การจัดกลุ่มเอกสารโดยใช้เครือข่ายภูมิคุ้มกันเทียมด้วยวิธีการวัดความเหมือนแบบโคไซน์

DOCUMENT CLUSTERING USING ARTIFICIAL IMMUNE NETWORK WITH COSINE SIMILARITY

บัณฑิต ปุญญวัฒนะ

BUNDIT PUNYAVADTANA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2553

KMITL 2010-EN-M-070-122

DOCUMENT CLUSTERING USING ARTIFICIAL IMMUNE NETWORK WITH COSINE SIMILARITY

BUNDIT PUNYAVADTANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

FACULTY OF ENGINEERING

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2010

KMITL 2010-EN-M-070-122

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การจัดกลุ่มเอกสารโดยใช้เครือข่ายภูมิคุ้มกันเทียมด้วยวิธีการวัดความเหมือนแบบโคไซน์

Thesis Title Document Clustering using Artificial Immune Network with Cosine Similarity

นักศึกษา นายบัณฑิต ปุญญวัฒนะ

รหัสประจำตัว 48060728

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.บุญวัฒน์ อัตชู

หมายเลขวิทยานิพนธ์ KMITL-2010-EN-M-070-122

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.บุญธีร์	เครือตราชู	
ผศ.ดร.เกียรติกูล	เจียรนัยธนะกิจ	
รศ.ดร.เอื้อน	ปิ่นเงิน	
ผศ.ดร.สมศักดิ์	วลัยรัชต์	
รศ.ดร.บุญวัฒน์	อัตชู	

วัน / เดือน/ ปี ที่สอบ วันจันทร์ที่ 7 มิถุนายน พ.ศ. 2553 เวลา 11.00-13.00 น.

สถานที่สอบ ณ อาคาร A ชั้น 3 ห้องประชุม 2

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

คณะวิศวกรรมศาสตร์ รับรองแล้ว

(รองศาสตราจารย์ ดร.กอบชัย เดชหาญ)

คณบดี คณะวิศวกรรมศาสตร์

วันที่ 7 มิถุนายน พ.ศ. 2553

สำนักทะเบียนและประมวลผล สจล.
วันที่ส่งเล่มวิทยานิพนธ์ฉบับสมบูรณ์
วันที่ 17 เดือน มิ.ย. พ.ศ. 53
ลงชื่อ.....

หัวข้อวิทยานิพนธ์	การจัดกลุ่มข้อมูลโดยใช้เครือข่ายภูมิคุ้มกันเทียมด้วยวิธีการวัดความเหมือนแบบโคไซน์
นักศึกษา	นายบัณฑิต ปุญญวัฒนะ
รหัสนักศึกษา	48060728
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2553
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.บุญวัฒน์ อัตชู

บทคัดย่อ

ในปัจจุบันมีการนำการทำงานของระบบภูมิคุ้มกันมาประยุกต์ใช้แก้ปัญหาต่างๆ ในด้านการเรียนรู้ด้วยคอมพิวเตอร์ งานวิจัยนี้ได้นำเสนอการใช้ Artificial Immune Network (aiNet) ซึ่งเป็นอัลกอริทึมแบบหนึ่งของการทำงานของระบบภูมิคุ้มกันเพื่อใช้จัดกลุ่มเอกสาร การทำงานของ aiNet จะมีการคำนวณค่าแอฟฟินิตี (Affinity) โดยทั่วไปใช้การวัดระยะทางแบบยูคลิด (Euclidean distance) กับข้อมูลที่เป็นค่าแบบ Real value สำหรับงานวิจัยนี้ได้มีการปรับปรุงโดยนำวิธีการวัดความคล้ายคลึงของเอกสารโดยใช้การวัดความเหมือนแบบโคไซน์ (Cosine Similarity) มาคำนวณค่าแอฟฟินิตีของ aiNet แทนการใช้การวัดระยะทางแบบยูคลิด โดยทดลองกับเอกสารที่ถูกจัดกลุ่มไว้แล้ว ซึ่งผลที่ได้แสดงให้เห็นว่าวิธีการที่นำเสนอมีประสิทธิภาพในการจัดกลุ่มเอกสารดีกว่า

Thesis Title	Document Clustering Using Artificial Immune Network with Cosine Similarity
Student	Mr. Bundit Punyavadtana
Student ID.	48060728
Degree	Master of Engineering
Program	Computer Engineering
Year	2010
Thesis Advisor	Assoc. Prof. Dr. Boonwat Attachoo

ABSTRACT

It has recently been shown that Artificial Immune Network (aiNet) provides inspiration for solving a wide range of machine learning problems. In this research we propose the application of aiNet for document clustering. Traditional aiNet algorithm determines the affinity of real value data set by using Euclidean distance. Cosine Similarity is used to determine the affinity instead of Euclidean distance in this research. The experiment results show that our proposed technique gets better results.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้อย่างดี ด้วยคำแนะนำ และคำปรึกษาจาก รศ.ดร.บุญวัฒน์ อัตชู ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์, รศ.ดร.เอื้อน ปิ่นเงิน ที่ได้ให้คำแนะนำแนวคิดต่างๆ ในระหว่างที่ข้าพเจ้าศึกษา ข้าพเจ้ารู้สึกทราบซึ้งในความอนุเคราะห์จากท่านอาจารย์ทั้งสองท่าน ข้าพเจ้าขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบพระคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุก ๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกคนที่ให้คำแนะนำต่างๆ และคอยให้กำลังใจเสมอมา

ขอขอบคุณบัณฑิตศึกษาคณะวิศวกรรมศาสตร์ที่ให้ความช่วยเหลือ ในเรื่องต่างๆ

ขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่เป็นกำลังใจ และให้การสนับสนุนในทุกเรื่องๆ ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี

สุดท้ายนี้คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน หากวิทยานิพนธ์ฉบับนี้มีข้อผิดพลาดประการใดข้าพเจ้าขอน้อมรับไว้เพียงผู้เดียว

บัณฑิต ปุญญวัฒนะ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเทคโนโลยีสารสนเทศมีข้อมูลข่าวสารปริมาณมากขึ้น ดังเช่น ในอินเตอร์เน็ตมีเอกสารมากมายในรูปแบบต่างๆ เทคนิคในการจัดลำดับความสำคัญของเอกสาร (Ranking) ไม่เพียงพอในการที่จะเพิ่มประสิทธิภาพของการค้นคืน การจัดกลุ่มเอกสาร (Document Clustering) การกรองสารสนเทศ (Information Filtering) หรือการกลั่นกรองเอกสาร (Information Extraction) เข้ามามีบทบาทในการช่วยค้นคืนสารสนเทศให้กับผู้ใช้มากขึ้น การจัดกลุ่มเอกสารมีจุดประสงค์คือแยกเอกสารออกเป็นกลุ่มตามความคล้ายคลึงและความสัมพันธ์กันซึ่งขึ้นอยู่กับข้อความที่ปรากฏในเอกสารแต่ละฉบับ และพยายามคิดค้นวิธีที่จะจัดกลุ่มเอกสารปริมาณมากๆ แบบอัตโนมัติโดยมีงานวิจัยพัฒนาขั้นตอนและวิธีการในการจัดกลุ่มเอกสารออกมาอย่างต่อเนื่อง

ในช่วงไม่กี่ปีมานี้ระบบภูมิคุ้มกันทำให้เกิดแนวความคิดมากมายสำหรับวิธีการแก้ปัญหาที่เปลี่ยนแปลงไป กลไกการทำงานของระบบภูมิคุ้มกันได้ถูกนำมาประยุกต์ใช้กับปัญหาในด้านต่างๆ เช่น การตรวจจับผู้บุกรุก (Anomaly Detection) [1], การรู้จำแบบ (Pattern Recognition) [2], การจำแนกประเภทเอกสารเว็บ (Web Document Classification) [3], การจัดกลุ่มข้อมูล (Data Clustering) [4,5,14] เป็นต้น

จากงานวิจัย [14] ได้ทดลองจัดกลุ่มเอกสารโดยตัวแทนเอกสารเป็นข้อมูลในรูปแบบของไบนารี (binary) และได้ทดลองเปรียบเทียบการจัดกลุ่มด้วยวิธี Hierarchical Agglomerative Clustering (HAC) , K-means และ aiNet ผลการเปรียบเทียบ aiNet มีคุณภาพในการจัดกลุ่มเอกสารดีกว่า งานวิจัย [5] ได้ทดลองจัดกลุ่มเอกสารโดยตัวแทนเอกสารเป็นข้อมูลในรูปแบบของไบนารีและ real value โดยทดลองเปรียบเทียบการจัดกลุ่มด้วยวิธี HAC , K-means และ aiNet ที่มีการปรับรูปแบบในแบบต่างๆ วัดคุณภาพการจัดกลุ่มเอกสารโดยค่าความถูกต้องและค่าเอฟ-เมเชอร์ (F-measure) [21] ผลการจัดกลุ่มที่ใช้ตัวแทนเอกสารในรูปของ real value และ aiNet ให้คุณภาพในการจัดกลุ่มดีกว่า จากงานวิจัยที่กล่าวมาจะเห็นได้ว่า aiNet มีประสิทธิภาพที่ดีในการจัดกลุ่มเอกสาร

หลักการสำคัญในการจัดกลุ่มเอกสารโดย aiNet คือ การคำนวณค่าแอฟฟินิตี (affinity) (ความสามารถในการยึดเกาะระหว่างแอนติบอดีกับแอนติบอดีหรือระหว่างแอนติบอดีกับแอนติเจน) ระหว่าง system unit และ input data ของ aiNet โดยปกติ aiNet ใช้การวัดระยะทางแบบยูคลิดในการคำนวณค่าแอฟฟินิตี กับข้อมูลชนิด real value ปัญหาก็คือการวัดระยะทางแบบยูคลิดนั้นเมื่ออินพุตเป็นเวกเตอร์ที่มีมิติจำนวนมากจะทำให้ค่าของระยะทางระหว่างเวกเตอร์ทั้ง

สองมีค่ามากทำให้การวัดความคล้ายคลึงเอกสารมีข้อผิดพลาดเกิดขึ้นส่งผลให้การจัดกลุ่มเอกสารมีคุณภาพด้อยลง

ดังนั้นงานวิจัยนี้ได้นำเสนอการปรับปรุง aiNet โดยใช้การคำนวณค่าแอฟฟินิตีด้วยการวัดความเหมือนแบบโคไซน์(Cosine Similarity) [8] ซึ่งเป็นวิธีที่นิยมใช้วิธีหนึ่งในการวัดความคล้ายคลึงของเอกสารโดยวัดมุมระหว่างเวกเตอร์แทนเพื่อลดข้อผิดพลาด และได้ทำการทดลองจัดกลุ่มเอกสาร โดยเปรียบเทียบผลของทั้งสองวิธีโดยวัดที่ความถูกต้องและค่าเอฟ-เมเชอร์ด้วยวิธีการที่นำเสนอนี้จะทำให้ aiNet ที่ปรับปรุงให้ผลความถูกต้องและค่าเอฟ-เมเชอร์สูงกว่า

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาการทำงานของ aiNet และข้อจำกัดของโมเดล aiNet
2. เพื่อศึกษาแนวทางในการพัฒนา aiNet และนำเสนอการปรับวิธีวัดค่าแอฟฟินิตีของโมเดล aiNet
3. เพื่อศึกษาการจัดกลุ่มเอกสารและการนำ aiNet มาประยุกต์ใช้เพื่อช่วยให้การจัดกลุ่มเอกสารมีประสิทธิภาพมากยิ่งขึ้น
4. เพื่อเปรียบเทียบประสิทธิภาพโมเดล aiNet แบบเดิมและแบบที่ปรับปรุงที่ผู้วิจัยได้พัฒนาขึ้นในการจัดกลุ่มเอกสาร

