การจัดกลุ่มข้อมูลด้วยขั้นตอนวิธี เคอินเวอร์สฮาร์โมนิกมีน
K - Inverse Harmonic Means Clustering Algorithm

กาญจนา จรัญศิริไพศาล (Kanjana Charansiriphaisan)*
ดร.สิรภัทร เชี่ยวชาญวัฒนา (Dr.Sirapat Chiewchanwattana) คำรณ สุนัติ (Kamron Sunut)*****

บทคัดย่อ

การจัดกลุ่มถูกใช้เป็นกระบวนการสำคัญในการวิเคราะห์ข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอน โดยพิจารณาจากความคล้ายกันของข้อมูล วิธีการจัดกลุ่มแบบเคมีน (K-Means: KM) เป็นเทคนิคหนึ่งที่เข้าใจง่าย และยังเป็นที่นิยมใช้งานกันอยู่ในปัจจุบัน วิธีการจัดกลุ่มแบบเคฮาร์โมนิกมีน (K-Harmonic means: KHM) พัฒนาจากวิธีการจัดกลุ่มเคมีนโดยคำนวณหาระยะห่างระหว่างข้อมูลด้วยมาตรวัดระยะทางแบบยูคลิเดียน (Euclidean distance) แล้วใช้ฟังก์ชันคิดค่าเฉลี่ยแบบค่าเฉลี่ยฮาร์โมนิกแทนการใช้ฟังก์ชัน Min ในการเลือกกลุ่ม ผลให้ประสิทธิภาพในการจัดกลุ่มดีกว่าการจัดกลุ่มแบบเคมีน สำหรับงานวิจัยนี้ได้นำเสนอวิธีการจัดกลุ่มแบบใหม่ที่เรียกว่าเคอินเวอร์สฮาร์โมนิกมีน (K-Inverse harmonic means : KIHM) โดยคำนวณหาค่าระยะห่างระหว่างข้อมูลด้วยฟังก์ชันเรเดียลเบสิสแบบผกผัน (Inverse radial basis function) สามารถช่วยลดจำนวนรอบในการหาคำตอบลงได้และเมื่อนำมาทดสอบกับชุดข้อมูลจริงคือข้อมูลการสืบค้นเว็บภาษาไทย ให้ผลการทดลองดีกว่าการจัดกลุ่มแบบเคฮาร์โมนิกมีน

ABSTRACT

Data clustering is an important process for data analysis. It is an unsupervised learning utilizing the similarity of data. An ordinary clustering algorithm is K-Means clustering (KM). Its computation is simple and it is showing satisfaction with clustering performance. The K-Harmonic means (KHM) is developed from the KM. The Min function of KM is replaced by the harmonic mean function. This yields the better clustering results. This paper presents an improvement of KHM called K-Inverse harmonic means clustering algorithm (KIHM). The harmonic means of the Euclidean distance is replaced by the harmonic means of reverse radial basis function of the distance from the cluster center. Experimental results show that the iterations of computation are reduced. Once the experiments are conducted on the Thai web snippets data, KIHM shows better clustering results than that of KHM.

คำสำคัญ : การจัดกลุ่ม เคฮาร์โมนิกมีน เคอินเวอร์ฮาร์โมนิกมีน

Key Words : Clustering, K-Harmonic means, K-Inverse harmonic means

---

* *นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

** *ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

*** *อาจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร*

บทนำ

ปัจจุบันด้วยความก้าวหน้าทางด้านเทคโนโลยีสารสนเทศ จะเห็นได้ว่าข้อมูลเป็นปัจจัยสำคัญในการวิเคราะห์หาคำตอบต่าง ๆ เพื่อนำมาพิจารณาตัดสินใจวินิจฉัยหรือทำนายเหตุการณ์บนพื้นฐานของเหตุและผล การจัดกลุ่มข้อมูล (Cluster Analysis) เป็นอีกวิธีการหนึ่งที่ถูกนำมาใช้เพื่อช่วยวิเคราะห์ข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอนโดยพิจารณาความคล้ายของข้อมูล ได้มีการศึกษาและพัฒนาขั้นตอนวิธีการจัดกลุ่มมากมายหลายวิธีด้วยกัน ซึ่งวิธีพื้นฐานที่สามารถเข้าใจง่ายคือการจัดกลุ่มแบบเคมีน โดยจะพิจารณาหาค่าระยะห่างระหว่างข้อมูลกับค่ากลางด้วยมาตรวัดยูคลิเดียน โดยหากข้อมูลอยู่ใกล้กับค่ากลางกลุ่มใดก็จะถูกจัดไว้ในกลุ่มนั้น แล้วทำการปรับค่ากลางแต่ละกลุ่มใหม่จากการคำนวณค่าเฉลี่ยของข้อมูลในกลุ่ม ซึ่งจะพบว่าวิธีเคมีน ยังมีข้อเสียตรงที่ต้องกำหนดจำนวนกลุ่มก่อนทุกครั้ง และการสุ่มค่ากลางขึ้นมาในแต่ละครั้งส่งผลให้คำตอบที่ได้ในการจัดกลุ่มแต่ละครั้งอาจไม่เหมือนเดิมและคำตอบที่ได้เป็นคำตอบเฉพาะที่ (local optimal) ต่อมาได้มีการพัฒนาขั้นตอนวิธีการจัดกลุ่มแบบเคฮาร์โมนิกมีน (KHM) โดยคำนวณหาค่าความเป็นสมาชิกของข้อมูลแล้วใช้ฟังก์ชันคิดค่าเฉลี่ยแบบค่าเฉลี่ยฮาร์โมนิกแทนการใช้ฟังก์ชัน Min ช่วยให้การจัดกลุ่มมีประสิทธิภาพมากขึ้น

