Advances in Language and Knowledge Engineering

Research in Computing Science

Series Editorial Board

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 123,** noviembre 2016. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. : 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 123**, November 2016. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Volume 123

Advances in Language and Knowledge Engineering

David Pinto
Darnes Vilariño (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2016

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional “Adolfo López Mateos”, Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

Indexed in LATINDEX, DBLP and Periodica

Printing: 500

Printed in Mexico

Editorial

This volume of the journal “Research in Computing Science” contains selected papers related to the topic of Language and Knowledge Engineering and their applications. The papers were carefully chosen by the editorial board on the basis of the at least two double blind reviews by the members of the reviewing committee or additional reviewers. The reviewers took into account the originality, scientific contribution to the field, soundness and technical quality of the papers. It is worth noting that various papers for this special issue were rejected (rejected rate was 34%).

The volume contains a selection of the best papers presented in the 4th International Symposium on Language & Knowledge Engineering (LKE’2016), an academic conference organized in the Faculty of Computer Science of the Benemérita Universidad Autónoma de Puebla (BUAP) which has been created and organized for the fourth time by the Language & Knowledge Engineering Lab with the aim of offering an academic platform in which experts in related areas may exchange experiences and publish their recent research advances.

We would like to thank Mexican Society for Artificial Intelligence (Sociedad Mexicana de Inteligencia Artificial) and the Thematic Academic Network named “Language Technologies” (Red Temática en Tecnologías del Lenguaje) for their invaluable support in the construction of this volume.

The entire submission, reviewing, and selection process, as well as preparation of the proceedings, were supported for free by the EasyChair system (www.easychair.org).

David Pinto
Darnes Vilariño
Guest Editors
Benemérita Universidad Autónoma de Puebla,
LKE-FCC-BUAP, Mexico

November 2016

Table of Contents

Authorship Verification: A Review of Recent Advances

Efstathios Stamatatos

University of the Aegean
83200 - Karlovassi, Greece
stamatatos@aegean.gr

Abstract. Authorship verification attempts to decide whether the author of a given set of texts is also the author of a disputed text. In comparison to closed-set and open-set attribution, the most popular tasks in relevant literature, the verification setting has some important advantages. First, it is more general since any attribution problem can be decomposed into a series of verification cases. Then, certain factors that affect the performance of closed-set and open-set attribution, like the candidate set size and the distribution of training texts over the candidate authors have limited impact on authorship verification. It is, therefore, more feasible to estimate the error rate of authorship attribution technology, needed in the framework of forensic applications, when focusing on the verification setting. Recently, there has been increasing interest for authorship verification, mainly due to the PAN shared tasks organized in 2013, 2014, and 2015. Multiple methods were developed and tested in new benchmark corpora covering several languages and genres. This paper presents a review of recent advances in this field focusing on the evaluation results of PAN shared tasks. Moreover, it discusses successes, failures, and open issues.

Keywords. Authorship Analysis, Authorship Verification, Text Categorization

1 Introduction

Authorship attribution is the line of research dealing with the quantification of writing style in texts and revealing the identity of their authors using computational methods [23, 58]. Applications closely related with this area are mainly from the humanities (e.g., revealing the author of novels published anonymously, verifying the authorship of literary works, etc.) [24, 34, 64] and forensics (e.g., discovering authorship links between proclamations of different terrorist groups, resolving copyright disputes, revealing multiple aliases of the same user in social media, verifying the authorship of suicide notes, etc) [1, 36, 65].

Authorship attribution can be seen as a single-label multi-class text categorization task. In each authorship attribution case, a candidate set (i.e., suspects) and samples of their writing are given. Then, the task is to find the most likely candidate based on the similarities of their personal style with the text under investigation. Other factors, like topic and genre of text or sentiment polarity should not affect this procedure. However, this is particularly challenging, since it is not yet possible to extract stylometric measures that are only determined by the personal style of author and are immune to changes in topic or genre [59].

The setting examined in the majority of the published studies refers to *closed-set attribution*, where a well-defined set of suspects is given and one of them is necessarily the true author of the disputed text [11, 47, 50, 53, 56]. This scenario matches the requirements of many forensic cases (i.e., traditionally solved by forensic linguists) where police investigators are able to provide a list of suspects based on the assumption that they are the only ones with access to certain resources, having knowledge of certain facts, etc. An alternative, more robust, setting is *open-set attribution* where it is possible the true author not to be included in the list of likely suspects [31, 54]. This is more appropriate in cases where it is not possible to rule out any likely author (e.g., a post in social media could be written by anyone).

A special case of open-set attribution is *authorship verification* where the candidate set is singleton [33, 35, 45]. In other words, given a set of texts by the same author, the task is to determine whether a text under investigation is by that author or not. This is essentially a one-class classification problem since the negative class is chaotic (i.e., all texts by all other authors).

Until recently, there were limited research studies dealing with authorship verification either exclusively [10, 16, 20, 33, 43] or in parallel with closed-set attribution [38, 62]. The recent influential studies of Koppel [32, 35] highlighting the significance of verification as a fundamental problem in authorship attribution and, mainly, a series of PAN shared tasks organized in 2013, 2014, and 2015 radically increased interest and research teams working in this area [25, 61, 60]. PAN evaluation campaigns provided benchmark corpora covering several natural languages and genres as well as an experimentation and evaluation framework to assess the performance of multiple verification methods. Since 2013, significant progress has been reported and multiple studies improved state-of-the-are methods, provided a better understanding of their strengths and weaknesses [6, 7, 9, 18, 22, 30, 45], and highlighted their applications in humanities and forensics [48, 63, 64]. This paper presents a review of recent advances in this field, focusing on the evaluation results of PAN shared tasks.

In the remaining of this paper, Section 2 discusses the advantages of verification setting over closed-set and multi-class open-set attribution. Then, Section 3 presents an overview of PAN shared tasks in authorship verification. Sections 4 and 5 review recent methods focusing on the stylometric features and the properties of the verification models they use, respectively. Section 6 briefly presents main evaluation results of PAN shared tasks and, finally, Section 7 summarizes main conclusions and discusses open issues.

2 Verification vs. Attribution

Authorship verification is a fundamental problem in authorship attribution since any problem, either a closed-set or open-set case, can be decomposed into a set of verification problems. However, it is quite challenging in comparison to both closed-set and open-set attribution since a verification model should estimate whether the disputed text is *similar enough* with respect to the given texts by a certain author while an attribution model should estimate who the *most similar* candidate author is.

As already mentioned, authorship attribution is associated with significant forensic applications. However, it is questionable whether it can be used as evidence in court. Certainly, this technology can be used by investigators to guide their focus on specific suspects and then collect other admissible evidence (e.g., DNA samples) to be presented in court. In United States federal courts, the *Daubert* standard that determines the admissibility of scientific expert testimony requires the estimation of the error rate of a scientific method [49]. Although it is possible to estimate the error rate of specific forensic methods, like DNA analysis [28], how could the error rate of authorship attribution be determined? Certainly, there are several factors that affect the performance of an attribution model, including the number of candidate authors, the distribution of training texts over the candidate authors, the length of text samples, and whether the texts under investigation match in genre and topic, not to mention factors like style ageing (when the personal style of someone changes over time). This is not unusual in forensic science, since the accuracy of many technologies used to provide forensic evidence is affected by specific factors. For example, fingerprint matching performance deteriorates in the case of latent fingerprint identification [12] while speaker recognition accuracy is affected by the duration of audio samples, the number of samples, cross-channel conditions, voice ageing etc. [8]

The estimation of error rate of authorship attribution technology is certainly more feasible if we adopt the verification setting. An inherent problem in closed-set and open-set attribution is that the performance of the attribution model deteriorates by increasing the size of candidate set [31, 38]. On the other hand, in authorship verification the candidate set is always singleton and therefore the error rate is easier to be estimated. Another crucial issue in closed-set and multi-class open-set attribution is that the performance of attribution models depends on the distribution of training texts over the candidate authors. The so called *class imbalance* problem causes attribution models to prefer majority authors in their predictions [57]. However, in a forensic case, the fact that many text samples are available for a certain candidate author should not make that suspect the most likely author of texts under investigation. Authorship verification is more robust to class imbalance since each candidate author is examined separately.

3 PAN Evaluation Campaigns

PAN is a series of shared tasks in digital text forensics [1]. Since 2009 several authorship analysis tasks have been organized including closed-set and open-set attribution, author profiling, author clustering and author obfuscation. In three consecutive editions of PAN (2013, 2014, and 2015) a shared task in authorship verification was organized and attracted the participation of multiple research teams (18 in 2013, 13 in 2014 and 18 in 2015). PAN organizers built new benchmark corpora covering several languages (English, Dutch, Greek, and Spanish) and genres (essays, novels, reviews, newspaper articles) and provided an online experimentation framework for software submissions and evaluation [15]. [2]

[1] http://pan.webis.de

[2] http://www.tira.io/

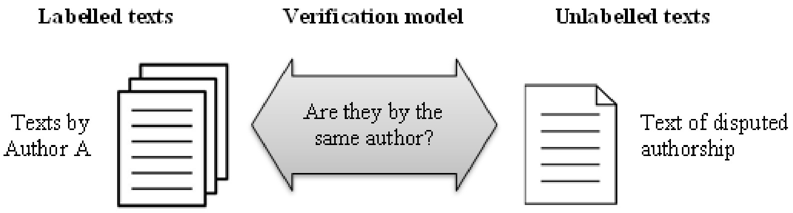

Fig. 1. An authorship verification problem as defined in PAN shared tasks

The definition of the verification task is as follows: *Given a small set of documents by the same author, is an additional (out-of-set) document also by that author?*. This definition is different than the one adopted by Koppel et al. [35] where they attempt to determine whether two documents are by the same author. The latter can be seen as an unsupervised task, where all documents are unlabelled, in terms of authorship. On the other hand, the PAN definition corresponds to a semi-supervised task where some documents are labelled by authorship (the documents by a certain author). The verification process as it is used in PAN tasks is demonstrated in Figure 1.

Each PAN corpus comprises a set of verification problems and within each problem a set of labelled (or *known*) documents (all by the same author) and exactly one unlabelled (or *unknown*) document are given. PAN participants should provide a binary answer (the unknown document is (not) by the same author) and a score in [0,1] indicating the probability of a positive answer (0 means it is certain that the unknown and known documents are not by the same author and 1 means the opposite). In case the verification method founds a specific problem too hard to solve, it is possible to leave it unanswered by providing a score value of exactly 0.5.[3] Evaluation of submissions is performed based on two measures: the Area Under the ROC curve (AUROC) and $c@1$, that is a modification of accuracy that takes into account the problems left unanswered [44]. The final ranking of participants is provided by the product of AUROC and $c@1$. [4]

An overview of the PAN corpora for authorship verification can be seen in Table 1. [5] The training part of each corpus was given to participants in order to develop and fine-tune their approaches while the evaluation corpus was released after the final deadline of submissions. It is important to notice that each corpus, either in training or evaluation set, is balanced with respect to the distribution of positive and negative verification problems. In other words, the prior probability of a positive (or negative) answer is 0.5. This is a general condition that can be applied to any real authorship verification case where there is no additional evidence that favours positive (or negative) answers.

[3] In PAN 2014 and 2015 editions the binary answers are omitted. Any score value greater than 0.5 corresponds to a positive answer and any score lower than 0.5 corresponds to a negative answer.

[4] In PAN 2013 two separate rankings were produced, one based on AUROC and another based on F_1 score

[5] All corpora can be downloaded from http://pan.webis.de/

Table 1. Authorship verification corpora used in PAN shared tasks

	Corpus	Training problems	Evaluation problems	Mean labelled texts / problem	Mean text length (words)
PAN 2013	English (Textbooks)	10	30	3.98	1058
	Greek (Articles)	20	30	5.16	1823
	Spanish (Editorials+Fiction)	5	25	3.07	849
PAN 2014	Dutch (Essays)	96	96	1.89	405
	Dutch (Reviews)	100	100	1.02	114
	English (Essays)	200	200	2.62	841
	English (Novels)	100	200	1.00	5115
	Greek (Articles)	100	100	2.77	1470
	Spanish (Articles)	100	100	5.00	1129
PAN 2015	Dutch (Cross-genre)	100	165	1.75	357
	English (Cross-topic)	100	500	1.00	508
	Greek (Cross-topic)	100	100	2.87	717
	Spanish (Mixed)	100	100	4.00	950

However, when such evidence exists, this parameter should be taken into account in the evaluation process.

In PAN 2013 and PAN 2014 tasks, all documents within a verification problem are in the same language, belong to the same genre, and there are thematic similarities. This means that the disputed text and the known texts have certain similarities. In PAN 2015 the only valid assumption is that all documents within a problem are in the same language. The disputed and the known documents may belong to different genres and their themes can be quite distant. This makes the latter edition of PAN very challenging since it is well known that genre and topic affect stylometric measures considerably. On the other hand, the assumption that all texts should have thematic similarities and belong to the same genre is not realistic since in many forensic cases this is certainly not possible (e.g., imagine the case of verifying the authenticity of a suicide note).

In general, the contribution of PAN shared tasks to the progress of authorship verification research is undoubted. PAN attracted the attention of multiple research teams in this task and provided benchmark corpora that became the standard in this field. Moreover, alternative verification methods were systematically compared and the state-of-the-art performance was estimated. It is also important that based on the fact that PAN required software submissions, a library of verification models is now available and can be used in future evaluations on new corpora as well as in the framework of new tasks, for example *author obfuscation* [46].

On the other hand, there are certain weaknesses of PAN tasks. The volume of some of the provided benchmarks is limited (e.g., the Spanish part of the PAN 2013 corpus). Evaluation results and associated conclusions are corpus-specific due to the lack of homogeneity in corpora properties (i.e., number of problems, known documents per problem, words per document). In addition, the quality of some submissions is questionable since they might be based on naive methods and hasty software implementations while some of the notebook papers do not provide a detailed description of their approach.

4 Stylometric Analysis

In authorship attribution research there is a wide variety of measures that attempt to capture nuances of the personal style of the authors [58]. Most of the measures proposed in the relevant literature correspond to lexical features, e.g., function word frequencies, word *n*-grams, word-length distribution, vocabulary richness measures, etc. Another effective type of features operates on the character level, e.g., character *n*-grams, punctuation mark frequencies, etc. Such features are language-independent and capture intra-word and inter-word information. More sophisticated measures require the analysis of texts by natural language processing tools and then syntactic (e.g., part-of-speech frequencies, rewrite rule frequencies, syntactic *n*-grams, etc.) or semantic features (e.g., semantic dependencies, use of synonyms, etc.) can be extracted. These higher-level features are usually noisy due to errors performed by NLP tools but usually they are useful complement of other, lower-level and more powerful features. Finally, in case all texts share some properties, for instance, they belong to the same genre (e.g., e-mails), they are about a certain thematic area (e.g., computer sales), they are in a specific format (e.g., html), it is possible and very effective to define application-specific measures for that particular domain.

An early study in authorship verification showed that sophisticated syntactic features can enhance the performance of simple lexical features, however, the gain in performance is not significant [16]. Most of the verification approaches submitted to PAN are based on low-level (lexical and character) features and avoid the use of syntactic, semantic, or application-specific features. One reason for that is that measures like character *n*-grams or very frequent word frequencies can practically be applied to any natural language with minimal requirements for text pre-processing. The use of NLP tools by PAN participants was limited to POS tagging and full syntactic parsing. This language-dependent analysis sometimes could not be applied to all languages covered by PAN corpora [60, 61]. Certainly, existing NLP tools, most probably not specifically trained for the texts under investigation, are expected to provide quite noisy stylometric measures.

Another common practice is to combine several features in an attempt to compromise for the weaknesses of a specific feature type. It is also remarkable that in some cases the proposed methods had to select the most suitable feature set for a given collection of verification problems [55]. That way, the features that seem to be more effective (using the training corpus) for a particular language, genre, or topic are selected.

5 Verification Models

The verification model decides whether the disputed text and the known texts are by the same author based on the degree of similarity in terms of the stylometric representation of texts. There are two main categories of verification models:

- *Intrinsic models*: They provide a decision based only on the analysis of the texts found in a given verification problem (the provided known and unknown texts of a certain problem). They exclude the use of external texts by other authors. Therefore, intrinsic models handle the authorship verification problem as a one-class classification case avoiding the use of any external, either labelled or unlabelled, data. Typical examples of intrinsic models are described by Jankowska et al. [22], Potha and Stamatatos [45], Layton [37], Halvani et al. [18], and Bartoli et al. [4]. Intrinsic models are usually faster since they are limited in the analysis of the known and unknown texts. Moreover, they are more robust since their performance does not depend on external factors.
- *Extrinsic models*: They use external documents by other authors to estimate if the similarity of the disputed texts with the known texts is significant enough. Such models actually transform the authorship verification problem to a binary classification case where the known texts form the positive class and the external documents form the negative class. Typical examples of extrinsic models are described by Koppel et al. [35], Seidman [55], Bagnall [2], Veenman & Li [66], and Kocher & Savoy [30]. Extrinsic models are usually more effective than intrinsic ones since a binary classification problem is easier to handle in comparison to a one-class problem. One crucial issue with this kind of methods is the use of an appropriate set of external documents. It is very important to use external documents that belong to the same genre with the ones under investigation [35].

From another point of view, the verification models can be distinguished by the type of learning they use.

- *Eager learning models*: They attempt to extract a general model of authorship verification based on the training corpus. Each verification problem is seen as an instance, either positive (when the known and unknown texts are by the same authors) or negative (in the opposite case). The set of instances of the training corpus is used to train a binary classifier which can then be used to guess the most likely class of any given verification case. Typical examples of this category are described by Frery et al. [13], Bartoli, et al. [4], Pacheco et al. [42], Hürlimann et al. [19], and Brocardo et al. [7]. Such models are effective only when the training corpus is representative of the verification cases that we are going to solve. Their effectiveness and complexity depend on the size and characteristics of the training corpus. Moreover, they can take advantage of powerful supervised learning algorithms, like SVM, neural networks, etc. and they are usually very fast in application phase.
- *Lazy learning models*: They handle every verification case separately. During the training phase they are practically resting. Once a verification case is available in the application phase, they perform all necessary kinds of analysis to estimate their answer. Typical examples of lazy learning models are described by Koppel et al. [35], Khonji and Iraqi [27], Bagnall [2], Jankowska et al. [22], Potha and Stamatatos [45], and Halvani et al. [18]. Such models require higher time cost in the application phase in comparison to eager learning models. However, a big strength is that they do not depend too much on the properties of the training corpus.

Finally, another distinguishing characteristic of verification models refers to the way they handle the labelled examples (known documents by the same author) within each verification problem.

- *Profile-based models*: They concatenate all known documents and then compare the concatenated text with the disputed text. Essentially, they attempt to capture the stylistic properties of the author by discarding any differences between the provided texts. Typical examples of profile-based models are described by Potha and Stamatatos [45], Kocher and Savoy [29], Pacheco et al. [42], Halvani et al. [30], and Kocher & Savoy [30]. A significant strength of such methods is that when text length is limited, by concatenating all available labelled texts they increase the robustness of stylometric representation. On the other hand, concatenated text may have a quite distant representation with respect to its constituent texts especially when the topic and genre of these texts do not match.
- *Instance-based models*: They handle each labelled text separately and compare it with the disputed text. Such models consider each text as a separate instance of author's style. When multiple labelled texts are available, they combine the answers to provide the final decision. Typical examples of this category are described by Seidman [55], Jankowska et al. [22], Moreau et al. [41], Brocardo et al. [7], and Castro-Castro et al. [9]. Another variation is to first concatenate all labelled texts and then split the resulting text into samples of equal size [5, 17]. Instance-based models are better able to exploit significant differences among labelled texts given that they can effectively handle the set of answers (one for each labelled text). On the other hand, they are affected by text length limitations.

It is also notable that some approaches attempt to combine profile-based and instance-based paradigms by first analysing each labelled text separately and then combining the extracted representations of all labelled texts [26, 51]. Such *hybrid* methods practically fail to combine the strengths of the two paradigms.

Table 2 shows the distribution of PAN participants over the types of verification models as defined above. [6] It is clear that the majority of PAN participants follow an intrinsic, lazy, and instance-based methodology. Eager learning method began to be popular in late editions of PAN when the size of the training corpus allowed the development of relatively effective models [13]. Moreover, extrinsic models gain popularity over the years based on the excellent results achieved by Seidman [55], Khonji and Iraqi [27], and Bagnall [2].

6 PAN Results

Analytical evaluation results of PAN participants in benchmark corpora including tests of statistical significance are provided in [25, 61, 60]. Table 3 shows the best results achieved by PAN participants for each corpus and the average performance of all PAN

[6] PAN shared tasks in authorship verification received 18 submissions in 2013, 13 submissions in 2014 and 18 submissions in 2015. All but two (in 2013) research teams also submitted a notebook describing their method.

Table 2. Distribution of PAN participants over the verification model categories (defined in Section 5)

Verification model	**PAN 2013**	**PAN 2014**	**PAN 2015**
Intrinsic	13	10	11
Extrinsic	3	3	7
Eager	2	3	10
Lazy	14	10	8
Profile-based	4	1	4
Instance-based	11	12	12
Hybrid	1	0	2

participants. It is clear that factors like language and genre do not affect the performance of verification models significantly. For instance, the results of Dutch essays are very high while the performance on another corpus in the same language (Dutch reviews) or in the same genre (English essays) are relatively low. Other factors, like the number of labelled texts per verification problem or text length (see Table 1) are certainly significant. In general, when there is a low number of labelled texts (1 or 2) of limited text length (less than 500 words), the performance of verification models worsens.

On the other hand, it is not always possible to explain high or low performance of verification models on a specific corpus based on the quantitative properties of corpus exclusively. There are other qualitative properties that are more useful. For instance, the English novels corpus consists of parts of novels on a specific subgenre of horror fiction that is characterized by an unusual vocabulary and extremely florid prose. This makes similarities between different authors to seem more significant than in normal prose. Verification results for that corpus are poor despite the relatively high text length of its documents.

It should be underlined that the performance of verification models is not heavily affected when the texts within a verification model do not match in genre and thematic area, as it happens in PAN 2015 corpora. Although the average performance of PAN participants is relatively low for those corpora, there were certain submissions capable of reaching impressively high results in these challenging cases [2, 4, 41].

A summary of characteristics of the best-performing systems in the 3 editions of PAN can be seen in Table 4. All submissions that achieved the best performance result (either c@1 or AUROC) in any of the PAN evaluation corpora are presented. For each submission the properties of its verification model as well as its requirements for elaborate analysis to extract stylometric features are described. As can be seen, most of the best-performing models use only low-level stylometric measures, like character and word n-grams. Only a few methods require more sophisticated analysis like POS tagging, or topic modeling (e.g., LSA, LDA). With respect to the verification model properties, extrinsic models, although a minority in PAN participants (see Table 2), are well represented in best-performing submissions and, actually, all three PAN overall winner submissions for 2013, 2014, and 2015 shared tasks belong to this category [2,

Table 3. Evaluation results (best and average performance in terms of $c@1$ and AUROC) of PAN participants on authorship verification corpora

		c@1		**AUROC**	
	Corpus	**Best**	**Average**	**Best**	**Average**
PAN 2013	English (Textbooks)	0.80	0.66	0.84	0.61
	Greek (Articles)	0.83	0.52	0.82	0.60
	Spanish (Editorials+Fiction)	0.84	0.59	0.93	0.67
PAN 2014	Dutch (Essays)	0.91	0.75	0.93	0.76
	Dutch (Reviews)	0.69	0.55	0.76	0.59
	English (Essays)	0.71	0.58	0.72	0.60
	English (Novels)	0.72	0.57	0.75	0.61
	Greek (Articles)	0.81	0.60	0.89	0.67
	Spanish (Articles)	0.78	0.68	0.90	0.71
PAN 2015	Dutch (Cross-genre)	0.77	0.55	0.83	0.60
	English (Cross-topic)	0.76	0.56	0.81	0.62
	Greek (Cross-topic)	0.85	0.54	0.89	0.67
	Spanish (Mixed)	0.83	0.59	0.93	0.66
	Average	0.79	0.60	0.85	0.64

27, 55]. It is also notable that none of the best-performing methods adopts the profile-based paradigm.

An important conclusion extracted from PAN shared tasks is that it is possible to combine different verification models and provide a robust approach with enhanced performance. PAN organizers report the results of a heterogeneous ensemble that combines that answers of all participants (by averaging the scores in each verification problem) and in many cases the performance of this ensemble is better than or competitive with the best-performing PAN participant [25, 61, 60]. Figures 2, 3, and 4 depict illustrative examples for three PAN corpora: the English essays corpus and the Spanish articles corpus from PAN 2014 as well as the Greek articles corpus from PAN 2015, respectively. In more detail, ROC curves on the evaluation parts of these corpora are shown for two methods: (i) the best-performing PAN model for that particular corpus and (ii) the ensemble combining answers by all PAN participants. As can be seen, in the case of English essays, the performance of the ensemble is better than the best individual participant in almost the whole ROC space. Concerning the Spanish articles corpus, the picture is more complicated since both the best PAN participant and the ensemble are competitive and each one of them is the best choice in a certain area of ROC space. When false positives have higher cost the best PAN participant is more effective while when the false negatives are more important the ensemble is a better choice. Finally, when examining the Greek articles corpus, the best PAN participant clearly outperforms the ensemble except in the case the cost of false negatives is extremely high.

In general, the performance of the ensemble in PAN 2015 corpora was lower in comparison to PAN 2014 corpora [61, 60]. This can be partially explained by the consid-

Table 4. Brief description of best-performing models in PAN shared tasks

PAN participant	Verification model	Elaborate stylometric analysis
Bagnall et al. 2015 [2]	extrinsic, lazy, instance-based	none
Bartoli et al. 2015 [4]	intrinsic, eager, instance-based	POS tagging
Frery et al. 2014 [13]	intrinsic, eager, instance-based	none
Ghaeini et al. 2013 [14]	intrinsic, lazy, instance-based	POS tagging
Halvani 2013 [17]	intrinsic, lazy, instance-based	none
Jankowska et al. 2013 [21]	intrinsic, lazy, instance-based	none
Khonji & Iraqi 2014 [27]	extrinsic, lazy, instance-based	none
Mayor et al. 2014 [39]	extrinsic, lazy, instance-based	none
Modaresi & Gross 2014 [40]	intrinsic, eager, instance-based	none
Moreau et al. 2015 [41]	extrinsic, eager, instance-based	LDA, POS tagging
Satyam et al. 2014 [52]	intrinsic, lazy, instance-based	LSA
Seidman 2013 [55]	extrinsic, lazy, instance-based	none
Veenman & Li 2013 [66]	extrinsic, lazy, instance-based	none

erably low performance scores of several participants in PAN 2015 corpora. Certainly, more sophisticated models for combining different methods can provide better results. So far, there is limited research regarding the optimal way to combine heterogeneous verification models [41].

7 Discussion

Since 2013 there has been a significant progress in authorship verification research mainly due to PAN evaluation campaigns that focused on this task. There are multiple research teams around the world that conduct research in this area and multiple methods and variations of them are nowadays available. Based on benchmark corpora in several languages and genres, produced in the framework of PAN shared tasks, systematic evaluation of proposed methods has been performed.

Certainly, there are several factors that affect the performance of verification methods as can be seen in the results of Table 3 in combination with the properties of each corpus (Table 1). However, these factors are less than the ones that are considered in both closed-set and multi-class open-set attribution, since the candidate set size and the distribution of texts over the candidate authors have limited effect in authorship verification.

The average performance of best systems in all PAN corpora (see Table 3) indicates that the error rate of state-of-the-art methods in authorship verification is around 20%. [7] Although this is too high in comparison to the most effective technologies used to provide forensic evidence (e.g., the error rate of DNA analysis is less than 1% [28]), it is comparable to other technologies that analyse noisy data, like latent fingerprint matching [12] or speaker identification [8]. The relatively higher AUROC scores indicate that the verification models are able to rank answers more effectively and they need to be

[7] An average c@1 value of 0.79 indicates an accuracy of around 80%.

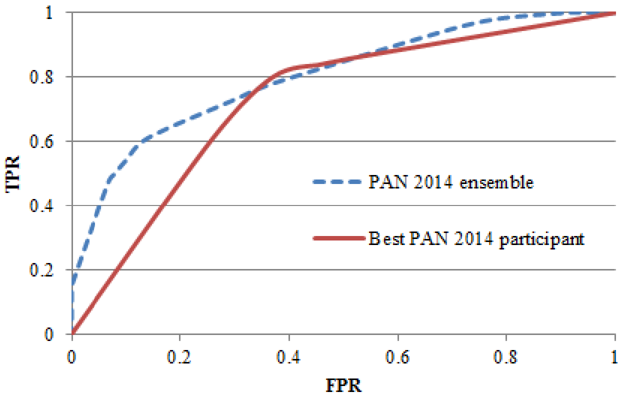

Fig. 2. ROC curves on the evaluation corpus of English essays in PAN 2014 of the best-performing participant for that particular corpus [13] and the ensemble of all PAN 2014 participants

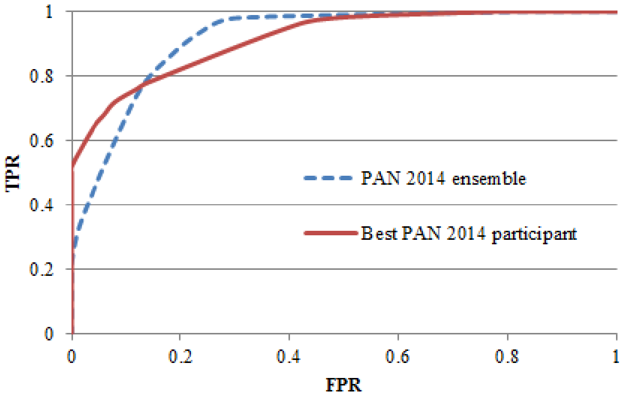

Fig. 3. ROC curves on the evaluation corpus of Spanish articles in PAN 2014 of the best-performing participant for that particular corpus [27] and the ensemble of all PAN 2014 participants

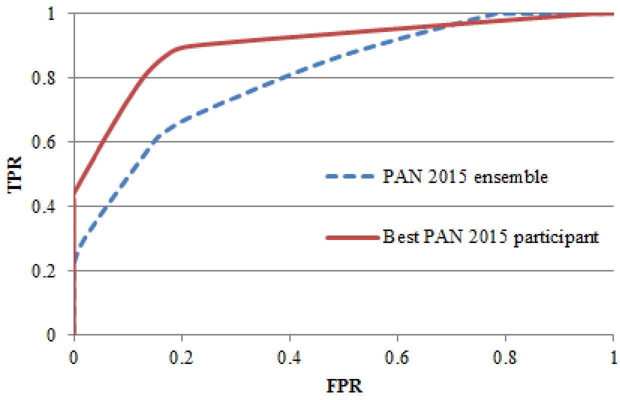

Fig. 4. ROC curves on the evaluation part of Greek cross-topic corpus in PAN 2015 of the best-performing participant for that particular corpus [2] and the ensemble of all PAN 2015 participants

further improved in transforming this ranking to binary answers. Taking into account the prior probability of positive answers is crucial in this direction [2]. All PAN corpora consider an equal prior probability for positive and negative instances. However, this assumption is not true in all application scenarios and more extensive evaluation experiments are needed by controlling this parameter.

The most effective models in PAN evaluation campaigns follow the extrinsic approach where the verification problem is transformed to a binary classification task by considering external documents by other authors [2, 27, 55]. An inherent problem of such methods is the appropriate selection of external documents for a given verification case. So far, existing approaches do not use sophisticated methods to select external documents and they just use texts from the training corpus or texts found my search engines based on queries extracted from specific seed documents [55, 66]. Considerable improvement can be expected if the most suitable set of external documents is found for a given verification problem [35].

