การรวบรวมรูปแบบการพิมพ์ของอีเมลล์ที่ไม่ปรากฏชื่อสำหรับการสืบสวนทางนิติวิทยาศาสตร์

Mining writeprints from anonymous e-mails for forensic investigation.

บทคัดย่อ

อาชญากรรมหลายประเภทได้ใช้ประโยชน์ในการไม่ปรากฏชื่อหรือการใช้นามแฝงในโลกไซเบอร์ในการกระทำที่ผิดกฎหมายต่างๆ การใช้อีเมลล์ก็เป็นอีกตัวอย่างหนึ่งที่มีการใช้อย่างแพร่หลายในหลายๆ กิจกรรม การดึง knowledge และข้อมูลจากอีเมลล์ ความสำคัญในการสืบสวนทาง cybercrime และการเก็บหลักฐาน นั่นเป็นสิ่งที่ท้าทายที่สุด และใช้เวลามากที่สุด เนื่องจากลักษณะพิเศษของ e- mail dataset . ในการศึกษานี้มุ่งประเด็นไปที่ปัญหาของการทำเหมืองสไตล์การเขียนจากอีเมลล์ที่ถูกเขียนโดยผู้เขียนที่ไม่ปรากฏชื่อหลายๆ คน โดยมีวัตถุประสงค์หลักคือ การจัดกลุ่มของอีเมลล์ที่ไม่ปรากฏชื่อโดย stylometric features และจากนั้นดึงเอาลักษณะ writeprint ออกมา เช่น เอกลักษณ์สไตล์การเขียน จากแต่ละกลุ่มโดยให้ความสำคัญการแก้ปัญหาที่คณะผู้จัดทำทำขึ้นนี้มีความแตกต่างจากการแบ่งแบบเดิม ซึ่งสัณนิษฐานได้ว่าข้อมูลที่ทำการทดลองนี้สามารถนำไปใช้สร้างรูปแบบการจัดกลุ่มได้ วิธีที่เรานำเสนอสามารถใช้ได้ในขั้นตอนแรกเริ่มของการสืบสวน ซึ่งนักสืบมักจะมีข้อมูลไม่มากเกี่ยวกับ case และผู้เขียนตัวจริงของอีเมลล์ต้องสงสัย การทดลองบนชุดข้อมูลจริงแสดงให้เห็นว่าการแบ่งกลุ่มโดยใช้ writing style มีความน่าสนใจในการนำไปใช้แบ่งกลุ่มอีเมลล์ที่ถูกเขียนโดยผู้เขียนคนเดียวกัน

1. ความเป็นมา

โลกคอมพิวเตอร์จัดให้เป็นรูปแบบที่อำนวยความสะดวกสำหรับอาชญากรรมที่ไม่ปรากฏชื่อ จะนำไปปฏิบัติในทางผิดกฎหมาย เช่น อีเมลล์ขยะ การหลอกลวงทางอินเตอร์เน็ต เป็นต้น อีเมลล์เป็นเครื่องมือสื่อสารที่ใช้กันปกติมากที่สุด ทำให้สูญเสียทั้งด้านการเงินและด้านศีลธรรมจึงกลายเป็นเหยื่อของอาชญากรรมทางอินเตอร์เน็ต ในอีเมลล์ขยะนั้น ยกตัวอย่างเช่น ผู้ร้ายพยายามที่จะปกปิดเอกลักษณ์ของเขา ในทำนองเดียวกันการหลอกลวงทางอินเตอร์เน็ตที่ปลอมเป็นธนากรเพื่อหลอกลวงเหยื่อหรือแม้กระทั่งกลุ่มผู้ก่อการร้ายก็ใช้ประโยชน์จากอีเมลล์เป็นช่องทางในการติดต่อสื่อสารข้อมูลความลับด้วย

เทคนิคในการวิเคราะห์หาผู้เขียนเพื่อประโยชน์ในการพิจารณาคดีนั้น ปัจจุบันอยู่บนพื้นฐานการตรวจพิสูจน์หลักฐานทางดิจิตอล เทคนิคนี้จะสร้างแบบการจำแนกประเภท stylometric features จากตัวอย่างการเขียนของผู้ต้องสงสัยที่อาจเป็นได้และใช้แบบนี้ชี้บ่งเอกลักษณ์ของเอกสารที่ผู้เขียนไม่ปรากฏชื่อ การศึกษาส่วนใหญ่จากผู้เขียนที่แท้จริงจะขัดแย้งกับข้อความที่ไม่ปรากฏชื่อ จึงจำเป็นต้องกำหนดผู้ต้องสงสัยที่อาจเป็นได้ สมมุติฐานอีกอย่างของข้อมูลที่ได้เข้าอบรมนั้นเพียงพอต่อการสร้างแบบการจำแนกประเภท ในทำนองเดียวกันเทคนิคผู้เขียนที่เป็นเอกลักษณ์ ถูกประยุกต์เป็นวัฒนธรรมและมีความเป็นเอกลักษณ์ในกลุ่มประชากร เช่น เพศ อายุ และระดับการศึกษา ของเอกสารที่ไม่ปรากฏชื่อ อย่างไรก็ตามเทคนิคเหล่านี้ต้องการ ข้อมูลขนาดใหญ่ของกลุ่มประชากรตัวอย่างในการจำแนกผู้เขียนออกเป็นหมวดหมู่สำหรับเพศ อายุ ฯลฯ

การศึกษานี้มุ่งประเด็นที่แผนการร้ายทั้งรายชื่อผู้ต้องสงสัยไม่ใช่ตัวอย่างจากการอบรมที่หาสามารถหาได้ง่ายของผู้สืบสวน ตัวอย่างเช่น ในระหว่างขั้นแรกของการสืบสวน ผู้สืบสวนอาชญากรรมอาจจะไม่มีเบาะแสเกี่ยวกับผู้ต้องสงสัยที่อาจเป็นได้ในการไม่ปรากฏชื่อส่งอีเมลล์ การเก็บรวบรวมอีเมลล์ต้องสงสัยที่ไม่ปรากฏชื่อ การเขียนที่อาจเป็นไปได้โดยกลุ่ม unknown ของผู้ต้องสงสัยไม่ใช่เป็นลายลักษณ์อักษรของตัวอย่าง ยิ่งไปกว่านั้นการพิจารณาในศาล อาจจะรู้หรือไม่รู้ชื่อที่แท้จริง

ของผู้เขียน การได้รับรู้ในรูปบบการเขียนของอีเมลล์ที่ไม่ระบุชื่อ ผู้สืบสวนสามารถใช้วัตถุประสงค์ดังนี้ ขั้นแรกแบ่งกลุ่มจากรูปแบบการเขียน ซึ่งผู้วิจัยได้ศึกษาการแบ่งกลุ่มที่เป็นผู้เขียนคนเดียวกัน ดังนั้นการแยก writeprint จากแต่ละกลุ่มอีเมลล์ writeprint เป็นกลุ่มลักษณะรูปแบบที่มีจุดเด่นเฉพาะตัวเพียงพอที่จะทำให้แยกผู้เขียนคนหนึ่งออกจากคนอื่น

ส่วนใหญ่สิ่งที่เขียนลงกระดาษจะแสดงการแบ่งกลุ่มโดยรูปแบบการเขียนจะแสดงว่าเป็นกลุ่มการเขียนโดยผู้เขียนคนเดียวกัน ทฤษฎีของผู้วิจัยจัดให้มีการสืบสวนอาชญากรรมในที่ส่วนลึกลับของรูปแบบการเขียนที่พบในอีเมลล์ที่ไม่ปรากฏชื่อ ซึ่งการแบ่งกลุ่มและแยก writeprint สามารถช่วยการป้อนข้อมูลสำหรับ data mining ระดับสูง การสืบหาความสัมพันธ์การแยกจุดเด่นของ stylometric การแบ่งกลุ่มจะนำมาประยุกต์การแยกแต่ละชนิด(คำหรือศัพท์ การสร้างประโยค โครงสร้างประโยค และเนื้อความที่เฉพาะเจาะจง) ในการทดลองของผู้วิจัยได้ประเมินผลความหลากหลายของจำนวนผู้เขียนและขนาดของกลุ่มการเข้าอบรม การใช้จุดเด่นของ visualization และ browsing ของผู้วิจัยที่พัฒนาเครื่องมือเครื่องใช้ ผู้สืบสวนสามารถค้นกระบวนการของข้อมูลและคุณภาพกลุ่มการแยกได้

ผู้วิจัยได้สรุปข้อสนับสนุนงานวิจัยดังต่อไปนี้

1.1 Clustering based on stylometric features

ได้มีหลักการที่เคยใช้กันมาแบ่งการระบุหัวข้อบทสนทนาจากการเก็บรวบรวมเอกสาร ในทางตรงกันข้าม งานวิจัยนี้จะแนะนำการแบ่งกลุ่มข้อความอีเมลล์ที่ไม่ระบุชื่อโดยรูปแบบการเขียนเป็นผลมาจากผู้เขียนคนเดียวกัน

