individuals. This operator is the main mechanism of the GA to explore the search space, by creating new points obtained by the exchange of genetic information between individuals. In this work we use a crossing rate corresponding to 70% of the population and the crossover point is chosen randomly.

During the crossing operation, can be born individuals who

violate constraints of dependence between parameters imposed by the classification algorithm. For example, the parameter *number of folds for reduced error pruning* are relevant if the parameter *use reduced error pruning* is true; i.e., first parameter (son) depends on second (parent). To address this problem, the chromosomes was designed to correct such violations using value of the parent parameter. In the previous example, if the parameter *use reduced error pruning* was true, then the parameter value *number of folds for reduced error pruning* would be disregarded by assigning a null value or zero depending on the domain parameter.

The mutation is a genetic operator which serves to introduce random variations in the population, or even to recover some features lost in operations such as crossover. Furthermore, mutation is used to prevent premature convergence to local optima, by sampling new points in the search space. This operator is applied with a given probability, after the crossover operator. In this work, the mutation rate varies according to the size X of the population according to the ratio $1/X$; i.e., on average 1 in X individuals is mutated [18].

Similarly crossover operator, the mutation operator can also generate individuals that violate some restrictions of parameters values of classification algorithm, therefore the same strategy adopted with crossover operator was used for handling this problem.

D. ExpDB

ExpDB[2] is a database designed to store a large number of data mining experiments, containing detailed information on the datasets, the algorithms and the parameter settings used, as well as the evaluation procedures and the obtained results [10]. Currently this database contains about $600,000$ experiments on 130 datasets, in varied subjects such as healthcare, finances, education, etc. The main idea is to facilitate large-scale experimentation, guarantee repeatability and reusability of experiments, and help to clarify the conditions under which certain results are valid. We use ExpDB to obtain already tested initial values for the parameters of the classification algorithms. These collected values will form the initial population to be improved via the application of GA.

IV. EMINER TOOL

Based on GA and ExpDB, we have proposed a methodology for handling the CASH optimization problem. This methodology is divided into three phases: *initial values definition*, *parameter optimization* and *algorithm selection*. Figure 1 ilustrates the main phases and steps of the proposed methodology.

In the first phase we have the definition of classification algorithms to be used and the extraction of the initial values for their parameters from the base ExpDB. The extraction occurs by first getting raw parameter sets from ExpDB. Then a

[2]http://expdb.cs.kuleuven.be/

filtering is performed to remove the repetead sets. Finally, the distinct parameter sets are encoded in different chromosomes to generate the initial population of the GA.

The second phase is concerned with the optimization of the classification algorithm parameters by using GA. Here it is performed the cross-validation (Eq. (1)) of each individual (representing a classifier for a given dataset) in the population. Each cross-validation uses k folds to test an individual and calculate their mean accuracy (fitness). Then the genetic operators (selection, crossover and mutation) are applied to the current generation to give rise to a new generation. After a given number of generations the best classifier is selected.

Finally, the last phase deals with selecting the classification algorithm for a given dataset. So, it repeats the previous phases for each classification algorithm in order to built a rank of used algorithms and parameter values.

A first implementation of our approach to the CASH problem was performed in the context of the ECADeG Project. It is a prototype tool called EMiner (ECADeG Miner), available for download at the address `http://lsd.ufma.br/joomla/index.php/projetos/8-ecadeg`.

V. EXPERIMENTAL EVALUATION

A. Computing Environment

The genetic algorithm implemented in EMiner was written in JAVA using the framework JGAP [18]. JGAP is a Genetic Algorithms and Genetic Programming component provided as a Java framework. It provides basic genetic mechanisms that can be easily used to apply evolutionary principles to problem solutions. Regarding the implementation of the decision trees and classification rules algorithms, and the k-folds cross-validation process, we have used the WEKA [19] API. WEKA is a collection of machine learning algorithms for data mining tasks that contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. All of out experiments were run on a InteGrade desktop grid formed by Linux machines with Intel Core I5, running at 3GHz and 4GB of RAM.

B. Datasets

We performed several experiments using datasets from the UCI database [20]. We have chosen datasets that have been frequently used as benchmarks for evaluating classification methods in the literature. These datasets consist of discrete and continuous features. Table II summarizes the number of classes, number of discrete and continuous attributes found in the datasets.

C. Experiments

Initially—in order to get a baseline for comparison—we conducted experiments to measure the performance of the 6 classification algorithms of Section III-B, in terms of mean accuracy (Eq. (1)), when each one of these is configured with
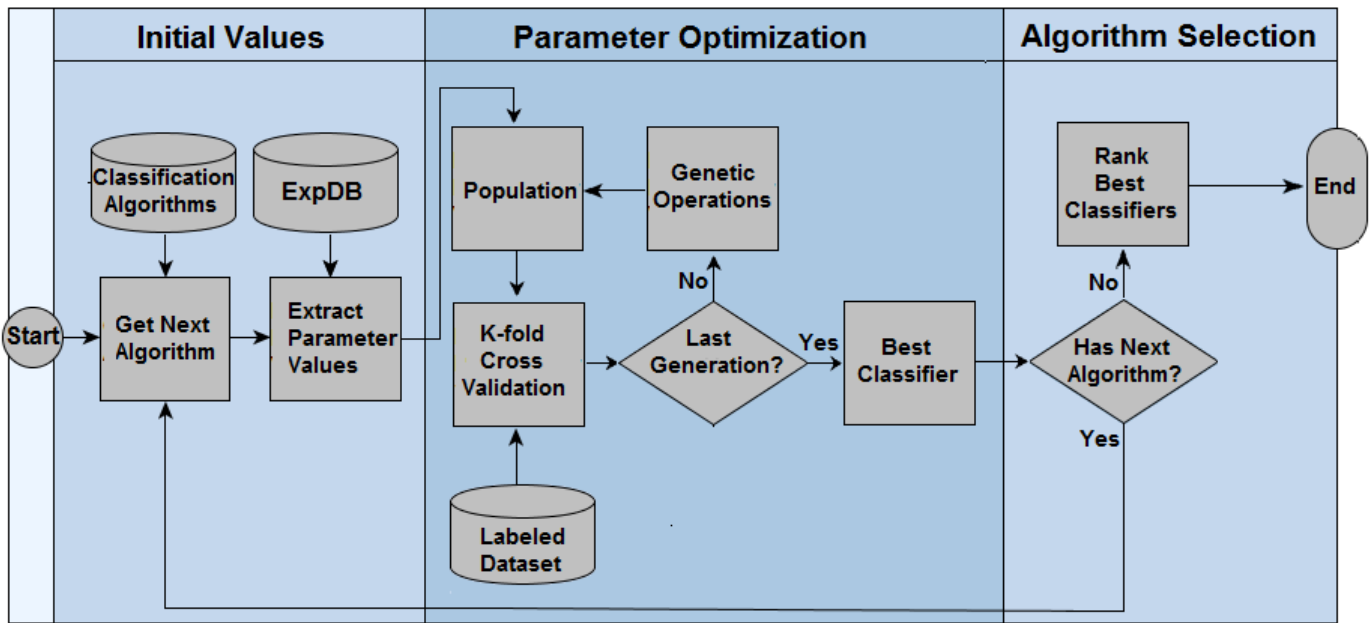
Fig. 1. Methodology for parameter optimization and algorithm selection.

TABLE II
DATASETS FROM THE UCI REPOSITORY.

ID	Name	Num. Discr.	Num. Cont.	Num. Classes	Num. Training	Num. Test
1	Abalone	1	7	28	2923	1254
2	Amazon	10000	0	50	1050	450
3	Cars	6	0	4	1210	518
4	Convex	0	784	2	50000	8000
5	Dexter	20000	0	2	420	180
6	German Credit	13	7	2	700	300
7	Gisette	5000	0	2	4900	2100
8	KDD09-Appentency	190	40	2	35000	15000
9	KR-vs-KP	37	0	2	2238	958
10	Madelon	500	0	2	1820	780
11	Secom	0	591	2	1096	471
12	Semeion	256	0	10	1115	478
13	Shuttle	9	0	7	43500	14500
14	Waveform	0	40	3	3500	1500
15	Wine Quality	0	11	11	3425	1469
16	Yeast	0	8	10	1038	446

TABLE III
NUMBER OF PARAMETER SETS OBTAINED FROM EXPDB.

ID	Algorithm	Avaliable ExpDB Parameter Sets	Used ExpDB Parameter Sets
A1	J48	15479	500
A2	Random Tree	21	21
A3	REPTree	21	21
A4	JRip	21	21
A5	PART	21	21
A6	Ridor	21	21

The motivation for these is to show that the use of the parameter values from ExpDB are really useful in obtaining classifiers with greater performance. The results obtained are shown in Table IV, fourth and eighth columns.

As a last round of experiments, we used our approach (Section 4). As long as it combines ExpDB and GA, each experiment consisted in performing 10-folds cross-validation (Eq. (1)) in each individual of a population of candidate classifiers (algorithm + parameters) over a number of generations. The population size and number of generations were fixed in 20 and 50 respectively. The final results can be seen in Table IV, fifth and ninth columns, for the training and test instances, respectively.

D. Discussion

The results obtained in the test performance show that the proposed approach has been effective. When we analyse line by line the Table IV we note that, in general, the parameters values from ExpDB overcome the results obtained by using default values. The exception occurs in dataset 13.

their default parameter values. The results obtained using the trainning instances (column "Num. Training" of Table II) are summarized in the second column of Table IV. In the sixth column of Table IV, we present the results obtained using the test instances (column "Num. Test" of Table II).

Next we performed experiments by using the sets of parameter values obtained from ExpDB. The number of parameters sets considered is shown in Table III. In this case, for each combination of algorithm and dataset, we performed as many experiments as the number of parameters sets present in Table III. In Table IV, third and seventh columns, we summarize the results obtained for trainning and test instances, respectively.

Thirdly we ran the GA for the classification algorithms and datasets being considered using 20 randomly chosen parameter sets as the initial population, over 100 generations.

Rayrone Zirtany Nunes Marques, Luciano Reis Coutinho, Tiago Bonini Borchartt, Samyr Béliche Vale, and Francisco José da Silva e Silva

TABLE IV

PERFORMANCE ON BOTH 10-FOLD CROSS-VALIDATION AND TEST DATA. EXPERIMENTS USING DEFAULT VALUES, EXPDB VALUES, GENETIC ALGORITHM AND GENETIC ALGORITHM WITH EXPDB VALUES (EMINER APPROACH). WE REPORT RESULTS AS MEAN ACCURACY ACROSS ALL CLASSIFIERS TESTED.

Dataset	10-Fold C.V. Performance (%)				Test Performance (%)			
	Default	ExpDB	GA	EMiner	Default	ExpDB	GA	EMiner
1	24.56	26.16	27.84	**28.39**	23.30	24.50	26.50	**27.69**
2	35.52	36.48	36.57	**36.76**	33.11	33.78	33.11	**36.67**
3	94.79	94.79	**96.69**	**96.69**	97.10	97.10	98.91	**98.91**
4	53.88	53.91	**53.94**	**53.94**	53.86	53.86	53.79	**53.79**
5	85.48	86.67	**92.38**	**92.38**	87.22	86.67	**88.33**	**88.33**
6	75.14	75.43	**77.43**	**77.43**	72.33	72.00	73.00	**73.00**
7	94.41	94.65	**95.31**	94.90	94.48	94.71	94.52	**94.76**
8	**98.20**	**98.20**	**98.20**	**98.20**	98.26	98.26	98.26	**98.26**
9	99.15	99.29	**99.69**	**99.69**	99.06	99.37	99.37	**99.58**
10	75.71	77.80	72.64	**79.75**	79.62	80.26	72.95	**83.26**
11	93.89	93.98	**94.44**	**94.44**	92.13	94.13	95.66	**95.66**
12	75.99	75.99	**78.58**	**78.58**	76.10	76.10	**76.52**	**76.52**
13	99.98	99.97	**99.99**	**99.99**	99.98	99.98	99.99	**99.99**
14	79.40	79.40	**81.11**	80.29	78.07	78.07	**78.20**	**78.20**
15	57.25	58.33	58.73	**59.38**	59.29	58.34	**61.23**	60.54
16	59.19	61.31	**61.50**	61.31	58.43	**59.78**	59.33	**59.78**

When we compare the columns of ExpDB values and the EMiner approach, we see that the combination of ExpDB and GA obtain better results than the use of the ExpDB parameter values alone (and better than the default parameter values, by transitivity). In all databases, we obtained classifiers with greater or equal accuracy by initiating with the ExpDB parameters values and evolving these parameter values by means of the GA.

When we analyse the simple GA approach with random initialization versus EMiner approach, we observe that the gain obtained by the combination ExpDB and GA is not only due to the GA, but it is also influenced by the use of the ExpDB parameter values to form the initial population for the GA. In general, an initial population of randomly chosen parameter values fared worse than one based on the ExpDB values. In our experiments, this observation failed only in datasets 7, 14 and 16 where we obtained a better classifier by starting from a randomly initial population. In most cases, when the genetic algorithm is started with the values of ExpDB, the best configuration of the parameters is found with a smaller number of generated populations. This advantage can be seen in the left chart in Fig. 2, where we compare the average number of generated populations until find the best accuracy for each dataset. The right chart in Fig. 2 show the relative performance improvement when the GA is started with ExpDB values. Among the sixteen datasets used, only in the *convex* dataset the use of ExpDB values has not resulted in a performance gain.

Given all these results we plot on Fig. 3 a bar chart showing the distribution of algorithm selection over the four group of experiments executed. Looking at the chart we see that the decision tree algorithms (J48, RandomTree and REPTree) dominated the algorithms based on classification rules (JRip, PART and Ridor). Of all, the most frequently used algorithm was J48 and PART, which were selected 4 in the experiments involving GA with ExpDB; 6 and 3 times (resp.) in GA alone; 7 and 6 time (resp.) in ExpDB alone, and 4 times in the

TABLE V
COMPARISON WITH OTHER WORK.

Dataset	EMiner	Thornton, et al. [6]
1	**27.69**	27.29
2	36.67	**62.44**
3	98.91	**100.00**
4	53,79	**77.95**
5	88.33	**92.78**
6	**73.00**	72.33
7	94.76	**97.76**
8	**98.26**	**98.26**
9	**83.26**	79.23
10	**95.66**	92.13
11	76.52	**94.97**
12	**99.99**	**99.99**
13	78.20	**85.80**
14	60.54	**66.44**
15	**59.78**	59.55

default values experiments. Notwithstanding this dominance, all algorithms had their change to be selected as the best one in at least one experimental setup, what shows the importance of automated solutions to the CASH optimization problem.

Finally, to complement the empirical evaluation of the proposed approach, we also compared the best results we obtained in Table IV with the best result found in literature [6], which use in their experiments the same datasets as we did. Table V summarizes the comparison. We observe that the proposed approach outperforms many time the results we could found on the literature.

Compared to the literature, we conclude that the main contribution of our work is the use of GA to select the most promising combination of the algorithm and parameter values for a given classification task, starting the search with experimental data retrieved from the public ExpDB repository [10]. Our approach to the CASH optimization problem is an alternative to the approaches presented in [6] and [7].
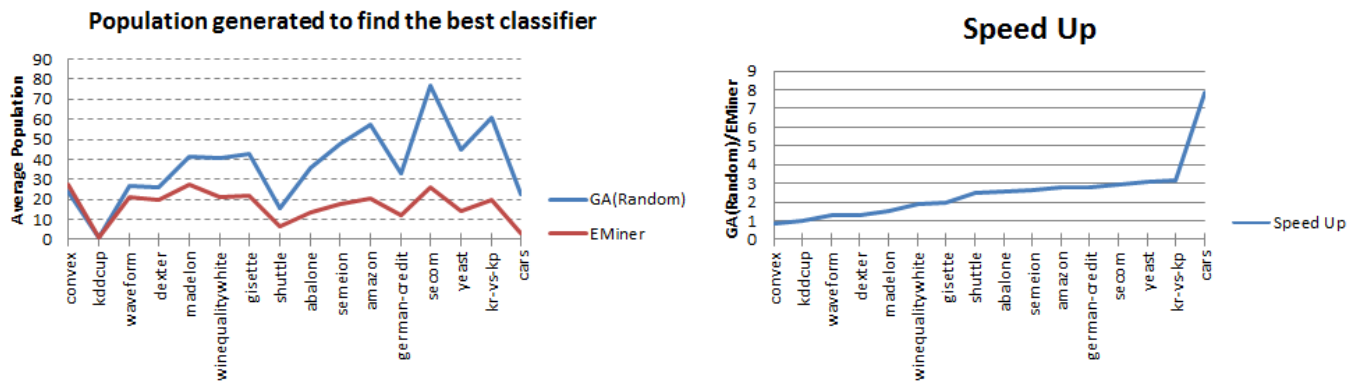
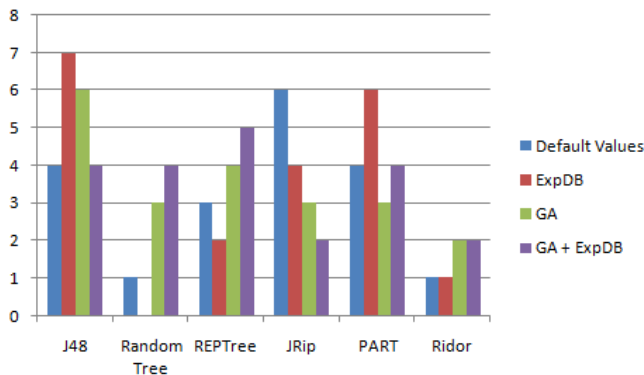Fig. 2. Speedup comparison between simple GA approach and EMiner.

Fig. 3. Distribution of algorithm selection.

VI. CONCLUSION

In this work, it was shown how the problem of algorithm selection and (hyper)parameter optimization, or CASH optimization problem, can be approached in an automated manner. We have proposed an approach that combines GA with data extracted from the ExpDB to search for a set of suitable parameter values, for a given list of classification algorithms, when we want to find a classification model with maximum accuracy/generality on a specific data mining application. The classification algorithms experimented in this study were 3 algorithms for decision tree induction (J48, Random Tree and REPTree) and 3 for classification rule induction (JRip, PART and Ridor). The proposed approach was implemented in tool called EMiner, built on the context of the ECADeG project.

Regarding the CASH optimization problem, the main insight of this work is the use of GA with a non-random initial population. This non-random initial population consists of data from various experiments previously conducted in datasets in different areas, and stored in ExpDB. Regarding the application of data mining in medicine and healthcare, the main contribution is the development of an automated tool to help medical researchers to come up with good classification models, despite not being experts in data mining algorithms and techniques.

To validate our approach, several experiments have been performed. These experiments allowed us to evaluate the gain in accuracy due to the introduction of ExpDB and GA, the two basic elements of the proposed approach. Also, these experiments allowed us to compare the approach with some work reported in the literature. We have concluded that the proposed approach is effective in some of the cases analysed.

As future work we envisage several research paths. The first one is to further test the approach with other data mining algorithms. Another is to conduct experiments to assess in more detail the real effectiveness and efficiency of starting from an initial population based on data from ExpDB instead of starting from a random one. Finally, as we strive for the construction of practical tools to help researchers in the area of medicine and healthcare, we are thinking in conducting qualitative experiments to see how our EMiner tool fares when evaluated by these professionals.

REFERENCES

[1] N. Lavrac, "Selected techniques for data mining in medicine," *Artificial Intelligence in Medicine*, vol. 16, pp. 3–23, 1999.

[2] M. Adnan, W. Husain, and N. Rashid, "A hybrid approach using naïve bayes and genetic algorithm for childhood obesity prediction," in *2012 International Conference on Computer Information Science (ICCIS)*, vol. 1, June 2012, pp. 281–285.

[3] P. K. Srimani and M. S. Koti, "Medical diagnosis using ensemble classifiers – a novel machine-learning approach," in *Journal of Advanced Computing*, 1st ed. Columbia International Publishing, 2013, pp. 9–27.

[4] S. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *2013 IEEE Conference on Information Communication Technologies (ICT)*, April 2013, pp. 1227–1231.

[5] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Pratical Machine Learning Tools and Techniques*, 3rd ed. Elsevier, 2011.

[6] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD-13. New York, NY, USA: ACM, 2013, pp. 847–855. [Online]. Available: http://doi.acm.org/10.1145/2487575.2487629

[7] R. Leite, P. Brazdil, and J. Vanschoren, "Selecting classification algorithms with active testing," in *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition*, ser. MLDM-12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 117–131. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31537-4_10

[8] F. Maia, R. Araújo, L. C. Muniz, R. Zirtany, L. Coutinho, S. Vale, F. J. Silva, P. Cincilla, I. Chabbouh, S. Monnet, L. Arantes, and M. Shapiro, "A grid based distributed cooperative environment for health care research," in *Foundations of Health Information Engineering and Systems Lecture Notes in Computer Science Volume*, vol. 7789. Springer Berlin Heidelberg, 2013, pp. 142–150. [Online]. Available: http://link.springer.com/chapter/10.1007%2F978-3-642-39088-3_9

[9] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

[10] H. Blockeel, J. Vanschoren, B. Pfahringer, and G. Holmes, "Experiment databases," *Machine Learning*, vol. 87, no. 2, pp. 127–158, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10994-011-5277-0

[11] F. Samadzadegan, A. Soleymani, and R. Abbaspour, "Evaluation of genetic algorithms for tuning SVM parameters in multi-class problems," in *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, Nov 2010, pp. 323–328.

[12] J. Zhou, O. Maruatona, and W. Wang, "Parameter optimization for support vector machine classifier with IO-GA," in *2011 First International Workshop on Complexity and Data Mining (IWCDM)*, Sept 2011, pp. 117–120.

[13] K.-K. Seo, "A GA-based feature subset selection and parameter optimization of support vector machine for content-based image retrieval," in *Proceedings of the 3rd International Conference on Advanced Data Mining and Applications*, ser. ADMA-07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 594–604. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73871-8_57

[14] A. Sureka and K. Indukuri, "Using genetic algorithms for parameter optimization in building predictive data mining models," in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science, C. Tang, C. Ling, X. Zhou, N. Cercone, and X. Li, Eds. Springer Berlin Heidelberg, 2008, vol. 5139, pp. 260–271. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88192-6_25

[15] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2*, ser. IJCAI-95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[16] J. R. Quinlan, "Improved use of continuous attributes in c4.5," *J. Artif. Int. Res.*, vol. 4, no. 1, pp. 77–90, Mar. 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622737.1622742

[17] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to select machine learning scheme for classification of biomedical datasets," Islamabad, Pakistan, 2009.

[18] D.-Y. Chen, T.-R. Chuang, and S.-C. Tsai, "JGAP: A Java-based graph algorithms platform," *Softw. Pract. Exper.*, vol. 31, no. 7, pp. 615–635, Jun. 2001. [Online]. Available: http://dx.doi.org/10.1002/spe.379

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[20] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database

Fabio Rinaldi, Oscar Lithgow-Serrano, Alejandra López-Fuentes, Socorro Gama-Castro,
Yalbi I. Balderas-Martínez, Hilda Solano-Lira, and Julio Collado-Vides

Abstract—As part of a large-scale biocuration project, we are developing innovative techniques to process the biomedical literature and extract information relevant to specific biological investigations. Biological experts routinely extract core information from the scientific literature using a manual process known as scientific curation. The aim of our activity is to improve the efficiency of this process by leveraging upon natural language processing technologies in a text mining system. There are two lines of investigation that we pursue: (1) finding information relevant for curation and present it in an adaptive interface, and (2) use sentence-similarity techniques to create interlinks across articles in order to allow a process of knowledge discovery.

Index Terms—Text mining, natural language processing, biocuration.

I. INTRODUCTION

THANKS to novel technological developments in genomics and the emergence of multiple high-throughput (H-T) strategies, we live in a time when studies are producing a tsunami of data. With the next generation sequencing technology, the amount of genomic data is now growing faster than the computational power [1]. In spite of the large number of databases and bioinformatics resources, a critical barrier in the field is how to accelerate access to and processing of such large amounts of information and knowledge. H-T-generated data produces large collections of individual, disconnected elements. On the other hand, pregenomic scientific papers tend to discover several interrelated elements with experiments that support more integrated perspectives, but limited to specific biological systems. The gathering in an organized and accessible database of detailed, manually curated collections of such well-studied biological systems provides the framework for an integrated understanding that is fundamental in genomics research.

RegulonDB [2] is a database, with manually curated knowledge, extracted from the literature, describing information related to transcriptional regulation in *Escherichia coli* K-12.[1] It contains biological objects such as genes, promoters, transcription factors (TFs), transcription factor binding sites (TFBSs), terminators and operons. It also contains relations of regulation among TFs and genes, promoters and operons. RegulonDB, first published in 1998, marked an effort that continues to this day for continuous and expanded curation [2], [3]. Briefly, RegulonDB facilitates access to organized information on the mechanisms of transcription initiation and it has been successful in this work; however, currently it does not facilitate access to fundamental concepts, generalizations, and knowledge of regulation of transcription initiation in *E. coli* (frequently found in reviews).

Several technical limitations have restricted the work to do so, and as a consequence, RegulonDB has only captured the knowledge contained in an estimated 10 to 15% of all sentences available in the literature of 5,244 original scientific papers supporting this database (version 8.6). This estimate is based on the number of sentences behind knowledge about TFs, TFBSs and their functions affecting promoters, the regulated TUs, and operons encoded in the database. Based on this diagnosis, we decided to improve the efficiency of biocuration process by leveraging upon natural language processing technologies in text mining systems.

Biomedical text mining can be used to partially automate the process of biomedical literature curation by using sophisticated algorithms for discovering biomedical entities together with interaction and events in which they participate. A successful biomedical text mining system is typically based on a pipeline which first discovers entities of interest in the text of a scientific article and subsequently looks for interactions between them. As described above, finding the unique database identifiers of the entities in focus is an important step in this process. Which database identifiers are used in this process depends largely on the application for which a text mining system is built, or in other words, the database for which the system is designed to extract information.

In order to accomplish the goal of digitally-assisted curation, we are working simultaneously on two main lines of research: (1) finding information relevant for curation and present

[1]http://regulondb.ccg.unam.mx/

Fabio Rinaldi, Oscar Lithgow-Serrano, Alejandra López-Fuentes, Socorro Gama-Castro, Yalbi I. Balderas-Martínez, Hilda Solano-Lira, et al.

it in an adaptive interface, and (2) use sentence-similarity techniques to create interlinks across articles in order to allow a process of knowledge discovery. These steps are described in detail in this paper, together with the preliminary design of the integrated system.

II. The OntoGene Text Mining System

We use an existing biomedical text mining system (OntoGene) in order to process a collection of documents relevant to the curation purposes of RegulonDB. Ontogene offers a powerful and flexible entity recognition module based on a dictionary lookup approach allowing for some variants and a post-annotation filtering module based on maximum entropy techniques. The aim of this section is to provide a brief description of the OntoGene Text Mining pipeline, which in the RegulonDB application is used to provide the basic preprocessing capabilities as well as for the identification and normalization of domain entities. For additional details the reader is invited to consult some of the related publications [4], [5], [6]. The publications used in the experiments described in this paper were either downloaded from PubMed Central (an open access repository of biomedical literatures) in XML format, when available, or converted from PDF using the PDFlib Text Extraction Toolkit[2].

The full sequence of processing steps offered by the OntoGene pipeline is the following, however the last three steps where not used for the applications described in this paper.

- Transformation of input format into the OntoGene-specific XML format
- Zoning: partition of the document in sections such as title, abstract, references, etc.
- Sentence splitting
- Tokenization
- Part of Speech Tagging
- Lemmatization
- Stemming
- Named Entity Recognition
- Chunking
- Dependency Parsing
- Detection of Interactions

The named entity recognition step is based on a large internal database of domain terms, sourced from life science databases, and customizable by the end user of the application. Several life science databases can be considered a rich terminological resources, since they provide not only concept descriptions, but also the terms that are actually used by researchers to refer to a particular concept. The OntoGene database can be automatically generated from a subset of such resources, taking from them the preferred names and synonyms of user-selected term categories. As term names are stored together with their original database identifiers, it is always possible to retrieve all information associated with

[2]http://www.pdflib.com/?id=12

a given concept. The OntoGene system takes automatically into account a number of possible minor variants of the terms (e.g. hyphen replaced by space), thus increasing the flexibility of term recognition. The annotation step automatically adds to the XML representation of the document a list of possible database identifiers for each term where a match was found.

The OntoGene pipeline is also optionally capable of generating candidate interactions among the detected domain entities. The approach is based on a preliminary generation of potential interactions by combinatorial pairwise combinations of entities in a given text span (typically a paragraph). In order to balance the low precision of such an approach, a machine-learning based reranking is performed after the initial combination. The reranking takes into account lexical features, such as word stems and PoS tags, syntactic features, such as dependency parses, and global distribution features, such as relative frequencies of terms in the specific paper compared to the average distribution in the whole collection.

The system is trained using a distant learning approach taking a reference database as provider of the "ground truth". For example, in a recent industrial application aimed at large-scale detection of protein-protein interactions from the literature, the BioGrid database was used as a reference. BioGrid is a very large scale manually curated resource about protein interactions and genetic interactions. In the specific application described in this paper, the RegulonDB database itself is taken as the ground truth reference.

In practice the OntoGene system uses a supervised machine learning method (based on a maximum entropy classifier) in order to compute a probability of a term/concept pair to be part of a relationship in the reference database. This probability score can be used to weed out false positive entities erroneously provided by the high-recall dictionary-based annotation system. Additionally, given that each annotated term can be associated to several identifiers in the reference database, the most likely association can be selected, thus leading to the disambiguation of the possible meanings of the term.

As a second step, the probabilities of the two concepts that participate in a candidate interactions are combined using their harmonic mean, producing a score for the relationship. These scores allow a ranking of the candidate relationships, and therefore either an automated selection based on a threshold, or a manual selection based on the inspection of the most likely candidates by domain experts.

The end result of the processing steps described above is an XML version of the original document enriched with information coming from the various modules. In particular, the information of relevance for the end users is the annotations of the domain entities, and (optionally) a set of candidate interactions. This rich XML format can be browsed through a specifically designed interface called OntoGene Document Inspector (ODIN) [7], [5], [8].

For the particular application discussed in this paper, the first step of processing consists in annotation of the
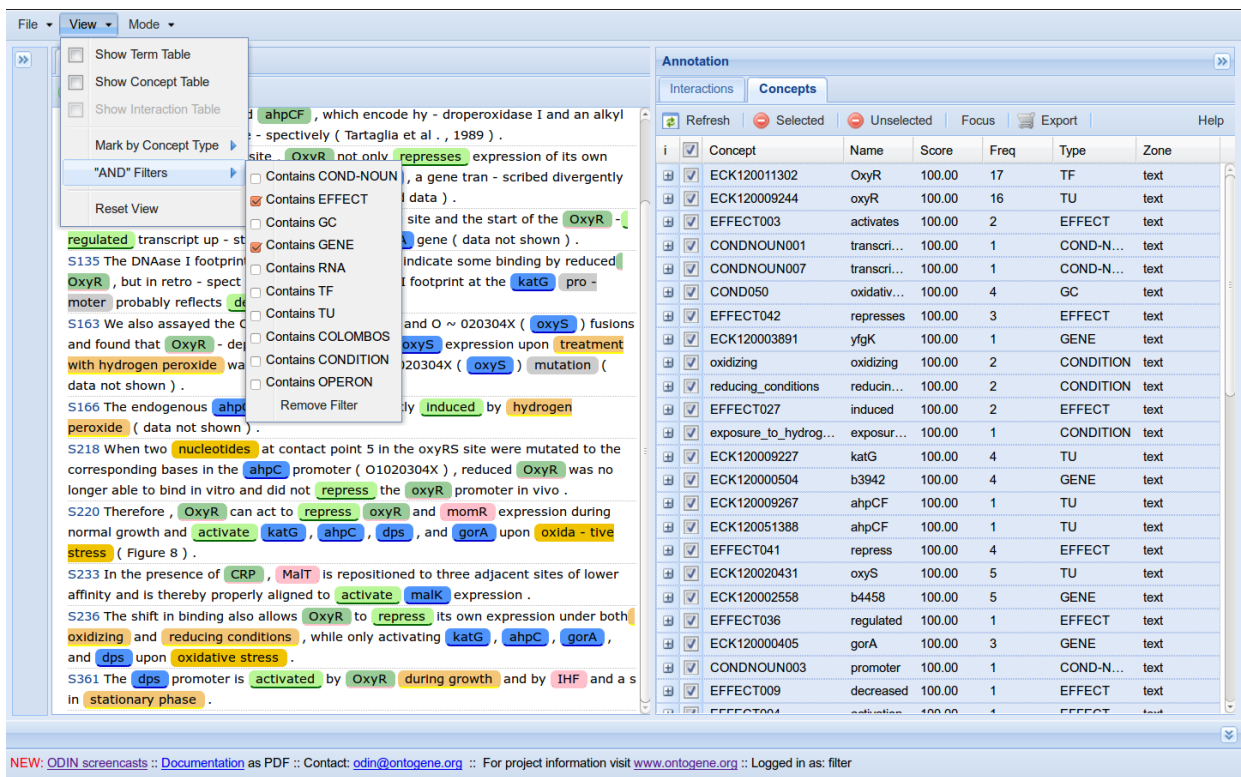
Fig. 1. ODIN customized for RegulonDB

domain entities which express the core knowledge of interest for RegulonDB curators: genes, transcription factors, growth conditions, etc. The list of genes of *E. coli* is derived from GenBank. In a recent experiment [9], we analysed a small set of articles relevant for the topic of genetic response to oxydative stress. All articles were annotated by the OntoGene pipeline and inspected by RegulonDB curators through the ODIN interface. In particular, ODIN offers a functionality called "sentence filteres", which allow the curators to select sentences which satisfy a simple logical condition defined by the user. Typically such a filter is defined by the presence in the same sentence of entities of two predefined types (e.g. "gene" and "effect"). Such condition is defined in order to locate sentences which are likely to contain the information that the curators need to extract from the documents. The experiment mentioned above showed that the curators could identify the desired items by reading fragments of the papers equivalent to only 11% of the total material that they would have had to read if they worked without the support of the assisted curation tool.

Fig. 2. Exemplification of the process of semantic linking

III. LINKING SENTENCES ACROSS ARTICLES BASED ON THEIR SIMILARITY

In the context of Natural Language Processing (NLP), Semantic Similarity between two texts is the task of evaluate the likeness of their meaning. It is a recurrent and important approach to address the na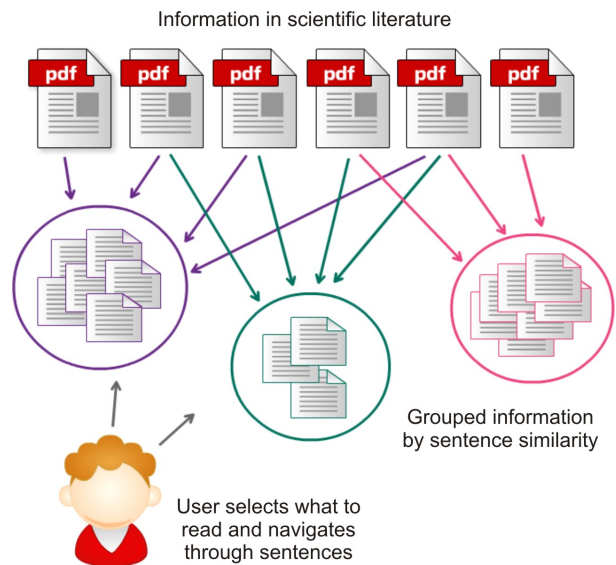tural language understanding issue in tasks such as summarization, paraphrasing, QA-Systems, conversational agents, etc.

Common approaches to its computation are based on the combined use of syntactic and semantic features of the texts being compared, such as: the lexical class, ontological similarity, syntactic category, etc. These features correspond to

Fabio Rinaldi, Oscar Lithgow-Serrano, Alejandra López-Fuentes, Socorro Gama-Castro, Yalbi I. Balderas-Martínez, Hilda Solano-Lira, et al.

knowledge of experts encoded in ontologies, lexical resources and syntactic parsers, among others.

Other kinds of features, also used, are those extracted from the automated analysis and statistical interpretation of large amount of texts (corpora) [10]. This distributional perspective relies on the hypothesis that the meaning of word and texts can be inferred by the context where it is used or, put in other words, that text occurring in similar context have similar meaning [11].

When doing biocuration, the experts read one by one a set of topic-related articles to annotate relevant information. This technique works well in the sense that relevant information is identified but having to read the whole articles sequentially is very time consuming. So based on the fact that the documents have several topics in common, we designed a system that uses sentence similarity to link sentences about the same subject across all the articles in the set. For instance, complex sentences (like examples a, b and c) will be related, since they are about the same topic:

a. *The oxidized form of Oxy is a transcriptional activator of a multitude of genes that assist in protecting the cell from oxidative damage* [12].

b. *Activated Oxy then induces transcription of a set of antioxidant genes, including katG (hydroperoxidase I), ahpCF (alkylhydroperoxidase), dps (a nonspecific DNA binding protein), gorA (glutathione reductase), grxA (glutaredoxin I), and oxyS (a regulatory RNA)* [13].

c. *A hallmark of the E. coli response to hydrogen peroxide is the rapid and strong induction of a set of Oxy -regulated genes, including dps, katG, grxA, ahpCF, and trxC* [14].

This way, the "linear reading" is modified, allowing the expert to choose one sentence of interest and jump/ navigate through other articles, guided by the current topic of interest. The system is formed by 4 components which are:

- A user-friendly web interface
- A web service layer
- A relational database
- A semantic similarity engine

The 3 former components are orchestrated by RestFul Web services which respond to the user's petition to search on the database, either key words or semantic similar sentences. On the other hand, the semantic similarity engine is an off-line process that is in charge of processing the articles and registers the results, sentences and relations, in the database.

The semantic similarity engine was built as a micro-framework where each involved step is an interface that can be re-implemented or extended in order to test different strategies. Besides, the implementation classes and external libraries are dynamically loaded from a configuration file; this alleviates the need of recompiling the framework in order to test different strategies.

In the current version, the processing sequence is the following:

- Apply Part-Of-Speech (POS) tagging using the Stanford POS tagger [15].

- Apply stemming using the snowball stemmer [16].
- Apply a rule based sentence selector, i.e. regular expressions that are based on the words' POS-tags. The motivation for this is to restrict the set of candidates and to keep those more informative. For example one of the rules is to select those sentences that contains the pattern "Noun-Verb-Noun with other optional tags like determinants, adjectives, etc.":

```
[IN] [DT] [JJ] NN [RB] VB [IN] [DT]
[JJ] NN
```

- Create a multidimensional vector representation of the sentence. Selected sentences are represented by vectors with as many dimensions as the length of the global vocabulary. Each vector dimension embodies a word of the vocabulary and the dimension's magnitude is the word frequency in the sentence. It is worth noting that counts are not normalized.
- Measure the similarity between each two sentences using the cosine similarity. This measure is particularly convenient because the length of the vectors is irrelevant in its computation. Moreover, it provides a confined similarity measure that ranges between -1 and 1, being 0 when the vectors are orthogonal (i.e. not related) and 1 when both vectors are identical. The similarity is computed using Efficient Java Matrix Library [17].

The novel web interface (see figure 3), currently in the implementation phase, provides to the user the means to search key words on one or several articles. Once the results are listed and the user decide to enter to an specific article, the article's sentences (content) are displayed and those which have semantic relations with other sentences, either in the same article or in others, are decorated with hyperlinks. When the user selects a hyperlink the related sentences that are located inside the current article are highlighted, and those which resides in others are listed in a panel along with the article name and the corpus to which it belongs to. In that way the user is provided with an instrument to navigate across different articles and corpus through the following up of a specific idea.

IV. EVALUATION

There have been several separate evaluations of the modules described in this paper. As a way to verify the quality of the core text mining functionalities, the underlying text mining pipeline has been used to perform several tasks which have been formally evaluated within the context of community-organized text mining evaluations campaigns ("shared tasks"), such as BioCreative [18]. Some of most interesting results include best results in the detection of protein-protein interactions in BioCreative 2009 [19], top-ranked results in several tasks of BioCreative 2010 [20], best results in the triage task of BioCreative 2012 [21].

The usefulness of ODIN as a curation tool, leveraging upon the results of the text mining system, has been demonstrated through an experiment aimed at making more efficient the
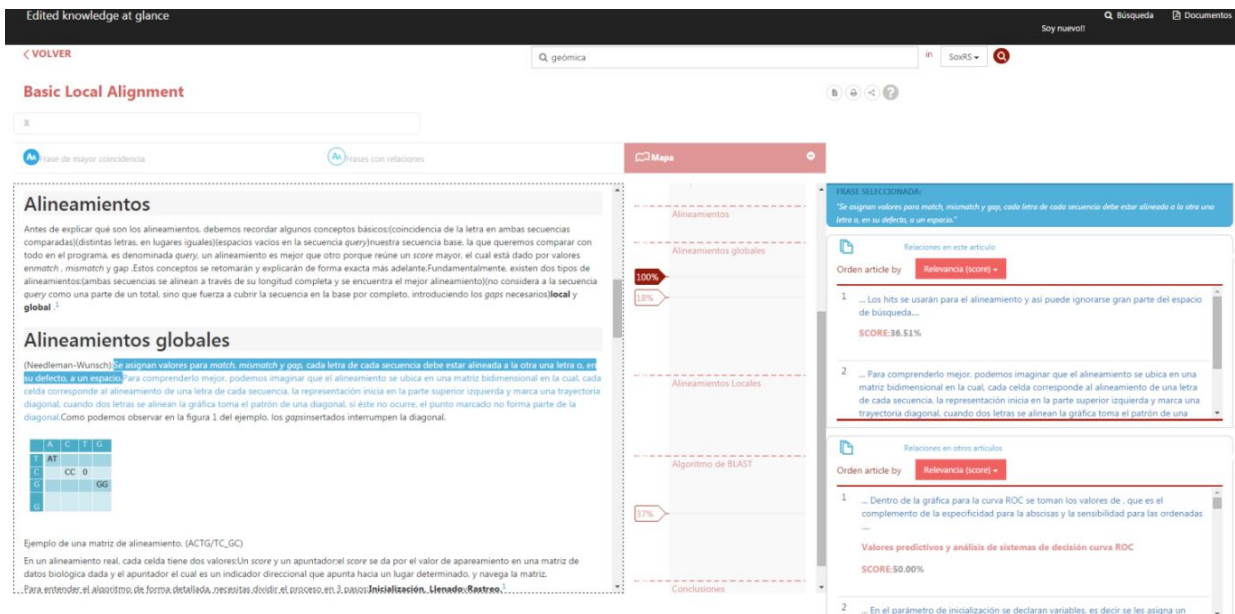
Fig. 3. Preliminary version of the interface of the system

identification by the curators of specific types of entities (*"growth conditions"*) which had not been curated before, and were therefore not part of the reference database. An initial set of articles, likely to be relevant for the given task, was identifies through conventional IR techniques. The curators used ODIN filters to restrict their view of the selected articles to the set of sentences satisfying a given logical condition (e.g. containing an entity of type *"transcription factor"* and an entity of type *"effect"*). The manual analysis of the selected sentences allowed them to identify the missing information in 75% of the cases, but having to inspect only about 10% of the articles, thus providing a considerable improvement in efficiency, as described in section II.

