

**Mining writeprints from
anonymous e-mails
for forensic investigation**



**ร.ต.ท.ดลเดช พงศ์เพิ่มทรัพย์
รหัส 52312351**

introduction

- Clustering based on stylometric features
- Preliminary information
- Cluster analysis
- Leading to authorship analysis



Related work

- Stylometric features
- E-mail characteristics
- E-mail cluster analysis



Table 1 – Lexical and syntactic features.

Features type	Features
Lexical: character-based	1. Character count (N) 2. Ratio of digits to N 3. Ratio of letters to N 4. Ratio of uppercase letters to N 5. Ratio of spaces to N 6. Ratio of tabs to N 7. Occurrences of alphabets (A-Z) (26 features) 8. Occurrences of special characters: < > % { } [\ / \ @ # ~ + - * \$ ^ & ÷ (21 features)
Lexical: word-based	9. Token count(T) 10. Average sentence length in terms of characters 11. Average token length 12. Ratio of characters in words to N 13. Ratio of short words (1–3 characters) to T 14. Ratio of word length frequency distribution to T (20 features) 15. Ratio of types to T 16. Vocabulary richness (Yule's K measure) 17. Hapax legomena 18. Hapax dislegomena
Syntactic features	19. Occurrences of punctuations, . ? ! : ; ' " (8 features) 20. Occurrences of function words (303 features)

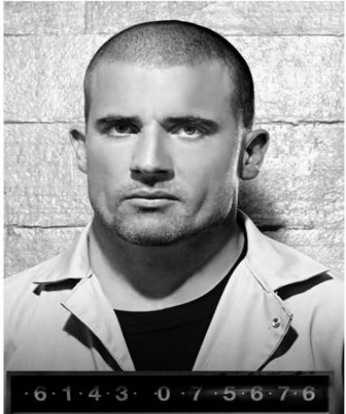
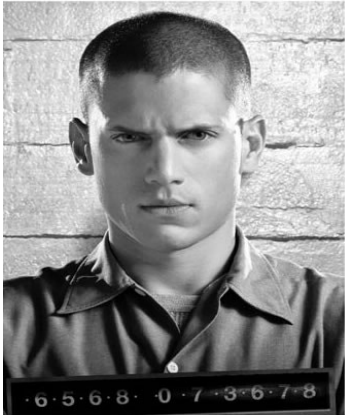
The problem

- Investigator may or may not know the number of suspects in advance.

WANTED FOR PRISON BREAK

MASTERMINDS BEHIND THE ESCAPE
FROM FOX RIVER STATE PENITENTIARY
AND CRIMES AGAINST THE FEDERAL GOVERNMENT

MICHAEL SCOFIELD LINCOLN BURROWS



6 · 5 · 6 · 8 · 0 · 7 · 3 · 6 · 7 · 8 6 · 1 · 4 · 3 · 0 · 7 · 5 · 6 · 7 · 6

Place of Birth: CHICAGO, IL
Height: 6'
Weight: 175 LBS.
Eyes: BLUE
Hair: DARK BROWN
Scars/Tattoos: MISSING ONE TOE ON LEFT FOOT, INTRICATE TATTOO COVERING FRONT AND BACK TORSO/ARMS
Warrant Issued: JOLIET, DISTRICT OF ILLINOIS
Date of Warrant: MONDAY, AUG. 21, 2006, 8:00PMET/PT

Place of Birth: CHICAGO, IL
Height: 6' 2"
Weight: 180 LBS.
Eyes: BLUE
Hair: BROWN
Scars/Tattoos: SCAR UNDER LEFT JAW
Warrant Issued: JOLIET, DISTRICT OF ILLINOIS
Date of Warrant: MONDAY, AUG. 21, 2006, 8:00PMET/PT

NOTICE:
IF YOU HAVE ANY INFORMATION THAT COULD LEAD TO
THE CAPTURE OF THESE TWO MEN, PLEASE CALL:
1-888-808-SEEN

method

- Pre-treatment
- Stylometric Features extraction
- Stylometry-based clustering
- Frequent patterns mining
- Writeprint mining



Disputed anonymous e-mails ensemble E



Phase 1: Pretreatment: cleaning, tokenization, and stemming

Bag of words representation of e-mails using vector space model

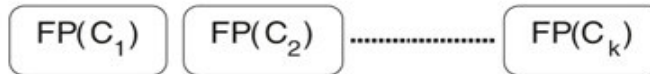
Phase 2: Feature Extraction: lexical, syntactic, structural, and domain-specific features

