Applications of Language & Knowledge Engineering

Research in Computing Science

Research in Computing Science, Año 19, Volumen 149, No. 4, abril de 2020, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: 1870-4069, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 30 de abril de 2020.

Research in Computing Science, year 19, Volume 149, No. 4, April 2020, is published monthly by the Center for Computing Research of IPN.

Applications of Language & Knowledge Engineering

Beatriz Beltrán
Yolanda Moyao
Andrés Vázquez
David Pinto (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2020

ISSN: 1870-4069

Formerly ISSN: 1665-9899.

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

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Editorial

This volume of the "Research in Computing Science" journal contains selected papers related to the topic of Language and Knowledge Engineering and their applications, with topics covering natural language processing, computational linguistics, knowledge engineering, pattern recognition, artificial intelligence and other paper covering computer science in general.

Language engineering is an area of artificial intelligence and applications aiming to bridge the gap between traditional computational linguistics research and the implementation of potentially real-world applications. Knowledge engineering, on the other hand, refers to all technical, scientific and social aspects involved in designing, building, maintaining and using knowledge-based systems.

The papers published in this special issue were carefully chosen by the editorial board based on 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 60%).

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).

Beatriz Beltrán
Yolanda Moyao
Andrés Vázquez
David Pinto
Guest Editors
Benemérita Universidad Autónoma de Puebla,
LKE-FCC-BUAP, Mexico

February 2020

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Statistically Based Authorship Identification

Angel Kuri Morales

Instituto Tecnológico Autónomo de México,
Mexico

akuri@itam.mx

Abstract. Presently very large volumes of information are being regularly produced in the world. Most of this information is unstructured, lacking the properties usually expected from relational databases. One interesting issue in computer science is how we may achieve data mining on such unstructured data. Intuitively, its analysis has been attempted by devising schemes to identify patterns and trends through means such as statistical pattern learning. The basic problem of this approach is that the user has to decide, a priori, the model of the patterns and, furthermore, the way in which they are to be found in the data. This is true regardless of the kind of data, be it textual, musical, financial or otherwise. In this paper we explore an alternative paradigm in which raw data is categorized by analyzing a large corpus from which a set of categories and the different instances in each category are determined, resulting in a structured database. Then each of the instances is mapped into a numerical value which preserves the underlying patterns. This is done using CESAMO, a statistical algorithm. Every categorical instance is then replaced by the adequate numerical code. The resulting numerical database may be analyzed with the usual clustering algorithms. In this work we exemplify with a textual database and apply our method to characterize texts by different authors and present experimental evidence that the resulting databases yield clustering results which permit authorship identification from raw textual data.

Keywords. Databases, encoding, statistics, optimization.

1 Introduction

The problem of analyzing large sets of unstructured data sets has grown in importance over the last years. As of June 2019, there were more than 1,690,000 web sites in the world [1]. Due to corrections to Metcalfe's law [2] (which states that the value of a telecommunications network is proportional to $n\ (log_2\ n)$ of the number of connected users of the system) there is a network world value of O(35e+08). The associated amount of data generated may be inferred from this number and even conservative metrics yield very large estimates. Most of these data are unstructured and recent commercial [3] approaches to the problem attest to the increasing importance assigned to this fact.

In the computer science community, data mining of texts [4], music [5] and general information [6, 7] is being approached with growing interest. In the vast majority of the cases, information extraction is highlighted and emphasizes the use of anaphora: the use of an expression the interpretation of which depends upon another expression in context. This approach, although intuitive and natural, suffers from the obvious disadvantage of being case-based. That is, a method developed for, say, texts in English will not be directly applicable to other languages and much less to other kinds of information. For example, even when limiting our range of study to texts stemming from corporate finance (i.e. mining industry literature for business intelligence), "horizontal" test mining (i.e. patent analysis) or life sciences research (i.e. mining biological pathway information) every one of the lines just mentioned relies on a case-by-case determination of the anaphoric usage.

The problem at the very heart of this issue is the fact that we must preserve the patterns underlying the information and it had not been treated with success in the past. In [8], however, a method to encode non-numerical data in mixed (numerical and categorical) databases was shown to be effective in preserving the embedded patterns. To do this we must consider the fact that there is a limited subset of codes which will preserve the patterns in those databases consisting of both numerical and non-numerical data (i.e. *mixed* databases or MD). To identify such pattern-preserving codes (PPC) we appeal to a statistical methodology. It is possible to statistically identify a set of PPCs by selectively sampling a bounded number of codes (corresponding to the different instances of the CAs) and demanding the method to set the size of the sample dynamically. Two issues have to be considered for this method to be defined in practice: a) How to set the size of the sample and 2) How to define the adequateness of the codes.

A note is in order: the PPCs are NOT to be assumed as an instance applicable to DBs other than the original one. That is to say: a set of PPCs (say PPC1) obtained from a DB (say DB1) is not applicable to a different DB (say DB2) even if DB1 and DB2 are structurally identical. In other words, PPC1 ≠ PPC2 for the same DB when the tuples of such DB are different.

Consider a set of *n*-dimensional tuples (say *U*) whose cardinality is *m*. Assume there are *n* unknown functions of *n-1* variables each, which we denote with $f_k(v_1,...,v_{k-1},v_{k+1},...,v_n)$;$k=1,...,n$.

Let us also assume that there is a method which allows us to approximate f_k (from the tuples) with F_k. Denote the resulting *n* functions of *n-1* independent variables with F_i, thus:

$$F_k \approx f(v_1,...,v_{k-1},v_{k+1},...,v_n);\ k=1,...,n\ . \tag{1}$$

The difference between f_k and F_k will be denoted with ε_k such that, for attribute *k* and the *m* tuples in the database:

$$\varepsilon k = \max\left[\, abs(\,fki - Fki)\right];\ i=1,...,m\ . \tag{2}$$

Our contention is that the PPCs are the ones which minimize ε_k for all *k*.

This is so because only those codes which retain the relationships between variable *k* and the remaining *n-1* variables AND do this for ALL variables in the ensemble will preserve the whole set of relations (i.e. patterns) present in the database, as in (3):

$$\Xi = min[\ max\ (\varepsilon k\ ;\ k=1,...,n)]\ . \quad (3)$$

Notice that this is a multi-objective optimization problem because complying with condition k 0in (2) for any given value of k may induce the non-compliance for a different possible k. Using the min-max expression of (3) equates to selecting a particular point in Pareto's front [9]. To achieve the purported goal we must have a tool which is capable of identifying the F_k's in (1) and the codes which attain the minimization of (3).

For this purpose we designed a new algorithm (called "CESAMO": Categorical Encoding by Statistical Applied Modeling) which relies on statistical and numerical considerations.

1.1 CESAMO Algorithm

In what follows we denote the number of tuples in the DB by t and the number of categorical attributes by c; the number of numerical attributes by n; the i-th categorical variable by vi; the value obtained for variable i as a function of variable j by $yi(j)$.

We will sample the codes yielding yi as a function of a sought for relationship. This relationship and the model of the population it implies will be selected so as to preserve the behavioral patterns embedded in the DB.

Two issues are of primordial importance in the proposed methodology:

a) How to determine the number of codes to sample.

b) How to define the function which will preserve the patterns.

Regarding (b), we use a mathematical model considering high order relations. Regarding (a), we know that the distribution of the means of the samples of yi (yi_{AVG}) will become Gaussian. Once it becomes Gaussian we know that further sampling of the yi's will not significantly modify the characterization of the population.

The general algorithm for CESAMO is as follows:

- Specify the mixed database MD.
- Specify the sample size (ss).
- MD is analyzed to determine n, t and $ci(i)$ for $i=1,\ldots,c$.
- The numerical data are assumed to have been mapped into [0,1). Therefore, every ci will be, likewise, in [0,1).

```
for i ← 1 to c
   Do until the distribution of yiAVG  is Gaussian
        Randomly select variable j (j ≠i)
        Assign random values to all instances of vi.
        yiAVG ← 0
        For k ←1 to ss
             yi ←  f(vj)
             yiAVG ← yiAVG+yi
        endfor
```

```
10.        yi_AVG = yi_AVG/ss
11.    enddo
12.    Select the codes corresponding to the best value of yi
13.  endfor
```

Hence, we sample enough codes to guarantee the statistical stability of the values calculated from $yi \leftarrow f(vj)$. The codes corresponding to the best approximation will be those inserted in MD. CESAMO relies on a double level sampling: only pairs of variables are considered and every pair is, in itself, sampling the multivariate space. This avoids the need to explicitly solve the multi-objective optimization underlying problem. The clustering problem may be, then, numerically tackled.

Notice that *vj* may be, itself, categorical. In that cases every categorical instance of *vj* is replaced by random codes so that we may calculate $f(vj)$.

One of the key points of CESAMO is how to define the functional relation specified in step 7 (i.e. $yi \leftarrow f(vj)$). This selection defines the way in which our intent of preserving the patterns in the data is understood. In our experiments we set $yi \leftarrow P_{11}(x) \sim \beta_0 + \sum_{i=1}^{6} \beta_i x^{2i-1}$ as a universal polynomial approximation. In [10] it was shown that any continuous function may be approximately realized with a linear combination of monomials which has a constant plus terms of odd degree. In the case of $P_{11}(x)$ the relationships are not limited *a priori*. The β_i of $P_{11}(x)$ were found with the so-called *Ascent Algorithm* (AA) [11]. The codes obtained from AA are called *functional* codes.

An example of a Mixed Database and the resulting Numerical Database after CESAMO is shown in Figure 1. All numerical values have been mapped into [0,1]. The corresponding codes are shown in Figure 2.

Once having shown that CESAMO does find the correct codes for categorical attributes in mixed databases any classic numerical clustering algorithm (such as Fuzzy C-Means [12] or SOMS [13]) may be used and, furthermore, any text (indeed, any collection of tokenizable data) may be treated as a set of categorical variables provided categories and their corresponding instances are identified.

The rest of the paper is organized as follows. In section 2 we describe the method we applied to tokenize English texts. In section 3 we describe the process of encoding the tokenized database to obtain the corresponding clusters. In section 4 we present an algorithm developed to identify the correspondence of the clusters in a labeled database. In section 5 we describe the experiments we performed to test the viability of our method. In section 6 we present our conclusions.

2 Tokenizing English Texts

To tokenize unstructured data, we first have to find a representative sample which adequately characterizes the universe of data (*U*). Once having done so we select the number of categories (*c*). The next step is to determine the instances within every category (*t*). The number of tokens (*k*) per tuple determines the form of the structured database. The final step is to select the text to tokenize and identify the tokens which

Age	Place	Educa-tion	Race	Sex	Income
55	North	9	White	M	2932.49
62	North	7	Asian	F	23453.37
57	South	24	Indian	F	1628.61
56	Center	18	White	M	4069.62
49	South	22	Indian	F	3650.23

Age	Place	Educa-tion	Race	Sex	Income
0.4928	0.0002	0.2000	0.8304	0.1332	0.0226
0.5942	0.0002	0.1200	0.0668	0.1283	0.1896
0.5217	0.2209	0.8000	0.4084	0.1283	0.0120
0.5072	0.2691	0.5600	0.8304	0.1332	0.0318
0.4058	0.2209	0.7200	0.4084	0.1283	0.0284

Fig. 1. Example of Mixed and Numerical Data after CESAMO.

North	0.0002
South	0.2209
Center	0.2691
White	0.8304
Asian	0.0668

Indian	0.4084
Other	0.7472
M	0.1332
F	0.1283

Fig. 2. Code for categorical instances.

Category	Instance
IM	IM
IM	IN
IM	JU
IM	LI
LU	LU
LU	ME
LU	MO
LU	NI

Category	Instance
OP	OP
OP	PA
OP	PE
OP	PO
PR	PR
PR	AN
PR	RE
PR	RE

Fig. 3. An example of categories and instances.

will populate the structured database. Notice that these steps are independent of the nature of *U*.

2.1 Tokenizing the Universe of Data

We started by selecting a collection of English texts from different authors: James Joyce, Carl Sagan, Lord Byron, William Shakespeare and English translations of Gabriel García Márquez and Julio Cortázar. 125,882 words were extracted, from which 15,031 distinct ones were identified (i.e. $|U| = 15{,}301$). We then made c=12 and t=4.

This meant that the words in *U* were separated into 48 intervals. Every interval consists of ≈313 words which were ordered alphabetically. A category, therefore, has 1,252 different words. Categories "IM" and "LU"; "OP" and "PR" are illustrated in Figure 3.

We say that a word starting with an instance in the (now) structured database is a token for the category. For example, the English word "*and*" is a token in category "PR". Under this classification, the sentence "*I will take a bus at twelve and arrive later tonight*" becomes the set of tokens <TE> <EN> <AU> <A> <JU> <DW> <VI> <LI> <OP> <JU> <SU>. A final specification consists of assigning a number of tokens to every tuple in the structured database. We set k=10. Hence, a "sentence" (which we will

Table 1. Structure of a zentence.

Zentence No.	Token No.									
1	1	2	3	4	5	6	7	8	9	10
2	1	2	3	4	5	6	7	8	9	10

AN	nul2	CO	nul4	EN	GR	LI	nul8	nul9	QU	SE	TE
AL	nul2	CO	DI	DW	nul6	LI	nul8	nul9	QU	RO	TR
A	nul2	CO	CR	ES	FA	nul7	ME	nul9	nul10	SU	TE
nul1	nul2	nul3	DE	EN	HE	LI	LÚ	nul9	RE	nul11	TE
A	nul2	nul3	DE	nul5	nul6	LI	MO	nul9	QU	SI	TE
nul1	AR	CI	DE	nul5	FR	JU	nul8	PO	nul10	nul11	nul1
A	nul2	nul3	DE	DW	nul6	JU	NI	PE	RE	nul11	TE

Fig. 4. A segment of a tokenized database.

call "zentence") has a fixed size of 10 tokens. A zentence has the structure illustrated in Table 1.

2.2 Obtaining the Tokenized Database

Not all zentences include instances of all categories and, therefore, a special token denoted by "nul_i" may appear in the *i-th* position of the zentence. That is, if no representative of category 1 appears in the zentence, it will be filled-in with nul1; if no representative of category 2 appears it will be filled-in with nul2 and so on. In other words, there are $(c+1)t$ symbols present in the database. In our case, therefore, there are up to 60 different categorical instances present in the database. This is illustrated in Figure 4.

Once U is determined and categorized any given selected text in English may be mapped into a relational database formed of zentences. We selected three texts by James Joyce and three by Carl Sagan and tokenized them accordingly. These texts were then encoded by CESAMO and finally clustered using Self-Organizing Maps. The resulting clusters were labeled and then tested for similarity. Clusters whose tuples are similarly labeled indicate the same author, different authors otherwise.

3 Coding and Clustering the Tokenized Database

Once the tokenized database has been obtained, we proceed to encode the attributes, which correspond to the tokens determined as above. The tokenized database is one in which all attributes are categorical. We find a set of codes (one for each different instance of all categories) such that the structures present in the non-numerical data set are preserved when every instance of every category is replaced by its numerical counterpart. CESAMO is based on the premise that patterns are statistically preserved once the distribution of the average approximation error after encoding is Gaussian. Let us assume that we have a hypothetical *perfect* set of pattern preserving codes. Assume, also, that there are n attributes and p tuples. Given such *perfect* set it would be, in

Age	Place of Birth	Years of Study	Race	Sex	Salary
55	North	9	White	M	2,932.49
62	North	7	Asian	F	23,453.37
57	South	24	Indian	F	1,628.61
56	Center	18	White	M	4,069.62
49	South	22	Indian	F	3,650.23

Fig. 5a. A mixed database (MD).

North	Center	South	White	Asian	Indian	Other	M	F
0.0002	0.2691	0.2209	0.8304	0.0668	0.4084	0.7472	0.1332	0.1283

Fig 5b. Instances of MD and possible codes.

Age	Place of Birth	Years of Study	Race	Sex	Salary
0.4928	0.0002	0.2000	0.8304	0.1332	0.0226
0.5942	0.0002	0.1200	0.0668	0.1283	0.1896
0.5217	0.2209	0.8000	0.4084	0.1283	0.0120
0.5072	0.2691	0.5600	0.8304	0.1332	0.0318
0.4058	0.2209	0.7200	0.4084	0.1283	0.0284
0.0870	0.0002	0.8400	0.0668	0.1332	0.2306

Fig 5c. Numerical database (ND) with codes from Figure 2b.

principle, possible to express attribute *i* as a function of the remaining *n-1* with high accuracy since this *perfect* code set will lend itself to a close approximation.

Every tuple in the DB consists of a set of possible codes which are to be assigned to every instance in the database. An encoded tuple for the database illustrated in Figure 5a is shown in Figure 5b. Figure 5c illustrates the database resulting from replacing the instances of MD with the codes of Figure 5b. Numerical variables are mapped into [0, 1] so that all numbers in the DB lie in the same range. This strategy guarantees that the resulting set of codes corresponds to the best global behavior.

That is, the final set of codes encompasses the best combinations of the f_i's minimizing the approximation error and the multi-objective optimization is solved Since the codes are arrived at from a stochastic process any two runs of CESAMO will result in different sets of codes, say S1 and S2.

This fact allows us to verify that, as postulated, patterns will be preserved. This is done by applying a clustering algorithm which yields an assignment of every tuple to one of *m* clusters. Under the assumption of pattern preservation, clustering with S1 and S2 should yield analogous clusters. This is, indeed, the fact, as was shown in [8].

4 Identification of Cluster Matching in Labeled Databases

The correct identification of analogous clusters is compulsory if, as intended, we are to determine whether two texts correspond (or not) to the same author. Texts T1A and T2A both authored by A should correspond to similar clusters whereas texts T1A and

												LABEL SET 1			
V1	V2	V3	V4	V5	V6	V7	V8	V9	VA	VB	VC	C11	C12	C13	C14
0.513	.41	.39	.37	.00	.67	.83	.03	.83	.57	.83	0.007	1	0	0	1
0.513	.41	.39	.81	.51	.19	.27	.08	.36	.95	.83	0.288	0	1	0	0
0.513	.41	.11	.81	.51	.11	.56	.09	.83	.07	.63	0.028	0	1	0	0
0.513	.16	.89	.20	.00	.11	.27	.03	.36	.57	.63	0.288	0	0	0	1
0.513	.41	.39	.20	.51	.19	.56	.03	.36	.57	.63	0.606	0	1	0	0
0.160	.64	.11	.20	.51	.11	.56	.08	.36	.95	.63	0.028	0	0	0	1
0.160	.41	.16	.20	.87	.11	.27	.09	.01	.03	.83	0.028	1	0	0	0
0.284	.92	.11	.20	.08	.19	.27	.41	.95	.95	.39	0.288	0	0	0	1
0.160	.41	.11	.81	.00	.11	.56	.09	.95	.57	.36	0.288	1	0	0	0
0.513	.12	.16	.20	.51	.11	.56	.03	.95	.57	.83	0.007	0	0	0	1
0.160	.41	.39	.81	.51	.11	.56	.08	.01	.07	.36	0.007	0	1	0	0
0.513	.41	.11	.81	.51	.11	.27	.09	.95	.57	.16	0.288	0	1	0	0

Fig. 6a. A segment of labeled Numerical Database T1.

T1B (authored, respectively, by A and B) should correspond to different clusters. To test the purported cluster similarity poses the technical problem whose solution we describe in what follows.

4.1 The Problem of Cluster Matching

Assume that tables T1 and T2 consisting of attributes V1, ..., VC have been classified into 4 clusters and labeled as illustrated in Figures 6a and 6b. This labeling convention is convenient since one may easily count the number of matches between two clusters. However, clustering algorithms do not necessarily yield the same order of the label columns.

For example, in Figure 7 we have compared column C11 to C21, on the one hand and C11 to C24 on the other. The first choice yields a count of 6 matches leading us to the conclusion that sets C11 and C21 do not match and that, therefore, T1 and T2 do not share the same clusters.

The second choice, however, yields the full 12 matches.

Therefore, in this instance one must conclude that column C11 (from set 1) actually corresponds to column C24 (from set 2). Accordingly, we should also conclude that T1 and T2 correspond to the same set for cluster 1. The correct pairing has to be achieved in similar fashion for all clusters.

If there are m clusters and p tuples there are m^p possible combinations of valid labeling sets. We need to investigate which of these does actually correspond to the proper matching of the m clusters in T1 with those of T2.

Only then we may compare T1 and T2 and determine their similarity. To achieve this identification, we designed the following algorithm.

4.2 Algorithm for Optimization of Cluster Matching

The proposed algorithm for optimization of cluster matching is as follows.

---

1. Create a matching table "MT" of dimensions $m \times m$.
   Make MT(i,j) $\leftarrow$ 0 for all i, j.
2. For $i \leftarrow 1$ to m
   - For $j \leftarrow 1$ to m
     - If column i = column j
       - MT(i,j) $\leftarrow$ *MT(i,j) +1*
     - endif
   - endfor

   endfor

   MT(i,j) will contain the number of matches between cluster i of table T1 and cluster j of table T2.
3. Create a table "Scores" of dimension Q (Q >>0).
4. For $i \leftarrow$ 1 to Q
   - 4.1. Set a random valid sequence S_i of m possible matching sequences between the clusters of T1 and those of T2.
   - 4.2. Find the number of matches M_i between T1 and T2 from table MT as per S_i.
   - 4.3. Make Scores(i) $\leftarrow M_i$.

   endfor
5. $I \leftarrow$ index of max(Scores(i)) for all i.
6. Select S_I. This is the matching set which maximizes the number of coincidences between the clusters of T1 and T2.

The core of the algorithm lies in step 4.1 where the valid matching sequences are determined. This algorithm will find the sequence which maximizes the number of matches between the clusters of T1 and T2 with high probability provided Q is large enough. In our experiments we made Q =1000. Given the large number of possible pairings between the clusters of T1 and T2 the algorithm is a practical way to select which cluster of T1 should be paired with which cluster of T2.

5 Experimental Authorship Identification

At this point we are in a position which allows us to test the initial hypothesis of authorship identification. As already stated, we selected 3 texts from James Joyce (JJ) and 3 from Carl Sagan (CS). These were, consecutively, tokenized, CESAMO-encoded, clustered with SOMs and labeled. Previous analysis led us to the conclusion that there were 4 clusters, as may be seen from the graph in Figure where we display the results of having trained texts from JJ and CS for up to 6 clusters with SOMs.

												LABEL SET 2			
V1	V2	V3	V4	V5	V6	V7	V8	V9	VA	VB	VC	C21	C22	C23	C24
0.053	.32	.06	.02	.04	.27	.53	.05	.08	.69	.75	0.852	1	0	0	1
0.053	.32	.06	.02	.47	.27	.48	.05	.08	.29	.20	0.852	0	0	1	0
0.717	.32	.02	.56	.04	.27	.48	.05	.31	.31	.58	0.295	0	0	1	0
0.017	.32	.06	.02	.47	.65	.53	.13	.08	.32	.01	0.260	1	0	0	0
0.053	.32	.02	.02	.47	.26	.48	.93	.08	.29	.75	0.295	0	0	1	0
0.717	.09	.06	.56	.04	.27	.48	.05	.36	.69	.01	0.295	1	0	0	0
0.017	.20	.06	.56	.04	.26	.48	.13	.03	.29	.01	0.852	0	0	0	1
0.017	.32	.06	.03	.01	.65	.48	.08	.31	.29	.29	0.852	1	0	0	0
0.053	.20	.06	.56	.01	.65	.48	.08	.36	.69	.58	0.295	0	0	0	1
0.053	.20	.06	.56	.12	.26	.48	.05	.36	.29	.01	0.852	1	0	0	0
0.053	.32	.02	.02	.01	.65	.48	.00	.31	.29	.20	0.852	0	0	1	0
0.017	.32	.06	.03	.47	.65	.53	.93	.08	.29	.75	0.260	0	0	1	0

Fig. 6b. A segment of labeled Numerical Database T2.

C11	C21	Same
1	1	1
0	0	1
0	0	1
0	1	0
0	0	1
0	1	0
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	0	1

C11	C24	Same
1	1	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
1	1	1
0	0	1
1	1	1
0	0	1
0	0	1
0	0	1

Fig. 7. Similarity for different choices of cluster columns.

The maximum errors relative to the mean were smallest for 4 clusters and the calculated standard deviation with a 0.05 *p-value* [14] was, likewise, smallest for the same number. The texts we selected were roughly the same size (≈16,000 words). They were then tokenized and CESAMO-encoded. The resulting databases were clustered and labeled. Next cluster matching was performed and adjusted when needed. We then proceeded to obtain a matrix of coincidences for the 15 possible combinations of the 6 texts. These are shown in Table 2. JJ_i denotes the *i-th* text by James Joyce; likewise CS_i denotes the texts by Sagan.

The texts pairings were ordered according to the percentage of matches we obtained. We observed the following behavior:

- All matching text matches were higher when the authors were the same, with the exception of item 6, where the texts by JJ vs CS had higher matches than expected.
- The correct assessment of authorship matches for the first 5 couples remains very close or above 75%. Therefore, a matching percentage of 75% appears to be sufficient to ascertain similar authorship.

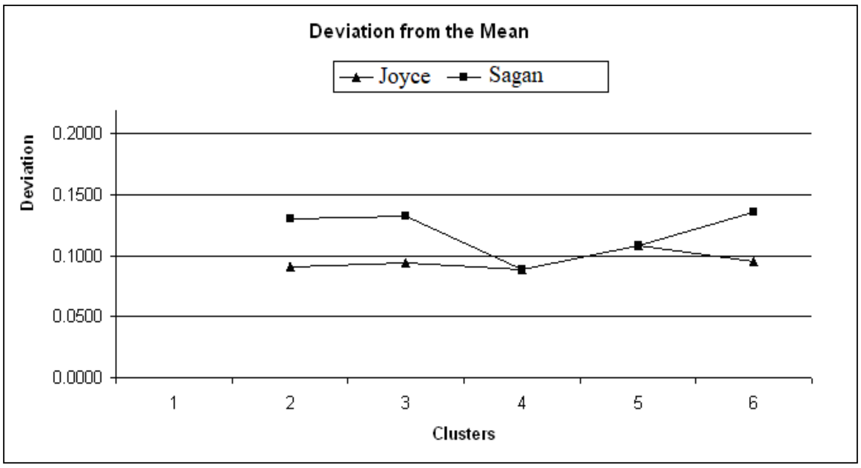


Fig. 8. Standard deviation for training errors.

Table 2. Comparison of clusters obtained.

	Text 1	**Text 2**	**Cluster Matches %**
1	CS1	CS2	98.23 %
2	CS2	CS3	86.43 %
3	CS1	CS3	82.50 %
4	JJ1	JJ2	78.54 %
5	JJ1	JJ3	74.69 %
6	JJ3	CS3	68.60 %
7	JJ2	JJ3	66.77 %
8	JJ1	CS3	65.71 %
9	JJ2	CS1	57.14 %
10	JJ3	CS2	57.14 %
11	JJ1	CS2	54.29 %
12	JJ2	CS3	54.29 %
13	JJ3	CS1	54.29 %
14	JJ1	CS1	51.52%
15	JJ2	CS2	48.57 %

– There appears to be no possible definitive conclusions of the purported authorship in the borderline percentages for items 6-8.

- Matching percentages below 60% seem to imply negative authorship for the analyzed couples.
- The identification percentage falls smoothly so that there is not a clear cut threshold dividing correctly assessed authorship from the alternative.

6 Conclusions

We have described a method to identify the authorship of selected English texts which is not based on linguistic considerations. It relies on the identification and preservation of the patterns embedded in the texts by the intelligent encoding of the data. There are several parameters which were heuristically determined and have to be further explored: e.g. the size of the zentences, the number of categories and the corresponding instances, the selected texts and their lengths. Setting them after systematic experimental tests might improve the algorithm significantly. Finally, the results, which seem to be promising, are only valid if we assume that the method will behave similarly if the restricted number of authors we selected were to be expanded. Experimental work remains to be done.

At any rate this seems to be a promising and novel alternative; particularly in view of the fact that, as pointed out in the introduction, it may be applied to any kind of unstructured data. We expect to report on the application of our method to more general and non-textual data in the near future.

References

1. Internetlivestats: https://www.internetlivestats.com/total-number-of-websites/ (2019)
2. Odlyzko, A., Tilly, B.: A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections. Manuscript (2005)
3. IBM: http://www-03.ibm.com/press/us/en/pressrelease/46205.wss (2015)
4. Tan, A.H.: Text mining: The state of the art and the challenges. In: Proceedings of the PAKDD, Workshop on Knowledge Discovery from Advanced Databases, 8(65) (1999)
5. Pachet, F., Westermann, G., Laigre, D.: Musical data mining for electronic music distribution. In: Web Delivering of Music, Proceedings, First International Conference on IEEE, pp. 101–106 (2001)
6. Chen, L., Sakaguchi, S., Frolick, M.N.: Data mining methods, applications, and tools (2000)
7. Feldman, R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press (2007)
8. Kuri-Morales, A.F.: Minimum Database Determination and Preprocessing for Machine Learning. In: Innovative Solutions and Applications of Web Services Technology, IGI Global, pp. 94–131 (2019)
9. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. In: IEEE transactions on evolutionary computation, 6(2), pp. 182–197 (2002)
10. Kuri-Morales, A., Cartas-Ayala, A.: Polynomial multivariate approximation with genetic algorithms. In: Canadian Conference on Artificial Intelligence, Springer, Cham, pp. 307–312 (2014)

11. Cheney, E.W.: Introduction to approximation theory (1966)
12. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy C-means clustering algorithm. Computers & Geosciences, 10(2), pp. 191–203 (1984)
13. Kohonen, T.: Self-organizing maps. Springer Science & Business Media, 30 (2001)
14. Westfall, P.H., Young, S.S.: Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons, 279 (1993)

ISSN 1870-4069

Approaches to EEG-based Brain-Computer Interfaces: A Survey

Rafael Gallardo García, Beatriz Beltrán Martínez, Darnes Vilariño Ayala, Rodolfo Martínez, Meliza Contreras González

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

rafael.gallardo@alumno.buap.mx
{bbeltran,darnes,beetho,mcontreras}@cs.buap.mx

Abstract. This survey presents a new classification of the *state of the art* approaches to electroencephalography-based brain-computer interfaces, resulting in twelve EEG-based BCI paradigms. A historical overview of the BCIs was included as complementary information. This work also presents the corresponding description and features of each approach. Also, this survey presents a comparison of the most popular electroencephalography recording methods, highlighting their advantages and showing their disadvantages. In order to clarify the similarities and differences of the analyzed approaches and systems, two comparative tables are presented, one for the paradigms and one for the recording systems.

Keywords. Brain-computer interfaces, electroencephalography, paradigms.

1 Introduction

Brain-computer interfaces (BCI) are real-time computer-based systems that translate brain signals into predefined and useful commands that can improve the human-computer interaction as well as improve the communication with other people [10,17]. BCIs are commonly developed to provide an alternative communication tool for people with severe neuromuscular disorders such as amyotrophic lateral sclerosis, spinal cord injury, and brainstem stroke [10]. There are several methods to acquire useful data from the brain, those methods could be invasive or non-invasive. Some examples of the invasive methods are the electrocorticography (ECoG) and the microelectrode arrays (MEAs). On the other side, non-invasive methods are the widely used due to its noninvasiveness, high temporal resolution, portability and reasonable cost [10,33], this category includes electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI) and near-infrared spectroscopy (NIRS), being the non-invasive EEG-based BCIs the objective of this survey, this approach is the most widely researched due to their minimal risk and the relative convenience of conducting studies and recruiting participants [17].

The purpose of this survey is to analyze the different approaches reported in the *state-of-the-art* brain-computer interfaces, as well as present an updated classification of these approaches, this work does not aim to present the results of the different paradigms, however the results of the papers that belongs to each paradigm were evaluated in depth. This survey is organized as follows: First, the *State of The Art* section presents a historical point of view. Second, an analysis of all the existing EEG-based paradigms is presented in the *Paradigms of EEG-based BCIs* section, based on the proposals of Cervantes et al. [4], Fernandez-Fraga et al. [6] and Hwang et al. [10], adding some updates and a summary table of paradigms and their corresponding description. Third, the section *Data Acquisition* presents a summary of the actual EEG recording methods used in EEG-based BCIs. The last sections contain *Conclusions* and *References*.

2 A Short History of Brain-Computer Interfaces

In 1988, Farwell and Donchin [5] presented the first P300-based BCI, they measured a positive potential in the EEG about 300 ms after the subject attended the target stimulus, this potentials serves as the control signal. The stimulus that elicit the P300 is detected by averaging the EEG responses to relatively rare presentations of the target stimulus interspersed with several non-target stimuli [6,21]. The Farwell and Donchin experiments consisted in a 6 x 6 matrix, that contained letters and other symbols, the subjects could select items from the matrix by the average response to the flash of the target item, which differed from the average responses to the other items [6]. Many P300-based studies have been published [3,10,16,20,21,26,28]. The P300 potentials can also be elicited by auditory stimuli and several research groups have explored this option [8,15,13].

Another approach to brain-computer interfaces explodes the sensorimotor rhythms (SMRs) as control signals in the BCI. This approach was first reported by Wolpaw et al. in 1991 [31]. SMR signals are μ and β oscillations, that are recorded over sensorimotor cortices, the μ and β oscillations change in amplitude with movement, imagined movement or preparation for movement [6]. SMR-based BCIs can control a cursor to hit targets on a screen or perform several computer-based tasks, users can learn how to control the SMR amplitudes by increasing or decreasing the amplitude of the rhythms. Many of the works with SMR-based BCIs consist in asking the user to generate specific mental states (by imagining physical movement), which is commonly called motor imagery (MI). Several research articles have been published around this approach [2,12,27,30,31].

Over time new approaches and techniques emerged, such as visually-evoked potentials (VEPs), steady-state visually evoked potentials (SSVEPs), error-related potentials (ErrPs), each of these having sub-categories with important results [10]. VEPs and SSVEPs are based in potentials that are elicited by visual content, the first approach to the analysis of these potentials was reported by Sutter E. [25] in 1992, and his publication was followed by many others in the area [1,11,20,24,23,29].

Successful brain-computer interfaces systems are commonly an hybrid between various paradigms, typically researchers try to eliminate the weaknesses of one approach by combining the strengths of various approaches.

In 2012, Spüler presents an hybrid BCI based on visually evoked potentials and improving the performance with an adaptation based on error-related potentials [24]. This decade shows a trend in the research of this hybrid systems can be seen: Yin et al. proposed a hybrid BCI that incorporates SSVEP into the P300 paradigm [32], also Li et al. reported an controlled wheelchair by combining P300 and SSVEP paradigms [14]. Most recent works in hybrid BCIs are interesting too, in 2019, Niknamian S. proposed a hybrid system between SSVEP and P300 to enhance the accuracy of a speller system [19]. Also, in 2019, Machado M. presented a hybrid system that combines visual and auditory stimuli [15]. In late 2019, Oralhan Z. published a paper with another hybrid method, an approach for a speller BCI based on P300 and SSVEP.

3 Paradigms of EEG-based BCIs

BCI systems may be classified in two main categories: endogenous and exogenous systems [4]. Endogenous systems are dependent of the user's ability to control their electrophysiological activity, sensorimotor rhythms (motor imagery or MI) based systems and Slow Cortical Potential (SCP) based systems belong to this category. MI systems are based on the imagery of performing motor actions to evoke signals similar to those observed in actual movement and SCP systems involve slow changes in voltage generated on the cerebral cortex, with a duration between 0.5 s and 10 s, SCP-based systems are also associated with movement. Endogenous systems require a period of intensive training [6]. The exogenous systems obtain the data from evoked related potentials (ERP), these systems depends on the electrophysiological activity triggered by external stimuli [6]. Exogenous systems are easily to master than endogenous systems and are based in the acquisition of data from evoked related potentials (ERP): P300 events, visual evoked potentials (VEP), steady-state visual evoked potentials (SSVEP) or auditory evoked potentials (AEP).

Also, BCI systems may be classified according to the experimental paradigm employed to elicit different kinds of brain activities [6,10]. Hwang et al. [10] proposed the following seven categories: motor imagery, visual P300, steady-state visual evoked potential (SSVEP), non-motor mental imagery, auditory, hybrid and other paradigms. Hwang et al. includes covert attention, motion-onset visual evoked potentials (MOVEP), flash onset and offset visual evoked potentials (FOVEP), and error related potentials in their ”other paradigms” section. Fernandez-Fraga et al. [6] proposed a little different classification with five categories: evoked potential by P300 events, visually evoked potential (VEP), steady-state visual evoked events (SSVEP) and auditory evoked potential (AEP). Table 1 summarizes all the paradigms and its corresponding description.