1.2 Preliminary information

บ่อยครั้งที่ผู้สืบสวนต้องเก็บข้อมูลผู้ต้องสงสัยเพื่อรวบรวมเป็นหลักฐานในการสืบสวนรูปแบบเดิมที่ใช้กันคือ การจัดกลุ่มข้อมูลที่มีรูปแบบเหมือนๆ กัน โดยผู้เขียนแต่ละคนก็จะมีลักษณะเฉพาะจากรูปแบบการเขียน สมมุติฐานว่าผู้เขียนเป็นคนๆ เดียวกัน(หรือเกือบจะเป็นคนๆ เดียวกัน) รูปแบบการเขียนและการแบ่งกลุ่มโดยจุดเด่น stylometric สามารถบอกได้ว่าเป็นอีเมลล์จากผู้เขียนคนเดียวกัน

1.3 Cluster analysis

จะวิเคราะห์โดยใช้ความสัมพันธ์ของความแตกต่างกันของ algorithms ในการประมวลผล ผลของจำนวนผู้ต้องสงสัยเท่ากับจำนวนของข้อความต่อผู้ต้องสงสัยในการแบ่งกลุ่มที่ถูกต้องเป็นที่อยู่ในการศึกษา

1.4 Leading to authorship analysis

Writeprint ของผู้ต้องสงสัยแบ่งโดยการอ้างเหตุผลที่มีขัดแย้งกันในอีเมลล์การไม่ปรากฏชื่อ ส่วนที่เหลือของงานวิจัยนี้จะแสดงดังนี้ : ส่วนที่ 2 ทบทวนรายละเอียดในการวิเคราะห์ผู้เขียน ส่วนที่ 3 ปัญหาทั่วไป ส่วนที่ 4 แสดงเค้าโครงการทำงานของการแบ่งกลุ่มอีเมลล์ ด้วยรูปแบบการเขียน และ writeprint ส่วนที่ 5 พิจารณาประโยชน์ของหลักการที่แท้จริงส่วนที่ 6 สรุปงานวิจัย

2. Related work

ผู้วิจัยได้อธิบายการทบทวนจุดเด่น stylometric ในส่วนที่ 2.1 และแสดงรายละเอียดลักษณะเฉพาะเจาะจงของอีเมลล์ dataset ในส่วนที่ 2.2 เทคนิคการใช้ศิลปะการเขียน ในส่วนที่ 2.3

2.1 Sttylometric features

รูปแบบของลายนิ้วมือ ถูกใช้ให้แสดงความเป็นเอกลักษณ์บุคคล ในปัจจุบันเป็นยุคคอมพิวเตอร์ ธรรมชาติของอาชญากรรมส่วนใหญ่มีเครื่องมือในการกระทำความผิดจึงได้เปลี่ยนไปด้วย เครื่องมือที่ใช้กันมาไม่นานนักก็จะมีการประยุกต์ใช้ในการพิจารณาคดีอาชญากรรมของโลกคอมพิวเตอร์ในชั้นศาล รูปแบบหรือการศึกษาจุดเด่น stylometric จะแสดงความเป็นอัตลักษณ์บุคคลที่จะแสดงความสัมพันธ์ที่สอดคล้องกันของการใช้คำที่จะบอกตัวตน การจัดรูปประโยคและย่อหน้าและองค์ประกอบของประโยคในย่อหน้า และในย่อหน้าของเอกสารด้วย

แม้ว่าไม่มีกลุ่มที่มีลักษณะเด่นที่เหมาะสมและความสามารถในการประยุกต์ใช้กับคนทุกคนและในกลุ่มคนส่วนใหญ่ แต่อย่างไรก็ตามการศึกษาตัวผู้เขียนก่อนที่จะดูเกี่ยวกับคำหรือศัพท์ ,การสร้างประโยค โครงสร้าง และ ลักษณะเด่นเฉพาะรายละเอียดโดยย่อและความสัมพันธ์ของการจำแนกในแต่ละชนิดของลักษณะดังแสดงข้างล่าง

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)

ลักษณะเกี่ยวกับคำหรือศัพท์จะทำให้เรียนรู้เกี่ยวกับการให้ความสำคัญลำดับแรกของการแยกความเป็นตัวตน และคำของปัจเจกบุคคล บางครั้งจะใช้ลักษณะเด่นพื้นฐานปัจเจกบุคคล แสดง 1-8 ในตารางที่ 1 ความถี่ของตัวอักษรเฉพาะบุคคล (ภาษาอังกฤษ 26 ฉบับ) ด้านบนทั้งหมดเป็นกรณีศึกษา จดหมายสำคัญที่ใช้ในการเริ่มต้นประโยค ค่าเฉลี่ยจำนวนความเฉพาะตัวต่อคำ และค่าเฉลี่ยจำนวนความเฉพาะตัวต่อประโยค การใช้ลักษณะเด่นบ่งชี้ความชื่นชอบของปัจเจกบุคคลสำหรับความพิเศษของความเฉพาะตัวที่แน่นอน หรือสัญลักษณ์หรือสิ่งที่ชอบสำหรับทางเลือกที่แท้จริงเพียงหนึ่งเดียว ตัวอย่างเช่น บางคนชอบใช้สัญลักษณ์ '$' แทน 'dollar', '%' แทน 'percent', และ '#' แทนการเขียนคำว่า 'number'.

ลักษณะเด่นของพื้นฐานคำจะรวมคำแยกความยาวของประโยค คำต่อประโยค และศัพท์ที่หรู ได้เคยมีการศึกษาถึงผู้เขียนมาก่อนแล้ว ปัจจุบันได้มีการศึกษาอีเมลล์วิเคราะห์หาผู้เขียนระบุคำพื้นฐาน เช่น ศัพท์ที่หรูไม่ได้มีแค่สองเหตุผลนี้ ข้อหนึ่งข้อความของอีเมลล์และเอกสารสั้นมากเมื่อเทียบกับอักษรและโคลงกลอน ข้อสองลักษณะเด่นสำนวนของคำส่วนมากดูที่คำแวดล้อมและสามารถรู้ว่าถูกควบคุมโดยประชากร

ลักษณะเด่นเกี่ยวกับการสร้างประโยคถูกเรียกการเน้นรูปแบบที่ประกอบด้วยวัตถุประสงค์การใช้คำ เช่น 'though', 'where', 'your', การใช้เครื่องหมายวรรคตอน เช่น '!' และ ':', ประเภทของคำลงท้าย และยัติภังค์ (ดูตารางที่ 1). Mosteller และ Wallace (1964)เป็นกลุ่มแรกที่แสดงประสิทธิผลของสิ่งที่เรียกว่าคำที่มีความหมายตามโครงสร้างของ addressing ฉบับของงานวิจัย Federalist Burrows (1987) เคยใช้ 30-50 ตัวอย่าง คำที่มีความหมายตามโครงสร้างสำหรับผู้เขียน . ต่อมาภายหลังได้ศึกษาการจำแนกให้ถูกต้องของการใช้เครื่องหมายวรรคตอน และคำที่มีความหมายตามโครงสร้าง Zheng et al. (2006) เคยพบคำที่มีความหมายตามโครงสร้างมากกว่า 300 คำ Stamatatos et al. (2000) เคยพบความถี่ประเภทของคำลงท้าย, การพิจารณาและการเปลี่ยนรูปเป็นสรรพนาม สำหรับการวิเคราะห์ผู้เขียน และประเภทของเอกสารที่ใช้ในการระบุตัวบุคคล

ลักษณะเด่นของโครงสร้างประโยคจะช่วยในการเรียนรู้เกี่ยวกับการรวบรวมความเป็นปัจเจกบุคคลอย่างไรให้เป็นแผนงานและเป็นโครงสร้างเอกสารของเขา/ของเธอ เช่น การรวบรวมประโยคอย่างไรภายในย่อหน้าและย่อหน้าภายในเอกสาร ลักษณะเด่นของโครงสร้างประโยคขั้นแรกจะถูกแนะนำ โดย de Vel et al สำหรับผู้เขียนที่เป็นเจ้าของอีเมลล์ ในทำนองเดียวกันลักษณะเด่นของโครงสร้างทั่วไป ผู้วิจัยเคยใช้ลักษณะเด่นเฉพาะของอีเมลล์ เช่น การมี/ไม่มี คำทักทายและคำลา เป็นที่น่าสังเกตและตำแหน่งภายในตัวของอีเมลล์ ยิ่งไปกว่านั้นบางคนจะใช้ชื่อ ก่อน/หลัง เป็นลายเซ็นต์ ซึ่งจะมีมากกว่าชื่ออาชีพและที่อยู่ภายในอีเมลล์ อีเมลล์ที่ประสงค์ร้ายจะไม่มีลายเซ็นต์ และในบางกรณีจะเป็นลายเซ็นต์ปลอม

ลักษณะเด่นของความเฉพาะของหัวข้อเป็นการใช้กิจกรรมเฉพาะที่แท้จริง การโต้แย้งในที่ประชุมหรือกลุ่มที่สนใจ โดย keywords หรือ terms 2-3 คำ ตัวอย่างเช่น คนที่เกี่ยวข้องในอาชญากรรมทางคอมพิวเตอร์ โดยทั่วไปจะใช้ 'sexy', 'snow', 'download', 'click here' and 'safe' ฯลฯ โดยปกติ term การแบ่งออกเป็นกลุ่มจะสร้าง domain หนึ่งๆ ที่ไม่สามารถประยุกต์ใช้ใน domain อื่น และการแปรผันที่เท่ากันจากบุคคลหนึ่งไปอีกบุคคลในบาง domain . Zheng et al. ใช้ 11 keywords จากอาชญากรรมทางคอมพิวเตอร์ taxonomy ในการวิเคราะห์การทดลองหาผู้เขียน ยิ่งไปกว่านั้นรายการที่ครอบคลุมลักษณะรูปแบบการเขียนที่เฉพาะเจาะจง จะใช้ใน Abbasi and Chen (2008).

