Levels of similarity in user profiles based cluster techniques and multidimensional scaling

Gustavo Rodríguez Bárcenas, Alex Cevallos Culqui, Jorge Rubio Peñaherrera, Segundo Corrales Beltrán, Enrique Torres Tamayo

Abstract—User profiles collected a set of distinctive features that characterize it, each user has their own interests and needs, according to their cognitive development, their experience of life, which makes them unique, user profiles can be derived uncountable studies, research principles and methods are used to build user profiles and taking into consideration the basic lexical semantics contained in these profiles could be identified levels of similarity and compatibility between users. It applies to a specific case study in a research center, it was proved that the vector space model, cluster analysis and multidimensional scaling are methods that can be integrated ICT with the aim of obtaining the perceptual relations different users of the system and the identification of Collective Knowledge Communities.

***Keywords*—**Similarity, cluster analysis, user profiles, multidimensional scaling, vector space model.

I. Introduction

THIS key element of any system of information and rationale for any entity engaged in providing information services is the user who satisfies these needs, interests and demands for information. For all offer information becomes a critical knowledge of the user, who is considered the alpha and omega of such offers. The user is the main character of the computer screen, it is the beginning and end of the cycle of transfer of information: it asks, analyzes, evaluates and recreate the information [1-5].

Gustavo Rodriguez Barcenas is PhD. Research Professor, Department of Computer Engineering and Computer Systems at the Cotopaxi Technical University, Ecuador (phone: +593 987 658 959, email: gustavo.rodriguez@utc.edu.ec).

Alex Cevallos Culqui is Research Professor, Department of Computer Engineering and Computer Systems at the Cotopaxi Technical University, Ecuador (email: alex.cevallos@utc.edu.ec).

Jorge Rubio Peñaherrera is Research Professor, Department of Computer Engineering and Computer Systems at the Cotopaxi Technical University, Ecuador (email: jorge.rubio@utc.edu.ec).

Second Corrales Beltran is Research Professor Department of Computer Engineering and Computer Systems at the Cotopaxi Technical University, Ecuador (email: segundo.corrales@utc.edu.ec).

Enrique Torres Tamayo is PhD. Research Professor, Department of Electrical Engineering at the Cotopaxi Technical University, Ecuador (email: enrique.torres@utc.edu.ec).

Today, organizations face a market that simultaneously becomes more competitive, specialized, global and entrenched on the Internet. The Information Technology and Communications are increasingly a focus for policy makers and corporate strategists concerned with development issues. Therefore, the implications of information technology beyond the way how are offered, distributed, sold and consumed services.

The term User Information are set forth in different ways in general, it can be classified information to the user as an individual who needs information for the continued development of its activities.

According to [2, 4-7] means the user as:

- Related actual or potential person, with the use of information systems.
- Interacting social and communication in a changing society and conflict Stars.
- Humans socially related, belonging to different social classes and possess cultural capital, habits and different worldviews.

Their information needs and seeking behaviors emerge in epistemological, social, cultural processes, and make a different use of information systems, collective, interactive, communications, construction and social transformation processes.

Obviously all computer system in some way are developed to meet training or information needs of users today working to implement new methods that allow better identification and representation of information and knowledge and on the other hand people who have such knowledge either implicitly or explicitly, in order to enable users to infer positioning and establish relations based on the analysis resulting from the representations obtained in the same, following the premises identifying the ICT context Today and the reference to a new paradigm of representation and visualization based on Web 2.0 and Internet 2.0.

The research aims to solve the related problems with the identification of the similarity which may exist between users of a system, on the fundamental patterns, fields of interest and social development, for it therefore seeks to determine the levels of similarity in profiles based user clustering techniques, multidimensional scaling in order to establish perceptual patterns and relationships among a community of users.

II. Definition of user profiles in Computer Systems

To Samper (2005) profile is a word that comes from the Latin pro filare, which means designing contours. A profile is a model of an object, a compact representation that describes its most important features, which can be created in the memory of a computer and can be used to represent the object in the computational tasks. Popular applications that create and manage profiles include personalization, knowledge management and data analysis.

The source profile, derived from psychology, understood as a set of different measures of a person or group, each of which is expressed in the same unit of measurement is also recognized. That is, certain characteristics of an individual are measured by tests that give different scores, these scores are its profile, which is used for diagnostic purposes [8]. Considering the above approach can understand the user's profile as a set of distinctive features that characterize.

In the case of a user profile of a software system, it can understand both personal data and characteristics of the computer system, as well as behavior patterns, personal interests and preferences. This user model is represented by a data structure suitable for analysis, recovery and use. In computer terms: a user profile is the representation of a set of characteristics that describe a person, in his role as an adaptive system user. A user profile is stored in most cases in the form of attribute-value pairs. The system stores, analyzes and makes available this information to the adaptive part (Corti, 2000).

The profile is built from the characteristics that identify and characterize a user on another and the factors of influence that surround [9, 10].

Each user has their own interests and needs, according to their cognitive development, the environment in which it operates and their life experience, which make them unique, user profiles can be derived innumerable studies for determining the level interaction between them, depending on the expertise fields collected in your profile, compatibility level of similarity or distance between them, clusters of users responding to the parameters defined in your profile.