4 Data Acquisition

A functional EEG-based BCI requires reliable, robust and high-quality EEG recording systems. There are several varieties of recording methods, the standard recording uses wet electrodes. Wet electrodes use a conductive gel that maintains good electrode contact with the scalp, what provides an excellent EEG recording. Unfortunately, wet electrodes are not optimal or practical for long-term daily use, this kind of electrodes require

Table 1. Summary of the existing BCI paradigms. The table presents a short description of each paradigm. The listed paradigms are not necessarily different, some paradigms in the table are subsets of others but they have characteristics that must be distinguished.

Paradigm	Description
Motor imagery	Based on the imagery of performing motor actions to evoke signals in the brain. Commonly consists in the imagination of kinesthetic movements of several parts of the body. The origin of the signals depends on the imagined activity.
P300 Events	Works over event-related potentials evoked by infrequent and task-relevant stimuli. Some of the stimulus have a relationship with the intention of the subject. The potential appears around 300 ms after each stimulus. These signals are measured most strongly in the parietal lobe.
Visual P300	Visual P300 is a kind of P300 event-related potential, with the difference that the stimuli are strictly visual.
VEP	Visually evoked potentials are detected on the EEG after the presentation of the visual stimulus. These kind of responses are usually originate from the occipital cortex of the brain.
SSVEP	Steady-state visual evoked potential is a periodic brain response evoked by a special stimuli: repetitive presentation of flickering or reversing visual stimulus. These potentials are also detected on the EEG after the presentation of the visual stimulus.
Non-motor MI	Consists in mental imagery tasks excluding motor imagery tasks. Mental calculations, remember images or faces, internal singing or speech and spatial navigation are good examples of the stimuli in this paradigm.
AEP	Auditory event potentials are perceived after the presentation of the auditory stimulus, commonly sounds at different frequencies, when the subject concentrates on any of them, a potential of the same frequency as the stimulus is generated.
Covert attention	Covert attention is defined as paying attention without moving the eyes to the point of that location. These tasks often require participants to observe a number of stimuli, but attend only one.
MVEP	Motion onset/offset visual evoked potentials are visual evoked potentials related to global motion during a visual motion discrimination task. These tasks consist in the discrimination of onset or offset motion in the stimuli.
FVEP	Flash onset/offset visual evoked potentials are the signals generated by a flashing stimuli, such as digits or letters that are displayed on a screen. Subjects can shift their gaze to the flashing target to induce a FVEP.
ErrP	Error related potentials (ErrPs) are the responses generated by the brain when the subject recognizes an error during a task. ErrPs are widely studied for error correction or adaptation in BCIs.
Hybrid	This category includes the simultaneous use of more than two paradigms mentioned in this table.

a careful application; the gel is sometimes messy and needs periodic replenishment; the cap or other apparatus that hold the electrodes in their side may be uncomfortable, awkward or unattractive [17]. Furthermore, wet electrode-based EEG systems are susceptible to a variety of artifacts due to the non-brain activity, such as electromyographic signals (EMG), bodily movements or nearby electrical equipment [17].

Table 2. Summary of EEG recording methods. The first column indicates the commercial name of the recording system as well as its model, the second column describes the type of electrodes that the system uses, the third column presents a set of advantages taken from several papers where these systems were used and the fourth column presents disadvantages reported in the literature.

System	Type of electrode	Advantages	Disadvantages
BioSemi ActiveTwo	Active wet	Excellent EEG recording quality Up to 280 channels Does not limit electrode location Very low impedance	Uncomfortable Not optimal for long-term use Very expensive
Emotiv EPOC	Moistened felt pads	Semi-rigid support Fast electrode placement Low-cost 14 channels	Less accurate than wet methods Restricted electrode placement Susceptible to EMG signals
g.SAHARA	Active dry	Does not limit electrode location Similar quality to wet electrodes Easy to use 8-64 channels	Could be uncomfortable
g.SCARABEO	Wet	Excellent EEG recording Does not limit electrode location 8-64 channels	Messy conductive gel Sensible to noise Not optimal for long-term use
B-Alert X10	Wet	Good EEG recording quality Ambulatory Does not limit electrode location	Maximum 9 channels
Wearable Sensing DSI-Hybrid	Active dry	Easy to use Fast electrode placement Similar quality to wet electrodes Analyzes BOLD actibity	Non-cosmetic Could be uncomfortable Awkward
QUASAR DSI 10/20	Dry	Comfortable Ambulatory Long-term comfort 21 channels	Less accurate than wet methods Awkward Non-cosmetic

4.1 Electrode Types

Due to the weaknesses of the wet electrode-based EEG recordings, several wet alternatives as well as dry electrodes have appeared in the last decade. One of the most used

EEG recording system is the Emotive EPOC: a 14-channel system that uses moistened felt pads instead the common conductive gels, mounted on a semi-rigid support that allows a quick and comfortable placement of the electrodes, but is less accurate than the conventional supports. Emotive EPOC systems are relatively cheap in comparison with other recording systems. An alternative to wet electrodes is the g.SAHARA system, a dry electrode that consists of a set of 8 pins, which mounted in a conventional cap does not limit electrode locations and provide similar results to the systems that uses wet electrodes. The BioSemi ActiveTwo systems use active wet electrodes, and provide excellent EEG recording quality, these systems allow up to 280 channels. Nijboer et al. [18] reported that a 32 channel version produced a P300-based accuracy higher than the g.SAHARA and EPOC systems. The EPOC and g.SAHARA electrodes rely on low impedance resistive contact with the scalp, while the dry electrodes of the QUASAR systems use a hybrid combination of high impedance resistive and capacitive contact with the scalp [22]. Hairston et al. reported that the EPOC and QUASAR systems could produce uncomfortable pressure points and movement artifacts [7], they also reported that dry electrode-based systems can be more difficult to secure to the scalp, crating a trade-off between comfort and recording quality. Table 2 summarizes several EEG recording systems, classified by the type of electrodes they use.

4.2 Electrode Holders

On the other hand, the device that holds the recording electrodes on the scalp is extremely important in the long-term home use. An ideal electrode holder device should allows electrodes to be accurately positioned in the scalp, allowing every possible position and ensuring that the electrodes will be firmly placed, all of this without sacrificing comfort and being non-intrusive and cosmetic. ”Insecure electrode placement can lead to noise due to sudden changes in impedance (”electrode pops”) and variable placement can increase day-to-day variations in the EEG features used by a BCI”, McFarland and Wolpaw [17] say.

5 Conclusions

Although the main focus of this survey is to establish a new classification of the existing paradigms to build brain-computer interfaces, the systems that use invasive EEG-recordings were not treated in depth. The invasive techniques that uses epidural or subdural electrodes or intracortical microelectrodes have more disadvantages than advantages. Although these techniques offer more secure placement and better spatial resolution that non-invasive techniques, their invasive nature and increased costs and risks are not well justified since there are methods with zero-invasiveness that can produce comparable quality and target acquisition times [17], making them a more viable, economical and accurate option. In addition, these invasive methods have not yet demonstrated reliable long term stability [9].

Focusing on the non-invasive methods, which are primarily treated in this survey (due to its excellent invasiveness-quality-price ratio), the trend is that these systems sometimes sacrifice comfort for quality and vice versa, and sometimes sacrifice accessibility for quality and accuracy as in BioSemi systems (which are very expensive). The Emotiv EPOC systems are less accurate than conventional electrode holders and restricts the possible electrode placements in the scalp, what results in a higher susceptibility to EMG contamination but that disadvantages are compensated by their connectivity, comfort, cosmetic design and relatively low cost. Newer technology as the g.SAHARA dry electrodes can provide a similar results to those provided by wet electrodes by sacrificing comfort and aesthetic. The g.SCARABEO systems provide an excellent EEG recording, with a good channel range and permissive electrode location, but is slower to set up when comparing with other systems and the conductive gel could be messy, resulting in a non-suitable system for long therm use. The selection of any acquisition technique will depends on a balance of a variety of variables: target quality, noise tolerance, comfort, long-term capabilities and budget.

Most EEG-based BCIs use evoked potentials as the control signals, primarily P300 evoked potential, sensorimotor rhythms or steady-state visual evoked potentials. Even so, there are several interesting approaches such as non-motor mental imagery, auditory evoked potentials or error-related potentials. Hybrid systems that used two or more of these paradigms together seem to have good results since the researchers use different paradigms to eliminate weaknesses.

Better EEG-recording systems that provide stable high-quality signals, that are comfortable and easy to use will improve the actual brain-computer interfaces. On the other hand, better algorithms and signal processing techniques that improve the performance and the accuracy of the BCIs are needed. Dry electrode systems and new machine learning algorithms have considerable promise.

References

1. Ahmadi, S., Borhanazad, M., Tump, D., Farquhar, J., Desain, P.: Sensor tying, optimal montages for VEP-based BCI (2019)
2. Alimardani, M., Nishio, S., Ishiguro, H.: Brain-computer interface and motor imagery training: The role of visual feedback and embodiment. In: Evolving BCI Therapy (10 2018)
3. Aloise, F., Aricò, P., Schettini, F., Riccio, A., Salinari, S., Mattia, D., Babiloni, F., Cincotti, F.: A covert attention P300-based brain–computer interface: Geospell. Ergonomics 55, 538–51 (03 2012)
4. Cervantes, G.S., Martínez, N.H.M., Pérez, M.A.R.: Interfaz cerebro computadora para el posicionamiento de un robot virtual. In: XII Simposio Mexicano en Cirugía Asistida por Computadora y Procesamiento de Imágenes Médicas, México, DF (2008)
5. Farwell, L., Donchin, E.: Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials. Electroencephalography Clinical. Neurophysiolgy 70, 510–523 (01 1988)
6. Fernandez Fraga, S., Aceves-Fernandez, M., Pedraza Ortega, J.C.: EEG data collection using visual evoked, steady state visual evoked and motor image task, designed to brain computer interfaces (BCI) development. Data in Brief 25, 103871 (03 2019)
7. Hairston, W.D., Whitaker, K.W., Ries, A.J., Vettel, J.M., Bradford, J.C., Kerick, S.E., McDowell, K.: Usability of four commercially-oriented EEG systems. Journal of neural engineering 11(4), 046018 (2014)

8. Halder, S., Hammer, E.M., Kleih, S.C., Bogdan, M., Rosenstiel, W., Birbaumer, N., Kübler, A.: Prediction of auditory and visual P300 brain-computer interface aptitude. PloS one 8(2) (2013)
9. Homer, M., Nurmikko, A., Donoghue, J., Hochberg, L.: Sensors and decoding for intracortical brain computer interfaces. Annual review of biomedical engineering 15, 383–405 (07 2013)
10. Hwang, h.j., Kim, S., Choi, S., Im, C.H.: EEG-based brain-computer interfaces: A thorough literature survey. International Journal of Human-Computer Interaction 29 (12 2013)
11. Kapeller, C., Hintermüller, C., Abu-Alqumsan, M., Prückl, R., Peer, A., Guger, C.: A BCI using VEP for continuous control of a mobile robot. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5254–5257 (2013)
12. Kevric, J., Subasi, A.: Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system. Biomedical Signal Processing and Control 31, 398–406 (2017)
13. Klobassa, D., Vaughan, T., Brunner, P., Schwartz, N., Wolpaw, J., Neuper, C., Sellers, E.: Toward a high-throughput auditory P300-based brain-computer interface. Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology 120, 1252–61 (07 2009)
14. Li, Y., Pan, J., Wang, F., Yu, Z.: A hybrid BCI system combining P300 and SSVEP and its application to wheelchair control. IEEE Transactions on Biomedical Engineering 60(11), 3156–3166 (2013)
15. Machado, M.F.C.: Hybrid BCI based on visual and auditory stimuli. Ph.D. thesis, Universidade de Coimbra (2019)
16. Mak, J., Arbel, Y., Minett, J.W., McCane, L.M., Yuksel, B., Ryan, D., Thompson, D., Bianchi, L., Erdogmus, D.: Optimizing the P300-based brain–computer interface: current status, limitations and future directions. Journal of neural engineering 8(2) (2011)
17. McFarland, D., Wolpaw, J.: EEG-based brain-computer interfaces. Current Opinion in Biomedical Engineering 4, 194 – 200 (2017), synthetic Biology and Biomedical Engineering / Neural Engineering
18. Nijboer, F., Laar, B., Gerritsen, S., Nijholt, A., Poel, M.: Usability of three electroencephalogram headsets for brain–computer interfaces: A within subject comparison. Interacting with Computers 27, 500–511 (07 2015)
19. Niknamian, S.: The introduction of designing a hybrid brain computer interface system. Biomedical Journal of Scientific and Technical Research 16 (03 2019)
20. Oralhan, Z.: A new approach for hybrid BCI speller based on P300 and SSVEP. Avrupa Bilim ve Teknoloji Dergisi (17), 1392–1398 (2020)
21. Sellers, E., Arbel, Y., Donchin, E.: BCIs that use P300 event-related potentials. Brain-Computer Interfaces: Principles and Practice (01 2012)
22. Sellers, E.W., Turner, P., Sarnacki, W.A., McManus, T., Vaughan, T.M., Matthews, R.: A novel dry electrode for brain-computer interface. In: International Conference on Human-Computer Interaction. pp. 623–631. Springer (2009)
23. Shao, L., Zhang, L., Belkacem, A.N., Zhang, Y., Chen, X., Li, J., Liu, H.: EEG-controlled wall-crawling cleaning robot using SSVEP-based brain-computer interface. Journal of Healthcare Engineering 2020 (2020)
24. Spüler, M., Rosenstiel, W., Bogdan, M.: Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning. Plos One 7 (12 2012)
25. Sutter, E.E.: The brain response interface: communication through visually-induced electrical brain responses. Journal of Microcomputer Applications 15(1), 31–45 (1992)

26. Takano, K., Komatsu, T., Hata, N., Nakajima, Y., Kansaku, K.: Visual stimuli for the P300 brain–computer interface: a comparison of white/gray and green/blue flicker matrices. Clinical neurophysiology 120(8), 1562–1566 (2009)
27. Toppi, J., Risetti, M., Quitadamo, L., Petti, M., Bianchi, L., Salinari, S., Babiloni, F., Cincotti, F., Mattia, D., Astolfi, L.: Investigating the effects of a sensorimotor rhythm-based BCI training on the cortical activity elicited by mental imagery. Journal of neural engineering 11(3), 035010 (2014)
28. Wang, W., Chakraborty, G.: Selection of effective probes for an individual to identify P300 signal generated from P300 BCI speller. International Journal of High Performance Computing and Networking 15(1-2), 72–79 (2019)
29. Wang, Y., Wang, R., Gao, X., Hong, B., Gao, S.: A practical vep-based brain-computer interface. IEEE Transactions on neural systems and rehabilitation engineering 14(2), 234–240 (2006)
30. Wierzgała, P., Zapała, D., Wójcik, G., Jolanta, M.: Most popular signal processing methods in motor-imagery BCI: A review and meta-analysis. Frontiers in Neuroinformatics p. 78 (11 2018)
31. Wolpaw, J.R., McFarland, D.J., Neat, G.W., Forneris, C.A.: An EEG-based brain-computer interface for cursor control. Electroencephalography and clinical neurophysiology 78(3), 252–259 (1991)
32. Yin, E., Zhou, Z., Jiang, J., Chen, F., Liu, Y., Hu, D.: A novel hybrid BCI speller based on the incorporation of SSVEP into the P300 paradigm. Journal of neural engineering 10(2) (2013)
33. Zander, T., Kothe, C.: Towards passive brain–computer interfaces: applying brain–computer interface technology to human-machine systems in general. Journal of neural engineering 8 (03 2011)

ISSN 1870-4069

Extraction of Body Posture Characteristics as a Correlation Variable with the Level of Attention

Alfredo Garcia[1], Juan Manuel Gonzalez[1], Amparo Palomino[2]

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

alfredo_amigo18@hotmail.com,
{jumagoca78,ampalomino}@gmail.com

Abstract. Nowadays, several factors have been associated that intervene in the behavior of the level of attention of people. These factors can be internal or external to the context in which the user develops. This paper addresses one of the factors inherent to human behavior: Body posture. Starting from this variable, the extraction of 8 statistical characteristics (mean, variance, obliquity, kurtosis, standard deviation, maximum, minimum, and rank) is obtained, which are processed mathematically to eliminate out-of-range measurements and measurements produced by errors or noise present in the experiment. For the visualization process, the characteristics extracted are correlated with the mean of the percentage of attention that is obtained from brain waves to finally show their shape in 2D graphics. This analysis is performed using statistical tools that graphically demonstrate some differences in the level of attention among a child with ADHD and a child without ADHD . With the results obtained in this work and as future work, behavior patterns and mathematical models will be sought which describe the level of attention based on the user's body posture.

Keywords. Attention level, brain signals, body posture, diagnostic ADHD, statistic analysis.

1 Introduction

ADHD is characterized by lack of attention, impulsivity and hyperactivity. Recently it has been estimated that it affects 3.5% of school-age children worldwide and is said to be one of the most common psychiatric disorders among young people. Children with these problems are often unpopular and lack reciprocal friendships, but are not always aware of their own unpopularity. Although these symptoms tend to decrease with age, at least 50% of children with ADHD still have symptoms that decrease in adulthood. Despite the vast literature that supports the efficacy of stimulant medication in the treatment of attention deficit / hyperactivity disorder (ADHD), several limitations of pharmacological treatments highlight the clear need for effective alternative psychosocial treatments. There is also evidence of interventions that involve both the school and the training of parents that have resulted in classifying them as "empirically validated treatments" [1]. Attention deficit hyperactivity disorder (ADHD) is a common neurobiological condition that affects school-age children. One of the main symptoms is the lack of attention, which is a key factor of low academic performance, especially in tasks that require a lot of concentration time [2]. Children with Attention Deficit Hyperactivity Disorder (ADHD) experience a deficit in cognitive processes responsible for behaviors aimed at specific objectives, known as executive functioning (FE) [3]. The biggest challenge for adults with attention deficit hyperactivity disorder (ADHD) is the management of information and tasks [4].

2 Systems for Measuring the Level of Attentıon

The study of physiological signals such as brain waves, heart rate, body temperature, among others, has revealed great advances in recent times obtaining significant results in applications from different fields of study such as medicine, robotics, psychology, among others. Currently there are low-cost commercial devices to obtain the reading of brain signals with which it is possible to know the level of attention of the user in an unreliable way. The performance of these devices is limited to the software and hardware established by the manufacturer for a specific task, added to the lack of accuracy in reading the signals, because the devices do not have a robust system for data acquisition and processing. The performance of these devices is limited, since their manufacture is oriented to simple tasks or didactic games. Another cause of the low performance of these devices is that they are invasive or intrusive; Tiaras, helmets, blood samples are used to obtain the signals of the user's physiological variables, any error in the calibration could generate an error in the final diagnosis [5].

To know the degree of affectation that ADHD produces in people, it is necessary to have tools that can provide a feedback of the percentage of attention when executing a specific task. Currently there is a variety of commercial devices that quantitatively provide the level of concentration, meditation, relaxation and user care, but in some cases are achieved in an invasive way, affecting the response of the user and consequently the final diagnosis. These devices usually use a physiological variable to infer the levels of attention in people, they are of the single-user type and of an accessible cost. The performance of these devices is limited since they have restrictions on the part of the manufacturer regarding the software and hardware implemented.

Data acquisition and processing speeds of MINDWAVE, EMOTIV EPOC, MUSE devices, among others; they have delays and can not obtain a reading of the acquired variables in a time approximated to the real time. Another disadvantage presented by this type of device is its low usability and versatility in practice, since the user requires a long time for the devices to recognize the physiological signals that are desired to be acquired. Some devices have a graphic interface designed by the manufacturer, whose feedback is based solely on the indication of the level of attention graphically.

Various applications in areas such as: psychology, education, business, health, among others require a system that accurately identifies the level of attention in people, and that in turn provides an instant response of what happens , as well as a reliable final

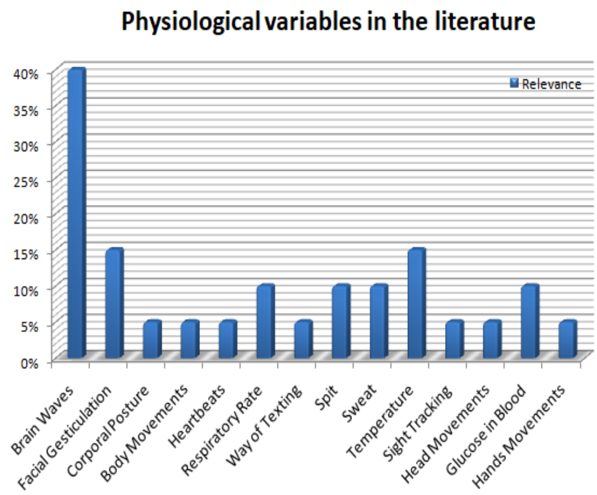


Fig. 1. Relevance of the physiological variables related to the level of attention in the literature.

Fig. 2. Correct positioning of the Neurosky MindWave headband.

diagnosis for decision making. It is also desirable to obtain a feedback that encourages the user to raise the level of attention at the same time as executing a specific task [6].

Biofeedback training systems foster a specific mental or physical state in a user through a closed cycle of bio-feedback. These systems gather the physiological state of a person through the detection of hardware, integrate this state into a computer-based interactive system and present the comments so that the user can work to adjust their status [7].

In this research work, we propose to implement a system to measure the level of attention in children with ADHD, generating an analysis on the samples obtained, with the purpose of characterizing the behavior of their brain waves and obtaining statistical patterns that allow us to identify the presence of ADHD in students through the use of this device. The system is non-invasive and has an interface centered on the user, with the aim of obtaining a final diagnosis that reliably describes the level of attention.

2.1 Physiological Variables Related to the Level of Attention

The variety of techniques applied in the field of research to quantify the level of attention of people leads toward a descriptive analysis that is presented in this work [8].

Starting from the review in the literature of the variables used to relate the level of attention in people, has been found that the brain waves are the physiological variables with greater relevancy due to the cognitive relationship that exists between think-ing and brain activity. Therefore its implication is direct [9, 10, 11].

The graph of the Fig. 1 shows the relevance that each one of the physiological variables has on the level of attention of the people. This analysis is ob-tained from the state of the art of the related works.

Depending on the physiological variable used, the device is chosen to perform the data acquisition. Within the most devices used in the literature are: the electroencephalogram, WEB cam, motion sensors, gyroscopes, electrodes, mouse, electrocardiogram, electrochemical sensors, keyboard, transducers, cam-eras and optical sensors [12].

2.2 MindWave Headband

To realize the implementation of the electronic system in this work we use the MindWave device developed by the manufacturer Neurosky (www.neurosky.com), which allows obtaining EEG signals through a headband type interface that is placed on the head and It is powered by a 1.5 V type AAA battery.

It uses a wireless interface to communicate with the computer and acquires the signals through passive bio-sensors connected to an electrode that makes contact with the forehead. In addition, it has a reference terminal is connected to the earlobe. This feature is used to determine the origin of a signal. In the brain-computer interfaces, the location of the electrodes allows obtaining different representations of the EEG. The MindWave device has only one terminal placed on the front of the subject, in what is formally known as a pre-frontal zone. Fig. 2 shows the correct way in which the device is used [13].

2.3 Body Band to Measure Body Posture

The results of the experiments based on the user's brain waves are considered as a dependent variable and we proceed to relate the level of attention of a person using body posture as a second physiological variable.

The sensor was initially placed in a vest to be able to perform the measurement (Fig. 3A), this involved a longer time in the sampling so it was replaced by a band as shown in Fig. 3B.

The orientation of the MPU-6050 sensor with respect to the three-dimensional system (X, Y, Z), can be seen in Fig. 4. The measured angles are only those that form the movements around the “X” axis and the movements around the Axis y".

This is because they are the only angles where static tonic support reactions are considered which occur to maintain a normal and upright posture against the force of

Fig. 3. A) Posture sensor placed in the vest. B) Posture sensor placed on the body band.

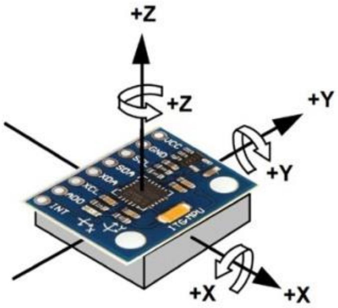


Fig. 4. Reference of position of the MPU-6050 sensor in the band.

gravity causing the user to require a neuromotor balance in relation to their neuroaxis which produces a direct impact on their level of attention and concentration [14].

3 Affectation of ADHD in the Brain

The dispersed attention deficit (whose abbreviation is ADD and ADHD if it is with hyperactivity that is the most frequent) is a disorder of unclear cause, probably with the intervention of genetic and environmental factors, in which there is an alteration at the system level central nervous system, manifesting itself through an increase in activity, impulsivity and lack of attention, and frequently associating other alterations.

The genetic factor is demonstrated, since ADHD is 5 to 7 times more frequent in siblings and 11 to 18 times more frequent in twin siblings. Several genes possibly involved have been described.

ADHD is one of the most frequent causes of school failure and social problems in children.

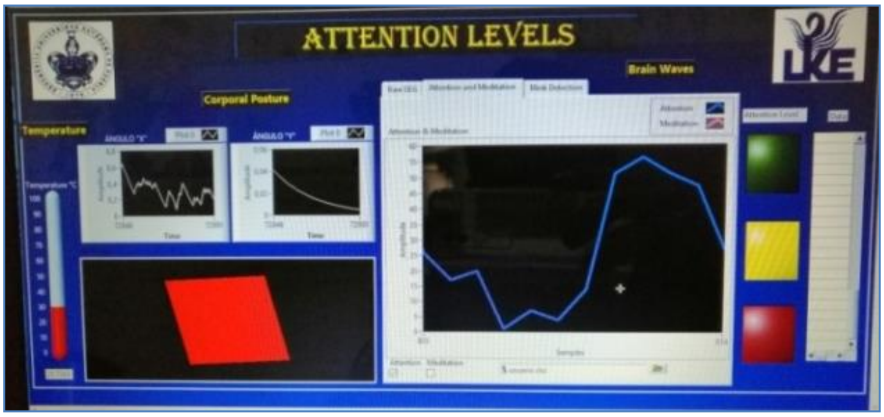


Fig. 5. Graphical interface implemented in LABVIEW.

Fig. 6. Development of the test in students with ADHD.

4 Development of the Attention Test

The experimental tests were conducted using the MindWave commercial device of the Neurosky company, to detect the level of attention in Mexican primary level students.

A sample of 22 students with diagnosed ADHD and 11 students without ADHD was evaluated whose ages are between 6 and 12 years.

The test consisted of a test to identify colors, which was obtained from the demos of the company Brain HQ (https://www.brainhq.com/why-brainhq/about-the-brainhq-exercises/attention).

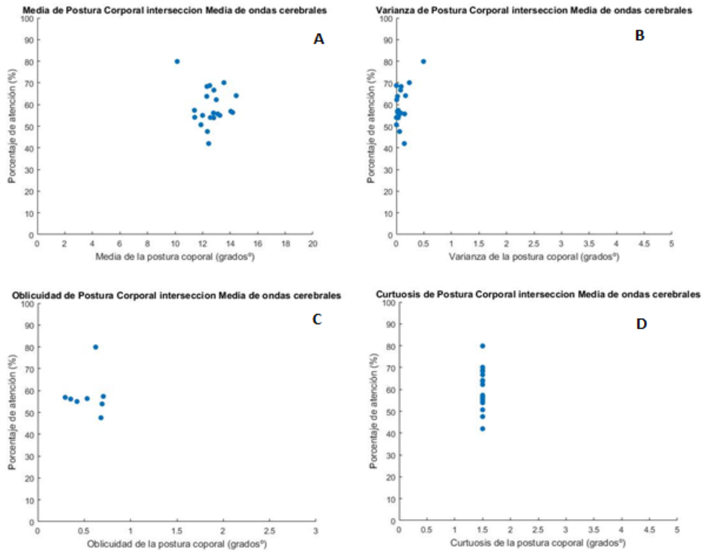


Fig. 7. Scatter plots 2D A) (mean PC, mean OC). B) (variance PC, average OC). C) (PC oblique, mean OC). D) (PC curtuosis, medium OC).

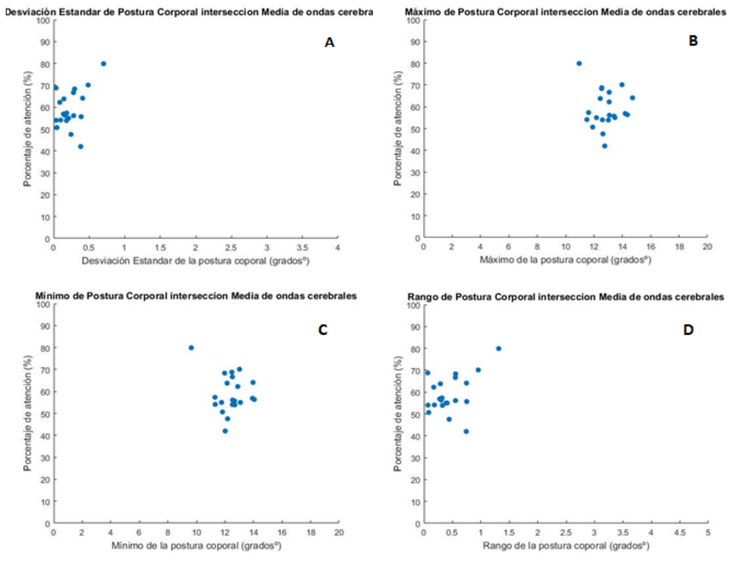


Fig. 8. Scatter plots 2D A) (PC deviation, OC average). B) (maximum PC, average OC). C) (minimum PC, average OC). D) (PC range, average OC).

To obtain the data of brain signals, a graphical interface was implemented, using the LABVIEW software. Fig. 5 illustrates the graphic interface where you can observe the behavior of brain signals, body posture, temperature, a traffic light as feedback , a vector where the sampled data and the variation of the user's attention level are stored.

The test was done in the tablet modality for both cases (students with diagnosed ADHD and students without ADHD). The practical development is shown in the Fig. 6.

The experiment was developed in a classroom where the environment was controlled and adequate to avoid distracting agents and obtain a natural response and a better user performance.

5 Analysis and Results

In the visualization of the data, 2D dispersion diagrams are plotted, which are a fundamental tool for finding patterns and correlation trends between the percentage of attention and the result of body posture.

The vectors are plotted in Fig. 7A (mean PC vector, OC average vector), in Fig. 7B the vectors are plotted (PC variance vector, OC average vector), in Fig. 7C the vector is plotted (PC obliqueness, OC average vector), in Fig. 7D the vectors are plotted (PC tannery vector, OC average vector), in Fig. 8A the vectors are plotted (PC standard deviation vector, OC average vector), in Fig. 8B the vectors are plotted (maximum vector PC, average vector OC), in Fig. 8C the vectors are plotted (minimum vector PC, average vector OC) and in Fig. 8D the vectors are plotted (vector range PC, OC vector).

Using the information analyzed, the correlation factor (ρ_{xy}) is obtained between each independent characteristic extracted from the body posture (X) and the dependent characteristic extracted from the percentage of attention (Y), both for children with ADHD and for children without ADHD (shown in table 1).

Table 1. Correlation coefficients by characteristic extracted in children with ADHD.

Extracted Feature	**Correlation coefficient (ρ_{xy}) in children with ADHD**	**Correlation coefficient (ρ_{xy}) in children without ADHD**
Mean	0.1737	0.2225
Variance	0.1098	0.1413
Obliquity	0.1059	0.1746
Curtuosis	0.2999	0.1622
Standard deviation	0.1628	0.1609
Maximum	0.2183	0.2466
Minimum	0.2040	0.2678
Rank	0.1611	0.1379

6 Conclusion

The results obtained on the data analysis show that the most influential statistical characteristics on the correlation between body posture and the level of attention are the minimum, maximum and curtuosis.

In the 2D dispersion diagrams, you can see patterns typical of each of the features extracted where from these you can make a classification where 2 membership groups are obtained which would be: children with ADHD and children without ADHD.

Acknowledgments. Special recognition to teacher "Claudia Gonzalez Calleros" for her valuable collaboration in taking samples with students with ADHD.

References

1. Pascual, M.F., Begoña, Z., Buldian, K.M.: Adaptive cognitive rehabilitation interventions based on serious games for children with ADHD using biofeedback techniques: assessment and evaluation. In: COMPUTE'10 Proceedings of theThirdAnnual ACM Bangalore Conference, pp. 1–4 (2010)
2. Asiry, O., Shen, H., Calder, P.: Extending attention span of ADHD children through an eye tracker directed adaptive user ınterface. In: ASWEC'15, Proceedings of the ASWEC 24th Australasian Software Engineering Conference, 1, pp. 149–152 (2015)
3. Weisberg, O., Galoz, A., Berkowitz, R., Weiss, N., Peretz, O., Azoulai, S., Rubin, D.K., Zuckerman, O.: TangiPlan: designing an assistive technology to enhance executive functioning among children with ADHD. In: IDC'14 Proceedings of the conference on Interaction design and children, 1, pp. 293–296 (2014)
4. Sonne, T., Jensen, M.M.: Evaluating the chillfish biofeedback game with children with ADHD. In: IDC'16 Proceedings of the The 15th International Conference on Interaction Design and Children, 1, pp. 529–534 (2016)
5. Domínguez, C.: Las ondas binaurales y sus efectos. Tesis de Investigación Experimental 1, Ciudad Cooperativa Cruz Azul, pp.1–22 (2015)
6. Aballay, L., Aciar, S., Reategui, E.: Propuesta de un método para detección de emociones en e-learning. In: ASAI´15, 16º Simposio Argentino de Inteligencia Artificial, pp. 121–128 (2015)
7. Sonne, T., Jensen, M.M.: ChillFish: a respiration game for children with ADHD. In: TEI '16 Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction, 1, pp. 271–278 (2016)
8. Marín, E.J.: Detección de emociones del usuario. Tesis Pontificia Universidad Católica de Valparaíso, 1, pp. 1–67 (2014)
9. Hernández, A., Vásquez, R., Olivares, B.A., Cortes, G., López, I.: Sistema de detección de emociones para la recomendación de recursos educativos. Programación Matemática y Software, 8(1), pp. 58–66 (2016)
10. Saneiro, M.M.: Apoyo psico-educativo y afectivo en entornos virtuales de aprendizaje. Badajoz International Journal of Developmental and Educational Psychology, 1(2), De INFAD Base de datos, pp. 233–241 (2015)
11. Campazzo, E., Martinez, M., Guzmán A.E., Agüero, A.: Entornos virtuales de aprendizaje integrado a tecnología móvil y detección de emociones. Secretaría de Ciencia y Tecnología/Departamento de Ciencias Exactas Físicas y Naturales/Universidad Nacional de La Rioja, 1, pp. 1–5 (2014)

12. García, A.E.: Análisis de ondas cerebrales para determinar emociones a partir de estímulos visuales. Universidad Veracruzana Facultad de Estadística e Informática, 1, pp. 1–137 (2015)
13. Torres, F., Sánchez, C., Palacio, B.: Adquisición y análisis de señales cerebrales utilizando el dispositivo MindWave. MASKANA, I+D+ingeniería, 1, pp.1–11 (2014)
14. Dorbessan, L., Rodríguez, C.A.: La postura corporal en el deporte simétrico y asimétrico. Tesis Universidad Abierta Interamericana, Argentina, 1, pp. 1–15 (2004)

ISSN 1870-4069

How Objects Categorize the Human Brain: EEG and fMRI as Analysis Point

Rigoberto Cerino[1], Sergio Vergara[2]

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

[2] Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Electrónica,
Mexico

{cerino_rigoberto, svergara2}@hotmail.com

Abstract. This paper aims to study the state of the art of research based on the reading and processing of brain signals, which allow categorization of different concepts, in multiple aspects, such as images or texts. An analysis of works that perform semantic classifications of objects such as houses, faces, tools, buildings is presented. Existing methodologies of the general process of reading and adaptation of brain waves, the types of filters implemented for the elimination of noise in the signals, and types of signal classifiers are exhibited. The study is oriented to non-invasive methods of brain wave reading, such as electroencephalography (EEG) through electrode headbands, and functional magnetic resonance imaging (fMRI).

Keywords. Categorization, classification, concepts, EEG, filter, fMRI, neuroimaging, semantics.