ลักษณะเด่นของ Idiosyncratic include ประกอบด้วยการสะกดคำผิดทั่วๆ ไป เช่น จาก 'f' เป็น 'ph' ในการหลอกลวงทางอินเตอร์เน็ต และ โครงสร้างไวยกรณ์ที่ผิด เช่น ประโยคที่รูปแบบคำกริยาไม่ถูกต้อง รายการของความเฉพาะตัวที่มากมายหลากหลายจากบุคคลยังบุคคล และเป็นการยากในการควบคุม Gamon (2004) อ้างว่ามีความแม่นยำสูง โดยการรวมลักษณะเด่นที่แท้จริงกับประเภทของคำ ,ความถี่ของคำที่มีความหมายตามโครงสร้าง และลักษณะเด่นที่ได้จากความหมายทางภาษา

2.2 E-mail characteristics

การประยุกต์เทคนิคการวิเคราะห์ผู้เขียนเป็นการท้าทายความสามารถมากกว่าประวัติศาสตร์และเอกสารเกี่ยวกับการประพันธ์วรรณคดี ซึ่งมีขนาดใหญ่ในการเก็บรวบรวมจึงประกอบด้วยหลายส่วน , ส่วนย่อยๆ และย่อหน้า ตามข้อจำกัดของไวยกรณ์และรูปแบบการจัดวางองค์ประกอบ อีเมลล์ในทางกลับกันที่สั้นในส่วนที่ยาวโดยปกติแล้วจะมี2-3 ประโยคหรือคำ ดังนั้นจึงเป็นการยากที่จะเรียนรู้ถึงพฤติกรรมการเขียนของคนจากอีเมลล์ของเขาเหล่านั้น ตัวอย่างเช่น การวิเคราะห์ผู้เขียนทำให้เป็นที่ยอมรับจะไม่จำเป็นสำหรับเนื้อความที่น้อยกว่า 500 คำ

อีเมลล์ที่ไม่เป็นทางการจะมีการเขียนในรูปแบบที่ไม่ค่อยใส่ใจที่จะสะกดให้ถูกต้องและหลักไวยกรณ์ที่ไม่ถูกต้อง ดังนั้นเทคนิคการวิเคราะห์นั้นจะประสบความสำเร็จในการวิเคราะห์ผู้เขียนสำหรับวรรณกรรมและการเก็บรวบรวมในอดีตอาจไม่มีการเก็บเป็น dataset

ลักษณะบางอย่างของเอกสารเป็นแหล่งข้อมูลมากมาย หัวอีเมลล์จะมีข้อมูลที่บอกที่มา, เวลา,ข้อมูลอีเมลล์ที่ส่ง, ผู้ส่งและที่อยู่ผู้รับ และการโต้ตอบของผู้รับ บางข้อความจะประกอบด้วยเนื้อความที่แนบมาด้วยหนึ่งหรือมากกว่านั้น ดังนั้นข้อมูลในทำนองเดียวกันนี้จะช่วยในการเรียนรู้เกี่ยวกับรูปแบบการเขียนและพฤติกรรมของผู้ใช้

2.3 E-mail cluster analysis

การเก็บหลักฐานที่น่าเชื่อถือของอาชญากรรมทางคอมพิวเตอร์ จะตรวจพิจารณาการประพฤติปฏิบัติที่หลากหลายประเภทในการวิเคราะห์ ตัวอย่างเช่น การกู้คืนข้อมูลจากข้อมูลเหล่านั้นเกี่ยวกับยา หนังสือลามก การhacking หรือ ลัทธิก่อการร้าย ฯ ซึ่งได้มาจากค้น keyword หรือ ประสิทธิภาพมากมายโดยใช้แบบดั้งเดิม เช่นเดียวกันกับผู้สืบสวนต้องการจินตนาการรูปแบบข้อมูลทั่วไปในการสื่อสารของผู้ต้องสงสัย การระบุผู้เขียนที่ถูกต้องขัดแย้งกับอีเมลล์ที่ไม่ปรากฏชื่อจึงต้องเรียนรู้กลไกที่แตกต่างกันของเทคนิค

Holmes and Forsyth (1995) และ Ledger and Merriam (1994) ได้ริเริ่มและประยุกต์การจัดเทคนิคการแบ่งกลุ่ม dataset หลังจาก Baayen et al. (1996) แสดงการแบ่งกลุ่ม stylometric ในการกำหนดผู้เขียน ผู้วิจัยพิจารณลักษณะเด่นของ data-driven เท่านั้น ในส่วนของ Aaronson (1999),จะประกอบด้วยคำที่ใช้บ่อย ,จดหมายที่เขียนบ่อย และความยาวของประโยคเป็นต้น และได้ศึกษาผลของลักษณะ data-driven, ลักษณะการสร้างประโยคโดยจะดูหลักไวยกรณ์โดยการใช้โปรแกรมตัดคำภาษา ที่ได้อ้างว่ามีความแม่นยำกว่าสมัยก่อน

Abbasi and Chen (2008) ได้ศึกษาผลของลักษณะเด่นสำหรับ stylometric ในการวัดความคล้ายคลึงโดยการใช้ Principal Component Analysis (PCA) และเทคนิคใหม่ที่เรียกว่า Writeprints. การเข้าใจที่ดีที่สุดไม่ใช่เพียงแค่การหาที่อยู่ทั้งหมด ที่จะบอกข้อคำถามในปัญหาที่มี

Li et al. (2006a) ได้ประยุกต์การแบ่งกลุ่มอีเมลล์โดยการใช้ algorithm ใช้ป้อนหัวข้อเป็น โปรแปรมตัดคำ Natural Language (NL) ซึ่งการประมวลผลเป็นการใช้วัตถุประสงค์ algorithm ที่กล่าวสรุปและเรียกว่า Generalized Sentence Patterns (GSP). การใช้ GSP เป็นตัวแบ่งชั้นความเท็จ การแบ่งกลุ่มนี้จะเป็นการจัดการวิธีหนึ่ง การทำงานของ Li et al. (2006a) ทำให้จำกัดความซ้ำซ้อนของอีเมลล์จาก GSP

ระบบของอินเตอร์เน็ตที่ใช้สั่งซื้อสินค้า online , มีการใช้นามแฝงมากมายในคนๆ เดียว วัตถุประสงค์ใหม่ในการระบุ algorithm เมื่อสองนามแฝงเป็นคนเดียวกัน เทคนิคนี้ประสบความสำเร็จในการแยกตำแหน่งได้แม่นยำมากกว่า 90% ที่อยู่ของอี

เมลล์บางฉบับไม่มีลักษณะเฉพาะ เป็นวัตถุประสงค์ของเทคนิคใหม่ ที่เรียกว่า writeprints สำหรับระบุผู้เขียนและการตรวจวัดความคล้ายคลึงกัน ผู้วิจัยเสนอลักษณะเด่นรายการที่ประกอบด้วย idiosyncratic ในการทดลอง การตรวจวัดส่วนที่คล้ายคลึงกันของการไม่ปรากฏชื่อและเปรียบเทียบกับเอกลักษณ์อื่นๆ และการคำนวณคะแนน ถ้าคะแนนสูงค่าก่อนกำหนดเอกลักษณ์จะมีเอกลักษณ์ตรงกัน

3. The problem

ปัญหาที่พบในการเก็บหลักฐานจากอีเมลล์ที่ไม่ปรากฏชื่อนั้นต้องใช้ความเข้าใจกับรูปแบบการเขียนอีเมลล์และการเก็บรวบรวมต้องระบุกลุ่มหลักของรูปแบบการเขียนที่เรียกว่า writeprints $\{WP_1,..,WP_k\}$ ในงานวิจัยนี้ผู้วิจัยพัฒนาเครื่องมือสำหรับผู้สืบสวนมองภาพพจน์และค้นหารูปแบบการเขียน ,ค้นหาในการเก็บข้อมูลของอีเมลล์ที่ไม่ปรากฏชื่อ

