A Novel Approach of Mining Write-Prints for
Authorship Attribution in E-mail Forensics

By
Farkhund Iqbal, Rachid Hadjidj, Benjamin Fung, Mourad Debbabi

Presented At
The Digital Forensic Research Conference
DFRWS 2008 USA  Baltimore, MD (Aug 11th - 13th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

http://dfrws.org
A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics

Farkhund Iqbal
Benjamin C. M. Fung
Rachid Hadjtidj
Mourad Debbabi

Computer Security Lab
Concordia Institute for Information Systems Engineering
Concordia University
Montreal, Canada
Authorship Identification

Informal problem description

• A person wrote an email, e.g., a blackmail or a spam email.

• Later on, he denied to be the author.

• Our goal: Identify the most plausible authors and find evidence to support the conclusion.
Cybercrime via E-mails

- My personal real-life example: Offering homestay for international students.

My home

Carmela in US

Same person

Anthony in Canada
Evidence I have

- Cell phone number of Anthony: 647-8302170
- 15 e-mails from “Carmela”
- A counterfeit cheque
The Problem

- To determine the author of a given malicious email $\mu$.

- Assumption #1: the author is likely to be one of the suspects $\{S_1, \ldots, S_n\}$.

- Assumption #2: have access to some previously written emails $\{E_1, \ldots, E_n\}$.

- The problem is
  - to identify the most plausible author from the suspects $\{S_1, \ldots, S_n\}$. 

Email $\mu$ from unknown author
Current Approach

E-mails $E_1$  
E-mails $E_2$  
E-mails $E_3$

Classification Model

- Capital Ratio:
  - $[0,0.3)$
  - $[0.3,0.5)$
  - $[0.5,1)$

- # of Commas:
  - $<0.5$
  - $>0.5$

- $S_1$
- $S_2$
- $S_3$
- …..
Related Work

- Abbasi and Chen (2008) presented a comprehensive analysis on the stylistics features.

- Lexical features [Holmes 1998; Yule 2000, 2001]
  - characteristics of both characters and words or tokens.
  - vocabulary richness and word usage.

- Syntactic features (Burrows, 1989; Holmes and Forsyth, 1995; Tweedie and Baayen, 2005, 2006)
Related Work

- Structural features
  - measure the overall layout and organization of text within documents.

- Content-specific features (Zheng et al. 2006)
  - collection of certain keywords commonly found in a specific domain and may vary from context to context even for the same author.
Related Work

1. Decision Tree (e.g., C4.5)
   - Classification rules can justify the finding.
   - **Pitfall 1:** Classification model is built from e-mails of all suspects. Suspects may share common writing styles, but the investigator may utilize those common styles as part of the evidence.
   - **Pitfall 2:** Consider one attribute at a time, i.e., making decision based on local information.

<table>
<thead>
<tr>
<th>Capital Ratio</th>
<th># of Commas</th>
<th>...</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Related Work

2. SVM
(Support Vector Machine)
(DeVel 2000; Teng et al. 2004)

- Accurate, because considers all features at every step.

- **Pitfall**: A black box. Difficult to present evidence to justify the...
Our Approach: AuthorMiner

**Phase 1:** Mining frequent patterns:

Frequent Pattern:
A set of feature items that frequently occur together in set of e-mails $E_i$.

Frequent patterns (a.k.a. frequent itemset)
- Foundation for many data mining tasks
- Capture combination of items that frequently occurs together
- Useful in marketing, catalogue design, web log, bioinformatics, materials...
Our Approach: AuthorMiner

Phase 2: Filter out the common frequent patterns among suspects.
Our Approach: AuthorMiner

Phase 2: Filter out the common frequent patterns among suspects.
Our Approach: AuthorMiner

Phase 3: Match e-mail $\mu$ with write-print.
Phase 0: Preprocessing

<table>
<thead>
<tr>
<th>E-mail</th>
<th>Feature A</th>
<th>Feature B</th>
<th>Feature C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1</td>
<td>A2</td>
<td>A3</td>
</tr>
<tr>
<td>ε₁</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ε₂</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ε₃</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ε₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ε₅</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ε₆</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ε₇</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ε₈</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ε₉</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ε₁₀</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε₁ = {A₂, B₁, C₁}</td>
</tr>
<tr>
<td>ε₂ = {A₂, B₁, C₁}</td>
</tr>
<tr>
<td>ε₃ = {A₂, B₁, C₁}</td>
</tr>
<tr>
<td>ε₄ = {A₁, B₁, C₁}</td>
</tr>
<tr>
<td>ε₅ = {A₄, B₁, C₁}</td>
</tr>
<tr>
<td>ε₆ = {A₃, B₂, C₂}</td>
</tr>
<tr>
<td>ε₇ = {A₄, B₁, C₂}</td>
</tr>
<tr>
<td>ε₈ = {A₃, B₂, C₂}</td>
</tr>
<tr>
<td>ε₉ = {A₂, B₁, C₂}</td>
</tr>
<tr>
<td>ε₁₀ = {A₁, B₁, C₂}</td>
</tr>
</tbody>
</table>
Phase 1: Mining Frequent Patterns

- An e-mail $\varepsilon$ contains a pattern $F$ if $F \subseteq \varepsilon$.
- The support of a pattern $F$, $\text{support}(F|E_i)$, is the percentage of e-mails in $E_i$ that contains $F$.
- $F$ is frequent if its support($F|E_i$) $> \text{min\_sup}$.
  - Suppose $\text{min\_sup} = 0.3$.
  - $\{A2,B1\}$ is a frequent pattern because it has support $= 4$. 