1.3 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

สำหรับการศึกษาวิธีการในการจัดกลุ่มเอกสารโดยใช้ aiNet จะต้องอาศัยหลักการทฤษฎีเหล่านี้

1. การหาตัวแทนเอกสารในแบบของเวกเตอร์สเปซโมเดล (Vector Space Model)
2. การวัดความคล้ายของข้อมูล
3. การประเมินผลคุณภาพของการจัดกลุ่ม (Evaluation of Clustering Quality)
4. อัลกอริทึม aiNet

1.4 ขอบเขตของการวิจัย

1. ศึกษาเปรียบเทียบประสิทธิภาพในการจัดกลุ่มข้อมูลของ aiNet แบบเดิมกับโมเดลใหม่ที่ผู้วิจัยได้ปรับวิธีวัดค่าแอฟฟินิตีจากแบบเดิม
2. เอกสารที่ใช้ทดสอบเป็นบทความภาษาอังกฤษโดยที่เอกสารในการทดสอบถูกจัดกลุ่มไว้เรียบร้อยแล้วและเลือกเอกสารโดยการสุ่มจากบางกลุ่มของเอกสารให้ได้จำนวนตามที่ต้องการ

3. ค่าของอินพุตที่ใช้ในการทดลองเป็นค่าชนิด real value เท่านั้น

1.5 ขั้นตอนของการศึกษา

1. ศึกษาความเป็นมาของระบบภูมิคุ้มกันเทียมและการทำงานของ aiNet
2. ศึกษางานวิจัยที่เกี่ยวข้องกับการประยุกต์ใช้ aiNet ในงานจัดกลุ่มเอกสาร
3. ปรับปรุงโมเดลของ aiNet โดยปรับวิธีการหาค่าแอฟฟินิตี
4. ทำการทดสอบโมเดล aiNet แบบเดิมกับโมเดลที่ปรับปรุงวิธีการในการหาค่าแอฟฟินิตี
5. สรุปผลการทดลองพร้อมจัดทำบทความตีพิมพ์และวิทยานิพนธ์

1.6 รายละเอียดในแต่ละบท

วิทยานิพนธ์ฉบับนี้ได้แบ่งเนื้อหาออกเป็น 5 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความเป็นมาของงานวิจัย ความมุ่งหมายและวัตถุประสงค์ สมมติฐานของการศึกษา รวมทั้งทฤษฎีหรือแนวคิดที่ใช้ในการศึกษา ขอบเขตและขั้นตอนของการศึกษาที่ใช้สำหรับงานวิจัยนี้

บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานของระบบภูมิคุ้มกันตามธรรมชาติและระบบภูมิคุ้มกันเทียม การจัดกลุ่มเอกสารและเทคนิคต่างๆ การประเมินคุณภาพของการจัดกลุ่มเอกสาร

บทที่ 3 กล่าวถึง aiNet และศึกษาการประยุกต์ใช้ aiNet ในการนำไปประยุกต์ใช้งานและนำเสนอโมเดลของ aiNet ที่ปรับปรุงในงานวิจัยนี้

บทที่ 4 กล่าวถึงการทดลองเพื่อทดสอบโมเดล aiNet แบบเดิมกับโมเดลที่มีการปรับปรุงวิธีการหาค่าแอฟฟินิตีและเปรียบเทียบผลที่ได้จากการทดลอง

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ และแนวทางการทำวิจัยต่อ

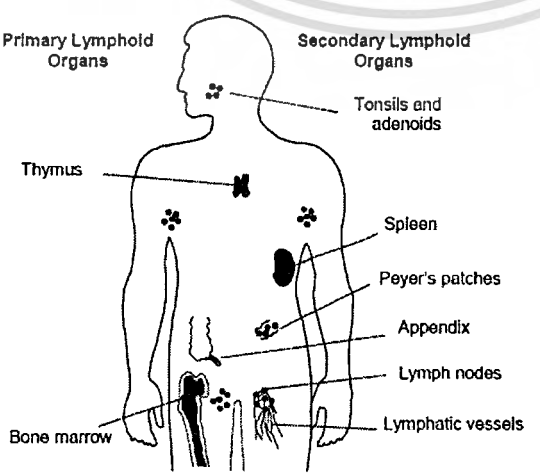
บทที่ 2
ทฤษฎีพื้นฐานที่ใช้ในการวิจัยและการจัดกลุ่มข้อมูล

ในบทนี้จะกล่าวถึงระบบภูมิคุ้มกันตามธรรมชาติและคุณสมบัติที่น่าสนใจ เพื่อเป็นพื้นฐานสำหรับทำความเข้าใจเกี่ยวกับระบบภูมิคุ้มกันเทียม ซึ่งนำกลไกการทำงานของระบบภูมิคุ้มกันตามธรรมชาติมาเป็นต้นแบบในการสร้างแนวทางใหม่ๆในการแก้ปัญหาทางคอมพิวเตอร์และการนำไปประยุกต์ใช้งานด้านต่างๆ และกล่าวถึงการจัดกลุ่มข้อมูล(Data Clustering) การดึงคุณลักษณะของเอกสารเพื่อใช้ในการจัดกลุ่ม ตัวอย่างอัลกอริทึมที่ใช้ในการจัดกลุ่ม การประเมินผลคุณภาพการจัดกลุ่มเอกสาร

2.1 ระบบภูมิคุ้มกันทางชีววิทยา

ระบบภูมิคุ้มกัน (Immune System) เป็นกลไกตามธรรมชาติของร่างกายที่ทำหน้าที่คอยป้องกันไม่ให้เชื้อโรคหรือสิ่งแปลกปลอมที่เป็นอันตรายเข้ามาทำอันตรายต่อร่างกายหรือเมื่อหลุดเข้ามาแล้ว ระบบภูมิคุ้มกันก็จะพยายามกำจัดสิ่งแปลกปลอมให้หมดไปจากร่างกายโดยเร็วและมีประสิทธิภาพ

สิ่งแปลกปลอมหรือแอนติเจน (Antigen; Ag) คือสิ่งแปลกปลอมหรือสารอะไรก็ตามที่สามารถชักนำ (induce) ให้ร่างกายสร้างแอนติบอดีหรือลิมโฟไซต์ที่ถูกกระตุ้นแล้ว และพร้อมจะตอบสนองต่อแอนติเจนที่มากระตุ้นโดยเฉพาะ (specifically sensitized lymphocyte) ซึ่งอาจเป็นชนิด ภูมิคุ้มกันในน้ำเหลือง (humoral immune response) หรือ ภูมิคุ้มกันชนิดพึ่งเซลล์ (cell-mediated immune response) และสามารถทำปฏิกิริยาจำเพาะกับแอนติบอดีหรือ lymphocyte นั้นๆ ได้ โดยมีการเปลี่ยนแปลงของเซลล์บี (B lymphocyte ,B cell) และเซลล์ที (T lymphocyte ,T cell)



รูปที่ 2.1 กายวิภาคศาสตร์ของระบบภูมิคุ้มกัน

2.1.1 หน้าที่หลักของอวัยวะ ในระบบภูมิคุ้มกัน

ระบบภูมิคุ้มกันมีอวัยวะที่เกี่ยวข้องอยู่ 2 ระดับซึ่งมีตำแหน่ง และหน้าที่หลักของอวัยวะดังนี้

2.1.1.1 Primary lymphoid organs

1. ไขกระดูก (Bone Marrow) เป็นต้นกำเนิดของเซลเม็ดเลือด เป็นอวัยวะที่สำคัญอวัยวะหนึ่ง มีหน้าที่สร้างเม็ดเลือด เลือดของคนเราประกอบด้วยเม็ดเลือดและพลาสมา เม็ดเลือดมี 3 ชนิด คือ เม็ดเลือด แดงเม็ดเลือดขาว และเกล็ดเลือด เม็ดเลือดแดงเป็นเม็ดเลือดส่วนใหญ่ที่มีในเลือดทำให้เลือดมีสีแดง เม็ดเลือดแดงมีฮีโมโกลบินเป็นส่วนประกอบ ทำหน้าที่นำออกซิเจนจากปอดไปเลี้ยงส่วนต่างๆ ของร่างกาย รวมทั้งนำคาร์บอนไดออกไซด์และของเสียต่างๆ กลับไปทิ้งยังปอด เม็ดเลือดอีก 2 ชนิด ซึ่งมีในเลือดในปริมาณที่น้อยกว่า คือเม็ดเลือดขาว และเกล็ดเลือด เม็ดเลือดขาวทำหน้าที่ป้องกัน ต่อสู้ และทำลายเชื้อโรค

2. ต่อมไทมัส (Thymus) เป็นต่อมที่อยู่ในทรวงอก บริเวณขั้วของหัวใจ ต่อมไทมัสจะเจริญเติบโต และมีขนาดใหญ่ที่สุดในวัยเด็ก มีหน้าที่สำคัญ ในระบบภูมิคุ้มกันของเรา คือ เมื่อไขกระดูกสร้างเม็ดเลือดขาว (Lymphocyte) ออกมาแล้ว เม็ดเลือดขาวเหล่านี้ ยังไม่สามารถใช้การได้ จะต้องเดินทางมายังต่อมไทมัส เพื่อเปลี่ยนแปลงเป็นเม็ดเลือดขาว ที่โตเต็มวัยสามารถต่อสู้กับสิ่งแปลกปลอมที่มารุกรานร่างกายของเรา และจะหลั่งฮอร์โมน ซึ่งทำให้เม็ดเลือดขาว แบ่งตัวเพิ่มจำนวนให้มากขึ้นได้ เมื่อเราโตขึ้น ต่อมไทมัสจะหยุดเจริญ และฝ่อไปจนเหลือขนาดเล็กๆ เท่านั้น เนื้อที่ส่วนใหญ่ของต่อม จะถูกแทนที่ด้วยเนื้อเยื่อไขมัน (adipose tissue) และเนื้อเยื่อเกี่ยวพันชนิดหลวม (loose or areolar connective tissue)

2.1.1.2 Secondary lymphoid organs

1. ต่อมทอนซิลและต่อมอดีนอยด์ (Tonsil and adenoids) เป็นต่อมน้ำเหลือง 2 ต่อมที่ตั้งอยู่ในช่องปาก มีหน้าที่หลักคือ การจับและทำลายเชื้อโรค ที่จะเข้าสู่ร่างกายทางช่องทางเดินอาหารซึ่งเป็นด่านแรก หน้าที่รองคือ สร้างภูมิคุ้มกันแต่ไม่ใช่ส่วนที่สำคัญ หน้าที่หลักคือ การทำลายเชื้อโรคในช่องปากมากกว่าเป็นกับดักของเชื้อโรค ต่อมทอนซิลจะทำงานร่วมกับต่อมน้ำเหลืองอีก 2 ต่อมบริเวณคอ คือ ต่อมอดีนอยด์และต่อมน้ำเหลืองที่โคนลิ้น ต่อมอดีนอยด์และต่อมทอนซิลจะหลั่งอิมมูโนโกลบูลิน ซึ่งทำหน้าที่ดักจับเชื้อโรคที่ลงมาในลำคอ และคอยต่อสู้กับเชื้อโรคที่มาทางจมูกและลำคอด้วย ต่อมทอนซิลจะทำหน้าที่ด้านระบบภูมิคุ้มกันมากที่สุดเมื่ออายุ 4-10 ปี หลังจากนั้นจะมีขนาดเล็กลง แต่ยังทำงานเกือบตลอดชีวิต ถ้าต่อมทอนซิลอักเสบบ่อยๆ การอักเสบจะทำให้เม็ดเลือดขาวในต่อมทอนซิลลดลง ต่อมทอนซิลจะฆ่าเชื้อโรคและสร้างภูมิคุ้มกันได้ลดลง และในบางครั้งแทนที่ต่อมทอนซิลจะเป็นที่กินเชื้อโรค แต่กลับกลายเป็นที่เก็บเชื้อโรคแทน ทำให้เกิดการอักเสบขึ้นมาใหม่ ซึ่งเป็นสาเหตุของการกลับมาเป็นซ้ำบ่อยๆ