Another important conclusion is that verification methods that apply eager learning can be very effective when the training corpus is large enough and comprises similar cases with the evaluation problems [13, 41]. On the other hand, if the training corpus is not representative of the difficulties found in evaluation set, eager learning models fail [60]. In practice, this means that if we want to apply such models in forensic applications, for any given verification problem, we should prepare an appropriate training corpus with cases of similar characteristics. In case there are certain suspects and labelled texts by them, it is possible to build such a corpus. It remains to be seen whether general-purpose corpora covering specific languages and genres can be useful in this respect.

PAN evaluation campaigns demonstrated that combining heterogeneous verification models is a very effective choice [25, 61]. Heterogeneous ensemble achieve consistently high performance in most corpora. Challenging cases where the genre or topic of texts within a verification problem do not match can be handled by more sophisticated ensemble models [41] that select the most appropriate models for each verification problem separately. The existence of a library of verification methods makes this research direction very promising.

Authorship verification tasks at PAN provided the necessary background to explore other relevant tasks. Based on the implementations of verification methods submitted to PAN shared tasks, another task focusing on author obfuscation (i.e., attempting to modify the style of a document so that a verification method does not recognize its author) was recently organized [46]. In another recent PAN shared task in author clustering (grouping documents by authorship) a variation of an authorship verification model was the best-performing participant [3]. All these indicate that verification is a fundamental task in authorship attribution and if we are able to deal with verification effectively it is possible to solve practically any case. There is a lot of room for improvement towards this direction.

References

1. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE 20(5), 67–75 (2005)
2. Bagnall, D.: Author Identification using multi-headed Recurrent Neural Networks. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
3. Bagnall, D.: Authorship Clustering Using Multi-headed Recurrent Neural Networks. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2016)
4. Bartoli, A., Dagri, A., Lorenzo, A.D., Medvet, E., Tarlao, F.: An Author Verification Approach Based on Differential Features. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
5. Bobicev, V.: Authorship Detection with PPM. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
6. Boukhaled, M.A., Ganascia, J.G.: Probabilistic anomaly detection method for authorship verification. In: Besacier, L., Dediu, A.H., Martín-Vide, C. (eds.) Proceedings of Statistical Language and Speech Processing: Second International Conference. pp. 211–219. Springer International Publishing (2014)
7. Brocardo, M.L., Traore, I., Woungang, I.: Authorship verification of e-mail and tweet messages applied for continuous authentication. J. Comput. Syst. Sci. 81(8), 1429–1440 (2015)
8. Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.F., Matrouf, D.: Forensic speaker recognition. IEEE Signal Processing Magazine 26(2), 95–103 (2009)
9. Castro-Castro, D., Arcia, Y.A., Brioso, M.P., Guillena, R.M.: Authorship verification, average similarity analysis. In: Recent Advances in Natural Language Processing. pp. 84–90 (2015)
10. Escalante, H.J., y Gómez, M.M., Pineda, L.V.: Particle swarm model selection for authorship verification. In: Proceedingsof the 14th Iberoamerican Conference on Pattern Recognition. pp. 563–570 (2009)
11. Escalante, H.J., Solorio, T., Montes-y-Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298 (2011)

12. Feng, J., Jain, A.K., Nandakumar, K.: Fingerprint matching. Computer 43, 36–44 (2010)
13. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at clef in author identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CLEF and CEUR-WS.org (2014)
14. Ghaeini, M.: Intrinsic Author Identification Using ModifiedWeighted KNN. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
15. Gollub, T., Potthast, M., Beyer, A., Busse, M., Pardo, F.M.R., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In: Proceedings of the 4th International Conference of the CLEF Initiative. pp. 282–302 (2013)
16. van Halteren, H.: Linguistic profiling for author recognition and verification. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL (2004)
17. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship Verification via k-Nearest Neighbor Estimation. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
18. Halvani, O., Winter, C., Pflug, A.: Authorship verification for different languages, genres and topics. Digital Investigation 16, S33 – S43 (2016)
19. Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., Nissim, M.: GLAD: Groningen Lightweight Authorship Detection. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
20. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: e-mail authorship verification for forensic investigation. In: Proceedings of the 2010 ACM Symposium on Applied Computing. pp. 1591–1598. ACM (2010)
21. Jankowska, M., Keselj, V., Milios, E.: Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
22. Jankowska, M., Milios, E.E., Keselj, V.: Author verification using common n-gram profiles of text documents. In: Proceedings of COLING, 25th International Conference on Computational Linguistics. pp. 387–397 (2014)
23. Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval 1, 234–334 (2008)
24. Juola, P.: How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. Scientific American (2013)
25. Juola, P., Stamatatos, E.: Overview of the author identification task at PAN 2013. In: Working Notes for CLEF 2013 Conference (2013)
26. Kern, R.: Grammar Checker Features for Author Identification and Author Profiling. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
27. Khonji, M., Iraqi, Y.: A slightly-modified gi-based author-verifier with lots of features (asgalf). In: CLEF 2014 Labs and Workshops, Notebook Papers. CLEF and CEUR-WS.org (2014)
28. Kloosterman, A., Sjerps, M., Quak, A.: Error rates in forensic DNA analysis: Definition, numbers, impact and communication. Forensic Science International: Genetics 12, 77 – 85 (2014)
29. Kocher, M., Savoy, J.: UniNE at CLEF 2015: Author Identification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
30. Kocher, M., Savoy, J.: A simple and efficient algorithm for authorship verification. Journal of the Association for Information Science and Technology (2016)

31. Koppel, M., Schler, J., Argamon, S.: Authorship Attribution in the Wild. Language Resources and Evaluation 45, 83–94 (2011)
32. Koppel, M., Schler, J., Argamon, S., Winter, Y.: The fundamental problem of authorship attribution. English Studies 93(3), 284–291 (2012)
33. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
34. Koppel, M., Seidman, S.: Automatically identifying pseudepigraphic texts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1449–1454 (2013)
35. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the American Society for Information Science and Technology 65(1), 178–187 (2014)
36. Lambers, M., Veenman, C.: Forensic authorship attribution using compression distances to prototypes. In: Geradts, Z., Franke, K., Veenman, C. (eds.) Computational Forensics, Lecture Notes in Computer Science, vol. 5718, pp. 13–24. Springer Berlin Heidelberg (2009)
37. Layton, R.: A simple Local n-gram Ensemble for Authorship Verification. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
38. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING 2008). pp. 513–520 (2008)
39. Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., , Meza, I.: A Single Author Style Representation for the Author Verification Task. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
40. Modaresi, P., Gross, P.: A Language Independent Author Verifier Using Fuzzy C-Means Clustering. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
41. Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
42. Pacheco, M., Fernandes, K., Porco, A.: Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
43. Pavelec, D., Oliveira, L.S., Justino, E., Batista, L.V.: Using conjunctions and adverbs for author verification. Journal of Universal Computer Science 14(18), 2967–2981 (2008)
44. Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. ACL (2011)
45. Potha, N., Stamatatos, E.: A profile-based method for authorship verification. In: Artificial Intelligence: Methods and Applications - Proceedings of the 8th Hellenic Conference on AI, SETN. pp. 313–326 (2014)
46. Potthast, M., Hagen, M., Stein, B.: Author Obfuscation: Attacking the State of the Art in Authorship Verification. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CLEF and CEUR-WS.org (2016)
47. Qian, T., Liu, B., Chen, L., Peng, Z.: Tri-training for authorship attribution with limited training data. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL. pp. 345–351 (2014)

48. Roffo, G., Cristani, M., Bazzani, L., Minh, H.Q., Murino, V.: Trusting skype: Learning the way people chat for fast user recognition and verification. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 748–754 (2013)
49. Saks, M.J., Koehler, J.J.: The coming paradigm shift in forensic identification science. Science 309(5736), 892–895 (2005)
50. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. pp. 93–102 (2015)
51. Sari, Y., Stevenson, M.: A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
52. Satyam, Anand, Dawn, A., , Saha, S.: Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
53. Savoy, J.: Authorship attribution based on a probabilistic topic model. Information Processing and Management 49(1), 341–354 (2013)
54. Schaalje, G.B., Blades, N.J., Funai, T.: An open-set size-adjusted bayesian classifier for authorship attribution. Journal of the American Society for Information Science and Technology 64(9), 1815–1825 (2013)
55. Seidman, S.: Authorship Verification Using the Impostors Method. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
56. Seroussi, Y., Zukerman, I., Bohnert, F.: Authorship attribution with topic models. Computational Linguistics 40(2), 269–310 (2014)
57. Stamatatos, E.: Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management 44(2), 790 – 799 (2008)
58. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
59. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21, 421–439 (2013)
60. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the author identification task at PAN 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum (2015)
61. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference. pp. 877–897 (2014)
62. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational Linguistics 26(4), 471–495 (2000)
63. Stolerman, A.: Authorship Verification. Ph.D. thesis, Drexel University (2015)
64. Stover, J.A., Winter, Y., Koppel, M., Kestemont, M.: Computational authorship verification method attributes a new work to a major 2nd century african author. Journal of the American Society for Information Science and Technology 67(1), 239–242 (2016)
65. Sun, J., Yang, Z., Liu, S., Wang, P.: Applying stylometric analysis techniques to counter anonymity in cyberspace. Journal of Networks 7(2), 259–266 (2012)
66. Veenman, C., Li, Z.: Authorship Verification with Compression Features. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)

Training NAO using Kinect

Michalis Chartomatsidis, Emmanouil Androulakis, Ergina Kavallieratou

Dept of Information & Communications Systems,
University of the Aegean Samos, Greece

kavallieratou@aegean.gr

Abstract. This paper describes how the motions of the humanoid NAO robot can be controlled using the Microsoft sensor, Kinect. An application is implemented by which the robot can be controlled using real time tracking. An option to capture and save some motions is also included in order to test if it is possible to train the robot to execute automated motions. The basic question to answer by this work is whether NAO is able to help the user by lifting up an object. To answer that, a series of experiments were performed to validate if the robot could mimic both the captured and the real time motions successfully.

Keywords. Kinect, NAO, Skeleton Tracking, Verification

1 Introduction

The Robotics is a modern technology, which includes the study, design and operation of robots, as well as researches for their further development. The definition given to the robot by the Robot Institute of America is: "A reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks". Nowadays, there are various types of robots:

- Industrial robots: used in industrial manufacturing environment.
- Household robots: This category includes a variety of devices such as vacuum cleaners robot, robot pool cleaners etc.
- Medical Robots: used in medicine, e.g. surgical robots.
- Service Robots: these have a specific and unique use as demonstration of specific technologies or robots used for research
- Military robots: used to neutralize bombs or in other fields such as search and rescue.
- Entertainment Robot: this consists of robotic games to motion simulators.
- Space robots: robots used in space stations.

Among the above categories, the anthropoid robots are extremely popular since they are the human dreams since many decades. However, the ability to imitate human is not natural. It requires a lot of work and research.

In this paper, the Microsoft Kinect Sensor is used to give natural human motion to the Aldebaran NAO anthropoid.

Next, in section 2, previous research is mentioned, while in section 3 Kinect and NAO are shortly described. Our methodology is presented in section 4, while in section 5, several experiments are described. Finally, our conclusion and future work are drawn in section 6.

2 Related Work

There are many works that combine Kinect and NAO and explain how these can help in many different situations.

In [1], a study on learning sign language to children with vision and hearing problems was presented. Both NAO and Kinect were used. Initially the Kinect device was used in order to record the movements of a man and then these movements were transferred to the memory of the robot. Also existing techniques for learning as well as a new web technique were presented. The results showed that learning through the internet is accurate up to 96%, comparatively much larger than the other techniques under study.

In [2], a system that allowed NAO to perform movements made by man in real time was presented. The movements were captured by the Xsens MVN system. The robot was called to perform a series of complex movements, predominantly balance so that they could come to some conclusions about the ability of the robot to adjust the center of its gravity without falling.

Guo, Melissa et al. [4], presented a platform by which children with autism can interact with a robot. This study was based on studies that showed that most children with autism communicate better with robots than with people. By the help of Kinect some movements were performed and then incorporated in the robot’s memory. The children were able to play with the robot as they could make movements and the robot in turn to repeat these movements in real time. The results showed that this could be a better therapy for children with autism.

Lopez Recio et al. [5] studied whether the NAO could help to elderly physiotherapy. Patients instead of following the guidance of a therapist look at the robot that moves and repeat. Moreover, beyond the physical robot a virtual one was used and the performance of the patients was compared. The results initially showed that by the natural Robot patients had better performance compared to the virtual one. Furthermore, in some cases the robot spent much time to complete a movement and the patients were not focusing on the robot. On the other hand, when the robot executed the motions in normal speed, patients were more successful.

3 Hardware

3.1 Kinect Sensor

The Kinect sensor (Fig.1) is a device created by Microsoft for the Xbox 360, Xbox one and PCs. Initially, it was used for video games in Xbox but later it became a strong tool for programmers. It can track human skeleton, recognize voice and provide depth or color data that can be used in many applications. The hardware consists of an RGB camera, an IR emitter, IR depth sensor, a tilt motor and a microphone array. Since Kinect was released, Microsoft SDK has been the official tool for developing applications. Besides that, there are some other tools, like OpeNI & Nite, CL NUI and libfreenect. Here, the Microsoft’s SDK is used, which allows programming in many program languages such as C++ and C#, it does not require a calibration pose to track the user and it is able to track more joints than any other developing tool.

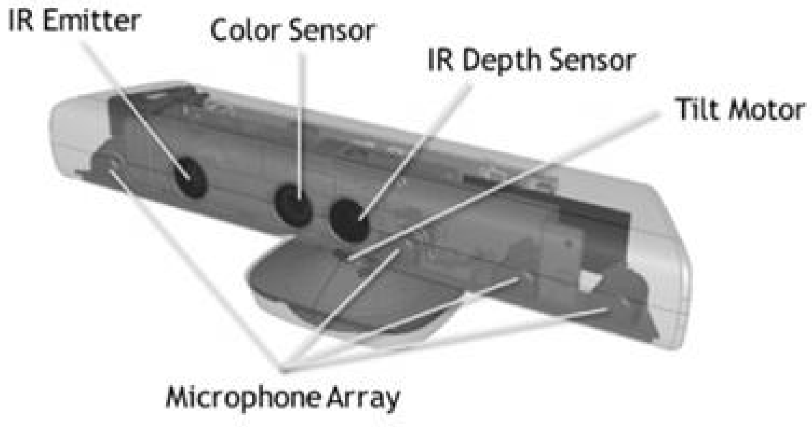

Fig. 1. Kinect sensor's hardware

3.2 NAO Robot

Manufactured by the French robotic company, Aldebaran Robotics, NAO (Fig.2) is a widely used robot in many research institutes. In this thesis the academic version was used that has 25 joints, resulting 25 degrees of freedom (DOF), two cameras that allow it to interact with the environment, four microphones and loudspeakers for listening to voice commands and numerous sensors to enable it to perceive the environment.

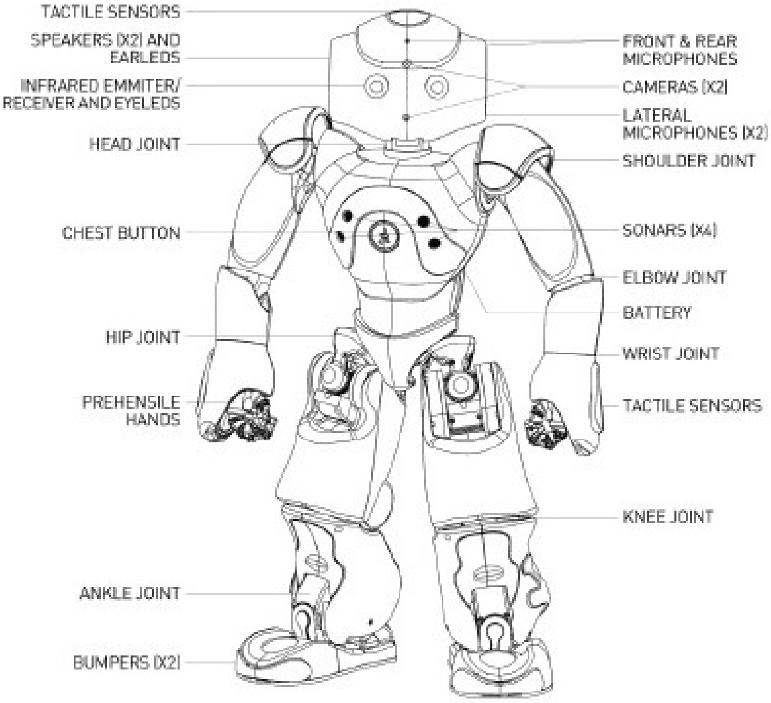

Fig. 2. NAO robot hardware

In order to move, NAO is using a dynamic model of square programming. In particular, NAO is receiving information from joint's sensors becoming more stable and resistant in walking. Random torso oscillations are being absorbed. The mechanism that controls its movement is based on reverse kinematics, which manages Cartesian coordinates, joint control, and balance. For example if during a motion, NAO realizes that is about to lose balance then it stops every motion.

In case it loses balance, it features a fall manager mechanism which is responsible to protect it in case it falls. The main purpose of this mechanism is to detect any change in the center of mass which is determined by the position of the feet. When NAO is about to fall down every other motion is being terminated and the hands are positioned depending on the fall direction. Also the center of mass is reduced and robot's inflexibility decreases.

Furthermore, NAO is equipped with sensors. These sensors allow NAO to have access to information through touching objects. Furthermore two sonar channels provide information which is used to calculate the distance between the robot and an obstacle. The detection range varies from 1cm up to 3 meters but for distance less than 15cm the distance from the obstacle cannot be calculated.

NAO is using an operating system, called NAOqi and it is based on natural interaction with the environment. Here, the NAOqi was used on computer, in order to test its behavior through simulations. It allows homogeneous communication between modules such as movement, sound or video. Also it can be used on many different platforms such as Windows, Linux or Mac. The software development is possible in different programming languages such as C++ and Python. For the purpose of this thesis NAOqi was used in Windows by using Python. Moreover, the Choregraphe was used, a software that allows to observe NAO's behavior.

4 The Proposed Methodology

In this section, it is described how the Kinect sensor can send the tracked data to NAO and how NAO is able to receive this data and repeat the motions of the user. Moreover, the proposed methodology to train NAO will be described.
The basic idea can be described in four steps:

1. the Kinect tracks the user and saves his skeleton joints,
2. the angle that is formed between three of the joints is calculated,
3. an offset is being calculated for this angle and
4. the value of the angle is being sent to robot.

Before presenting how Kinect and NAO are communicating, it should be indicated that in order to have a better understanding about robot's behavior, is is separated in seven body parts: head, left arm, left hand, right arm, right hand, left leg and right leg. For each of the body parts, it was further studied the motion of the joints that were considered important for this thesis as shown in the table below (Table 1).

Table 1. Joints and body parts that were studied

Body Part	Joint (code name)
Left Arm	LShoulderPitch
	LShoulderRoll
	LElbowRoll
	LHand
Right Arm	RShoulderPitch
	RShoulderRoll
	RElbowRoll
	RHand
Head	HeadRoll
Left Leg	LHipPitch
Right Leg	RHipPitch

4.1 Communication

In order for NAO to communicate with the Kinect sensor, the best way proved the use of the socket technology. Each of the body part mentioned above was acting as a server and the Kinect sensor was acting as a client. Thus, the Kinect sensor was connected to these seven body parts.

Furthermore, in order to have all the servers online and ready to receive the values from Kinect, threads were used. Thus, every server was able to run regardless the other.

4.2 Angle Calculation

In order to calculate the angle formed between three joints, simple mathematics are used. Basically, four parameters are used, the detected human skeleton and three skeleton joints. Two vectors per two joints are created and converted to units, and then the dot and cross product are calculated. Here's an example in order to understand deeply the calculation (Fig.3). Let’s suppose that the elbow angle is to be calculated. From the detected joints the x, y, z coordinates of the shoulder (j1), elbow (j2) and wrist joint (j3), are used. The j2j1 and j2j3 vectors are created normalized to units and the cross and dot products are calculated. Using the last two, the atan2 of the angle is calculated and converted in radians.

After the angle calculation, an offset is applied to the angle and the new value is sent to the robot. The methodology of how this offset is calculated is described in the next paragraph.

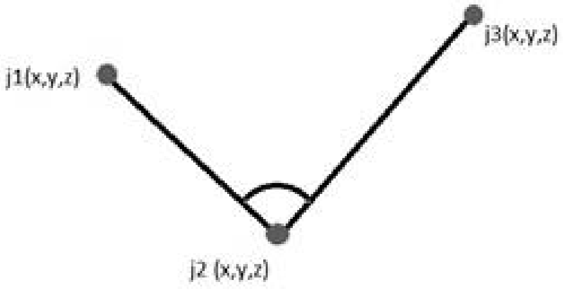

Fig. 3. Calculation of angle

4.3 Offset Calculation

The determination of the appropriate offset for every angle was a quite challenging procedure. Many trials were performed before end up with the correct values. The main difference that makes this procedure difficult is that the value range that NAO can accept, is different than the value range the Kinect returns. The basic procedure to find the correct values can be described through the stages:

- Every joint of the robot is studied for the selected body parts through Choregraphe and considered a starting pose, a mid and a final. Then the angle value for each pose is kept.
- Then the same steps are used to find the angle values for the user using Kinect's skeleton tracking.
- Finally, by the corresponding values an offset value is calculated for each joint.

4.4 Training NAO

As already mentioned, it is implemented a program that trains the robot by doing some motions. In order to succeed that the user records a motion of his own and stores it into a file of skeleton objects. After that NAO is taking the command to execute the saved motion through a simple procedure: The saved file is read and for every three joints in the file the procedure that was described above is repeated.

This way NAO can be trained in doing some motions that may require to be executed many times. It's clear that using real time motion the user would have to excute the same motion many times and the robot would follow. This would be exhausting for the user and could not help in any way.

4.5 Avoiding Loss of Balance

The ability of NAO to be able to execute the motions without the danger of falling was one of the most important things to be implemented. Balance is a factor that cannot be ignored. Even if the robot holds an object there is the danger that it may fall down. To avoid this situation when the robot connects and is ready to receive the values from skeleton tracking it sets its center of mass to both legs. Therefore it is able to execute any kind of motions or hold any object without falling down.

4.6 Receiving Values

As mentioned, each server is responsible for some joints of each body part. In order to command NAO to move, some stages have to be followed:

- First, the message that is received through socket from the client is a sting type containing values for each joint of the body part
- To use these values, they have to be converted into a list of float type numbers that can be accepted by the NAO, using the Python's *split()* method.
- After that, by the use of the *setAngles*() function, NAO takes the command to execute the move.

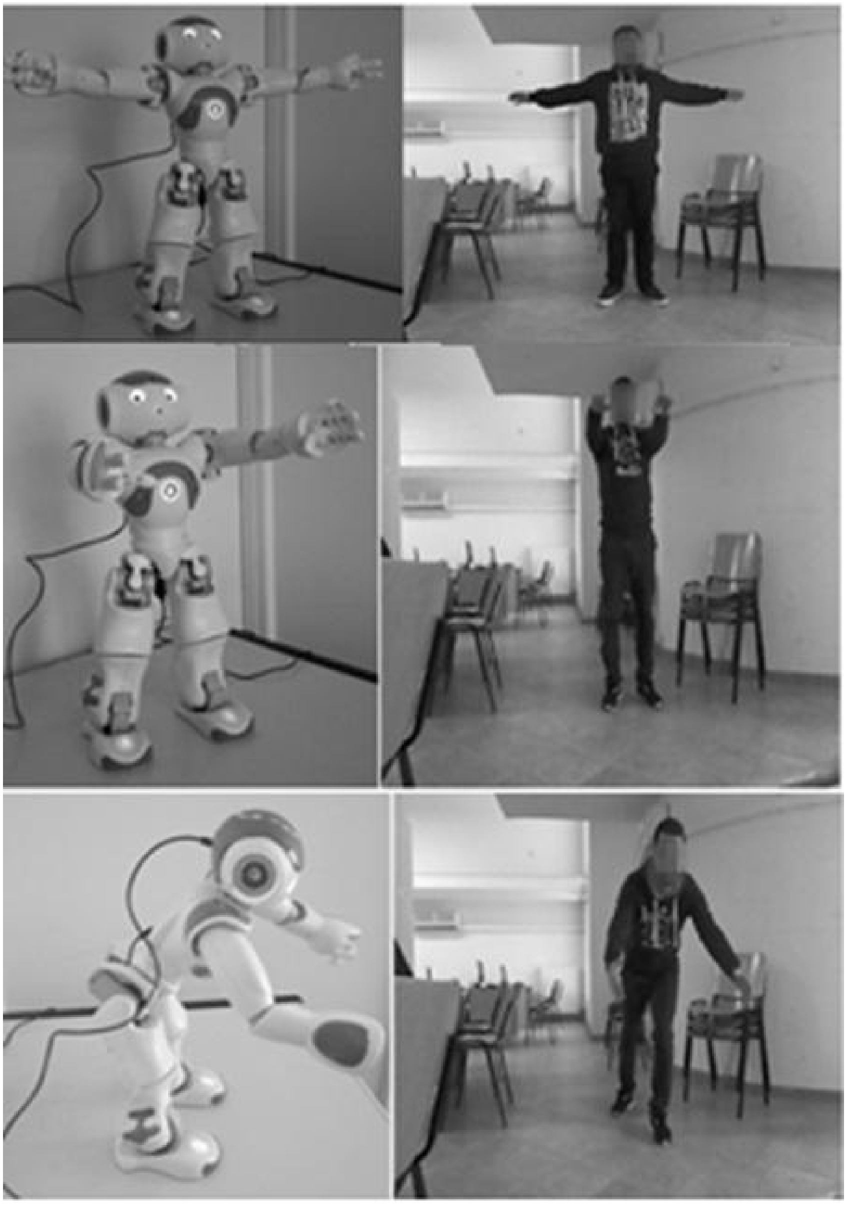

Fig. 4. Testing real time tracking

5 Experimental Results

In order to verify our technique, it was tested through a series of experiments. First, the real-time tracking and sending data to NAO was tested and then the robot was trained to execute a motion that was recorded earlier by the user.

5.1 Real Time Experiment

The first experiment is testing if NAO is able to follow the user in real-time. It's an experiment to test whether our angle and offset calculations were correct and the NAO can follow successfully the human motion in executing time and accuracy.

Our experiments (Fig.4) proved that NAO performs the motion with high success rate. However, it is difficult to control its palms because during some moves Kinect was unable to track whether the user's hand was open or closed. Any other move, which did not include opening or closing the palms, was performed accurately

5.2 Picking Up with One Hand

The second experiment (Fig.5) includes the recording of a motion by the sensor and then the imitation of this motion by the robot. Specifically, NAO picks up an object and leaves it a little further. It should be indicated that NAO does not recognize the object. As a result, in order this experiment to be successful; the object had to be at a specific location.

As mentioned in the previous experiment, the most important problem to face was to establish a way to show the robot it should close the palms throughout the movement. Using the angle between the joints left (right) elbow, left (right) wrist and left (right) hand, respectively, it was not possible to execute always the desired movement. The detection of the sensor proved significant errors in the calculation. The only way for the successful registration of the movement, was to keep a hand stable at the point where the sensor accurately understands the angle and control the opening-closing of the other hand. Specifically, in order for NAO to be able to close the right hand and grab the object, the user had to control it by using his left hand. In particular the angle that is formed between the joints Left Elbow – Left Wrist – Left Hand of the user is controlling whether the right hand of the robot is open or closed.

5.3 Helping the User

The main goal for our third and last experiment (Fig.6) was to find out if NAO is able to collaborate with the user, in order to help him lift an object.

Our first step was to train NAO into executing the proper move to lift the object. In order to perform that, a proper way to command NAO to close his palms had to be planned. As mentioned before closing NAO's palms was not something very easy. To succeed in that, a proper set of joints of the human skeleton, able to be tracked by Kinect, for the entire time that the motion was taking place had to be established.

After a lot of trials, the conclusion was that the best way to make NAO's palms to close was by checking the movement of the user's head. In particular, if the user's head was looking down then NAO's hands were closing and if the user's head was looking up, NAO's hands were opening.

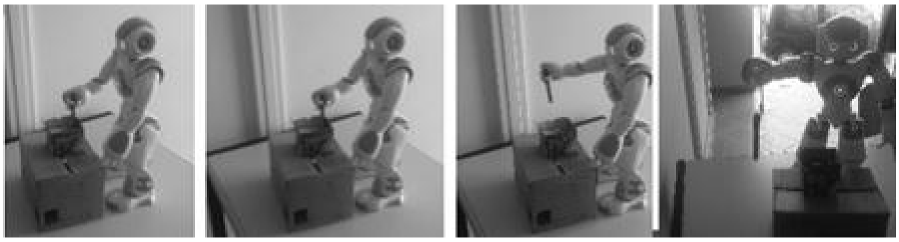

Fig. 5. NAO picks up an object

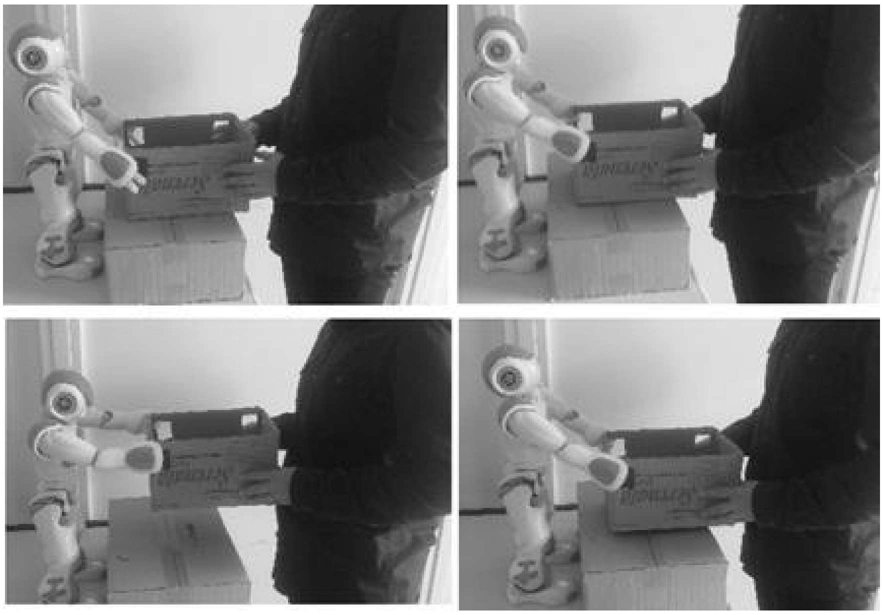

Fig. 6. NAO helps the user pick up an object

After that, the execution of the movement could be performed. As you can see in the fig.6 the motion consists of four stages:

1. Nao and user touch the object,
2. They both grab the object and start lifting it,
3. The object is at maximum distance from the ground and,
4. The object is put back to the ground.

It should be also mentioned here, that for the previous experiments NAO does not recognize its environment, the object needs to be placed at specific spot in order for the move to be successful.

It was experimented and proved that NAO is able to be programmed in order to help the user lift an object of certain size and weight in regular human execution time. This could mean that NAO robot is able to help people in their ordinary life.

6 Conclusions

In this paper, an easy way has been proposed to control the humanoid robot NAO by using the Microsoft Kinect sensor. Two methods, one for real time tracking movements and another for recording the motion and let the robot imitate it later, have been experimented. Furthermore, it has been tested successfully, the capability of NAO to help people in their ordinary life by lifting and moving light or collaborating with a person to extend or lift bigger and long objects.

As a future work, it is planned to introduce the vision of NAO or Kinect in the procedure for object recognition and introduce inversion of the human movement.

