Applications

Multidimensional Data Analysis to assess interactions in an e-learning community

Giuseppe Giordano and Maria Prosperina Vitale

Dipartimento di Scienze Economiche e Statistiche
Università di Salerno

ggiordan@unisa.it; mvitale@unisa.it

Abstract

The quality of the cooperation and collaboration between members is one of the crucial factors in the development of an online learning community. In this paper we focus on the analysis of the quantity and type of interaction and cooperation between students in the asynchronous discussion forum of a virtual classroom. In order to describe both the qualitative and quantitative measures of the interrelationships in the net structure we propose to adopt the theoretical framework of Multidimensional Analysis of Textual Data in connection with the theoretical framework of Social Network Analysis. The tools made available by Correspondence Analysis of the lexical table are used to derive a semantic reference space in which to locate the nodes and arcs of the communication network. The underlying interrelation structure and the evolution of the conversational themes are shown by visualizing the students that share the same vocabulary and patterns of frequent lemmas used in the forum. The role of each student in the communication process is highlighted by suitable statistical indicators defined in the framework of Social Network Analysis.

Edizioni Erickson – Trento
www.erickson.it

Je-LKS Journal of e-Learning and Knowledge Society — *Vol. 2, n. 2, giugno 2006 (pp. 191-204)*

1. Introduction

The particular characteristics of e-learning activities lead to the definition of new performance indicators for evaluating how network technologies impact on teaching processes. The evaluation of an online course deals not only with contextual factors in student learning — such as types of resources available and changes in teaching and learning practices — but also the computer-mediated communication in the virtual classroom. The latter is made up of vertical (teacher/student) and horizontal (student-to-student) relationships. The quality of the cooperation and collaboration between members is one of the crucial factors in the development of an online learning community.

In this paper we focus on the analysis of the quantity and type of interaction and cooperation between students in the asynchronous discussion forum of a virtual classroom. The vocabulary used by the students in the forum constitutes our raw data.

The main goal is to identify indicators that can assess the quality and the quantity of the relationships between students, according to the following steps: a) recording of the messages database provided by the discussion forum; b) lemmatization of the vocabulary as a whole; c) definition of secondary data structures which allow to obtain graphical representations by means of Correspondence Analysis (Lebart, Morineau & Piron, 1997).

The factorial approach is profitably employed to evaluate the relationships, to underline the main themes in the conversations, and to distinguish «cooperative learning» in the language classroom from «informal social interactions». Furthermore, the importance of the relationships between interacting units will be assessed through structural measures and notions of Social Network Analysis.

In this paper, after a brief review of the assessment process and the different definitions of *community* in e-learning environments, we will describe the methods used to explore and analyse the interactions and collaboration in virtual communities. The proposed methodology will be exemplified by using the data from the online Statistics course at the University of Salerno. Finally, concluding comments will briefly indicate the relevance of virtual communities in the quality of e-learning experiences.

2. Assessment issues in the e-learning environment

Recently, the development of Web-based technologies and Computer Mediated Communication has led to noticeable transformations in the learning and teaching processes.

The assessment process of e-learning experiences requires different tools from the ones used for assessment in the traditional face-to-face learning environment, due to the fact that online courses require different learning styles, course contents,

technical support and technology knowledge (Khan, 2004). The interest in the e-learning evaluation process is highlighted by the numerous research projects that explore the use of students' questionnaires (Bangert, 2004; Thompson & MacDonald, 2005) or individualized virtual focus groups (Monolescu& Schifter, 2000) and indicators (Sonwalkar, 2002; Aymerich, Fenu, Masala & Poddi, 2005) as powerful research tools to assess online programs.

Furthermore, the assessment deals not only with contextual factors in student learning but also communication in the virtual classroom. In particular, the quality of the cooperation and collaboration between members in virtual learning communities is one of the crucial factors in the quality of an e-learning experience (Gatti & De Luca, 2005; Thompson & Macdonald, 2005). In the virtual learning community joint learning tasks and outcomes motivate a group of learners who come together for a period of time to engage in a common e-learning experience (Trentin, 2001; Johnson, 2001; Pudelko & Pudelko, 2003).

In order to better understand how virtual community enhances the quality of online courses in a university context, some researches evalueted the interactions in cooperative learning environments such as knowledge building through interactions in the asynchronous group discussion (Rovai & Barnum, 2003; Vonderwell, 2003; Tateo, 2004; Mazzoni, 2004), the nature of the interaction within a networked learning community (De Laat, Lally & Lipponen, 2004) and the new methodologies for analyzing participation, learning and interaction in the electronic forum (Temdee, Thipakorn, Sirinaovakul & Schelhowe, 2003). Some tests to assess graduate students' sense of classroom community have been also proposed (Rovai, 2002a; 2002b; Piave & Iadecola, 2005).

