DIGITAL FORENSIC RESEARCH CONFERENCE

E-mail Authorship Attribution Using Customized Associative Classification

By

Michael Schmid, Farkhund Iqbal and Benjamin Fung

Presented At

The Digital Forensic Research Conference

DFRWS 2015 USA Philadelphia, PA (Aug 9th - 13th)

E-mail Authorship Attribution using Customized Associative Classification

Michael R. Schmid, Concordia University

Farkhund Iqbal, Zayed University

Benjamin C. M. Fung, McGill University

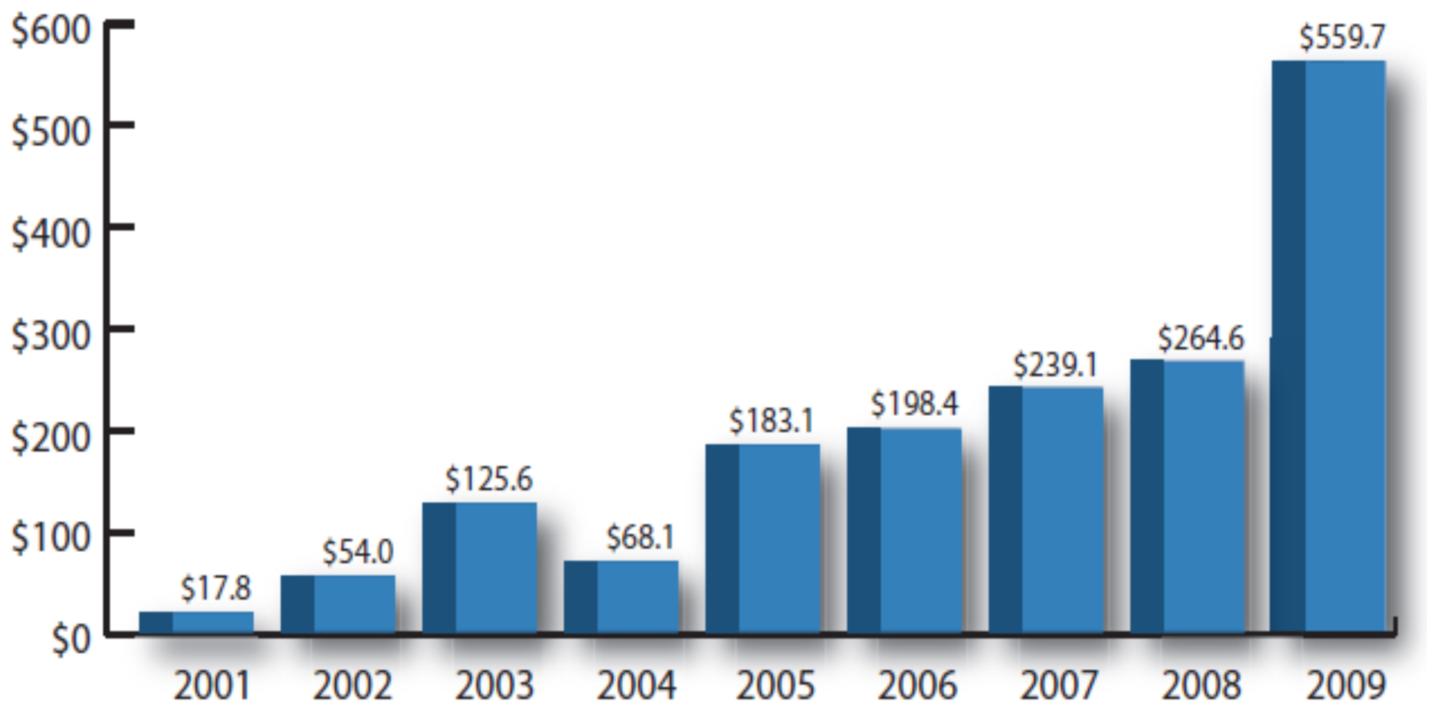
Agenda

- Motivation
- Problem Scope
- State-of-the-Art
- CMARAA-Proposed Approach
- Contributions
- Conclusion

Cybercrime Statistics

The Cyber Crime Report By IC3 US

- Number of complains = 336,655 (22.3% increase in 2009)
- Financial loses = $559.7 m (doubled in 2009)
- E-mail scams that used the FBI's name was biggest offense (16.6%)



Cybercrime Statistics

New Generation Malware

- Mariposa bots: 13m infections in 192 Countries
- Zeus bots: 3.6m infections in US, 44% in banks
- Stuxnet: 38000 infections

Norton Cybercrime Report 2011

- Annual losses : $388B
- More than the black market, $288B
- One Million victims per day (14 victims/sec.)