2. ต่อมน้ำเหลือง (Lymph nodes) มีหน้าที่เป็นด่านกรองเชื้อโรคและช่วยสร้างเม็ดเลือดขาว เมื่อเม็ดเลือดขาวกำจัดเชื้อโรคได้ไม่หมด เชื้อโรคจะแพร่กระจายเข้าสู่ต่อมน้ำเหลือง ต่อมน้ำเหลือง จะทำหน้าที่กำจัด เชื้อโรคต่อไป โดยการกรองเชื้อโรคไว้แล้วทำลายทิ้ง ต่อมน้ำเหลืองตั้งอยู่ตามบริเวณต่าง ๆ ของร่างกาย

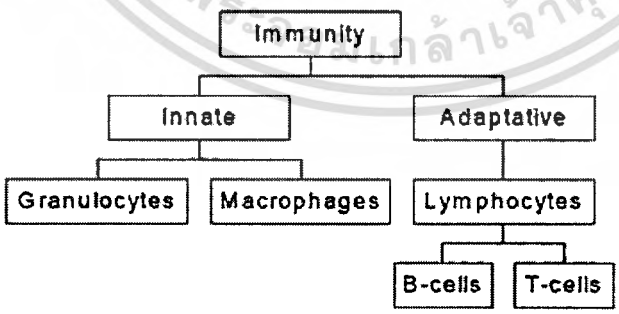
3. ไส้ติ่ง (Appendix and Peyerispatches) มีหน้าที่เป็นอวัยวะที่เกี่ยวข้องกับน้ำเหลือง ซึ่งช่วยให้ เซลล์บี พัฒนาได้เต็มที่ (เซลล์บี เป็นเม็ดเลือดขาวประเภทหนึ่ง) และช่วยในการผลิตแอนติบอดีประเภท immunoglobulin A (IgA)อีกด้วย นักวิจัยแสดงให้เห็นว่าไส้ติ่งเกี่ยวข้องกับการผลิตโมเลกุลที่ช่วยกำกับการเคลื่อนที่ของ lymphoidcyte ในร่างกายหลายที่

4. ม้าม (Spleen) ทำหน้าที่ในการดึงเอาธาตุเหล็กจากฮีโมโกลบินของเซลล์เม็ดเลือดแดงมาใช้ในร่างกาย และยังเอาของเสียออกจากกระแสเลือดในรูปของน้ำปัสสาวะเช่นเดียวกับตับ ม้ามสร้างแอนตีบอดี ในการต่อต้านเชื้อโรค

5. หลอดน้ำเหลือง (Lymphatic vessels) ภายในประกอบด้วยน้ำเหลืองซึ่งจะเชื่อมต่อระหว่างต่อมน้ำเหลืองแต่ละแห่ง

2.1.2 ชนิดและการทำงานของระบบภูมิคุ้มกัน

ในร่างกายมนุษย์ประกอบด้วยระบบภูมิคุ้มกันต่าง ๆ มากมายหลายระบบ ซึ่งแต่ละระบบจะมีกลไกการทำงานที่แตกต่างกัน โดยสามารถจำแนกลักษณะการทำงานของระบบภูมิคุ้มกันตามความจำเพาะเจาะจงในการป้องกันสิ่งแปลกปลอมได้เป็น 2 ลักษณะ คือ ระบบภูมิคุ้มกันแบบไม่จำเพาะเจาะจง (Nonspecific immune response หรือ Innate immune response) และระบบภูมิคุ้มกันแบบจำเพาะเจาะจง (Specific immune response หรือ Adaptative immune response) ซึ่งแสดงดังรูปที่ 2.2



รูปที่ 2.2 แสดงชนิดของระบบภูมิคุ้มกัน

2.1.2.1 ระบบภูมิคุ้มกันแบบไม่จำเพาะเจาะจง เป็นการกำจัดสิ่งแปลกปลอมออกจากร่างกายโดยวิธีการง่ายๆ มีความสามารถในการป้องกันหรือทำลายเชื้อจุลินทรีย์หรือสิ่งแปลกปลอมไม่สูงนัก อาจกำจัดเชื้อจุลินทรีย์ได้เพียงระดับหนึ่งเท่านั้น ระบบภูมิคุ้มกันนี้จะมีการ

เปลี่ยนแปลงไปตามอายุ พันธุกรรม ฮอร์โมน และภาวะโภชนาการของแต่ละบุคคล เกิดขึ้นเมื่อร่างกายได้รับสิ่งแปลกปลอมนั้นเป็นครั้งแรก หรือแม้ได้รับอีกในคราวต่อมา ร่างกายก็อาจใช้วิธีการนี้กำจัดสิ่งแปลกปลอมร่วมกับระบบภูมิคุ้มกันแบบจำเพาะเจาะจง

1. Barrier หรือเครื่องกีดขวางตามธรรมชาติ ซึ่งได้แก่ผิวหนัง เยื่อเมือก ซึ่งบุตามอวัยวะต่างๆ ขนอ่อน(cilia) เมื่อสิ่งแปลกปลอมนั้นสามารถผ่าน barrier นี้เข้าไปได้จะถูกร่างกายกำจัดโดยใช้ inflammatory response และ phagocytosis

2. Inflammatory response เป็นการเคลื่อนย้ายของ phagocytic cell (neutrophilic granulocyte และ macrophage) มายังบริเวณที่มีสิ่งแปลกปลอม บริเวณนั้นจะมีลักษณะจำเพาะคือ ปวด บวม แดง ร้อน และจะพบว่าประมาณ 30-60 นาที หลังจากที่สิ่งแปลกปลอมเข้าไป เม็ดเลือดขาวจำพวก neutrophilic granulocyte จะเป็นพวกแรกที่มาถึงบริเวณนี้ โดยการลอดตัวผ่านออกทางรอยต่อของ endothelial cell ของเส้นเลือดออกมาในเนื้อเยื่อ เพื่อจะมากินและทำลายสิ่งแปลกปลอมนั้นประมาณ 4-5 ช.ม. หลังจากนั้นเซลล์อีกพวกหนึ่งคือ mononuclear cells ซึ่งได้แก่ Iymphocyte จึงจะผ่าน endothelial cell ออกมาแล้ว monocyte จะเปลี่ยนเป็น macrophage ส่วนเม็ดเลือดขาว Iymphocyte จะมาทำหน้าที่ specific immune response ดังจะได้กล่าวต่อไป

3. Phagocytosis หรือ cell-eating เมื่อพวก neutrophilic granulocytes และ macrophage มาถึง จะเคลื่อนตัวไปหาสิ่งแปลกปลอมนั้น (chemotaxis) แล้วประกบติด (attachment) ต่อมาจะกลืน(ingestion) แล้วจึงมีการย่อย (intracellular digestion) ด้วยกลไกหลายอย่างในเซลล์ แล้วจึงปล่อยสิ่งแปลกปลอมที่ถูกทำลายแล้วออกไปจากเซลล์ (elimination)

2.1.2.2 ระบบภูมิคุ้มกันแบบจำเพาะเจาะจง เป็นการกำจัดสิ่งแปลกปลอมที่ต้องอาศัยกลไกที่ยุ่งยากกว่าวิธีแรกเกิดขึ้นเมื่อร่ายกายไม่สามารถใช้วิธีระบบภูมิคุ้มกันแบบจำเพาะเจาะจงกำจัดสิ่งแปลกปลอมนั้นออกไปได้ เซลล์ที่มีหน้าที่รับผิดชอบในด้านนี้คือ lymphocytes สิ่งแปลกปลอมในที่นี้มีชื่อเรียกใหม่ว่า แอนติเจน (antigen) หรืออิมมูโนเจน (immunogen) การตอบสนองดังกล่าว แบ่งออกเป็น 2 ส่วน คือ

1. ระบบภูมิคุ้มกันจากกระแสเลือดและสารคัดหลั่ง (Humoral Immune Response; HIR) คือ ระบบภูมิคุ้มกันที่เกิดจากเซลล์บีตอบสนองต่อแอนติเจนแต่ละชนิดอย่างจำเพาะเจาะจง ทำให้มีการสร้างแอนติบอดีขึ้น เพื่อกำจัดแอนติเจนต่าง ๆ ที่เข้ามาในร่างกาย เรียกว่าแอนติบอดีที่สร้างขึ้นอย่างจำเพาะนี้ว่า อิมมิวโนโกลบูลิน (immunoglobulin) หรือ การตอบสนองทางอิมมูนโดยการใช้สารน้ำ ซึ่งหมายถึง แอนติบอดี (antibody) เซลล์ที่รับผิดชอบในเรื่องนี้คือ B Lymphocyte และ plasma cell นอกจากนี้ยังมีสารน้ำอื่นๆ ช่วยส่งเสริมการทำงานของ specific immunity คือ complement

เมื่อร่างกายเราได้รับแอนติเจน ร่างกายของเราจะสามารถสร้างแอนติบอดีได้ภายใน 14 วัน ทั้งนี้ขึ้นกับชนิดของแอนติเจน ปริมาณของแอนติเจนที่ได้รับ และวิธีการเข้าสู่ร่างกาย โดย

ระบบภูมิคุ้มกันจากกระแสเลือดและสารคัดหลั่ง เป็นระบบที่สามารถถ่ายทอดจากผู้ที่มีภูมิคุ้มกัน (immunized donor) ไปยังผู้ที่ยังไม่มีภูมิคุ้มกัน (native host) ได้ ด้วยการส่งผ่านทางกระแสเลือด

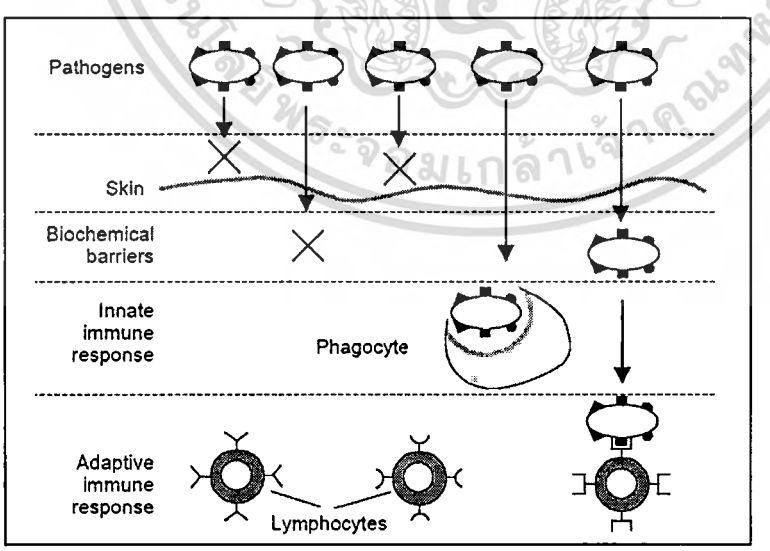
2. ระบบภูมิคุ้มกันจากเซลล์ (cell-mediated immune response; CMIR หรือ cell-mediated immunity; CMI) คือ ระบบภูมิคุ้มกันที่เกิดจากการตอบสนองทางภูมิคุ้มกันของเซลล์ ซึ่งเซลล์ที่ทำหน้าที่รับผิดชอบ คือ เซลล์ลิมโฟไซต์ที่มีการตอบสนองต่อสารจำเพาะ (specificcally sensitized lymphocyte; SSL) หรือ T lymphocyte ซึ่งมีหน้าที่ผลิตสาร lymphokines เซลล์ที่ซึ่งมีการพัฒนาผ่านทางต่อมไทมัส จนได้เป็นเซลล์ที่สมบูรณ์ 3 ชนิด คือ เซลล์ทีทำลายสิ่งแปลกปลอม เซลล์ทีผู้ช่วย และเซลล์ทีกดระงับ

