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)

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

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

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

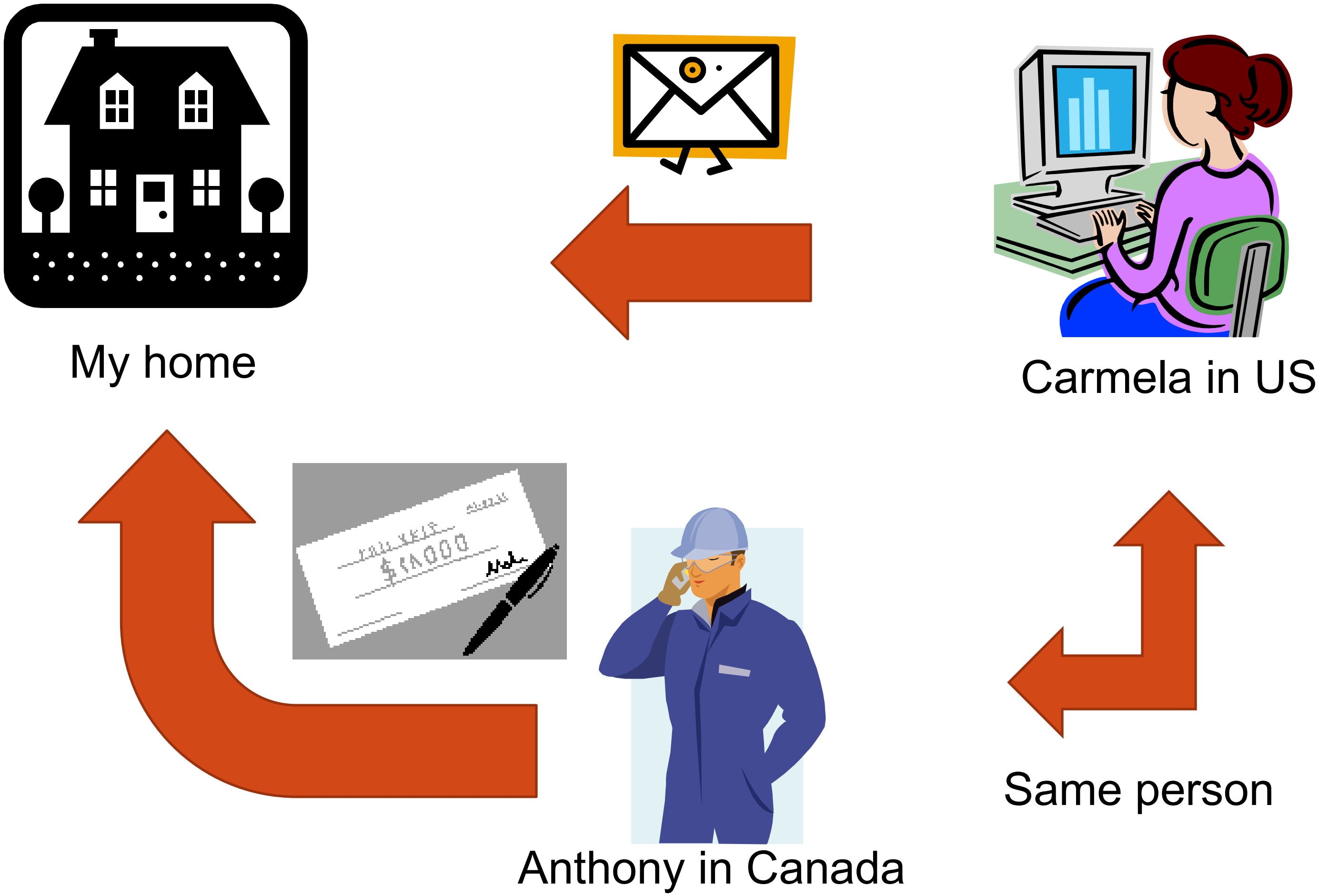
Authorship Identification

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.