References

1. Isong, Itauma, Hasan Kivrak, and Hatice Kose: Gesture imitation using machine learning techniques. In: Signal Processing and Communications Applications Conference (SIU) 2012 20th. IEEE (2012)
2. Koenemann, Jonas, Felix Burget, and Maren Bennewitz: Real-time imitation of human whole-body motions by humanoids. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE (2014)
3. Gouda, Walaa, and Walid Gomaa: Nao humanoid robot motion planning based on its own kinematics. In: Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On. IEEE (2014)
4. Guo, M., Das, S., Bumpus, J., Bekele, E., & Sarkar, N.: Interfacing of Kinect Motion Sensor and NAO Humanoid Robot for Imitation Learning (2013)
5. Recio, López D., Márquez Segura, E., Márquez Segura, L., & Waern, A.: The NAO models for the elderly. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction. pp. 187-188. IEEE Press. (2013)
6. Khoshelham, Kourosh: Accuracy analysis of kinect depth data. In: ISPRS workshop laser scanning. Vol. 38. No. 5. (2011)
7. Khoshelham, Kourosh, and Sander Oude Elberink: Accuracy and resolution of kinect depth data for indoor mapping applications. In: Sensors 12.2. 1437-1454. (2012)
8. Oikonomidis, Iason, Nikolaos Kyriazis, and Antonis A. Argyros: Efficient model-based 3D tracking of hand articulations using Kinect. In: BMVC. Vol. 1. No. 2. (2011)
9. Tang, Matthew: Recognizing hand gestures with microsoft's kinect. In: Palo Alto: Department of Electrical Engineering of Stanford University. (2011)

10. Weise, T., Bouaziz, S., Li, H., & Pauly, M. : Realtime performance-based facial animation. In: ACM Transactions on Graphics (TOG). Vol. 30, No. 4, p. 77. ACM. (2011)
11. Kondori, F. A., Yousefi, S., Li, H., Sonning, S., & Sonning, S: 3D head pose estimation using the Kinect. In: Wireless Communications and Signal Processing (WCSP), 2011 International Conference on. pp. 1-4. IEEE. . (2011)
12. Nirjon, S., Greenwood, C., Torres, C., Zhou, S., Stankovic, J. A., Yoon, H. J., & Son, S. H. : Kintense: A robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3d skeleton data. In: Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on. pp. 2-10. IEEE. (2014)

Applying an Incremental Satisfiability Algorithm to Automatic Test Pattern Generation

Guillermo De Ita, Meliza Contreras, and Pedro Bello

Faculty of Computer Sciences, BUAP
deita@cs.buap.mx, mcontreras@cs.buap.mx,pbello@cs.buap.mx

Abstract. We propose a novel method to review the satisifiability of $(K \wedge \phi)$, where K is a two conjunctive form and ϕ is a three conjunctive form, both formulas defined on the same set of variables. We extend our method to solve the incremental satisfiablity problem (ISAT), and we present different cases where ISAT can be solved in polynomial time. Our proposal is adequate to solve the 2-ISAT problem, and our method allows to recognize tractable instances of 2-ISAT. We illustrate a practical application of our algorithm in the area of recognizing faults on combinatorial circuits.

Keywords. Satisfiability Problem, Incremental Satisfiability Problem, 2-SAT, Propositional Entailment Problem, Efficient Satisfiability Instances

1 Introduction

A central issue in determining these frontiers has centered in the satisfiability problem (SAT) in the propositional calculus [3]. The case 2-SAT, that determines the satisfiability of propositional two Conjunctive Normal Forms (2-CF), is an important tractable case of SAT.

SAT is an important theoretical problem since it was proved as the first problem in the NP-complete complexity class. Despite the theoretical hardness of SAT, current state-of-the-art decision procedures for SAT, known as SAT solvers, have become surprisingly efficient. Subsequently these solvers have found many industrial applications. Such applications are rarely limited to solving just one decision problem, instead, a single application will typically solve a sequence of related problems. Modern SAT solvers handle such problem sequences as an instance of the incremental satisfiability problem (ISAT) [10].

We will consider the ISAT problem as a dynamic incremental set of clauses: $F_0, F_1, \ldots, F_n$, starting with an initial satisfiable formula F_0. Each F_i results from a change in the preceding one F_{i-1} imposed by the 'outside world'. Although the change can be a restriction (add clauses) or a relaxation (remove clauses), we will focus in the restriction case, so we consider adding a new set of clauses to F_{i-1} in order to form F_i. The process of adding new clauses is finished when F_i is unsatisfiable or there are no more clauses to be added.

One idea used on ISAT methods, is to preserve the structures formed when previous formulas were processed, allowing the recognition of common subformulas that they were previously considered. More importantly, it allows the solver to reuse information across several related consecutive problems. The resulting performance improvements make ISAT a crucial feature for modern SAT solvers in real-life applications [10].

ISAT is of interest to a large variety of applications that need to be processed in an evolutive environment [3]. This could be the case of applications such as reactive scheduling and planning, dynamic combinatorial optimization, reviewing faults in combinatorial circuits, dynamic constraint satisfaction and machine learning in a dynamic environment [9].

In [3], we designed an algortihm for reviewing $Sat(K \wedge \phi), K$ and ϕ being CF's. In this work, we adapt our initial algorithm considering that K is a 2-CF and ϕ a 3-CF. We present here, a study about the threshold for the 2-ISAT problem that could be relevant to understand the border between P and NP complexity classes. We also show the practical relevance of our algorithm in the area of automatic test pattern generation (ATPG) systems that consists in differentiating defective components from defect-free components.

2 Preliminaries

Let $X = \{x_1, \ldots, x_n\}$ be a set of n *Boolean variables.* A *literal* is either a variable x_i or a negated variable $\overline{x}_i$. As usual, for each $x \in X$, $x^0 = \neg x = \overline{x}$ and $x^1 = x$.

A *clause* is a disjunction of different and non-complementary literals. Notice that we discard the case of tautological clauses. For $k \in I\!N$, a *k-clause* is a clause consisting of exactly k literals, and a $(\leq k)$-clause is a clause with at most k literals.

A *conjunctive normal form* (CNF, or just CF) F is a conjunction of non-tautological clauses. We say that F is a monotone positive CF if all of its variables appear in unnegated form. A k-CF is a CF containing only k-clauses. $(\leq k)$-CF denotes a CF containing clauses with at most k literals.

A variable $x \in X$ *appears* in a formula F if either x or $\neg x$ is an element of F. The *size* of a CF F is defined as the total number of literals appearing in the CF F. We use $\upsilon(X)$ to represent the variables involved in the object X, where X can be a literal, a clause, or a CF. For instance, for the clause $c = \{\overline{x_1}, x_2\}$, $\upsilon(c) = \{x_1, x_2\}$. $Lit(F)$ is the set of literals involved in F, i.e. if $X = \upsilon(F)$, then $Lit(F) = X \cup \overline{X} = \{x_1, \overline{x}_1, ..., x_n, \overline{x}_n\}$. Also, we used $\neg Y$ as the negation operator on the object Y.

An *assignment* s for F is a function $s : \upsilon(F) \rightarrow \{0, 1\}$. An *assignment* s can also be considered as a set of literals without a complementary pair of literals, e.g., if $l \in s$, then $\overline{l} \notin s$, in other words s turns l *true* and $\overline{l}$ *false* or viceversa. Let c be a clause and s an assignment, c is *satisfied* by s if and only if $c \cap s \neq \emptyset$. On the other hand, if for all $l \in c$, $\overline{l} \in s$, then s falsifies c.

Let F be a CF, F is *satisfied* by an assignment s if each clause in F is satisfied by s. F is *contradicted* by s if any clause in F is falsified by s. A model of F is an assignment for $\upsilon(F)$ that satisfies F. A falsifying assignment of F is an assignment for $\upsilon(F)$ that contradicts F. If $n = |\upsilon(F)|$, then there are 2^n possible assignments defined over $\upsilon(F)$. Let $S(F)$ be the set of 2^n assignments defined over $\upsilon(F)$. $s \vdash F$ denotes that assignment s is a model of F, while that $s \nvdash F$ denotes that s is a falsifying assignment of F.

If $F_1 \subset F$ is a formula consisting of some clauses from F, and $\upsilon(F_1) \subset \upsilon(F)$, an assignment over $\upsilon(F_1)$ is a *partial* assignment over $\upsilon(F)$. If $n = |\upsilon(F)|$ and $n_1 = |\upsilon(F_1)|$, any assignment over $\upsilon(F_1)$ has 2^{n-n_1} extensions as assignments over $\upsilon(F)$. If s has logical values determined for all variables in F then s is a *total assignment* of F.

The SAT problem consists of determining whether F has a model. SAT(F) denotes the set of models of F, then SAT(F) $\subseteq S(F)$. The set FAL(F) $= S(F) \setminus SAT(F)$ consists of the assignments from $S(F)$ that falsify F. Clearly, for any propositional formula F, $S(F) = SAT(F) \cup Fals(F)$.

3 Reducing Conjunctive Normal Forms

Let K be a CF, i.e., $K = \bigwedge_{i=1}^{m} C_i$, where each $C_i, i = 1, \ldots, m$ is a disjunction of literals. Let us introduce the main problem to be considered here.
Instance: Let K be a 2-CF and ϕ be a 3-CF, such that $\upsilon(\phi) \subseteq \upsilon(K)$.
Problem: To determine SAT($K \cup \phi$).

We will present here, an efficient algorithm to solve this problem. But first, we introduce some common rules used to simplify a conjunctive normal form F, keeping just the necessary subformulas that determine the satisifiability of F. For example, for a CF it is common to delete all redundant clauses as: tautological clauses, repeated clauses and clauses with pure literals.

Subsumed clause Rule: Given two clauses c_i and c_j of a CF F, if $Lit(c_i) \subseteq Lit(c_j)$ then c_j is subsumed by c_i, and c_j can be deleted from F, because all satisfying assignment of c_j is a satisfying assignment of c_i, that is $Sat(c_j) \subseteq Sat(c_i)$. Thus, it is enough just to keep c_i (the clause which subsumes) in the CF.

Furthermore, subsumed clause rule can be combined with resolution in order to simplify ϕ, as we show in the following lemma.

Lemma 1. *Let $(x, y) \in K$ and a clause $(\neg x, y, z) \in \phi$ then its resolvent (y, z) is a binary clause that can be added to K and its father $(\neg x, y, z)$ can be deleted from ϕ.*

Proof. We have that $(x \vee y) \wedge (\neg x \vee y \vee z) \equiv y \vee (x \wedge (\neg x \vee z)) \equiv y \vee ((x \wedge \neg x) \vee (x \wedge z)) \equiv y \vee (x \wedge z) \equiv (x \vee y) \wedge (y \vee z)$, and if these last two clauses are preserved in K then the clause $(\neg x, y, z)$ can be deleted from ϕ, because it is subsumed by $(y \vee z) \in K$ and therefore, the set of models of $(K \wedge \phi)$ are preserved without changes.

Resolution is also useful in our purpose to move clauses from ϕ to K. For example, if ϕ contains clauses type: (x, y, z) and $(\neg x, y, z)$, then they can be deleted from ϕ and the clause (y, z) is added to K. The justification of this rule comes from the distributive property, since $(x \vee y \vee z) \wedge (\neg x \vee y \vee z) \equiv (y \vee z) \vee (x \wedge \neg x) \equiv (y \vee z)$.

On the other hand, other common rules used to simplify formulas have to be adapted for our purpose.

Rule of Pure Literal: Let F be a CF, $l \in Lit(F)$ is a pure literal if l appears in F but $\overline{l}$ does not appear in F.

If a clause contains a pure literal, that clause can be eliminated from F, keeping the logical value of F. Because if the literal l is set to $True$, the clause containing l is also $True$, therefore it can be deleted from F. However, this rule has to be applied carefully for our problem, since in the process of adding new clauses ϕ to K, the initial pure literals in K could be not longer pure in $K \cup \phi$. Therefore, to delete clauses with pure literals must be applied into a local reach, working in each instance $(K \wedge \phi)$. But, if a new set of clauses ϕ_{i+1} has to be considered, then all clause with pure literals deleted from $(K \wedge \phi)$ must be returned to ϕ_{i+1}.

It is easy to build $Fals(K)$ since each clause C_i determines a subset of falsifying assignments of K. For example, $Fals(K) = \bigcup_{i=1}^{m} Fals(C_i)$. The following lemma expresses how to form the falsifying set of assignments of a CF.

Lemma 2. *Let $K = \bigwedge_{i=1}^{m} C_i$ be a CF, then $Fals(K) = \bigcup_{i=1}^{m} \{\sigma \in S(K) \mid Fals(C_i) \subseteq \sigma\}$*

Lemma 3. *If a CF K is satisfiable, then $\forall K' \subseteq K$, K' is a CF satisfiable.*

Proof. If K is satisfiable, then $Fals(K) = \bigcup_{C_i \in K} Fals(C_i) \subset S(F)$. Clearly, if we discard some clauses from K, forming K', then $Fals(K') = \bigcup_{C_i \in K'} Fals(C_i) \subseteq \bigcup_{C_i \in K} Fals(C_i) \subset S(F)$. Thus, K' is satisfiable.

Corollary 1 *If a CF K is unsatisfiable, then $\forall$ CF K' such that $K \subseteq K'$, K' remains unsatisfiable.*

Proof. An unsatisfiable CF K holds that $Fals(K) = \bigcup_{C_i \in K} Fals(C_i) = S(F)$. Then, if we aggregate more clauses to K forming K', then $Fals(K) = \bigcup_{C_i \in K} Fals(C_i) \subseteq \bigcup_{C_i \in K'} Fals(C_i) = S(F)$. Thus, K' is also unsatisfiable.

4 The Transitive Closure of a 2-CF

The fact that in a 2-CF formula a clause is equivalent to a pair of implications can be straightforward established as follows: if $\{x, y\} \in F$ then $\{x, y\}$ is equivalent to both $\overline{x} \rightarrow y$ and $\overline{y} \rightarrow x$. The arrow → has the usual meaning of implication in classical logic.

Definition 1 *Let F be a 2-CF and L its set of literals. The relation $\rightarrow_R \subset L \times L$ is defined as follows: $x \rightarrow_R y$ if and only if $x \rightarrow y$.*

Definition 2 *Let F be a 2-CF, a partial assignment s of F is a feasible model for F, if s does not falsify any clause in F.*

We consider now the transitive closure of $\rightarrow_R$, denoted by "$\Rightarrow$". This new relation $\Rightarrow$ can always be constructed inductively from $\rightarrow_R$. For any feasible model s of F where x and y occur in F; if $x \Rightarrow y$ and x is true in s then it is straightforward to show that y is true in s. It is said that y is *forced* to be true by x. Let $T(x)$ be the set of literals forced to be true by x, that is $T(x) = \{x\} \cup \{y : x \Rightarrow y\}$.

It is clear that, if x is a literal occurring in a formula F, and if $\bar{x} \in T(x)$, then x cannot be set to true in any model of F. Analogously, if $x \in T(\bar{x})$ then x cannot be set to false in any model of F.

Definition 3 *Let F be a 2-CF, for any literal $x \in F$, it is said that $T(x)$ is inconsistent if $\overline{x} \in T(x)$ or $\perp \in T(x)$, otherwise $T(x)$ is said to be consistent.*

Unit clauses in 2-CF can be expressed as implications, that is, if F has unit clauses $\{u\}$ then $u \equiv u \vee \perp$, hence $\perp \in T(\overline{u})$. As a consequence, in formulas with unit clause $\{u\}$ follows that $T(\overline{u})$ is inconsistent. Let F be a 2-CF with n variables and m clauses, it has been shown that for any literal $x \in F$, $T(x)$ and $T(\overline{x})$ are computed in polynomial time over $|F|$, in fact, for all $l \in Lit(F)$, $T(l)$ is computed with time complexity $O(n \cdot m)$ [5].

For any literal x in a 2-CF, the sets $T(x)$ and $T(\overline{x})$ allow to determine which variables have a fixed logical values in every model of F, that is to say, the variables that are true in every model of F and the variables that are false in every model of F. The properties of the sets $T(x)$ and $T(\overline{x})$ will be established as a lemma.

Lemma 4. *Let F be a 2-CF and x a variable in F.*

1. *If $T(x)$ is inconsistent and $T(\overline{x})$ is consistent then $\overline{x}$ is true in every model of F.*
2. *If $T(\overline{x})$ is inconsistent and $T(x)$ is consistent then x is true in every model of F.*
3. *If both $T(\overline{x})$ and $T(x)$ are inconsistent then F does not have models and F is unsatisifiable.*
4. *If both $T(\overline{x})$ and $T(x)$ are consistent then x does not have a fixed valued in each model of F.*

Proof. 1. Suppose $\overline{x}$ is false in a model of F, so x should be true in that model of F. However, $T(x)$ is inconsistent, so $x \Rightarrow \overline{x}$ and x cannot be true in the model of F contradicting the assumption. Hence, any model of F has to assign false to x and true to $\overline{x}$. The other cases are proved similarly.

From properties (1) and (2) of lemma 4 we formulate the following definition

Definition 4 *A base for the set of models of a 2-CF F, denoted as $S(F)$, is a partial assignment s of F which consists of the variables with a fixed truth value.*

We denote by *Transitive_Closure(F)* to the procedure which computes the sets $T(x)$ and $T(\bar{x})$ for each $x \in \upsilon(F)$. The transitive procedure applied on a

2-CF F allows to build bases for the set of models of F. If a base $S(F)$ is such that $|S(F)| = |\upsilon(F)|$, then each variable of F has a fixed truth value in every model of F, so there is just one model.

Definition 5 *Let F be a 2-CF and x a literal of F. The reduction of F by x, also called $forcing$ x and denoted by $F[x]$, is the formula generated from F by the following two rules*

a) removing from F the clauses containing x (subsumption rule),
b) removing $\overline{x}$ from the remaining clauses (unit resolution rule).

A reduction is also sometimes called a *unit reduction.* The reduction by a set of literals can be inductively established as follows: let $s = \{l_1, l_2, \ldots, l_k\}$ be a partial assignment of $\upsilon(F)$. The reduction of F by s is defined by successively applying definition 5 for l_i, $i = 1, \ldots, k$. That is reduction of F by l_1 gives the formula $F[l_1]$, following a reduction of $F[l_1]$ by l_2, giving as a result the formula $F[l_1, l_2]$ and so on. The process continues until $F[s] = F[l_1, ..., l_k]$ is reached. In case that $s = \emptyset$ then $F[s] = F$.

Example 1. Let $F = \{\{x_1, \overline{x}_2\}, \{x_1, x_2\}, \{x_1, x_3\}, \{\overline{x}_1, x_3\}, \{\overline{x}_2, x_4\}, \{\overline{x}_2, \overline{x}_4\}, \{x_2, x_5\}, \{x_3, \overline{x}_5\}\}$. If $s = \{x_2, \overline{x}_3\}$, $F[x_2] = \{\{x_1\}, \{x_1, x_3\}, \{\overline{x}_1, x_3\}, \{x_4\}, \{\overline{x}_4\}, \{x_3, \overline{x}_5\}\}$, and $F[s] = \{\{x_1\}, \{x_1\}, \{\overline{x}_1\}, \{x_4\}, \{\overline{x}_4\}, \{\overline{x}_5\}\}$.

Let F be a 2-CF formula and s a partial assignment of F. If a pair of contradictory unitary clauses is obtained while $F[s]$ is being computed, then F is falsified by the assignment s. Furthermore, during the computation of $F[s]$, new unitary clauses can be generated. Thus, the partial assignment s is extended by adding the already found unitary clauses, that is, $s = s \cup \{u\}$ where $\{u\}$ is a unitary clause. So, $F[s]$ can be again reduced using the new unitary clauses. The above iterative process is generalized, and we call to this iterative process $Unit\_Propagation(F, s)$. For simplicity, we will abbreviate $Unit\_Propagation(F, s)$ as $UP(F, s)$.

As a result of applying $UP(F, s)$, we obtain a new assignment s' that extend to s, and a new subformula F' formed by the clauses from F that are not satisfied by s'. We denote $(F', s') = UP(F, s)$ to the pair resulting of the application of Unit Propagation on F by the assignment s. Notice that if s falsifies F then s' could have complementary literals and F' contains the null clause. And when s satisfies F, then F' is empty.

5 Incremental Satisfiability Problem

The incremental satisfiability problem (ISAT) involves checking whether satisfiability is maintained when new clauses are added to an initial satisfiable knowledge base K. ISAT is considered as a generalization of SAT since it allows changes of the input formula over time. Also, it can be considered as a prototypical Dynamic Constraint Satisfaction Problem (DCSP) [7].

Different methods have been applied to solve ISAT, among them, variations of the branch and bounds procedure, denoted as IDPL methods, which are usually based in the classical Davis-Putnam-Loveland (DPL) method. In a IDPL procedure, when adding new clauses, the procedure maintains the search tree generated previously for the set of clauses K. IDPL performs substantially faster than DPL for a large set of SAT problems [6]. Rather than solving related formulas separately, modern solvers attempt to solve them *incrementally* since many practical applications require solving a sequence of related SAT formulas [2, 4].

Assuming an initial KB K, and a new CF ϕ to be added, let us consider some cases where SAT$(K \wedge \phi)$ can be determined efficiently.

1. If K and ϕ are 2-CF's then $(K \wedge \phi)$ is a 2-CF that is the input of ISAT. In this case, 2-ISAT is solvable in linear-time by applying the well known algorithms for 2-SAT [5, 1]
2. For monotone formulas, ISAT keeps satisfiable formulas. If each variable maintains a unique sign in both K and ϕ then $(K \wedge \phi)$ is always satisfiable.
3. If ϕ consists of one clause and we have the searching graph of K, we only have to review which consistent path of the graph falsifies ϕ, and this can be done in linear time on the number of literals of K and the number of consistent paths of the searching graph.

It is clear that a set of changes over a satisfiable KB K in 2-CF could change K into a general CF, in which case, K will turn into a general CF K', $K \subset K'$, where the SAT problem on K' is a classic NP-complete problem.

From now on, let us consider that K is a 2-CF and ϕ is a 3-CF, both of them do not match the previous cases presented in this section. Therefore, we consider that ϕ consists of clauses that effectively decrease the set of models of K.

First, we show the relevance of our method to determine SAT$(K \wedge \phi)$ for applying it in a practical area. We consider automatic test pattern generation (ATPG) systems that consist in differentiating defective components from defect-free components. We start considering the method proposed by Larrabee [8]. Her method is based in the formation of cojunctive forms to express test patterns for single stuck at faults in combinatorial circuits.

For example, all binary and unary gates are expressed via CF's. Considering a logic gate with two inputs X, Y and output Z, basic gates are expressed as: *And* Gate: $(\overline{Z}+X)(\overline{Z}+Y)(Z+\overline{X}+\overline{Y})$, *Or* Gate: $(\overline{X}+Z)(\overline{Y}+Z)(X+Y+\overline{Z})$. The *Not* Gate: $(\overline{X}+Y)(\overline{Y}+X)$, and the *Xor* Gate is $(\overline{X}+Y+Z)(X+\overline{Y}+Z)(\overline{X}+\overline{Y}+\overline{Z})(X+Y+\overline{Z})$.

In Larrabee's method, a CF is used to generate test patterns on combinatorial circuits, considering the construction of such Boolean formula with true and false outputs of the circuit. In order to generate a test pattern for a single fault on the circuit, a CF that detects the fault in the circuit is extracted, and then, is needed to apply a procedure for reviewing the satisfiability of the formed formula.

The most important steps of the Larrabee's system are illustrated with the combinational circuit that appears in Figure 1:

1. A transformation process is applied on each gate forming the circuit and transforming it into a CF, according to the patterns described previously. For

example, the first circuit in Figure 1 is equivalent to the CF: $\{\{C, E\}, \{\overline{C}, \overline{E}\}$ $, \{X, \overline{D}\}, \{X, \overline{E}\}, \{\overline{X}, D, E\}, \{\overline{D}, A\}, \{\overline{D}, B\}, \{D, \overline{A}, \overline{B}\}\}$.

2. It is needed to represent a faulted version of the initial circuit by making a copy of the original circuit, renaming variables and inserting two new nodes representing the presumed disrupted connection in the faulted circuit. In our example, we consider D' as a disrupted connection with a new output, denoted by X', on the final gate. In this case, D represents stuck-at 1, D' represents a faulted behavior at the fault site, and $\overline{D}$ is the node representing the correct behavior at the fault site. Then, we obtain the following CF: $\{\{C, E\}, \{\overline{C}, \overline{E}\} \{X', \overline{D}\}, \{X', \overline{E}\},$ $\{\overline{X'}, D, E\}, \{\overline{D}, A\}, \{\overline{D}, B\}, \{D, \overline{A}, \overline{B}\}, \{D'\}\}$.
3. The two circuits(faulted and unfaulted) are joined by a XOR-gate to represent that only one of them will be satisfiable. $F = \{\{X', \overline{D'}\}, \{X', \overline{E}\}, \{D'\},$ $\{\overline{X'}, D', E\}, \{C, E\}, \{\overline{C}, \overline{E}\} \{Z, Z'\}, \{\overline{Z}, \overline{Z'}\}, \{X, \overline{D}\}, \{X, \overline{E}\}\{\overline{D}, B\}, \{\overline{D}, A\},$ $\{\overline{X}, D, E\}, \{D, \overline{A}, \overline{B}\}, \{\overline{X}, X', Z\}, \{X, \overline{X'}, Z\}, \{\overline{X}, \overline{X'}, Z'\}, \{X, X', Z'\}\}$.
4. Our procedure requires that $\upsilon(\phi) \subseteq \upsilon(K)$, so we can redefine variables in ϕ that initially they do not appear in K, using binary clauses. In our example, $Z' = \overline{Z}$ and then the clauses $\{\overline{Z}, \overline{Z'}\}$ and $\{Z, Z'\}$ guarantee that only one of the two variables Z or Z' will be valid into the XOR gate: $\{\{\overline{X}, X', Z\}, \{X, \overline{X'}, Z\}, \{\overline{X}, \overline{X'}, Z'\}, \{X, X', Z'\}\}$.
5. At the end of the Larrabee's method, a final CF is obtained $F = \{\{X', \overline{D'}\},$ $\{X', \overline{E}\}, \{\overline{X'}, D', E\}, \{D'\}, \{C, E\}, \{\overline{C}, \overline{E}\} \{Z, Z'\}, \{\overline{Z}, \overline{Z'}\}, \{X, \overline{D}\}, \{X, \overline{E}\},$ $\{\overline{X}, D, E\}, \{\overline{D}, A\}, \{\overline{D}, B\}, \{D, \overline{A}, \overline{B}\}, \{\overline{X}, X', Z\}, \{X, \overline{X'}, Z\}, \{\overline{X}, \overline{X'}, Z'\},$ $\{X, X', Z'\}\}$

Now, we describe how our proposal works to review the satisifiability of $(K \wedge \phi)$, being K a 2-CF and ϕ a 3-CF. We illustrate our proposal for considering the final CF obtained via the Larrabee's method.

Let $S = S(K)$ be the base for the initial 2-CF K. First, we apply the simplification rules described in section 3, in order to reduce ϕ and extend K, keeping the logical value of $(K \wedge \phi)$. After that, we consider $(\phi, s) = UP(\phi, S)$ because the partial assignment common to all model of K must remain in any model of ϕ. Furthermore, the new binary clauses in ϕ can be considered to be part of K and they can be deleted from ϕ.

Applying the rules to reduce formulas, we obtain the new two formulas: $K = \{\{X', \overline{D'}\}, \{X', \overline{E}\}, \{D'\}, \{C, E\}, \{\overline{C}, \overline{E}\} \{Z, Z'\}, \{\overline{Z}, \overline{Z'}\}, \{X, \overline{D}\}, \{X, \overline{E}\},$ $\{\overline{D}, A\}, \{\overline{D}, B\}$ and $\phi = \{\{\overline{X'}, D', E\}, \{D, \overline{A}, \overline{B}\}, \{X, \overline{X'}, Z\}, \{\overline{X}, \overline{X'}, Z'\},$ $\{X, X', Z'\}\}$.

When $(\phi', S') = UP(\phi, S)$ is applied, new binary clauses are generated from clauses of ϕ. Those clauses are moved to K and deleted from ϕ. The process of moving binary clauses from ϕ to K, obligate us to update the closures and the base of K. Let us consider K' the set of new binary clauses to be added to K, such that $\upsilon(K') \subseteq \upsilon(K)$. $\forall c = \{x, y\} \in K'$, we consider the pair of implications: $\neg x \Rightarrow T(y)$ and $\neg y \Rightarrow T(x)$. Therefore, the original closures for x and y are updated as: $T(\neg y) = T(\neg y) \cup T(x)$ and $T(\neg x) = T(\neg x) \cup T(y)$.

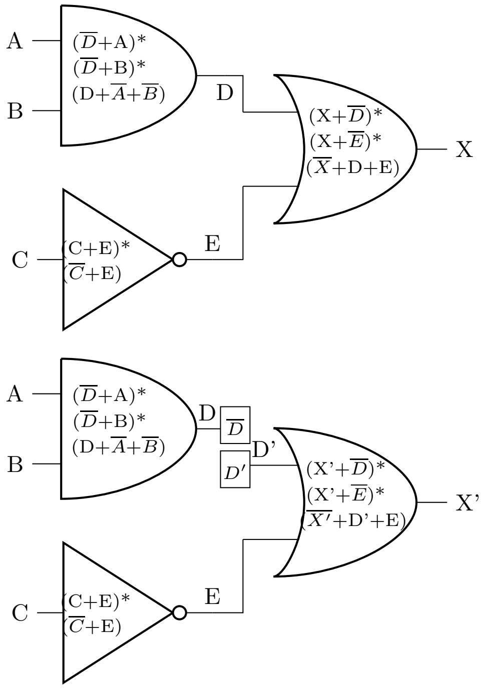

Fig. 1. Testing a combinatorial Circuit

Furthermore, $\forall T(l)$ where $\neg x \in T(l)$, it updates as $T(l) = T(l) \cup T(y)$, and $\forall T(l)$ where $\neg y \in T(l)$, it updates as $T(l) = T(l) \cup T(x)$. After updating the transitive closures, the new base for $K \cup K'$ has to be recomputed, and K is updated as: $K = K \cup K'$.

In the case of our example, we have that the base of our system of transitive closures is: $S(F) = \{D'\}$. By applying the rule of Pure Literal, D' is deleted from F, then $F[D'] = \{\{X'\}, \{X', \overline{E}\}, \{C, E\}, \{\overline{C}, \overline{E}\} \{Z, Z'\}, \{\overline{Z}, \overline{Z'}\}, \{X, \overline{D}\}, \{X, \overline{E}\}, \{\overline{X}, D, E\}, \{\overline{D}, A\}, \{\overline{D}, B\}, \{D, \overline{A}, \overline{B}\}, \{\overline{X}, X', Z\}, \{X, \overline{X'}, Z\}, \{\overline{X}, \overline{X'}, Z'\}, \{X, X', Z'\}\}$.

As $\{X'\}$ is now a unitary clause, then X' is added to the base $S(F) = \{D', X'\}$, and $F[X']$ is computed. $F[X'] = \{\{C, E\}, \{\overline{C}, \overline{E}\} \{Z, Z'\}, \{\overline{Z}, \overline{Z'}\}, \{X, \overline{D}\}, \{X, \overline{E}\}, \{\overline{X}, D, E\}, \{\overline{D}, A\}, \{\overline{D}, B\}, \{D, \overline{A}, \overline{B}\}, \{X, Z\}, \{\overline{X}, Z'\}\}$.