Symantec Intelligence Report February 2012

- Spamming 81.3%, phishing 0.462%
- China most spammed: 86.2%, US and Canada, 81.4%

E-mail (in) Security

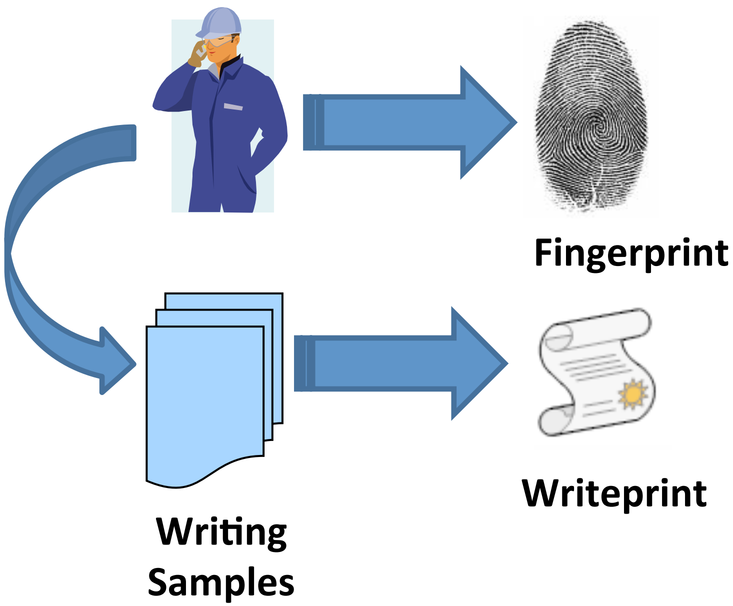
E-mail was not originally designed with security in mind.

- e-mail metadata can be spoofed
- malware is trivially transmitted by e-mail
- even clean metadata is insufficient for drawing a conclusion on authorship

Authorship Analysis

From Fingerprint to Writeprint

Stylometry shows that a person can be identified from his writing style



Used as evidence in courts of US, Europe, & Australia. [Chen et al. 2003]

Authorship: Writeprint/Wordprint

Welcome greetings

Using name as a signature

Hi,

I have several pretty cheap CD to sell. They are brand new ☺, and only $1 for each. Please contact jim@gmail.com if you are interested.

Cheer…

Nick

Closing remarks

Use email for correspondence

Stylometric Features

- ❑ Lexical features
- ❑ Syntactic features
- ❑ Structural features (layout)
- ❑ Content-specific features
- ❑ Idiosyncratic features

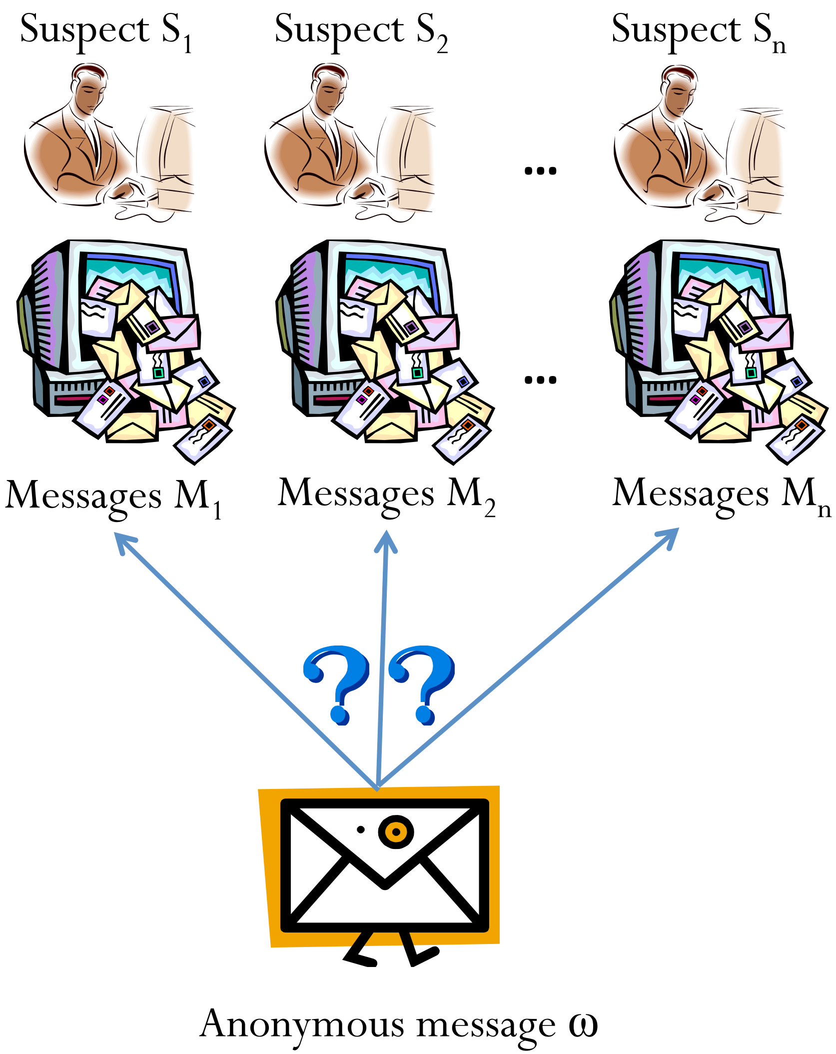
Problem: Authorship Attribution

Given

- An anonymous message ω
- Suspects $\{S_1, \ldots, S_n\}$
- Sample messages $\{M_1, \ldots, M_n\}$ of $\{S_1, \ldots, S_n\}$

The problem is

- to identify the most plausible author of message ω, and
- to gather convincing evidence to support the finding.