Evidence I have

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


SCOTT'S REAL ESTATE LIMITED PARTNERSHIP
161 BAY STREET, SUITE 2300
CT TOWER, BCE PLACE
TORONTO, ONTARIO M5J 2S1

BANK OF MONTREAL
ONE FIRST CANADIAN PLACE
TORONTO, ONTARIO M5X 1H3

1783

1783

Date 0 5 1 5 2 0 0 8
M M D D Y Y Y Y

Four Thousand Five Hundred and Zero Cents

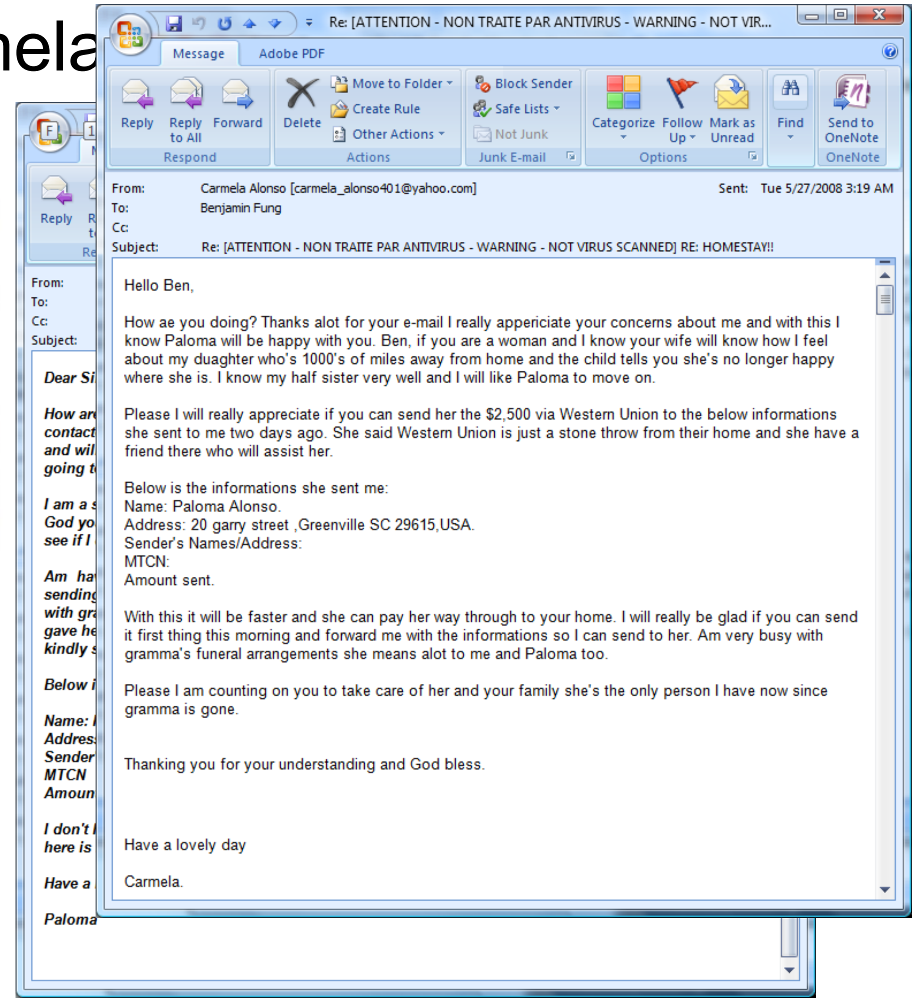
PAY

$ *****4,500.00

TO THE ORDER OF BENJAMIN FUNG

SCOTT'S REAL ESTATE LIMITED PARTNERSHIP
PER
PER

⑈001783⑈ ⑆00022⑉001⑆ 1501⑉883⑈


Re: [ATTENTION - NON TRAITE PAR ANTIVIRUS - WARNING - NOT VIR...

From: Carmela Alonso [carmela_alonso401@yahoo.com] Sent: Tue 5/27/2008 3:19 AM
To: Benjamin Fung
Cc:
Subject: Re: [ATTENTION - NON TRAITE PAR ANTIVIRUS - WARNING - NOT VIRUS SCANNED] RE: HOMESTAY!!

Hello Ben,

How ae you doing? Thanks alot for your e-mail I really appericiate your concerns about me and with this I know Paloma will be happy with you. Ben, if you are a woman and I know your wife will know how I feel about my duaghter who's 1000's of miles away from home and the child tells you she's no longer happy where she is. I know my half sister very well and I will like Paloma to move on.

Please I will really appreciate if you can send her the $2,500 via Western Union to the below informations she sent to me two days ago. She said Western Union is just a stone throw from their home and she have a friend there who will assist her.

Below is the informations she sent me:
Name: Paloma Alonso.
Address: 20 garry street ,Greenville SC 29615,USA.
Sender's Names/Address:
MTCN:
Amount sent.

With this it will be faster and she can pay her way through to your home. I will really be glad if you can send it first thing this morning and forward me with the informations so I can send to her. Am very busy with gramma's funeral arrangements she means alot to me and Paloma too.

Please I am counting on you to take care of her and your family she's the only person I have now since gramma is gone.

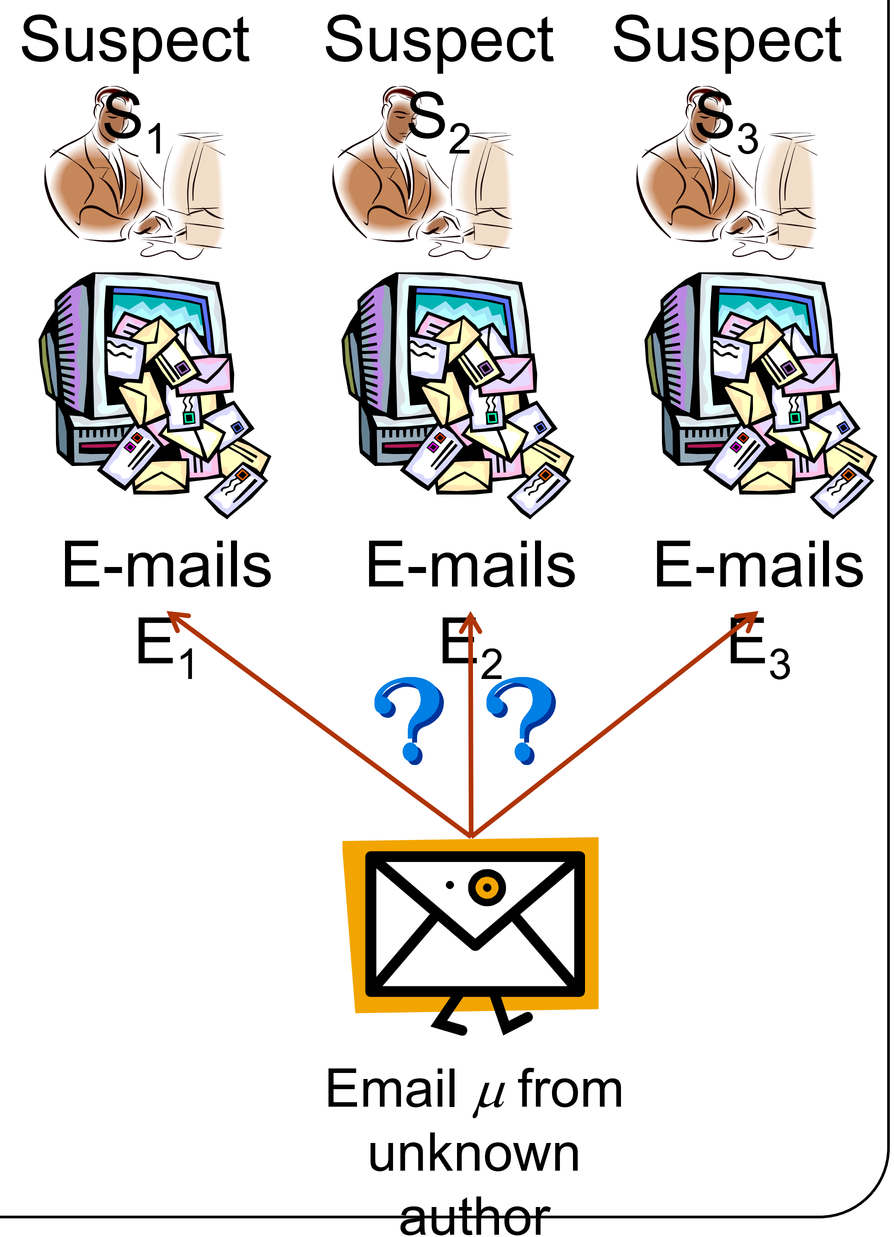
Thanking you for your understanding and God bless.