The transitive closures are computed over the new K:

$T(A) = \{A\}, T(\overline{A}) = \{\overline{A}, \overline{D}\}$ $T(B) = \{B\}, T(\overline{B}) = \{\overline{B}, \overline{D}\}$

$T(C) = \{C, \overline{E}\}, T(\overline{C}) = \{\overline{C}, E, X, Z', \overline{Z}\}$

$T(D) = \{D, A, B, X, Z', \overline{Z}\}, T(\overline{D}) = \{\overline{D}\}$

$T(E) = \{E, \overline{C}, X, Z', \overline{Z}\}, T(\overline{E}) = \{\overline{E}, C\}$

$T(X) = \{X, Z', \overline{Z}\}, T(\overline{X}) = \{\overline{X}, \overline{D}, \overline{E}, Z, C, \overline{Z'}\}$

$T(Z) = \{Z, \overline{Z'}, \overline{X}, \overline{E}, C, \overline{D}\}, T(\overline{Z}) = \{\overline{Z}, Z', X\}$

$T(Z') = \{Z', \overline{Z}, X\}, T(\overline{Z'}) = \{\overline{Z'}, Z, \overline{X}, \overline{E}, C, \overline{D}\}$.

And the last 3-CF is $\phi = \{\{\overline{X}, D, E\}, \{D, \overline{A}, \overline{B}\}\}$.

When $(\phi', S') = UP(\phi, S)$ is applied, then any variable $x \in v(S')$ does not appear more in ϕ', because if $x \in S'$ and $C = \{x, y, z\} \in \phi$ then C is satisfied and C does not appear more in ϕ'. Otherwise, if $C = \{\neg x, y, z\} \in \phi$ then the clause $\{y, z\}$ is generated instead of C, and it is added to K because this is a new binary clause. In whatever case, any variable in $S = Base(K \wedge \phi)$ does not appear more in ϕ'. Notice that during this step, a new base $S' \neq Base(K)$ could be generated since S' is looking for future feasible assignments for $(K \wedge \phi)$.

Our procedure looks afterwards for possible feasible assignments for $(K \cup \phi)$, in the next way: Let $S_1 = S \cup T(x)$ and $S_2 = S \cup T(\neg x)$ which are consistent because in other case x or $\neg x$ must be in S.

Let $(F_4, S_4) = UP(\phi, S_1)$ and $(F_5, S_5) = UP(\phi, S_2)$. And let K_1, K_2 be the 2-CF's forming F_4 and F_5, respectively. Our algorithm consists of the following steps:

1. If (($Nil \in F_4$ or $(K \cup K_1)$ is unsatisfiable) and ($Nil \in F_5$ or $(K \cup K_2)$ is unsatisfiable)) then $(K \wedge \phi)$ is unsatisfiable. Because any feasible assignment can not be extended with value for the variable x without falsifying $(K \cup \phi)$.
2. Else If ($Nil \in F_4$ or $(K \cup K_1)$ is unsatisfiable) then $T(x)$ can not be part of any model of $(K \cup \phi)$, then we can extend the base S as: $S = S \cup T(\neg x)$.
3. Else If ($Nil \in F_5$ or $(K \cup K_2)$ is unsatisfiable) then $T(\neg x)$ can not be part of any model of $(K \cup \phi)$, then we can extend the base S as: $S = S \cup T(x)$.
4. Otherwise, all new unitary clause, generated via UP, allows to update the transitive closures. That is, $\forall \{l\}$ generated by $UP(\phi, S_1)$ or $UP(\phi, S_2)$, we have that $T(x) = T(x) \cup T(\neg l)$, and $T(l) = T(l) \cup T(\neg x)$. These four steps are iterated until determine the satisfiability of $(K \wedge \phi)$.

Applying these last steps to our example, we have that $UP(\phi, T(\overline{D})) = \phi[\overline{D}] = \{\{\overline{X}, E\}, \{\overline{A}, \overline{B}\}\}$ so that ϕ is reduced from a 3CF to a 2CF. Afterwards, the transitive closures have to be updated. This step finishes until K is unsatisfiable, or ϕ is empty (and then $K \cup \phi$ is satisfiable), or $UP(\phi, S)$ does not generate new binary or unitary clauses. Let us consider now that $UP(\phi, S)$ does not generate neither new unitary nor binary clauses, then as second step in our procedure, we select a literal $x \in ((Lit(K) \cap Lit(\phi)) - S)$ such that $T(x)$ and $T(\neg x)$ are consistent and $|T(x)| + |T(\neg x)|$ is maximum into the set of common literals of K and ϕ.

Again, considering our example, we have that the last closure system is:
$T(A) = \{A, \overline{B}, \overline{D}\}$, $T(\overline{A}) = \{\overline{A}, \overline{D}\}$ $T(B) = \{B, \overline{A}, \overline{D}\}$, $T(\overline{B}) = \{\overline{B}, \overline{D}\}$
$T(C) = \{C, \overline{E}, \overline{X}, \overline{D}, Z, \overline{Z'}\}$, $T(\overline{C}) = \{\overline{C}, E, X, Z', \overline{Z}\}$
$T(D) = \{D, A, B, X, Z', \overline{Z}, \overline{B}, \overline{A}\}$ inconsistent
$T(\overline{D}) = \{\overline{D}\}$ thus $\overline{D}$ is added to S
$T(E) = \{E, \overline{C}, X, Z', \overline{Z}\}$, $T(\overline{E}) = \{\overline{E}, C, \overline{X}, \overline{D}, Z, \overline{Z'}\}$
$T(X) = \{X, Z', \overline{Z}, E, \overline{C}\}$, $T(\overline{X}) = \{\overline{X}, \overline{D}, \overline{E}, Z, C, \overline{Z'}\}$
$T(Z) = \{Z, \overline{Z'}, \overline{X}, \overline{E}, C, \overline{D}\}$, $T(\overline{Z}) = \{\overline{Z}, Z', X, E, \overline{C}\}$
$T(Z') = \{Z', \overline{Z}, X, E, \overline{C}\}$, $T(\overline{Z'}) = \{\overline{Z'}, Z, \overline{X}, \overline{E}, C, \overline{D}\}$

Note that in this case, the transitive closures obtained indicate that $\overline{D}$ is added to the base since an inconsistency was found for $T(D)$. Therefore, K has various models that satisfy ϕ. But, if we consider D as part of a model of K, all ternary clause contained in ϕ is transformed into a binary clause, and the process is finished indicating that $(K \wedge \phi)$ is satisfiable.

The great advantage of this method is that at least two thirds of the clauses generated are binary clauses (the 2-CF). This is true because each two-input unate gate contributes two binary clauses and one ternary clause. Unate gates with more than two inputs contribute more than two thirds of binary clauses. Fanout points, buffers, and inverters contribute with only binary clauses. In practice, applying this method, at most 80% to 90% of the clauses are binary clauses. Thus, our algorithm provides an efficient method to solve this class of problems.

6 Conclusions

We have designed a novel method for the incremental satisfiability (ISAT) problem. Thus, we have shown different cases where the ISAT problem can be solved in polynomial time.

Especially, considering an initial base K in 2-CF, we present an algorithm for solving the 2-ISAT problem that allows us to determine the satisifiability for $(K \wedge \phi)$, where ϕ is a 3-CF. Furthermore, we have established some tractable cases for the 2-ISAT problem.

We have illustrated the usefulness of our method in the area of automatic test pattern generation (ATPG) systems that allows to distinguish defective components from defect-free components in combinatorial circuits.

References

1. Buresh-Oppenheim J., Mitchell D.: Minimum 2CNF resolution refutations in polynomial time. Proc. SAT'07 - 10th int. Conf. on Theory and applications of satisfiability testing, (2007), pp.300-313
2. Cabodi G., Lavagno L., Murciano M., Kondratyev A., Watanabe Y.: Speeding-up heuristic allocation, scheduling and binding with SAT-based abstraction/refinement techniques. ACM Trans. Design Autom. Electr. Syst., 15(2), (2010)
3. De Ita Guillermo, Marcial-Romero R., Hernández J. A.: The Incremental Satisfiability Problem for a Two Conjunctive Normal Form, *Ceur-WS Lanmr 2016*, Vol.1659, (2016), pp.25–32
4. Eén N., Sorensson K.: An Extensible SAT-solver. In Enrico Giunchiglia and Armando Tacchella, editors, Selected Revised Papers of 6th International Conference on Theory and Applicationsof Satisfiability Testing (SAT), Santa Margherita Ligure, Italy, LNCS Vol. 2919, (2003), pp. 502-518
5. Gusfield, D., Pitt, L.: A Bounded Approximation for the Minimum Cost 2-Sat Problem. Algorithmica 8, (1992), pp. 103-117
6. Hooker J.N.: Solving the incremental satisfiability problem. Journal of Logic Programming 15, (1993), pp.177-186
7. Gutierrez J., Mali A.: Local search for incremental Satisfiability. Proc. Int. Conf. on AI (IC-AI'02), Las Vegas, (2002), pp. 986-991
8. Larrabee T.: Test Pattern Generation Using Boolean Satisfiability. IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, Vol. 11:1, (1992), pp. 4-15
9. Mouhoub M., Sadaoui S.: Systematic versus non systematic methods for solving incremental satisifiability, Int. J. on Artificial Inteligence Tools, Vol. 16:1, (2007), pp.543-551
10. Wieringa S.: Incremental Satisfiability Solving and its Applications. PhD thesis, Department of Computer Science and Engineering, Alto University Pub., (2014)

Automatic Detection of Diabetic Retinopathy using Image Processing

Avila-Hernandez J. J., Guzmán-Cabrera R.*, Torres-Cisneros M.
Gómez-Sarabia C. M. and Ojeda-Castañeda J.
Campus Irapuato-Salamanca, Universidad de Guanajuato, México
*guzmanc@ugto.mx

Abstract. We present a reliable, automatic digital procedure for diagnosing and following-up Diabetic Retinopathy. Specifically, we implement a procedure that records retinal areas of interest, which are suitably magnified and binarized employing fractal image techniques. Then, from two well-known retinal data banks (ROC and DIARETDB), we create an automatic Bayesian classifier for the decision making stage. We show that the reported procedure is capable of detecting 96.7 % of the Diabetic Retinopathy lesions.

Keywords. Retinopathy Diabetic, Image Processing, Aneurysm Format

1 Introduction

The World Health Organization considers Diabetes Mellitus (DM) as a public health problem, due to its high prevalence, its associated morbidity and its mortality. It is known that within 15 years of the first diagnosis, some patients develop diabetic retinopathy; which affect anatomically and functionally the performance of the human retina. Hence, there is a need for an early detection of DM, as well as for following up any retinal [1].

For early detection, a trained expert performs the diagnosis; and after the diagnosis is made, at fixed periods, a physician needs to examine the retina for identifying the evolution of the lesions. Within 15 years of the first diagnosis, approximately 98 % of people with diabetes type 1, and 78 % with type 2, develop some form of retinopathy [1].

The World Health Organization reports that by 2030 the prevalence will be 300 million people with DM type 2. In Mexico the number of people with DM fluctuates between 6.5 and 10 million [2]. Furthermore, the World Health Organization estimates that Retinopathy Diabetic (RD) produces 4.8% of the 37 million blind people in the world [3]. Hence, DM is a serious health problem.

Here, we discuss an automatic procedure for identifying and for extracting abnormalities, as a hemorrhages and aneurysms. We believe that the procedure will provide a tool, to the specialists, for early diagnosis timely following-ups. To this end, we employ fractal analysis and image segmentation. In the section 2, we revise the fundamentals of image processing, including a brief discussion on fractal analysis. In section 3, we describe the proposed procedure. In section 4, we show the main results of this work. And in section 5, we summarize our contribution.

2 Basics on Image Processing

Digital image processing is a set of methods and tools used often in several branches of science and engineering. As part of these set of tools, image segmentation is used for detecting edges between different regions in a picture. And then after edge detection, by using other set of tools, one can isolate different regions.

For medical applications, Bhattacharya and Das have employed discrete wavelet transformations, due to its multiresolution properties [4]. Sung et al. have used mathematical morphology operations, together with wavelet transforms, for locating Regions of Interest (ROI) in a picture [5].

Rather recently, a procedure was tested for diagnosing cancer, from screening mammographs. This procedure employs morphological operators, machine learning techniques and a clustering algorithm for intensity-based segmentation [6]. By using this procedure, it is possible to distinguish masses and micro calcifications from background tissue.

It is convenient to recognize that image segmentation has been found applications for detecting textures, for identifying tumors and lesions in photo-acoustic images [7], and in thermographic images [8].

2.1 Fractal Geometry and Analysis

The term fractal is commonly used to describe the family of non-differentiable functions that are infinite in length. Fractal objects contain structures that are nested within one another. There are 2-D fractal pictures, as well as 3-D fractal images. In all cases, the main feature of a fractal image is that it characterizes the way in which a quantitative dataset grows in mass, with linear size [9].

Fractal geometry offers an alternative model for seeking regularities in picture, by looking at its parts at different scales. To this end, the object is expressed as the geometric limit of an iterative process. However, there is a risk of generating fractures, which lead to the absence of differentiability at the boundaries of the regions in a picture [10].

Various natural phenomena display self-similarity, which is characterized by a parameter called fractal dimension, *D*. This real number is used as an exponent for describing how the object structure is repeated *N* times, at different scales here denoted by the real number *r*. These quantities are interrelated by Equation 1a:

$$N=1/r^{D}. \tag{1a}$$

Equivalently, we have that

$$D=log\ (N)/Log(1/r). \tag{1b}$$

In this contribution, the fractal dimension D characterizes a pathological tissue. The rationale being that within the human body healthy tissues have a high degree of similarity. Fractal descriptors can be found in [11, 12].

3 Proposed Procedure

In figure 1, we depict the flowchart of the proposed methodology. As a first step of the proposed methodology, the input is a RGB image. We select the green channel, since the green channel offers a high contrast; when identifying lesions in the retinal images.

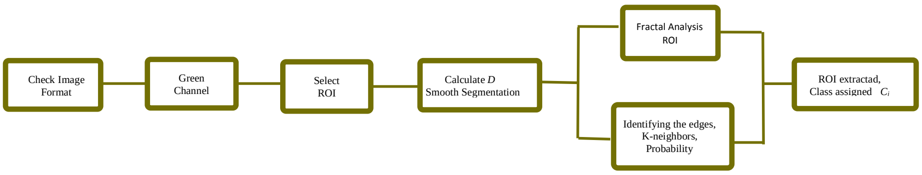
Check Image Format
Green Channel
Select ROI
Calculate D Smooth Segmentation
Fractal Analysis ROI
Identifying the edges, K-neighbors, Probability
ROI extractad, Class assigned C_i

Fig. 1. Flowchart of the proposed methodology

Then, we select the region for performing the fractal analysis of the image. Next, we evaluated the parameter *D* for performing the segmentation of the image. Once that the image is segmented, we complete the fractal analysis over the selected region. After that, we extract the ROI; and simultaneously we classify the image.

4 Main Results

In figure 2, we exemplify the use of the proposed methodology. In Figure 2 A), we display the original image; we use a frame for locating area of interest. In Figure 2 B), we show a magnified version of the selected area. In Figure 2 C), we present the image after fractal analysis, which describes the ROI extraction process.

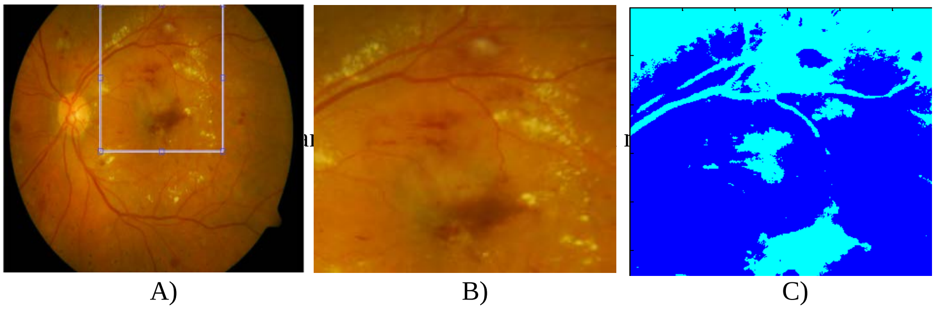

Fig. 2. Example of use of the proposed methodology

In Figure 3 A), we report again the process of selecting an area of interest. In Figure 3 B), we show a magnified picture of the selected area. And in Figure 3 C) we display an image after fractal analysis. This latter picture indicates the absence of retinopathy.

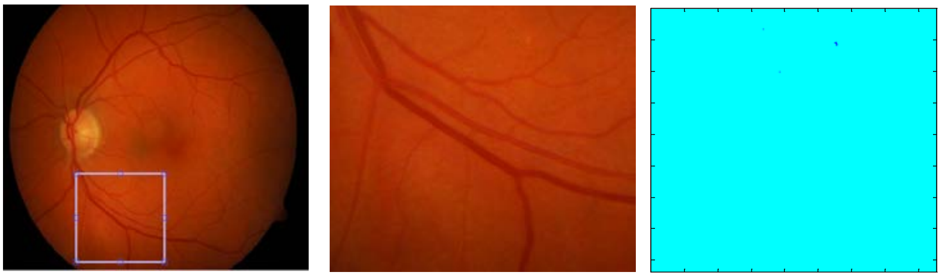

Fig. 3. Result of processing in an image without retinopathy

For our present work, we employ 75 retinal images. Of these images 40 were pictures for training our procedure. And the remaining 35 pictures were used for testing the trustworthiness of our procedure. In the test pictures, we select at least one area of interest. For obtaining the ROI, we employ a Bayesian classifier for assigning the class with retinopathy or without retinopathy. The classifier was applied to each of the 35 testing pictures. In 96.7% of the cases, our procedure could assign the right class.

5 Conclusions

We have focused our attention on a pathological disorder of the eye, known as Diabetic Retinopathy; which is considered the third cause of irreversible blindness in the world. We have shown that image processing techniques can provide tools for both early diagnoses, as well as for following-ups. Furthermore, we have indicated that image processing techniques are also useful for storing and for sharing relevant medical data. We have shown that by using a Bayesian classifier, one can have an automatic decision making procedure; which can detect 96.7 % of the Diabetic Retinopathy lesions.

References

1. WHO: Prevention of blindness from diabetes mellitus, Report of a consultation in Geneve, Switzerland, 2005
2. IMSS: Diagnóstico y tratamiento de la retinopatía diabética, guía de referencia rápida, Catálogo Maestro de Guías de Práctica Clínica: IMSS-171-09, 2015
3. The Royal College of Ophthalmologists: Diabetic Retinopathy Guidelines, 2012
4. M. Bhattacharya and A. Das: Fuzzy logic based segmentation of microcalcification in breast using digital mammograms considering multiresolution. International Machine Vision and Image Processing Conference, pages 98–105, 2007

5. Y. Sung-Nien, L. Kuan-Yuei, and H. Yu-Kun.: Detection of microcalcifications in digital mammograms using wavelet filter and Markov random field model. Computerized Medical Imaging and Graphics, 30(3):163–173, 4 2006

6. R. Guzmán-Cabrera, J. Guzmán-Sepúlveda, M. Torres-Cisneros, D. May-Arrioja, J. Ruiz-Pinales, O. Ibarra-Manzano, G. Aviña-Cervantes and A. González-Parada.: Digital image processing technique for breast cancer detection. International Journal of Thermo-physics, 34(8):1519–1531, 2013

7. R. Guzmán-Cabrera, J. R. Guzmán-Sepúlveda, M. Torres-Cisneros, D. May-Arrioja, J. Ruiz-Pinales, O. Ibarra-Manzano, and G. Aviña-Cervantes: Pattern recognition in photo-acoustic dataset. International Journal of Thermo-physics, 34(8):1638–1645, 2013

8. R. Guzmán-Cabrera, J. Guzmán-Sepúlveda, A. González-Parada, J. Rosales-García, M. Torres-Cisneros, and D. Baleanu.: Digital processing of thermographic images for medical applications. Revista de Chimie, 67(1):53–56, 2016

9. P. Mignot Jacques Levy-Vehl and Jean-Paul Berror: Multifractal, texture, and Image analysis. CVPR, pages 661-664, 1992.

10. A. Silvetti, C. Delrieux: Análisis Multifractal Aplicado a Imágenes Médicas, SeDiCI, Argentina, 2014

11. C. Evertsz and B. Mandelbrot. Multifractal Measures. Chaos and Fractals. Springer, Amsterdam,1992

12. A. Fournier, D. Fussell and L.: Carpenter. Computer Rendering of Stochastic Models. Communications of the ACM, 25(6): 371-384,1992

13. J.J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever and B. Van Ginneken: Ridge based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging, 2004

14. S. Jiménez, P. Alemany, I. Fondón, A. Foncubierta, B. Acha y C. Serrano: Automatic detection of vessels in color fundus images, Arch. Soc. Esp. Oftalmol. v.85 n.3, 2010

Analyzing the Effect of Family Factors over the Scholar drop out at Elementary School using Classification and Association Rules Techniques

Silvia Beatriz González Brambila and Josué Figueroa González

Universidad Autónoma Metropolitana,
Av. San Pablo 180, Ciudad de México, México
{sgb,jfgo}@correo.azc.uam.mx

Abstract. Data Mining offers great opportunities for analyzing data related with several themes, one of the most interesting is the educative environment, which has a lot information about many areas which can be improved. Scholar drop out, it's one of the biggest problems that education faces, but the great amount of factors that cause it, make it difficult for analyzing. This work presents an analysis of the elementary school drop out problem using techniques like decision tree and generation of association rules for obtaining models that allow to identify the most important family aspects that causes scholar desertion.

Keywords. Association Rules, Decision Trees, Educational Data Mining, Mining Educational Data, Studies Drop Out

1 Introduction

Data Mining (DM) focuses in analyzing big volumes of information in order to obtain patterns or knowledge about different topics for explaining, classifying or predicting some kind of phenomena and helping in making decisions. In the last decades, there has been a rise in the quantity of available information related with sectors like commerce, finances or consumer preferences. Also, the use of technology in the educative environment has also risen, so the amount of data and information related with several aspects of education. Applying technology over traditional educative systems, and other like e-learning systems have allowed gathering a lot of information about students, courses, schools and other elements; processing this information using DM techniques has created a relatively new branch of DM called Educational Data Mining (EDM). EDM includes the use of DM techniques for discovering patterns in large amount of information generated in an educative environment. This concept first appeared in year 2000 with a set of conferences, but it has been a rise in the amount of researching and works related with this topic [12]. Educational environment offers several opportunities for applying techniques for discovering knowledge, like: predicting students performance, planning courses, clustering students according certain features, etc.; being one of the most worked areas, the prediction of students' performance over their studies, a single class or even an exam or exercise.

In any educative system, probably the biggest problem is related with the scholar drop out, which is present since the elementary, until superior level. The main issue, is that there are a lot of factors which can cause that a student leaves its studies. Family, personal, social, academic or labor factors affects all students, and identifying the most important could be really difficult due to the great amount of information which should be processed. Considering this, the main goal of this work is to apply two of the most common DM techniques, classification and generation of association rules, for obtaining predictive models and rules among events or situations which allow to understand the effect of family factors over the scholar drop out at elementary school.

2 Knowledge Discovery using Data Mining

Discovering knowledge requires a set of stages that makes easier the process of finding interesting patterns over data. This process, called Knowledge Discovery on Databases (KDD) is composed by different steps [7]: obtaining information, cleaning process or data cooking, use of DM techniques and interpreting for obtaining knowledge. In the DM stage, the applied techniques depend on the goal to be reached, in general these techniques are: classification, clustering, regression and association rules, having each on of these, different algorithms and techniques for generating results [8].

2.1 Classification

Classification or supervised learning, is on the most used DM techniques; it uses a learning scheme where, using a set of classified data (training data), a model which allows predicting new not classified data, is generated. This two steps are called training or learning stage and the classification stage. Most common algorithms used in classification technique are: decision trees, neural networks and Naive Bayes; once a model has been generated, its efficiency it's measured using another set of data (test data).

2.2 Decision Trees

A Decision Tree is one of the most used predictive models in DM. In a Decision Tree, leafs represents a decision or classification and branches a set of characteristics that lead to a particular decision. As a tree, it has a root node, called the best predictor, this means that it's the most important factor or variable for taking a decision about a classification. Trees are used in DM for obtaining models which predicts the value of a variable, called decision. Once the model is constructed, the characteristics of a non classified data are taken and depending their values, a path is followed from the root to a leaf that indicates the corresponding class.

At the moment of constructing a Decision Tree, the main problem is to decide which of the variables will be the root node, and then, decide the order of the rest of the variables. The nearer a node is to the root, represents that it is more important for assigning a classification. This problem is solved by finding a variable that better divides the target destiny considering the purity of its children set. Purity refers to how mixed are the goal values in a node. Purity measure is known as information, and the concept of impurity is known as Entropy, and is defined as follows:

$$\sum_{i=1}^{k} P(C_i|D)\, log_k\, P(C_i|D)) \tag{1}$$

Where:

$$P(C_i|D) = \frac{amount\, of\, oservations\, in\, D\, with\, value\, C_i}{amount\, of\, oservations\, in\, D} \tag{2}$$

Another concept used during the generation of a Decision Tree besides Entropy is the Information Gain and it's based in the decrease of Entropy after a set of data has been divided. Information Gain is defined as:

$$G(S, A) = Entropy(S) - \sum_{v=values(A)} \frac{S_v}{S}\, Entropy\,(S_v) \tag{3}$$

Where *values(A)* represents the set of all the possible values of the attribute or property A and Sv is the sub set of elements in S from which the attribute A has a value of *v*.

2.3 Measuring the Classification Efficiency

Data for generating a Decision Tree are be divided in two: training or learning data and test data. The model is generated using the training data and then, it's tested using the test data for determining how efficient is. The recommended percentage for each one is 70% to 80% for training and 20% to 30% for testing. The model determines a value for the decision variable (predicted value) of a test data and this is compared with the real value. Comparing predicted value and real value for all the elements of the test data set, can be calculated the efficiency of the tree.

2.4 Association Rules

Association rules are used for showing the relationships that exists in a set of items. In a formal way, an association rule is defined as: let I = $\{I_1,I_2,\ldots,I_m\}$ a set of attributes known as items, and T a set of transactions $\{t_1,\, t_2,\, \ldots,\, t_n\}$ represented as t[k] = 1 if t is related with I k and t[k] = 0 otherwise. Let X a set of some of the elements in I, a transaction satisfies X if for all the elements I$_k$ in X, t[k] = 1. An association rule is an implication represented as X ⇒ I ij where X is a set of some of the elements in I and I_{ij} is an element of I which is not present in X. In this way, the rule X ⇒ I_{ij} is satisfied in the set of transactions T, if certain percentage of transactions in T that satisfy X, also satisfy I_j[1].

2.5 Measuring the Importance of a Rule

It's very common that a great amount of association rules are generated, most of them can be redundant or not significant. For this reason, measures for knowing the importance of a rule have been developed [13]. From this measures, the most used are: Support, Confidence and Lift.

Support of a set of elements A represents the percentage of transactions which contains A in a set of transactions T. Support is defined as:

$$support(A) = \frac{|A|}{|T|} \tag{4}$$

Confidence is the amount of transactions which contains A as an antecedent and B as a consequence. A can represent a single element or a set of elements. Formally, confidence is defined as:

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \tag{5}$$

Lift represents the occurrence frequency of A and B respect an expected value, considering that A and B are independent. Lift is defined as:

$$lift(A \Rightarrow B) = \frac{support(A \Rightarrow B)}{support(A) * support(B)} \tag{6}$$

The value of lift shows how strong is the relationship among A and B according to:

- If Lift $<$ 1, it's considered that A and B have a weak relationship and are called substitutes
- If Lift $=$ 1, it's considered that a relationship among A and B is because a random behavior
- If Lift $>$ 1, it's considered that A and B have a strong relationship and are called complements

3 Related Works

One of the most recurrent researches in EDM is related with using classification techniques for generating models which classify or predict the behavior of students during their studies, a course, exam or other activities.

In [3] it's analyzed the data from students, which is obtained from a data warehouse, using a decision tree and other techniques for generating a model that predicts their performance considering their verbal and mathematical abilities. The efficiency of all the techniques is compared. In [11] some classification techniques are used for generating models which predict the final score of students in a certain course based on their personal, social and academic characteristics. In [4] some of the algorithms for creating decision trees are used for generating models which predict the performance of students considering academic, familiar and demographic factors. In [10] generation of association rules are used for analyzing the effect of teaching in different languages over the amount of regis tered students in several courses. In [5] are identified the elements which have in common the best students in a course using association rules and other DM techniques. In [9], the phenomena of dropping out an on-line course is analyzed using decision trees. A model for identifying the students which have greater risk of dropping the course is generated.

4 Knowledge Discovery

For obtaining the prediction model and the association rules, were followed the steps considered in the KDD process. Data was obtained from the National Poll about Scholar Drop out at Medium High Level [14]. Despite of the poll was focused in medium high level students, there were a lot of surveyed which didn't finish the elementary school, so they were considered for this work. Were considered 699 people, 407 didn't finish elementary school and 292 did it.

The decision tree was generated using Classification and Regression Trees algorithm (CART) [6], and for generating the association rules, was used the Apriori algorithm [2], this was performed with the software R.

4.1 Pre-processing the Information

The goal of this stage is preparing the information for applying DM techniques. The poll had questions grouped in three aspects: personal and family, academic and labor. For this work, only were considered the personal and family factors; initially were considered 19 variables and one decision variable. For helping the interpretation of the models, it was assigned a letter for each variable, variables, including their letter and possible values are shown at Table 5 at the end of document.

4.2 Classification using Decision Trees

It was chosen the variable ZZ (Studies finished) to be the predicted variable with two possible values, Yes, if the student finished their studies or No, otherwise. 75% of the data (524) were used for the training process and 25% (175) for testing. The set of rules obtained from the model is presented in Table 1, and the Decision Tree obtained is shown in Figure 1.

The represented variable by each letter can be reviewed in Table 5. The values of S correspond to:

- NO: Didn't finish Elementary School
- EL: Finished Elementary School
- ME: Finished Medium School
- MES: Finished Medium High School

Table 1. Rules obtained from Decision Tree

Rule
IF L = ”NO”, THEN Finish = ”NO”
IF L = ”YES” AND S = ”ME” OR ”MES” OR ”SUP”, THEN ZZ = ”YES”
IF L = ”YES” AND S = ”NO” OR ”EL” AND F = ”YES”, THEN ZZ = ”NO”
IF L = ”YES” AND S = ”NO” OR ”EL” AND F = ”NO” AND O = ”YES”, THEN Z = ”NO”
IF L = ”YES” AND S = ”NO” OR ”EL” AND F = ”NO” AND O = ”NO”, THEN Z = ”YES”

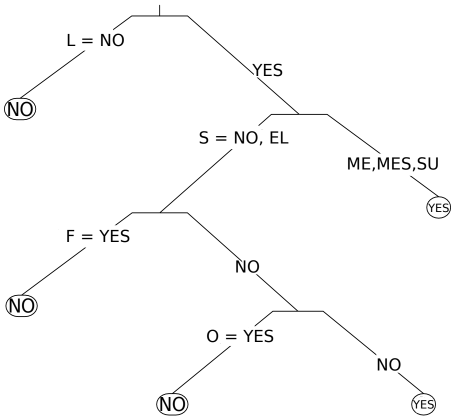

Fig. 1. Decision tree obtained after processing the information

- SU: Started or Finished Superior School or higher

The efficiency of the model was measured comparing the predicted value using the tree, against the real value for the test data. The results are shown on Table 2.

Table 2. Measuring the efficiency of the model

	No	Yes
No	80	20
Yes	27	48

This represents that, from 100 cases with ”NO” value, 80 were classified in a right way, and 20 in a wrong way. 27 cases with ”YES” value were assigned by the model to ”NO” and 48 were predicted as ”YES”. Considering these results, the efficiency of the model was 73.142%.

