An Evolutionary-Based Method for Reconstructing Conversation Threads in Email Corpora
Outline

- Introduction
- Problem Definition
- Previous Works
- Our Method
- Evaluation
Email

- Email has become one of the most popular tools for handling conversation among people

- This type of Web data is produced in enormous quantities
  - Email Overloading
    - Email data analysis
Problem Definition

- Emails are not completely independent of each other
  - They could be written as a response to other emails

- **Conversation Thread:**
  Cluster of exchanged emails, among the *same group of people* by reply or forwarding on *the same topic*.
  - There are Parent-Child relations between emails
    - Conversations have *tree* structure
Application

• Several useful applications beside improving user experience in email management:
  • Discussions search in public mailing lists
  • Email summarization
  • Email classification
  • Automatic question answering
  • Expert finding
  • Visualization
Previous work

• **Metadata based**
  • Use email headers fields such as *In-Reply-To* and *References*
    - *These information, are not always available.*
    - *are not completely true in real conversations.*

• **Content based**
  • Use content of emails
The problem of reconstructing conversation threads could be mapped to an optimization problem:

- **Finding a jungle of best rooted trees in the space of any possible rooted tree**

First solution: **exhaustive search**

\[
J(n) = \sum_{k=1}^{n} \binom{n-1}{k-1} n^{n-k}
\]

- That is greater than Catalan number
- For example, we have 16 different jungles of rooted trees with just 3 nodes.
Previous Solutions

- **Second solution:** Ignoring the tree structure (Erara et al. 2008)
  - Find conversations as clusters of emails
  - Arrange emails that belong to the same conversation in a chronological order

- **Third solution:** Making some restricting and unreal assumptions (Wu et al. 2006, Joshi et al. 2011)
  - Considering all of the emails of same conversation to have the same subject line
  - Considering fixed max time window for conversations
Our Evolutionary Approach

- Intelligent evolutionary search in the whole solution space to find global optimum instead of some approximate solution:

Genetic Programming
Pillars of Genetic Programming

- **Representation**
  - Each conversation thread is characterized by a single chromosome
  - Chromosome: A rooted tree

- **Initialization**
  - Traditional tree generation algorithm for genetic programming
    - RAMPED HALF-AND-HALF
  - Prune the unfeasible space
    - Parent emails always have earlier time relative to their children
**Pillars of Genetic Programing**

- **Fitness Function**
  - Multiple objective function:
    \[
    F(C) = W_t F_t(C) + W_s F_s(C) + W_p F_p(C) + W_c F_c(C)
    \]
  - Time fitness:
    \[
    F_t(C) = \frac{n}{\text{diff}(C)}
    \]
  - Subject fitness:
    \[
    F_s(C) = \frac{\sum_{e_i, e_j \in C \text{ and } e_i \rightarrow e_j} \text{Sim}_S(e_i, e_j)}{\log n}
    \]
Pillars of Genetic Programming

- Participants fitness:

\[ F_p(C) = W_{F1} F_{1p}(C) + W_{F2} F_{2p}(C) \]

- Local closeness of participants

\[ F_{1p}(C) = \frac{\sum_{e_i, e_j \in C} Sim_p(e_i, e_j)}{2 \log^2 n} \]

- Global closeness of participants

\[ F_{2p}(C) = \frac{\sum_{p_i, p_j \in p(C)} \text{closeness}_{SN}(p_i, p_j)}{2 \log |p(C)|} \]
Global closeness of participants:

People who are closer to each other in terms of communication, are more probable to contribute in a special conversation

- \[ \text{closeness}_{SN}(p_i, p_j) = \frac{|\text{adj}(p_i) \cap \text{adj}(p_j)|}{|\text{adj}(p_i) \cup \text{adj}(p_j)|} \]
- Content fitness:

\[ F_C(C) = \frac{\sum_{e_i, e_j \in C \text{ and } e_i \rightarrow e_j} \text{Sim}_C(e_i, e_j)}{\log n} \]

- Selection
  - Combination of *Roulette Wheel* method and *Elitism*

- Reproduction
  - Duplication, Mutation and Crossover
Overall Procedure

- Regarding the explained elements:
  - First an initial population has been chosen
  - Then, these steps are repeated until termination.
    1. calculating the fitness of each individual
    2. selecting the best-fit individuals
    3. breeding new individuals through crossover and mutation to generate a new generation.
  - Select best trees from last generation as the final result
Experiments

- **BC3** corpus has been used as dataset for experimental evaluations.
  - BC3 is a subset of W3C
  - B3C corpus contains 269 emails and 40 conversation threads with an average length of 6.752 emails

Conversation Threads of BC3
Experiments

- Measures:

\[
\text{Precision} = \frac{N_{GP}}{N_{TC}} \quad \text{Precision}_{PC} = \frac{PC_{GP}}{PC_{TC}}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Precision_PC</th>
<th>Recall (Regardless of the structure)</th>
<th>Recall (Regarding to the structure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Evolutionary Method</td>
<td>0.747</td>
<td>0.796</td>
<td>0.937</td>
<td>0.876</td>
</tr>
<tr>
<td>Lewis and Knowels method(1998)</td>
<td>0.575</td>
<td>-</td>
<td>0.425</td>
<td>-</td>
</tr>
<tr>
<td>Wu and Oard method(2006)</td>
<td>0.314</td>
<td>-</td>
<td>0.511</td>
<td>-</td>
</tr>
<tr>
<td>Joshi et al. method(2011)</td>
<td>-</td>
<td>0.873</td>
<td>-</td>
<td>0.632</td>
</tr>
</tbody>
</table>
**Summery**

- We present a new approach based on genetic programming for reconstruction of conversation threads in emails data.

- We map the problem to an optimization problem.
- We Exploite genetic programming
  - Intelligent search in the space of possible solutions.
Future Work

- Representing chromosome by a jungle of rooted trees instead of single rooted tree
  - Some parameters are considered as a part of solution and it is determined automatically
- Using semantic similarity among the content of emails
- Using better closeness measures to estimate similarity of people regarding their social network
- Using learning approach to parameter estimation
Question?