A user profile is a set of data, mostly textual nature, though technological developments have led to incorporate text pictures, graphic, etc.

The range of information it collects a user profile is steadily growing, research textual reference to nature or terminology that will be collected in user profiles will.

User profiles will be stored in the database system, the database is a matrix in which each row represents a user and each column indicates the presence or not of a given term in its corresponding profile.

We can consider a database User Profiles *(U)*, users comprising u_i, where they have been given a set of terms *(T)*, formed by n terms t_j, in which each user u_i It contains a number of terms, as a result of the fields entered in the profile. Thus, it is possible to represent each user as a vector belonging to an n-dimensional space, the number of terms entered in the profile forming the set T n being:

$$u_i = (t_{i1}; t_{i2}; t_{i3}; \dots\dots\dots; t_{in}) \quad (1)$$

Where each of the elements t_{ij} this vector can represent the presence, absence or term relevance t_j in the user u_i on your profile.

III. Methods

For this research, a system that performs a set of actions as shown below develops:

- Creation of the user profile.
- Determining the similarity or proximity to other users.
- Determination of user groups from cluster techniques and Multidimensional Scaling (Multidimensional Scaling, MDS).
- Determining the level of similarity and distance between system users.

For the development of this system and represent it in a case study the following criteria are taken into consideration:

1. Addressing the functional aspects for the development of the system, defining the fundamental processes by means of user stories.
2. Determine or establish the aspects related to the design and implementation of the system. Present engineering tasks each system module.
3. Carry out performance tests of the system, acceptance tests. The tests are performed by modules for the acceptance of each independently.

A. *Considerations for creating user profile*

To create the user profile the premises described by Samper (2005) are taken, it is taken as a standard to follow the explicit method because it is required that the profile is built from own analysis and assessment made by the user himself same, according to their interests and motivations, for it also considered the following criteria:

1. Acquisition of data: the acquisition of the data is taken as reference method explicit information.
2. Representation Profile: inductive reasoning method is used as the inductive reasoning is progress from the particular to the general, so the user interaction with the system is monitored, this will reuse the information in your profile.
3. User Feedback: be considered the method of explicit feedback, because it is obtained according Samper (2005) asking the user directly. You may be asked to complete a questionnaire or make a value judgment about something, or simply edit your profile by adding new parameters related to their interests and activities that are essential elements for performance.

B. *User profile fields*

They shall refer to a type user, or a user who operates in a research context for this data which will define the user profile in your computer. Next to them is based on the total, labor and education, experiences of a person or user.

For the preparation of the matrix of terms will be used fields that describe the user profile where more relevant there, as are the identification of their knowledge, information needs and user interests, specialties, etc., these are listed below and in Fig. 1and 2 and 3:

- Name of the education.
- Name of additional training.
- Specialties.
- Topics of interest and subject descriptors.
- Keywords of investigations.
- Keywords of published articles.
- Keywords of papers presented at events, seminars and conferences.

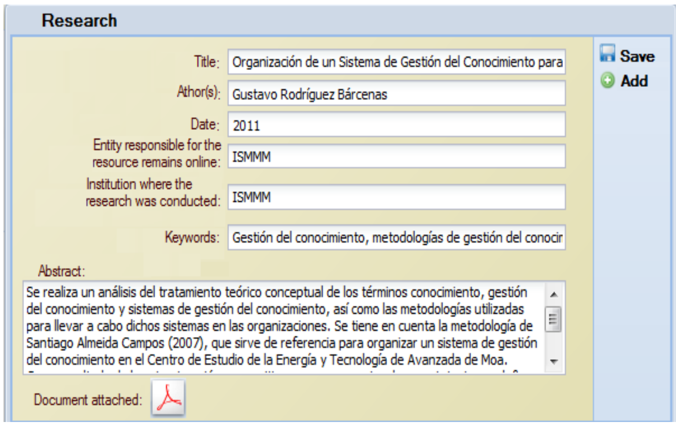


Fig. 1 Profile Section, where research data are input made.

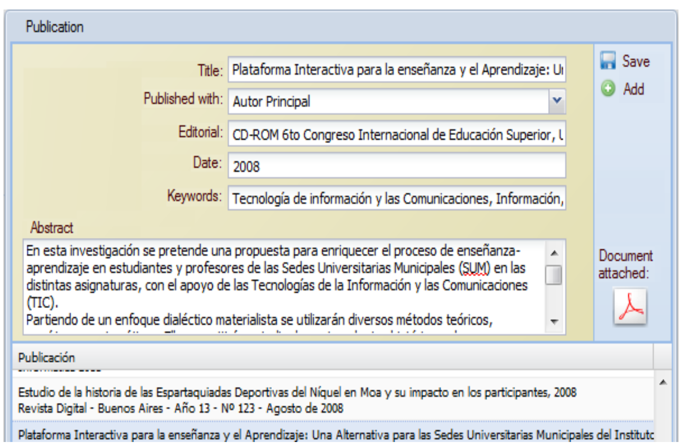


Fig. 2 Section of the profile, where the data are entered in publications.

Fig. 3 Compatible link Fields with a specific user.

