An Exploratory Study into Automated Real-Time Categorisation of Engineering E-Mail

James A. Gopsill
Dept. of Mechanical Engineering
University of Bath
Bath, United Kingdom

Stephen J. Payne
Dept. of Computer Science
University of Bath
Bath, United Kingdom

Ben J. Hicks
Dept. of Mechanical Engineering
University of Bristol
Bristol, United Kingdom

***Abstract*—For large, spatially and temporally distributed engineering projects, e-mail is a central means for the discussion of engineering work and sharing of digital assets that define the product and its production process. The importance of communication and the value of its content for resolving issues *post facto* are universally accepted. More recently, the potential value of its content to predict events, issues and states *a priori* has been explored with some success. However, while in the former context (*post facto*) trends and patterns can be established through iteration and refinement over time; for prediction, heuristics need to be established in advance and closer to real-time analysis becomes necessary due to the critical and very often short timescales. It is this challenge of making predictions from the content of e-mail that is considered in this paper. In particular, the paper deals with engineering e-mail and the ability to automatically predict its purpose from its content rather than relying solely on the subject line.**

The work builds upon previous studies by the authors concerning the characterisation of the content of e-mail: what they are about, why they were sent and how the content is expressed. The paper summarises the previous work and looks at the potential of identifying the purpose of e-mail through the use of Naive Bayes and an adapted Latent Semantic Analysis ap- proach. While the techniques have only been applied to an initial exploratory study of 98 e-mails, the results suggest the potential for automated real-time categorisation of engineering e-mails through achieving an accuracy of 66%. Such a capability would both support prioritisation of e-mail for engineers and macro level characterisation of project e-mail dynamics. The latter provides the opportunity for real-time analysis of an engineering projects status and correspondingly, modes of management intervention.

Keywords—Engineering Communication, E-Mail, Naive Bayes, Latent Semantic Analysis.

I. INTRODUCTION

It is self-evident that e-mail has become a central means for the discussion of engineering work and sharing of digital assets[1] that define the product and its production process [1]. This is especially the case when teams become larger, increasingly multi-disciplinary and more distributed both spatially and temporally [2]. Delinchant et al. [3] argues that the prominence of e-mail is due to engineering companies offering support for the communication tool and its ubiquity across the engineering domain.

[1]Examples include: reports, calculations, photographs and results from simulations.

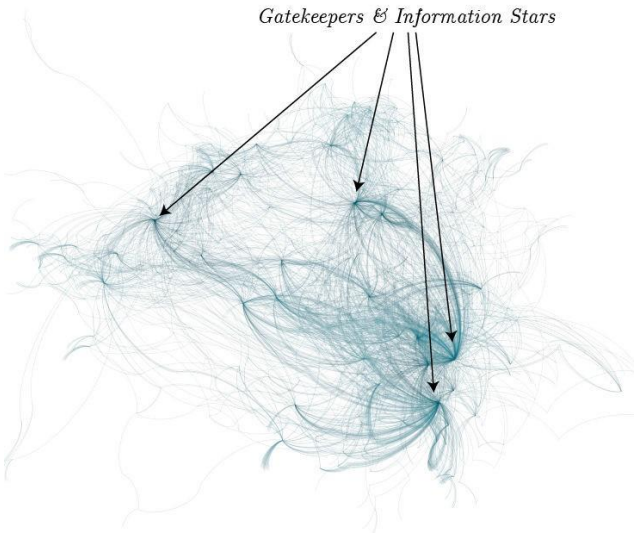

Fig. 1. Visualisation of an Engineering E-Mail Network

Engineering communication research has shown that the volume of communication is indicative of progress being made within an engineering project [4], [5]. In addition, Dong [6] reveals that almost all successful design teams have high-levels of communication as this helps maintain a shared understanding between the engineers. Although it may seem a positive step to encourage increased communication between engineers, there are a number of limitations of e-mail that need to be addressed both from Personal Information Management and Project Management perspectives.

A. Personal Information Management

Engineers typically have to send e-mails through a hierarchy of personnel before being able to reach the *right* people to share knowledge with [7], [8]. Engineering projects also contain key figurehead/expert engineers who are the *'go to people'* within the project. These are often referred to in the literature as *gatekeepers* or information *stars* as they fill one of two roles; 1) to know 'who knows' and therefore direct engineers to the relevant expert or 2) are experts in a particular field themselves [9], [10]. If one were to visualise the network of communication within an engineering project, it would appear similar to the visualisation in Figure 1, which highlights four *gatekeepers* from a project involving approximately 670 people