4.3 Obtaining Association Rules

Were generated rules for the possible values of ZZ (Studies finished). The antecedents for each possible values are shown in Table 3 for Finishing and, in Table 4 for Not Finishing. The importance of a rule was measured using the lift value.

Table 3. Antecedents related with Finishing the Elementary School

Rule	Confidence	Lift
{F=”NO”,L=”YES”,N=”NO”,T=”ME”}	0.8301887	1.987335
{C=”NO”,L=”YES”,N=”NO”,T=”ME”}	0.8301887	1.987335
{L=”YES”,N=”NO”,O=”NO”,T=”ME”}	0.8269231	1.979518
{L=”YES”,N=”NO”,O=”NO”,T=”ME”},	0.8200000	1.962945

Table 4. Antecedents related with Not Finishing the Elementary School

Rule	Confidence	Lift
{E=”NO”,L=”NO”,O=”YES”,S=”NO”}	0.8409091	1.444215
{B=”NO”,F=”YES”,R=”NO”,T=”EL”}	0.8409091	1.444215
{B=”NO”,D=”NO”,F=”YES”,S=”NO”}	0.8367347	1.437046
{L=”DIRECT”,B=”YES”,O=”NO”,T=”NO”},	0.8367347	1.425803

The represented variable for each letter can also be reviewed in Table 5. The values for T are the same than the ones for S in the Classification with Decision Trees section.

4.4 Interpreting Models

Once the models and rules were obtained from the decision tree and the association rules, the results must be analyzed. It’s interesting that the variable L (considering studies few important) with a value of ”NO” has a lot of importance for dropping out the school, it is expected that the opposite occurs, it’s supposed that considering the studies important leaves to continue them, but at elementary school, maybe the interest for studying of a kid it’s not a decisive factor. In this case, it’s more relevant aspects related with the parents, both techniques show the importance of the studies of the parents. In the decision tree, values for S (studies of the father) that have a higher level than elementary school are related with finishing school, the same occurs with association rules, where the value of T (studies of the mother) is higher than elementary school it's related with the consequence of finishing. Having not studies or only the elementary school for both or any of the parents is related with dropping the school. The F variable (lack of money) is also related with not finishing the school in both techniques; indicating that it could be a relevant factor in the scholar drop out. The same occurs with variable O (brothers that dropped out the school),in the decision tree and the rules associated with not finishing the studies, this value is present indicating that if a brother or sister abandoned it's studies, its probable that the other can do it. Other variables which appear on the rules are not present in the decision tree, but several of the ones with the highest value of lift do it. Considering this, it can be concluded that both models are right and can help in studying the problem of scholar drop out at elementary school.

Although the results of both models are congruent, the percentage of efficiency of the Decision Tree can be considered low, a good percentage should be above 90%. This may be because of the small quantity of data, with more data, there are more combinations for generating better models.

5 Conclusions

Applying Data Mining Techniques to an educative environment, offers a big amount of opportunities for studying a lots of aspects that occurs in education. From the performed work, it can be concluded that depending the desired goal, the appropriate technique of Data Mining must be chosen. Following the the KDD process can be helpful for easing the whole process. It's important to fulfill a right cleaning and pre-processing of the data for obtaining better results independently the used technique. Also, testing different sets of data it's necessary for obtaining better models, this specially related with Decision Trees, but also testing different parameter values for applying them to the association rules algorithm helps in obtaining better results. As mentioned, the amount of information is an important topic, having more quantity would help in obtaining better models. Related with the quality of the models, having an adequate way for measuring the efficiency of each model or set of rules it's essential for acquiring the right knowledge. In the obtained Decision Tree, the efficiency can be considered low, however some of their nodes also appear as relevant rules in the association rules model, which validate the obtained results.

Having the models or rules it's not the end of the process, a stage of interpreting those results it's necessary for finally obtaining knowledge that helps in making decisions about a certain problem.According the obtained results in each model, can be concluded that promoting a higher level of studies in the parents (before, or even, having children) could reduce the elementary school desertion problem. Also politics for improving the income for homes can be a good solution.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Acm sigmod record vol. 22, No. 2, 207–216. ACM, (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: 20th int. conf. very large data bases, VLDB, pp. 487–499. (1994)
3. Agarwal, S., Pandey, G. N., Tiwari, M. D.: Data mining in education: data classification and decision tree approach. International Journal of e-Education, e-Business, e-Management and e-Learning, 2(2), pp. 140–144, (2012)
4. Bhardwaj, B. K., Pal, S.: Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418. (2012)
5. Baradwaj, B. K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417 (2012)
6. Fayyad, U. M., Irani, K. B.: The attribute selection problem in decision tree generation. In: AAAI. pp. 104–110, (1992)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine, 17(3), pp. 37–54, (1996)
8. Goebel, M., Gruenwald, L.: A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, 1(1), pp. 20–33, (1999)
9. Kotsiantis, S. B., Pierrakeas, C. J., Pintelas, P. E.: Preventing student dropout in distance learning using machine learning techniques. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer Berlin Heidelbergpp. pp. 267–274. (2003)
10. Pandey, U. K., Pal, S.: A Data mining view on class room teaching language. arXiv preprint arXiv:1104.4164. (2011)
11. Ramesh, V., Parkavi, P., Ramar, K.: Predicting student performance: a statistical and data mining approach. International journal of computer applications, 63(8), pp. 35–39. (2013)
12. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27 (2013)
13. Sheikh, L. M., Tanveer, B., Hamdani, M. A.: Interesting measures for mining association rules. In: Multitopic Conference, Proceedings of INMIC 2004. 8th International pp. 641–644. IEEE, (2004)
14. Subsecretaría de Educación Media Superior. Encuesta Nacional de Deserción en la Educación, `http://www.sems.gob.mx/sems/encuesta_nacional_desercion_ems`

Table 5. Personal factors for rule generation

Key	Variable	Value
A	While studying, lived with	Alone Direct Other relatives Friends Own family
B	Desire for continuing studying	Yes, No
C	Influence of parents for dropping out	Yes, No
D	Influence of other relatives for dropping out	Yes, No
E	Influence of friends for dropping out	Yes, No
F	Lack of money in home	Yes, No
G	Low desire for studying	Yes, No
H	Bullying	Yes, No
I	Problems in home	Yes, No
J	Preference for studies of other relatives	Yes, No
K	Serious illness or decease of a relative	Yes, No
L	Considering studies few important	Yes, No
M	Low self steem	Yes, No
N	Closest friends dropped out school	Yes, No
O	Brothers or sisters dropped out school	Yes, No
P	Cigar consume	High Medium Low No
Q	Alcohol consume	High Medium Low No
R	Drugs consume	High Medium Low No
S	Father's level studies	No Elementary Medium Medium Superior Superior
T	Mother's level studies	No Elementary Medium Medium Superior Superior
ZZ	Studies finished	Yes No

Use of Text Patterns for Evaluating Concepts in Corpora of Restricted Domain*

Mireya Tovar[1], David Pinto[1], Azucena Montes[2], and Gabriel Serna[3]

[1]Benemérita Universidad Autónoma de Puebla,
Faculty Computer Science, Puebla, Mexico,
{mtovar,dpinto}@cs.buap.mx
http://www.lke.buap.mx/

[2]Instituto Tecnológico de Tlalpan, TecNM
ing_tlalpan@tecnm.mx

[3]Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET),
Department of Computer Science, Mexico
gabriel@cenidet.edu.mx

Abstract. In this paper we present an approach for evaluating concepts in ontologies of restricted domain. The aim is to validate those concepts by discovering the linguistic patterns that have been used in the process of constructing the ontology. For this purpose, we have evaluated the concepts stored in three different ontologies from the following restricted domains: oil recovery, artificial intelligence and e-Learning. Two types of procedures were used for evaluating the ontological concepts: first, an intrinsic process using only the target ontology was carried out, whereas the second one procedure employed an extrinsic evaluation in which a reference corpus supports the validation process. The experimental results show a satisfactory performance when the approach proposed is executed in each ontology evaluated.

Keywords. Ontology Domain, Concept Evaluation, Linguistic Patterns

1 Introduction

An ontology can be defined as “an explicit formal specification about a shared conceptualization” [1]. In general, this type of semantic resources are made up of concepts or classes, relations, instances, attributes, axioms, restrictions, rules and events.

Nowadays, there are plenty of computational systems claiming to automatically generate ontologies, however, in the major of the cases, those systems lack of the particular process of automatic evaluation, therefore, the quality of the semantic resources generated is unknown.

* This research work has been partially supported by PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, by VIEP-BUAP project 00570 and CONACyT project 257357.

The evaluation of ontologies task aims to measure the quality of such linguistic resources. The final aim is to facilitate the work of the ontological engineer or the domain expert when they verify the quality of ontologies with a considerable high number of items inside it. This analysis is time consuming (hours/person). The evaluation process is not a trivial one, because it is necessary to chose the items of the ontology that should be considered when evaluating it, so as the particular criteria to employ in the evaluation (see [2]).

In this research work, we propose an automatic method based on Natural Language Processing (NLP) for the evaluation of concepts of an ontology of restricted domain. The proposed methodology assumes that the ontology has been constructed in an automatic, semi-automatic or manual way, and that the reference corpus (collection of documents) is semantically associated with the target ontology. The aim is to "validate", en a first stage, the quality of the ontological concepts. The evaluation of these concepts is carried out by means of two independent ways: 1) Using a reference corpus for the target ontology, and 2) Using the same ontology as training and test set, i.e., without any reference corpus.

The remainder of this paper is structured as follows. In Section 2 we present research works related with the automatic extraction of candidate terms. We also discuss the approaches for the automatic identification of restricted domain ontological concepts (see Section 3), including the construction and evaluation of morpho-syntactical patterns (see Section 4). Finally, conclusions and findings are given in Section 5.

2 Related Work

Most of the evaluation approaches for semantic resources in literature are focused on the evaluation of the ontology structure, assuming that the ontological concepts have been correctly defined by the ontological engineer. However, with the aim of provide a much more wide view of the evaluation process, we consider different approaches, even if some of those approaches are not so popular. The reviewed research works are firstly categorized in terms of the type of process employed when the ontology was constructed: automatic, semi-automatic or manual, and secondly, in terms of the practical creation purpose for the ontology. The major of the literature works may be categorized as follows [2]:

Human-Based Evaluation following Criteria, Standards and Requirements This type of evaluation allows to evaluate certain characteristics of the ontology, providing a numeric score [3]. Some features considered in this kind of evaluation are: Completeness, Correctness, Readability and flexibility [4, 5].

Application-Based Evaluation This type of evaluation consists of testing the performance of the ontology in a given application. For example, answering user questions using an ontology [6], [7].

Gold Standard-Based Evaluation In an evaluation based on Gold Standard, the quality of the ontology is expressed by means of the similarity of it with respect to another ontology built manually, i.e., a gold standard ontology [7], [8], [9]. The comparison of both ontologies can be on two levels: lexical (similarity between concepts), and conceptual (similarity between relationships and taxonomies) [10]. In [11] different evaluation measures for lexical and semantic levels of ontologies are presented.

Reference Corpus-Based Evaluation In this case, the quality of the ontology is represented by the degree of a corpus topic covered by the ontology. For example, precision and recall metrics were used in [12] to evaluate the degree of lexical similarity of the ontology triplets with respect to elements extracted from the reference corpus. Furthermore, in [13] a probabilistic approach is used to compare the labels of an ontology with respect to a set of important terms identified in the reference corpus (extended by adding two levels of hyperonyms from WordNet).

We are interested in the automatic extraction of terms in a corpus domain from morpho-syntactical patterns identified in the words that form the concepts defined in the ontology. The morpho-syntactic patterns are built automatically from the words that integrated the concept using clustering. These patterns are used to extract the terms in the corpus. Next, we validate these terms with the concepts of the ontology.

3 The Proposed Approach for Evaluation of Ontological Concepts

In this section, we present an approach for the evaluation of restricted domain ontological concepts. First, we present the target ontologies to be evaluated (see Table 1), together with the metrics employed for the evaluation. Thereafter, we introduce an approach based on morpho-syntactical pattern for evaluating the concepts by means of a v-fold validation process, using the same ontology. These patterns are employed in another stage of the complete process for extracting and evaluating the candidate concepts by using a restricted domain corpus associated to the ontology.

In Table 1 we show the total number of concepts (C) and hierarchical relations (R) of three restricted domain ontologies that were evaluated following the proposed approach of this research work. In the same table is also shown the total number of documents (D) of each reference corpus, the number of tokens (T) and the vocabulary size (V). The restricted domains considered in this paper are: oil improved recovery methods (OIL), artificial intelligence (AI), and standard e-Learning SCORM (SCORM)[1] [14].

The evaluation of the proposed approach is carried out by means of metrics traditionally employed in information retrieval, such as precision (P), recall (R)

[1] AI and SCORM ontologies are freely available at http://azouaq.athabascau.ca/

Table 1. The restricted domain ontologies and reference corpora employed in the experiments

Dominio	Ontology		Corpora		
	C	*R*	*D*	*T*	*V*
OIL	48	37	575	9,727,092	188,047
AI	276	205	8	10,805	2,180
SCORM	1461	1038	36	32,644	2,154

and F-measure (F). Precision measures the proportion of candidate concepts that really belong to the target ontology between the number of terms identified as candidate concepts by the system. Recall measures the proportion of candidate concepts identified by the system as ontological concepts between the number of all the real ontological concepts of the ontology. F-Measure is an harmonic measure that combines precision and recall.

The evaluation method for ontological concepts considers firstly the automatic identification of morpho-syntactical patterns employed by the ontological engineer (a human being or a computational system) in the original construction of the ontological concepts. Thus, we present the method used in the construction of linguistic patterns and the validation of it.

3.1 Construction of Morpho-Syntactical Patterns

The method employed for constructing the patterns and its use for extracting candidate terms, either from the ontology or from the reference corpus, is presented as follows:

1. To apply a Part-Of-Speech (PoS) tagger to the ontological concepts. In this case, we use TreeTagger [15].
2. To identify the morpho-syntactic tags for each ontological concept.
3. To cluster the morpho-syntactical PoS tags. We intercepted the morpho-syntactic PoS tags for forming clusters.
4. To construct regular expressions for the cluster of morpho-syntactical tags.
5. To extract candidate terms by applying the mentioned regular expressions to the reference corpus or ontology.

The results obtained by the aforementioned approach when it was applied to each ontology and corpus are shown in Tables 2, 3, and 4. The first column indicates the number of patterns identified in the ontology, the second column indicates the frequency of the pattern in the ontology, the third column shows the identified pattern itself, whereas the fourth column indicates the number of repeated terms in the corpus that match the pattern. The PoS tags are indicated by the following letters: P for preposition, A for adverb, N for noun, J for adjective, V for verb, C for number, F as a foreign word, and S is the symbol ".".

The most frequent morpho-syntactic pattern is the one starting with a noun, followed by nouns and adjectives (pattern 1). This result is expected because the ontological concepts mostly will have this behavior. Actually, the number of multi-word expressions matching this pattern in the reference corpus is extremely high, thus they should be considered only as “candidate” concepts and filtered through some kind of term reduction technique.

Interestingly, some unexpected patterns such as pattern 10 of Table 3 have appeared in the AI ontology. The frequency of this pattern is 1 in the ontology, but it was not possible to find multi-word terms in the reference corpus. We suppose this problem is caused by the morphological tagger, because it does not have contextual information for correctly assigning the PoS tag to the concept. In order to avoid having incorrect morpho-syntactical patterns, this problem needs to be solved.

Table 2. Morpho-syntactical patterns identified in the OIL ontology

N	Fr	Pattern	Fr
	Ont		corpus
1	21	$N^+J?$	1,823,294
2	11	$(NV)((J?\vert(N^+)?)$	125,308
3	11	$J(N^+)?$	646,029
4	5	$A(N?\vert V?)$	223,301
5	4	$V(J\vert N)$	71,358
6	1	$P(N^+)$	192,879
		Candidate terms (without repetition):	378,465

Table 3. Morpho-syntactical patterns identified in the AI ontology

N	Fr	Pattern	Fr
	Ont		corpus
1	243	$(N^+)((VN)?\vert(V^+)?)$	2,693
2	84	$(J)^+(N^+)?$	1,000
3	35	$N((JN)\vert(PJN)\vert(C)\vert(PV)\vert(VJN)\ \vert(PN^+))$	400
4	17	$(V^+)(N^+)?$	1,582
5	10	$(AN)((PJN?)\vert(VN?)\vert(JN))$	135
6	6	$J((NPJN?)\vert(VN?))$	43
7	3	$A((VJN)\vert(JN))$	27
8	2	$(VP)(N\vert(JN))$	105
9	1	PJN	151
10	1	$FNPJN$	0
		Candidate terms (without repetition):	3,581

Table 4. Morpho-syntactical patterns identified in the SCORM ontology

N	Fr Ont	Pattern	Fr corpus
1	9,238	$(N^+)((PJ^+(N^+?))?\vert(PJ^+N^+)?\vert(JN^+)?\vert(VN^+)?\vert(J(N^+)?)?\vert(PN^+)?\vert(PV^+N^+)?\vert(PNJ)?\vert(V^+)?)$	7,422
2	500	$V((JN^+)?\vert(PJN^+)?\vert(NPJN)?\vert(PJ^+)?\vert(NPN^+)?\vert(NJN)?\vert(NPVN^+)?\vert(NPJ^+)?\vert(J^+)?\vert(PJVN)?\vert(N^+V)?\vert(PV^+J)?\vert\ (PVN)?)$	5,560
3	327	$(J^+)((N^+PN^+)?\vert(N^+(PN^+)?)?)$	2,108
4	123	$V^+(PN^+\vert N^+)$	1,105
5	84	$(J)(V(PN)?\vert NVN\vert NPJ^+(N^+)?\vert NPNPJN\vert V^+N^+\vert NPVN\vert N^+J\vert V(PN)?\vert NPJVN\vert NPNPN^+\vert N^+V)$	336
6	34	$(NP)(V\vert NPJN?\vert JNPNPN\vert JV^+\vert NVN^+\vert NVJN\vert NPN^+\vert JN^+V\vert JVN^+\vert NCSCS\ CN^+)$	230
7	17	$(N)(PV\vert VJN?\vert JN^+PJ^+N^+\vert CJN\vert VPN^+\vert CSCSCJ\vert JN^+V\vert VPVN)$	313
8	9	$(J)(JVN\vert V^+N\vert VJ)$	95
9	7	$(JV)((PN)?\vert(NJ)?)$	74
10	5	$(AJ)(VN\vert NV?)$	27
11	1	PJN	273
		Candidate terms (without repetition):	5,552

4 Evaluation of the Morpho-Syntactical Patterns

In a first phase of the evaluation of the ontological concepts, the regularity of the morpho-syntactical patterns employed by the ontological engineer (either being a human or a computer program) is considered. We evaluate the concepts using a v-fold cross validation process under the same target ontology. The explanation of this procedure together with the obtained results is given in the following section. Thereafter, we show how to validate the ontology by a second type of procedure which uses the morpho-syntactical patterns for findings concepts in a reference corpus with the aim of comparing those with the ones already stored in the ontology to be evaluated.

4.1 Intrinsic Evaluation of Ontological Patterns for Restricted Domain Ontologies

The intrinsic evaluation of ontological concepts considers only the target ontology, i.e., no other external resources such as a reference corpus is used. The procedure can be described as follows:

1. The concepts of the ontology to be evaluated are considered into two sets: training and test.
2. We identify regular morpho-syntactical patterns on the training set

3. The patterns identified are then used to validate the test set

We execute this procedure k times by changing the training and test set in a leave-one-out process, thus we execute a k-fold cross validation process [16] over the same ontology, in which $k - 1$ sets are used as a training set, and the remaining one as the test set. The average of the obtained results in each iteration is given as a final result.

In a strict sense, the morpho-syntactical patterns should be employed for matching complete strings of words, however, in some cases, a sub-string may match with a particular pattern. For example, the concept "long term planning system" is already stored in the AI ontology, however, there are some patterns which permit to determine that the string "planning system" may be also a candidate concept, even if this multiword term is not stored in the target ontology.

In Table 5 we show the obtained results when this particular method of sub-string matching is employed. Precision (P), recall (R) and F-Measure (F) evaluate the quality of the ontological concepts (OC) stored in the ontology. As can be seen, we were able to find every concept of the ontology ($R = 1$) besider a number of new concepts (candidates to be concepts (CC)) that also match with the morpho-syntactical patterns discovered. From our particular point of view, this process may be interesting, since it will allow to suggest the inclusion of new concepts to the ontology. Taking into account the total number of candidate concepts found, we obtain an F-measure greater than 0.8 in all the ontologies evaluated.

Table 5. Results obtained with the intrinsic evaluation of ontological concepts

Ontology	$\|OC\|$	$\|CC\|$	$\|OC \cap CC\|$	P	R	F
OIL	48	68	48	0.70588	1	0.82759
AI	276	346	276	0.79769	1	0.88746
SCORM	1461	1874	1461	0.77962	1	0.87616

Table 6 shows the results obtained by averaging the different v folds of cross-validation when it was applied to the different restricted domain ontologies evaluated in this research work. We have also evaluated the ontologies by taking into account the total number of candidate concepts found. The F-Measure obtained is about 70%, a value that indicates the degree of regularity when constructing the ontological concepts of the target ontologies. The obtained results show that the SCORM ontology presents a more stable construction schema than the other two ontologies evaluated, i.e., the SCORM ontology has been constructed employing a uniform set of morpho-syntactical patterns when the ontological concepts were identified by the ontological engineer.

The following research question arises when we obtain this kind of results: Are there universal morpho-syntactical linguistic structures employed by the ontological engineers when constructing restricted domain ontologies?. In order to answer this important question, we have executed an experiment in which the k-

Table 6. V-fold cross-validation for intrinsic evaluation of ontologies

Ontology	P	R	F
OIL	0.58881	0.84893	0.69103
AI	0.57950	0.92216	0.71113
SCORM	0.60909	0.96251	0.74575

fold cross-validation process has been carried out among different domains, i.e., we have used one domain as a training set and another one as the test set (k-domain cross-validation). Table 7 presents the evaluation results obtained when we employed the morpho-syntactical patterns discovered using the ontology of the first column (Ont) for discovering candidate terms using the ontologies Ont_1 and Ont_2. Thus, in the first row the OIL ontology has been used as training set, whereas the other two ontologies (AI and SCORM) were used as test set. The second row shows the results obtained when the AI ontology is used as a training set and the other two ontologies are used as the test set (Ont_1=OIL and Ont_2=SCORM). Finally, in the third row we present the evaluation values obtained when the SCORM ontology is used as the training set, whereas Ont_1=OIL and Ont_2=AI are used as a test set.

As expected, the performance of the approach is lower than the previous experiments, because in this case we are dealing with a knowledge transfer process from a domain X to a different domain Y. Again, we observed that the concepts of the OIL ontology do not have homogeneous linguistic patterns, while the SCORM ontology seems to have a set of linguistic patterns much more generic than the other two ontologies. Since the F-Measure is not greater than 70% for all the evaluations, we consider that more investigation need to be done in future work with the aim of deeply understand the manner the ontological engineers employ the morpho-syntactical patterns when they construct ontologies of restricted domain.

Table 7. v-domain cross-validation evaluation results

	Ont$_1$			Ont$_2$		
Ont	P	R	F	P	R	F
OIL	0.50795	0.69335	**0.58581**	0.43975	0.56912	**0.49594**
AI	0.53304	0.76214	**0.62455**	0.51982	0.82250	**0.63700**
SCORM	0.61797	0.96233	**0.75121**	0.57237	0.89571	**0.69452**

4.2 Extrinsic Ontological Concepts Evaluation (using a Reference Corpus)

The extrinsic evaluation is carried out when we have a corpus associated to the same domain of the target ontology to be evaluated (reference corpus). In this case, we use the ontology as a source for identifying the morpho-syntactical patterns employed by the ontological engineer when constructed the ontological concepts. These patterns are then used for extracting candidate concepts from the reference corpus and compared them against the original concepts in order to evaluate the quality of the original ontological concepts.

The methodology employed in this case consists of the following steps:

1. To apply a Part-Of-Speech (PoS) tagger to the ontological concepts of the target ontology.
2. To identify the morpho-syntactic tags for each ontological concept.
3. To apply a Part-Of-Speech (PoS) tagger to the reference corpus.
4. To extract candidate concepts in the reference corpus by using the morpho-syntactic patterns previously discovered.
5. To compare the candidate concepts against the original ontological concepts and provide a measure of the ontology quality.

Table 8 shows the results obtained with this kind of ontology validation. We observe that by applying this methodology, we are able to discover 87.5% of the OIL ontological concepts, 74.27% of the AI ontological concepts, and 59% of the SCORM ontological concepts. The proposed approach allows then to evaluate the quality of the methodology when a reference corpus is available. Of course, the better the reference corpus (amount and quality of the documents), the higher the reliability of this procedure of evaluation.

Some patterns obtaines a extremely generic, thus generating in some cases huge numbers of candidate concepts (CC). It is then important to implement a candidate concept reduction schema that allows to improve the evaluation procedure, limitating the candidate concepts to those associated to the ontology domain. Again, the new concepts discovered may be used for suggesting to the ontological engineer their inclusion into the ontology.

Table 8. Extrinsic evaluation of ontological concepts

Ontology	$\|C\|$	$\|CC\|$	Enc	P
OIL	48	364,033	42	0.87500
AI	276	3,581	205	0.74275
SCORM	1,461	5,552	864	0.59138

5 Conclusions

In this research work we have presented an approach based on morpho-syntactical patterns for evaluating ontological concepts stored in ontologies of restricted domain. Intrinsic and extrinsic evaluation procedures were carried out, showing that it is possible to evaluate the quality of the ontology when a reference corpus is available, but also when this collection of documents is not available.

By evaluating the target ontologies employing the intrinsic procedure we were able to analyze the regularity of the morpho-syntactical patterns employed by the ontological engineers when constructing the ontological concepts. According to the results obtained, the SCORM ontology is the one that has been constructed with more care, despite of the high number of ontological concepts that this ontology has, because the linguistic patterns discovered in this ontology are homogeneous and less specific than the ones used in the other two ontologies.

On the other hand, when we employed the extrinsic procedure, we were able to validate the original ontological concepts by using candidate concepts discovered in a reference corpus. The approach proposed obtained a minimum of 59% of accuracy and a maximum of 87.5%, whereas the recall was 100% in all the cases.

Even if the purpose of the method proposed is to validate ontological concepts stored in ontologies of restricted domain, it is also possible to use it for suggesting or recommending the inclusion of new concepts in the target ontology which we consider an important contribution of the experiment carried out.

One of the methods proposed for discrimination of candidate terms, as future work, is the identification of taxonomic relationships and not taxonomic relationships associated with these terms in the corpus.

The resuts has been applied to suport the semiautomatic build of ontological conceptual model. The next step is to integrate this approach in order to obtain relevant term extraction on specific domain.

References

1. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Guarino, N., Poli, R., eds.: Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, Kluwer Academic Publishers (1993)
2. Brank, J., Grobelnik, M., Mladenić, D.: A survey of ontology evaluation techniques. In: Proc. of 8th Int. multi-conf. Information Society. (2005) 166–169
3. Pak, J., Zhou, L.: A framework for ontology evaluation. In Sharman, R., Rao, H.R., Raghu, T.S., eds.: WEB. Volume 52 of Lecture Notes in Business Information Processing., Springer (2009) 10–18
4. Gómez-Pérez, A.: Ontology Evaluation. International Handbooks on Information Systems. Springer (2004)
5. Cantador, I., Ferández, M., Castells, P.: A collaborative recommendation framework for ontology evaluation and reuse. In: Actas de International Workshop on Recommender Systems, en la 17th European Conference on Artificial Intelligence (ECAI 2006), Riva del Garda, Italia. (2006) 67–71
6. Salem, S., AbdelRahman, S.: A multiple-domain ontology builder. In: Proceedings of the 23rd International Conference on Computational Linguistics. COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 967–975
7. Reyes Ortiz, J.A.: Creación automática de Ontologías a partir de Textos con un Enfoque Lingüístico. PhD thesis, Dept Ciencias Computacionales, Cenidet, Cuernavaca, Mor., Mex. (2013)

8. Sabou, M., Lopez, V., Motta, E., Uren, V.: Ontology selection: Ontology evaluation on the real semantic web. In: Proceedings The 4th International EON Workshop, Evaluation of Ontologies for the Web. (2006)
9. Zavitsanos, E., Paliouras, G., Vouros, G.A.: Gold standard evaluation of ontology learning methods through ontology transformation and alignment. IEEE Trans. Knowl. Data Eng. **23**(11) (2011) 1635–1648
10. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: Proceedings of European Knoeledge Ackquisition Workshop (EKAW). (2002)
11. Dellschaft, K., Staab, S.: Strategies for the evaluation of ontology learning. In Buitelaar, P., Cimiano, P., eds.: Bridging the Gap between Text and Knowledge Selected Contributions to Ontology Learning and Population from Text, Amstedam, IOS Press (2008)
12. Spyns, P., Reinberger, M.L.: Lexically evaluating ontology triples generated automatically from texts. In Gómez-Pérez, A., Euzenat, J., eds.: ESWC. Volume 3532 of Lecture Notes in Computer Science., Springer (2005) 563–577
13. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of International Conference on Language Resources and Evaluation. (2004)
14. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M., eds.: WOP. Volume 929 of CEUR Workshop Proceedings., CEUR-WS.org (2012)
15. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, Manchester, UK (1994) 44–49
16. Murty, M., Devi, V.: Pattern Recognition: An Algorithmic Approach. Undergraduate topics in computer science. Springer London, Limited (2011)

Knowledge-based Workflow Ontology for Group Organizational Structure

Mario Anzures-García[1], Luz A. Sánchez-Gálvez [1], Miguel J. Hornos[2], and Patricia Paderewski-Rodríguez[2]

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Avenida San Claudio y 14 Sur, Ciudad Universitaria. 72570 Puebla, México

[2] Departamento de Lenguajes y Sistemas Informáticos, E.T.S.I. Informática y de Telecomunicación, Universidad de Granada, C/ Periodista Saucedo Aranda, s/n, 18071 Granada, Spain.

{mario.anzures, sanchez.galvez}@correo.buap.mx; {mhornos,patricia}@ugr.es

Abstract. The workflow model represents the set of steps performed by different entities and their execution ordering to control and manage the carried out process on organization. In which, the structure organizational determines the division of labor in order to persons perform the organization process in an appropriated manner. In collaborative applications, this structure determines how the interaction among users, as well as between the users and the application, are carried out. For this reason, a workflow to manage the group organizational structure will be ideal. Although, workflow lacks the expressive power to represent the domain knowledge and the sequence of operations. An ontology describes the knowledge domain through concepts, relations, axioms and instances, but ontology does not specify how these entities should be used and combined. Therefore, in this paper a workflow ontology to control the group organizational structure is proposed. A case study is presented to show the use of knowledge-based workflow ontology for this structure.

Keywords. Workflow Model, Group Organizational Structure, Ontology, Knowledge Base, Workflow Ontology

1 Introduction

Nowadays, knowledge must be processed computationally to be used not only as individuals but to groups. This leads to require a knowledge base to represent the problem domain. This base can aid to understand, manage and control every performed process by the organizations. In the which, the group work is determined by an organizational structure that is governed by a set of rules that establish its configuration. On the one hand, in the collaborative applications, this structure determines how the communication, collaboration, and coordination among the group members are performed. On the other hand, this configuration can change in accordance with tasks and group needs at a given moment. It is very important that the structure can adapt dynamically to cope with changing organization and own application conditions. For this reason, this structure is modeled by an ontology, since, it can adjust for the changes within the group and to the different working styles of several groups. Moreover, it is one of the strategies for the knowledge structured representation in a formal way, helping to remove ambiguity and redundancy, detecting errors and allowing automated reasoning [1, 2].