More recently, we tested the sentence similarity component using three set of scientific articles:

1) 42 articles of SoxRS: oxidative stress in *Escherechia coli* K12
2) 35 articles of *Salmonella typhimurium* pathogenicity island SP1
3) 10 articles of role of EZH2 gene in cancer

We had six domain experts that worked with these sets (two per set). The goal for the test exercise was to read the articles looking for specific information, as it's done in the curation process. Then to extract and save all the information they could find within 2 hours of reading the literature. One of the experts from each set had access to the system, the other didn't and used the PDF files instead. The users with access to the system were able to review more articles thus they extracted more sentences in total with similar information. The users with the files couldn't finish all articles but they extracted more sentences per reviewed article.

The general opinion form the experts was that the system could be very powerful if the similarity is improved to detect more topic-related sentences and also made some suggestions to the web interface in order to be more intuitive. The system has proved to be useful for the curation process, we are now working to add more capabilities, improve the interface design by implementing User eXperience (UX) techniques, and integrate all components in a single unified system. Figure 4 shows a schematic representation of the system that we are now in the process of implementing.

V. CONCLUSION

The work described in this paper takes place in the context of a large-scale NIH-funded[3] four-year project, started in 2015, which has as one of its goals the implementation of a process of digital assisted curation, which involves the integration of advanced text mining techniques within the curation pipeline of a biomedical database. Human experts will be able to leverage upon the best results of text mining technologies in order to improve the effectiveness of the curation process without sacrificing its quality.

This paper describes some of the components that will be used in order to reach that goal: an advanced text mining pipeline for entity extraction and relation detection, a customizable flexible user-interface, and a way to interlink information across several papers. The new system will constitute a very powerful curation tool that will allow semiautomatic data annotation, and a new way of knowledge discovery reducing reading time without affecting understanding.
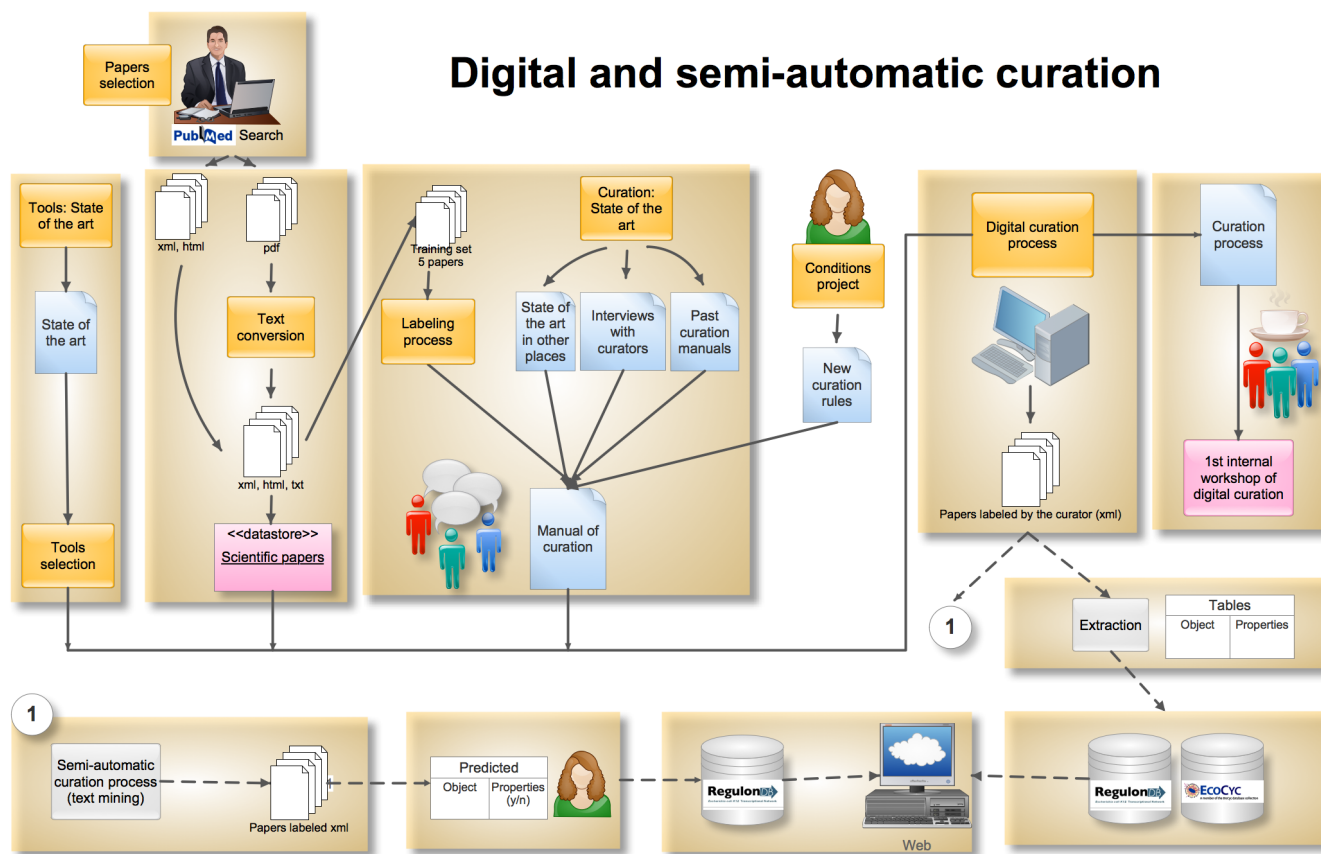
[3]National Institute of Health, US

Fig. 4. Diagram illustrating the process of digital curation

REFERENCES

[1] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, p. 207, May 2010. [Online]. Available: http://dx.doi.org/10.1186/gb-2010-11-5-207

[2] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, "RegulonDB: A database on transcriptional regulation in escherichia coli," *Nucleic Acids Research*, vol. 26, no. 1, pp. 55–59, 1998. [Online]. Available: http://dx.doi.org/10.1093/nar/26.1.55

[3] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muniz-Rascado, J. S. Garcia-Sotelo, V. Weiss, H. Solano-Lira, I. Martinez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernandez, K. Alquicira-Hernandez, A. Lopez-Fuentes, L. Porron-Sotelo, A. M. Huerta, C. Bonavides-Martinez, Y. I. Balderas-Martinez, L. Pannier, M. Olvera, A. Labastida, V. Jimenez-Jacinto, L. Vega-Alvarado, V. D. Moral-Chavez, A. Hernandez-Alvarez, E. Morett, and J. Collado-Vides, "RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D203–D213, 2013.

[4] F. Rinaldi, "The ontogene system: An advanced information extraction application for biological literature," *EMBnet.journal*, vol. 18, no. Suppl B, pp. 47–49, 2012. [Online]. Available: http://journal.embnet.org/index.php/embnetjournal/article/view/546/755

[5] F. Rinaldi, S. Clematide, Y. Garten, M. Whirl-Carrillo, L. Gong, J. M. Hebert, K. Sangkuhl, C. F. Thorn, T. E. Klein, and R. B. Altman, "Using ODIN for a PharmGKB re-validation experiment," *Database: The Journal of Biological Databases and Curation*, vol. 2012, pp. 1–12, 2012. [Online]. Available: http://database.oxfordjournals.org/content/2012/bas021.full

[6] F. Rinaldi, S. Clematide, H. Marques, T. Ellendorff, R. Rodriguez-Esteban, and M. Romacker, "Ontogene web services for biomedical text mining," *BMC Bioinformatics*, vol. 15, no. Suppl 14, p. S6, 2014.

[7] F. Rinaldi, S. Clematide, and G. Schneider, "Odin: Advanced text mining in support of the curation process," in *Pacific Symposium on Biocomputing (PSB)*, Big Island, Hawaii, Jan. 2012.

[8] F. Rinaldi, A. P. Davis, C. Southan, S. Clematide, T. R. Ellendorff, and G. Schneider, "ODIN: a customizable literature curation tool," in *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, vol. 1, 2013, pp. 219–223.

[9] S. Gama-Castro, F. Rinaldi, A. López-Fuentes, Y. I. Balderas-Martínez, S. Clematide, T. R. Ellendorff, A. Santos-Zavaleta, H. Marques-Madeira, and J. Collado-Vides, "Assisted curation of regulatory interactions and growth conditions of OxyR in E. coli K-12," *Database: The Journal of Biological Databases and Curation*, vol. bau049, pp. 1–13, 2014. [Online]. Available: http://database.oxfordjournals.org/content/2014/bau049

[10] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proceedings of the 21st National conference on Artificial Intelligence*, vol. 1, 2006, pp. 775–780. [Online]. Available: http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf

[11] H. Schutze, "Dimensions of meaning," in *Proceedings Supercomputing 1992*, 1992, pp. 787–796. [Online]. Available: http://ieeexplore.ieee. org/lpdocs/epic03/wrapper.htm?arnumber=236684

[12] A. Wallecha, J. Correnti, V. Munster, and M. van der Woude, "Phase variation of ag43 is independent of the oxidation state of oxyr," *Journal of Bacteriology*, vol. 185, no. 7, pp. 2203–2209, 2003.

[13] M. Zheng, B. Doan, T. D. Schneider, and G. Storz, "Oxyr and soxrs regulation of fur," *Journal of Bacteriology*, vol. 181, no. 15, pp. 4639–4643, 1999. [Online]. Available: http://www.ncbi.nlm.nih. gov/pmc/articles/PMC103597/

[14] M. Zheng, X. Wang, L. J. Templeton, D. R. Smulski, R. A. LaRossa, and G. Storz, "DNA microarray-mediated transcriptional profiling of the escherichia coli response to hydrogen peroxide," *Journal of Bacteriology*, vol. 183, no. 5, pp. 4562–4570, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC95351/

[15] K. Toutanova, D. Klein, and C. D. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, vol. 1, 2003, pp. 252–259. [Online]. Available: http://dl.acm.org/citation.cfm?id=1073478

[16] M. Porter, "Snowball: A language for stemming algorithms," 2001. [Online]. Available: http://www.snowball.tartarus.org/texts/introduction. html

[17] A. Peter, "Efficient java matrix library (EJML)." [Online]. Available: http://ejml.org

[18] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. Wilbur, L. Rocha, H. Shatkay, A. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. Dogan, J.-F. Fontaine, M. Andrade-Navarro, and A. Valencia, "The protein-protein interaction tasks of biocreative III: Classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, vol. 12, no. Suppl 8, p. S3, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/ 12/S8/S3

[19] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, and M. Romacker, "OntoGene in BioCreative II.5," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 472–480, 2010.

[20] G. Schneider, S. Clematide, and F. Rinaldi, "Detection of interaction articles and experimental methods in biomedical literature," *BMC Bioinformatics*, vol. 12, no. Suppl 8, p. S13, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S8/S13

[21] F. Rinaldi, S. Clematide, S. Hafner, G. Schneider, G. Grigonyte, M. Romacker, and T. Vachon, "Using the OntoGene pipeline for the triage task of BioCreative 2012," *The Journal of Biological Databases and Curation*, vol. bas053, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3568389/

Warnings and Recommendation System for an E-Learning Platform

Camilo Peñuela, Elizabeth León, and Jonatan Gómez

Abstract—A warning messages and recommendation system for an E-Learning system is proposed, the goal is to identify which students are likely to have a poor academic performance, and give them timely feedback by showing alerts and recommended material. The proposed system uses a set of profiles previously identified by a student profiling model, using socio-economic (age and gender) and web navigation data on the system (number of accesses to resources, percentage of accesses in class, average absence time and average session length). Each profile is analyzed and a warning message is assigned to each one; also, the sequences of consultations performed by students with a high academic performance are recognized and used to choose which resources are recommended. Based on the sequence performed by a student in a current session, the platform may recommend access specific resources.

Index Terms—Learning management system, web log, student profile, educational data mining.

I. Introduction

THE traditional educational system is built upon a long-established set of customs: the professor transmits the knowledge by giving a talk to his/her passive students, in a well-established place (class room), linear built (topics are sorted from start to end), standardized tests, and usually with a high student/teacher ratio. This system supposedly guarantees the knowledge transmission, but the standardized tests determine in a general way if the students are acquiring the knowledge, regardless if memorizing or appropriating it.

Thanks to computing and communication technologies, educational models have transformed from monolithic linear models to flexible non-linear models centered in the skills and abilities of each student, they allow the students to learn in their own way and at their own rhythm, guided by an online tool independent of location or student's time [1], [2]. Several educational processes, including both classroom attendance and distance education (radio, TV, internet) have been included [3].

E-learning is a distance educational model which integrates information technologies with pedagogical elements in order to teach online [1], [2], [4], [5]. An e-learning system is comprised of several subsystems, it may include [6]:

Learning Management System, Content Management System, Intelligent Tutoring System, Computer Adaptive Test, and Topic interaction and simulation tools. In particular if contents, tests, and tutors are able to adapt according to the skills of each student, the e-learning system is called adaptive e-learning system [7].

These E-learning systems are producing a huge amount of data that can be used for evaluating and improving the learning strategies [8], and there are several research works trying to link data related to a student with his/her academic performance, e.g., trying to discover some student profiles [9]. Profiles are formal structures of pieces of information related to users, usually defined to represent categories of shared common features [10]. By identifying student profiles, an e-learning system can, among others: present personalized content [10], customize the activities to encourage participation [11], predict and prevent academic failure [12], and gather information that allows to improve the course for reducing the dropout ratio and increase the students' performance [1]. Several factors that affect students' academic performance have been found using data mining techniques: familiar, personal, economic, and geographic [13], [14], [15], [16]. However, results of these studies depend on the kind of data used to identify the profiles, e.g., socio-economic, number of accesses to the course material [17], time spent by students in consulting the material [18], and web server logs [19]. Web logs can be used to identify users navigation profiles.

Web mining is the application of data mining to data collected from the web [20]. It has three main categories: content mining (which analyses the web content), structure mining (which analyses the links between pages), and usage mining (which extracts useful information from server logs) [20]. Regarding usage mining, several factors must be taken into account: not all pages across the web are of equal importance to students [21], many pages and documents are consulted just for location rather than for importance [21], and not every access to a page is registered (due to cache loading), such lost accesses can be recovered through the identification of possible paths based on the website structure tree [22].

In this paper, a warnings and recommendation system to give timely feedback to students who are likely to have a poor academic performance, is proposed, this system uses the profiles found by a student profiling model, which is further described in [23], this model uses the data registered by a LMS called "Virtual Intelligent Learning Platform": socio-economic (gender and birth date) and behavioral (accesses to the course

Camilo Peñuela, Elizabeth León, Jonatan Gómez

TABLE I
AMOUNT OF DATA BY TERM/TEST

Test	Students	Sessions	Accesses
First. 2014-1	544	4675	25205
Second. 2014-1	469	3313	10734
Third. 2014-1	360	2272	5549
First. 2014-2	469	5022	24721
Second. 2014-2	412	3593	12190
Third. 2014-2	312	1698	3795

TABLE II
FEATURES USED FOR THE CLUSTERING PROCESS

Feature	Description
ag	Student's age
gd	Student's gender
nD	Number of accesses to documents
nE	Number of accesses to exercises
nV	Number of accesses to videos
dC	Percentage of accesses to documents in class time
eC	Percentage of accesses to exercises in class time
vC	Percentage of accesses to videos in class time
sl	Average time spent by session
ta	Average absence time between sessions

material). Each cluster is a potential profile, it is analyzed to find a connection between its properties and the academic performance, if the performance was low, a warning message based on the profile properties is assigned to the profile. Once the profiles, each assignment of student to a profile and the warning messages were stored on the knowledge base, the warnings system assigns one profile to each active student (who is currently taking the course), if the student is assigned to a profile with an associated warning message, that message is shown in the Platform.

This paper is organized as follows: In Section 2, a brief introduction to the student profiling model is presented, including the preprocessing, and clustering. In Section 3, the warnings and recommendation system proposed in this paper is presented. In Section 4, conclusions and future work are presented.

II. STUDENT PROFILING AND COMMON SEQUENCES IDENTIFICATION MODEL

Student profiles and common sequences are determined according to web navigation through the Virtual Intelligent Learning Platform [6], age and gender. The model is shown in Figure 1 [23].

The profiling model has several input data, coming from the platform database [23]:

– Web usage log: Any access performed by each student to each resource of the course material.
– Students' birth date and gender.
– Test attempts: Any attempt performed by each student to each test, including the grade.

The web usage log contains data of around 1.000 students who have taken the "Computer Programming" course in 2014 (around 20 groups per term) and have taken three tests (three tests are applied each term). Each test of a term (e.g. First test) has covered different topics, and the amount of data differs among tests (Table I), because of this, three independent profiling processes were performed, obtaining a set of profiles per test.

The data are pre-processed by removing students who have never taken a test, and multiple accounts of the same student are mixed into only one [23].

A. Data Aggregation

Its goal is to count the number of accesses, carried out by each student for each test taken. The input data of *Web usage log* were aggregated one row by student. By each student, all of his/her accesses are sorted and split according to test attempts dates, any access carried out between the first day of the academic term and the day when the first test was taken, was assigned to the first test, any access carried out between the day when the first test was taken and the day when the second test was taken, was assigned to the second test, and so forth. The values presented in Table II are calculated for each set of accesses. *Absence time* is the time between two sessions.

If a value could not be calculated because of lack of data, which implies a division by zero (e.g. calculate the average absence time with only one session), the null value was changed by the average of another students' values, of the same term.

B. Data Preparation

This process transforms each set of tables returned by *Data aggregation* process, of a given test (e.g. all tables of first test), into one unique minable table, which is then used by *Clustering* process to find student profiles.

For each table returned by *Data aggregation* process, all rows are normalized using range transformation to [0.0 1.0], then a density based algorithm [24] is used to identify a discard outliers, any point which is farther away than 0.7 compared with at least 70% of all other points, is marked as outlier and discarded. Then, the grades of test attempts of the remaining points (i.e. all points that were not considered as outliers) are curved by subtracting the mean from the grade of all terms, and finally, the fields are normalized again to [0.0 1.0], to avoid a concentration of points to few values, done by a outlier which was too far away. The students' gender may have the value of 0.0 (Female) or 0.3 (Male).

Once all tables have been processed, each set of tables of a given test (e.g. First test) are mixed into only one, by concatenating all tables of that set. As a result, a minable table for each test is obtained.
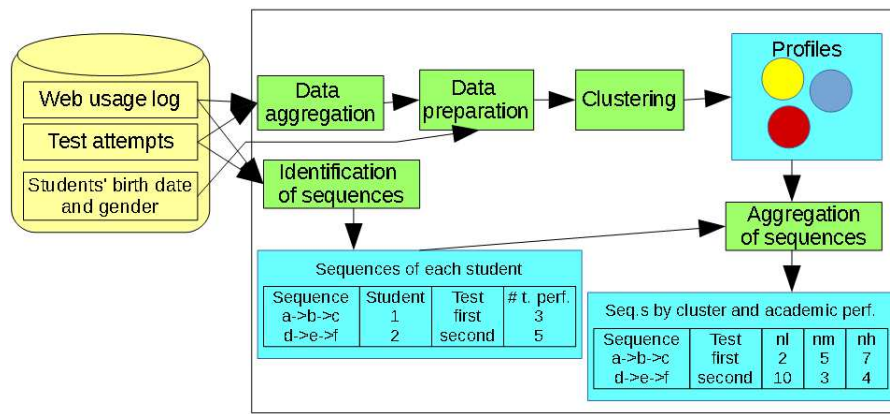
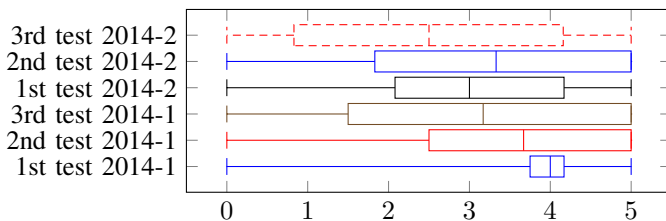Fig. 1. Profiling and common sequences identification model

Fig. 2. Box plot of grades gotten in attempts of each test

Fig. 3. Average centroid distance vs number of clusters

Once the set of clusters for each test was obtained, all the fields are discretized by frequency in order to improve the cluster descriptions, all the legend shown in Figures 4 to 6 were discretized in this way.

C. Clustering

For each table returned by *Data preparation* process, the K-Means algorithm (with Euclidean distance) is applied, using all fields presented in Table II, each cluster is a potential profile. In order to determine the k value, experiments varying it between 2 and 20, are executed 1000 times, each process has performed up to 10000 optimization steps.

The values of sl and ta were used in milliseconds, and ag in days, by the clustering algorithm, however, ag is shown in years, sl in minutes, and ta in hours, in the cluster descriptions.

D. Clustering Model Results

The profiling model was applied to the "Computer Programming" course in 2014, around 1000 students have taken the course, and performed around 20500 sessions. The grades use the numeric scale from 0.0 to 5.0, any grade equal to or greater than 3.0 is considered as a passing grade. The clustering model results are also presented in [23].

The Figure 2 shows the box plot of attempt grades per test. It was noticed a high concentration of grades in first test of term 2014-1, namely most students have gotten high grades regardless of their behavior, because of this, they were discarded.
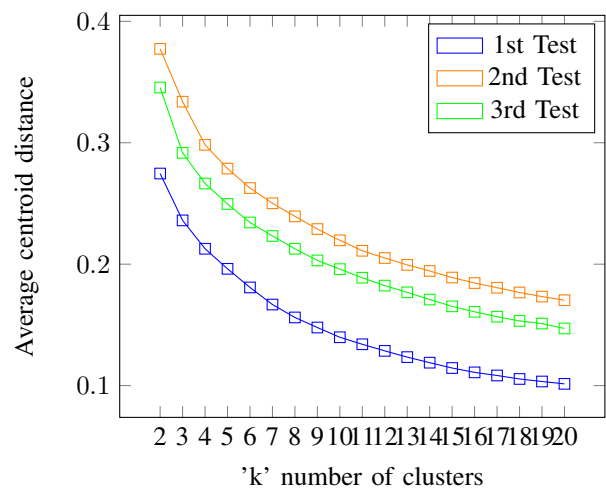
The Figure 3 shows the convergence of the k value, it indicates that the marginal gain tends to drop at 6, this value was assigned to k.

For the first test, 70 students were assigned to profile 0, 23 to profile 1, 73 to profile 2, 151 to profile 3, 74 to profile 4 and 140 to profile 0. For the second test, 246 students were assigned to profile 0, 33 to profile 1, 204 to profile 2, 257 to profile 3, 123 to profile 4 and 112 to profile 5. For the third test, 89 students were assigned to profile 0, 64 to profile 1, 137 to profile 2, 277 to profile 3, 74 to profile 4 and 187 to profile 5.

The clusters found are described in Figures 4 (First test), 5 (Second test), and 6 (Third test). The values of second and third tests are normalized because the profiles were identified with data coming from both terms.

Based on these descriptions of first test, a connection between the consultation volume of the course material and the academic performance was discovered, this is most noticeable in clusters 2 (many accesses and high academic performance) and 4 (few accesses and low academic performance). Another

clusters had a moderated number of accesses and a balanced academic performance. Regarding second and third tests, the academic performance behavior is more balanced among clusters. In second test, the academic performance is slightly better in cluster 3, it is characterized by a moderated number of accesses to all resource types, with preference for documents, which were accessed mostly out of class, namely access documents out of class may improve the academic performance. In third test, the academic performance is slightly better in cluster 2, it is characterized by a lot of accesses to documents, a moderated number of accesses to exercises and only few accesses to videos. Many accesses to documents were performed out of class. Regarding accesses to exercises and videos, the variables describe two behaviors, located at each extreme: a group contains students who have performed most of their accesses (at least 90%) in class, while students of the other group have performed most of their accesses (at least 90%) out of class.

It can be concluded that the platform is an useful tool for developing the course, because every test has a connection between the consultation volume and the academic performance, although the academic performance is more balanced in second and third tests and that connection is not always a direct connection, e.g. in second test the profile 1 has the highest number of accesses and the lowest academic performance.

E. Identification of Sequences

A numeric code was assigned to each unique resource (document, exercise or video). The resources accessed by one student in one session are sorted according to time, obtaining an array, e.g. "1 3 1 1 2" indicates that the student logged in, the he accessed the resource with code "1", then "3", and so forth, then he stopped consulting for at least 30 minutes. The *Identification of sequences* process identifies those sequences and counts the number of times that a given sequence was performed by the same person in order to study for a given test (e.g. The first test), that count is assigned to "# t. perf." (Number of times performed).

F. Aggregation of Sequences

Three grade groups are identified, all the grades gotten by all students of a given test (e.g. the first test) are discretized by frequency in three binds: nh (High), nm (Medium) and nl (Low) academic performance. The number of times that a sequence was performed by students who have gotten a low, medium or high academic performance were counted, and these values are assigned to nl, nm and nh respectively. The sequences with a nh value greater than that of nm and greater than that of nl are identified as of high academic performance, they are selected to choose which resources are going to be recommended.

III. WARNINGS AND RECOMMENDATION SYSTEM DESIGN

The main goal is to give timely feedback to the current student in order to improve his/her academic performance, by suggesting actions based on the profile assigned to him/her. The warnings and recommendation system design is shown in Figure 7.

This system uses the same data that were used by the clustering and common sequences identification model, also the clean data (used to identify the profiles), the profiles and the common sequences (with their nh,nm and nl values), which were stored in the knowledge base.

The warnings system is executed in the background through a scheduled task, which is comprised of several processes shown in Figure 7, this task is responsible for assigning a profile to each active student, based on recent behavioral and socio-economic data. When a student logs in and start consulting, the platform queries the most recent profile assigned, and based on it, the platform may show a warning.

The first thing that must be considered, is that the warnings system works with incomplete data, i.e., the current student still has enough time to carry out more accesses, while the data of the knowledge base were built with complete lapses, because the *Profiles in knowledge base* were filled immediately after the historic students have taken each test. Because of this, the behavioral data of active students must be prepared to allow to compare them with those of historic students.

A. Query Accesses of Previous Sessions

It processes all the active students, one at a time. For each one, it checks if the student has taken all tests with reached deadlines, e.g. define if the student did not take the second test, but now it is too late for taking that, in such case the student has dropped out the course, otherwise the date of the last test attempt indicates the beginning date of the lapse of study for the next scheduled test, all the accesses carried out since that date, were performed in order to study for that future test, they are the data used for assigning a profile to the current student and warn him/her if necessary.

B. Data Aggregation

In a similar way (but using recent activity data of active students) to the *Data aggregation* process of the profiling model, the data returned by the *Query accesses of previous sessions* process is aggregated into one row per student, by calculating the features presented in Table II.

C. Data Preparation

This process prepares the data to allow them be comparable with the data stored in *Profiles in knowledge base*, namely the data must be of equal weight for each field, and the outliers must be discarded. The *Data aggregation* process returns three tables (one per test), for each one, the *Data preparation* process normalizes them using range transformation to [0.0
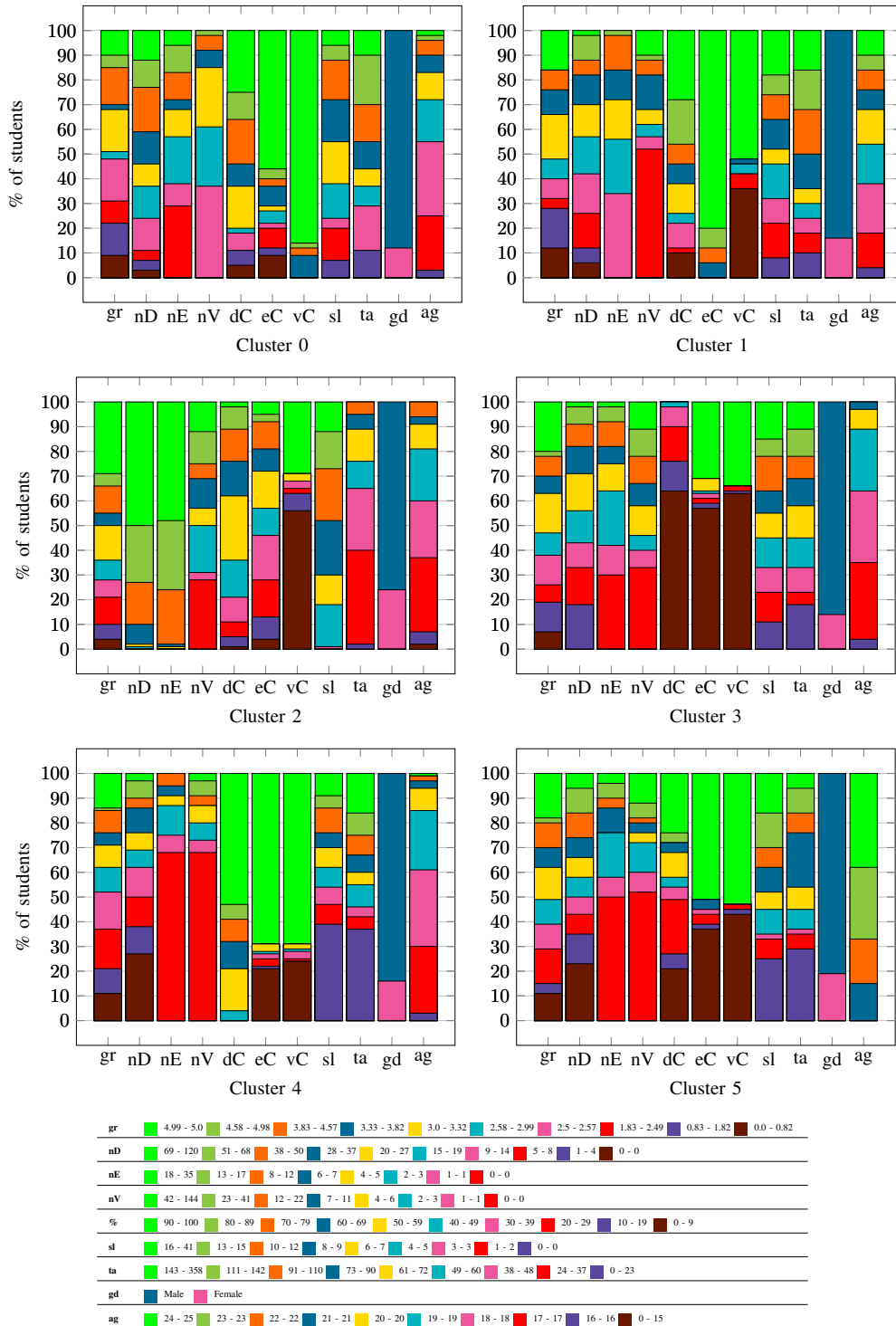
Fig. 4. Variable descriptions of clusters of the first test in 2014-2
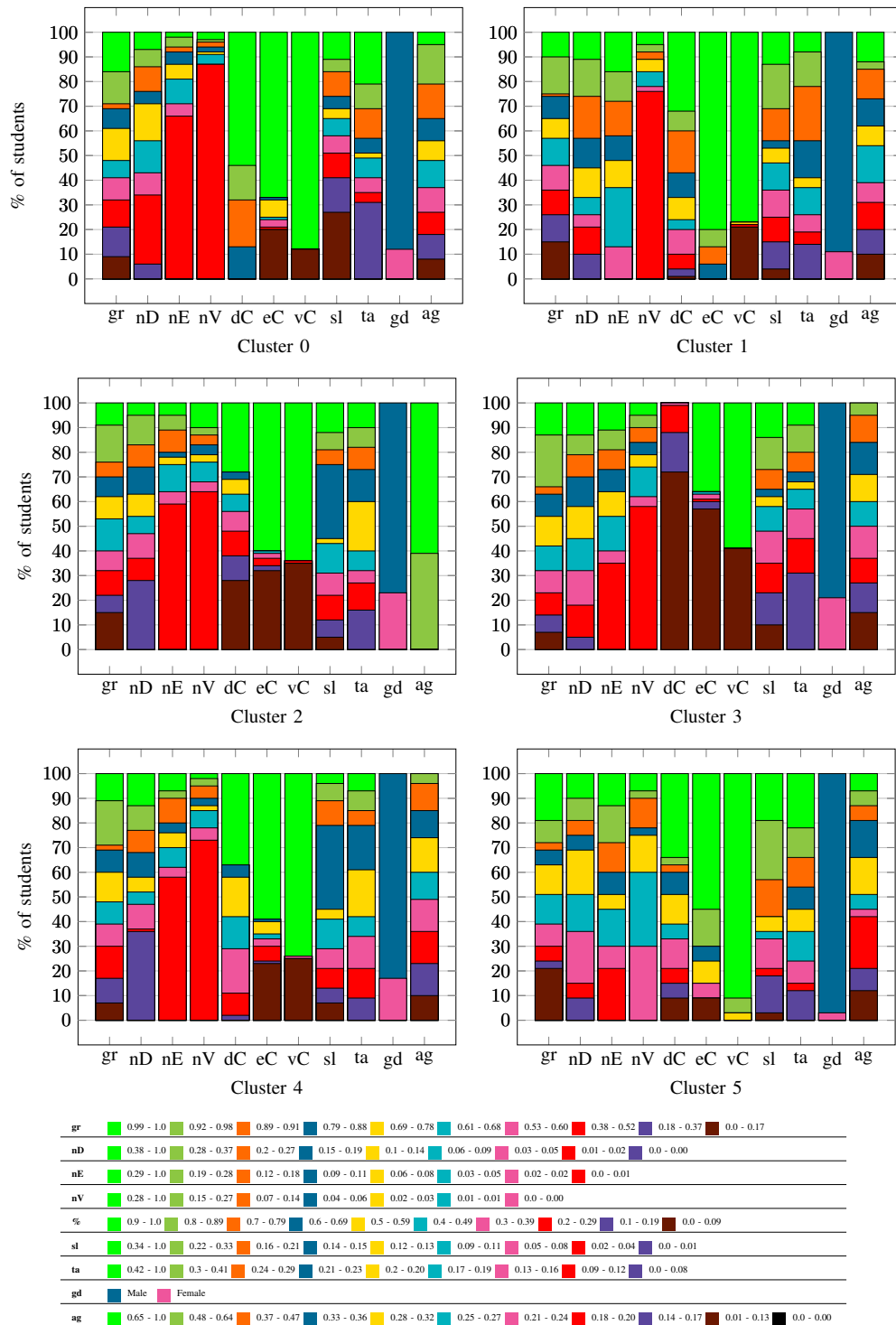
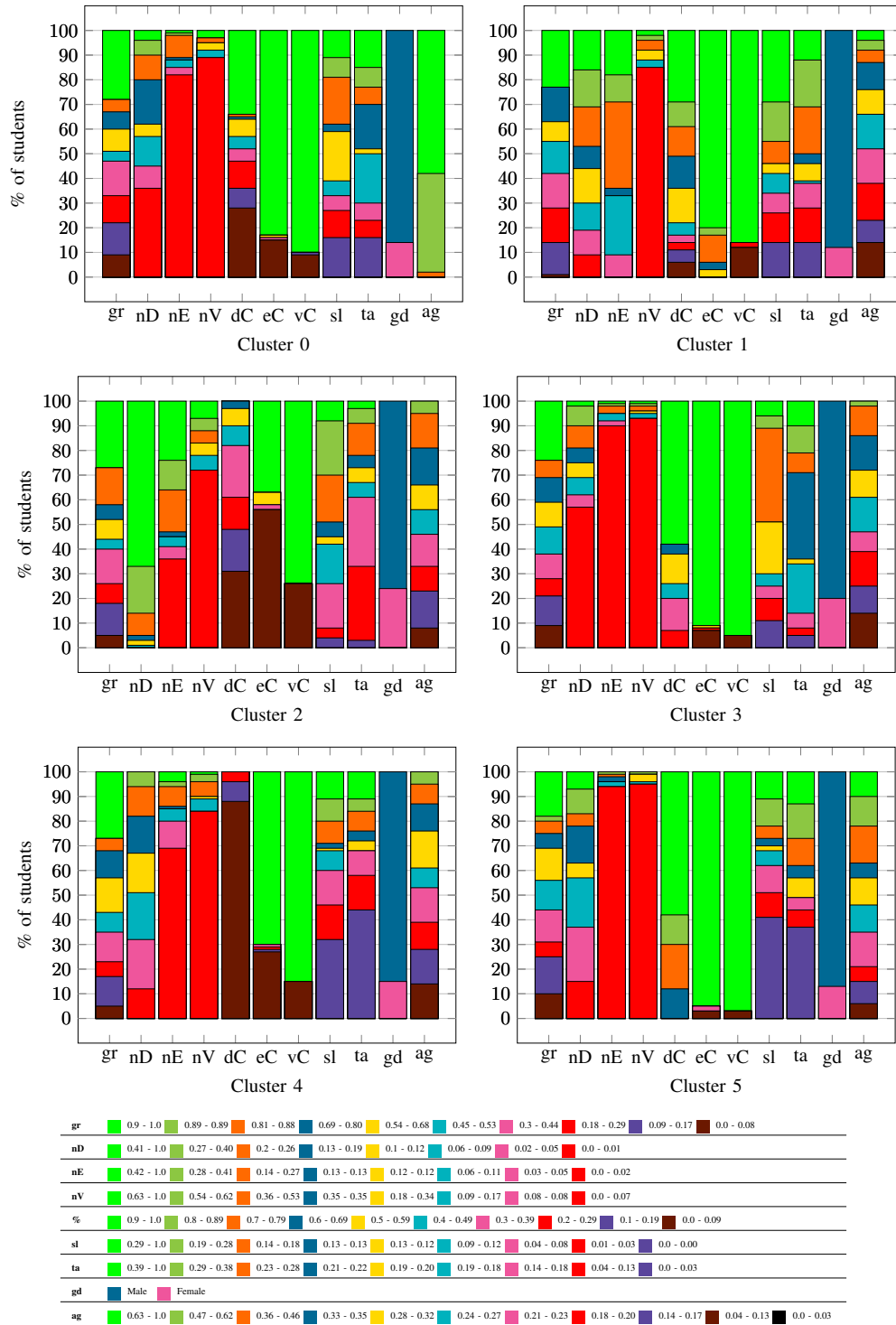Fig. 5. Variable descriptions of clusters of the second test

38

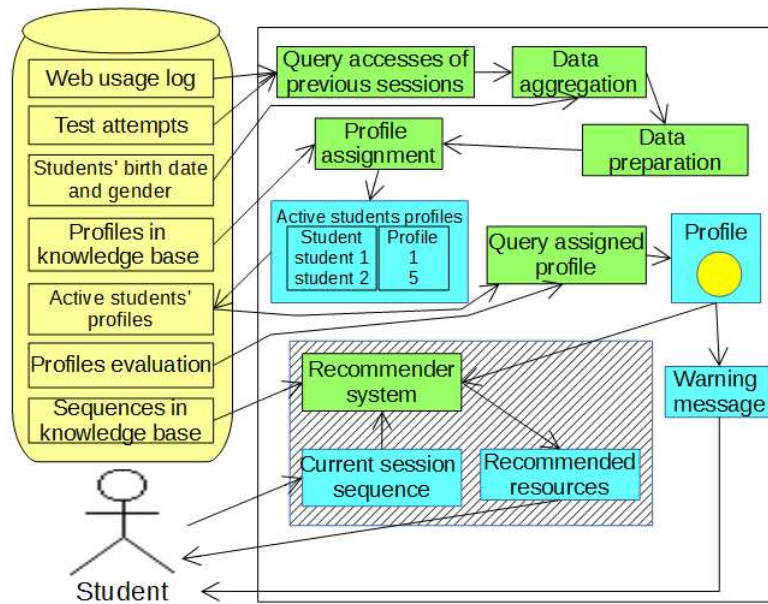Fig. 6. Variable descriptions of clusters of the third test

Camilo Peñuela, Elizabeth León, Jonatan Gómez

Fig. 7. Warnings and recommendation system

TABLE III
ACADEMIC PERFORMANCE AND WARNING MESSAGE ASSIGNED TO
CLUSTER

Warning message	Cluster	Test
You are encouraged to make a stronger use of the course material	3,4	First
You are encouraged to watch more videos	1	Second
You are encouraged to access more exercises and videos	2	Second

1.0], then a density based algorithm was run over the table to identify and discard any row that is farther away than 0.7 to at least 70% of all other rows, after calculating the euclidean distance with the fields: nD, nE, nV, sl and ta. All rows that were not discarded, are normalized using range transformation to [0.0 1.0]. This process is run to avoid concentrations of non discarded rows on few values, due to few outliers that were far away.

D. Profile Assignment

This process, assigns a profile previously identified by the *Clustering* process of the profiling model, to each active student, by processing each minable view returned by *Data preparation* process (it returns one view per test), a row at a time. This process, uses a K-Nearest-Neighbor classifier to assign a profile to the active student, the label is the profile, the labeled data are stored in *Profiles in knowledge base*, and k was set to the 10% of the number of historic students who have taken the test that the current student is going to take. The assignment is stored in *Active students' profiles* overwriting the row registered by the past scheduled task.

E. Giving Feedback to Current Student

Each time the student logs in and watches the topic list (after the selection of the course), the platform queries the last assigned profile by the *Profile assignment* process, and checks if that profile has a warning message, if so, that message is shown in the platform. The process to be run is *Query assigned profile*, it queries an assignment of a cluster to the current student in *Active students' profiles* table, if a profile is found, the process queries a warning message assigned to this profile, in *Profiles evaluation* table, if that message is found, it is shown in the platform. The *Recommender System* process identifies sequences registered on the *Sequences in knowledge base* which are similar to the one performed by the current student in his/her session so far. Those sequences were found for the test that the current student is going to take, and for each sequence, its *nh* value is greater than that of *nm* and than that of *nl*. Those sequences which contain resources not yet accessed by the current student, are selected to choose which resources are going to be recommended.