1 Introduction

The identification of the neural processes that underlie semantic representations is a key challenge in cognitive neuroscience. Different hypotheses have been proposed on how representations of particular concepts establish a conceptual knowledge system. The general acceptance is that the properties of shared objects are reflected in the organization of the semantic system and that the system is generalized through concepts that belong to a particular category (such as animals, tools or buildings). The notion of category specificity in the organization of object knowledge arose in the 1980s, when Warrington and his colleagues reported for the first time on patients with selective disabilities for a semantic category compared to other semantic categories [1]. Since these initial investigations, a large number of studies have confirmed the phenomenon of semantic deficits specific to the category.

It has been reported that patients have impediments to all kinds of knowledge about a particular category, such as living things, for example, [2, 3].

Differences in brain activity related to the category have been demonstrated with various neuroimaging methods in healthy subjects, for living beings versus man-made objects, and for various categories of specific objects, such as faces, body parts, animals, fruits, vegetables, buildings, tools and furniture [4-6]. For some types of objects, the functional organization by semantic category has been demonstrated within a given modality, for example, category-specificity in the visual path for faces [7, 8] or for living versus non-living entities [9, 10]. It has also been shown that objects and their sensory or functional attributes (such as actions associated with tools) activate the same neuronal regions [11, 12, 1], suggesting that these regions are implicitly involved in the conceptual representation.

Achieving a clearer picture of the categorical distinctions in the brain is essential for the understanding of the conceptual lexicon, but much more precise investigations both in categorical distinctions and in other aspects of the conceptual re-presentation [13, 14] will be necessary for this evidence contribute to lexical research. Although semantics clearly plays a central role in the ability of human language, since the transfer of meaning is the goal of a purposeful communication, our understanding of its instantiation and functional location in the brain is far from complete. These types of systems can have very specific applications in real life, for example, as support for people who lost speech, people who do not listen, people who have had a stroke and need to verify if the concepts are still present in their mind.

2 Representation of Object Concepts in the Brain

Object concept: “Memory representations of a class or category of objects. Necessary for numerous cognitive functions, including the identification of an object as a member of a specific category and making inferences about the properties of the object” [15].

Evidence of the functional neuroimaging of the human brain indicates that information about the outstanding properties of an object, such as its appearance, how it moves and how it is used, is stored in the active sensory and motor systems when that information was obtained. As a result, the concepts of objects belonging to different categories, such as animals and tools, are represented in neural networks based on partially distinct sensory and motor properties. This suggests that object concepts are not explicitly represented, but arise from weighted activity within brain regions based on properties. However, some property-based regions seem to show a categorical organization, thus providing consistent evidence with domain-specific formulations based on categories as well.

The central idea is that knowledge of the object is organized by sensory characteristics (Form, movement, color) and motor properties associated with the use of the object (and in some models, other functional properties, verbally measured, such as where typically finds an object, its social meaning, etc.) [5]. Most studies examine only distinctions in very distant semantic fields, for example comparing abstract and concrete concepts, verbs and nouns, or natural and artifact types [1, 10, 16-19].

Convergent evidence of monkey neurophysiology, neuropsychology and functional brain imaging has established that object recognition critically depends on the current of ventral occipitotemporal processing (see [20]). In addition, functional studies of brain imaging of object recognition have provided convincing evidence that the occipitotemporal cortex is not a homogeneous object processing system, but has a fine-grained structure that appears to be related to the object category. The most tested categories have been human faces, houses, animals and tools [1, 10, 16, 17, 19, 21-24]. The direct comparison of one category of object with another has revealed different activity groups (for example, the fusiform area of the faces, (FFA); the region of the brain called Parahippocampal Place Area (PPA) (Fig. 1). In addition, pattern analysis techniques have identified different activity patterns related to the category of objects that discriminate between a relatively large number of object categories [10, 25-31]. These patterns related to the category of objects extend over a large area of occipitotemporal cortex, are stable both within and between subjects, and can be identified even when subjects freely view complex scenes [32].

[33] provided evidence that category-related activity groups in occipitotemporal cortices associated with visualization of object images are also seen when subjects participate in a verbal conceptual processing task.

3 Conceptual Processing and Subsequent Temporal Cortex

Functional brain imaging studies on conceptual and semantic-lexical processing (that is, using word stimuli) have constantly isolated two key brain regions: left ventrolateral prefrontal cortex (VLPFC) and the ventral and lateral regions of the posterior temporal cortex, generally stronger in the left hemisphere than in the right (Fig. 1) [34, 35].

Activity in VLPFC has been strongly associated with semantic memory control; specifically, recovery guide and post-recovery selection of conceptual information stored in subsequent temporal areas and perhaps in other cortical areas [1].

A large amount of functional neuroimaging evidence has implicated the temporal lobes, particularly the posterior region of the left temporal lobe, as a critical site for stored representations, especially on concrete objects. Recent studies have provided additional support for this view by demonstrating that the regions of the left posterior temporal cortex that are known to be active during conceptual processing of images and words (fusiform gyrus and lower and middle temporal gyrus) (Fig. 1) were also active during the listening comprehension of the sentences [36-38].

Another recent approach to investigate functional neuroanatomy of conceptual processing has been to use stimulus repetition tasks. It is well established that previous experience with a stimulus results in a more efficient process (repetition primacy) and a reduced hemodynamic response, typically known as repetition suppression, but also as adaptation, neural primacy and repetition attenuation, when that stimulus is found later [39].

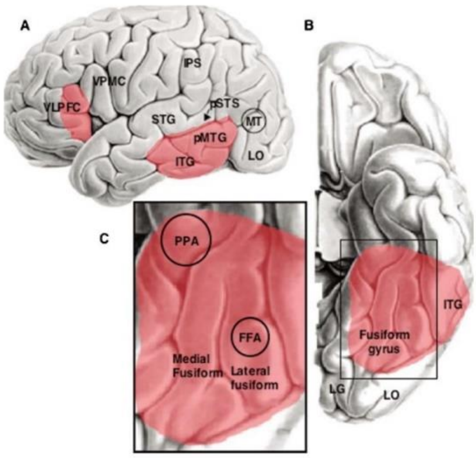


Fig. 1. Schematic lateral view of the left hemisphere (A) and ventral view of the frontal and right temporal lobes (B). The fusiform turn is shown in greater detail in (C). The red regions show the approximate location of areas typically involved in conceptual processing tasks, especially with specific objects. ITG, Inferior temporal gyrus; LG, lingual gyrus [5].

Activity patterns related to the category of objects have been observed in the ventral and lateral regions of the posterior temporal cortex using a variety of stimuli (images, written names, sounds associated with objects, names heard) [9, 39, 40].

The ventral temporal cortex shows strong category effects, but these effects were not modulated by movement. In contrast, the lateral temporal areas responded more strongly to the movement than to the static images, supporting the hypothesis that the lateral temporal cortex is the cortical site of complex motion processing.

4 Identification of Object Categories from EEG Related to Events

First, electroencephalography (EEG) has a well-documented ability to characterize certain brain states, in particular the processing of different semantic categories. Second, the high temporal resolution of the EEG allows an accurate characterization of the concept's recovery in terms of the electrophysiological patterns that make decoding possible. Thirdly, the development of semantic decoding algorithms based on EEG is interesting from the perspective of applications, since the temporal resolution of EEG allows decoding in real time. There are multiple pattern analysis techniques that allow the decoding of conceptual information [1].

[1] investigates the possibility of identifying conceptual representations of EEG related to events based on the presentation of an object in different modalities: its spoken name, its visual representation and its written name. Bayesian logistic regression is implemented with a multivariate Laplace before classification, to identify the neuronal activity related to the concepts from ERP. The highest accuracies (89% of

correctly classified tests) were obtained by classifying the drawings of objects (their visual representation).

In [16], a set of advanced data extraction techniques is presented that allows deciphering the category of individual concepts from individual tests of EEG data.

A comparison between information measures and ERPs revealed a reliable correlation between the N400 amplitude and a surprise word [5]. These findings suggest that different measures of information quantify cognitively different processes and that readers do not use the hierarchical structure of a sentence to generate expectations about the next word.

[19] bases his research on studying the semantic relationship between pairs of nouns of concrete objects such as "horse-sheep", "swing-melon" and how this activity relationship is reflected in the EEG signals. The authors perform an analysis focused on feature extraction algorithms. They train different classifiers to associate a set of signals to a previously learned human response, belonging to two classes: semantically related or not semantically related. Although the previous studies showed an influence of the perception of the object in the tasks related to the action, in [35] it is verified if the representations of the action facilitate the recognition of visual objects.

In [30], twelve different categories are selected as visual stimuli and the subjects were presented in a controlled task and an analysis of different ERP calculations is performed where the user distinguishes whether the stimulus presented is an “objective (category detected) / non-objective” or “objective / rest”, and the results provide useful information about the channels and the part of the signals that are affected by different categories of objects in terms of brain signals. In research [8] it is intended to untangle activities at the node and network level in milliseconds of time scale of perception and decision making. Clear and noisy images of faces and houses are used for the task of categorizing images, and EEG records combined with source reconstruction techniques to study when and how oscillatory activity is organized within the FFA, PPA and DLPFC.

In [31], the dynamics of human vision are studied using a combination of rapid rates of stimulus presentation, electroencephalography and multivariate decoding analysis. The representative structure of a large number of stimuli is obtained, and the emergent abstract categorical organization of this structure is presented. In addition, it is possible to separate the temporal dynamics of perceptual processing from the effects of selecting higher-level objectives.

A particularly relevant component for semantic processing is the N400. Subsequent research has shown that N400 components are generated whenever stimulus events induce semantic or conceptual processing. As such, many researchers have used the N400 component of brain waves as a dependent variable in psycholinguistic experiments. [39] investigates how speech and gesture affect interpretation processes in real time, and addresses the cognitive and neuronal processes that mediate speech-gesture integration.

In [28], the author evaluates the contribution of mid-level characteristics to the decoding of conceptual category using EEG and a new paradigm of rapid periodic decoding. It uses a stimulus set consisting of intact objects of the animated categories (for example, fish) and inanimate (for example, chair) and coded versions of the same objects that they were. However, animation decoding for encoded objects was only possible at the slowest periodic presentation speed.

In [26], the EEG signals together with a multivariate pattern recognition technique were used to investigate the possibility of identifying the conceptual representation based on the presentation of 12 semantic categories of objects (5 examples per category). The attentional facilitation of the constituent characteristics does not spread automatically within an object, but depends on the relevance of the specific task of each characteristic. In [37] a novel experimental design is used, which allows simultaneous electrophysiological measurements of the allocation of attention to two different characteristics (rotation and color) within an object (a square). This was possible by presenting a square that evokes two visual evoked potentials in steady state (SSVEP) for rotation and color changes, respectively. Given the continuous oscillatory nature of the SSVEPs, it was possible to investigate the temporal course of neuronal activity in the early visual cortex of the human brain when the subjects attended one of the two characteristics.

5 Identification of Object Categories from fMRI

It has been a lasting challenge to establish the correspondence between a simple cognitive state (such as the thought of a hammer) and the underlying brain activity. In addition, it is unknown if the correspondence is the same between individuals. A recent approach to study brain function uses machine learning techniques to identify the neuronal pattern of brain activity that underlies several thought processes. Previous studies that used a machine learning approach have been able to identify the cognitive states associated with the visualization of an object category, such as houses [1-8]. The central feature of this approach is its identification of a multivariate pattern of voxels and its characteristic activation levels that collectively identify the neural response to a stimulus.

These machine learning methods have the potential to be particularly useful for discovering how semantic information about objects in the cerebral cortex is represented because they can determine the topographic distribution of activation and distinguish the information content in various parts of the cortex. Multivariate pattern analysis is a technique that allows the decoding of conceptual information, such as the semantic category of an object perceived from neuroimaging data. Impressive results of single-trial classification have been reported in studies that used functional magnetic resonance imaging (fMRI) [1].

[1] focused on identifying the cognitive state associated with the visualization in 4 seconds of an individual line drawing (1 of 10 family objects, 5 tools and 5 houses, such as a hammer or a castle). It is able to identify the category of the object, and for the first time, identify both the individual objects and the category of the object that the participant was seeing, based only on the activation patterns of other participants.

In [9], the neural patterns associated with individual objects as well as with categories of objects were identified using a machine learning algorithm applied to activation distributed throughout the cortex. This study also investigated the degree to

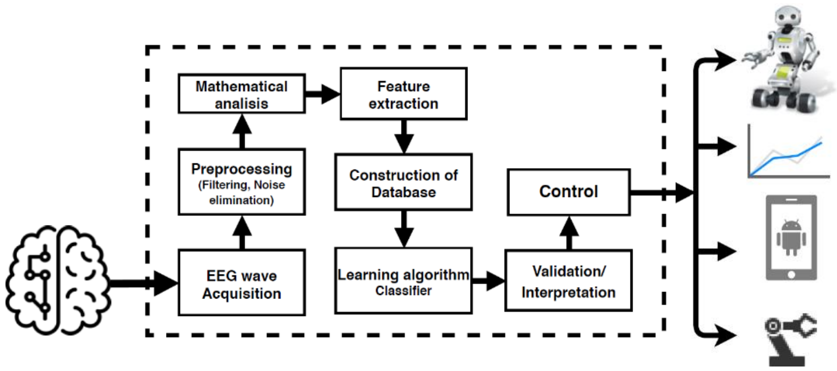


Fig. 2. Methodology required for the analysis of EEG signals oriented to the process of categorization of objects in the human brain.

which objects and categories are similarly represented at the neural level in different people.

[17] trains classifiers to identify which of the ten examples of objects and two categories of objects a participant was seeing.

A common neuronal pattern was discovered among the participants, and was implemented to train a classifier to identify the correct object category and the object example of the fMRI data of new participants who did not participate in the classifier training. In [41] an investigation base on the question of whether it is posible that a coding model based on five semantic attributes directly related to sensory motor experience can successfully predict brain activation patterns caused by word sets. The results show that a lexical concept is not represented identically in different brains, but is a reflection of the unique life of each participant.

In [23], an analysis of lexical categories in the brain is performed, through digital image processing for the detection of regions of the brain that are activated when the user reads abstract and concrete nouns or verbs, in order to classify them into semantic categories. [37] summarizes the evidence of temporal and spatial brain imaging studies that have investigated the emotional effects on the lexical, semantic and morphosyntactic aspects of language during the understanding of individual words and sentences. The revised evidence suggests that emotion is represented in the brain as a set of semantic characteristics in a distributed sensory, motor, language and affective network.

fMRI studies have revealed that DLPFC calculates higher level cognitive functions, including image categorization [8].

Based on the previous analysis, where the research presents its main methodologies, the process that a brain signal requires to be interpreted, and the elements to take into account the process, an approach is made towards the EEG signals and it is proposed that there may be a sufficiently generic methodology for the process of interpretation of the EEG signals oriented to the image classification (Fig. 2).

Tables 1 and 2 present the distribution of the main articles found in the literature that provide distinctive elements of the process of categorization of objects in the human

Table 1. Main methods of analysis, filters and feature extraction processes in the literature for the object categorization process.

	Semantic classification	Method	Filter	Extraction
[17]	Tools and buildings	fMRI	190 s high-passes	PCA
[39]	Animals and tools	EEG	1–30 Hz band-pass filter	
[16]	Mammals and tools	EEG	1–120 Hz band-pass filter to eliminate slow deviations in the signal and high frequency noise, and then sampled at 300 Hz	CSP
[40]	Nouns (sound, color, manipulation, visual movement and Shape)	fMRI		
[23]	Nouns and verbs	fMRI	128 s high-passes	
[19]	Nouns	EEG	Filter passes bands. Finite response filter (FIR) with a lower cutoff frequency of 20 Hz and a high cutoff frequency of 1 Hz.	LPC, PCA, ICA, SEGN FDTW y CSP
[1]	Written words (actions and objects)	fMRI		
[8]	Faces and houses	EEG	Fast Fourier Transformations (FFT), 1-100 Hz band-pass, and a digital noise filter at 60 Hz	
[30]	Animated and inanimate categories	EEG	FIR with Hamming window with 0.1 Hz-100 Hz, and reduced to 250 Hz.	
[27]	Animated vs. inanimate, faces vs. bodies, human bodies vs. non-human bodies, human vs. non-human faces			
[28]	Animated and inanimate images	EGG	FIR with Hamming window 0.1 Hz-100 Hz	
[26]	12 categories of different objects (animals, flowers, body parts, etc.)	EEG	FIR with Hamming window 0.1–150 Hz	

brain, such as the method of analysis of identification of important brain areas, the neural mechanism interpreted, the main component analyzed of the signal, the type of filter applied, the methods of extraction of characteristics, the methods of classification and the categories that classify.

6 Conclusions

The main conceptual advances offered by these findings are that there is an identifiable neuronal pattern associated with the perception and contemplation of individual objects, and that, depending on the type of stimulus, part of this pattern is shared among

Table 2. Main classification methods, brain regions studied, neural mechanisms and components analyzed in the literature for the object categorization process).

	Classification	Region	Neural mechanism	Components analyzed
[17]	Grouped Gaussian-Naive Bayes (GNB) variance classifier	Left hemisphere, ventral premotor cortex and posterior parietal cortex, right parahippocampal gyrus		
[39]		Parieto-occipital central, occipitotemporales	ERP	NI-P2 waves, N400
[16]	SVM		ERP	Ondas NI-P2
[23]		Frontotemporal cortex, inferior occipital cortex, precetral cortex, (IFG)		
[19]	Decision Tree (DT), Naive Bayes (NB), Decision (RL), Artificial Neural Networks (ANN), k-Nearest Neighbors (kNN) and Support Vector Machines.		ERP	P300, N400
[1]		Lateral occipito-temporal cortex, middle temporal area (MT) and medial superior temporal area (MST), left frontal cortex, ventral occipito-temporal cortex		
[8]			ERP	
[30]	LDA with the Representation Similarity Analysis (RSA) framework.			
[27]	SVM with the Representation Similarity Analysis (RSA) framework		ERP	Primary visual area V1 and inferior temporal cortex (IT).
[28]	LDA		ERP	N300, N400
[42]	SVM, regularized least squares with linear and Gaussian cores			
[26]	Naive-Bayesian Classification (NBC)			

various participants. This neural pattern is characterized by an activation distribution across many cortical regions, which involves locations that encode various object properties. The analysis performed gives information about visually represented objects.

The human visual system recognizes objects quickly and the neuronal activity of the human brain generates signals that provide information on the categories of objects seen by the subjects, the results provide useful information about the channels and the part of the signals that are they are affected by different categories of objects in terms of brain signals.

In terms of neural processes of semantics, we have identified scalp locations, time intervals and frequency bands that are especially informative about category differences. The ultimate goal of interaction and cooperation between humans and machines is that a reasonable response be achieved directly to the user's intention.

The technology and brain-computer interface processes are based on direct access to physical activity information in the thinking processes of the human brain, providing an effective neuro-path that allows the interpretation of brain signals. This has become an important development direction in the field of natural language processing.

7 Discussion

Current data shows that the object category can be successfully decoded from the first visual components of the scalp EEG. This contribution is relevant for the investigation of the brain-computer interface. The neuroimaging findings reviewed here provide strong support for models based on sensory-motor properties by revealing a considerable overlap in the neural circuits that support the perception, performance and knowledge of objects. For data analyzes involving fMRI, the high cost of studies may make such systematic explorations impractical.

Fortunately, the studies that demonstrate that conceptual knowledge and semantic category-analysis can be analyzed using EEG, represents a potential research area, in which you can deepen in methods of filtering the acquired signals, in the combination of existing methodologies in order to create hybrid analysis procedures and evaluate their accuracy. Since EEG studies can be performed at a much lower cost than fMRI, they can be a more feasible methodology for large-scale lexical research and categorization.

In contrast, some authors emphasize that EEG techniques involve numerous efforts to improve the accuracy of the location of the neuronal source, and that the information obtained in many of the processes performed is not sufficient to provide a complete and robust estimate of the distribution spatial of the neuronal responses that underlie the perception of different kinds of objects; but that there are specific methods and contributions for this type of signals, in which pattern classification techniques are involved through machine learning.

We believe that large-scale systematic explorations of mental lexicon and categorization with neural data are necessary, which involve both a more careful analysis of conceptual distinctions and a greater range of categories, because most of the papers present comparative analyzes between reduced number of categories, in addition to certain categories are completely isolated from each other.

In this sense, it would be interesting to carry out studies and experiments that allow to know if it is possible to identify concepts through neuronal signals, associated to the process of human communication, that is, to see the possibility of detecting a complete action related to human language.

References

1. Simanova, I., Gerven, M., Oostenveld, R., Hagoort, P.: Identifying object categories from event-related eeg: toward decoding of conceptual representations. PLOS One, pp. 1–12 (2010)
2. Patterson, K., Nestor, P.J., Rogers, T.T.: Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neuroscience, 8, pp. 976–987 (2007)
3. Mahon, B.Z., Caramazza.: A Concepts and categories: a cognitive neuropsychological perspective. Annual Review Psychology, 60, pp. 27–51 (2009)
4. Gerlach, C.: A review of functional imaging studies on category specificity. J Cogn Neurosci, 19, pp. 296–314 (2007)
5. Martin, A.: The representation of object concepts in the brain. Annual Review Psychology, 58, pp. 25–45 (2007)
6. Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L.: Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19(12), pp. 2767–2796 (2009)
7. Kanwisher, N., Yovel, G.: The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci., 361, 2109–2128 (2006)
8. Chand, G. B., Lamichhane, B., Dhamala, M.: Face or house image perception: beta and gamma bands of oscillations in brain networks carry out decision-making. Brain Connectivity, 6(8), pp. 621–631 (2016)
9. Noppeney, U., Price, C.J., Penny, W.D., Friston, K.J.: Two distinct neural mechanisms for category-selective responses. Cereb Cortex 16, pp. 437–445 (2006)
10. Kozunov, V., Nikolaeva, A., Stroganova, T.: Categorization for faces and tools-two classes of objects shaped by different experience-differs in processing timing, brain areas involved, and repetition effects. Frontiers in Human Neuroscience (2018)
11. Kiefer, M., Sim, E.J., Herrnberger, B., Grothe, J., Hoenig, K.: The sound of concepts: four markers for a link between auditory and conceptual brain systems. The Journal of Neuroscience, 28, pp. 12224–12230 (2008)
12. Hoenig, K., Sim, E.J., Bochev, V., Herrnberger, B., Kiefer, M.: Conceptual flexibility in the human brain: dynamic recruitment of semantic maps from visual, motor, and motion-related areas. Journal of Cognitive Neuroscience, 20, pp. 1799–1814 (2008)
13. Maguire, M., Brier, M., Ferree, T.: EEG theta and alpha responses reveal qualitative differences in processing taxonomic versus thematic semantic relationships. Brain and Language, 114, pp. 16–25 (2010)
14. Sachs, O., Weis, S., Zellagui, N., Huber, W., Zvyagintsev, M., Mathiak, K.: Automatic processing of semantic relations in fMRI: Neural activation during semantic priming of taxonomic and thematic categories. Brain Research, 1218, pp. 194–205 (2008)
15. Martin, A.: The representation of object concepts in the brain. Annual Review of Psychology, 58(1), pp. 25–45 (2007)
16. Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., Lakany, H.: EEG decoding of semantic category reveals distributed representations for single concepts. pp. 12–22 (2011)
17. Shinkareva, S., Mason, R., Malave, V., Wang, W., Mitchell, T. M., Just, M. A.: Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PloS One, 3(1) (2008)
18. Pulvermüller, F.: The neuroscience of language: On brain circuits of words and serial order. Cambridge: Cambridge University Press (2002)
19. Calvo, H., Paredes, J., Figueroa-Nazuno, J.: Measuring Concept Semantic Relatedness through Common Spatial Pattern Feature Extraction on EEG Signals. Cognitive Systems Research (2018)

20. Grill-Spector, K., Malach, R.: The human visual cortex. Annual Review of Neuroscience 27, pp. 649–677 (2004)
21. Kanwisher N, Downing P, Epstein R, Kourtzi Z.: Functional neuroimaging of visual recognition. In: Handbook of Functional NeuroImaging of Cognition, Eds. Cabeza, R., Kingstone, A., pp. 109–152 (2004)
22. Martin A.: Functional neuroimaging of semantic memory. In: Handbook of Functional NeuroImaging of Cognition, Eds. Cabeza, R., Kingstone, A., pp. 153–186 (2001)
23. Chand, G., Lamichhane, B., Dhamala, M.: Face or house image perception: beta and gamma bands of oscillations in brain networks carry out decision-making. Brain Connectivity (2016)
24. Moseley, R., Pulvermüller, F.: Nouns, verbs, objects, actions, and abstractions: Local fMRI activity indexes semantics, not lexical categories. Brain and Language, 132C, pp. 28−42 (2014)
25. Cox, D.D., Savoy R.L.: Functional magnetic resonance imaging (fMRI) ''brain reading'': detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage, 19, pp. 261–270 (2003)
26. Spiridon, M., Kanwisher, N.: How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron, 35, pp. 1157–1165 (2002)
27. Behroozi, M., Daliri, M., Shekarchi, B.: EEG phase patterns reflect the representation of semantic categories of objects. Medical & Biological Engineering & Computing (2015)
28. Grootswagers, T., Robinson, A., Shatek, S., Carlson, T.: Untangling featural and conceptual object representations. NeuroImage, 202 (2019)
29. Helbig, H., Graf, M., Kiefer, M.: The role of action representations in visual object recognition. Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale (2006)
30. Mohammad, R., Mitra, T., Kavous, S.: EEG signature of object categorization from event-related potentials. Journal of Medical Signals & Sensors, pp. 37−44 (2012)
31. Grootswagers, T., Robinson, A., Carlson, T.: The representational dynamics of visual objects in rapid serial visual processing streams (2018)
32. Hanson, S.J., Matsuka, T., Haxby J.V.: Combinatorial codes in ventral temporal lobe for object recognition: Haxby revisited: is there a ''face'' area? NeuroImage 23, pp. 156–166 (2001)
33. Chao L.L, Haxby J.V, Martin A.: Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nature Neuroscience, 2, pp. 913–919 (1999)
34. Bookheimer, S.: Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annual Review Neuroscience, 25, pp. 151–188 (2002)
35. Grill-Spector, K., Malach, R.: The human visual cortex. Annual Review Neuroscience, 27, pp. 649–677 (2004)
36. Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ MO.: Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cerebral Cortex, 14, pp. 247–255 (2004)
37. Hinojosa, J.A., Moreno, E., Ferré, P.: Affective neurolinguistics: towards a framework for reconciling language and emotion. Language, Cognition and Neuroscience, pp. 1−27 (2019)
38. Schwarzlose, R.F., Baker, C.I., Kanwisher, N.: Separate face and body selectivity on the fusiform gyrus. Journal of Neuroscience, 25(47), pp. 11055–11059 (2005)
39. Grill-Spector, K., Henson, R., Martin, A.: Repetition and the brain: neural models of stimulus-specific effects. Trends Cognitive Science, 10, pp. 14–23 (2006)
40. Wu, Y., Coulson, S.: How iconic gestures enhance communication: An ERP study. Brain and Language, 101, pp. 234−245 (2007)
41. Fernandino, L., Humphries, C.J., Seidenberg, M.S., Gross, W.L., Conant, L.L., Binder, J. R.: Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsychologia, 76, 17–26 (2015)

42. Isik, L., Meyers, E., Leibo, J., Poggio, T.: The dynamics of invariant object recognition in the human visual system. Journal of neurophysiology (2013)

ISSN 1870-4069

A Conversational Model for the Reminiscence Therapy of Patients with Early Stage of Alzheimer

Víctor Manuel Morales de Jesús, María Josefa Somodevilla García

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

victor.morales@cs.buap.mx, mariajsomodevilla@gmail.com

Abstract. Reminiscence therapy is a treatment that helps mitigate unstable psychological states in patients with Alzheimer's disease, where past experiences are evoked through conversations between the patient and their caregiver, stimulating autobiographical episodic memory. It is highly recommended that people with Alzheimer's receive this type of therapy constantly. This article describes the initial work in the development and implementation of a conversational model that can support caregivers of patients with this type of dementia to provide therapy periodically and enhance the benefits it offers. It is intended that the model allows generating personalized conversations between a prototype of the conversational system and patients. The proposed model collects patient information related to their preferences, history and lifestyle to personalize the conversation according to the profile of them. The modules that integrate this architecture are the following: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialog Management (DM), Dialog Model, Natural Language Generation (NLG) y Text-to-Speech (TTS).

Keywords. Alzheimer, conversational model, dialog model, reminiscence therapy.

1 Introduction

Dementia is a neurodegenerative and progressive condition that is characterized by the alteration of cognitive processes, behavior, emotional state and the limitation in the ability to develop activities of daily living [1]. The most common type of dementia is Alzheimer's, covering up to 80% of cases worldwide. It is also one of the main reasons for disability in elderly people over 65, generating dependence on those who suffer from it. In 2010, there were 35.5 million people with dementia worldwide, and it is estimated that by 2030 this rate will increase up to 65.7 million, while by 2050 the worrying amount of 115.4 million is anticipated [2].

According to this trend, an accelerated growth in the population of older adults is expected in Mexico within the next 20 years. Some estimates indicate that during the period from 2000 to 2030 the population between 0 and 14 years will present a decrease of 14%, the population from 15 to 64 years will have a growth of 47%, while the population over 64 years will increase 300% [3]. The analysis of these data is relevant as it shows a need in the development and implementation of the necessary care for this population group. Almost a third part of this population will be older adults with various levels of dependence, and as mentioned above, the main cause of dependence in the elderly is precisely dementia.

Although dementia is currently not curable, it is necessary that people suffering from some type of dementia can preserve a good quality of life as long as possible. In this sense, there are both pharmacological and non-pharmacological treatments focused on mitigating the psychological, behavioral and cognitive impairment symptoms. It is preferable that the treatment of a patient with Alzheimer's disease begins with a non-pharmacological intervention, since this type of treatment promotes the use of different methods and techniques to provide emotional and physical stability to patients without the side effects of pharmacological treatments [4]. Through the implementation of appropriate environments, stimulating tasks and a diverse kind of therapies according to the needs of each patient, the benefits of this type of interventions can be maximized [5].

Reminiscence therapy is among these types of interventions; therapy is based on conversations between the patient and his caregiver, which are guided by the caregiver and focus on activities and experiences of the past, where the patient is chronologically guided through his life story [6]. This type of therapy focuses on the long-term memory of the patient using information that is familiar to the patient, and relatively easy to remember, since during the early stage of Alzheimer's the short-term memory is mainly affected. The benefits of subjecting Alzheimer's patients to reminiscence therapy have been analyzed in several studies and it has been observed that carrying it out constantly supports improving unstable psychological and emotional states in patients, as well as increasing their sociability and trust [6-8].

However, it is difficult for most patients with Alzheimer's to consult a specialist or therapist periodically. Similarly, due to the overburden in Alzheimer’s patient caregivers, they are quite limited in the time they can use to provide therapy constantly. In this sense, the objective of this work is focused on the creation of a conversational model capable of generating personalized conversations in order to support caregivers and therapists to provide reminiscence therapy in a constant way.

2 Related Work

It is relevant to mention that a large number of proposes focus on the diagnosis of dementia through the classification of symptoms and identification of distinctive characteristics for early detection. Thus, in [9] they propose a virtual avatar with spoken dialogue functionalities that conducts an interview to patients, which is based on the MMSE (Mini-Mental State Examination), Wechsler's memory scale and other related neuropsychological tests. Authors record the interaction between the avatar and the patient, extract diverse audiovisual characteristics and subsequently implement classification algorithms, achieving an accuracy of 0.93 in the detection. In the approach proposed by [10], they analyze three main linguistic characteristics that are considered as verbal indicators of confusion in people with Alzheimer's disease, which are: the richness in the vocabulary, the structure of the syntactic trees of the sentences and acoustic signals. Finally, they apply various machine learning algorithms to identify confusion within the dialogue reaching 82% accuracy.

On the other hand, there are proposals that focus on the development of assistive technology to help people with dementia to carry out their daily life activities [4, 11-13], as well as to support different types of therapy for the management of psychological symptoms, emotional and behavioral [14-16].

In [13] a system that provides occasional reminders and reminiscence conversations remotely is implemented. Through video calls they aim improving psychological stability and being able to assist people with dementia to perform simple tasks. They observed that after the intervention, the psychological stability of a patient persisted for up to three hours after the conversation and the success rate when completing a task was increased to 80%. This type of proposal is compared in more detail against other instruction monitoring strategies in [17]. [11] describes the development of a mobile robot, called ED, which supports Alzheimer's patients in the performance of tasks through visual monitoring and verbal reminders. They analyze the voice interactions between ED and each of the Alzheimer's patients involved in the study (n = 10). Their analysis reveals that for patients who have high levels of confusion, it is very likely that they ignore the robot at the time it provides assistance to perform a task, which represented 40% of the behavior of the group under study.

In the work described in [12] they analyzed the interaction between people with Alzheimer and a simulated intelligent cognitive assistant to support patients to develop their daily activities. The assistant is able to detect when there is a problem to perform the task on the part of the patient and then offers assistance based on the context of the moment of confusion. The authors performed the evaluation of the system in three focus groups: Alzheimer's patients, their caregivers and people without Alzheimer's. Finally, the analysis of the data obtained from each of these groups showed that the interaction style and the type of voice have a greater relevance for Alzheimer's patients. It is desirable that there may be a personalization in the interaction that exists between the cognitive assistant and the patient.

Similarly, in [4] an assisted cognition system was designed and the results on the use and adoption of the system are analyzed to support in providing occupational therapy and assess the effectiveness of the system to mitigate behavioral changes in Alzheimer's patients. They carry out an evaluation of the system with two couples (patient-caregiver) and observe that the personalization of the intervention and the tactile interface for interaction facilitate the adoption of the system. They also mention that implementing this type of systems in the treatment of Alzheimer's patients reduces the workload of their caregivers.

Alternatively, in [14] a study is conducted which describes how people with Alzheimer's disease can help develop a companion robot. In their work, they identify elements of importance for patients, which make a company robot have a higher level of acceptance. They mention that a relevant aspect is the need for the robot to know

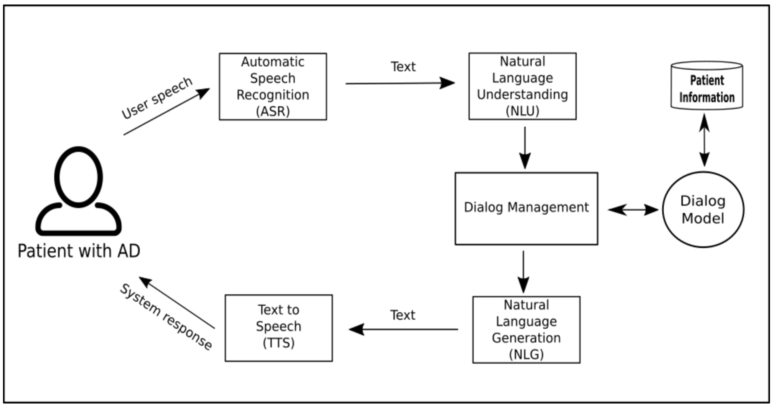


Fig. 1. Conversational model architecture.

interests, preferences and the patient's life history and use this information to motivate the conversation and reminiscent about events that patients can still remember. Similarly, in [15] they design a semi-autonomous agent (Eva), which is capable of engaging in simple conversations by controlling a human operator to be perceived as an autonomous agent.

Eva implements direct interventions with Alzheimer's patients, through simple conversations and music therapy. After conducting a preliminary evaluation with a group of caregivers of patients with Alzheimer's disease (n = 8), they obtain quantitative and qualitative results that were used to evaluate the robot's interaction skills.

Because the production and understanding of language is relatively well preserved in the early stage of Alzheimer's, proposals that use spoken dialogue systems to support early diagnosis and the development of non-pharmacological treatments have shown be well adopted by patients and caregivers, so they can be a valuable tool in the treatment of this condition.

Most of the proposals analyzed implement generic dialogues, that is, the same conversation takes place for all patients and therefore they do not provide the benefits that have been observed in conversations where the dialogue is personalized to each patient as established by the Reminiscence therapy Therefore, this paper proposes a model to generate a personalized dialogue focused on reminiscence therapy, such as the one described in section 3.

3 Proposed Conversational Model

In this section, the proposed architecture of the conversational model is presented. The modules of this architecture are shown in Figure 1 and they are as follows: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialog Management (DM), Dialog Model, Natural Language Generation (NLG) and Text-to-Speech (TTS).

3.1 Automatic Speech Recognition (ASR)

The ASR module is responsible for recognizing the user's (Alzheimer's patient) utterances and transcribing them in written sentences (plain text). Because the purpose of the work is focused on the development of the conversational model and not on a method for automatic speech recognition, at this initial stage the Google Cloud Speech-to-Text[1] web service is used, since it shows a good performance when recognizing speech in the Spanish language.