ผู้วิจัยวัดความสามารถในการจำแนกของลักษณะเด่นของ stylometric ในข้อมูลของอีเมลล์ ตัวอย่างเช่น ถ้าเก็บข้อมูลอีเมลล์แตกต่างก็เป็นการเขียนหัวข้อที่แตกต่างกันชัดเจน ลักษณะเด่นเฉพาะของเนื้อความอาจให้ผลการแบ่งกลุ่มที่ดีกว่าสัญลักษณ์บอกรูปแบบ การศึกษานี้มุ่งประเด็นการคำนวณความแตกต่างของการแบ่งกลุ่ม algorithms และการตวจวัด algorithm ที่เหมาะสมมากในแผนงานที่เฉพาะเจาะจง ยิ่งไปกว่านั้น ผู้วิจัยจะช่วยให้ผู้ใช้เข้าใจโครงสร้างภายในคลังข้อมูลของอีเมลล์ใน term ของผู้เขียนที่มีลักษณะแตกต่างกัน และตัดสินใจทำอย่างไรให้ลดข้อจำกัดลงในการสืบสวน

4. Our Method

แนวคิดทั่วไปอธิบายไว้ในรูปที่ 1 สามารถสรุปได้ 5 phase 1) Pretreatment : ประกอบด้วยการแยกตัวอีเมลล์ และประยุกต์มาตรฐานเทคนิคขบวนการที่สมบรูณ์แสดงเครื่องหมายและที่มาที่สุดท้ายของ phase แรก 2) Stylometric feature extraction : เป็นการใช้การระบุการเข้ารูปแบบการเขียนที่พบในอีเมลล์ที่ไม่ปรากฏชื่อแต่ละอีเมลล์เปลี่ยนเป็นปัจจัยด้านจำนวน 3) Stylometry-base clustering : เป็นการประยุกต์ในการระบุกลุ่มใหญ่ของรูปแบบการเขียนของผู้เขียนที่แตกต่างกัน 4) Frequent patterns mining : เป็นการเปิดเผยการซ่อนความเชื่อมโยงในรูปแบบการเขียนที่แตกต่างกัน 5) Writeprint mining : จัดให้มีการเขียนอีเลลีที่ได้จากสามphaseนี้ เป็นการเขียนโดยคนๆ เดียวกัน ผู้วิจัยก็จะสามารถแยก writeprint จากแต่ละกลุ่มออกเป็นผู้เขียนคนเดียวกัน

4.1 Pre-treatment

แต่ละอีเมลล์มีแนวโน้มจะเปลี่ยนความเฉพาะตัว การใช้ java สัญลักษณ์ API , แต่ละแนวโน้มความเฉพาะตัวจะเปลี่ยนเป็นสัญลักษณ์หรือคำ ต่างจากพื้นฐานหัวข้อการแบ่งกลุ่มในลักษณะเด่นของการสร้างประโยค ที่ผู้วิจัยคำนวณลักษณะเด่น ในการทดลอง ผู้วิจัยใช้คำที่มีความหมายตามโครงสร้างมากกว่า 300 คำ แสดงรายการในตารางที่ 1 คำๆหนึ่งจะเห็นรูปแบบที่แตกต่างจึงเป็นการเพิ่มแบบในการจัดเก็บข้อมูล ไปยังจุดหมายเดียวกันที่การเปลี่ยนแปลงรากศัพท์บางคำ, การประยุกต์การค้นรากศัพท์ของคำๆหนึ่ง ซึ่งมีการเปลี่ยนแปลงตามไวยกรณ์ และตามการเปลี่ยนแปลงคำ ซึ่งเป็นที่นิยมโดยการใช้ data mining และ การสื่อสารแบบ Natural language Processing(NLP) ผู้วิจัยประยุกต์โดยวางแผนการเพิ่มหลักเกณฑ์

แน่นอนว่าการจัดลำดับคำมักจะเป็น'United States of America' และ 'United Arab Emirates' เป็นต้น ซึ่งบ่อยครั้งที่จะพบพ้อมกัน ดังนั้นผู้วิจัยจึงพัฒนาเกณฑ์ในการวัดอัตโนมัติจัดลำดับและปฏิบัติเป็นสัญลักษณ์เดียวกัน จะช่วยลดแบบในการจัดเก็บข้อมูล การใช้เกณฑ์การวัดการเว้นระยะแทน แต่ละอีเมลล์ μ_iเปลี่ยนเป็นปัจจัยการออกแบบในการจัดเก็บข้อมูล μ_i $\{F1,..,Fn\}$ หนึ่งในอีเมลล์ทั้งหมดจะเปลี่ยนเป็นปัจจัยของลักษณะเด่น กระบวนการปรับปรุงโครงสร้างข้อมูลของฐานข้อมูลที่มี

ความซ้ำซ้อนให้อยู่ในรูปแบบที่เป็นบรรทัดฐาน และวัตถุประสงค์คือจำกัดค่าของลักษณะเด่นแท้[0,1] และหลีกเลี่ยงการประเมินบางคุณลักษณะอื่นๆ ที่มากเกิน

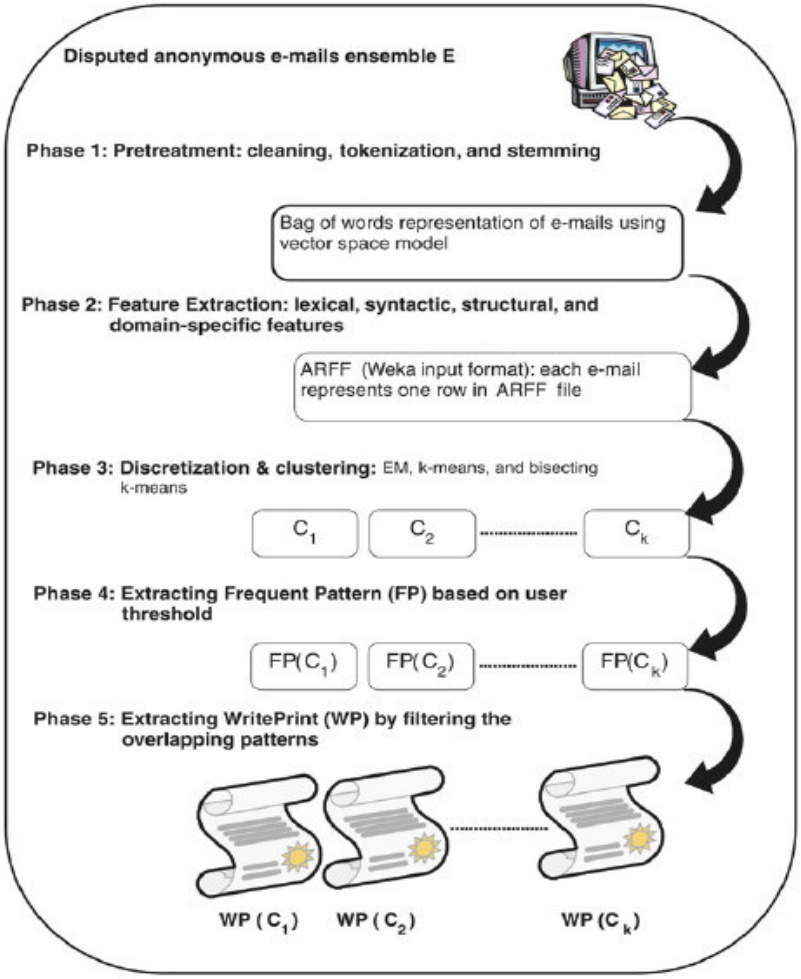

Fig. 1 – Mining WritePrints {WP_1,…,WP_k} from anonymous e-mails E.