สำหรับงานวิจัยนี้ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มแบบใหม่ที่เรียกว่าขั้นตอนวิธีเคอินเวอร์ส-ฮาร์โมนิกมีน (K-Inverse harmonic means clustering : KIHM) โดยนำฟังก์ชันเรเดียลเบสิสแบบผกผันมาช่วยในการคำนวณหาค่าระยะห่างของข้อมูลแทนมาตรวัดยูคลิเดียน ซึ่งเมื่อทำการทดลองกับข้อมูลคำอธิบายเว็บภาษาไทย ให้ผลดีกว่าวิธี KM และ KHM อีกทั้งยังใช้จำนวนรอบในการทำงานน้อยกว่าแต่ยังพบปัญหาในการจัดกลุ่มตรงที่การกำหนดค่าพารามิเตอร์เบตา (β) ส่วนที่เหลือของบทความได้แบ่งออกเป็นห้าส่วนดังนี้ คือ ทฤษฎีที่เกี่ยวข้อง วิธีการวิจัย ผลการวิจัยและการอภิปรายผล สรุปผลการวิจัย และข้อเสนอแนะ

ทฤษฎีที่เกี่ยวข้อง

1. Euclidean Distance

มาตรวัดระยะทางแบบยูคลิเดียน (Euclidean Distance) เป็นการคำนวณหาระยะห่างระหว่างจุดใด ๆ สองจุด เมื่อกำหนดให้

- j แทน ปริภูมิ j มิติ
- x แทนข้อมูลชุดที่ 1 $x = [x_{i1}\ x_{i2}\ \dots\ x_{im}]^T$
- y แทนข้อมูลชุดที่ 2 $y = [y_{i1}\ y_{i2}\ \dots\ y_{im}]^T$
- $d_{x,y}$ แทนระยะทางระหว่างข้อมูล x และ y

จะได้สมการคำนวณหาระยะทางแบบยูคลิเดียนดังนี้

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}(x_j - y_j)^2} \quad (1)$$

2. K-Means (KM)

การจัดกลุ่มข้อมูลแบบเคมีน (K-means) (Zhang, Hsu and Dayal, 1999) เป็นขั้นตอนวิธีการจัดกลุ่มแบบง่าย มีขั้นตอนวิธีดังนี้

ขั้นตอนที่ 1 สุ่มค่ากลางมาจำนวน j ชุด แทนด้วย C_j เมื่อ j แทนจำนวนกลุ่ม

ขั้นตอนที่ 2 คำนวณหาค่าระยะทางแบบยูคลิเดียนระหว่างชุดข้อมูลเข้า (x_i) กับค่ากลาง (C_k) แล้วนำข้อมูล x_i นั้นไปใส่ไว้เป็นสมาชิกในกลุ่ม θ_j ที่มีระยะทางที่น้อยที่สุด ถือเป็นกลุ่มที่ชนะ

$$j = \text{argmin}\|x_i - c_j\| \quad (2)$$

ขั้นตอนที่ 3 คำนวณค่ากลางสำหรับแต่ละกลุ่มใหม่ โดยเฉลี่ยจากข้อมูลที่เป็นสมาชิกของกลุ่มนั้น ๆ ทุก ๆ c_j

$$c_j = \frac{1}{q_j}\sum_{x_i \epsilon \theta_j} x_i \quad (3)$$

ขั้นตอนที่ 4 คำนวณหาค่าระยะทางรวมทั้งหมดแทนด้วย D โดยเป็นค่าระยะทางระหว่างค่ากลางของกลุ่มกับสมาชิกของกลุ่มนั้น ๆ เมื่อให้ I_{ij} คือ ค่าความเป็นสมาชิกของข้อมูล x_i ต่อกลุ่มที่ j ดังนี้

$$I_{ij} = 1 \quad ,x_i \text{ เป็นสมาชิกใน } \theta_j \quad (4)$$

$$D = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} 1/d_{ij}} \quad (5)$$

ขั้นตอนที่ 5 ทำซ้ำ (2-4) จนกระทั่งเงื่อนไขระยะทางรวมไม่มีการเปลี่ยนแปลงแล้วหรือเปลี่ยนแปลงน้อยมาก

$$1 - \frac{D^{(old)}}{D^{(new)}} \leq \alpha \quad \text{เมื่อ } \alpha \text{ มีค่าน้อย} \qquad (6)$$

3. K-harmonic means

การจัดกลุ่มข้อมูลแบบเคฮาร์โมนิกมีน (K-harmonic means) (Zhang *et al*, 2000) เป็นวิธีที่ปรับปรุงพัฒนามาจากวิธี KM โดยใช้ฟังก์ชันคิดค่าเฉลี่ยแบบค่าเฉลี่ยฮาร์โมนิกแทนการใช้ฟังก์ชัน Min วิธีการจัดกลุ่มแบบ KHM มีขั้นตอนวิธีดังต่อไปนี้

กำหนดค่าเริ่มต้นดังนี้

- จำนวนกลุ่มแทนด้วย j
- ค่ากลางของกลุ่มแทนด้วย C_j
- ข้อมูลแทนด้วย X_i
- จำนวนข้อมูลแทนด้วย N

ขั้นตอนที่ 1 สุ่มค่ากลางจำนวน j กลุ่ม แทนด้วย C_j

ขั้นตอนที่ 2 คำนวณหาค่าฟังก์ชั่นเป้าหมาย (Objective function value according) โดยแทนระยะห่างระหว่างชุดข้อมูลเข้าและค่ากลางด้วยสมการดังนี้

$$d_{ij} = \|X_i - C_j^{p}\| \qquad (7)$$

$$KHM(X,C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} 1/d_{ij}} \qquad (8)$$