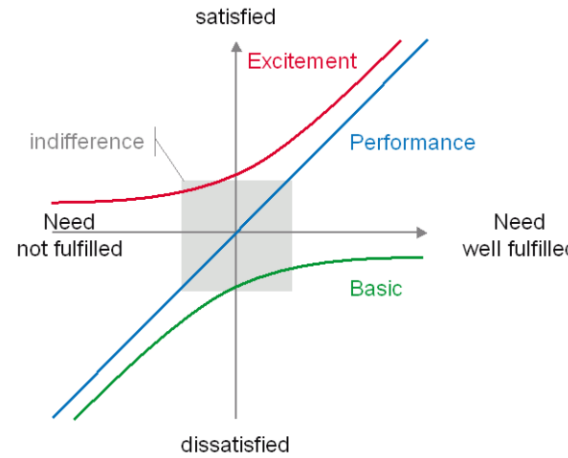


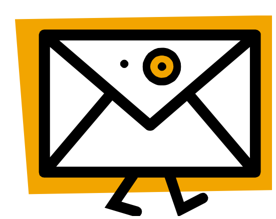
Authorship Analysis: State-of-the-Art

Suspects

Message samples

Extract Features





Message ω

Plausible Author

Authorship Analysis: State-of-the-Art

1 Naïve Bayesian

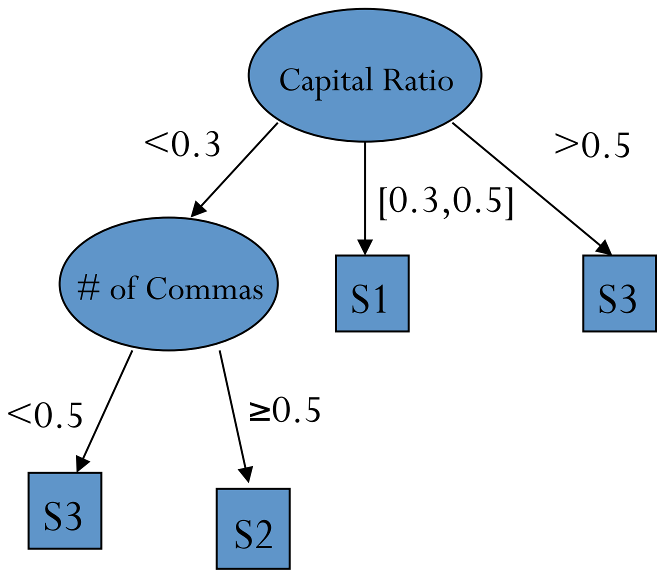
$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)}$$

Efficient

Pitfalls:
Lower classification accuracy in authorship analysis

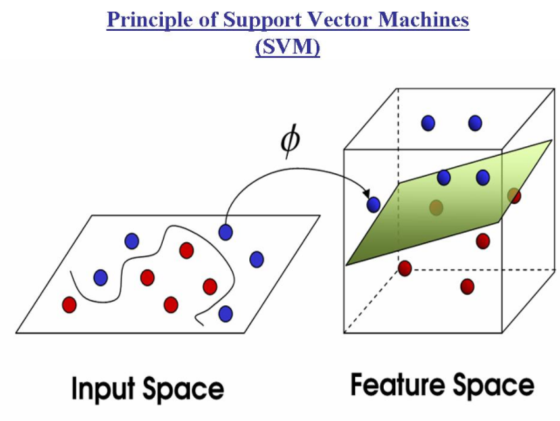
(Sebastiani 2002, Diederich et al. 2003)
Dong et al. 2006

2 Decision Trees



Efficient, interpretable results
Pitfalls: Making decision based on local information
Zhao & Zobel 2005, Sabastiani 2002