ARFF (Weka input format): each e-mail represents one row in ARFF file

Phase 3: Discretization & clustering: EM, k-means, and bisecting k-means



Phase 4: Extracting Frequent Pattern (FP) based on user threshold



Phase 5: Extracting WritePrint (WP) by filtering the overlapping patterns

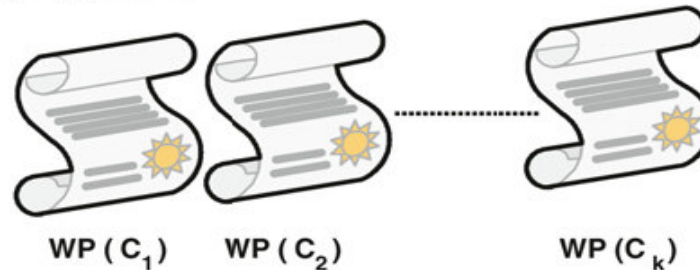


Table 2 – Structural features.

Features type	Features
Structural features	<ol style="list-style-type: none">21. Lines in an e-mail22. Sentence count23. Paragraph count24. Presence/absence of greetings25. Has tab as separators between paragraphs26. Has blank line between paragraphs27. Presence/absence of separator between paragraphs28. Average paragraph length in terms of characters29. Average paragraph length in terms of words30. Average paragraph length in terms of sentences31. Use e-mail as signature32. Use telephone as signature33. Use URL as signature
Domain-specific features	<ol style="list-style-type: none">34. agreement, team, section, good, parties, office, time, pick, draft, notice, questions, contracts, day (13 features)