In general, the ASR module recognizes and transcribes emissions in real time. Subsequently, the transcribed emissions are processed by the NLU module using various techniques in order to obtain the semantic attributes present in the emissions.

3.2 Natural Language Understanding (NLU)

Once the user's broadcasts have been transcribed, it is necessary to analyze the text to obtain the semantic interpretation of the information contained in the received message, so that it can be processed by the conversational model. This process will be carried out in the NLU module by implementing various language processing techniques to perform this task. Some of these techniques involve: syntactic analysis, recognition of named entities, and classification of speech acts, to name a few. Currently we are working on the development of a method for the identification and classification of speech acts, which are described below.

Speech acts classification. The purpose of the identification and classification of speech acts is to determine the intention of an utterance. A speech act is an utterance that serves as a function in communication and this function could be a greeting, request, complaint, invitation, compliment or statement. The process for creating a model for classifying speech acts is the one shown in Figure 2.

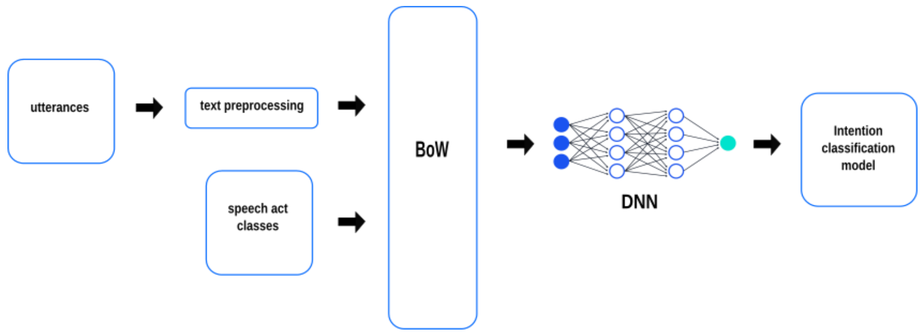


Fig. 2. Creation of the intention classification model.

[1] https://cloud.google.com/speech-to-text/

Table 1. Types of speech acts.

Speech act	Function
Assertives	The speaker affirms or denies something (affirming, suggesting, informing, concluding, etc.)
Directives	The sender tries to get the receiver to perform some action (inviting, ordering, asking)
Commissives	The issuer agrees to varying degrees to perform an action or to do something in the future (swearing, promising, bet)
Expressives	The speaker expresses an affective state, which can be emotional or physical (greeting, congratulating, complaining)
Declarations	Its execution produces a change in some state of affairs, so the issuer is usually an authority (marrying, baptizing, etc.)

Initially, a corpus of 130 phrases was created, which is considered to be likely emissions by users when interacting with the system. Subsequently, five types (classes) of speech acts were established that the system could recognize and classify which are shown in Table 1.

In this way, each phrase within the created corpus is associated with a type of speech act, but it is also associated with a specific act, which could be understood as a subclassification.

For example, the phrases ("Hello", "Good morning", "How are you"), are labeled as expressive_gretting, that is, they are the type of expressive speech act, but specifically is an act of greeting; the phrases ("What is your name?", "What are you like?") are labeled as directive_quetion_personal, which in this case is considered a directive speech act, but specifically they are an act of question and in this case they can be specified further as a personal question. Until now the labeling process was carried out manually.

Once each phrase is associated with a type (class) of speech act, a vector (Bag of Words) of all the phrases in the corpus is created. The created BoW representation is then used as input data in a neural network to obtain a speech act classification model or intention classification model (ICM). The neural network used follows a simple (feed-forward) scheme with two hidden layers. The goal of the neural network is to assign a class label for each phrase represented by BoW.

Until now, the model created offers an accuracy of 0.89, however, it is contemplated to increase the number of phrases, as well as experiment with other classification algorithms to try to obtain a better precision. It is also important to mention that although, with the current classification model, it is intended to know the intention of the emission, it is still necessary to extract the semantic attributes that will allow the system to understand not only the type of intention (speech act) detected, but the purpose of the intention.

A chatbot was created to implement the speech act classification model. It uses the model to identify the type of speech act of a new sentence and select a possible response from a set of candidate responses. The tasks performed by the chatbot are the following:

1. Initially, the chatbot receives a written sentence, which is preprocessed.
2. A vector (BoW) is created by comparing the words of the input sentence with the patterns that have been used to train the classification model. If the word is within the established patterns, then it is setting in the vector.
3. Using the BoW representation, the model is implemented to predict the speech act in the entry sentence. That is, the model determines which pattern the sentence is associated with.
4. Finally, the model returns the set of classes along with the probability of the sentence to belong to each class. In this case, to establish if the sentence belongs to a class is determined by setting a probability threshold $\mu = 0.75$ to determine if the sentence can belong to one of the classes or not, this process is carried out as follows:

   a. From the set of classes *C* and probabilities *P*, the one with the highest probability value *P(c)* is selected. If the probability of the selected class is greater than the established threshold, that is, $P(c) > \mu$, then the sentence is classified with the speech act denoted by the class. Subsequently, a response is selected from the set of response templates associated with that class.
   b. On the other hand, if the probability value of the selected class is less than the threshold $P(c) < \mu$, then the sentence cannot be associated to any class and the chatbot displays the message *"Sorry, I did not understand"*.

An example of a chatbot dialogue is shown in Table 2.

Table 2. Example of a dialog context with the chatbot.

Dialog between the user (U) and the system (S)
S: Hola, buenos días.
U: Hola, ¿quién eres? (*expresivo_saludo, directivo_pregunta_personal*)
S: Soy María y me gusta platicar con las personas.
U: ¿de qué te gusta platicar? (*directivo_pregunta_gustos*)
S: …*
Candidate response templates for the speech act: *directivo_pregunta_gustos*
* Puedo platicar de diversos temas como música, películas y varias cosas más.
* Me gusta mucho platicar sobre la familia. ¿Te gustaría que platicar sobre tu familia?
* Me agradaría platicar sobre qué tipo de música te gusta. ¿Te gustaría?

As can be observed, the set of templates are currently static, therefore, during the development of this work these response templates there will be modified, so that they will be dynamic responses in order to customize the response according to the user

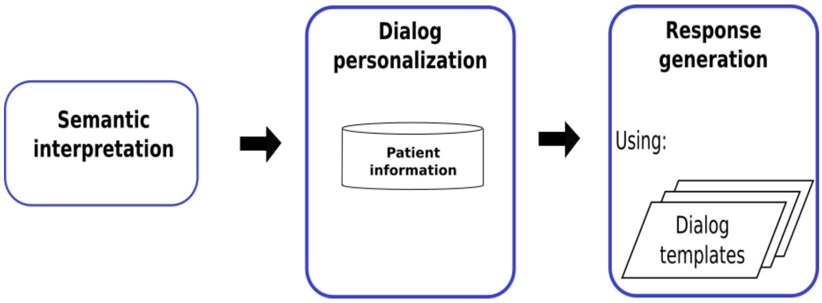


Fig. 3. Development of the dialog model.

information. In addition, it is important to mention that although at the moment a response is selected randomly according to the type of intention detected in the sentence, it is subsequently proposed that the response will be selected based on the conversation history.

3.3 Dialog Management (DM)

The DM module will be in charge of controlling the conversation shifts between the user and the system, determining the actions to carry out a performance during the interaction according to the current state of the dialogue, but will also analyze the history of the dialogue to determine which will be the most appropriate response. This module will interact directly with the Dialogue Model module to provide it with the semantic information that can be used to select relevant patient information and personalize the conversation.

This module will implement a dialog manager to determine which of the potential moves will be optimal in each state, to meet the goals or interests of the user. The dialog manager consists of two sub-components: the dialogue strategy and the adaptation strategy. The dialog manager is based on a state machine. Each state specifies a transition state from the current state and the condition to validate that transition, as well as a syntax for what the user can use as response in that state (for example, single-choice option or an open-answer option in the case when a question is posed to the user). State definitions also include the specification of agent schemes as templates with variables that are instantiated each time the state is entered. Schemes are a type of protocols, which directs a specific type of dialogue and its completion. The implementation of the DM is contemplated to be developed in a future version of this work.

3.4 Dialog Model

During the construction of the dialogue model, a set of stages have been considered according to the particular tasks that need to be developed in each one. Figure 3 shows the general process for the construction of the dialogue model.

Initially, it is necessary for the dialogue model to obtain a semantic interpretation of the user's utterances. This interpretation as mentioned above is done in the NLU module. The information collected from the patient is then used as it is intended that the created model be able to generate a personalized dialogue for each patient.

The collection of patient information is carried out through a set of forms proposed in [18, 19], which focus on collecting data on family relationships, history and lifestyle of patients with Alzheimer's. The information is provided by the patient's caregiver, who is usually a close relative. Then, once the patient information is available, the dialogue model can use this information to generate a response according to the patient's information and the context of the dialogue to personalize the conversation.

3.5 Natural Language Generation (NLG)

The response has to be appropriate to the context of the dialog, this means that it is necessary include the information provided by the Dialog Management module and this information will be formatted using lexical elements and an adequate syntactic structure to be understood by the user. That is, it has to be given in natural language and in this particular case in the Spanish language.

In this case, the use of predefined templates to generate the response is considered. Each template will have pre-established words and phrases and empty fields that will be filled with information provided by the Dialogue Management module as mentioned above. Because Alzheimer's patients usually prefer the use of simple and short sentences to communicate, the templates will be developed according to these characteristics.

3.6 Text-to-Speech (TTS)

Finally, the response generated will be synthesized through the TTS module using the IBM Watson Text-to-Speech[2] web service that provides a more natural voice generation. However, like the ASR module, it is not possible to make a comparison with different tools that allow this task to be carried out.

At this moment, the architecture of the conversational system and the objective of each of the modules that integrate it have been described in a general way. In this sense, the model is currently being developed to recognize and classify speech acts used by the user (person with Alzheimer's) when interacting with the system.

4 Conclusions and Future Work

Due to the cognitive impairment that affects people with some type of dementia, in this case specifically with Alzheimer's dementia, in early stages the patients present various types of symptoms, where one of the most common is short-term memory loss.

[2] https://www.ibm.com/watson/services/text-to-speech/

This situation significantly affects their behavior towards the people around them and generates unstable psychological states such as anxiety and depression. However, it has been observed that providing patients with non-pharmacological treatments such as reminiscence therapy helps to mitigate this type of symptoms and improve their quality of life. Therefore, it is not strange that there is a great effort from the computational approach in order to develop tools that support the different types of therapies that seek to mitigate the psychological, emotional and behavioral symptoms generated by dementia. Nevertheless, in the case of conversational models or systems focused on providing therapy to Alzheimer's patients, most of these proposals are based on an interaction with the patient through generic dialogues without taking into account the life history and preference of the patients; this means that a personalization during the intervention is not achieved.

This document presents the proposal for the creation of a conversational model that serves as a support tool in providing reminiscence therapy to patients with Alzheimer's disease, through the generation of personalized conversations according to life history and interests of each patient. This would allow us to take advantage of the benefits of providing the therapy constantly.

Because the proposed proposal is in an initial development, as future work it is contemplated to continue with the development and improvement of the method of classification of speech acts and to implement compression techniques of natural language that allow to extract semantic attributes of the emissions, as well as to develop each one of the methods described in the architecture of the conversational model. Finally, it is considered to carry out an initial experimentation of the system to evaluate the adoption of the system by patients with early stage Alzheimer's.

References

1. Gutierrez, L.: Demencias en México. In: Demencias, una visión panorámica, Universidad Potosina, 9–20 (2016)
2. Wimo, A., Prince, M.: World Alzheimer report 2010. Technical Report, Alzheimer´s Disease International (2010)
3. Fernández, P.H., Velarde, S.I., Hernández, M.F., Murgía, V.: Dinámica demográfica 1990-2010 y proyecciones de población 2010-2030. Consejo Nacional de Población (CONAPO) (2014)
4. Navarro, R.F., Rodríguez, M.D., Favela, J.: Use and adoption of an assisted cognition system to support therapies for people with dementia. Computational and Mathematical Methods in Medicine (2016)
5. Gitlin, L.N., Hodgson, N., Jutkowitz, E., Pizzi, L.: The cost-effectiveness of a nonpharmacologic intervention for individuals with dementia and family caregivers: The tailored activity program. The American Journal of Geriatric Psychiatry, 18(6), 510–519 (2010)
6. Woods, B., O'Philbin, L., Farrell, E. M., Spector, A. E., Orrell, M.: Reminiscence therapy for dementia. Cochrane Database of Systematic Reviews (2018)
7. González-Arévalo, K.A.: Terapia de reminiscencia y sus efectos en los pacientes mayores con demencia. Psicogeriatría, 5, pp. 101–111 (2015)
8. Neal, M., Barton-Wright, P.: Validation therapy for dementia. The Cochrane Library (2003)

9. Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., Nakamura, S.: Detecting dementia through interactive computer avatars. IEEE Journal of Translational Engineering in Health and Medicine, 5, pp. 1–11 (2017)
10. Chinaei, H., Currie, L.C., Danks, A., Lin, H., Mehta, T., Rudzicz, F.: Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. Computational Linguistics, 43(2), pp. 377–406 (2017)
11. Rudzicz, F., Wang, R., Begum, M., Mihailidis, A.: Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease. ACM Transactions on Accessible Computing, 7(2), pp. 1–22 (2015)
12. Wolters, M. K., Kelly, F., Kilgour, J.: Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. Health Informatics Journal, 22(4), pp. 854–866 (2016)
13. Yasuda, K., Kuwahara, N., Kuwabara, K., Morimoto, K., Tetsutani, N.: Daily assistance for individuals with dementia via videophone. American Journal of Alzheimer's disease and other Dementias, 28(5), pp. 508–516 (2013)
14. Casey, D., Felzmann, H., Pegman, G., Kouroupetroglou, C., Murphy, K., Koumpis, A., Whelan, S.: What people with dementia want: designing Mario an acceptable robot companion. In: International Conference on Computers Helping People with Special Needs, pp. 318–325. Springer (2016)
15. Cruz-Sandoval, D., Favela, J.: Semi-autonomous conversational robot to deal with problematic behaviors from people with dementia. In: Ochoa, S. F., Singh, P., and Bravo, J. (Eds.), Ubiquitous Computing and Ambient Intelligence, Cham. Springer International Publishing, pp. 677–688 (2017)
16. Sakakibara, S., Saiki, S., Nakamura, M., Yasuda, K.: Generating personalized dialogue towards daily counseling system for home dementia care. In: Duffy, V. G. (Ed.), Digital Human Modeling. Applications in Health, Safety, Ergonomics, and Risk Management: Health and Safety, Cham. Springer International Publishing, pp. 161–172 (2017)
17. Perilli, V., Lancioni, G. E., Hoogeveen, F., Caffó, A., Singh, N., O'Reilly, M., Sigafoos, J., Cassano, G., Oliva, D.: Video prompting versus other instruction strategies for persons with Alzheimer's disease. American Journal of Alzheimer's disease and other Dementias, 28(4), pp. 393–402 (2013)
18. Rokkaku, R.: Care management sheet pack for the elderly with dementia: The center method ver. 03. Nihon Ronen Igakkai zasshi. Japanese journal of geriatrics, 42(3), pp. 318–319 (2005)
19. Phair, L.: Caring for people with dementia in acute care settings. Nursing Older People, 22 (1), pp. 9–10 (2010)

ISSN 1870-4069

Systematic Review of the State of the Art Regarding the Identification of Cancer Cells of the Leukemia Type with Digital Image Processing

José de Jesús Moya Mora, Manuel Martín

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

moyamora@me.com, mmartin@cs.buap.mx

Abstract. The present diagnosis is based on a review of the literature of several research projects carried out on the use of tools in the detection of leukemia cancer. This paper presents a discussion of the existing literature to know the current state of knowledge about the identification of leukemia using digital images and areas of opportunity for future work are identified.

Keywords. Acute leukemia, machine learning, deep learning, digital image processing, pattern recognition.

1 Introduction

Machine learning and deep learning tools can provide alternatives regarding how medical problems are solved when giving a diagnosis to a patient. The expansion of the data generated by our systems, the medical literature and the inefficiencies of health care systems will require the use of the power of artificial intelligence tools [4, 31].

The integration of computational tools into medical practice, such as machine learning and deep learning, has begun to be widely used. The U.S. Food and Drug Administration of North America has tested much artificial intelligence-based software since 2017 for medical use [6, 31].

The introduction of digital pathology has revolutionized and provided many opportunities in the improvement of traditional pathology and opened up new research opportunities, such as telemedicine [10, 14].

Recently, the use of digital pathology has allowed the use of machine learning algorithms in the automation of the diagnosis of medical treatments [9, 27]. The challenges facing the use of machine learning, as well as deep learning algorithms in the pathology, are diverse, from the digitalization of cell samples, manual labeling in case of supervised learning, initial and maintenance costs, advanced equipment, technical experience, and ethical consideration. However, the possible opportunities to implement tools in pathology are quite a lot [2, 29].

Leukemia is a hematological neoplasm that confers mortality throughout different ages since this disease includes very early ages as is the case of babies in some cases until appearing in adults. It was estimated that there were around 350,000 thousand new cases in 2012 worldwide [7].

2 Works and Methods

In this article, we review the literature related to the use of machine learning and deep learning in the diagnosis of acute and chronic leukemia, both lymphoid and myeloid. The objective of this review was to understand current trends and limitations in the diagnosis of leukemia. The characteristics included in the designs of the reviewed works, techniques used, properties, as well as identification techniques of the different types of leukemia were analyzed.

Applying the search strategies in the aforementioned databases, 30 publications were found, which were completely reviewed and considered of great interest since they provide information on the methodologies in their studies, as well as the results they obtained.

The evaluation metrics used in the research reviewed in the literature varied, however, all studies gave at least one evaluation metric such as accuracy or sensitivity, which were the most common among the diagnoses reviewed.

2.1 Machine Learning

In the literature reviewed for the identification of leukemia the largest number of studies studied was found lymphoid leukemia as the most studied, which is done through a database of images of the disease, goes through a pre-processing of it to obtain better results as read in the literature which allows to obtain characteristics that better define the cell, later reaches a classification and validation process to mention the result obtained by the methodology used.

Fig. 1. Block representation of the methodologies used in the literature for the identification of leukemia.

The acquisition of images in most of the diagnoses found uses a database, Image Database (IDB). This digital image library contains two sets of data, the cells of a set are in the original format, meaning that they are not segmented, while the cells of the second set are already segmented.

Image analysis and pattern recognition methods have been widely used in the field of pathological analysis to help specialists study different cell patterns in microscopic images.

There are different types of cells available in the diagnosis made in the blood smear or tissue sections. These include red blood cells, white blood cells, and platelets. They all exist in different types with different characteristics of shape, color, and texture. The diversity of the cells, the existence of staining artifacts and complex scenes, are some of the examples presented to the specialist to determine the type of leukemia that a patient could suffer when reviewing the samples obtained.

Overlapping or clustering of cells could cause segmentation problems, as well as variations in color, contrast, and background due to non-standard staining techniques applied to leukocytes, different thicknesses and lighting conditions [16].

Leukocytes tend to migrate to tissue sections of blood vessels to eliminate harmful agents and begin the healing process [18]. Leukocytes are divided into two main categories according to the structure of the nuclei: polymorphonuclear cells (granular), mononuclear cells (non-granular) Granulocytes have granules in their cytoplasm and are of three types: neutrophils, basophils, eosinophils.

On the other hand, lymphocyte and monocyte cells are non-granular cell types, since they have only one nucleus. It should be mentioned that leukocytes have no color, but they acquire it when they are stained with chemicals to make them visible under the microscope [22], as can be seen in Fig. 2.

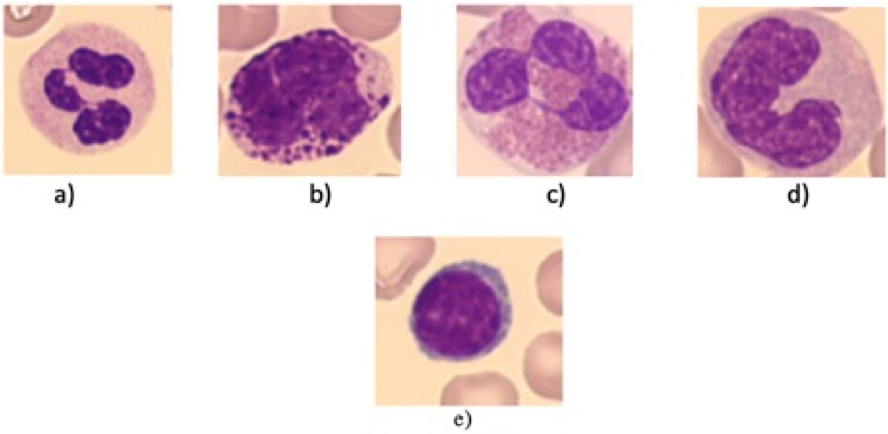


Fig. 2. Representation of white blood cell cells, (a) neutrophil, (b) basophil, (c) eosinophils, (d) monocyte, and (e) lymphocyte [5].

To highlight the colorless leukocytes available in the tissue section or the blood smear, special types of dyes or chemicals are used, and this process is known as staining. The different types of methods used for staining are: Wright staining [33].

Photomicrographs of the tissue section or blood smear may have variations in their color intensity due to the concentration of staining, aging of the staining solution or stained slides, to name a few.

Images of blood smears or tissue sections for clinical and preclinical analysis are widely acquired through bright field microscopy. Image quality is also affected by the use of various types of illuminators such as LED, HBO and XBO [3]. The included works used different approaches to carry out their preprocessing of their databases. The methodology used in the most common cell segmentation algorithm was the recognition of patterns such as geometric or texture, followed by threshold-based methodologies such as Otsu.

Table 1. Different characteristics used to identify leukocytes mentioned in the literature.

Article	Number of images	Feature Extraction Method	Clasificator	Subtypes of cancer cells identified
[17]	98	Area, perimeter and circularity of the lymphocyte nucleus	SVM (Support Vector Machine)	2 classes, normal lymphoid and diseased lymphoid
[18]	220	Gaussian approach, Otsu algorithm, geometric, center and radius of the cells.	MCA (Center point circle algorithm)	2 classes, monocytes and neutrophils, both sick
[19]	106	Segment the white blood cell, HSL color space,	K-means clustering	2 classes basophils and lymphocytes, both sick
[20]	108	Color-based method with search engine of the nearest neighbor with Euclidean distance.	SVM	2 classes, diseased lymphocytes, lymphocyte 1, lymphocyte 2 and lymphocyte 3
[21]	149	Color, geometry and texture analysis	SVM	2 monocyte and neutrophil classes.
[22]	108	Color (green), Otsu?s, morphological operations.	Hough circular transformation	3 classes sick lymphocytes, lymphocyte 1, lymphocyte 2, lymphocyte 3
[23]	300	Color spaces, RGB, HSL, CMYK	K-means clustering	basophilic and monocyte classes both sick.
[24]	288	Characteristics of some blood cell flow cytometry data provided by Beckman-coulter corporation.	SVM	4 normal and sick monocyte classes, healthy and diseased neutrophils.
[25]	150	Geometric and texture characteristics.	SVM y K-PCA	3 monocyte, basophilic and eosinophilic classes all sick.
[26]	160	Characteristics of color space and geometries.	K-means y SVM	4 classes basophils and neutrophils sick and healthy.
[27]	150	Characteristics geometries.	KNN y linear Bayes normal	3 classes diseased lymphocytes, lymphocyte 1, lymphocyte 2 and lymphocyte 3
[28]	256	Statistical and texture characteristics.	EMC-SVM	4 monocyte, lymphocyte, basophil and eosinophil classes all sick.

[29]	166	Statistical characteristics	ELM, RVM	6 monocyte, lymphocyte and basophilic classes, sick and healthy.
[30]	108	Characteristics geometries.	k-nearest neighbor	3 classes, diseased lymphocytes, lymphocyte 1, lymphocyte and lymphocyte 3.
[31]	65	Color and morphological space	SVM	3 classes, basophilic, neutrophils and lymphocyte all sick.
[32]	138	Statistical, texture and geometric characteristics	SVM	2 classes, basophilic, sick and healthy.
[33]	145	Characteristics of color space	Bayesian	2 classes basophilic and eosinophil.
[34]	168	Geometric characteristics and statistics	SVM	4 classes, neutrophils, basophilic, eosinophils and lymphocyte.
[35]	222	Geometric characteristics and statistics	k-nearest neighbor and SVM	4 classes, monocytes, basophilic, neutrophils and eosinophil
[36]	480	Morphological characteristics	Neural network	6 classes, monocytes, lymphocyte, basophilic, healthy and sick
[37]	568	Color spaces	Deep neural network	4 classes monocytes, basophilic, healthy and sick
[38]	345	Geometric, color	Bayesian networks	3 classes lymphocytes sick
[39]	440	Statistics	Neural network	6 classes lymphocytes, monocytes, eosinophil, healthy and sick
[40]	560	Features, geometric, color.	SVM, k-means	4 classes monocytes, basophilic, healthy and sick
[41]	380	Texture, geometric, color	Neural network	4 classes lymphocytes healthy and sick

Most of the works reviewed in table 1 usually segment both the nucleus and the cytoplasm, few studies carry out segmentation in the nucleus of the cell. In the reviewed diagnoses there is not much difference in the evaluation metrics given, from the models in those works that only segmented the nucleus and the cytoplasm to those that only segmented the nucleus. At the time of extracting characteristics, geometry or texture techniques were used.

Some factors that play an important role in the accuracy and classification of leukemia, is how cells are segmented, as well as representations of characteristics used. Characteristic representations must contain useful information, while robust, the background, color, size, location or uneven illumination of the images. Feature extraction representations have been used with different machine learning techniques to classify cells in the blood.

Color and lighting variations can reduce the efficiency of the manual or automated identification system that can lead to a skewed analysis. Therefore, these images should be normalized to minimize variations. This facilitates the best segmentation and, consequently, improves classification accuracy [1]. Several methods of leukocyte classification are proposed in grayscale images that eliminate the need to normalize color but lead to loss of cell color information.

2.2 White Blood Cell Segmentation

In general, the segmentation process involves the removal of leukocytes from the background of the image which then presents a noisy background, where the cell is located, which makes the most vital process difficult. For tissue section images, segmentation becomes more complex due to its complex morphological structures, variable staining, lighting variations, out of focus image components and variability in objects of interest [20].

Cellular segmentation methods can be classified into threshold-based methods [19, 23], methods based on pattern recognition [17], deformable models [12] and metaheuristic-based methods [11, 30].

Threshold methods have also been used as previous steps to locate expected cell locations. Hamghalam [21] used the Otsu threshold method [24] to detect the precise limit of the nuclei in the peripheral blood smear images before applying the active contour method for the detection of the cytoplasm limit [13].

Table 2. Methods of cell segmentation.

Article	Category	Subcategory	Description
17, 18, 19, 20, 21, 22, 23, 34	Thresholding based method	Watershed, Otsu threshold method, growth region	Fast and reliable methods for uniform but consistent images
25, 26, 27, 30, 31, 34, 35, 36	Method based on pattern recognition	SVM, ANN, k-means clustering algorithm, fuzzy c-means	Categorized as unsupervised or supervised methods. Unsupervised methods produce poor results for images that have a complex.
28, 29	Methods metaheuristics	10 point, bold	The segmentation problem is considered an optimization problem and they tried to find the global optimum to segment the objects.

Other variants of threshold-based methods are region-based methods and watershed methods. The performance of the region-based methods is based on the image intensity information. In the region's growth methods, the connected regions of the image are found using seed points and predefined pixel intensity information or border information [28].

In general, seed points are selected manually, which is a major disadvantage of regional cultivation methods [25]. Bread and cabbage [8] used the entropy-based region growth method for leukocyte segmentation. In this work, the regions of interest (of its acronym in English ROI) were located using the special color, and then the entropy of the regions was continuously improved by the expansion of the ROI.

Since leukocytes in microscopic images can be treated as objects, pattern recognition methods can be used to perform segmentation that can be classified as unsupervised or supervised methods.

Unsupervised methods, also known as grouping methods, extract the objects from the data itself. On the contrary, supervised methods use learning-based methods to classify objects. Clustering methods divide the n objects into k groups using the optimization of a criterion function designed for a particular problem. The most used methods in this context are the k-means [26], the diffuse c-means [15].

2.3 Leukocyte Feature Extraction

The outstanding characteristics of the objects to be analyzed are generally used for classification. In microscopic images for the classification of healthy or diseased cells, most of the recent literature focused on texture, size and shape characteristics [26, 32]. In addition to these characteristics, methods based on principal component analysis (PCA) have also been used for feature extraction [8, 26]. In general, the characteristics used in leukocyte analysis can be grouped into geometric characteristics and texture characteristics as shown in table 3.

Table 3. Characteristics used to classify leukocytes.

Article	Category	Features used
17, 18, 19, 20, 21, 22, 23, 24, 3031, 32, 33	Geometric features	Area, radius, convex area, perimeter form factor, concavity, elongation, circularity, symmetry, rectangularity, area ratio, solidity, compactness, concavity
26, 27, 28, 30, 31, 32, 33, 34, 41	Texture characteristics	Variance, standard deviation, correlation, entropy, color, mean, homogeneity, energy, smoothness

For cell discrimination, geometric characteristics play a vital role as they describe the structure and size of leukocytes. The area and perimeter of the leukocytes are the characteristics used to represent the size of the cells, while the shape characteristics can be grouped into characteristics based on regions and boundaries. To extract the characteristics, the cell image is converted into a binary image where the pixels of the cell are represented with a non-zero value.

The characteristics of shape and size are briefly analyzed below:

– Area: the area of the cell is represented by the total number of non-zero pixels within the cell limit.
– Area ratio: the area ratio is defined as the ratio between the number of pixels in the cytoplasm and the number of pixels in the nucleus.
– Convex area: in some cases, the convex hull is calculated, and its area is called the number of pixels within its limit.
– Symmetry: symmetry represents the difference between lines that are perpendicular to the axis greater than the cell limit in both directions.
– Concavity: Concavity is defined as the extent to which the actual limit of a cell is within each chord between non-adjacent limit points.
– Elongation: elongation is the measure of the relationship between the maximum distance and the minimum distance from the center of gravity to the core limit.
– Number of lobes: nuclear polymorph cells have a different number of nuclei (lobes) in their cytoplasm that can be used as one of the most prominent features for cell sorting.
– Orientation: the angle between the x axis and the major axis of the cell is known as orientation.
– Circularity: the circularity of the cell is defined by the ratio of the perimeter of the narrowest bounding circle to the perimeter of the cell.
– Rectangularity: the rectangularity of the core is represented as the ratio between the perimeter of the narrowest bounding rectangle and the perimeter of the core.
– Perimeter: calculated by measuring the sum of the distances between successive limit pixels.

The different properties mentioned were used by different researchers for leukocyte differentiation [1, 5, 20, 23, 33]. These characteristics also play an important role in reducing different noisy elements present in the images. Piuri and Scotti [15, 21] used 13 different characteristics for the identification of leukocytes in blood smear images. But most researchers used texture features along with geometric features to increase the accuracy of the classifier.

3 Discussion

Leukemia is a major hematological malignancy, with high prevalence and incidence, leukemia diagnosis is facing multiple changes.

Leukocytes classification is the complex process for computer systems as compared to human observer which has motivated the researchers to study this field in the perspective of artificial intelligence.

Automated leukocytes classification can help the pathologists in the disease identification and drug development. Although, a significant amount of work has been done in this field in past two decades, there are still some challenges which lead to lower accuracy of cell classification.

Majority of the included studies in this review used supervised learning algorithms. A drawback of using supervised learning in pathology is the need to label samples which is time-consuming and might introduce errors. A solution to that would be to use unsupervised learning methodologies, in which the patterns are determined by the data itself.

One of the major problems in the analysis of tissue section images is its complex morphological structure. It is difficult to differentiate leukocytes from other histological structures such as cells of hair follicles and basal cells of epidermis, red blood cells and other artifacts produced during processing and staining. Another limitation is the variability in the shape of leukocytes mainly due to the plane of sectioning and age of leukocyte in the inflamed tissues, which impose criticality in cell identification.

4 Conclusions

Automated diagnostic studies used a variety of segmentation methodologies of both the nucleus and cytoplasm. The most used method was the method based on pattern recognition, with diffuse c-averages as the most used methodology. Fuzzy c-mean has proven to be more accurate than the grouping of k-means according to the literature investigated. All studies have extracted geometric and texture characteristics. The included studies represented many limitations in research, both in machine learning and in deep learning. These limitations include issues such as sample size, generalization and prospective analysis.

The models presented in this work will reach high accuracy, commonly more than 90%. This is a very common result in the field of machine learning and deep learning research in the field of research that was on cancer cells. This can be attractive; However, it can pose different problems. First, the models presented in this review are generally based on a small sample size and, in many studies, the data come from a single center, which raises the question of how machine learning models are proposed. Therefore, these studies must use more robust databases that will need records and large digital libraries with the ability to avoid the limitation of overfitting.

References

1. Abbas, K., Banks, J., Chandran, V., Tomeo-Reyes, I., Nguyen, K.: Classification of White Blood Cell Types from Microscope Images:Techniques and Challenges, pp. 17–25 (2018)
2. Acs, B., Rimm, D.L.: Not Just Digital Pathology, Intelligent Digital Pathology. JAMA Oncology 4(3), 403–404 (03 2018)
3. Adjouadi, M., Zong, N., Ayala, M.: Multidimensional pattern recognition and classification of white blood cells using support vector machines. Particle & Particle Systems Characterization - PART PART SYST CHARACT 22 (2005)
4. Beam, A.L., Kohane, I.S.: Big Data and Machine Learning in Health Care (2018)

5. Bhavnani, L., Jaliya, U., Joshi, M.: Segmentation and counting of wbcs and rbcs from microscopic blood sample images. International Journal of Image, Graphics and Signal Processing 8, 32–40 (2016)
6. Digital Health Criteria: Food and Drug Administration. https://www.fda.gov/MedicalDevices/Digital Health/ucm575766.htm (2019)
7. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F.: Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. International Journal of Cancer 136(5), E359–E386 (2015)
8. Ghaisani, F., Wasito, I., Faturrahman, M., Mufidah, R.: Deep belief networks and bayesian networks for prognosis of acute lymphoblastic leukemia. In: ICACS '17: Proceedings of the International Conference on Algorithms, Computing and Systems. pp. 102–106 (2017)
9. Ghaznavi, F., Evans, A., Madabhushi, A., Feldman, M.: Digital imaging in pathology: Whole-slide imaging and beyond. Annual Review of Pathology: Mechanisms of Disease 8(1), 331–359 (2013)
10. Golden, J.: Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. JAMA 318, 2184 (2017)
11. Jiang, K., Liao, Q.M., Dai, S.Y.: A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. vol. 5, pp. 2820 – 2825 Vol.5 (2003)
12. Joshi, M.M.D., Karode, A.H., Suralkar, P.S.R.: White blood cells segmentation and classification to detect acute leukemia (2013)
13. Kim, K., Jeon, J., Choi, W., Kim, P., Ho, Y.S.: Automatic cell classification in human's peripheral blood images based on morphological image processing. vol. 2256, pp. 225–236 (2001)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60 – 88 (2017)
15. Macawile, M.J., Quinones, V., Ballado, A., Cruz, J., Caya, M.V.: White blood cell classification and counting using convolutional neural network. pp. 259–263 (2018)
16. Ramoser, H., Laurain, V., Bischof, H., Ecker, R.: Leukocyte segmentation and classification in blood-smear images. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. pp. 3371–3374 (2005)
17. Ravikumar, S.: Image segmentation and classification of white blood cells with the extreme learning machine and the fast relevance vector machine. Artificial Cells, Nanomedicine, and Biotechnology 44(3), 985–989 (2016)
18. Ruberto, C.D., Loddo, A., Putzu, L.: A leukocytes count system from blood smear images segmentation and counting of white blood cells based on learning by sampling (2016)
19. Sajjad, M., Khan, S., Jan, Z., Muhammad, K., Moon, H., Kwak, J.T., Rho, S., Baik, S.W., Mehmood, I.: Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities. IEEE Access 5, 3475–3489 (2017)
20. Salah, H., Muhsen, I., Salama, M., Owaidah, T., Hashmi, S.: Machine learning applications in the diagnosis of leukemia: Current trends and future directions. International Journal of Laboratory Hematology 41 (2019)
21. Sanei, S., Lee, T.: Cell recognition based on pca and bayesian classification. In 4th International Symposium, ICA (2003)
22. Sarrafzadeh, O., Rabbani, H., Talebi, A., Banaem, H.: Selection of the best features for leukocytes classification in blood smear microscopic images. vol. 9041 (2014)
23. Scotti, F.: Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. pp. 96–101 (2005)
24. Segura, M., Rivero, R., Suárez, V., Machado, M., Martínez, E., Otero, A., Abraham, C., Hernández Ramírez, P.: Inmunofenotipaje en el diagnóstico de síndromes linfo y mielo-

proliferativos. Revista Cubana de Hematologia, Inmunologia y Hemoterapia 16, 198–205 (2000)
25. Shafique, S., Tehsin, S.: Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technology in Cancer Research & Treatment 17 (2018)
26. Teman, C.J., Wilson, A.R., Perkins, S.L., Hickman, K., Prchal, J.T., Salama, M.E.: Quantification of fibrosis and osteosclerosis in myeloproliferative neoplasms: A computer-assisted image study. Leukemia Research 34(7), 871 – 876 (2010)
27. Theera-Umpon, N.: White blood cell segmentation and classification in microscopic bone marrow images. vol. 3614, pp. 787–796 (2005)
28. Tizhoosh, H., Pantanowitz, L.: Artificial intelligence and digital pathology: Challenges and opportunities. Journal of Pathology Informatics 9(1), 38 (2018)
29. Tomasz, M., Leszek, M.: Analysis of features for blood cell recognition. pp. 42–45 (2004)
30. Topol, E.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25 (2019)
31. Tran, T., Vununu, C., Atoev, S., Lee, S.H., Kwon, K.R.: Leukemia blood cell image classification using convolutional neural network. International journal of computer theory and engineering 10, 54–58 (2018)
32. Vogado, L.H., Veras, R.M., Araujo, F.H., Silva, R.R., Aires, K.R.: Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification. Engineering Applications of Artificial Intelligence 72, 415 – 422 (2018)
33. Zhang, C., Xiao, X., Li, X., Chen, Y.J., Zhen, W., Chang, J., Zheng, C., Liu, Z.: White blood cell segmentation by color-space-based k-means clustering. Sensors (Basel, Switzerland) 14, 16128–16147 (2014)

ISSN 1870-4069

Prototype of a Recommendation System of Educative Resources for Students with Visual and Hearing Disability

Carmen Cerón, Etelvina Archundia, Beatriz Beltrán

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Puebla, Mexico

{mceron,etelvina,bbeltran}@cs.buap.mx

Abstract. The purpose of this article is to present the design of a recommendation system of educative resources for the students with visual or hearing disability to support the learning in the subjects of the Human Formation area. The recommendation is based in the perception of the utility of the educative resource through a vote using collaborative filtering and the K-nearest neighbor algorithm. For the implementation was integrated to a platform using PHP, MySQL, JavaScript and the library Mahout Apache. Finally, are presented the results obtained of the tests realized to a focal group of students with disability and the evaluation of the system.