3 Support Vector Machines

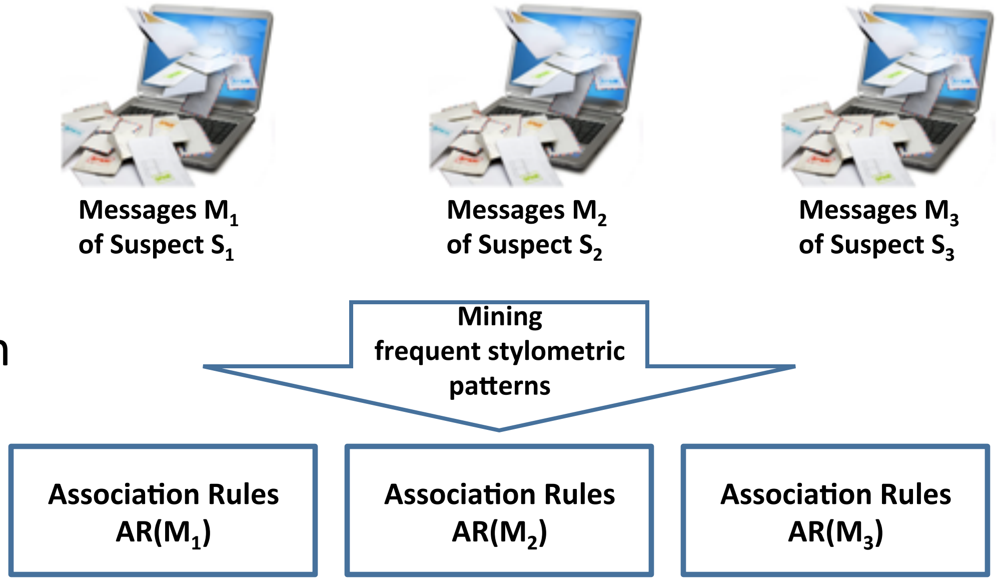


More accurate, can handle sparse data, feature combination
Pitfalls:
- Black box
- Dimensionality problem still exists

(de Vel et al. 2000-3)

CMARAA- Classification by Multiple Association Rule for AA

Phase 1: Mining Association Rules



A *class association rule* has the form

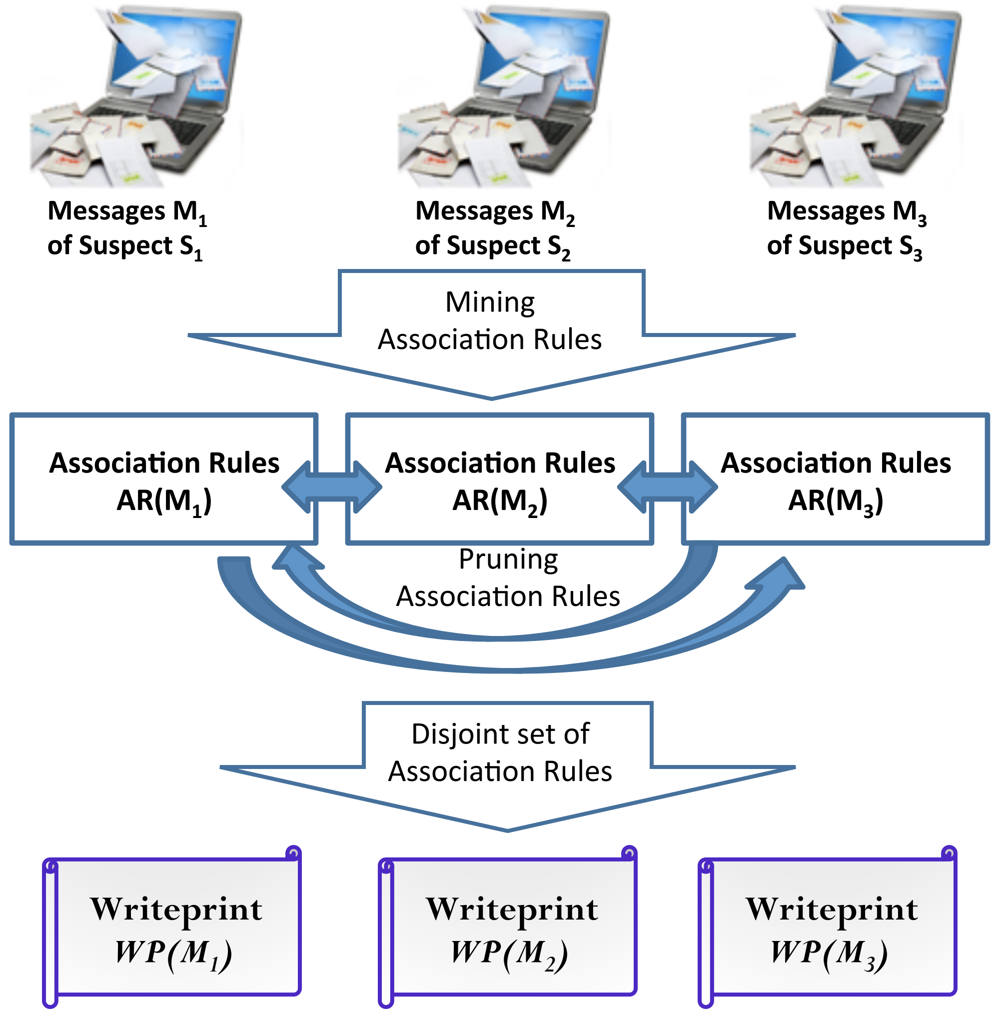
$A \rightarrow B$, where $A \subseteq V$, $B \in S$,
$sup(A \rightarrow B) \geq min\_sup$, and
$conf(A \rightarrow B) \geq min\_conf$,

Where V is set of rules and S is set of suspects and *min_sup* and *min_conf* are the minimum support and minimum confidence thresholds specified by the user.

