การจัดกลุ่มบทคัดย่อทางวิทยาศาสตร์ของยีสต์

Clustering of Scientific Abstracts of Yeast

จีรยุทธ ไชยจารุวณิช ธิติพล เทียมจันทร์

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

อ.เมือง จ. เชียงใหม่ 50202 โทรศัพท์ 053-943409 โทรสาร 053-943433

Email: jeerayut@science.cmu.ac.th compsci18@hotmail.com

ABSTRACT - This paper uses the covariance-based clustering proposed by the author in [1] to analyze the scientific abstract corpus of yeast given in KDD Cup 2002 (Task 2) competition to show the effectiveness of the algorithm even dealing with data of large size.
KEY WORDS - Text clustering, Spectral data clustering

บทคัดย่อ - บทความนี้นำวิธีการจัดกลุ่มสมาชิกของเมตริกซ์จำนวนจริงแบบไม่สมมาตร โดยวิธี covariance-based clustering ที่เสนอใน [1] มาใช้ในการวิเคราะห์กลุ่มเอกสารบทคัดย่อทางวิทยาศาสตร์ของยีสต์ ที่กำหนดให้ในการแข่งขันการสืบค้นความรู้ KDD Cup 2002 (Task2) เพื่อแสดงให้เห็นว่าวิธีการจัดกลุ่มดังกล่าวสามารถประยุกต์ใช้งานได้แม้ข้อมูลจะมีขนาดใหญ่
คำสำคัญ - การจัดกลุ่มเอกสารข้อความ การจัดกลุ่มข้อมูลด้วยวิธีสเปกตรัล

1. บทนำ

การทำเหมืองข้อมูลมีจุดประสงค์หลักคือ การวิเคราะห์ค้นหาความรู้หรือรูปแบบความสัมพันธ์ที่แฝงอยู่ในข้อมูลที่สามารถนำไปใช้ให้เกิดประโยชน์ ในปัจจุบันมีเทคนิคที่ใช้วิเคราะห์ข้อมูลหลายเทคนิคด้วยกัน ได้แก่ เทคนิคการแบ่งกลุ่มข้อมูล (data clustering) เทคนิคการหากฎความสัมพันธ์ (association rule) เทคนิคการจัดกลุ่มและทำนายกลุ่มข้อมูล (data classification and prediction) [2] เป็นต้น อีกทั้งยังมีการนำเอาหลายเทคนิคไปประยุกต์ใช้ร่วมกันเพื่อให้เหมาะสมกับงานต่างๆ สำหรับเทคนิคการแบ่งกลุ่มหรือคลัสเตอร์ข้อมูลนั้น เป็นเทคนิคที่ได้รับความนิยมในการนำไปประยุกต์ใช้กับเทคนิคอื่น และเหมาะที่จะนำไปแบ่งกลุ่มข้อมูลในกรณีที่ไม่มีการระบุประเภทหรือจำนวนกลุ่มที่มีอยู่ของข้อมูลมาก่อน โดยหลักการของเทคนิคคลัสเตอร์ข้อมูลคือ การจัดข้อมูลให้เป็นสมาชิกของกลุ่มเดียวกันหากมีความคล้ายคลึงกัน และ ข้อมูลที่เป็นสมาชิกของกลุ่มหนึ่งจะไม่คล้ายคลึงกับข้อมูลที่เป็นสมาชิกของกลุ่มอื่น

เทคนิคคลัสเตอร์ข้อมูลมีขั้นตอนวิธีหรืออัลกอริทึมที่รู้จักกันอย่างแพร่หลาย เช่น อัลกอริทึม k-means อัลกอริทึม agglomerative and divisive hierarchical clustering [2] เป็นต้น บทความนี้จะนำอัลกอริทึม covariance-based clustering ที่ผู้เขียนเคยเสนอใน [1] มาใช้ในการวิเคราะห์จัดกลุ่มเอกสารข้อความโดยจะใช้ข้อมูลตัวอย่างที่กำหนดให้ใน KDD Cup 2002 (Task2) [5] เพื่อแสดงให้เห็นว่าวิธีการจัดกลุ่มดังกล่าวสามารถประยุกต์ใช้งานได้แม้ข้อมูลจะมีขนาดใหญ่

2. การจัดกลุ่มข้อมูลด้วยวิธี Covariance-based clustering

การจัดกลุ่มข้อมูลด้วยวิธี covariance-based clustering [1] อาศัยเทคนิคในการนำเอเกนเวคเตอร์ หรือสเปกตรัมของเมตริกซ์มาใช้ช่วยในการจัดรูปลักษณ์ของเมตริกซ์ใหม่ โดยสลับแถวและสดมภ์ของเมตริกซ์ข้อมูล ให้สมาชิกที่มีค่าใกล้เคียงกันปรากฏในบริเวณใกล้ ๆ กันมากที่สุด การเปลี่ยนรูปลักษณ์ของเมตริกซ์ดังกล่าว ทำให้

สามารถเห็นถึงการจัดกลุ่มของสมาชิก ซึ่งหมายถึงการจัดกลุ่มของข้อมูลแต่ละแถวและการจัดกลุ่มของแอททริบิวต์ในแต่ละสดมภ์นั่นเอง