-เซลล์ทีทำลายสิ่งแปลกปลอม หรือเซลล์ทีไซโททอกซิก (cytotoxic T cell; Tc) ทำหน้าที่ทำลายแอนติเจนที่เข้าสู่ร่างกาย ซึ่งได้แก่ เซลล์จุลินทรีย์ เซลล์ร่างกายที่ติดเชื้อ หรือเซลล์มะเร็ง เซลล์ทีจะหลั่งโปรตีนออกมาทำลายเซลล์ติดเชื้อให้แตกสลายและตายในที่สุด

-เซลล์ทีผู้ช่วย หรือเซลล์ทีเฮลเปอร์ (helper T cell; TH) ทำหน้าที่กระตุ้นลิมโฟไซต์ชนิดบี ให้สร้างแอนติบอดีที่จำเพาะต่อชนิดแอนติเจน ทั้งยังทำหน้าที่กระตุ้นการทำงานของเซลล์ทีชนิดอื่น ๆ ด้วย

-เซลล์ทีกดระงับ หรือเซลล์ทีซัพเพรสเซอร์ (supressor T cell; Ts) ทำหน้าที่ควบคุมการทำงานของลิม โฟไซต์ชนิดบีและชนิดทีที่เป็นเซลล์ทีผู้ช่วย หรือเซลล์ทีทำลายสิ่งแปลกปลอมให้อยู่ในสภาวะสมดุล

ซึ่งเซลล์ทีต่าง ๆ เหล่านี้จะไปสะสมอยู่ตามอวัยวะต่าง ๆ ได้แก่ ต่อมน้ำเหลือง ต่อมทอนซิลและม้าม รวมถึงกระแสเลือดทั่วร่างกาย



รูปที่ 2.3 แสดงกลไกการป้องกันร่างกายจากการรุกรานทำลายจากสิ่งแปลกปลอม

จากรูปที่ 2.3 แสดงถึงกลไกการป้องกันร่างกายจากการรุกรานทำลายจากสิ่งแปลกปลอม ซึ่งมีหลายระดับชั้น ดังนี้

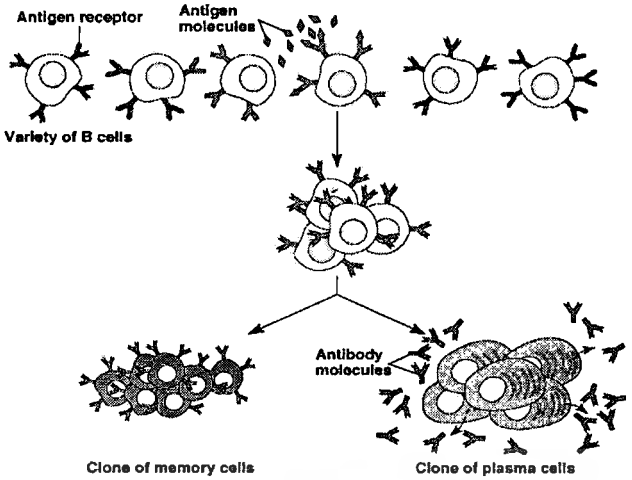
Physical barriers : ผิวหนัง(Skin) เป็นด่านป้องกันที่อยู่ด้านนอกของร่างกาย มีบทบาทในการป้องกันเชื้อจุลินทรีย์ ฝุ่นละอองรวมทั้งสิ่งแปลกปลอมต่าง ๆ ไม่ให้เข้าสู่ร่างกาย โดยที่ผิวหนังจะมีความชุ่มชื้นต่ำ ทำให้เชื้อจุลินทรีย์ต่าง ๆ ที่มาเกาะตามผิวหนังขาดความชุ่มชื้นและตายได้ในที่สุด นอกจากนี้ที่ผิวหนังยังมีสารกลุ่มเคอราติน (keratin) ซึ่งช่วยป้องกันการติดเชื้อ และผิวหนังยังสามารถขจัดเชื้อจุลินทรีย์ออกไปได้ ด้วยการหลุดลอกของผิวหนังชั้นนอก เยื่อบุผิว เป็นส่วนที่มีเยื่อเมือกช่วยดักจับเชื้อจุลินทรีย์ด้วยการหุ้มเคลือบ โดยประกอบกับการทำงานของขนที่มีขนาดเล็ก (cilia) ซึ่งสามารถพบได้ตามระบบทางเดินหายใจ เช่น โพรงจมูก ช่วยกวาดสิ่งแปลกปลอมหรือเชื้อจุลินทรีย์ให้เคลื่อนที่ไปทางหลอดลมหรือโพรงจมูก และขับออกจากร่างกายโดยการไอ จาม หรือขับออกในรูปเสมหะ ที่อาจคายออกหรือกลืนลงสู่กระเพาะอาหารแล้วถูกขับออกทางอุจจาระได้ นอกจากโพรงจมูกแล้ว กลไกการป้องกันสิ่งแปลกปลอมเช่นนี้ อาจพบได้ตามช่องเปิดของร่างกายส่วนต่าง ๆ อีกด้วย

Biochemical barriers: การป้องกันโดยสารเคมีในร่างกาย (chemical factor) คือ กลไกการป้องกันสิ่งแปลกปลอมเข้าสู่ร่างกายที่เกิดขึ้นจากสารเคมีต่าง ๆ ที่ร่างกายหลั่งออกมา ทำให้เกิดสภาพที่ไม่เหมาะสมต่อการเจริญเติบโตของเชื้อจุลินทรีย์ เช่น เอนไซม์บางชนิดที่สามารถยับยั้งการเจริญเติบโตของเชื้อจุลินทรีย์ สารคัดหลั่งบางชนิดที่ทำให้ร่างกายมีสภาพความเป็นกรด-เบสสูงจนไม่เหมาะต่อการเจริญเติบโตของเชื้อจุลินทรีย์ เป็นต้น

สำหรับชั้นของระบบภูมิคุ้มกันแบบไม่จำเพาะเจาะจง (Innate immune response) และระบบภูมิคุ้มกันแบบจำเพาะเจาะจง (Adaptative immune response) นั้นรายละเอียดได้กล่าวไว้แล้วในหัวข้อที่ 2.1.2.1 และ 2.1.2.2

2.1.3 ทฤษฎีการโคลน (Clonal Selection Theory)

Sir Macfarlane Burnet [9] กล่าวไว้ว่าเซลล์บีทุกตัว จะมียีน (genes) กำหนดการทำหน้าที่ของมันอยู่แล้วภายในเซลล์ ตั้งแต่ยังไม่ได้พบกับแอนติเจน ซึ่งยีนนั้นจะเป็นตัวกำหนดชนิดของแอนติเจนที่ลิมโฟไซต์ และจะตอบสนองเมื่อมีแอนติเจนเข้าสู่ร่างกายจะไปจับกับเซลล์บีที่จำเพาะแล้วเซลล์บีจะตอบสนองโดยการแบ่งตัว (proliferation) และเปลี่ยนแปลงรูปร่าง (differentiation) จนเกิดเป็นกลุ่ม (clone) ของเซลล์ที่ทำหน้าที่ผลิตแอนติบอดีที่จำเพาะต่อแอนติเจนนั้นๆ



รูปที่ 2.4 แสดงการทำงานแบบโคลนอลซีเล็คชัน (Clonal selection)

ขั้นตอนการทำงานแบบโคลนอลซีเล็คชันแสดงดังรูปที่ 2.4

1. แอนติเจนจับกับตัวรับแอนติเจนบนเซลล์บี
2. เซลล์บีที่มีตัวรับที่จำเพาะต่อแอนติเจนนั้นจะเพิ่มจำนวนเป็นกลุ่ม
3. บางเซลล์พัฒนาไปเป็น short-lived plasma cell และหลั่งแอนติบอดี
4. บางเซลล์พัฒนาไปเป็น long-lived memory cell ที่จะทำให้เกิดการตอบสนองอย่างรวดเร็วเมื่อร่างกายได้รับแอนติเจนเดิม

2.1.4 แอฟฟินิตีมัทรูเรชัน (Affinity Maturation)

แอฟฟินิตีมัทรูเรชัน (Affinity Maturation) เป็นกระบวนการเปลี่ยนแปลงของโมเลกุลผิวเซลล์ที่ทำหน้าที่ในจับกับแอนติเจนจะเลือกจับคู่กับแอนติเจนที่แรงกว่า แอฟฟินิตี หมายถึงความสามารถในการยึดเกาะของโมเลกุลผิวเซลล์กับแอนติเจน การจับที่มีแอฟฟินิตีสูงกว่าหมายถึงว่ามีการจดจำและมีการตอบสนองที่ดีกว่า ภูมิคุ้มกันมีการปรับตัวเนื่องจากการเปลี่ยนแปลงของโมเลกุลผิวเซลล์ที่ทำหน้าที่จับกับแอนติเจน ซึ่งจะต่อต้านแอนติเจนได้สำเร็จและนำไปสู่การตอบสนองในการต่อต้านแอนติเจนที่มีประสิทธิภาพมากขึ้น

ลิมโฟไซต์ (lymphocytes) เป็นเซลล์ของระบบภูมิคุ้มกันภายในร่างกายที่ไม่ใช่เซลล์สืบพันธุ์ การเปลี่ยนแปลงที่เกิดขึ้นระหว่างแอฟฟินิตีมัทรูเรชันเรียกว่า โซมาติกมิวเตชัน (somatic mutation) อัตราการเปลี่ยนแปลงจะสูงขึ้นระหว่างการเพิ่มจำนวน เราเรียกว่าโซมาติกไฮเปอร์มิวเตชัน (somatic hypermutation) การเปลี่ยนแปลงจะผกผันกับความแรงของการจับของโมเลกุลผิวเซลล์ที่ทำหน้าที่ในการจับกับแอนติเจน ความสามารถในการยึดเกาะที่สูง จะมีอัตราการเปลี่ยนแปลงน้อยและถ้าความแรงของการจับต่ำก็จะมีอัตราการเปลี่ยนแปลงมาก ซึ่งเป็นกลยุทธ์ของระบบภูมิคุ้มกันเพื่อรักษาความแรงในการจับไว้ และในเวลาเดียวกันมันจะมีโอกาสที่สร้างโมเลกุลผิวเซลล์ที่ทำหน้าที่ในการจับสูงกว่า

ความผกผันของการเปลี่ยนแปลงกับอัตราการขยายของเซลล์ในการจับโดยตรงกับแอนติเจนจะเป็นสัดส่วนกัน เมื่อสิ่งแปลกปลอมบุกรุกจำนวนเซลล์ภูมิคุ้มกันที่จดจำแอนติเจนจะมีระดับความแรงในการจับที่แตกต่างกัน เซลล์เหล่านี้จะเกิดกระบวนการเพิ่มจำนวนแบบโคลนอลซีเล็คชันและแอฟฟินิตีมัทรูเรชัน จำนวนเซลล์ที่ได้มีสัดส่วนกับความแรงของการจับกับแอนติเจน ความแรงสูงจะมีการเพิ่มจำนวนของเซลล์ที่ได้รับการจดจำมากกว่าเซลล์ที่มีความแรงในการจับกับแอนติเจนต่ำ กระบวนการเพิ่มจำนวนมีผลต่อเซลล์บี และ เซลล์ที แต่แอฟฟินิตีมัทรูเรชันจะกล่าวเฉพาะในเซลล์บี