CMARAA- Classification by Multiple Association Rule for AA

Phase 1: Mining Association Rules

Phase 2: Apply pruning-dropping common Association Rules



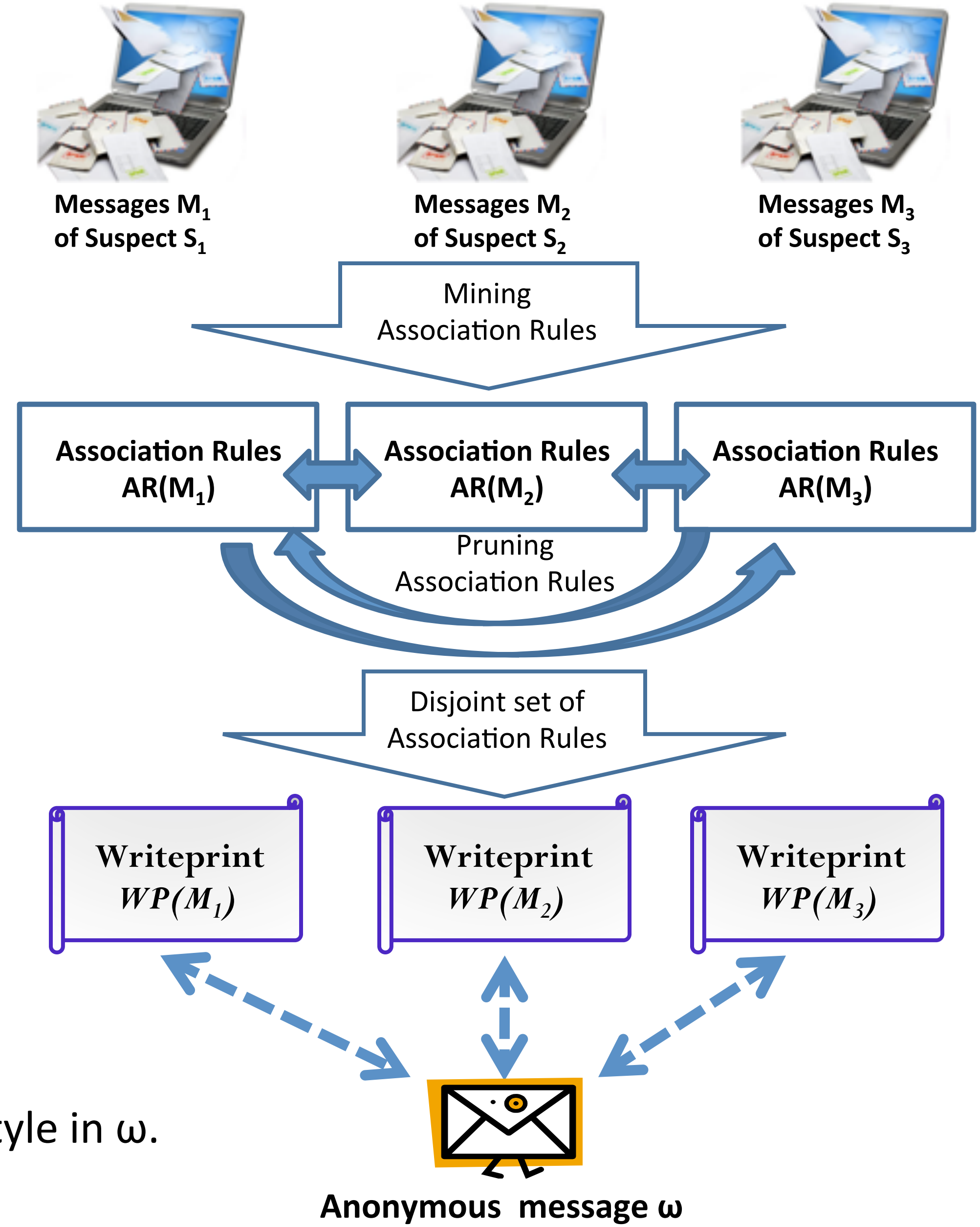
CMARAA- Classification by Multiple Association Rule for AA

Phase 1: Mining Association Rules

Phase 2: Apply pruning-dropping common Association Rules

Phase 3: Match message ω with writeprints.

$WP(M_i)$ is similar to ω
if many ARs in *$WP(M_i)$* matches the style in ω.