In the assessment process it has to be considered that several tools of distance education models are based on asynchronous learning networks. Students use computers and communications technologies to communicate with other learners without having to be online at the same time. These are very important in describing how students feel about being involved in the learning community and developing relationships with other members of the community. The quality and the intensity of communication tool use reflect the students' attitude towards sharing experiences, opinions, issues and observe to which level their educational goals are satisfied. In particular, the quality of interaction, through the information exchange, is a relevant factor that influences the building of community in e-learning environments, increases participation, collaborative thinking and promotes social construction of knowledge.

3. A Strategy for evaluating the virtual learning community

In this paper we explore the interaction patterns in online courses focusing on the quality and quantity of textual communication flows between students. Our

strategy combines factorial analysis of the textual data with Graph Theory and the assessment tools of Social Network Analysis in order to analyze participation, learning and interaction in an asynchronous discussion forum. The main steps of the strategy are shown in figure 1.

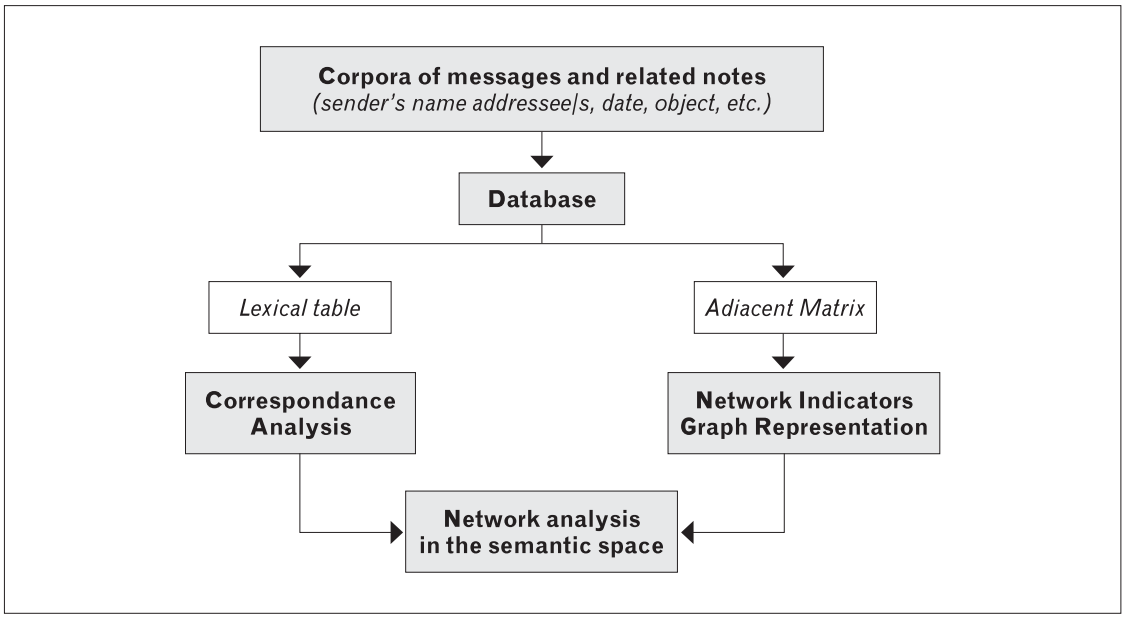

Figure 1 The steps of the strategy.

The first step consists of statistical data analysis based on the corpora of the messages sent by the students. The goal is to extract and emphasize peculiar schemes of dialogs, highlighting the vocabulary used and distinguishing between formal and informal language. This can be profitably employed to define collaborative behaviours, sociability, informal interactions, etc. The use of factorial methods in textual data analysis is extremely useful in the exploration of the main behavioural patterns between the actors of the community. The analysis outcomes provide information about actor prestige, the centrality of a theme and, most importantly, a spatial representation of actors and themes according to the underlying metric of Correspondence Analysis (as described in section 3.1).

In the second step we assess the strength of the interrelationship in the virtual community. We introduce the traditional concepts of Social Network Analysis and its usual graphical representation through Graph Theory (described in section 3.2). In this context we are able to use some statistical indicators such as: *Popularity, Participation* and *Density* of the net. Accordingly, graph theory defines the equivalent measures based on the definitions of *Mean Degree, In-degree, Out-degree, Density, Reachability,* etc. (for more thorough definitions see Wasserman & Faust, 1999).