4.2 Feature extraction

จำนวนทั้งหมดของ stylometric ถูกค้นพบเกินกว่า 1000 ลักษณะ ในการทดลองผู้วิจัยใช้ 419 ลักษณะ (แสดงในตารางที่ 1 และ ตารางที่ 2) โดยทั่วไปแล้วจะแบ่งเป็น 2 ชนิดของลักษณะ ชนิดแรกเป็นค่าเกี่ยวกับตัวเลข เช่น ความถี่ของความเฉพาะตัวของบางปัจเจกบุคคล และเครื่องหมายวรรคตอน ค่าเกี่ยวกับตัวเลขเป็นกระบวนการปรับปรุงโครงสร้างข้อมูลของฐานข้อมูลที่มีความซ้ำซ้อนให้อยู่ในรูปแบบที่เป็นบรรทัดฐานถึง [0,1] โดยการแบ่งสิ่งที่เกิดขึ้นทั้งหมดของรายการลักษณะเด่นที่สูงสุด กระบวนการปรับปรุงโครงสร้างข้อมูลของฐานข้อมูลที่มีความซ้ำซ้อนให้อยู่ในรูปแบบที่เป็นบรรทัดฐานเป็นการประยุกต์ตลอดการเก็บรวบรวมของอีเมลล์ทั้งหมด ชนิดที่สองเป็นค่าของเลขฐานสอง เช่น ไม่ว่าจะเป็นอีเมลล์ที่มีการทักทายหรือไม่แน่นอนว่าลักษณะเด่นนั้นจะคำนวณโดยการประยุกต์ใช้การวัด Yule's K ของคำศัพท์ที่หรูหรา

บางลักษณะจะถูกสกัดแยกโดยการคำนวณอัตราส่วนของลักษณะอื่นๆ ที่ทราบ ตัวอย่างเช่น การคำนวณอัตราส่วนความยาวของคำ ความถี่ในการกระจายตัวทั้งหมดของจำนวนคำ(W) เป็นการพิจารณาแยกลักษณะ ลักษณะหนึ่งในการสกัดแยก, แต่ละอีเมลล์เป็นปัจจัยสำหรับค่าลักษณะ ในการศึกษานี้ผู้วิจัยมุ่งประเด็นมากที่ลักษณะโครงสร้างซึ่งมีหน้าที่สำคัญในการแยกรูปแบบการเขียน

การชี้ลักษณะเด่น 1-8 ที่เกี่ยวข้องในการคำนวณความถี่ของการเขียนเฉพาะตัว ส่วนบนของจดหมายในการเริ่มต้นประโยคมีความสำคัญกับการแยก ต่างกับคำที่มีช่วงความยาว 1-3 ตัวตน ส่วนมากเป็นบริบทอิสระและจะพิจารณาการแยกลักษณะ ความถี่ของคำสำหรับหลายความยาว 1-20 ตัวตน มีความสำคัญกับการแยก Hepax Legomena และ Hepax dislegomena เป็น term ที่ใช้สำหรับการเกิดขึ้นครั้งที่หนึ่งและการเกิดขึ้นครั้งสองของคำ การกล่าวเริ่มต้นผู้วิจัยเคยใช้มากกว่า 300 ของคำที่มีความหมายตามโครงสร้าง

Table 2 – Structural features.

Features type	Features
Structural features	21. Lines in an e-mail 22. Sentence count 23. Paragraph count 24. Presence/absence of greetings 25. Has tab as separators between paragraphs 26. Has blank line between paragraphs 27. Presence/absence of separator between paragraphs 28. Average paragraph length in terms of characters 29. Average paragraph length in terms of words 30. Average paragraph length in terms of sentences 31. Use e-mail as signature 32. Use telephone as signature 33. Use URL as signature
Domain-specific features	34. agreement, team, section, good, parties, office, time, pick, draft, notice, questions, contracts, day (13 features)

ลักษณะโครงสร้างประโยคแสดงที่ 21 ในตารางที่ 2 เป็นชนิดของข้อมูลที่แสดงถึงการตัดสินใจแบบตรรกะ การตรวจสอบไม่ว่าอีเมลล์จะเป็นคำกล่าวต้อนรับและคำทักทายคำอำลา ตัวแยกย่อหน้าสามารถเว้นว่าง หรือต้องเว้นวรรค หรืออาจจะไม่แยกระหว่างย่อหน้า สำหรับลักษณะเฉพาะของหัวข้อ ผู้วิจัยเลือก 13 คำที่ใช้บ่อยมาก จาก enron e-mail dataset แสดงรายการที่ 34 ในตารางที่ 2

4.3 Clustering

Cluatering เป็นกระบวนการจัดกลุ่มที่เหมือนกันไว้ด้วยกัน โดยสัญชาตญาณผลของการแก้ปัญหากลุ่มจะมีความคล้ายคลึงกันภายในกลุ่มสูง แต่ภายนอกมีความคล้ายคลึงกันน้อย ในการวิจัยการศึกษาของผู้วิจัย, อีเมลล์จะมีความคล้ายคลึงกันในรูปแบบการเขียนเหมือนกัน

วัตถุประสงค์ของผู้วิจัยคือคำนวณโดยการใช้การแบ่ง 3 กลุ่ม algorithms : Expectation Maximization (EM) , k-mean และ bisecting k-mean โดยทั่วไปแล้วจะใช้สูตร ส่วน Expectation Maximization (EM) วัตถุประสงค์แรกจะพบการใช้บ่อยครั้งในการพยากรณ์ค่า K ตัวอย่างเช่น การวิเคราะห์พิสูจน์อีเมลล์ที่ไม่ปรากฏชื่อ ผู้สืบสวนอาจจะไม่ทราบจำนวนผู้เขียนทั้งหมดภายในข้อมูลที่เก็บรวบรวมไว้ ยิ่งกว่านั้น ผู้ใช้ต้องทำให้ผลถูกต้องที่ได้มาจาก k-mean หรือ bisecting k-mean

การวัดที่ไม่มีสิ่งเจือปนสำหรับผลที่มาจากตัวแบ่งกลุ่มและผลที่สมบูรณ์ของการทดลอง โดยทั่วไปจะใช้สูตรที่เรียกว่า F-measure ซึ่งได้มาจากความแม่นยำและการกู้กลับโดยที่การวัดความเที่ยงตรงจากการใช้ขอบข่าย Information Retrieval (IR) สามสิ่งนี้จะแสดงตามสมการทางคณิตศาสตร์ ดังนี้

$$recall(N_p, C_q) = \frac{O_{pq}}{|N_p|} \quad (1)$$

$$precision(N_p, C_q) = \frac{O_{pq}}{|C_q|} \quad (2)$$

$$F(N_p, C_q) = \frac{2 * recall(N_p, C_q) * precision(N_p, C_q)}{recall(N_p, C_q) + precision(N_p, C_q)} \quad (3)$$

ซึ่ง O_{pq} เป็นจำนวนสมาชิกกลุ่มแท้จริงของ N_p ในตัวแบ่งกลุ่มC_q , N_p เป็นกลุ่มแท้จริงของข้อมูล O_{pq} และ C_q เป็นกำหนดกลุ่มO_{pq}

ผู้วิจัยได้พัฒนา software toolkit สามารถช่วยปฏิบัติบอกรูปแบบการเขียนทั้งหมดเป็นกระบวนการ โดย GUI เป็นตัวประสานช่วยในการเลือกลักษณะเด่น การเลือก algorithms และการเลือก parameter (เช่น จำนวนของการแบ่งกลุ่ม เป็นต้น) สิ่งนี้จะช่วยประเมินความสัมพันธ์ที่แน่นอนสำหรับแต่ละชนิดของรูปแบบการเขียน ในการจำแนกรูปแบบที่แตกต่างกันของบุคคล เครื่องมือ software ของผู้วิจัยใช้เปรียบเทียบความแตกต่างของการแบ่งกลุ่ม algorithms และเลือก algorithms สำหรับอีเมลล์ dataset โดยกำหนดการปรับปรุง algorithms ภายในเนื้อความ

Table 3 — Feature items extracted from e-mail clusters of ensemble E.

Cluster C	Message μ	Feature F_1			Feature F_2			Feature F_3		
		$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	$F_{3,1}$	$F_{3,2}$	$F_{3,3}$
C_1	μ_1	0	1	0	0	0	1	0	0	1
C_1	μ_2	0	1	0	0	0	1	0	0	1
C_1	μ_3	0	1	0	0	1	0	0	0	1
C_1	μ_4	1	0	0	0	0	1	0	0	1
C_2	μ_5	1	0	0	0	1	0	0	1	0
C_2	μ_6	1	0	0	0	1	0	0	0	1
C_2	μ_7	1	0	0	1	0	0	0	0	1
C_3	μ_8	0	1	0	1	0	0	1	0	0
C_3	μ_9	0	0	1	1	0	0	1	0	0
C_3	μ_{10}	0	1	0	1	0	0	0	1	0
C_3	μ_{11}	0	1	0	1	0	0	1	0	0

4.4 Mining Frequent Patterns (FP)

ครั้งหนึ่งตัวแบ่งกลุ่มในรูป $\{C_1,\ldots,C_k\}$ แต่ละตัวแบ่งกลุ่มจะใช้กำหนดรูปแบบการเขียนที่ได้ใน C_i โดยสัญชาตญาณ" รูปแบบการเขียน" ในชุดของอีเมลล์ ซึ่งเป็นการรวมกลุ่มย่อยสำหรับลักษณะรายการ ความบ่อยที่เกิดขึ้นในอีเมลล์ที่แท้จริง $\{\mu_1,,,,,\mu_n\}$ ตัวอย่างเช่น คนที่ใช้คำเป็นแบบที่ตายตัวด้วยสัดส่วนที่เหมือนกันมากที่สุดในอีเมลล์ของเขา โดยลักษณะรายการผู้วิจัยจัดค่า mean ในลักษณะของย่อหน้าถัดไป ผู้วิจัยยึดรูปแบบความถี่ที่เกิดขึ้นโดยใช้แนวคิด frequent itemset ในแนวทางที่คล้ายคลึงกันของหนึ่งรายละเอียด ในกระบวนการนี้จะประกอบด้วยสองขั้นตอนใหญ่ (1) การสกัดแยกรูปแบบ(P) และ(2) การคำนวณความถี่ของรูปแบบ(FP) แสดงด้านล่าง ผู้วิจัยจำกัดขั้นแรกที่สกัดแยกรูปแบบการเขียน และ ค่า meanรูปแบบที่เกิดบ่อย