C. *Weight terms related to user profiles*

According to the expression (1) the process of construction of the vectors - user database of user profiles include the removal of the terms in which the representation of users will be made by removing the contents of profile information. The main task of this method is given by the automatic association of the representation of each user based on the content of information, that is, determine the weights of each term taken from your user profile in the vector u_i. Its role will be:

$$F: U \times T \rightarrow [0, 1]$$

The representation of each vector-user will component, of which they are referenced in the profile will have a different value of 0, while those that are not referenced will have a null or 0 value.

The frequency of occurrence of a term in a profile of some form determines its importance in suggesting that these frequencies can be used to summarize the area of knowledge in which the user or the main interests of the same moves.

Following what describes the vector space for Recovery Systems Model, and a continuation of the methods used to store the terms contained in the profile of each user, continue with the selection process, this is followed to determine the importance or weight of each term in the vector-user. Calculating the weight or importance of each term it is called weighting term.

Gerald Salton weight using this concept in his recovery model based on the vector space. In this model, a matrix term / document representing the database is formed. Each vector of the matrix represents a document; each element of the vector will have value 0 (zero) if the document does not contain the term; weight or value of the term if they contain [3, 11-17].

A first approach is based on counting the occurrences of each term in a document, as it is often called the term *ith* the *j-th* document, and it shows as *tfi, j*. A second measure of the importance of the term is known as inverse document frequency of a term in the collection, usually known by its acronym *idf* (inverse document frequency), as reflected [12, 18] and responding to the following expression:

$$w_{i,j} = tf_{i,j} \times Log\left(\frac{N}{n_i}\right) \quad (2)$$

Where N is the number of documents in the collection, and n_i the number of documents that mention the *i-th* term, if we associate the case of this research with *(U) (N)* as the number of users of the database user profiles, and n_i as the number of users contained in i the term profile, then it is possible to determine the importance or weight of each term in the profile of each user.

D. *Similarity between system users*

Similarity calculation is taken into account between the vectors making up the weight matrix, which are essentially vector-users, for the degree of relevance of a user u_i by profile with respect to the others that compose the matrix, you may establish the similarity between vectors of this matrix, or as each vector be a user and will ascertain the similarity of each user with respect to the other. The system takes a real value will be greater the more similar the users are analyzed.

There are different functions to measure the similarity between vectors, all of which are based on considering both as points in an n-dimensional space, the function cosine is one of them:

Cosine function:

$$F\ \cos(A,B) = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A_j^2 \cdot \sum_{j=1}^{n} B_j^2}} \qquad (3)$$

Where A_j y B_j are respectively the weights associated with the term t_j in the representation of users A y B.

Typical functions generate similarity values between 0 to elements without similarity, and 1 for completely equal elements.

A similarity matrix may be displayed symmetrically, each δ_{ij} element M represents the similarity between stimulus i and j stimulus as shown in M:

$$M = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & & \delta_{1n} \\ \delta_{21} & \delta_{22} & \delta_{23} & \cdots & \delta_{2n} \\ \delta_{31} & \delta_{32} & \delta_{33} & & \delta_{3n} \\ & \vdots & & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \delta_{n3} & \cdots & \delta_{nn} \end{pmatrix}$$

E. Multidimensional Scaling to perceptually represent users

The MDS is a technique of spatial representation that is displayed on a map a set of stimuli whose relative position you want to analyze.

In researching his objective will be focused on obtaining a spatial representation that is a map that displays the perceptual relationship between the various users of the system, so that they can see what users are near and far including from its setting on your user profile. This is possible due to the transformation of the similarity distances between them that can be represented in a multidimensional space.

The procedure, in very general terms, follow some basic ideas in the most technical. The starting point is a matrix of similarity between n objects, with δ_{ij} element in row i and column j, which represents the similarity of object i to object j. The number of dimensions, p, is also set to make the graph of objects in a particular solution. Generally it follows the path as [19-30] is:

Fix the n objects in an initial configuration in p dimensions, that is, assume for each object coordinates ($x1$, $x2$, ..., xp) in the space of p dimensions.

Calculate the Euclidean distances between objects in that configuration, that is, calculate d_{ij}, which are the distances between the object i and object j.

$$d(O_i, O_j) = \sqrt{\sum_{k=1}^{n} \left(x_k(O_i) - x_k(O_j)\right)^2} \qquad (4)$$

Where O_i y O_j are the objects for which you want to calculate the distance, n is the number of characteristics of objects in space and $x_k(Oi)$, $x_k(Oj)$ is the value of the k-th attribute in Oi y Oj, respectively.

So you should also check the following three axioms:

- $\mathrm{d(x,y)} \geq 0 \quad \forall\, \mathrm{x,y} \in \mathrm{X}$, y $\mathrm{d(x,y)} = 0$ If and only if $\mathrm{x} = \mathrm{y}$
- $d(x,y) = d(y,x)\ \forall\, x,y \in X\ (symmetry)$
- $d(x,z) \leq d(x,y) + d(y,z)\ \forall\, x,y,z \in X\ (triangle\ inequality)$

Make a regression of d_{ij} over δij. This regression can be linear, polynomial or monotonous. Using the method of least squares estimates of the coefficients a y b are obtained, and hence can be obtained which is known generically as a "disparity".

$$\hat{d}_{ij} = \hat{a} + \hat{b}\delta_{ij} \qquad (5)$$

If a monotonic regression was assumed, an exact relationship between d_{ij} and δ_{ij} it does not fit but simply assumed that if δ_{ij} grows, then d_{ij} grows or remains constant.