ขั้นตอนที่ 3 คำนวณหาค่าความเป็นสมาชิกของข้อมูลทุกตัว X_i กับค่ากลางแต่ละกลุ่ม C_j แทนด้วย $m(C_j|X_i)$ คำนวณได้จากสมการดังต่อไปนี้

$$m(C_j|X_i) = \frac{1/d_{ij}^{2}}{\sum_{j=1}^{k} 1/d_{ij}^{2}} \qquad (9)$$

ขั้นตอนที่ 4 คำนวณหาน้ำหนักของข้อมูล

$$W(X_i) = \frac{\sum_{j=1}^{k} 1/d_{ij}^{2}}{(\sum_{j=1}^{k} (1/d_{ij}))^{2}} \qquad (10)$$

ขั้นตอนที่ 5 คำนวณหาค่ากลางใหม่จากสมการดังต่อไปนี้

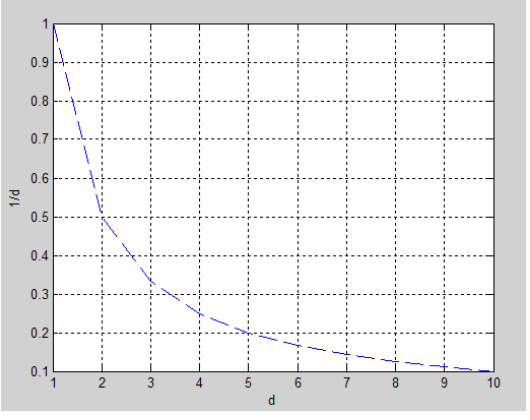
$$C_j = \frac{\sum_{i=1}^{N} m(C_j|X_i)\ W_i X_i}{\sum_{i=1}^{N} m(C_j|X_i)\ W_i} \qquad (11)$$

ขั้นตอนที่ 6 วนรอบทำซ้ำจนกระทั่งค่า KHM(X,C) ไม่เปลี่ยนแปลง ให้ข้อมูล X_i อยู่กลุ่ม j เมื่อค่าข้อมูลมีค่าความเป็นสมาชิกในกลุ่มนั้นมากที่สุด

4. K-Inverse harmonic means

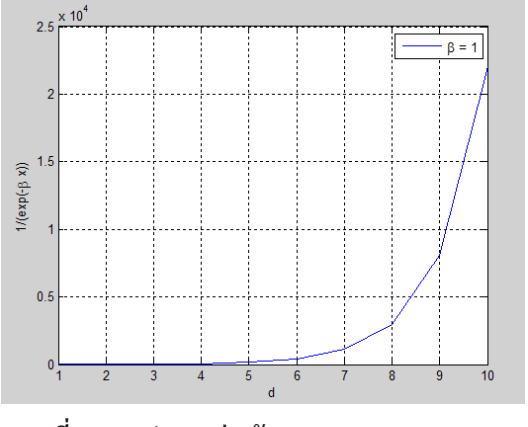
การจัดกลุ่มข้อมูลแบบเคอินเวอร์ฮาร์โมนิกมีน (K-Inverse harmonic means) เป็นขั้นตอนวิธีที่ใช้ฟังก์ชันเรเดียลเบสิสแบบผกผัน (Inverse radial basis function) มาคำนวณหาระยะห่างระหว่างข้อมูลแทนมาตรวัดยูคลิเดียน

เมื่อนำสมการ $1/d_{ij}$ โดย d_{ij} แทนค่าระยะห่างระหว่างข้อมูลโดยใช้มาตรวัดยูคลิเดียนตามสมการ (1) สามารถวาดกราฟแสดงความสัมพันธ์ระหว่างระยะทางแทนด้วยแกน x และค่าที่คำนวณจากสมการ $1/d_{ij}$ แทนด้วยแกน y ดังภาพที่ 1



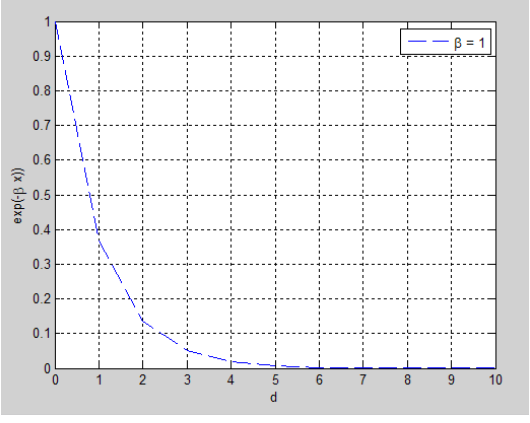
ภาพที่ 1 กราฟระยะห่างวัดจาก $1/d_{ij}$

เมื่อแทนค่าระยะทางด้วยฟังก์ชันเรเดียลเบสิสดังสมการ $1/\exp((-\beta) * d_{ij})$ แทนค่า β เท่ากับ 1 นำมาวาดกราฟจะได้กราฟดังภาพที่ 3 จะเห็นว่ากราฟเมื่อระยะทางน้อยจะได้ค่าน้อยแต่เมื่อระยะทางมากจะได้ค่ามาก



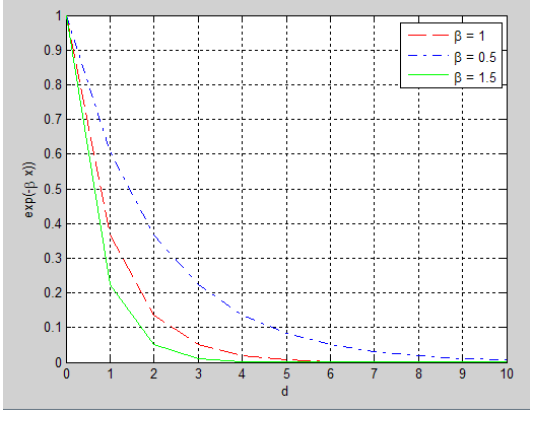
ภาพที่ 2 กราฟระยะห่างวัดจาก $1/\exp((-\beta) * d_{ij})$