The third step involves a graphical representation, the factorial map, derived from the Correspondence Analysis of the lexical table as the reference space in which to map the connectivity of the underlying communication network by means of Graph Theory. In this way we describe both qualitative and quantitative measures of the net interrelationships analysed by the two different theoretical frameworks.

3.1 The Multidimensional Analysis of Textual Data

The multidimensional analysis of textual data is typically carried out through several steps. In the first phase, the corpora (the bodies of the e-mails) are selected and labelled according to different criteria (e.g. *the sender's name, the addressee/s, the date, the object, etc.*), then text segments are created, lemmatised and disambiguated; finally, the whole vocabulary is obtained.

The main statistical information is given by the occurrence of each distinct lexical form. This distribution can be further classified according to each of the previous selected criteria. For instance, we are interested in cross-classifying the vocabulary with the message sender (i.e. each student). In this way a large contingency table (lexical table) is built and the statistical analysis involves a factorial decomposition and graphical representation of the most meaningful factors. In the framework of multidimensional statistical data analysis, Correspondence Analysis (Greenacre, 1984) of the lexical table is carried out (Lebart, Salem & Berry, 1998).

The resulting factorial plan can be described as a map where each student and each lemma are represented as points on the plan (figure 2). The distance between each pair of student-points takes into account the similarity of the used vocabulary. Students near the axes' origin are those who share the more common language (the more frequent lemmas also lie in proximity of the axes' origin). Whereas, student-

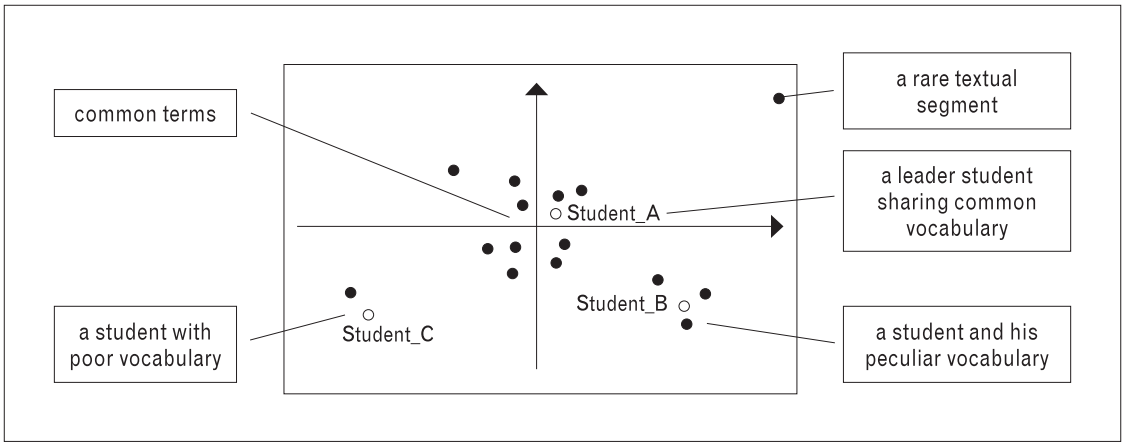

Figure 2 Graphical Representation and Interpretation of Correspondence Analysis on Textual Data.

points which tend to spread out on the plan are characterized by an idiosyncratic vocabulary and tend to use rare lemmas.

The distance between the points on the factorial plan is a function of the statistical association between the actors and their vocabulary, so the geometric distance between student-points is explicative of the personal and active role in the community net.

This factorial plan makes it possible to discover the association between textual units and message senders. According to the interpretative rule of Correspondence Analysis we are able to distinguish:

a) the commonly used expressions versus the rare terms;
b) usual and unusual patterns of textual units characterising one or more students, as well as particular themes (technical problems, colloquial forms, jargons, etc.);
c) students who are isolated on the plan as they are characterised by their own particular language;
d) students who cluster together according to a common used vocabulary.

3.2 The role of Social Network Analysis

In order to describe the interactions between the online students in a forum and how they share knowledge, we exploit the tools of Social Network Analysis and Graph Theory. This phase comprises several steps. Firstly, we indicate each student with a *node* and each relation, established by sending e-mails, with an *arc*. Let us note that we consider a *directed graph* (without *loops*) and an *incident arc* to be directed from the sender to the addressee. Moreover, the directed graph can be weighted by the number of messages sent or received; such an *integer weighted directed graph* is explicitly referred to as a *Network* by some Authors (see Harary, 1969; Roberts, 1976).

The resulting data structure is arranged in a matrix ($N \times N$) whose rows represent the sending nodes and the columns the addressee nodes.

