Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 293–304

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Improving the Real-Time Searching in the Organizational Memory

María Laura Sánchez Reynoso, Mario Diván*

Economic and Law School, National University of La Pampa, Coronel Gil 353 1st floor, Santa Rosa, CP6300, Argentina

Abstract

The real-time data processing constitutes a critical area when talking about real-time decision making. Strong decisions are based on recommendations for describing the associated course of actions, but the real-time processing gives a very short time for searching them. The Processing Architecture based on Measurement Metadata is a data stream engine oriented to measurement projects, which supports the decision making through an organizational memory. The search space related to the organizational memory is initially in-memory limited using the structure of the measurement projects. Given a project, the related projects are ordered based on a given scoring from its structural definition. Here, a new structural coefficient based on the text similarity, which is computed from the textual definition of each descriptive attribute of a project is introduced. This allows better scoring of the related projects, even when its definitions could be affected by human errors or multiple definitions. The scoring is critical when in a given situation, a project has not specific experience for recommending, in such context, the recommendations from the near projects are served. The pabmm_sh library is outlined and a simulation on its associated processing times for the similarity computing are introduced based on the token definition for a measurement project. The library adds a new alternative perspective in the processing architecture for driving the searches into the organizational memory. It can update 2000 projects less than 1 second, keeping the individual processing time of each project under 1 millisecond.

Keywords: Real-Time Searching; Organizational Memory; Measurement Projects; Structural Coefficient

* Corresponding author. Tel.: +54 (2954) 451637; fax: +54 (2954) 451637.
E-mail address: mjdivan@eco.unlpam.edu.ar

10.1016/j.procs.2019.06.043

1. Main text

The data processing is cheaper every day thanks to the technology evolution jointly with the scale economy, even the big data repositories and the data streams are fed allowing the data-driven decision making as a natural aspect in the different organizations (e.g. the governments) [1]. The measurement and evaluation constitute a logic step for determining the current state of a concept under monitoring and its posterior evaluation. One of the applications of the data-driven decision making is the monitoring of entities, which is especially useful for keeping track of an entity (e.g. a person), jointly with the characterization of its behavior [2].

The Processing Architecture based on Measurement Metadata (PAbMM) [3] is a data stream engine based on the Apache Storm (storm.apache.org), which is specialized in the Measurement and Evaluation (M&E) projects. The M&E projects are defined in terms of a measurement and evaluation framework named C-INCAMI (Context – Information Need, Concept Model, Attribute, Metric and Indicator) [4]. The framework allows establishing the concepts, terms and their relationships before receiving any measure from the data sources. The data sources could be heterogeneous, and they are out of the control of the processing architecture. For that reason, it constitutes an auspicious environment for the using of Internet of Things (IoT) devices [5].

Using the M&E project definition, the measures are interchanged among the data sources and the processing architecture by mean of a measurement adapter. The measurement adapter is located on the mobile devices and it is responsible for the translating between the original data format from each data source and the expected data format at the PAbMM. Once the measures are received at the PAbMM, the architecture has three operative lines: i) The first operative line is related to the data stream replication to thirds by subscription; ii) The second line is associated with the real-time data analysis, decision-making and recommending when some typified situation is detected; and iii) the last operative line is responsible for storing the data, once the synthesis algorithm was run (e.g. in place of write each temperature of the day, just the changes of temperature along the day are kept).

For detecting the typified situations and for supporting the decision-making in real-time, the processing architecture uses an organizational memory. The organizational memory is integrated by i) The historical data; ii) The M&E projects definitions; and iii) The previous experiences and knowledges from the experts related to each project. The historical data are useful for answering ad-hoc queries and training the incremental classifiers based on Hoeffding Trees, which allows detecting typified situations [6]. The M&E project definitions allows informing to the measurement adapter and PAbMM, the details about the origin of each data and the associated entity's characteristic to be processed. Finally, the previous experience and knowledge from the experts are useful for carrying forward the recommendations of the courses of actions related to typified situations.

However, it is possible that a typified situation has no previous experience (e.g. it could be a new situation) and in that case initially there is not available recommendations. Thus, and for avoiding this situation, a pair of structural and behavioral coefficients was proposed based on the project definition [7]. The structural coefficient basically looking for common attributes which characterize the entities under analysis, and for there, an initial order allows find the similar entities and reuse its experience. For improving the precision, the behavioral coefficient analyzes the common attributes of the entities based on the correlation between the data distribution. Thus, a scoring contemplating the structure and its associated behavior allows looking for similar entities for reusing its experience and knowledge.

The concern related to the structural coefficient is that the idea of "common" is established based on the attribute ID of each entity. That is to say, two entities have a common attribute if they use the same attribute ID, but what happens if they do not do that? When PAbMM had a little number of M&E projects, the characterization of each entity under analysis was easy to see, that is to say, it could be visually reviewed. However, when the number of projects grew-up, the experts who define each project, could define the same attribute similarly or using different ID, but not exactly equal. This provoked that the structural coefficient was affected because the idea of the same attribute was redundant and even with different IDs, impacting directly in the search strategy on the organizational memory (i.e. two projects left to be similar, because of their attribute IDs).