CMARAA- Example

❑ Phase-0: Preprocessing

- Feature extraction: We used lexical, syntactic, structural, and domain-specific features.
- Feature discretization:

Messages (μ)	Feature X			Feature Y		Feature Z	
	X_1	X_2	X_3	Y_1	Y_2	Z_1	Z_2
μ_1	0	1	0	0	1	0	1
μ_2	0	1	0	1	0	1	0
μ_3	0	1	0	1	0	1	0
μ_4	1	0	0	1	0	1	0
μ_5	0	0	0	1	0	0	1
μ_6	0	0	1	0	1	0	1
μ_7	0	0	0	1	0	0	1
μ_8	0	0	1	0	1	0	1
μ_9	0	1	0	1	0	1	0
μ_{10}	0	0	0	1	0	0	1

Messages (μ)	Feature items
μ_1	$\{X_2, Y_2, Z_2\}$
μ_2	$\{X_2, Y_1, Z_1\}$
μ_3	$\{X_2, Y_1, Z_1\}$
μ_4	$\{X_1, Y_1, Z_1\}$
μ_5	$\{Y_1, Z_2\}$
μ_6	$\{X_3, Y_2, Z_2\}$
μ_7	$\{Y_1, Z_2\}$
μ_8	$\{X_3, Y_2, Z_2\}$
μ_9	$\{X_2, Y_1, Z_1\}$
μ_{10}	$\{X_1, Y_1, Z_2\}$

CMARAA- Example (Cont.)

Each 1-Item Frequent Pattern is then inserted in the CR list:
{X2} -> B
{Y1} -> A
{Y2} -> B
{Z2} -> A
{Z3} -> B
{X1} -> A

Each Rule is checked against certain conditions in a practice called pruning.

FP-Growth as implemented in WEKA by specifying minimum support and minimum confidence

E-mail	Items	Author
e1	{X1, Y1, Z1}	A
e2	{X2, Y2, Z3}	B
e3	{X2, Y2, Z3}	B
e4	{X1, Y1, Z2}	A
e5	{X2, Y2, Z3}	B
e6	{X2, Y2, Z3}	B
e7	{X3, Y1, Z2}	A
e8	{X3, Y1, Z2}	A
e9	{X1, Y1, Z1}	A
e10	{X2, Y2, Z2}	B

CMARAA- Example (Cont.)

Now, building from our 1-Item Frequent Pattern set, we can look for 2-Item sets.

For brevity, we won't exhaustively process this table.

Some obvious 2-Item sets are:
{X2, Y2} with a support of 0.5
And
{Y1, Z2} with a support of 0.3

E-mail	Items	Author
e1	{X1, Y1, Z1}	A
e2	{**X2, Y2**, Z3}	B
e3	{**X2, Y2**, Z3}	B
e4	{X1, **Y1, Z2**}	A
e5	{**X2, Y2**, Z3}	B
e6	{**X2, Y2**, Z3}	B
e7	{X3, **Y1, Z2**}	A
e8	{X3, **Y1, Z2**}	A
e9	{X1, Y1, Z1}	A
e10	{**X2, Y2**, Z2}	B

CMARAA- Example (Cont.)

Each 2-Item Frequent Pattern is then inserted in the (partial) CR list:

{X2, Y2} -> B

~~{X2} -> B~~

{Y1} -> A

~~{Y2} -> B~~

{Y1, Z2} -> A

{Z2} -> A

{Z3} -> B

{X1} -> A

E-mail	Items	Author
e1	{X1, Y1, Z1}	A
e2	{**X2, Y2**, Z3}	B
e3	{**X2, Y2**, Z3}	B
e4	{X1, **Y1, Z2**}	A
e5	{**X2, Y2**, Z3}	B
e6	{**X2, Y2**, Z3}	B
e7	{X3, **Y1, Z2**}	A
e8	{X3, **Y1, Z2**}	A
e9	{X1, Y1, Z1}	A
e10	{**X2, Y2**, Z2}	B

CMARAA- Example (Cont.)

Now, building from our 2-Item Frequent Pattern set, we can look for 3-Item sets.

We can easily identify one 3-Item set:
{X2, Y2, Z3} with a support of 0.4

E-mail	Items	Author
e1	{X1, Y1, Z1}	A
e2	**{X2, Y2, Z3}**	B
e3	**{X2, Y2, Z3}**	B
e4	{X1, Y1, Z2}	A
e5	**{X2, Y2, Z3}**	B
e6	**{X2, Y2, Z3}**	B
e7	{X3, Y1, Z2}	A
e8	{X3, Y1, Z2}	A
e9	{X1, Y1, Z1}	A
e10	{X2, Y2, Z2}	B

CMARAA- Example (Cont.)

The 3-Item Frequent Pattern is then inserted in the CR list

{X2, Y2} -> B
{Y1} -> A
{X2, Y2, Z3} -> B
{Y1, Z2} -> A
{Z2} -> A
{Z3} -> B
{X1} -> A

E-mail	Items	Author
e1	{X1, Y1, Z1}	A
e2	**{X2, Y2, Z3}**	B
e3	**{X2, Y2, Z3}**	B
e4	{X1, Y1, Z2}	A
e5	**{X2, Y2, Z3}**	B
e6	**{X2, Y2, Z3}**	B
e7	{X3, Y1, Z2}	A
e8	{X3, Y1, Z2}	A
e9	{X1, Y1, Z1}	A
e10	{X2, Y2, Z2}	B