Have a lovely day

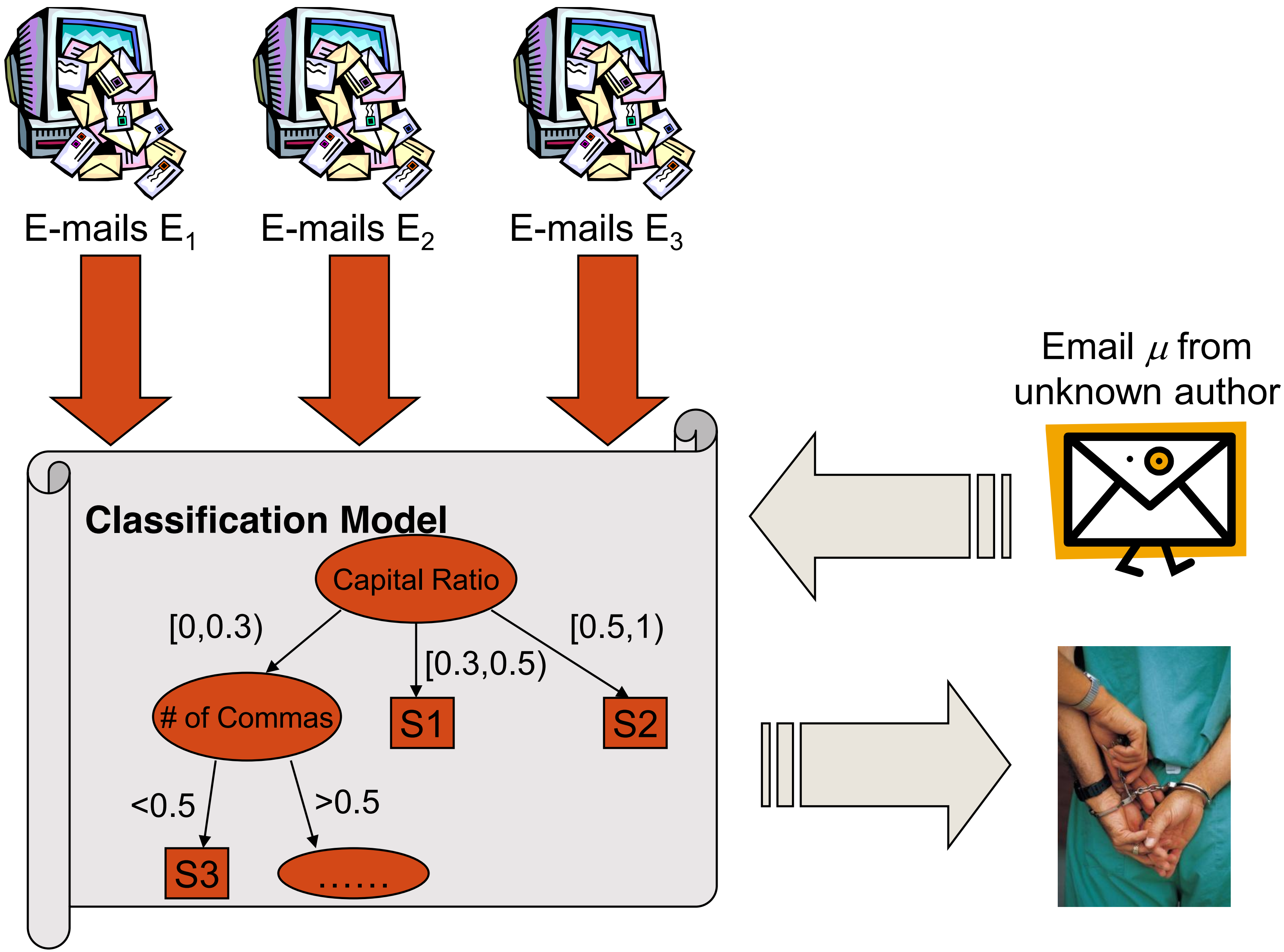
Carmela.