เมื่อแทนค่าระยะทางด้วยฟังก์ชันเรเดียลเบสิสแบบผกผันดังสมการ $1/(\frac{1}{\exp(-\beta*d_{ij})})$ หรือเท่ากับ $\exp((-\beta)*d_{ij})$ เมื่อแทนค่า β เท่ากับ 1 นำมาวาดกราฟจะได้กราฟดังภาพที่ 3 จะเห็นว่ากราฟในภาพที่ 1 และ 3 มีลักษณะคล้ายกันคือ เมื่อระยะทางมีค่าน้อยจะให้ค่ามากและเมื่อระยะทางมากจะได้ค่าน้อยลง



ภาพที่ 3 กราฟระยะห่างวัดจาก $\exp((-\beta)*d_{ij})$

เมื่อกำหนดค่า β เป็น $\{0.5, 1, 1.5\}$ ในสมการ $\exp((-\beta)*d_{ij})$ จะเห็นว่า *β เพิ่มขึ้นจะส่งผลให้ค่า* $\exp((-\beta)*d_{ij})$ มีค่าน้อยลงและระยะทางจะแคบเข้ามา*ในขณะที่เมื่อค่า β มีค่าน้อยจะส่งผลให้ค่า* $\exp((-\beta)*d_{ij})$ มีค่ามากขึ้นและระยะทางจะไกลออกไปด้วย



ภาพที่ 4 กราฟระยะห่างวัดจาก $\exp((-\beta)*d_{ij})$ เมื่อกำหนดค่า β ={0.5, 1, 1.5}

ทำการปรับสมการใหม่โดยแทนค่าระยะทางด้วยฟังก์ชันเรเดียลเบสิสแบบผกผัน จะได้ขั้นตอนวิธีดังต่อไปนี้

กำหนดค่าเริ่มต้นดังนี้

- จำนวนกลุ่มแทนด้วย j
- ค่ากลางของกลุ่มแทนด้วย C_j
- ข้อมูลแทนด้วย X_i
- จำนวนข้อมูลแทนด้วย N
- ค่าพารามิเตอร์เบต้า แทนด้วย β

ขั้นตอนที่ 1 สุ่มค่ากลางจำนวน j กลุ่ม แทนด้วย C_j

ขั้นตอนที่ 2 คำนวณหาค่าฟังก์ชั่นเป้าหมาย (Objective function value according) โดยแทนระยะห่างระหว่างชุดข้อมูลเข้าและค่ากลางด้วยสมการดังนี้

$$KIHM(X,C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \exp((-\beta)*d_{ij})} \quad (12)$$

ขั้นตอนที่ 3 คำนวณหาค่าความเป็นสมาชิกของข้อมูลทุกตัว X_i กับค่ากลางแต่ละกลุ่ม C_j แทนด้วย $m(C_j|X_i)$ คำนวณได้จากสมการดังต่อไปนี้

$$m(C_j|X_i) = \frac{\exp((-\beta)*d_{ij})}{\sum_{j=1}^{k} \exp((-\beta)*d_{ij})} \quad (13)$$

ขั้นตอนที่ 4 คำนวณหาน้ำหนักของข้อมูล

$$W(X_i) = \frac{\sum_{i=1}^{N} m(C_j|X_i)W_iX_i}{\sum_{i=1}^{N} m(C_j|X_i)W_i} \quad (14)$$

ขั้นตอนที่ 5 คำนวณหาค่ากลางใหม่จากสมการดังต่อไปนี้

$$C_j = \frac{\sum_{i=1}^{N} m(C_j|X_i)W_iX_i}{\sum_{i=1}^{N} m(C_j|X_i)W_i} \quad (15)$$

ขั้นตอนที่ 6 วนรอบทำซ้ำจนกระทั่งค่า KIHM(X,C) ไม่เปลี่ยนแปลง ให้ข้อมูล X_i อยู่กลุ่ม j เมื่อค่าข้อมูลมีค่าความเป็นสมาชิกในกลุ่มนั้นมากที่สุด

ตัวอย่างการคำนวณด้วยขั้นตอนวิธีเคอินเวอร์ฮาร์โมนิกมีน

กำหนดค่าเริ่มต้น ดังนี้

$$X_i = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 6 & 5 \\ 7 & 5 \\ 8 & 5 \end{bmatrix}$$

$\beta = 1$

ขั้นตอนที่ 1 สุ่มค่ากลางขึ้นมา C_1 = [3 1]

C_2 = [1 1]

ขั้นตอนที่ 2 คำนวณหาค่าฟังก์ชั่นเป้าหมายด้วยสมการที่ 12

$$KIHM(X,C)=\sum_{i=1}^{N}\frac{k}{\sum_{j=1}^{k}\exp\left((-\beta)*d_{ij}\right)}$$

ตารางที่ 1 แสดงค่าระยะทางระหว่างชุดข้อมูล (X_i) กับค่ากลาง C_1 , C_2

ข้อมูล \ ค่าระยะทาง	C_1	C_2
X_1	2.00	0.01
X_2	1.00	1.00
X_3	0.01	2.00
X_4	5.00	6.40
X_5	5.65	7.21
X_6	6.40	8.06

คำนวณค่า KIHM(X,C) = 1.7772

ขั้นตอนที่ 3 คำนวณหาค่าความเป็นสมาชิกของข้อมูลทุกตัว X_i กับค่ากลาง C_1 ด้วยสมการที่ 13

$$m(C_j|X_i) = \frac{\exp\left((-\beta)*d_{ij}\right)}{\sum_{j=1}^{k}\exp\left((-\beta)*d_{ij}\right)}$$

ตารางที่ 2 แสดงค่าความเป็นสมาชิกของข้อมูลกับค่ากลาง C_1 , C_2

ข้อมูล \ ค่า $m(C_j\|X_i)$	$m(C_1\|X_i)$	$m(C_2\|X_i)$
X_1	0.1203	0.8797
X_2	0.5000	0.5000
X_3	0.8797	0.1203
X_4	0.8027	0.1973
X_5	0.8255	0.1745
X_6	0.8401	0.1599