Keywords. Usability, semantic web, inclusive learning, learning objects repository.

1 Introduction

The recommendation systems are applied in various areas such as the marketing, medicine, technology and education. In these systems the user realizes an object or item selection based on his or her needs and preferences, giving recommendations even to others. The design of adaptive recommendation systems to support the education and especially to users from different modalities as on-site or e-learning actually require of digital resources that allow uphold the learning processes according to their academic preferences and requirements of the students even with any disability.

Which entails to promote materials and digital resources that underpin the learning process in the programs that are offered in the distinct modalities, where it seeks to intensify the learning strategies and auto learning in order to obtain better results in the academic performance of the students. As well as an inclusive approach for people with disabilities.

This investigation has as purpose the design of a recommendation system of accessible learning resources like learning objects or digital resources designed to support the students with disabilities learning in Human Formation area.

The Recommendation System (RS-OLR) uses the collaborative filtering technique and the K-nearest neighbor algorithm, and to find the similarity between users with the objective to find a major quantity of resources in tune with their needs. The system is a search and recommendation tool of accessible learning resources to the subjects of the university human formation in the learning programs of BUAP in on-site or e-learning modality.

The article is organized as is indicated next: In the section 2, are presented the foundation and theoretical review of the research work. In the section 3, is defined the methodology for the analysis and the recommendation system design. In the section 4, are shown the results of the pilot test of the focal group of students with disability. Finally, are presented the conclusions and future work of the research.

2 Related Work

In this section are reviewed the topics of open learning resources, the learning objects, repositories and the recommendation systems that give support this research.

2.1 Open Learning Resources, Learning Objects and Repositories

The Open Learning Resources (OLR), are designated as materials for the teaching-learning, since 2002 the UNESCO [1], pronounced in favor of the generation of these resources, which include:

- Learning contents: Complete courses, materials for courses, modules, contents, learning objects, etc., which are didactic.
- Tools: Software for the creation, delivery, use and improvement of the open learning content, including searching and organization of the content, learning management systems (LMS), tools of content development, and communities of e-learning.
- Resources of implementation: Copyright licenses that promote the open publication of materials, design principles and local adjustment of content.

All of this also entails an inclusion process therefore the subjects must be accessible for all people and achieving a universal design, but still require to work on accessible materials.

According to IEEE “A Learning Object (LO) is any entity, digital, which can be used, re-used or referenced during technology supported learning” [2]. Such objects are utilized in e-learning, interactive and collaborative environments. Examples of LO include a lesson, multimedia contents, a video, simulations, and animations.

The Learning Objects are used as a digital learning resource that seeks to favor the access to learning contents and facilitate the learning to the specific development of the disciplinary competence.

They are composed of two part: the content and on the other the metadata, being esencial to be able to integrate to a repository of LO, that is to say a LOR.

The Learning Objects have as functional requisites: accessibility, reutilization and interoperability. In addition, they possess advantages over other digital learning resources such as manageability, durability and scalability [3].

The Learning Objects possess metadata that describe and identify the learning resources and facilitate their searching and recovery. Currently, exist various standards of metadata, which look to easy the exchange and accessibility of the learning objects amid the most used are the IEEE-LOM, DublinCore, Can Core y OBAA. The LO require an instructional design to guarantee to promote the learning by the dimensions: pedagogical, didactic and technological [4]. Concerning the Learning Objects Repositories are specialized digital libraries, that lodge multiple types of educative resources of learning objects and/or its metadata, which are used in different e-learning environments that help the teaching-learning process [5]

The LOR arise to establish a solution that would allow the users find easily digital resources oriented to education, solving the problems of content management and can establish norms for the quality of its resources, also must offer facilities to the creation, management of the metadata and stored resources [6].

2.2 Recommendation Systems

The Recommendation Systems (RS) emerged in the middle of '90s (Li, 2010) with the objective of toast the users those results of searches of information near to their necessities [7].

For Chesani [8] defines a recommendation system as "that one able to realize predictions starting from the fact that a user likes or do not an item, object, mean, or information that would access". A RS is associated with a group of items, where I = i1,...,in and its objective is to recommend to the users items of I that can be of their interest. The system is able to treat each user individually, bias a product or service to attend the necessities, is a priority form to improve the quality of the searches.

The recommendation systems are the allies of the personalization of computational systems, principally in the web, by its ability to identify preferences and suggest relevant items for every user: for which is needed of profiles that store the information and preferences of every user [9]. Such is the case of the open learning resources as the LO s that have associated metadata to be able to identify and classify. According to [10] the recommendation systems are classified in: RS Based on the Content, Collaborative RS, and Hybrid RS.

Recommendation Systems based in Collaborative Filtering The collaborative RS (C-RS) have two classifications: based on memory (employ similarity metrics to determine the similarity between a pair of users and calculate the items that have been voted for both users and compare such votes to calculate the similarity) and based on models (utilize the matrix of votes to create a model through which establish a group of similar users to the active user) and other classifiers for this category are the Bayesian, clustering and based in regressions models.

The algorithms that often are used to implement the collaborative filtering techniques based on memory, are methods based on neighborhood (K-nearest neighbor algorithm). These function selecting an appropriated group of users, according the similarity of the same regarding the active user, and use the values of such users to generate the values of the active user.

User Profile and Student Model In the case of a user profile of a software system, this can comprehend both personal data, behavior patrons, personal interests and preferences [11]. The profiles created explicit or implicit, regarding to understand how such characteristics of the student that are relevant in the educative process, and the interrelation within these, as they are the level of comprehension of a topic, learning styles, users likes, stratum, psychological characteristics like state of mind, the goals and at least their environment or context. All this characteristics related between them form what is called the student model, which allows to construct a profile and model the system to the user needs.

2.3 Related Works

The different contributions about the recommendation systems of learning objects in repositories, such as:

In the National University of Colombia [12], give an input about the development of adaptive smart courses, using a tutorial system and proposing a model of seven modules to search, recover and recommend learning objects, to support the learning virtual environments.

Likewise, there exist the works based on the recovery of thematic content and of the defined context [13] in the different repositories, which have permeate in support the education at distance.

Another investigation of Colombia [14] has applied the collaborative filtering technique to value the perception of the use of learning objects by the students as a support to their activities and the results favor the recommendations by an adaptation of the K-nearest neighbor algorithm.

The contribution of recovery systems of LO in repositories [15] have attached to the use of learning styles, the processing of natural language, ontologies and the hybrid recommendation systems have facilitated the management of the LO, but lack of the description of the real use of users.

According to Cechinel et al. [16] in the RS developing utilizing the library Apache Mahout to calculate recommendations of learning resources generated by different algorithms of collaborative filtering based in memory applied over sets of obtained data of the MERLOT repository, where more than 3500 resources available for the e-learning.

3 Methodology

The methodology in this research is exploratory with a qualitative approach, for what was applied the model of prototypes and design centered in the user for the proposal of the recommendation system.

3.1 Analysis and Design of the Proposed Recommendation System

The architecture client-server of three layers to integrate LMS Moodle and the system of recommendation called "SR-Recursos Educativos Accesibles Universitario" (SR-REAU). The objective of the proposal is to develop a recommendation system of digital accessible learning resources like are the LO or other resource, according to the student with disability profile and learning results. The recommendation system has at least two main tasks: prediction and recommendation. In Figure 1, is showed the architecture of the system and regarding the design of RS is based on:

- The representation of the recommendations: through the evaluation of the contents by a unique value (like or do not) and the Likert scale 1 to 5, regarding to the utility to learn.
- Regarding the evaluation are added over the items to generate and classify the recommendations and finally is showed an ordered list, according to the profile of the student with disability, with the data: tuition, name, last name, user, code, e-mail and career. Likewise level or degree of visual or hearing disability.

3.2 Development of the Recommendation System

For the implementation it was used the collaborative filtering technique and the K-nearest neighbor algorithm through the Mahout Apache library. The recommendation system was developed in PHP, HTML, Javascript and the database MySQL, in a Linux server with Apache. The database stores information of the educative resources which can be digital resources: pdf files, documents, videos, LO or any digital educative resource, where the principal objective is to store a valuation of the resource (see Figure 2).

For which, the four steps that were implemented to achieve the recommendations in the system of collaborative filtering are the following:

1. Calculate the similarity between users: In first place, is inserted the user profile: id_student, disability, id_resource and the valuation that gave the student, this will allow to determine through the metrics, the similarity between a pair of users.
2. Calculate the K-neighbors of the active user: Making use of the selected metric of similarity, in the case of the API of Apache Mahout counts with: Pearson correlation similarity, Euclidean distance similarity, among others, what is calculated is the similarity between pair of users, the occupied metric

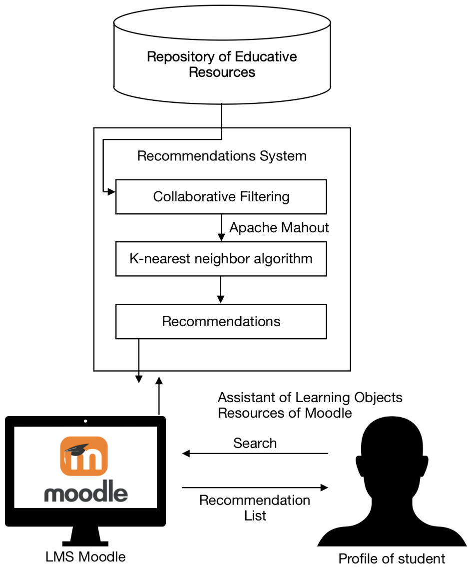


Fig. 1. Architecture of the recommendation system SR-REAU in three layers.

was the coefficient of Pearson coefficient similarity. So that are obtained the k users more similar to the active user. These users are called as the k-neighbors of the active user.

3. Calculate the predictions of the items: Selecting the subset of users which valuations are going to be used and, so will have influence in the generation of the prediction for the active user. From the k-neighbors of the active user, are determined the possible valuations that the active user would do about the items not yet voted, that is to say, is predicted how the user would value that items, for which in Apache Mahout it has various algorithms to calculate the neighbors closer to the active user where was selected the k-nearest neighbors methods, being this the one who reduces the effect of noise in the classification, but creates limits between similar classes.
4. Show N better recommendations: Behind the calculation of the predictions, is chosen N items more suitable to be recommended to the user, that is to say, the higher, more voted predictions, held.

In Figure 3, it is visualized the options of the Recommendation System of University Accessible Educative Resources where the user registers, searches educative resources for the subject and subsequently evaluates. Likewise consults the retting list of the recommended resources.

In Figure 4, is showed the list of the results of the search and after the selection of an educative resource for people with disability, the recommendation system will allow the user to order them according to the criteria: use and evaluation.

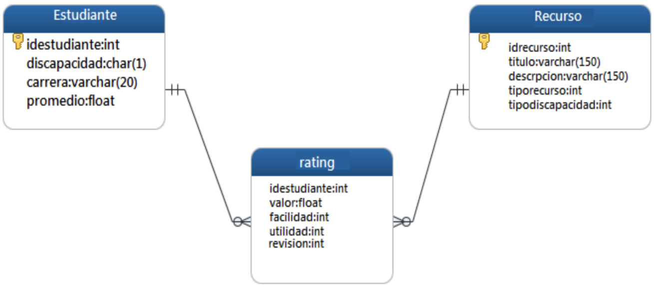


Fig. 2. Relation of the database for the vocation of educative resources.

Fig. 3. Prototype of the Recommendation System-REAU.

In order to evaluate, in Figure 5, users cast their vote according to three criteria, in a Likert scale of 1 to 5, where 1 is "Nothing", 2 is "Little", 3 is "Regular", 4 is "Quite" and 5 is "Totally": spacing

- Value, which represents a general grade of the resource.
- Academic relevance, if it was useful for the learning.
- Ease of use, refers to the usability in general to navigate and use.

4 Tests and Results

4.1 Pilot Test

For the test of the system, it was applied the technique of inspection and inquiry with the user. The participants of the test were selected according to three

Fig. 4. List of recommendations of the system.

criteria: a) be a student, b) have some hearing or visual disability and c) be in a course of the area of university human formation.

For which, it was conformed a focal group of 6 users with visual and hearing disability: two students with medium hearing disability, as well as three with deafness, from the Degree in Graphic Designing and one student with blindness from the Degree in Law. Regarding the functionality tests, were presented two possible scenarios for the final users, hereunder, are described:

- Situation 1: To the user was given a brief explanation of the use of the system and was accompanied in every activity actively.
- Situation 2: To the user was explained the use of the system and only was accompanied in the beginning of a selected activity and only in a much needed case.

For every situation the users must accomplish certain tasks as are:

- Task 1. Register as user of the system and find the navigation menu and select an option.
- Task 2: Realize a search and select the item of the interest of the user and value the educative resource
- Task 3. Consult the list of recommendations and value again.

The results demonstrate that the usability of the recommendation system to support people with disability is intuitive and easy to manage, achieving to interact according to their disability, because is included a plug-in “webkit-speechrecognition" which allows to listen and interact with the system to the users with visual disability.

Fig. 5. Review for the value of the educative resource (item).

Table 1. Obtained results of the pilot test.

Tasks	Situation 1	Situation 2
No. 1	100%	95%
No. 2	100%	95%
No. 3	95%	90%
Average	98.33%	91.6%

4.2 Evaluation of Metrics

The metrics of accuracy or decision, evaluate the certainty of the prediction system and value if the recommendations are appreciable and effective for the user.

According to Cleverdon [17], there are five metrics to be considered:

1. Delay. Period of time passed since is done the demand until is given the answer.
2. Presentation. The physic format of the output of the system.
3. Exhaustiveness. Capacity of the system to present the relevant items.
4. Precision. Capacity of the system to hide items which are not relevant.

For this test, only was used two metrics that help to measure the effectiveness regarding the list of recommendations as Precision (P) and Recall (R). These are calculated from the table of contingency which classifies the items regarding to the necessities of information distinguishing two groups: relevant o no relevant of a total of 237 educative resources in the system.

In addition, also the items are classified according to how they have been recommended to the user (selected) or not (not selected). For which was taken the data of the six users U= u1, u2, u3, u4, u5, u6 with their values of 1 to

5 that are assigned to the educative resources in different sessions during the scholar semester. Hereunder, is presented the Table of contingency (Table 2), which shows the proportion of selected relevant items (Nrs=178) regarding to the total of selected items (Ns=200), that is to say, measures the probability that a selected item would be relevant. Moreover, the exhaustiveness or recall, use the proportion of selected relevant items (Nrs=178) regarding to the total Nr =190:

$$P = \frac{Nrs}{Nr} = 0.89. \tag{1}$$

Table 2. Table of contingency of precision and recall of the system.

	Selected	Not selected	Total
Relevant	178	12	190
Irrelevant	22	25	47
Total	200	37	237

In (1), it is indicated that the probability of the system to recommend an educative resource would be relevant is of 8% for the student meanwhile that the exhaustiveness (2) os of 0.93, the probability that the system presents the relevant items recommended to the user:

$$R = \frac{Nrs}{Nr} = 0.93. \tag{2}$$

5 Conclusions and Future Work

One of the principle contributions was the design and development of the recommendation systems applying the technique of collaborative filtering using the API Apache Mahout, PHP, Ajax, JavaScript, HTML5, achieving to apply the K-nearest neighbor algorithm and calculate the similarity measures to ease the searches and processes of recovery of information with major precision to deliver or recommend trustable educative resources and useful for the learning of the students according to the hearing or visual disability, achieving to strengthen digital abilities and promote relevant and significant searches that allow to support the subjects of Human Formation and the disciplinary competences with accessible digital materials so that the student with disability has access to the resources according to the profile and utility that perceives for the learning, giving a valuation to the resource, and by this way to strengthen the learning processes in face-to-face or distance learning.

The work to future, is to amplify the sample and incorporate other algorithms and strategies of evaluation to improve the recommendations and predictions with major depth in the system. Finally, the trend of the recommendation systems is total and immerse in the educative systems in all the modalities as a tool to support the quality of the didactic digital resources and materials, for which is demanded websites or repositories that generate recommendations more conscious and complete to support the education with new strategies and scenarios in the high level.

References

1. Miao, F., Mishra, S., McGreal, R.: Open educational resources: Policy, costs and transformation, United Nations Educational, Scientific and Cultural Organization, pp. 1-3. UNESCO (2016)
2. IEEE Learning Technology Standards Committee, LTSC http://ieeeltsc.org/
3. Garcia, A.: Objetos de aprendizaje: características y repositorios. BENED (2005)
4. Betancur, C., Moreno, C. J., Ovalle, C., Demetrio, A.: Modelo para la recomendación y recuperación de objetos de aprendizaje en entornos virtuales de enseñanza/aprendizaje. Revista Avances en Sistemas e Informática, 6, pp. 45–56 (2009)
5. Downes, S.: The learning marketplace. Meaning, Metadata and Content Syndication in the Learning Object Economy. Moncton (2004)
6. Alfano, C., Henderson, S.: Repositories. In Lea, In Learning Objects for Instruction: Design and Evaluation, pp. 16–28 (2007)
7. Casali, A. Godo, L. Sierra, C.: Modelos graduados de BDI para arquitecturas de agentes. En: Leite J., Torroni P. (eds) Lógica computacional en sistemas de múltiples agentes. CLIMA04. Lecture Notes in Computer Science, 3487, pp. 126–143 Springer, Berlín, Heidelberg (2005)
8. Chesani, F.: Recommendation Systems. Corso di laurea Ing. Inform, pp. 1–32 (2004)
9. Cazella, S., Reategui, E., Nunes, A: Ciencia de Opinion: Estado da arte en Sistemas de Recomendação. JAI: Jornada de Atualização em Informática da SBC. Rio de Janeiro, pp. 161–216 (2010)
10. Bafoutsou, G., Mentzas, G: Review and Functional Classification of Collaborative Systems. International Journal of Information Management, 22, pp. 281–305 (2002)
11. Dagostino, E., Casali, A.: Sistema de Apoyo al Aprendizaje Diagnóstico Utilizando Perfiles de Usuario: EndoDiag EIU.UG.ES, pp. 1–14 (2005)
12. González, G., Duque, D., Ovalle, D.: Modelo del Estudiante para Sistemas Adaptativos de Educación Virtual. Revista Avances en Sistemas e Informática, 5, pp. 199–206 (2008)
13. Deco, C., Bender, C., Saer, J.: Sistema recomendador de recursos educativos para la enseñanza de las ciencias. Energia, 8(8) (2010)
14. Capuano, N., Iannone, R., Gaeta, M., Miranda, S., Ritrovato, P., Salerno, S.: Un sistema de recomendación para objetivos de aprendizaje. Sistemas de información, E-learning, y Knowledge Management Research, pp. 515–521, Springer (2013)
15. Cechinel, C., Sánchez-Alonso, S.: Analyzing associations between the different ratings dimensions of the MERLOT repository. Interdisciplinary Journal of E-Learning and Learning Objects, 7, pp. 1—9 (2011)
16. Cechinel, C., Sicilia, M.Á., Sánchez, S., García, E.: Evaluating collaborative filtering recommendations inside large learning object repositories. Information Processing& Management, 49(1), pp. 34–50 (2013)
17. Cleverdon, C., Keen, M.: Factors Determining the Performance of Indexing Systems, 2, Test Results. ASLIB Cranfield Res. Proj., Cranfield, Bedford, England (1966)

ISSN 1870-4069

Analyzing Students' Performance in a Mathematics Course Sequence using Educational Data Mining

Beatriz González Beltrán, Silvia González Brambila, Lourdes Sánchez Guerrero, Josué Figueroa González

[1]Universidad Autónoma Metropolitana, Unidad Azcapotzalco, Mexico City, Mexico

{bgonzalez,sgb,lsg,jfgo}@azc.uam.mx

Abstract. One of the main concerns in institutions of higher education is the time their students need to finish their studies. Students tend to invest more time than the established in approving a particular course or a group of them. This problem is more evident when there exists a sequence of related courses that are related through a prerequisite schema, because it is not clear whether the knowledge acquired in a previous course is appropriate for the next one. This work considers this problem and presents the use of Educational data mining for determining if there is an influence of some academic performance aspects in previous courses over the tries needed for approving a later one in a sequence of mathematics subjects. The performance of different measures for the predictive generated models using decision trees was not very high, and results shown that considering only the immediate previous course before the analyzed had little better results than considering all the ones in the sequence. Analyzing all the course sequence, many of the courses had as essential variables the performance not in the immediate, but two or three previous ones.

Keywords. Courses performance influence, educational data mining, information processing, prerequisite courses' influence, sequence of course analysis.

1 Introduction

Educational institutions generate a lot of information about several aspects of their operation that can be analyzed for finding patterns that lead to solve problems or making decisions about a certain aspects. One of the principal concerns in any educational institution is related to the performance of students across their studies. This is an aspect that concerns different authorities which are responsible of academic topics. Academic performance in a superior education institution can be measured in different ways like the percentage of graduated students, time invested in finishing a bachelor's degree and so on.

Also, the academic performance of students can be affected by several reasons: personal, labor, and academic which can be related to low preparation, study habits, even how students take their courses can be an influential factor for having a good or bad performance.

Discovering knowledge from educational data is an area that has been applied for more than two decades. The main goal of this area, known as Educational Data Mining (EDM) [4], is improving academic performance, reducing the failure rate in the studies, a particular topic or even in a single exam [3]. A way for achieving this is related with predicting student performance based on several personal or academic characteristics.

A concern for superior educational institutions is the time invested by students in finishing their university careers. Many students need twice or more the time for finishing, this represents both an operational and financial problem for institutions. One of the reasons for this is related to the time needed for approving one or more topics in a study plan because students invest more time than the established for approving a group of topics, primarily when these topics are related through a prerequisite schema, where students can not take a course if they have not approved the previous one.

A prerequisite schema is established considering that in a chain of courses, the previous ones will prepare students to take the next ones succesfully. It is considered that having a good performance in a previous course will imply having a successful result in the later one, however, this is not always true, there could be other factors that influence in the results obtained in later courses and not only the obtained mark in the previous ones.

Also, the results in, for example, a first course could have a certain impact not only in the immediate next one, but in two or more courses forward.

In the university where this study was performed, it is common that students invest more than twice the time needed for approving a group of topics, especially in the mathematical chain of courses which is composed by five with a prerequisite schema. Also, the way in which students can register their topics allows them not taking the next course in the following scholar period even if they have approved the previous one. This causes that many students do not take a course the next scholar period after approving the previous one, this could also be a factor that affects their performance in the following course.

It is necessary an analysis that considers several aspects of students' performance in a course for determining if these have an impact in the results obtained in later ones.

The goal of this work is to determine a relationship between the students' performance and the math courses sequence in engineering programs. Considered aspects involve:

- Approving mark.
- Amount of tries for approving.
- Number of scholar periods needed for approving.
- Time passed between approving a previous course and taking for first time the next one.

The performance measure is related to the number of tries to approve each course in the sequence more than the obtained mark. Former variables are used for predicting the number of tries that students will need for approving a later course after approved a previous one.

The results of this work could help to guide the student in choosing the moment of taking a course considering the performance over the previous ones. Also, it could help to determine if academic aspects have a student impact on the performance and even analyze the pertinence of the prerequisite schema in a study plan.

The structure of this paper is as follows: Section 2 reports some related works about predicting the academic performance of students considering different aspects. Section 3 presents the application of a data mining process for analyzing the data. Section 4 presents the analysis of the obtained results. Finally, section 5 offers the conclusions and future works.

2 Related Works

Predicting student's performance in a course or through its studies is one of the most studied topics in EDM. Many works focus in predicting the results of students considering personal, social or academic factors.

In [1], the authors use data mining to predict secondary school student performance in Mathematics and Portuguese courses. The database was build from school reports (grades and number of absences) and questionnaires with closed questions related to demographic, social, and school-related variables. They had three different DM algorithms: binary, 5-level classification, and regression. They applied a Naive Predictor (NV) as a baseline comparison and four DM methods: Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM). They obtained a high predictive accuracy when the first and second school period grades are known (91.9% with NV and 93% with DT on Mathematics and Portuguese, respectively). Moreover, the number of past failures is the most important factor when no student scores are available. There are other relevant factors, such as the number of absences, extra school support, travel time, the mother's job, going out with friends, and alcohol consumption.

In [6], is proposed a model to monitor student progress in e-learning systems to predict performance, progress, and potential. A student is represented by a Student Attribute Matrix, with attributes for performance and non-performance. Was considered learning styles, learning models, and Bloom's Taxonomy. To analyze the potential of a student was proposed a descriptor with higher-level attributes of performance and non-performance attributes. For the student performance estimator and the attribute causal relationship indicator was used a Back Propagation Neural Network. Results show correct and accurate student progress in high school students. Using this tool is possible to generate feedback indicators to understand and improve the performance of the students.

A comparative study on the effectiveness of educational data mining techniques to early predict student failure rate in introductory programming courses is presented in [2]. The authors tried to reduce the failure rate identifying problems at early stages and performing data preprocessing and algorithms fine-tuning. The comparison included Neural Networks, Decision Trees, Support Vector Machine (SVM) and Naïve Bayes, using an F score to evaluate the effectiveness. The SVM technique outperformed the other ones. The data sources were from Brazilian Public University in distance and on-campus education, with 21 and 16 characteristics, respectively.

In [5], the authors designed a classification model by utilizing data mining techniques for predicting the likelihood of a student to pass the Licensure Examination for Teachers (LET). The authors compared five data mining techniques (Neural Network, Support Vector Machine, C4.5 Decision Tree, Naïve Bayes, and Logistic Regression). This work identified the predictive variables on measuring the student's performance in the LET at the Bulacan Agricultural State College. As a result of the analysis, C4.5 Decision Tree had the highest accuracy of 73.10% followed by Neural Network with 65.67%. For that reason the authors used this algorithm for design the classification model. Using the model, the Bulacan Agricultural State College could be able to identify students who will likely fail the LET. These students will be given higher priority during their mock board review and be able to pass the board examination. Studies show that mock board performance has a significant relationship in passing the board examination.

3 Looking for Relationships between Courses

For determining if there was a relationship between the performance in a previous mathematics course and the next one, was used the data mining methodology which involves the general stages of: data exploration and preprocessing, modeling and analysis of the results. The last stage is presented in Section 4 of the document.

3.1 Data Exploration and Preprocessing

Processed data were obtained from the students' marks database, known as *kardex*, which contains information about the courses that students have taken during their studies. These data include student id, course id, obtained mark (approved or not), and scholar period when the course was taken. Were used the information of the five courses of the mathematical branch:

- Mathematics Workshop (MW),
- Essential Calculus (EC),
- Differential Calculus (DC),
- Integral Calculus (IC),
- Differential Equations (DE).

All of these courses have a prerequisite schema, meaning that for taking one, it is necessary to have approved the previous one. The relationship between these courses is shown in Figure 1.

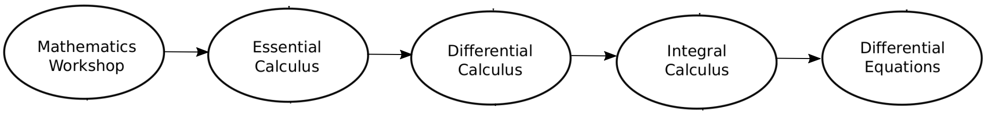


Fig. 1. Mathematics courses sequence.

Information used for evaluating the performance in each group, as mentioned in the Introduction section, was:

- Obtained mark: the mark with the one the student approved the course.
- Number of tries needed for approving a course: amount of tries the student needs for approving the course.
- Scholar periods invested in approving a course: total of scholar periods needed for approving a course, consider that not necessarily correspond to the number of tries. A student could have approved in their second try, but invested more scholar periods.
- Number of scholar periods passed after approving one course and taking for the first time the next one: it is common that students do not take the later course the next scholar period after approving the previous one.

Table 1 presents the variables names, their description and domain.

For the study, were considered engineering students from all the study plans offered by the university which already have approved the five courses, the total of processed students was 2,119.

Kardex file only contains obtained marks and scholar periods, so it was necessary to process data for obtaining the first scholar period when the student took the course and calculate the number of periods invested in approving it. The number of tries was obtained, for each course, counting the total of non-approving marks obtained plus the one when it was approved. The time elapsed between approving one course and taking the following was measured considering the scholar period when a previous course was approved and the one in which the student took the following course for the first time.

3.2 Modeling

Were used decision trees for analyzing the information, the predicted variable was the total of tries needed for approving each course. Were considered two scenarios, using only the performance in the previous course, and considering the performance in all the previous courses according to Figure 1. In this way, were analyzed the following relationships:

- Relationship 1. TEC $\sim$ (MMW, PNMW, TMW, TPMW),
- Relationship 2. TDC $\sim$ (MEC, PNEC, TEC, TPEC),
- Relationship 3. TIC $\sim$ (MDC, PNDC, TDC, TPDC),
- Relationship 4. TDE $\sim$ (MIC, PNIC, TIC, TPIC).

Table 1. Description and domain of used variables.

Variable name	Description	Domain
MMW	Mathematics Workshop mark	Very Good, Good, Sufficient
TMW	Number of tries for approving Mathematics Workshop	1, 2 or more
PNMW	Number of periods needed for approving Mathematics Workshop	1 to 13
TPMW	Number of scholar periods passed after approving Mathematics Workshop and studying Essential Calculus	1 to 15
MEC	Essential Calculus mark	Very Good, Good, Sufficient
TEC	Number of tries for approving Essential Calculus	1, 2 or more
PNEC	Number of periods needed for approving Essential Calculus	1 to 17
TPEC	Number of scholar periods passed after approving Essential Calculus and studying Differential Calculus	1 to 13
MDC	Differential Calculus mark	Very Good, Good, Sufficient
TDC	Number of tries for approving Differential Calculus	1, 2 or more
PNDC	Number of periods needed for approving Differential Calculus	1 to 15
TPDC	Number of scholar periods passed after approving Differential Calculus and taking Integral Calculus	1 to 6
MIC	Integral Calculus mark	Very Good, Good, Sufficient
TIC	Number of tries for approving Integral Calculus	1, 2 or more
PNIC	Number of periods needed for approving Integral Calculus	1 to 16
TPIC	Number of scholar periods passed after approving Integral Calculus and taking Differential Equations	1 to 11
TDE	Number of tries for approving Differential Equations	1, 2 or more

Considering the sequence of previous topics for each course (EC only has one previous course), the analyzed relationships were:

- Relationship 5. TDC $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC),
- Relationship 6. TIC $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC, MDC, PNDC,TDC, TPDC),
- Relationship 7. TDE $\sim$ (MMW, PNMW, TMW, TPMW, MEC, PNEC, TEC, TPEC, MDC, PNDC,TDC, TPDC, MIC, PNIC, TIC, TPIC).

Decision trees were generated using the CART algorithm and a ten-fold cross-validation was used for each relationship. Although the number of tries has values from 1 to 5 (5 is the maximum of tries for approving a course in the university), most of the cases correspond to 1 or 2. The first results considering the five categories (one per try) have an accuracy less than 40%. For this reason, 2 to 5 tries were considered as a single category: 2 or more (2M). Besides the accuracy of prediction, also was measured the importance of each variable for each relationship.

The performance of the model was measured using accuracy, precision, recall and F_1 score metrics based on confusion matrices. Instead accuracy is the most intuitive measure, it is only valid when there are symmetric data sets, in this case, the amount of students with 1 try was greater that those ones with 2, 3, 4 or 5 (2 or more), so accuracy did not give a good measure of the results. other measures offers a better measurement of model performance where there is an unbalanced class distribution.

3.3 Obtained Results

Accuracy of the predictive models considering only the relationship between a course and the previous one are presented in Table 2.

Table 2. Performance measures for immediate course relationships.

Relationship	Accuracy	Precision 1 try	Precision 2M tries	Recall 1 try	Recall 2M tries	F_1 score 1 try	F_1 score 2M tries
1	0.652	0.790	0.431	0.689	0.563	0.7365	0.487
2	0.650	0.878	0.230	0.677	0.508	0.764	0.316
3	0.656	0.905	0.159	0.686	0.441	0.778	0.226
4	0.659	0.859	0.271	0.696	0.499	0.769	0.348

The importance of each variable for the relationships that only considers the immediate previous course performance are:

- For relationship 1: MMW 51.709, PNMW 33.456, TMW 10.686, and TPMW 8.453,
- For relationship 2: MEC 13.491, PNEC 24.588, TEC 14.951, and TPEC 8.116,
- For relationship 3: MDC 9.211, PNDC 32.812, TDC 22.904, and TPDC 2.806,
- For relationship 4: MIC 20.39, PNIC 30.760, TIC 25.383, and TPIC 8.742.

Figure 2 shows an evolution of the importance of the variables mark, periods needed, tries and time elapsed at the moment of predicting the total of tries needed for approving a course considering only the immediate previous one.

Table 3 shows the results of the models generated considering the influence of all the courses that compose the sequence before the predicted one.

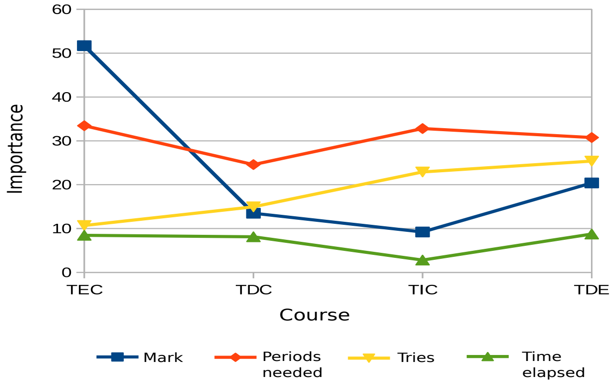


Fig. 2. Evolution of crietria importance in the prediction.

Table 3. Performance measures for all courses sequence relationships.