The Problem

- To determine the author of a given malicious e-mail μ.
- 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 e-mails $\{E_1, \ldots, E_n\}$.
- The problem is
  - to identify the most plausible author from the suspects $\{S_1, \ldots, S_n\}$



Current Approach



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

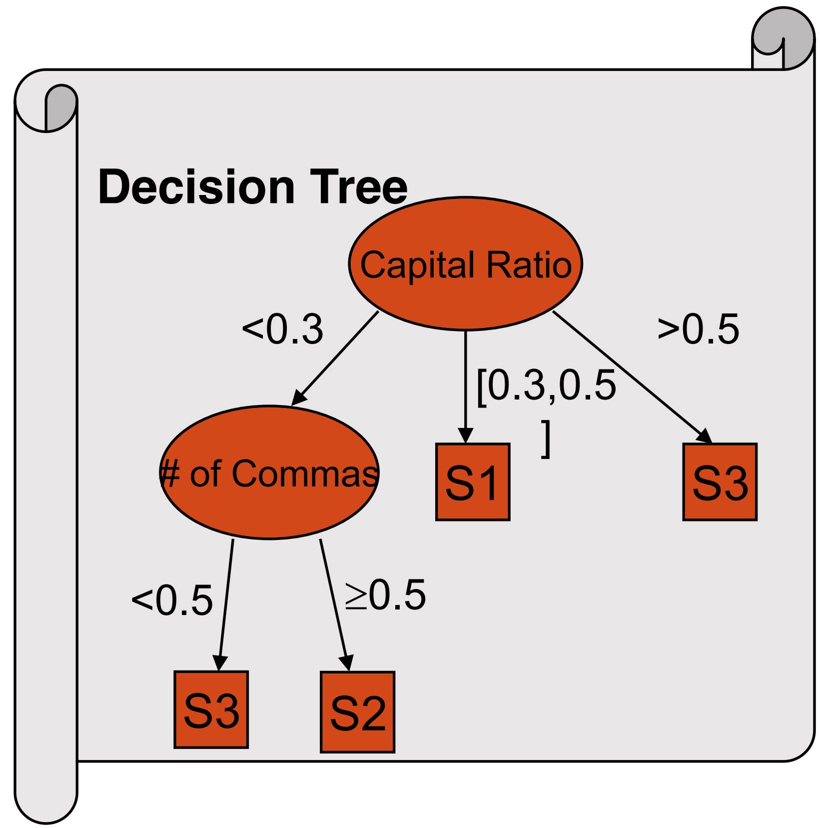
Related Work

- 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

Capital Ratio	# of Comma s	…	Class
…	…	…	…

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

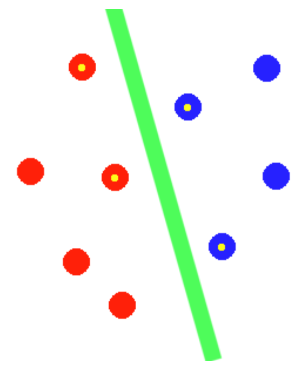


Related Work

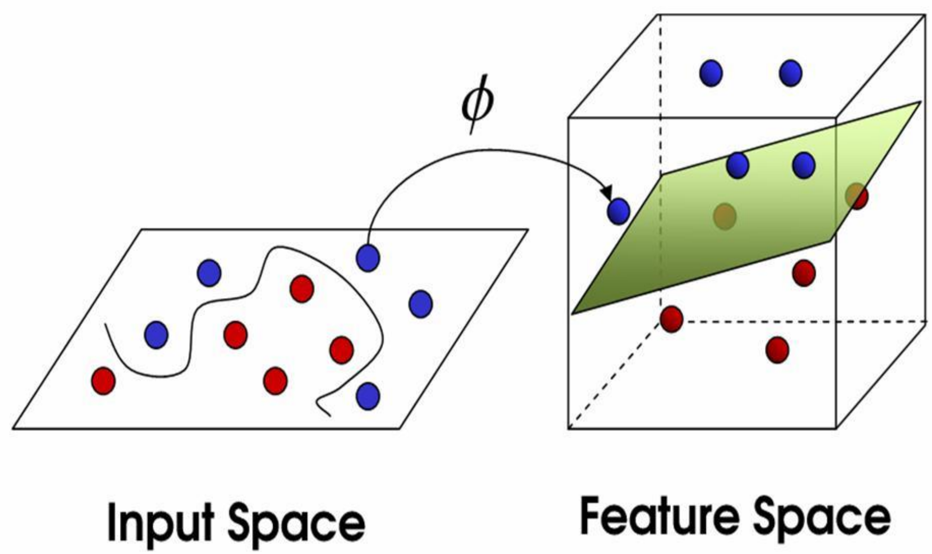
2. SVM

(Support Vector Machine)

(DeVel 2000; Teng et al. 2004)



Principle of Support Vector Machines (SVM)