ลำดับการจัดเรียงดังกล่าว ได้จากการวิเคราะห์สเปกตรัลของเมตริกซ์โควาเรนซ์ C (covariance matrix) ของเมตริกซ์ข้อมูล U ที่พิจารณาโดยลำดับการจัดเรียงสดมภ์ π_c ได้จากลำดับการเรียงค่าสมาชิกของเอเกนเวคเตอร์ ν_n ที่สอดคล้องกับค่าเอเกนที่มากที่สุดของเมตริกซ์โควาเรนซ์ และ ลำดับการจัดเรียงแถว π_r ได้จากลำดับการเรียงค่าสมาชิกของเวคเตอร์ผลคูณ $U\nu_n$ ระหว่างเมตริกซ์ข้อมูลที่พิจารณากับเอเกนเวคเตอร์ดังกล่าว ในส่วนย่อยต่อไปนี้จะอธิบายที่มาของขั้นตอนวิธีข้างต้น

2.1 การจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูลด้วยวิธีสเปกตรัล

กำหนดให้ U เป็นเมตริกซ์ของจำนวนจริงขนาด m x n การจัดกลุ่มสมาชิกของ U หมายถึงการจัดลำดับแถวและสดมภ์ของ U ใหม่ ให้สมาชิกของ U อยู่ใกล้เคียงกันมากที่สุด นั่นคือเราสามารถนิยามปัญหาการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูล U ได้ดังนี้

$$\max_{\pi_c} \frac{1}{m}\sum_{j=1}^{n-1} < U_j^{\pi_c}, U_{j+1}^{\pi_c} > + \max_{\pi_r} \frac{1}{n}\sum_{i=1}^{m-1} < U_i^{\pi_r}, U_{i+1}^{\pi_r} >$$

โดยที่ พจน์ซ้ายมือ หมายถึง การหาลำดับสดมภ์ที่ทำให้สมาชิกมีการจัดกลุ่ม

พจน์ขวามือ หมายถึง การหาลำดับแถวที่ทำให้สมาชิกมีการจัดกลุ่ม

π_c, π_r หมายถึง เวคเตอร์การจัดเรียงลำดับของสดมภ์และแถวตามลำดับ

$U_j^{\pi_c}$ หมายถึง สดมภ์ที่ j ของ U หลังการจัดเรียงสดมภ์ด้วย π_c

$U_i^{\pi_r}$ หมายถึง แถวที่ i ของ U หลังการจัดเรียงแถวด้วย π_r

<u, v> หมายถึง ความคล้ายคลึงกันของเวคเตอร์จำนวนจริง $u, v \in R^p$ ซึ่งนิยามโดย

$$< u, v >= \sum_{k=1}^{p} u_k v_k$$

จะเห็นว่าปัญหาการจัดกลุ่มสมาชิกดังกล่าวสามารถแยกพิจารณาอย่างเป็นอิสระต่อกัน และแต่ละปัญหาย่อยเป็นปัญหา quadratic combinatorial optimization ซึ่งจัดได้ว่ามีความซับซ้อนของปัญหาอยู่ในกลุ่ม NP-Complete [3] ดังนั้น การหาเวคเตอร์ลำดับที่ดีที่สุดในเวลา polynomial time จึงทำได้ยาก บทความนี้ขอเสนอวิธีสเปกตรัลที่นำเอเกนเวคเตอร์ของเมตริกซ์มาใช้เพื่อหาเวคเตอร์ลำดับ π_c, π_r ที่ให้การจัดกลุ่มของสมาชิกได้ดีภายในเวลาอันสมควร โดยจะแยกพิจารณาการจัดลำดับสดมภ์ในส่วนที่ 2.2 และพิจารณาการจัดลำดับแถวในส่วนที่ 2.3

2.2 การจัดลำดับสดมภ์เพื่อการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูล

เมื่อกำหนดเมตริกซ์ของจำนวนจริงแบบไม่สมมาตร U ขนาด m x n มาให้ เราสามารถนิยามปัญหาการจัดลำดับสดมภ์เพื่อการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูลได้ ดังนี้

$$\max_{\pi_c} \frac{1}{m}\sum_{j=1}^{n-1} < U_j^{\pi_c}, U_{j+1}^{\pi_c} >$$

เราสามารถหาความคล้ายคลึงกันของสมาชิกในสดมภ์ U_j, $U_{j'}$ ได้ด้วยค่าผลคูณเวคเตอร์

$$< U_j, U_{j'} >= \sum_{i=1}^{m} U_{i,j}\, U_{i,j'}$$

ดังนั้นเราสามารถหาความคล้ายคลึงกันเฉลี่ยระหว่างสดมภ์ใดๆด้วยเมตริกซ์ C โดยที่

$$C_{j,j'} = \frac{< U_j, U_{j'} >}{m} \qquad \forall j, j' \in \{1,.., n\}$$

นั่นคือ $$C = \frac{1}{m}U^T U$$

จะสังเกตได้ว่า C เป็นเมตริกซ์ของจำนวนจริงแบบสมมาตร ค่า $C_{jj'}$ จะบ่งบอกถึงความคล้ายคลึงกันระหว่างสดมภ์ j และ j' ในเมตริกซ์ U ดังนั้นถ้าต้องการให้สดมภ์ในเมตริกซ์ U ที่มีลักษณะคล้ายกันมาอยู่ใกล้กันจะต้องจัดลำดับเมตริกซ์ C ให้สมาชิกบริเวณเส้นทแยงมุมมีค่าให้มากที่สุด นั่นคือ การจัดกลุ่มสมาชิกของ U โดยการลำดับสดมภ์ใหม่สามารถทำได้โดยการหาลำดับของการจัดเรียงแถวและสดมภ์ของ C ใหม่ ให้สมาชิกที่มีค่ามากอยู่บริเวณเส้นทแยงมุมของเมตริกซ์ C ให้ได้มากที่สุด ปัญหาดังกล่าวมีนิยามที่เทียบเท่าได้ดังนี้

$$\max_{\pi_c}\left(\pi_c{}^T C \pi_c\right) = \max_{\pi_c} \frac{1}{m}\sum_{j=1}^{n-1} < U_j^{\pi_c}, U_{j+1}^{\pi_c} >$$

หากพิจารณาค่าเอเกนและเอเกนเวคเตอร์ของ C แล้วจะเกิดสมการดังนี้

$$Cv_i = \lambda_i v_i$$

โดยที่ λ_i คือ ค่าเอเกนที่ i ของเมตริกซ์ C

v_i คือ เอเกนเวคเตอร์ที่สอดคล้องกับค่า λ_i โดยที่ $\|v_i\| = 1$

หากคูณสมการข้างต้นด้วย v_i^T ทั้งสองข้างจะได้

$$v_i^T C v_i = v_i^T \lambda_i v_i = \lambda_i$$

หากกำหนดให้ค่าเอเกนของเมตริกซ์ C เรียงลำดับดังนี้

$$\lambda_1 \le \lambda_2 \le ... \le \lambda_n$$

เอเกนเวคเตอร์ v_n ที่สอดคล้องกับค่าเอเกน λ_n ที่มากที่สุด จะเป็นคำตอบของปัญหา $\max_{x \in R^n} x^T Cx$ โดย x เป็นเวคเตอร์ที่มีค่าสมาชิกเป็นจำนวนจริง

อย่างไรก็ตาม ในปัญหาของการจัดเรียงลำดับสดมภ์ของเมตริกซ์ C ที่พิจารณา $\left[\max_{\pi_c}\left(\pi_c{}^T C \pi_c\right)\right]$ เราต้องการ เวคเตอร์ของการจัดลำดับที่มีค่าสมาชิกเป็นจำนวนเต็มบวกที่อยู่ในช่วง 1 ถึง n แบบไม่ซ้ำ ใน [1] มีการเสนอให้ใช้วิธีการ approximation เอเกนเวคเตอร์ v_n ให้เป็น π_c โดยใช้ลำดับของการจัดเรียงค่าสมาชิกของ v_n เป็นคำตอบ นั่นคือ $\pi_c = order\ (v_n)$

วิธีสเปกตรัลในการจัดลำดับสดมภ์เพื่อการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูล จึงมีขั้นตอนวิธีดังต่อไปนี้

Algorithm CovarianceBasedColumnClustering(U, π_c)

Input: $U \in R^{m*n}$ เมตริกซ์ของจำนวนจริงแบบไม่สมมาตร

Output: π_c เวคเตอร์การจัดเรียงลำดับสดมภ์ของเมตริกซ์ U เพื่อให้สมาชิกของเมตริกซ์มีการจัดกลุ่ม

(1) คำนวณหาโควาเรนซ์เมตริกซ์ของเมตริกซ์ข้อมูล $C = \frac{1}{m} U^T U$

(2) คำนวณหาเอเกนเวคเตอร์ที่สอดคล้องกับ ค่าเอเกนเวคเตอร์ที่มากที่สุด ของเมตริกซ์โควาเรนซ์ $v_n = eigenvector\ (C, n)$

(3) คำนวณหาลำดับการจัดเรียงสดมภ์ของเมตริกซ์ข้อมูล $\pi_c = order\ (v_n)$

2.3 การจัดลำดับแถวเพื่อการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูล

เมื่อกำหนด U เป็นเมตริกซ์ของจำนวนจริงแบบไม่สมมาตรขนาด m x n ปัญหาการจัดลำดับแถวเพื่อการจัดกลุ่มสมาชิกของเมตริกซ์ข้อมูล U สามารถนิยามได้ ดังนี้

$$\max_{\pi_r} \frac{1}{n} \sum_{i=1}^{m-1} < U_r^{\pi_r}, U_{i+1}^{\pi_r} >$$

โดยที่ π_r เป็นเวคเตอร์ของการจัดลำดับแถว

เราสามารถพิจารณาปัญหาดังกล่าวในลักษณะเดียวกับการพิจารณาในส่วน 2.2 นั่นคือ หากความคล้ายคลึงกันเฉลี่ยระหว่างแถว $i, i' \in \{1,.., m\}$ ใน U กำหนดได้โดยเมตริกซ์ R โดยที่ $R_{i,i'} = \frac{< U_i, U_{i'} >}{n}$ นั่นคือ $R = \frac{1}{n} U U^T$ แล้ว เราสามารถนิยามปัญหาการจัดลำดับแถวที่เทียบกันได้ดังนี้

$$\max_{\pi_r}\left(\pi_r{}^T R \pi_r\right) = \max_{\pi_r} \frac{1}{n} \sum_{i=1}^{m-1} < U_i^{\pi_r}, U_{i+1}^{\pi_r} >$$

เราสามารถแก้ปัญหาดังกล่าวได้ในลักษณะคล้ายกันกับกรณีการจัดลำดับสดมภ์ นั่นคือการหาเอเกนเวคเตอร์ที่สอดคล้องกับค่าเอเกนที่มากที่สุดของ R และหาลำดับการจัดเรียงของสมาชิกของเอเกนเวคเตอร์ดังกล่าวเพื่อนำไปจัดลำดับแถวของเมตริกซ์ U

อย่างไรก็ตาม โดยส่วนใหญ่ เมตริกซ์ U มักมีจำนวนแถวสูงกว่าจำนวนสดมภ์มาก การคำนวณหาเอเกนเวคเตอร์ของเมตริกซ์ R จึงไม่สามารถทำได้ง่ายนัก ใน [1] มีการเสนอวิธีการนำเอเกนเวคเตอร์ของเมตริกซ์ C ซึ่งคำนวณได้จากส่วนที่แล้วมาใช้ เพื่ออนุมานหาการจัดลำดับแถวโดยไม่ต้องคำนวณหาเอเกนเวคเตอร์ของเมตริกซ์ R ใหม่

จะสังเกตุได้ว่า $R = \frac{1}{n} U U^T$ และ $C = \frac{1}{m} U^T U$ ดังนั้นเราจะได้ความสัมพันธ์ดังต่อไปนี้

$$U^T R\, U = \frac{1}{n} U^T U\, U^T U = \frac{m^2}{n} C\, C$$

ซึ่งทำให้ปัญหาต่อไปนี้เทียบเท่ากัน

$$\max_{x \in R^m} \left(x^T U^T R U x \right) = \frac{m^2}{n} \max_{x \in R^m} \left(x^T C^2 x \right)$$

สำหรับทุกๆ เอเกนเวคเตอร์ v_i ที่สอดคล้องกับค่าเอเกน λ_i ของเมตริกซ์ C เราจะได้

$$C v_i = \lambda_i v_i$$

หากคูณเมตริกซ์ C ทั้งสองข้างของสมการข้างต้น จะทำให้ได้

$$C^2 v_i = \lambda_i C v_i = \lambda_i^2 v_i$$

จากสมการข้างต้น ทำให้เราทราบว่า เมื่อ λ_n เป็นค่าเอเกนที่มากที่สุดสำหรับเมตริกซ์ C แล้ว λ_n^2 จะเป็นค่าเอเกนที่มากที่สุดสำหรับเมตริกซ์ C^2ด้วย ดังนั้น v_n จึงเป็นเอเกนเวคเตอร์ที่สอดคล้องกับค่าเอเกนที่มากที่สุดสำหรับเมตริกซ์ C^2

นั่นหมายความว่า v_n เป็นคำตอบของปัญหา $\max_{x \in R^n} x^T C^2 x$ และปัญหา $\max_{x \in R^n} x^T U^T R U x$ นั่นคือ $U v_n$ เป็นคำตอบของปัญหา $\max_{x \in R^m} x^T R x$

จากความจริงดังกล่าว เราสามารถอนุมานได้ว่า เวคเตอร์ของการจัดลำดับแถว π_r สามารถหาได้จากลำดับการจัดเรียงค่าสมาชิกของเวคเตอร์ $U v_n$ โดยที่ v_n เป็นเอเกนเวคเตอร์ที่สอดคล้องกับค่าเอเกนที่มากที่สุดของเมตริกซ์ C เมื่อรวมการพิจารณาทั้งส่วนที่ 2.2 และ 2.3 ขั้นตอนวิธีสเปกตรัลของการหาการจัดลำดับสดมภ์และแถวของเมตริกซ์ U เพื่อการจัดกลุ่มสมาชิก จึงมีดังนี้

Algorithm CovarianceBasedClustering (U, π_c, π_r)

Input: $U \in R^{m*n}$ เมตริกซ์ข้อมูลที่เป็นจำนวนจริงแบบไม่สมมาตร

Output: π_c, π_r เวคเตอร์การจัดเรียงลำดับสดมภ์และแถวของเมตริกซ์ U เพื่อให้สมาชิกเมตริกซ์มีการจัดกลุ่ม

(1) คำนวณหาโควาเรนซ์เมตริกซ์ของเมตริกซ์ข้อมูล $C = \frac{1}{m} U^T U$

(2) คำนวณหาเอเกนเวคเตอร์ที่สอดคล้องกับค่าเอเกนเวคเตอร์ที่มากที่สุด ของเมตริกซ์โควาเรนซ์ $v_n = eigenvector\ (C, n)$

(3) คำนวณหาลำดับการจัดเรียงสดมภ์ของเมตริกซ์ข้อมูล $\pi_c = order\ (v_n)$

(4) คำนวณหาลำดับการจัดเรียงแถวของเมตริกซ์ข้อมูล $\pi_r = order\ (U v_n)$

การคำนวณหาเอเกนเวคเตอร์ในขั้นตอนที่ (2) เป็นการคำนวณหาเอเกนเวคเตอร์เฉพาะของเมตริกซ์จำนวนจริงแบบสมมาตร ซึ่งสามารถทำได้โดยวิธีทั่วไป เช่น Lanczos's method, Re-orthogonalization, etc [4].

จะสังเกตได้ว่าความซับซ้อนของขั้นตอนวิธีที่เสนอ ขึ้นอยู่กับความซับซ้อนของการคำนวณในแต่ละขั้นตอน โดยสามารถประมาณอย่างคร่าว ๆ ได้ตามลำดับดังนี้ $O(mn^2)$, $O(n^3)$, O(n.log n), และ O(mn+n.log n) จะเห็นได้ว่า หาก m >> n ความซับซ้อนทั้งหมดของวิธีที่เสนอจะเท่ากับ $O(mn^2)$ และหาก m << n ความซับซ้อนจะเท่ากับ $O(n^3)$

3. ข้อมูลที่กำหนดให้ใน KDD Cup 2002

ในส่วนนี้เราจะนำเทคนิคการจัดกลุ่มข้อมูลด้วยวิธี covariance-based clustering มาประยุกต์ใช้ในการศึกษาลักษณะของข้อมูลชุดเอกสารต่างๆ ที่กำหนดให้ใน KDD Cup 2002 (Task2) [5] ซึ่งมีส่วนประกอบต่างๆ ออกเป็นดังนี้

3.1 Gene–Abstracts file

Gene–Abstracts เป็นแฟ้มข้อมูลที่ระบุให้ทราบว่ายีน (gene) แต่ละยีนถูกกล่าวถึงในบทคัดย่อ (abstracts) ใดบ้าง โดยในบทคัดย่อจะมีคำสำคัญ (keyword) อธิบายคุณลักษณะต่างๆที่เกี่ยวกับยีนนั้น ภายในแฟ้มข้อมูล Gene–Abstracts จะมีลักษณะดังรูปที่ 1

YML034W	10734188
YML034W	10894548
YHR051W	207698
:	:

รูปที่ 1 แสดงตัวอย่างข้อมูลภายในแฟ้ม Gene-Abstracts

ในแต่ละบรรทัดภายในแฟ้มข้อมูลแสดงความสัมพันธ์เชิงคู่ลำดับระหว่างรหัสของยีน(สดมภ์แรก) กับชื่อแฟ้มข้อมูลของบทคัดย่อ (สดมภ์ที่สอง) ยีนที่พิจารณาเป็นยีนของเซลยีสต์โดยระบุไว้ในที่นี้ทั้งสิ้น 3329 ยีน

3.2 Abstracts Files

บทคัดย่อจะมีทั้งหมด 15,234 แฟ้มข้อมูล อยู่รวมในโฟลเดอร์เดียวกัน แต่ละบทคัดย่อถูกกำหนดชื่อด้วยรหัสตัวเลขซึ่งสอดคล้องกับที่อ้างอิงในแฟ้มข้อมูล Gene–Abstracts บทคัดย่อเหล่านี้ สามารถดาวน์โหลดได้จากฐานข้อมูล MEDLINE database (www.ncbi.nlm.nih.gov/entrez/query.fcgi) ภายในแฟ้มข้อมูลของบทคัดย่อจะมีลักษณะเป็นดังรูปที่ 2

> *Some strains of Saccharomyces cerevisiae exhibit a specific transport system for ureidosuccinic acid, which is regulated by nitrogen metabolism. ...*

รูปที่ 2 แสดงตัวอย่างข้อมูลภายในแฟ้มบทคัดย่อเลขที่ 13831

3.3 Function-Hierarchy file

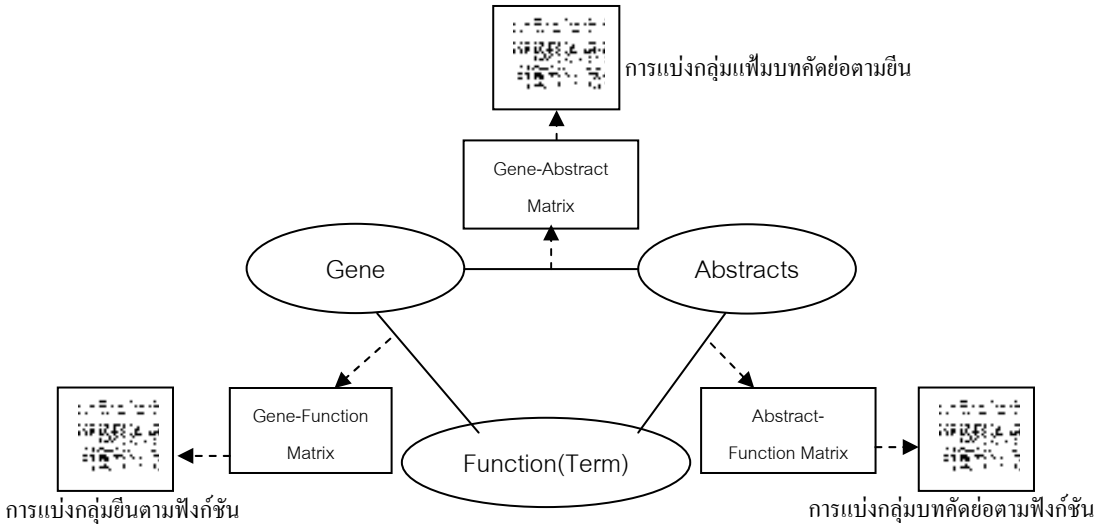
ในแฟ้มข้อมูลนี้เป็นการอธิบายข้อมูลในลักษณะลำดับขั้น (hierarchy) ของฟังก์ชันต่างๆ (function) ของยีน ในการจัดกลุ่มเอกสารบทคัดย่อ (abstract clustering) นั้น เราสามารถนำ Function-Hierarchy มาใช้สำหรับกำหนดคำศัพท์ทางเทคนิค (technical term) โดยอาจมีการปรับแฟ้มข้อมูล Function-Hierarchy ใหม่ เพื่อให้เหมาะสมกับการทำหน้าที่เป็นพจนานุกรมตัดคำศัพท์ (dictionary) ด้วยการคัดคำบางคำที่ไม่สนใจหรือไม่สำคัญออกจากแฟ้มเอกสารนี้ ภายในแฟ้มข้อมูล Function-Hierarchy ใหม่จะมีลักษณะดังตัวอย่างในรูปที่ 3 ซึ่งจะได้ฟังก์ชันของยีนทั้งสิ้น 342 ฟังก์ชัน

> *METABOLISM*
>
> *amino acid metabolism*
>
> *amino acid biosynthesis*
>
> *biosynthesis aspartate*
>
> *biosynthesis lysine*
>
> *biosynthesis cysteine-aromatic .....*

รูปที่ 3 แสดงตัวอย่างข้อมูลภายในแฟ้ม Function-Hierarchy

4. การวิเคราะห์จัดกลุ่มข้อมูล KDD Cup 2002

การจัดกลุ่มเอกสารงาน KDD Cup Task2 สามารถนำขั้นตอนต่างๆ มาเขียนเป็นแผนผังได้ดังรูปที่ 4



รูปที่ 4 แสดงแผนผังขั้นตอนการวิเคราะห์การจัดกลุ่มเอกสาร KDD Cup 2002

4.1 การจัดกลุ่มของยีนและแฟ้มบทคัดย่อ

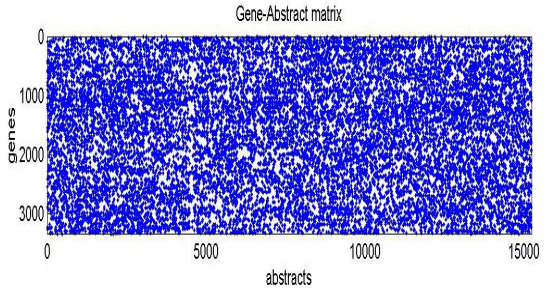
ขั้นตอนการจัดกลุ่มข้อมูลเริ่มจากการนำแฟ้มข้อมูล Gene–Abstracts มาจัดรูปแบบใหม่ให้เหมาะสมเพื่อนำไปสร้างเป็น เมตริกซ์สำหรับการนำไปจัดกลุ่ม โดยทำการกำหนดดรรชนีให้กับยีนและแฟ้มบทคัดย่อ ซึ่งจะได้คู่ลำดับของความสัมพันธ์ระหว่างยีนกับแฟ้มบทคัดย่อ ทำให้ได้รูปแบบดังตัวอย่างนี้

```
[3306]    'YPR173C'    '10393249'    [  604]
[3306]    'YPR173C'    '1097413'     [ 2561]
[3306]    'YPR173C'    '11329380'    [ 3568]
[3306]    'YPR173C'    '11559748'    [ 4160]
[3306]    'YPR173C'    '11563910'    [ 4176]
[3306]    'YPR173C'    '9155008'     [12497]
[3306]    'YPR173C'    '9431454'     [13546]
```

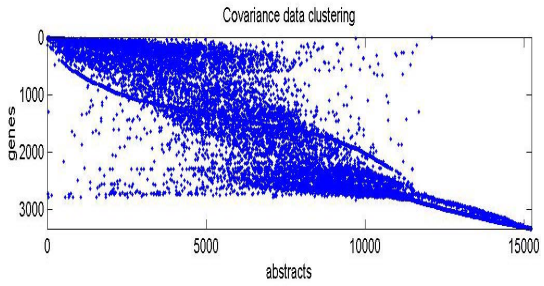
ในสดมภ์แรกจะเป็นดรรชนีของยีน (gene index) และสดมภ์สุดท้ายเป็นดรรชนีของแฟ้มบทคัดย่อ (abstract index) เมื่อมีการกำหนดดรรชนีเรียบร้อยแล้วจึงสร้างเป็นเมตริกซ์ ซึ่งในที่นี้เราใช้เมตริกซ์แบบ sparse matrix ที่สามารถสร้างเมตริกซ์ขนาดใหญ่และรองรับข้อมูลมากๆได้เนื่องจากจำนวนของชื่อแฟ้มข้อมูลของบทคัดย่อมีจำนวนมากถึง 15,234 แฟ้มข้อมูล

หากใช้การจัดการเมตริกซ์ในรูปแบบปกติ จะทำให้เกิดปัญหาในเรื่องของหน่วยความจำและเวลาในการทำงานของโปรแกรม เนื่องจากเมตริกซ์แบบเต็มรูปของความสัมพันธ์ระหว่างยีนกับข้อมูลบทคัดย่อจะมีขนาดเป็นจำนวนของยีนคูณกับจำนวนของแฟ้มบทคัดย่อ ข้อดีของการใช้ sparse matrix คือ สามารถจัดการข้อมูลในลักษณะคู่อันดับ ทำให้สามารถนำดรรชนีที่กำหนดมาระบุลงในเมตริกซ์รูปแบบนี้ได้ทันทีและสะดวกกับการจัดการ

