Survey Text Clustering Algorithm

นางสาวสุภาพร โสภาจร 545020112-1 นางสาวณิชาภา หวังอ้อมกลาง 545020253-3
นางสาวไหมคำ ตันติปทุม 545020138-3 นางสาวณัฏฐนันท์ ว่องสาริกัน 545020285-0
คณะวิทยาศาสตร์ ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น

Abstract- ด้วยปัจจุบันเครือข่ายอินเตอร์เน็ตเป็นที่นิยมและส่วนมากจะเป็นข้อมูลสารสนเทศทรัพยากรองค์กร แหล่งสารสนเทศส่วนมากในการวิจัยจะเกี่ยวกับ text mining และการสืบค้นสารสนเทศ ด้วยการนำวิธีการจัดกลุ่มมาใช้ในการจัดกลุ่ม text และสารสนเทศต่างๆ การจัดกลุ่มเป็นวิธีที่สนใจข้อมูลจาก data mining ในงานวิจัยนี้จึงได้รวมอัลกอริทึมการจัดกลุ่ม การวิเคราะห์และเปรียบเทียบวิธีการของการจัดกลุ่มต่างๆ ซึ่งเนื้อหาหลักจะกล่าวถึงขอบเขตค่าพารามิเตอร์เริ่มต้น,สภาพแวดล้อมทั่วไป ซึ่งอัลกอริทึมที่นำมาใช้ในเนื้อหานี้ประกอบด้วย k-means, hierarchical, Self-organizing maps algorithm(SOM) และ combine k-means and Self-organizing maps algorithm

Key word- *text clustering algorithm, hierarchical clustering, k-means algorithm, self-organization maps, cluster text*

I. Introduction

การแบ่งกลุ่มข้อความเป็นกระบวนการที่ดำเนินการเกี่ยวกับองค์กรหรือหน่วยงานไปยังชุดข้อความภายใต้เงื่อนไขของการยกเลิกการเรียนรู้ พื้นฐานแนวคิดเป็นเพื่อที่จะแบ่งข้อความที่คล้ายกันไปยังคลาสที่เหมือนกัน การใช้เทคนิคการจัดกลุ่มข้อความ คุณสามารถหาระบบการจำแนกชุดข้อความที่มีขนาดใหญ่และให้มุมมองที่กว้างขึ้นสำหรับชุดข้อความ จะประยุกต์ใช้กับการแยกข้อมูลข่าวสารและเว็บเหมืองข้อมูล อัลกอริทึมการจัดกลุ่มสามารถแบ่งเป็นประเภทได้ดังต่อไปนี้

(1) Hierarchical clustering

(2) Partitioned clustering

(3) Density-based algorithm

อัลกอริทึม Self-organizing maps ในที่นี้ขอเรียกย่อๆ ว่า SOM ขณะเดียวกันปัญหาการจัดกลุ่มมีความพิเศษเฉพาะตัว ในแง่หนึ่งเวคเตอร์ข้อความเป็นเวคเตอร์ high-dimension มักเป็นหนึ่งพันหรือแม้กระทั่งเป็นหมื่น ในทางกลับกันเวคเตอร์ข้อความมักจะเป็นเวคเตอร์การแพร่กระจาย ดังนั้นมันเป็นเรื่องยากสำหรับทางเลือกของกลุ่มศูนย์กลาง เป็นเครื่องที่ไม่มีการสอนกระบวนการเรียนรู้เพราะว่าไม่จำเป็นต้องมีกระบวนการสอนและเอกสารคู่มือการใช้ในประเภทในขั้นสูง การจัดกลุ่มความยืดหยุ่นที่แน่นอนและสามารถจัดการได้อัตโนมัติ มันกลายเป็นสิ่งสำคัญที่หมายถึงสิ่งที่ให้ความสำคัญสำหรับผู้วิจัยมากขึ้น

II. Text Feature

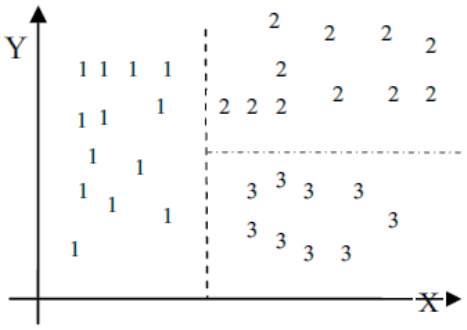
เพราะว่าข้อมูลข่าวสารของข้อความถูกจำกัดโครงสร้างหรือไม่มีโครงสร้าง ในขณะที่เนื้อหาเอกสารจะถูกใช้ภาษาธรรมชาติของมนุษย์ ดังนั้นคำถามพื้นฐานของการจัดกลุ่มข้อความ คือการแบ่งแยกเนื้อหาข้อความทำได้อย่างไร ซึ่งเป็นการวิเคราะห์ทางคณิตศาสตร์และกระบวนการดำเนินการจากรูปแบบต่าง ๆ ศาสตราจารย์ Salton ได้นำเสนอ Vector Space Model (VSM) และถูก