ขั้นตอนที่ 4 คำนวณหาน้ำหนักของข้อมูลด้วยสมการที่ 14

$$W(X_i) = \frac{\sum_{j=1}^{k}1/d_{ij}^{\,2}}{(\sum_{j=1}^{k}(1/d_{ij}))^2}$$

ได้ ค่าน้ำหนักของข้อมูลดังนี้

ตารางที่ 3 แสดงค่าน้ำหนักของข้อมูล(X_i)

ข้อมูล \ ค่า $W(X_i)$	$W(X_i)$
X_1	0.9901
X_2	0.5000
X_3	0.9901
X_4	0.5076
X_5	0.5073
X_6	0.5066

ขั้นตอนที่ 5 คำนวณหาค่ากลางใหม่ด้วยสมการที่ 15 ได้ค่ากลางใหม่ ดังนี้

$$C_j = \frac{\sum_{i=1}^{N} m(C_j|X_i)W_i X_i}{\sum_{i=1}^{N} m(C_j|X_i)W_i}$$

C_1 = [4.8208 3.0094]

C_2 = [2.3823 1.7144]

วนซ้ำทำงานจนกระทั่งค่า KIHM ไม่เปลี่ยนแปลงหรือเปลี่ยนแปลงน้อยมาก

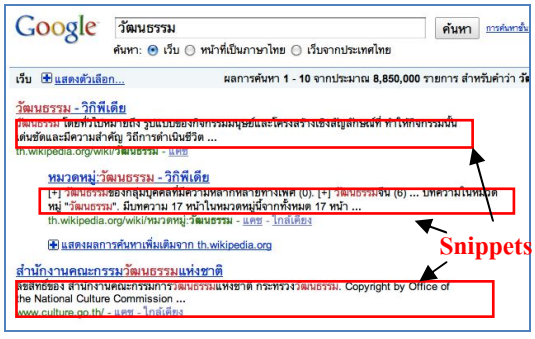
จากตัวอย่างข้างต้น กำหนดให้ค่า β= 1 และสุ่มค่ากลาง C_1 = [3 1] และ C_2 = [1 1] เมื่อวนรอบเพื่อปรับค่ากลางใหม่จะได้คำตอบ C_1 = [7.0003 4.9836] และ C_2 = [1.9997 1.0164] ในรอบที่ 7

วิธีการวิจัย

1. ชุดข้อมูลที่ใช้ในการทดลอง

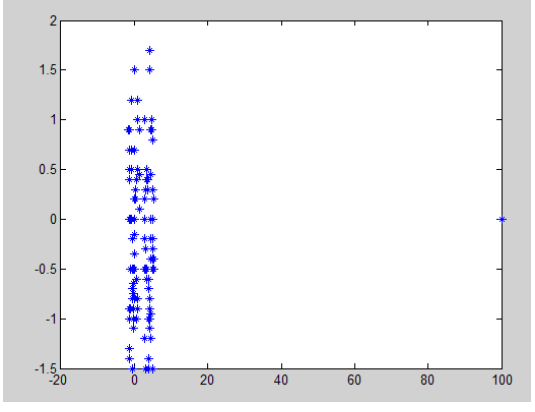
ชุดข้อมูลที่ใช้ในการทดลองส่วนแรกเป็นข้อมูลอธิบายเว็บภาษาไทย (พัชระ, 2553)โดยข้อมูลเป็นลักษณะคำอธิบายสั้น ๆ (Snippets) โดยนำมาจัดเรียงใหม่กำหนดให้แต่ละแถวแทนเอกสารและ

คอลัมน์แทนความถี่ของคำซึ่งมีทั้งหมด 118 แถว 665 คอลัมน์ สำหรับเอกสารที่นำมาทดสอบนี้มาจากสารบบหรรษา(H01) โดยแบ่งเป็น 3 หมวดคือ หมวดสังคม หมวดศิลปะ และหมวดวัฒนธรรม ตัวอย่างข้อมูลดังภาพที่ 5

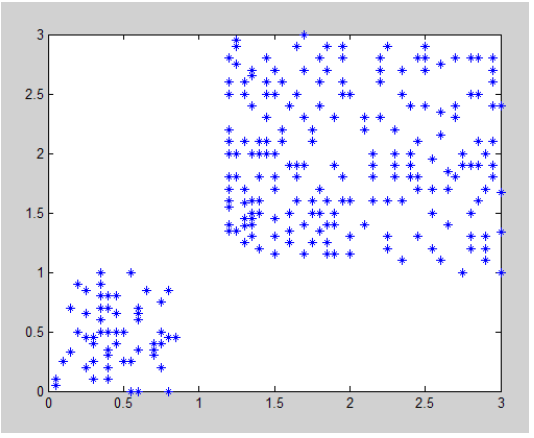


ภาพที่ 5 คำอธิบายสั้น ๆ ในเว็บ google.com

ชุดข้อมูลส่วนที่สองมีทั้งหมด 5 ชุดด้วยกัน เป็นข้อมูลจริงจากฐานข้อมูล UCI Machine Learning Repository จำนวน 3 ชุด คือ ชุดข้อมูล Iris ประกอบด้วย 4 แอตทริบิวต์ 150 เรคคอร์ด 3 กลุ่ม ชุดข้อมูล Wine ประกอบด้วย 13 แอตทริบิวต์ 178 เรคคอร์ด 3 กลุ่ม และชุดข้อมูล Soybean (Small) ประกอบด้วย 35 แอตทริบิวต์ 47 เรคคอร์ด 4 กลุ่ม เป็นข้อมูลที่สร้างขึ้นเองจำนวน 2 ชุดคือ DataSet1 ขนาด 2 มิติ 100 เรคคอร์ด 2 กลุ่ม สำหรับข้อมูลชุดนี้จะมีข้อมูลที่อยู่ไกลออกไปจากกลุ่มข้อมูล (outliers) และ DataSet2 ขนาด 2 มิติ 250 เรคคอร์ด 2 กลุ่ม