IV. RESULTS

The warning messages system is aimed to students who are likely to have a low academic performance. Based on the descriptions shown in Figures 4 (First test), 5 (Second test), and 6 (Third test), the profiles characterized by a low academic performance are: 3 and 4 of first test, 60% and 70% of students failed the test, respectively. 1 and 2 of second test, around 60% of students failed the test in both cases. The warning messages assigned to these profiles, are presented in Table III.

A lot of common sequences were identified, consider a student who is going to take the second test, he was assigned to profile 2, he logs in and perform the sequence

TABLE IV
RESOURCE CODES OF CURRENT SESSION SEQUENCE

Code	Resource	Topic
082	Flows exercises	Flows
083	Flows chapter	Flows

TABLE V
RESOURCE CODES OF SEQUENCES REGISTERED ON KNOWLEDGE BASE

Code	Resource	Topic
060	Input and output flows and persistence. Video	Flows
072	Logic exercises	Logic
084	Recursion exercises	Recursive functions
085	Recursion chapter	Recursive functions
087	Loops chapter	Loops
102	Languages presentation	Languages

"082,083", the resources involved in this sequence are shown in Table IV. The platform queries the similar sequences registered on the knowledge base, which were performed in order to study for the second test, and whose nh value was greater than that of nm and than that of nl, for the profile 2. They are "082,083,060,082,083,082,083" and "082,083,083,060,060,085,084,102,072,085,085,087". The resources involved in those sequences are shown in Table V, they are the resources chosen for recommendation.

V. CONCLUSIONS AND FUTURE WORK

A warnings system was proposed, it is aimed to give timely feedback to students who are likely to have a poor academic performance, it uses recent activity data of active students (who are currently taking the course), profiles and normalized data of historic students (who have taken the course in past terms) to assign a profile to the active student, and based on it, show a warning message if necessary.

The profiling model uses students' age and gender and behavioral data (number of accesses to each resource type, percentage of accesses in class, average session length and average absence time), then, it was possible to analyze the clusters in order to notice a connection with the academic performance and assign a warning message to a profile if necessary. The model finds the groups of similar students using the k-means algorithm (to choose the appropriated *k-Value* the average centroid distance was used). Profiles were identified and related to academic performance, it was noticed that some behaviors in navigation through the resources are reflected in the grades of the test attempts.

A recommender system is also proposed, it uses the common sequences found for historic students, and their *nh,nm* and *nl* values, to identify the sequences connected to a high academic performance, and select them to choose which resources are going to be recommended.

As future work, it is necessary to get feedback from students about warning messages and recommendations, to validate the system, however this validation takes at least one term, to obtain a complete set of students' perceptions.

Most common sequences were short, up to three accesses, it was due to the low granularity of topics covered by each document and exercises file, each one covers a whole chapter (e.g. Flows). As future work, the course material could be split into atomic topics, this action will allow identify more specific sequence patterns.

The Virtual Intelligent Learning Platform can be used to support more courses, the profiling and common sequences identification model can be applied to those courses, and their students would receive warning messages and recommendations.

The profiling model could be enriched with more data coming from another systems (e.g. the Academic Information System, which contains more academic data, or the welfare division, which contains more socio economic data), those data will be used by the profiling model as they become available.

REFERENCES

[1] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," *European Journal of Scientific Research*, vol. 43, pp. 24–29, 2010.

[2] D. R. Garrison, T. Anderson, and R. Garrison, *E-Learning in the 21st Century: A Framework for Research and Practice*, 1st ed. New York, NY, 10001: Routledge, 2003.

[3] X. Zhao, "Adaptive content delivery based on contextual and situational model," Ph.D. dissertation, The University of Electro-Communications, Tokyo, Japan, 2010.

[4] G. Kearsley, *Online education: Learning and teaching in Cyberspace.* Wadsworth Publishing Company, 2000.

[5] G. Salmon, *E-Tivities: The Key to Active Online Learning.* Routledge, 2002.

[6] J. Goméz, E. León, A. Rodriguez, E. C. Cubides, J. Mahecha, J. C. Rubiano, and W. Prado, "A didactic e-learning platform with open content navigation and adaptive exercises," in *2012 International Conference on Education and e-Learning Innovations (ICEELI)*. IEEE, 2012, pp. 1–6.

[7] V. M. García-barrios, F. Mödritscher, and C. Gütl, "Personalisation versus adaptation? a user-centred model approach and its application," 2005.

[8] A. K. Hamada, M. Z. Rashad, and M. G. Darwesh, "Behavior analysis in a learning environment to identify the suitable learning style," *International Journal of Computer Science and Information Technology*, vol. 3, no. 2, pp. 48–59, apr 2011. [Online]. Available: http://dx.doi.org/10.5121/ijcsit.2011.3204

[9] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, no. 1, pp. 368–384, 2008. [Online]. Available: http://dblp.uni-trier.de/db/journals/ce/ce51.html\#RomeroVG08

[10] C. Mencar, C. Castiello, and A. M. Fanelli, "A profile modelling approach for e-learning systems," in *ICCSA (2)*, ser. Lecture Notes in Computer Science, O. Gervasi, B. Murgante, A. Laganà, D. Taniar, Y. Mun, and M. L. Gavrilova, Eds., vol. 5073. Springer, 2008, pp. 275–290. [Online]. Available: http://dblp.uni-trier.de/db/conf/iccsa/iccsa2008-2.html\#MencarCF08

[11] J. Quevedo, E. M. nés, J. Ranilla, and A. Bahamonde, "Automatic choice of topics for seminars by clustering students according to their profile," *International Journal of Social, Behavioral, Educational, Economic and Management Engineering*, vol. 3, no. 9, pp. 473–477, 2009. [Online]. Available: http://waset.org/Publications?p=30

[12] V. P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C.-A. Comes, "Determining students' academic failure profile founded on data mining methods," *ITI 30th Int. Conf. on Information Technology Interfaces*, 2008.

[13] C. López, "Data mining model to predict academic performance at the universidad nacional de colombia," Master's thesis, Universidad Nacional de Colombia, 2013.

[14] J. P. Vandamme, N. Meskens, and J.-F. Superby, "Predicting academic performance by data mining methods," *Education Economics*, vol. 15, no. 4, pp. 405–419, 2007. [Online]. Available: http://EconPapers.repec.org/RePEc:taf:edecon:v:15:y:2007:i:4:p:405-419

[15] E. Y. Fethi A. Inan and M. M. Grant, "Profiling potential dropout students by individual characteristics in an online certificate program," *Int'l J of Instructional Media*, vol. 36, 2009.

[16] B. A. Chansarkar and A. Michaeloudis, "Student profiles and factors affecting performance," *Int. J. Math. Educ. Sci. Technol*, vol. 32, pp. 97–104, 2001.

[17] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A clustering methodology of web log data for learning management systems," *Educational Technology & Society*, vol. 15, pp. 154–167, 2011.

[18] I. K. Nagy and C. Gaspar-Papanek, *User Behaviour Analysis Based on Time Spent on Web Pages*. Springer-Verlag Berlin Heidelberg, 2009.

[19] H. Liu and V. Kešelj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data Knowl. Eng.*, vol. 61, no. 2, pp. 304–330, May 2007. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2006.06.001

[20] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.

[21] M. Kantardzic, *Data Mining: Concepts, Models, Methods and Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[22] J. L. Ortega Priego and I. F. Aguillo Caño, "Minería del uso de webs: Web usage data mining," *El Profesional de la Información*, vol. 18, no. 1, pp. 20–26, 2009.

[23] C. Penuela, "Student profiling model for the "Computer Programing" course," Master's thesis, Universidad Nacional de Colombia, 2015.

[24] R. I, *Rapid Miner Operator Reference*, 2014. [Online]. Available: http://docs.rapidminer.com/studio/operators/

Bi-variate Wavelet Autoregressive Model for Multi-step-ahead Forecasting of Fish Catches

Nibaldo Rodriguez and Lida Barba

Abstract—This paper proposes a hybrid multi-step-ahead forecasting model based on two stages to improve monthly pelagic fish-catch time-series modeling. In the first stage, the stationary wavelet transform is used to separate the raw time series into a high frequency (HF) component and a low frequency (LF) component, whereas the periodicities of each time series is obtained by using the Fourier power spectrum. In the second stage, both the HF and LF components are the inputs into a bi-variate autoregressive model to predict the original time series. We demonstrate the utility of the proposed forecasting model on monthly sardines catches time-series of the coastal zone of Chile for periods from January 1949 to December 2011. Empirical results obtained for 12-month ahead forecasting showed the effectiveness of the proposed hybrid forecasting strategy.

Index Terms—Wavelet analysis, bi-variate regression, forecasting model.

I. Introduction

MULTI-STEP-AHEAD forecasting of pelagic species time series is one of the main goals of the fishery industry and the government. To the best of our knowledge, very publications exist on one-step-ahead forecasting models for fisheries time series based on both autoregressive integrated moving average (ARIMA) models [1], [2] and multilayer perceptron (MLP) neural network models [3], [4]. On the one hand, the disadvantage of models based on linear regression is the supposition of stationarity of the fishes catches time series. However, the fisheries time series are non-stationary due to climatic fluctuations. On the other hand, although MLP neural networks allow modeling the non-linear behavior of a time series, they also have some disadvantages such as slow convergence speed and the stagnancy of local minima due to the steepest descent learning method. To improve the convergence speed and forecasting precision of anchovy catches off northern Chile, Gutierrez [3] proposed a hybrid model based on a MLP neural network combined with an ARIMA model, whose model gave an explained variance of 87%.

In this paper, a multi-step-ahead forecasting model of monthly fishes catches is proposed to achieve a more

accurate model than a MLP neural network model. Our proposed forecasting model is based on two phase. In the first phase, the haar stationary wavelet transform (SWT) is used to extract a high frequency (HF) component of intra-annual periodicity and a low frequency (LF) component of inter-annual periodicity. The wavelet decomposition was selected due to its popularity in hydrological [5], [6], electricity market [7], financial market [8] and smoothing methods [9], [10], [11]. In the second stage, both the HF and LF components are the inputs into a bi-variate autoregressive (BAR) model to predict the original time series. Besides, the proposed BAR model is compared with a MLP neural network model with N_i input nodes, N_h hidden nodes and two output nodes.

This paper is organized as follows. In the next section, we present hybrid multi-step-ahead forecasting model. The simulation results are presented in Section 3 followed by conclusions in Section 4.

II. Proposed Multi-step-ahead Forecasting

In order to predict the future values of time series $x(n)$, we can separate the raw time series $x(n)$ into two components by using Haar SWT. The first extracted component x_H of the time series is characterized by fast dynamics, whereas the second component x_L is characterized by low dynamics. Therefore, in our forecasting model a time series is considered as a functional relationship of several past observations of the components x_L and x_H as follows:

$$\hat{x}(n + h) = f(x_L(n - m), x_H(n - m));$$

the h value represents forecasting horizon and $i = 1, 2, \ldots m$ denotes lagged values of both the LF and HF components. Besides, the functional relationship $f(\cdot)$ in this paper is estimated by using a BAR model and a MLP neural network model. The following three subsections present the SWT, BAR forecasting model and MLP forecasting model.

A. Stationary Wavelet Transform

Let $x(n)$ denote the value of a time series at time n, then $x(n)$ can be represented at multiple resolutions by decomposing the signal on a family of wavelets and scaling functions [9], [10], [11]. The approximation (scaled) signals are computed by projecting the original signal on a set of orthogonal scaling functions of the form:

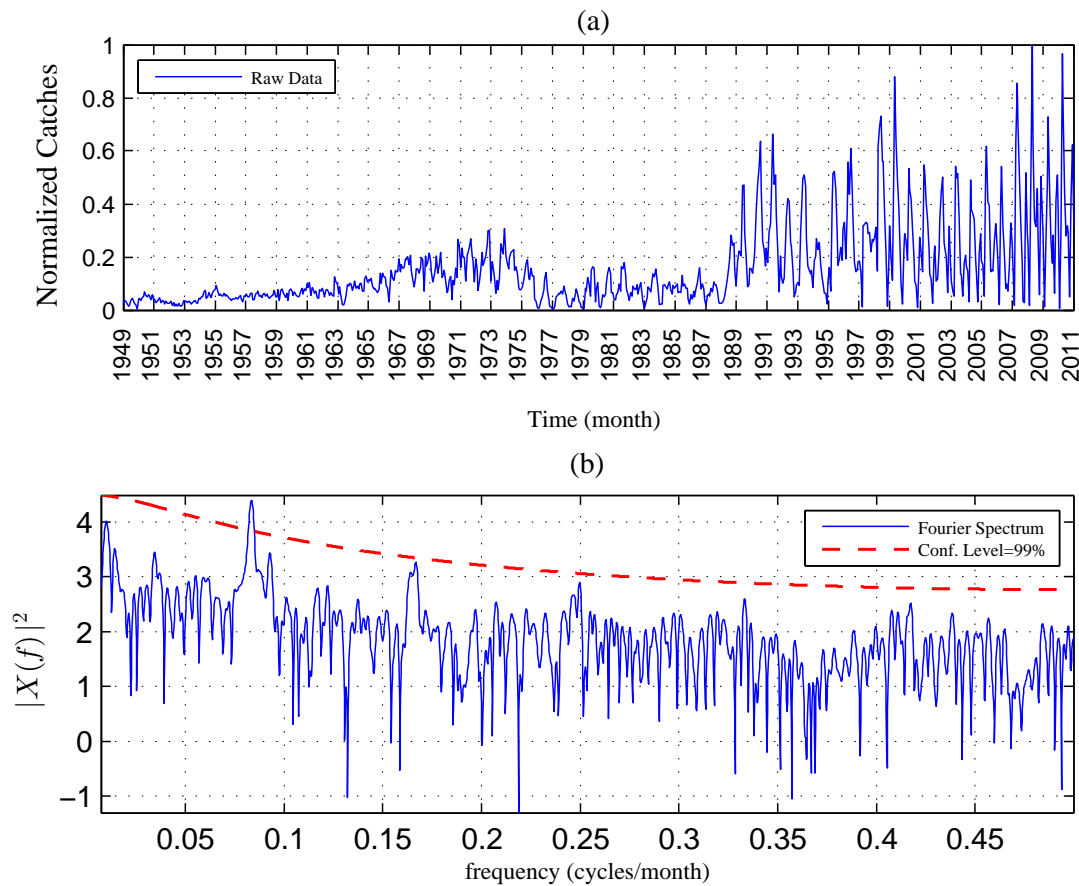$$\phi_{jk}(t) = \sqrt{2^{-j}}\phi(2^{-j}t - k),$$

Nibaldo Rodriguez and Lida Barba

Fig. 1. Monthly sardines catches

or equivalently by filtering the signal using a low pass filter of length r, $h = [h_1, h_2, ..., h_r]$, derived from the scaling functions. On the other hand, the detail signals are computed by projecting the signal on a set of wavelet basis functions of the form

$$\psi_{jk}(t) = \sqrt{2^{-j}}\psi(2^{-j}t - k),$$

or equivalently by filtering the signal using a high pass filter of length r, $g = [g_1, g_2, ..., g_r]$, derived from the wavelet basis functions. Finally, repeating the decomposing process on any scale J, the original signal can be represented as the sum of all detail coefficients and the last approximation coefficient. In time series analysis, discrete wavelet transform (DWT) often suffers from a lack of translation invariance. This problem can be tackled by means of the un-decimated stationary wavelet transform (SWT). The SWT is similar to the DWT in that the high-pass and low-pass filters are applied to the input signal at each level, but the output signal is never decimated. Instead, the filters are up-sampled at each level.

Consider the following discrete signal $x(n)$ of length N where $N = 2^J$ for some integer J. At the first level of SWT, the input signal $x(n)$ is convolved with the $h_1(n)$ filter to

obtain the approximation coefficients $a_1(n)$ and with the $g_1(n)$ filter to obtain the detail coefficients $d_1(n)$, so that:

$$a_1(n) = \sum_k h_1(n - k)x(k),$$
$$d_1(n) = \sum_k g_1(n - k)x(k),$$

because no sub-sampling is performed, $a_1(n)$ and $d_1(n)$ are of length N instead of $N/2$ as in the DWT case. At the next level of the SWT, $a_1(n)$ is split into two parts by using the same scheme, but with modified filters h_2 and g_2 obtained by dyadically up-sampling h_1 and g_1.

The general process of the SWT is continued recursively for $j = 1, ..., J$ and is given as:

$$a_{j+1}(n) = \sum_k h_{j+1}(n - k)a_j(k)$$
$$d_{j+1}(n) = \sum_k g_{j+1}(n - k)a_j(k)$$

where h_{j+1} and g_{j+1} are obtained by the up-sampling operator inserts a zero between every adjacent pair of elements of h_j
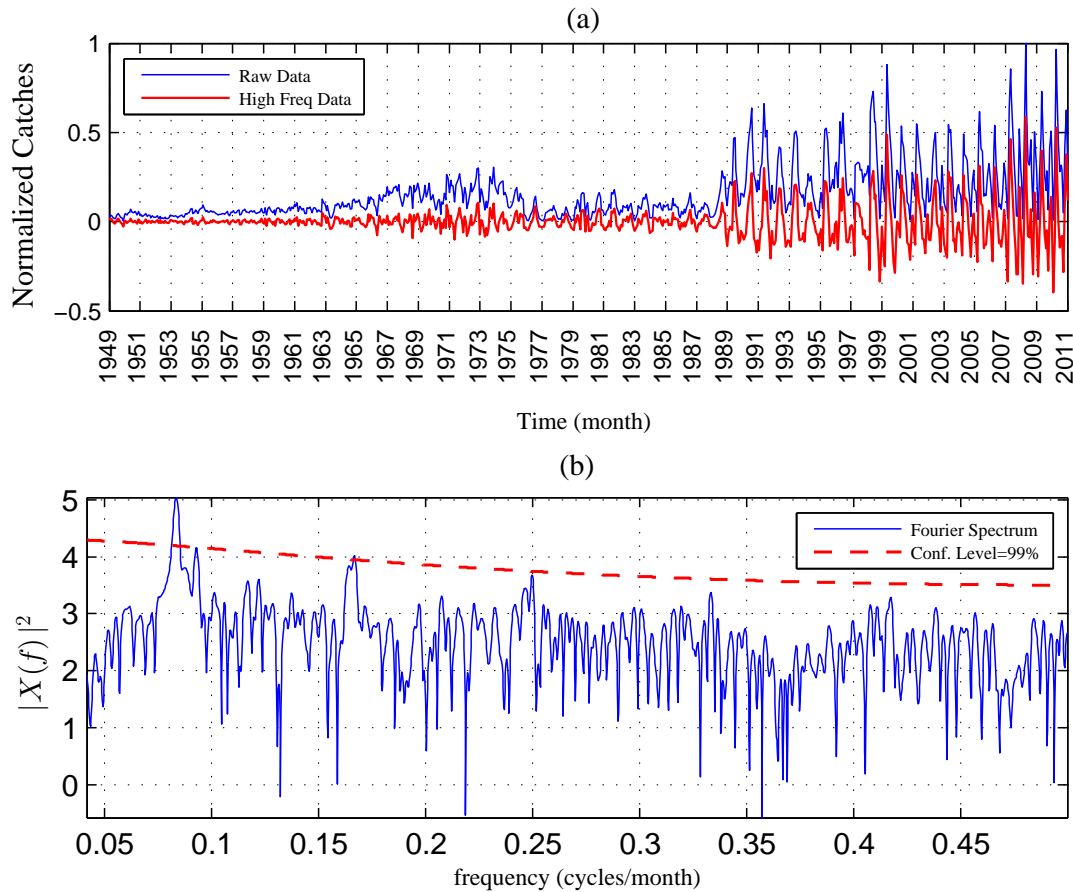
Fig. 2. High frequency sardines catches

and g_j; respectively. Therefore, the output of the SWT is then the approximation coefficients a_J and the detail coefficients $d_1, d_2, ..., d_J$. The wavelet decomposition method is fully defined by the choice of a pair of low and high pass filters and the number of decomposition steps J. Hence, in this study we choose a pair of Daubechies Db2 filters (has two wavelet and scaling coefficients) [12].

B. Bi-variate Forecasting Model

A bi-variate wavelet autoregressive (BWAR) model is used to estimate the function $\hat{f}(\cdot)$, which is given as

$$U = ZA,$$
$$u_{i,1} = X_H(i+h),$$
$$u_{i,2} = X_L(i+h),$$
$$z_{i,j} = X_H(i-j), j = 0, \ldots, m-1,$$
$$z_{i,m+j} = X_L(i-j), j = 0, \ldots, m-1,$$

where U is the matrix dependent variables of M rows by 2 columns, M is the set of input-output samples, Z is the regressor matrix of M rows by $2m$ columns and A is the parameters matrix of $2m$ rows by 2 columns. In order to

estimate the parameters A the linear least squares method is used, which is given as

$$A = Z^\dagger U;$$

$(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse [13].

C. Neural Network Forecasting Model

A single-hidden neural network with two output nodes is used to estimate the function $\hat{f}(\cdot)$, which is defined as

$$u_k(n) = \sum_{j=1}^{N_h} b_j \phi_j(z_i, v_j), k = 1, 2,$$
$$\hat{x}(n+h) = u_1(n) + u_2(n),$$

where N_h is the number of hidden nodes, $z = [z_1, z_2, \ldots z_{2m}]$ denotes the input regression vector containing $2m$ lagged values, $[b_1, \ldots b_{N_h}]$ represents the linear output parameters, $[v_{j,1}, v_{j,2}, \ldots v_{j,2m}]$ denotes the nonlinear parameters, and
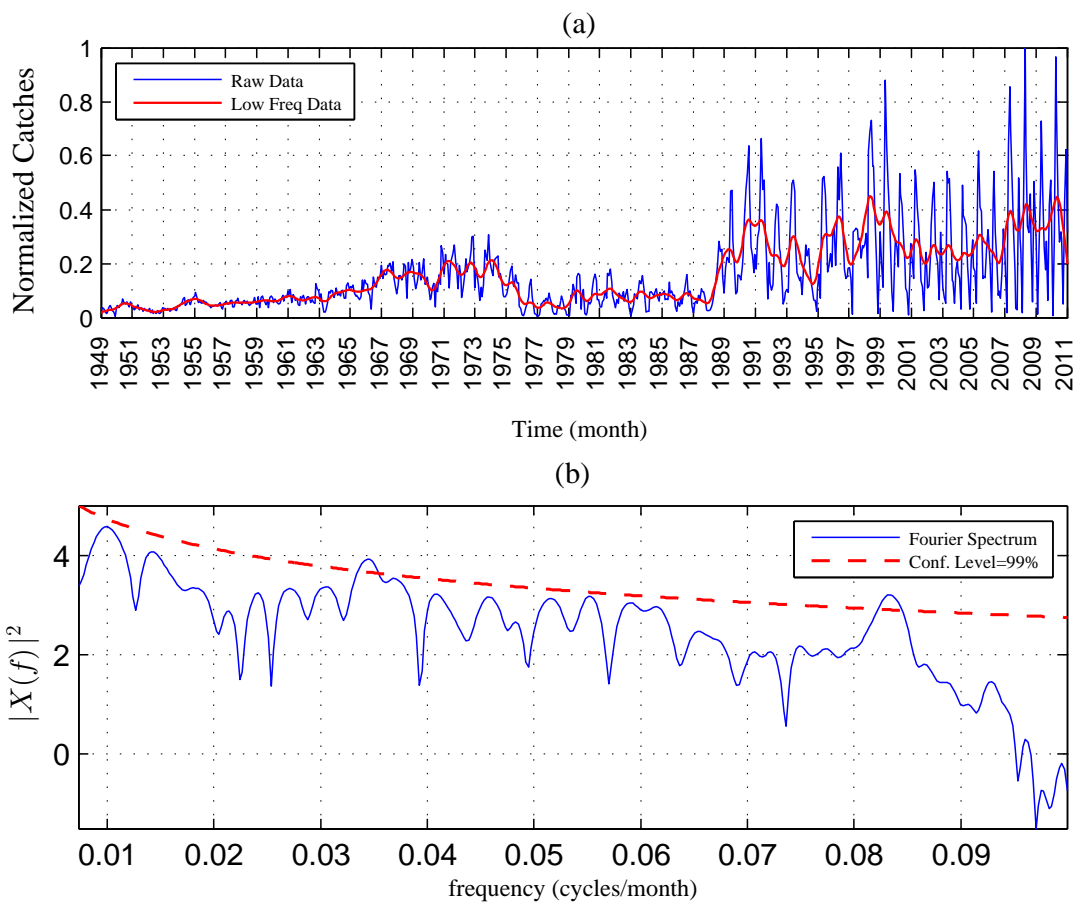
(a)

(b)

Fig. 3. Low frequency sardines catches

$\phi_j(\cdot)$ are hidden activation functions, which are derived as:

$$\phi_j(z_i) = \phi\Big(\sum_{i=1}^{2m} v_{j,i} z_i\Big),$$

$$\phi(z) = \frac{1}{1 + exp(-z)}.$$

In order to estimate both the linear and nonlinear parameters of the MLP, we use the Levenberg-Marquardt (LM) algorithm [14]. The LM algorithm adapts the $\theta = [b_1, \ldots b_{N_h}, v_{j,1}, \ldots v_{j,2m}]$ parameters of the neuro-forecaster minimizing mean square error, which is defined as:

$$E(\theta) = \frac{1}{2} \sum_{i=1}^{M} \Big(e(\theta_i)\Big)^2$$

Finally, the LM algorithm adapts the parameter θ according to the following equations:

$$\theta = \theta + \Delta\theta,$$

$$\Delta\theta = (\Upsilon\Upsilon^T + \mu I)^{-1} \Upsilon^T e,$$

where Υ represents the Jacobian matrix of the error vector evaluated in θ_i and the error vector $e(\theta_i) = x(n+h) - \hat{x}(n+$

$h)$ is the error of the MLP neural network for i pattern, I denotes the identity matrix and the parameter μ is increased or decreased at each step of the LM algorithm.

III. EXPERIMENTS AND RESULTS

In this section, we apply the proposed BWAR model for 12-month-ahead sardines catches forecasting. The data set used corresponded to landing of sardines in the south of Chile. These samples were collected monthly from 1 January 1949 to 31 December 2011 by the National Fishery Service of Chile (www.sernapesca.cl). The raw sardines data set have been normalized to the range from 0 to 1 by simply dividing the real value by the maximum of the appropriate set. On the other hand, the original data set was also divided into two subsets. In the first subset the 85% of the time series were chosen for the calibration phase (parameters estimation), whereas the remaining data set were used for the testing phase.

The normalized raw time series and the Fourier power spectrum are present in Figures 1(a) and 1(b); respectively. The red thick line in Figure 1(b) designates the confidence level against red noise spectrum. From Figure 1(b) it can be observed that there are one peaks of significant power, which
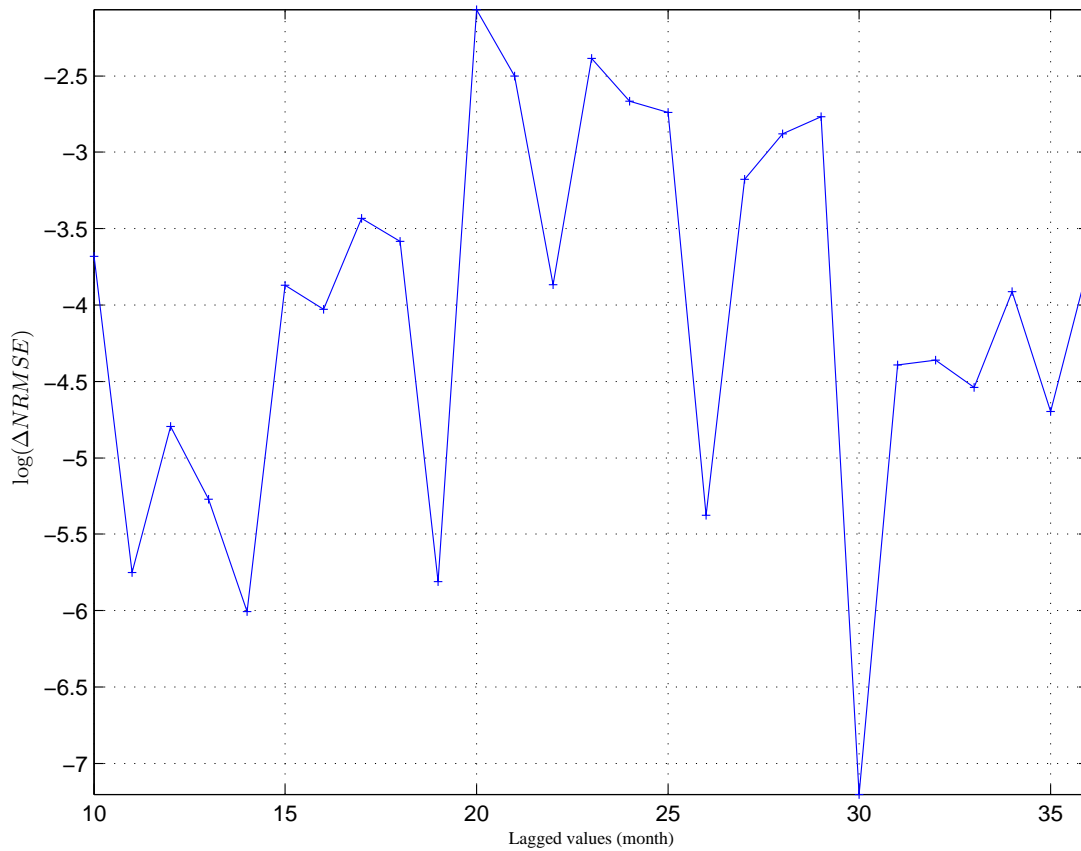
Fig. 4. NRMSE versus Lagged values

has an annual periodicities of 12 months ($freq = 0.083$). After we applied the Fourier power spectrum to the raw time series, we decided to use 3-level wavelet decomposition due to the significative peak of 12 months. Both the HF and LF components (times series) are presented in Figures 2(a) and 3(a); respectively, whereas the power spectrum of both time series are illustrated in Figure 2(b) and 3(b); respectively.

In this study, three criteria of forecasting accuracy called normalised root mean squares error (NRMSE), modified Nash-Sutcliffe efficiency coefficient (MNSE) and coefficient of determination (R2) were used to evaluate the forecasting capabilities of the proposed hybrid forecasting models, which are defined as

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{L}\left(x(i) - \hat{x}(i)\right)^2}{\sum_{i=1}^{L}\left(x(i) - \bar{x}\right)^2}}$$

$$MNSE = 1 - \frac{\sum_{i=1}^{L}\left|x(i) - \hat{x}(i)\right|}{\sum_{i=1}^{L}\left|x(i) - \bar{x}\right|}$$

$$R2 = 1 - \frac{\sum_{i=1}^{L}\left(x(i) - \hat{x}(i)\right)^2}{\sum_{i=1}^{L}\left(x(i) - \bar{x}\right)^2},$$

where $x(i)$ is the actual value at time i, $\hat{x}(i)$ is the forecasted value at time i, $\bar{x}$ is the mean of observed data and L is the number of forecasts.

Find the order of the bi-variate autoregressive model is a complex task, but here we will use the following metric to evaluate different lagged values, which is given as

$$\Delta(NRMSE) = NRMSE(m) - NRMSE(m - 1),$$

where m denotes values with m=2,...,36 months.

Figure 4 shows the results of testing data for lagged values between 10 and 36 months due to significant periods of the low frequency component, whose best result was achieved with $m = 30$ months, whereas Figures 5 show the results obtained with the best BWAR(30) forecasting model during the testing phase. Figure 5(a) provides data on observed monthly sardines catches versus forecasted catches; this forecasting behavior is very accurate for testing data with a NRMSE of 11% and a MNSE of 98%. On the other hand, from Figure 5(b) it can be observed a good fit to a linear curve with a coefficient of determination of 98%.

Once evaluated the BWAR(30) forecasting model perform calibration of the neural network with $N_i = 60$ input nodes, $N_h = \sqrt{N_i + N_o} = \sqrt{60 + 2} = 8$ hidden nodes and $N_o = 2$ output nodes. In the training process, overall weights were initialized by a Gaussian random process with
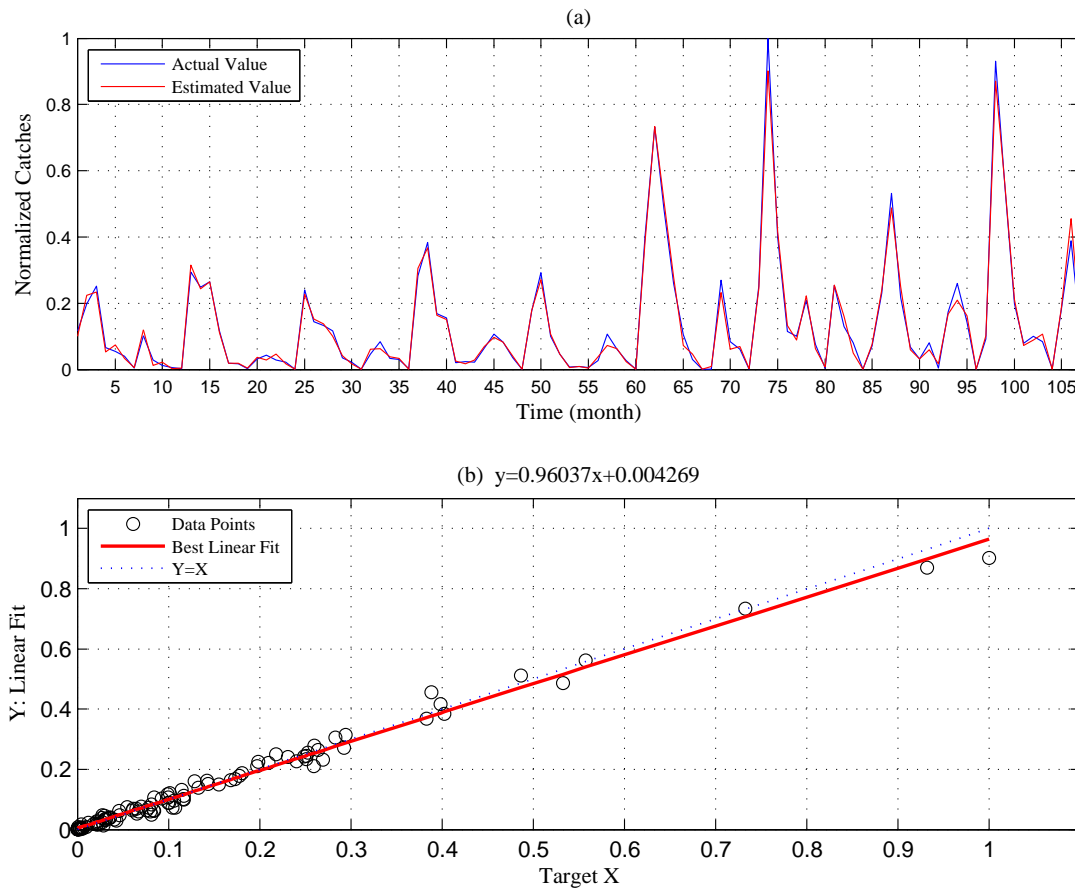
Nibaldo Rodriguez and Lida Barba

Fig. 5. Twelve-month-ahead BWAR forecasting for test data set

a normal distribution $N(0,1)$ and the stopping criterion was a maximum number of iterations set at 500. Due to the random initialization of the weights, we used 30 runs to find the best MLP neural network with a low prediction error. Figures 6(a) and 6(b) show the results obtained with the MLP(60,8,2) forecasting model during the testing phase. Figure 6(a) illustrates the observed data set versus forecasted data set, which obtains a NRMSE and a MNSE of 35% and 68%; respectively. On the other hand, Figure 6(b) shows the scatter curve between observed values and forecasted values with a R2 of 89%.

IV. CONCLUSIONS

In this paper was proposed a multi-step-ahead forecasting model to improve prediction accuracy based on Haar stationary wavelet decomposition combined with a bi-variate autoregressive model. The reason of the improvement in forecasting accuracy was due to use Daubechies SWT to separate both the LF and HF components of the raw time series, since the behavior of each component is more smoothing than raw data set. It was show that the proposed hybrid forecasting model achieves 11% and 98% of NRMSE and MNSE; respectively. Besides, the experimental results demonstrated a better performance of the proposed model when compared with a MLP neural network prediction model. Finally, hybrid forecasting model can be suitable as a very promising methodology to any other pelagic species.

REFERENCES

[1] S. KI., "Prediction of the mullidae fishery in the eastern mediterranean 24 months in advance," *Fisheries Research*, vol. 9, pp. 67–74, 1996.
[2] S. K.I. and C. E.D., "Modelling and forecasting annual fisheries catches: comparison of regression, univariate and multivariate time series methods," *Fisheries Research*, vol. 25, pp. 105–138, 1996.
[3] J. C. Gutierrez, S. C., Y. E., R. N., and P. I., "Monthly catch forecasting of anchovy engraulis ringens in the north area of chile: Nonlinear univariate approach," *Fisheries Research*, vol. 86, no. 188-200, 2007.
[4] S. P. Garcia, L. B. DeLancey, J. S. Almeida, and R. W. Chapman, "Ecoforecasting in real time for commercial fisheries: The Atlantic white shrimp as a case study," *Marine Biology*, vol. 152, pp. 15–24, 2007.
[5] J. F. Adamowski, "Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis," *Journal of Hydrology*, vol. 353, no. 3-4, pp. 247–266, 2008.

Fig. 6. Twelve-month-ahead WMLP forecasting for test data set

[6] O. Kisi, "Stream flow forecasting using neuro-wavelet technique," *Hydrological Processes*, vol. 22, no. 20, pp. 4142–4152, 2008.

[7] N. Amjady and F. Keyniaa, "Day ahead price forecasting of electricity markets by a mixed data model and hybrid forecast method," *International Journal of Electrical Power Energy Systems*, vol. 30, pp. 533–546, 2008.

[8] B.-L. Z., R. C., M. A. J., D. D., and B. F., "Multiresolution forecasting for futures trading using wavelet decompositions," *IEEE Trans. on neural networks*, vol. 12, no. 4, pp. 765–775, 2001.

[9] R. Coifman and D. L. Donoho, "Translation-invariant denoising, wavelets and statistics," *Springer Lecture Notes in Statistics*, vol. 103, pp. 125–150, 1995.

[10] G. Nason and B. Silverman, "The stationary wavelet transform and some statistical applications, wavelets and statistics," *Springer Lecture Notes in Statistics*, vol. 103, pp. 281–300, 1995.

[11] J.-C. Pesquet, H. Krim, and H. Carfantan, "Time-invariant orthonormal wavelet representations," *IEEE Trans. on Signal Processing*, vol. 44, no. 8, pp. 1964–1970, 1996.

[12] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge, England: Cambridge University Press, 2000.

[13] D. Serre, *Matrices: Theory and Applications*. Springer,NewYork, NY, USA, 2002.

[14] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE transactions on neural networks*, vol. 5, no. 6, pp. 989–993, 1996.

Location Privacy-Aware Nearest-Neighbor Query with Complex Cloaked Regions

Tran Khanh Dang and Chan Nam Ngo

Abstract—The development of location-based services has spread over many aspects of modern social life. This development brings not only conveniences to users' daily life but also great concerns about users' location privacy. In such services, location privacy aware query processing that handles cloaked regions is becoming an essential part in preserving user privacy. However, the state-of-the-art cloaked-region-based query processors only focus on handling rectangular regions, while lacking an efficient and scalable algorithm for other complex region shapes. Motivated by that problem, we introduce enhancements and additional components to the location privacy aware nearest-neighbor query processor that provides efficient processing of complex polygonal and circular cloaked regions, namely the Vertices Reduction Paradigm and the Group Execution Agent. We also provide a new tuning parameter to achieve trade-off between answer optimality and system scalability. Experiments show that our query processing algorithm outperforms previous works, in terms of processing time and system scalability.

Index Terms—Complex cloaked region, database security and integrity, group execution, location-based service, location privacy, nearest-neighbor query.

I. INTRODUCTION

TO PRESERVE the LBS user's location privacy, the most trivial method is to remove the direct private information such as identity (e.g., SSID). However, other private information, such as position and time, can also be used to violate the user's location privacy [1]. In preventing that, the researchers have recently introduced the Location Middleware [2]. It acts as a middle layer between the user and the LBS Provider to reduce the location information quality in the LBS request. The quality reduction is performed by the obfuscation algorithm which transforms the exact location to be more general (i.e., from a point to a set of points [3], a rectilinear region [4], [5], [6], [7], or a circular region [8], [9], etc.). The request is then sent to the LBS Provider to process without the provider knowing the user's exact location. Due to the reduction in location quality, the LBS Provider returns the result as a candidate set that contains the exact answer. Later, this candidate set can be filtered by the Location Middleware to receive the request's exact answer for the LBS user.

Consequently, to be able to process those requests, the LBS Provider's Query Processor must be able to deal with

the cloaked region rather than the exact location. In this paper, we propose a new Privacy Aware Nearest-Neighbor (NN) Query Processor that extends Casper* [10]. Our query processor can be embedded inside the untrusted location-based database server [10], [11], or plugged into an untrusted application middleware [2]. The query processor is completely independent of the location-based database server in the LBS Provider (as long as it supports basic functions such as range query) and underlying obfuscation algorithms in the Location Middleware (as long as the output is a cloaked region of n-gon shape). Moreover, it also supports various cloaked region shapes, which allows more than one single obfuscation algorithm to be employed in the Location Middleware [2]. In addition, we introduce a new tuning parameter to achieve trade-off between candidate set size and query processing time. Finally, we propose an additional component for the Location Middleware, the Group Execution Agent, to strongly enhance the whole system's scalability. The Group Execution Agent will group queries of the same filter parameters (POI type, keywords, etc.) that have adjacent or overlapped cloaked region into one to reduce the number of queries sent to and executed by the LBS Provider.

The contributions in this paper can be summarized as follows:

- We introduce an enhanced Privacy Aware Nearest-Neighbor Query Processor. With its Vertices Reduction Paradigm (VRP), complex polygonal and circular cloaked regions are handled efficiently in reasonable query processing time. In addition, the performance can be tuned through a new parameter to achieve trade-off between candidate set size and query processing time.
- We propose an additional component for the Location Middleware, the Group Execution Agent (GEA) and its Group Execution (GE) algorithm, to enhance the whole system's scalability. The basic idea of the GEA is that we group several queries of same filter parameters that have adjacent or overlapped cloaked region into one to reduce computational and communicational cost in the system.
- The applications of both Vertices Reduction Paradigm and Group Execution Agent are not limited to the scope of Privacy Aware Nearest-Neighbor Query. In fact, VRP is able to serve as enhancing component for any algorithms that need to process irregular shapes as general polygons and GEA can be applied to scale up systems that process multiple regions concurrently.