In order to manage the organizational structure, it is necessary to specify how the entities should be used and combined, which the ontology does not make. Consequently, a workflow can be used to this, because it refers to coordinated execution of multiple tasks or activities [3, 4]. Nevertheless, it lacks of necessary expressive to knowledge representation. So a solution is a workflow ontology, which supplies a formal knowledge representation in order to specify the elements and control the steps ordered set of the organizational structure.

Therefore, in this paper, a knowledge-based workflow ontology along with a set of rules is proposed to manage the Group Organizational Structure. In such a way, the knowledge about this structure and special workflow, are formal, and explicitly modeled. Using this knowledge representation scheme and rules, the application can adapt to frequent changes in organizational structure, rules and procedures.

The rest of the paper is organized as follows. Section 2 describes briefly the schemas of knowledge representation. Section 3 explains the ontologies. Section 4 presents the workflow model, and workflow ontology. Section 5 details the workflow ontology of the group organizational structure, and conceptual proof in according to a case study focused on academic virtual space. Section 6 summaries the conceptual results obtained. Section 7 outlines the conclusions and future work.

2 Knowledge Base

Given that knowledge is a portion of all human activities, it is necessary to store it —seizing its meaning— organize it and make it available. So, it requires a representation scheme to provide a set of procedures, which allows the knowledge, to be stored, organized, and to represent the problem naturally. This leads to require a knowledge base to represent the problem domain, as well as can reason and draw conclusions through an inference mechanism for the contents of the knowledge base [5]. The representation scheme must be denoted by a model of some domain of interest in which symbols assist as substitutes for real world artifacts. These symbols must be stored as interest domain statements. The knowledge representation schemes are [6]:

- *Semantic Network* is appropriate for capturing the taxonomic structure of categories for domain objects and for expressing general statements about the domain of interest. Nevertheless, the representation of concrete individuals or even data values does not fit well the idea of semantic networks.
- *Frames* represent a concept, consisting of slots for which fillers are specified. The reasoning in frame-based systems involves both intentional and

extensional knowledge contained in the knowledge base of the frame. However, the frames provide more expressive power but less capacity to infer.

- *Rules* come in the form of IF-THEN-constructs and allow to express various kinds of complex statements. Rule-based knowledge representation systems are especially suitable for reasoning about concrete instance data. Complex sets of rules can efficiently derive implicit such facts from explicitly given ones. They are problematic if more complex and general statements about the domain shall be derived, which do not fit a rule's head [6].
- *Logic* is the dominant form of knowledge representation, since is used to provide a precise formalization and axiomatization of problem domain, which is ideal for representing and processing knowledge within computers in a meaningful way. Nowadays, all symbolic knowledge representation and reasoning formalisms can be understood in their relation to First-order (predicate) logic, therefore, this is the prevalent and single most important knowledge representation and reasoning formalism. First-order logic allows one to describe the domain of interest as consisting of objects, i.e. things that have individual identity, and to construct logical formulas around these objects formed by predicates, functions, variables and logical connectives [7]. Description logic [8] is essentially a set of decidable fragments of first-order logic and is expressive enough such that it has become a major knowledge representation and reasoning paradigm. A description logic theory consists of statements about concepts, individuals, and their relations. Individuals correspond to constants in first-order logic, and concepts correspond to unary predicates. Concepts can be named concepts or anonymous (composite) concepts. Named concepts consist simply of a name, which will be mapped to a unary predicate in first-order logic. Composite concepts are formed from named concepts by using concept constructors, similar to the formation of complex formulas out of atomic formulas in first-order logic [9].

The ontology is an ideal solution to represent the knowledge domain using description logic symbols, which allow to specify it of a simple way, and readable for both human and machines; as well as perform much deeper reasoning through the machine. It facilitates a knowledge base in order to provide semantic, common understanding, communication and knowledge sharing on the domain of interest and a knowledge reasoning, carrying out an inference process to reach conclusions on the knowledge base by means on a reasoner, inference rules and query languages.

3 Ontologies

Ontology, according to Gruber, *is a formal and explicit specification of a shared conceptualization* [9]. *Conceptualization* refers to an abstract model of some phenomenon in the world by identifying the relevant concepts of this. *Explicit specification* means that the type of concepts used, and the constraints on their use are explicitly defined. Thus the ontology is a high level formal specification of a certain knowledge domain, providing a simplified and well defined view of domain.

3.1 Ontology Structure

The specification of the ontology is defined through of the following components [9]:

- *Classes*: Set of classes (or concepts) that belong to the ontology. They may contain individuals (or instances), other classes, or a combination of both with their correspondents attributes.
- *Relations*: These define interrelations between two or several classes (object properties) or a concept to a data type (data type properties).
- *Axioms*: These are used to impose constraints on the values of classes or instances. Axioms represent expressions in ontology (logical statement) and are always true if used inside the ontology.
- *Instances*: These represent the objects, elements or individuals of an ontology.

Nowadays, the ontologies (particularly in OWL —Ontology Web Language) have been extended with rules by Semantic Web Rule Language (SWRL), which use other predicates than just class or property names:

- *class expressions:* These are arbitrary class expressions, not just named classes.
- *property expressions:* The only operator available in OWL 2 for creating property expressions is inverse of object property; however, the same effect can be achieved by exchanging the property arguments, so there is no need to use property expressions in SWRL
- *data range restrictions*: They specify a type of data value, like integer, date, union of some XML Schema types, enumerated type.
- *sameIndividual and differentIndividuals*: These are used for specifying same and different individuals
- *core SWRL built-ins:* They are special predicates defined in SWRL proposal which can manipulate data values, for example, to add numbers custom SWRL built-ins —it can define own built-ins using Java code.

3.2 Ontology Languages

Like the knowledge representation and reasoning, ontologies require a logical and formal language to be expressed. In the area of Artificial Intelligence many languages have been developed for this purpose, some based on First-order (predicate) logic as KIF and Cycl providing modeling primitives and the possibility of redoing formulas that enable them to become in terms of other formulas. Other Frames-based languages with more expressive power but less inference capability as Ontolingua and F-Logic; others based on descriptive logic that are more robust in the power of reasoning as a Loom, OIL, DAML + OIL and OWL. OWL [10, 11] is an ontology language recommended by the W3C for use in the Semantic Web. The OWL representational facilities main are directly based on Description Logics. This basis confers upon OWL a logical framework, including both syntax and model-theoretic semantics, allowing it is a knowledge representation language capable of supporting a knowledge base and a practical reasoning and effective. Moreover, the Description Logic provides readily available reasoners such as Pellet [12] and HermiT [13], both of which have been extended to handle all of OWL. OWL ontologies can also be combined with rules using the new W3C Rule Interchange Format (RIF) standard [14]. For the development of ontologies are used tools, which provide graphical interfaces that facilitate the knowledge representation and reasoning. This article focuses on Protégé, which is an engineering tool open source ontology and a knowledge-based framework. Protégé is widely used for the development of ontologies, due to the scalability and extensibility with lots of plugins; and by facilitate inference knowledge through reasoners, query languages and rules. Ontologies in Protégé can be developed in a variety of formats, including OWL, RDF (S), and XML Schema.

Summarizing, ontology establishes the vocabulary used to describe and represent knowledge, and to facilitate machine reasoning.

4 Workflow Model

Workflow is seen as an automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another, for action, according to a set of procedural rules [15, 16]. The workflow management, commonly, is achieved through three constructs: routes (it represents task sequences), rules (it defines routing and role constructs), and roles (it represents one who is responsible for a task). A suitable management of a workflow requires the following aspects [17]:

- *Expressiveness:* It should provide constructs to represent conditional mapping relationships between roles and actors based on the organizational model as well as complex business rules, including exceptional rules.
- *Model verification:* It should allow analysis that assures the correctness of workflow specification, together with checking the occurrence of inconsistent, redundant, and incomplete rules as well as non-terminality of processes.
- *Change management:* It should allow easy development of the propagation mechanisms against changes on the organizational structure and rules as well as organizational procedures to assure the correctness of a workflow model.

The ontology provides enough expressivity by the supplied structure (concepts, relations, instances, and axioms); the ontology verification is accomplished through reasoners (Pellet, and HermiT); and the change management on the ontology can be made modifying, adding, and/or deleting concepts, relations, and/or instances. Consequently, the ontology is ideal to workflow management. So, recently, it has paid special attention to the development of workflow ontologies. It presents a collaborative workflow for terminology extraction and collaborative modeling of formal ontologies using two tools Protege and OntoLancs [18]; it allows the development of ontology cooperatives and distributed based on dependencies management between ontologies modules [19]; it shows an ontology-based workflows for ontology collaborative development in Protégé [20], it presents the combination of workflows with ontologies to design way formal protocols for laboratories [21], and it proposes a workflow ontology for the preservation digital material produced by an organization or a file system [22].

All these works focused on building workflow ontologies to represent collaborative work in different areas, however, this paper presents a workflow ontology to manage the group structure organizational in the collaborative domain.

5 Workflow Ontology for Group Organizational Structure

In the collaborative applications, the shared work is supported for sessions, which denote the shared workspace. This type of applications typically provides a shared workspace by a session manager. On the one hand, this manager allows to establish the session (i.e., it permits to set up the connection, to create and manage meetings, and to enable a user to join and leave a session using a simple user interface). On the other hand, this manager allows defining the group organizational structure that states how sessions are organized to accomplish the shared work. In general, collaborative applications do not separate the mechanisms to establish the shared workspace from the group organizational structure. Therefore, in this paper, it is considered the proposed separation in [1]; because it allows us to support changes in the group at runtime and specify this organizational structure through a policy.

This structure can be hierarchical (one member has a greater status than other members, such as in meetings with department chief) or not hierarchical (the participants have equal status, for example, informal meetings of university professors). These structures are ruled by a policy, which determines how the group members will be organized. In groupware, this policy is called session management policy, which establishes the group organizational structure in terms of the functions that group members will carry out. This policy has been modeled by ontology [1, 2], adjusting to group dynamic nature and evolving needs of the same.

However, it is required the organizational structure management, thus, a workflow ontology of the group organizational structure is proposed.

5.1 Workflow Ontology Description

This ontology (see Fig. 1) defines that: the group organizational structure (GOS) is made up of users, and is governed by one policy (Pcy), which establishes a hierarchical organizational structure or not-hierarchical organizational structure by means of the roles (one or more —Rls) that users can play. Each role designing one status (Stt —which founds the role priority in the group), one right/obligation (R/O — set privileges for the user in the application), and a tasks set (Tsk—which are role functions) and they can be composed of one or more activities (Atv —which are operations that allow users to achieve a given goal) that use resources (Rsc). For each task indicates the event (Evt) that triggers it, its precedence (PTk —i.e., tasks order), and its type (*Sequential-task —SqT;* one activity follows the other. *Parallel-Task —*

PrT; these happen at the same time, but they use different objects, and no interference between them can occur. *Partially-Concurrent-Task —PCT;* it refers to tasks that can be active at the same time but there is no simultaneous modification of any object). *Fully-Concurrent-Task —FCT;* it occurs when two or more simultaneous tasks to modify rights to same set of objects). It establishes the application stages (Stg —it reflects each of the collaboration moments). For each stage determines the order of them (Stage Precedence —SPc), the tasks that correspond to these, and precedence of the tasks (STk) in the same.

The specification of the Workflow Ontology is carried out through of the following steps (WOS):

1. Starting Workflow (StW)
2. Defining the GOS name.
3. Determining the Policy name.
4. Establishing the Roles of the groupware.
   - 4.1. Designing a Status to role.
   - 4.2. Signalizing a Right/Obligation to role.
   - 4.3. Specifying the tasks that each role carries out in the groupware.
     - 4.3.1. Designating the event that triggers each Task.
     - 4.3.2. Indicating the task type (sequential, parallel, partially concurrent, and fully concurrent).
     - 4.3.3. Mark out the Activities of each Task.
       - 4.3.3.1. Defining the resources to the activity.
       - 4.3.3.2. If there are more resources go to step 4.3.3.1, else go to step 4.3.3.
     - 4.3.4. If there are more activities of one task go to step 4.3.3, else go to step 4.3.
   - 4.4. If there are more tasks for one role go to step 4.3, else go to step 4.
5. If there are more roles for the application go to step 4, else go to step 6.
6. Establishing the Stages of the collaborative application.
     - 6.1.1. Determining the order of the stage.
     - 6.1.2. Assigning tasks to a stage.
     - 6.1.3. Indicating the tasks' precedence in each stage.

5.2 Proof Conceptual of the Workflow Ontology

The study case consists in the development of a groupware for Managing Departmental Test (MDT) of the *Facultad de Ciencias de la Computación de la Universidad Autónoma de Puebla*. The Departmental Test (DET) homogenizes the teaching of a subject, i.e. it guarantees that all teachers encompass the same percentage of the academic program. For this reason, it requires a shared workspace that allows professors to manage and apply a DET.

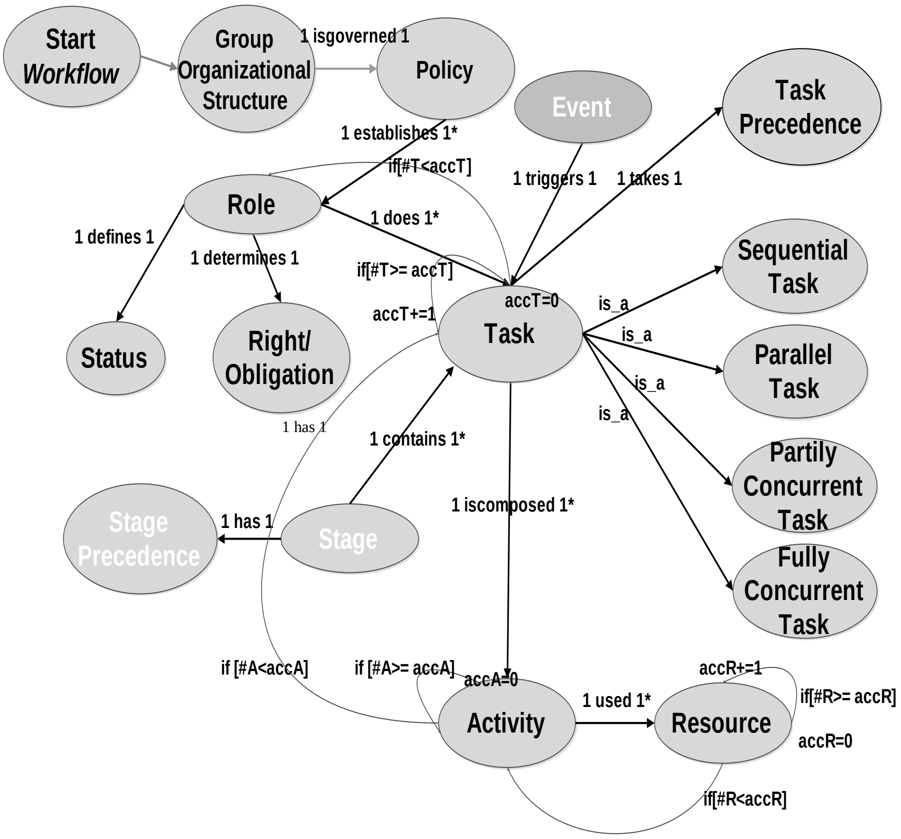

Fig. 1. Workflow ontology of the group organizational structure

For reasons of space, the workflow ontology, that displays the knowledge representation in a conceptual, and formal manner; and a Table, that shows the workflow ontology elements, will be presented in a partial form.

Several roles are considered in MDT: The Manager (Mgr) who configures the application (CfA) and has status equal to 1, so, he/she registers the users, who play the other four roles, the knowledge areas, and the subjects that are a part of them. The Area Coordinator (ArC) with status 2, who manages the test (MaT), so, he/she registers the TeC and schedules the professors' meetings, related with the same subject. The Test Coordinator (TeC) with status 3, who organizes the test (OgT), so, he/she put in order the completion of each test, requesting and agreeing the number of tests to be applied, as well as on the dates and the number of questions, which will be included; then he/she will post the test and the classroom, where each Professor (Pfs) will apply it. The Pfs with status 4, who generates the test (GeT), thus, he/she will propose and vote the date when the test will be performed, so as the number of questions contained in the exam. The Students (Sds) with status 5, who views scores (ViS) of the test, so, he/she will look up the information about the date and classroom, where the test will be carried out, as well as to find the grades obtained for each subject. In general, the five roles must register to join at the session.

The collaborative application for managing the MDT has four stages (Stg):

1) Test Configuration (TCf) with stage precedence (SPc) equal to 1. In this stage only the role Mgr participates; executing the tasks of: Authenticate (Aut) him/herself (which is triggered by the Event accesses to the system), Create, Read, Update, and Delete (CRUD) for ArC (which are activated by the Event manages to ArC), Area (which is initiated by the Event manages to Area), and Subject (which is originated by the Event manages to Subject).
2) Test Preparation (TeP) with SPc equal to 2. In this stage, the roles Mgr and TeC enter to the same; the former performing the tasks of Aut, CRUD TeC, Proposing Meeting Date, and Setting Date (StD) with the activities of writing date (Wrd —that used the resources label, and calendar) and sending date (sed), these two tasks are triggered by the Event schedule meeting (ScM); the latter executing the tasks of Aut, and CRUD Pfs. The tasks Aut, Crud TeC and Pfs are triggered by the same Events of similar tasks corresponding to the role Mgr.
3) Test Elaborating (TeE) with SPc equal to 3, and two roles joining: 1) The role TcC carries out the tasks of: Aut him/herself with PTk equal to 1, Proposing Date (PD) with PTk equal to 2, Setting Date (SD) with PTk equal to 3, Proposing Number of Questions (PNQ) with PTk equal to 4, Setting Number of Questions (SNQ) with PTk equal to 5, and Posting Questions (PQ) with PTk equal to 6. The tasks 2 and 3 are activated by Event called defining the test date (DTD), while the 4, 5, and 6 by Event establishes test questions. On the other hand, the task 3 is composed of the activities: selecting date (ed), confirming date (CD), and submit date (uD). The first use the resource calendar; the second use the resource confirmation bottom, and thee third the acceptation bottom (AB). 2) The role Pfs executes the tasks: Aut him/herself, Consulting Proposals, Choosing Date, Loading Proposal of questions, Downloading Exercises of the test, Choosing Questions of the test, Posting Notice, and Posting Message.
4) Test Results (TeR) with SPc equal to 4. Four roles (ArC, TeC, Pfs, and Stu) participate in this Stg. The role ArC implements the tasks of Downloading Scores, Generating Reports, Loading Statistics, Posting Messages, Posting Notices, and Scheduling Test. The role TeC to performing the same tasks that the role ArC; in addition, he/she carries out the tasks of Loading DET, and Posting classroom. The role Pfs effects the tasks of: Loading Scores, Download Scores, Register in Group, Loading Scores, Posting Messages, Posting Notices, and Scheduling Test. The role Stu carries out the task of Download Scores, Posting Messages, and Scheduling Test.

The Table 1 shows the workflow ontology elements that constitute the knowledge base and that are gotten of the case study with respect to the Stage TeE and the role TcC. This table is expressed in terms of the ontology specification, as well as, of the rules that determine the execution flow of the steps to be performed by this workflow.

Table 1. Workflow ontology specification

WOS	Cpt	CoI	CAI	Relation	Cdy	TaC	TCI	TCA	TAI	Rule
1	STW									
2	GOS	GOS-MDT	1	isgoverned	1	Pcy	PMDT	NPcy	1	if [[GOS](?X) and [Pcy](?Y)] (?X,?Y)] then [isgoverned Pcy] (?X,?Y)
3	Pcy	PMDT	1	establishes	1*	Rls	Mgr, ArC, TcC, Pfs, Sds	#R, accR	5, 0	If [[Pcy](?X) and [Rls](?Y)] (?X,?Y)] then if [accR<=#R] then establishes Rls and accR+=1 else if [[Stg](?Z) and [Tsk](?W)](?Z,?W)] then contains Tsk] (?Z,?W)] (?X,?Y)
4	Rls	TcC	1	defines	1	Stt	3			if [[Rls](?X) and [Stt](?Y)] (?X,?Y)] then defines Stt] (?X,?Y) and con+=1;
4.1	Rls	TcC	1	determines	1	R/O	OgT			if [[Rls](?X) and [R/O](?Y)] (?X,?Y)] then determines R/O] (?X,?Y)
4.2	Rls	TcC	1	does	1*	Tsk	Aut, PD, SD, PNQ, SNQ, PQ	#T, accT	5, 0	if [[Rls](?X) and [Tsk](?Y)] (?X,?Y)] then if [[accT<=#T] then does Tsk and accT+=1; else if [[Pcy](?W) and [Rls](?Z)] (?W,?Z)] then establishes Rls] (?X,?Y).
4.2.1	Evt	ScM	1	triggers	1	Tsk	DTD			if [[Evt](?X) and [Tsk](?Y)] (?X,?Y)]then [trigger Tsk] (?X,?Y)
4.2.2	Tsk	SD	1	is_composed	1*	Atv	eD, CD, uD	#Atv, accA	3, 0	if [[Tsk](?X) and [Atv](?Y)] (?X,?Y)] then [is_composed Atv] (?X,?Y)
4.2.2	Atv	uD	1	uses	1*	Rsc	AB	#A, accC	1, 0	if [[Act](?X) and [Rsc](?Y)](?X,?Y)] then [uses Rsc and accA+=1] (?X,?Y); if [[Rsc](?Z) and [accC>=Rsc]](?Z) then [conA*=1 and goes [[Act](?X) and [Rsc](?Y)](?X,?Y)]
4.2.3	Tsk	SD	1	takes	1	PTk	3			if [[Tsk](?X) and [PTk](?Y)](?X,?Y)] then [takes PTk] (?X,?Y)
5	Stg	TeE	1	contains	1*	Tsk	Tcf, TeE, TcC, TEr	#S, accS	4,0	if [[Stg](?X) and [Tsk](?Y)](?X,?Y)] then contains Tsk] (?X,?Y)
6	Stg	TeE	1	supports	1	SPc	3			f [[Stg](?X) and [SPc](?Y)](?X,?Y)] then supports SPc] (?X,?Y)

Therefore, this table presents the following columns; allowing us to proof the ontology: Concepts (Cpt), Concept Instance (CoI), Concept Attribute (CAt), Concept Attribute Instance (CAI), Relation (Rel), Cardinality (Cdy), Target Concept (TaC), Target Concept Instance (TCI), Target Concept Attribute (TCA), Target concept Attribute Instance (TAI), and Rules (Rul).

6 Conceptual Results

Summarizing, the workflow ontology allows us to know: the roles participate in the interaction (fully concurrent, partially concurrent, parallel, and/or sequential) and in what order do; the resources used by each user for accomplishment each activity. This is possible, thanks that this ontology establishes:

1) The roles that access to each stage.
2) The priority of each stage.
3) The task executed in each stage and its priority on it.
4) The task carried out by each role.
5) The activities that compose each task.
6) The resources used in each activity.
7) Who performs each type task (SqT, PrT, PCT, and FCT).

7 Conclusions and Future Work

This paper has presented a workflow ontology to manage the group organizational structure. On the one hand, this ontology is created of the knowledge base (which aid to understand, manage and control every performed process by the organizations) provides by the session management policy ontolog;, allowing us to adjust for the changes within the group and to the different working styles of several groups, and helping to remove ambiguity and redundancy. On the other hand, a workflow is specified through this ontology in order to model how the organizational structure entities should be used and combined, as well as; it should be controlled the steps ordered set of the organizational structure.

The future work is orientated to specify a methodology to develop groupware, starting with workflow ontology described in this article.

References

1. Anzures-García, M.; and Sánchez-Gálvez L.A.: Policy-based group organizational structure management using an ontological approach. In Proc. International Conference on Availability, Reliability and Security (ARES), pp. 807-812, (2008)
2. Anzures-García, M.; Sánchez-Gálvez, L.A.; Hornos, M.J.; and Paderewski-Rodríguez, P.: Ontology-Based Modelling of Session Management Policies for Groupware Applications. Lecture Notes in Computer Science, Vol. 4739, pp. 57–64, Springer, Heidelberg, (2007)

3. Fischer, L.: Workflow Handbook. Future Strategeis Inc., Lighthouse Point, FL, (2004)
4. Marinescu, D.: Internet-Based Workflow Management: Toward a Semantic Web. Wiley, New York, (2002)
5. Anzures-García, M.; Sánchez-Gálvez, L.A.; Hornos, M.J.; and Paderewski-Rodríguez, P. A.: Knowledge Base for the Development of Collaborative Applications. Engineering Letters, vol. 23, no.2, pp. 65-71, (2015)
6. Grimm, S.; Hitzler, P.; and Abecker, A.: Knowledge Representation and Ontologies. In: Semantic Web Services: Concepts. Technologies and Applications, Springer-Verlag. pp. 51-105. (2007)
7. Russel, S.; and P. Norvig.: Artificial Intelligence – A Modern Approach. Prentice-Hall, 1995
8. Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D. and Patel-Schneider, P. F.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, (2003)
9. Gruber, R.: A translation approach to portable ontology specification. Knowledge Acquisition. Vol. 5, pp. 199-220, (1993)
10. Patel-Schneider, P.F.; Hayes, P.; and Horrocks, I.: OWL Web Ontology Language semantics and abstract syntax. W3C Recommendation, 10 February 2004
11. Horrocks, I.; Patel-Schneider, P.F.; and van Harmelen, F.: From SHIQ and RDF to OWL: The making of a web ontology language. J. of Web Semantics, Vol. 1-1, pp. 7-26, (2003)
12. Pellet: OWL reasoner. Maryland Information and Network Dynamics Lab, 2003. http://www.mindswap.org/2003/pellet/index.shtml
13. Motik, B.; Shearer, R.; and Horrocks, I.: Optimized reasoning in description logics using hypertableaux. In Proc. of the 21st Int. Conf. on Automated Deduction (CADE-21), volume 4603 of Lecture Notes in Artificial Intelligence, Springer pp. 67-83, (2007)
14. RIF RDF and OWL Compatibility. W3C Recommendation, 22 June 2010. Available at http://www.w3.org/TR/rif-rdf-owl/
15. Allen, R.: Workflow: An introduction. In Workflow Handbook, L. Fisher, Ed. Future Strategies, Lighthouse Point, FL, pp. 15–38, (2001)
16. Marshak, R.T.: An overview of workflow structure. Proc. of Workflow, pp. 13-42. San Jose, USA: Future Strategies Inc. (1994)
17. Lee, H.B.; Kim. J.W.; and Park, S.J.: KWM: Knowledge-based Workflow Model for Agile Organization. Journal of Intelligent Information Systems Vol. 13 (3), pp. 261-278. November (1999)
18. Gacitua, R.; Arguello-Casteleiro, M.; Sawyer, P.; Des, J.; Perez, R.; Fernandez-Prieto, M.J.; and Paniagua, H.: A collaborative workflow for building ontologies: A case study in the biomedical field. *Research Challenges in Information Science*, 2009. RCIS 2009. Third International Conference on, vol., no., pp.121,128, 22-24 April 2009
19. Kozaki, K.; Sunagawa, E.; Kitamura, Y.; and Mizoguchi, R.: A Framework for Cooperative Ontology Construction Based on Dependency Management of Modules. ESOE, Vol. 292 of CEUR Workshop Proceedings, pp. 33-44. CEUR-WS.org,(2007)
20. Sebastian, A.; Noy, N.F.; Tudorache, T.; and Musen, M.A.: A Generic Ontology for Collaborative Ontology-Development Workflows. Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns, (2008)
21. Maccagnan, A.; Riva, M.; Feltrin, E.; Simionati, B.; Vardanega, T.; Valle, G.; and Cannata, N.: Combining ontologies and workflows to design formal protocols for biological laboratorios, Automated Experimentation, Vol. 2-3, (2010)
22. Mikelakis, M.; and Papatheodorou, C.: An ontology-based model for preservation workflows. In Proceedings of the 9th International Conference on Digital Preservation, (2012)

A Methodology for Location-Allocation Problem

María Beatriz Bernábe Loranca[1], Rogelio González Velázquez[1], Mario Bustillo Díaz[1], Martín Estrada Analco[1], Jorge Cerezo Sánchez[2], Griselda Saldaña González[2]
Abraham Sánchez López[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación

[2] Universidad Tecnológica de Puebla
Puebla, Puebla. México
beatriz.bernabe@gmail.com

Abstract. In this work we propose a methodology that allows establishing the relationships between the location of the facilities and the clients' allocation with a dense demand. The use of this application lets us know the optimal location of production facilities, warehouses or distribution centers in a geographical space. We solve as well, the customers' dense demand for goods or services; this is, finding the proper location of the facilities in a populated geographic territory, where the population has a demand for services in a constant basis. Finding the location means obtaining the decimal geographical coordinates where the facility should be located, such that the transportation of products or services costs the least. The implications and practical benefits of the results of this work have allowed an enterprise to design an efficient logistics plan in benefit of its supply chain. Firstly, the territory must be partitioned by a heuristic method, due do nature combinatory of the partitioning. After this process, the best partition is selected with the application of factorial experiment design and the surface response methodology. Once the territory has been partitioned into k zones, where the center of each zone is the distribution center, we apply the continuous dense demand function by solving the location-allocation for an area where the population has a dense demand for services.

Keywords. Dense Demand, Location-Allocation, Methodology, Response Surface

1 Introduction

In a broad sense, it is understood that logistics is found inside the supply chain, and, in general the logistical networks in the supply chain are a system that manage the merchandise network and physical flow among the members of the supply chain influenced by the territorial distribution and by the transportation systems to reduce the logistical expenses and coordinate the production-distribution activities. On this point, the supply chain can be defined as the set of enterprises that comprise providers, manufacturers, distributors and sellers (wholesale or retail) efficiently coordinated by means of collaborative relationships between their key procedures to place the inputs or products requirements in each link of the chain at the right time and at the lowest cost, looking for the biggest impact on the value chains of the members with the end of satisfying the final consumers' requirements. However, in this work we focus on the service location-allocation aspect. This way, the goal of this work is presenting a methodology to support the strategic decision-making process of an enterprise, primarily to locate the facilities for the planning of a logistical network. We have made a special emphasis on the case of georeferenced zones and dense demands, which is a territorial design problem combined with a location-allocation problem.

The methodology is based on finding the proper location of the facilities in a populated geographic territory, where the population continuously demands services. Obtaining the locations means finding the longitude, latitude coordinates of each location point in such a way that the transportation of products and services has a minimum cost.

The problem implies solving the territorial design partition. This partition is obtained with a methodology that begins with the selection of the method of partitioning according to the results generated by the experiment factor and response surface (design of experiments statistical). Then the minimizing the Weber function that has a demand function multiplied by the Euclidean distances as weights is done. The demand function represents the population's demand in every territory, whereas the Euclidean distance is calculated between the potential location points and the demand points.

2 Statistical Design Methodology for Partitioning

The methodology proposed suggests partitioning the territory with a territorial design method that generates compact groups or clusters [10], which has to be obtained first to be able to solve the Location-Allocation Problem (LAP) for a Territorial Design Problem (TDP) with dense demand. The Distribution Centers (DC) allocated will provide services to a group of communities that are found in every geographic area, and each of them is represented by its centroid. The location should be the one that minimizes the travelling expenses by finding the geographical coordinates of the center of the centroids. The populations from these communities represent the potential clients of the DC, and the demand is modeled with continuous demand functions with two variables based on the population density of every group [5].

Due to the numeric nature of the solutions obtained, this problem addresses a continuous case of the LAP. Additionally with the mathematical approach associated, we use a geographical information system (GIS) to create maps of the territories designed [13].

