Topic trend detection in newsgroups

Bettina Hoser, Jan Schröder, Andreas Geyer-Schulz,
Maximilian Viermetz, Michal Skubacz

In this paper we report on a project for Corporate Technology of Siemens AG, Munich, which aimed at the use of a social network analysis approach for trend and trend shift detection over time in a technically oriented newsgroup on mobile phones. The analysis was based on the assumption that a trend shift is relevant only if relevant (central) members of the newsgroup initiated this shift. A shift could occur in one of two ways. Either within a subgroup the topic shifted, e.g. other words were used, or the relevance of a group within the newsgroup shifted, and the now more relevant subgroup discussed a different topic. We used eigensystem analysis as a method and could show that as groups shifted over time, so did topics.

1 Introduction and Data

In a world of short product life cycles, especially in the telecommunication and mobile phone markets, anticipation of trends is of utmost importance for companies that want to survive in this highly contested markets. The anticipation of trends has been based on marketing research efforts alone in the past. In the age of the Internet users share their information, perceptions and thoughts with other like-minded users for example in newsgroups. Methods based on social network analysis can be used to evaluate this freely available and information rich communication to observe what users feel and think about products, techniques and or services. Forecasting trends from these communications and developing the right products and services at the right time could bring a crucial advantage to any company capable of performing this kind of analysis.

To achieve this goal, not only does one have to take a look at what users write about, but who is writing. A topic may not be worthwile if the author has no influence within the group. But if a highly esteemed author in a group brings up a new topic he claims as relevant it may be of relevance to the investigating company. Thus it does not suffice to look at words or phrases used in the context of the newsgroup, but one also needs to incorporate the knowledge about the social network and its structure underlying this newsgroup.

In this project a joint analysis of the social network and the topic structure was developed and tested. The objective of Corporate Technology was to find a way to incorporate the knowledge of the social structure of a group of mobile phone users with the knowledge about the topics about which they communicate. The data set which was collected and preprocessed by two of the authors is from a newsgroup on mobile phones. The preprocessed data set was imported into a SQL data base specifically adapted for the further analysis of this study. Overall the data set consists of 2808 messages written by 709 authors during the observation period from July 2005 to October 2005.

The rest of this paper is organized as follows. In the next section we will give a very brief introduction into the analysis of Eigensystems of graphs using complex Hermitian adjacency matrices. Section 3 will show the results of a time dependent analysis of the author-to-author network as well as the authors-use-of-words network. The final section will round off this paper with a conclusion and an outlook.

2 Method

One of the standard indices of social network analysis is the eigenvector centrality index. It defines a node in a graph or network as central if it is connected to other central nodes. Thus this is a recursive approach. Mathematically it is described by a system of linear equations, which can be solved by eigensystem analysis. For further reading we refer to Wasserman and Faust [6], Katz [5] and Bonacich [1].

Based on that standard approach we use the complex eigenvector centrality as defined by Hoser and Geyer-Schulz [3]. It is based on a complex hermitian adjacency matrix H that is derived from the real valued weighted adjacency matrix A of the underlying network by Eq.(1).

$$H = (A + iA^t)e^{-\frac{i\pi}{4}} \quad (1)$$

where $i^2 = -1$ is the imaginary unit and A^t is the transpose of A. Thus directional information about the communication is kept. Hermitian matrices have a complete orthogonal eigensystem, all eigenvalues are real, the eigenvector components may be complex valued. The set of all eigenvalues is called spectrum.

The complex valued eigenvector components are interpreted as a rank prestige index of each subgroup member, following the interpretation of the real valued eigenvector centrality index described for example in [6]. In addition, each member has for each subgroup structure / eigenvector in the spectral representation a different rank prestige index. This index depends on his relation to the respective anchor of the subgroup. The eigenvectors themselves represent different patterns of communication.

The interpretation of a complex eigensystem can be based on two extreme examples (based on an adjacency matrix with $h_{kk} = 0$). The first is a star graph. The spectrum is symmetric in the sense that the non-zero eigenvalues come in pairs of $+\lambda_k, -\lambda_k$, with k defining the number of star centers, and the corresponding eigenvectors show a phase shift of π between components that are in contact with each other, e.g. belong to the same star. The other extreme is the complete graph. Here the spectrum is organized