– We provide theoretical and experimental evidences to prove that our Privacy Aware Query Processor outperforms previous ones in terms of both processing time and system scalability.

The rest of the paper is organized as follows. In Section 2 and 3, we highlight the related works and briefly review the cloaked-region-based Casper* Privacy Aware Nearest-Neighbor Query Processor. Section 4 is dedicated for the underlying system architecture. In Section 5, we present two versions of the the proposed Vertices Reduction Paradigm. For the sake of clarity of the first version, we also provide proofs of accuracy and running examples. The Group Execution Agent is discussed in Section 6. Then we present our experimental evaluations in Section 7. Lastly, Section 8 will finalize the paper with conclusion and future works.

II. RELATED WORK

A. Location Privacy Preservation

To protect user privacy in LBS is to protect the user's location information, includes: identity, position and path. Based on this classification, researchers have proposed algorithms to preserve LBS user's privacy against attackers. The most trivial approach is to remove the information that directly reveals the personal and private information, such as identity (e.g., SSID), but other information, usually location and time, the so-called Quasi-Identifier [1] can also indirectly reveal the user identity. Thus, more complicated and advanced algorithms are introduced for more privacy protection. The algorithms can be classified into four categories as follows:

– Anonymity: to make user indistinguishable among $(k-1)$ other users for user identity protection [2], [12], [5], [7]. However, this category of algorithm has weaknesses where k does not necessarily determine the privacy of users [13]. In anonymous LBS (e.g., navigation), this category of algorithms aims for the protection of identity privacy.

– Obfuscation: to make user location imprecise or inaccurate or generalize the location into a region [8], [4], [3], [6], [14], [15]. Generally, this process is done by the means of perturbation, adding dummies, reducing precision or location hiding. In LBS that require user identity (e.g., paid LBS), this category of algorithms aims at position privacy.

– Transformation: to map user location into another location and repeatedly issue queries to process proximity services [16].

– Encryption: this class of algorithms is based on cryptography, the database server cannot know the data in the query and result [17].

Each privacy algorithm has its own characteristics and application domains. Hence, depending on what kind of objects needing to be protected, privacy algorithms should be chosen carefully and wisely to give the best protection to LBS users. In that manner, [2] proposes a framework that maximize
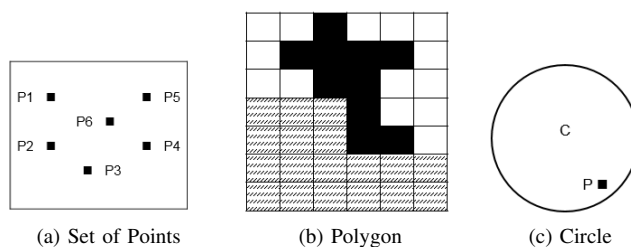

Fig. 1. Cloaked Region Classification

LBS users' privacy level by considering the user and service context factors. Among the algorithms above, obfuscation algorithms have obtained much interest because of their intuitive concepts and simple implementations. Obfuscation algorithms aim at decreasing LBS users' location quality to protect their privacy. To achieve that, from the exact LBS user location (position), the algorithms generate a more general cloaked region, of which the region shape can be very various, from a discrete set of points to a polygonal region or a circular region. Furthermore, the state-of-the-art algorithms deal with not only the geometry of the region, but also its geographic and semantic features. Recently, semantic aware obfuscation algorithm in the PROBE framework [4], or the database level obfuscation algorithm Bob-Tree [6], [15] also consider sensitive features and unreachable region area in the obfuscation process.

B. Cloaked Region Classification

As discussed in previous subsection, to protect the LBS user's location privacy, various obfuscation algorithms have been proposed in order to reduce the quality of the user's location. This is done by transforming the user's exact location as a point to a more general location as a region. In general, the cloaked regions' shapes can be categorized into 3 different types as follows:

– A set of points: as depicted in Figure 1a, this kind of cloaked regions is proposed by [3], in which the exact location p is transformed to a set of n points {p1, p2, p3, ..., pn} based on the concept of imprecision and inaccuracy. In case of inaccuracy, the set of points will not include the original exact location p.

– A polygonal region: this kind of cloaked regions is generated by most of the obfuscation algorithms. The exact location p will be generalized to a polygon represented by an ordered list of points {p1, p2, ..., pn, p1}, which are the vertices of the polygon's outer ring. The polygonal region can be as simple as a rectangular region [5], [14], [7] or as complex as an n vertices rectilinear region [4], [6], [15]. For example, in Figure 1b, the complex rectilinear cloaked region (solid black area) is generated by Bob-Tree [15], a database level and geographic aware cloaking algorithm, which considers some parts of the map are unreachable (hatched area)

such as lakes, mountains, etc. So the cloaked regions will not include the unreachable parts, leads to their indeterminate rectilinear shape.

– A circular region: as depicted in Figure 1c. In the works of [8], the authors propose various obfuscation algorithms to generalize an exact location p to a (c, r) circle, in which c is the center and r is the radius. The radius r then can be reduced, enlarged and the center c can be moved to achieve better obfuscation for LBS user. Or in the works of [9], the authors also provide an obfuscation algorithm to generate a circular cloaked region from the LBS user's exact location with a random method.

C. Location Privacy Aware Nearest-Neighbor Query Processor

To support the privacy aware queries, recent research efforts have been dedicated to deal with cloaked area processing. The location privacy aware query processor is either embedded directly into an untrusted location-based database server or application middleware to process queries from LBS users or a trusted third party (e.g., the Location Middleware). In addition, to reinforce a higher level of user location privacy, the query processor must support four basic types of query [10] as follows:

– Public Query over Public Data.
– Private Query over Public Data.
– Public Query over Private Data.
– Private Query over Private Data.

In these circumstances, public query or data refers to objects or their information that everyone knows or are willing to be revealed (e.g., the position of POIs, the position of patrol police car, ambulance or public services) while private query or data mentions those whose access is limited (e.g., the user's exact position or private places).

Recently, [3] have proposed an obfuscation algorithm that transforms an exact user location to a set of points in a road network based on the concepts of inaccuracy and imprecision. They also provide a nearest-neighbor query processing algorithm. The idea is that the user will first send the whole set of points to the server, the server will send back a candidate set of nearest-neighbors. Based on that candidate set, the user can either choose to reveal more information in the next request for more accurate result or terminate the process if satisfied with the candidate set of nearest-neighbors. The other works by [11] and [9] respectively propose algorithms to deal with circular and rectilinear cloaked regions. Those works find the minimal set of nearest-neighbors based on continuous nearest-neighbor search [18]. In a different approach, Casper* only computes a superset of the minimal set of nearest-neighbors that contains the exact nearest-neighbor, in order to achieve trade-off between query processing time and candidate set size for system scalability [10]. In addition, Casper* also supports two more query types: Private Query over Public Data and Public Query over Private Data.

TABLE I
CASPER* QUERY PROCESSING ALGORITHM NOTATIONS

Cloaked region A. Vertex v ($v_1, v_2, \ldots, v_i, \ldots v_n$).		
Public	Filter (NN) of v_i	t_i
Object	Distance $v_i t_i$	$dist(v_i, t_i)$
Private	Filter (NN) of v_i	At_i
Object	Distance $v_i At_i$	$min - max - dist(v_i, At_i)$

Among previous works, only Casper* supports Query over Private Data, while the others either only support Query over Public Data [3] or lack the trade-off for system scalability [11], [9]. Moreover, Casper* keeps the exact location of LBS users secret from the LBS Providers. However, Casper* is only efficient in dealing with rectangular regions. While it can handle polygonal cloaked regions, the application into these cases needs evaluations and modifications. Moreover, in case of systems like OPM [2], the Query Processor must have the ability to deal with various kinds of cloaked region because the system supports more than one single obfuscation algorithm. Motivated by those problems, our proposed Privacy Aware Nearest-Neighbor Query Processor offers the ability to efficiently handle complex polygonal and circular cloaked regions with its Vertices Reduction Paradigm and a new tuning parameter for system scalability. Furthermore, we provide an addition component for the Location Middleware, the Group Execution Agent, to strongly enhance the whole system's scalability. Simply saying, the GEA groups several queries that have adjacent or overlapped cloaked region into one to reduce computational and communicational cost in the system.

III. THE CASPER* PRIVACY AWARE NEAREST-NEIGHBOR QUERY PROCESSOR

In this section, let us briefly review the Casper* algorithm, start with its terms and notations (Table I). For each vertex v_i of the cloaked region A, its nearest-neighbor is called a $filter$, denoted as t_i if that nearest-neighbor is a public object (public nearest-neighbor) (Figure 2b, t_1, t_2 of v_1, v_2). In case the nearest-neighbor is private, it is denoted as At_i. A private object is considered as a private nearest-neighbor if it has the minimum distance from its cloaked region's furthest corner to v_i (Figure 2d, At_1). The distance between a vertex and its filter is denoted as $dist(v_i, t_i)$ ($= v_i t_i$) (Figure 2b, distance $v_1 t_1$ and $v_2 t_2$ in case of public nearest-neighbor) or $min - max - dist(v_i, At_i)$ ($= v_i At_i$) (Figure 2d, distance from v_1 to the furthest corner of At_1, in case of private nearest-neighbor). For each edge e_{ij} formed by adjacent vertices v_i, v_j, a $split - point$ s_{ij} is the intersection point of e_{ij} and the perpendicular bisector of the line segment $t_i t_j$ (Figure 2b, s_{12}). For the purpose of the Casper* Nearest-Neighbor Query Processing algorithm [10], given a cloaked region A, it is to find all the nearest-neighbors of all the points (1) inside A and (2) on its edges. The algorithm for Query over Public Data can be outlined in the three following steps below.

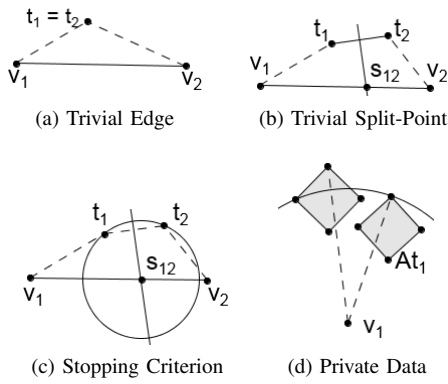STEP 1 (Filter Selection): We find the filters for all of cloaked region A's vertices.

(a) Trivial Edge   (b) Trivial Split-Point

(c) Stopping Criterion   (d) Private Data

Fig. 2. The Casper* algorithm

STEP 2 (Range Selection): For each edge e_{ij} of A, by comparing v_i, v_j's filters t_i and t_j, we consider four possibilities to find the candidate nearest-neighbors and range searches that contain the candidate nearest-neighbors:

- Trivial Edge Condition: If $t_i = t_j$ (Figure 2a, $t_1 = t_2$), t_i (t_j) is the nearest-neighbor of all the points on e_{ij}, so we add t_i (t_j) into the candidate set.
- Trivial Split-Point Condition: In this case, $t_i \neq t_j$, but split-point s_{ij} of e_{ij} takes t_i, t_j as its nearest-neighbors (Figure 2b). This means t_i and t_j are the nearest-neighbors of the all points on $v_i s_{ij}$ and $s_{ij} v_j$ respectively. So we add t_i, t_j into the candidate set.
- Recursive Refinement Condition: If two conditions above fail, we will consider to split the edge e_{ij} into $v_i s_{ij}$ and $s_{ij} v_j$, then we apply STEP 2 to them recursively. A parameter $refine$ is used to control the recursive calls for each edge. It can be adjusted between 0 and ∞ initially in the system. For each recursive call, refine will be decreased by 1, and when it reaches 0, we will stop processing that edge. In this case, $refine > 0$, we decrease it by 1 and process $v_i s_{ij}$ and $s_{ij} v_j$ recursively.
- Stopping Criterion Condition: When $refine$ reaches 0, we add the circle centered at s_{ij} of a radius $dist(s_{ij}, t_i)$ as a range query into the range queries set R and stop processing current edge (Figure 2c).

STEP 3 (Range Search): we execute all the range queries in R, and add the objects into the candidate set. As a result, the candidate set contains nearest-neighbors for all the points (1) inside cloaked region A and (2) on its edges. After that, the candidate set will be sent back to the Location Middleware to filter the exact nearest-neighbor for the LBS user.

For Query over Private Data, STEP 2 is similar to Query over Public Data, with some modifications. Instead of adding At_i directly into the candidate set, we will have to add a circle centered at v_i of a radius $min-max-dist(v_i, At_i)$ as a range query into the range queries set R (Fig. 2d, center v_1 and the radius of distance from v_1 to the furthest corner of At_1). The same behavior is applied to v_j and s_{ij} of edge e_{ij}.

IV. SYSTEM ARCHITECTURE

Our underlying system architecture consists of two main components, (1) the Location Middleware and (2) the LBS Provider, as depicted in Figure 3.

The Location Middleware manages users' exact positions and their privacy profiles. Upon joining the system, the user agrees to entrust the Location Middleware with the user's precise positions. To preserve the user's location privacy, the user will specify a privacy profile which indicates the desired privacy requirements in interacting with other untrusted components (e.g., the LBS Provider). This privacy profile may include some information such as k (k-anonymity), the minimum and the maximum area of the cloaked regions, or other parameters required by the privacy protection algorithm and the context (time, place, component, etc.) to activate the corresponding profile, etc. In our system, the Location Middleware, according to the privacy profile, will transform the user's precise position into a cloaked region. More than one spatial cloaking algorithm can be employed at the Location Middleware to provide the user with context aware spatial cloaking [2], which selects the best cloaking algorithm given current context of the user. When the user sends an LBS request to the Location Middleware, the request will be forwarded to the LBS Provider with the cloaked region as the input. The LBS Provider will calculate a candidate set of answers based on the cloaked region and send the set back to the Location Middleware. The Location Middleware will then filter the above set for the exact answer and send it to the user. Besides, our Location Middleware also include an additional component, the Group Execution Agent, which will group the LBS queries before sending them to the LBS Provider to strongly enhance the whole system's scalability.

The LBS Provider provides LBS query answers for the Location Middleware. Its main component is the Location Privacy Aware Query Processor. The processor accepts the cloaked region as input instead of the precise position. It will compute a candidate set of answers using the cloaked region and the spatial objects in the database. After that, the LBS Provider will send the set back to the Location Middleware. In this manner, the LBS Provider receives only the cloaked regions, not the exact positions of LBS users. Thus, even if the adversaries gain control over the LBS Provider, they still cannot compromise the LBS users' location privacy. Our Privacy Aware Query Processor is an extended version of the state-of-the-art location privacy aware query processor Casper*. Besides the ability to deal with the four privacy aware query types, it can also efficiently handle complex polygonal cloaked regions, by utilizing the Vertices Reduction Paradigm. Our processor also supports one more type of cloaked region, the circular cloaked region, by transforming it into a polygonal cloaked region. Moreover, the performance of the processor can be tuned through a new parameter, to achieve a trade-off in system scalability, in terms of both query processing time and query candidate set size.
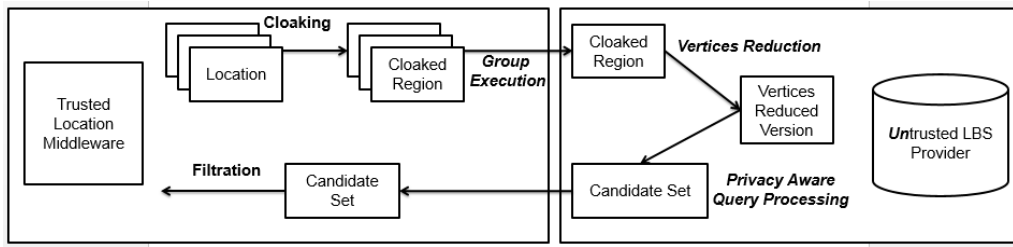
Fig. 3. System Architecture

TABLE II
QUERY PROCESSING ALGORITHM RUNTIME COMPLEXITY UTILIZING
VERTICES REDUCTION PARADIGM

Rectangle	$(4Qt_4 + 4(2^{refine} - 1)Qt_4)$	(1)
Polygon	$(nQt_n + n(2^{refine} - 1)Qt_n)$	(2)
VRP	$(mQt_m + m(2^{refine} - 1)Qt_m)$	(3)
# of vertices (4, m, n)	$4 \leq m \leq n$	
Query runtime (Qt_*)	$Qt_4 \leq Qt_m \leq Qt_n$	

In our system, only the Location Middleware and the users themselves are trusted with the exact location information. The LBS provider only receives cloaked regions, which will ensure the users' location privacy. Furthermore, to simplify the system architecture, we assume that the networks between parties are secured and all the pre-processes (including service and user profile registration, trade-offs between participating parties, pay-offs, validation, authentication, authorization, etc.) are already well prepared. It is also worth noting that our proposed enhancements do not affect (not improve and not worsen) the security and privacy aspect of the underlying privacy protection algorithms and query processing processor as we mainly focus on improving the performance and scalability of the system.

V. VERTICES REDUCTION PARADIGM

A. The Polygonal Cloaked Region Problem

As discussed previously, among the polygonal cloaked regions, there are cloaked regions that are complex, which means they have much more than 4 vertices. They can either (1) come directly from the semantic and geographic aware obfuscation algorithms, or (2) they are the group queries' regions that the Group Execution Agent of the Location Middleware sends to the LBS Provider.

For a rectangular cloaked region, the Casper* private query processing algorithm's runtime complexity can be calculated as follows. (1) STEP 1 needs to run exactly 4 nearest-neighbor queries to find 4 filters for the 4 vertices of the rectangular cloaked region, so the runtime is $4Qt_4$ (with Qt_4 is the time of a nearest-neighbor query). (2) In STEP 2, for each edge of the rectangular cloaked region, we need $(2^{refine} - 1)$ nearest-neighbor queries to find the candidates and range searches for STEP 3, so the runtime is $4(2^{refine} - 1)Qt_4$. Hence, the total algorithm runtime is $(4Qt_4 + 4(2^{refine} - 1)Qt_4)$ (worst case). Accordingly,

to process a polygonal cloaked region with n vertices, the algorithm's runtime complexity is $(nQt_n + n(2^{refine} - 1)Qt_n)$.

Although the Casper* Privacy Aware Query Processor can deal with polygonal cloaked region A that has n vertices (n-gon), its runtime significantly depends on A's number of vertices (STEP 1) and edges (STEP 2). As shown in Table II's formula 1 and 2, to process an n-gon, Casper* suffers from two aspects. (1) Under the same number of objects in the spatial database and the same access method, the processing time of STEP 1 increases. Because the algorithm has to find n filters for n vertices of the n-gon, it needs to run more nearest-neighbor queries ($4Qt_4 \leq nQt_n$). Besides, the calculation of $min - max - dist(v_i, At_i)$ in Query over Private Data also increases the nearest-neighbor query runtime for the n-gon because we have to process n vertices of the targets before we can determine their furthest corners ($Qt_4 \leq Qt_n$). (2) With the same refine value, the processing time of STEP 2 increases as it has to process more edges ($4(2^{refine} - 1)Qt_4 \leq n(2^{refine} - 1)Qt_n$).

B. The Vertices Reduction Paradigm

To ease the polygonal cloaked region problem, we introduce the Vertices Reduction Paradigm, in which we simplify the polygon so that it has as fewer vertices as possible before processing it with the Casper* algorithm. For that purpose, the Ramer-Douglas-Peucker (RDP) algorithm is employed [19], [20]. For each private object (n-gon) in the database, we maintain only a vertices reduced version (VRV, m-gon, $m \leq n$) of that private object. The vertices reduced version is generated by the Ramer-Douglas-Peucker algorithm and it will be stored inside the database until invalidated, e.g. the user gets a new original cloaked region. For nearest-neighbor query processing, we will use the vertices reduced versions instead of the original ones to reduce processing time ($m \leq n$ and $Qt_m \leq Qt_n$, as depicted in formula 3 of Table II). Accordingly, in the Vertices Reduction Paradigm, the query processing time is only $(mQt_m + m(2^{refine} - 1)Qt_m)$. By using the Vertices Reduction Paradigm, we can achieve at most (nQt_n/mQt_m) speedup in query processing time. In Query over Public Data, $Qt_n = Qt_m$ so the speedup is at most (n/m). In Query over Private Data, $Qt_n/Qt_m = n/m$, so the maximum speedup is $(n/m)^2$.

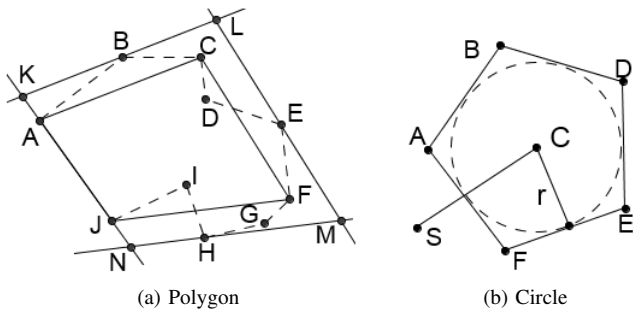The purpose of the Ramer-Douglas-Peucker algorithm, given an n-gon (ABCDEFGHIJ in Figure 4a), is to find a

(a) Polygon   (b) Circle

Fig. 4. The Polygonal and Circular Vertices Reduction Versions

(a) Trivial Condition $(+d)$ (b) $+\epsilon$ for Query over Private Data

(c) Range searches

Fig. 5. The Query Processing algorithm modifications

subset of fewer vertices from the n-gon's list of vertices. That subset of vertices forms an m-gon that is simpler but similar to the original n-gon $(m \leq n)$. The inputs are the n-gon's list of vertices and the distance dimension $\epsilon > 0$. First, we find the vertex that is furthest from the line segment with the first and last vertices as end points. If that furthest vertex is closer than ϵ to the line segment, any other vertices can be discarded. Otherwise, given the index of that vertex, we divide the list of vertices into two: [1..index] and [index..end]. The two lists are then processed with the algorithm recursively. The output is the m-gon's list of vertices (Figure 4a, $ACFJ$).

In next subsections, we will discuss two different approaches to utilize the vertices reduced versions. The first one is to use the vertices reduced version directly. The second one, which is better than the first, is to use the bounding polygon of the vertices reduced version. In both approaches, the Ramer-Douglas-Peucker algorithm's overhead in computing the vertices reduced versions is insignificant compared to the total processing time of the query. As depicted in Figure 4a, the dotted polygon $ABCDEFGHIJ$ is the original n-gon while $ACFJ$ and $KLMN$ is the m-gon of the first and second approach respectively. For a circular region, the vertices reduced version is its regular bounding polygon (Figure 4b, $ABDEF$) and we use the distance from another vertex to its center plus the radius as $min - max - dist$ of it and that vertex in private nearest-neighbor search ($SC + r$ in Figure 4b, where S is the processing vertex, C is the center of the processing private target object and r is the private target object's radius).

C. The Direct Vertices Reduction Paradigm

In this approach, by using the m-gons as the cloaked regions of the query and the private objects, we reduce the query processing time in STEP 1 and STEP 2 of the Casper* algorithm (Table II, formula 3). However, since we use an approximate version of the original cloaked region, we need some modifications in STEP 2 to search for nearest-neighbors of the parts of n-gon that are outside the m-gon (e.g., ABC in Figure 4a). During the Ramer-Douglas-Peucker algorithm's recursive divisions, for each simplified edge, we maintain the distance of the furthest vertex that is not inside the m-gon (B, E and H in Figure 4a). The list's size is at max m. We denote
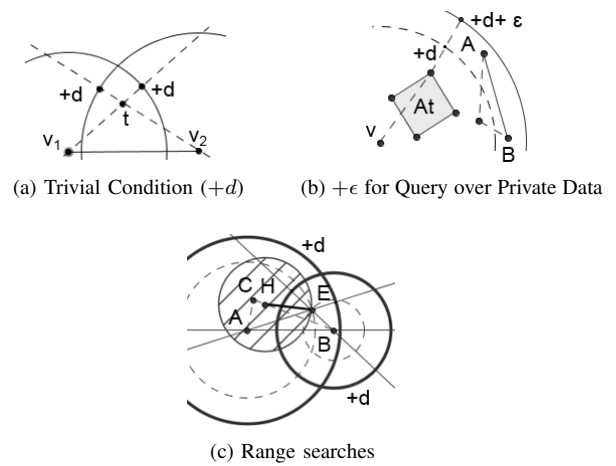
those distances as d (Figure 4a, distance from H to FJ). The modifications (Table III) make use of the distance d and only apply to the simplified edges that the discarded vertices are not all inside the m-gon, e.g. AC, CF and FJ in Figure 4a.

1) Modifications for query over public data: They can be summarized as follows:

- Trivial Edge (**TE**) and Split-Point (**TSP**) Condition: using the corresponding distance d maintained above, we add two range queries centered at v_i, v_j of radii $dist(v_i, t_i) + d$, $dist(v_j, t_j) + d$ into the range queries set R. As in Figure 5a, given t is the nearest-neighbor of simplified edge $v_1 v_2$, d is the distance of the furthest discarded vertex of edge $v_1 v_2$, we need to add two range queries center at v_1, v_2 with respective radius of $v_1 t + d$ and $v_2 t + d$. For Trivial Split-Point Condition, we add one more range query centered at sij of a radius $dist(s_{ij}, t_i) + d$ into R. As shown in Figure 5c, given that E is the nearest-neighbor of the simplified edge AB (from edge AC and AB), the nearest-neighbor E' of any point H on BC (C is a discarded vertex outside the m-gon) must be inside the hatched circle centered at H of radius HE ($HE' \leq HE$), which is always inside the two bold circles created by the enlarged $(+d)$ range queries. It is also the same for any points in ABC.

- Stopping Criterion Condition (**SC**): similarly, we increase the range query's radius by d to ensure including the nearest-neigbors of the outside parts of the original n-gon.

2) Modifications for query over private data: Because the private objects are also simplified, we will increase the search radius by $d + \epsilon$ for not missing them as candidate nearest-neigbors. As depicted in Figure 5b, consider that other private objects are also vertice-reduced versions, regarding vertex v of the m-gon in the query processing phase, the range query $(+d + \epsilon)$ reaches the simplified edge AB of another private object while the range query $(+d)$ does not.

3) Proof of accuracy: Let us consider a simplified edge as in Figure 6a, where the set of vertices A, B, C, D, E is reduced

TABLE III
ALGORITHM MODIFICATIONS SUMMARY

	Cases	Ranges $(center, radius)$
Public Data	TE	$(v_i, v_i t + d), (v_j, v_j t + d)$
	TSP	$(v_i, v_i t_i + d), (v_j, v_j t_j + d), (s_{ij}, s_{ij} t_i + d)$
	SC	$(v_i, v_i t_i + d), (v_j, v_j t_j + d), (s_{ij}, s_{ij} t_i + d)$
Private Data	TE	$(v_i, v_i At + d + \epsilon), (v_j, v_j At + d + \epsilon)$
	TSP	$(v_i, v_i At_i + d + \epsilon), (v_j, v_j At_j + d + \epsilon), (s_{ij}, s_{ij} At_i + d + \epsilon)$
	SC	$(v_i, v_i At_i + d + \epsilon), (v_j, v_j At_j + d + \epsilon), (s_{ij}, s_{ij} At_i + d + \epsilon)$

to edge AE. The box $AFGE$ ($AF = EG = furthest - distance - d$) contains all the outside parts of the simplified edge in this case. If the nearest-neighbors are inside the m-gon, they are already included in the candidate set in STEP3. So we only prove the correctness of STEP2, to find the candidate outside.

Query over Public Data In the trivial edge condition, where there is only one nearest-neighbor T for edge AE, we show that if T is not the nearest-neighbor of all the points in $AFGE$, the real nearest-neighbor Y must also be included in the range searches. There are three possibilities of T's position:

- On the edge AE (1).
- Inside the box $AFGE$ (2).
- On the edge FG or outside the box $AFGE$ (3).

First of all, in three cases, the range searches always cover the whole box $AFGE$, so we only prove that for edge FG, if T is not the nearest-neighbor, the real nearest-neighbor Y must be included in the range searches.

For vertex F, if there is another nearest-neighbor Y, it must satisfy $FY < FT$. Thus, the point Y is included in the range search centered at A, radius $AF + AT$.

As shown in Figure 6a, A, F, T_2 and T_3 are aligned, given $AT_3 = AT_1$, $FT_2 = FT$, $TT_1 = AF$, the segment FT_2 is always in the segment AT_3. We can see that the circle center at A, radius AT_3 always contains the circle center at F, radius FT_2, which always contains the circle center at F, radius FY. Hence, for any point Y that is closer to F than T, it will always be included in the circle center at A, radius AT_1.

It is similar to vertex G and all the points on edge FG. So if there is another nearest-neighbor Y of edge FG, it must be included in the range centered at A, radius $AT + AF$ and the range centered at E, radius $ET + EG$.

In the cases of Split-Point or Stopping Criterion Condition, the problems can be considered as the case of Trivial Edge Condition above, by dividing the problem into two sub-problems of edges AS and SE.

Query over Private Data We will provide two proofs in this part, (1) the real nearest-neighbor is included in the range searches and (2) the vertices-reduced private target objects will not be missed in comparison to the original ones.

For the first proof, similarly, the cases of Split-Point or Stopping Criterion Condition can also be considered as the case of Trivial Edge condition. Thus, we only show this case to prove the correctness of the modifications:

(a) Public Data

(b) Private Data: Case 1

(c) Private Data: Case 2

Fig. 6. Proof of accuracy: Query over Public/Private Data

- On the edge AE (1).
- Inside the box $AFGE$ (2).
- On the edge FG or outside the box $AFGE$ (3).

Similar to the proof of query over public data, the range searches cover the whole box and for both vertex F and G. As shown in Figure 6b, the real nearest-neighbor Y is always included in the range searches. So there is no candidate missing in the algorithm.

For the second proof, given the private target object $OPRS$, which is a vertice-reduced version of $OPQRS$. As the set Q, R, S is simplified to edge QS. Let us consider two cases, as shown in Figure 6c:

- Without the enlargement $+\epsilon$, originally, the private target $OPQRS$ will intersect with this range search. But the vertice-reduced version $OPQS$ will not.
- With the enlargement $+\epsilon$, both the private target $OPQRS$ and the vertice-reduced version $OPQS$ will intersect with this range search.

4) Running example: Let us briefly go through a simple running example to illustrate the algorithm more clearly. First, as depicted in Figure 7a, given an n-gon $ABCDEFG$, we process it with RDP, the result is the m-gon $ADFG$. Among all the edges (AD, DF, FG and GA) of the new m-gon, only AD and DF are simplified edge. Hence, the Casper*

(a) RDP Algorithm   (b) Public Objects

(c) edge AD   (d) edge DF

Fig. 7. Running Example: Private Query over Public Data

(a) Private Objects

(b) Enlarged radius

Fig. 8. Running Example: Private Query over Private Data

algorithm can be applied normally to edge FG and GA. The modifications will only be applied for edge AD and DF.

Query over Public Data Figure 7b shows the public objects related to the m-gon $ADFG$. In STEP1, the filters are found as follows.
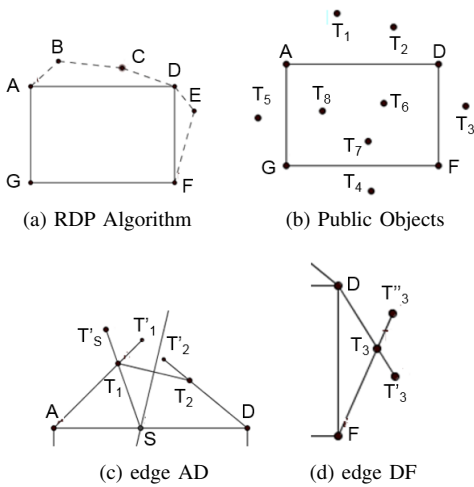
A gets the filter T_1 and T_5 ($AT_1 = AT_5$), D gets the filters T_2 and T_3 ($DT_2 = DT_3$), F gets the filter T_3 and T_4 ($FT_3 = FT_4$) and G gets the filter T_4 and T_5 ($GT_4 = GT_5$).

Figure 7c depicts the processing result of edge AD in STEP2. The distance from B to edge AD is used as distance d for this edge processing step. As A and D get different filters (T_1 and T_2), a split-point S is found. We add 3 search ranges (1) centered at A, radius AT_1', (2) centered at D, radius AT_2' and (3) centered at S, radius AT_S', where $T_1T_1' = T_2T_2' = T_1T_S' = d$. Figure 7d depicts the processing result of edge DF. Similarly, distance d is the distance from E to edge DF. As D and F get the same filter T_3, we add 2 search ranges centered at D, radius DT_3' and centered at F, radius FT_3'', where $T_3T_3' = T_3T_3'' = d$.

For edge FG and AG, as the two vertice of both edges get the same filters (T_4 and T_5), T_4 and T_5 are added into candidate set.

In STEP3, we add all objects inside $ADFG$ (T_6, T_7, T_8) into the candidate set. We also find all the objects inside the range searches found in STEP2.

Query over Private Data Figure 8a shows the private objects related to $ADFG$. The filters found in STEP1 are (At_1 and At_4 for A), (At_1 and At_2 for D), (At_2 and At_3 for F) and (At_3 and At_4 for G).

Figure 8b shows the range searches added in STEP2 for edge AD and DF. We add 2 range searches for each edge, (centered at A, radius AT_1', $TT_1' = d + \epsilon$) and (centered at D, radius DT_2', $TT_2' = d + \epsilon$) for edge AD, (centered at D, radius DT_3', $TT_3' = d + \epsilon$) and (centered at F, radius DT_4', $TT_4' = d + \epsilon$) for edge DF.

For edge FG and GA, as shown in Figure 8b, we process them normally with Casper* and gets another 2 range searches for each edge, (centered at F, radius FT_5) and (centered at G, radius GT_6) for edge FG, (centered at G, radius GT_7) and (centered at A, radius AT_8) for edge GA.

In STEP3, we add all private objects that intersect with the range searches added in STEP2 into the candidate set.

D. The Bounding Vertices Reduction Paradigm

One characteristic of the m-gon generated by the Ramer-Douglas-Peucker algorithm is that it may not contain the original n-gon. In this approach, we want to ensure the m-gon contains the original one. During the Ramer-Douglas-Peucker algorithm's recursive divisions, for each simplified edge (m edges), we maintain the furthest vertex that is not inside the m-gon (B, E and H in Figure 4a). After that, we calculate the m lines that are parallel to the respective edges of the m-gon and through the respective furthest vertices in the list (e.g., KL, LM, MN and NK in Figure 4a). The intersection of those lines forms a new m-gon that contains the original n-gon inside it (Figure 4a's $KLMN$). Therefore, the candidate set of the simplified m-gon is a superset of the original n-gon without directly modifying the Casper* query processing algorithm.

Although the first approach reduces the query processing time much, it suffers from the moderate increase of the candidate set size. Differently, the second approach achieves both better candidate set size and query processing time than the first one. Firstly, we can add the filters directly into the candidate set without the risk of missing the exact

nearest-neighbor because the simplified m-gon contains the original n-gon (no outside parts). Secondly, although the range query's radius is indirectly enlarged through the enlargement of the original n-gons to the bounding m-gons, it is kept minimum as $+d + d$, an indirect $+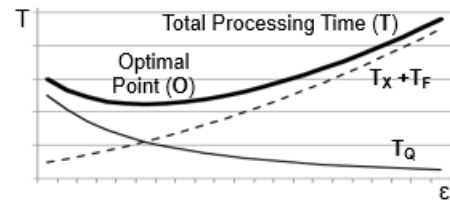d$ of the cloaked region and another $+d$ of the private object. Thus the number of results for each range query is also reduced in comparison to the first approach. Furthermore, the reduction in number of range queries also leads to a slight reduction of processing time.

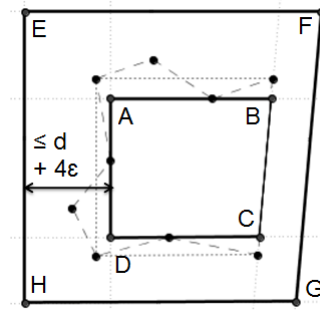### E. The Distance Dimension ϵ as Vertices Reduction Paradigm Tuning Parameter

The Total Processing Time (T) of a query consists of three components. (1) The Query Processing Time (T_Q), which is for the Query Processor to compute the candidate set. In the Vertices Reduction Paradigm, T_Q also includes the overhead of the Ramer-Douglas-Peucker algorithm to simplify the query's cloaked region. (2) The Data Transmission Time (T_X), which is for the candidate set to be read from the objects database and transmitted to the Location Middleware for nearest-neighbors filtration. (3) The Answers Filtration Time (T_F), which is for the candidate set to be filtered for the exact nearest-neighbor of the query request. T_Q is monotonically decreasing with the decrease of the number of vertices, while T_X and T_F are monotonically decreasing with the decrease of candidate set size. Thus, we can utilize the distance dimension ϵ as a tuning parameter for Vertices Reduction Paradigm since it affects the number of vertices in the vertices reduced versions and the search radius of range queries in the range query set R. As a result, a large value of ϵ leads to less number of vertices and less query processing time, but larger candidate set size. We will consider two cases in respect to the ϵ value: (1) $T_Q > T_X + T_F$. Initially, the ϵ is too small that query processing takes too much time. To resolve this, we must increase ϵ for fewer vertices and edges in query processing. (2) $T_Q < T_X + T_F$. This indicates the candidate set size is too big because the range searches' radius is too large that $(T_X + T_F)$ is longer than T_Q. We have to decrease ϵ to reduce the radii. Thus, in order to find an optimal value of ϵ for the best T, we increase ϵ until it reaches the optimal point O (Figure 9a).

As discussed previously, by utilizing VRP, we can achieve at most (n/m) speedup of Query Processing Time in Query over Public Data and $(n/m)^2$ speedup in Query over Private Data. Thus, the total decrease in Query Processing Time T_Q is accordingly $(1 - m/n)T_Q$ (Query over Public Data) and $(1 - m/n)^2 T_Q$ (Query over Private Data).

Let us consider a uniform-distributed data set, if the distance between public objects is u, the distance from a vertex to its nearest public object is $(u\sqrt{2}/2)$, we denote this distance as d_{pub} ($dist$). Accordingly, if the longest edge of the bounding rectangle of the private objects is b, so the distance from a vertex to its nearest private object is $((u + b)\sqrt{2}/2)$, we denote this distance as d_{pri} ($min-max-dist$). In the

(a) The tuning parameter ϵ

(b) The bound of candidate set

Fig. 9. Performance tuning

Vertices Reduction Paradigm, the increase in candidate set size depends on the difference of the areas covered by itself and the original Casper*. In the original Casper*, the area covered is at least less than the inner bounding polygon of the original polygon, as depicted in Figure 9b ($ABCD$). Meanwhile, in the Vertices Reduction Paradigm, the area covered is at most the polygon enlarged from the outer bounding polygon of the original one, as depicted in Figure 9b ($EFGH$). The reason of this enlargement is to cover the area of the range queries in STEP 2. As the distance in the respective parallel lines of the inner and the outer bounding polygon is at most 2ϵ and the range query's radius is at most $(d_{pub} + \epsilon)$ (for Query over Public Data) and $(d_{pri} + 2\epsilon)$ (for Query over Private Data). Thus the area difference is at most the difference between the two polygons above, which is at least $C(d_{pub} + 3\epsilon)$ for Query Over Public Data and $C(d_{pri} + 4\epsilon)$ for Query over Private Data, with C is the perimeter of the enlarged polygon ($EFGH$ in Figure 9b). Therefore, for Query over Public Data, the total time increase is $(C/u+1)((d+3\epsilon)/u+1)(t_X + t_F)$, with t_X and t_F is time to transmit and filter one candidate. Respectively, the total time increase is $(C/u + 1)((d + 4\epsilon)/u + 1)(t_X + t_F)$ for Query over Private Data. By collecting the required information (u, average of b, and average of C) of the formula above, we can explain and consider the impact of the increase in candidate set size of the Vertices Reduction Paradigm and adjust the distance dimension ϵ for better system performance. In general, the impact of the Vertices Reduction Paradigm is summarized in Table IV.

It is worth noting that the Vertices Reduction Paradigm works best when the query processing time is much longer than the data transmission and data filtration time ($T_Q \gg$

TABLE IV
SYSTEM PERFORMANCE IMPACT UTILIZING
VERTICES REDUCTION PARADIGM

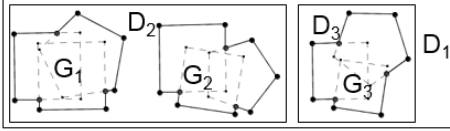| Public	$\dfrac{(1-m/n)T_Q \text{ -}(C/u+1)((d+3\epsilon)/u+1)(t_X+t_F)}{d=d_{pub}=(u\sqrt{2}/2)}$	(1)
Private	$\dfrac{(1-m/n)^2 T_Q \text{ -}(C/u+1)((d+4\epsilon)/u+1)(t_X+t_F)}{d=d_{pri}=((u+b)\sqrt{2}/2)}$	(2)

Fig. 10. Effective grouping with R-Tree

$T_X + T_F$), which means the system cannot provide high computational power. In that case, the paradigm will simplify the polygons to reduce the query processing time. But it will also increase the candidate set size to ensure the inclusion of the exact answer, leads to the increase in data transmission and data filtration time. Hence, the best situation to apply the Vertices Reduction Paradigm is that the system has fast read access and high network bandwidth so the increase in data transmission time will not cancel out the benefits of the paradigm. The faster read access and the higher the network bandwidth is, the better effect the Vertices Reduction Paradigm can achieve.

VI. GROUP EXECUTION AGENT

As shown in Figure 10, there are many queries of same filter parameters with adjacent and overlapped regions at a time (the dotted regions), or even better, a query's region is contained inside another's. Obviously, such queries share a part of or the whole candidate set. To take advantage of that, we add one additional component, the Group Execution Agent (GEA), into the Location Middleware. We also propose the Group Execution (GE) algorithm for the Agent. The Group Execution Agent will group as many queries as possible for one query execution before sending them to the query processor (group N queries into K group queries, $K \leq N$, i.e., group 9 queries to 3 groups as shown in Figure 10, the bold $G_{1,2,3}$ are used as cloaked regions in nearest-neighbor queries). The query processor will calculate the candidate set for the whole group as one single query. After that, the Location Middleware will filter the exact answer from the group query's candidate set for each query in the group.