ให้ $F=\{F_1,\ldots,F_n\}$ เป็นกลุ่มของลักษณะที่แสดงในตารางที่ 1 และ ตารางที่ 2 ทฤษฎีสำหรับ frequent itemset ผู้วิจัยแยกแต่ละลักษณะของ F_i โดยสิ้นเชิงในบางช่วงเวลา $\{F_{i,1},\ldots,F_{i,j}\}$ ซึ่งแต่ละ $F_{i,b} \square \{F_{i,1},\ldots,F_{i,j}\}$ หมายถึง ลักษณะรายการ b สำหรับ F_i (แสดงในตารางที่ 3) ซึ่งต่างจากค่าลักษณะโดยสิ้นเชิงที่มีช่วงเวลาที่เท่ากัน ผู้วิจัยถามผู้ใช้ถึงรายละเอียดจำนวนมากของจำนวนที่เกิดขึ้นต่อช่วงเวลา สำหรับแต่ละลักษณะที่เกี่ยวกับตัวเลข ผู้วิจัยจะแบ่งค่าเป็นสองกลุ่มG1 และ G2 แต่ละกลุ่มจะแบ่งเป็นสองกลุ่มย่อยตามข้อจำกัดของจำนวนที่เกิดขึ้น(ภายในกลุ่ม)เกินขีดจำกัด กระบวนการนี้ทำซ้ำจนกระทั่งเป็นกลุ่มลักษณะที่เกิดขึ้นทั้งหมด วัตถุประสงค์ของทฤษฎีขนาดช่วงเวลาเท่ากับจำนวนช่วงเวลาทั้งหมดเป็นตัวกำหนดความไม่หยุดนิ่งของแต่ละลักษณะ

Table 4 – Patterns extracted from ensemble E.

Cluster(C)	E-mail(μ)	Pattern(P)
C_1	μ_1	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
	μ_2	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
	μ_3	$\{F_{1,2}, F_{2,2}, F_{3,3}\}$
	μ_4	$\{F_{1,1}, F_{2,3}, F_{3,3}\}$
C_2	μ_5	$\{F_{1,1}, F_{2,2}, F_{3,2}\}$
	μ_6	$\{F_{1,1}, F_{2,2}, F_{3,3}\}$
	μ_7	$\{F_{1,1}, F_{2,1}, F_{3,3}\}$
C_3	μ_8	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$
	μ_9	$\{F_{1,3}, F_{2,1}, F_{3,1}\}$
	μ_{10}	$\{F_{1,2}, F_{2,1}, F_{3,2}\}$
	μ_{11}	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

ให้ $P \subseteq F$ เป็นกลุ่มของรายการลักษณะที่เรียกว่า pattern อีเมลล์ μ ประกอบด้วย pattern P ถ้า $P \subseteq \mu$ รูปแบบมีลักษณะรายการเป็น q-pattern ตัวอย่างจะแสดงในตารางที่ 4 , pattern $P=\{F_{1,2},F_{2,3},F_{2,3}\}$, จะสกัดแยกจากอีเมลล์ μ_1 เป็น 3 pattern สนับสนุน pattern P เป็นร้อยละของอีเมลล์ใน E_i ที่อยู่ใน P ซึ่ง pattern P เป็น frequent pattern ในกลุ่มของอีเมลล์ E_i ถ้าสนับสนุน P มากกว่า หรือเท่ากันในบางผู้ใช้ที่มีความเฉพาะตัวสูง รูปแบบการเขียนจะพบในกลุ่ม C_i แทนกลุ่ม frequent patterns, แสดงโดย $FP(C_i) = \{F_{1,1},..,F_{m,n}\}$, สกัดแยกจากอีเมลล์ E_i ในกลุ่ม C_i ซึ่งจำนวนเต็ม m และ n แทนลักษณะตัวเลข และตัวเลขช่วงเวลา ตามลำดับผู้วิจัยใช้ตัวอย่างในการอธิบายสูงกว่าแนวคิดของวัตถุประสงค์ writing style mining .การคาดคะเนจุดสุดท้าย 3 clusters, C_1 ด้วยอีเมลล์ $\{\mu_1,\mu_2,\mu_3,\mu_4\}$, C_2 ด้วยอีเมลล์ $\{\mu_5,\mu_6,\mu_7\}$ และ C_3 ด้วยอีเมลล์ $\{\mu_8,\mu_9,\mu_{10},\mu_{11}\}$ แสดงในตารางที่ 4 เสนอลักษณะ item ภายในอีเมลล์ชี้บ่งด้วย a '1' ในลำดับ cell และในทำนองกลับกัน การสกัดแยกรูปแบบของแต่ละอีเมลล์ μ_s และ สัมพันธ์กับcluster C_i แสดงในตารางที่ 4 ซึ่งมีค่าอ้างอิงที่แตกต่างกันโดยสิ้นเชิงสำหรับการสกัดแยก $\{F_1, F_2, F_3\}$ แต่ละitems เสร็จเรียบร้อยหลังจาก clustering phase

ขณะนี้การคำนวณ frequent patterns แต่ละ cluster, ผู้วิจัยยอมรับการจำกัดผู้ใช้ min_sup = 0.4 ค่ากลางของ pattern $P = \{F_{1,1},..,F_{m,n}\}$ เป็นความถี่อย่างน้อย 40% ของอีเมลล์ภายใน cluster C_i ที่มีลักษณะ items ทั้งหมดใน P ตัวอย่างเช่น pattern $\{F_{1,2}, F_{2,3}, F_{3,3}\}$ เป็น frequent pattern เพราะว่าน้อยกว่า 3 และหรือ 4 อีเมลล์สำหรับ cluster C_1 ที่มี pattern นี้ pattern $\{F_{2,2}\}$ เป็นเพียงหนึ่งอีเมลล์เท่านั้นสำหรับ cluster เดียวกันและเป็นรูปแบบที่เกิดขึ้นไม่บ่อย ในทำนองเดียวกัน , pattern $\{F_{1,2}, F_{2,1}, F_{3,1}\}$ จะปรากฏที่อย่างสามจากสี่อีเมลล์ของ cluster C_3 และเป็นรูปแบบที่พบบ่อย

ในทางตรงกันข้ามแต่ละ patterns $\{F_{1,3}\}$ และ $\{F_{3,2}\}$ จะปรากฏเพียงหนึ่งอีเมลล์ที่สัมพันธ์กับ cluster และเป็นรูปแบบที่เกิดขึ้นไม่บ่อย $\{F_{1,2}, F_{2,1}, F_{3,1}\}$ และ $\{F_{1,3}\}$ เป็น 3-รูปแบบที่พบบ่อย และ 1-รูปแบบที่พบบ่อย ตามลำดับ. ในตัวอย่างของผู้วิจัย การใช้ min_sup = 0.4 เป็นค่ากลางของรูปแบบที่พบบ่อย ถ้ามีอย่างน้อย2ใน3 และหรือ 4 อีเมลล์ ความถี่ทั้งหมดและความสัมพันธ์อีเมลล์/clusters, สกัดแยกจากการนำเข้ามารวมชุดกัน ซึ่งแสดงในตารางที่ 5.

Table 5 – Frequent patterns (FP) extracted from ensemble E.

Cluster (C)	Frequent Patterns (FP)
C_1	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
C_2	$\{F_{1,1}, F_{2,2}, F_{3,3}\}$
C_3	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

4.5 Writing styles

Writeprint น่าจะระบุความเป็นหนึ่งเดียวของปัจเจกบุคคลได้ รูปแบบอาจจะแบ่งโดยมากกกว่าหนึ่ง clusters เช่น ตัวอย่างของผู้วิจัย $F_{1,2}$ แบ่งจาก cluster C_1 และC_3 ขณะที่ $\{F_{3,3}\}$ อยู่ระหว่าง C_1 และ C_2. ดังนั้นทั้ง patterns $\{F_{1,2}\}$ และ $\{F_{3,3}\}$ ถูกลบจาก clustersที่เกี่ยวข้องกัน รูปแบบที่เกิดขึ้นบ่อยประกอบขึ้นเป็นหนึ่งเดียว(หรือเกือบเป็นหนึ่งเดียว) writeprints $\{WP_1, WP_2, WP_3\}$ เป็นข้อมูลที่เก็บจาก clusters C_1, C_2, และ C_3, แสดงในตารางที่ 6. จากสามเหตุผลผู้วิจัยสรุปภาพรวมของสามอีเมลล์ต้องสงสัย writeprints $\{WP_1,..,WP_k\}$ ที่ขัดเจนถูกใช้ระบุผู้เขียนที่แท้จริงของอีเมลล์ประสงค์ร้าย

Table 6 – Writing styles (WS) mined from ensemble E.