According to Social Network theory, some statistical indicators are defined, for example the *popularity, participation* and *density* of the net. Degree analysis — which measures the direct connections between nodes in the network — is carried out by considering the nodal *In-Degree* and *Out-Degree* as the number of lines that are incident to a node and from a node, respectively. Furthermore, the nodes are defined as *Isolate* when the degree is equal to 0; as *Pendant* nodes when the degree is equal to 1 and finally, the measure of *Inclusiveness* is defined as the percentage of the number of nodes minus the number of isolates divided by the total number of nodes.

The graphical representation of a net by a graph (also called *sociogram*) can be enhanced by looking for a suitable metric that allows the n nodes in a map to be

located according to pairwise proximity criteria. In order to study the underlying structure of a social network, Factorial Analysis of sociometric data is in widespread use in the context of Social Network Analysis.

3.3 Network Analysis in the Semantic Space

The first two steps of the strategy defined let us consider different features of the same framework. These two facets correspond to information about the semantic richness and the density of the links defining the net topology; the common framework is communication in the e-learning environment.

In the third step we bring this information together in order to provide a common reference space in which to visualize and analyse the qualitative and quantitative nature of the communication features.

Different spatial representations are used to visualise the proximities of the actors in the net. Multidimensional Scaling techniques, Factor Analysis and Cluster Analysis are suitable to this aim (Kruskal & Wish, 1978; Aldenderfer & Blashfield, 1984; Weller & Romney, 1990). These techniques are based on different data structures which are mainly related to the similarity or dissimilarity indices; for example, a proximity index can be provided by the adjacency matrix.

What we propose comes from the steps of the analysis procedure in figure 1. The first step provides a factorial representation of net actors which takes into account the information of this procedure recorded in the lexical table and describes the communication style of the students. In other words, the proximity of students in such a map does not depend on the network structure as defined in step 2. Then, by superimposing the net structure onto the factorial map, we can simultaneously analyse the actors' connectivity along with the metric induced by the semantic space. In this way, we are able to analyze both the communication flows and their contents.

4. A case study: the assessment of the online Statistics course at the University of Salerno

The above illustrated strategy is applied in order to assess the interrelationships between online students in an asynchronous discussion forum. In this study the statistical units are the «messages» posted by each of the 38 students enrolled in the Statistics online course of the Undergraduate Programme in Social Sciences at the University of Salerno. The time period under study ranges from October 2004 to January 2005. A total of 742 *learner-learner* messages are analyzed in order to highlight the main themes and study the interactions in the text-based asynchronous discussion group. For each message the *corpus*, the *sender*, the *addressee* and the *date of sending* are recorded.

The corpora are pre-processed according to the usual lexicometric approach through the lemmatisation phase and the choice of repeated segments as textual units. The lexical table is the contingency table combining the segments and the message senders, its generic cell represents the occurrence of textual segments for each student. The statistical association structure of the lexical table is studied using Correspondence Analysis.

The graphical representations obtained by Correspondence Analysis reveal the underlying interrelation structure and the evolution of the conversational themes by visualizing students that share the same vocabulary (figure 3) and patterns of frequent lemmas used in the forum (figure 4).

In particular, in figure 3, the students using a common vocabulary are located in the neighbourhood of the axes' origin (*Stud_13; Stud_31; Stud_19; Stud_9; Stud_6; Stud_22*); whereas the decentralized student-points are characterised by a peculiar vocabulary due to atypical discussion themes (*Stud_21; Stud_25; Stud_29; Stud_17; Stud_20*).

