Mining Enron Emails Using Data Streaming Techniques

Semih Sahin
20801301

Outline

- Problem Description
- Enron Email Dataset
- Motivation/Importance
- Methodology
- Expected Results

Problem Description

- Social Network Analysis
  - Relationship extraction
  - Relationship type with sentiment analysis
- Data Stream Mining

•Enron Email Dataset

- Enron Email Set
  - Contains 0.5M messages
  - Collection of real email that is public
- Berkeley Enron Email Analysis Project
- Emails are labeled by hand
  - Genre
  - Primary Topic
  - Tone

Motivation/Importance

- More specific community detection
  - Time
  - Relationship type

Methodology

- Natural Language Processing
    - classification
- Sentiment Analysis
- Data Stream Mining
    - no batch processing
    - single scan algorithms
    - pipeline, task and data parallelism (if possible)

Expected Results

- Effectiveness
  - several algorithms will be compared
- Efficiency
  - depends on the available parallelism
  - much faster than sequential mining algorithms

Thank you !