Through a convenient statistic, the goodness of fit between the distances of the configuration and disparities measured. There are different definitions of this statistic, but the majority comes from the definition of so-called stress index.

One of the criteria used is as follows:

$$STRESS1 = \sqrt{\frac{\sum\sum\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum\sum d_{ij}^2}} \qquad (6)$$

All summations over i and j ranging from 1 to p and disparities depend on the type of regression used in the third step of the process.

STRESS1 is the formula introduced by Kruskal who offers the following guidance for interpretation in table 1:

Table 1. Interpretation of Stress. Source: Kruskal (1964).

STRESS1 Size	Interpretation
0.2	Poor
0.1	Regular
0.05	Good
0.025	Excellent
0.00	Perfect

The coordinates ($x1$, $x2$, ..., x_t) of each object are changed slightly so that the extent of adjustment is reduced.

The distance matrix (D), matrix coordinates (X) of the stimuli are represented in a space of n dimensions (in the case of research just 2 dimensions).

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} & & d_{1n} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & d_{33} & & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & d_{nn} \end{pmatrix}$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

F. Cluster analysis to identify clusters of users

The cluster may establish the hierarchy in terms of user groups, based on similarity matrices and obtained away. Thus, each user can identify with the group it belongs according to how distant it is, or the like.

Agglomerative hierarchical algorithm based on distance:

1. Start with N clusters (the initial number of elements) and an $N \times N$ symmetric matrix of distances.
2. Within the distance matrix, find clusters that between the U and V which is the lowest among all, d_{uv}.
3. Joining the U and V clusters into one. Update distance matrix:
   I. Deleting rows and columns of U and V cluster.
   II. Forming the row and column distances new cluster (UV) and other clusters.

Repeat steps (2) and (3) a total of (N-1) times, that is if all the points are in the same cluster, finish; but, again steps (2) and (3).

While hierarchical groups gradually build algorithms, algorithms try to discover cluster partition iteratively relocating points between subsets.

K-means algorithm as [31-34] is one of the simplest and known clustering algorithms. It is based on the square error optimization, following an easy way to divide a given database into k groups fixed a priori. The main idea is to define k centroids (one for each group), and then locate the remaining points in the class of its nearest centroid. The next step is to recalculate the centroid of each cluster and relocate the points again in each group. The process is repeated until no changes in the distribution of the points from one iteration to the next.

IV. Results

The results correspond to the study of a specific case involving the Center for Energy Studies and Advanced Technologies (CESAT) responsible in several issues of the Energy Efficiency and Rational Use of Energy (EERUE), user profiles they have been created with reference to several researchers in the study center. Your needs and the priority knowledge in this area are reflected in the profile built by the actor himself or the person responsible for administering the system.

As a result of the implementation of the system they are recorded multiple users, many of them members and supporters of the study center; for better understanding and comprehension, they were only considered some actors who respond to CESAT system so that it can be displayed more legibly expose what is intended.

Table 2 shows some fields from multiple users are displayed in the system. ID (numeric identifier in the database) it represents not achieve, because these are just an intentional sample in order to reveal the system functionality as the number of terms and other elements constituting calculations procedures similarity distance and described in methods. Initials are to identify users in the investigation.

Table 1. Some of the users of the system

id	Username	Initials	Specialty
39	egongora	U_1	Thermodynamics and air conditioning specialist
40	rmontero	U_2	Total specialist efficient energy management
41	iromero	U_3	Specialist electrical machines
42	alegra	U_4	Specialist in mathematical modeling, simulation and research methodology
43	lrpuron	U_5	Specialist in artificial intelligence applied to industrial processes
44	yretirado	U_6	Specialist ore drying using solar energy
47	grbarcenas	U_7	Specialist Information Technology and Communications Processes
49	yaguilera	U_8	Specialist Computer Networks and Communications
50	dgonzalezr	U_9	Computer specialist 1
51	eromero	U_{10}	Computer specialist 2

From the selection of fields taken into account in the 10 users previously selected in Table 2, a total of 470 lexical-semantic elements made between terms and phrases that identify, specialty knowledge domains, keywords are obtained, among others.

Counting occurrences of each term in the profiles of the selected users, is the parent frequency of terms in these profiles, its magnitude is represented by Table 3.

Table 3. Matrix user profiles

$$\begin{array}{c|c} & t_1 \quad t_2 \quad t_3 \quad \dots \quad t_n \\ \hline \begin{array}{c} User_1 \\ User_2 \\ User_3 \\ \\ User_n \end{array} & \begin{pmatrix} \mathbf{1} & \mathbf{2} & \mathbf{1} & & \boldsymbol{n} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \cdots & \boldsymbol{n} \\ \mathbf{0} & \mathbf{2} & \mathbf{1} & & \boldsymbol{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{n} & \boldsymbol{n} & \boldsymbol{n} & \cdots & \boldsymbol{n} \end{pmatrix} \end{array}$$

Expression (2) assuming that the number of selected users (N) is equal to 10, the weight matrix (W) of lexical elements contained in the profiles of the users of the system as shown in Table 4 was obtained , by dimensions of the table only a representation of the structure it is shown.