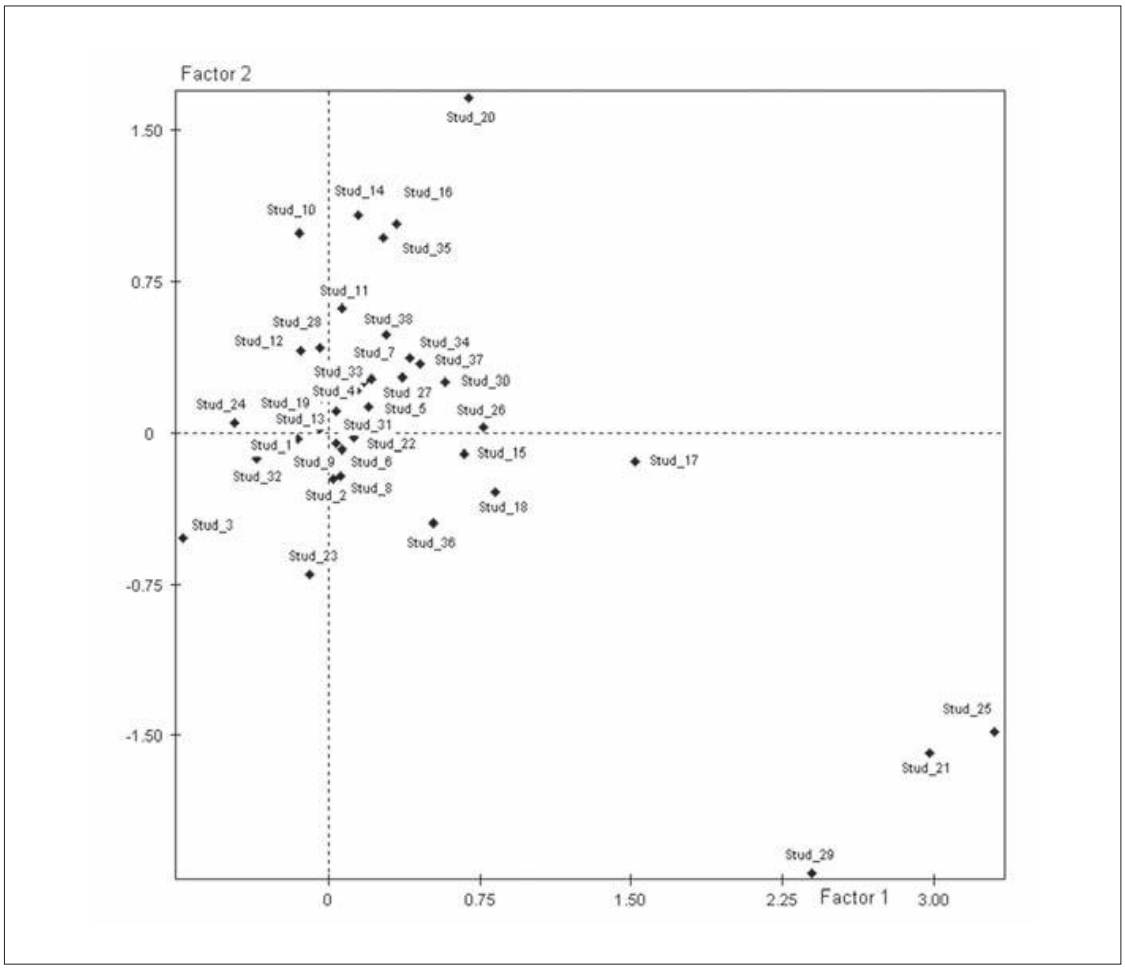

Figure 3 First factorial plan of Correspondence Analysis: the students' representation.

The patterns described by textual segments are shown in figure 4. The factorial plan points out different themes and interaction styles that can be categorised as:

- *Start-up Collaboration*
  This theme is represented by the messages sent by students in the first few days of the course. It deals with technical issues about the software and the platform tools. The textual segments are located on the bottom left of the factorial plan.
- *Cooperative Learning*
  This is the dominant theme of the forum. It is based on messages about *Statistical* topics in order to give details on theoretical concepts. The textual segments lie all around the origin of the coordinate axes and this vocabulary is shared by the majority of students.
- *Exam Communication*
  Several students share information about procedures and dates of the intermediate and final tests. The textual segments are found on the right of the first axis.
- *Informal Conversation*
  This is represented by few students (*Stud_21; Stud_25; Stud_29*) located on the right of the first axis. They exchange greetings during the Christmas holidays and are characterized by a particular vocabulary.

The role of each student in the communication process is highlighted by graphs and statistical indicators defined in the framework of Social Network Analysis (Table 1). The analyzed network is composed of 38 nodes, with a total number of 230 links. The network density is 0,16 (i.e. active links represent 16% of all the potential links). In our analysis some students (*Stud_1, Stud_2, Stud_3, Stud_5, Stud_6*) possess high *in-degree* and *out-degree* measures, stressing their centrality role in the network.

Table 1
STATISTICAL INDICES OF THE COMMUNICATION NETWORK

	IN-DEGREE	OUT-DEGREE
Sum	230	230
Mean	5,897	5,897
Std.Dev.	6,617	5,62
Min.	0	0
Max.	33	22
# of Isolate	3	1
# of Pendant	10	11
Inclusiveness (%)	92,31	97,44

