MINING ORGANIZATIONAL EMAILS FOR SOCIAL NETWORKS WITH APPLICATION TO ENRON CORPUS

By

Yingjie Zhou

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Decision Sciences & Engineering Systems

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Dr. William A. Wallace, Thesis Adviser

Dr. Wai Kin (Victor) Chan, Member

Dr. Mark Goldberg, Member

Dr. Malik Magdon-Ismail, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2008
(For Graduation August 2008)

ABSTRACT

With the ever increasing amount of electronic data available, the need for tools and techniques to clean and analyze massive data sets has grown. Email plays an important role in communications. In this research, the efforts are concentrated on inter organizational emails, i.e., emails among employees in an organization. The goal of this research is to develop a series of algorithmic methods to clean organizational emails and analyze the social networks constructed from the organizational emails.

Like other forms of data, email data can be noisy and need to be cleaned before any analysis is conducted. However, few studies have investigated the cleaning of archived organizational emails. In this study, systematic cleaning procedures and methods are provided and implemented for Enron email data. A strategy is then summarized and proposed to serve as a guide in cleaning any archived organizational email data set. Algorithms for name disambiguation and misdirected email detection are developed for further cleaning the email data set. A generative model is developed to determine the informativeness of a message using the conversation threads it generates.

Since people use emails to exchange thoughts, opinions, and feelings, emails represent certain relations among them. A social network is constructed from the emails with nodes representing individuals and links representing emails. Statistical and Social Network Analysis methods, including factor analysis, Singular Value Decomposition, degree distribution, node centrality, disjoint clustering technique, and overlapping clustering algorithm, are applied to study the network from various perspectives. The results are provided, interpreted, and compared. Finally the text analysis for values in organizational emails is implemented to test two proposed hypotheses.