From the application of the cosine function in equation (4) and the weight values obtained through its representation in Table 4, obtained as results a symmetric matrix of similarities between users, as seen in Table 5.

Table 5. Matrix similarity using the cosine function.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}
U_1	1	0.081	0.085	0.062	0.023	0.158	0.010	0.002	0.000	0.001
U_2	0.081	1	0.142	0.016	0.432	0.078	0.011	0.019	0.000	0.000
U_3	0.085	0.142	1	0.011	0.158	0.018	0.006	0.014	0.001	0.001
U_4	0.062	0.016	0.011	1	0.012	0.003	0.016	0.013	0.012	0.008
U_5	0.023	0.432	0.158	0.012	1	0.063	0.038	0.023	0.000	0.000
U_6	0.158	0.078	0.018	0.003	0.063	1	0.037	0.005	0.000	0.000
U_7	0.010	0.011	0.006	0.016	0.038	0.037	1	0.259	0.219	0.175
U_8	0.002	0.019	0.014	0.013	0.023	0.005	0.259	1	0.647	0.391
U_9	0.000	0.000	0.001	0.012	0.000	0.000	0.219	0.647	1	0.690
U_{10}	0.001	0.000	0.001	0.008	0.000	0.000	0.175	0.391	0.690	1

Given the similarities obtained and empirical assessments made by the author, totally intentional, a level of compatibility between system users is raised, as shown in Table 6.

Table 6. Variables and linguistic labels for compatibility.

List of variables and linguistic labels for compatibility (ES = similarity)		
Interval value	Linguistic variables Compatibility	Linguistic label
$S = 0$	No compatibility	I
$0 < S < 0.1$	Compatibility Extremely Low	CEL
$0.1 \leq S < 0.25$	Compatibility Very Low	CVL
$0.25 \leq S < 0.5$	Compatibility Moderately Low	CML
$S = 0.5$	Media compatibility	MC
$0.5 < S < 0.75$	Moderately High Compatibility	MHC
$0.75 \leq S \leq 0.99$	Compatibility Very High	CVH
$S = 1$	Compatibility	C

Table 7. Level of compatibility between selected users of the system.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}
U_1										
U_2	CEL									
U_3	CEL	CVL								
U_4	CEL	CEL	CEL							
U_5	CEL	CML	CVL	CEL						
U_6	CVL	CEL	CEL	CEL	CEL					
U_7	CEL	CEL	CEL	CEL	CEL	CEL				
U_8	CEL	CEL	CEL	CEL	CEL	CEL	CML			
U_9	CEL	I	CEL	CEL	I	I	CVL	MHC		
U_{10}	CEL	I	CEL	CEL	I	I	CVL	CML	MHC	

Table 7 shows the result of interpolation of linguistic labels set in table 6, showing the level of compatibility between selected users from the system. The following aspects are seen in general:

- Compatibilities are very low (CVL) gongora between users (specialist in refrigeration and air conditioning) and yretirado (specialist ore drying using solar thermal); between rmontero (specialist in total efficient energy management) and iromero (specialist in electrical machines); between iromero and lrpuron (specialist in artificial intelligence applied to industrial processes); between dgonzalezr (computer specialist 1) and grbarcenas (specialist ICT and knowledge management) and between iromero (computer specialist 2) and grbarcenas.
- Extremely low compatibility (CEL): egongora between the user and other users except for yretirado; between rmontero and alegra (specialist in mathematical modeling, simulation and research methodology), yretirado, grbarcenas and yaguilera (specialist in computer networks); between iromero and other users except lrpuron; between alegra and other users; between lrpuron and yretirado, grbarcenas and yaguilera and between yretirado and yaguilera.
- Incompatibility (I): rmontero between users and yretirado lrpuron with dgonzalesr and eromero.
- Compatibility moderately low (CML) between users and lrpuron rmontero; between grbarcenas and yaguilera; between yaguilera and eromero.
- Compatibility moderately high (MHC) between users and dgonzalezr yaguilera and between eromero dgonzalezr and users.

Another result is the existence of a number of selected actors are graduates of the same specialties, but represent something distant domains of knowledge, example of this are the yaguilera and rmontero users and both are graduates of Electrical Engineering respectively, however rmontero represents the domain of EERUE and yaguilera the domain of telematic systems, only joins their training and therefore extremely low compatibility with a similarity of 0.019, other cases are the grbarcenas user regarding yretirado egongora and the three are graduates of Mechanical Engineering with a similarity of 0.010 and 0.037 respectively grbarcenas about them represents a different domain knowledge, however yretirado between egongora and there is a similarity of 0.158 representing both domains of similar knowledge.

In fig. 4 the compatibility level display for the user (grbarcenas) users with greater similarity (yaguilera, dgonzalez and eromero) shown, the remaining others exhibit compatibility Extremely Low, or similarity below 0.100. This level of support reflects the relationship between knowledge and interests that actors have to mind these are the users of the system.

