Using data mining and social networks to analyze the structure and content of educative on-line communities

Conference Paper · January 2005

CITATIONS
8

READS
175

3 authors:

Robert Rallo
Pacific Northwest National Laboratory
98 PUBLICATIONS **2,682** CITATIONS
SEE PROFILE

Jesus Salinas
University of the Balearic Islands
538 PUBLICATIONS **4,468** CITATIONS
SEE PROFILE

Mercè Gisbert
Universitat Rovira i Virgili
270 PUBLICATIONS **2,064** CITATIONS
SEE PROFILE

USING DATA MINING AND SOCIAL NETWORKS TO ANALYZE THE STRUCTURE AND CONTENT OF EDUCATIVE ON-LINE COMMUNITIES

R. Rallo, M. Gisbert[#] and J. Salinas[§]

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Catalunya, SPAIN.

[#] Departament de Pedagogia, Universitat Rovira i Virgili, Tarragona, Catalunya, SPAIN

[§] Departament de Ciències de l'Educació, Universitat de les Illes Balears, Illes Balears, SPAIN

1. Introduction

The use of Computer Mediated Communication (CMC) systems generates the development of virtual communities. The analysis of the structure of these communities is a challenging task. The e-mail interactions can be used as an indicator of social relationships between community members. Based upon these interactions, a social network could be developed to model the dynamics and structure of the virtual community.

2. Objectives

The main goal of the present work is to extract all the "hidden information" behind the e-mail based information exchange produced in on-line communities. This goal is achieved through:

- The development of a data mining process to extract social information from a discussion list.
- The characterization of the structure and dynamics of the community using social network analysis.
- The assessment of the proposed methodology using a well-known discussion list such as EDUTEC-L.

3. Methodology

3.1. Extraction of social relationships indicators.

The main problem when dealing with e-mail lists is the lack of information about the recipient of the message due to the fact that all messages are always directed to the whole list. In this work two different approaches are used:

- Direct Interactions: can be detected using the *In-Reply-To* field of the e-mail header. The main drawback of this technique is its dependency on the use of the "reply" option in the e-mail client.
- Threaded interactions: Two subscriptors are tied by a social relation if they participate in the same threads of discussion. The main problem with this approach is the automatic detection of threads.

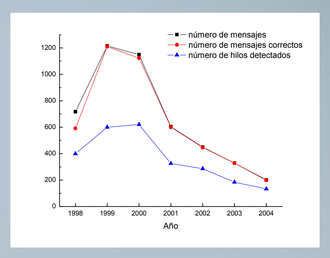
3.2. Analysis of the list.

- Annual. Independent, year by year analysis.
- Accumulated interactions. All the previous years are considered in the model.

4. Case Study: EDUTEC-L and its associated virtual community

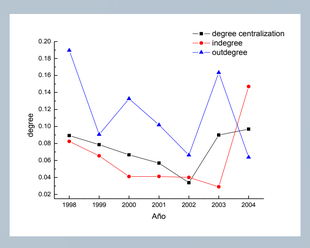
EDUTEC-L is a discussion list initiated in 1996 and hosted by RedIris (the Spanish NREN).

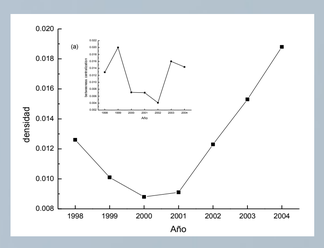
The data used in this work covers a 7 years period: 1998-2004.