TABLE V
QUERY PROCESSING ALGORITHM RUNTIME COMPLEXITY UTILIZING
GROUP EXECUTION AGENT

VRP	$N(T_Q + T_X) + NT_F$	(1)
GEA	$K(T_Q + T'_X) + NT'_F$	(2)
# of queries (N,K)	$K \leq N$	
Average T_X time (T_X, T'_X)	$KT'_X \leq NT_X$	

Algorithm 1 Group Execution Algorithm ($list\_region$, max_A)

1: $list\_group \leftarrow \{\}$
2: **for all** region r_i **in** $list\_region$ **do**
3:      $min\_a \leftarrow \infty$; $min\_d \leftarrow \infty$; $min\_r_j \leftarrow$ **null**
4:      **for all** region $r_j \neq r_i$ **do**
5:          $a \leftarrow bound\_area(r_i, r_j)$;
             $d \leftarrow intersect\_area(r_i, r_j)$
6:          **if** $a < min\_a$ **and** $a < max_A$ **then**
7:              $min\_a \leftarrow a$; $min\_d \leftarrow d$; $min\_r_j \leftarrow r_j$
8:          **end if**
9:      **end for**
10:     $list\_group \leftarrow list\_group \cup \{(r_i, min\_r_j, d, min\_a)\}$
11: **end for**
12: $sort(list\_group)$ **by** d descending **and** a ascending
13: $list\_grouped\_region \leftarrow \{\}$
14: **for all** group (r_i, r_j, d, a) **in** $list\_group$ **do**
15:     **if** r_i **is not** grouped **then**
16:         $grouped\_region \leftarrow$ **null**
17:         **if** r_j **is not** grouped **then**
18:             $grouped\_region \leftarrow group(r_i, r_j)$;
                 $mark\_grouped(r_i)$; $mark\_grouped(r_j)$
19:         **else**
20:             $grouped\_region \leftarrow r_i$; $mark\_grouped(r_i)$
21:         **end if**
22:         $list\_grouped\_region \leftarrow$
             $list\_grouped\_region \cup \{grouped\_region\}$
23:     **end if**
24: **end for**
25: **return** $list\_grouped\_region$

The Group Execution algorithm is outlined in Algorithm 1. Its purpose, given a list of query regions (of size N) and a parameter max_A, is to group the regions in the list into K grouping regions ($K \leq N$) of which areas are smaller than max_A. First, we find the best region pairs (r_i, r_j) that have the least enlargement in area when we group them together, the area a of the grouping regions must be less than max_A. For each pair, we also compute the intersection area d, and put all of them into $list(r_i, r_j, a, d)$. Then we sort the list by $d : descending$, $a : ascending$ and put the queries into group queries accordingly. The main reasons for this list computation and sorting are as follows. (1) $a \leq max_A$. We want to control the areas of the final grouping regions so that they will not be too large, leads to a large candidate set for each group query and long transmission time. (2) Sort $list(r_i, r_j, a, d)$ by $d : descending$, $a : ascending$. We want the cloaked regions of the queries in a group has maximum overlapping area since the larger intersection area, the larger number of common candidates. If two or more groups have the same intersection areas, we will prioritize the group that has smaller union area, according to reason 1.

The grouping regions are the bounding regions of each

group's regions and will be used as the cloaked regions of those group queries in nearest neighbor query processing. The group query's candidate set is a superset of the candidate sets of the queries in the group (as the final region contains all the regions in the group), so the Group Execution Agent does not miss any exact nearest neighbor of those queries. Thus, the Group Execution Agent does not affect the accuracy of the query processing algorithm. To sum up, from N queries $(q_1, q_2, q_3, ..., q_n)$, we create K group queries $(g_1, g_2, g_3, ..., g_k)$. Then we send the K group queries to the query processor to compute their candidate sets. From each group query's candidate set, we calculate the exact answer for each query in the group and send it back to the LBS user.

The system benefits from the Group Execution Agent as shown in Table V. (1) The query processing time for each group queries is the same as a single query (TQ) because we only execute the group query once with the grouping region as input. Thus, the sum of all queries' processing time decreases $(KT_Q \leq NT_Q)$. This also leads to the decrease of average query processing time. (2) The group query's candidate set size increases because it is a superset of the candidate sets of those queries in the group, but the average transmission time decreases as we only transmit the common candidates once (as $KT'_X \leq NT_X$). The average filtration time increases $(T'_F \geq T_F)$, but it is minor in comparison to the benefits above.

Furthermore, to achieve reasonable runtime, the algorithm's input list must satisfies two conditions: (1) the list's regions are adjacent to each other for easier grouping, (2) the list size is small enough to avoid scalability problem because the algorithm's runtime complexity is $O(N^2)$. To find those suitable lists, we maintain an R-Tree [21] in the Location Middleware. The R-Tree is not the fastest spatial access method, but its index criterion is that the directory rectangle's area is minimized which suits the Group Execution Agent the best. When a query is sent to the Middleware, its cloaked region is indexed by the R-Tree. By finding the R-Tree's nodes of which the directory rectangle's area is smaller than a predefined area value $kmax_A$, we will get the suitable lists from those nodes' regions. For example, in Figure 10, we find two suitable lists from the nodes D_2 and D_3's regions (given D_1's area exceeds $kmax_A$). Later, the algorithm returns grouping regions G_1, G_2 and G_3, that reduces the number of queries from 9 to 3. The runtime complexity is now only $O(n_1^2 + n_2^2 + ... + n_d^2)$, with n_1 to n_d is the number of regions in the directory rectangles. Because $(n_1 + ... + n_d = N)$, so $O(n_1^2 + n_2^2 + ... + n_d^2) < O(N^2)$. In this case, we sacrifice the actual best combinations because we only run the Group Execution algorithm on each suitable list instead of all queries in the system at that time. But by that trade-off, we achieve reasonable runtime with nearly best results.

In fact, the Group Execution Agent's speedup effect is dependent of how much overlapped the regions are. The worst case could be that we cannot group any query but still have the overhead of the R-Tree and the Group Execution algorithm. However, in most cases, when the number of queries is large

enough, the Group Execution Agent does strongly reduce the system's average query processing and transmission time and improve the system scalability. Furthermore, it is also worth noting that the grouping regions are complex polygonal regions since they are the bounding of all cloaked regions in each group. After the grouping regions are sent to the query processor, to reduce total processing time, the Vertices Reduction Paradigm will simplify them before calculating the candidate set. In other words, the Group Execution Agent and the Vertices Reduction Paradigm are complements for each other. The Group Execution Agent eases the drawback of the candidate set size of the Vertices Reduction Paradigm. The Vertices Reduction Paradigm reduces the total processing time when the query processor calculates the candidate set for the grouping regions. Together, they strongly enhance the whole system's performance and scalability.

VII. Experimental Evaluations

A. Overview

In this section, we evaluate both two Vertices Reduction Paradigm approaches and the Group Execution algorithm for the Private Query over Public and Private Data. The Vertices Reduction Paradigm approaches are evaluated with respect to the tuning parameter ϵ. For all two types of private query, we compare our algorithms with the Casper*, the performance evaluations are in terms of total processing time and candidate set size. We will also examine the system's average total processing time speedup and candidate set size reduction ratio in case of with and without the Group Execution Agent in the Location Middleware.

TABLE VI
EXPERIMENTAL SETTINGS

Cell area	$10000m^2$	**GEA** $kmax_A$	$300000m^2$
Grid size	100x100	**GEA** max_A	$100000m^2$
Circle radius	$180m$-$200m$	**Public Data**	9801
Object size	$< 1KB$	**Circular Private**	3200
Mobile Network	$14.4Mbits/s$	**Polygonal Private**	3200

Table VI summarizes our experimental settings. For all our experiments' private data, first we generate the private moving objects' precise positions using Siafu, a context simulator and the map of the crowded district 1 in Ho Chi Minh city, Vietnam. Siafu, the context simulator takes input of the map (consists of reachable and unreachable regions) and a list of moving objects (with start positions, target positions, etc.) and simulates the movement of those objects over time. We then capture the moving objects' position at a certain random time for uses in the experiments. The total number of private moving objects is 6400, which is big enough to show the effects of the improvements clearly. Then 50% of the precise positions are cloaked into complex rectilinear regions with Bob-Tree [15]. In our experiments, the map is divided into a grid of $100x100$ cells with each cell's area is of $10000m^2$. All the moving objects are obfuscated to reach a total of at least 10

cells in their cloaked region. As a result, the complex cloaked regions' numbers of vertices are very high and all range from 15 to 30. The remaining 50% of the private moving objects are cloaked into circular regions utilizing the works in [8] with the radius size at least $180m$ and less than $200m$ so that the area of the circle will be at least the area of the polygon of Bob-Tree. For the public data, we distribute 9801 objects uniformly in all our experiments ($99x99$, each at the center of the Bob-Tree cell above). The sizes of both the public and private data objects are at max $1KB$. We consider the network between Location Middleware and LBS Provider is optimistically high-speed of $1Gbits/s$ while the mobile network is HSPA 3G network of $14.4Mbits/s$ download rate.
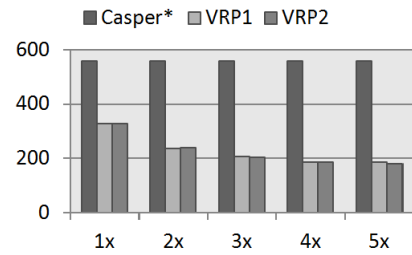
B. Vertices Reduction Paradigm

We will investigate the effect of the tuning parameter ϵ in this section. The baseline algorithm is the original Casper* query processing algorithm. Since Casper* cannot process circular regions, we only compare the total processing time of the polygonal regions. The ϵ values are respectively 1 time to 5 times of the Bob-Tree grid edge size.
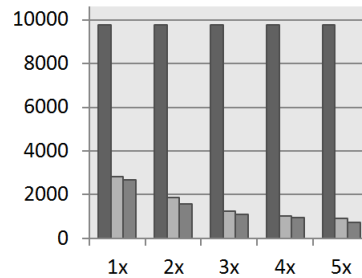
For the charts, the x-axis shows the respective values of the tuning parameter ϵ, as the ϵ increases from 1 time (1x) to 5 times (5x) of the Bob-Tree grid cell edge size. When the ϵ exceeds $5x$, the effects do not increase clearly since we cannot make the cloaked regions simpler than the previous ones in our experiments. The total processing times are measured in milliseconds (ms) and the candidate sets' sizes are measured as the number of data in the result. The abbreviation VRP1 is for the Direct Vertices Reduction Paradigm, while VRP2 is for the Bounding Vertices Reduction Paradigm and Casper* for the original algorithm.

For Private Query over Public Data, as depicted in Figure 11a, when the ϵ value increases, the total processing time decreases. At best (5x), the total processing time is only 32.19% of the original Casper* algorithm's processing time. As explained in Section 5, the main reason for this improvement in total processing time to happen is that the number of range queries is significantly reduced in Vertices Reduction Paradigm. In our experiment, at max reduction (5x), the number of vertices in Vertices Reduction Paradigm is only 24.85% of the original algorithm. Thus we can achieve at most 4 times speedup in Query over Public Data and 16 times speedup in Query over Private Data. For Private Query over Private Data, Figure 11b describes the effect of Vertices Reduction Paradigm. Total processing time in this case is also significantly improved. At best (5x), the total processing time is only 7.34% of the original algorithm's processing time. Besides the reduction in the number of the range queries, using the vertices reduced versions of the cloaked regions also simplifies the process of determining the furthest corners in finding the nearest private objects, thus the total processing time decreases greatly.

However, as a drawback, the candidate set sizes increase in both cases of Private Query over Public Data and Private

(a) Total Processing Time (ms) of Private Query over Public Data

(b) Total Processing Time (ms) of Private Query over Private Data

(c) Candidate Set Size of Private Query over Public Data

(d) Candidate Set Size of Private Query over Private Data

Fig. 11. Vertices Reduction Paradigm

Data. In our cases, as illustrated in Figure 11c and Figure 11d, at worst (5x), the candidate set sizes increase by 2.6 times (Private Query over Public Data) and 2.1 times (Private Query over Private Data). But this is not a big problem, since the data size is small (less than $1KB$) and the network between Location Middleware and LBS Provider is high-speed ($100Gbits/s$).

To sum up, due to the application of the Vertices Reduction Paradigm, as the tuning parameter ϵ increases, the total processing time decreases significantly while the candidate set size only increases slightly.

C. Group Execution Agent

We will examine the improvement when we integrate the Group Execution Agent in the Location Middleware in addition to the VRP included in the LBS Provider. In this experiment, we investigate the effect of the Group Execution Agent in the case of grouping 2 to 10 queries (the charts' x-axis is the number of queries grouped), as well as the average benefits in the whole system, in terms of average total processing time and average candidate set size. Similar to the Vertices Reduction Paradigm experiment, we perform this experiment on two private query types, i.e. the Private Query over Public Data and Private Query over Private Data. The pre-defined system parameters max_A and $kmax_A$ are respectively 1.5 and 3 times of the minimum cloaked region area. In other words, $max_A = 150000 m^2$ and $kmax_A = 300000 m^2$. Meanwhile, the concurrent query number is 500, which means 500 users are issuing the same query content (but with different cloaked regions) at the same time. In this experiment, the total query processing time also includes the overhead of the R-Tree indexing time and region grouping time.

As depicted in Figure 12a and Figure 12b, the system receives great speedup in average total processing. On average, the speedups are 5.24 times (Private Query over Public Data) and 3.84 times (Private Query over Private Data). In other words, the average total processing time decreases at least 3.84 times when the Location Middleware is integrated with the Group Execution Agent. The main reason for this benefit is that the actual number of queries is reduced as we group the queries before sending them to the Privacy Aware Query Processor in the Location-based Database Server.

Figure 12c and Figure 12d illustrate the average candidate set size reduction ratios when we apply the Group Execution Agent in the Location Middleware. The reduction ratios are at max the number of queries in a group, but this is not always guaranteed, as the actual reduction ratio depends on two main factors, (1) the data distribution and (2) the union and intersection area of the cloaked regions. On average, the total numbers of data to transmit are reduced by 5.12 times (Private Query over Public Data) and 3.33 times (Private Query over Private Data).

Generally speaking, when the Location Middleware is included with the Group Execution Agent, the whole system scalability increases and we gain speedups in average total processing time and reduction ratios in average candidate set size.

VIII. Conclusions and Future Works

This paper introduces an extended version of the state-of-the-art Location Privacy Aware Query Processor, to

(a) Average Total Processing Time Speedup in Private Query over Public Data

(b) Average Total Processing Time Speedup in Private Query over Private Data

(c) Average Candidate Set Size Reduction Ratio in Private Query over Public Data

(d) Average Candidate Set Size Reduction Ratio in Private Query over Private Data

Fig. 12. Group Execution

efficiently deal with complex polygonal cloaked regions, by proposing the Vertices Reduction Paradigm. We also support one more type of cloaked region, the circular shape ones, by transforming them into polygonal ones. The extended query processor can be embedded inside a location-based database server or an untrusted application middleware, allows users to completely keep their exact locations private, by sending only the cloaked regions to the server. In addition, the performance of the query processor can be tuned through a new parameter, to achieve a trade-off in system scalability, in terms of both

query processing time and query candidate set size.

As the experimental results show, our works reduce the query processing in both private queries over public and private data for complex polygonal and circular cloaked regions alike, while the increase in candidate set size is very small, especially in private queries over private data. However, the drawback of increase in candidate set size is insignificant since both transmission time and filtration time is very small in comparison to query processing time.

For further development, we are working on improving the Vertices Reduction Paradigm and the Group Execution Agent, i.e. faster algorithms and better effects, supports for continuous queries. In addition, the support of k nearest-neighbor should also be taken into account.

In LBS, location privacy aware query processing that handles cloaked regions has become an essential part in preserving user privacy. However, the state-of-the-art cloaked-region-based query processors only focus on handling rectangular regions, while lacking an efficient and scalable algorithm for other complex region shapes. Motivated by that problem, we introduce enhancements and additional components to the location privacy aware nearest-neighbor query processor that provides efficient processing of complex polygonal and circular cloaked regions, namely the Vertices Reduction Paradigm (VRP) and the Group Execution Agent (GAE).

Generally, our processor provides better support for the Application Middleware in systems that utilize more than one single obfuscation algorithm, as it supports various cloaked region shapes. Besides, as it provides efficient processing of indeterminate rectilinear regions, this result also allows and encourages the development of cloaking algorithms that utilize the geographic, semantic and sensitive features to protect user location privacy, e.g., Bob-Tree. The basic idea of the VRP is that we only maintain a set of vertices-reduced (VR) regions that are computed from the set of private objects' original cloaked regions. The processor will only use the VR versions to reduce computational cost. However, that improvement in total processing time also leads to the increase in candidate set size. To ease that problem, we introduce a new tuning parameter ϵ to achieve trade-off for total processing time and candidate set size. On the other hand, the GAE groups the queries before sending them to the Privacy Aware Query Processor in Location-based Database Server. To achieve that, the queries' cloaked regions are first indexed with the in-memory R-Tree. Then the GAE will scan the R-Tree and perform the group operations. This component strongly enhances the whole system's scalability as it reduces the number of queries to process.

As the experimental results show, our works reduce the query processing in both private queries over public and private data for complex polygonal and circular cloaked regions alike, while the increase in candidate set size is very small, especially in private queries over private data. However, the drawback of increase in candidate set size is insignificant

since both transmission time and filtration time is very small in comparison to query processing time.

The applications of both Vertices Reduction Paradigm and Group Execution Agent are not limited to the scope of Privacy Aware Nearest-Neighbor Query. In fact, VRP is able to serve as enhancing component for any algorithms that need to process irregular shapes as general polygons and GEA can be applied to scale up systems that process multiple regions concurrently.

For further development, we are working on improving the Vertices Reduction Paradigm and the Group Execution Agent, i.e., faster algorithms and better effects. In addition, the support of k nearest-neighbor should also be taken into account. Moreover, current private query processing algorithms only focus on position privacy, however, more attention and effort should be put into path privacy too, as the privacy breach in user trajectory also does harms to the development of LBS.

REFERENCES

[1] C. Bettini, X. S. Wang, and S. Jajodia, "How anonymous is k-anonymous? look at your quasi-ID," in *Proceedings of the 5th VLDB Workshop on Secure Data Management*, ser. SDM-08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 1–15.

[2] T. K. Dang, C. N. Ngo, T. N. Phan, and N. N. M. Ngo, "An open design privacy-enhancing platform supporting location-based applications," in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC-12. New York, NY, USA: ACM, 2012, pp. 59:1–59:10. [Online]. Available: http://doi.acm.org/10.1145/2184751.2184824

[3] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," in *Proceedings of the Third International Conference on Pervasive Computing*, ser. PERVASIVE-05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 152–170.

[4] M. L. Damiani, E. Bertino, and C. Silvestri, "The PROBE framework for the personalized cloaking of private locations," *Trans. Data Privacy*, vol. 3, no. 2, pp. 123–148, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1824401.1824404

[5] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, ser. MobiSys-03. New York, NY, USA: ACM, 2003, pp. 31–42. [Online]. Available: http://doi.acm.org/10.1145/1066116.1189037

[6] T. T. B. Le and T. K. Dang, "Semantic-aware obfuscation for location privacy at database level," in *Proceedings of the 2013 International Conference on Information and Communication Technology*, ser. ICT-EurAsia-13. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 111–120.

[7] Q. C. Truong, A. T. Truong, and T. K. Dang, "Memorizing algorithm: Protecting user privacy using historical information of location-based services," *IJMCMC*, pp. 65–86, 2010.

[8] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. di Vimercati, and P. Samarati, "Location privacy protection through obfuscation-based techniques," in *Proceedings of the 21st Annual IFIP WG 11.3 Working Conference on Data and Applications Security*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 47–60. [Online]. Available: http://dl.acm.org/citation.cfm?id=1770560.1770566

[9] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 12, pp. 1719–1733, Dec. 2007. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.190662

[10] C.-Y. Chow, M. F. Mokbel, and W. G. Aref, "Casper*: Query processing for location services without compromising privacy," *ACM Trans. Database Syst.*, vol. 34, no. 4, pp. 24:1–24:48, Dec. 2009. [Online]. Available: http://doi.acm.org/10.1145/1620585.1620591

[11] H. Hu and D. Lee, "Range nearest-neighbor query," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 78–91, Jan 2006.

[12] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous location-based queries in distributed mobile systems," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW-07. New York, NY, USA: ACM, 2007, pp. 371–380. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242623

[13] R. Shokri, C. Troncoso, C. Diaz, J. Freudiger, and J.-P. Hubaux, "Unraveling an old cloak: K-anonymity for location privacy," in *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, ser. WPES-10. New York, NY, USA: ACM, 2010, pp. 115–118. [Online]. Available: http://doi.acm.org/10.1145/1866919.1866936

[14] Q. C. To, T. K. Dang, and J. Kung, "OST-tree: An access method for obfuscating spatio-temporal data in location based services," in *2011 4th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Feb 2011, pp. 1–5.

[15] Q. C. To, T. K. Dang, and J. Küng, "A hilbert-based framework for preserving privacy in location-based services," *Int. J. Intell. Inf. Database Syst.*, vol. 7, no. 2, pp. 113–134, Apr. 2013. [Online]. Available: http://dx.doi.org/10.1504/IJIIDS.2013.053546

[16] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu, "SpaceTwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ser. ICDE-08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 366–375. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2008.4497445

[17] A. Khoshgozaran and C. Shahabi, "Private information retrieval techniques for enabling location privacy in location-based services," in *Privacy in Location-Based Applications*, C. Bettini, S. Jajodia, P. Samarati, and X. S. Wang, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 59–83.

[18] Y. Tao, D. Papadias, and Q. Shen, "Continuous nearest neighbor search," in *Proceedings of the 28th International Conference on Very Large Data Bases*, ser. VLDB-02. VLDB Endowment, 2002, pp. 287–298. [Online]. Available: http://dl.acm.org/citation.cfm?id=1287369.1287395

[19] D. H. Douglas and T. K. Peucker, *Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature*. John Wiley & Sons, Ltd, 2011, pp. 15–28. [Online]. Available: http://dx.doi.org/10.1002/9780470669488.ch2

[20] J. Hershberger and J. Snoeyink, "Speeding up the douglas-peucker line-simplification algorithm," University of British Columbia, Vancouver, BC, Canada, Canada, Tech. Rep., 1992.

[21] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD-84. New York, NY, USA: ACM, 1984, pp. 47–57. [Online]. Available: http://doi.acm.org/10.1145/602259.602266

Cipher Image Damage: An Application of Filters

Rolando Flores-Carapia, Víctor Manuel Silva-García, Benjamín Luna-Benoso and Carlos Rentería-Márquez

Abstract—In this paper, color images are encrypted and subsequently damage occlusion is made to the encrypted figures, with different sizes; the intention is to simulate an attack. In this research, two aspects are discussed, namely: the first is to encrypt images with quality; that is, the figures encrypted pass randomness tests proposed in this paper. The second aspect deals with the problem of recovering the encrypted figure information when it has been damaged. To retrieve information from encrypted images, the enryption of images is carried out in two steps: in the first a permutation is applied to the entire image and the second uses the AES cryptosystem with variable permutations. To perform this task an algorithm is used that utilizes the pi number to generate the permutations. To improve the sharpness of the deciphered figures with damage two filters are applied; median and average. To measure the degree of improvement in the damaged images two tests are proposed; the first is the correlation coefficient between adjacent pixels in the horizontal, vertical and diagonal directions. The second is based on the information entropy.

Index Terms—AES with variable permutations, goodness-of-fit test, fourier transform, correlation, entropy of information, median and mean filters.

I. INTRODUCTION

THE purpose of this paper is to analyze the attacks problem on encrypted file images. Such attacks may be in communication or storage, and are important in the context of real-time decisions [1]. That is, when an encrypted message is damaged – by attack or not – and then is decrypted the risk is not knowing the original message and sometimes there is not much time to decide what was the original file, say, less time than to ask for it again. So, the first point is to realize an image encryption with quality. In this sense, there are different image encryption works: there are some recent methods using the Hilbert transform [2], other Chaos [3], [4] and Hyper-chaos [5] or even AES cryptosystem [6] with CBC mode encryption [7], although in this latter case the encryption process is sequential. In the Hilbert transform and Chaos cases, there is the difficulty of not knowing specifically, the size of the keys set. In the case of Hyper-chaos, the keys set is 2^{167} and only the brute force attack is mentioned. The chosen-plaintext attacks as linear [8] or differential [9] are not addressed. In addition, there is a research of color image encryption [10], where the set size of the keys is not specified. Also, there are investigations in optics [11], [12], [13] where the number of keys is also

not noted. There is an interesting research in this field [14], however, the original image is modified when the figure is decrypted. On the other hand, there is an important encryption research of color images [15]. This work does not use as proof of the encryption quality any of the proposals in NIST Special Publication 800-22.

As for encrypted damage figures [16], the proposal is to use filters [17] in order to improve the sharpness of images with decrypted damage. This article only uses median and mean filters [17]. It is left for further research to propose other filters using mathematical morphology [18]. Then, for damage encrypted figures two actions that solve this type of problem are proposed, namely: as a first step encrypting images must be such that when these have a failure, it does not appear focused on the decoded figure, see figure 3. Therefore, applying an initial permutation over all image pixels, randomly constructed is recommended. The intention is to spread the figure data, so that the pixels remain randomly distributed. In this sense, when an encrypted figure suffers damage, it does not appear focused on the deciphered image. As a second step, AES with variable permutations is applied over the permutated image. It is shown later that the images encrypted by this method pass the tests of randomness proposed, which are: horizontal, vertical and diagonal correlation [19]; Discrete Fourier Transform (DTF) [20]; entropy; and the proposed "Goodness-of-fit test" [19]. In addition, a sensitivity analysis for the proposed k key is performed; that is, the study of the Correlation coefficient between the figure encrypted with k and $k + 1$, in order to show that no relationship between them is carried out. A filter is used in the figure decrypted with damage, to improve sharpness. Two types of filters are proposed. The first is the median with different sizes of mask and the second is the average. In addition, two instruments to measure the improvement after applying the filters are used. These instruments to measure the quality of the result are the following: the first is the Correlation coefficient between adjacent pixels in the directions: horizontal, vertical and diagonal. The second corresponds to the information Entropy. Surely there are several filters that could do a better job than suggested here. But, this is not the objective of this research which shows the improvement in the decrypted figures when some of the aforementioned filters are applied.

Damage to the encrypted images is carried out by concentric rectangles, see Figure 6, and the fault sizes are 35%, 40% and 45% of the total figure area. Of course, the dimensions of these faults may be higher or lower depending on the "quality" required to decrypt images with damage. However, in general it can be said that for the bigger the faults, the

TABLE I
A 3×3 MASK

$(x-1, y-1)$	$(x-1, y)$	$(x-1, y+1)$
$(x, y-1)$	(x, y)	$(x, y+1)$
$(x+1, y-1)$	$(x+1, y)$	$(x+1, y+1)$

lower the sharpness of the deciphered image. Furthermore, the damage or noise utilized in this paper is the occlusion noise; other noises are not discussed, for example, additive or multiplicative [21], [22].

It is important to clarify that this article does not use the compression process in the encryption image, because there are some countries whose safety areas do not allow the compression process in encryption images [23]. In other words, the process employed is Encryption $\rightarrow$ Decryption. The images encrypted in this work appear in many other investigations. These figures are: Peppers, Baboon, Barbara and Lena. A fifth figure is proposed according to the criteria described below.

II. PRELIMINARIES

The discussion begins with the filters, and as mentioned above, only two of them are discussed, namely, the median and the mean or average. It starts with the median filter which is a statistical non-linear type [17]. Two types of masks are used, which are 3×3 and 5×5 elements. In general, it can be said that the filters applied in this research conducted a manipulation in space $(n \times m)$ of image pixels. In the case of the mean filter, it only used the 3×3 size. In both filters the process is as follows: given any figure pixel (x, y) the analysis is made in the neighbor elements. For example, in the particular case of the 3×3 mask the (x, y) pixel has as adjacent pixels all the (x, y) considered in Table 1.

Regarding the median filter the process is as follows: the gray levels of the additive primary colors are arranged red, green and blue for each of the neighbor cells at (x, y) and including the pixel. Denote each of these values, as $M_{r,i}$, $M_{g,i}$ and $M_{b,i}$ for each of the additive primary colors respectively with $i = 1, \ldots, n$; where n is the total number of cells in the mask. In the case of one color figures, the different gray levels are worked. To illustrate the point, as a case study an array of 3×3 is taken, then, for each of the nine cells $(x-1, y-1)$ $(x-1, y) \ldots (x+1, y+1)$ each of the amounts $M_{r,i}$, $M_{g,i}$ and $M_{b,i}$ with $i = 1, \ldots, 9$ are sorted. After the amounts of cells are arranged and the median is denoted as δ; then, the following must comply: δ is greater or equal than the $\lceil n/2 \rceil - 1$ first elements; i.e 50%, and less than the remaining elements. The size of the arrays to be analyzed in this research are 3×3 and 5×5, since for a larger mask the process is slow, although, the quick sort algorithm was used [24].

The damage of the figure does not necessarily achieve greater sharpness. The average filter has equal probability weights; i.e. the same values. The formula is $\sum_{i=1}^{n} M_{c,i} P(M_{c,i})$, where $M_{c,i}$ is the color value c in the

cell i, which could be red, green, or blue, and $P(M_{c,i})$ is the probability or relative weight of $M_{c,i}$ value, which for this particular situation is $1/n$. The $P(M_{c,i})$ can be different amounts in different cells, taking into account that $\sum_{i=1}^{n} P(M_{c,i}) = 1$. This filter is classified as linear [17].

The algorithm used for image encryption is AES [25] for the following reasons: because it is a recent symmetric encryption system and as the international standard at this time, it makes this algorithm the most studied in the world. On the other hand, an efficient method for breaking it has not yet been found [26]. It is also noted that AES is a symmetric algorithm [7], which makes it very fast to encrypt information. There are different encryption protocols with AES [7]. In this work the ECB mode is used for the following two reasons: First, it is important to evaluate how variable permutations work in the encryption figures. Second, this mode allows parallelizing the encryption process and thus reduces time.

However, when the ECB protocol is used to encrypt figures that have low randomness degree in their bits and does not apply the variable permutations process, and also uses the same key for all 128 bit blocks of the image, it may be that the encrypted figure could give us information; i.e, the distribution of the different shades of basic colors follows a certain pattern; see Figure 2. This is the reason why an additional element is used in the algorithm, in this case a different permutation in each 128 bits block. This permutation is applied in the first round after the operation x-or rather than at the entrance of the first round as Triple-DES does [27]. It also shows that the keys set of AES cryptosystem can be up to 2^{256} elements.

The "quality" aspect of a figure encryption has to do with the randomness degree in the distribution of the encrypted image bits. In this sense, several methods have been used to measure the degree of randomness [20], although in this research the following are used: Correlation; horizontal, vertical and diagonal; Entropy; Discrete Fourier Transform and a different form to measure the degree of randomness of the bits of an encrypted image is proposed, using a "Goodness-of-fit test" [19]. Transcendental numbers are utilized, which have the characteristic of not being a solution of any polynomial with the form $a_n x^n + a_{(n-1)} x^{(n-1)} + \ldots + a_0$ with $a_i \in \mathbb{Z}$ [28], but also have the property that the decimal point to the right does not follow any regularity, so they are good candidates for use in generating pseudo-random numbers. The pi number is the transcendent to be employed in this article because it has been well studied [29].

The generating of permutations is dependent on the AES key. This is according to the following procedure: if we denote by m the integer that represents the string of 128 bits of the AES key. Then the product $m \times pi$ is also a transcendental number and from it is possible to get the constants that are used to generate permutations based on the following procedure. Given a non-negative integer $l \geq 2$ it is possible to define the set $N_l = \{n \in \mathbb{N} | 0 \leq n \leq l! - 1\}$ and on the other hand, according to the division algorithm of Euclid [30] for

all $n \in N_l$, it can be written uniquely as follows:

$$n = C_0(l-1)! + C_1(l-2)! + \ldots + C_{l-2}(1)! + C_{l-1}(0)!.$$

Then, using modular arithmetic it is possible to get the C_i for $i = 1, \ldots, l-1$ in a pseudo-random way. Later it will be shown, how from the expression (1) and using an algorithm there is a way to obtain the pseudo-random permutations: for both situations; the whole image and for each block of 128 bits.

The Entropy is measured according to the formula $-\sum_{x \in X} P_r(x) log_2 P_r(x)$; in Section 5, a more detailed explanation is given. Regarding the color image, each of the primary additives –red, green and blue– is described by one byte; that is, 256 gray levels are sufficient for each of them. Then, if a particular primary color has uniform distribution; i.e, all points are equally likely, the entropy value is 8 [31]. Although, one case in which the Entropy is 8 can be constructed, and the distribution of values is not random. However, in practice this is not so. Thus, values as close as possible to 8 for each of the primary colors, in an encrypted figure are sought.

A statistical test to evaluate the randomness of a bits sequence is formulated by means of a null hypothesis H_0, which states that the bit string is random versus the alternative hypothesis H_a which indicates that this is not so. To accept or reject the null hypothesis a statistical and threshold to define a rejection region is used; so, if the value of the statistical based on the data yields an amount that is in the rejection zone, this implies that the null hypothesis is rejected, otherwise H_0 is accepted.

In any hypothesis test scheme there are two types of errors, namely: type I error and type II error. Type I error is one that is committed when H_0 is true and it is rejected. Type II error is accepting H_0 when this hypothesis is false. The error that is controlled is the type I because it is considered that H_0 is the more important of the two hypotheses. The amount used in this research for the type I error is $\alpha = 0.01$, although $\alpha = 0.001$ can also be used [19].

The probability distributions that are used in the randomness tests are: Chi-square χ^2, the standard normal distribution and Complementary Error Function $erfc(z) = (\frac{2}{\sqrt{pi}}) \int_z^\infty e^{-u^2} du$ [32].

Correlation between two random variables x, y, is performed using the adjacent pixels of an encrypted image for each primary color additive; red, green and blue. The reasoning is this: if the encryption figure has good "quality"; then it is expected that the Correlation coefficient between adjacent pixels horizontal, vertical and diagonal is a number close to zero.

III. Algorithm to Generate Permutations

Suppose for the moment, that in the expression (1) the constants $C_0, C_1, \ldots, C_{l-2}$ are known and based on them the following algorithm is constructed:

Step 0. An array in ascending order is defined as follows: $X[0] = 0, X[1] = 1, \ldots, X[l-1] = l-1$.

Step 1. The condition $C_0 < (l)$ is observed; then, $X[C_0]$ is one element of the array in step 0. So, $X[C_0]$ is removed from the arrangement in step 0, and instead is replaced by $X[l-1]$; that is, the last element. If $X[C_0]$ is the last element, then this is replaced for the penultimate element. Note, only two operations are performed; removal and replacement. That is, the other elements of the array remain unchanged.

Step 2. In the same way as in the previous step, the condition $C_1 < (l-1)$ is fulfilled, so, $X[C_1]$ is an array element of step 1. Thus, following the same logic of step 1, $X[C_1]$ is removed and instead is replaced for the last item. Of course, if $X[C_1]$ is the last element, then, the process is the same as step 1.

Step $l-1$. If this process is repeated at the end the following result will appear: $X[C_{l-2}]$ and $X[C_{l-1}] = k$ with $0 \le k \le l-1$. The number $X[C_{l-1}]$ appears automatically since it is the last; that is, $C_{l-1} = 0$. The array of positive integers $X[C_0], X[C_1], \ldots, X[C_{l-2}]$ is a permutation of the array $0, 1, \ldots, l-1$. This procedure is performed in steps $l-1$. Regarding the complexity to perform this algorithm is $\mathcal{O}(l)$, since in each step removal and replacement of an element is carried out, the others remaining unchanged. In this paper two sizes of permutation are generated, namely: the 128 positions and that which permutes the pixels in the entire image. If the figure is about 500,000 pixels, 960×540, the size of the problem to be solved is $\mathcal{O}(500,000)$ for this particular case, which means it can be resolved speedily. Actually, the time spent on the construction of the switch from one image of dimensions 960×540 is less than ten milliseconds using software.

Now, it is relevant to show how the constants $C_0, C_1, \ldots, C_{l-2}$ are obtained. Note that it is not important to know the number n, fortunately, because otherwise the integers for the simplest case have a magnitude of $128! - 1 \approx 10^{215}$, and for the more complex case the permutation is over the whole image size, the numbers could be $500,000! - 1$, which is huge.

In this sense, the quantities $(l-1)!, (l-2)!, \ldots$ are only used as marks; that is, it is not necessary to write them with all their digits. Then, the next question to address is: how to choose the pseudo-random values for the C_i for $i = 1, 2, \ldots, l-1$. As mentioned above, this paper used the pi number as follows:

(1) The symmetric cryptosystem key Advanced Encryption Standard-AES, is a string of zeros and ones, which represents a positive integer. Denote this integer as m; then, this paper proposes to multiply m by pi, such that the product is itself a transcendent number. Particularly, in this investigation the symmetric cryptosystem AES-128 is used, although there is the possibility of employing up to 256 bit keys.

(2) After completing the multiplication $l \times pi$, and after the decimal point to the right, one Byte strings are taken which are denoted as: $b_0, b_1, \ldots, b_{126}$: For each plaintext of 128 bits, it will utilize 127 strings of one Byte. It follows that, the sets

number of 127 chains of one Byte corresponds to the blocks number of 128-bits that have the image to be ciphered.

Each $C_i = b_i \bmod (128 - i)$, for $i = 0, 1, \ldots, 126$ is defined. Remember, the constant $C_{127} = 0$ since it is the last.

(3) Once the constants C_i, for $i = 0, 1, \ldots, 126$ were calculated for each block of 128 bits the algorithm described above is applied to get the permutations of 128 positions.

When the whole image has to be permutated, the number of positions to exchange, l, can be $500,000$ or more elements. In such case, the procedure is similar to $l = 128$. But, there are some differences, namely, the size chains a_i is 24 bits, or 3 Bytes. This, since many current images do not exceed 2^{24} bits in the spatial resolution. Furthermore, the three-Byte blocks also represent integers. So, it is proposed to calculate the constants C_i as follows: $C_i = a_i \bmod l - i$, for $i = 0, 1, \ldots, l - 2$ y $C_{l-1} = 0$.

Sometimes in the image to be encrypted some bytes are subtracted, according to the following criteria: If $24 \times l \bmod 128 \not\equiv 0$ where l is the number of pixels of the figure, then, a minimum amount of Bytes is subtracted, say n, such that $24(l) - 8(n) \bmod 128 \equiv 0$. It is important to note that $8n < 128$ and the $8n$ bits are not encrypted. Once, the $C_0, C_1 \ldots C_{l-2}$ values are known, the π_l permutation over a l elements array is calculated, according to the procedure described previously.

IV. AES ENCRYPTION ALGORITHM WITH A VARIABLE PERMUTATION

This section explains how the tool developed in the previous section in the process of image encryption is used. There are two steps to encrypt a figure; the first is to generate a permutation of the size image and later apply it over the whole figure. The second step is to use the AES algorithm with a modification, i.e., to utilize a different permutation for each 128-bits block after the x-or operation in the first round. Permutation on the whole image is intended to disperse the information so, when the encrypted figure is damaged, it does not appear in a focused manner in the decrypted figure. Actually, this is the intention of permuting the entire image. On the other hand, the algorithm described in the previous section defines a Bijective function [33], from the integers set to the permutations set, which is denoted as I_m, thus, if I_m is a Bijective function, it follows that it is a one to one function. This is important, since two different sets of constants C_i have associated two different permutations, which means in a general way the resulting block from the $Input \oplus k_1$ operation is modified in a different manner, where the entrance string in the first round is $Input$, and k_1 is the first from the keys schedule.

Another question that may arise is why after the x-or operation? And why in the first round? Regarding the first question, the reason it is not used at the entrance of the first round as with Triple-DES or Triple DES-96 [34], is because some images have areas of the same color; for example, black or white. In this situation a permutation applied to

ones or zeros strings does not make any modifications to them. But, when it is used after the x-or operation, this allows modifying the bit strings. Furthermore, why in the first round? Given that the information is mixed in each round, then any changes made in the first round there is more opportunity to scramble information, and at the end of the encryption process the ones and zeros will appear randomly. It is important to explain whether the multiplication of the integer associated to the AES key with pi, somehow affects the communication between two people. Actually, this is irrelevant in a secure communication scheme, such as Public Key Infrastructure (PKI) [35]. The key can be encrypted and transmitted using an asymmetric encryption cryptosystem, for example, ElGamal [36], RSA [37] or Elliptical curve [38]. The receiver can know the key with its private key and compute $l \times pi$. Later, the receiver can compute the variable permutations and the permutation of the whole image.

V. RANDOMNESS ANALYSIS OF ENCRYPTED IMAGES

As mentioned at the beginning of this work tests of randomness are carried out to find out what is the "quality" of the encrypted images; that is, what is the randomness degree in the colors of encrypted figures. Clearly, when the images are with different shades the process is the same.

A. Correlation, Entropy and Discrete Fourier Transform tests

Randomness analysis of the following tests is started: Correlation in directions; horizontal, vertical and diagonal; Entropy and Discrete Fourier Transform. Also, as mentioned earlier, the encryption of the images is performed without compression, or more specifically lossless information. In any image encryption it is important that the distribution of its bits should be random, in order to avoid bias that might lead to attacks to discover the key or plaintext.

With respect to the Correlation between adjacent pixels of an encrypted image with "quality," it is expected that there is a Correlation coefficient close to zero between adjacent pixels, that is, the linear relationship between them must be very weak [39]. Adjacent pixels are considered in three directions, namely, horizontal, vertical and diagonal.

The process of computing the Correlation between two random variables; x and y, is carried out as follows:

A pixel of the encrypted image is selected randomly. This pixel has a value for red, green and blue which is denoted as x_r, x_g and x_b. After selecting a random pixel, the next pixel is taken in adjacent directions horizontal, vertical or diagonal as appropriate. Similarly, as in the previous case, the adjacent pixel selected has a value for red, green and blue. These amounts are denoted as follows: y_r, y_g and y_b.