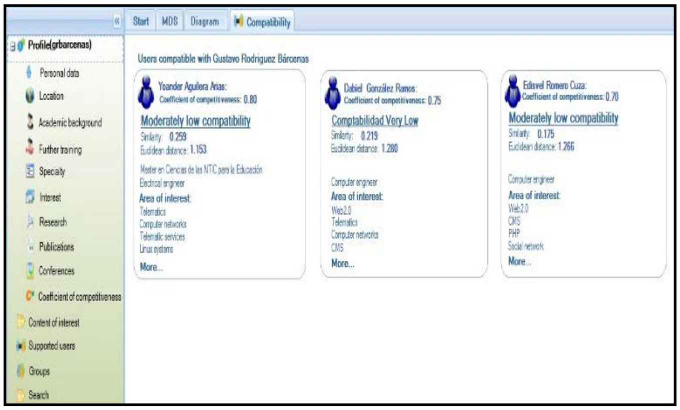


Fig. 4 Level of support for the grbarcenas user.

From the methodological procedures for viewing MDS was obtained as shown in figure 5, obtained mainly two groups, one (A- left) comprised dgonzalezr, eromero, yaguilera and grbarcenas, representing a collective community ICT-related knowledge and its application; the other group comprising the remainder (B) represent a collective community knowledge linked to EERUE in both groups two users who in some way are borders, these are grbarcenas and joy are displayed, this is the result of the heterogeneity of fields knowledge in both incursions.

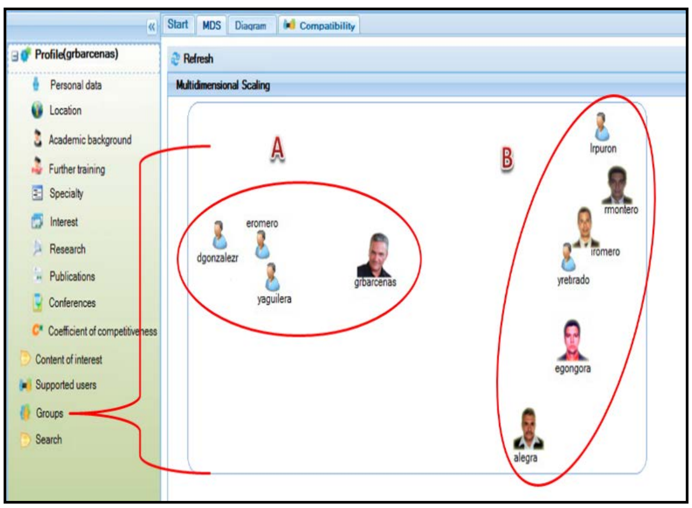


Figure 5. MDS representation of the chosen system users.

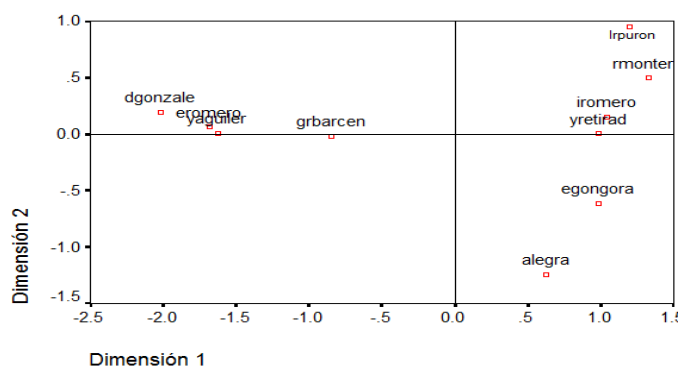
The test results of the system compared with professional software SPSS, the distance matrix between the selected players (Table 8), from this and the methodological procedures is obtained coordinates in two dimensions are obtained (Table 9), resulting in the representation of figure 1, so compared with that obtained by the system is perceived to have similar distribution and location in the plane formed by the two established dimensions.

Table 8. Matrix of Euclidean distance between the actors.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}
U_1	0	1.368	1.31	1.34	1.432	1.195	1.461	1.631	1.725	1.641
U_2	1.368	0	1.245	1.462	0.806	1.363	1.508	1.659	1.77	1.691
U_3	1.31	1.245	0	1.413	1.228	1.396	1.467	1.618	1.724	1.642
U_4	1.34	1.462	1.413	0	1.467	1.416	1.44	1.601	1.693	1.613
U_5	1.432	0.806	1.228	1.467	0	1.386	1.479	1.651	1.765	1.687
U_6	1.195	1.363	1.396	1.416	1.386	0	1.423	1.621	1.719	1.637
U_7	1.461	1.508	1.467	1.44	1.479	1.423	0	1.153	1.28	1.266
U_8	1.631	1.659	1.618	1.601	1.651	1.621	1.153	0	0.585	0.867
U_9	1.725	1.77	1.724	1.693	1.765	1.719	1.28	0.585	0	0.51
U_{10}	1.641	1.691	1.642	1.613	1.687	1.637	1.266	0.867	0.51	0

Table 9. Stimulus coordinates of each actor in two dimensions.

Stimulus coordinates		
	Dimension	
Actor	1	2
egongora (A_1)	0.9866	-0.6119
rmontero (A_2)	1.3264	0.4978
iromero (A_3)	1.0421	0.1517
alegra (A_4)	0.6250	-1.2414
lrpuron (A_5)	1.2008	0.9497
yretirado (A_6)	0.9863	0.0060
grbarcenas (A_7)	-0.8449	-0.0180
yaguilera (A_8)	-1.6258	0.0074
dgonzalezr (A_9)	-2.0146	0.1956
eromero (A_{10})	-1.6820	0.0632



Graphic 1. Configuration stimuli resulting in two dimensions.