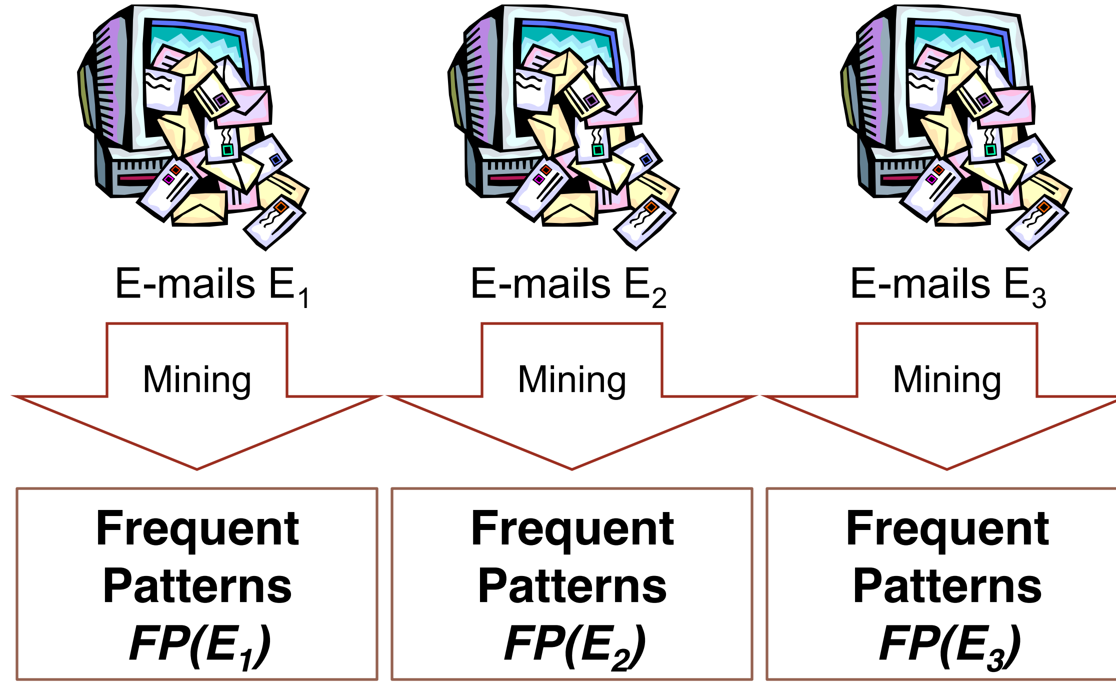
- Accurate, because considers all features at every step.

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

Source: http://www.imtech.res.in/raghava/rbpred/svm.jpg

Our Approach: AuthorMiner

Phase 1: Mining frequent patterns:



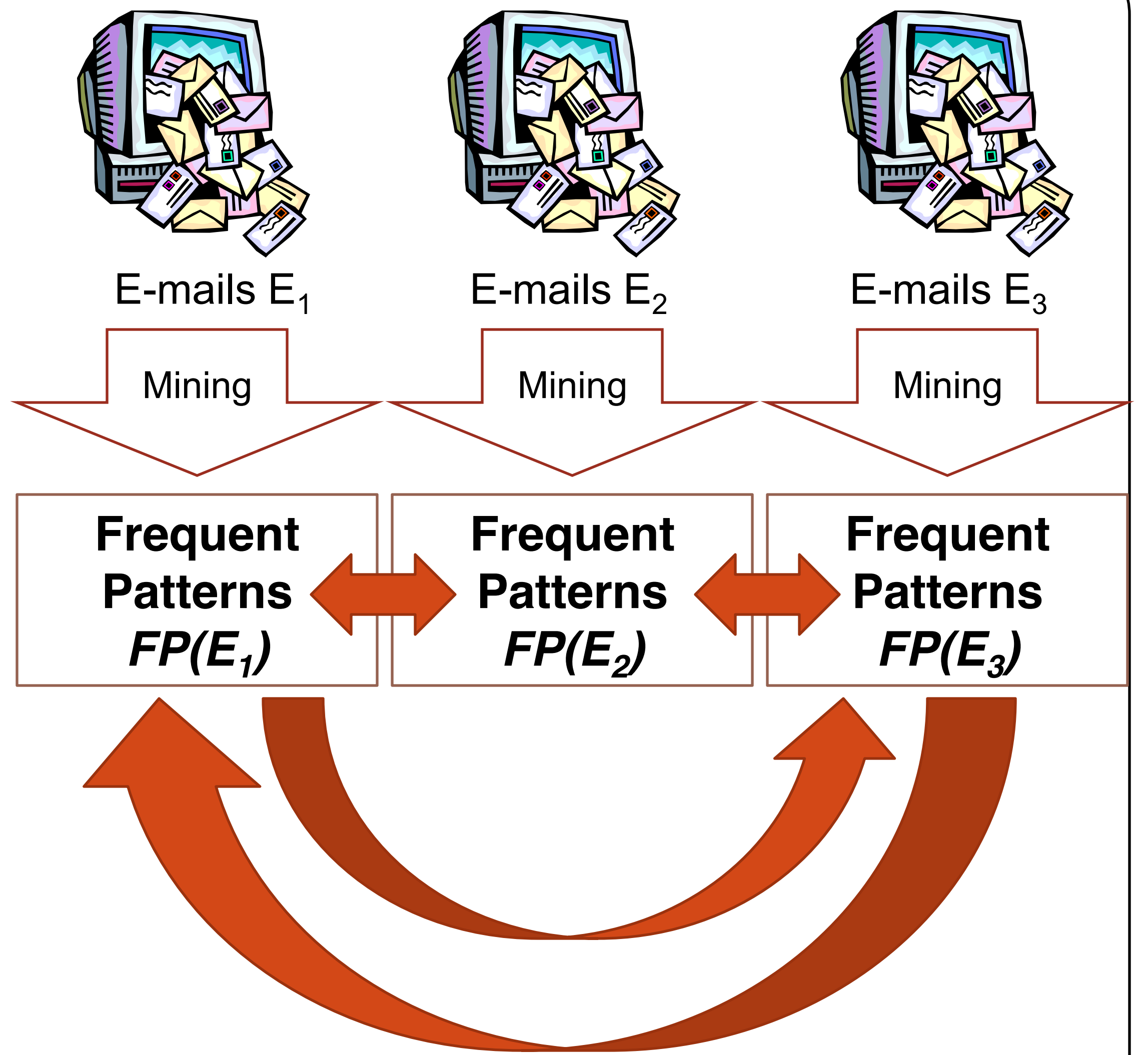
Frequent Pattern:

A set of feature items that frequently occur together in set of e-mails E_i.