There are many efforts to solve problems related with the location-allocation of services, and the state of the art for it deserves a work of its own [1, 2 ,3, 4, 5 ,6, 7, 8]. However, we can say that our contribution focuses on presenting a methodology that begins with a territory partitioning process that employs a P-median method and partitioning restrictions to find p-centers by incorporating a metaheuristic [9]. Due to the fact that the population's demand for services is dense, once the distribution centers have been determined, their geographical location (x = longitude, y = latitude) in R^2

must be found such that the services or products transportation has a minimum cost, using the Weber function as stated in the previous section.
For example, let’s assume that we wish to know where to locate a healthcare center (assuming as well that all the associated conditions are met). Then for this particular application we could locate clinics in the centroids of every geographic unit and in the center-most point of all the communities, locating a general hospital as DC such that transferring patients requires a minimum amount of time.

First we have obtained the territory partition, where the cluster formation is based on geometric compactness of territorial design and the minimum distances between centroids [10], i.e, the first part of the methodology consists of the selection of the partition and the heuristic method to continue the statistical experiment (Design of experiments) which will indicate the number of suitable partitions. Design of experiments allows to analyze data using statistical models to observe the interaction between the independent variables and as affect the dependent variable. This methodology is based on experimentation. At the time of these experiments is obtain replicas and randomize data. Using replicas we have an estimate of experimental error, if higher the number of replicas is, the experimental error is lower. This means that the experiments are given in the same conditions. Randomization during the experiment is essential to avoid the dependence between samples and ensure that results are actually caused by the dependent variables and not by the experimenter.

The first part of the methodology is showed. (Selection of the partition using design of experiments):

1. Select the partition method
2. Select the candidates of heuristic methods
2.1 Develop a design of experiments to select the best heuristic method (the one that reach the best cost function).
2.1.1 Estimate effects of the factors (heuristics parameters)
2.1.2. Form an initial model
2.1.3. Develop testing statistics
2.1.4. Redesign the model
2.1.5. Analyze the residuals
2.1.6. Interpretation of results
3. Determine the parameters in the selected metaheuristics by identify the number of partitions (groups)
3.1 Select the initial model (Box Bhenken, central composed , etc.)
3.2 Develop experimental tests
3.3Analise the regression model in order exits statistical evidence for the reliability of the experiment.
3.3.1 Verification of the experimental model
3.3.2 Validation of the parameters
4 Select the partition to better optimize the cost function

3 Demand Dense Methodology

Diverse territorial design applications are very useful to solve location problems for services and sales points [11, 12]. For this work, the logistical network design problem with dense demand over a geographical region implies defining (finding) market or services areas, this is a TD application.

The partition of one territory into zones generates *k* zones; this is understood of course as geographical zones grouping. Then, from a logistical point of view, we have available zones to locate facilities that provide services to satisfy the clients' demand. We have chosen a partition of 5 zones for our case study with the goal of allocating Demand Density Functions that we'll denote as DDF, proposed in [6], where DL1,..,DL6 are linear functions and as DNL1 as DNL2 are no linear functions and they are shown in the following table:

Table 1. Demand Density Functions (DDF)

DDF	$D(x, y)$
DL1	$100+7.5x+7.5y$
DL2	$100+10x+5y$
DL3	$100+(100/7)x+(5/7)y$
DL4	$600+(10/3)x+(5/3)y$
DL5	$600+2.5x+2.5y$
DL6	$100+(100/21)x+(5/21)y$
DNL1	$100+(9/80)x^2+(9/80)y^2$
DNL2	$100+(3/1.6\times10^5)x^4+(1.6\times10^5)y^4$

In integral calculus (Stewart 1999) the center of mass of an aluminum sheet that has a density function ρ(x, y) that occupies a region can be obtained (see Fig. 1).

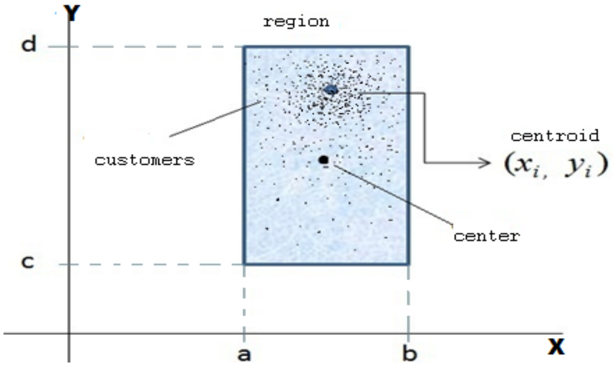

Fig. 1. Center and centroid of a rectangular region in R^2 based on the density of the points

It’s possible to assume that a geographical region in the plane is like a sheet and to obtain its center of mass or centroid through the use of (1), where the density function is a population density function, said point has geographical coordinates for rectangular regions, but in this project we are applying it to irregular regions. The centroids can be interpreted as location points for facilities because the point is located in a place with high population density [1, 2].

$$x = \frac{\iint x\rho(x,y)dxdy}{\iint \rho(x,y)} \quad ; \; y = \frac{\iint y\rho(x,y)dxdy}{\iint \rho(x,y)} \qquad (\mathbf{1})$$

The formulas to obtain the longitude, latitude coordinates of the location point of every geographical unit, require the use of double integrals over a region R, with density functions with two variables$\rho(x,y)\rho(x,y)$, where R is a geographical region and the density function can be any of the functions from table 1. On the other hand, we need to obtain the integration limits of every geographical unit, by means of a geographical information system (GIS) [1, 8]. The Romberg method has been chosen for double integration. The support software we’ve employed is free and is known as X numbers, which is an Excel add-on. In this way the coordinates of every centroid for every geographical unit, of territory is obtained.

From a logistical view, each territory that has a population possesses a demand for goods and services and it can defer per zones due to multiple factors. Let’s assume that the demand can be modeled by a DDF that associates a demand volume for a certain service to every geographic point. The demand density term is associated to the population density in the zone. Each of the five groups of geographical unit, will be associated with a DDF from table 1 to integrate them to the location model that is described below as a minimization problem.

Let’s consider that a centroid of an geographical unit, is a point with geographical coordinates which location depends on the density of population, we can say that $c_j = (x_j, y_j)$ is the representative of the geographical unit.

The solution consists in finding the coordinates $(x,y)(x,y)$ of the point $q \in Gi$ $q \in Gi$ such that the transportation cost from each community to the central facility is minimized.

The mathematical model that represents the conditions mentioned above is written in the following way:

$$\underset{(\mathrm{x},\mathrm{y}) \in \mathrm{G_i}}{\text{Minimize}} \quad \sum_{\mathrm{j}=1}^{|\mathrm{Gi}|} |\mathrm{D}(\mathrm{x},\mathrm{y})| \sqrt{\left(\mathrm{x}-\mathrm{x_j}\right)^2 + \left(\mathrm{y}-\mathrm{y_j}\right)^2} \qquad (\mathbf{2})$$

The objective function represented in (2) is the total transportation cost TC, known as the Weber function.

Parting from what we have exposed, we express the final model and contribution of our problem:

Given a set of customers distributed within a territory $\mathrm{T} \subseteq \mathrm{R}^2$ and $\mathrm{P} = \{\mathrm{G}_1,\ \mathrm{G}_2, \dots, \mathrm{G}_\mathrm{k}, \dots \mathrm{G}_\mathrm{p}\}$ a partition of T into p clusters, each G_k is a cluster of Agebs for $k = 1,2,\dots,p$. Each geographical object has a representative called centroid $c_j = (x_j\ , y_j,)$ from which each community is served. Each point $\mathrm{q} = (\mathrm{x},\mathrm{y}) \in \mathrm{T}$, has a density of demand given by $\mathrm{D(q)} = \mathrm{D(x,y)}$. Let $d(q, c_j)$, be the Euclidean distance from any point to the centroid. The cost of transportation from a point q to the centroid $\mathrm{c_j}$ is defined as $D(q)d(q, c_j)$.

The solution consists in finding the coordinates from a point $(\mathrm{x},\mathrm{y}) \in \mathrm{G_i}$ such that the transportation cost is minimized from each community to a central facility (like equation 2). The mathematical model that represents the mentioned conditions is the following:

$$\underset{(\mathrm{x},\mathrm{y}) \in \mathrm{G_i}}{\text{Minimize}} \quad \mathrm{Z} = \sum_{\mathrm{j}=1}^{\mathrm{Gi}} |\mathrm{D(x,y)}| \sqrt{(\mathrm{x}-\mathrm{x_j})^2 + (\mathrm{y}-\mathrm{y_j})^2} \qquad (\mathbf{3})$$

Subject to
$$\bigcup_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{G_i} = \mathrm{T} \qquad (\mathbf{4})$$

and
$$\bigcup_{\mathrm{j}=1}^{|\mathrm{G_i}|} \mathrm{A_j} = \mathrm{G_i} \qquad \forall\ \mathrm{i} = 1, 2, \dots, \mathrm{p} \qquad (\mathbf{5})$$

The objective function Z represented in (3) is the total transportation cost and (4) and (5), are the constraints of the partitioning of the territory T and the sub territories $\mathrm{G_i}$. The sequence of necessary steps to obtain the coordinates of the central facility in a cluster is as follows:

1. Define the parameters to partition the territory T.
2. Generate the partition with the VNS metaheuristic.
3. With the file obtained generate a map inside a map with a GIS.
4. Associate to the chosen cluster $P_i, i = 1,2,\dots,p$ $P_i, i = 1,2,\dots,p$, a demand density function $D(x,y).D(x,y)$.
5. Calculate the centroids of each AGEB, using (4).
6. Apply (3) for the chosen cluster.

$$x_i = \frac{\iint x\rho(x,y)dxdy}{\iint \rho(x,y)dxdy} \quad \text{and} \quad y_i = \frac{\iint y\rho(x,y)dxdy}{\iint \rho(x,y)\, dxdy} \qquad (\mathbf{6})$$

Equations (6) were rewritten to have an order in the methodology. This equation is also (1) and the equations above are the classical formulas of calculus used to calculate the centroid of a metallic plate with density ρ, in this paper we take the Agebs as metallic plates and a population density given by $\rho(x,y)$

4 Results

We have applied this methodology to one territory and we can establish that the solutions obtained for the LAP model for TDP with dense demand are consistent with the geographic location of the region. We obtained the coordinates of eight possible location points, depending on the demand density function, which associates itself to the geographical units belonging to the cluster under study selected from the partition of the territory. Fig. 3 shows the location of nine points. Therefore, we can select any point as the location point of the cluster; we can also state that the demand density function does not influence the location of the centers of the centroids. The ninth point which coordinates are $P_c = (x_{cc}, y_{cc})$ this point is calculated as the point which distance to any center is the minimum we can consider as the center of centers, and as the best point of location of the whole cluster of geographical units, without losing generalities. In a more general way, the location point can be any point within the circle $(x - x_{cc})^2 + (y - y_{cc})^2 = r^2$, where $r = max\{ d(P_c, c_j) \mid j = 1, 2, 3, \dots, n\}$. The proposal in this article gives a structure to solve location allocation models based on geographic information systems as it's reflected in the any case study.

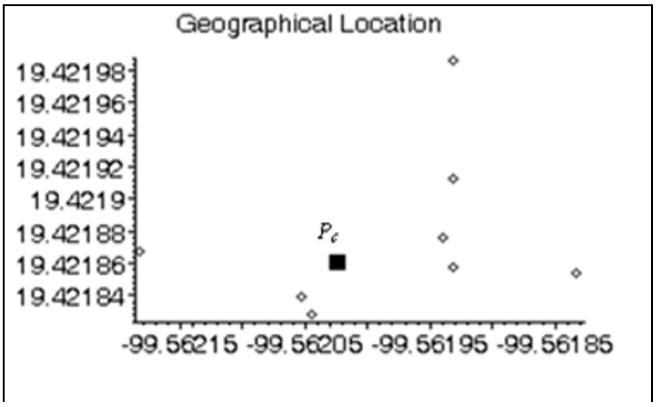

Fig. 2. Location that minimizes the objective function

The methodology was tested in demand regions of irregular shape in comparison with previous papers where the regions are rectangular or convex polygons. According to the analysis of the results obtained, the inclusion of the territorial design aspects with the use of density functions in location-allocation models, gives a greater range of possible applications to real problems, for example in the design of a supply chain, among others. The integration of diverse tools such as metaheuristics, geographic information systems and mathematical models provide a strong methodology in visual environments such as maps. Another contribution of this paper is the consideration of three relevant aspects: territorial design, location-allocation, and dense demand.

In general, the proposal presented contributes to the decision-making process in logistical problems when the population's demand is implicit. As a case study we chose Metropolitan Zone, however an advantage of our methodology consists in that it can handle other kinds of geographical data such as blocks, districts or states.

References

1. Zamora, E.: Implementación de un Algoritmo Compacto y Homogéneo para la Clasificación de AGEBs bajo una Interfaz Gráfica. Tesis de Ingeniería en Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Puebla, México, 18-27, (2006)
2. Bernábe, B., Espinosa, J., Ramírez, J., Osorio, M.A.: Statistical comparative analysis of Simulated Annealing and Variable Neighborhood Search for the Geographical Clustering Problem. Computación y Sistemas, vol. 42-3, pp. 295-308, (2009)
3. Bernábe, B., Pinto, D., E. Olivares, J., Vanoye, González, R., Martínez.J.: El problema de homogeneidad y compacidad en diseño territorial. XVI CLAIO Congreso Latino-Iberoamericano de Investigación Operativa, (2012)
   Bernábe, B., Coello, C., Osorio, M.A.: A Multiobjective Approach for the Heuristic Optimization of Compactness and Homogeneity in the Optimal Zoning. JART Journal of Applied Research and Technology, vol.10-3, pp. 447-457, (2012)
4. Díaz, J., Bernábe, B., Luna, D., Olivares, E., Martínez, J.L.: Relajación Lagrangeana para el problema de particionamiento en datos geográficos. Revista de Matemática Teoría y Aplicaciones vol. 19-2, pp. 43-55, (2012)
5. Pizza, E., A. Murillo, A., Trejos, J.: Nuevas técnicas de particionamiento en clasificación automática. Revista de Matemática: Teoría y Aplicaciones, vol. 6-1, pp.1–66, (1999)
6. Vicente, E.; Rivera, L.; Mauricio, D.; Grasp en la resolución del problema de cluster. ISSN: 1815-0268, vol. 2- 2, pp. 16-25, (2005)
7. MapX Developer´s guide. MapInfo Corporation. Troy, NY., www.mapinfo.com
8. Kaufman, L.; Rousseeuw, P.; Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm, North-Holland, Amsterdam , pp. 405-416, (1987)

An Effect of Term Selection and Expansion for Classifying Short Documents

Christian Sánchez-Sánchez, Héctor Jiménez-Salazar

Departamento de Tecnologías de la Información,
División de Ciencias de la Comunicación y Diseño,
Universidad Autónoma Metropolitana Unidad Cuajimalpa. México D.F.
{csanchez, hjimenez}@correo.cua.uam.mx

Abstract. Many web sites(blogs) over the Internet provide the users the possibility of sharing information like: opinions, news, even their profiles. The peculiarity of this information is that usually the description contains few words. Currently exist a great interest in developing tools that help to process this information in order to organize or categorize it, for helping decision making. Due the importance of this task, in this paper it is explored, through a set of experiments the effect of simple expansion and term selection over two Data Sets. It is applied the Absolute Term Frequency (ATF) term selection technique over this kind of documents, and it is showed that using a percentage of the terms, to represent the information, the classification result could be improved. At the end of the paper it is showed the classification phase where the document expansion could improve the number of classified instances.

Keywords. Term Selection, Document Expansion, Document Categorization

1 Introduction

Many Websites(blogs) over the Internet provide the users the possibility of sharing information like: opinions, news, even their profiles. The peculiarity of this information is that usually the description contains few words. For example Twitter allows to write messages at the most 140 characters.

Currently exist a great interest in developing tools that help to process Web Information in order to organize(for instance sentiment analysis) or categorize it(for example topic detection), for helping decision making. Regarding user profiles in some proposals authors try to identify leadership characteristics or classify them according the activity they do. As an example, In 2014 the Reputation Laboratory (REPLAB) [1] it was proposed a task were the objective was to categorize Twitter user profiles according to the domain activities the users do. The categories were: editors, public institutions and so on.

In some proposals for solving those problems supervised classification algorithms have been used. The common phases for classification process are: document indexing, vector dimension reduction, training, classification and evaluation [5]. Term or feature selection could help to: improve accuracy, delete redundancies and reduce computational cost. On the other hand, Term selection also has been used for query expansion in Relevance Feedback for Information Retrieval.

Nevertheless, regarding to the information contained in Blogs (seen as a collection of short documents) some questions arise: 1) *¿Can Term Selection benefit the categorization of this kind of information?*, 2) if this is so *¿Which term selection technique could work better to improve the classification outcome?*, 3) *¿Can Document expansion improve the classification results?*, 4) *¿What could be more convenient expanding short documents or selecting terms for improving classification results?*

In the interest of finding an answer to these questions, in this paper are employed and tested two different techniques for selecting terms, before classification: Information Gain (IG) and term selection based on taking percentage of term, sorted by Absolute Term Frequency. For expanding documents was used the technique of adding synonyms of all the terms contained in the Data Set.That way, through a set of experiments, it is showed the comparison of the results obtained using the term selection techniques and document expansion. Evaluating the classification outcome (Precision, Recall and F-Measure) using SMO algorithm [11] over two Data Sets (REPLAB14 and 20 News Groups).

It is important to mention that is not the focus of the designed experiments to improve the results of other approaches, but it is to try to identify the effect of term selection and document expansion over short documents during classification process.

The rest of the paper is organized as follow: in next section related work and some techniques for term selection are exhibited, section 3 describes the used Data Sets, section 4 define the proposed experiments, in section 5 the results of the experiments are shown, meanwhile last section present conclusions and future work.

2 Related Work

One of the term selection techniques, usually used for Text Categorization, is Document Frequency (DF). The DF results could be compared with classic techniques like X^2 or Information Gain (IG) [16].

While proposals like the reported by Joachims et al. [3] argues that term selection could weaken the efficiency of Classifiers like Support Vector Machines, also proposals for clustering and classification of Twitter Information have showed that term selection improved the results of those tasks. Such is the case of the work presented by Sánchez-Sánchez et al. [13] where tweets are clustered according to their topic, or the proposed by Villatoro et al. [14] in which Twitter author profiles are classified and ranked, both cases they use DF term selection to improve the results. Similarly, Pinto et al. [10] acceptable results in the clustering of scientific texts (abstracts) using the Transition Point Technique.

Li et al. [5] proposed to obtain the discriminability and coverage of terms in order to select them, using a combination of measures like DF, a probability ratio and the Average Vector Length. The last measure because it is believed that the poor accuracy at a low dimensionality is imputed to the small average vector length of the documents. They showed that this proposal improved the results gotten using X^2 in two different data sets.

Similarly, Peters et al. [9] presented a uncertainty-based mechanism to discriminate the noisy terms and then select the rest of the terms. Here it is showed how to calculate the uncertainty according to a relative frequency of terms and DF. Such that the model calculates value-uncertainty tuples with the purpose of evaluate the quality of information through a k factor (the value mean divided by uncertainty mean). Small values of k, according to a given value Q, represent noisy terms. It is shown a comparison, against other methods, where it were gotten competitive results over three Data Sets.

The work generated by Lam-Adesina [4] term selection was used in order to tackle the Relevance Feedback IR feature. In Its proposal the first results gotten (using a query) are summarized, The summarizing is done using employing Lunh's keywords clustering [7] with and without considering the query terms. Then the probabilistic model BM25 [12] is applied in order to weight terms and join the heavier terms to a new query.

On the other hand, it is proposed a term selection mechanism based in calculating a set of features: term distribution, query term co-occurrence, pair query term co-occurrence, weighted term proximity, query and expansion DF. Finally, each result is classified as "good" or "bad" to obtain a model that helps the selection.

3 The DataSets

Two collections were used: Twitter Profile and 20 news groups. The first one is composed by a set of user's profiles divides in Training and Test. This data set was generated for been used in REPLAB competition in 2014 (REPLAB14). The data set was designed for solving the task of categorization of users according to a certain domain of activities, classifying users as: publishers, public institutions, athletes, etc.

Below it is given more information about the collections:

3.1 REPLAB14 Training and Test DataSets

Training collection has 10 user's profile categories: celebrity, company, employee, investor, journalist, ngo, professional, public institutions, sportmen, undecidable.

Each category is formed by the next number of documents: celebrity (61 profiles), company (145 profiles), employee (4 profiles), investor (3 profiles), journalist (466 profiles), ngo (102 profiles), professional (594 profiles), public institutions (40 profiles), sportmen (57 profiles), undecidable (1027 profiles).

Test Data Set has the same number of categories that training, but each category has more profiles. Next more details are given: celebrity (208 profiles), company (222 profiles), employee (14 profiles), investor (7 profiles), journalist

(992 profiles), ngo (233 profiles), professional (1543 profiles), public institutions (90 profiles), sportmen (208 profiles), undecidable (1412 profiles).

3.2 20 News Groups

The data set is organized into 20 different newsgroups (corresponding to a different topic). The characteristic of this collection is that some of the topics are very closely related to each other, while others are not. The newsgroups are: comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.-mac.hardware , comp.windows.x, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, misc.forsale, talk.-politics.misc, talk.politics.guns, talk.politics.mideast, talk.religion.misc, soc.religion.christian and alt.atheism. The version of the data set used has 18846 documents sorted by date(divided into Training and Test) ; duplicates and headers were removed.

4 Experiment Configuration

Some considerations were taken into consideration for the experiments design, these are explained below.

Currently, Support Vector Machines (SVM) [2] had become a learning algorithm very popular for Test Categorization Task [6], due it's consistent execution and capacity for handling big dimension space of inputs. That is why for the proposed experiments was decided to utilize it applying Sequential Minimal Optimization (SMO), using Weka [15]. For tackling multi-class classification was used pairwise classification and the predicted probabilities are coupled using Hastie and Tibshirani's pairwise coupling method in WEKA.

In order to standardize the information and eliminate some elements that may be irrelevant to the classification, the following pre-processing was performed: all text was transformed to lowercase, eliminating URL's, deleting all punctuation marks, removing words and truncating with Porter algorithm.

For representing the documents was used a Boolean weighting model, whereby the presence or absence of the term in the document is indicated. So that with the terms reduction or expansion, the dimension of the vectors changes.

Concerning to answer some of the questions previously stated, the following experiments described in next section were proposed.

4.1 Experiment 1-Classifying Data Sets without Term Selection or Expansion

In the interest to have a baseline, and to know if the reduction or expansion of terms could benefit or affect the classification, as it was stated in Question 1 (which also is immersed in all experiments). The classifier was trained (using cross-validation) to get the models for each of the collections.

That is, for the first five experiments was used REPLAB14 Data Set, the collection was divided into training and test data (5 times, One for each experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**2a)** it was done the same using 20 News Groups Data Set.

4.2 Experiment 2-Classifying Expanded Documents in Data Sets

In this experiment **2a)** each of the documents contained in the REPLAB2014 collection was expanded, by adding to each document synonyms of the terms that comprise it. In order to get the synonyms Wordnet [8] resource was used. Similarly to the previous experiment 5 experiments were performed, the collection was divided into training and test data (5 times, One for each experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**2b)** it was done the same using 20 News Groups Data Set.

The results gotten in Experiment **1a** con **2a** were compared with **1b** and **2b** respectively. This in order to help answer questions 3 and 4.

4.3 Experiment 3-Classifying Data Sets formed by Reduced Documents, by Term Selection using Information Gain

The third experiment aimed to help answering questions 2 and 4. In the first part, **3a)** Information Gain method (GI) was used to select a subset of terms, with these terms each document was represented. The representation is based on leaving, inside the Document, only the terms found in the subset obtained. Having done this, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment) The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**3b)** it was done the same using 20 News Groups Data Set. In **3c)** the documents were expanded and using GI some terms are selected to represent the documents. Finally, **3d)** it was done the same using 20 News Group collection.

4.4 Experiment 4- Classifying Data Sets formed by Reduced Documents, by Term Selection using Absolute Term Frequency(ATF)

Similarly to the previous experiment trying to find answers to questions 2 and 4. **4a)** A list is formed with tuples term-ATF, ATF is the number that the each term t_k appears in the whole Data Set. Where d_i represents a document, T the Data Set and t_k each term.

$$ATFt_k = \sum_{d_i \in T} t_k \tag{1}$$

Once all frequencies were obtained the terms were ordered from highest to lowest ATF. Then, it were taken an amount of terms from 10% to 90%. Subsequently, all the documents inside the Data Set were represented with each percentage of terms. After that, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. **4b)** it was done the same using 20 News Groups Data Set.

For each classification obtained(using the percentages of terms to represent documents), the results were compared in order to identify the best representation, that which improved precision. In some cases finest results were searched, through exploring nearest percentages of terms, taking one percent of terms more (of the best) each time.

Results from experiments **3a** and **4a**, and **3b** and **4b** were compared respectively, in order to identify the representation that improves precision.

4.5 Experiment 5- Classifying reduced Documents after Expansion

This experiment was designed looking for finding an answer to the question 3. Using REPLAB14 Data Set it was sought to test if **5a)** the winning term selection technique(reduction) worked after Document expansion. That's why also, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. **5b)** it was done the same using 20 News Groups Data Set.

4.6 Experiment 6- Training the Classifier with the Best Representations for the Data Sets and Testing

With regard to integrate and use results and answering questions 1-4.In this experiment it was sought to use the best representation of the Training part in order to generate the model and evaluate the results of classifying the test part, for both collections: **6a1)** REPLAB14 and **6a2)** 20 News Group.

For making this possible, the Test part of both Data Sets must be represented with the same terms of the winner reduction (selection) technique. Meanwhile, in the last experiment the documents inside the Test part of each Data Sets:**6b1)**REPLAB14 and **6b2)**20 News Groups are expanded and then represented with the winner reduction technique.

5 Experiment Results

Next the results of each experiment are shown, as well as the comparisons among: term selection techniques, term selection vs expansion, and term selection vs no selection.

Experiment 1

Those results were the baseline to compare, wondering if term selection improves classification. After pre-processing and indexing Data Sets, the classification result (P=Precision, R= Recall, F= F-Measure) was:

1a)

P=0.299, R=0.205, F=0.216

1b) While using 20 News Groups result was:

P=0.767, R=0.751, F=0.756

Experiment 2

2a) After expanding documents, including synonyms of each term inside each document, and then classifying REPLAB14, the outcome was:

P=0.241, R=0.192, F=0.201.

If this result is compared to the previous **1a** it can be observed that the classification was not improved after expansion. Actually, F-Measure was lower, $0.201 < 0.216$.

2b) Afterwards expanding 20 News Groups, results were:

P=0.717, R=0.746, F=0.731 Comparing F-Measure to the previous **1b** it can be noted that the classification neither was improved, , $0.731 < 0.756$.

Experiment 3

3a) After applying Information Gain technique, for selecting terms in REPLAB14, the results were:

P=0.304, R=0.201, F=0.213 Comparing results to the reported in **1a** neither was an improvement in F-Measure, though the precision was higher.

3b) Applying GI for term selection to represent 20 News Groups Documents results were:

P=0.769, R=0.75, F=0.759.In this case there was improvement in classification results compared to**1b**.

3c) If GI is applied after expanding Documents the result was:

P=0.223, R=0.19, F=0.197 It can be observed that the classification result is worst than the reported in **1a**.

3d) If a similar process **3c** is applied to 20 News Groups Data Set the result is not favorable:

P=0.761, R=0.75, F=0.755

Experiment 4

4a) After calculating the Absolute Term Frequency for all terms, the terms are listed from lowest to highest frequency. The results of taking a percentage of those terms and represent the collection REPLAB14 and classify it is showed in the next table (see Table.1). As it can be seen the classification outcome, compared with the result reported in **1a**, is highest if it is taken from 60% to 70% where it reach the peak (P=0.317, R=0.224, F=0.242) and then it start decreasing.

4b) Applying the same term selection technique (ATF), in order to classify 20 News Groups the results are depicted in Table 2.

It can be detected that the classification outcome is improved, compared with the reported in **1b**, if the percentage of selected terms is 90% where it is the

Table 1. Classification results gotten employing ATF term selection, over REPLAB14

Percentage	Precision	Recall	F-measure
10%	0.249	0.218	0.227
20%	0.268	0.219	0.232
30%	0.259	0.215	0.226
40%	0.28	0.223	0.237
50%	0.297	0.221	0.236
60%	0.271	0.208	0.219
70%	**0.317**	**0.224**	**0.242**
80%	0.286	0.212	0.224
90%	0.286	0.205	0.217
100%	0.299	0.205	0.216

Table 2. Classification results gotten employing ATF term selection, over 20 News Groups

Percentage	Precision	Recall	F-measure
10%	0.737	0.728	0.731
20%	0.747	0.736	0.74
30%	0.748	0.736	0.74
40%	0.75	0.737	0.742
50%	0.75	0.739	0.743
60%	0.754	0.743	0.747
70%	0.758	0.743	0.748
80%	0.759	0.744	0.748
90%	**0.771**	**0.754**	**0.759**
100%	0.767	0.751	0.756

maximum (P=0.771,R=0.754,F=0.759). The term selection technique applied in both Data Sets helped to improve the results.

Experiment 5

5a) If document expansion is applied before ATF term selection, with the purpose of classifying REPLAB14, the next results were gotten (See Table.3).

Comparing the results reported in **1a**, it can be seen that representing documents with the 80% of the terms the classification is better (P=0.242,R=0.2,F=0.21). In spite of the documents where previously expanded the result was the opposite than the reported in **3c**. Nevertheless, the result was lower than the reported in **4a**.

5b) If it is done the same process, described in **5a**, over 20 News Groups the maximum outcome was obtained selecting 77% of the terms, (P=0.633,R=0.629, F=0.630). Nonetheless, it is not a better result that the reported in **4b**.

It is important to say that for those experiments the ATF term selection worked better than IG, and it has benefited the classification results of such documents.

Table 3. Classification results gotten employing ATF term selection after document expansion, over REPLAB14

Percentage	**Precision**	**Recall**	**F-measure**
10%	0.179	0.174	0.174
20%	0.181	0.177	0.177
30%	0.208	0.199	0.201
40%	0.222	0.202	0.208
50%	0.222	0.202	0.208
60%	0.212	0.192	0.198
70%	0.227	0.194	0.201
80%	**0.242**	**0.2**	**0.21**
90%	0.232	0.19	0.198
100%	0.241	0.192	0.201

Experiment 6

6a1) Due that selecting the 70% of the terms sorted by ATF, and representing the REPLAB14 documents, in order to classify them, gave the best result. Then all the documents of the Test Data Set part were represented with those terms. It was used the model gotten using the REPLAB14 training part and the results of classifying the test was:
P=0.333 R=0.141 F=0.125

It is important to mention that selecting the 70% of the terms 4619 documents of a total of 4929, contained in the Test Part. And the number of correctly classified instances was 1638 (35.5 %)

6a2) The result of applying the previous process over 20 News Groups was:
P=0.769 R=0.757 F=0.762

6b1) Nevertheless if the REPLAB14 Test documents are represented using the 70% of the terms after Document expansion, using the same model used in **6a1** for classification. The results were:
P=0.260 R=0.174 F=0.172

Although the precision was lower indeed more documents could be represented, compared to **6a1**, 4749 documents of a total of 4929. Where 1873 Documents were classified correctly (39.5%).

In this case it can be said that the expansion in the test documents helped to classify correctly more documents.

6b2) The result of applying the previous process over 20 News Groups was:
P=0.748 R=0.793 F=0.0.769

6 Conclusions and Future Work

As a conclusion it could be argued that employing ATF Term Selection and GI benefited classification results. This using SMO Algorithm and a Boolean Weight Representation, of the vector members for representing the documents. The ATF Term selection technique was better although other terms where included in the documents (by expansion).