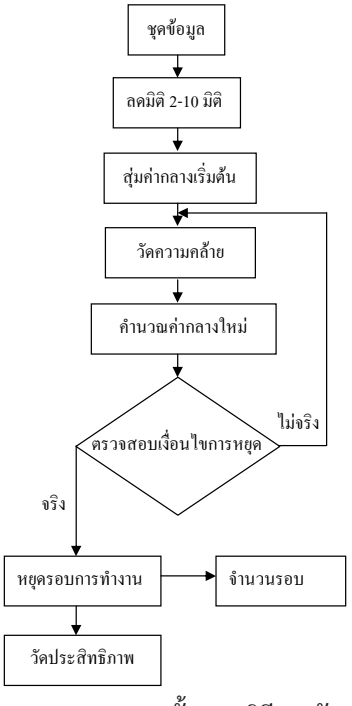
ภาพที่ 5 แสดงชุดข้อมูล DataSet1



ภาพที่ 6 แสดงชุดข้อมูล DataSet2

วิธีการทดลอง

นำข้อมูลในส่วนแรก(H01) มาลดมิติด้วยวิธีSingular Value Decomposition (SVD) (Jing and Jun, 2004) ให้เหลือ 2-10 มิติแต่ชุดข้อมูลส่วนที่สองไม่ต้องผ่านกระบวนการลดมิติ แล้วนำข้อมูลเข้าสู่กระบวนการจัดกลุ่มด้วยขั้นตอนวิธี 3 วิธี คือ KM, KHM และ KIHM นำคำตอบที่ได้มาวัดประสิทธิภาพและนับจำนวนรอบการทำงาน ดังแผนภาพแสดงขั้นตอนการทดลองดังภาพที่ 6



ภาพที่ 7 แผนภาพแสดงขั้นตอนวิธีการจัดกลุ่มข้อมูล

ผลการวิจัยและการอภิปรายผล

การทดลองได้ใช้วิธีการวัดค่าความถูกต้องโดยพิจารณาจากข้อมูลที่ถูกจัดไว้ในกลุ่มได้ถูกต้องจากการทดสอบวัดค่าความถูกต้องในแต่ละรอบจะให้ค่าความถูกต้องไม่เท่ากัน สำหรับงานวิจัยนี้ได้ทดสอบวัดค่าความถูกต้อง 100 รอบ แล้วนำค่าความถูกต้องที่ได้มาหาค่าเฉลี่ยดังแสดงในตารางที่ 4-6

ตารางที่ 4 แสดงผลการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคมีน (KM) เมื่อทดสอบกับข้อมูลคำอธิบายเว็บภาษาไทย (H01)

จำนวนมิติ	ค่าความถูกต้อง		
	ค่ามากสุด	ค่าน้อยสุด	ค่าเฉลี่ย
2	66.95	31.35	58.05
3	63.56	25.42	54.15
4	64.41	26.27	51.61
5	67.79	39.83	54.23
6	64.41	22.88	49.4
7	71.19	33.05	47.96
8	56.78	42.37	49.49
9	66.1	42.37	51.44
10	62.71	38.13	52.12

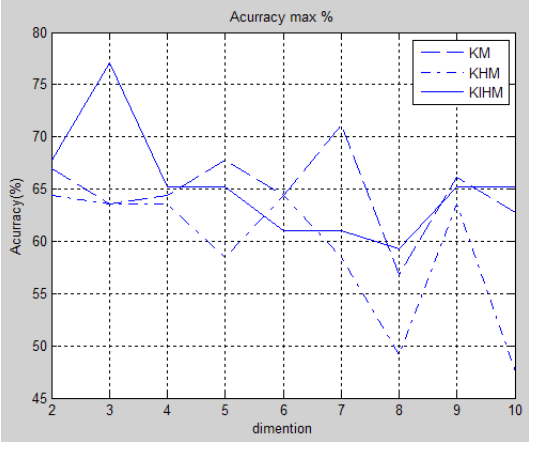
ตารางที่ 5 แสดงผลการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคฮาร์โมนิกมีน (KHM) เมื่อทดสอบกับข้อมูลคำอธิบายเว็บภาษาไทย (H01)

จำนวนมิติ	ค่าความถูกต้อง		
	ค่ามากสุด	ค่าน้อยสุด	ค่าเฉลี่ย
2	64.4	63.55	64.15
3	63.55	63.55	63.55
4	63.55	63.55	63.55
5	58.47	50	55.08
6	64.4	44.91	52.79
7	58.47	26.27	54.32
8	49.15	16.94	39.15
9	63.55	53.38	58.47
10	47.46	21.18	39.74

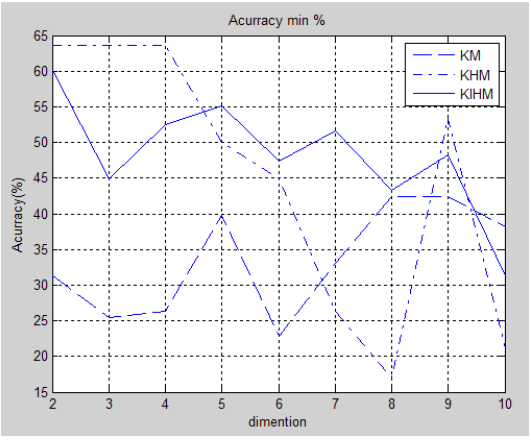
ตารางที่ 6 แสดงผลการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคอินเวอร์ฮาร์โมนิกมีน (KIHM) เมื่อทดสอบ กับข้อมูลคำอธิบายเว็บภาษาไทย (H01)

