Project proposal

Email Mining

By -Shruti Garg (2003cs10189)
Rohan Choudhary (2003cs50220)

Aim of the Project

1) Email Classification
2) Mining Social Networks

Project Description

1) Email Classification can be applied to several different applications including filtering messages based on priority, assigning messages to user created folders or identifying spam [3]. We will focus on the problem of assigning messages to a user's folders based on that user's foldering strategy. Later on, we will build an automatic tagging mechanism which will classify a user's emails into certain categories without any input from the user using topic discovery methods. One major consideration in the classification is that of how to represent the messages. Specifically, one has to decide which features to use, and how to apply those features to the classification. In past, various kinds of features have been considered – unstructured text (text in subject and body), categorical text (text in 'to' and 'from' fields), numeric data (message size, number of recipients), relationship data (connections between an email message and other types of objects, such as users, folders, or other emails).We plan to explore these techniques and use them to build our automatic classifier which could be used as a plug-in for desh.

2) Mining Social network

Social network Detection means finding patterns of interaction among members of a society. [1],[2] The Emailing network can be naturally modeled as a directed graph, consisting of a set of abstract nodes (all addresses in 'to' and 'from' fields of all emails form separate nodes) joined by directional edges (an edge is put from one node to other if the previous one has sent an email to the latter). Once this social network graph based on emails is formed, first thing to do will be finding overlapping clusters. These can be viewed as sub-communication networks within the society. After this some social network metrics like – In degree, out degree, between nesses (high between ness means that this node forms an important link (bottleneck) in many of the shortest paths in graph) can be defined. These metrics can be then used to compute the equivalent of page rank which in this case will be rank assigned to each node (emailer) and one can then identify hubs (persons who frequently send email to many other people) and authorities (important people who receive email from lot of people) in a social network.

After this, if time permits we will try to analyze the temporal evolution of social networks as is being done in case of Enron corpus dataset by many researchers. [4]

Dataset Required

1) Desh logs (who send to whom data over a few months)
2) Enron corpus (publicly available dataset)

References

1) MSR-challenge report: Mining email social networks in Postgres
Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, Anand Swaminathan
2) Beyond source code: Mining email social networks
Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, Anand Swaminathan
3) Email Classification with Co-Training
Svetlana Kiritchenko and Stan Matwin
4) The Enron Corpus: A New Dataset for Email Classification Research
Bryan Klimt and Yiming Yang