Cluster (C)	Writing styles (WS)
C_1	$\{F_{2,3}\}$
C_2	$\{F_{1,1}, F_{2,2}\}$
C_3	$\{F_{2,1}, F_{3,1}\}$

5. Experiments and evaluation

ความมุ่งหมายของผู้วิจัยในส่วนนี้จะประเมินค่าวัตถุประสงค์และทำการวิเคราะห์ ไม่ว่าจะสามารถระบุได้ถูกต้องหรือไม่ ความแตกต่างในรูปแบบการเขียนของอีเมลล์ที่เก็บรวบรวมไว้ กลุ่มสำหรับการทดลองจำเป็นต้องออกแบบเพื่อหาคำตอบให้กับคำถามที่เกิดขึ้น ซึ่งการแบ่งกลุ่ม algorithm ดีกว่าอย่างอื่นสำหรับอีเมลล์ dataset อะไรที่เป็นความสัมพันธ์แน่นอนสำหรับแต่ละความแตกต่างทั้ง 4 ชนิด ของลักษณะการเขียน ? อะไรที่มีผลกับความหลากหลายของจำนวนผู้เขียนในผลที่ได้จากการทดลอง ? ในการทดลองผู้วิจัยพิจารณาผลความหลากหลายของจำนวนข้อความอีเมลล์ต่อผู้เขียนที่ได้คุณภาพ ผู้วิจัยทดลอง 3 กลุ่ม 1) ประเมินลักษณะ stylometric ในเทอมของ F-measure ผู้วิจัยประยุกต์การแบ่งกลุ่มรวมได้เกินกว่า 9 ลักษณะที่แตกต่าง

2) ความหลากหลายของจำนวนผู้เขียนซึ่งเก็บจากค่าคงที่ parameters อื่น (ข้อความต่อผู้เขียนและลักษณะเด่น) 3)ในกลุ่มการทดลองที่สามผู้วิจัยตรวจดูผลของจำนวนข้อความต่อผู้เขียน

ในทั้งหมด 3 กลุ่มของการทดลองสามความแตกต่าง การแบ่งกลุ่ม algorithm ได้แก่ EM , k-mean และ bisecting k-mean การรวมลักษณะที่แตกต่างคือ {T_1, T_2, T_3, T_4,$T_1 + T_2$, $T_1 + T_3$, $T_2 + T_3$, $T_1 + T_2 + T_3$, $T_1 + T_2 + T_3 + T_4$, }, ซึ่ง T_1, T_2, T_3 และ T_4 แทน เกี่ยวกับคำหรือศัพท์, การสร้างประโยค, โครงสร้างประโยค, และ ความเฉพาะเจาะจงของหัวข้อตามลำดับ

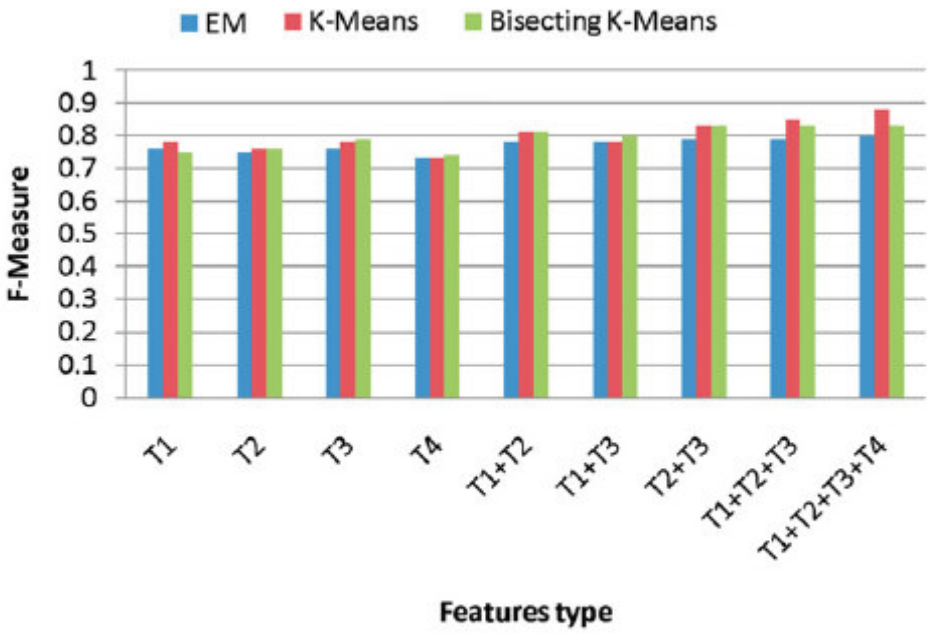

Fig. 2 – F-Measure vs. Feature Type and Clustering Algorithms (*Authors* = 5, *Messages* = 40).

ผู้วิจัยใช้ข้อมูลอีเมลล์ในชีวิตจริง : Eron E-mail dataset บรรจุ 200,399 อีเมลล์ของ 150 ผู้ใช้ของบริษัท Eron ผู้วิจัยซุ่มเลือกผู้ใช้ h จาก Eron E-mail dataset การแทนผู้เขียน h {A_1,.,A_h} สำหรับผู้เขียน A_i, ผู้วิจัยเลือก x สำหรับ A_i's e-mails. ซึ่ง h แปรผันจาก 3-10 ซึ่งค่าของ x เลือกจาก {10, 20, 40, 80, 100}.

ในการทดลองกลุ่มแรก ผู้วิจัยเลือก 40 e-m ails จากแต่ละหนึ่งของห้าผู้เขียน ผลจากสามการแบ่งกลุ่ม algorithms แสดงในรูป 2 ซึ่งจะอธิบายค่าของ F-measure ช่วง EM จาก 0.73 -0.80 , k-means จาก 0.73- 0.88 , และ bisecting k-means จาก 0.75 -0.83 .ผลที่ดีกว่าของ k-means และ bisecting k-means มากเกิน EM (ในกลุ่มการทดลอง) แสดงว่าเป็นการทราบจำนวนของ clusters K, ประการหนึ่งสามารถได้มาจากผลที่ดีกว่า. ผลของ k-means ดีกว่า bisecting k-means. ในตอนแรกไม่ได้คาดไว้ถึงผลที่ทำให้มีเหตุผลภายหลังจากการทำกลุ่มการทดลองทั้งหมดให้สมบรูณ์ K-means ดีกว่า เมื่อเทียบกับ

bisecting k-means มากถึง 40 อีเมลล์ต่อผู้เขียน การเพิ่มอีเมลล์นอกจาก 40 อีเมลล์ของแต่ละผู้เขียน ความถูกต้องสำหรับ bisecting k-means เริ่มเพิ่มมากขึ้น. คล้ายกับว่า bisecting k-means สามารถวัดได้มากว่า EM และ k-means.

การมองลักษณะเฉพาะบุคคล, T_4 (ลักษณะความเฉพาะของเนื้อความ) ทำได้น้อยขณะที่ T_3 (ลักษณะโครงสร้าง) ทำได้ดีมาก. สองแนวโน้มนี้ตรงกันในการที่จะศึกษา stylometric. ผลที่ดีที่สุดได้จาก k-means ที่ , $T_1 + T_2 + T_3 + T_4$ การรวมกันของลักษณะทั้งหมด 4 ชนิด โดยการเพิ่มลักษณะเนื้อความเฉพาะ, $T_1 + T_2 + T_3$, ผู้วิจัยไม่ได้เห็นเด่นชัดในการปรับปรุงผลของ EM และ bisecting k-means ทำให้เลือก keywords อาจเป็นไปได้ทั่วไปในระหว่างอีเมลล์ของการเลือกผู้เขียน ส่วนสำคัญอื่นๆ สังเกตุที่ { $T_2 + T_3$ } ผลที่ดีกว่าผลอื่นของการรวมสองลักษณะ (เช่น $T_1 + T_2$ และ $T_1 + T_3$)

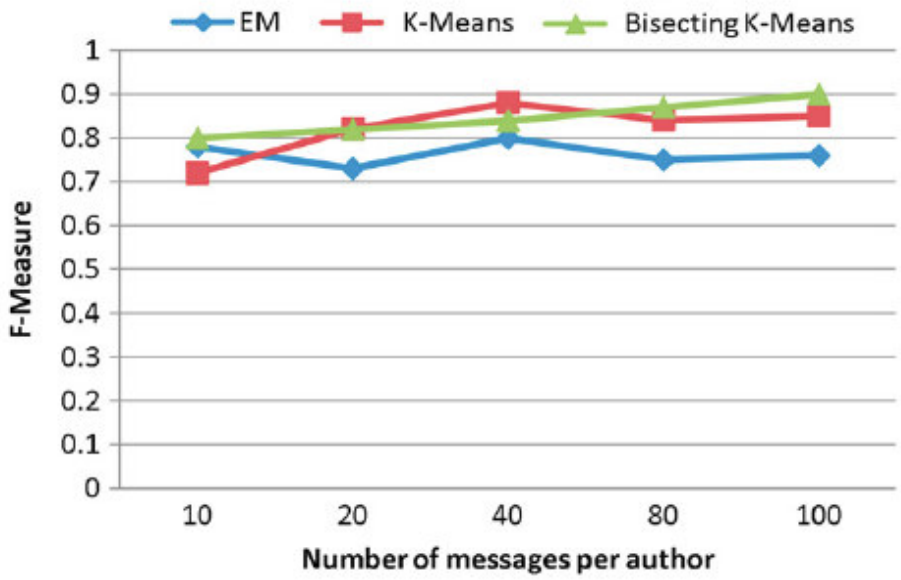

Fig. 3 – F-Measure vs. Features Type and Clustering Algorithms (*Authors* = 5, *Features* = $T_1 + T_2 + T_3 + T_4$).