Relationship	Accuracy	Precision 1 try	Precision 2M tries	Recall 1 try	Recall 2M tries	F_1 score 1 try	F_1 score 2M tries
5	0.646	0.824	0.284	0.701	0.445	0.757	0.346
6	0.629	0.806	0.284	0.686	0.429	0.741	0.341
7	0.624	0.730	0.366	0.704	0.446	0.726	0.4

The importance of each variable was also measured for relationships that consider not only the previous course but also the ones that compose the sequence until the studied one. Table 4 presents the importance of variables across the complete sequence, consider that the importance of zero represents that a variable corresponds to a course that is taken after the one analyzed.

Figure 3 presents a radar chart showing the importance of each variable in the analyzed courses of the mathematical course sequence.

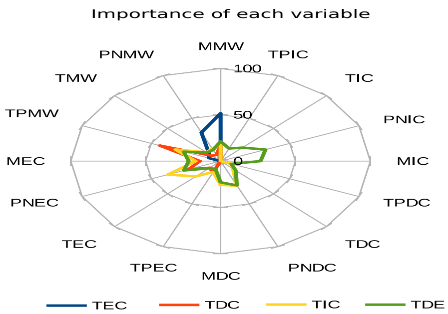


Fig. 3. Importance of all variables in the complete courses sequence.

Considering that predicted value for Relationship 2 was the same that for

Table 4. Importance of each variable through the courses sequence.

Variable	TEC	TDC	TIC	TDE
MMW	51.709	17.222	19.520	21.352
PNMW	33.465	7.190	14.349	13.102
TMW	10.686	8.267	13.439	12.921
TPMW	8.453	44.012	32.664	26.652
MEC	0	13.710	16.553	21.271
PNEC	0	23.262	37.900	26.525
TEC	0	10.093	23.360	10.983
TPEC	0	13.394	12.673	9.025
MDC	0	0	25.704	22.910
PNDC	0	0	28.670	28.406
TDC	0	0	11.729	14.363
TPDC	0	0	8.123	7.738
MIC	0	0	0	26.693
PNIC	0	0	0	32.471
TIC	0	0	0	20.604
TPIC	0	0	0	14.729

Relationship 5 (TDC), similar for Relationship 3 and Relationship 6 (TIC) and for Relationships 4 and 7 (TDE), a comparison of performance measures obtained using both schema of relationship was performed, results are shown in Table 6.

In the same way, Table 5 presents the most important variable in the immediate course and in the complete course sequence relationship. Also the place that the most important variable in the immediate course relationship occupied considering all the courses sequence is presented in the last column. Consider that the TEC relationship is discarded because it is the same for immediate and full sequence relationship.

Table 5. Most important variable comparison for both schemes.

Variable	Immediate relationship	Full sequence relationship
TDC	PNEC	TDC
TIC	PNDC	PNEC
TDE	PNIC	PNIC

4 Results' Analysis

Results' analysis involved the analysis of the values of used performance measures of the obtained models for both relationships, the immediate course and the full course sequence. Also the importance of each variable in the prediction was analyzed.

Table 6. Comparison of immediate and full course sequence performance measures.

Measure	Rel. 2	Rel. 5	Rel. 3	Rel. 6	Rel. 4	Rel. 7
Accuracy	0.650	0.646	0.656	0.629	0.659	0.624
Precision 1 try	0.878	0.824	0.905	0.806	0.859	.730
Precision 2M tries	0.230	0.284	0.159	0.284	0.271	0.366
Recall 1 try	0.677	0.701	0.686	0.686	0.696	0.704
Recall 2M tries	0.508	0.445	0.441	0.429	0.499	0.446
F1 score 1 try	0.764	0.757	0.778	0.741	0.769	0.702
F1 score 2 tries	0.316	0.346	0.226	0.341	0.348	0.4

4.1 Immediate Course Relationship

From the relationships that involved only the immediate course, were obtained a low accuracy (65% as maximum), however this could be due to an unbalanced distribution in the classes, where most of them had the value of 1 try. Other measures had a better performance, precision which shows how many classes labeled as one try actually correspond to that value had the biggest value with 90%, which corresponds to the third relationship, the number of tries needed for approving Integral Calculus (TIC). Other relationships also had acceptable values with a minimum of 79%. However, the precision for classes labeled with 2M class was very poor with a maximum of 43% and a minimum of 15%. This shows that the generated model fails at the moment of classifying those students which need 2 or more tries for approving a course.

Recall value indicates the classes that correspond to a specific value and were labeled in the same way. Cases that correspond to the one try class did not obtain very good results, with values a little higher than the ones of the accuracy, a maximum of 69% and a minimum of 67%. For the classes labeled as two or more tries, results were also very low, with a maximum of 56% and a minimum of 44%.

For the F_1 score, which is a weighted average of recall and precision, the results can be considered barely appropriate, with a maximum of 77% and a minimum of 73% for the ones labeled as 1 try. Again, cases of 2 or more class, obtained poor results with a minimum of 22% and a maximum of 48%.

About the importance of each variable in the prediction of the total of tries needed for approving a course, as Figure 2 shows, the most important variable for the second course of the sequence, EC, was the mark obtained in MW. However, the obtained mark importance decreased through the sequence of courses.

Two variables shew a continual behavior, the number of scholar periods needed for approving and the number of tries. The number of scholar periods had values between 25 and 35 through all the sequence. Number of tries shew an increase across the course sequence.

The variable that represents the number of scholar periods passed after approving a course and taking the later one always had the least importance in the prediction.

4.2 Full Course Sequence Relationship

Results considering all the courses in the sequence had a similar behavior that the ones obtained with immediate course relationship, even, including all the courses in the sequence before the predicted one, reduced a little the value of the performance measures.

Accuracy of models had values from 62% to 64% which is considered low, again the best results were given by precision with a minimum of 73% and a maximum of 82% in the classification on cases labeled with one try. Similar to the former analysis, cases labeled as two or more tries had a low value. Recall measure also had better results for one trial class with a minimum of 68% and a maximum of 70% meanwhile cases labeled as of two or more tries had a minimum of 42% and a maximum of 44%.

F_1 score also shew a similar behavior than for immediate course relationship with a minimum of 72% and a maximum of 75% for one try class and 34% as minimum and a maximum of 40% for two or more classes.

As Figure 3 shows, for EC, the most important variable at the moment of prediction was MMW; for DC, the most important variable was not one of the previous course (EC), but the time passed after approving MW, followed by the scholar periods for approving the previous course EC and then, the mark in MW. This represents that two of the three most important variables were not from the immediate previous course, but from the previous one of this.

For TIC, only one of the three most important variables corresponded to the immediate previous course (PNDC), and the other two to different courses (MW and EC). Notice than the most important variable (TPMW) was from the first course of the sequence.

Finally, for TDE, two of the three most important variables corresponded to the previous course (IC), including the most important (PNIC), then one of DC and finally, another from the previous course (MIC). It is interesting that the fourth in importance, corresponded again to the first course of the sequence (TPMW).

4.3 Comparing Both Immediate and Full Sequence Courses Schemes

Table 5 shows that from the three relationships, two of them change their most important variable in the prediction. However, analyzing Table 4, was found that:

- PNEC, the most important variable in immediate relationship for TDC, in the full sequence relationship occupied the second place.
- PNDC, which was the most important for predicting TIC in the immediate relationship, had the third place in the full sequence one.
- PNDC, remains as the most important variable for both relationships in the prediction of TDE.

It is also of interest notice that in the immediate next course, three of four courses has as the most important variable the periods needed for approving the previous one. Also, in the three full course sequence relationships, the periods needed for approving a course (immediate previous or not) was the most important in all of them.

As values of Table 6 show, better results for almost all the performance measures were obtained considering only the previous course, meaning that adding more variables corresponding to the students' performance in all the previous courses of the sequence did not really help to improve the prediction of the number of tries.

5 Conclusions

The goal of this work was determining if there exists an influence of the academic performance in a course over the number of tries needed for approving the next one. For this, were considered several performance measures and tested two relationships, one considering only the immediate previous course and another which considered the complete sequence of previous courses. This, for a sequence of five mathematics courses.

According to the results of the different performance measures used for evaluating the predictive model generated using decision trees, the results did not show a convincing evidence that analyzed criteria had an important relationship with the amount of periods. Best results were obtained using precision as performance measure, but only for those students that approve in their first try.

Comparing the results of both relationships, considering only the immediate course obtained a little better results for almost all the performance measures.

Analyzing the importance of each variable, it is of interest noticing that the obtained mark in a previous course had not a significant relevance over approving the next course at the first try, only in one of the four relationships, this variable was the most important.

The variables which appeared more times as the most important ones were related to the number of scholar periods needed for approving a previous course, from the seven studied relationships, a variable related to this aspect, appeared in six.

Something remarkable is related with the time passed after approving a course and taking the next one, it was expected that at greater this time, the number of tries tend to rise, however, this variable always appeared as the least important in all the relationships.

A similar analysis for the complete courses sequence shows that it is common that variables related to previous courses more than the immediate one, had certain importance in the amount of tries needed for approving one. Interestingly, that all of the courses have a variable considered important which correspond to the Mathematics Workbench course, the first of the sequence, which at the university is a leveling course.

It can be said that considering only the performance in a immediate or all the courses of a sequence did not offer a definitive idea about how would be the performance of students in later ones.

Future works related to this study will be directed in obtaining better results for the performance measures of the models. A first option is using different predictive algorithms, however, it is considered that a bigger set of data could offer better results. Adding other characteristics like personal ones, or the behavior during the courses including for example, use of different platforms, amount of study hours, number of exercises, level of interaction with professors and other could help to model in a better way the behavior of the student, so that the models could generate better results.

References

1. Cortez, P., Silva, A.: Using Data Mining to Predict Secondary School Student Performance. In: Brito, A., Teixeira, J. (eds.) Proceedings of the 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008). pp. 5–12. EUROSIS, Porto, Portugal (Apr 2008)
2. Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., Rego, J.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior 73, 247–256 (2017)
3. Kumar, V., Chadha, A.: An empirical study of the applications of data mining techniques in higher education. International Journal of Advanced Computer Science and Applications 2(3) (2011)
4. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6), 601–618 (2010)
5. Rustia, R.A., Cruz, M.M.A., Burac, M.A.P., Palaoag, T.D.: Predicting student's board examination performance using classification algorithms. In: Proceedings of the 2018 7th International Conference on Software and Computer Applications. pp. 233–237. ACM (2018)
6. Yang, F., Li, F.W.: Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education 123, 97–108 (2018)

ISSN 1870-4069

Ontology Reusing: A Review

Cecilia Reyes Peña, Mireya Tovar Vidal

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
Mexico

reyesp.cecilia@gmail.com, mtovar@cs.buap.mx

Abstract. Building ontologies from scratch is a process that requires many resources as time, knowledge about the domain, among others. Sometimes ontologies are developed in the same domain with similar designs and the same propose causing a repetitive work. Ontology reusing allows us to build models with lower resources. There are several techniques for ontology reusing as: mapping, alignment, fusion, ontology integration, ontology networks, among others. In this paper, we analyze three techniques oriented to ontology reusing from the use of entire ontologies (no modularization).

Keywords. Ontology reusing, ontology mapping, ontology alignment, ontology networks.

1 Introduction

Many ontologies have been developed in order to build models that can facilitate the semantic interoperability. The term ontology is defined by Gruber [5]: *an explicit, formal specification of a shared conceptualization*, that has a defined and legible vocabulary to express main concepts and relationships about a specific domain [19, 13]. The ontologies allow the computers and humans understand the relationships about a domain using formal language [12].

Ontology reusing is based on builds models faster and at with a lower cost than build a model from scratch using traditional ontology design methodologies [17]; this task increases the interoperability of involved ontologies [22]. In general, ontology reusing belongs to the Ontology Learning area, and is based on building a new ontology by assembling, extending, specializing and adapting other ontologies [10].

There are techniques for ontology reusing as: mapping, alignment, fusion, ontology integration, ontology networks, among others. In order to select an optimal technique for building a model from ontological resources, we must consider the answer of the next questions:

- Does model include different domains?
- Is there an ontology for each model component?
- Is it necessary design ontologies from scratch for complementing the model or only some components (relationships, axioms, etc.)?

Further, there are mismatches for considering in order to a correct ontology reusing, some of them are classified in according to differences from design and construction stages (see Fig. 1)[1].

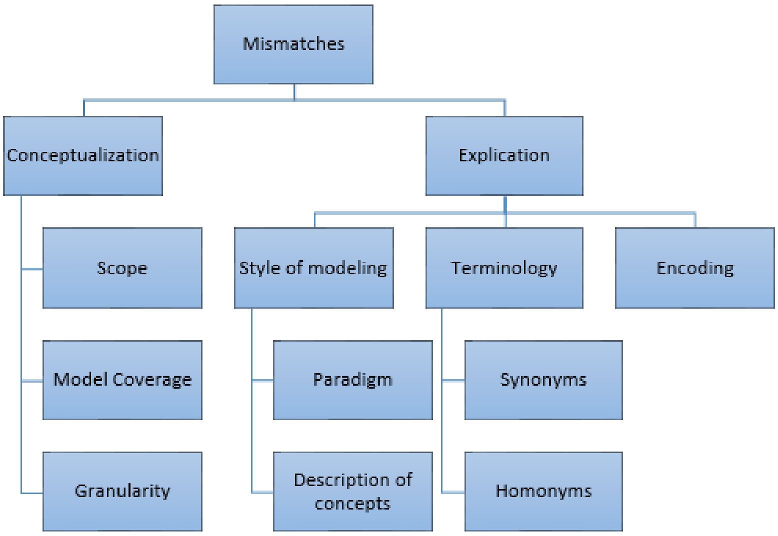


Fig. 1. Mismatches in ontology reusing.

For evaluating the resulted ontology, there are many criteria as modularity, which indicates if exist sets of reusable components [4]; connectivity, it measures the most important concepts based on the amount of relationships [20]; and coupling, this is about the number of external concepts that are referenced or imported [7].

In this work, the alignment, mapping and ontology networks techniques for ontology reusing are analyzed in order to offer a global vision about this approaches. This document has divided into five sections; the section 2 shows the ontology mapping task, in the section 3 the ontology alignment task is described; the section 4 contains information about the ontology network; and finally, the conclusion and future work are explained in the section 6.

2 Ontology Mapping

Ontology Mapping or Ontology Matching is the proccess to find correspondences between two ontologies, this correspondences are not belonging part of any ontology [1, 9]. The resulted correspondences are affected by the granularity level into the information of the ontologies [3]. The ontology mapping is defined as a 4-tuple $\langle e, e', r, c\rangle$ where e and e' are entities of two ontologies (O_1 and O_2

respectively); r is a semantic relation $\in \{\sqsubseteq$ more general, $\sqsupseteq$ more specific, and $\equiv$ equivalence$\}$; and the confidence value c $(0, 1]$, i.e. 0=not reliable and 1= reliable [6]. There are three phases for mapping: mapping discovery, mapping representation, and mapping execution [1].

In the Fig. 2 it shows two ontologies, after a mapping process the result is an equivalence correspondence between the car and vehicle classes.

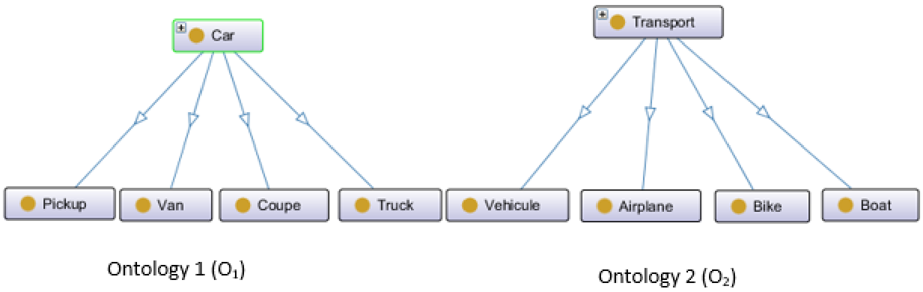


Fig. 2. Example of two ontologies.

3 Ontology Alignment

In ontology learning, the ontology alignment is the task that puts different models in correspondences by discovering similarities between discrete entities from two ontologies by a semi-automatic process in order to give semantic interoperability [1, 18]. Ontology alignment allows to visualise correspondences, resource transformation, and querying in two ontologies at the same time [21]. The ontology alignment is defined as $O^M = O_1 \bigcup O_2 \bigcup M$ [6]; where M is the set of correspondences as a result of application of Match operator. Fig. 3 illustrates the result of ontology alignment using the ontologies shown in the Fig. 2.

The Match operator has schema-based or instance-based matching; in schema-based, it considers the concepts and relations in order to determine correspondences using a similarity measure; in instance-based, the instances that belong to different concepts are consider for discovering a similarity [1].

For alignment evaluation using user validation approach, is necessary that users are acquainted with the formal representations of involved ontologies [2]. In [6] recommend if the evaluation is based on precision, recall and F-measure, using Gold standard and Silver standard approaches to guaranty the correct results.

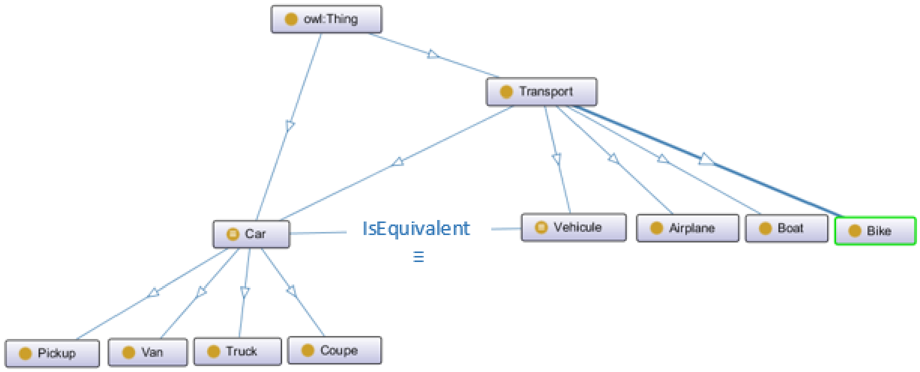


Fig. 3. Result of ontology alignment.

4 Ontology Networks

Unlike mapping and alignment ontology, an ontology network is a collection of ontologies related through meta-data that indicate the dependence between them [18], and sometimes these ontologies do not share the same domain. Specifically, the ontology network design refers to create meta-relationships between ontological entities and represents the opposite process of ontology modularization, where the main task is finding parts into ontology that can work as modules. The meta-relationships in the ontology network (versioning, inclusion, inconsistency, similarity, among others) are relationships whit a meaning that depends of the meanings of the others ontologies used in the network [16], i.e. the relations in an ontology depend of the domain, while, the meta-relations are explicit and independent of the domain [15].

An example of ontology network is presented in [11], in this paper we can see the representation of modules is through triangles and contains the name of the reused ontologies; the relations do not belong to a particular domain, but indicate the way of the modules are related by general terms (see Fig. 4).

5 Related Works

Zulkarnain et al. [22] propose a methodology for ontology reusing in medical area, this methodology is focused to use one or more ontological modules by term extraction, ontology recommendation and mapping; the evaluation of resulted ontology is based on the resources from domain. Pinto et al. [10] developed a methodology for reusing ontology by integration focus; where the ontologies for integration must be compatible with the requirements of the domain according the knowledge pieces for changing or removing. Finally, in the last step, the ontology structure using full or partial ontology integration is built.

In [3], it is presented a proposal of the union of three medical domain ontologies through the mapping of ontologies, which represent information about

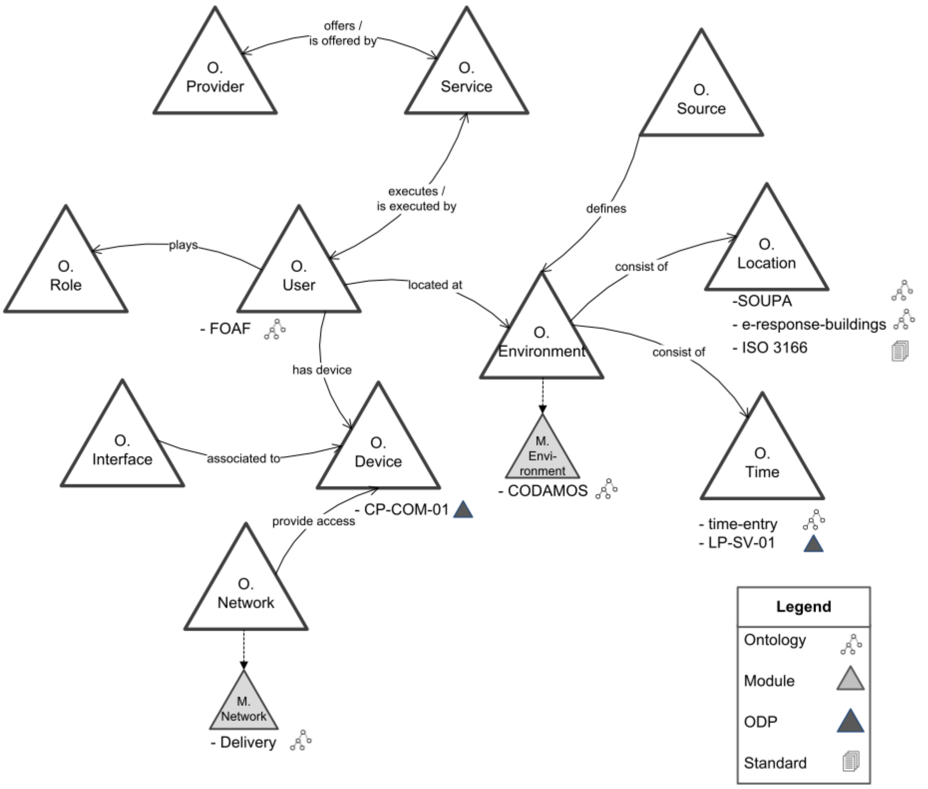


Fig. 4. Example of ontology network [11].

the Diseases (*Disease Ontology* , DO), Human Phenotypes (*Human Phenotype Ontology*, HPO) and X-Ray (*Radiology Gamuts Ontology*, RGO).

They looked for elements that were synonyms and the longer text strings between the main RGO classes with respect to DO and HPO. In the case of relationship mapping, three types were identified: sameAs, subsumption and mayCause, in order to integrate a relationship scheme based on elements coming in the following order DO-RGO-HPO.

Also, it has been proposed to use an ontology to integrate information in a heterogeneous way in order to create a repository using a detection similarity algorithm; for the creation of the main ontology, it was necessary to map features from ontologies already known to extend the domain [14].

Dragisic et al. [2] proposed to evaluate the ontology alignment results by a user validation. In this, using three main aspects: user profile (familiarity into the domain and the elements of knowledge representation); system services; and user interface (explanation of results).

In [15], an ontology network was developed in order to describe the pedagogy evaluation domain and the semi-automatic generation of evaluation, using ontologies focused on evaluation process. Poveda-Villalón et al. [11] built an ontology network for mobile environments (mIO) using the scenarios 2,3,7,8, and 9 of NeOn Methodology [8]; they reused many ontologies in order to create modules as time, location, user, among others; the modules were connected by relations extracted from textual natural language that describes use cases.

6 Conclusions and Future Work

For knowledge representation is very important reusing ontological resources, because many of them are evaluated and maintained by domain experts; it allows building model at a low cost. Mapping ontology finds correspondences between two ontologies that represent the same domain or only a part of it; meanwhile, the ontology alignment uses these correspondences for merging the ontologies. On the other hand, the ontology networks use some not related ontologies and design new relations for the interoperability independently of the ontology domain.

In future work, we will develop a methodology for ontology network design that includes mapping and alignment task in order to include ontologies with correspondences.

Acknowledgment. This work is supported by the Sectoral Research Fund for Education with the CONACyT project 257357, and partially supported by the VIEP-BUAP project.

References

1. De Bruijn, J., Ehrig, M., Feier, C., Martín-Recuerda, F., Scharffe, F., Weiten, M.: Ontology mediation, merging and aligning. Semantic web technologies pp. 95–113 (2006)
2. Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., Pesquita, C.: User validation in ontology alignment. In: International Semantic Web Conference. pp. 200–217. Springer (2016)
3. Finke, M.T., Filice, R.W., Kahn Jr, C.E.: Integrating ontologies of human diseases, phenotypes, and radiological diagnosis. Journal of the American Medical Informatics Association 26(2), 149–154 (2019)
4. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: European Semantic Web Conference. pp. 140–154. Springer (2006)
5. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies 43(5-6), 907–928 (1995)
6. Harrow, I., Jiménez-Ruiz, E., Splendiani, A., Romacker, M., Woollard, P., Markel, S., Alam-Faruque, Y., Koch, M., Malone, J., Waaler, A.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. Journal of biomedical semantics 8(1), 55 (2017)

7. Hlomani, H., Stacey, D.: Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. Semantic Web Journal 1(5), 1–11 (2014)
8. ingGroup, O.E.: Neon methodology. http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/methodologies/59-neon-methodology, accessed: 2020-02-05
9. Noy, N.F.: Ontology mapping. In: Handbook on ontologies, pp. 573–590. Springer (2009)
10. Pinto, H.S., Martins, J.: Reusing ontologies. In: AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes. vol. 2, p. 7. Karlsruhe, Germany: AAAI (2000)
11. Poveda-Villalón, M., Suárez-Figueroa, M.C., García-Castro, R., Gómez-Pérez, A.: A context ontology for mobile environments. In: CIAO@ EKAW (2010)
12. Rahimi, A., Liaw, S.T., Taggart, J., Ray, P., Yu, H.: Validating an ontology-based algorithm to identify patients with type 2 diabetes mellitus in electronic health records. International journal of medical informatics 83(10), 768–778 (2014)
13. Ramos, E., Núñez, H., Casañas, R.: Esquema para evaluar ontologías únicas para un dominio de conocimiento. Enlace 6(1), 57–71 (2009)
14. Ren, S., Lu, X., Wang, T.: Application of ontology in medical heterogeneous data integration. In: Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on. pp. 150–155. IEEE (2018)
15. Romero, L.: Marco de trabajo basado en una red de ontologías para dar soporte a la generación de evaluaciones en entornos de e-learning. Ph.D. thesis, Universidad Tecnológica Nacional. Facultad Regional Santa Fe (2015)
16. Savić, M., Ivanović, M., Jain, L.C.: Complex Networks in Software, Knowledge, and Social Systems, vol. 148. Springer (2018)
17. Suárez-Figueroa, M.C.: NeOn Methodology for building ontology networks: specification, scheduling and reuse. Ph.D. thesis, Informatica (2010)
18. Suarez-Figueroa, M.C., Gomez-Perez, A., Motta, E., Gangemi, A.: Introduction: Ontology engineering in a networked world. In: Ontology engineering in a networked world, pp. 1–6. Springer (2012)
19. Tello, A.L.: Ontologías en la web semántica. España: Universidad De Extremadura (2001)
20. Yu, J., Thom, J.A., Tam, A.: Ontology evaluation using wikipedia categories for browsing. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 223–232. ACM (2007)
21. Zagal Flores, R.E.: Alineación de ontologías usando el método boosting (M. Sc. thesis, IPN) (2008)
22. Zulkarnain, N.Z., Meziane, F., Crofts, G.: A methodology for biomedical ontology reuse. In: International conference on applications of natural language to information systems. pp. 3–14. Springer (2016)

ISSN 1870-4069

Use of Data Flow Diagrams for Building Process with Message Passing: A Parallel Design Proposal

Mario Rossainz López, Mireya Tovar Vidal, Nallely Morales Lozada,
Jesús Alberto Islas Fuentes

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

{rossainz,mtovar}@cs.buap.mx,
{nalle.ml29, albertisfu}@gmail.com

Abstract. The present work shows a method to design parallel programs with message passing using Data Flow Diagrams (DFDs), which are graphs that are used within the classic structured design of Software Engineering. It shows the modification and semantic adaptation of the graphic elements of the DFDs to the semantics of the elements that are used in the passing of messages of the concurrent / parallel programming to map the processes, communication channels, geometric parallelism, parallel composition, generalization , specialization and nesting of processes to the graphic elements of DFDs so that the design made with them turns out to be a transparent design in the coding of a parallel system that uses processes, communication channels, composition and process nesting. The complete parallel design assisted with DFDs of an application called Parallel Generator of Natural Numbers is shown to demonstrate the usefulness of the proposal.

Keywords. Data flow diagram, parallel programming, message passing, distributed memory, parallel design, structured design.

1 Introduction

There is interest by a large part of the community dedicated to computational discipline, to develop increasingly powerful systems taking advantage of the benefits of current computer architectures. The most important approach in the development of such systems is the use of parallel programming at various levels

Parallel systems are found today in practically every type of device, not only in computers, the same are present in embedded devices in household appliances (screens, audio components, refrigerators, video players, etc.), than in supercomputers of institutes, universities, the army or government agencies [1].

But the way in which this parallel software is developed in order to make efficient use of these components takes great relevance with respect to the parallel algorithms designed and implemented, adapting them to the different existing platforms.

Within this area, the knowledge that is had on the concepts related to the design of programs and parallel algorithms takes great importance to implement efficient parallel programs in its logic, in its execution and in its performance. The parallel program design model adopted in this work is the so-called channel-process model [1]. This model consists of a parallel message passing architecture whose components are two: the processes that have tasks to be performed and the means of communication between said processes to share information through the concept of the channel [2]. A process consists of a sequential program to which a task is associated, a locality of non-shared local memory and a collection of ports of inputs and / or outputs that are the communication channels [3].

The channel used is a zero capacity storage structure (that is, only one piece of data travels through the channel, sent by one process to another that are connected to each other by the same channel; where a sending process sends a message through the shared channel and the receiving process receives it, so that while the message or data is within the channel, the sender will not be able to send any other data until he knows that the receiver has already received the message and the channel will be empty and ready to store another message of the same type as the previous one). A process that is connected to another through a communication channel and that is a receiver will be blocked in its execution if it tries to receive a message that has not yet been sent. In the same way the sending process will be blocked in its execution if when sending a message through its communication channel it has not been received by the receiving process. Therefore, the communication between the sending process and the receiving process through the communication channel is synchronous, thus generating the concept of rendezvous, [3].

The classic design methodology for this type of parallel programming as indicated by [4] is the one proposed by Ian Foster: to divide the computation (tasks) and the data into pieces, determine the communication patterns between the tasks that they will be the processes, generate the composition of tasks / processes (nesting tasks) and assign each process to a processor or thread of execution. The idea is to divide data as independent as possible and then determine how many processes should be created and how to associate those processes with the data. A functional decomposition is generated that consists of dividing the total calculation into several processes and associating the data with them. There will be processes that can be divided into simpler processes generating more decompositions and there will be processes that can be joined and nested in a composition of more processes. The objective is to determine primitive processes that can no longer be divided and that indicate the starting point of the more general processes that make up the system, giving rise to a degree of fine or coarse parallelism; that depends on the problem to be solved and the input data [4]. Ideally, determine as many primitive processes as possible in order to maximize the degree of parallelism.

Finally, the communication patterns between primitive and non-primitive processes must be determined through the creation of communication channels.

The processes can carry out both their communications and the execution of their associated algorithms in parallel. If we analyze and relate the design methodology of Ian Foster with the method of use and creation of a data flow diagram or DFD we can see a one-to-one mapping of the components of the Foster methodology with the graphic components of the DFDs , so that this type of graphs can help the novel programmer in parallel to propose a good design of a parallel application.

This is the proposal of this research work that is organized as follows: section 2 shows the characteristics of DFDs, their components, the types of connections and how to build them, section 3 talks about parallel programming with message passing, section 4 shows how to design a parallel system using DFDs and applies the methodology proposed in a case study, and finally in section 5 the conclusions and future work are shown.

2 State of the Art

According to [5] the design of a software system describes the organization of the system, expressed in terms of its components, the relationships between them, the relationships with its environment, and the fundamental principles that govern the design and evolution of the system.

The design of a system focuses on what is known as "viewpoint", which is a form of abstraction that focuses on some specific aspects of the system, abstracting from the rest. As there are different viewpoint based on the generality and particularities that each software engineer uses to define a design, it is necessary to use unified graphic models that correctly represent the logical, process, development and physical point of view of scenarios etc., of the system to implement. In the case at hand, the viewpoint of the design of parallel applications using message passing can be unified with different graphic models, from contextual diagrams to represent the components that make up the system, generating a software architecture, to the use of diagrams of transition of states in UML, but without a doubt an ad-hoc graphic model for such a design is the use of Data Flow Diagrams (DFD's) that is used in the classic structured design of Software Engineering.

Currently there are several works in the literature that make use of DFDs to show the design of the parallel systems that they propose. In [6] shows the Design of Applications in Distributed Systems with the two possible environments, conventional Operating Systems and the Internet, using DFDs to model the design of the concept of specialization of a service in the client-server architecture. In [7] the design and implementation of a distributed video on demand application based on the client-server architecture is shown, where it uses the DFDs to design the exchange of RTP and RTCP packets in the request of the video service.

3 Data Flow Diagram (DFD)

A DFD is a network-shaped diagram that represents the flow of data and the transformations that are applied to them when moving from the entrance to the exit of a software system [8]. DFDs are commonly used to model the functions of a software system and the data flow between them at different levels of abstraction in a concept of structured design within software engineering. The software system is modeled through a set of level DFDs in which the upper levels define the system functions in general and the lower levels define these functions in detail [8].

3.1 Components

- **Procedures:** They represent the functions of a software system. A process represents a function, procedure or operation system, that transforms the input data streams into one or more output streams [9]. Its graphic representation is a circle and inside it includes a number and the name that represents the function which must be unique within the DFD (see Fig. 1).
- **Storage:** Represent the data stored in a specific structure such as a database or a file. A data storage represents system information stored temporarily or permanently [9]. The storage is a logical repository in the DFD that can physically represent a file, a database, a file cabinet drawer, etc. In a DFD there may be more than one different data storage. Its graphic representation is two parallel bars (see Fig. 1).
- **External Entities:** They are the sources or destinations of the system information. It represents a source (generator) or destination (consumer) of information for the system but that does not belong to it [9]. It can represent a subsystem, a person, a department, an organization, etc., that provides data to the system or that receives it from it. They are represented in the DFD by a square with a representative name of the external entity inside (see Fig. 1).
- **Data Flow:** they represent the information flows that flow between the functions of the system. It is a path through which data travels from one part of the system to another. They represent the moving data within the system. Data flows are the means of connection of the DFD components [9]. They are represented with directed arcs, where the arrow indicates the direction of the data, (see Fig. 1).

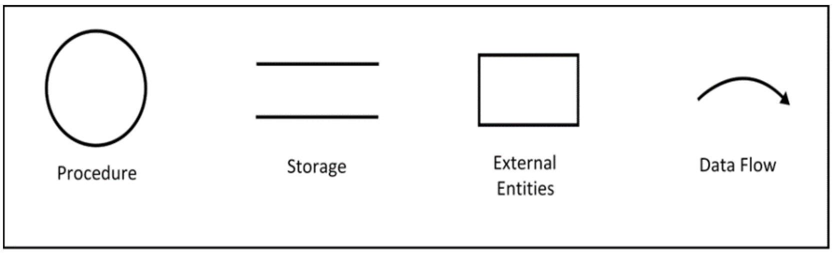


Fig. 1. Graphical representation of the components of a DFD.

Table 1 shows the connections allowed in DFDs [10].

Table 1. Allowed connections between the components of a DFD.

Destination/ Source	Procedure	Storage	External Entitie
Procedure	TRUE	TRUE	TRUE
Storage	TRUE	FALSE	FALSE
External Entitie	TRUE	FALSE	FALSE

3.2 Connections

The Procedure-Procedure connection: The direct connection between two procedures through a data flow is possible if the information is synchronous, that is, that the target procedure begins at the moment when the source procedure ends its function. If this is not the case, it is necessary that there is a temporary storage that saves the data of the origin procedure [10, 11]. The target procedure will then capture this data when needed (see Fig. 2). **The Procedure-Storage connection:** There are different types of connections that can be made between procedures and storages in a DFD. The Fig. 3 illustrates the following cases:

- **Query Flow:** shows the use of the storage information by the procedure for one of the following actions: use the values of one or more attributes of a storage occurrence or check if the values of the selected attributes meet certain criteria.

- **Update Flow:** Indicates that the procedure will alter the information that is in the storage to: create a new occurrence of an existing entity or interrelation in the storage, delete one or more occurrences of an entity or interrelation and modify the value of an attribute.

- **Dialogue Flow:** Between a procedure and a storage it represents at least one query flow and one update flow that have no direct relationship. Dialog flows are also used to simplify the interface between two components of a DFD.