From the methodological procedures in section methods linked to hierarchical cluster analysis dendrogram in Figure 6, where you can highlight a certain way and to corroborate the results obtained in the MDS is obtained, a cluster is observed more accentuated in distance between dgonzalezr, eromero and yaguilera and these linked to grbarcenas; Likewise the link alegra with clusters formed by egongora, yretirado, rmontero, lrpuron and iromero, hierarchically seen the link between all these users with different levels of compatibility.

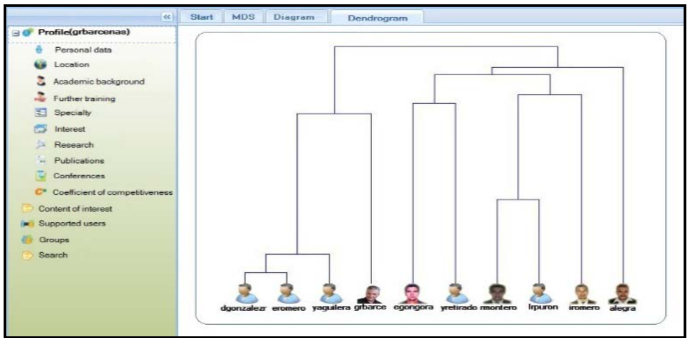


Figure 6. Representation of a dendrogram of users of the system selected from the hierarchical cluster analysis.

V. Discussion of Results

In this section, a tool for viewing relationships between actors CESAT, knowledge, collective knowledge communities, which are used in an intuitive way, to help users easily understand their current status regarding presents others, as well as access their explicit knowledge and level of compatibility. This tool is based on distance and similarity measures, and with the implementation of MDS and clustering algorithms identify and represent the different groups of people with similar characteristics.

The reconciliation process involves extracting terminology profiles for the relationship between the actors in the CEETAM. Therefore, to fully automate this process was a complex task due to the high number of interactions required. However it is noteworthy that these actions make use of ICT, mainly describing the World Wide Web, for viewing from distances and levels of similarity between actors compatibility, responding to new trends in virtual environments in this area, know which facilitates informal networks of actors in the organization, as referred to [35, 36] in their work related to social networks, collective intelligence and social capital.

VI. Conclusions

A tool as a result of the combination of theoretical and technological aspects that allows the link between the transfer of knowledge and collective or shared intelligence developed.

The vector space model, the cluster analysis and multidimensional scaling are methods that can be integrated ICT with the aim of obtaining the similarity distance, conglomerates, compatibility, map of perceptual relationship between users of a system, as It was demonstrated in the case of CESAT.

In the case study could identify levels of support among researchers Study Center users, being able to visualize the relationship between them and possible knowledge communities.

References

1. Du, J.T. and A. Spink, *Toward a web search model: Integrating multitasking, cognitive coordination, and cognitive shifts.* Journal of the American Society for Information Science and Technology, 2011. **62**(8): p. 1446-1472.
2. Salazar, P.H. *The user profile information*. E-Journal 1993 [cited 2009 December 12]; Available from: http://www.ejournal.unam.mx/.
3. Samper, J.J., *Study and evaluation of an intelligent system for recovery and internet filtering information.*, in *Department of Architecture and Computer Technology*. 2005, Granada University: Granada.
4. Cuza, E.R., *Automated System Recovery Information in Virtual Environments based on User Profiles*, in *Department of Computer Science*. 2010, Metallurgical Mining Higher Institute of Cuba: Moa.
5. Day, R.E., *Death of the user: Reconceptualizing subjects, objects, and their relations.* Journal of the American Society for Information Science and Technology, 2011. **62**(1): p. 78-88.
6. Ramírez, D. *Recovery and Information Organization. Recovery models*. 2007 [cited 2012 January 12]; Available from: http://modelos-recuperacion.50webs.com/recuperacion-modelo-booleano.html.
7. Sun, J., *Why different people prefer different systems for different tasks: An activity perspective on technology adoption in a dynamic user environment.* Journal of the American Society for Information Science and Technology, 2012. **63**(1): p. 48-63.
8. Corti, R. *Learning Support System Diagnosis Using User Profiles: EndoDiag II*. 2000 [cited 2011 December 12]; Available from: http://www.fceia.unr.edu.ar/~acasali/publicaciones/endodiag2.pdf.
9. Naranjo, E. and D. Álvarez. *Information literacy: a way to encourage reading*. 2003 [cited 2009 December 20]; Available from: http://docencia.udea.edu.co/bibliotecologia/seminario-estudios-usuario/unidad2/unidad2.html.
10. Ahn, J., *The effect of social network sites on adolescents' social and academic development: Current theories and controversies.* Journal of the American Society for Information Science and Technology, 2011. **62**(8): p. 1435-1445.
11. Broncano, R.G., *Recovery Models*, in *Recovery and access to information*. 2006: University of Madrid Carlos III.
12. López-Herrera, A.G., *Models of Information Retrieval Systems Based Document Fuzzy Linguistic Information*, in *Department of Computer Science and Artificial Intelligence*. 2006, Granada University: Granada. p. pp. 237.