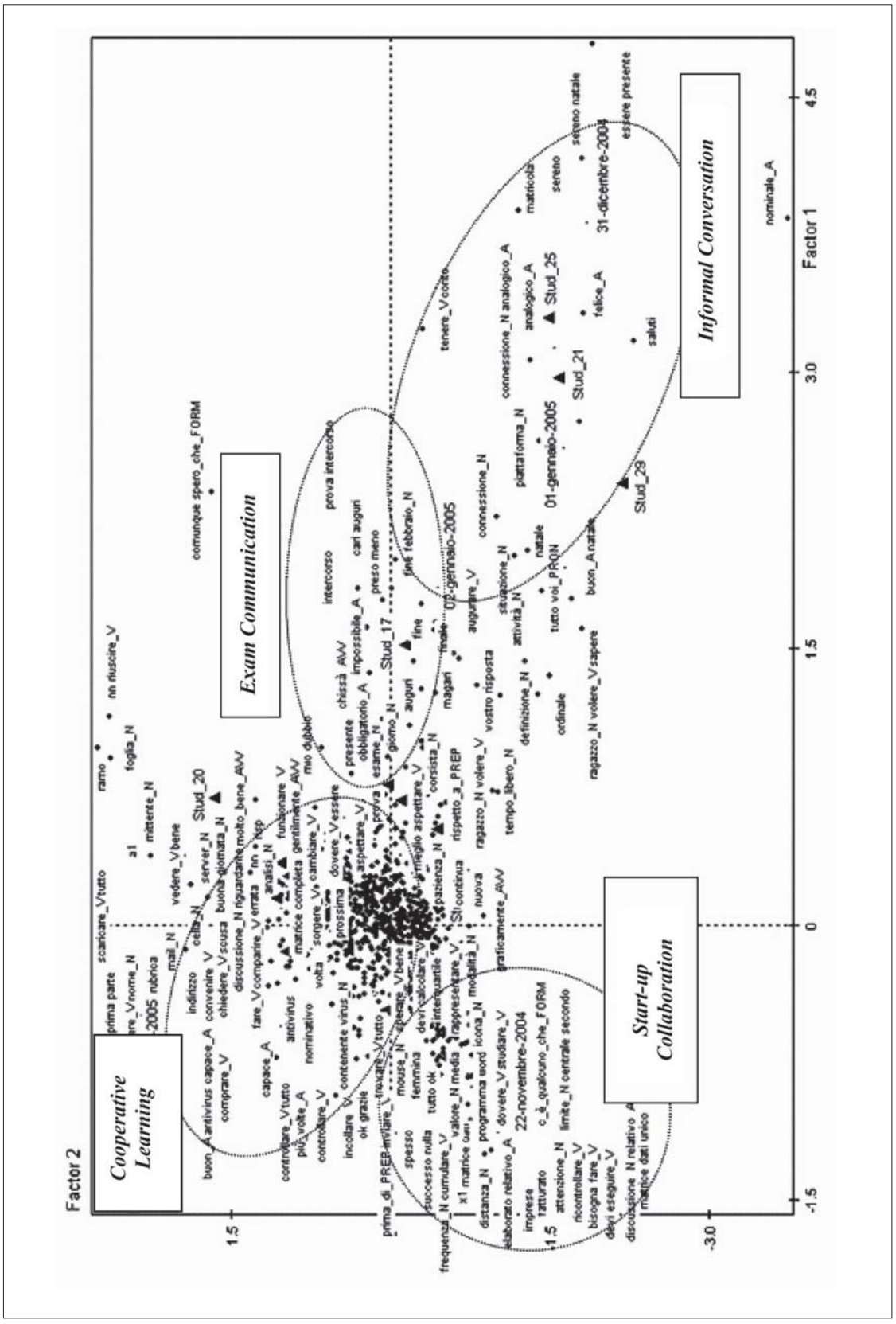

Figure 4 First factorial plan of Correspondence Analysis: the textual segments representation.

Moreover, the graph in figure 5 visualises the links between students. Each straight-line represents a communication flow between two students (both incoming and outgoing) with different weights given by the number of posted messages. The relative position of each node is derived from the association measure used in Correspondence Analysis as a function of the common vocabulary shared by the students. In this representation, the semantic space of Correspondence Analysis is used to visualize the underlying communication network.

Finally, the association between the collaborative attitude and the final examination scores reported by students has been analysed. In Table 2 the average and the standard deviation of final scores are given for each of the groups described in the previous steps. On the one hand, let us note that the «Cooperative Learning» group takes a higher average score than the other groups. On the other hand, the isolated students in the «Informal Conversation» and «Exam Communication» groups (*Stud_21; Stud_25; Stud_29; Stud_17; Stud_20*) do not pass the final exam at all. Even if the experiment has not been designed for a comparison with a control group, our experience let us affirm that students who attend a traditional course show worst performance in the final examinations.