2.1.5 ทฤษฎีเครือข่ายภูมิคุ้มกัน (Immune Network Theory)

ทฤษฎีเครือข่ายภูมิคุ้มกัน [10] แสดงถึงปฏิกิริยาระหว่างแอนติบอดีกับแอนติบอดีอื่นๆ ซึ่งไม่ใช่ปฏิกิริยาระหว่างแอนติบอดีกับแอนติเจน แอนติบอดีจะเชื่อมต่อกันเป็นเครือข่าย โดยแอนติบอดีที่เชื่อมต่อกันนั้นจะมีรูปร่างเหมือนกับแอนติเจนที่มากระตุ้น เครือข่ายจะเกิดปฏิกิริยาตอบสนองหรือไม่ตอบสนองก็ได้ ถ้าเกิดการตอบสนองต่อแอนติเจนก็จะทำการแบ่งตัวเพิ่มจำนวน เกิดการเคลื่อนไหวและสร้างแอนติบอดีออกมา ถ้าไม่เกิดการตอบสนองเซลล์จะถูกกำจัดออกจากเครือข่าย ปฏิกิริยาของจำนวนเซลล์ภายในเครือข่ายมีความแตกต่างกันมาก แต่อธิบายโดยรวมได้ดังสมการที่ (2.1)

$$RVP = \textit{Influx of new cells} - \textit{Death of unstimulated cells} + \textit{reproduction of stimulated cells} \qquad (2.1)$$

เมื่อ RPV(Rate of population variation) คืออัตราการเปลี่ยนแปลงของแอนติบอดีภายในเครือข่าย

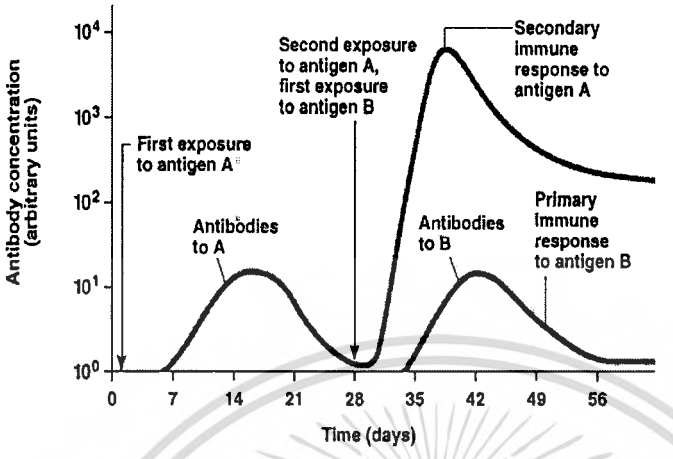
2.1.6 การตอบสนองทางภูมิคุ้มกัน

การตอบสนองทางภูมิคุ้มกันมีลักษณะจำเพาะ ดังนี้คือ

1.มีความสามารถจำแนกได้ว่าสิ่งใดเป็นสิ่งแปลกปลอมและสิ่งใดเป็นของตัวเอง (differentiation of self from non-self) โดยที่จะมีการตอบสนองทางภูมิคุ้มกันกันเฉพาะต่อสิ่งแปลกปลอม (non-self) เท่านั้น

2.มีความจำเพาะ (specificity) การตอบสนองจะเกิดขึ้นจำเพาะต่อสิ่งแปลกปลอม หรือแอนติเจนที่เข้ามาเท่านั้น

3.มีความจำ (memory) เมื่อได้รับแอนติเจนชนิดเดียวกันเป็นครั้งที่ 2 หรือครั้งที่ 3 จะมีการตอบสนองที่รวดเร็วและด้วยปริมาณที่มากกว่าการตอบสนองที่เกิดขึ้นเมื่อได้รับแอนติเจนเป็นครั้งแรกแสดงดังรูปที่ 2.5



รูปที่ 2.5 แสดงการตอบสนองทางภูมิคุ้มกัน

ในการเพิ่มจำนวนของลิมโฟไซต์ที่ถูกคัดเลือก หลังจากเผชิญกับแอนติเจนเป็นครั้งแรก ใช้เวลานานประมาณ 10-17 วัน เรียกการตอบสนองในระยะแรกนี้ว่า primary immune response ได้เซลล์ 2 ชนิดคือshort-lived effector cell (plasma cell(เซลล์บี) &effector T cell(จากเซลล์ที)) และ long-lived memory cells ถ้าร่างกายมีการเผชิญกับแอนติเจนเดิมอีก เป็นครั้งที่ 2 จะเกิดการตอบสนองเรียก secondary immune response ซึ่งจะใช้เวลาในการตอบสนองสั้นลง เพียง 2-7 วัน

2.1.7 คุณสมบัติระบบภูมิคุ้มกัน

กลไกการทำงานของระบบภูมิคุ้มกันก่อให้เกิดแรงบันดาลใจในการแก้ปัญหาในการเรียนรู้ เหตุผลที่ทำให้ระบบภูมิคุ้มกันได้รับความสนใจในเชิงการคำนวณ สรุปได้ดังนี้

- การจดจำ (Recognition) ระบบภูมิคุ้มกันมีความสามารถในการจดจำ ระบุ และการตอบสนองอย่างมีแบบแผนต่อสิ่งแปลกปลอมและการจำแนกได้ว่าสิ่งใดเป็นสิ่งแปลกปลอมและสิ่งใดเป็นของตนเอง

-การแยกลักษณะ (Feature Extraction) ระบบภูมิคุ้มกันสามารถใช้เซลล์สารกระตุ้นในการแยกลักษณะของสารกระตุ้นโดยกรองเกี่ยวกับโมเลกุลรบกวนที่เป็นสาเหตุให้เกิดโรคที่เรียกว่า แอนติเจน ก่อนที่จะถูกกระทำโดยเซลล์ภูมิคุ้มกันที่มีอยู่รวมถึงลิมโฟไซต์

-ความหลากหลาย (Disversity) มีกระบวนการหลักสองกระบวนการคือ กระบวนแรกการสร้างของโมเลกุลตัวรับเนื่องจากการรวมกันของกลุ่มยีนจากยีนไลบารี่ที่มีจำกัด ระบบภูมิคุ้มกันสามารถสร้างตัวรับได้ไม่มีที่สิ้นสุด ดังนั้นระบบภูมิคุ้มกันสามารถรองรับแอนติเจนที่เกิดขึ้นได้อย่างมาก กระบวนการที่สองเซลล์ภูมิคุ้มกันจะทำสำเนาตัวเองเพื่อตอบสนองต่อ

แอนติเจน ระหว่างการทำซ้ำ จะทำให้เกิดการเปลี่ยนแปลงภายในร่างกายในอัตราที่สูงทำให้มีการสร้างรูปแบบของโมเลกุลตัวรับ ซึ่งทำให้เกิดความหลากหลายของตัวรับภูมิคุ้มกัน

-การเรียนรู้ (Learning) ภายในร่างกายมีความสามารถในการเลือกการยึดเกาะของระบบภูมิคุ้มกันเพื่อปรับการตอบสนองกระบวนการนี้เรียกว่า แอฟฟินิตีมัทรูเรชัน ซึ่งแสดงให้เห็นว่าระบบภูมิคุ้มกันมีการจดจำเพิ่มขึ้น ทฤษฎีเครือข่ายภูมิคุ้มกันเป็นอีกตัวอย่างที่การเรียนรู้ในระบบมีภูมิคุ้มกันซึ่งมีการเปลี่ยนแปลงระหว่างเซลล์ที่จดจำและโมเลกุลตลอดเวลา และแสดงถึงแอนติเจนที่บุกรุกที่เป็นสาเหตุให้เกิดการก่อกวนในเครือข่ายภูมิคุ้มกัน ดังนั้นไดนามิกอิมมูนเน็ตเวิร์ก (dynamic immune network) แสดงให้เห็นสภาวะที่เครือข่ายไม่มีการกระตุ้นของแอนติเจนจะมีแบบแผนในการปรับตัวเองอีกครั้งเพื่อทำเครือข่ายให้เหมาะสม เพราะฉะนั้นแอนติเจนที่บุกรุกจะทำให้เครือข่ายภูมิคุ้มกันปรับตัวเองใหม่ตลอดเวลา

-มีความจำ (Memory) เมื่อภูมิคุ้มกันตอบสนองกับแอนติเจนกลุ่มโมเลกุลของเซลล์และจะมีชีวิตยืนยาวมากขึ้นเพื่อที่จะตอบสนองกับแอนติเจนชนิดเดิมหรือที่คล้ายชนิดเดิมเพื่อในอนาคตจะตอบสนองได้เร็วและรุนแรงขึ้น กระบวนการนี้เป็นกระบวนการตอบสนองต่อภูมิคุ้มกันที่จะรักษาเซลล์และโมเลกุลที่สมบูรณ์ในการจดจำแอนติเจน

-การป้องกันที่กระจาย (Distributed detection) จะสืบทอดไปในระบบภูมิคุ้มกัน ซึ่งไม่มีจุดหนึ่งจุดใดควบคุมได้ทั้งหมด แต่ละเซลล์ภูมิคุ้มกันทำหน้าที่เฉพาะและตอบสนองต่อแอนติเจนใหม่ที่บุกรุกร่างกายในส่วนอื่นๆ ได้

-เซลฟ์ออร์แกนไนสซิ่ง (Self-organization) เมื่อแอนติเจนเข้ามากระตุ้นระบบภูมิคุ้มกัน ถ้าเซลล์ภูมิคุ้มกันไม่มีเซลล์ภูมิคุ้มกันเดิมหรือโมเลกุลที่มีลักษณะคล้ายกับแอนติเจนนั้น จะเกิดขั้นตอนการคัดเลือกเพื่อเพิ่มจำนวน และแอฟฟินิตีมัทรูเรชันโดยจะเลือกและการขยายเซลล์ที่จะป้องกันและจะมีชีวิตเป็น เมมโมรี่เซลล์ (memory cells) ต่อไป

-เมตาไดนามิก (Metadynamics) ระบบภูมิคุ้มกันจะทำการสร้างเซลล์และโมเลกุลใหม่ที่คงที่และกำจัดเซลล์เก่าที่ไม่ถูกใช้ เมตาไดนามิก คือ กระบวนการสร้างที่ต่อเนื่อง การรับสมาชิกใหม่ การตายของเซลล์มีภูมิคุ้มกันและโมเลกุล

2.2 ระบบภูมิคุ้มกันเทียม (Artificial Immune system : AIS)

ระบบภูมิคุ้มกันเทียม เป็นการนำกลวิธีการทำงานของระบบภูมิคุ้มกันมาเป็นต้นแบบในการสร้างแนวทางใหม่ๆ ในการแก้ปัญหาทางคอมพิวเตอร์ เนื่องจากการที่ร่างกายของคนเรานั้นมีระบบป้องกันร่างกายจากสิ่งแปลกปลอมที่เข้ามาทำอันตรายที่มีความสามารถในการจดจำและจำแนกความแตกต่างของสิ่งผิดปกติในลักษณะต่างๆ ของร่างกายได้อย่างถูกต้องซึ่งนับว่าเป็นระบบที่อัจฉริยะ ซึ่งมีผู้ให้คำนิยามของระบบภูมิคุ้มกันเทียม [11] ไว้หลายคำนิยามดังนี้

Definition 2.1 Artificial immune systems are data manipulation, classification, representation and reasoning methodologies which follow a biological plausible paradigm: that of the human immune system

Definition 2.2 An artificial immune system is a computational system based upon metaphors of the natural immune system

Definition 2.3 The artificial immune systems are intelligent methodologies inspired by the naturalimmune system,toward real-world problems solving

จากความแตกต่างระหว่างทฤษฎีทางคณิตศาสตร์ของระบบภูมิคุ้มกันกับระบบภูมิคุ้มกันเทียมเราสามารถสรุปเป็นความหมายทั่วไปได้ว่า

Definition 2.4 Artificial immune system (AIS) are adative systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving.