13. Pérez, C.A., et al., *Evaluation of Algorithms Based on Fuzzy Logic Applied to Processing of Open Hole Log Data.* Engineering and Region, 2010. **6**(1).
14. Salton, G., A. Won, and C.S. Yang, *A Vector Space Model for Automatic Indexing.* Comunication of the ACM, 1975. **18**(11).
15. Salton, G., *The SMART Retrieval System*. 1971: Prentice-Hall.
16. Salton, G. and M.J. McGill, *Introduction to Modern Information Retrieval*, in *Computer Science Series*. 1983, McGraw-Hill.
17. Salton, G., *Automatic Text Procesing – The Analysis, Transformation and Retrieval of Information by-Computer*, Addison-Wesley, Editor. 1989.
18. Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*. 1999: ACM Press Books & Addison-Wesley.
19. Assent, I., R. Krieger, and B. Glavic, *Clustering multidimensional sequences in spatial and temporal databases.* Knowledge Information System, 2008. **16**: p. pp. 29-51.
20. Borg, I. and P. Groenen, *Modern multidimensional scaling*, in *MDS Aplications. New York: Springer Verlag*. 1997: New York: Springer Verlag.
21. De Leeuw, J. and P. Mair. *Multidimensional scaling using majorization: SMACOF in R.* Statistics Preprint Series 2008 12/02/2011]; Available from: http://preprints.stat.ucla.edu/537/smacof.pdf.
22. Diaz, J.O., R.S.M. Castellanos, and J.V. Mallou, *Escalamiento Unidimensional y Multidimensional de Diseños Creativos.* Psicothema, 1992. **4**(1): p. pp. 291-296.
23. Guerrero-Casas, F.M. and J.M. Ramírez-Hurtado, *Multidimensional scaling analysis: an alternative and complement to other multivariate techniques*. 2002: Department of Economics and Business, University Pablo de Olavide, Seville, Spain.
24. Kruskal, *Multidimensional scaling by opti-mizing goodness of fit to a nonmetric hypothesis.* Psychometrika, 1964a. **29**: p. pp. 1-27.
25. Kruskal, *Nonmetric multidimensional scaling: A numerical method.* Psychometrika, 1964b. **29**: p. pp. 115-129.
26. Linares, G., *Multidimensional Scaling Concepts and Approaches.* Operational Research Jornal, 2001. **22**(2).
27. López, o.M.M. and J.G. Herrero, *Data analysis techniques practical applications using Microsoft Excel and Weka*. 2006, University Carlos III: Madrid.
28. López-González, E. and R. Hidalgo-Sánchez, *No Metric Multidimensional Scaling. An example with R, using the algorithm SMACOF.* Education Studies, 2010. **18**: p. pp. 9-35.
29. O'Toole, A.J., et al., *Partially dis-tributed representations of objects and faces in ventral tempo-ral cortex.* Journal of Cognitive Neuroscience, 2005. **17**: p. pp. 580–590.
30. Torguerson, W.S., *Multidimensional scaling: Theory and method.* Psychometrika, 1952. **17**: p. pp. 401-419.
31. González, D.P., *Clustering algorithms based on density and clusters Validation*, in *Departament de Llenguatges I Sistemas Informátics*. 2010, Universitat Jaume I
32. Queen, M. and J. Some, *Methods for Classification and Analysis of Multivariate Observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967. p. pp. 281-297.
33. Hartigan, J. and M. Wong, *Algorithm AS136: A k-means clustering algorithm.* Applied Statistics, 1979. **28**: p. pp. 100-108.
34. Chen, C.W., J.B. Luo, and K.J. Parker, *Image Segmentation via Adaptive K-Means Clustering and Knowledge based Morphological Operations with Biomedical Operations.* IEEE Trans. Image Processing, 1998. **7**(12).
35. Sacaan, S. *Social networking and collective intelligence, IV Congress of CyberSociety*. 2009 [cited 2013 July 25]; Available from: http://www.cibersociedad.net/congres2009/es/coms/las-redes-sociales-y-la-inteligencia-colectiva-nuevas-oportunidades-de-participacion-ciudadana/879/.
36. Marteleto, M.R. and A. Braz *Redes e capital social: o enfoque da informação para o desenvolvimento local*. Ci. Inf., 2004. **33**.

First author:

Gustavo Rodríguez Bárcenas (M' 2015). Master of Science in Computer Systems for Education through the Metallurgical Mining Higher Institute of Cuba, Cuba. Master of Science in Information Science from the University of Havana, Cuba. Diploma of Advanced Studies in Scientific Documentation and Information at the University of Granada, Spain. Doctor of Science (PhD.) In Information Sciences from the University of Granada, Spain. He has been full professor at the Metallurgical Mining Higher Institute of Cuba. Currently Professor at the Cotopaxi Technical University, Ecuador. Professor of several undergraduate courses related to the specialty of Information and Computer Systems, Computer Aided Design and Data Transmission Networks. Professor of various postgraduate courses in specialties related to Computer Science and Energy Management Systems. He has published articles in high impact journals such as "Journal of American Society for Information Science and Technology", He has written several books, including lead author and co-author.