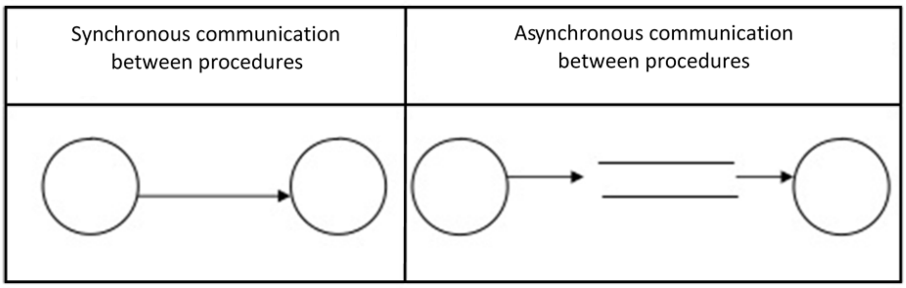


Fig 2. Procedure-Procedure Connection in a DFD.

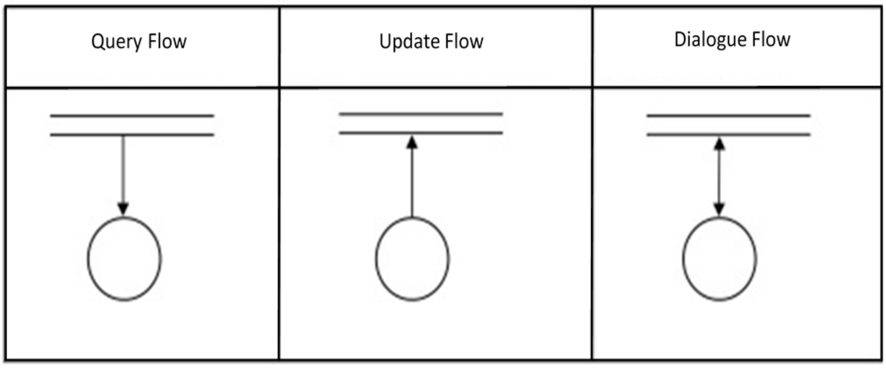


Fig. 3. Procedure-Store connection in a DFD.

3.3 Construction

The construction of a DFD is based on the principle of decomposition by levels of detail. The idea is to generate a model of a system represented by DFDs through layers.

The decomposition by levels allows the system to be analyzed from the general scope to the detail through successive intermediate levels (top-down approach), according to. This form of design of a system provides us with a number of advantages:

- it helps to build the specification from top to bottom,
- the different levels can be addressed to different people,
- independent functions of the system can be modeled at the same time,
- it facilitates system documentation since each diagram can be explained separately.

Thus, the decomposition of a process in a DFD produces another DFD (see Fig. 4):

- **Context Diagram:** It is the highest level of the hierarchy in the design of a system. In this diagram there is only one procedure that represents the complete system.
- **System Diagram [LEVEL 0]:** It is the decomposition of the DFD of the Context diagram into another DFD in which the main functions of the system or subsystems are represented.
- **Middle Level Diagrams [LEVEL 1, 2, 3, …]:** These are the DFDs that result from the decomposition of each of the DFD procedures of the System Diagram into new diagrams that represent simpler functions.
- **Primitive Function Diagrams [Level n]:** These are DFDs that represent functions that are detailed enough that the creation of new DFDs is not necessary.

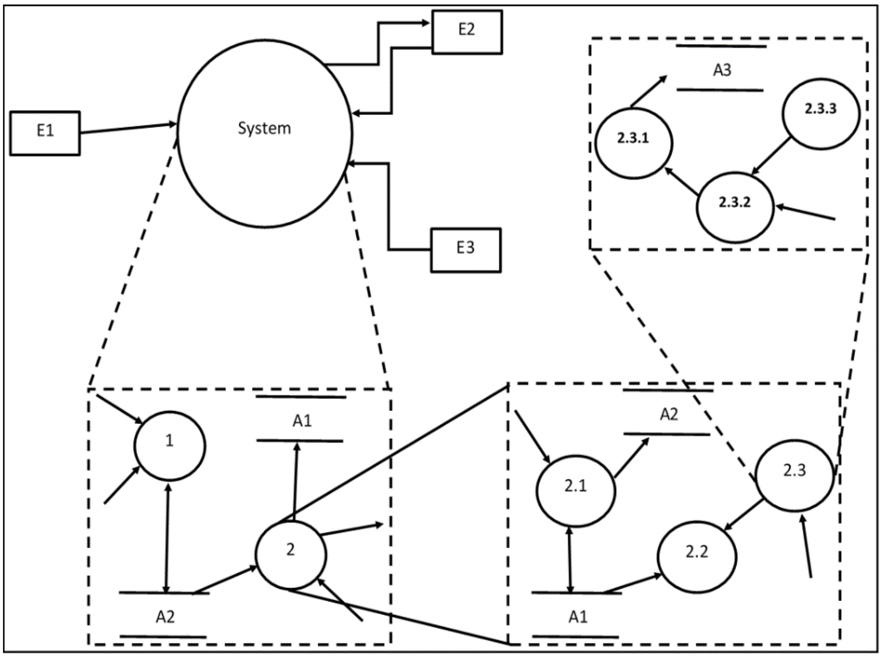


Fig 4. Decomposition by levels of a DFD.

4 Parallel Programming with Message Passing

The natural way to communicate and synchronize processes in this type of systems is using message passing: The processes exchange messages with each other through explicit send and receive operations that constitute the basic primitives of any system of communication of this type [12]. The fundamental elements involved in the communication in systems with message passing are: a sending process (transmitter), a receiving process (receiver), a communication channel (channel), the message to be sent/received (message) and operations of sending (send ()) and reception (receive ()), see Fig. 5.

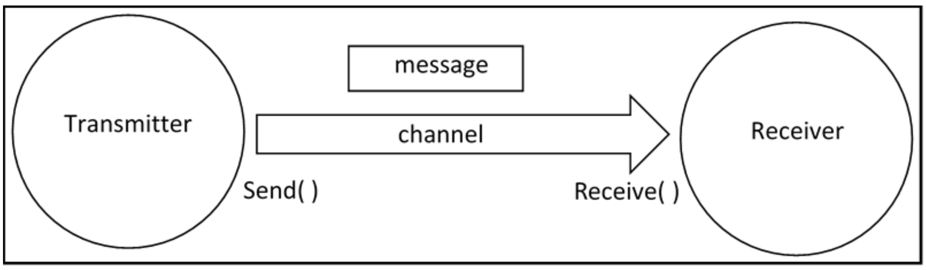


Fig 5. Message passing.

4.1 Types of Communication Between Processes

It is the way in which the sender indicates to whom the message is addressed and vice versa, that is, the way in which the receiver indicates from whom he expects a message. We talk about direct or indirect communication [12].

- **Direct-Symmetric Communication:** It is characterized in that the sender explicitly identifies the recipient of the message in the sending operation. The receiver, in turn, identifies the sender of the message and establishes a communication link between them.
- **Direct-Asymmetric Communication:** The sender continues to identify the receiver, but the receiver does not identify a specific sender.
- **Indirect Communication:** The sender and receiver processes are not explicitly identified. Communication is done by depositing messages in an intermediate storage that is supposed to be known for processes interested in communication. That intermediate store is called the mailbox.
- **Indirect Communication with Channels:** The sending and receiving operations are carried out through the specification of a channel (communication link), which generally has an associated type and on which only data of the same type can be sent. In addition, a channel cannot be used by multiple transmitters and receivers and is unidirectional. The concept of channel arises in languages like Occam and Pacal-FC.
- **Asynchronous Communication:** The sender can carry out the sending operation without it being necessary to coincide in time with the receiving operation by the receiver. It is necessary to store the messages in buffers until the receiver removes them.
- **Synchronous Communication:** The sending and receiving operations by senders and receivers must coincide in time, that is, the sender will be blocked in the send operation until the receiver is ready to receive the message, or vice versa, the receiver will be blocked in the receive operation until the sender sends the message. The processes will be blocked until the coincidence of both occurs in time (rendezvous).

4.2 Channels and Messages

- **Data flow.** Once a communication channel between sender and receiver is established, this, according to the flow of data that passes through it, can be of two types: unidirectional or bidirectional. For the first, information always flows in one direction between the two interlocutors, while for the second, information flows in both directions.
- **Channel capacity.** It is the possibility that the channel has to store the messages sent by the sender when they are not received immediately by the receiver. The channels can be of zero capacity (where there is no buffer where messages are stored), channels of finite capacity (where the existing buffer has a fixed size) and channels of infinite capacity (where the buffer associated with the link of communication is assumed infinite in its capacity).

- **Message size**. The messages that travel through the channel can be of fixed length or of variable length.
- **Channels with type or without type.** Some communication schemes require defining the type of data that will flow through the channel, imposing the restriction of sending data only of the type for which the channel was declared.
- **Message passing by copy or by reference.** Communication through message passing requires sending information between the processes involved in the communication. This can be done in two ways: make an exact copy of the data (message) that the sender wants to send from his address space to the address space of the receiving process (passing by copy or value) or simply send the recipient the address in the address space of the sender where the message is located (passing by reference). The latter requires that processes share memory.

5 Designing a Parallel System with Data Flow Diagrams

A parallel program consists of several processes that run at the same time [13]. A process is a sequential program in execution. The parallel program has control over its processes. Within the human cognitive aspect, the composition of processes through more processes, that is, through simple processes build more complex processes; It becomes natural when parallel programs based on distributed memory are developed [13]. The processes are highly independent because they are limited to the task they have to solve and provide the fundamentals of the program structure. In the field of object orientation, processes can be considered as active objects because they have the capacity to execute themselves. During the analysis phase of the development of a parallel software, the tasks to be performed by the program are determined, among other things [14].

However, it is in the design phase that these tasks are usually represented with graphic models that represent both the structure and semantics of the processes associated with the system and for this, we use activity diagrams, or state transition diagrams in the OO design. However, the structured design gives us through the data flow diagrams an important graphic model that we can adopt in the design of a parallel program with message passing in a completely transparent way. A data flow diagram, as defined in section 2, is a directed graph that can be adapted in the design of a parallel system to define processes, as well as the degree of parallelism that we intend to program by generating Intermediate level DFDs and program structure as a whole using the context diagram DFD.

Each circle of a DFD model represents in the design of a parallel program with message passing a process or a control flow, while the arrows will now indicate the communication channels that the processes use to communicate with each other and thus define the flow of data or its express communication. The rectangles that represent external entities in the DFD will now represent processes of input or output of information that reach other processes modeled with the circles and data storage of the

DFDs will now represent mailboxes where the processes can leave or collect information stored through communication channels, if required and according to the type of communication used in the processes (see section 3.1). The graphical representation of a data flow diagram is therefore a powerful model in the design phase, which represents processes at different levels of nesting, that is, specialization and generalization.

5.1 Case Study: Parallel Generator of Natural Numbers

The design of a parallel system with message passing that generates natural numbers in sequence is shown. This case study has been taken from [15, 16]. The design is done using the DFDs. First the context diagram (more general level in the DFD hierarchy), then the level 0 DFD diagram which is the decomposition of the context diagram in the main processes of the system, along with the communication between them through channels and Medium-level DFDs that represent the decomposition of the processes identified in the system DFD and finally the design of the DFDs of the primitive processes that are those that can no longer be broken down into more process.

The context DFD of this case study is shown in fig. 6. The process represented by the circle in the figure generates consecutive natural numbers starting from the number zero, although the user can indicate the starting number. The numbers generated are sent through the communication channel represented by the arrow or flow to another process called “Display” in charge of receiving them in sequence and printing them on the screen. This process is responsible for defining the limit of reception of natural numbers. Both the Parallel Generator of Natural Numbers process and the Display process are created in a “Parallel Composition of Processes” in concurrent execution [17]. The communication between these processes occurs through the communication channel in accordance with the principle of rendezvous, see Synchronous Communication of section 3.1 for details.

The Display process of fig. 6, is a primitive process and cannot be broken down into more sub-processes. On the contrary, the Parallel Generator of Natural Numbers process can be broken down into more sub-processes which are designed through the Level 0 DFD of fig. 7. The DFD of level 0, consists of three processes (Prefix (0), Delta and Successor or Suc) which in accordance with the logic of operation of the system are connected by their respective channels and whose communication between them is carried out through of the rendezvous concept. Each process is an active object to which a control thread is associated, while a channel is a passive object without its own life.

Of the three DFD processes of Level 0, the Suc process is a primitive process. Its operation consists in increasing by one the value that enters through its Channel b and then sending it to the Prefix process through Channel c. The other two processes, Prefix (0) and Delta, are processes composed of more processes so we can design their corresponding DFDs of Middle Levels (see Fig. 8).

The level 1 DFD of the Prefix process has two channels, one input (Channel c) and one output (Channel a) that connect to the Id process that composes it. The first time the Prefix process is executed, it sends the N value provided by the user through its output channel. N is the initial natural number. In its next executions the process

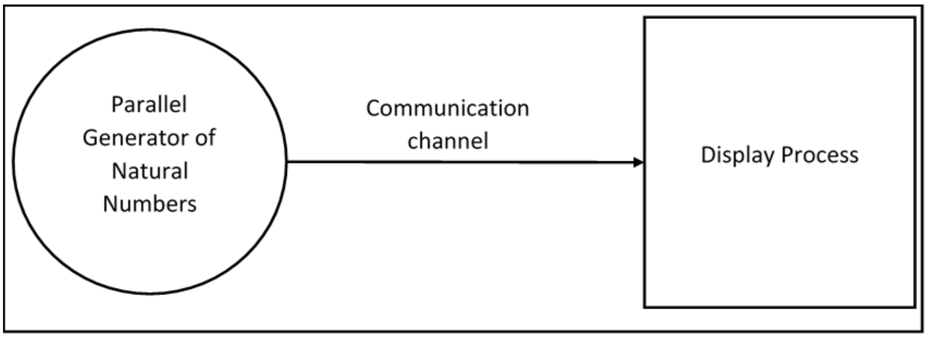


Fig 6. Context DFD.

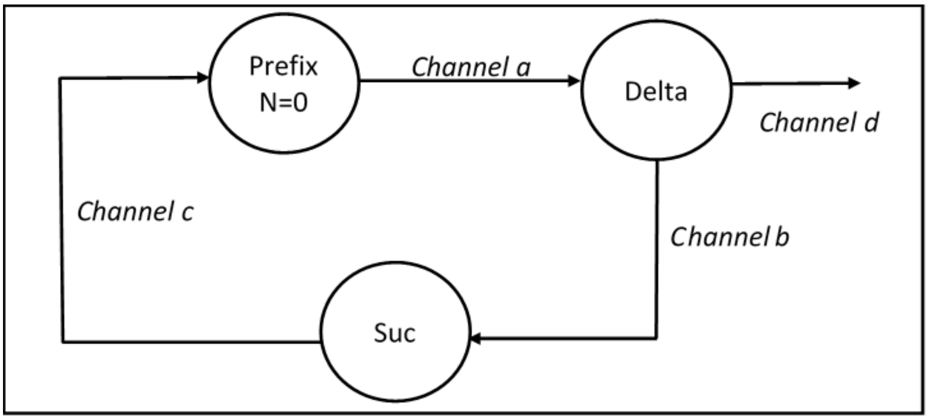


Fig 7. Level 0 DFD.

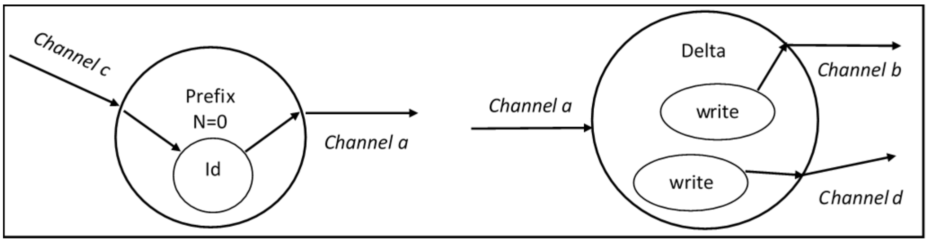
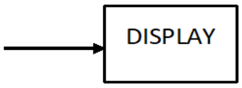
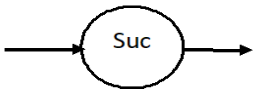
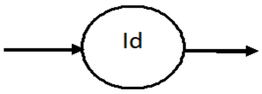
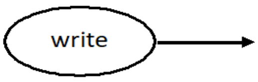


Fig 8. Level 1 DFD (Middle Levels).

executes the behavior of its internal process Id (it is said to behave like him). The internal process Id is a primitive process. In the level 1, DFD (fig. 8) the Delta process has three communication channels, one input channel and two output channels.

Delta's behavior is as follows: in an infinite execution, the number received by its input channel (channel a) is sent by the internal processes (processes Write), which are executed in parallel, so that they send through their respective channels output (Channel b and Channel d) the data received without any modification. Write processes are primitive processes. Finally, the design is completed with the graphic representation of the level 2 DFDs of the primitive processes (see table 2).

Table 2. Level 2 DFDs (Primitive Processes).

TYPE OF EXECUTION	LEVEL 2 DFD (PRIMITIVE PROCESS)	DESCRIPTION
Infinite 		Process that receives from its input channel a data (natural number) and then prints it on the screen.
Infinite 		Process that receives a natural number through its input channel and then generates the successor of that number and the new value is sent through its output channel.
Infinite 		Process that receives a natural number through its input channel and is immediately sent by its output channel without modification.
Finite 		Process that receives a natural number as a parameter and is then sent through its output channel without any modification.

6 Conclusions

A way to design parallel programs using Data Flow Diagrams of the structured design of classical software engineering has been proposed. The same graphic symbols of the DFD were used by changing their semantics of the structured design by the semantics of parallel programming with message passing, so that their uses in the design of a parallel system result in a transparent mapping of processes, communication channels, geometric parallelism, logical partition of processes, generalization and specialization in different levels of nesting that correspond to the different levels of generalization and particularization of DFDs (context, system, levels 1, 2, 3, etc ... until reaching processes primitives).

To demonstrate its usefulness, this proposal was used in the parallel design of a Natural Number Generator. This application has been programmed in JAVA using a particular class library called JPMI based on the process algebra of Hoare or CSP.

References

1. Moyano, J.: Programación Paralela - Conceptos y Diseño de Sistemas Distribuidos (2016)
2. Fujimoto: Parallel and Distributed Simulation Systems. USA, Wiley-Interscience (2000)
3. Palma-Méndez, J.T., Garrido-Cabrera, M.C., Sánchez, F., Quesada-Arencibia, A.: Programación Concurernte. Paraninfo (2003)

4. Wilkinson, B., Allen, M.: Parallel programming techniques and applications using networked workstations and parallel computers. Prentice Hall (2000)
5. Salguero, E.: Diseño de Aplicaciones Distribuidas. GitHub Gist (2018)
6. Martínez, G.E.: Diseño de Sistemas Distribuidos. LWP, Comunidad de Programadores (2019)
7. Montalvo, O.P., Byron, P.V.: Diseño e Implementación de una aplicación distribuida de video bajo demanda basada en la arquitectura cliente-servidor. En: XXV Jornadas en Ingeniería Eléctrica y Electrónica, 25, pp. 344–355 (2014)
8. Sumano, M.A.: Análisis estructurado moderno (2012)
9. Cillero, M.: Diagrama de flujo de datos (2018)
10. García, S., Morales, E.: Análisis y Diseño Detallado de Aplicaciones Informáticas de Gestión. Madrid, Thomson Paraninfo (2003)
11. De Amescua-Seco, A.: Análisis y Diseño Estructurado y Orientado a Objetos de Sistemas Informáticos. Madrid, McGraw Hill (2003)
12. Capel, M., Rodriguez, V.S.: Sistemas Concurrentes y Distribuidos. Granada, Copycentro Editorial (2012)
13. Hoare, C.A.R.: Communicating Sequential Processes. London, Prentice Hall (2003)
14. Kendall & Kendall: Análisis y Diseño de Sistemas, México, Pearson (2011)
15. Hilderink, G., Broenink, J., Vervoort, W., Bakkers, A.: Communicating Java Threads. IOS Press (1999)
16. Hilderink, G., Broenink, J., Bakkers, A., Schaller, N.C.: Communicating Threads for Java, Architectures, Languages and Techniques. IOS Press (2000)
17. Friborg R.M., Brian V.: PyCSP – Controlled Concurrency. International Journal of Information Processing and Management, 1(2), pp. 40–49 (2010)

ISSN 1870-4069

Parameter Estimation for an ontology Evaluation Metric

Mireya Tovar Vidal, Gerardo Flores Petlacalco, Hugo Raziel Lasserre Chávez, Aimee Cecilia Hernández García, Emmanuel Santos Rodríguez, Irvin Yair Cabrera Moreno, Juan Carlos Flores Molina, Jesús Flores Castillo, Christian Martínez Cuamani

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, Mexico

mtovar@cs.buap.mx, {gfloresp93, hugohp}@outlook.com, {zvok59, e.ss.rdz, yair.cb, fmjuancarlos, j.jikamaru}@gmail.com, cmc_2012@hotmail.com

Abstract. Ontologies are semantic resources that capture the knowledge of a particular domain, by means of the elements that comprise it. Despite the manner the ontologies are created (manually, automatically or semiautomatically), there still the fact that the process is not perfect and, therefore, an additional step for ontology quality validation is needed. The ontology evaluation task aims to measure the quality of these resources. Among other approaches, the corpus-based evaluation attempts to validate the ontology components by means of an external data source which usually is a collection of documents associated to the same domain of the ontology to be evaluated. In this paper, a metric that involves results of different methods for identifying semantic relations and concepts in the domain corpus is presented and for evaluating the quality of an ontology. So, an overall evaluation score is provided and a parameter estimation for the metric proposed is applied. The experimental results show a satisfactory performance, which it is considered interesting for the task of measuring the quality of ontologies of restricted domain.

Keywords. Ontology, ontology evaluation, metric, natural language processing.

1 Introduction

Daily, the information on the current web grows and much of this information is generated without a structure that can be understood by both machines and humans, which makes it a dificult task to process. The semantic web, proposed by Tim Berners-Lee [1], searches to give structure and knowledge to the conventional web. Ontologies are used to represent knowledge in a structured way on the semantic web.

An ontology is defined as "an explicit and formal specification of a shared conceptualization" [4].

This type of semantic resource, that allows to capture the explicit knowledge in data, is formed by concepts or classes, relations, instances, attributes, axioms, restrictions, rules, functions and events. Domain ontologies are a system of representation of the knowledge that it is possible to organize in taxonomic and non-taxonomic structures of concepts of some area or domain of specific knowledge. If the ontology has been designed manually or automatically, it is necessary to evaluate its quality.

In the literature the evaluation is classified depending on the form of evaluation used, which is: compare the ontology with a gold standard, apply the ontology in an application and evaluate the results, make comparisons against source data of the domain ontology, and finally evaluations made by humans to determine what criteria the ontology satisfies [2]. Gómez-Pérez [3] presents two terms for the ontology evaluation: verification and validation. The verification ensures that the definitions meet the requirements correctly. The validation ensures that the meaning of the definitions correctly model the phenomena of the world. In this work, a metric is applied to evaluate an ontology in the artificial intelligence domain and a parameter estimation for the ontology evaluation metric is applied. The aim of the experiment is to decide if the result of the automatic evaluation system is within the confidence interval of an estimator.

In this work, a metric is applied to evaluate an ontology in the artificial intelligence domain and a parameter estimation for the ontology evaluation metric is applied. The aim of the experiment is to decide if the result of the automatic evaluation system is within the confidence interval of a statistical estimator.

The purpose of statistics is to use the information contained in a sample to make inferences about the population from which the sample is taken [7]. Populations are characterized by numerical descriptive measures called parameters. The main objective of this investigation is to perform statistics to calculate the value of one parameter or more relevant parameters. Among the most important estimators are the mean, variance and standard deviation.

An estimator is expressed through a formula to calculate the value of an estimation based on the measurements contained in a sample. Two types of estimates will be used, the point estimate and the interval estimate. The point estimate includes the parameter estimate in a single value or point.

Interval estimators are also called confidence intervals. The upper and lower endpoints of the confidence interval are named as the upper and lower confidence limits. The probability that a confidence interval will include a fixed amount is called the confidence coefficient.

This document is divided into four sections; in the section 2 the proposed metric is described; the section 3 contains information about the experimental results; and finally, the conclusion and future work are explained in the section 4.

2 Proposed Metric

This work is a complement to the work done in [5, 6], in this a parameters estimation of the designed metric is presented.

First the methodology is presented after the parameters used are presented. The methodology considers the following phases:

1. **Automatic Pre-Processing of Information.** The concepts and ontological relationships are extracted from the ontology using Jena. After, the documents or sentences of the corpus of domain are associated with the concepts and relationships through an information retrieval system. In addition, natural language processing tools are used to pre-processing the text such as: elimination of punctuation symbols, parts of speech and others.
2. **Automatic Discovery of Candidate Terms and/or Ontological Relations.** In this phase the approaches used for the discovery of concepts and ontological relations in the corpus of domain are submitted, some approaches are: lexical-syntactic pattern, formal concept analysis, similarity, dependency analysis and latent semantic analysis. The purpose of this phase is to find evidence of the relation and concepts in the corpus.
3. **Evaluation of the Ontology.** In this phase, a metrics to evaluate the domain ontology is proposed.

In the third phase of the methodology, the quality of the ontology is determined usign the metric of Equation 1. The metric receives the O ontology as a parameter and the output is the result of the evaluation. The metric (M) is formed by the product of three matrices: $Matrix_C$, $Matrix_E$ and $Matrix_I$:

$$
\begin{aligned}
&M(O) = Matrix_C \; Matrix_E \; Matrix_I, \\
&Matrix_C = \begin{bmatrix} A(E_1, CI) \; ... \; A(E_n, CI) \\ A(E_1, NT) \; ... \; A(E_n, NT) \end{bmatrix}, \\
&A(E_i, R) = \frac{\sum_{i=1}^{|R|} Reliability(T_i)}{|R|}, \\
&Reliability(T_i) = \begin{cases} 1 \; If \; \alpha * qual(C_{i,1}) + \\ \quad \beta * qual(C_{i,2}) + \\ \quad \gamma * qual(R_i) > 0.75 \\ 0 \text{ otherwise}, \end{cases} \\
&qual(R) = \frac{|E_i(R)|}{|R|}, \\
&Matriz_E = \begin{bmatrix} a_1 \; b_1 \\ a_2 \; b_2 \\ ... \; ... \\ a_n \; b_n \end{bmatrix}, \\
&Matrix_I = \begin{bmatrix} d_1 \; d_2 \end{bmatrix}.
\end{aligned}
\tag{1}
$$

$Matrix_C$ contains the results of the measure of accuracy (A) of each approach E_i that was applied in phase 2 of the methodology to the domain corpus and to the semantic relationships of the ontology.

In equation 1, $T_i = (C_{i,1}, R_i, C_{i,2})$ is a triplet of ontology; $C_{i,1}$ and $C_{i,2}$ are concepts; and R_i is the ontological relation.

In the case of the quality of the semantic relation ($qual(R)$) we consider the measure of accuracy that, considering the total of relations proposed by the approach ($E_i(R)$) and the total ontology relation ($n = |R|$), where R are the class-inclusion (CI) or non-taxonomic relations (NT). Finally, the matrices $Matrix_E$ and $Matrix_I$ are matrices of external and internal coefficients respectively, which normalize the values between 0 and 1.

As a second part of the evaluation metric, parameter estimation is included. We take the metric value for a point estimate and a confidence interval estimation, for one of the parameters, e.g., mean.

In the case of the confidence interval, we consider the confidence coefficient of $0.95 = 1 - \alpha$. In the Equation 2 and in the Equation 3 the mean and variance are defined as unbiased estimators:

$$\overline{Y} = \mu = \frac{\sum_{i=1}^{n} Y_i}{n-1}, \tag{2}$$

$$S^2 = \sigma^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}. \tag{3}$$

The confidence interval observed for μ is obtained through equation 4:

$$\overline{Y} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}, \tag{4}$$

where $t_{\alpha/2}$ is determined by $n - 1$ degree of freedom. The t distribution has a density function very similar to the normal density.

The standard deviation of θ is the square root of its variance. It is a measure of the amount of variation of a set of values.

3 Results

This section presents the experimental results of the implementation of the evaluation metric. To validate the phases of the methodology, an evaluation is carried out by domain experts and a baseline based on mutual information that measures the degree of co-relation of the semantic relations. Section 3.1 presents the amount of information evaluated by the domain experts and by the automatic system. Considering that the validation is done manually and this requires a high person-hour cost, the experts only evaluated a subset of the sentences from the corpus.

3.1 Description of the Dataset

The domain of ontology considered in the experiments is artificial intelligence (AI)[1] [8]. The ontology contains a number of concepts (C), class-inclusion (S)

[1] Ontologies and their corpus are available on the page: http://www.site.uottawa.ca/ azouaq/goldstandards.htm

relations, and non-taxonomic (R) relations (see Table 1). In Table includes information over the total of documents (D), the amount of sentences (O), total tokens or words (T) of these sentences, the vocabulary (V) of the sentences, the number of filtered sentences (Of) by the information retrieval system; the sentences reviewed by the experts to validate class-inclusion relations (OSE) and non-taxonomic relations (ORE).

Table 1. Dataset.

Domain	Ontology			Corpus						
	C	S	R	D	O	T	V	O_f	OSE	ORE
IA	276	205	61	8	475	11,370	1,510	415	312	110

3.2 Experimental Results for Class-Inclusion Relations

In Table 2, the results obtained from the approach using the accuracy criteria (A); the quality (C) in the prediction of the approach, according to three human experts (H_1, H_2 and H_3); and the baseline is presented. Table 2 also includes the results of the estimators: mean (μ), variance σ^2 and standard deviation σ. The standard deviation of the accuracy of the results of the approaches is presented in Fig. 1 and the Fig. 2 the standard deviation of the $Avg(H_i)$ is presented, too. As can be seen, the approach that is outside the limits of the standard deviation in both figures is FCA sfd_0, which suggests that it could be omitted as a parameter in the calculation of the ontology evaluation metric.

Table 2. Accuracy of the Ontology AI and the Quality of Predictions of Approaches for Class-Inclusion Relations.

Approach	A	$C(H_1)$	$C(H_2)$	$C(H_3)$	$Avg(H_i)$
1 LSP	88.78	89.76	84.39	88.29	87.48
2 Sim-cos	90.24	83.41	80.98	87.80	84.07
3 Sim-cos_u	98.05	90.24	86.83	95.61	90.89
4 FCA min	95.61	89.76	85.37	94.15	89.76
5 FCA sfd_0	100.00	92.20	88.78	97.56	92.85
6 LSA-cos	94.15	90.24	89.76	92.68	90.89
Baseline	56.00	57.00	51.00	55.00	54.00
μ	94.47	89.27	86.02	92.68	89.32
σ^2	19.01	9.06	10.15	15.53	9.72
σ	4.36	3.01	3.19	3.94	3.12

3.3 Experimental Results for Non-Taxonomic Relations

In Table 3, the results obtained from the approach using the accuracy criteria (A); the quality (C) in the prediction of the approach, according to three human

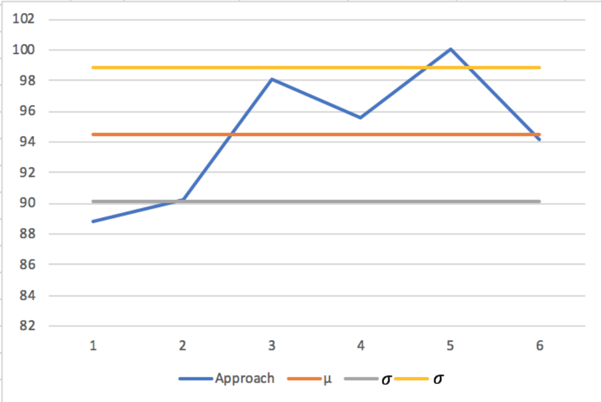


Fig. 1. Standard deviation (σ) of approaches (A) for class-inclusion relations.

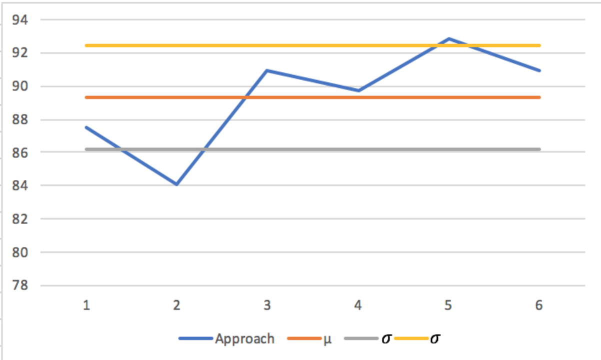


Fig. 2. Standard deviation (σ) of human averages (Avg for class-inclusion relations).

experts (H_1, H_2 and H_3); and the baseline is presented. Table 3 also includes the results of the estimators: mean (μ), variance σ^2 and standard deviation σ. The standard deviation of the accuracy of the results of the approaches is presented in Fig. 3 and the Fig. 4 the standard deviation of the $Avg(H_i)$ is presented, too.

3.4 Experimental Results of the Evaluation Metric and of the Automatic Evaluation System

The results of the approaches presented in Table 2 for class-inclusion relations and the results of Table 3 for non-taxonomic relations are used with the metric of evaluation. In Table 4 is shown the experimental results of the metric (M), where $M(S)$ is the result of the metric for the automated evaluation system, considering only the data validated by the experts, $M(H_i)$ with $i = 1, 2, 3$ and Avg is the average of the results obtained from the experts.

Table 3. Accuracy of the ontology AI and the quality of the predictions of approaches for non-taxonomic relations.

1 Approach	A	$C(H_1)$	$C(H_2)$	$C(H_3)$	$Avg(H_i)$
2 ADS	88.52	81.97	86.89	83.61	84.15
3 Sim-cos	93.44	86.89	88.52	85.25	86.89
4 Sim-cos_u	98.36	88.52	93.44	90.16	90.71
5 FCA min	93.44	80.33	88.52	85.25	84.70
6 FCA sfd_2	100.00	86.89	95.08	91.80	91.26
7 FCA sfd_3	95.08	81.97	90.16	90.16	87.43
8 LSA-cos	90.16	83.61	85.25	85.25	84.70
Baseline	48.00	51.00	46.00	52.00	50.00
μ	94.14	84.31	89.69	87.35	87.12
σ^2	16.91	9.72	12.27	10.47	8.45
σ	4.11	3.12	3.50	3.24	2.91

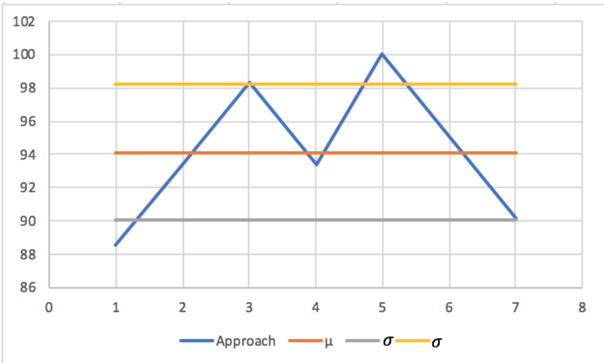


Fig. 3. Standard deviation (σ) of approaches (A) for non-taxonomic relations.

$M(A)$ is the result of the metric for the automated evaluation system considering all the domain corpus.

Table 4. Results of metric evaluation applied to AI domain ontology.

O	$M(A)$	$M(S)$	$M(H_1)$	$M(H_2)$	$M(H_3)$	$M(Avg)$
AI	90.80%	94.31%	86.79%	87.86%	92.52%	89.05%

Now, we use the Equation 4 to calculate the confidence interval, assuming that the evaluations Y_i are normally distributed. Using the data in Table 4, $M(Avg) = \overline{Y} = 89.05$, $s = 3.05$, $n - 1 = 2$ degrees of freedom and using $t_{\alpha/2} = t_{0.025} = 4.303$, then we get $(81.49, 96.63)$ as the observed confidence interval for μ. The estimation of the parameter μ indicates that the results by the experts and by the automatic system are within the allowed limits that determine the mean value of the evaluation of the ontology.

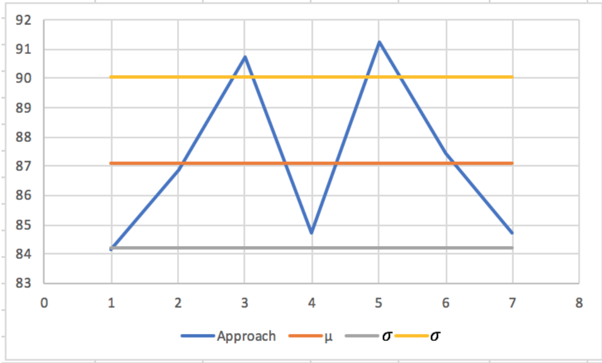


Fig. 4. Standard deviation (σ) of human averages ($Avg(H_i)$) for non-taxonomic relations.