The main contributions are synthesized as follows: i) A new structural coefficient computed from the text definition of each entity's attribute is incorporated. This allows determining the similarity or not of each attribute based on its textual definition contained in the CINCAMI/Project Definition [8] schema and using different kinds of text similarity coefficients for that aim. In addition, even with redundancy or different definitions, the attributes with

a similarity coefficient upper than a given threshold are considered equals or commons to the effect of looking for similar experiences; and ii) A new pabmm_sh library is released under the terms of the Apache 2.0 license for extending the functionality of PAbMM related to the structural coefficient, and for its free use in other projects.

The article is organized as follows. Section 2 introduces some related works. Section 3 synthesizes the project definition schema based on the measurement and evaluation framework. Section 4 describes the background related to the new structural coefficient and its relationship with the project definition used by PAbMM. Section 5 outlines the simulation result based on a typical and conceptual project definition. Section 6 synthesizes some conclusions and future works.

2. Related Works

Zhou, Wu, and others [9] propose a model based on the user navigation and the learning of the navigational preferences, the different kinds of social relationship, and the mapping for homogenizing the data structure keeping their properties under a supervised model. The idea related to the model is to try to predict the next link related to the navigational behavior of a given user. In this sense, our proposal is oriented to looking for similar M&E projects when the previous experience is not available.

Amüller, Christiani, Pagh and Silvestri [10] propose the distance-similarity hashing as a way to generalize the Locality-sensitive hashing (LSH). This kind of techniques is very interesting in problems in which the dimensionality is high by definition such as happens in the text processing. The underlying idea related to LSH is to get a collision when some texts are similar, making easy the document classification or the initial classification of them. However, DSH advance on a collision probability function which incorporates the possibility of distance among the points with the same hash. Our proposal uses different text similarity coefficients for determining the distance or similarity between the definitions related to each attribute (i.e. plain text) of a given project, such as Jaccard, Cosine, Sorensen-Dice, Jaro-Winkler, and Levenshtein [11, 12].

Chaidaroon, Ebesu, and Fang [13, 14] introduce a technique for deep semantic hashing oriented to identify similar text by the collision. Thus, the aim is to detect similar texts using neural networks. The underlying idea is similar to the semantic hashing; however, the original model requires a training dataset in which each data has its corresponding label for the training stage. Once the model is trained, the neural network is able to be applied, but the training is a bottleneck. For this reason, in [13] a method based on the unsupervised ranking methods is outlined capitalizing using a weak supervision.

Le and Mikolov [15] proposed a ParagraphVector model which in front of the high-dimensionality data such as text, propose a vectorization keeping the idea related to the text position and the meaning, deriving in the possibility of identifying text similarity based on its meaning. In [16] an extension of ParagraphVector is proposed with the aim of modeling the hierarchical data structure and be able to capture its associated semantic as a vector.

3. The Project Definition and its Impact in the Processing Architecture

The key idea related to a measurement and evaluation process is to reach comparability in its associated results (i.e. the measures), jointly with the possibility of repeat the process independently who perform it. In addition, the possibility of extending the measurement process aligned with new requirements is essential for keeping valid the project [4]. Thus, the aim of the C-INCAMI framework is to reach a common understanding which establishes the concepts, terms, and relationships among them necessaries for carrying forward and implementing a measurement project.

The processing architecture, or simply PAbMM in forward, is specialized in M&E projects mounted on the Apache Storm ecosystem. The basic idea is to define an M&E project in terms of an M&E framework, broadcast its definition between all the participants, and once that is managed, being ready to process each data stream related to the entity under monitoring. Thus, once the project has been defined, the definition is broadcasted between the data collectors, data pre-processors and the processing architecture for reaching a common understanding (e.g. all the participants know the entity under analysis, the attributes who characterize it, etc.). Once the common understanding has been reached, the data collector gather the data, the data pre-processors convert the data under an understandable

way for all the participants (i.e. from the raw data of each sensor to the CINCAMI/Measurement Interchange Schema [3]), and PAbMM receives and processes the data in real-time.

In this way, both the data collectors as the processing architecture depends directly from the M&E project definition for gathering the data and processing it too. Moreover, each stream is organized under the CINCAMI/MIS schema using the M&E project definition, and in the same way, the processing architecture interprets the data stream under the same schema using the project definition. Any change in the project definition must be replicated in each participant for reaching the common understanding.

For better understanding, this section is organized into two topics. The first topic is related to explain the idea of the project definition, the relationship with the M&E framework and its applicability in PAbMM. The second topic introduces the project definition schema and its role related to the processing architecture.