So, suppose M pairs of pixels x, y are chosen randomly. Then, it is possible to calculate the Correlations in the three directions for the three primary colors. The formula for calculating the Correlation coefficient in the horizontal

direction and for the red color is as follows:

$$r_{h;x_r,y_r} = \frac{\sum_{i=1}^{M}(x_{h;i,r} - \overline{x}_{h,r})(y_{h;i,r} - \overline{y}_{h,r})}{\sqrt{(\sum_{i=1}^{M}(x_{h;i,r} - \overline{x}_{h,r})^2)(\sum_{i=1}^{M}(y_{h;i,r} - \overline{y}_{h,r})^2)}},$$

where $\overline{x}_{h,r}$ and $\overline{y}_{h,r}$ are

$$\overline{x}_{h,r} = \frac{1}{M}\sum_{i=1}^{M}x_{h;i,r} \text{ and } \overline{y}_{h,r} = \frac{1}{M}\sum_{i=1}^{M}y_{h;i,r}.$$

Clearly, the expressions in the vertical and diagonal directions as well as for the green and blue colors are the same. In the case of a mono-color image, the process is the same as a color figure.

In case of the Entropy, the study of pixels dispersion in the images is performed by separating the primary colors at each pixel. When a single color is required, one Byte is necessary to calculate the Entropy, i.e., 256 gray levels. For color figures three Bytes are necessary, one for each basic color. In this vein, it is said that if the distribution of bits is totally random the Entropy is 8. To measure the randomness in strings of zeros and ones in practical cases the proceeding is as follows: When Entropy is near 8 it understands that the string of zeros and ones is random; otherwise it would mean that it is not.

To calculate the Entropy it is assumed that there is a string of pixels. In this regard, it is possible to separate each pixel in the chain into its basic color. Then, suppose that the bits string in the red color is divided into blocks of 8 bits, that is, one Byte; it follows that it has 256 possible values. Frequencies are recorded in a table of 256 classes according to their order of appearance. Therefore, each class is assigned a frequency f_i for $i = 0, 1, \ldots, 255$, so, an estimation of the probabilities for each of the classes is $P[x_i] = \frac{1}{f_i}$, where f_i is the class frequency x_i, $i = 0, 1, \ldots, 255$. In this vein, Entropy, say for red color, is calculated as follows: $H_c = -\sum_{x_i \in X} P_c(x_i)log_2[P_c(x_i)]$, where X is the set of all classes. In a simple manner, from the last expression it can be seen that for the green and blue colors the formulas are the same.

The Discrete Fourier Transform measures the degree of randomness of zeros and ones string, that is, there is no periodicity –repetitive patterns– one followed by another.

The following items appear in the calculation of the test statistical:

- N_0. It is a theoretical amount expected; $\frac{(0.95)n}{2}$, where n is the chain length.
- N_1. It is the number of values below a threshold h, which in turn depends on the string length n.

$f_j = \sum_{k=1}^{n} x_k e^{\frac{2(pi)j(k-1)i}{n}}$. If n is odd, just the last bit string is suppressed. Clearly, f_j has a real and another complex part. The $\|f_j\|$ module is calculated, which is real; later, it is compared with h. If $\|f_j\| < h$ a one is added to N_1 value. Otherwise N_1 remains with the previous value. With the latter

dates the quantity

$$d = \frac{N_1 - N_0}{\sqrt{\frac{n(0.95)(0.05)}{4}}}$$

and the test statistical $P - value = erfc(\frac{d}{\sqrt{2}})$; are calculated with $erfc(\frac{d}{\sqrt{2}}) = 2(1 - \Phi(d))$. The decision rule is: if $P - value$ is less than 0.01 the null hypothesis is rejected, otherwise it is accepted.

B. Randomness Test Proposed

Due to work with images testing randomness based on how the color bits are arranged in an encrypted figure, so the $\chi^2 = \sum_{i=1}^{k}[\frac{(o_i - e_i)^2}{e_i}]$ statistical is used for each primary color. The o_i is the observed value and e_i is the expected value. Using the χ^2 statistical it is possible to quantify the freedom degree that has the distribution of the different shades of colors; red, green and blue.

In all NIST 800-22 tests, to determine the degree of randomness of the bits in a string, this type of proof does not appear, that is, the tone distribution randomness of the basic colors is not measured, so, it is a different situation. In the same way as the tests of 800-22 NIST standard, the proposal proof uses the Goodness-of-fit test, utilizing the statistical, χ^2, which has a probability distribution of Chi-square with $n - 1$ degrees of freedom. The freedom degrees are obtained in the following way: the shades of each color of an image can be represented as a histogram whose abscissa has 256 divisions. Then, the degrees of freedom are 255 [19]. On the other hand, if it is considered that the random variable χ^2 approaches the normal distribution according to the center limit theorem, it follows that the mean and variance of χ^2 statistical are: $\mu = 255$ and $\sigma = \sqrt{2(255)} = 22.5831$ for each basic color.

With this information it is simple to calculate the threshold for a significance level of $\alpha = 0.01$ and $\alpha = 0.001$, considering that both size of significance gray levels are in the right tail of the normal distribution. So, the threshold for significance level $\alpha = 0.01$ is 307.61 and for $\alpha = 0.001$ is 324.78.

Then, the process of making the decision to accept or reject the null hypothesis, according to particular datas is as follows:

a) The statistical $\chi^2 = \sum_{i=1}^{k}[\frac{(o_i - e_i)^2}{e_i}]$ is calculated with specific values, where o_i and e_i are observed and expected values number i.

b) The probability to the right of the value χ^2 is calculated; if this probability is greater or equal to 0.01, then the null hypothesis is accepted, otherwise it is rejected. If the significance level is 0.001 the procedure is similar.

C. Sensitivity Analysis

It is important in image encryption to make the Correlation between two figures encrypted with two very close keys, for example, the difference between these two keys could be one. This Correlation should be close to zero. This means, that no
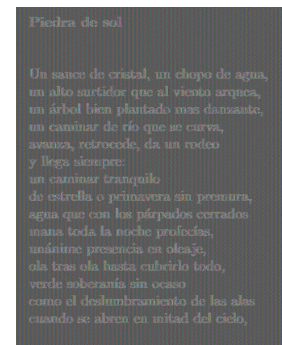
Rolando Flores-Carapia, Víctor Manuel Silva-García, Benjamín Luna-Benoso, and Carlos Rentería-Márquez

Fig. 1. Type image to be encrypted.

Fig. 2. The original (a) and cipher (b) image without variable permutation.

Fig. 3. (a) The original image, (b) the encrypted image with variable permutation and (c) deciphered image with damage.

matter the closeness between different keys, the result is that there is no relationship between the two encrypted images.

D. Proposed Image to be Encrypted

In the introduction to this article it was mentioned that a criterion would be presented to choose the figure to be encrypted. This criterion is based on a characteristic of the Goodness-of-fit test that tells us the following: if the tones distribution in each of the three basic colors was totally random $\chi^2 = 0$. In fact, this means that each color histogram is a uniform distribution.

However, when χ^2 has a very large value for each of the primary colors AES with variable permutations should be applied, otherwise the encryption is not efficient, see Figure 2. In addition, there are many relatively small images with χ^2 which can be applied directly to the AES cryptosystem, see Figure 5; i.e. it is not necessary to use a random permutation after the x-or operation in the first round for each 128-bits block of plaintext. Encrypted image passes all the aforementioned tests of randomness and also the proposal. So, in this research, it is proposed to choose an image that has a χ^2 as large as possible for each of the basic colors in order to show that the proposed method is effective for encrypt images.

In this investigation a figure with the following chi-square for the primary colors red, green and blue is used: $\chi_r^2 = 56{,}638{,}911.17$, $\chi_g^2 = 56{,}555{,}658.91$ and $\chi_b^2 = 55{,}396{,}932.18$. This image is of a piece of poetry by Octavio Paz (Mexican poet, 1914–1998) [40]. See Figure 1.

VI. RESULTS PRESENTATION OF ENCRYPTED IMAGES

As noted earlier in this article, first the results of the encryption process are shown to verify that the procedure observes the tests of randomness proposals, and also, it is compared with other researches. The images that are ciphered appear in many papers of encryption figures. Such images are five, namely: Peppers, Baboon, Barbara, Lena and the proposed figure. The first four images are presented in Figure 4, and the proposal in Figure 1. The 128-bits key of the AES cryptosystem is written in a hexadecimal system, and it is as follows:

$$k = 00112233445566778899AABBCCDDEEFF.$$

In fact, it can be assigned randomly.

As noted earlier, the k key is associated with a positive integer, which is as follows:

$$l = 88962710306127702866241727433142015.$$

Then, multiply the number by pi; that is, the product $l \times pi$ which in turn is a transcendental number. To the right of the decimal point the number of bits needed to cover all the sets of constants used to encrypt the image is taken. The tests: DFT, Goodness-of-fit proposal, Entropy and Correlation coefficient of adjacent pixels in the directions horizontal, vertical and diagonal, also, a sensitivity analysis for the k and $k + 1$ key are applied in this section.

A. The Discrete Fourier Transform, Entropy and the Proposed Test

It starts with DFT applied to Figures 4 and 1; later, the test for Goodness-of-fit. In both cases it is encrypted with

Fig. 4. Images: (a) Peppers, (b) Baboon, (c) Barbara and (d) Lena.

the k key. It is determined if the bit strings for each of the colors; red, green and blue approve the randomness criteria for significance level $\alpha = 0.01$.

The test results of the DTF are presented in Table 2 and the Goodness-of-fit test in Table 3.

Regarding the amounts $P - value_r$, $P - value_g$ and $P - value_b$ which are the threshold values for the primary colors; red, green and blue, they appear in Table 3.

Study of the Entropy, for each of the primary colors is conducted separately. As noted in this research earlier, an image is "well encrypted" if the Entropy of each basic color is closer to eight. Table 4 shows the results.

Entropy results presented in Table 4 are better than those shown in other studies [5].

B. Sensitivity Analysis

Concerning the sensitivity, the amounts of the Correlation coefficient between figures encrypted of image (a) Figure 4 are shown, one with the k key and the other with $k + 1$ key. It also clarifies that the way to make this task is by means of a randomly chosen sample as follows: three thousand pairs of pixels (x_i, y_i) are taken, where x_i is a randomly chosen pixel from the encrypted image with k, and y_i is the corresponding pixel in the image encrypted with $k + 1$.

Correlation analysis was performed for the three primary colors; red, green and blue. The results are presented in Table 5.

C. Correlations Between Adjacent Pixels

If an image is "well coded" it is expected that the Correlation coefficient between adjacent pixels in the directions: horizontal, vertical or diagonal, have a Correlation
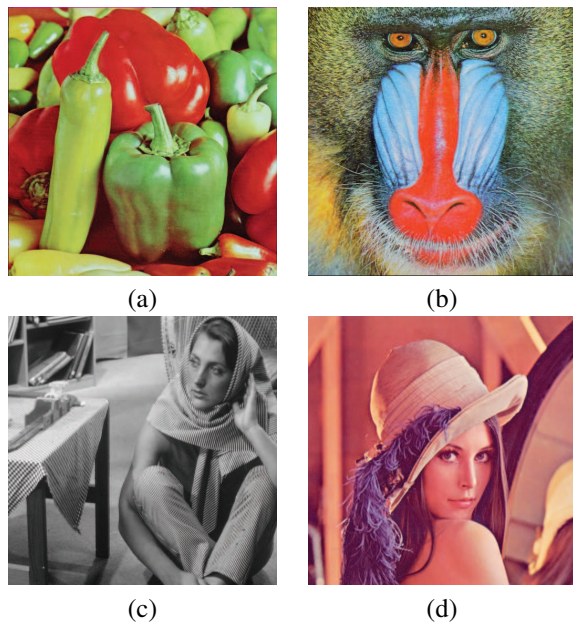
close to 0. This means that there is no linear relationship between adjacent pixels for the same three directions. The Correlations calculation takes a random 3000 pairs sample of pixels and for each basic color in all three directions the analysis is performed.

Images in Figure 4 and Figure 1 are used for this task. The results of the Correlations for the original images are shown in Table 6 and those encrypted in Table 7.

In both tables the values for each primary color are separated. Regarding the notation, the Correlation coefficient is denoted as r, which has two subscripts. The first one indicates the direction: h, v or d and the second is the color: r, g or b. For example, if the diagonal Correlation coefficient is desired for the blue color it is denoted as $r_{d,b}$.

VII. PRESENTATION OF RESULTS FOR IMAGES WITH DAMAGE

This part carries out the figures analysis with damage, whereas the faults are applied in concentric rectangles; in fact they may be of any other shape, since in the first stage of the encryption process the permutation over the whole image disperses the pixels in a pseudo-random manner. Clearly, when the figure is decrypted with failure the result is less sharp than the original image. This is due to noise which is introduced in the decoded image. The type of noise that is applied in this article is the occlusion; leaving to other research the additive and multiplicative noise.

In this order, two filters are applied, namely: the first is the median with 3×3 and 5×5 masks. The second is the mean, and the objective of both is to reduce the noise; that is, make the decipher images sharper. Also, the object of this work noted above is not to compare filters, but to show that the implementation of some well-known filters can improve the sharpness.

Damage sizes used in this article are shown hereunder: 30%, 40% and 45%. Moreover, the elimination of noise is measured in relation to decoded figures with damage, as the receiver ignores the original information. This research proposes to measure the improvement in sharpness according to the following instruments: the Correlation coefficients in three directions; horizontal, vertical, diagonal and the Entropy.

All these tests are conducted for the three primary colors. Obviously, in the case of images with different gray levels the process is similar. Table 8 shows the results of the decoded image with a fault, and corresponds to image (a) Figure 4. These results for each of the colors are written: red, green, blue and consider different damage sizes: 35%, 40% and 45%.

The mean values are shown in Table 9. These amounts are obtained for the three gray levels of faults mentioned, and for the three basic colors. It is noted that the procedure of encryption and decryption with damage is performed with the image (a) Figure 4, regarding the results of the Correlation coefficient in the three directions: horizontal, vertical, diagonal and for the three primary colors are expressed in Tables 10 and 11. This test applies for the two filters; that is, the median

Rolando Flores-Carapia, Víctor Manuel Silva-García, Benjamín Luna-Benoso, and Carlos Rentería-Márquez

TABLE II
THE DFT TEST RESULTS APPLIED TO ENCRYPT IMAGES OF FIGURE 1 AND FIGURE 4 WITH k KEY (✓ ACCEPTED, χ REJECTED).

Test name	Significance Label $\alpha = 0.01$	$P-value/Decision$				
		Figure 1	Figure 4			
			(a)	(b)	(c)	(d)
Spectral	Red	0.667/✓	0.308/✓	0.702/✓	0.613/✓	0.769/✓
DTF	Green	0.861/✓	0.023/✓	0.923/✓	0.760/✓	0.183/✓
	Blue	0.504/✓	0.602/✓	0.970/✓	0.531/✓	0.334/✓

TABLE III
THE PROPOSAL TEST RESULTS APPLYING TO ENCRYPT IMAGES OF FIGURE 1 AND FIGURE 4 WITH k KEY (✓ ACCEPTED, χ REJECTED).

Test name	Significance Label $\alpha = 0.01$	$P-value/Decision$				
		Figure 1	Figure 4			
			(a)	(b)	(c)	(d)
Proposal	Red	0.46/✓	0.33/✓	0.46/✓	0.38/✓	0.78/✓
Test	Green	0.04/✓	0.18/✓	0.52/✓	0.48/✓	0.71/✓
	Blue	0.46/✓	0.09/✓	0.21/✓	0.33/✓	0.68/✓

TABLE IV
ENTROPY OF ENCRYPTED IMAGES USING THE k KEY FOR FIGURE 1 AND FIGURE 4.

Entropy	Figure 1	Figure 4			
		(a)	(b)	(c)	(d)
Red	7.99934	7.99929	7.99929	7.99928	7.99934
Green	7.99925	7.99924	7.99930	7.99930	7.99931
Blue	7.99934	7.99921	7.99924	7.99927	7.99932

TABLE V
SENSITIVITY ANALYSIS FOR IMAGE (A) FIGURE 2. USING THE KEYS k KEY AND $k+1$

Correlation	Results of Sensitivity
	Image (a) figure 2
Red	0.0313
Green	0.0133
Blue	0.0098

Fig. 5. Image (a) Figure 4 encrypted, using the same key k for each 128 bits block

and the average. The reasoning here is as follows: Noise reduction is significant as can be seen, when the Correlation as an instrument of measure is used. Also, in general it can be said that for bigger damage, the filters are less efficient, i.e. both measuring instruments may be inaccurate.

In this regard, the size of the fault is important because if the damage is around 80% or 90% it is very difficult with this information to recover the original image.

Figure 6 and Figure 7, where the encrypted images have 40% damage are presented. In Figure 6, the 5×5 median filter for image (a) Figure 4 is applied; and in Figure 7 the average filter for the same image.

VIII. DISCUSSION OF THE RESULTS

The discussion of results is separated into two parts: the first deals with the aspect of image encryption, and the second addresses the problem of encrypted figures with damage.

Regarding the first point, encrypted images pass all the randomness tests proposed in this research. Images were encrypted with the k key, which can be chosen at random. Furthermore, the results of Entropy are better than in other studies.

Sensitivity testing was also performed in order to show that there is no relationship between encrypted images and keys in close proximity.

On the second point, it is shown that when these figures were encrypted and then decoded with damage, they can be improved by using filters. The filters used were: median and average and in both cases the noise decreased. Two measuring instruments to determine the noise reduction were used: Entropy and Correlation coefficient in three directions. In all cases it was possible to reduce noise as illustrated in Tables 8 to 11.

Generally, if the damage is large, say greater than 50%, it can be said that it is very difficult to gather the information of the original image again. Furthermore, the type of damage used was occlusion, leaving for other investigations the analysis of additive and multiplicative noise.

IX. CONCLUSIONS

This research has presented a different way of encrypting color images consisting of two steps, namely in the first a permutation is applied over all image pixels and in the second the AES cryptosystem with variable permutations is used. To show that the encrypted images are of "quality" four tests were applied. The result was that all suggested images that were encrypted passed the randomness tests, and in some cases the

TABLE VI
THE CORRELATION COEFFICIENT RESULTS IN DIRECTIONS; HORIZONTAL, VERTICAL, AND
DIAGONAL FOR COLORS RED, GREEN AND BLUE FOR ORIGINAL IMAGES FIGURE 1 AND FIGURE 4.

color	Correlation Coefficient	Figure 1	Figure 4			
			(a)	(b)	(c)	(d)
Red	*Horizontal*	0.60	0.99	0.86	0.89	0.97
	Vertical	0.73	0.99	0.77	0.95	0.98
	Diagonal	0.50	0.98	0.73	0.88	0.96
Green	*Horizontal*	0.63	0.98	0.90	0.90	0.97
	Vertical	0.80	0.98	0.85	0.95	0.98
	Diagonal	0.49	0.96	0.84	0.88	0.96
Blue	*Horizontal*	0.60	0.97	0.92	0.89	0.95
	Vertical	0.77	0.97	0.87	0.96	0.96
	Diagonal	0.47	0.96	0.85	0.88	0.93

TABLE VII
THE CORRELATIONS RESULTS IN DIRECTIONS: HORIZONTAL, VERTICAL AND DIAGONAL
FOR COLORS RED, GREEN AND BLUE, FOR CIPHER IMAGES WITH k OF FIGURE 1 AND FIGURE 4.

color	Correlation Coefficient	Figure 1	Figure 4			
			(a)	(b)	(c)	(d)
Red	*Horizontal*	0.031	0.033	0.009	0.004	0.000
	Vertical	0.003	0.016	0.041	0.061	0.011
	Diagonal	0.004	0.002	0.003	0.038	0.010
Green	*Horizontal*	0.002	0.004	0.006	0.016	0.008
	Vertical	0.002	0.002	0.037	0.019	0.026
	Diagonal	0.001	0.019	0.014	0.019	0.020
Blue	*Horizontal*	0.005	0.012	0.010	0.040	0.022
	Vertical	0.013	0.008	0.009	0.005	0.040
	Diagonal	0.014	0.008	0.000	0.003	0.015

TABLE VIII
ENTROPY RESULTS OF IMAGE (A) FIGURE 4 WITH SEVERAL DAMAGE SIZES, USING MEDIAN FILTER

Test name	Size damage Figure 4	Entropy of deciphered image with damage	Entropy with median filter 3×3	Entropy with median filter 5×5
Red color Entropy	35%	7.851	7.655	7.614
	40%	7.874	7.677	7.627
	45%	7.895	7.694	7.641
Green color Entropy	35%	7.756	7.287	7.258
	40%	7.795	7.279	7.244
	45%	7.829	7.272	7.228
Blue color Entropy	35%	7.649	7.185	7.105
	40%	7.703	7.221	7.113
	45%	7.751	7.261	7.122

TABLE IX
ENTROPY VALUES OF IMAGE (A) FIGURE 4 WITH SEVERAL DAMAGE SIZES, USING THE AVERAGE FILTER

Test name	Size damage Figure 4	Entropy of deciphered image with damage	Entropy with average filter
Red color Entropy	35%	7.851	7.551
	40%	7.874	7.499
	45%	7.895	7.438
Green color Entropy	35%	7.756	7.068
	40%	7.795	7.011
	45%	7.829	7.959
Blue color Entropy	35%	7.649	7.100
	40%	7.703	7.058
	45%	7.751	7.017

TABLE X
ENTROPY RESULTS OF IMAGE (A) FIGURE 4 WITH SEVERAL DAMAGE SIZES, USING MEDIAN FILTER

Test name	Direction	Size damage Figure 4	Correlation of decoded image with damage	Correlation with median filter $3 \times x3$	Correlation with median filter 5×5
Red color Correlation	Horizontal	35%	0.392	0.957	0.964
		40%	0.311	0.931	0.956
		45%	0.284	0.922	0.941
	Vertical	35%	0.374	0.961	0.955
		40%	0.339	0.927	0.954
		45%	0.258	0.913	0.946
	Diagonal	35%	0.380	0.938	0.934
		40%	0.321	0.926	0.931
		45%	0.253	0.896	0.913
Green color Correlation	Horizontal	35%	0.217	0.938	0.906
		40%	0.190	0.904	0.894
		45%	0.155	0.888	0.885
	Vertical	35%	0.255	0.947	0.878
		40%	0.227	0.929	0.871
		45%	0.191	0.910	0.882
	Diagonal	35%	0.237	0.919	0.830
		40%	0.207	0.880	0.819
		45%	0.163	0.871	0.812
Blue color Correlation	Horizontal	35%	0.162	0.877	0.939
		40%	0.141	0.842	0.892
		45%	0.128	0.786	0.909
	Vertical	35%	0.185	0.892	0.910
		40%	0.152	0.847	0.905
		45%	0.107	0.824	0.874
	Diagonal	35%	0.148	0.860	0.853
		40%	0.113	0.796	0.848
		45%	0.120	0.722	0.825

TABLE XI
CORRELATION VALUES OF IMAGE (A) FIGURE 4 WITH SEVERAL DAMAGE SIZES, WITH AVERAGE FILTER

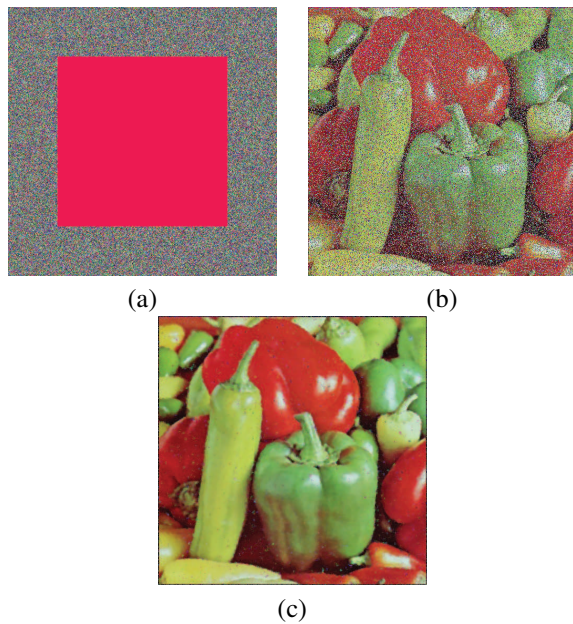| Test name	Direction	Size damage Figure 4	Correlation of decoded image with damage	Correlation with average filter
Red color Correlation	Horizontal	35%	0.392	0.957
		40%	0.311	0.918
		45%	0.284	0.900
	Vertical	35%	0.374	0.921
		40%	0.339	0.911
		45%	0.258	0.893
	Diagonal	35%	0.380	0.897
		40%	0.321	0.884
		45%	0.253	0.852
Green color Correlation	Horizontal	35%	0.217	0.929
		40%	0.190	0.841
		45%	0.155	0.816
	Vertical	35%	0.255	0.875
		40%	0.227	0.866
		45%	0.191	0.846
	Diagonal	35%	0.237	0.820
		40%	0.207	0.722
		45%	0.163	0.753
Blue color Correlation	Horizontal	35%	0.162	0.845
		40%	0.141	0.807
		45%	0.128	0.787
	Vertical	35%	0.185	0.831
		40%	0.152	0.816
		45%	0.107	0.782
	Diagonal	35%	0.148	0.772
		40%	0.113	0.749
		45%	0.120	0.687

Fig. 6. (a) Ciphered Figure 5 with 40% of damage, (b) deciphered image(a) and (c) median 5×5 filter applied to (b).

Fig. 7. Image (b) Figure 6 deciphered with 40% damage and average filter applied

results were better than others. Two popular filters were used to improve the deciphered figures with faults. The measuring instruments to determine the degree of improvement in the decoded images show that there is a reduction in noise, i.e. the Entropy amount was reduced and the Correlations coefficient value increased. Finally, the software was developed in C ++ and the time encryption of figures proposed in this research was around 85 milliseconds, and an Intel Core i7 processor was used.

REFERENCES

[1] S. Baruah, V. Bonifaci, G. D'Angelo, H. Li, A. Marchetti-Spaccamela, N. Megow, and L. Stougie, "Scheduling real-time mixed-criticality jobs," *IEEE Transactions on Computers*, vol. 61, no. 8, pp. 1140–1152, 2012.

[2] L. Xuemei, X. Tong, and L. Dai, "A novel scheme reality preserving image encryption," in *2011 Third International Conference Measuring Technology and Mechatronics Automation*. IEEE, 2011, pp. 218–221.

[3] J. Li and L. Gan, "Study on chaotic cryptosystem for digital image encryption," in *2011 Third International Conference Measuring Technology and Mechatronics Automation*. IEEE, 2011, pp. 426–430.

[4] C. Fu, J. Chen, H. Zou, W. Meng, Y. Zhan, and Y. Yu, "A chaos-based digital image encryption scheme with an improved diffusion strategy," *Optics Express*, vol. 20, no. 3, pp. 2363–78, 2012.

[5] H. Zhu, C. Zhao, and X. Zhang, "A novel image encryption-compression scheme using hyper-chaos and chinese remainder theorem," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 670–680, 2013.

[6] "Advanced encryption standard (AES)," 2001, FIPS PUB 197.

[7] D. R. Stinson, *Cryptography: Theory and Practice*. Chapman & Hall/ CRC Press, 2005.

[8] M. Matsui, "Linear cryptanalysis method for DES cipher," *Lecture Notes in Computer Science*, vol. 765, pp. 386–397, 1994.

[9] E. Biham and A. Shamir, "Differential cryptanalysis of the full 16-round DES," *Lecture Notes in Computer Science*, vol. 740, pp. 487–496, 1993.

[10] S. Keshari and S. G. Modani, "Color image encryption scheme based on 4-weighted fractional fourier transform," *Journal of Electronic Imaging*, vol. 21, no. 3, pp. 033 018–1–6, 2012.

[11] W. Chen, X. Chen, and C. Sheppard, "Optical color-image encryption and synthesis using coherent diffractive imaging in the fresnel domain," *Optics Express*, vol. 20, no. 4, pp. 3853–65, 2012.

[12] P. Refregier and B. Javidi, "Optical image encryption based on input plane and fourier plane random encoding," *Optical Letters*, vol. 20, no. 7, pp. 767–769, 1995.

[13] W. Chen and X. Chen, "Double random phase encoding using phase reservation and compression," *Journal of Optics*, vol. 16, no. 2, pp. 025 402–7, 2014.

[14] W. Chen, B. Javidi, and X. Chen, "Advances in optical security system," *Advances in Optics and Photonics*, vol. 6, no. 2, pp. 120–155, 2014.

[15] Y. Wu, G. Yang, H. Jin, and J. P. Noonan, "Image encryption using the two-dimensional logistic chaotic map," *Journal of Electronic Imaging*, vol. 21, no. 1, p. 013014, 2012.

[16] V. M. Silva-García, R. Flores-Carapia, C. Rentería-Márquez, B. Luna-Benoso, C. A. Jiménez-Vázquez, and M. D. González Ramírez, "Cipher image damage and decisions in real time," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 013 012–1–13, 2015.

[17] R. Gonzalez and R. Woods, *Digital Image Processing*. Prentice Hall, 2008.

[18] P. Soille, *Morphological Image Analysis*. Springer-Verlag, 2004.

[19] R. Wolpe and R. Myers, *Probability and Statistics for Engineers and Scientists*. Prentice Hall, 2007.

[20] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. NIST 800-22, 2010.

[21] J. Peters, D. Janzing, and B. Scholkopf, "Causal inference on discrete data using additive noise models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 12, pp. 2436–2450, 2011.

[22] Y. Yao and S.-S. Chen, "Multiplicative noise enhances spatial reciprocity," *Physica A: Statistical Mechanics and its Applications*, vol. 413, no. 1, pp. 432–437, 2014.

[23] "Prácticas comerciales – requisitos que deben observarse para la conservación de mensajes de datos," 2002, Nom-151. Norma Oficial Mexicana NOM-151-SCFI-2002.

[24] P. Dymora, M. Mazurek, and D. Strzalka, "Long-range dependencies in quick-sort algorithm," *Przeglad Elektrotechniczny*, vol. 90, no. 1, pp. 149–152, 2014.

[25] J. Daemen and V. Rijmen, "AES proposal: Rijndael," 1998.

[26] D. A. Osvik, A. Shamir, and E. Tromer, "Cache attacks and countermeasures: The case of AES," *Lecture Notes in Computer Science*, vol. 3860, pp. 1–20, 2006.

[27] "Data encryption standard (DES)," 1999, FIPS PUB 46-3.

[28] S. Michael, *Calculus: cálculo infinitesimal*. Barcelona, Spain: Reverte, 1993.

[29] "Jaohxv. pi world," http://ja0hxv.calico.jp/pai/estart.html, accessed: 2010-09-30.

[30] J. Gallian, *Contemporary abstract algebra*. Brooks/Cole, 2011.

[31] E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[32] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. Washington: National Bureau of Standards, 1968.

[33] T. M. Apostol, *Análisis Matemático*. Barcelona: Reverté, 1994.

[34] V. M. Silva-García, R. Flores-Carapia, C. Rentería-Márquez, and B. Luna-Benoso, "Triple-DES-96 cryptographic system," *International Journal of Contemporary Mathematical Sciences*, vol. 8, no. 19, pp. 925–934, 2013.

[35] A. Gómez, *Enciclopedia de la Seguridad Informática*. México: Alfaomega, 2007.

[36] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Transaction on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985.

[37] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[38] R. Azarderakhsh and A. Reyhani-Masoleh, "Parallel and high-speed computations of elliptic curve crytography using hybrid-double multipliers," *Transactions on Parallel and Distributed Systems*, vol. 26, no. 6, pp. 1668–1677, 2015.

[39] D. Jay, *Probabilidad y Estadística: para ingeniería y ciencias*. International Thompson, 2005.

[40] J. D. Argüelles, *Antología general de la poesía mexicana*. México: Océano, 2012.

An Implementation of Propositional Logic Resolution Applying a Novel Specific Algebra

Eduardo Zurek, Mayra Zurbaran, Margarita Gamarra, and Pedro Wightman

Abstract—This paper presents a methodology for evaluating propositional logic satisfiability using resolution-refutation. The method applies a strategy based on an algebra developed by the authors that estimates the possible outcomes of the expression and generates a logic value for refuting or accepting the satisfiability of the argument.

I. INTRODUCTION

ARTIFICIAL intelligence's main objective has been the construction of systems capable of replicating aspects that can be categorized as intelligent. Formally, "… is the set of techniques, methods, tools and methodologies that help us build systems that behave in a similar way to a human in the concrete problem resolution". To achieve this, it is used logic among other resources, logic can be understood as a combination of language and semantics, to represent the knowledge related with the ability to perform certain reasoning [1].

To give a better understanding to logic, the algebraic logic is introduced, it can be described in general terms as "the discipline that studies logics by associating with them classes of algebras, classes of logical matrices and other algebra related mathematical structures that relates the properties of logics with algebraic properties of the associated algebra" [2].

In this paper will be used the propositional logic or zeroth order logic; it is the simplest logic language, based in a numerable (not necessarily finite) of atomic propositions (AP). A proposition is an expression in natural language that can only be false (F) or true (T). A common form of representing all the evaluations is through the truth table; which is a complete enumeration of the value of the formula for all models. However, for evaluating larger expressions, the

E. Zurek is with the Research in Robotics and Intelligent Systems group at the Universidad del Norte, Barranquilla, Colombia (e-mail: ezurek@uninorte. edu.co).

M. Zurbaran is with the GReCIS group at Universidad del Norte, Barranquilla, Colombia (e-mail: mzurbaran@uninorte.edu.co).

M. R. Gamarra Gamarra is with the IET group, Department of Electronic and Telecommunication Engineering, Universidad Autónoma del Caribe, Barranquilla, Colombia (e-mail: margarita.gamarra@edu.co).

P. Wightman is with the GReCIS group at the Universidad del Norte, Barranquilla, Colombia (e-mail: pwightman@uninorte.edu.co).

truth table becomes infeasible since the complexity is of $O(2^n)$ for a formula containing n propositional variables, therefore, the tableau method was later introduced.

The Tableau method is a formal proof procedure existing for several logics. It is based in a refutation or indirect proof procedure; showing a formula X is valid by intending to assert it is not with a syntactical expression. The expression is broken down syntactically by splitting it in several cases; this may be referred as the tableau expansion stage, it can be thought of as a generalization of disjunctive normal form expansion [3]. Finally, there are rules for closing cases; this is deriving impossibility conditions or contradictions based on syntax. If each case closes, the tableau can be considered closed. A closed tableau is proof of the validity of the expression X it began with.

With the above in mind, it was developed an algebra of our own by enlarging the domain of the conventional Boolean logic of true, false, and indefinite.

The paper is structured as follows. Section II gives a general background on the classic Tableau methodology and related implementations. Section III explains the methodology introduced by the authors. Section IV explains the algorithm and its implementation and mentions the software used for comparing the results. Section V presents the results and explains how the comparison was evaluated. Finally, Section VI presents the conclusions.

II. BACKGROUND

The Tableau methodology was introduced in the 1950's by Beth and Hintikka and was later perfected by Smullyan and Fitting. It brings together the proof-theoretical and the semantical approaches to present a logical system, which is also very intuitive and has been broadly used according to [3]; the usages include implementation for circumscriptive reasoning as in [4], for interval temporal logic as seen in [5], for solving satisfiability check problems on Boolean circuits [6].

For the means of this article it is important to define propositional satisfiability and how it differs from propositional validity. Propositional satisfiability determines if there exists an interpretation that satisfies a given Boolean formula. This is, in words of Vardi [7] "The Boolean Satisfiability Problem (SAT, for short) asks whether a given Boolean formula, with Boolean gates such as AND and NOT,

has some assignment of 0s and 1s to its input variables such that the formula yields the value 1". This problem is said to be NP-Complete, proved by Steven Cook in 1971. In [8], [9] validity is defined as an argument is logically valid if and only if its conclusion is a logical consequence of its premises. If an argument whose conclusion is β and whose only premise is α is logically valid, then α is said to logically imply β. Other pertinent definitions are tautologies and contradictions; a well-formed formula is a tautology if and only if it is true for all possible truth-value assignments to the statement letters making it up. On the other hand, a well-formed formula is a self-contradiction if and only if it is false for all possible truth-value assignments to the statement letters making it up.

III. METHODOLOGY

The objective of this algorithm is to evaluate propositional logic satisfiability using a refutation proof and a novel algebra defined by the authors, to proof the satisfiability of an expression X by deducing that ¬X is not possible. Tableau proofs are originally presented as trees, whose nodes are formulas that are subcases and the tree structure gives the logical dependence between them [10], due to the structure the method uses, it is not trivial to implement in a programming language, requiring data structures advanced knowledge. With this in mind a simpler approach was conceived, which does not require graph structures and little processing. The presented method EZLogicUN is based on a general algebraic logic which considers not only the two Boolean states of *true* and *false*, but includes the following two expressions:

$$q \lor \neg q \qquad (1)$$
$$q \land \neg q \qquad (2)$$

From this point and by the means of simplifying the codification of the proposed method, the notation used replaces the operators: ∨ for +, ∧ for *, → for > and ¬ for −. In this way, the mentioned expressions are shown in Table I.

TABLE I.
NOTATION

Expression	Abbreviation
Indefinite	0
q	1
$-q$	−1
$q + -q$	2
$q * -q$	3

The functions used are: not (−), and (*) and or (+). The domain is limited to the expressions above and will be denominated by the numbers 0, 1, −1, 2, 3 accordingly. Table II shows the algebraic operations for an expression X and its negation.

TABLE II.
NOT OPERATOR

X	−X
−1	1
0	0
1	−1
2	3
3	2

The resulting expressions with the NOT operator can be inferred by propositional logic rules of negation. 2 and 3 specifically, are a case of the De Morgan axiom [11]:

$$-(q \pm q) = -q * - - q = -q * q = q * -q$$
$$\Rightarrow -2 = 3 \qquad (3)$$

$$-(q + -q) = -q * - - q = -q * q = q * -q$$
$$\Rightarrow -2 = 3 \qquad (4)$$

Table III presents the results of operating the expression X AND Y. Similarly to the NOT operator table, the results are deduced by applying simple propositional logic rules. The resulting inference that prevails is the one that avoids a contradiction if possible.

TABLE III.
AND OPERATOR

X	Y	X * Y
−1	−1	−1
−1	0	−1
−1	1	3
−1	2	3
−1	3	3
0	0	0
0	1	1
0	2	2
0	3	3
1	1	1
1	2	3
1	3	3
2	2	2
2	3	3
3	3	3

In the same way is calculated the operations for the OR function in Table IV. The OR function gives the possibility to select the case in which no contradictions (resulting in 3) arises if possible. This is particularly useful while analysing all the possibilities for the values of the expression to be a propositional satisfiability.

IV. IMPLEMENTATION

For the means of comparing the outcomes of our implementation, EZLogicUN, there will be evaluated the same logic arguments on both; our algorithm which was programmed in a MATLAB [12] script and the LoTREC Tableau Prover [13] which runs under the Java Runtime Environment (JRE) [14]. Even though the process varies in both approaches, the satisfiability of the argument should prove to be the same in both tools.

TABLE IV.
OR OPERATOR

X	Y	X + Y
−1	−1	−1
−1	0	0
−1	1	2
−1	2	−1
−1	3	−1
0	0	0
0	1	1
0	2	0
0	3	0
1	1	1
1	2	1
1	3	1
2	2	2
2	3	2
3	3	3

The proposed algorithm is a variation of the Tableau method, the principle of proving the validity of an argument X by deducing that ¬X is not possible remains. However, instead of looking for contradictions in the deductive process, this method aims to find a valid resolution, discarding the cases that lead to contradictions until there are no alternative cases or until the validity of ¬X has been proved, in which case X is invalid and therefore not propositionally satisfied.

The algorithm is of recursive nature and the use of data structures is limited to arrays instead of Tableau's tree structure. It consists of two phases, in the first phase the goal is to break down the expression and analyse its operators; for this, it is subdivided by identifying where there are parentheses, and then within each group, the operations are analysed, taking into account the precedence of the operators. The implication operator is transformed using the implication law, so that only $\cdot, +$ and $-$ are left in the expression.

After each case has been reduced to a simple operation, these are substituted by the notations of Table I, which covers all the possibilities. Next, the satisfiability of the whole expression is evaluated based on the algebra of Tables II, III, and IV for the corresponding operations.

The final outputs of the algorithm are squared arrays that stand for no contradictions if 3's are not found, or would proof the expression a contradiction if the opposite. The presence of 3 stands for a contradiction, implying that there is no possible outcome in which X is valid.

To summarize, the phases of the algorithm can be presented as follows:

1. Parentheses analysis.
2. Transform implications.
3. Divide the expression in simple operations (subcases).
4. Replace using notation from Table I and present the expression as an array.
5. Compare to check if there are contradictions (3's) with the proposed algebras (Tables II, III, and IV).
6. Output final arrays.

V. RESULTS

Table V shows the outputs of both programs evaluating the same logical expression for comparison purposes. As shown, the method introduced gives the same evaluation results as the classical tableau implementation of LoTREC.

TABLE V.
TESTS

Argument	EZLogicUN	LoTREC
$p \cdot -p$	Contradiction	Contradiction
$q + p > p + q$	Satisfied	Satisfied
$(p > -q) \cdot (p > q)$	Satisfied	Satisfied
$q \cdot p \cdot (p > q)$	Satisfied	Satisfied
$p \cdot q \cdot (r + -q)$	Satisfied	Satisfied
$p + -p$	Satisfied	Satisfied

The following images are introduced to show a step by step of a classic Tableau Tree for evaluating some of the expressions in Table V. The images were generated using LoTREC, firstly it is presented a general structure of the tree and then each leaf node is extended to show if any contradiction arises (the word FALSE will indicate this at the end of the analysis) and to which branch of the tree the node belongs. If there if just one node forming the Tableau Tree as Fig. 1 then the node is displayed directly.

The expression evaluated in Fig. 1, $p \cdot -p$, did not produce any branches during the tableau tree construction, rather it was concluded that the expression was FALSE in the first and only node of the tree.
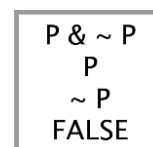
```
P & ~ P
   P
 ~ P
FALSE
```

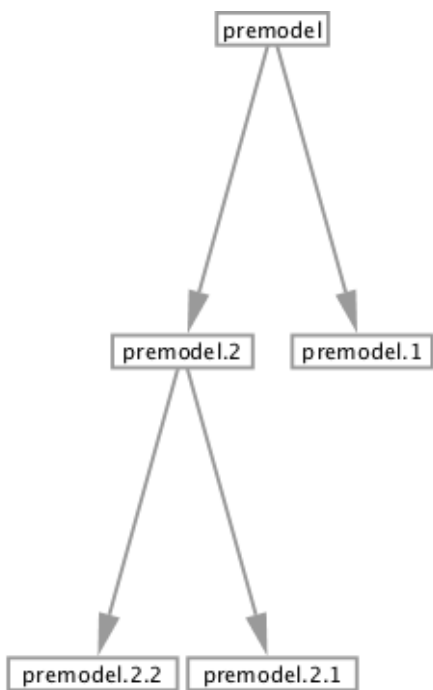Fig. 1. Tableau tree node output by LoTREC for the expression $p \cdot -p$

Fig. 2. Tableau tree structure output by LoTREC for the expression $q + p > p + q$

The next images depict the contents of each leaf node or premodel as stated by LoTREC of the tree in Fig. 2. These concluding leaf nodes are: premodel 1, premodel 2.1 and premodel 2.2.
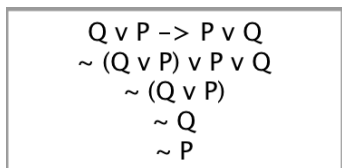

Q v P -> P v Q
~ (Q v P) v P v Q
~ (Q v P)
~ Q
~ P

Fig. 3. Tableau tree node output by LoTREC of **premodel 1** of the tree structure in Fig. 2.