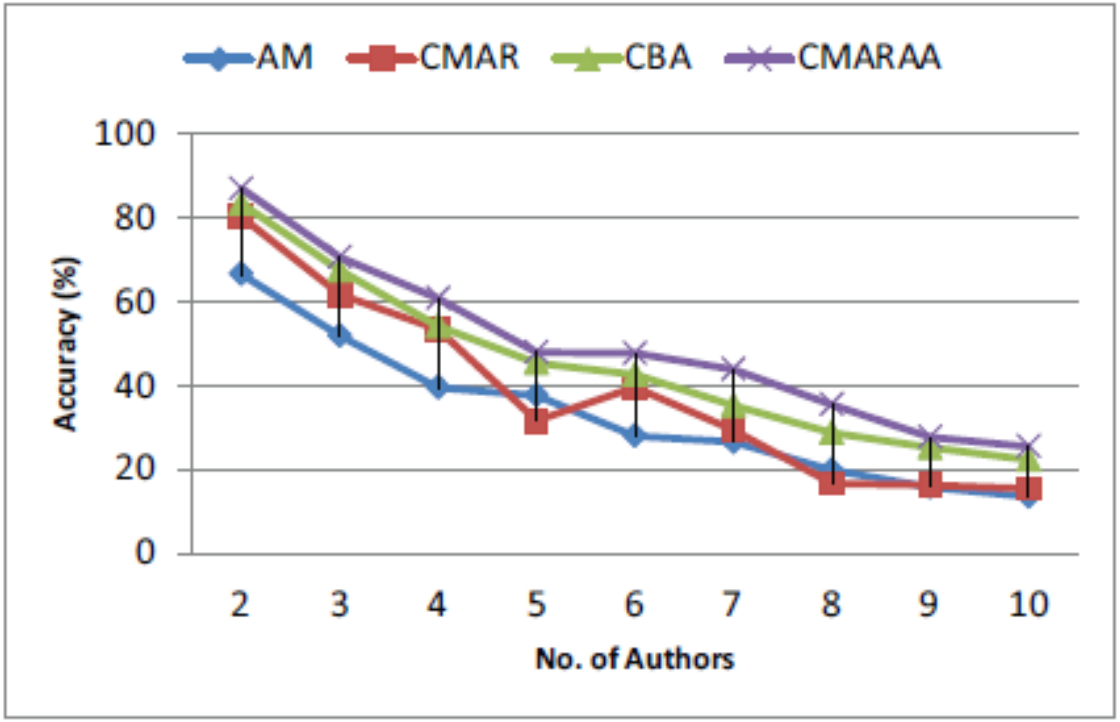
CMARAA: Evaluation

Objectives of evaluation:

- Evaluate the performance of our proposed method CMARAA
- Compare classification accuracy with other authorship analysis methods
- Dataset used in evaluation: Enron e-mail data set
  - 20MB of e-mails
  - 14 authors
  - 50-600 e-mails per author

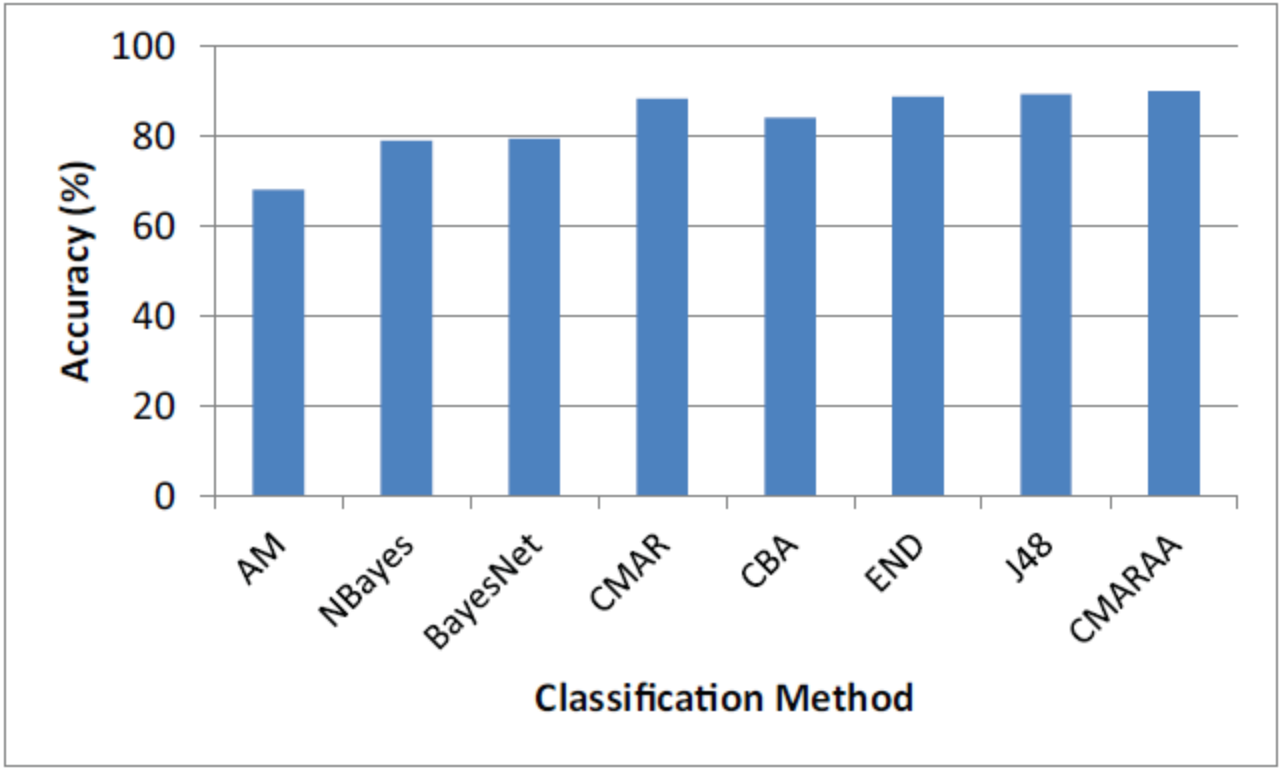
Enron Email Dataset: http://www.cs.cmu.edu/~enron/