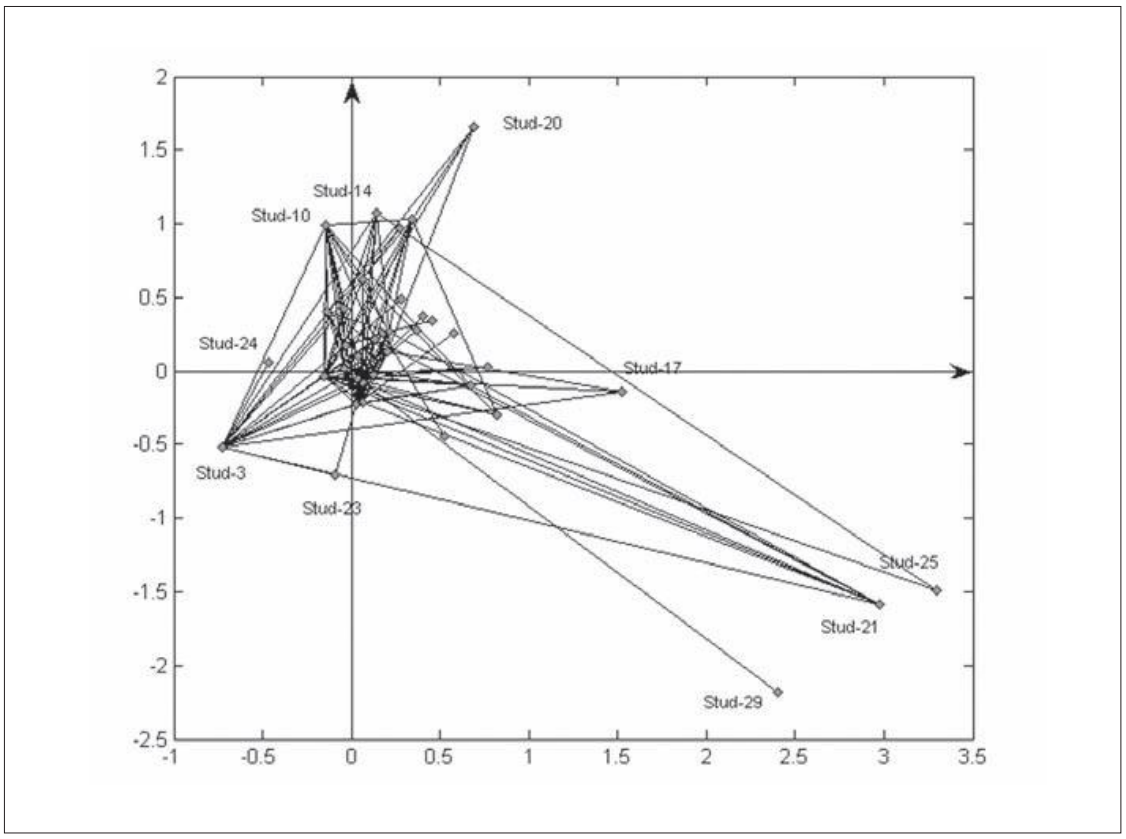

Figure 5 First factorial plan of Correspondence Analysis: the student representation.

Table 2

THE AVERAGE AND THE STANDARD DEVIATION OF FINAL SCORES FOR EACH OF THE GROUPS OF STUDENTS

Groups	*Average final scores*	*Standard Deviation final scores*
Cooperative Learning	25,63	2,74
Start-up Collaboration	23,56	2,83
Informal Conversation and Exam Communication	22,00	2,68

5. Concluding remarks

The evaluation of communication flows represents one of the fundamental issues in the e-learning processes. The study of virtual community characteristics is applied in order to enhance the traditional evaluation tools which are based on the assessment of the students' satisfaction through questionnaires and web access statistics.

In this paper, we have described a strategy that uses Multidimensional Textual Data Analysis in order to define different conversational themes. The factorial graphical representations have made it possible to highlight the cooperative behaviour of the students and the use of informal language for some of them. Furthermore, in order to describe the students' role in the communication process we have combined the factorial maps with Graph Theory and the related statistical indicators of Social Network Analysis.

Acknowledgement

The authors have been supported by the research program 2005 «Multivariate methods for evaluation in e-learning processes», coordinated by Prof. M.R. D'Esposito – Department of Statistics and Economic Sciences, University of Salerno – Italy.

BIBLIOGRAPHY

Aldenderfer M.S. & Blashfield R.K. (1984), *Cluster Analysis,* Newbury Park, CA, Sage.

Aymerich F.M., Fenu G., Masala V. & Poddi, G. (2005), *Definizione di macroindicatori per la valutazione comparativa di modelli di erogazione eLearning.* Proceedings of the Conference: Expo e-learning, 2005, Ferrara.