ประยุกต์ใช้อย่างกว้างขวางและเป็นอีกหนึ่งวิธีที่ให้ผลดีกว่าในช่วงไม่กี่ปีที่ผ่านมา แนวความคิดหลักของอัลกอริทึมเป็นช่องว่างของเอกสารที่พบเห็นเป็น vector space ที่ประกอบไปด้วย กลุ่มของเวคเตอร์ที่ตั้งฉากกันและแต่ละเอกสารจะถูกแทนที่เป็นลักษณะเวคเตอร์ปกติ

$$V(d) = (t_1, w_1(d); \ldots t_i, w_i(d); \ldots t_n, w_n(d);)$$

ซึ่ง t_i คือ ข้อมูลขาเข้า w_i (d) คือ t_i ที่อยู่ใน d ที่เป็นค่าน้ำหนัก ดังนั้น บทความเป็นการแสดงออกถึงเวคเตอร์ที่อยู่ใน high dimensional space

รูปที่ 1 ให้ประเภทของการแบ่งข้อมูลซึ่งถูกสร้างขึ้นโดยอัลกอริทึมต้นไม้การตัดสินใจ การจัดกลุ่มเริ่มต้นผลลัพธ์ประกอบด้วย 3 กลุ่ม



จุดข้อมูลในกลุ่มที่ 1 แทนที่ด้วย 1 จุดข้อมูลในกลุ่มที่ 2 แทนที่ด้วย 2 ข้อมูลจุดในกลุ่มที่ 3 แทนที่ด้วย 3

III. การเปรียบเทียบ Text Clustering Algorithm

วัตถุประสงค์ของการจัดกลุ่มข้อความที่เป็นชุดข้อมูลข้อความขนาดใหญ่ ที่สามารถแบ่งกลุ่มไปยังหลาย ๆ ประเภท และกระทำระหว่างข้อมูลข่าวสารข้อความในคลาสที่เหมือนกัน ซึ่งมีความคล้ายคลึงกันสูง แม้ว่าความแตกต่างของข้อความระหว่างประเภทที่แตกต่างกันมันเป็นเรื่องง่ายสำหรับคนที่ใช้ข้อมูลข่าวสารที่เป็นข้อความ ดังนั้น การเปรียบเทียบอัลกอริทึมการจัดกลุ่มข้อความควรจะขึ้นอยู่กับหลักเกณฑ์ 6 ข้อดังต่อไปนี้

(1) มีการขยายขีดความสามารถให้สูงขึ้นกว่าเดิม อัลกอริทึมการจัดกลุ่มไม่เพียงแต่ใช้ในชุดข้อมูลอย่างง่าย แต่ยังใช้ในชุดข้อมูลที่จริงที่มีขนาดใหญ่ซึ่งควรจะเกิดผลดี

(2) สามารถจัดการข้อมูลที่มีมิติสูงได้ ชุดข้อมูลข้อความจะถูกแสดงออกโดย VSM โดยทั่วไปจะมีหนึ่งพันหรือคู่มิติเว็บ ดังนั้นอัลกอริทึมสำหรับการจัดกลุ่มข้อความจะสามารถจัดการข้อมูลมิติสูง ๆ ได้

(3) มันสามารถค้นหารูปทรงของกลุ่มหลาย ๆ กลุ่มได้ เนื่องจากการพัฒนาการมองข้ามกฎระเบียบข้อบังคับและครอบคลุมกฎทุกกฎของการมองข้าม ขอบเขตระหว่างคลาสหลาย ๆ คลาส

(4) ซึ่งแต่ละคลาสจะถูกเพิ่มความเบลอ ทำให้รูปทรงของกลุ่มไม่ถูกจำกัดว่าจะเป็นวงกลมหรือรูปทรงอื่น ๆ มันต้องการทำอัลกอริทึมการจัดกลุ่มข้อความให้สามารถหารูปทรงของคลาสได้