<table>
<thead>
<tr>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_1 = {A2, B1, C1}$</td>
</tr>
<tr>
<td>$\varepsilon_2 = {A2, B1, C1}$</td>
</tr>
<tr>
<td>$\varepsilon_3 = {A2, B1, C1}$</td>
</tr>
<tr>
<td>$\varepsilon_4 = {A1, B1, C1}$</td>
</tr>
<tr>
<td>$\varepsilon_5 = {A4, B1, C1}$</td>
</tr>
<tr>
<td>$\varepsilon_6 = {A3, B2, C2}$</td>
</tr>
<tr>
<td>$\varepsilon_7 = {A4, B1, C2}$</td>
</tr>
<tr>
<td>$\varepsilon_8 = {A3, B2, C2}$</td>
</tr>
<tr>
<td>$\varepsilon_9 = {A2, B1, C2}$</td>
</tr>
<tr>
<td>$\varepsilon_{10} = {A1, B1, C2}$</td>
</tr>
</tbody>
</table>
Phase 1: Mining Frequent Patterns

- **Apriori property**: All nonempty subsets of a frequent pattern must also be frequent.
  - If a pattern is not frequent, its superset is not frequent.
- Suppose \( \text{min} \_ \text{sup} = 0.3 \)
- \( C_1 = \{A1,A2,A3,A4,B1,B2,C1,C2\} \)
- \( L_1 = \{A2, B1,C1,C2\} \)
- \( C_2 = \{A2B1,A2C1,A2C1,A2C2,B1C1,B1C2,C1C2\} \)
- \( L_2 = \{A2B1,A2C1,B1C1,B1C2\} \)

<table>
<thead>
<tr>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varepsilon_1 = {A2, B1, C1} )</td>
</tr>
<tr>
<td>( \varepsilon_2 = {A2, B1, C1} )</td>
</tr>
<tr>
<td>( \varepsilon_3 = {A2, B1, C1} )</td>
</tr>
<tr>
<td>( \varepsilon_4 = {A1, B1, C1} )</td>
</tr>
<tr>
<td>( \varepsilon_5 = {A4, B1, C1} )</td>
</tr>
<tr>
<td>( \varepsilon_6 = {A3, B2, C2} )</td>
</tr>
<tr>
<td>( \varepsilon_7 = {A4, B1, C2} )</td>
</tr>
<tr>
<td>( \varepsilon_8 = {A3, B2, C2} )</td>
</tr>
<tr>
<td>( \varepsilon_9 = {A2, B1, C2} )</td>
</tr>
<tr>
<td>( \varepsilon_{10} = {A1, B1, C2} )</td>
</tr>
</tbody>
</table>
Phase 2: Filtering Common Patterns

Before filtering:

\[ FP(E_1) = \{ A2, B1, C1, C2, A2B1, A2C1, B1C1, B1C2, A2B1C1 \} \]
\[ FP(E_2) = \{ A1, B1, C1, A1B1, A1C1, B1C1, A1B1C1 \} \]
\[ FP(E_3) = \{ A2, B1, C2, A2B1, A2C2 \} \]

After filtering:

\[ WP(E_1) = \{ A2, A2C1, B1C2, A2B1C1 \} \]
\[ WP(E_2) = \{ A1, A1B1, A1C1, A1B1C1 \} \]
\[ WP(E_3) = \{ A2, A2C2 \} \]
Phase 3: Matching Write-Print

- Intuitively, a write-print \( WP(E_i) \) is similar to \( \mu \) if many frequent patterns in \( WP(E_i) \) matches the style in \( \mu \).
- Score function that quantifies the similarity between the malicious e-mail \( P \) and a write-print \( WP(E_i) \):

\[
\text{Score}(\mu \approx WP(E_i)) = \frac{\sum_{j=1}^{p} \text{support}(MP_j|E_i)}{|WP(E_i)|}
\]

- The suspect having the write-print with the highest score is the author of the malicious e-mail \( \mu \).
Major Features of Our Approach

- **Justifiable evidence**
  - Guarantee the identified patterns are frequent in the e-mails of one suspect only, and are not frequent in others' emails

- **Combination of features (frequent pattern)**
  - Capture the combination of multiple features (cf. decision tree)

- **Flexible writing styles**
  - Can adopt any type of commonly used writing style features
  - Unimportant features will be ignored.
Experimental Evaluation

- Dataset: Enron E-mail
- 2/3 for training, 1/3 for testing, 10-fold cross validation

![Graphs showing accuracy vs min_sup](image)

21
Experimental Evaluation

- Example of write-print:

\{\text{regrds, u}\}
\{\text{regrds, capital letter per sentence } = 0.02\}
\{\text{regrds, u, capital letter per sentence } = 0.02\}
Conclusion

- Most previous contributions focused on improving the classification accuracy of authorship identification, but only very few of them study how to gather strong evidence.

- We introduce a novel approach of authorship attribution and formulate a new notion of write-print based on the concept of frequent patterns.
References

- O. De Vel. Mining e-mail authorship. paper presented at the workshop on text mining. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2000.
References

- G. Yule. On sentence length as a statistical characteristic of style. Biometrika 25
References