3.1. The Project Definition based on the Measurement and Evaluation Framework

The C-INCAMI framework is guided by the information need, that is to say, previous to identify the entity under monitoring, it is necessary to establish the information need related to a given project (e.g. Avoid severe damages through the prevention of risks with a direct impact in the outpatient health). Once the information need has been established, the entity category could be defined (for example, an outpatient, person, etc.). It seems to be a simple step, but it is critical, because in these two concepts the project (and all of the actors) must understand why the project is necessary and what is the related aim [4].

Thus, the next step is to define the way in which an entity category could be characterized. In this step, each entity category is described by mean of a set of attributes who represent an entity characteristic aligned with the information need of the project. For example, if the entity category be a person, an attribute could be the corporal temperature. The idea of the set of attributes is to characterize an entity through the different discrete aspects that conform to it. Because the entity is immersed in an environment, such environment is defined under the idea of the context. In this way and in analogous way, the context is described by their attributes known as the context properties. Each context property is to the context, the same that the attribute is for the entity [17].

Each attribute has an associated metric who defines the way in which it is quantified. Each metric incorporates the scale, unit, calculation method, etc. Because a context property is a specialization of the attribute for the context, each context property is quantified through a metric as well. In this way, both the attributes as context properties are quantified by mean of the metrics, but just one metric must be chosen for a given attribute or context property. Once the metric was defined for an attribute or context property, each actor along the data processing system knows the association between the metric with the attribute for a given entity, or between the metric and a context property.

However, when the metric provides the measures from the data source (i.e. the instrument for measuring) to the PAbMM, the architecture receives the value of a given metric (i.e. the measure) but it does not know anything about the meaning of the value. That is to say, even when the relationship between metric and attribute is clearly known, the meaning of each value and its interpretation requires the knowledge from the experts in the specific domain. For this reason and with the aim of interpreting each value of the metrics, the indicators are defined and associated with a metric. Each indicator incorporates the knowledge from the experts under the way of the decision criteria, which allows interpreting the value at the light of criteria established by experts with deep knowledge around the entity monitored and its associated context.

For example, following the initial example of this topic, given an outpatient under monitoring, the experts could recommend monitoring its heart rate and the axillary temperature, jointly with the environment temperature, environment pressure, and humidity for its immediate context. Once more time, the value of the axillary temperature alone does not say much, for example, 38ºC is good or not for the outpatient? Here is where the indicators are defined, precisely they are useful for interpreting the value of the metrics using the knowledge from the experts through the decision criteria. Thus, a specialist could define that any temperature out of the range (36.0; 37.4) should be notified to the medical center because would be abnormal for this kind of outpatient.

In this way, starting by the M&E project, the information need, the entity under monitoring, attributes, the associated context jointly with the context properties, and ending with the indicators all of them have a specific ID which allows its individual identification along the different components related to the PAbMM. The next topic related to the project definition schema synthetically describes the way in which this kind of definition is

interchanged among the components, fostering its interoperability and the common understanding required at the moment in which the data should be processed.

3.2. The Impact of the CINCAMI/Project Definition Schema at PAbMM

The project definition schema based on the CINCAMI framework was defined with the aim of fostering the interoperability related to the M&E projects. Basically, this proclaimed common understanding related to the M&E project definition should be understood by the different actors involved along the measurement process. The CINCAMI/Project Definition (CINCAMI/PD) is the way in which the M&E project definitions are interchanged not just among the components of the processing architecture, but also among the different systems which require the use of this kind of library [8].

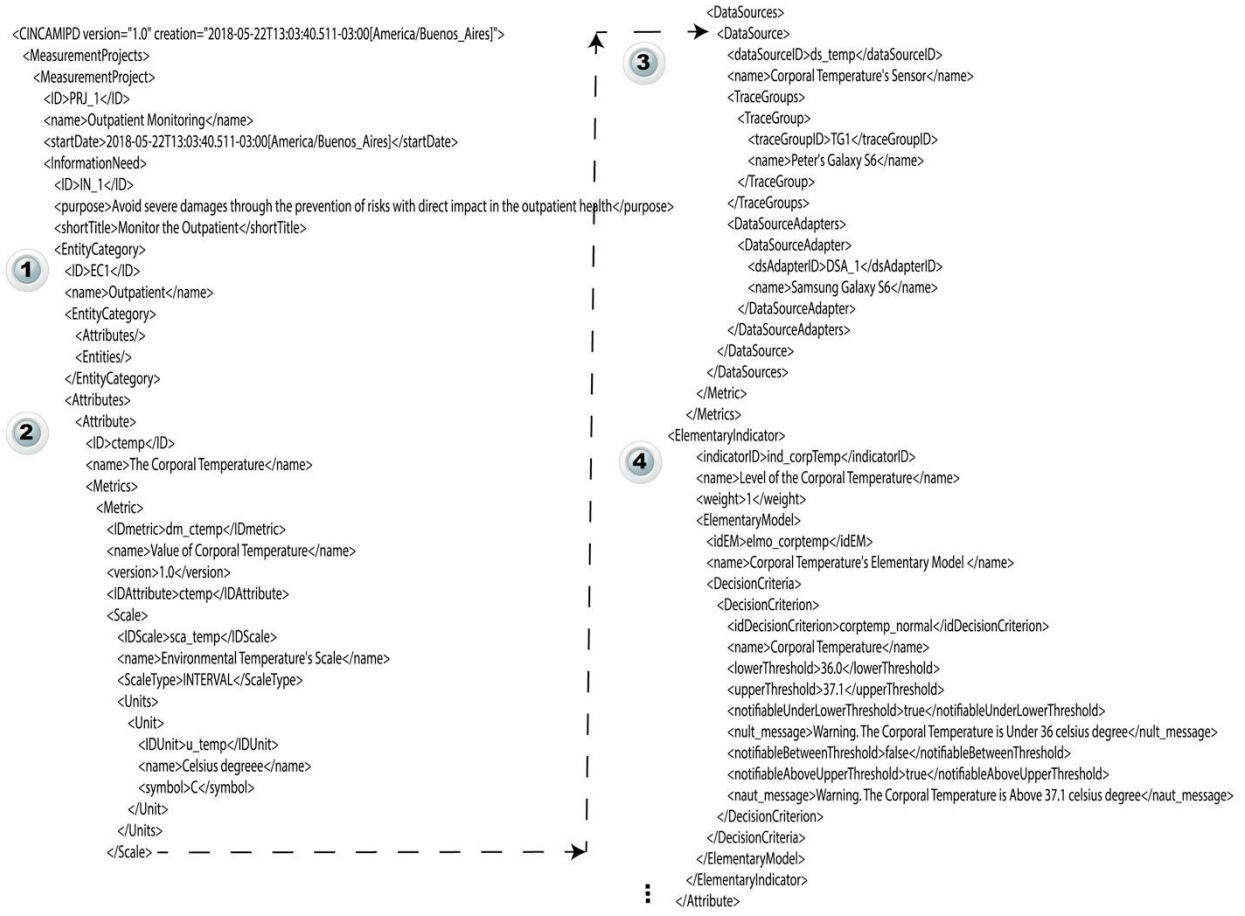

Fig. 1. Partial View of a CINCAMI/PD message related to the Outpatient.

Figure 1 synthesize a partial view of a project definition file related to the outpatient example. Each CINCAMIPD message could have many projects associated, and each one has a specific ID (e.g. PRJ_1 in Fig. 1) which is unique among all the measurement projects under monitoring in PAbMM. Inside each project, the entity under monitoring is defined (See circle with number one in Fig.1). As It was said before, each entity is described by its attributes, the example shows at the right of the circle with the number 2 (See Fig.1) the definition for the attribute "*The corporal temperature*", which is accompanied with the specification for the unit, scale, etc.

The circle with the number 3 in Fig. 1 indicates the identification for the data source responsible for providing the values for the indicated metric "*Value of the corporal Temperature*", jointly with the measurement adapter (MA). It

is responsible for the translation between raw data and the measurement interchange schema. Nowell, for the interpretation of each value, a basic definition for an indicator is described at the right of the circle with the number four (See fig. 1). As it is possible to see, it defines a simple decision criterion, the associated messages, jointly with its relationship with the metric (i.e. the value of the corporal temperature). Thus, the project definition schema allows interchanging the project definitions based on the C-INCAMI framework among all the automatized components related to the processing architecture fostering a common understanding related to the entity under analysis and its context.

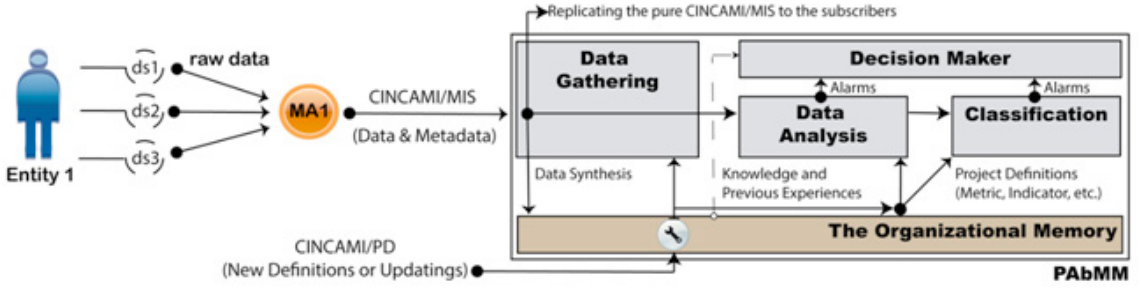

Fig. 2. The Relationship between the CINCAMI/Project Definition and the Processing Architecture

The importance of the project definition in PAbMM is such, that before receiving and processing any measure, a project must be loaded at the PAbMM by mean of its corresponding CINCAMI/PD schema. This loading just happens at the beginning or when the project is updated (See figure 2). Once the project has been loaded, the data processing happens in real-time keeping the definition in a stable way (i.e. without changes while the data are processed and analyzed). That is to say, from each entity under analysis, the associated data sources (e.g. ds1,ds2, and ds3 in Figure 2) collect the measures and inform them to the measurement adapter under the native data format (i.e. raw data). Once the measurement adapter has collected the measures, it generates a measurement interchange document using the project definition (i.e. CINCAMI/MIS). In this sense, the MA knows that the ds1 informs a value and using the project definition, it knows the attribute of an entity associated with that data source. Thus, each CINCAMI/MIS message incorporates the data (i.e. the measures) jointly with the tags that allow identifying the associated attribute and entity (i.e. the metadata). Finally, each time that the PAbMM receives a CINCAMI/MIS document from the MA, it knows the association among attributes, metrics, indicators and the other concepts because the project definitions are loaded in memory and shared with all the actors related to PAbMM. This is very useful because just using the metadata (e.g. the scale and unit related to a given metric), it is possible to detect miscalibration from the data sources without requires an statistical analysis and avoiding the computational cost. The data collecting in PAbMM is the responsibility of the gathering component (Data Gathering in Figure 2), who: i) It informs the data for its real-time analysis; ii) It replicates the cincami/mis message without changes to the subscribers; iii) It applies the synthesizes algorithm for storing the data into the organizational memory. In addition, the Organizational Memory contains: i) The previous experience and knowledge which is served to the decision maker when the recommendations need to be attached to a given decision (See Fig. 2); ii) The project definitions used for the classifiers and the data analyzers; and iii) The training data set for training the classifiers. The previous experiences and the knowledge from the organizational memory are swapped between the available memory and the disk depending on the scoring calculated from the project definition in terms of the similarity for the entities under analysis. This is especially useful when an alarm is launched and the entity under monitoring has no specific experience for bringing recommendations. In such case, the similar projects are used for recommending, avoiding a situation without recommendation.

4. The New Structural Coefficient based on the Definitions of the Attributes

The Organizational Memory (OM) is a big data repository implemented on the Apache HBase technology in PAbMM. With the aim of keeping in memory the closest recommendations in terms of the active measurement

projects and the received data by the architecture, the structural coefficient determines in the architecture the most similar projects based on its definition. In addition, the behavioral coefficient is jointly used with the structural coefficient for calculating a similarity scoring among the M&E projects based on the data distribution behavior (using the Spearman Kendall or Pearson correlation) [7, 18]. Thus, By the use of the structural and behavioral coefficients, the most related projects are kept in memory for bringing recommendations when some real-time alarm is received by the decision maker (See Figure 2). The original structural coefficient is defined like is shown in equation 1:

$$str_{sim(e1,e2)} = \frac{|e_1 \cap e_2|}{|e_1 \cup e_2|} \quad (1)$$

Where:

- $|e_1|$: It represents the quantity of the attributes which characterize to the entity one
- $|e_1 \cup e_2|$: It represents the total quantity of different attributes between the entity one and two
- $|e_1 \cap e_2|$: It represents the total quantity of common attributes between the entity one and two

The main concern of this structural coefficient is associated with the way in which the attribute is considered common or not between the entities because the comparison is based on the attribute ID.

Table 1. A typical mistake of the data quality related to the M&E project definitions

Attribute ID	Attribute Name	Definition	Outpatient 1	Outpatient 2
ctemp	The Corporal Temperature	Value of the axillary temperature in Celsius degree	Used	
heartrate	The Heart Rate	Number of beats per minute (bpm)	Used	
pc_humi	The Environmental Humidity	Amount of the water vapor in the air	Used	
pc_temp	The Environmental Temperature	Value of the environmental temperature in Celsius degree	Used	
pc_press	The Environmental Pressure	Pressures resulting from human activities which bring about changes in the state of the environment	Used	
c_temp	The Corporal Temperature	Value related to the axillary temperature in Celsius degree		Used
heart_rate	The Heart Rate	Quantity of beats per minute (bpm)		Used
pc_hum	The Environmental Humidity	Volume of the water vapor in the air		Used
pc_tem	The Environmental Temperature	Quantity related to the environmental temperature in Celsius degree		Used
pc_pressure	The Environmental Pressure	Pressures derived from human activities which bring about changes in the state of the environment		Used

Table 1 is an example derived from the project definition shown in Figure 1. Using the current structural coefficient, there is no coincidence among the attribute's IDs when the comparing between outpatient 1 and 2 is made (i.e. the current coefficient structural would be 0 because of 0 is the common attributes and 10 the total number of different attributes -0/10-). However, it is possible to appreciate the definition for outpatient 1 and 2 are semantically analogous but syntactically different. Because the M&E project definition is made by the experts, and even when they can use the organizational memory for considering the previous experience and knowledge, they are human beings. This implies that is possible to define the same attribute in a similar way or using different IDs with the same definition. Both situations are solved using the new text similarity-driven structural coefficient, because the

current coefficient use the textual definition from the attribute and the comparison is made through the text similarity coefficients, establishing a threshold for equivalence. Equation 2 defines the individual text similarity coefficient as follows:

$$\forall i = 1..|e_1 \cup e_2| \wedge j = 1..|e_1 \cup e_2| \mapsto sim_{text}(a_i, a_j) \subseteq [0,1] \tag{2}$$

Where:

$|e_1 \cup e_2|$: It represents the total quantity of different attributes between the entity one and two

a_i, a_j: It represents the text definition for the attributes i and j respectively

$sim(a_i, a_j)$: It represents a text similarity coefficient computed through Jaccard, Cosine, Sorensen-Dice, Jaro-Winkler or Levenshtein coefficients

$$\forall i = 1..|e_1 \cup e_2| \wedge j = 1..|e_1 \cup e_2| : selective(a_i, a_j) \begin{cases} 1 & sim_{text}(a_i, a_j) \geq threshold \\ 0 & sim_{text}(a_i, a_j) < threshold \end{cases} \tag{3}$$

A threshold is a number between 0 and 1, which is arbitrarily defined by the processing architecture manager and it is common for all the M&E projects. Thus, the triangular text similarity coefficient matrix is integrated as follows:

$$structural\_matrix(e_1, e_2) = \begin{pmatrix} selective(a_1, a_1) = 1 & selective(a_1, a_2) & \cdots & selective(a_1, a_{|e_1 \cup e_2|}) \\ selective(a_2, a_1) & \ddots & & selective(a_2, a_{|e_1 \cup e_2|}) \\ \vdots & & & \vdots \\ selective(a_{|e_1 \cup e_2|}, a_1) & selective(a_{|e_1 \cup e_2|}, a_2) & \cdots & selective(a_{|e_1 \cup e_2|}, a_{|e_1 \cup e_2|}) = 1 \end{pmatrix} \tag{4}$$

Equation 4 shows the similarity matrix for the entities e_1 and e_2, which dimensionality is $|e_1 \cup e_2| x |e_1 \cup e_2|$. Because the $selective(a_i, a_j) = selective(a_j, a_i)$ the matrix is triangular. Finally, given two entities e_1 and e_2, the new text-similarity-driven structural coefficient is computed as:

$$str_{sim(e_1, e_2)} = \frac{\sum_{i=1}^{|e_1|} \max(\sum_{j=1}^{|e_2|} selective(a_i, a_j))}{|e_1 \cup e_2| - \sum_{i=1}^{|e_1|} \max(\sum_{j=1}^{|e_2|} selective(a_i, a_j))} \tag{5}$$

As it is possible to see in equation 5, the $selective(a_i, a_j)$ are pre-computed in the structural matrix. As proof of concept, the attributes for the outpatient 1 and 2 from the Table 1 are used for computing the new structural coefficient using the cosine distance as follows:

Table 2. An Example of the Structural Matrix for the New Text-Similarity-Driven Structural coefficient based on cosine distance, limited to the comparison between the outpatient 1 and outpatient 2 projects

Attribute ID	c_temp	heart_rate	pc_hum	pc_tem	pc_pressusre
ctemp	**0.8457**	0.0773	0.2998	0.5608	0.2095
heartrate	0.0238	**0.7812**	0.1118	0.0219	0.0507
pc_humi	0.2132	0.0838	**0.8499**	0.1765	0.2877
pc_temp	**0.6422**	0.0721	0.2581	**0.7764**	0.3388
pc_press	0.1359	0.0486	0.2898	0.2501	**0.9131**

Given a threshold of 0.6, Each attribute of the outpatient 1(rows) are contrasted with each attribute of the outpatient 2 (columns) in terms of the value located in its intersection. In this case, the cosine distance was used as the sim_{text}. Thus, when at least there exists one value upper or equal to the threshold for an attribute, the max value (i.e. one by definition in equation 3) is returned by the attribute. For example, for the attribute ctemp (outpatient 1) a 1 is obtained because the cosine distance between ctemp and c_temp (outpatient 2) is 0.8457 and it is upper than 0.6 (the threshold). However, for the pc_temp attribute (outpatient 1), two attributes from outpatient 2 give 0.6422 and 0.7764 for the attributes c_temp and pc_tem respectively. In the last case and independently that two value are upper than 0.6, the value 1 is returned for the pc_press attribute. Finally, the structural coefficient comparing the outpatient 1 and 2 is:

$$str_{sim(e_1,e_2)} = \frac{\max(1)+\max(1)+\max(1)+\max(2)+\max(1)}{5+5-(\max(1)+\max(1)+\max(1)+\max(2)+\max(1))} = \frac{5}{5+5-5} = 1 \quad (6)$$

Comparing the result of the application of the original equation 1 in contrast with the result of equation 6, it is possible to observe the impact in the scoring related to the organizational memory. Why? Because the behavioral coefficient just is computed for the similar attributes and because previously the structural coefficient was cero, the outpatient 1 and 2 were considered different. Now, thanks to the text similarity coefficient computed from the attribute definition in the project definition, the behavioral coefficient will be computed over a wide range of the similar entities based on its definition, improving the scoring for the recommendations (See figure 2).