Frequent patterns (a.k.a. frequent itemset)

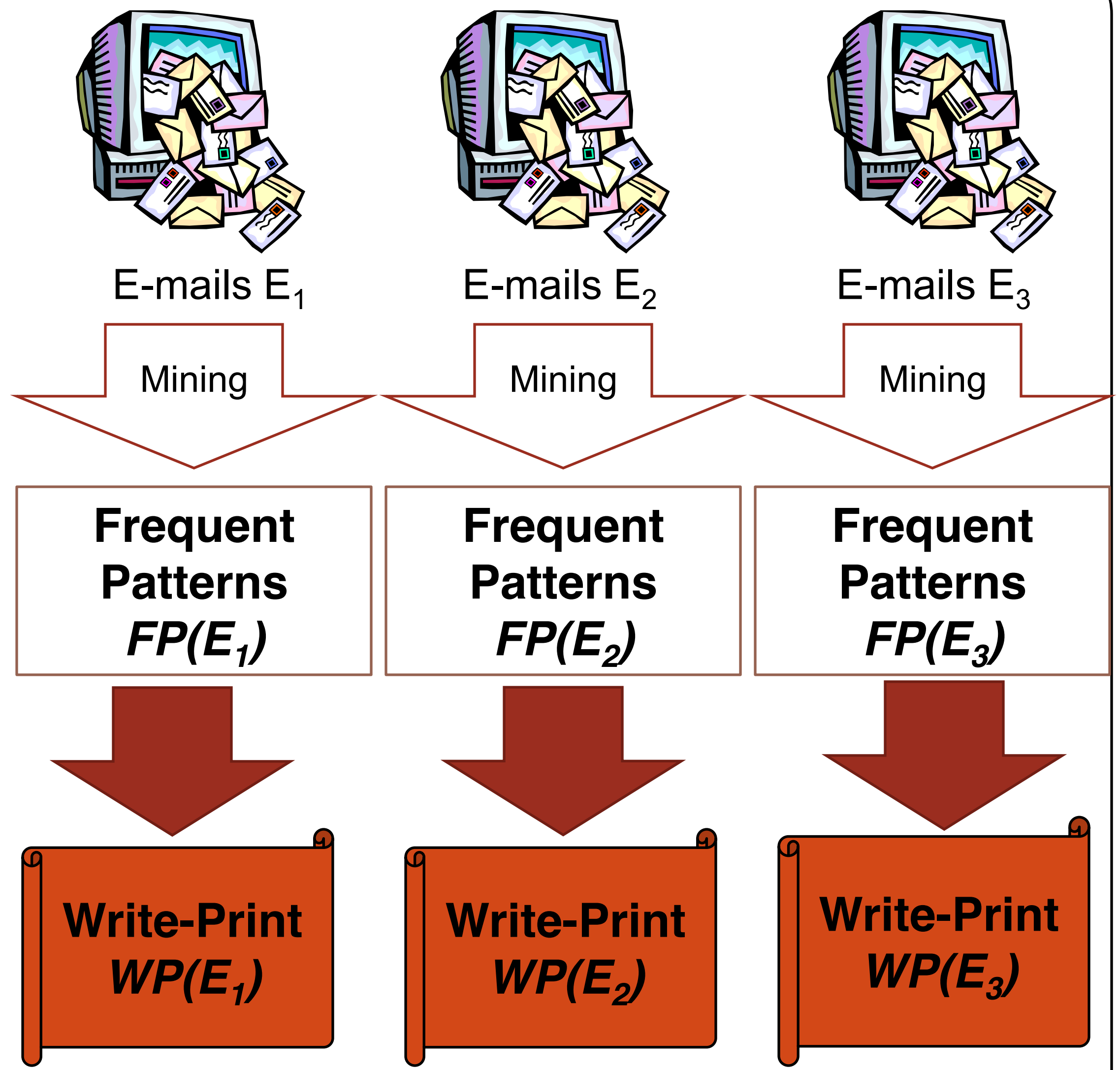
- Foundation for many data mining tasks
- Capture combination of items that frequently occurs together
- Useful in marketing, catalogue design, web log, bioinformatics, materials