CMARAA: Evaluation (Cont.)



AM- AuthorMiner: Frequent Item based AA
CMAR: Classification by Multiple Association Rule
CBA: Classification By Association

CMARAA: Evaluation (Cont.)



Enron Email Dataset: http://www.cs.cmu.edu/~enron/

Sample Writeprint

Combination of stylometric features

Writeprint of Fossum-d (Enron Dataset) contains 86 patterns; two of them are:

{f91:low, f92:low} with support = 83%
{f243:high, f244:high} with support = 78%

where

f91: *ratio of distinct words and total words,*
f92: ratio of special symbols with total characters,
f243: *frequency of the function* word “where",
f244: *frequency of the function word “whether"*

Contributions

- First attempt to utilize associative classification on authorship analysis.
- **Class-based** associative classification ensures that each author is duly represented in the classifier.
- Association rules offer presentable and intuitive evidence.
- Comparable accuracy to other state-of-the-art authorship analysis methods.

References

- Enron Email Dataset: http://www.cs.cmu.edu/~enron/
- R. Agrawal, T. Imieli´nski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD International Conference on Management of Data, pages 207–216, 1993.
- F. Iqbal, B. C. M. Fung, H. BinSalleeh, and M. Debbabi. A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications. Information Sciences: Special Issue on Data Mining for Information Security, Elsevier. (In press) Impact Factor: 3.29
- F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1-2):56-64, October 2010. Elsevier.
- F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Investigation, 5(1):42-51, September 2008. Elsevier.

Conclusion

- Comparable accuracy with interpretable and convincing evidence
- Best suited for authorship verification whereas it can applied to authorship characterization/ profiling
- Applicable to tiny documents such as tweets and SMS.
- Applicable to authorship of other languages such as Arabic and Chinese

Thank You

You may please send your questions to authors

Michael R.Schmid (michael.schmid@concordia.ca)
Farkhund Iqbal (farkhund.iqbal@zu.ac.ae)
Benjamin Fung (ben.fung@mcgill.ca)

Evaluation

❑ Lexical Features

1. character count excluding space characters (M)
2. ratio of digits to M
3. ratio of letters to M
4. ratio of uppercase letters to M
5. ratio of spaces to M
6. ratio of spaces to total characters
7. ratio of tabs to M

8-33. alphabets frequency (A-Z) (26 features)

34-54. frequency of special characters: < > % | { } [ ] / \ @ # ~ + - * $ ^ & _ (21 features)

55. Word count (W)
56. Average word length
57. Average sentence-length in terms of characters
58. Ratio of short words (1-3 characters) to W

69-88. Ratio of word length frequency distribution to W (30 features)

89. Ratio of function words to W
90. Vocabulary richness, i.e., T/W
91. Ratio of Hapax legomena to W
92. Ratio of Hapax legomena to T
93. Ratio of Hapax dislegomena to W

Stylometric Features

❑ Syntactic Features

94-101 Occurrences of punctuations , . ? ! : ; ’ " (8 features)
102. Ratio of punctuations with M
103-252. Occurrences of function words (150 features)

❑ Structural Features

253. Ratio of blank lines/total number of lines within e-mail
254. Sentence count
255. Paragraph count
256. Presence/absence of greetings
257. Has tab as separators between paragraphs
258. Has blank line between paragraphs
259. Presence/absence of separator between paragraphs
260. Average paragraph length in terms of characters
261. Average paragraph length in terms of words
262. Average paragraph length in terms of sentences
263. Contains Replied message
264. Position of replied message in the e-mail
265. Use e-mail as a signature
266. Use telephone as signature
267. Use URL as a signature

Stylometric Features

- Content-specific Features

  268-280. deal, HP, sale, payment, check, windows, software, offer, microsoft, meeting, conference, room, report
  (13 features)