Q v P -> P v Q
~ (Q v P) v P v Q
P v Q
P

Fig. 4. Tableau tree node output by LoTREC of premodel 2.1 of the tree structure in Fig. 2

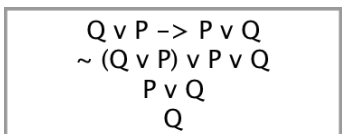Q v P -> P v Q
~ (Q v P) v P v Q
P v Q
Q

Fig. 5. Tableau tree node output by LoTREC of **premodel 2.2** of the tree structure in Fig. 2

Since none of the leaf nodes presents (Fig 3, 4, and 5) any contradiction as the premodel in Fig. 1. Then it is concluded that the expression $q + p > p + q$ is TRUE.

By using the script introduced by the authors for the expressions previously evaluated and explained in LoTREC, the script outputs the following arrays for expression $p. -p$ in Fig. 6 from the MATLAB software [12]. The presence of 3's in the output indicates that the expression is a contradiction and that there is no possible values for the formula that make it propositionally satisfied. Fig. 7 on the contrary does not include any 3's in the output arrays, this indicates that there are values that satisfy the expression $q + p > p + q$, which is in accordance to the output by the LoTREC software in Fig. 2 to 5.

The same analysis is done for expressions: $(p > -q) \cdot (p > q)$ and $p + -p$. Analogously, the other expressions in Table V. were compared and analyzed to identify whether they were propositionally satisfied or contradictions.

```
t1 =

       1     0     0     0     0

e.v == .
      -1     0     0     0     0

       1     0     0     0     0

       3     0     0     0     0

       3     0     0     0     0
```

Fig. 6. Output array by EZLogicUN of the expression $p. -p$.

```
t1 =

       0     1     0     0     0

e.v == +
       1     0     0     0     0

       0     1     0     0     0

       0     0     0     0     0

e.v == +
       0     0     0     0     0

       0     0     0     0     0

       0     0     0     0     0

       0     0     0     0     0
```

Fig. 7. Output array by EZLogicUN of the expression $q + p > p + q$.

Fig. 8. Output tree by LoTREC of the expression $(p > -q) \cdot (p > q)$.

Fig. 9. Pre-model 1 of the tree by LoTREC in Fig. 8 of the expression $(p > -q) \cdot (p > q)$.

Fig. 10. Pre-model 2.1 of the tree by LoTREC in Fig. 8 of the expression $(p > -q) \cdot (p > q)$.

Fig. 11. Pre-model 2.2 of the tree by LoTREC in Fig. 8 of the expression $(p > -q) \cdot (p > q)$.

Fig. 12. Pre-model 3 of the tree by LoTREC in Fig. 8 of the expression $(p > -q) \cdot (p > q)$.

Fig. 13. Output array by EZLoginUN for the expression $(p > -q) \cdot (p > q)$.

Fig. 14. Output array by EZLoginUN for the expression $p + -p$.

Eduardo Zurek, Mayra Zurbaran, Margarita Gamarra, and Pedro Wightman

VI. Conclusions

This paper has presented a novel approach for verifying the veracity of a zeroth order Boolean expression. The results obtained with this approach were compared to LoTREC. The implementation of the proposed approach requires simple data structures and its programming is straightforward due to the replacement policies based on the presented tables. Our method could improve the performance of decision-making systems based on logic sentences.

References

[1] J. T. Palma Méndez and R. L. Marín Morales, *Inteligencia artificial: técnicas, métodos y aplicaciones*. Aravaca: McGraw-Hill Interamericana de España, 2008.

[2] R. Jansana, "Propositional consequence relations and algebraic logic," in *The Stanford Encyclopedia of Philosophy*, 2011.

[3] M. D'Agostino and D. M. Gabbay, *Handbook of tableau methods*. Springer, 1999.

[4] I. Niemelä, "Implementing circumscription using a tableau method," in *ECAI*, 1996, pp. 80–84.

[5] P. Wolper, "The tableau method for temporal logic: An overview," *Logique et Analyse*, vol. 28, no. 110–111, pp. 119–136, 1985.

[6] T. A. Junttila and I. Niemelä, "Towards an efficient tableau method for Boolean circuit satisfiability checking," in *Computational Logic — CL 2000*, Springer, 2000, pp. 553–567.

[7] M. Y. Vardi, "Boolean satisfiability: Theory and engineering," *Commun. ACM*, vol. 57, no. 3, pp. 5, 2014.

[8] "Propositional Logic | Internet Encyclopedia of Philosophy".

[9] P. D. Magnus, "An Introduction to formal logic," *University at Albany: State University of New York*, 2008.

[10] A. J. A. Robinson and A. Voronkov, *Handbook of Automated Reasoning*. Elsevier, 2001.

[11] H. K. Büning and T. Lettmann, *Propositional Logic: Deduction and Algorithms*. Cambridge University Press, 1999.

[12] "MATLAB—The language of technical computing." [Online]. Available: http://www.mathworks.com/products/matlab/. [Accessed: 23-Jan-2015].

[13] L. F. del Cerro, D. Fauthoux, O. Gasquet, A. Herzig, D. Longin, and F. Massacci, "Lotrec: The generic tableau prover for modal and description logics," in *Automated Reasoning*, Springer, 2001, pp. 453–458.

[14] "Java Software | Oracle." [Online]. Available: https://www.oracle.com/java/index.html. [Accessed: 23-Jan-2015].

Identification of Central Points in Road Networks using Betweenness Centrality Combined with Traffic Demand

Rodrigo de Abreu Batista and Ana Lucia Cetertich Bazzan

Abstract—This paper aims to identify central points in road networks considering traffic demand. This is made with a variation of betweenness centrality. In this variation, the graph that corresponds to the road network is weighted according to the number of routes generated by the traffic demand. To test the proposed approach three networks have been created, which are Porto Alegre and Sioux Falls cities and a regular 10×10 grid. Then, trips were microscopically simulated and the results were compared with the proposed method.

Index Terms—Traffic assignment, complex networks, centrality, betweenness centrality.

I. Introduction

METROPOLITAN regions are currently facing major problems regarding urban mobility. For example, in 2014 Brazil had 78.1 million private vehicles, which represents an increase of 229.3% of fleet of vehicles. This means one private vehicle per 2.6 inhabitants in comparison to 2004.

Dealing with such growth in the fleet requires duly planning the highway system in order to reduce the effects of traffic on population and environment. Planning transportation systems involves, among other factors, analyzing the distribution of traffic flow throughout road networks.

There is evidence that the measure of betweenness centrality, as proposed by Freeman [1], is capable of predicting traffic flow in road networks. However, this measure ought to be adapted because it does not suitably represent the distribution of demand [2], [3], [4]. The objective of this paper is to analyze how betweenness centrality may be adapted and used in identifying central points in a road network. In this paper the term central points refers to those points that are most often traversed by road users when traveling along their routes.

II. Definitions

A. Transport Networks

Transport networks can be defined as networks that are composed of roadways and junctions between roadways (*e.g.*

intersections). Such networks are typically represented as weighted directed graphs $G = \{V, E\}$, composed of a set of vertices V (junctions) and edges E (roadways) and a cost function $C(e)$ which associates weight to each edge. In the context of transport networks, the length, the travel time or the capacity are commonly used as the weight of the edges.

Demand for traffic represents the behavior of users in using the network infrastructure. By behavior, it is understood the decisions made by the users that are relevant to the problem which is being modeled (*e.g.* choice of routes).

The locations of origin and destination of demand are grouped in districts. Districts can be defined based on information obtained through sociodemographic studies, data of georeferencing and urban statistics, so that these variations may be the least possible within a given district [5]. In general, and also in this work, each district is associated to a location within the network and it is composed of a set of vertices and edges without there being overlapping with other districts.

The demand of a network is commonly represented by a matrix that relates districts of origin to destinations, associating each possible combination, to a figure that corresponds to the intensity in which these trips occur. As it relates origins to destinations, it is called an origin-destination matrix, or OD matrix.

A driver who wishes to travel from district s to district t, represented by *(s,t)*, may use more than one series of edges that lead from s to t. Each these possible paths is called a route. Since each of these edges has an associated cost, there is particular interest in the path with the lowest cost, which in this article is called the path with the least length, or the shortest path. Therefore, creating the routes consists of associating series of edges to trips that are specified by the OD matrix.

B. Betweenness Centrality

Betweenness centrality is based on the idea that a vertex is more central as more low cost paths pass through it. The shortest paths between all pairs of vertices in the network are considered in this calculation. The traditional method for calculating betweenness, as well as other centrality measures, were originally developed in the scope of studying social networks, and they have recently been highlighted in the literature [6].

Betweenness for a vertex u is defined according to Equation 1, in which σ_{ij} is the number of shortest paths between i and j, and $\sigma_{ij,u}$ is the number of these paths through which u passes. The shortest paths are calculated on the basis of cost of the edges, thus centrality is sensitive to the function of cost that has been chosen.

$$B_u = \sum_{i \in V} \sum_{j \in V \setminus \{i\}} \frac{\sigma_{ij,u}}{\sigma_{ij}}. \tag{1}$$

III. Methods

This section presents a method for calculating betweenness by taking into consideration the demand that is going to use the network. The method consists of constructing a graph that represents the network and weighting it by using the demand. In order to calculate betweenness, the occupation of the roadways is taking into consideration and it results from the demand that is represented by the OD matrix.

A. Assigning Weights to the Edges

Before considering betweenness by considering the demand, it is necessary to determine the routes that correspond to the demand. As an OD matrix has only origins and destinations, calculating the routes is necessary for obtaining a table that adds up the amount of routes that pass through each edge. The routes were obtained by calculating the path of lowest cost between the origin and the destination of each trip that is presented in the OD matrix. That is why sequences of edges that form the paths with the lowest cost were found for each OD pair, by using the Dijkstra algorithm [7].

Once the routes that correspond to the demand have been calculated, it is possible to construct a table of the occurrence of the amount of routes which pass through each edge. This procedure is a basic stage for defining the weights of the edges in the network graph, since the amount of routes will serve as input for the functions used to calculate the cost of the edges. Weight to the edges can be carried out in distinct manners. Several studies use the length of the roadway as the cost of the edges [3], [8], [2], [9]. In this paper we have used the occupation rate of the roadways as the weight of the edges, as it is understood that it reflects the use of the network related to the demand. Furthermore, the unitary function, in which the lowest cost routes refer to the number of hops necessary to go from the origin to the destination, was also considered.

In this study, we decided to use decreasing cost functions to attribute costs to the edges. Figure 1 illustrates the situation that brought about this decision. A network in which 10 routes pass along each of the edges is shown in Figure 1a; the values of betweenness for each of the vertices are shown in the same figure. One may notice that the betweenness values are evenly distributed since this is a regular network.

Supposing that the number of trips between vertices A and B is increased by 5 trips, the natural logic is to increase the number of trips on the edge to 15. Figure 1b illustrates whats happens to the values of betweenness when the cost of edge

AB is increased. In this case, as the betweenness algorithm considers the paths of lowest cost, paths that previously passed through AB have ceased to do so. Thus, vertices C and D have come to receive the greatest betweenness values, while vertices A and B were those that received a real increase in demand. This would require the reader to use and inverted interpretation of the measure, so that the vertices with the lowest betweenness values are central in relation to the demand.

In order to solve this problem and have the greatest values of betweenness be attributed to the vertices with the greatest volume of demand, we decided to use decreasing cost functions. In this case, an increase of demand between vertices A and B causes a decrease in the weight of the edge, as this was attributed by a decreasing function. Figure 1c shows that in this case the vertices with the highest values of betweenness coincide with the vertices that have the highest demand.

The experiments were guided by taking the following cost functions into consideration:

F_1: **Decreasing Exponential** It is possible to use an exponential function for modeling the cost of an edge according to its occupation. Likewise, assuming that the cost of an edge also decreases exponentially in relation to its rate of occupation, the cost attributed to the edges is defined as function C, as defined in range $(0; 1]$. This weight is calculated in accordance with the decreasing exponential function that is shown in Equation 2, a particular case of the family of equations $y = a(1-b)^x$, and it only considers the amount of trips n that pass through a given edge.

$$C(n) = (1 - 0.001)^n. \tag{2}$$

F_2: **Rational Function** The rational cost function shown in Equation 3 was also considered in the experiments. This function was chosen to explore the behavior of betweenness distribution when cost decreases faster than the exponential function had previously explained.

$$C(n) = 1/n. \tag{3}$$

F_3: **Decreasing Linear Function** The linearly decreasing function exhibited in Equation 4, in which $k > n$, was also considered in the experiments. The main objective of using a linear cost function is to study the behavior of the proposed method when edge costs are diminishing more smoothly than the decreasing exponential function.

$$C(n) = k - n. \tag{4}$$

F_4: **Number of Hops** Considers the number of hops that were performed. This function represents the number of edges there are in the path with the lowest cost calculated for an OD pair.

F_5: **Length of an Edge** Considering that several studies have used the length of a roadway as weight of the edges, this was used in the cost function F_5 as a way to compare this study to previous studies.
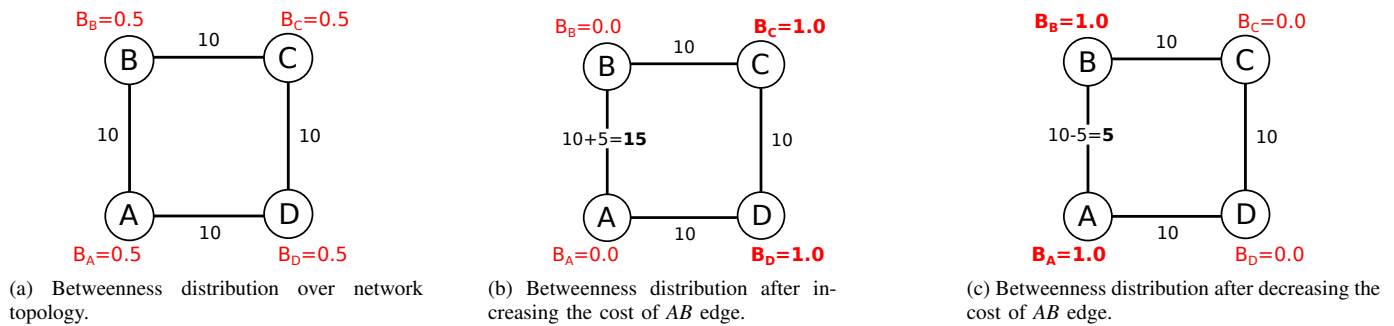
(a) Betweenness distribution over network topology.

(b) Betweenness distribution after increasing the cost of *AB* edge.

(c) Betweenness distribution after decreasing the cost of *AB* edge.

Fig. 1. Example of influence of the cost of the edges in betweenness.

B. Calculation of Betweenness Centrality Considering Demand

After the graph that represents roadways has been constructed and its edges have been properly weighted by using the routes generated from the OD matrix and cost function, it is possible to calculate the betweenness. As the interest here lies in identifying the central points, the results of the betweenness algorithm shows high values for vertices that have greater demand.

Algorithm 1 lists the steps involved in calculating betweenness centrality considering the demand.

Algorithm 1 Betweenness considering demand

Require: Network R, OD Matrix M, Cost Function C
Ensure: Betweenness from all vertices
1: **procedure** DEMANDBETWEENNESSCENTRALITY(R, M, C)
2:     Construct graph G with the same topology as R
3:     Calculate the routes from M over G
4:     Create a table T of routes passing through each edge
5:     Attribute costs to the edges of G by using T and C
6:     Calculate betweenness over G
7:     **return** betweenness from the vertices

IV. EXPERIMENTS AND RESULTS

The proposed approach was tested by means of experiments on three networks. Two of these are abstractions of real networks in the cities of Porto Alegre and Sioux Falls, and a third network consists of a regular 10x10 grid. All three networks are shown in Figure 2. Furthermore, with the aim of analyzing the behavior of betweenness at different occupation levels, demands with volumes of 10%, 25%, 50% and 75% of the total capacity of each network are used.

A. Traffic Demand

For the Porto Alegre Arterials network a pattern of demand was specified with the aim of reproducing the flow patterns of drivers that are observed in the city at the beginning of the day, in which they leave the outskirts of the city and go downtown. In this demand, which is called Non-Uniform Outskirts→Downtown Demand (NUODD), seven distinct

points in the outskirts of the city and one point in the central region of the city were used as origin and destination respectively. Considering that the capacity of the Porto Alegre Arterials network is 127,320 vehicles, demands of volume of 10% (12,372 trips), 25% (31,830 trips) and 50%[1] (63,660 trips) were defined.

For the Sioux Falls network, the same model of demand used in the paper by Chakirov and Fourie [10] was used. In their study the authors based their work on census data to create a de-aggregate demand and a microscopic model of the Sioux Falls network, based on the network that was originally used in the study by LeBlanc *et al.* [11]. In this study only the volume of demand corresponding to the morning rush hour was used. Thus, the demand used in this network has a volume of 44,652 trips and it was generated by an iterative model in order to achieve the stochastic user equilibrium. See [10] for details.

For the regular 10x10 grid two regions were determined— edge and center—on which three types of demand were defined. The first of these, uniform demand (UD), shows uniform distribution of the origins and destinations of the trips that were generated, and its aim is to create random trips within the 10x10 grid. The second, Non-Uniform Edge→Center Demand (NUECD), is composed of trips that go from the border toward the center of the grid, and which have the aim of creating congestion in the central region. The third type of demand which is called Non-Uniform Center→Edge Demand (NUCED), is composed of trips that leave the center and go toward the edge of the network. Considering that the grid has a capacity of 4,890 vehicles, demands of volume equivalent to 10% (489 trips), 25% (1,223 trips) and 50% (2,445 trips) are used.

B. Comparing the Proposed Method to a Microscopic Simulation

Since access to the real measurements that were carried out on roadways of the cities of Porto Alegre and Sioux Falls were not available, it was decided to test the proposed

---

[1]The total volume is equal to 50% of the maximum capacity of the network. Demands with volumes over 50% higher were not simulated.

Rodrigo de Abreu Batista and Ana Lucia Cetertich Bazzan

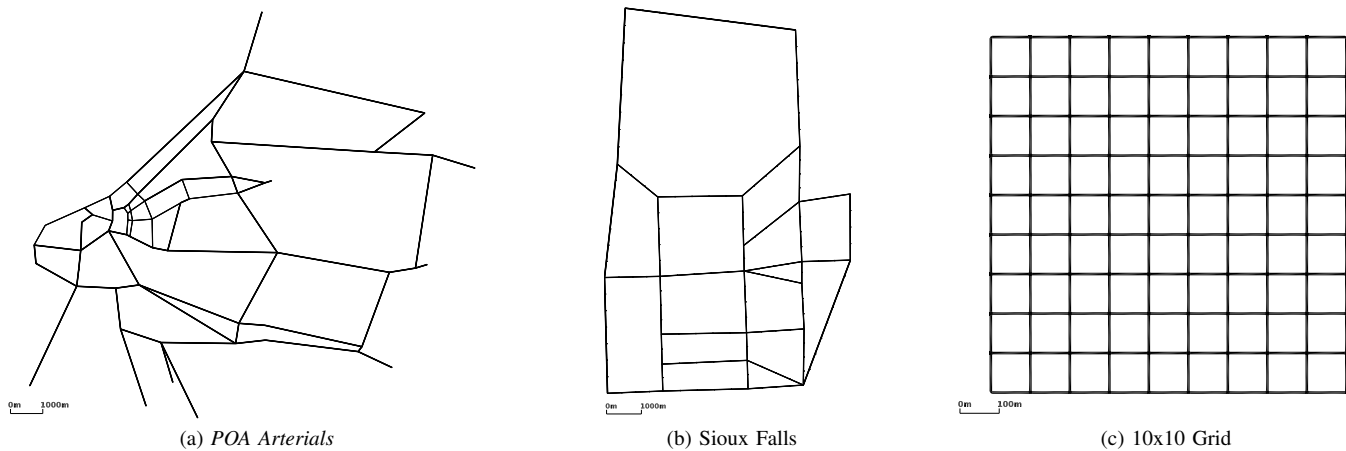| (a) *POA Arterials* | (b) Sioux Falls | (c) 10x10 Grid |

Fig. 2. Networks used in the experiments.

technique by means of comparing it to a simulation. In this case, a microscopic simulation performed in the SUMO [12] simulator was used, and it applies the routes calculated from an OD matrix to a given network. Figure 3 shows the steps involved in the microscopic simulation process and the steps of the proposed method. Both methods receive the OD matrix and the road network files as input and calculate the betweenness of the vertices at the end. For the microscopic simulation, additional steps for trips and routes generation are required to produce the SUMO related files.

By using SUMO it was possible to obtain information about occupation of the edges at the peak of the occupation of the network. Figure 4 shows the mean occupation curve of the Sioux Falls network along time, highlighting the time-step of the peak of mean occupation. The rates of occupation obtained were used to weight a graph that represents the network, on which betweenness which serves as a basis for comparing it to the model proposed was calculated. Table I shows the vertex-by-vertex details of betweenness that were calculated by the proposed method and the betweenness values calculated at the peak of occupation of the network. The five most significant values of each case study were highlighted to make visualization easier.

C. Results

With the aim of comparing the results of betweenness obtained by the proposed method to the results obtained through simulation, Pearson's correlation coefficient was used. Thus, the correlation between the results generated by the proposed method with the results obtained by the simulation were calculated for each of the experiments. This correlation was calculated between the results of betweenness over each set set of vertices. Table II shows the results that were detailed by the experiments and the cost function.

In the case of the Porto Alegre Arterials network, decreasing linear function and rational function showed the best results.

Considering that hops and edge length functions disregard demand, it is possible to note that even so, the former showed results that were significantly better than the latter. It is also possible to note that, for lower volumes, the correlation values obtained were greater, which may be attributed to the fact that microscopic simulation considers factors which the static model does not consider.

In the experiment on the Sioux Falls network, the decreasing exponential functions and the rational functions were those that showed the best results. This experiment was the one that showed the lowest rates of correlation. This can be attributed to the fact that the routes of this demand were generated by a different process than the others. In this case, the routes were generated by the model developed by Chakirov and Fourie [10], while in the other case studies, the routes were calculated by considering the shortest paths.

In the case of the 10x10 grid, hops and edge length functions were the ones that showed the best results. Specifically for this case study, the fact that shortest paths with same value exist between a given origin and destination showed a deviation that may have distorted the results. Another point to be noticed is the strong correspondence between the hops and edge length columns, due to the regularity of the grid, which makes the edge length function just as accurate to the hops function.

We observed that the decreasing exponential function and the decreasing linear function showed the best results when the instances of Sioux Falls and Porto Alegre Arterials were considered. As the occupation peak of the network is being considered, many edges have occupation rates that are near 1. The hops function also showed significant results, exceeding the others in some cases.

V. RELATED WORK

In Holme's paper [6], he investigates the relation between traffic flows in communication networks and centrality measures. In this model, particles are moved along the
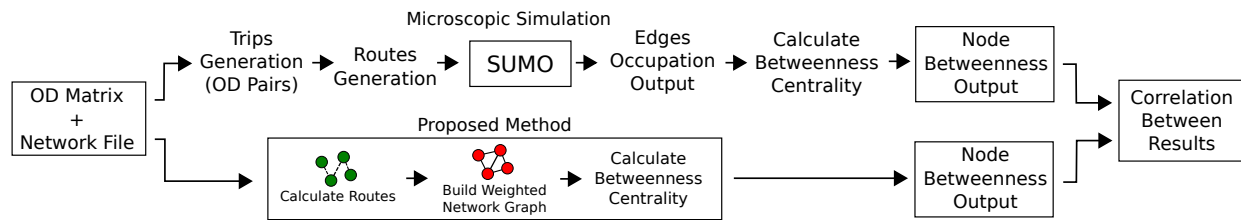
Fig. 3. Steps involved in the betweenness centrality calculation using the proposed method and the microscopic simulation.

TABLE I
BETWEENNESS CALCULATED BY USING DIFFERENT FUNCTIONS FOR ATTRIBUTING COST AND THE SIMULATION OVER THE SIOUX FALLS NETWORK, DETAILED BY VERTICES. THE TOP FIVE BETWEENNESS VALUES ARE SHOWN IN RED FOR EACH CASE.

Vertex	Cost Function					Simulation
	F_1	F_2	F_3	F_4	F_5	
4	19,294	17,562	17,038	19,117	17,006	12,907
5	6,554	10,559	15,610	18,557	24,134	7,572
6	15,091	18,700	15,519	15,242	16,639	14,279
8	18790	18808	12784	12961	13774	16894
9	13,674	13,623	19,573	20,462	27,655	14,953
10	42,980	37,882	36,784	30,539	33,664	31,824
11	35,269	33,075	31,314	25,068	23,418	42,001
14	4,657	6,185	9,400	12623	15,148	29,303
15	16,300	15,283	17,327	18,509	17,621	29,817
16	25,777	21,679	12,148	12,148	13,699	29,798
…	…	…	…	…	…	…

TABLE II
CORRELATION BETWEEN BETWEENNESS CENTRALITY CALCULATED BY THE PROPOSED METHOD AND SIMULATION, DETAILED BY EXPERIMENTS AND COST FUNCTIONS.

Instance	Cost Function				
	F_1	F_2	F_3	F_4	F_5
Sioux Falls	0.61	0.61	0.59	0.54	0.45
POA Arterials NUODD Vol. 10%	0.96	0.92	0.99	0.98	0.69
POA Arterials NUODD Vol. 25%	0.80	0.84	0.84	0.81	0.58
POA Arterials NUODD Vol. 50%	0.76	0.78	0.77	0.74	0.69
Grade 10x10 NUECD Vol. 10%	0.84	0.79	0.88	0.89	0.89
Grade 10x10 NUECD Vol. 25%	0.90	0.83	0.91	0.86	0.86
Grade 10x10 NUECD Vol. 50%	0.91	0.85	0.93	0.80	0.80
Grade 10x10 UD Vol. 10%	0.67	0.61	0.67	0.89	0.89
Grade 10x10 UD Vol. 25%	0.63	0.59	0.63	0.91	0.91
Grade 10x10 UD Vol. 50%	0.73	0.73	0.73	0.75	0.75
Grade 10x10 NUCED Vol. 10%	0.90	0.73	0.90	0.88	0.88
Grade 10x10 NUCED Vol. 25%	0.84	0.82	0.85	0.87	0.87
Grade 10x10 NUCED Vol. 50%	0.78	0.67	0.78	0.83	0.83

edges of a graph, constrained by the restriction that two particles may not occupy the same vertex at the same time. The particles move along between their randomly defined origins and destinations, and three different updating policies are considered, which are: random walk, in which particles randomly choose a position; detour-at-obstacle, in which a particle randomly chooses a position among their neighbors that are nearest the destination; and wait-at-obstacle, in which if no vertices are free near the destination, the particle does not move.

In order to monitor the traffic density regarding betweenness, the author chose the scale-free network model of Barabási-Albert, because it shows a wide distribution of betweenness values. Regarding betweenness, the author noted that the vertices with low or average betweenness rates showed steady occupation rates, and concluded that betweenness itself

cannot estimate the capacity of a vertex. At this point, our work differs from Holme's work since we use a microscopic simulation to compare to the betweenness. We also consider nonuniform demands that differ from randomly defined OD pairs used in Holme's work. Beyond that, the subject of study in our work was road networks, while in Holme's work the author focused on communication networks. It influences basically the network types studied: communication networks can be explained by scale-free models, while road networks can be better explained by random graphs.

In the study of Kazerani and Winter [2], the issue related to the capacity of betweenness for explaining traffic flows was analyzed. In their study, they came to the conclusion that the traditional betweenness measure [1] is unable to explain traffic flows significantly because it does not consider the traffic demand that flows in a network, nor its dynamics.
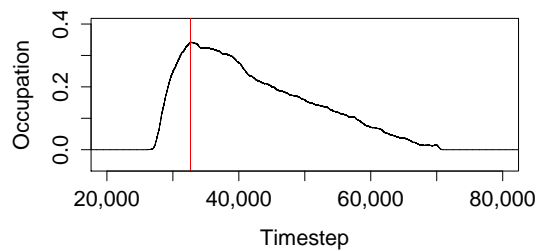
Fig. 4. Mean occupation of the Sioux Falls network related to time, with the peak of occupation occurring at the time-step 32,671.

Still regarding the study of Kazerani and Winter, the authors suggest that an adaptation of the traditional centrality measure is necessary, in which the physical and temporal aspects would be considered, so that a significant correlation with the traffic observed in the network may be attained. In this work, we proposed a variation of the betweenness that, by considering traffic demand, achieved higher correlation with the observed traffic in microscopic simulation.

In the study of Gao *et al.* [3], the authors investigate the capacity of betweenness to preview traffic flows by analyzing the correlation between the centrality measure and real traces collected from GPS. For that, data collected from GPS installed on 149 taxis of the city of Qingdao (China) were used. From this analysis, the authors concluded that the betweenness measure does not explains the traffic flows well, and they attribute it two main reasons: first, when calculating betweenness, origins and destinations of trips are vertices of the graph, while in real OD pairs origins and destinations are associated with edges; second, the OD pairs distribution is not uniform, being associated to the distribution of human activity, which, is influenced by factors like area occupation.

In a tentative to explain the differences between betweenness and traffic flows, the authors established a comparison between the centrality measure and a model developed in three steps: in the first step it is assumed that the demand occurred uniformly over geographic space; in the second, the model was extended to consider the distribution of human activity using data collected from the use of cellphones; in the third step, the demand model developed in the previous step was extended to support the distance decay factor, that models the behavior of people of seeking resources in nearest places.

When compared with betweenness values, the model developed in the second step has shown the greatest correlation. The study suggests that the betweenness presents low correlation with the model developed in the third step because it does not consider the distance decay factor. However the authors do not suggest modifications to the betweenness, which is known to consider uniform demand and disregard distance decay factors.

In this sense, we have also addressed the problem of the distance decay factor found in the model presented by Gao *et al.* by using an OD matrix. Since the OD matrix represents in fact the traffic demand, only the OD pairs related to the desired trips will be found in this matrix.

In the study of Galafassi and Bazzan [4] a betweenness variation that considers traffic demand is suggested. Differently from the metrics proposed by Freeman [1], in their study only the routes that belongs to the specified traffic demand were considered. The authors compare the correlation between the modified metrics of betweenness with the amount of waiting vehicles on the edges, and show that the proposed method explains traffic flows better than the original metrics. The experiments were executed over a 6x6 regular grid and Porto Alegre network, both considering different demand volumes and types (uniform vs nonuniform).

The method proposed here is an extension of the study developed by Galafassi and Bazzan. In this work we kept the betweenness calculation module unchanged, modifying the way of how weights are attributed to the edges to consider the traffic demand. We also compare the proposed method with the occupation of edges during simulation, instead of the waiting vehicles' queue and extended the experiments to consider Sioux Falls network.

VI. Conclusions and Future Work

The problem addressed along this work is central points identification in road networks using betweenness centrality. We have noticed that some authors tried to explain traffic flows using betweenness and failed because this metric itself assumes a uniform distribution of demand. Our method, on the other hand, consisted in combining the betweenness algorithm with the traffic demand so that higher values of betweenness were attributed to the vertices with higher demand.

The proposed method was tested in three networks and a microscopic simulation was performed for each one of them. The occupation of the edges was extracted from the simulations and the results were correlated with the betweenness values calculated by the proposed method. In general, the exponential decay and linear cost functions showed the best results among the studied functions.

The improvement in the proposed method was basically caused by two factors. First, the shortest paths calculated by the betweenness algorithm were influenced by the demand that uses a route. The second point is credited to the use of decreasing cost functions, which caused the weight of an edge to decrease as a function of the number of routes that pass through it.

Despite the fact that the technique proposed in this study is able to help identify central points in transport networks, it is only a small step towards a larger goal, which is to improve the road users' travel times. Hence, a possible extension of this study would be to assess whether traffic light operations at points with high values of betweenness would improve the average travel time for drivers, and whether these points are, in fact, the most critical ones.

Another aspect that could be investigated is the use a weighted correlation coefficient that is calculated considering the capacity of each vertex. Vertex capacity could be estimated by the capacities of its incident edges. Thus, vertices that have

large capacity would receive greater weight in the calculation of the correlation.

Another possible extension of this work is to analyze the occupation of links individually during the microscopic simulation, and to approximate a function that models its behavior. This function could be used in the algorithm proposed in this work so that its performance could be compared with other functions considered here.

ACKNOWLEDGMENT

REFERENCES

[1] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[2] A. Kazerani and S. Winter, "Can betweenness centrality explain traffic flow," in *Proceedings of the 12th AGILE International Conference on GIS*, 2009.

[3] S. Gao, Y. Wang, Y. Gao, and Y. Liu, "Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality," *Environment and Planning B: Planning and Design*, vol. 40, no. 1, pp. 135–153, 2013.

[4] C. Galafassi and A. L. C. Bazzan, "Analysis of traffic behavior in regular grid and real world networks," in *Proceedings of the Fifth International Workshop on Emergent Intelligence on Networked Agents (WEIN)*, 2013. [Online]. Available: http://www.inf.ufrgs.br/maslab/pergamus/pubs/GalafassiBazzan2013-wein.pdf

[5] A. L. Bazzan and F. Klügl, "Introduction to intelligent systems in traffic and transportation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 7, no. 3, pp. 1–137, 2013.

[6] P. Holme, "Congestion and centrality in traffic flow on complex networks," *Advances in Complex Systems*, vol. 6, no. 2, pp. 163–176, 2003.

[7] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[8] P. Crucitti, V. Latora, and S. Porta, "Centrality in networks of urban streets," *Chaos: an interdisciplinary journal of nonlinear science*, vol. 16, no. 1, p. 015113, 2006.

[9] S. Derrible and C. Kennedy, "Applications of graph theory and network science to transit network design," *Transport reviews*, vol. 31, no. 4, pp. 495–519, 2011.

[10] A. Chakirov and P. Fourie, "Enriched sioux falls scenario with dynamic and disaggregate demand. working paper," Future Cities Laboratory, Singapore-ETH Centre (SEC), Singapore, Tech. Rep., 2014.

[11] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, "An efficient approach to solving the road network equilibrium traffic assignment problem," *Transportation Research*, vol. 9, no. 5, pp. 309–318, 1975.

[12] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo-simulation of urban mobility-an overview," in *SIMUL 2011, The Third International Conference on Advances in System Simulation*, 2011, pp. 55–60.

Project Scheduling: A Memetic Algorithm with Diversity-Adaptive Components that Optimizes the Effectiveness of Human Resources

Virginia Yannibelli and Analía Amandi

Abstract—In this paper, a project scheduling problem is addressed. This problem supposes valuable assumptions about the effectiveness of human resources, and also considers a priority optimization objective for project managers. This objective is optimizing the effectiveness levels of the sets of human resources defined for the project activities. A memetic algorithm is proposed for solving the addressed problem. This memetic algorithm incorporates diversity-adaptive components into the framework of an evolutionary algorithm. The incorporation of these components is meant for improving the performance of the evolutionary-based search, in both exploitation and exploration. The performance of the memetic algorithm on instance sets with different complexity levels is compared with those of the heuristic search and optimization algorithms reported until now in the literature for the addressed problem. The results obtained from the performance comparison indicate that the memetic algorithm significantly outperforms the algorithms previously reported.

Index Terms—Project scheduling, human resource assignment, multi-skilled resources, memetic algorithms, evolutionary algorithms, simulated annealing algorithms.

I. INTRODUCTION

IN MOST companies and organizations, project scheduling is recognized as a really central, difficult and time-consuming task [1, 2].

A project scheduling problem usually implies defining feasible start times (i.e., the precedence relations between the project activities must not be violated) and feasible human resource assignments (i.e., the human resource requirements of project activities must be satisfied) for project activities so that a given optimization objective is reached. In this context, the available knowledge of the effectiveness of human resources in relation to project activities is considered in order to define feasible human resource assignments for project activities. This is really important because both the development and the results of project activities depend

mainly on the effectiveness of the human resources assigned to such activities [1, 2].

In the past 30 years, different kinds of project scheduling problems have been formally described and also addressed in the literature. However, to the best of the authors' knowledge, only few project scheduling problems take into consideration human resources with different effectiveness levels [3, 4, 5, 6, 10, 22, 23, 26], a main characteristic of real project scheduling problems. Such project scheduling problems supposes very different assumptions about the effectiveness of human resources.

The project scheduling problems formally described in [3, 4, 5, 26] suppose that human resources master one or several skills and have an effectiveness level for each mastered skill. Then, the effectiveness level of a human resource in a given project activity is determined by considering only the effectiveness level of the human resource in respect of one of the skills required for such activity. Therefore, it is supposed that the effectiveness level of a human resource depends only on their skills.

Unlike the project scheduling problems above mentioned, the project scheduling problem described in [6] suppose that the effectiveness level of a human resource depends on several factors inherent to its work context (i.e., the project activity to which the resource is assigned, the skill to which the resource is assigned within the project activity, the set of human resources assigned to the project activity, and the attributes of the resource). This assumption concerning the effectiveness levels of human resources is really important since human resources usually have very different effectiveness levels regarding different work contexts, and thus, the effectiveness level of a human resource is typically determined in respect of the factors inherent to its work context [1, 2]. In addition to this assumption, the project scheduling problem described in [6] considers a priority optimization objective for project managers. This objective implies optimizing the effectiveness of the sets of human resources defined for the project activities.

The project scheduling problem described in [6] is a variant of the RCPSP (Resource Constrained Project Scheduling Problem) [9] and so is recognized as an NP-Hard optimization problem. For this reason, heuristic search and optimization

algorithms are required to solve problem instances with different complexity levels in a reasonable period of time.

In this paper, a memetic algorithm is proposed to solve the project scheduling problem described in [6]. This memetic algorithm incorporates diversity-adaptive components into the framework of an evolutionary algorithm. The incorporation of these diversity-adaptive components is meant for improving the performance of the evolutionary-based search [18, 19, 20].

This memetic algorithm is proposed for solving the problem described in [6] mainly because of the following. Memetic algorithms with diversity-adaptive components have been proven to be really effective in the resolution of a variety of NP-Hard optimization problems, and also have been proven to be much more effective than evolutionary algorithms and memetic algorithms with non diversity-adaptive components in the resolution of different NP-Hard optimization problems [18, 19, 20]. Therefore, the memetic algorithm proposed could outperform the heuristic search and optimization algorithms reported until now in the literature for solving the problem. To the best of the authors' knowledge, three heuristic search and optimization algorithms have been reported until now for solving the problem: a traditional evolutionary algorithm [6], a memetic algorithm [7] which incorporates a hill-climbing algorithm within the framework of an evolutionary algorithm, and a hybrid evolutionary algorithm [8] which incorporates a simulated annealing algorithm within the framework of an evolutionary algorithm.

The remainder of the paper is organized as follows. In Section 2, a review of reported project scheduling problems which consider the effectiveness of human resources is presented. In Section 3, the project scheduling problem addressed here is described. In Section 4, the memetic algorithm proposed for the problem is presented. In Section 5, the computational experiments carried out to evaluate the performance of the memetic algorithm are presented and also an analysis of the results obtained. Finally, In Section 6, the conclusions of the present work are presented.

II. Related Works

In the past 30 years, different kinds of project scheduling problems have been formally described and addressed in the literature. However, to the best of the authors' knowledge, only few of these project scheduling problems take into account human resources with different effectiveness levels [3, 4, 5, 6, 10, 22, 23, 26], a main characteristic of real project scheduling problems. Such project scheduling problems suppose very different assumptions about the effectiveness of human resources. This section reviews the assumptions supposed in different project scheduling problems described in the literature.

In [12, 13, 14, 17, 25], multi-skill project scheduling problems with different optimization objectives are formally described. In these problems, project activities require a given

set of skills for their development, and also a given number of human resources for each skill into such set. The human resources available for the project activities master one or several skills, and these problems supposes that human resources that master a given skill have the same effectiveness level in respect of such skill.

In [15, 16, 24], skilled workforce project scheduling problems with different optimization objectives are formally described. In these problems, each project activity requires only one human resource with a given skill. Besides, the human resources available for project activities master one or several skills. These problems suppose that human resources that master a given skill have the same effectiveness level in respect of such skill.

In [3], a multi-skill project scheduling problem is formally described which considers hierarchical levels of skills. In this regard, this problem supposes that human resources that master a given skill have different effectiveness levels with respect to such skill. Besides, the project activities of this problem require a given set of skills for their development, a given minimum level of effectiveness for each skill of the set, and a given amount of human resources for each pair of skill and level. Then, this problem supposes that the human resource sets feasible for a given project activity have the same effectiveness level regarding the development of such activity.

In [4, 5, 26], multi-skill project scheduling problems with different optimization objective sets are formally described. In these problems, most project activities require only one human resource with a given skill. The human resources available for project activities master one or several skills. These problems suppose that human resources that master a given skill have different effectiveness levels in relation to such skill. Besides, these problems suppose that the effectiveness level of a human resource in a given activity only depends on the effectiveness level of the human resource in respect of the skill required for such activity.

Unlike the problems previously mentioned, the project scheduling problem formally described in [6] supposes that the effectiveness of a human resource depends on several factors inherent to its work context, and then different effectiveness levels can de defined for each human resource in relation to different work contexts. This assumption concerning the effectiveness of human resources is really important since human resources usually have very different effectiveness levels regarding different work contexts, and thus, the effectiveness level of a human resource is typically determined in respect of the factors inherent to its work context [1, 2]. Considering the previously mentioned, the project scheduling problem described in [6] supposes important assumptions concerning the effectiveness of human resources in the context of project scheduling problems.

94

III. PROJECT SCHEDULING PROBLEM DESCRIPTION

In this paper, the project scheduling problem introduced in [6] is addressed. A description of this project scheduling problem is presented below.