คำนิยามที่ 2.4 อธิบายว่าระบบภูมิคุ้มกันเทียมนั้นจะต้องมีคุณสมบัติพื้นฐานของส่วนประกอบของระบบภูมิคุ้มกันตามธรรมชาติ และออกแบบจากการทำงานตามทฤษฎีและการทดลองทางภูมิคุ้มกัน โดยมุ่งไปในการแก้ปัญหา

ดังนั้นคุณสมบัติเบื้องต้นของระบบภูมิคุ้มกันเทียมจะต้องมี การกำหนดของรูปแบบอินพุตเป็นแอนติเจนและรูปแบบของอินพุตอีกอันหนึ่งเป็นแอนติบอดี และจะต้องมีลำดับของกระบวนการในการทำงานของระบบภูมิคุ้มกันด้วย เช่น กระบวนการในการจดจำ ,กระบวนการในการคัดเลือกเพื่อเพิ่มจำนวนแอนติบอดี ,กระบวนการในการกำจัดสิ่งแปลกปลอม ,กระบวนการของเครือข่ายภูมิคุ้มกัน เป็นต้น จึงจะถือได้ว่าเป็นระบบภูมิคุ้มกันเทียม

2.2.1 ขอบข่ายงานทางวิศวกรรมของระบบภูมิคุ้มกันเทียม (An Engineering Framework for AIS)

การสร้างขอบข่ายงานของระบบภูมิคุ้มกันเทียมแต่เดิมนั้นเป็นสิ่งที่ยาก เนื่องจากเหตุผลหลายประการดังนี้

1. จำนวนของผู้วิจัยยังเป็นกลุ่มเล็กๆ แต่ในรอบหลายปีมานี้เพิ่มมากขึ้น

2. นักวิจัยพบความยากในการกำหนดการทำงานระหว่างระบบภูมิคุ้มกันเทียมและระบบภูมิคุ้มกันทางชีววิทยา

3. การประยุกต์ใช้ระบบภูมิคุ้มกันเทียมยังค่อนข้างกว้าง

4. หนังสือที่เกี่ยวกับการกำหนดขอบข่ายงานในการออกแบบระบบภูมิคุ้มกันเทียมเพิ่งจะมีเมื่อเร็วๆ นี้

ข้อจำกัดอันหนึ่งของการกำหนดขอบข่ายงานของระบบภูมิคุ้มกันเทียมในปัจจุบัน การประยุกต์หลักการของระบบภูมิคุ้มกันเทียมเป็นการกำหนดระบบการคำนวณตามหลักการและการทำงานของระบบภูมิคุ้มกันตามธรรมชาติ หลักการและรูปแบบถูกนำมาประยุกต์ใช้ในการแก้ปัญหา

การกำหนดจะครอบคลุมบางรูปแบบที่ถูกอ้างถึงเกี่ยวกับการแบ่งระหว่างระบบภูมิคุ้มกันเทียมและทฤษฎีที่เป็นประโยชน์ของระบบภูมิคุ้มกัน ขณะที่การทำงานบนระบบภูมิคุ้มกันเป็นการกำหนดรูปแบบและการทำความเข้าใจหน้าที่และการทดลอง ส่วนระบบภูมิคุ้มกันเทียมนั้นเป็นการแก้ปัญหาในทางการคำนวณและวิศวกรรม มีตัวอย่างในการคำนวณที่คล้ายกันหลายตัวอย่าง การกำหนดขอบข่ายงานในการออกแบบระบบภูมิคุ้มกันเทียมบางส่วนยังเป็นปัญหาอยู่ว่าเราจะกำหนดขอบข่ายงานอย่างไร มีตัวอย่างที่มีการคำนวณลักษณะใกล้เคียงกันในทางชีววิทยา เช่น นิวรอลเน็ตเวอร์ก (ANN) ขอบข่ายงาน ก็คือกลุ่มของอาร์ทิฟิเชียลนิวรอล (artificial neurons) ส่วนอัลกอริทึมเชิงวิวัฒนาการ(Evolutionary algorithm :EAs) ก็คือ กลุ่มของ อาร์ทิฟิเชียลโคโมโซม (artificial chromosomes) เป็นต้น

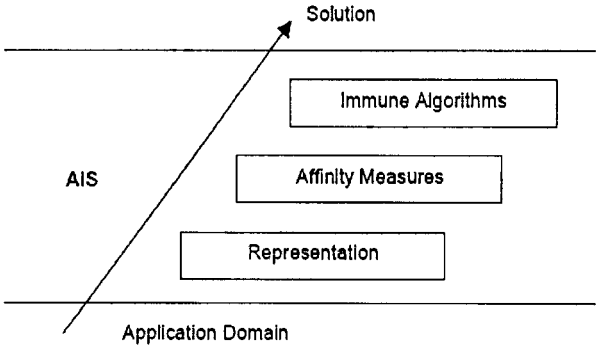
ดังนั้นขอบข่ายงานการออกแบบขั้นตอนการทำงาน การหาค่าด้วยการคำนวณ จะต้องทำตามหลักการเบื้องต้นดังนี้

-การแทน (Representation) สำหรับส่วนประกอบของระบบภูมิคุ้มกันเทียม

-การวัดค่าแอฟฟินิตี (Affinity measure) เป็นขั้นตอนของกลไกที่จะหาค่าของปฏิกิริยาที่มีกับสภาวะแวดล้อม โดยทั่วไปจะจำลองกลุ่มของอินพุตที่กระตุ้นเป็นฟังก์ชันความเหมาะสม (Fitness Function) หรือ สิ่งอื่นๆ

-อัลกอริทึมของระบบภูมิคุ้มกัน (Immune Algorithms) เป็นขั้นตอนการปรับการเคลื่อนไหวของระบบว่ามีพฤติกรรมอย่างไรเมื่อเวลาเปลี่ยนไป

หลักการนั้นจะเป็นแนวทางในการออกแบบระบบภูมิคุ้มกันเทียมได้เป็นอย่างดี การแทนอวัยวะของภูมิคุ้มกันโดยการสร้างรูปแบบที่ไม่มีตัวตน เซลล์และโมเลกุล ,กลุ่มของฟังก์ชัน,การกำหนด affinity function การกำหนดปริมาณของปฏิกิริยาที่กระทำต่อกันขององค์ประกอบและกลุ่มของขั้นตอนที่กำหนดทั่วไปในการควบคุมการทำงานของระบบภูมิคุ้มกันเทียม ในรูปที่ 2.6 แสดงขอบข่ายงานทางวิศวกรรมของระบบภูมิคุ้มกันเทียม โดยจะเป็นลำดับชั้นและขั้นตอนสำหรับ Application Domain ก็คือ ขอบเขตของแอพพลิเคชั่นที่จะนำไปประยุกต์ใช้ในการแก้ปัญหาเพื่อหาทางออกของปัญหา (Solution) และกระบวนการต่างๆ ที่จะอธิบายต่อไปนี้

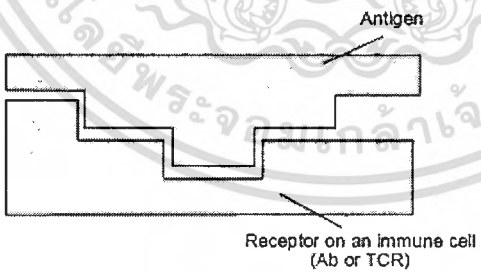


รูปที่ 2.6 แสดงขอบข่ายงานทางวิศวกรรมของระบบภูมิคุ้มกันเทียม

2.2.2 การแทน (Representation)ส่วนประกอบของระบบภูมิคุ้มกันเทียม

ในระบบภูมิคุ้มกันเห็นได้ว่า เซลล์บี และ เซลล์ที มีความสำคัญ มันมีรูปร่างพื้นผิวของโมเลกุลที่จะทำปฏิกิริยาเหมือนกับแอนติเจนที่มันสามารถจดจำและทำลาย ในระบบภูมิคุ้มกันเทียมนั้นจะต้องมีการแสดงหน้าที่ในการกระตุ้นและมีองค์ประกอบของรูปแบบเซลล์ภูมิคุ้มกันและเซลล์โมเลกุลที่สร้าง

เชป-สเปซ (Shape-Space) คือ สิ่งที่ใช้ในการคำนวณเพื่ออธิบายระดับความผูกพันระหว่างตำแหน่งย่อยๆ ของโมเลกุลของผิวเซลล์กับตำแหน่งย่อยๆ ของแอนติเจนในระบบภูมิคุ้มกันเทียม หลักการแสดงถึงกลไกในการจดจำโดยโมเลกุลของผิวเซลล์ที่ทำหน้าที่ในการจับกับแอนติเจนและปริมาณของปฏิกิริยาระหว่างโมเลกุลของผิวเซลล์ที่ทำหน้าที่ในการจับกับแอนติเจน ระดับของความแรงในการจับระหว่างตำแหน่งย่อยๆ ของเชป-สเปซของแอนติเจนจะคำนวณด้วยส่วนที่ประกอบที่ตรงกันแสดงดังรูปที่ 2.7



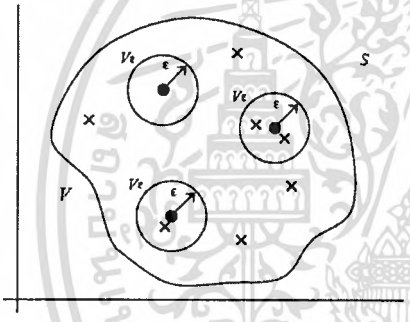
รูปที่ 2.7 แสดงเชป-สเปซของระบบภูมิคุ้มกันเทียม

คุณลักษณะที่อธิบายคุณสมบัติของโมเลกุลของเซลล์ในการจดจำที่เห็นโดยทั่วไปเรียกว่าเจนเนอร์รัลไลซ์เชป (generalized shape) รูปร่างทั่วไปของแอนติชอดีอธิบายโดยใช้การกำหนดค่า L ดังนั้นจุดหนึ่งจุดของ L-dimemsional shape-space, S^L เป็นสิ่งที่ระบุลักษณะทั่วไปของความสามารถในการยึดเกาะแอนติบอดีกับแอนติเจน

จำนวนของเซลล์รีเซปเตอร์ (cell receptor) ที่มีอยู่ในปริมาตร V มีจำนวน N จุด ปฏิกิริยาในการยึดเกาะระหว่างแอนติเจนและแอนติบอดีที่มีลักษณะและรูปร่างเหมือนกันในปริมาตร V

เปรียบเหมือนลูกกุญแจกับแม่กุญแจ การยึดเกาะที่สมบูรณ์ก็เหมือนกับเปิดล็อกกุญแจ แต่อาจพบว่ากุญแจที่ใส่เข้าไปในล็อกได้แต่ไม่สามารถเปิดได้ เนื่องจากมีความสามารถในการยึดเกาะต่ำ และถึงแม้ว่าแอนติเจนและแอนติบอดีจะไม่ยึดเกาะกันโดยสมบูรณ์ ก็อาจจะมีการยึดเกาะกันอยู่แต่ความสามารถในการยึดเกาะต่ำ