จำนวนมิติ	ค่าความถูกต้อง		
	ค่ามากสุด	ค่าน้อยสุด	ค่าเฉลี่ย
2	67.79	60.17	64.06
3	77.11	44.91	62.37
4	65.25	52.54	60.42
5	65.25	55.08	59.15
6	61.01	47.45	54.66
7	61.01	51.69	55.93
8	59.32	43.22	53.05
9	65.25	48.3	55.93
10	65.25	31.35	46.27

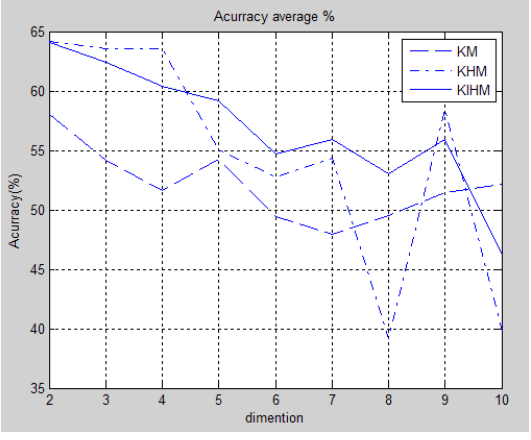
จากผลการทดลองตามตารางที่ 4-6 นำมาแสดงค่าความถูกต้องเพื่อเปรียบเทียบขั้นตอนวิธีทั้ง 3 วิธี ได้ดังภาพที่ 7-9



ภาพที่ 7 เปรียบเทียบค่าความถูกต้องที่ได้ค่ามากที่สุด



ภาพที่ 8 เปรียบเทียบค่าความถูกต้องที่ได้ค่าน้อยที่สุด



ภาพที่ 9 เปรียบเทียบค่าความถูกต้องเฉลี่ย

จากภาพที่ 9 ได้แสดงการเปรียบเทียบค่าความถูกต้องของขั้นตอนวิธีทั้ง 3 วิธี โดยเมื่อหาค่าเฉลี่ยความความถูกต้องแล้วจะเห็นว่าขั้นตอนวิธี KIHM จะให้ค่าความถูกต้องมากที่สุดยกเว้นเมื่อทดสอบกับข้อมูล 3 มิติ 4 มิติและ 9 มิติ ที่ให้ค่าความถูกต้องน้อยกว่าวิธี KHM ส่วนขั้นตอนวิธี KM ให้ค่าความถูกต้องน้อยที่สุดยกเว้นเมื่อทดสอบกับข้อมูล 8 มิติ และ 10 มิติ ที่ให้ค่าความถูกต้องสูงกว่า KHM

ตารางที่ 7 แสดงการเปรียบเทียบจำนวนรอบการทำงานเฉลี่ยของวิธีการจัดกลุ่ม 3 วิธี

จำนวนมิติ	ขั้นตอนวิธี		
	KM	KHM	KIHM
2	4	16	8.8
3	6	28	7.6
4	5	25	10.9
5	5	36	13.9
6	5	51	16.9
7	4	137	24
8	3	43	33.2
9	5	35	26.1
10	5	22	51

จากตารางที่ 7 จะเห็นว่าขั้นตอนวิธี KHM มีจำนวนรอบในการทำงานมากที่สุด ส่วนวิธี KIHM ใช้จำนวนรอบในการทำงานน้อยกว่า KHM แต่มีจำนวนรอบมากกว่า KM ซึ่งทำงานด้วยจำนวนรอบน้อยที่สุด สามารถเรียงลำดับขั้นตอนวิธีที่ทำงานด้วยจำนวนรอบจากน้อยไปหามากได้ดังนี้

KM < KIHM < KHM

ตารางที่ 8 แสดงการเปรียบเทียบค่าความถูกต้องเฉลี่ยในการจัดกลุ่มและจำนวนรอบการทำงานเฉลี่ยของวิธีการจัดกลุ่ม 3 วิธี ด้วยชุดข้อมูล Iris ,Wine, Soybean (small), Dataset1 และ Dataset2

DataSet	**Algorithm**					
	KM		KHM		KIHM	
	Accur.	Iter	Accur.	Iter	Accur.	Iter
Iris	86.06	2.6	89.33	17.7	89.33	16
Wine	67.97	3.4	66.51	35	70.78	11
Soybean	81.27	2.5	70.85	42.5	72.57	25
Dataset1	50.5	10	96	25	96	16
Dataset2	100	6	96.8	24	100	10

จากผลการทดลองในตารางที่ 8 จะเห็นว่าขั้นตอนวิธีเคอินเวอร์ฮาร์โมนิกมีนให้ค่าความถูกต้องมากที่สุดเมื่อจัดกลุ่มด้วยข้อมูล Iris, Wine, Dataset1 และ Dataset2 ส่วนวิธีเคมีให้ค่าความถูกต้องมากที่สุดเมื่อจัดกลุ่มด้วยข้อมูล Soybean และให้ค่าความถูกต้องน้อยที่สุดเมื่อจัดกลุ่มในข้อมูล Dataset1 เนื่องจากข้อมูลใน Dataset1 มีข้อมูลที่เป็น outliers อยู่ จึงทำให้การจัดกลุ่มได้ผลไม่ดี ซึ่งเป็นข้อเสียของการจัดกลุ่มด้วยวิธีการเคมีน สำหรับขั้นตอนวิธีเคฮาร์โมนิกมีนจะให้ค่าความถูกต้องที่ใกล้เคียงกับวิธีเคอินเวอร์สฮาร์โมนิกมีนและใช้จำนวนรอบในการทำงานมากกว่าวิธีเคอินเวอร์สฮาร์โมนิกมีน