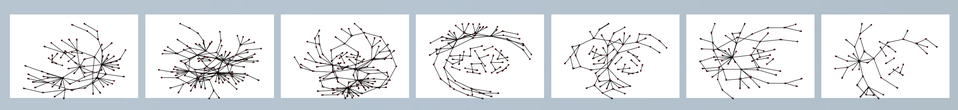


5. The network of "direct interactions" (In-Reply-To)

5.1. Annual evolution.

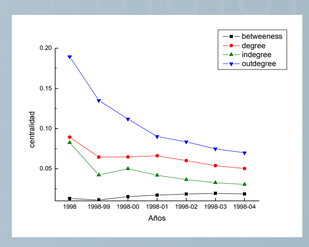


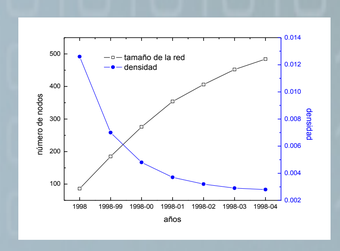




1998 1999 2000 2001 2002 2003 2004

5.2. Accumulated evolution (1998 – 2004).

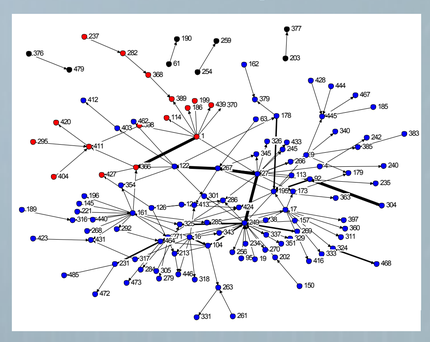




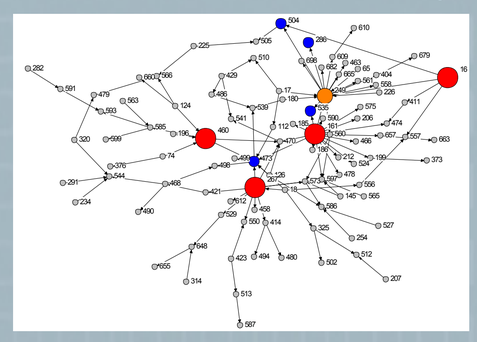
1998 1998-1999 1998-2000 1998-2001 1998-2002 1998-2003 1998-2004

5.3. Community substructure and user roles.

The Newman-Girvan algorithm reveals the presence of some structure into the community.

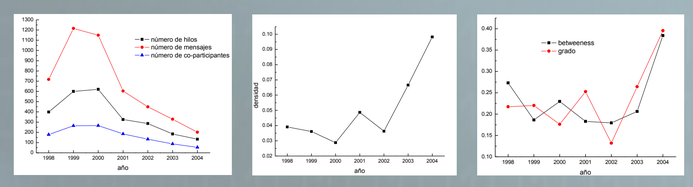


1999. The analysis shows the existence of three communities of users. It is important to note that the connection between the main communities (blue and red) is done through some specific nodes.

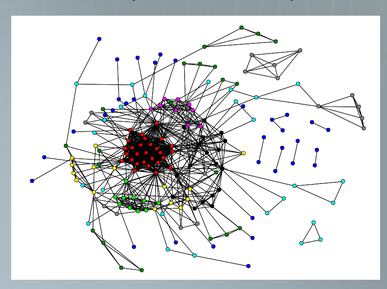


2000. The hubs and authorities analysis of the network confirms the existence of different user roles (i.e. list moderators, facilitators, …). The size of the nodes is proportional to its centralization.

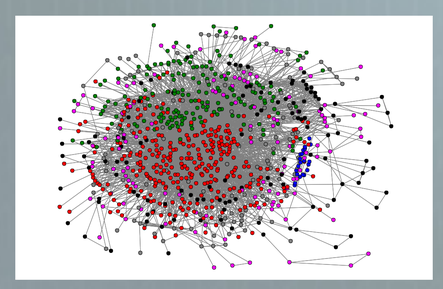
6. The network of Threaded co-participation



The structure of the social network corresponding to threaded interactions is more complex than in the previous case.



2001. Main communities detected



1998-2004. Main communities detected

7. Conclusions

Social network analysis is a powerful tool to extract knowledge from on-line communities.

Certain community members acquire relevant roles that evolve as the community develops. These roles could be described using hubs & authorities or cut-points analysis.

There exist some structure into the community. The Newman-Girvan algorithm or the faction analysis allows the discovery of these sub-communities and the identification of structurally important nodes.

Contact information:

- Robert Rallo: robert.rallo@urv.net
- Mercè Gisbert: merce.gisbert@urv.net
- Jesús Salinas: jesus.salinas@uib.es