จากความสัมพันธ์ของยีนกับแฟ้มบทคัดย่อข้างต้น เราสามารถวิเคราะห์หาการจัดกลุ่มของยีนที่ถูกกล่าวถึงในแฟ้มบทคัดย่อด้วยการจัดกลุ่มข้อมูลโดยวิธีการ covariance-based clustering ผลลัพธ์แสดงดังรูปที่ 5 และ 6 โดยในรูปที่ 5 และ 6 แกนนอนจะหมายถึงแฟ้มบทคัดย่อ แกนตั้งจะหมายถึงยีน จากรูปเราสามารถวิเคราะห์ได้อย่างคร่าวๆ ว่ายีนแบ่งออกเป็นสองกลุ่มใหญ่ๆ โดย กลุ่มหนึ่งมีจำนวนบทคัดย่อที่เกี่ยวข้องจำนวนน้อย และอีกกลุ่มหนึ่งมีจำนวนบทคัดย่อที่เกี่ยวข้องจำนวนมาก เราสามารถจำแนกรายชื่อของยีนและแฟ้มบทคัดย่อของทั้งสองกลุ่มดังกล่าวเพื่อประโยชน์ในการสืบค้นต่อไป



รูปที่ 5 แสดงแผนภาพก่อนการจัดกลุ่มยีนกับบทคัดย่อ

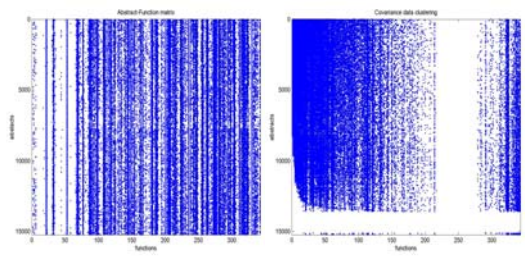


รูปที่ 6 แสดงแผนภาพหลังจากการจัดกลุ่มยีนกับบทคัดย่อ

4.2 การจัดกลุ่มของฟังก์ชันของยีนและแฟ้มบทคัดย่อ

ในส่วนของการหาการจัดกลุ่มของบทคัดย่อตามความคล้ายของจำนวนฟังก์ชันที่ปรากฏอยู่ภายในบทคัดย่อสามารถทำได้โดยการสร้างเมตริกซ์ความสัมพันธ์ในรูปแบบ Abstracts-Function คำศัพท์ที่สำคัญในแต่ละแฟ้มบทคัดย่อหมายถึงฟังก์ชันที่ระบุไว้ในแฟ้มข้อมูล *Function-Hierarchy* ที่ได้กล่าวข้างต้น เราจะนำคำศัพท์ชุดนี้มาทำการคัดเลือกเฉพาะคำที่สำคัญออกจากบทคัดย่อโดยวนเปิดแฟ้มบทคัดย่อต่างๆ แล้ว ตรวจหาฟังก์ชัน จากนั้นจะได้เป็นเมตริกซ์ ที่มีขนาดเป็นจำนวนของแฟ้มบทคัดย่อคูณกับจำนวนของฟังก์ชัน โดยสมาชิกของเมตริกซ์นี้ จะระบุจำนวนครั้งที่ปรากฏของฟังก์ชันในแต่ละแฟ้มบทคัดย่อ

เราสามารถวิเคราะห์หาความสัมพันธ์ของฟังก์ชันในแต่ละแฟ้มข้อมูล โดยใช้วิธีการจัดกลุ่มข้อมูลแบบ covariance-based clustering อีกครั้ง ซึ่งผลลัพธ์เป็นดังรูปที่ 7 และทำให้สามารถสรุปอย่างคร่าวได้ว่าบทคัดย่อส่วนใหญ่กล่าวถึงฟังก์ชันต่างๆ ของยีนและจะมีบางฟังก์ชันที่ไม่ได้ถูกกล่าวถึงเลย

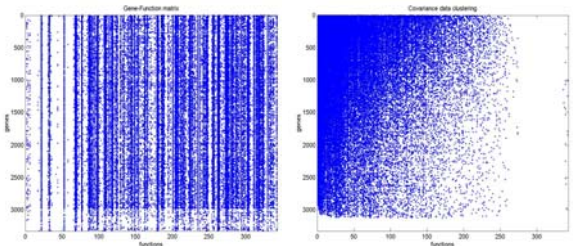


รูปที่ 7 แสดงแผนภาพก่อนการทำการจัดกลุ่มและหลังจากการจัดกลุ่มคำศัพท์เฉพาะในแต่ละแฟ้มบทคัดย่อ โดยแกนตั้งหมายถึงแฟ้มบทคัดย่อ แกนนอนจะหมายถึงฟังก์ชันของยีน