4 Conclusions

This paper presents a study to determine the ideal parameters that indicate the importance of each approach used in the evaluation metric. Based on the experimental results, when using mean, variance and standard deviation how parameters and the validation of human experts, inference that some approaches can be eliminated in the system evaluation process. On the other hand, based on the mean as a parameter, the confidence interval is determined, on which the results of the evaluations can vary. It is observed that the automatic system restricted to the data evaluated by human experts and by the complete corpus is maintained within these limits, which implies that the automatic system provides an acceptable ontology evaluation result. In future work, we will develop other approaches for evaluating the ontology and extend the proposed metric. In addition to including other ontologies and their corpora in the experiments

Acknowledgment. This work is supported by the Sectoral Research Fund for Education with the CONACyT project 257357, and partially supported by the VIEP-BUAP project.

References

1. Berners-Lee, T., Hendler, J.: The semantic web. Scientific American 284, 34–43 (2001)
2. Brank, J., Grobelnik, M., Mladenić, D.: Automatic evaluation of ontologies. In: Kao, A., Poteet, S.R. (eds.) Natural Language Processing and Text Mining, pp. 193–219. Springer London (2007)
3. Gómez-Pérez, A.: Ontology evaluation. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 251–273. Springer Berlin Heidelberg (2004)

4. 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. Kluwer Academic Publishers, Deventer, The Netherlands (1993)
5. Tovar Vidal, M., Pinto Avendaño, D., Montes Rendón, A., González Serna, J.G., Vilariño Ayala, D.: Evaluation of ontological relations in corpora of restricted domain. Computación y Sistemas 19(1) (2015)
6. Tovar Vidal, M., Pinto Avendaño, D.E., Montes Rendón, A., González Serna, J.G.: A metric for the evaluation of restricted domain ontologies. Computación y Sistemas 22(1) (2018)
7. Wackerly, D., Mendenhall, W., Scheaffer, R.: Estadística matemática con aplicaciones. Grupo Editorial Iberoamérica (2002)
8. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., Suárez-Figueroa, M.d.C. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)

ISSN 1870-4069

The Fuzzy Relations between Intervals in a Convolution-based Depiction

Krystian Jobczyk, Antoni Ligęza

AGH University of Science and Technology,
Poland

Abstract. This paper is aimed at the proposing a new convolution-based approach to the representation of fuzzy Allen's relations between fuzzy intervals. It refers to the earlier attempt of H-J. Ohlbach to represent these relations is terms of integrals. The next a framework of a theory for the convolution fuzzy Allen's relations is put forward.

Keywords. Depiction, convolution, fuzzy relations.

1 Introduction

In [1], J. Allen introduced the 13 possible relations between intervals – described later in [2] in modal terms. The intervals from the original Allen's work – as compact subsets of a real line $\mathbb{R}$ – form operationally convenient objects and do not fuzzify any Allen's relations between them. The situation may change radically, when these intervals are exchanged for fuzzy intervals. They form two-dimensional objects in $\mathbb{R}^2$ of a (usually) trapezoidal form. These fuzzy intervals sometimes 'fuzzify' Allen's relations between them. It holds, when some points of an initial fuzzy interval remain in a given Allen's relation ('before', 'later', etc.) with the points of a second fuzzy interval, but some points do not. Anyhow, all the situations ellucidate only a qualitative side of both Allen's and fuzzy Allen's relations.

An interesting attempt to grasp quantitative (computational) aspects of Allen's and fuzzy Allen's relations was put forward in such works of De Cock-Schocker's school as: [15, 14, 13][1]. This proposal forms a kind of a sophisticated calculus, in which fuzzy Allen's relations are expressed in terms of minima, maxima, suprema and infima. An alternative approach to the representation of fuzzy Allen's relations was proposed by H-J. Ohlbach in [12, 11, ?]. In the conceptual framework of his approach, fuzzy Allen's relations are represented by normalized integrals. Some ideas of the paper steem from the earlier integral-based approaches to fuzziness from [5, 9, 4]. The Ohlbach's ideas were adopted and referred to the so-called *Simple Temporal Problem under Uncertainty with Preferences* (STPU) in [7] and developed in [3, 8].

[1] Fuzzy-temporal aspects of Allen's relations was also discussed from a more engineering perspective in [6] and implicitly mentioned in [10].

Motivation and Objectives of the Paper. Meanwhile, one has an impression that Ohlbach's integral-based interpretation is not sufficient. It follows from the following reasons: A) fuzzy Allen's relations are viewed here as fuzzy values of integrals, what seems to be an excessive simplification, B) Ohlbach's approach 'escapes' towards reasonings based on probability theory and statistics instead of real-analysis and algebra-based reasonings and C) it seems that this approach does not (completely properly) emphasize a sense of some definitions of fuzzy Allen's relations. It motivates us to propose an alternative approach to the representation of fuzzy Allen's relations in terms of (normalizable) convolutions as a more adequate solution. According to it – the main objective of the paper is to propose an outline of a convolution-based approach to the representation of fuzzy Allen's relations.

The rest of the paper is organized as follows. In Section 2 a terminological background of the paper analysis is put forward – the Ohlbach's approach in particular. In Section 3, the convolution-based approach to the representation of fuzzy Allen's relations is proposed. Section 4 contains concluding remarks and a brief description of further perspectives.

2 Terminological Background

2.1 Ohlbach's Integral-based Approach to Fuzzy Allen's Relations in a Nutshell

. Ohlbach's approach to the representation of fuzzy Allen's interval relations is two-stages. In the first stage, the so-called *fuzzy Allen's relations of the point-interval type* are considered. In the second one – fuzzy Allen's relations are extended to the so-called *fuzzy Allen's relations of the fuzzy interval- fuzzy interval type*[2]. Both types of relations may be briefly specified as follows.

– *Fuzzy Allen's relations of the point- interval* type. They assert that a point, say p, remains in R-relation to a fuzzy interval j. Symbolically: $R_p(j)$, where R is a chosen Allen's relation and may be represented as a distribution function.
– *Fuzzy Allen's relations of the fuzzy interval- fuzzy interval* type arise, if we blow points p's to a new fuzzy interval, say i. Since each $R_p(j)$ may be interpreted as a distribution function, it also has a density-based representation. Finally, such a newly created fuzzy interval-interval relation $R^{Fuzz}(i,j)$ (if i,j are fuzzy intervals) may be interpreted as expected values of the general form:

$$R(i,j)^{Fuzzy} = \int_{-\infty}^{\infty} i(x)\widehat{R_p}(j)(x)\mathrm{d}x, \qquad (1)$$

where $\widehat{R_p}(j)(x)$ forms a density-based representation of $R_p(j)(x)$ (see:[12, 11, **?**]).

To cut the long story short, Ohlbach proposes to see fuzzy Allen's relations as *normalized integrals of a single variable.*

[2] Fuzzy intervals are representable here as two-dimentional trapezoidal objects.

2.2 Ohlbach's Integral-based Approach to Fuzzy Allen's Relations in Detail

Fuzzy Allen's relations of the points-fuzzy interval type. Assume that some point p, a fuzzy interval j and one of 13 Allen's relations, say R, are given. Observe that one can put the point-interval relation, $R_p j(x)$, which asserts that p is located in a position defined by R with respect to the interval j, etc.

Example 1 *Taking a point p and an interval j, the relation $B_p(j)$ will asserts that p lies 'before' the interval j and $D_p(j)$ asserts that p is 'during' j.*

Let us preface further considerations by some useful observation that each point-interval relation (of Allen's sort) determines its corresponding function.

Collolary 1 *([12, 11])*[3] *A fuzzy point-interval relation R(t, i) is a function that maps a time point t and the interval i to a fuzzy value. Conversely, if i is a fuzzy interval and $R^{'}$ is a function, then R defined as follows:*

$$R(t, i) =^{def} R^{'}(i)(t), \tag{2}$$

is the corresponding fuzzy point-interval relation.

Fuzzy Allen's relations for two fuzzy intervals. Observe now that each such a point-interval Allen's relation may be extended to its corresponding interval-interval relation over a new fuzzy interval – as depicted in Fig. 22. For example, taking a time point t, an interval (not necessary a fuzzy one) j and a point-interval *R(t, j)* we can put:

$$R(i,j) = i(t)R(t, j). \tag{3}$$

It allows us to write: $R(i, j)(t) = f^{i(x)}(t)R(j)(t)$, where $f^{i(x)}(t)$ is a function characterizing an interval i.

Example 2 *In this way one can specify after(i,j) = i(t)after(t, j), before(i,j) = i(t)before(t, j) and all interval-interval Allen's relations.*

Fuzzy Allen's relations for fuzzy intervals. In order to define fuzzy Allen's relations – due to Ohlbach's ideas – let us return to Allen's point-interval relations and consider, say 'before'-relation $B_p(j)(x)$, for a fuzzy interval j and i such that $p \in i$. Due to – [11], this relation may be rendered in terms of the so-called *extend* function E^+ and the complementation operator $N(E^+)$. This function 'behaves' as the functions depicted in Fig. 20 for $L^{\leq}$-relation. Namely, $N(E^+)$ decreases in the right neighborhood of a given point (b-α in Fig.20a) and it increases for arguments being far from it. Let us try to think about $B_p(j)(x)$ determined by $N(E^+)$ in terms of probability theory now.

[3] This corollary was introduced as a definition by Ohlbach in [12].

Allen's point-interval relations in terms of probability theory. Therefore, assume that a probability space Ω of elementary events with a probability measure $P : \Omega \to [0, 1]$ are given. For a given fuzzy interval i we define points of its $\mathbb{R}$-support (See: Figure 18) as elements of Ω. Define also a *random variable* $X : \Omega \to \mathbb{R}$ such that $X(\omega) = X(p) = x \in \mathbb{R}$. In other words, we associate each point p of i-support to a single variable x of a real line. It enables to view $N(E^+)$ as a *distribution* function for i. In fact, $N(E^+)$ in Fig. 20 a) 'represents' a probability the event: $-\infty \leq X = x < b - \alpha$. Formally, $N(E^+)(x) = P(-\infty \leq X = x < b - \alpha))$.

Observe also that such a $N(E^+)$ is a continuous function and $N(E^+) < \infty$. Thus, there exists a function f_X to be called *probability density* – co-definable with $N(E^+)$ as follows:

$$N(E^+)(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x. \tag{4}$$

Summing up, the (fuzzy) point-interval relation $B_p(j)$ in terms of $N(E^+)$ may be interpreted as a distribution for the second fuzzy interval i, that contains p's points. It remains to decide, what might represent the fuzzy interva-interval relation $B(i, j)$.

H-J. Ohlbach postulates to consider a unique expected value for X for in this role, although he did not render this postulate explicitely. In a general case, having a function $\phi : \mathbb{R} \to \mathbb{R}$, the expected value $E(\phi(X))$ is defined as:

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(x)\mathrm{d}(F(x)), \tag{5}$$

where $F(x)$ is a distribution.

Fuzzy Allen's relations of the interval-interval type. It remains to specify this expected value of (38) in our case, or taking $N(E^+)(x) = F(x)$ (as a distribution) in (38). Therefore:

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(x)\mathrm{d}(N(E^+)(x)). \tag{6}$$

Because of (5), we have:

$$\int_{-\infty}^{\infty} \phi(x)\mathrm{d}N(E^+)(x) = \int_{-\infty}^{\infty} \phi(x)\mathrm{d}(\int_{-\infty}^{x} f_X(x)\mathrm{d}x) = \int_{-\infty}^{\infty} \phi(x) f_X(x)\mathrm{d}x,$$

Thus:

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(x) f_X(x)\mathrm{d}x. \tag{7}$$

If put $\phi(x) = i(x)$ as a function characterizing the interval i, we can obtain the required form in our case:

$$E(i(x)) = \int_{-\infty}^{\infty} i(x)\widehat{B_p}(x)\mathrm{d}x, \tag{8}$$

where $\widehat{B_p(x)}$ denotes the fuzzy point-interval 'before' as a density function.

Example 3 *If assume that:*

$$F(x) = N(E^+) = \begin{cases} 0, & \text{for } x \leq a, \\ \frac{x-a}{b-a}, & \text{for } a < x \leq b, \\ 1, & \text{for} x > b, \end{cases} \quad \textit{then } f(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{1}{b-a}, & \text{for } a \leq x \leq b, \\ 1, & \text{for } x > b, \end{cases}$$

$$\textit{and } B(i,j) = E^j_{before}(i(x)) = \tfrac{1}{b-a} \int_{-\infty}^{\infty} i(x)\mathrm{d}x.$$

Taxonomy of fuzzy Allen's relations in Ohlbach's depiction. We have just emphasized how fuzzy Allen's relations may be rendered in terms of integrals. In addition, a general form of them was elaborated for 'before'-relation. In this moment, a complete taxonomy of fuzzy Allen's relations in Ohlbach's depiction will be introduced – due to [12, 11].

Normalization. Obviously, all the integrals above, in particular (37) and (42), are finite. In particular, (37) holds, if and only if the integral on the right side of (37) is finite. It is warranted by the fact that $N(E^+)$ is assumed to be finite. Whereas finiteness of the integrals constitutes a sufficient condition from a purely mathematical point of view, it is unsatisfactory to consider (42) and similar conditions as the adequate representations for fuzzy Allen's relations. In fact, we expect that these integrals will take fuzzy values from $[0, 1]$, so they should be normalized.

Different methods of normalization is known and used. For 'before'-relation, the factors $|i|$ and $|i|_a^b$ defined as follows:

$$|i| = \int_{-\infty}^{\infty} i(x)\mathrm{d}x, \quad |i|_a^b = \int_a^b i(x)\mathrm{d}x. \tag{9}$$

By contrast, $meet(i, j), start, finishes$ require the normalization factors of the type $N(i, j)$, so as dependent on both i and j. Ohlbach argues in [12], p. 26) for a choice of the following two factors of this type, as they admit also 1 as a possible value:

$$N(i,j) = \max_a \int_{-\infty}^{\infty} i(x-a)j(x)\mathrm{d}x, \quad N(i,j) = \min(|i|, |j|). \tag{10}$$

An outline of Ohlbach's taxonomy. These general assumptions and observations allows us to introduce the whole taxonomy of the integral-based representation of Allen's relations. Putting aside its detailed presentation, we illustrate how fuzzy Allen's relations are defined in this approach in two cases.

1. Meet^{Fuzzy} (see [12, 11]): This definition is based on the observation that 'meet'-relation holds between two fuzzy intervals if and only if there are some functions, say *Fin(i)* and *St(j)*, that 'cut' the initial points from the first interval i and the final ones from the second j interval. It allows us to note the following.

– If i or j are empty, they cannot meet, so the relation yields 0.

– Similarly, if i is $[a,\infty)$-type, j is $(-\infty,a)$ and conversely, for a given fixed a,
– Otherwise, one can define this relation as a statement that the 'end' part *Fin(i)* of the interval i touches the initial part *St(i)* of the interval j. The factor *N(Fin(i), S(j))* normalizes this integral to be smaller than 1.

It leads to the following depiction:

$$\textit{meet(i,j)}^{Fuzzy} = \begin{cases} 0 & \text{if } i=\emptyset \text{ or } j=\emptyset \\ & \text{or } i=[a,\infty)\ \wedge j=(-\infty,a), \\ & \text{or } j=[a,\infty)\ \wedge i=(-\infty,a), \\ \displaystyle\int_{-\infty}^{\infty} \frac{Fin(i)St(j)dx}{N(Fin(i),St(j))} & \text{otherwise.} \end{cases}$$

2. Before. A similar way of reasoning enables of defining other Allen's relations. For example, 'before' $B(i,j)^{Fuzzy}$ is defined as follows (a detailed justification may be found in $[12,11,\textbf{?}]$. As earlier, $(i\cap j)\widehat{B(j)}$ forms a density-based representation of $B_pj(x)$ blown up over the intersection of fuzzy intervals i and j.):

$$\textit{B(i,j)}^{Fuzzy} = \begin{cases} 0 & i=\emptyset \text{ or } i=[a,\infty) \text{ or } j=\emptyset, \\ 1 & i=(-\infty,a] \text{ or } i\cap_{\min} j=\emptyset, \\ \displaystyle\int_{-\infty}^{\infty}(i\cap j)\widehat{B(j)}dx/|i\cap_{\min} j| & i=(-\infty,a], j \text{ is bounded} \\ & \text{or } j=[a,\infty), \\ \displaystyle\int_{-\infty}^{\infty} i\widehat{B(j)}dx/|i| & \text{otherwise.} \end{cases}$$

Further examples of fuzzy integral Allen's relations may be found in $[12,11]$.

2.3 Conceptual Framework of Further Analysis

The notion of convolution, that we need, requires a new conceptual framework. Its determined by a class of Lebesgue integrable functions on $\mathbb{R}$, denoted by $L(\mathbb{R})$. This class forms a unique example of the so-called *Banach spaces*.

In order to describe both types of spaces, assume that X is a given vector (linear) space. Each vector space is defined over a scalar field, say K. This fact is denoted by $X(K)$. The usual scalar fields are: the field R of real numbers $\mathbb{C}$, or over the field $\mathbb{C}$ of complex numbers. We write then: $X(\mathbb{R})$ and $X(\mathbb{C})$ (*resp.*) to render the fact that X is defined over $\mathbb{R}$ or $\mathbb{C}$.

Assume also that $X(\mathbb{R})$ is given. Let us introduce now a new function $\|\bullet\|$: $X(\mathbb{R}) \mapsto [0,\infty)$ that respects the following conditions:

$$\|x\| = 0 \iff x = 0,\ \ \|\alpha x\| = |\alpha|\|x\|,\ \text{for } \alpha\in\mathbb{R}, \tag{11}$$

$$\|x+y\| \le \|x\| + \|y\|. \tag{12}$$

This function is to be called *a norm* and the whole space $(X(\mathbb{R}), |\bullet|)$ forms a *normed space*. A *Banach space* is such a *normed vector space* X, which is complete with respect to that norm, that is to say, each Cauchy sequence $\{x_n\}$ in X converges to an element x in X, i.e. $\lim_{n\to\infty} x_n = x$.

Example 1. $(\mathbb{R}, |\bullet|)$ with the norm $\|x\| = |x|$, for each $x \in \mathbb{R}$ is a Banach space.

Some special examples of Banach spaces – that are especially interesting for us – are presented in the table below.

Type of spaces	**Abbr.**	**Elements of the space**	**Norms**
The space of Lebesgue integrable functions on $\mathbb{R}$ 'in square'	$L^2(\mathbb{R})$	Functions $f, g, h, \ldots$ - Lebesgue integrable on $\mathbb{R}$	$\|f\| = \left(\int |f|^2 \mathrm{d}x\right)^{\frac{1}{2}}$
The space of Lebesgue integrable functions on $\mathbb{R}$ 'in p'	$L^p(\mathbb{R})$	Functions $f, g, h, \ldots$ - Lebesgue integrable on $\mathbb{R}$	$\|f\| = \left(\int |f|^p \mathrm{d}x\right)^{\frac{1}{p}}$, $1 < p < \infty$

Definition 1 ***(Convolution.)*** *Let us assume that functions f and g are Lebesgue integrable in $\mathbb{R}$, i.e they belong to $(L^1(\mathbb{R}, \|\bullet\|)$. Then the convolution of f and g – denoted by $f * g$ – is usually defined as follows:*

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x - t)g(t)dt\,, \tag{13}$$

where the right side is an improper (Riemann) integral.

The proofs will also exploit the concept of (Lebesque) measurable functions and the Beppo-Levi's monotone convergence theorem.

Definition 2 ***Beppo-Levi's theorem*** *Assume that a non-decreasing sequence $\{f_k\}$ of measurable non-negative functions $f_k : X \to [0, +\infty]$, for a given measure space (Ω, σ, μ), $X \in \sigma$, is given. If $f(x) = \lim_{k\to\infty} f_k(x)$, for each $x \in X$, then:*

$$\int_X f(x)\mathrm{d}\mu = \lim_{k\to\infty} \int_X f_k(x)\mathrm{d}\mu.$$

3 Fuzzy Allen's Relations in a Computational Depiction

Before we introduce the convolution-based representation of fuzzy Allen's relations in order to grasp the computational aspects of these relations, let us return to the initial case of 'meet'-relation in the Ohlbach's integral depiction. It was already said that this definition is based on two functions: $Fin(i)$ and $St(j)$ that 'cut' the initial part of i and the final part of j (resp.).

In fact, both intervals should be integrated together, but both functions should be considered as 'running' in different directions: $Fin(i)$ running towards the second j interval and $St(j)$ as running in the inverse direction – towards the interval i. It means that $St(j)$ should be rather consider as a function of a new argument, say t, and $Fin(i)$ as $Fin(i)(x-t)$[4]. Meanwhile, such a combination of $Fin(i)$ and $St(j)$ of mutually independent functions is given by a *convolution* of them[5].

Finally, there exists another argument for the convolution-based representation of fuzzy Allen's relations. Namely, it is a known fact of real analysis that convolutions take finite values. It means that their use would make a mathematical discussion on fuzzy Allen's relations less conditional than the Ohlbach's integral approach. In fact, a success of Ohlbach's approach is only possible provided that the appropriate integrals are finite. In the convolution-based approach this problem disappears thanks to this elementary property of convolutions.

3.1 Fuzzy Allen's Relations in the Convolution-based Depiction

In this convention, fuzzy Allen's relations from [12, 11, **?**] should be rather rendered as follows:

$$meet(i,j)^{Fuzzy}(x) = \int_{-\infty}^{\infty} \frac{Fin(i)(x-t)St(j)(t)dt}{N(Fin(i)(x-t), St(j)(t))}. \tag{14}$$

(Note that $\mathrm{meet(i,j)}^{Fuzzy}$ is a function of x-argument as we integrate with respect to the second argument t.) Similarly, one could modify, for example, 'before' -relation:

$$before(i,j)^{Fuzzy} = \int_{-\infty}^{\infty} i(x-t)\widehat{B}(j)(t)\mathrm{d}t/|i|. \tag{15}$$

Fuzzy Allen's relations as norms of convolutions. It seems that an idea to represent fuzzy Allen's relations in terms of convolution is already the appropriate one. As illustrated, convolutions better 'encode' an idea to combine two functions and they are computationally convenient – as they are finite. Finally, they might be also normalized.

Nevertheless, one can argue that convolutions still are not ideal in this role. In order to illustrate this fact, let us consider the following paradoxical dichotomy.

A On one hand, we can have *two different* fuzzy Allen's relations, say $R_1^{Fuzzy}(i,j)$ – R_2^{Fuzzy} given by two normalized convolutions C_1 and C_2 (resp.)– that take the same values $\alpha \in [0,1]$ in some area. (They diagrams are identical in this area).

---

[4] The argument $x-t$ ensures that i and j meet together. Thus, $Fin(i)$ should be seen as a function of an argument x in t-translation. $|i|$ is a normalization factor

[5] Note that this situation may be seen as a 'combination' of two independent signals running in inverse directions that are represented in physics by convolutions.

B On the other hand, we can have *a single* fuzzy Allen's relation, say $R^{Fuzzy}(i, j)$ – depicted by a normalized convolution C – that may be multiplied by a scalar $\alpha \in \mathbb{R}$. This new convolution αC would take another values than C alone, although it represents the same relation $R^{Fuzzy}(i, j)$.

Obviously, we are willing to consider C and αC as mutually linked (as they represent $R^{Fuzzy}(i, j)$ and $\alpha R^{Fuzz}(i, j)$). Simultaneously, we want to see C_1 and C_2 as mutually independent – as they represent different relations $R_1^{Fuzzy}(i, j)$ and $R_2^{Fuzzy}(i, j)$ – even though their diagrams are partially identical.

It seems that considering *norms from convolutions* as the alternative representation of fuzzy Allen's relations (instead of convolutions themselves) allows us to avoid these difficulties. For a confirmation of this hypothesis let us assume that $R^{Fuzzy}(i, j)$ is represented now by two norms, say $\|\|_1$ and $\|\|_2$. Formally, we postulate:

$$R^{Fuzzy}(i, j) = \|C\|_1, \quad R^{Fuzzy}(i, j) = \|C\|_2. \tag{16}$$

Let us state that $\|C\|_1$ and $\|C\|_2$ may be viewed as *mutually equal* provided that there are such real constants $a, b < \infty$ that:

$$a\|C\|_1 \leq \|C\|_2 \leq b\|C\|_1{}^6. \tag{17}$$

This condition allows us to identify two norm-values associated to a given fuzzy Allen's relation. Since such a mutual equivalence of norms is sufficient in our approach, it delivers an argument to represent fuzzy Allen's relations by *norms of convolutions*. Formally, if $R(i, j)$ denotes a convolution 'basis' of $R^{Fuzzy}(i, j)$, then $R^{Fuzzy}(i, j)$ can be written:

$$R^{fuzzy}(i, j) = \|R(i, j)\|_{L(\mathbb{R}^1)}. \tag{18}$$

Example 2. Assume that Allen's relation 'before' in a convolutive representation $B(i, j) = f^i \star B_p(j)$ is given, for some functions f^i and $B_p(j)$. Then fuzzy 'before' B^{fuzzy}:

$$B^{fuzzy}(i, j) = \|B(i, j)\|_{L(\mathbb{R}^1)}. \tag{19}$$

But $B(i, j)$ is a convolution, so $B(i, j)(x) = \int_{-\infty}^{\infty} f^i(x - t)B_p(j)(t)\mathrm{d}t$. Assuming that $B(i, j)(x) \in L(\mathbb{R})$ with the norm $\|f\| = \int |f(x)|\mathrm{d}u$, for each $f \in \Big(L(\mathbb{R}^1), \|\ \|\Big)$, we have:

$$B^{fuzzy}(i, j) = \| \int_{-\infty}^{\infty} f^i(x-t)B_p(j)(t)\mathrm{d}t\|_{L(\mathbb{R}^1)} = \int | \int_{-\infty}^{\infty} f^i(x-t)B_p(j)(t)\mathrm{d}t|\mathrm{d}\mu. \tag{20}$$

[6] This property is a known property of norms.Note that this condition is satisfied in our case. In fact, it is enough to put $a = b = \alpha$.

The key idea of our proposal is briefly expressed as follows.

Type of relations:	Given by:	Examples:
Fuzzy Allen's relations of the 'interval-interval' type	L($\mathbb{R}$)-norms of convolutions	$R^{fuzzy}(i,j) = \|R(i,j)(x)\|_{L(\mathbb{R})} = \int (\int_{-\infty}^{\infty} f^i(x-t)R_p(j)(t)\mathrm{d}t)\mathrm{d}\mu$

In last part of this section, we intend to prove two computational features of fuzzy Allen's relations in a convolution depiction. The first theorem shows that these relations are normalizable, the second one – that their diagrams are, somehow, predictable as they are uniformly continuous.

Theorem 1. *Convolution-basis of fuzzy Allen's relations are normalizable.*

Proof: It follows from the above fact that:
$\int_{-\infty}^{\infty} f^{i(x)}R(t,j)dt < \infty$ and the fact that – due to Fubini's theorem – this integral belongs to $L^1(\mathbb{R})$. It means that there is an upper bound for it, say M. It is enough now to put a normalization factor N to ensure that $0 \leq \frac{M}{N} \leq 1$. □

Theorem 2. *Let* $1 \leq p < \infty$, $f_i, R(j) \in L_p(\mathbb{R})$, $\|R(j)\| \leq M$, *for some* M *where* i_f *is a characteristic function for a fuzzy interval* i *and* $R(j)$ *is a functional representation of a point-interval relation (of Allen type) with respect to a fuzzy interval* j. *Assume also that* f_i *is uniformly continuous on* $\mathbb{R}$. *Then* $f_i \star R(j)$ *is* ***uniformly continuous*** *on* $\mathbb{R}$, *too:*

Proof: Since f_i depends on $x - t$ and $R(j)$ dependent on t, let us establish $x - t = z$. It easy to see now that there is a $\rho > 0$ such that $z_1, z_2 \in \mathbb{R}$ and $\|z_1 - z_2\| < \rho$ implies $\|f_i(z_1) - f_i(z_2)\| < \epsilon$. However, f_i is assumed to be uniformly continuous in $\mathbb{R}$, i.e. there is a $\rho > 0$ such that for all $z_1, z_2 \in \mathbb{R}$, $|z_1 - z_2| < \rho$ implies $\|f_i(z_1 - f(z_2)\| < \epsilon$. Then $\|z_1 - z_2\| < \rho$ also implies:

$$\|i_f \star R(j)(z_1) - f_i \star R(j)(z_2)\| = \Big(\int (|i_f(z_1) - f_i(z_2)| \bullet |R(j)(t)|)^p dt\Big)^{\frac{1}{p}}$$

$$= \|(f_i(z_1) - f_i(z_2)) \bullet R(j)\|_p \leq \|f_i(z_1) - f_i(z_2\| \bullet \|R(j)\|_p < \epsilon_1,$$

where $\epsilon_1 = M\epsilon$. Obviously, $\epsilon_1 \to 0$. The last inequality follows from the fact that $\|R(j)\|_p \leq M$ and from Schwartz's inequality $\|xy\| \leq \|x\|\|y\|$, for each $x, y \in L^p(\mathbb{R})$, $1 < p < \infty$.

The next theorem illustrates a computational power of the convolution-based approach. In fact, it forms a unique version of *Borel's Convolution Theorem* for convolutions used for defining fuzzy Allen's relations. It will be briefly called: 'Convolution Theorem for fuzzy Allen's relations'.

Theorem 3. ***(Convolution Theorem for fuzzy Allen's relations)****. Let* i, j *be fuzzy intervals. Let also* $f^i, R(t,j) \in L^1[-\infty, \infty]$ *be a function characterizing a* *fuzzy interval i and a point-interval Allen relation (resp.). Then their convolution* $h = f^i * R$ *has the following property:*

$$h(x) = \int_{-\infty}^{\infty} |f^{i(x)}(x-t) * R(j)(t)|\mathrm{d}t = \int_{-\infty}^{\infty} |f^{i(x)}(x-t)|\mathrm{d}x \int_{-\infty}^{\infty} |R(j)(t)|\mathrm{d}t.$$

Simply:

$$\mathcal{F}(f^i * R(j)) = \mathcal{F}(f^i)\mathcal{F}(R(j)). \tag{21}$$

Proof: Assume that $f^i, R \in L^1(\mathbb{R}^1)$ – as in the formulation of this theorem – are given. In order to compute their convolution:

$$h(x) = \int_{-\infty}^{\infty} |f^{i(x)}(x-t) * R(j)(t)|\mathrm{d}t, \tag{22}$$

we exploit Fourier transforms of f^i and R, i.e. $\mathcal{F}(f)$ and $\mathcal{F}(g)$ (respectively) – defined as follows:

$$\mathcal{F}(f^i) = \int_{\mathbb{R}} f^{i(x)}(x-t)e^{-2\pi ixv}\mathrm{d}x \ \text{ and } \mathcal{F}(R) = \int_{\mathbb{R}} R(j)(t)e^{-2\pi ixv}\mathrm{d}t.$$

Therefore, for all $y \in \mathbb{R}$, Fourier transform $F(f * g)$ is as follows.

$$\begin{aligned} \mathcal{F}(f^i * R(j)) = \mathcal{F}(h)(x) = \mathcal{F}(\int_{-\infty}^{\infty} |f^{i(x)}(x-t)R(j)(t)|dt) \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f^{i(x)}(x-t)R(j)(t)|\mathrm{d}te^{-2\pi ixy}\mathrm{d}x \\ = \int_{-\infty}^{\infty} R(j)(t)\int_{-\infty}^{\infty} f^{i(x)}(x-t)e^{-2\pi ixy}\mathrm{d}x\mathrm{d}t. \end{aligned} \tag{23}$$

By substitution $x - t = u$, or $x = t + u$ (and $\mathrm{d}x = \mathrm{d}u$) we obtain:

$$\begin{aligned} \int_{-\infty}^{\infty} R(j)(t)\int_{-\infty}^{\infty} f^{i(x)}(x-t)e^{-2\pi ixy}\mathrm{d}x\mathrm{d}t \\ = \int_{-\infty}^{\infty} R(j)(t)\Big(\int_{-\infty}^{\infty} f^{i(x)}(x-t)e^{-2\pi i(t+u)y}\mathrm{d}u\Big)\mathrm{d}t. \end{aligned}$$

Applying Fubini theorem in order to interchange the order of limitation we can write:

$$\begin{aligned} \int_{-\infty}^{\infty} R(j)(t)\Big(\int_{-\infty}^{\infty} f^{i(x)}(x-t)e^{-2\pi i(t+u)y}\mathrm{d}u\Big)\mathrm{d}t \\ = \int_{-\infty}^{\infty} R(j)(t)e^{-2\pi ity}\mathrm{d}t\int_{-\infty}^{\infty} f^{i(x)}(x-t)e^{-2\pi iuy}\mathrm{d}x = \mathcal{F}(f^i)\mathcal{F}(R(j)). \end{aligned}$$

Therefore,

$$\mathcal{F}(f^i * R(j)) = \mathcal{F}(f^i)\mathcal{F}(R(j)), \tag{24}$$

what finishes the proof.

4 Conclusions

It has already been shown how fuzzy Allen's relations may be depicted in a computational, convolution-based approach. It seems that this approach serves a kind of an improvement and a further development of the Ohlbach's integral approach towards a theoretic well-founded calculus. This solution allows us to omit some theoretic difficulties of Ohlbach's approach, such as an unexpected infiniteness of his integrals. Research on fuzzy Allen's relations may be extended in (at least) two different directions. At first, some approximation methods (such as Hardy-Littlewood Theorem) for fuzzy Allen's relations in the convolution-based depiction may be directly adopted. Secondly, the convolution-based approach may be developed towards a new algebraic reinterpretation of Allen's algebra in terms of incidence algebra. Nevertheless, it requires a deeper analysis.

References

1. Allen, J.: Maintaining knowledge about temporal intervals. In: Communications of ACM. pp. 832–843 (1983)
2. Halpern, J., Shoham, Y.: A propositional modal logic of time intervals. Journal of the ACM 38, 935–962 (1991)
3. Jobczyk, K.: Temporal planning with fuzzy constraints and preferences. Phd thesis, University of Caen (2017)
4. Jobczyk, K., Ligeza, A.: Fuzzy-temporal approach to the handling of temporal interval relations and preferences. Proceeding of INISTA pp. 1–8 (2015)
5. Jobczyk, K., Ligeza, A.: Multi-valued Halpern-Shoham logic for temporal Allen's relations and preferences. In: Proceedings of the Annual International Conference of Fuzzy Systems (FuzzIEEE). pp. 217–224 (2016)
6. Jobczyk, K., Ligeza, A.: Systems of temporal logic for a use of engineering: Toward a more practical approach. In: Intelligent Systems for Computer Modelling. pp. 147–157 (2016)
7. Jobczyk, K., Ligeza, A.: Towards a new convolution-based approach to the specification of STPU-solutions. In: FUZZ-IEEE. pp. 782–789 (2016)
8. Jobczyk, K., Ligeza, A.: The hybrid plan controller construction for trajectories in Sobolev space. In: ICAISC'17. pp. 532–543 (2017)
9. Jobczyk, K., Ligeza, A.: An epistemic Halpern-Shoham logic for gradable justification. In: Proceedings of the Annual International Conference of Fuzzy Systems (FuzzIEEE) (2018)
10. Jobczyk, K., Ligeza, A.: Temporal traveling salesman problem – in a logic- and graph-theory-based depiction. In: Proceeding of ICAISC'18. pp. 544–556 (2018)
11. Ohlbach, H.: Relations between time intervals. In: 11th Internal Symposium on Temporal Representation and Reasoning. vol. 7, pp. 47–50 (2004)
12. Ohlbach, H.J.: Fuzzy time intervals and relations-the FuTIRe library. Research Report PMS-04/04, Inst. f. Informatik, LMU Munich (2004)
13. Shockert, S., Cock, M.: Temporal reasoning about fuzzy intervals. Artificial Intelligence 172(8-9), 1158–1193 (2008)
14. Shockert, S., Cock, M., Kerre, J.: Efficient algorithms for fuzzy qualitative temporal reasoning. IEEE Transactions on Fuzzy Systems (2000)
15. Shockert, S., Cock, M., Kerre, J.: Imprecise temporal interval relations. pp. 108–113 (2006)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx