International Journal of
Information Systems and Project Management
ISSN (print):2182-7796, ISSN (online):2182-7788, ISSN (cd-rom):2182-780X
Available online at www.sciencesphere.org/ijispm

Thematic grouping for messages in major events

Wallace Pinheiro
Centro de Desenvolvimento de Sistemas
QG do Exército, Bloco G, 2.º Andar, SMU, Brasília, DF, CEP 70630-901
Brazil
www.shortbio.net/wallace@cds.eb.mil.br

Ricardo Fernandes
Centro de Desenvolvimento de Sistemas
QG do Exército, Bloco G, 2.º Andar, SMU, Brasília, DF, CEP 70630-901
Brazil
www.shortbio.net/ricardo@cds.eb.mil.br

Luciene Souza
Centro de Análises de Sistemas Navais
Pr. Barão do Ladário, s/n - Ilha das Cobras, ed. 8 do AMRJ, 3.º andar, RJ, CEP 20091-000
Brazil
www.shortbio.net/luciene.carvalho@casnav.mar.mil.br

Abstract:
The process of information evaluation may compete with the decision making process by requiring the limited cognitive resources. In the case of Major Events, such as the Football World Cup or the Olympic Games, the thematic grouping of every information may be overwhelming. The theme switching caused by information associated with a different thematic group may be modelled as an interruption in multitasking set-ups. Thus, automating the thematic grouping of information may facilitate the decision making process by reducing the theme switching for a decision maker when he reads a set of messages. In this paper were used clustering techniques with multi-criteria to group the messages in themes. These criteria were implemented as configurable operators. To achieve a better comprehension of those parameters, we introduced the concepts of Thematic Strength (TS) and Thematic Density (TD). The evaluation of the strategy was made over a set of operational messages available in the Pacificador system during the 2014 FIFA World Cup.

Keywords:
decision-making; thematic grouping; multi-criteria clustering; major events; messages.

DOI: 10.12821/ijispm040403

Manuscript received: 20 April 2016
Manuscript accepted: 8 September 2016

1. Introduction

The human decision making process is a cognitive process of choice among alternatives [1],[2]. It means the decision making process is bounded by the human cognition constraints. Considering this and the higher demands for information processing and agile decision, the human brain becomes a bottleneck for the decision flow. We can conclude that the human cognition is the main asset to be preserved for critical decision operations.

For that reason, it is mandatory to avoid unnecessary cognition overload in order to preserve agility. The process of information evaluation, for example, may compete with the decision making process by requiring the limited cognitive resources. In the case of Major Events, such as the Football World Cup or the Olympic Games, the thematic grouping of all information may be overwhelming. Under this scenario, the communication among the operational people is made through message-enabled devices, such as cell phone and tablets. But the messages presentation order is not based on thematic association. It means that the decision maker may use much of the cognitive resources only to organize the information on the proper theme.

The theme switching caused by information associated with a different thematic group may be modelled as an interruption in multitasking set-ups [3].

We propose a solution for the cognitive overload caused by high information processing demands in Major Events based on the automation of the thematic grouping process. The thematic grouping process can be understood as the process of clustering information related to a same theme in a same group using one or more features of the information. In this paper, the information is materialized as one or more message and a set of features extracted from the messages.

Automating the thematic grouping of information may facilitate the decision making process by reducing the theme switching for the decision maker when he reads a set of messages. To group the messages in themes, we used some clustering techniques with multi-criteria. These criteria were implemented as configurable operators. To achieve a better comprehension of those parameters, we introduced the concepts of Thematic Strength (TS) and Thematic Density (TD). The evaluation of the strategy was made over a set of operational messages available in the Pacificador system during the 2014 FIFA World Cup. The information processing from different sources must deal with information duplication and correlation. The information correlation is understood as the relation among information that refers to the same object, event or theme. In our case, we propose a way to group information according to their themes.

From Peng et al. [4], it is possible to remark the importance of the grouping process for supporting the analysis of large amount of data. It points out that the grouping contributes by reducing the number of alternatives as it offers some representative elements to be used in the process of alternative choices.

The application of multi-criteria may contribute to the effectiveness of message recovery as reported by Schuff et al. [5]. In this work, the themes are approximated by groups built by a computer program, using multi-criteria.

These groups need to contain at least two messages to be considered a theme. Those groups aim at facilitating the message association recovery as a good way to reduce the complexity of message analysis [4],[5]. Therefore, we denote those groups as thematic groups or themes because they should provide a set of messages related to a similar topic.

We remark that theme recovery, in the comprehension cognitive level, based on the thematic grouping, might demand human evaluation, however it is out the scope of this work. Moreover, it is important to

observe that, according to Roussinov and Chen [6], the computer grouping resembles the human grouping made by specialists.

We also remark that the number of messages in each theme depends on operational factors. We put those factors aside of our approach, allowing the number of messages in each theme be variable. The statistical analysis of each theme is also excluded from the present article.

The thematic grouping proposed in this work evolves some concepts from Souza and Pinheiro [7],[8] for multi-criteria clustering of data from different sources. Specifically, we propose new ideas for the architecture, the multi-criteria clustering, the sequential filtering based on attributes and the configuration techniques for the operators. Furthermore, we apply the proposed methodology as an experiment over operational messages from the 2014 FIFA World Cup, obtained from the Pacificador system.

The remainder of this work is organized as follows: Section 2 describes the related works; Section 3 presents the proposed methodology; Section 4 shows the experiments results; and Section 5 presents the conclusion and future works.

2. Background and Related Works

2.1 Techniques for Analyzing and Grouping Information

One way of connecting pieces of information with correlation is using a strategy known as record linkage [9]. The demands presented by the growing amount of information impose scalability and performance requirements over the record linkage process. And that imposition leads to the development of many indexing techniques to reduce the number of comparison among the set of information. It is done by the identification of sets of information with explicit no correlation. It effectively reduces the number of comparisons, maintaining the high quality of record linkage and reducing the complexity of the algorithms [10].

The Traditional Blocking indexing technique is used since 1960 [9]. It is done assigning each piece of information to a given block according to some criteria. These criteria are given to increase the potential relation among the information inside the blocks. The comparison reduction is achieved by restricting the search among the information in the same block. The sensitive part of this strategy is the definition of the criteria, which may separate the information correlated in different blocks. To solve these issues, many different approaches were proposed [9],[10]: Sorted Neighborhood Indexing; Q-Gram Based Indexing; Suffix Array-based Indexing; String-Map-based Indexing; Canopy Clustering; and Hierarchical Clustering. This last one received a special attention in this work and it is used to cluster elements composed by texts. On the Hierarchical Clustering, the algorithms may change the number of groups after each iteration. The proximity or similarity measure of that kind of grouping method is known a linkage metric. Dendograms may be built by agglomeration (bottom-up) or division (top-down), being the latter more computationally complex. Therefore, this work implements the agglomerative clustering strategy.

Others frequent techniques used to group information involve partitioning clustering algorithms, such as K-means and K-medoids [11]. However K-means and K-medoids demand the user provides, a priory, the number of clusters, which makes this strategy unsuitable for the scenario discussed in this paper. The hierarchical clustering is relevant for this work because it uses tree of groups, denoted as dendograms, that allows choosing the number of final groups according the chosen similarity level for the groups. Similarity out of these limits means explicit no possible correlation among groups, being quite intuitive to users. The definition of the limits may depend on expert knowledge and can be done by parameters and attributes in a multi-criteria way, as proposed in this work.

Multi-criteria decision methods are also important in the context of this work, because they allow comparing different attributes (or critters) based on people judgment. As this work uses different attributes to group messages from different sources (people), these methods enable the valuation of these attributes using pair-wise comparisons between the attributes. In this scenario, Analytic Hierarchy Process (AHP) [12],[13] is one of the most used multi-criteria decision, because it is simple and effective in many scenarios. In this method, people indicate the relative significance between pair of attributes. The attributes correspond to features extracted from alternatives to be ordered. For example, alternatives can be represented by cars and attributes can be represented by price, category, boot space, etc. The responsible by the decision can specify their preferences using comparisons (to pair of attributes and alternatives) using values that can range from 1/9 to 9 (1 represents attributes with the same value). These values are used to populate a matrix that is used to calculate the weight of attributes and alternatives.

2.2 Pacificador System

The Pacificador system, developed by the Centro de Desenvolvimento de Sistemas (CDS), is the Brazilian Army Command and Control system employed by the Brazilian Defence Ministry. It was used in the 2012 Rio+20 United Nations Conference, the 2013 Confederations Football Cup, the 28th World Youth Day (2013) and the 2014 FIFA World Cup.

The operations of support and security of Major Events involve people that must work in harmony. The coordination of the operational staff requires communications to serve different possible media. The text messages are of great value because they are relatively cheap and because they tend to be efficient on information codification, synthesizing the ideas and lowering the communication infrastructure.

The operational messages carry some relevant attributes, for instance: description in natural language; time stamp; and position of the author of the message. These attributes were used for the information grouping. In order to achieve this goal, this paper proposes an architecture for correlation data from distinct sources, in a Command and Control environment (C2), whose data follow the "Joint Consultation, Command and Control Information Exchange Data Model (JC3IEDM)" specification that aims for the international interoperability between C2 Systems. Therefore, messages provided by the Pacificador system were exported to the JC3IEDM format (http://mipsite.lsec.dnd.ca/).

3. Thematic grouping

3.1 Overview

In this article, the grouping of information is divided in two major phases: information normalization and information grouping. The decision diagram of the proposed strategy is given in Fig. 1 using BPM (http://www.bpmn.org/) notation through Bizage Modeler (http://www.bizagi.com/).

The first phase is responsible for preparing information incoming from different sources. It deals with the aspects of natural languages that may interfere on the grouping algorithms. It also understands different information formats in order to extract the attributes to be used in the next phase. Among the tasks of this process, it can be cited [14],[15]: tokenization; filtering normalization; discarding of stopwords; lemmatization; and stemming. These tasks are important to correlate data from different sources, but the user can choose which of them should be executed. For this reason, in Fig. 1, each one of the tasks can or cannot be executed, according to the analyses of the input data.

The second phase is responsible for identifying the information with higher similarity under multi-criteria building groups of similar messages. This grouping is meant to recover the theme of every information piece, for that reason we denoted it as thematic grouping. Under the Major Events scenario, the information relation should not only consider the descriptions, which are related to the semantic dimension. This

approach can be seen in the information relation processes as in Christen [9]. We propose that other attributes must be considered. We base our approach in the 5W2H approach where: WHAT corresponds to the description of the message; WHERE corresponds to the location from the message is sent; WHEN corresponds to the data-hour of the message creation; WHO corresponds to the author of the message; WHY corresponds the reason of the message sending; HOW corresponds to the context of the message (type of object or event being described and the environment where is located the object or event being described); and HOW MUCH corresponds to the relevance of the message. These attributes offer a better explanation of the message to leverage the quality of the thematic grouping.

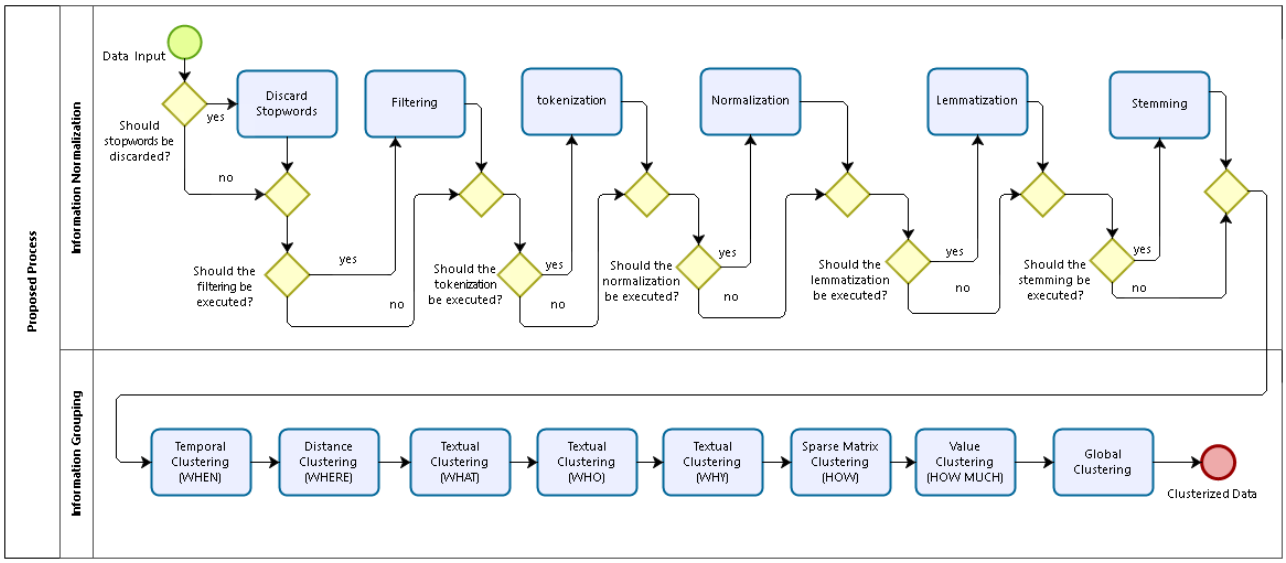

Fig. 1. Proposed Strategy

The attributes classified according to the 5W2H action plan are related to the JC3IEDM entities that provide the pattern for the messages obtained from the Pacificador system. In this process, this paper follows the idea detailed in Souza and Pinheiro [8]: the OBJECT-ITEM entity is associated to the WHAT attribute, indicating the textual description of the object or event. The attribute WHERE is related to the LOCATION entity and their connected classes: GEOGRAPHIC-POINT and OBJECT-ITEM-LOCATION, which provide the object position (latitude and longitude) and velocity. The REPORTING-DATA-ABSOLUTE-TIMING entity provides the date/hour of a report that is resulted from object or event visualization. It is associated to the attribute WHEN. The attribute HOW is associated to the OBJECT-TYPE entity and provides a combination of information related to the type of the objects (for instance: aircraft, ship, etc.) and the environment where objects were found (for instance: air, water, etc.). The attribute WHO provides information about the source of the message and it is related to the entity REPORTING-DATA. At this point, this paper extends this idea using the entity REPORTING-DATA from JC3IEDM to provide information about the source of the message (WHO sends the message) as well information about the message, such as: accuracy (accuracy-code), credibility (credibility-code) and reliability (reliability-code). Thus, these parameters can be used to infer the value of the message (HOW MUCH) through a correspondence between the string values used in the JC3IEDM for each parameter and numeric that can be set to them. For instance, the parameter accuracy could receive the following values: Confirmed = 100; Probable = 80; and Doubtful = 40. Finally, the attribute WHY is associated to the entity ACTION that can describe the goal of the message.

In this work, an operator that combines pair of messages implements each attribute. Similarity functions present in each operators link messages and this link is established when the similarity between two messages is higher than a minimum pre-established level of similarity. Groups of messages are composed of messages that have some relation of similarity with at least one member of a group. Members of different groups do not have any relation with any member of other group. In other words, the groups do not have overlap.

Operators are used in sequence and only pairs of messages that are related by the previous operator need to be processed by the next operator, allowing that each operator acts as a filter for the next. Thus the first operator analyzes all possible pair combinations of messages. The following operators just need to analyze the pair related by the previous operator.

The operators consider five types of similarities functions to link messages: based on temporal difference (attribute WHEN); based on distance (attribute WHERE); based on textual similarity (attributes WHAT, WHO, WHY); based on value difference (attribute HOW MUCH); and based on a sparse matrix similarity (attribute HOW). This last attribute, differently from the others, has to deal with two parameters: type of object and environment. The textual similarity (attributes WHAT, WHO, WHY) is implemented by the Hierarchical clustering approach, discussed in Section 2.

The parameters follow a pattern provided by the JC3IEDM specification. For instance, the types of the objects can be described by: aircraft, ship, car, troop, among others; and the environment of the object can be described by: air, water, road, trail, etc. When the user sends a message, he can choose the type and environment related to the object or event being described in the message. Thus, we proposed for this last attribute the use of a sparse matrix to obtain the similarity level between two messages. An example of a sparse matrix, based on Souza and Pinheiro [8], to calculate similarities related to the HOW attribute is showed in Table 1.

Table 1. Sparse Matrix

	Type: Aircraft Environment: Air	Type: Aircraft Environment: Land	Type: Vehicle Environment: Road	Type: Vehicle Environment: Field	Type: Troop Environment: Road	Type: Troop Environment: Field
Type: Aircraft Environment: Air	1	0.4				
Type: Aircraft Environment: Land	0.4	1				
Type: Vehicle Environment: Road			1	0.6		
Type: Vehicle Environment: Field			0.6	1		
Type: Troop Environment: Road					1	0.7
Type: Troop Environment: Field					0.7	1

As the final stage, results of the operations are combined to produce a global similarity between each pair of messages. The decision maker can set a minimum global similarity value that relates two messages. If the similarity between a pair of message is higher than the global similarity, then the messages of this pair belongs to a same group. To calculate the global similarity, it is used the formula showed in (1), where: *w(ATTRIBUTE)* corresponds to the weight of an attribute of the 5W2H approach and *w*o(ATTRIBUTE)* corresponds to the weight multiplied by the similarity between two messages using an attribute.

$$GlogalSimilarity = (w*o(what)+w*o(where)+w*o(when)+w*o(who)+w*o(why)+w*o(how)+w*o(howmuch))/ (w(what)+w(where)+w(when)+w(who)+w(why)+w(how)+w(howmuch)) \quad (1)$$

To estimate the weights of the other operators, it can be used a multi-criteria decision making method, such as Analytic Hierarchy Process (AHP) [12],[13]. In this case, AHP should be used just to calculate the weight of each attribute (critter) using pair-wise comparisons between the attributes. Thus, we have a way to calculate the importance of each attribute considering the opinions of several users. The approach proposed in this paper considers that new message can be processed on demand. It allows the creation of new groups or the inclusion of new messages in the existent groups, without having to process all combinations of messages again.

The analysis of complexity is important because it allows knowing the cost of each operator. Thus, the less expensive operators can be applied first, considering that the operators have the commutative property. The correct ordering of the operators may reduce the total time of the clustering process. The first operator generally processes a higher number of messages combinations than any other operator. It happens because it discards some combinations of messages according to its similarity levels. These combinations do not need to be evaluated by the next operator. Following this idea, the second operator processes a higher number of messages combinations than the third operator and so on.

The complexity operator analysis studies the time and the space used by the algorithms. The asymptotic complexity analysis considers the relation between time and space used by algorithm depending on a large number of input information [16]. That analysis is directly linked to the operator cost.

Other perspective that could be considered to apply operators sequentially is the selectivity analysis. However, this analysis is not considered in this work.

3.2 Operators configuration

The use of acceptable levels of similarity brings advantages, because they can be interpreted as maximum errors admitted to correlate each attribute of a message. However, it is not easy to define suitable error values. In certain cases, it is easier to define the characteristics of the message groups that should be created. Because of that, it is proposed two new concepts that are related to the features of the clusters:

Thematic Strength (TS) and Thematic Density (TD). These concepts are represented by the expressions (2) and (3), where: the variable numClusters corresponds to the number of clusters with two or more messages and the variable numMsgs indicates the total number of messages grouped in the clusters with two or more messages, being considered themes the clusters composed of two or more messages. Clusters composed of just one message are not considered in these expressions.

$$TS(numClusters, numMsgs) = numClusters * numMsgs \quad (2)$$

$$TD(numClusters, numMsgs) = numMsgs/numClusters \quad (3)$$

In order to adjust the right values according to these concepts considering a set of messages, it is necessary to calibrate them through a small subset of the messages considered as a training set (it is proposed a training set composed of approximately 10% of the total messages number). Thus, each operator needs to be applied to this subset using different configuration values. The results will provide the suitable values for the operators' parameters. Thus, from this process, it is possible to define the acceptable levels of similarity.

It can be noticed that TS increases when the number of clusters and the number of messages in the clusters increases. The maximum theoretical value of this measure is reached when the clusters are composed of just two messages and the total of messages is equal to the number of processed messages. It follows that higher values of this measure may produce higher number of clusters with few messages in each one. This situation makes difficult to analyze the themes for a big number of messages. Otherwise, lower values of TS may produce lower number of cluster with few messages, which is not appropriate to evaluate the themes.

On the other hand, TD increases when the average size of the cluster increases. The maximum theoretical value of this message is reached when all processed messages belongs to the same cluster. It follows that

higher values of this measure are produced by a small number of cluster with many messages, which does not facilitate the processing to be executed by the decision maker. Otherwise, lower values of TD indicate cluster with few messages, which is also not interesting for the user. In a way, TD gives a counterpoint to TS.

Therefore, it is proposed to combine these two measures trying to find a balance between them. It is proposed to use the maximum value given by the weighted arithmetic mean of TS and TD. If correctly set, this measure may reduce the number of cluster composed of just one message and may increase the size and number of clusters accordingly. This measure, named maximum Thematic Clustering (maxTC), is given by the expression (4), where the terms alpha and beta represent weights and allow adjusting this measure to different scenarios. The terms maxTS and maxTD correspond to the maximum values of TS and TD.

$$maxTC = ((alpha * (TS/maxTS) + beta * (TD/maxTD))/(alpha + beta)) \quad (4)$$

It is possible, from the maxTC, to find the number of clusters and the total number of messages in the clusters and, consequently, the parameters of each operator.

4. Experiments

It is important to highlight that the number of themes depends on clusterization process based on attributes extracted from the messages. As seen previously, this process was based on messages temporal difference, distance of senders, similarity among objects or events descriptions, messages from the same sender (grouping), similarity of messages goals, similarity of the object types and environments where the objects were found, and similarity of the items: accuracy, credibility and reliability from the messages.

To evaluate the strategy proposed, it is used a subset of the attributes related to the 5W2H approach. They are: time (data-hour of sending); position (latitude-longitude of the data sender); and description (detailed of observed object or event). These attributes are considered sufficient for this analysis because when they are used in a combined way they provide suitable results in a scenario of operations in major events, as demonstrated in the experiments. In future works, all 5W2H attributes will be analyzed together what should improve still more the results.

The three chosen attributes are implemented by operators working with: temporal distance (based on time); spatial distance (based on distance); and textual similarity (based on description). The operators used the agglomerative hierarchical clustering that has at least quadratic order. However, each operator uses a different number of parameters, what implies they have different costs.

In the case of the text operator, the similarity functions, like Jaccard or TF-IDF functions [14], need to compare all terms (words) from the data items. Thus, the cost of this operator is related to the average number of words. According to this idea, the texts related to descriptions are more expensive than the texts that identify authors (normally shorter).

In the case of the distance operator, the function used to calculate the distance receives two parameters: latitude and longitude. To calculate the distance, it is necessary to compare these two values for each data pair. As it just compares these two values, in most part of the cases, this function is less expensive than textual similarity functions.

In the case of the operator based of temporal difference, the function used to calculate the difference receives one parameter: the message timestamp. To calculate the time difference, it is necessary to compare this value for each item data pair. As it just compares this value, function is less expensive than the others analyzed until now.

Consequently, the optimized ordering of the operators is: 1) Temporal Clustering; 2) Distance Clustering; and 3) Textual clustering.

Additionally, after applying the operators sequentially, it was applied the global similarity to obtain the final results. This operation is set to consider the same weight for the three criteria (temporal difference, distance, textual similarity) and the minimum global similarity equal to 1%. This value represents the minimum acceptable similarity in this experiment. In other scenarios, it is suggested to use AHP to set the weights and the minimum global similarity.

In this paper, it was executed experiments using data from real operations that happened in the 2014 FIFA World Cup. The messages representing the data were registered using the Pacificador system.

The methodology proposed in this paper was implemented through components developed in java. These components were encapsulated as plugins of the Apatar Tool, an open source application that aims to provide connectivity to several databases and applications [17].

Field agents wrote most messages used in the experiments, but all people involved in the military operations could write messages. The messages normally described potential and real incidents, orders and monitoring of planned activities.

The methodology of evaluation considered, firstly, each operator isolated using 500 messages and, after, all operators combined using 2000 messages. The first experiment is used to set the parameters of the operators and the second experiment evaluated the behavior of the operators using these parameters. Therefore, the first experiment considered several level of similarities and the second experiment considered one level of similarity to each operator.

Fig. 2 and Fig. 3 describe the results obtained by the operator based on textual similarity.

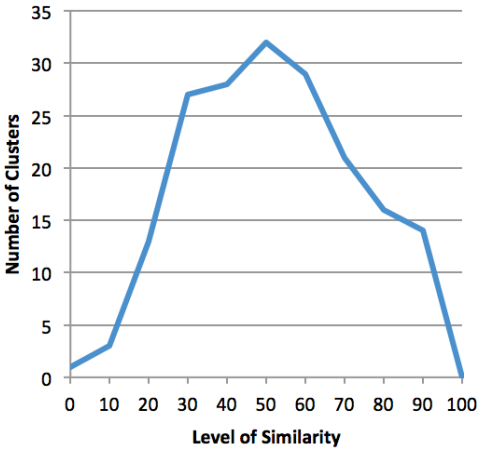

Fig. 2. No. of Cluster versus Level of Similarity

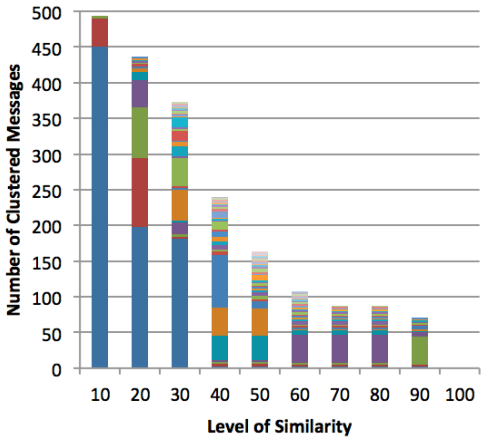

Fig. 3. No. of Clustered Messages versus Level of Similarity

Fig. 2 shows number of cluster (composed of two or more messages) versus level of textual similarity. It can be noticed that the number of cluster increases up to the similarity level reaches approximately 50%. After that, the number of cluster starts reducing. This happens because the clusters are fragmented in smaller pieces when the similarity level increases. When this value reaches approximately 50%, the clusters start splitting in clusters with one element that were not considered as themes.

Fig. 3 shows total number of clustered messages (composed of two or more messages) versus textual similarity level, completing the analysis of this operator. Thus, messages not grouped with other messages are not considered. It can be seen that the total number of messages in groups of two or more messages reduces when the similarity level increases. Besides, the number of clusters (represented by the colored bars) starts with a small number of clusters, reaches their maximum value approximately on 50% and, after, reduces this value when the similarity level increases.

Fig. 4 and Fig. 5 describe the results obtained by the operator based on temporal difference.

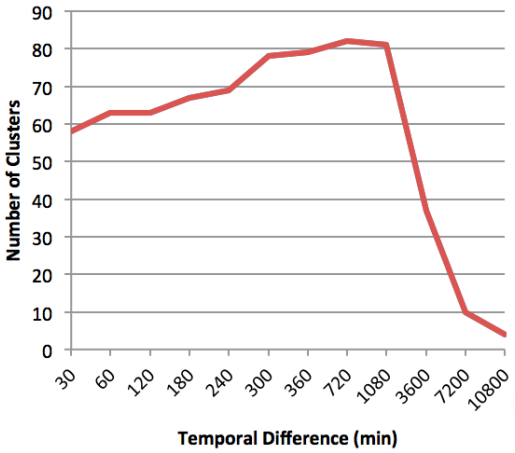

Fig. 4. No. of Cluster versus Temporal Difference

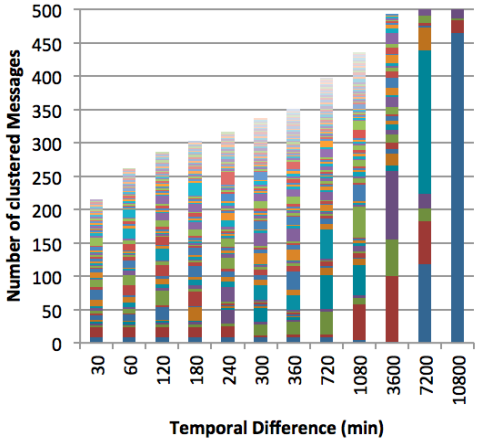

Fig. 5. No. of Clustered Messages versus Temporal Difference

Fig. 4 shows number of clusters (composed of two or more messages) versus temporal difference. It can be observed that the number of clusters increases up to the temporal difference reaches approximately 720 minutes. After that, the number of clusters starts reducing. This happens because the clusters are combined in bigger clusters when the temporal difference increases. Initially, clusters with two elements are produced more frequently. When the temporal difference reaches approximately 720 minutes, the clusters start combining more frequently in clusters with more than two elements, reducing the total number of clusters and, consequently, themes.

Fig. 5 shows total number of clustered messages (composed of two or more messages) versus temporal difference. It can be observed that the total number of messages in groups of two or more messages increases when the temporal difference increases. This behavior is the inverse of the operator based on textual similarity. The maximum number of clusters is reached when the temporal difference is about 720 minutes.

Fig. 6 and Fig. 7 describe the results obtained by the operator based on distance.

Fig. 6 shows number of cluster (composed of two or more messages) versus distance among the points of messages sending. It can be seen that the number of clusters increases up to the maximum near 50 meters. After, it reduces and increases twice: near 5 kilometers and 50 kilometers. After that, it reduces continuously. It can be inferred those three maximum points represent limits of cluster (and sub-clusters) of messages considering distance. 50 meters limit small groups, such as: people that provided security around stadiums, displacements. 5 kilometers limit the people that took care of major event places, such as: places of games, training, resting, etc. 50 kilometers correspond to the combination of groups in different cities

where the events happened. Thus, these different sizes of clusters represent the groups of entities existent in the operations.

Fig. 7 shows total number of clustered messages (composed of two or more messages) versus distance.

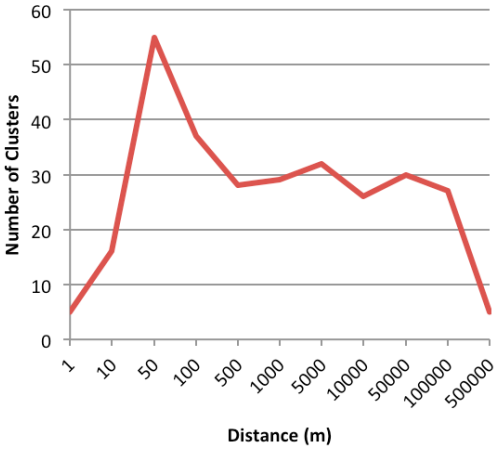

Fig. 6: No. of Cluster versus Distance

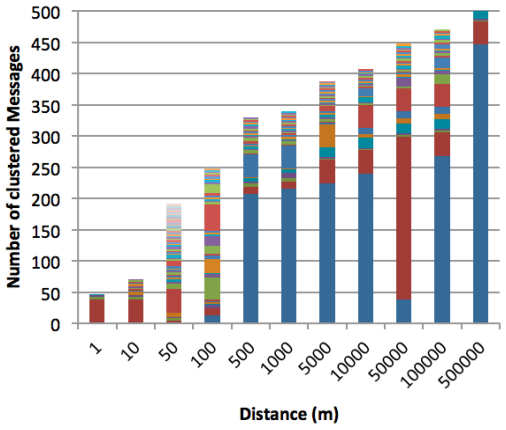

Fig. 7: No. of Clustered Messages versus Distance

It can be seen that the total number of messages in groups of two or more messages increases when the distance increases. As discussed previously the number of cluster (represented by the colored bars) varies according to the distance. The maximum number of cluster is reached when the distance is about 50 meters.

To set parameters of the operators used previously, it was used the concepts of Thematic Strength, Thematic Density and Maximum Thematic Clustering, given by the expressions (1), (2) and (3). We recommend that a calibration of the parameters values be done using a small subset of the messages to be grouped. In this paper the set of 500 messages initially used in the experiments were considered as a set of calibration for the operators. The parameters obtained were: 30% for textual similarity; 1080 minutes for temporal difference; and 50 kilometers for distance.

To set parameters of the operators used previously, it was used the concepts of Thematic Strength, Thematic Density and Maximum Thematic Clustering, given by the expressions (1), (2) and (3) applied to a small subset of the messages to be processed (set of calibration). In this paper, the set of 500 messages, used in the previous experiments, were considered as a set of calibration for the operators. The better values for the parameters, as demonstrated previously, were: 30% for textual similarity, 1080 minutes for temporal difference and 50 kilometers for distance. These values can be reused in different scenarios. However, to better calibrate the parameters values, it is recommended to use a small subset of the messages to be processed.

Fig. 8 shows the result of three operators combined using the parameters. Four sets of messages were used with: 500, 1000, 1500 and 2000 messages. The percentages of grouped messages were: 23% to 500 messages; 35% to 1000 messages; 42% to 1500 messages; and 47% to 2000 messages. Independently of the processed messages quantity, the number of clusters corresponded to approximately 10% of grouped messages. It is important to highlight that not all messages will form groups with other messages, remaining isolated. In our experiment, using 2000 messages, approximately 53% remained isolated. For these messages, the decision maker still needs to do a traditional evaluation, having to analyze these messages and correlate them manually. However, if it is possible to observe that increasing the number of messages, the percentage

of grouped messages (composed of two or more messages) is increased while the percentage of created clusters is kept, reducing the number of isolated messages. It happens because more and more messages under the same theme are inserted in the clusters that have already been created or are grouped with the isolated messages. This fact is very interesting, because it indicates that, increasing the number of processed messages, in a certain point, the creation of new clusters (themes) becomes less frequent and new messages are normally included in the existent clusters that will provide the themes for the decision maker. It is important to point out that the themes are built according to the parameters of the operators that can be defined by the decision maker.

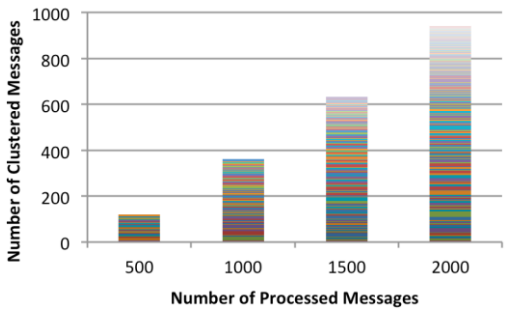

Fig. 8: No. of Clustered Messages versus No. of Processed Messages

Based on the experimental results, it is possible to see that instead of dealing with messages related to similar themes scrambled in a big bag of messages, the user can deal with groups containing related or similar messages. As demonstrated in the experiments, considering 2000 messages, the users have almost 50% of messages grouped in approximately 100 clusters. Thus, the theme switching caused by information associated with a different thematic group was reduced, facilitating the decision making process.

5. Conclusion and Future Works

This paper discusses a strategy for clustering textual messages using multi-criteria in a scenario of major events. The goal is to facilitate the decision making process once these messages would be grouped according to some criteria that may provide a proper contextualization for the messages. These message clusters, named themes, offer a better situational awareness for people involved in the operations that are executed in events. This way, a decision-maker can read the all messages related to a subject sequentially without having to change to another subject every time he read a different message. This strategy makes easier the decision process.

The proposed strategy uses an approach of multi-criteria clustering inspired in the 5W2H approach. To evaluate the strategy, it was used three features of the messages: temporal difference among the messages sending, distance among the messages senders and textual similarity. Each one of these attributes is used as a different operator. One operator acts as a filter for the next, therefore the order of the operators is also analyzed to optimize the global application of them.