Experiments and evaluation

- Em
- K-means
- Bisecting k-means



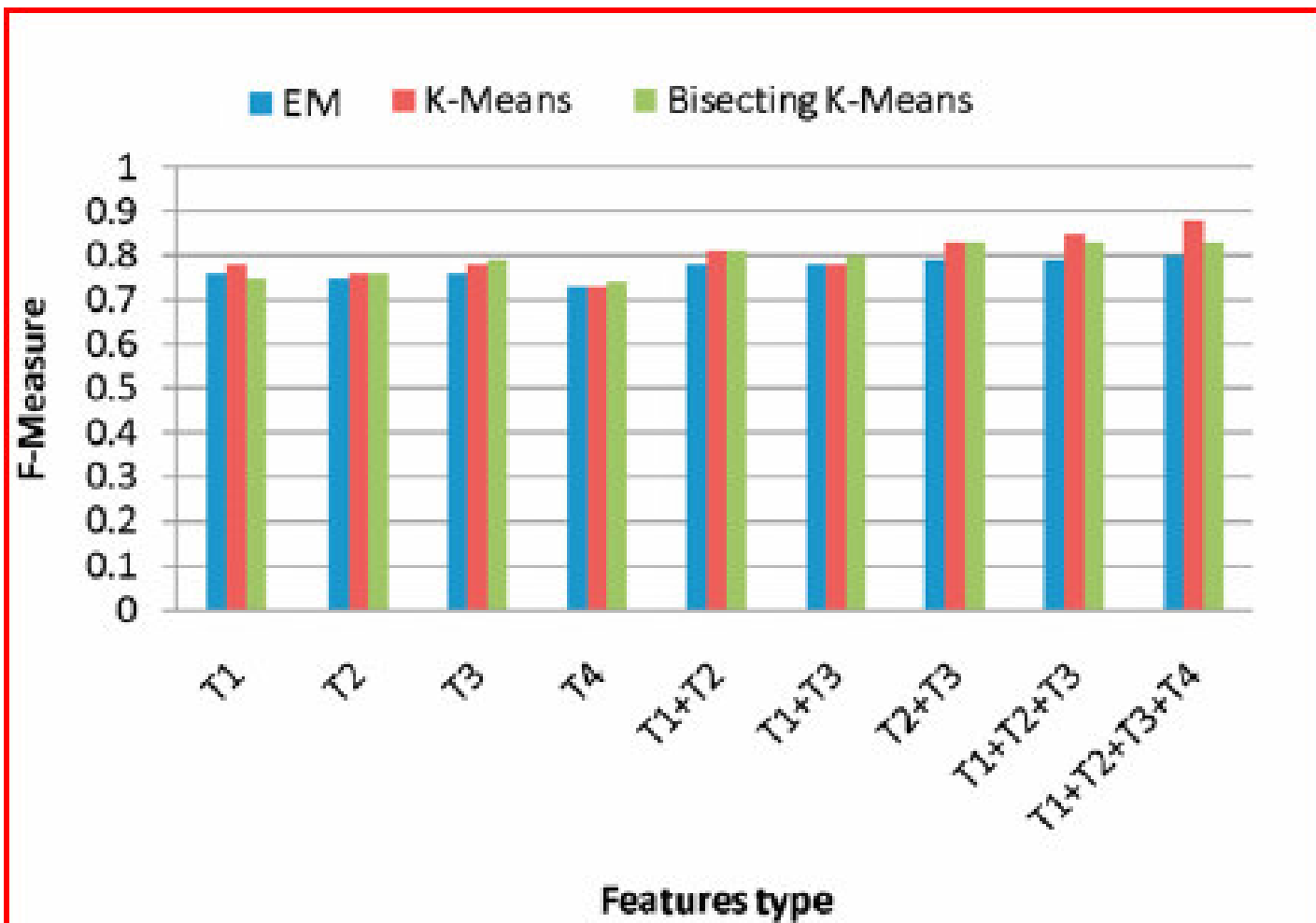


Fig. 2 – F-Measure vs. Feature Type and Clustering Algorithms (Authors = 5, Messages = 40).