ในกลุ่มการทดลองต่อไปจำนวนของผู้เขียน (5) และลักษณะกลุ่ม ($T_1 + T_2 + T_3 + T_4$) เก็บเป็นค่าคงที่ ค่าของ F-measure เพิ่มขึ้นด้วยการเพิ่มจำนวนของอีเมลล์ต่อผู้เขียน แสดงในรูปที่ 3 และ 4 โดย K-means และ bisecting k-means 90% ไม่มีการเจือปนสำหรับ 40 ข้อความต่อผู้เขียน ขณะที่ EM ให้ผลที่ขัดแย้ง การเพิ่มจำนวนข้อความต่อผู้เขียนที่นอกไปจาก 40 ผลในทางลบทั้งสาม algorithm ในระหว่าง, EM ที่ลดลงอย่างรวดเร็วกว่า 2 อย่างอื่น และ bisecting k-means เป็นเส้นที่มั่นคงมากกว่าเมื่อเทียบกับ k-means อย่างเดียว ผลการอธิบายความสัมพันธ์ของพฤติกรรม สำหรับ algorithms นี้ ในเทอมที่สามารถวัดได้

ในกลุ่มที่สามของการทดลอง In the third set of experiments (แสดงในรูปที่ 4), ผู้วิจัยพิจารณาลักษณะ $T_1 + T_2 + T_3 + T_4$ และ รับที่ 100 อีเมลล์ในแต่ละผู้เขียน . ค่า F-measure ของ bisecting k-means มีถึง 0.91 สำหรับการรวมทั้งหมดในกลุ่มการทดลอง ความแม่นยำทั้งหมดของสามการแบ่งกลุ่ม จะแสดงแบบที่ลดลง มากกว่าผู้เขียนในแผนการทดลอง

ความแม่นยำที่ดีประสบความสำเร็จโดยประยุกต์ k-means มากกว่าการรวม ลักษณะทั้ง 4 ชนิด ที่อีเมลล์ต่อผู้ใช้ที่จำกัด ถึง 40 Bisecting k-means เป็นทางเลือกที่ดีกว่า เมื่อมีผู้เขียนจำนวนมาก และกลุ่มการเข้าอบรมมีขนาดใหญ่ การนำจำนวนหัวข้อการอภิปราย เป็นผลที่ดีกว่าที่สามารถได้มาจากการเลือก domain-specific words อย่างระมัดระวัง อีกทางหนึ่งอาจระบุ keywords ที่เฉพาะตัวของผู้เขียนโดยประยุกต์เนื้อความพื้นฐานของการแบ่งกลุ่มในอีเมลล์แต่ละผู้เขียนอย่างเป็นอิสระ ผลของ EM มีน้อยและยากที่จะปรับปรุงโดยการปรับพารามิเตอร์

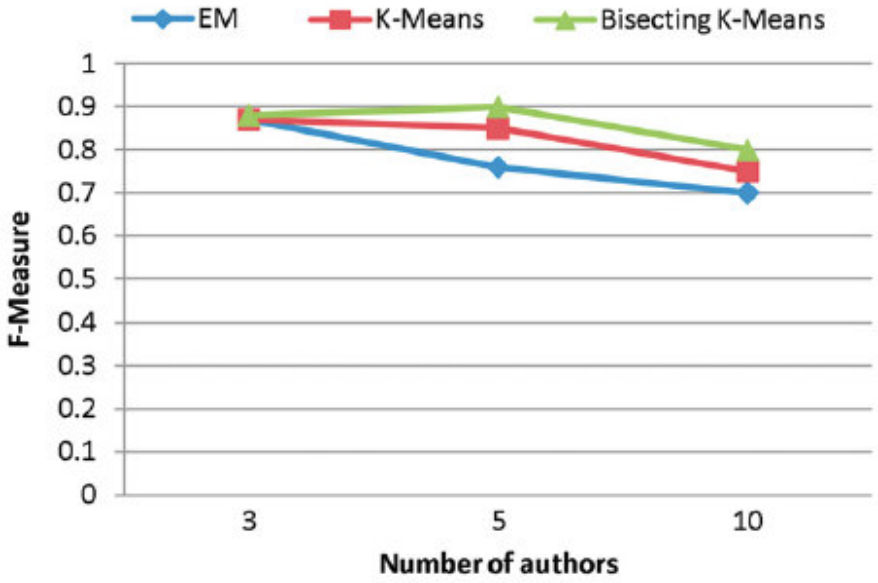

Fig. 4 – F-Measure vs. Number of authors and Clustering Algorithms (*Messages* = 100, *Features* = $T_1 + T_2 + T_3 + T_4$).

6. Conclusion

ผู้วิจัยพัฒนาการวิเคราะห์โครงสร้างอีเมลล์ที่จะแยกความแตกต่างของรูปแบบการเขียนจากการเก็บรวบรวมอีเมลล์ที่ไม่ปรากฏชื่อ วิธีการเสนอกลุ่มแรกให้ระบุชื่ออีเมลล์ตามคุณสมบัติ stylometric และหลังจากนั้น สกัดแยกความเป็นหนึ่งเดียว (เกือบจะเป็นหนึ่งเดียว) ด้วยรูปแบบการเขียนจากแต่ละ cluster สิ่งนี้จะช่วยผู้สืบสวนให้เรียนรู้เกี่ยวกับผู้เขียนที่ปกปิดตัวตนจากอีเมลล์ที่ไม่ปรากฏชื่อ รูปแบบการเขียนใน terms ของลักษณะรูปแบบจัดให้มีหลักฐานที่แน่ชัดมากกว่าผลบางตัวทางสถิติ ผลการทดลองของผู้วิจัยแสดงเทคนิคการแบ่งกลุ่ม สำหรับกลุ่มอีเมลล์ที่เป็นส่วนประกอบหลักของ stylometric

ความแม่นยำที่ลดลงของ 3 เทคนิคการแบ่งกลุ่ม เนื่องจากการเพิ่มจำนวนผู้เขียนที่ลงสมัคร และขนาดของตัวอย่างที่สามารถระบุประเด็นได้ ดังนั้น จึงจำเป็นต้องพิจารณาเทคนิคการแบ่งกลุ่มอย่างมาก ยิ่งไปกว่านั้นรายการลักษณะที่มีอยู่จำเป็นต้องอธิบายโดยการรวมลักษณะ idiosyncratic และการใช้จะรวมลักษณะความคล้ายด้วย

การศึกษาค้นคว้าที่มีอยู่ แสดงความเฉพาะของคำเนื้อความ ที่มีบทบาทสำคัญในรูปแบบ mining เมื่อใช้ในบทความเฉพาะที่เกิดขึ้นในการตรวจพิสูจน์ทางคอมพิวเตอร์ ดังนั้นมีความจำเป็นในการพัฒนาเทคนิคเสียงสำหรับการเลือกคำหลัก ซึ่งการเพิ่มประสิทธิภาพของลักษณะที่แน่นอนจะเป็นประโยชน์ในการกำหนดรูปแบบผู้เขียนที่แท้จริง นอกจากนี้การเปลี่ยนแปลงพฤติกรรมของมนุษย์จากสภาพแวดล้อมสู่สภาพแวดล้อม และจากคนสู่คน จำเป็นต้องมีการพัฒนาวิธีการจับรูปแบบลักษณะจะดีกว่าผลของการเขียนจริง ที่อยู่ภาษาที่เพิ่มขึ้นเป็นทิศทางการวิจัยอื่นๆ การวิจัย stylometric ยังคงอยู่ในขั้นแรกเริ่มของ forensics ซึ่งในอนาคตจะมีการพัฒนาที่ครอบคลุมในการวิเคราะห์ผู้เขียนที่เชื่อถือได้ ก่อนที่จะมีการยอมรับในกฎหมาย