5. Implementation of the Text Similarity-Driven Structural Coefficient

The library pabmm_sh is an open source library, written in the Java language and released under the terms of the Apache 2.0 license. The source code can be download from github.com/mjdivan/pabmm_sh. Basically, the library implements the functionality for extending the loading and updating of the project definitions in the Processing Architecture (See Figure 2) based on the new structural coefficient.

The project definitions are received under the CINCAMI/PD schema and then they are incorporated in a multi-threading queue. Each project in the queue is scanned by an attribute detector, searching for the used attributes along the project. The detected attributes are mapped to a Hashtable keeping its metadata and avoiding the redundancy. Once the projects were scanned, a triangular matrix implementation is created and the attributes from the projects are incorporated. Finally, the structural coefficient is computed for the all active entities and the structural matrix is made available to the decision maker, to the effects of the selective filtering related to the recommendations.

A simulation for determining the overhead related to the pre-computing of the new structural coefficient at the moment in which the projects were loaded or updated was performed. The simulation simulated from 100 active projects to 2000 active projects simultaneously updated in the processing architecture, running on a virtual machine with Windows 7 Professional with 8GB of RAM and a processor Intel Core i5. The host was a MacBook Pro with an Intel Core i7 2,9Ghz, 16GB RAM and macOS Mojave.

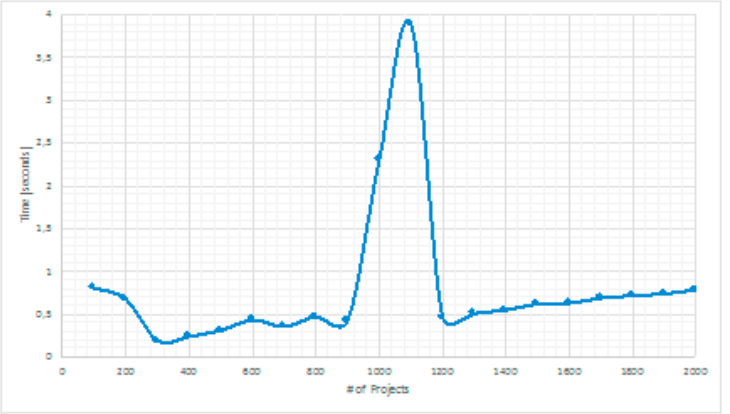

Fig. 3. The Time Evolution for the pabmm_sh library at the time in which the number of the active projects is incremented

As it is possible to appreciate in Figure 3, the required processing time for updating or loading a new definition is minimum (i.e. lesser than 1 second). Even, it should consider that the updating does not make interference in relation to the real-time data processing. That is to say before the data be processed by the PAbMM, the project definition must be loaded, but once loaded, the definition keeps stable along the data processing. In this sense, the referred time is mainly associated with the start-up of the processing architecture, and eventually, with some updating of the project definitions. The observed beak between 1000 and 1200 active projects in figure 3, it is a consequence of the garbage collector from the Java Virtual Machine.

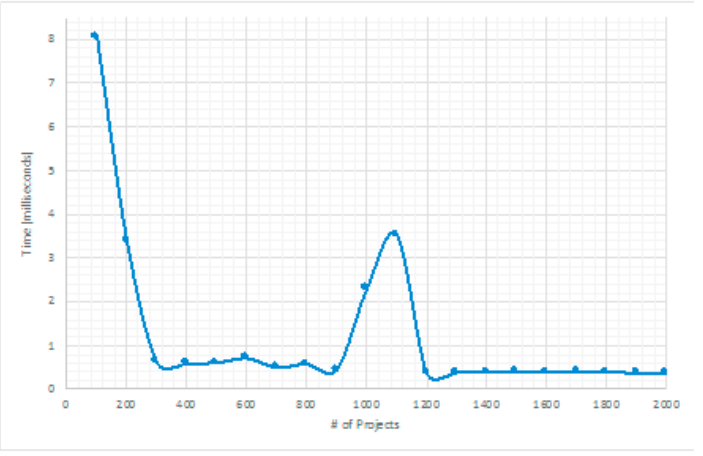

Fig. 4. The Project Processing Rate Evolution for the pabmm_sh

Figure 4 shows the project processing rate evolution derived from figure 3. In this sense, the obtained times indicates that the individual processing time of each CINCAMI/PD message is lesser than 1 millisecond, which does not represent an important cost for the architecture, mainly considering that this kind of operations (i.e. the updating or initial loading) happens at the start-up.

6. Conclusions

The CINCAMI/PD schema is oriented to the interchanging of the M&E project definitions based on the C-INCAMI framework and fostering the interoperability among different systems. On the one hand, the measurement and evaluation framework has the aim of reaching a common understanding of the concepts, terms and their

relationships at the moment in which a measurement project must be implemented. On the other hand, the CINCAMI/PD allows the free interchanging of the M&E project definitions among all the actors of the PAbMM for strengthening the metadata-driven data processing (i.e. the metadata are related to the project definition itself).