4.3 การจัดกลุ่มของยีนและฟังก์ชันของยีน

ในการศึกษาทางชีววิทยาระดับโมเลกุล เราต้องการทราบว่าแต่ละยีนของเซลสิ่งมีชีวิตมีฟังก์ชันอย่างไรบ้าง ในส่วนนี้เราจะนำผลจากการวิเคราะห์ข้อมูลในสองส่วนข้างต้นมาวิเคราะห์หาการจัดกลุ่มระหว่างยีนและฟังก์ชันของยีน จากเมตริกซ์ยีนกับบทคัดย่อ (Gene-Abstracts) และ เมตริกซ์คำศัพท์เฉพาะที่สำคัญในแต่ละแฟ้มบทคัดย่อ (Abstracts-Function) เมื่อเราต้องการหาความสัมพันธ์ระหว่าง ยีน กับคำศัพท์ในแต่ละแฟ้มบทคัดย่อ (Function) เราสามารถนำเมตริกซ์ข้างต้นทั้งสองมาคูณกัน เพื่อให้ได้เมตริกซ์ผลลัพธ์ Gene-Function = Gene-Abstracts * Abstracts-Function ค่าสมาชิกของเมตริกซ์ Gene-Function จะหมายถึง จำนวนบทคัดย่อที่กล่าวถึงยีนและฟังก์ชัน ซึ่งบ่งบอกความสัมพันธ์โดยนัยระหว่างยีนและฟังก์ชันนั้นๆ

เมื่อได้เมตริกซ์ Gene-Function เราสามารถใช้วิธี covariance-based clustering ในการจัดกลุ่ม ซึ่งทำให้ได้ผลสรุปออกมาดังรูปที่ 8 ผลจากการจัดกลุ่มในรูปที่ 8 ทำให้สามารถสรุปอย่างคร่าวได้ว่ายีนที่พิจารณามีกลุ่มฟังก์ชันเฉพาะร่วมกันอย่างเรียงราย และมีกลุ่มยีนที่ยังไม่มีการค้นพบหน้าที่หรือฟังก์ชันของมัน



รูปที่ 8 แสดงแผนภาพก่อนการทำการจัดกลุ่มและหลังจากการจัดกลุ่มยีนกับคำศัพท์ในแต่ละแฟ้มบทคัดย่อ โดยแกนนอนหมายถึงฟังก์ชันของยีน แกนตั้งหมายถึงยีน

ขั้นตอนวิธีต่างๆ ที่อธิบายข้างต้นได้รับการพัฒนาด้วยโปรแกรมแมทแลป และปฏิบัติการบนเครื่องคอมพิวเตอร์ส่วนบุคคลที่มีหน่วยประมวลผลรุ่น Pentium III ความเร็ว 650 MHz. และมีหน่วยความจำ 256 Mb. ตารางที่ 1 แสดงขนาดและเวลาการคำนวณของวิธีการจัดกลุ่มสมาชิกของเมตริกซ์ต่างๆ ด้วยวิธีสเปกตรัล

ตารางที่ 1 ขนาดและเวลาการคำนวณของวิธีการจัดกลุ่มสมาชิกของเมตริกซ์ต่างๆ ด้วยวิธีสเปกตรัล

เมตริกซ์	ยีนและแฟ้มบทคัดย่อ	แฟ้มบทคัดย่อและฟังก์ชันของยีน	ยีนและฟังก์ชันของยีน
ขนาด	3329*15234	15234*342	3329*342
เวลาการคำนวณ (วินาที)	1.28	23.22	5.05

5. สรุป

บทความนี้ได้แสดงการนำวิธีการจัดกลุ่มสมาชิกด้วยวิธี covariance-based clustering มาประยุกต์ใช้ในการวิเคราะห์ข้อมูลเอกสารขนาดใหญ่ วิธีการจัดกลุ่มสมาชิกดังกล่าวอาศัยการวิเคราะห์ปริมาณการจัดกลุ่มสมาชิกด้วยค่าเอเกนและเอเกนเวคเตอร์ของเมตริกซ์โควาเรนซ์ของเมตริกซ์ที่พิจารณา จากคุณสมบัติทางคณิตศาสตร์ต่างๆ เกี่ยวกับค่าเอเกนและเอเกนเวคเตอร์ ทำให้เราค้นพบลำดับการจัดเรียงแถวและสดมภ์ เพื่อการจัดกลุ่มสมาชิกได้อย่างมีประสิทธิภาพและสามารถประมวลผลได้กับข้อมูลขนาดใหญ่

เอกสารอ้างอิง

[1] จีรยุทธ ไชยจารุวณิช สุวิชยะ รัตตะรมย์ วิธีสเปกตรัลเพื่อการจัดกลุ่มสมาชิกของเมตริกซ์แบบไม่สมมาตร The Sixth National Computer Science and Engineering Conference (NCSEC 2002), October 29-31 2002, Pattaya, Thailand

[2] J.Han and M.Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. 2001

[3] A.L.Peressini, F.E.Sullivan, and J.J.Uhl Jr., The Mathematics of Nonlinear Programming, Springer-Verlag, New York, 1980

[4] J.H.Wilkinson, The algebraic eigenvalue problem, Clarendon press, Oxford, 1965

[5] KDD Cup 2002, The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, http://www.biostat.wisc.edu/~craven/ kddcup/

จีรยุทธ ไชยจารุวณิช สำเร็จการศึกษา BSc. (Computer Science) ปี 1995, MSc. (Computer Science) ปี 1996 จาก University of Grenoble I ประเทศฝรั่งเศส และ PhD. (Computer Science) ปี 2000 จาก University of Evry Val d'Essonne ประเทศฝรั่งเศส ปัจจุบันเป็นอาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่