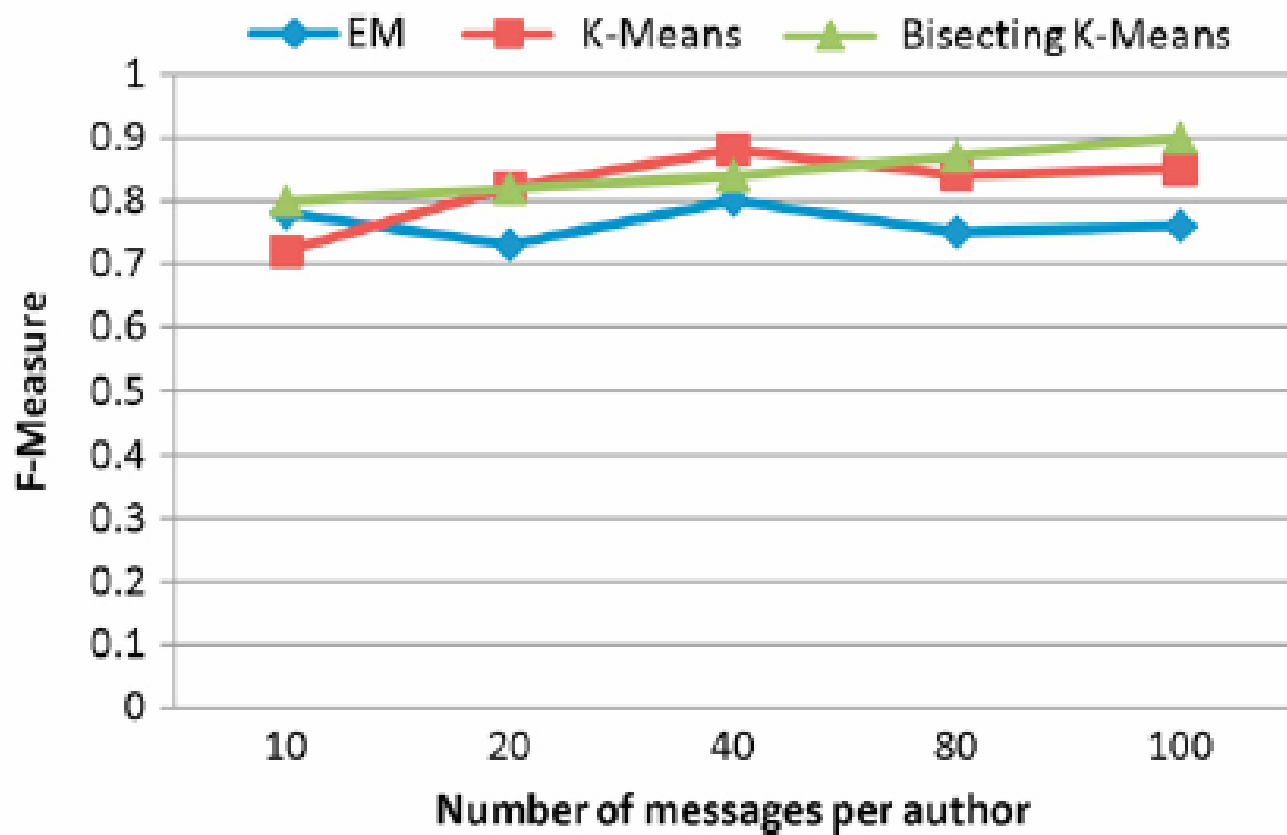


Fig. 3 – F-Measure vs. Features Type and Clustering Algorithms (Authors = 5, Features = $T_1 + T_2 + T_3 + T_4$).

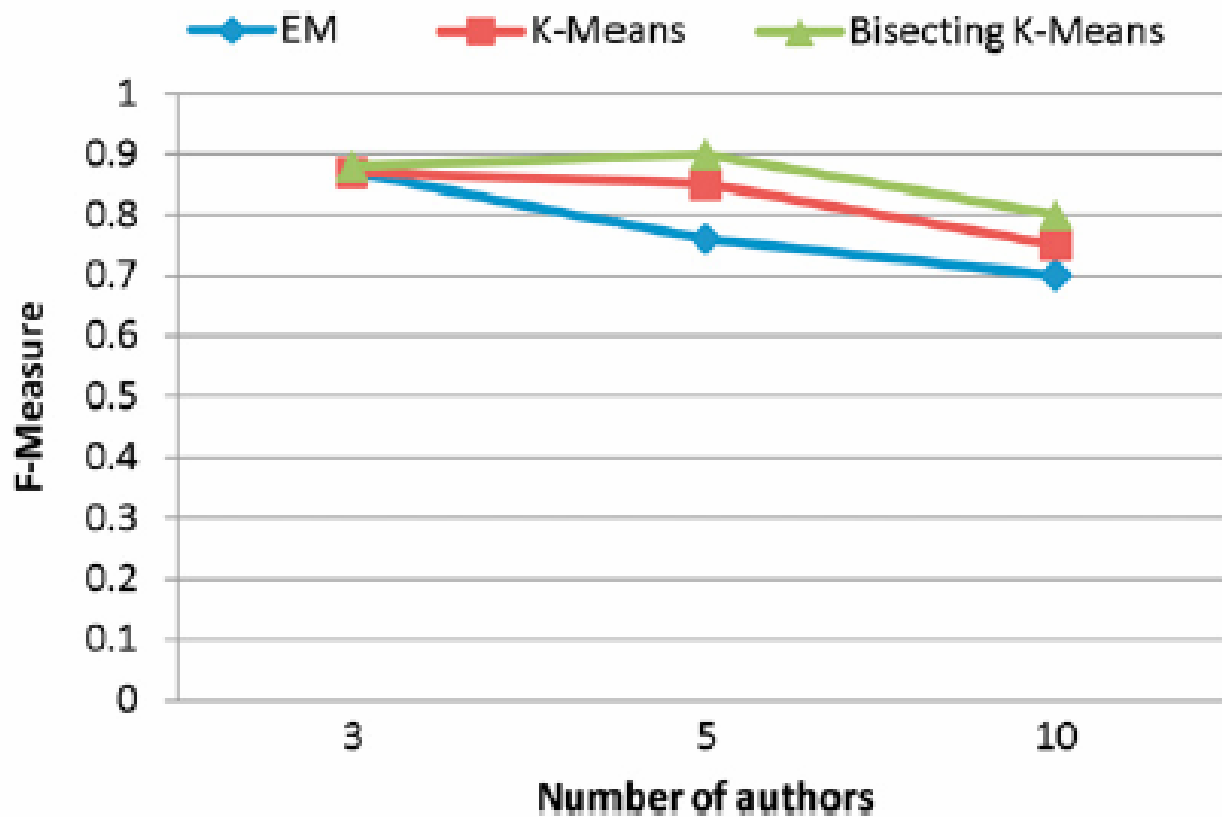


Fig. 4 – F-Measure vs. Number of authors and Clustering Algorithms (Messages = 100, Features = $T_1 + T_2 + T_3 + T_4$).

Conclusion