(5) การพึ่งพาพารามิเตอร์ input และ ความรู้หลัก ๆ ต่ำ หลายอัลกอริทึมจำเป็นที่ต้องให้บางพารามิเตอร์หลาย ๆ ตัวก่อน แต่ในความรู้ก่อนหน้านี้ พารามิเตอร์เหล่านั้นยากที่จะกำหนดค่าและผลลัพธ์การจัดกลุ่มจะละเอียดอ่อนมากของพารามิเตอร์เหล่านี้ จึงพยายามที่จะหลีกเลี่ยงมัน

(6) ลำดับข้อมูลเข้าไม่ละเอียด ข้อความจะถูกแสดงโดยการใช้ VSM คำศัพท์จะเป็นคุณสมบัติของหน่วยรายการ และค่าความถี่ของการใช้คำศัพท์เป็นค่ารายการคุณสมบัติ ดังนั้นลำดับข้อมูลของข้อมูลข้อความไม่มีผลกระทบต่อผลลัพธ์ของการจัดกลุ่มครั้งสุดท้าย

(7) มีความสามารถที่ดีในการจัดการกับข้อมูลรบกวนส่วนใหญ่ของฐานข้อมูลที่ประกอบด้วยการแยกจุดข้อมูลที่ไม่รู้จักและข้อมูลอื่น ๆ ถ้าอัลกอริทึมมีความไวต่อข้อมูลมันจะลดคุณภาพของผลการจัดกลุ่ม

IV. Common Text Clustering

A. Hierarchical[1]

การจัดกลุ่มข้อความเป็นปัญหาทั่วไปของ การเรียนรู้แบบไม่มีผู้สอน (unsupervised machine learning) ขั้นตอนวิธี Hierarchical clustering โดยการรวมกลุ่มตามคล้ายคลึงกัน ตัวชี้วัดความคล้ายคลึงเช่น ค่า cosine, Dice coefficient, Jaccard similarity coefficient ซึ่งได้กลายมาเป็นเทคโนโลยีหลักเกี่ยวกับการจัดกลุ่มเอกสาร Hierarchical clustering เป็นวิธีการจัดกลุ่มเอกสารโดยทั่ว ๆ ไป ซึ่งสามารถสร้าง hierarchical nested class วิธีการ Hierarchical clustering จะเป็นหมวดหมู่ของ Hierarchical อีกนัยหนึ่งที่มีการเปลี่ยนแปลงของหมวดหมู่ มีการเปลี่ยนแปลงที่สอดคล้องกัน

ผลของรูปแบบ hierarchical clustering รูปแบบเดียวประเภทต้นไม้ แต่ละ class node ประกอบด้วย child nodes หลาย ๆ โหนด brother node เป็นส่วนหนึ่งของ parent nodes ของมัน (Fig.2). ด้านล่างของต้นไม้มี 5 กลุ่ม ในชั้นสุดท้าย กลุ่ม 2 ประกอบด้วย ข้อมูล 5 จุด และ ข้อมูล 6 จุด กลุ่ม 4 ประกอบด้วย ข้อมูล 8 จุด และ ข้อมูล 9 จุด กับการท่องต้นไม้แบบล่างขึ้นบน จำนวนของกลุ่มคือน้อยและน้อย

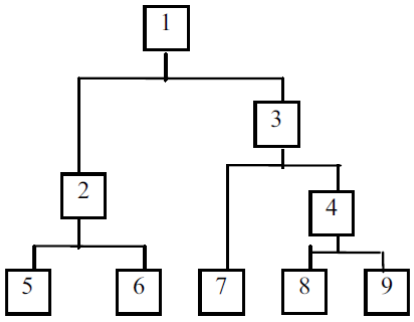


Fig.2. a sample of hierarchical clustering

ดังนั้นวิธีการนี้ช่วยในการจัดกลุ่มข้อมูลที่แตกต่างกัน สอดคล้องกับวิธีการสร้างประเภทต้นไม้ วิธีการ hierarchical clustering สามารถแบ่งออกได้ 2 ประเภทดังนี้ ประเภทที่ 1 คือ วิธีการแบบบูรณาการ (วิธีการล่างขึ้นบน) และ ประเภทอื่น ๆ คือ การแยกวิธีการ (วิธีการจากบนลงล่าง) ความถูกต้องของ Hierarchical clustering ค่อนข้างสูง แต่เมื่อนำแต่ละ class มารวมกันจะต้องมีการเปรียบเทียบทุก class ความคล้ายคลึงกันใน global และ เลือกตัวที่มีความคล้ายกันมากที่สุดของทั้งสอง class จึงค่อนข้างช้า ความบกพร่องของ Hierarchical clustering คือเมื่อขั้นตอน (รวมหรือแยก) เสร็จก็ไม่ต้องถูกยกเลิกดังนั้นจึงไม่สามารถแก้ไขการตัดสินใจที่ผิดได้ วิธีการ Hierarchical clustering โดยทั่วไปแบ่งเป็น วิธีการจัดกลุ่มจากด้านล่างขึ้นบนและวิธีการจัดกลุ่มลำดับชั้นจากด้านบนลงล่าง