Balbi S. & Giordano G. (2001), *A Factorial Technique for Analysing Textual Data with External Information,* in: S. Borra, R.Rocci, M. Vichi & M. Schader (eds), *Advances in Classification and Data Analysis*, Berlin, Springer-Verlag.

Bangert A.W. (2004), *The Seven Principles of Good Practice: A framework for evaluating on-line teaching,* «Internet and Higher Education», 7: 217-232.

Calvani A. & Rotta M. (2000), *Fare formazione in Internet. Manuale di didattica online,* Trento, Erickson.

De Laat M., Lally V. & Lipponen L. (2004), *Patterns of Interaction in a Networked Learning Community,* Proceedings of the World Conference on E-Learning in Corp., Govt., «Health., & Higher», 2004(1): 1846-1853.

Gatti F. & De Luca P. (2005), *Valutazione dell'Interazione Online nelle Comunità Virtuali di Apprendimento.* Proceedings of the Conference: Expo e-learning 2005, Ferrara.

Greenacre M. (1984), *Theory and Applications of Correspondence Analysis,* New York, Academic Press.

Harary F. (1969), *Graph theory*, Addison-Wesley series in mathematics, Massachusetts, USA.

Johnson C.M. (2001), *A survey of current research on online communities of practice, «*The Internet and Higher Education» 4(1): 45-60.

Khan B.H. (2004), *E-learning: progettazione e gestione*, Trento, Erickson.

Kruskal J.B. & Wish M. (1978), *Multidimensional scaling*. Newbury Park, CA, Sage.

Lebart L., Salem A. & Berry L. (1998), *Exploring Textual Data*, Dordrecht, Kluwer Academic Pulisher.

Lebart L., Morineau A. & Piron M. (1997), *Statistique Explorative Multidimensionnelle*, Paris, Dunod.

Mazzoni E. (updated 2004) *Strumenti per un approccio quantitativo allo studio delle interazioni. Il software Net Miner e i Log File*, [documento WWW] URL: http://formare.erickson.it/archivio/maggio_04/5mazzoni.html accessed on 1th October 2005.

Monolescu D. & Schifter C. (2000), *Online Focus Group: A Tool to Evaluate Online Students' Course Experience, «*Internet and Higher Education» 2(2-3): 171-176.

Piave N.A. & Iadecola G (2005), *Virtual Classroom Learning Evaluation Tool: a model and a test useful in measuring online course' effectiveness,* Proceedings of the II Sie-l Conference, Firenze.

Pudelko F.H. & Pudelko B. (2003), *Understanding and analysing activity and learning in virtual communities, «*Journal of Computer Assisted Learning», 19: 474-487.

Roberts F.S. (1976), *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems,* Englewood Cliffs, NJ, Prentice-Hall.

Rovai A.P. (2002a), *Development of an instrument to measure classroom community,* «Internet and Higher Education», 5: 197-211.

Rovai, A.P. (2002b), *Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks,* «Internet and Higher Education», 5: 319-332.

Rovai, A.P., Barnum, K.T. (2003), *On-Line Course Effectiveness: An Analysis of Student Interactions and Perceptions of Learning,* «Journal of Distance Education», 18(1): 57-73.

Sonwalkar N. (updated 2002), *A New Methodology for Evaluation: The Pedagogical Rating of Online Courses,* [documento WWW] URL: <http://www.syllabus.com/article.asp?id=5914> accessed on 2th November 2005.

Tateo L. (updated 2004), *Struttura delle relazioni e contenuto argomentativo dei messaggi nella comunicazione mediata da computer*, [documento WWW] URL: <http://formare.erickson.it/archivio/maggio_04/6tateo.html> accessed on 1th October 2005.

Temdee P., Thipakorn B., Sirinaovakul B. & Schelhowe H. (2003), *Of Collaborative Learning: An Agent Based Approach for Social Network Analysis,* Proceedings of the World Conference on E-Learning in Corp., Govt., «Health., & Higher Ed.», 2003, (1):1786-1789.

Thompson T.L. & MacDonald C.J. (2005), *Community building, emergent design and expecting the unexpected: Creating a quality eLearning experience,* «Internet and Higher Education», 8: 233-249.

Trentin G. (2001), *From formal training to communities of practice via network-based learning,* «Educational Technology», 41(2): 5-14.

Vonderwell S. (2003), An examination of asynchronous communication experiences and perspectives of students in an online course: a case study, *Internet and Higher Education* 6: 77-90.

Wasserman S. & Faust K. (1994), *Social network analysis: methods and applications,* Cambridge, Cambridge University Press.

Weller S.C. & Romney A.K. (1990), *Metric Scaling: Correspondence Analysis*, Newbury Park, CA, Sage.