In this case ATF term selection was better, selecting from 80% to 90% the classification accuracy is better than using all the terms. This was a constant in both Data sets.

ATF algorithm took the terms according to their popularity (frequency between documents and frequently in the document, jointly) and therefore did not discriminate some terms that other techniques may penalize or dismiss easily. This could be deduced reviewing the REPLAB14 Data Set but it is necessary to analyse other Data Sets.

Document expansion (adding the synonyms of all the term into the document) showed that can help to improve classification results if and only if it is applied in the Test part, similarly that it is done in Relevance Feedback for Information Retrieval.

As future work it is planned to: *a)* Perform tests of statistical significance for the results and subsequently, *b)* design experiments with; other collections of similar short documents, other types of representations and other classifiers. *c)* make comparisons to other term selection techniques like DF or Transition Point.

References

1. Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., Rijke, M., Spina, D.: Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, chap. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management, pp. 307–322. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-11382-1_24
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995), http://dx.doi.org/10.1023/A:1022627411411
3. Joachims, T.: Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings, chap. Text categorization with Support Vector Machines: Learning with many relevant features, pp. 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg (1998), http://dx.doi.org/10.1007/BFb0026683
4. Lam-Adesina, A.M., Jones, G.J.F.: Applying summarization techniques for term selection in relevance feedback. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1–9. SIGIR '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/383952.383953
5. Li, J., Sun, M.: Scalable term selection for text categorization. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 774–782. Association for Computational Linguistics, Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/D/D07/D07-1081
6. Liu, Y., Loh, H.T., Kamal, Y.T., Tor, S.B.: Natural Language Processing and Text Mining, chap. Handling of Imbalanced Data in Text Classification: Category-Based Term Weights, pp. 171–192. Springer London, London (2007), http://dx.doi.org/10.1007/978-1-84628-754-1_10

7. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (Apr 1958), http://dx.doi.org/10.1147/rd.22.0159
8. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995), http://doi.acm.org/10.1145/219717.219748
9. Peters, C., Koster, C.H.A.: Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25–27, 2002 Proceedings, chap. Uncertainty-Based Noise Reduction and Term Selection in Text Categorization, pp. 248–267. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-45886-7_17
10. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering Abstracts of Scientific Texts Using the Transition Point Technique, pp. 536–546. Springer Berlin Heidelberg, Berlin, Heidelberg (2006), http://dx.doi.org/10.1007/11671299_55
11. Platt, J.C.: Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep., Advances in kernel methods - Support vector learning (1998)
12. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (Apr 2009), http://dx.doi.org/10.1561/1500000019
13. Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W.A.: Uamclyr at replab2013: Monitoring task. In: CLEF (Working Notes) (2013)
14. Villatoro-Tello, E., Ramírez-de-la Rosa, G., Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W.A., Rodríguez-Lucatero, C.: Uamclyr at replab 2014: Author profiling task. In: CLEF (Working Notes). pp. 1547–1558 (2014)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
16. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 412–420. ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997), http://dl.acm.org/citation.cfm?id=645526.657137

Design of Silent Actuators using Shape Memory Alloy

Jaideep Upadhyay{1,2}, Husain Khambati{1,2}, David Pinto[1]

[1]Benémerita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación, Mexico

[2]Manipal University Jaipur, India
{jaideep1895@gmail.com,husainkhambati1895,davideduardopinto}@gmail.com
http://www.lke.buap.mx/

Abstract. In this paper we present a design for an actuator using a shape memory alloy called flexinol. This actuator can be used to mimic the movements of various muscle groups for robotic applications. This design takes advantage of the properties of shape memory alloys, especially flexinol and its high contraction force. The use of flexinol enables the device to be compact and be highly reliable. We also present an efficient way of controlling the movement of the shape memory alloy using pulse width modulation. The designs are to be embedded inside an artificial facial skin of a robot to expresses human like emotions in robots, in particular, as artificial muscle design for human cheek muscles and the muscles which are used only for linear movement of the human skin.

Keywords. Flexinol, Shape Memory Alloy, Actuators, Artificial Facial Skin

1 Introduction

Generating emotional facial expression in robots is a very challenging task which has been studied for different research teams. This interest has been motivated by the current progress of robots who has been evolved from only being useful in industrial settings to stay in close contact with human beings. In this way, emotion expression in robots is essential for human-robot interaction.

There have been developed different robots providing emotional expressions such as the ones introduced by the research team of Hiroshi Ishiguro. A number of realistic humanoid robots that include an adult woman and a child that can recite news reports in different languages and voices [3, 4].

There are, however, other research teams working in providing emotional expressions to realistic humanoids, for example, BINA48 [1], who is a humanoid robot created by Hanson Robotics and owned by Martine Rothblatt's Terasem Movement, Incorporated (TMI), which is a non-profit organization with the purpose of preserving, evoking, reviving and downloading human consciousness in non-biological or nantech bodies [2]. This robot was created with the aim of showing emotional expressions when having a dialog with human beings.

Some other devices exist around the world, however, we would like to talk about Arthur (see Figure 1), a humanoid robot capable of show emotional expressions in its face. This robot is held by the research team of the Language & Knowledge Engineering Lab. at the Faculty of Computer Science of Benemerita Universidad Autonoma de Puebla.

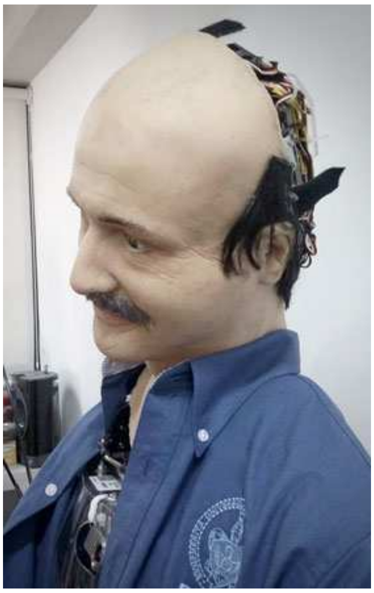

Fig. 1. Humanoid robot al LKE Laboratory

Arthur has more than 20 servomotors under the skin which allows it to move its artificial skin in different positions with the aim of expressing emotions such as anger, happiness or sadness. Currently, this robot uses plastic strings in order to move the skin, which we consider to be an obsolete kind of technology. In this way, we would like to explore new ways to allow the humanoid robot to express those kind of emotions but with much more reliable technologies. This paper explores the use of new actuators based on Flexinol for such purpose.

The remaining of this paper is structured as follows. Section 2 motivates our research work presenting advantages of the approach. Section 3 discusses the design of actuators using shape memory alloy. Section 4 presents an analysis of the type of Flexinol wires employed in the experiments with different test over them. Section 5 show a preliminar design of the actuators proposed. Finally, in Section 6, the conclusions of this research work are given.

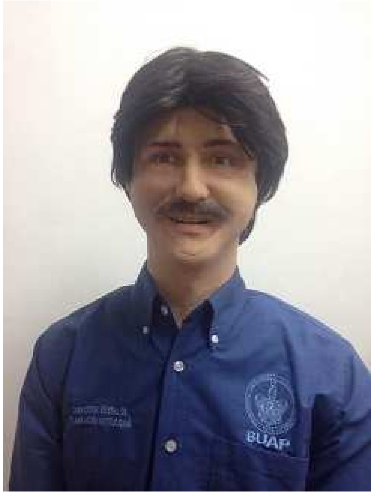

Fig. 2. Humanoid robot al LKE Laboratory

2 Motivation

The current technology which deals with generation of facial expressions in a robot makes use of a complex placement of servo motors and links to impart expressions in the artificial skin of the robot. This system consumes a lot of power and is very complex to rectify if a problem arises. The strings or links are attached at various positions inside the artificial skin of the robot and to the rotating parts of the servo motors, this enables the servo motors to pull the skin from inside to mimic the human expressions. This method makes the skin of the robot very weak when the strings pull it and also damages the inside layer of the skin over a period of time. Due to the complex arrangement of the servo motors and the links inside, it is a great problem if a part is damaged or not working properly as the complexity of the system makes it impossible for the user to repair or replace the part on his own. The cost of the system also increases drastically as a lot of servo motors are used. Servos are bulky and due to their gearing system produce noise. Our method provides a simple solution of using Shape memory alloys as the primary actuator. The advantages of these alloys are abundant. They are smaller, lighter, easy to control, cheaper and reliable. The manufacturing of these actuators is also a simple process.

3 Designing Silent Actuators

The aim of this Research Project is to develop an actuator design which can replace servo motors and solenoids in robotic application. The design presented in the paper have several advantages over servo motors and solenoids as our design is smaller, less noisy and can be replaced very easily if damage occurs. As flexinol is basically just a wire so the designs can be configured to suit any situation very easily. Flexinol is a form of shape memory alloy as Shape memory alloys display two distinct crystal structures or phases. Temperature and internal crystal structure determine the phase that the Shape memory alloy will be at. Martensite exists at lower temperatures, and austenite exists at higher temperatures. When a Shape memory alloy is in martensite form at lower temperatures, the metal can easily be deformed into any shape. When the alloy is heated, it goes through transformation from martensite to austenite. In the austenite phase, the memory metal "remembers" the shape it had before it was deformed. Flexinols austenite phase is such that it allows the structure compresses. The flexinol wire contracts to upto 2%-5% of its original length when it is heated to austenite phase and thus produces a pulling force whose magnitude depends upon the diameter of the wire used. The heating of the flexinol is achieved by passing a certain amount of current through the wire. The flexinol wire inherently possesses very little resistance to the current applied, so a current limiting system is required to control the amount of current flowing through the wire otherwise this may cause irreparable damage to the structure of the alloy. Generally, the shape memory alloy has over a million life cycles but this depends on the stress exerted on the crystal structure of the alloy. High current, high temperatures or high loading on the wire may drastically affect its life cycles. The designs discussed in this paper will mainly focus on the artificial muscle design for human cheek muscles and the muscles which are used only for linear movement of the human skin.

4 Analysis of Flexinol Wires

For the efficient design of the actuators it was necessary to know the behaviour of the flexinol wires so various basic experiments were carried out.

Table 1 shows the specifications of three different flexinol wires as all the experiments were carried out using these wires only at room temperature.

Table 1. Specifications of three different flexinol wires at room temperature

Feature	**Diameter Size (Inches)**		
	0.004	**0.008**	**0.015**
Resistance (Ohms/Inch)	3.0	0.8	0.2
Maximum Pull Force (grams)	150	590	2,000
Approximate* Current (mA)	180	610	2,750
Contraction* Time (seconds)	1	1	1
Off Time 90 C HT Wire (seconds)**	0.4	2.2	10.0

*Please note contraction time is directly related to electric current imposed. The guidelines are only approximations, since other factors like ambient temperature, air currents, and heat sinking will vary with specific devices.

** Approximate Cooling Time

4.1 Test-1

For the first test a 0.004-inch diameter wire was used having a phase transition temperature of 90C was used to determine the load bearing capacity of the wire. A current of 0.35 amps at a voltage of 5V was supplied to the wire. The contracting force of the wire was measured to be 420g approx. The life cycle of the wire was reduced to 10 cycles as the current supplied is approximately twice the amount of current for the safe operation of the wire as seen in the above table.

4.2 Test-2

For the second test a 0.015-inch diameter wire was used having a phase transition temperature of 90C was used to determine the load bearing capacity of the wire. A current of 2.91 amps at a voltage of 9V was applied to the wire. The contracting force of the wire was measured to be 3kg approx. The life cycle of the wire was reduced to 6-7 cycles and the cooling rate of the wire was also too low for rapid actuation.

5 Preliminar Design of Silent Actuators

In this paper we only sketch the general design of two actuators: a linear and a bending one.

The Linear Muscle Actuator should consist of two hollow shafts that should fit into one another. One shaft of the device must have a key and the other should have a keyway into which the keyed shaft will sit. This design is to prevent the two parts from twisting onto each other when it is actuated. A compression spring should be inserted into both the shafts. The flexinol wire has to be held using nuts and bolts on the outside of the structure so as to increase the cooling rate of the wire. The two flexinol wires should be connected in parallel in the circuit. The compression of the flexinol wire is expected to cause the two shafts to compress the spring and once the compression of the flexinol is removed the spring will allow the two shafts to return to their respective positions. When this device is going to be used as a facial muscle one end of the actuator should be fixed to a solid structure while the other end should be attached to the artificial skin in order to ensure the linear movement in one direction only. The amount of linear movement in this device shoulb be controlloed by an amount of current flowing through the flexinol wire by applying a PWM (pulse width modulation) signal of appropriate duty cycle.

The Bending Muscle Actuator should consist of a thin plastic plate shaped like a human facial muscle and a spring tightly attached at the middle of the top part of the plastic plate. The sole purpose of the plastic plate is to provide structure to the muscle actuator and to provide proper curvature while bending. Nuts and bolts are need for holding the flexinol wires (shape memory alloy, nitinol) in place and to also act as a medium of current flow in the wires. The flexinol wires should be wound between the nuts and bolts on either side of the plastic plate. The spring must be isolated from the nut and bolt assembly and the flexinol wires to stop the current flowing through the spring which can change the path of the current and the flexinol wires will fail to heat up and contract.

A spring having its compressed state as its natural state is going to be used to return the muscle actuator to its original position. The spring attached to the top part of the plastic plate will bend along with the plastic plate, thus causing an expansion in the spring, generating a contractive force in the spring and thus when the muscle actuator is switched off the spring will pull back the actuator to its original shape. The force exerted by the spring and the force exerted by the flexinol wire will act in opposite direction and the amount of bending will be controlled by varying the force exerted by the flexinol wire and thus different amount of bending can be achieved at different equilibrium points of the force exerted by the spring and the flexinol wire. The force exerted by the flexinol wire could be adjusted by varying the amount of current supplied to the it. The force exerted by the flexinol wire will bend the plastic plate and it will mimic the motion of a muscle.

6 Conclusions and Further Work

In this paper we have presented a general desing of silent actuators for a humanoid robot with the aim of having in a near future new technology for expressing emotions in this robot using silent and much more reliable actuators.

Some of the conclusions after analyzing the wire proposed to use follows:

1. Use of a current limiting circuit is a must as the wire tends to draw more current than its rated value. The greater the current drawn the greater the temperature in the wire which will cause a decrease in the life cycles of the shape memory alloy.
2. We would prefer thinner wires to their thicker counterparts. The thicker the wire the more the power is required for its actuation and also the cooling time of the wire is increased drastically. If the force required by these flexinol wires is greater, then a bundle of thinner wires should be preferred over the thicker wires.

Hence from the conclusions of the above experiments the 0.004-inch wire was used as an actuator wire for the artificial muscles as it has a very quick cooling rate thus a very good reaction time and a current limiting circuit was installed to keep the current of the wire below the rated current. The thicker wires would require a separate system for cooling which would increase the size of our device.

Our future work will involve the designing of the complete facial muscle groups. Also it will involve designing of the skeletal frame and skin for the proper actuation of the muscle groups. Designing of the complete control system for the entire facial expression system will also be a priority in future work. Testing and implementation of better grade quality of shape memory alloys or other such smart materials in the design for better efficiency and control.

References

1. Harmon, A.: Making friends with a robot named bina48. New York Times, The address of the publisher (July 4th 2010), retrieved October 10, 2016
2. Inc., T.M.: Terasem movement: Lives are good. http://www.terasemcentral.org/ (2016), retrieved October 10, 2016
3. Nishio, S., Ishiguro, H., Hagita, N.: Geminoid: Teleoperated android of an existing person. In: de Pina Filho, A.C. (ed.) Humanoid Robots: New Developments, pp. 343–352. I-Tech Education and Publishing, Vienna, Austria (Jun 2007), http://www.intechopen.com/articles/show/title/geminoid__teleoperated_android_of_an_existing_person
4. Nishio, S., Taura, K., Ishiguro, H.: Regulating emotion by facial feedback from teleoperated android robot. In: International Conference on Social Robotics. pp. 388–397. Chengdu, China (Oct 2012), http://link.springer.com/chapter/10.1007/978-3-642-34103-8_39

Analysis of Mammograms Using Texture Segmentation

J. Quintanilla-Domínguez[1], J.M. Barrón-Adame[1], J.A. Gordillo-Sosa[1], J. M. Lozano-Garcia[2], H. Estrada-García[2] and R. Guzmán-Cabrera[2]*

[1]Universidad Tecnológica del Suroeste del Estado de Guanajuato, UTSOE México
[2]Campus Irapuato-Salamanca, Universidad de Guanajuato, México

* guzmanc@ugto.mx

Abstract. Breast cancer is one of the most dangerous types of cancer among women around the world. It is also one of the leading causes of mortality in middle and old aged women. The World Health Organization's International Agency for Research on Cancer estimates that more than 1 million cases of breast cancer will occur worldwide annually, with 580,000 cases occurring in developed countries and the remainder in developing countries. In this paper, we present an effective methodology in order to detect clusters of Micro calcifications in digitized mammograms, based on the synergy of Image Processing, Pattern Recognition and Artificial Intelligence. The results obtained allow us to see the effectiveness of the proposed method.

Keywords. Segmentation, Cancer Detection, Image Processing

1 Introduction

The risk of a woman developing breast cancer during her life time is approximately 11% [1]. The early detection of breast cancer is of vital importance for the success of treatment, with the main goal to increase the probability of survival for patients. Currently the most reliable and practical method for early detection and screening of breast cancer is mammography. Mammography is a technique used to visualize normal and abnormal structures within the breasts. Mammography is a special type of X-ray imaging used to create detailed images of the breast (Mammograms). However, achieving this early cancer detection is not an easy task. Micro calcifications (MCs) can be an important early sign of breast cancer, they appear as bright spots of calcium deposits. Individual MCs are sometimes difficult to detect because of the surrounding breast tissue, their variation in shape, orientation, brightness and diameter size [2]. But it is still a hard task to detect all the MCs in mammograms, because of the poor contrast with the tissue that surrounds them.

However, many techniques have been proposed to detect the presence of MCs in mammograms: image enhancement techniques, Artificial Neural Networks (ANN), wavelet analysis, Support Vector Machines (SVM), mathematical morphology, image analysis models, fuzzy logic techniques, etc. Image enhancement algorithms have been utilized for the improvement of contrast features and the suppression of noise. In [3] proposed five image enhancement algorithms for the detection of MCs in mammograms.

Bhattacharya and Das [4] proposed a method based on discrete wavelet transform due to its multiresolution properties with the goal to segment MCs in digital mammograms. Morphological Top-Hat algorithm was applied for contrast enhancement of the MCs. Fuzzy C-Means clustering (FCM) algorithm was implemented for intensity-based segmentation. Sung et al. [5] proposed an approach by means of mathematical morphology operations and wavelet transform to locate the MCs in digital mammogram.

In [6] proposed an algorithm that was tested over several images taken from the digital database for screening mammography for cancer research and diagnosis, and it was found to be absolutely suitable to distinguish masses and microcalcifications from the background tissue using morphological operators and then extract them through machine learning techniques and a clustering algorithm for intensity-based segmentation. Segmentation processes for the detection of textures, ROIs, lesions, tumors have also been used on photo-acoustic images [7] and thermographic images [8].

The remaining sections of this work are organized as follows: Section 2, presents the details of the proposed method. Section 3 presents the details of the proposed method and experimental results while the conclusions are presented in sections 4, respectively.

2 Methodology

The mammograms used to train and test in this work were extracted from a mini-mammographic database provided by Mammographic Image Analysis Society (MIAS) [9]. This database contains 322 mammograms, 118 mammograms contain some abnormality, 66 are benign and 52 are malignant, and the remainder of the mammograms are diagnosed as normal.

The abnormalities found in these mammograms are MCs, Well-defined/circumscribed masses, ill-defined masses, speculated masses, architectural distortions and asymmetries. Each mammogram from the database is 1024 × 1024 pixels and with a spatial resolution of 200 µm/ pixel. These mammograms have been reviewed by an expert radiologist and all the abnormalities have been identified and classified. The place where these abnormalities such as MCs, have been located is known as, Region of Interest (ROI). The ROI images size in this work is 256 × 256.

The proposed approach is applied to each of the ROI images individually in order to show the obtained results by means of a segmented image. Several ROI images from mammograms with dense tissue and the presence of MCs were selected to train and test the proposed approach.

Next, the morphological Top-Hat transform is used in order to enhance the ROI image, with the goal of detecting objects that differ in brightness from the surrounding background, in this case is to increase the contrast between the MCs clusters and background. Then, the same structuring element at with same size is applied. The size of structuring element used is 3 × 3. Fig. 1 shows original ROI images processed by Top-Hat transform. In the next stage two window-based features such as, mean and standard deviation were applied. They are extracted from enhancement images within a rectangular window, in this work a 5×5-pixel block window was used.

To carry out the grouping and labeling, we use two methods well known in the state of the art: k-means and SMO, these methods are briefly described below:

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set Z in a d-dimensional space, through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k prototypes, one for each cluster. The next step is to take each point belonging to a given Z and associate it to the nearest prototype. When no point is pending, the first step is completed and an early group is done.

At this point it is necessary to re-calculate k new prototypes as centers of the clusters resulting from the previous step. After obtaining these k new prototypes, a new binding has to be done between the same data set points and the nearest new prototype. A loop has been generated. As a result of this loop we may notice that the k prototypes change their location step by step until no more changes are done. In other words, prototypes do not move any more.

The initial conditions used in this work for this method were:

- Cluster number takes values 2 to 6.
- Prototypes were initialized as random values.
- Euclidean distance function.
- Maximum iteration number: 100.

Self-organizing maps (SMO), are simple analogues to the brain’s way to organize information in a logical manner. The main purpose of this neural information processing is the transformation of a feature vector of arbitrary dimension drawn from the given feature space into simplified generally two-dimensional discrete maps. A SOM network performs the transformation adaptively in a topological ordered fashion. This type of neural network utilizes an unsupervised learning method, known as competitive learning, and is useful for analyzing data with unknown relationships. The basic SOM Neural Network consists of the input layer, and the output (Kohonen) layer which is fully connected with the input layer by the adjusted weights (prototype vectors). The number of units in the input layer corresponds to the dimension of the data.

There are several steps in the application of the algorithm. These are competition and learning, to get the winner in the process. In the training (learning) phase, the SOM forms an elastic net that folds onto the “cloud” formed by the input data. Similar input vectors should be mapped close together on nearby neurons, and group them into clusters. If a single neuron in the Kohonen layer is excited by some stimulus, neurons in the surrounding area are also excited. That means for the given task of interpreting multidimensional image data, each feature vector x, which is presented to the four neurons of the input layer, typically causes a localized region of active neurons against the quiet background in the Kohonen layer.

The initial conditions for this method were:

- Network structure [4 k], k takes values 2 to 16.
- Weight vector was initialized as random values.
- Topology function: Hexagonal layer.

- Distance function: Euclidean distance.
- Maximum epoch: 100.

3 Results

In this work, each image obtained after applying the image enhancement process as well as the images obtained by window-based features, are considered as a features to generate a set of patterns that represent the MCs and the normal tissue. Each pattern is constructed from gray level intensity of pixels of the obtained images, each pattern generated is called features vector, where these features vector represents a point in d-dimensional space. It is known each of the images used there are pixels belonging to MCs and to normal tissue, then each analyzed pattern belong to one of two possible classes, i.e. there are patterns that belong to the set Q_1 if correspond to MCs and patterns belonging to normal tissue Q_0.

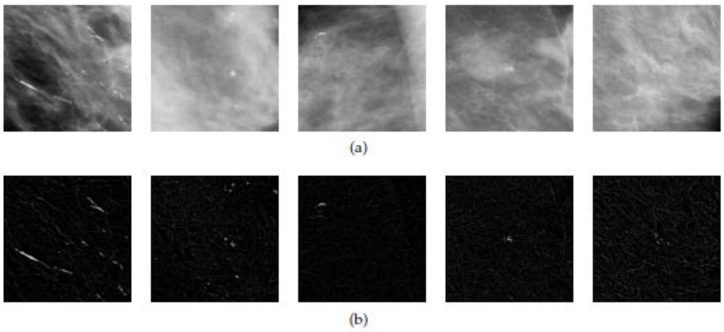

Fig. 1. (a)Original ROI images. (b) ROIs images processed by Top-Hat transform

In Table 1 shows the number of patterns assigned to classes Q_0 and Q_1 obtained.

Table 1. Number of patterns assigned to Q_1 and Q_0

Label	Number of patterns by k-means	Number of patterns by SOM
Q_0	588181	583081
Q_1	*1643*	6743

Due to the large number of patterns of the class that do not belong to MCs with respect to the number of patterns that belong to the class of MCs a balancing was performed, see Table 2, The network parameters such as network size and architecture (number of nodes, hidden layers etc), and gain parameters were kept the same. For all cases the neural network had one hidden layer with eight hidden nodes. In order to determine the network structure and metaplasticity parameters, the same network parameters applied in [10] and [11] were used.

Table 2. Results of balancing

Label	Number of patterns by k-means	Number of patterns by SOM
Q_0	8215	33715
Q_1	*1643*	6743

Network structure used in the experiments: Number of input neurons equal to the number of attributes of the records in the database (plus the bias input).

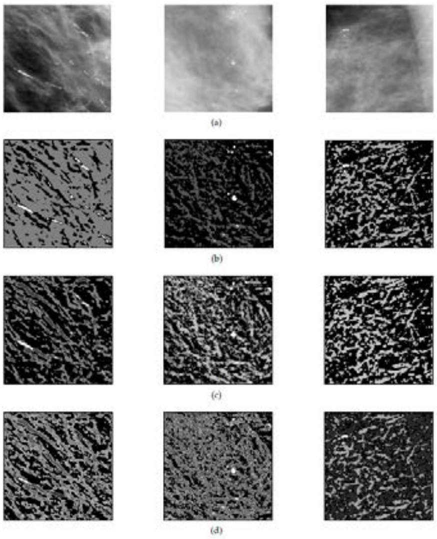

Fig. 2. Labelling of FV by SOM. (a)Original ROIs. (b)The obtained results of the 3rd partition. (c)The obtained results of the 4th partition. (d) The obtained results of the 6th partition

Table 3 shows the metaplasticity parameters A and B and the best network structures. Two different criterions to decide for the better network structure and metaplasticity parameters are considered:

1. Metaplasticity parameters: fixing a number of neurons in the hidden layer sufficiently high to presume that the ANN has sufficient processing units to perform the classification, begin to vary the metaplasticity parameters starting with A and finally with parameter B, until we achieve the mentioned value (MSE $\approx$ 0.01) in the minimum number of iterations.
2. Number of neurons in hidden layers: We vary the number of neurons in hidden layers until we achieve the Mean Squared Error (MSE) of approximately 0.01 (metaplasticity parameters are not changed) with the minimum number of neurons without degrading final performance.

Table 3. The best network structures and metaplasticity parameters

Data Set FV$_S$	Network Structure			Metaplasticity Parameters		Mean Squared Error
	I	H	O	A	B	
k-means	4	1	1	3	0	0
	4	1	1	3	0	0
SOM	4	1	1	3	0.25	0
	4	1	1	3	0	0

In this work different network structures were used; the activation function for all neurons is sigmoidal with scalar output in the range (0,1); and with the same metaplasticity parameters. A confusion matrix is built to determine the probability of the detection MCs vs. probability of false MCs. Table 4 shows the performance of the classifiers presented in this work.

Table 4. Confusion matrices and performance of the classifiers

Classifier AMMLP	Desired Results	Output Results		Sensitivity (%)	Specificity (%)	Total Classification Accuracy (%)
		MCs	Normal Tissue			
			k-means			
Structure1 4 : 15 : 1	MCs	483	0	100	99.67	99.72
	Normal Tissue	8	2466			
Structure2 4 : 12 : 1	MCs	476	7	98.55	99.63	99.45
	Normal Tissue	9	2465			
			SOM			
Structure3 4 : 15 : 1	MCs	1328	2	99.84	99.95	99.93
	Normal Tissue	3	6758			
Structure4 4 : 10 : 1	MCs	1317	13	99.02	99.94	99.78
	Normal Tissue	4	6757			

4 Conclusions

In this work two clustering algorithms, k-means and Self Organizing Maps, in order to detect MC clusters in digitized mammograms were used. Clustering algorithms help us to get a better comprehension and knowledge of data with the objective of segmenting the image into different areas (background and MC). After a learning process, the partitional clustering algorithms provide a set of centroids as the most representative elements of each group. As such, clustering algorithms partition the input images in homogeneous areas, each of which is considered homogeneous with respect to a property of interest.

Before applying the clustering algorithms, we applied a digital image processing technique as the image enhancement using mathematical morphology operations in order to improve the contrast between the MC clusters and the background in the ROIs. The mathematical morphology operations based on Coordinate Logic Filters given that the main features of this filters makes that their implementation is simple and fast hardware implementation, although in this work only simulation was carried out. Our methodology was tested in different ROIs images, with different kinds of tissue and different shapes of the MCs.

The experimental results show that the proposed method can locate Clusters of Microcalcifi- cations in an efficient way, moreover the method promises interesting advances in Medical Industry.

References

1. N. Pal, B. Bhowmick, S. Patel, S. Pal, and J. Das: A multi-stage neural network aided system for detection of microcalcifications in digitized mammograms. Neurocomputing, 71 (13-15):2625–2634, 2008
2. L. Wei, Y. Yang, and R. Nishikawa: Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. Pattern Recognition, 42(6):1126 – 1132, 2009
3. A. Papadopoulos, D. Fotiadis, and L. Costaridou: Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques. Computers in Biology and Medicine, 38(10):1045 – 1055, 2008
4. M. Bhattacharya and A. Das: Fuzzy logic based segmentation of microcalcification in breast using digital mammograms considering multiresolution. International Machine Vision and Image Processing Conference, pages 98–105, 2007
5. Y. Sung-Nien, L. Kuan-Yuei, and H. Yu-Kun: Detection of microcalcifications in digital mam- mograms using wavelet filter and markov random field model. Computerized Medical Imaging and Graphics, 30(3):163–173, 4 2006
6. R. Guzman-Cabrera, J. Guzman-Sepulveda, M. Torres-Cisneros, D. May-Arrioja, J. Ruiz-Pinales, O. Ibarra-Manzano, G. Avina-Cervantes, and A. Gonzalez-Parada: Digital image processing technique for breast cancer detection. International Journal of Ther- mophysics, 34(8):1519–1531, 2013. doi: 10.1007/s10765-012-1328-4
7. R. Guzman-Cabrera, J. R. Guzman-Sepulveda, M. Torres-Cisneros, D. May-Arrioja, J. Ruiz- Pinales, O. Ibarra-Manzano, and G. Avina-Cervantes: Pattern recognition in photoacoustic dataset. International Journal of Thermophysics, 34(8):1638–1645, 2013
8. R. Guzman-Cabrera, J. Guzman-Sepulveda, A. Gonzalez-Parada, J. Rosales-Garcia, M. Torres- Cisneros, and D. Baleanu: Digital processing of thermographic images for medical applications. Revista de Chimie, 67(1):53–56, 2016
9. J. Suckling, J. Parker, and D. Dance: The mammographic image analysis society digital mam- mogram database. Exerpta Medica International Congress Series., 1069:375–378, 1994
10. A. Marcano-Ceden˜ o, J. Quintanilla-Domínguez, and D. Andina: Wood defects classification using artificial metaplasticity neural network. 35th Annual Conference of the IEEE In- dustrial Electronics Society, IECON, pages 3422–3427, 2009
11. D. Andina, A. Álvarez-Vellisco, A. Jevtic´, and J. Fombellida: Artificial metaplasticity can im- prove artificial neural network learning. In Intelligent Automation and Soft Computing,15(4):681–694, 2009

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
noviembre de 2016
Printing 500 / Edición 500 ejemplares