Data Mining Acadeic Emails to Model Employee Behaviors and Analyze Organizaitonal Structure

Kayla M. Straub

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical Engineering

Robert W. McGwier, Chair

Aloysius A. Beex

Michael Buehrer

Bert Huang

March 25, 2016

Blacksburg, Virginia

Keywords:

Data Mining Academic Emails to Model Employee Behaviors and Analyze Organizational Structure

Kayla M. Straub

(ABSTRACT)

Email correspondence has become the predominant method of communication for businesses. If not for the inherent privacy concerns, this electronically searchable data could be used to better understand how employees interact. For example, after the Enron dataset was made available, researchers were able to provide great insight into employee behaviors based on the available data despite the many challenges with that dataset. The work in this paper demonstrates the application of a suite of methods to an appropriately anonymized email dataset created from volunteers' email metadata. This new dataset, from an internal email server, is first used to validate machine learning and feature extraction algorithms and then to generate insight into the interactions within the center. Based solely on email metadata, a random forest approach modeled behavior patterns and accurately classified employees by job title. The algorithm performed very well not only on participants in the study but also on other members of the center who were connected to participants through email. Furthermore, the data revealed relationships not present in the formal operating structure. The result is a much fuller understanding of the center's internal structure than can be found in the official organization chart.

Dedication

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Build up context for the report. Describe the problem and challenges associated with it. Why is this important? Brifely cover the process/solution I have implemented. Touch on potential applications of this work.

Chapter 2

Literature Review

Cover background works in similar studies, other email analyses [mostly Enron], and graph theory fundamentals. Highlight how few studies have been performed before in this domain as well as their lack of quantifiable results. Show the limitations of the Enron dataset.

Chapter 3

Data Collection

Discuss the need for a modern, clean dataset. Intimate knowledge of the environment that produced this data is to our advantage. Describe the steps taken to collect, anonymize, and store the data properly. Provide detailed descriptions/statistics of the dataset.

Subchapter: Enron issues Subchapter: Collection process Subchapter: Dataset statistics

Chapter 4

Feature Analysis

Go into depth explaining all features used. Cover the math behind all graph-based metrics.

Subchapter: Graph-based features Subchapter: Social-based features Subchapter: Feature selection

Chapter 5

Algorithm Design

Discuss model selection and features selection. Describe the design process and justify all choices made.

Subchapter: Learning algorithm

Chapter 6

Implementation

Cover the process by which everything was calculated and implemented. Talk about the software packages used and any special techniques.

Chapter 7

Testing and Results

Talk about the testing process. What did and did not work. Describe the results and what they mean.

Subchapter: Proving behavior is constant in time Subchapter: Classification results Subchapter: Hierarchy analysis

Chapter 8

Conclusions

Summarize the results. Discuss the inferences that can be drawn from this result.

Subchapter: discuss future work/what you could do with a larger dataset.

Chapter 9

Applications

Go into detail about what the applications are of this result. Cover: corporate espionage and alternative communication networks (call records, text message metadata, social networks, etc.).

Bibliography