แอนติบอดีที่มีปฏิกิริยากับแอนติเจนทั้งหมดที่อยู่ในพื้นที่เล็กๆโดยรอบพื้นที่ที่กำหนดโดยค่า ε, เรียกว่า แอฟฟินิตีเทรชโฮลด์ (affinity threshold) ปริมาตร V_ε ที่กำหนดโดยแอฟฟินิตีเทรชโฮลด์เรียกว่า รีคอกนิชันรีเจียน (recognition region) ซึ่งแอนติเจนเหล่านั้นจะมีรูปร่างแตกต่างกันเพียงเล็กน้อย กล่าวคือเปลี่ยนแปลงเล็กน้อยจากแอนติเจนอย่างเดียวกัน ซึ่งแต่ละแอนติบอดีสามารถจดจำแอนติเจนที่มีจำนวนแน่นอนในปริมาตร V ความสัมพันธ์นี้อยู่ในพื้นที่ของเชป-สเปซ ที่เรียกว่า ครอส-รีแอคติวิตี (cross-reactivity) คือ การที่แอนติบอดีที่สามารถทำปฏิกิริยากับแอนติเจนต่างๆ ได้เนื่องจากมีบางส่วนเหมือนหรือคล้ายคลึงกันกับแอนติเจน ภายใต้เงื่อนไขของแอฟฟินิตีเทรชโฮลด์ ε



รูปที่ 2.8 แสดงเชป-สเปซ S, แอนติบอดี, แอนติเจน และ แอฟฟินิตีเทรชโฮลด์

รูปที่ 2.8 อธิบายถึง ในเชป-สเปซ S มีปริมาตร V มีแอนติบอดี(●)และแอนติเจน(×) แอนติบอดีจะสามารถจดจำแอนติเจนใดๆ ที่อยู่รอบแอนติบอดีภายในปริมาตร V_ε ที่กำหนดโดยค่า ε

2.2.3 การวัดแอฟฟินิตี (Affinity Measures)

รูปแบบการแทนแอนติบอดีและแอนติเจนจะเป็นส่วนกำหนดวิธีการในการคำนวณ affinity ทางคณิตศาสตร์รูปร่างทั่วไปของโมเลกุลแทนด้วย (m) ทั้ง แอนติบอดี แทนด้วย Ab และ แอนติเจน แทนด้วยAg สามารถแสดงเป็นแอตตริบิวต์สตริง $m=<m_1,m_2,\ldots,m_L>$ $m \in S^L \subseteq \Re^L$,หรือแสดงในรูปแบบอื่น เช่น นิวรอลเน็ตเวิร์ก (neural network) หรือ Petri net ซึ่งในงานวิจัยนี้จะอธิบายเฉพาะในแอตตริบิวต์สตริง

ชุดของสตริง ประกอบด้วยชนิดของแอตตริบิวต์ (attribute) เช่น real value, integers, bits และ symbols ชุดของ แอตตริบิวต์ เหล่านี้ขึ้นอยู่กับหลักการของปัญหาของระบบภูมิคุ้มกันเทียม

และความสำคัญในการกำหนดรูปแบบ การหาค่าปริมาณของการกระทำ ชนิดของ attribute ที่เราจะกำหนดในแบบของ เชป-สเปซ มีดังนี้

-real-valued เชป-สเปซ แอตตริบิวต์สตริง ก็คือ real-valued vectors

-Integer เชป-สเปซ แอตตริบิวต์สตริง ประกอบด้วย integer value

-Hamming เชป-สเปซ ประกอบด้วยสตริงของตัวอักษรที่มีความยาวจำกัด

-Symbolic เชป-สเปซ ประกอบด้วยชนิดต่างๆ ของแอตตริบิวต์สตริง เช่น ชื่อ สี เป็นต้น

ดังนั้นถ้าพิกัดของแอนติบอดี เป็น $Ab=<Ab_1,Ab_2,\ldots Ab_L>$ และของแอนติเจนเป็น $Ag=<Ag_1,Ag_2,\ldots Ag_L>$ ภายใต้มุมมองของการจดจำปฏิกิริยาของแอนติบอดีหรือระหว่างแอนติบอดีและแอนติเจนสามารถวัดได้โดยการวัดระยะทางหรือเรียกการวัดความสามารถในการยึดเกาะว่า แอฟฟินิตีเมเซอร์ (affinity measure) ซึ่งเป็นความสัมพันธ์ระหว่างแอตตริบิวต์สตริง การวัดความสามารถในการยึดเกาะ กระทำระหว่างการจับคู่ของแอตตริบิวต์สตริงจะมีค่าเป็นบวก

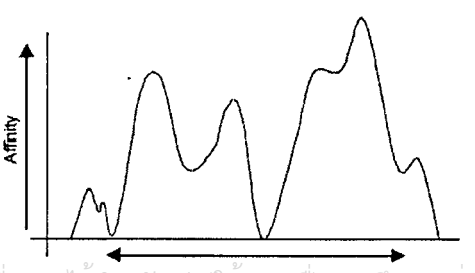
การวัดระยะทางที่ใช้ในการคำนวณแอฟฟินิตีระหว่างส่วนประกอบของระบบภูมิคุ้มกันเทียมที่เป็นแบบ real valued เชป-สเปซสามารถใช้ได้หลายวิธีทั้งแบบยูคลิด (Euclidean distance) และ แบบแมนฮัตตัน (Manhattan distance)ในกรณีของการใช้ยูคลิด แอฟฟินิตี D ระหว่างแอนติเจนและแอนติบอดี คำนวณได้ดังสมการที่ (2.2) ซึ่งเราเรียกว่ายูคลิดเชป-สเปซ (Euclidean shape-spaces) แต่ถ้าเราใช้ แบบแมนฮัตตันแทนในการคำนวณค่าแอฟฟินิตี เราจะเรียกว่า แมนฮัตตันเชป-สเปซ (Manhattan shape-spaces) ซึ่งแสดงได้ดังสมการที่ (2.3)

$$D = \sqrt{\sum_{i=1}^{L} (Ab_i - Ag_i)^2} \qquad (2.2)$$

$$D = \sum_{i=1}^{L} | Ab_i - Ag_i | \qquad (2.3)$$

เมื่อ D เป็นค่าแอฟฟินิตี $Ab=<Ab_1,Ab_2,\ldots Ab_L>$ เป็นแอนติบอดี และ $Ag=<Ag_1,Ag_2,\ldots Ag_L>$ เป็นแอนติเจน

ให้การแทนของแอตตริบิวต์สตริงของแอนติเจนและกลุ่มของแอนติบอดี ซึ่งแต่ละแอตตริบิวต์สตริงของแอนติบอดีจะสัมพันธ์กับความสามารถในการยึดเกาะบางแอนติเจนที่รูปร่างแตกต่างกันเราเรียกว่า แอฟฟินิตีแลนด์สเคป (affinity landscape) แสดงได้ดังรูปที่ 2.9



รูปที่ 2.9 แสดงแอฟฟินิตีแลนด์สเคป

อีกอันหนึ่งของ real-valued เชป-สเปซ คือแฮมมิ่งเชป-สเปซ (Hamming shape-spaces) ในกรณีที่แอนติเจนและแอนติบอดีจะถูกแทนด้วยลำดับของสัญลักษณ์ที่มีความยาวจำกัด สมการที่ 2.4 แสดงการวัดระยะทางแบบแฮมมิ่ง (Hamming distance) ใช้เพื่อหาค่าแอฟฟินิตี ระหว่างแอตตริบิวต์สตริง ที่มีความยาว L ในรูปแบบของแฮมมิ่งเชป-สเปซ ถ้าเป็นไบนารี สตริง $k \in \{0,1\}$ ในเทอมของบิตที่แทนโมเลกุลนั้นจะเรียกว่า "binary Hamming shape-space" หรือ "binary shape-space" ถ้าเป็น ternary สตริง ตัวอย่าง เช่น k=3 การแทนโมเลกุลจะเรียกว่า "ternary Hamming shape-space" หรือ "ternary shape-space" ขึ้นอยู่กับปัญหาในการศึกษา และใช้กับ Integer shape-space ก็คือ แอตตริบิวต์จะสัมพันธ์กับจำนวนเต็มดูได้จากกรณีของแฮมมิ่งเชป-สเปซ ซึ่งใช้อย่างกว้างขว้าง เช่น ปัญหาของ traveling salesman หรือ scheduling application

$$D = \sum_{i=1}^{L} \delta_i, \text{ where } \begin{cases} 1 & \text{if } A_b \neq A_g \\ 0 & \text{otherwise} \end{cases} \qquad (2.4)$$

ตัวอย่างการกำหนด เชป-สเปซแบบไบนารีสตริง แทนโมเลกุลด้วยบิตสตริง ในกรณีนี้ความสามารถในการยึดเกาะบิตสตริงระหว่างบิตสตริงของแอนติบอดีและบิตสตริงแอนติเจนสามารถใช้วิธีต่างๆ ดังนี้

1) ระยะทาง แบบแฮมมิ่ง (สมการ 2.4)

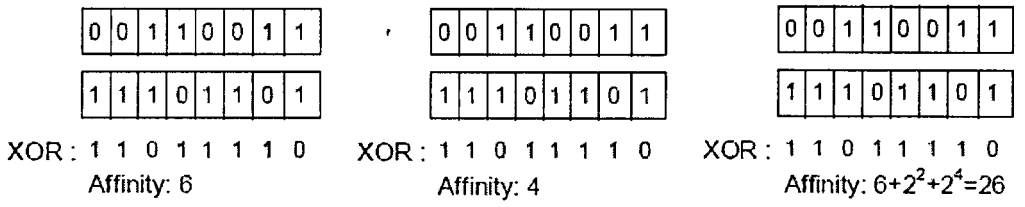
2) r-contiguous bitrule

3) multiple r-contiguous bit rule

ระยะทาง แฮมมิ่ง สามารถคำนวณโดยการประยุกต์ใช้ (XOR) กับไบนารีสตริงการวัด r-contiguous จำนวนของสัญลักษณ์ที่ตรงกันระหว่างสตริงทั้งสอง ในกฎ Multiple r-contiguous บิตสตริงจะต้องตรงกันและประกอบกันได้อย่างสมบูรณ์ เราจะสนใจลักษณะคล้ายกันของโมเลกุลที่กำหนดได้ตามสมการ (2.5)

$$D = D_H + \sum_i 2^{l_i} \qquad (2.5)$$

เมื่อ D_H เป็นระยะทางแฮมมิ่ง ทั้งหมดโดยสมการ(2.4) , l_i และความยาวส่วนที่ประกอบกันสมบูรณ์ อย่างน้อย 2 บิตติดต่อกัน ในรูปที่ 2.10 แสดงการวัดความสามารถในการยึดเกาะการทั้งสามแบบ



ก. จำนวน bit ที่ XOR ข.r-contiguous bit rule ค.multiple r-contiguous bit rule

รูปที่ 2.10 แสดงแอฟฟินิตีเมเซอร์แบบไบนารีแฮมมิ่งเชป-สเปซ

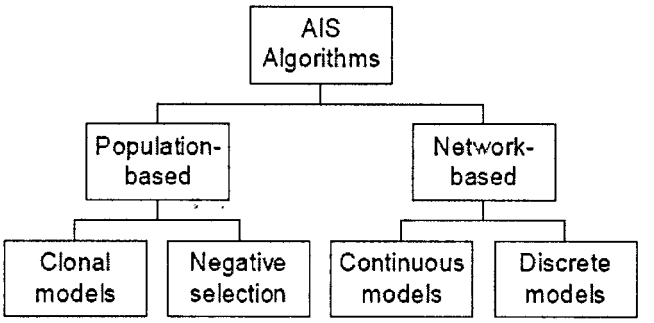
ลำดับสุดท้ายของเชป-สเปซคือ ซิมบอลิคเชป-สเปซ (Symbolic shape-space) ซึ่งแทนส่วนประกอบของระบบภูมิคุ้มกันเทียมได้ด้วยสัญลักษณ์ของแอตตริบิวต์ แสดงดังรูปที่ 2.11 ซึ่งจะมีทั้งสตริงที่เป็น Symbolic ,integer และ real-valued และการมีแอตตริบิวต์ที่ตรงกันระหว่างแอนติบอดีและแอนติเจนจะถูกแทนด้วย 1 หากไม่ตรงกันจะแทนด้วย 0