Additionally, to set the parameters of the proposed operators, it is proposed two news concepts: Thematic Strength and Thematic Density that are applied in a combined way to calibrate a set of messages. This approach allows calibrating the parameters of the operators through the required features desirable for the cluster, customizing the operators in different scenarios.

The Strategy is applied to messages obtained from real operations executed in the 2014 FIFA World Cup. The results show the approach is very promising. The clusters of messages that can be produced may facilitate the decision-making process in future events, as well as analysis of past events that may help the organization of forthcoming events.

In future works, the parameters configuration should also consider the multi-criteria decision making method AHP to calibrate the weight of the concepts that are used in a combined way. Besides, we intend to apply the operators considering selectivity analysis.

References

[1] O. M. Anshakov and O. M. T. Gergely, *Cognitive Reasoning: A Formal Approach*, 1st ed. Berlin, Germany: Springer Berlin Heidelberg, 2010.

[2] R. Hastie and R. M. Dawes, *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*, 2nd ed. California, USA: SAGE Publications, 2010.

[3] D. D. Salvucci and N. A. Taatgen, *The Multitasking Mind*, Oxford Series on Cognitive Models and Architectures. New York, USA: Oxford University Press, 2011.

[4] Y. Peng, Y. Zhang, Y. Tang and S. Li, "An incident information management framework based on data integration, data mining, and multi-criteria decision making," *Decision Support Systems*, vol. 51, no. 2, pp. 316-327, May, 2011.

[5] D. Schuff, O. Turetken and J. D'Arcy, "A multi-attribute, multi-weight clustering approach to managing "e-mail overload"," *Decision Support Systems*, vol. 42, no. 3, pp. 1350-1365, December, 2006.

[6] D. G. Roussinov and H. Chen, "Document clustering for electronic meetings: an experimental comparison of two techniques," *Decision Support Systems*, vol. 27, no. 1–2, pp. 67-79, November, 1999.

[7] L. C. C. Souza and W. A. Pinheiro, "A Strategy to Identify Data from the Same Object Based on MIP Data Model," in *Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI & CBIC)*, Ipojuca, Brazil, 2013, pp. 675-680.

[8] L. C. C. Souza and W. A. Pinheiro, "An Approach to Data Correlation using JC3IEDM Model," in *34th IEEE Military Communications Conference (MilCom)*, Tampa, USA, 2015, pp. 1099-1102.

[9] P. Christen, "Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1537-1555, July, 2012.

[10] G. D. Bianco, "Reduction of the User Effort to Deduplication Configuration in Big Databases (in Portuguese: Redução do Esforço do Usuário na Configuração da Deduplicação de Grandes Bases de Dados," M.S. thesis/PhD. dissertation, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, Brazil, 2013.

[11] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas and M. Teboulle, 1st ed. Berlin, Germany: Springer Berlin Heidelberg, 2006, pp. 25-71.

[12] L.-A. Vidal, F. Marle and J.-C Bocquet, "Using a Delphi Process and the Analytic Hierarchy Process (AHP) to Evaluate the Complexity of Projects," *Expert systems with applications*, vol. 38, no. 5, pp. 5388-5405, May, 2011.

[13] M. Velasquez and P. T. Hester, "An Analysis of Multi-Criteria Decision Making Methods," *International Journal of Operations Research*, vol. 10, no. 2, pp. 56-66, May, 2013.

[14] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, 2nd ed. New York, USA: ACM Press/Addison Wesley, 1999.

[15] T. Korenius, J. Laurikkala, K. Jarvelin and M. Juhola, "Stemming and Lemmatization in the Clustering of Finnish Text Documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04)*, New York, USA, 2004, pp. 625-633.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, 3rd ed. Massachusetts, USA: Massachusetts Institute of Tecnology, 2009.

[17] A. Khizhnyak. (2007, April 11). *Apatar Administration Guide* [Online]. Available: http://www.apatar.com/pdf/Apatar Administration.pdf.

Biographical notes

Wallace Anacleto Pinheiro
He received his D.Sc. in Computer Science from Federal University of Rio de Janeiro in 2010. He works at the Centro de Desenvolvimento de Sistemas (CDS), a Center of Technology and Science of the Brazilian Army, in Brasília, Brazil. His current research interest includes: information retrieval, query optimization, data quality and data mining.

www.shortbio.net/wallace@cds.eb.mil.br

Ricardo Queiroz de Araujo Fernandes
He received his D.Sc. in Computer Science from Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 2012. He works at the Centro de Desenvolvimento de Sistemas (CDS), a Center of Technology and Science of the Brazilian Army, in Brasília, Brazil. His current research interest includes: information retrieval and clustering.

www.shortbio.net/ricardo@cds.eb.mil.br

Luciene Carvalho C. Souza
She received her M. Sc. in Computer Sciences from the Military Institute of Engineering (IME) in 2014. He is currently working at the Centro de Análise de Sistemas Navais (CASNAV), a Center of Systems Analysis of the Brazilian navy in Rio de Janeiro, Brazil. His current research interests include information retrieval, clustering and information quality.

www.shortbio.net/luciene.carvalho@casnav.mar.mil.br