Our Approach: AuthorMiner



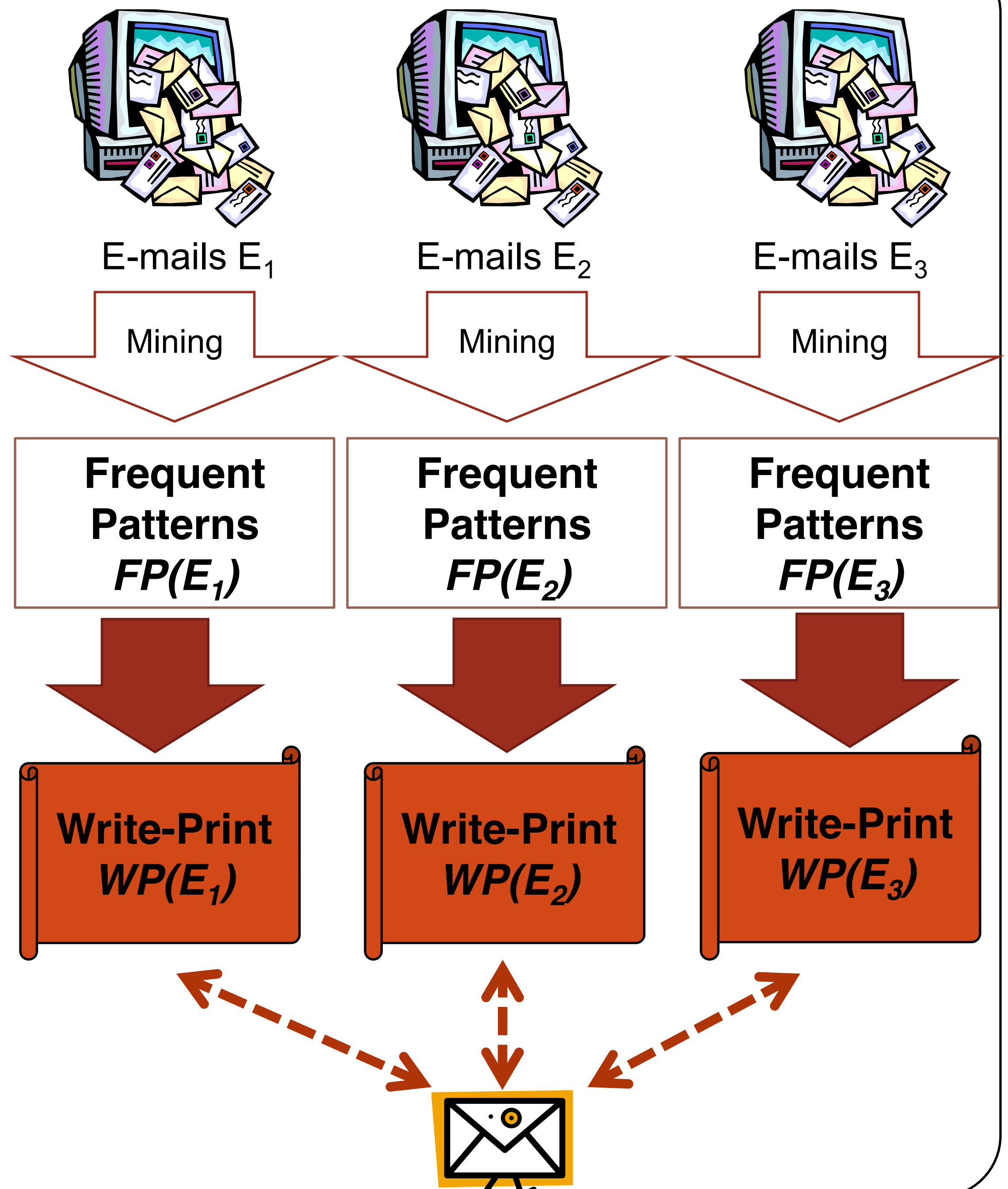
Phase 2: Filter out the common frequent patterns among suspects.

Our Approach: AuthorMiner



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

Our Approach: AuthorMiner



Phase 3: Match e-mail μ with write-print.

Phase 0: Preprocessing

	Feature A				Feature B		Feature C	
E-mail	**A1**	**A2**	**A3**	**A4**	**B1**	**B2**	**C1**	**C2**
ε_1	0	1	0	0	1	0	1	0
ε_2	0	1	0	0	1	0	1	0
ε_3	0	1	0	0	1	0	1	0
ε_4	1	0	0	0	1	0	1	0
ε_5	0	0	0	1	1	0	1	0
ε_6	0	0	1	0	0	1	0	1
ε_7	0	0	0	1	1	0	0	1
ε_8	0	0	1	0	0	1	0	1
ε_9	0	1	0	0	1	0	0	1
ε_{10}	1	0	0	0	1	0	0	1

E-mail
$\varepsilon_1 = \{A2, B1, C1\}$
$\varepsilon_2 = \{A2, B1, C1\}$
$\varepsilon_3 = \{A2, B1, C1\}$
$\varepsilon_4 = \{A1, B1, C1\}$
$\varepsilon_5 = \{A4, B1, C1\}$
$\varepsilon_6 = \{A3, B2, C2\}$
$\varepsilon_7 = \{A4, B1, C2\}$
$\varepsilon_8 = \{A3, B2, C2\}$
$\varepsilon_9 = \{A2, B1, C2\}$
$\varepsilon_{10} = \{A1, B1, C2\}$

Phase 1: Mining Frequent Patterns

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

E-mail
$\varepsilon_1 = \{\boxed{A2, B1}, C1\}$
$\varepsilon_2 = \{\boxed{A2, B1}, C1\}$
$\varepsilon_3 = \{\boxed{A2, B1}, C1\}$
$\varepsilon_4 = \{A1, B1, C1\}$
$\varepsilon_5 = \{A4, B1, C1\}$
$\varepsilon_6 = \{A3, B2, C2\}$
$\varepsilon_7 = \{A4, B1, C2\}$
$\varepsilon_8 = \{A3, B2, C2\}$
$\varepsilon_9 = \{\boxed{A2, B1}, C2\}$
$\varepsilon_{10} = \{A1, B1, C2\}$

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 min_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}
- C = {A2B1C1,B1C1C2}

E-mail
$\varepsilon_1 = \{A2, B1, C1\}$
$\varepsilon_2 = \{A2, B1, C1\}$
$\varepsilon_3 = \{A2, B1, C1\}$
$\varepsilon_4 = \{A1, B1, C1\}$
$\varepsilon_5 = \{A4, B1, C1\}$
$\varepsilon_6 = \{A3, B2, C2\}$
$\varepsilon_7 = \{A4, B1, C2\}$
$\varepsilon_8 = \{A3, B2, C2\}$
$\varepsilon_9 = \{A2, B1, C2\}$
$\varepsilon_{10} = \{A1, B1, C2\}$

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 μ if many frequent patterns in $WP(E_i)$ matches the style in μ.
- Score function that quantifies the similarity between the malicious e-mail