สรุปผลการวิจัย

จากการศึกษาขั้นตอนวิธีการจัดกลุ่มนั้น วิธีการจัดกลุ่มที่ยังเป็นที่นิยมและเข้าใจง่ายคือ วิธีเคมีน (KM) แต่มีข้อเสียตรงที่การสุ่มค่ากลางขึ้นมาในแต่ละครั้งส่งผลให้คำตอบที่ได้ไม่น่าเชื่อถือ เพราะในการจัดกลุ่มข้อมูลนั้นจะพิจารณาความคล้ายโดยวัดระยะทางที่ใกล้ที่สุด ต่อมาได้มีการพัฒนาขั้นตอนวิธีเคฮาร์โมนิคมีน (KHM) ซึ่งวิธีนี้จะมีการคำนวณหาค่าสมาชิกและค่าน้ำหนักให้กับข้อมูลจึงทำให้คำตอบที่ได้น่าเชื่อถือมากขึ้น และให้ผลที่ดีกว่าวิธีเคมีน (KM) ดังนั้นงานวิจัยนี้จึงได้พัฒนาขั้นตอนวิธีแบบใหม่ที่เรียกว่าเคอินเวอร์สฮาร์โมนิคมีน (KIHM) โดยใช้ฟังก์ชันเรเดียลเบสิสแบบผกผัน มาคำนวณหาค่าระยะห่างแทนมาตรวัดยูคลิเดียน เมื่อนำมาทดลองกับข้อมูลส่วนแรกคือข้อมูลอธิบายเว็บภาษาไทย (Snippets) และทำการเปรียบเทียบกับขั้นตอนวิธี 3 วิธี คือ KM, KHM และ KIHM แล้ววัดค่าความถูกต้อง จากผลการทดลองแสดงให้เห็นว่า วิธี KIHM สามารถจัดกลุ่มข้อมูลได้ค่าความถูกต้องเฉลี่ยสูงที่สุด ส่วนวิธี KHM ให้ค่าความถูกต้องเฉลี่ยสูงรองจาก KIHM และวิธี KM ให้ค่าความถูกต้องน้อยที่สุด และเมื่อมาพิจารณารอบการทำงานของแต่ละวิธี จะเห็นว่าวิธี KM มีรอบการทำงานน้อยที่สุด วิธี KIHM มีรอบการทำงานมากกว่าวิธี KM ส่วนวิธี KHM มีรอบการทำงานมากที่สุดซึ่งเมื่อทดสอบกับข้อมูล 7 มิติใช้จำนวนรอบการทำงานสูงถึง 137 รอบ นอกจากนี้จะเห็นว่าชุดข้อมูลที่ทำการลดมิติให้เหลือ 2 มิติเมื่อนำมาทดสอบจัดกลุ่มด้วยขั้นตอนวิธีทั้ง 3 วิธี ให้ค่าความถูกต้องในการจัดกลุ่มมากกว่าชุดข้อมูล 3–10 มิติ

สำหรับผลการทดลองกับชุดข้อมูลส่วนที่สอง คือ Iris, Wine, Soybean, Dataset1 และ Dataset2 ขั้นตอนวิธีขั้นตอนวิธีเคอินเวอร์สฮาร์โมนิกมีนให้ค่าความถูกต้องสูงที่สุดยกเว้นเมื่อทดลองกับชุดข้อมูล Soybean และความถูกต้องมีค่าใกล้เคียงกันกับวิธีเคฮาร์โมนิกมีนส่วนจำนวนรอบการทำงานนั้นวิธีเคอินเวอร์สฮาร์โม นิกมีนนั้นทำงานด้วยจำนวนรอบที่น้อยกว่าวิธีเคฮาร์โมนิกมีนส่วนวิธีเคมีนจะทำงานด้วยจำนวนรอบน้อยที่สุด

จากผลการวิจัยสรุปได้ว่า วิธีการจัดกลุ่มแบบ KIHM สามารถจัดกลุ่มข้อมูลได้ดีกับชุดข้อมูลอธิบายเว็บภาษาไทย (H01) ข้อมูล Iris, Wine, Dataset1 และ Dataset2 ยกเว้นข้อมูล Soybean และทำงานด้วยจำนวนรอบน้อยกว่าขั้นตอนวิธีเคฮาร์โมนิกมีน

ข้อเสนอแนะ

จากการทดลองจะเห็นว่าขั้นตอนวิธี KIHM จะมีการกำหนดค่าพารามิเตอร์เบต้า β ซึ่งค่า β จะมีผลต่อคำตอบและจำนวนรอบในการทำงาน ในการวิจัยครั้งนี้ได้กำหนดค่า β เองโดยไม่ทราบว่าค่า β ที่เหมาะสมคือค่าใด ดังนั้นจึงส่งผลให้การทดลองกับชุดข้อมูลบางชุดเช่น Soybean ได้ผลไม่ดี จึงควรมีการศึกษาเพื่อหาวิธีการกำหนดค่า β ที่เหมาะสมในการจัดกลุ่มเพื่อให้ได้คำตอบที่ดีและมีประสิทธิภาพมากยิ่งขึ้น

เอกสารอ้างอิง

พัชระ นาเสงี่ยม. 2553. การประสานคำตอบสำหรับการจัดกลุ่มผลการสืบค้นเว็บภาษาไทยที่มีความน่าเชื่อถือ. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น.

Zhang, B., Hsu, M., and Dayal, U. 2000. K-Harmonic Means. Proceedings of International Workshop on Temporal. Spatial and Spatio-Temporal Data Mining. TSDM 2000. Lyon.

Jing, G., and Jun, Z. 2004. Clustered SVD strategies in latent semantic indexing. Elsevier Ltd. USA. 1051-1063.

Zhang, B., Hsu, M., and Dayal.1999. K-harmonic-means-a data clustering algorithm. Technical Report HPL-1999-124. HP Laboratories, PaloAlto.