A project contains a set A of N activities, $A = \{1, \dots, N\}$, that has to be scheduled. Specifically, a starting time and a human resource set have to be defined for each project activity of the set A. The duration, human resource requirements, and precedence relations of each project activity are known.

The duration of each project activity j is notated as d_j. Besides, it is considered that pre-emption of project activities is not allowed. This means that, when a project activity starts, it must be developed period by period until it is completed. Specifically, the d_j periods of time must be consecutive.

Among the project activities, there are precedence relations. This is because usually each project activity requires results generated by other project activities. Thus, the precedence relations establish that each project activity j cannot start until all its immediate predecessors, given by the set P_j, have completely finished.

To be developed, project activities require human resources skilled in different knowledge areas. Specifically, each project activity requires one or several skills and also a given number of human resources for each skill required.

It is considered that a qualified workforce is available to develop the activities of the project. This workforce is made up of a number of human resources, and each human resource masters one or several skills.

Set SK contains the K skills required in order to develop the activities of the project, $SK = \{1,\dots, K\}$, and set AR_k contains the available human resources with skill k. Then, the term $r_{j,k}$ represents the number of human resources with skill k required for activity j of the project. The values of the terms $r_{j,k}$ are known for each project activity.

It is considered that a human resource cannot take over more than one skill within a given activity, and also a human resource cannot be assigned more than one activity at the same time.

Based on the assumptions previously mentioned, a human resource can be assigned different project activities but not at the same time, can take over different skills required for a project activity but not simultaneously, and can belong to different possible sets of human resources for each activity.

Therefore, different work contexts can be defined for each available human resource. It is considered that the work context of a human resource r, denoted as $C_{r,j,k,g}$, is made up of four main components. In this respect, the first component refers to the project activity j which r is assigned (i.e., the complexity of j, the domain of j, etc.). The second component refers to the skill k which r is assigned within project activity j (i.e., the tasks associated to k within j). The third component is the set of human resources g that has been assigned j and that includes r (i.e., r must work collaboratively with the other

human resources assigned to j). The fourth component refers to the attributes of r (i.e., his or her educational level regarding different knowledge areas, his or her level regarding different skills, his or her experience level regarding different tasks and domains, the kind of labor relation between r and the other human resources of g, etc.). In respect of the attributes of r, it is considered that these attributes could be quantified from available information about r (e.g., curriculum vitae of r, results obtained from evaluations made to r, information about the participation of r in already executed projects, etc.).

The four components previously mentioned are considered the main factors that determine the effectiveness level of a human resource. Because of this, it is assumed that the effectiveness level of a human resource depends on all the components of his or her work context. Then, different effectiveness levels can be considered for each human resource in respect of different work contexts.

The effectiveness level of a human resource r, in respect of a possible context $C_{r,j,k,g}$ for r, is notated as $e_{rCr,j,k,g}$. The term $e_{rCr,j,k,g}$ refers to how well r can take over, within activity j, the tasks associated to skill k, considering that r must work collaboratively with the other human resources of set g. The term $e_{rCr,j,k,g}$ takes a real value over the range $[0, 1]$. The values of the terms $e_{rCr,j,k,g}$ inherent to each human resource available for the project are known. It is considered that these values could be obtained from available information regarding the participation of the human resources in already carried out projects.

The problem of scheduling a project involves to determine feasible start times (i.e., the precedence relations between the project activities must not be violated) and feasible human resource assignments (i.e., the human resource requirements of project activities must be met) for project activities so that the optimization objective is reached. In this respect, an optimization objective priority for project managers is considered. This optimization objective implies optimizing the effectiveness of the sets of human resources assigned to the project activities. This objective is modeled by Formulas (1) and (2):

$$\max_{\forall s \in S} \left(e(s) = \sum_{j=1}^{N} e_{R(j,s)} \right) \tag{1}$$

$$e_{R(j,s)} = \frac{\sum_{r=1}^{|R(j,s)|} e_{rC_{r,j,k(r,j,s),R(j,s)}}}{|R(j,s)|} \tag{2}$$

Formula (1) optimizes the effectiveness of the sets of human resources assigned to the N project activities. In this formula, set S contains all the feasible schedules for the project in question. The term $e(s)$ refers to the effectiveness level of the sets of human resources assigned to the project activities by schedule s. The term $R(j,s)$ refers to the set of

human resources assigned to activity j by schedule s. The term $e_{R(j,s)}$ refers to the effectiveness level corresponding to $R(j,s)$.

Formula (2) estimates the effectiveness level of the set of human resources $R(j,s)$. This effectiveness level is estimated by calculating the mean effectiveness level of the human resources belonging to $R(j,s)$.

For a more detailed discussion of the project scheduling problem described here and, in particular, of Formulas (1) and (2), readers are referred to the work [6] which has introduced this problem.

IV. MEMETIC ALGORITHM WITH DIVERSITY-ADAPTIVE COMPONENTS

A memetic algorithm is proposed in order to solve the project scheduling problem addressed here. This memetic algorithm incorporates diversity-adaptive components into the framework of an evolutionary algorithm. The incorporation of these diversity-adaptive components is meant for improving the performance of the evolutionary-based search [18, 19, 20].

In the next sections, the general behavior and the main components of the memetic algorithm are described.

A. General Behavior of the Memetic Algorithm

Fig. 1 describes the general behavior of the memetic algorithm.

Memetic Algorithm

inputs: population_size, number_generations, AP_{cLA}, AP_{cUA}, AP_{mLA}, AP_{mUA}, λ (replacement factor), number_iterations, α (cooling factor)

outputs: best_solution

procedure:
1: population = generate_initial_population(population_size);
2: generation = 1;
3: **while** (generation $\leq$ number_generations) **do**
4:     mating_pool = parent_selection_process(population);
5:     offsprings = crossover_process(mating_pool, AP_{cLA}, AP_{cUA})
6:     mutation_process(offsprings, AP_{mLA}, AP_{mUA})
7:     population = survival_selection_process(population, offsprings, λ)
8:     simulated_annealing_stage(population, number_iterations, α)
9:     generation = generation + 1
10: **end while**
11: best_solution = get_best_solution_from(population)
12: **return** best_solution

Fig. 1. Main behavior of the memetic algorithm

As seen in Fig. 1, the memetic algorithm is an iterative process. This process starts from an initial population of solutions. Each solution of this population encodes a feasible schedule for the project to be scheduled. In addition, each solution has a fitness value which represents the quality of the related schedule with respect of the optimization objective of the addressed project scheduling problem. As was mentioned

in Section III, this optimization objective involves optimizing the effectiveness of the sets of human resources assigned to the project activities. The iterative process ends when a given number of generations is reached. After this happens, the iterative process provides the best solution of the last population or generation as a solution to the project scheduling problem.

In each iteration, the memetic algorithm develops the following stages. First, a parent selection process is used in order to determine which solutions of the current population will compose the mating pool. The solutions of the current population with the best fitness values will have more probability of being selected.

After the mating pool is composed, the solutions in the mating pool are paired, and a crossover process is applied to each pair of solutions with a diversity-adaptive probability AP_c in order to generate new feasible ones.

Then, a mutation process is applied to each solution generated by the crossover process, with a diversity-adaptive probability AP_m. The mutation process is applied in order to introduce diversity in the new solutions generated by the crossover process.

Then, a survival selection process is used in order to define which solutions from the solutions in the current population and the solutions generated from the mating pool will compose the new population.

Finally, a diversity-adaptive simulated annealing algorithm is applied to each solution of the new population, except to the best solution of this population which is maintained into this population. Thus, the simulated annealing algorithm modifies the solutions of the new population.

B. Components of the Memetic Algorithm

In the next sections, the main components of the memetic algorithm are described. These components are: the encoding and decoding of solutions, the fitness function, the parent selection process, the crossover process, the mutation process, the survival selection process, and the simulated annealing algorithm.

1) Encoding and Decoding of Solutions

To encode the solutions of the population, the encoding introduced in [6] for project schedules was used. By this encoding, each solution is encoded by two lists with a length equal to N, considering that N is the number of activities in the project to be scheduled.

The first list is a traditional activity list. Each position on this list contains a different activity j of the project. Each activity j of the project can appear on this list in any position higher than the positions of all its predecessor activities. The activity list represents a feasible order in which the activities of the project can be added to the schedule.

The second list is an assigned resources list. This list contains information about the human resources of each skill k

Project Scheduling: A Memetic Algorithm with Diversity-Adaptive Components that Optimizes the Effectiveness of Human Resources

ISSN 2395-8618

assigned to each activity of the project. Specifically, position j on this list contains a detail about the human resources of each skill k assigned to activity j of the project.

To build or decode the schedule related to the encoding above-described, the serial schedule generation method presented in [6] was used. This method adds the N project activities in the schedule one by one, considering the order defined by the activity list. For each project activity to be added in the schedule, this method defines the earliest feasible starting time. In this sense, the method considers that a project activity can start once all its predecessors have been completed and when all the human resources assigned to the project activity are available. Thus, the schedule obtained by this method from the encoding is always a feasible one.

With respect of the use of the serial schedule generation method, it is important to note that only one schedule can be obtained from a given encoded solution, however different encoded solutions could lead to the same schedule.

To generate the encoded solutions of the initial population considering the encoding previously-described, the random generation process introduced in [6] was used. By using this process, a very diverse initial population is obtained. This is meant in order to prevent the premature convergence of the evolutionary-based search developed by the algorithm.

2) Fitness Function

The fitness function is utilized in order to determine the fitness values of the encoded solutions. The fitness value of an encoded solution represents the quality of the related schedule with respect of the optimization objective of the addressed project scheduling problem. As mentioned in Section III, this optimization objective involves optimizing the effectiveness of the sets of human resources assigned to the project activities.

The detailed behavior of the fitness function is described as follows. Considering a given encoded solution i, the fitness function decodes the schedule s from the solution i by using the serial schedule generation method described in Section IV.B.1. Then, the fitness function calculates the value of the term $e(s)$ corresponding to s (Formulas (1) and (2)). This value defines the fitness value $f(i)$ of the solution i.

To calculate the value of term $e(s)$, the fitness function uses the values of the terms $e_{rCr,j,k,g}$ inherent to s (Formula 2). As was mentioned in Section III, the values of the terms $e_{rCr,j,k,g}$ inherent to each available human resource r are known.

Note that the term $e(s)$ takes a real value over $[0, …, N]$.

3) Parent Selection Process

The parent selection process is applied in order to determine which solutions of the current population will compose the mating pool. Then, the solutions in the mating pool, named parent solutions, will be utilized by the crossover process to generate new feasible solutions, named offspring solutions.

In order to carry out the parent selection process on the current population, the roulette wheel selection process [18] was applied here. This is one of the parent selection processes most applied in the literature [18].

The roulette wheel selection process considers the fitness values of the solutions of the current population and also is biased by a random factor. Thus, the solutions of the current population with the best fitness values will have more chances of being selected and so incorporated in the mating pool.

The roulette wheel selection process works as follows. Given the solutions of the current population, this process defines a selection probability $p(i)$ for each solution i of the current population by Formula (3), where $f(i)$ is the fitness value of solution i and M is the number of solutions into the current population.

Once defined the selection probabilities of the M solutions of the current population, the process calculates the cumulative probability vector v corresponding. Specifically, the process creates an empty vector v with a length equal to M, considering that position i on this vector v represents to solution i. Then, the process calculates a value $v(i)$ for each position i on vector v by Formula (4):

$$p(i) = \frac{f(i)}{\sum_{i=1}^{M} f(i)} \qquad (3)$$

$$v(i) = \begin{cases} p(i) & i = 1 \\ v(i-1) + p(i) & i = 2,…, M \end{cases} \qquad (4)$$

After the cumulative probability vector v is calculated, the process can start to select solutions from the M solutions of the current population.

To select one solution from the M possible solutions, the process randomly choose a real value p over the range $[0, 1]$. Then, the process goes through vector v in order to find the first position i on this vector with a value higher than p. Once such position i is found, solution i is considered by the process as the selected solution.

The above-described operation is repeated by the process until M solutions are selected to compose the mating pool.

4) Diversity-Adaptive Crossover Process

Once composed the mating pool, the solutions in the mating pool are paired, taking into account the order in which these solutions were incorporated into the mating pool. Then, a crossover process is applied to each of these pairs of solutions with a diversity-adaptive probability AP_c, in order to generate new feasible solutions. The crossover process applied here is described below.

Considering a given pair of solutions ($p1$ and $p2$) to be recombined, the crossover process develops the next stages. First, a crossover process feasible for activity lists is applied to the activity lists of $p1$ and $p2$, in order to generate two new

activity lists. The first of these new activity lists is assigned to the first offspring solution ($o1$) of the pair, and the second of these new activity lists is assigned to the second offspring solution ($o2$) of the pair. Then, a crossover process feasible for assigned resources lists is applied to the assigned resources lists of $p1$ and $p2$, in order to generate two new assigned resources lists. The first of these new assigned resources lists is assigned to the first offspring solution ($o1$) of the pair, and the second of these new assigned resources lists is assigned to the second offspring solution ($o2$) of the pair. Therefore, the crossover process generates two new solutions ($o1$ and $o2$) from the pair of solutions ($p1$ and $p2$).

With respect of the crossover process applied to the activity lists of $p1$ and $p2$, the two-point crossover process for activity lists [21] was applied. This process works as follows. First, the process randomly chooses two crossover points $x1$ and $x2$, considering that $1 \leq x1 < x2 < N$. After that, the activities on positions $[1, x1]$ of the list of $p1$ are positioned in the positions $[1, x1]$ of the list of $o1$, in the same order. Then, the process selects the first ($x2 - x1$) activities of the list of $p2$ that are not included in the list of $o1$, and copies these activities in positions $[x1+1, x2]$ of the list of $o1$, considering the order in which these activities appear in the list of $p2$. Finally, the process selects the ($N - x2$) activities of the list of $p1$ that are not included in the list of $o1$, and copies these activities in positions $[x2+1, N]$ of the list of $o1$, considering the order in which these activities appear in the list of $p1$. Thus, the process generates the activity list of $o1$ from the activity lists of $p1$ and $p2$.

The generation of the activity list of $o2$ is similar to the generation of the activity list of $o1$. However, the roles of $p1$ and $p2$ are inverted to generate the list of $o2$.

With respect of the crossover process applied to the assigned resources lists of $p1$ and $p2$, the uniform crossover process [18] was applied. This process works as follows. First, the process creates an empty vector u with a length equal to N. Then, the process randomly defines a real value on $[0, 1]$ for each position i of the vector u, considering $i = 1,…,N$. After that, the process uses the vector u to define the assigned resources lists of $o1$ and $o2$. Specifically, for each position i, if $u(i) \leq 0.5$, the resource assignment for position i of the list of $o1$ ($o2$) is inherited from $p1$ ($p2$). Otherwise, if $u(i) > 0.5$, the resource assignment for position i of the list of $o1$ ($o2$) is inherited from $p2$ ($p1$).

The described crossover process is applied with a diversity-adaptive probability AP_c. In this respect, the diversity-adaptive crossover probability proposed in [11] was considered. This diversity-adaptive crossover probability is defined by Formula (5), where f_{max} is the maximal fitness of the current population, f_{avg} is the average fitness of the current population, and the term ($f_{max} - f_{avg}$) is considered as a measure of the diversity of the current population. Then, f' is the higher fitness of the two solutions to be recombined, and AP_{cLA} and

AP_{cUA} are given values for the crossover probability, considering $0 \leq AP_{cLA} \leq 1$ and $0 \leq AP_{cUA} \leq 1$.

$$AP_c = \begin{cases} \dfrac{AP_{cUA}\ (f_{max} - f')}{(f_{max} - f_{avg})} & f' \geq f_{avg} \\ AP_{cLA} & f' < f_{avg} \end{cases} \quad (5)$$

By Formula (5), the crossover probability AP_c is adaptive according to the diversity of the current population. In this respect, when the diversity of the current population reduces, AP_c is increased in order to promote the exploration of the search space and thus to prevent the premature convergence of the evolutionary search developed by the memetic algorithm. Otherwise, when the current population is very diverse, AP_c is decreased in order to promote the exploitation of the search space. Thus, probability AP_c is adaptive according to the diversity of the current population, in order to promote either the exploitation or exploration of the search space.

5) Diversity-Adaptive Mutation Process

A mutation process is applied to each one of the solutions obtained by the crossover process, with a diversity-adaptive probability AP_m. The mutation process applied is described below.

Considering a given solution $p1$ to be mutated, the mutation process develops the next stages. First, a mutation process feasible for activity lists is applied to the activity list of $p1$. Thus, the activity list of $p1$ is mutated. Then, a mutation process feasible for assigned resources lists is applied to the assigned resources list of $p1$. Thus, the assigned resources list of $p1$ is mutated. By applying the described mutation process, solution $p1$ is mutated.

With respect of the mutation process applied to the activity list of $p1$, the adjacent pairwise interchange process [21] is applied. This mutation process works as follows. For each position $i = 1,…, N$-1, the process swaps the activities on positions i and $i+1$ with a diversity-adaptive probability AP_m, considering that these activities can be swapped only when the activity on position i is not a predecessor of the activity on position $i+1$.

With respect of the mutation process applied to the assigned resources list of $p1$, the random resetting process [18] is applied. This mutation process works as follows. For each position $i = 1,…, N$, the process defines a new resource assignment with a diversity-adaptive probability AP_m.

In relation to the diversity-adaptive probability AP_m utilized by the mutation process, the diversity-adaptive mutation probability proposed in [11] was considered. This diversity-adaptive mutation probability is defined by Formula (6), where f_{max} is the maximal fitness of the current population, f_{avg} is the average fitness of the current population, and ($f_{max} - f_{avg}$) is a measure of the diversity of the current population. Then, f'' is the fitness of the solution to be mutated, and AP_{mLA} and

AP_{mUA} are predefined values for the mutation probability, considering $0 \leq AP_{mLA}, AP_{mUA} \leq 1$.

$$AP_m = \begin{cases} \dfrac{AP_{mUA} \ (f_{max} - f'')}{(f_{max} - f_{avg})} & f'' \geq f_{avg} \\ AP_{mLA} & f'' < f_{avg} \end{cases} \quad (6)$$

By Formula (6), the mutation probability AP_m is adaptive based on the diversity of the current population. Specifically, when the diversity of the current population decreases, AP_m is increased, promoting the exploration of the search space and thus avoiding the premature stagnation of the evolutionary search developed by the memetic algorithm. In contrast, when the current population is diverse, AP_m is decreased, promoting the exploitation of the search space. Thus, probability AP_m is adaptive based on the diversity of the current population, to promote either the exploitation or exploration of the search space.

6) Survival Selection Process

The survival selection process is used in order to determine which solutions from the solutions in the current population and the solutions generated from the mating pool will compose the new population.

In order to carry out the survival selection process, the fitness-based steady-state selection process [18] was applied here. By applying this process, the best solutions achieved by the memetic algorithm throughout the evolutionary search are preserved [18].

The fitness-based steady-state selection process works as follows. First, the process sorts the M solutions of the current population according to their fitness values. After that, the process orders the M solutions generated from the mating pool, according to their fitness values.

Then, the process selects the best $(M - \lambda)$ solutions from the current population, where λ is a parameter that takes an integer value over the range $[1, M-1]$. After that, the process selects the best λ solutions from the solutions generated from the mating pool by the crossover and mutation processes.

Finally, the M selected solutions are used by the process to compose the new population.

7) Diversity-Adaptive Simulated Annealing Algorithm

Once a new population is obtained by the survival selection process, a diversity-adaptive simulated annealing algorithm is applied to each solution of this population, excepting the best solution of this population. The best solution of the population is maintained into the population.

This diversity-adaptive simulated annealing algorithm is a variant of the simulated annealing algorithm described in [8].

Fig. 2 below presents the general behavior of the diversity-adaptive simulated annealing algorithm.

As seen in Fig. 2, the diversity-adaptive simulated annealing algorithm is an iterative process. This process starts from a given encoded solution s and a given initial value t_i for the temperature parameter. The iterative process ends when a given number of iterations is reached.

Simulated Annealing Algorithm
inputs: *solution_s*, *temperature_ti*, number_iterations,   α (cooling factor)
outputs: new_solution
procedure:
1: s = *solution_s*;
2: t = *temperature_ti*;
3: c = 1;
4: **while** ((t > 0) and (c ≤ number_iterations)) **do**
5:    s' = generate_new_solution_from(s);
6:    **if** (fitness(s) < fitness(s')) **then**
7:      s = s';
8:    **else**
9:      Δ = fitness(s) - fitness(s');
10:      acceptance_probability = exp(-Δ/t);
11:      x = random(0, 1);
12:      **if** (x < acceptance_probability) **then**
13:       s = s';
14:      **end if**
15:    **end if**
16:    t = t x α;
17:    c = c + 1;
18: **end while**
19: new_solution = s;
20: **return** new_solution;

Fig. 2. Main behavior of the simulated annealing algorithm

In each iteration, the process generates a new encoded solution s' from the encoded solution s by applying a move operator. Then, the process analyzes if the solution s should be replaced by the new solution s'. When the fitness value of the solution s is lower than that of the solution s', the process replaces to the solution s by the solution s'. Otherwise, when the fitness value of the solution s is higher than or equal to that of the solution s', the process replaces to the solution s by the solution s' with an acceptance probability $exp(-\Delta/t)$. Note that the term Δ represents the difference between the fitness value of the solution s and the fitness value of the solution s', and the term t represents the current value of the temperature parameter. The used acceptance probability is proportional to the current value of the temperature parameter. At the end of each iteration, the current value of the temperature parameter is reduced by a given cooling factor α.

To apply the described simulated annealing algorithm to the solutions of the population obtained by the survival selection process, an initial value t_i is defined for the temperature parameter. This value t_i is defined based on the diversity of the population. In particular, this value t_i is inversely proportional to the diversity of the population, and is calculated as detailed

in Formula (7), considering that the term $(f_{max} - f_{avg})$ represents to the diversity of the population:

$$t_i = 1 / (f_{max} - f_{avg}). \qquad (7)$$

By Formula (7), when the population is very diverse, the value t_i is very low, and so the acceptance probability of the simulated annealing algorithm is also low. Thus, the algorithm fine-tunes the solutions to which it is applied, promoting the exploitation of the search space. When the diversity of the population reduces, the value t_i increases, and the acceptance probability of the algorithm also increases. Thus, the algorithm tries to move away from the solutions to which it is applied, promoting the exploration of the search space.

Considering the above-mentioned, the simulated annealing algorithm is adaptive based on the diversity of the population, in order to promote either the exploitation or exploration of the search space.

With respect to the move operator used by the simulated annealing algorithm to generate a new solution from a given solution, this move operator is described below.

Considering a given encoded solution s, the move operator develops the next stages. First, a move operator for activity lists is applied to the activity list of s. Thus, a new activity list is obtained. Then, a move operator for assigned resources lists is applied to the assigned resources list of s. Thus, a new assigned resources list is obtained. By the described move operator, a new solution is obtained from the given solution s.

In relation to the move operator applied to the activity list of s, the simple shift operator [21] is applied. This operator works as follows. The operator selects randomly only one activity of the list, and then moves the selected activity from its current position to a new feasible position for this activity. In this respect, the new position is randomly selected from the set of feasible positions for the activity on the list.

In relation to the move operator applied to the assigned resources list of s, an operator which is a variant of the random resetting process [18] is applied. This operator works as follows. The operator selects randomly only one position of the list, and then defines a new resource assignment for the selected position.

V. COMPUTATIONAL EXPERIMENTS

In this section, the computational experiments carried out in order to evaluate the performance of the proposed memetic algorithm are described. Then, the results obtained by these experiments are presented and analyzed in detail. Finally, the performance of the memetic algorithm is compared with those of the heuristic search and optimization algorithms reported until now in the literature for solving the addressed problem.

A. Instance Sets

In order to carry out the computational experiments, the six instance sets introduced in [7] were used. Each of these instance sets contains 40 different instances. Besides, these instance sets have no instances in common. Table I presents the main characteristics of these six instance sets.

Each instance of these six instance sets includes a number of activities to be scheduled and a number of human resources available for developing these activities. For each activity, the instance specifies the duration, the predecessor activities, the required skills, and the number of human resources required for each of these skills. For each available human resource, the instance specifies the skills mastered by the human resource. Besides, the instance specifies the effectiveness level of each available human resource r with respect of each one of the possible work contexts for r in the instance. In particular, the instance includes all the terms $e_{rCr,j,k,g}$ corresponding to each available human resource r and a random value over the range [0,1] for each of these terms.

Each instance of these six instance sets has a known optimal solution with respect of the effectiveness level of the sets of human resources assigned to the project activities. These know optimal solutions are considered here as references to evaluate the performance of the memetic algorithm.

TABLE I.
MAIN CHARACTERISTICS OF THE INSTANCE SETS

Instance Set	Activities per instance	Possible sets of human resources per activity
j30_5	30	1 to 5
j30_10	30	1 to 10
j60_5	60	1 to 5
j60_10	60	1 to 10
j120_5	120	1 to 5
j120_10	120	1 to 10

B. Parameter Setting of the Memetic Algorithm

The memetic algorithm has the following nine parameters: population size, number of generations, probabilities AP_{cLA} and AP_{cUA} for the crossover process, probabilities AP_{mLA} and AP_{mUA} for the mutation process, replacement factor λ for the survival selection process, number of iterations and cooling factor α for the simulated annealing algorithm.

The setting of these nine parameters affects the behavior of the memetic algorithm, and thus, could affect the performance of the memetic algorithm. Because of this reason, preliminary experiments were developed with the aim of determining the parameter setting by which the memetic algorithm reaches the best performance on the six instance sets. These preliminary experiments are described below.

Different values were considered for each one of the nine parameters. The second column of Table II presents the range of values considered for each one of these parameters. In this respect, two values were considered for the population size (i.e., 90 and 180), and three values were considered for the number of generations (i.e., 300, 600 and 900). Different

Project Scheduling: A Memetic Algorithm with Diversity-Adaptive Components that Optimizes the Effectiveness of Human Resources

ISSN 2395-8618

values were considered for the probabilities AP_{cLA} and AP_{cUA} (i.e., values over the range [0.5, 1]), and also different values were considered for the probabilities AP_{mLA} and AP_{mUA} (i.e., values over the range [0.01, 0.2]). Two values were considered for the replacement factor λ of the survival selection process (i.e., 45 and 75). Finally, three values (i.e., 25, 50 and 75) were considered for the number of iterations of the simulated annealing algorithm, and three values were considered for the cooling factor α (i.e., 0.7, 0.8 and 0.9). The combinations of these parameters values determined different possible settings for the parameters of the memetic algorithm.

The different possible settings were evaluated on the six instance sets. Specifically, each setting was run 10 times on each instance of the six instance sets. After that, the results obtained by the settings for each of the six instance sets were analyzed. In this respect, for each instance set, the average percentage deviation from the optimal solution was analyzed and also the percentage of instances for which the optimal solution is reached at least once among the 10 runs developed.

The third column of Table II presents the parameter setting by which the memetic algorithm reached the best performance on the six instance sets. This parameter setting is considered in the following sections.

TABLE II.
PARAMETER SETTING OF THE MEMETIC ALGORITHM

Parameter	Values considered	Values selected
Population Size	90, 180	90
Generations	300, 600, 900	300
Crossover		
AP_{cLA}	[0.5, 1]	0.9
AP_{cUA}	[0.5, 1]	0.6
Mutation		
AP_{mLA}	[0.01, 0.2]	0.1
AP_{mUA}	[0.01, 0.2]	0.05
Survival Selection		
λ (replacement)	45, 75	45
Simulated Annealing		
Iterations	25, 50, 75	25
α (cooling)	0.7, 0.8, 0.9	0.9

C. Main Results

The performance of the memetic algorithm was evaluated on the six instance sets. Specifically, the memetic algorithm was run a number of 30 times on each instance of the six instance sets. To develop these runs, the parameter setting detailed in the third column of Table II was used.

The results obtained by the memetic algorithm for each of the six instance sets were analyzed. In particular, for each instance set, the average percentage deviation from the optimal solution was analyzed and also the percentage of instances for which the optimal solution is reached at least once among the 30 runs carried out.

Table III presents the average percentage deviation from the

optimal solution (Av. Dev. (%)) obtained by the memetic algorithm for each of the six instance sets.

TABLE III
RESULTS OBTAINED BY THE MEMETIC ALGORITHM

Instance Set	Av. Dev. (%)
j30_5	0
j30_10	0
j60_5	0
j60_10	0
j120_5	0.1
j120_10	0.36

With respect of the Av. Dev. (%) values obtained by the memetic algorithm for the first four instance sets, the memetic algorithm obtained Av. Dev. (%) values equal to 0%. These results mean that the memetic algorithm reached an optimal solution in each of the 30 runs developed on each instance of these four instance sets. Thus, the memetic algorithm reached an optimal performance for the first four instance sets.

With respect of the Av. Dev. (%) values obtained by the memetic algorithm for the last two instance sets, the memetic algorithm obtained Av. Dev (%) values equal to 0.1% and 0.36%, respectively. The meaning of these Av. Dev (%) values was analyzed taking into consideration that the known optimal solutions of the instances of these two sets have a fitness level equal to 120. Thus, Av. Dev (%) values equal to 0.1% and 0.36% mean that the average fitness level of the solutions reached by the memetic algorithm is 119.88 and 119.57, respectively.

Therefore, the memetic algorithm has reached very near-optimal solutions for the instances of the two sets. In addition, the memetic algorithm reached an optimal solution at least once among the 30 runs developed on each instance of these two sets. Based on the above-mentioned, the memetic algorithm reached a near-optimal performance for the last two instance sets.

D. Comparison with the Heuristic Algorithms Reported in the Literature for the Addressed Problem

In this section, the performance of the memetic algorithm is compared with those of the heuristic search and optimization algorithms reported until now in the literature for solving the addressed problem.

To the best of the authors' knowledge, only three heuristic search and optimization algorithms have been reported until now in the literature for solving the addressed problem. In this respect, a traditional evolutionary algorithm was reported in [6], a traditional memetic algorithm was reported in [7] which integrates a hill-climbing algorithm within the framework of an evolutionary algorithm, and a hybrid evolutionary algorithm was reported in [8] which incorporates a simulated annealing algorithm within the framework of an evolutionary algorithm.

In [7, 8], the three above-mentioned algorithms have been evaluated on the six instance sets described in Section V.A. The results obtained by each of the three algorithms for the six instance sets are detailed in Table IV, as reported in [7, 8].

Based on the results in Table IV, the performance of the algorithm reported in [8] is better than those of the algorithms reported in [6, 7]. Thus, the algorithm reported in [8] may be considered as the best algorithm reported until now in the literature for solving the addressed problem.

Below, the performance of the algorithm reported in [8] is compared with that of the memetic algorithm proposed here. The algorithm reported in [8] will be referred as algorithm HEA.

TABLE IV
RESULTS OBTAINED BY THE HEURISTIC ALGORITHMS REPORTED IN THE LITERATURE FOR THE ADDRESSED PROBLEM

Instance Set	Evolutionary algorithm [6]	Memetic algorithm [7]	Hybrid evolutionary algorithm [8]
	Av. Dev. (%)	Av. Dev. (%)	Av. Dev. (%)
j30_5	0	0	0
j30_10	0	0	0
j60_5	0.42	0	0
j60_10	0.59	0.1	0
j120_5	1.1	0.75	0.64
j120_10	1.29	0.91	0.8

With respect of the Av. Dev. (%) values obtained by the algorithm HEA and the memetic algorithm proposed here for the first four instance sets, the two algorithms have obtained Av. Dev. (%) values equal to 0%. These results mean that the two algorithms reached an optimal solution in each of the 30 runs developed on each instance of these four instance sets. Thus, for the first four instances sets, the two algorithms have obtained an optimal performance in respect of the Av. Dev. (%) values.

However, with respect of the Av. Dev. (%) values obtained by the algorithm HEA and the memetic algorithm for the last two instance sets, the algorithms have obtained very different Av. Dev. (%) values. In this respect, the algorithm HEA has obtained Av. Dev. (%) values equal to 0.64% and 0.8%, whereas that the memetic algorithm has obtained Av. Dev. (%) values equal to 0.1% and 0.36%. These results mean that the fitness level of the solutions reached by the memetic algorithm for the last two instance sets is significantly higher than that of the solutions reached by the algorithm HEA. Therefore, for the last two instance sets, the performance obtained by the memetic algorithm in respect of the Av. Dev. (%) values is significantly better than that of the algorithm HEA. These results are mainly because of the reasons discussed below.

In the memetic algorithm proposed here, most components into the evolutionary framework are diversity-adaptive. The behavior of these components is adaptive based on the diversity of the evolutionary algorithm population, with the aim of improving the evolutionary-based search. Unlike the memetic algorithm proposed, in the algorithm HEA, most components into the evolutionary framework are non-adaptive. These components totally disregard the diversity of the evolutionary algorithm population and therefore the possibility of improving the evolutionary-based search. Based on the mentioned, the memetic algorithm has significant advantages to develop the evolutionary-based search.

VI. CONCLUSIONS

In this paper, the project scheduling problem introduced in [6] was addressed. This problem supposes really important assumptions concerning the effectiveness of human resources. In this respect, the problem supposes that the effectiveness of a human resource depends on several factors inherent to its work context, and then different effectiveness levels can de defined for each human resource in relation to different work contexts. Besides this assumption, the problem considers a priority optimization objective for project managers. This objective implies optimizing the effectiveness of the sets of human resources defined for the project activities.

To solve the addressed problem, a memetic algorithm was proposed. This memetic algorithm incorporates diversity-adaptive components into the framework of an evolutionary algorithm. The incorporation of these diversity-adaptive components is meant for improving the performance of the evolutionary-based search.

To evaluate the proposed memetic algorithm, exhaustive computational experiments were developed. Specifically, the performance of the memetic algorithm was evaluated on six instance sets with different complexity levels. After that, the performance of the memetic algorithm on the instance sets was compared with those of the heuristic search and optimization algorithms reported until now in the literature for solving the addressed problem.

According to the analysis of the results obtained by the experiments, it may be stated that the memetic algorithm has reached an optimal performance for the first four instance sets and a near-optimal performance for the last two instance sets. Besides, from the comparative analysis developed, it may be stated that the performance of the memetic algorithm is significantly better than those of the heuristic search and optimization algorithms reported until now in the literature for the addressed problem.

In our future work, new diversity-adaptive components will be evaluated into the framework of the memetic algorithm. In addition, new diversity measures will be evaluated.

REFERENCES

[1] G. R. Heerkens, *Project Management*. McGraw-Hill, 2002.
[2] R. K. Wysocki, *Effective Project Management*. Wiley Publishing, 3rd Edition, 2003.

[3] O. Bellenguez, E. Néron, "Lower Bounds for the Multi-skill Project Scheduling Problem with Hierarchical Levels of Skills". In: E. Burke and M. Trick (eds.) *PATAT 2004. Lecture Notes in Computer Science*, vol. 3616, Springer, pp. 229–243, 2005.

[4] T. Hanne, S. Nickel, "A multiobjective evolutionary algorithm for scheduling and inspection planning in software development projects", *European Journal of Operational Research* 167, 663–678, 2005.

[5] W. J. Gutjahr, S. Katzensteiner, P. Reiter, Ch. Stummer, M. Denk, "Competence-driven project portfolio selection, scheduling and staff assignment", *Central European Journal of Operations Research*, vol. 16, no. 3, pp. 281–306, 2008.

[6] V. Yannibelli, A. Amandi, "A knowledge-based evolutionary assistant to software development project scheduling", *Expert Systems with Applications*, vol. 38, no. 7, pp. 8403–8413, 2011.

[7] V. Yannibelli, A. Amandi, "A Memetic Approach to Project Scheduling that Maximizes the Effectiveness of the Human Resources Assigned to Project Activities", in: E. Corchado et al. (eds.) *HAIS 2012. Lecture Notes in Computer Science*, vol. 7208, Springer, pp. 159–173, 2012.

[8] V. Yannibelli, A. Amandi, "A Diversity-Adaptive Hybrid Evolutionary Algorithm to Solve a Project Scheduling Problem", in: E. Corchado et al. (eds.) *IDEAL 2014, Lecture Notes in Computer Science*, vol. 8669, Springer, pp. 412–423, 2014.

[9] J. Blazewicz, J. Lenstra, A. Rinnooy Kan, "Scheduling Subject to Resource Constraints: Classification and Complexity", *Discrete Applied Mathematics*, vol. 5, pp. 11–24, 1983.

[10] V. Yannibelli, A. Amandi, "Project scheduling: A multi-objective evolutionary algorithm that optimizes the effectiveness of human resources and the project makespan", *Engineering Optimization*, vol. 45, no. 1, pp. 45–65, 2013.

[11] M. Srinivas, L. M. Patnaik, "Adaptive probabilities of crossover and mutation in genetic algorithms", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 4, pp. 656–667, 1994.

[12] O. Bellenguez, E. Néron, "A branch-and-bound method for solving multi-skill project scheduling problem", *RAIRO—Operations Research*, vol. 41, no. 2, pp. 155–170, 2007.

[13] L. E. Drezet, J. C. Billaut, "A project scheduling problem with labour constraints and time-dependent activities requirements", *International Journal of Production Economics*, vol. 112, pp. 217–225, 2008.

[14] H. Li, K. Womer, "Scheduling projects with multi-skilled personnel by a hybrid MILP/CP benders decomposition algorithm", *Journal of Scheduling*, vol. 12, pp. 281–298, 2009.

[15] V. Valls, A. Pérez, S. Quintanilla, "Skilled workforce scheduling in service centers", *European Journal of Operational Research*, vol. 193, no. 3, pp. 791–804, 2009.

[16] U. Aickelin, E. Burke, J. Li, "An Evolutionary Squeaky Wheel Optimization Approach to Personnel Scheduling", *IEEE Transactions on evolutionary computation*, vol. 13, no. 2, 433–443, 2009.

[17] C. Heimerl, R. Kolisch, "Scheduling and staffing multiple projects with a multi-skilled workforce", *OR Spectrum*, vol. 32, no. 4, pp. 343–368, 2010.

[18] A. E. Eiben, J. E. Smith, *Introduction to Evolutionary Computing*, Springer, 2nd edition, 2015.

[19] F. J. Rodriguez, C. García-Martínez, M. Lozano, "Hybrid Metaheuristics Based on Evolutionary Algorithms and Simulated Annealing: Taxonomy, Comparison, and Synergy Test", *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 6, pp. 787–800, 2012.

[20] E. Talbi (ed.), *Hybrid Metaheuristics*, Studies in Computational Intelligence, no. 434, Springer, 2013.

[21] R. Kolisch, S. Hartmann, "Experimental Investigation of Heuristics for Resource-Constrained Project Scheduling: An Update", *European Journal of Operational Research*, vol. 174, pp. 23–37, 2006.

[22] P. De Bruecker, J. Van den Bergh, J. Beliën, E. Demeulemeester, "Workforce planning incorporating skills: State of the art", *European Journal of Operational Research*, vol. 243, no. 1, 1–16, 2015.

[23] J. Van den Bergh, J. Beliën, P. De Bruecker, E. Demeulemeester, L. De Boeck, "Personnel scheduling: A literature review", *European Journal of Operational Research*, vol. 226, no. 3, 367–385, 2013.

[24] K. Braekers, R. F. Hartl, S. N. Parragh, F. Tricoire, "A bi-objective home care scheduling problem: Analyzing the trade-off between costs and client inconvenience", *European Journal of Operational Research*, vol. 248, no. 2, pp. 428–443, 2016.

[25] H. Li, N. K. Womer, "Solving stochastic resource-constrained project scheduling problems by closed-loop approximate dynamic programming", *European Journal of Operational Research*, vol. 246, no. 1, pp. 20–33, 2015.

[26] T. Silva, M. De Souza, R. Saldanha, E. Burke, "Surgical scheduling with simultaneous employment of specialised human resources", *European Journal of Operational Research*, vol. 245, no. 3, pp. 719–730, 2015.

Journal Information and Instructions for Authors

I. JOURNAL INFORMATION

Polibits is a half-yearly open-access research journal published since 1989 by the *Centro de Innovación y Desarrollo Tecnológico en Cómputo* (CIDETEC: Center of Innovation and Technological Development in Computing) of the *Instituto Politécnico Nacional* (IPN: National Polytechnic Institute), Mexico City, Mexico.

The journal has double-blind review procedure. It publishes papers in English and Spanish (with abstract in English). Publication has no cost for the authors.

A. Main Topics of Interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research. The main topics of interest include, but are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary Algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces
- Multimedia and Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Robotics
- Geo-processing
- Database Systems

- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Virtual Instrumentation
- Computer Architectures

B. Indexing

The journal is listed in the list of excellence of the CONACYT (Mexican Ministry of Science) and indexed in the following international indices: LatIndex, SciELO, Periódica, e-revistas, and Cabell's Directories.

There are currently only two Mexican computer science journals recognized by the CONACYT in its list of excellence, *Polibits* being one of them.

II. INSTRUCTIONS FOR AUTHORS

A. Submission

Papers ready for peer review are received through the Web submission system on www.easychair.org/conferences/?conf= polibits1; see also updated information on the web page of the journal, www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish. In case of Spanish, author names, abstract, and keywords must be provided in both Spanish and English; in recent issues of the journal you can find examples of how they are formatted.

The papers should be structures in a way traditional for scientific paper. Only full papers are reviewed; abstracts are not considered as submissions. The review procedure is double-blind. Therefore, papers should be submitted without names and affiliations of the authors and without any other data that reveal the authors' identity.

For review, a PDF file is to be submitted. In case of acceptance, the authors will need to upload the source code of the paper, either Microsoft Word or LaTeX with all supplementary files necessary for compilation. Upon acceptance notification, the authors receive further instructions on uploading the camera-ready source files.

Papers can be submitted at any moment; if accepted, the paper will be scheduled for inclusion in one of forthcoming issues, according to availability and the size of backlog.

See more detailed information at the website of the journal.

B. Format

The papers should be submitted in the format of the IEEE Transactions 8x11 2-column format, see http://www.ieee.org/ publications_standards/publications/authors/author_templates. html. (while the journal uses this format for submissions, it is in no way affiliated with, or endorsed by, IEEE). The actual publication format differs from the one mentioned above; the papers will be adjusted by the editorial team.

There is no specific page limit: we welcome both short and long papers, provided that the quality and novelty of the paper adequately justifies its length. Usually the papers are between 10 and 20 pages; much shorter papers often do not offer sufficient detail to justify publication.

The editors keep the right to copyedit or modify the format and style of the final version of the paper if necessary.

See more detailed information at the website of the journal.