In this work, we have introduced a new structural coefficient based on the text similarity for improving the real-time recommending searching associated with the decision maker in PAbMM. That is to say, for avoiding redundancy and human mistakes at the moment in which a project is defined by the experts, the structural coefficient compares the attribute similarity based on its textual definition. It is possible that the same attribute has different IDs or even similar definitions, for that reason the text coefficients such as cosine, jaccard, among others were incorporated in the pabmm_sh library. Thus, the PAbMM manager could currently define a threshold which act as reference at the moment of determining if some attribute is equivalent or not each other.

A proof of concept was described for contrasting the old structural coefficient versus the new coefficient, and the new structural coefficient presents better tolerance in relation to human mistakes at the project definition. This implies that the scoring used for the recommendations between projects and the swapping associated with the organizational memory is improved because the behavioral coefficient is computed from a wider range of attributes than before.

A simulation on the pabmm_sh library was introduced for analyzing the potential impact of the previous computing of the structural coefficient along the processing architecture. In this sense, the related times associated with the updating or initial loading of the project definitions was less than 1 second for 2000 projects. Even, the individual project time processing was under 1 millisecond.

As future works, alternatives associated with Word2Vec will be analyzed for improving the structural coefficient in terms of semantic similarity.

Acknowledgement

Acknowledgement and Reference heading should be left justified, bold, with the first letter capitalized but have no numbers. Text below continues as normal.

References

1. Agbozo E and Spassov K, "Establishing Efficient Governance Through Data-Driven e-Government," in 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, 2018.
2. Dong G, Chen J, Wang H and Zhong N, “A Narrow-domain Entity Recognition Method Based on Domain Relevance Measurement and Context Information” in International Conference on Web Intelligence, Leipzig, Germany, 2017.
3. Diván M and Sánchez Reynoso M, "Real-Time Measurement and Evaluation as System Reliability Driver," in System Reliability Management: Solutions and Technologies, A. Anand and M. Ram, Eds., Florida, USA, CRC Press, 2018, pp. 161-188.
4. Olsina L, Papa F and Molina H, "How to Measure and Evaluate Web Applications in a Consistent Way," in Web Engineering: Modelling and Implementing Web Applications, G. Rossi, O. Pastor, D. Schwabe and L. Olsina, Eds., London, Springer-Verlag, 2008, pp. 385-420.
5. Samosir J, Indrawan-Santiago I and Haghighi P, “An Evaluation of Data Stream Processing Systems for Data Driven Applications” Procedia Computer Science, vol. 80, pp. 439-449, 2016.
6. Morales G and Bifet A, "SAMOA: Scalable Advanced Massive Online Analysis," Journal of Machine Learning Research, vol. 16, pp. 149-153, 2015.
7. Diván M and Sánchez Reynoso M, "Behavioural Similarity Analysis for Supporting the Recommendation in PAbMM" in 1st International Conference on Infocom Technologies and Unmanned Systems (ICTUS), Dubai, 2017.
8. Diván M and Sánchez Reynoso M, “Fostering the Interoperability of the Measurement and Evaluation Project Definitions in PAbMM” in 7nd International Conference on Realiability, Infocom Technologies and Optimization (ICRITO'2018), Noida, 2018.
9. Zhou F, Wu B, Yang Y, Trajcevski G, Zhang K and Zhong T, "Vec2Link: Unifying Heterogeneous Data for Social Link Prediction" in 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 2018.
10. Aumüller M, Christiani T, Pagh R and Silvestri F, “Distance-Sensitive Hashing” in 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, US, 2018.
11. Fender A, Emad N, Petiton S, Eaton J and Naumov M, “Parallel Jaccard and Related Graph Clustering Techniques” in 8th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, Denver, Colorado, 2017.
12. Hrytsenko Y, Daniels N and Schwartz R, “Efficient Distance Calculations Between Genomes Using Mathematical Approximation” in ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 2018.
13. Chaidaroon S, Ebesu T and Fang Y, “Deep Semantic Text Hashing with Weak Supervision” in 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 2018.

14. Chaidaroon S and Fang Y, “Variational Deep Semantic Hashing for Text Documents” in 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, 2017.
15. Le Q and Mikolov T, “Distributed Representations of Sentences and Documents” in 31st International Conference on Machine Learning, Beijing, China, 2014.
16. Pengfei, L, King Keung, W and Helen, M, “A model of extended paragraph vector for document categorization and trend analysis” de 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017.
17. Molina H and Olsina L, "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in Quality of Information and Communications Technology (QUATIC), Lisbon, 2007.
18. James J, Witten D, Hastie T and Tibshirani R, “An Introduction to Statistical Learning with Applications in R”, 8th ed., New York: Springer Science+Business Media, 2017.