B. K-Means [2]

เป้าหมายของอัลกอริทึม K-Means จะขึ้นอยู่กับการ input ค่า k ซึ่งชุดข้อมูลจะถูกแบ่งออกเป็นกลุ่มที่ k อัลกอริทึมจะเป็นวิธีการทำซ้ำโดยการปรับปรุงค่าในแต่ละรอบขึ้นอยู่กับจุด k โดยอ้างอิงจากจุดที่อยู่รอบ ๆ กลุ่มที่ k ค่ากลางแต่ละค่าจะเป็นจุดอ้างอิงของรอบถัดไป ในการทำซ้ำนั้นจะได้ค่ากลางจริงที่ทำให้ได้ผลการจัดกลุ่มที่ดีกว่า

อัลกอริทึม K-Means มีการทำงานดังนี้ : สมมติชุดข้อมูลที่จุด D คือ $\{x_1, x_2, \ldots x_n\}$, $x_i = (x_{i1}, x_{i2}, \ldots x_{ir})$ เป็นเวกเตอร์ที่อยู่ภายใน $X \subseteq \Re$ และ r แสดงจำนวน attribute ของข้อมูล (มิติของข้อมูล)

อัลกอริทึม K-Means(k,D)

(1) เลือกค่า k เริ่มต้นที่เป็นค่ากลางในการจัดกลุ่ม

(2) ทำซ้ำในแต่ละจุดข้อมูล เมื่อ $x \in D$

(3) คำนวณระยะทางจาก x ไปยังจุดกึ่งกลางแต่ละจุด และกำหนด x เพื่อหาค่ากลางที่ใกล้เคียงที่สุด

(4) คำนวณค่ากลางใหม่โดยใช้ค่ากลางปัจจุบันเป็นเกณฑ์และจะหยุดเมื่อเจอค่าที่ตรงกันกับจุดของข้อมูล

อัลกอริทึม K-Means มีความได้เปรียบคือ มีรูปทรงเรขาคณิตที่ดีและมีนัยสำคัญทางสถิติที่เป็นลักษณะของตัวเลข จะไม่ไวต่อการจัดลำดับ มีผลดีในการทำ convex cluster และสามารถทำงานในแบบขนานได้ รวมทั้งสามารถทำการ cluster ภายใต้บรรทัดฐานของการสุ่ม แต่ข้อเสียของ

มันคือต้องทราบจำนวนกลุ่มที่จะทำการ cluster ล่วงหน้า attribute ไม่สามารถประมวลผลข้อมูลได้แน่จัด มีความไวต่อการแยกจุด ไม่สามารถจัดกลุ่มข้อมูลที่ไม่ใช่ทรงกลมหรือกลุ่มข้อมูลที่ขนาดมีความแตกต่างกันเป็นอย่างมาก เป็นวิธีที่ใช้บ่อยที่สุด แต่ไม่ได้ดีที่สุดในโลก วิธีการมีความเสี่ยงต่อการเจอจุดที่มีความผิดปกติ ที่ขาดความยืดหยุ่น และในการจัดกลุ่มบางครั้งไม่มีความสมดุล

C. Self-Organizing Map (SOM)[3]

เป็น high-dimensional ของการจัดกลุ่มและการมองเห็นอัลกอริทึมการเรียนรู้แบบไม่มีผู้สอน อัลกอริทึม SOM เป็นการจำลองคุณสมบัติของสมองมนุษย์ไปเป็นการประมวลผลสัญญาณที่ถูกพัฒนาเป็นโครงข่ายประสาทเทียมแบบจำลองนี้ได้นำเสนอโดย Finnish Helsinki ศาสตราจารย์ Tuevo Kohonen ในปี 1981 ปัจจุบันนี้ได้ถูกนำมาใช้กันอย่างกว้างขวางของโครงข่ายประสาท self-organizing โครงข่ายประสาทจะถูกอธิบายเป็นโปรโตไทป์เวคเตอร์สำหรับแต่ละกลุ่มเป็นกลุ่มโปรโตไทป์ โปรโตไทป์เวคเตอร์ที่ไม่จำเป็นต้องสอดคล้องกับตัวอย่างข้อมูลและระบุวัตถุ ตามระยะการวัด วัตถุใหม่จะถูกกำหนดให้ผูกติดกับวัตถุนี้จะถูกแทนที่ด้วยเวคเตอร์โปรโตไทป์ที่คล้ายคลึงกันมากที่สุด