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

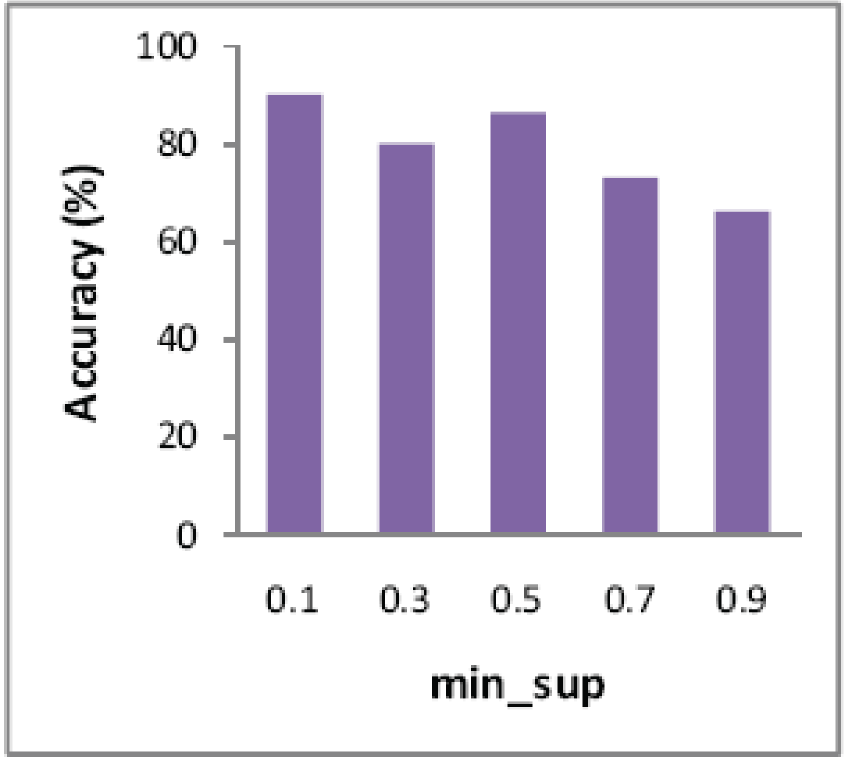
- The suspect having the write-print with the highest score is the author of the malicious e-mail μ.

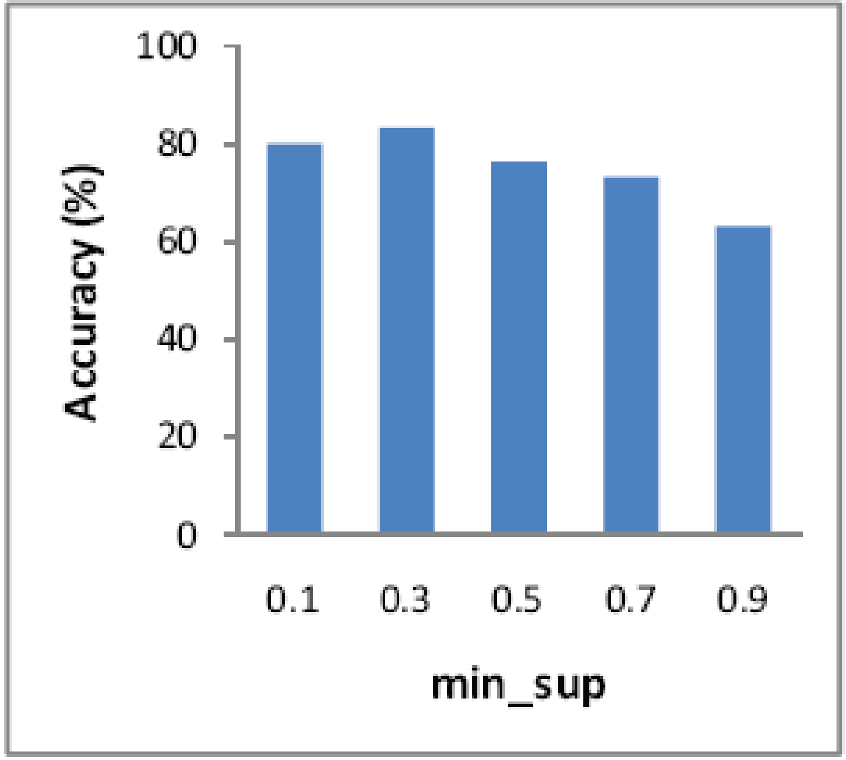
Major Features of Our Approach

- ***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





Experimental Evaluation

- Example of write-print:

{regrds, u}

{regrds, capital letter per sentence = 0.02}

{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

- J. Burrows. An ocean where each kind: statistical analysis and some major determinants of literary style. Computers and the Humanities August 1989;23(4–5):309–21.
- 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.
- B.C.M. Fung, K. Wang, M. Ester. Hierarchical document clustering using frequent itemsets. In: Proceedings of the third SIAM international conference on data mining (SDM); May 2003. p. 59–70

References

- I. Holmes I, R.S. Forsyth. The federalist revisited: new directions in authorship attribution. Literary and Linguistic Computing 1995;10(2):111–27.
- G.-F. Teng, M.-S. Lai, J.-B. Ma, and Y. Li. E-mail authorship mining based on svm for computer forensic. In *In Proc. of the 3rd International Conference on Machine Learning and Cyhemetics,* Shanghai, China, August 2004*.*
- J. Tweedie, R. H. Baayen. How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities 1998;32:323–52.
- G. Yule. On sentence length as a statistical

References

- G. Yule. The statistical study of literary vocabulary. Cambridge, UK: Cambridge University Press; 1944.
- R. Zheng, J. Li, H.Chen, Z. Huang. A framework for authorship identification of online messages: writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 2006;57(3):378–93.