อัลกอริทึม SOM สามารถอธิบายได้ดังต่อไปนี้

(1) การสุ่มค่าเริ่มต้นการเชื่อมต่อค่าน้ำหนัก จำนวนครั้งการฝึกสอนสำหรับ K ตัวนับจำนวนการฝึกสอน k=0

(2) การสุ่มเลือกโหมด input การคำนวณระยะห่าง Euclidean ของทุกหน่วย input

(3) เลือกรับโหนด

(4) การเชื่อมต่อน้ำหนักของโหนดผู้ชนะและโหนดหลักของตัวมันเองเพื่อทำการปรับ

(5) ตัวนับที่ถูกเพิ่มขึ้น ถ้า k<K รันในขั้นตอนที่ 2 หรือไม่ก็จบการฝึกสอน

(6) ผลลัพธ์ของค่า output

D. Combine K-Means and SOM[4]

ในอัลกอริทึมนี้มีการใช้ Vector spatial model ที่สนับสนุนกระบวนการทำงานของเท็กท์และใช้ตัวแปรค่าน้ำหนักมาคิดร่วมด้วย โดยกำหนดเวกเตอร์ d=(w1,w2,w3,...wn) เมื่อ w คือน้ำหนักของเวกเตอร์แต่ละตัวในเท็กท์ ความยาวของเท็กท์ แทน d โดย m คือจำนวนไอเท็มทั้งหมด แล้ว wi(i=1,2,3...,m) เป็นค่าน้ำหนักของไอเท็มนั้น(ti) ใน text(d)

การกำหนด characteristic items เริ่มจากตัดคำที่ไม่ต้องการ ให้อยู่ในรูปที่จำง่ายตามกฎ TF-DF ในลำดับถัดไป หลีกเลี่ยงเหตุการณ์การใช้ค่าน้ำหนักมาก เพราะจะทำให้ความถี่ของเท็กท์ (ค่า tf มีค่ามาก) โดยค่าน้ำหนักจะมาจากการคำนวณค่าน้ำหนักในสมการ(1) โดย wij คือ ค่าน้ำหนักไอเท็ม j ในข้อความ I และ coefij จะกำหนดมาในสมการ(2) ภายในสมการ(2) จะกำหนดให้ tfij คือค่าความถี่ของไอเท็ม j ในข้อความ i

$$w_{ij} = (coef_{ij}) \cdot (\log N - \log df_i) \quad (1)$$

$$coef_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } tf_{ij}=1 \\ 1.5 & \text{if } 1<tf_{ij}\leq 5 \\ 2 & \text{if } 5<tf_{ij}\leq 10 \\ 2.5 & \text{if } tf_{ij}>10 \end{array} \right\} \quad (2)$$

ดังนั้นกลุ่มของเวกเตอร์จะกำหนดจากการเซตของเท็กท์ โดยการสร้าง model เพื่อใช้ในการจัดกลุ่ม ระยะห่าง (d) คือ ระยะห่างระหว่างเวกเตอร์ของเท็กท์ โดยนำฟังก์ชันของ cosine มาใช้หาค่าระยะห่าง ซึ่งนิยามตามสมการ(3)

$$d(doc_i, doc_j) = 1 - sim(doc_i, doc_j) \quad (3)$$

จากสมการ(3) sim(doci,docj) จะคำนวณค่าได้จากสมการ(4) โดย sim(doci,docj) จะเรียกอีกแบบว่า

cosine similar function และค่าความเหมือนใดมีค่าสูงกว่า จะหาได้จาก text i และ text j นำมาเปรียบเทียบกัน ดังนั้น ค่า cosine ของระยะห่าง หาได้จากระยะห่างของเท็กท์ทั้งสองจากสมการ(4)

$$\sin(doc_i, doc_j) = \frac{\sum_{k=1}^{m}(w_{ik} \cdot w_{jk})}{\sqrt{\sum_{k=1}^{m}(w_{ik})^2 \cdot \sum_{k=1}^{m}(w_{jk})^2}} \quad (4)$$

หลักการทำงานของการจัดกลุ่ม text โดยการรวมกันของอัลกอริทึม k-means และ SOM (SOMK) มีขั้นตอนดังนี้

ขั้นตอนที่ 1 จะใช้ Vector spatial model มาใช้เป็น text information,ลบคำที่ไม่ต้องการด้วยกระบวน conventional และใช้กฎ TF-DF จัด characteristic items ให้อยู่ในรูปแบบที่จำง่าย และกำหนดเป็น text characteristic

ขั้นตอนที่ 2 คำนวณหาค่าน้ำหนัก (weight) ของ characteristic itemsแต่ละตัว และแสดง text ด้วยเวกเตอร์

ขั้นตอนที่ 3 ข้อมูลนำเข้ามาจากเวกเตอร์ของ text ของอัลกอริทึม SOM และจัดกลุ่มข้อความด้วยวิธีของ SOM (จำนวนของโหนดอินพุทจะต้องเท่ากับมิติเวกเตอร์ และจำนวนโหนดเอาท์พุทจะต้องเท่ากับจำนวนของประเภท text) และระบุกลุ่มของค่าน้ำหนักเอาท์พุท

ขั้นตอนที่ 4 เป็นขั้นตอนสุดท้าย ที่นำอัลกอริทึม k-means โดยหาค่ากลางจากกลุ่มน้ำหนักและนำมาปรับค่ากลุ่มของ text

V. Conclusion

จากการนำความรู้ที่เกี่ยวข้องกับการจัดกลุ่มมาใช้ในงานวิจัยนี้ โดยจะวิเคราะห์อัลกอริทึมและเปรียบเทียบวิธีการของอัลกอริทึมต่าง ๆที่นำมาประกอบใช้ในการจัดกลุ่มเทกท์และรวมตัวอักษร(Characteristics) ซึ่งจะต้องอาศัยการเข้าใจภาษาธรรมชาติและการวิจัย การจัดกลุ่มไม่เพียงแต่ประยุกต์ใช้กับภาษาธรรมชาติ แต่ยังใช้แอพพลิเคชั่นการพิสูจน์และการสอดคล้องกันของภาษาธรรมชาติ โดยทั่วไปจะสร้างอัลกอริทึมเป้าหมายของอ๊อปเจกต์ขึ้นมาใหม่ เช่น การใช้อัลกอริทึมการจัดกลุ่มแบบออนโทโลจี การใช้อัลกอริทึมการจัดกลุ่มแบบเซแมนติก

ด้วยเทคโนโลยีครือข่ายและการจัดการเอกสารออนไลท์มีการใช้งานที่แพร่หลายและเอกสารออนไลท์มีปริมาณที่มาก จึงได้เสนอเทคนิคการจัดกลุ่ม ถือเป็นเทคนิคใหม่ที่มีบทบาทสำคัญที่ถูกนำมาพัฒนาและประยุกต์ใช้ เพื่อเพิ่มประสิทธิภาพและอัลกอริทึมการจัดกลุ่มเป้าหมาย

Performance	Hierarchical clustering	K-Means	Self-Organization Maps
attribute value	no requirement	numeric attribute	numeric attribute
shape	arbitrary	convex	?
measure	any	distance of normal space	euclidean distance
granularity	flexible	k and initial point	parameters
results optimization	no optimization	rebuild an optimization on	optimization
initial condition	no	yes	yes
termination condition	not precise	precise	precise
adapt to dynamic data	no	yes	yes
noise	no influence	in influence	?

ตาราง 1 การเปรียบเทียบประสิทธิภาพของ Clustering Algorithm

References

[1] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical Clustering Algorithms for Document Datasets", *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141-168, Mar 2005.

[2] Li Xinwu, "Research on text clustering algorithm based on improved K-means", in *Computer Design and Applications (ICCDA), 2010 International Conference on*, 2010, vol. 4, pp. V4-573-V4-576.

[3] Zhonghui Feng, Junpeng Bao, and Junyi Shen, "Dynamic and adaptive self organizing maps applied to high dimensional large scale text clustering", in *Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on*, 2010, pp. 348-351.

[4] Li Xinwu, "Research on Text Clustering Algorithm Based on K_means and SOM", in *Intelligent Information Technology Application Workshops, 2008. IITAW '08. International Symposium on*, 2008, pp. 341-344.