	Description	Date	Flight	Country	From	To	Price(£)
Antibody(Ab1):	Business	1996	212	Brazil	Campinas	Greece	546.78
Antibody(Ab2):	Holiday	2000	312	U.K.	London	Paris	102.35
Antigen(Ag):	Holiday	2000	212	U.K.	London	Greece	543.78
Match Ag-Ab1:	0	0	1	0	0	1	1
Match Ag-Ab2:	1	1	0	1	1	0	0

รูปที่ 2.11 แสดงตัวอย่างการแทนส่วนประกอบของระบบภูมิคุ้มกันเทียมด้วยสัญลักษณ์

2.2.4 อัลกอริทึมของระบบภูมิคุ้มกัน (Immune Algorithms)

ส่วนแรกของการออกแบบระบบภูมิคุ้มกันเทียมได้อธิบายถึงรูปแบบของแอตตริบิวต์สตริงคือ ตัวแทนเซลล์รีเซปเตอร์และแอนติเจน โดยการประเมินผลความสามารถในการยึดเกาะและการสร้างรูปแบบสตริงประชากรเริ่มต้นเกิดในไขกระดูก ชนิดของแอตตริบิวต์บางส่วนจะกำหนดหน้าที่ในการประเมินความผูกพัน (ปริมาณการรับรู้) ของเซลล์รีเซปเตอร์กับสิ่งแวดล้อม (แอนติเจน) และ เซลล์รีเซปเตอร์อื่น ๆ ส่วนที่สองเป็นการประยุกต์ขอบข่ายให้สอดคล้องกับแอพพลิเคชัน (applications) บางขั้นตอนจะปรับการควบคุมระบบภูมิคุ้มกันเทียมให้ทำงานตามรูปแบบของกระบวนการที่ได้กล่าวมา เพื่อให้ชัดเจนขึ้นเราสรุปอัลกอริทึมระบบภูมิคุ้มกันเทียมดังรูปที่ 2.12 เราจำแนก Clonal models และขั้นตอนวิธีการเลือกลบ (Negative selection) เป็น Population-based และรูปแบบเครือข่าย (Network-based) แบ่งออกเป็นเครือข่ายภูมิคุ้มกันแบบต่อเนื่อง (A Continuous Immune Network Model) และเครือข่ายภูมิคุ้มกันแบบไม่ต่อเนื่อง (Discrete Immune Network Models) และส่วนต่อไปนี้จะกล่าวถึงหลักการโดยคร่าว ๆ ของอัลกอริทึมของระบบภูมิคุ้มกันเทียมและกระบวนการสำคัญ



รูปที่ 2.12 แสดงการแบ่งหมวดหมู่ของอัลกอริทึมของระบบภูมิคุ้มกันเทียม

2.2.4.1 Clonal Models

de Castro และ Von Zuben [12] ได้ใช้หลักการโคลนอลซีเล็คชันและกระบวนการแอฟฟินิตีมัทรูเรชัน ของการปรับการตอบสนองของภูมิคุ้มกันเพื่อหาวิธีที่เหมาะสมในการดำเนินการ เช่น การเรียนรู้โดยเครื่อง, การรู้จำ และการหาคำตอบที่ดีที่สุด อัลกอริทึมนี้เคยใช้ในการประเมินปัญหาการรู้จำตัวอักษรไบนารี, multimodal optimization, combinatorial optimization และปัญหา traveling salesman (TSP) โดยเรียกว่า Clonal Selection Algorithm และกำหนดชื่ออัลกอริทึมนี้ว่า CLONALG คือการเลือกและการเพิ่มจำนวนในการกระตุ้นเซลล์ตามส่วนของความสามารถในการยึดเกาะส่วนย่อย ๆ ของเซลล์ การตายของเซลล์ที่ไม่กระตุ้นแอฟฟินิตีมัทรูเรชัน การสร้างและการบำรุงรักษาต่าง ๆ ของเซลล์ ซึ่งพัฒนาโดยได้รับแรงบันดาลใจจากระบบภูมิคุ้มกัน อัลกอริทึม CLONALG [12] มีการทำงานดังนี้

1. Generate a set of 1 candidate solutions (antibody repertoire) in a shape-space to be defined by the problem under study;
2. Select Q1 highest affinity cells in relation to the antigen set to be recognized or to the function being optimized;
3. Clone (generate identical copies of) these Q selected cells. The number of copies is proportional to their affinities: the higher the affinity, the larger the clone size (number of offspring);
4. Mutate with high rates (hypermutation) these Q selected cells proportionally to their affinities: the higher the affinity, the smaller the mutation rate (see further discussion);
5. Re-select Q2 highest affinity mutated clones to compose the new repertoire;
6. Replace some low affinity cells by new ones;
7. Repeat Steps 2 to 6 until a given stopping criterion is met.

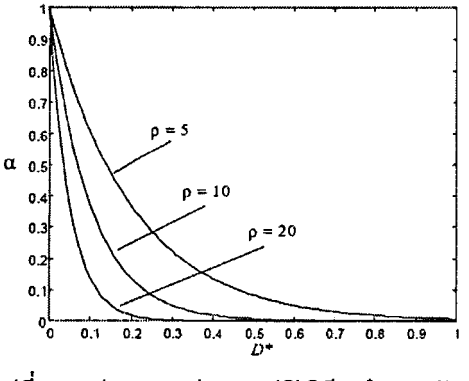
CLONALG จะมีคุณสมบัติเหมือน อัลกอริทึมแบบวิวัฒนาการ(Evolutionary algorithm) ที่ค้นหาประชากรตามกลไกความหลากหลายทางพันธุกรรม และการเลือก ซึ่งอัลกอริทึม CLONALG เป็นชนิดหนึ่งของอัลกอริทึมวิวัฒนาการ

สำหรับในกระบวนการของโคลนอลซีเล็คชันมีกลไกที่เรียกว่า ไฮเปอร์มิวเตชัน กลไกนี้ช่วยให้ระบบภูมิคุ้มกันเพิ่มความสามารถในการยึดเกาะ (recognition capabity) ของแอนติบอดีกับความสัมพันธ์ในการเลือกแอนติเจนกระบวนการนี้เรียกว่า แอฟฟินิตีมัทรูเรชัน (Affinity Maturation)

เซป-สเปซจะเป็นตัวแทนของเซลล์รีเซปเตอร์ใดๆ และแอนติเจนแทนด้วยแอตตริบิวต์สตริงขั้นตอนที่จะกำหนดความเปลี่ยนแปลงและการเข้ารหัสของส่วนประกอบของระบบมีหลายขั้นตอน ขั้นตอนเหล่านี้สามารถทำได้เหมือนกับการมิวเตชันในอัลกอริทึมแบบวิวัฒนาการ ลักษณะที่สำคัญอันหนึ่งของโซมาติกไฮเปอร์มิวเตชันก็คือ แอตตริบิวต์สตริงจะมีอิสระและเป็นสัดส่วนกับอัตราการมิวเตชันเพื่อความสามารถในการยึดเกาะกับแอนติเจน ดังนั้นเซลล์ใดมีค่าความสามารถในการยึดเกาะสูงกว่าจะมีอัตรามิวเตชันต่ำ ส่วนเซลล์ใดมีความสามารถในการยึดเกาะต่ำก็จะมีอัตราการมิวเตชันสูงกว่า เราประเมินความสัมพันธ์ในการยึดเกาะแต่ละช่วงเวลาของแต่ละเซลล์ที่แข่งขันโดยการปรับ (scale) ความสามารถในการยึดเกาะซึ่งเป็นส่วนกลับของ exponential function กำหนดความสัมพันธ์ระหว่างอัตราการเปลี่ยนแปลง α กับการปรับความสามารถในการยึดเกาะ D แสดงดังสมการที่ (2.6)

$$\alpha(D^*) = \exp(-\rho D^*) \tag{2.6}$$

เมื่อ ρ เป็นพารามิเตอร์ที่ควบคุมการปรับของส่วนกลับเลขยกกำลังกับ D* เป็นการปรับค่าความสามารถในการยึดเกาะให้ตรงกับความต้องการ โดย D*=D/Dmax ความสัมพันธ์ระหว่างแอฟฟินิตีและอัตราการมิวเตชัน α แสดงดังรูปที่ 2.13 ถ้าค่า D* น้อยค่าอัตราการมิวเตชัน α จะมาก



รูปที่ 2.13 แสดงการเปลี่ยนแปลงระหว่างแอฟฟินิตี D^* และอัตราการมิวเตชัน α สำหรับค่า ρ ที่เปลี่ยนแปลง

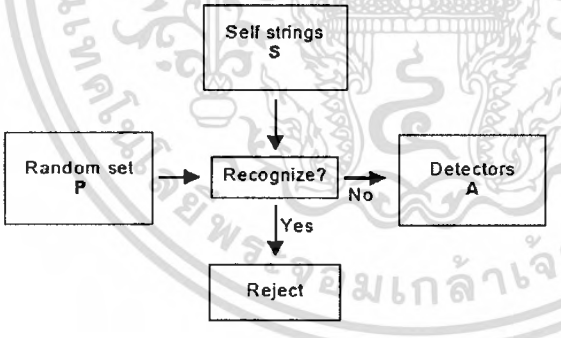
2.2.4.2 การคัดเลือกทางลบ (Negative Selection)

การคัดเลือกทางลบ เซลล์ทีจะมีการแบ่งตัวอย่างรวดเร็วกับแอนติเจนที่จำเพาะต่อสารประกอบของตัวเอง และจะคัดเลือกเซลล์ที ที่จำเพาะต่อสารประกอบที่เกิดจากสิ่งแปลกปลอม และจะมีส่วนหนึ่งถูกคัดเลือกให้ตายโดยกระบวนการการคัดเลือกทางลบ

แนวคิดนี้ Forrest et al [13] พัฒนาวิธีการตรวจสอบความผิดปกติตามการเลือกเชิงลบของเซลล์ทีภายในไทมัส ซึ่งเรียกชื่อขั้นตอนนี้ว่า negative selection algorithm และการประยุกต์ใช้ในการรักษาความปลอดภัย สิ่งที่น่าสนใจของอัลกอริทึมนี้คือมีการกระทำเหมือนกับการจดจำรูปแบบแต่เป็นรูปแบบที่ระบบไม่รู้จัก ขั้นตอนของการคัดเลือกทางลบมีขั้นตอนการทำงานดังนี้

กำหนดให้ S เป็นของสตริงที่ต้องการป้องกัน A เป็นเซตของรูปแบบในการจดจำเรียกว่า detectors ซึ่งไม่มีสตริงใดๆตรงกับชุดของสตริง S ลำดับขั้นตอนในการจดจำชุดสตริง A อธิบายได้ดังรูปที่ 2.14

1. สร้างสตริงโดยการสุ่มและกำหนดเป็นชุดสตริง P
2. ทำการหาค่า affinity ของชุดสตริงใน P กับชุดของสตริง S
3. ถ้าชุดสตริงใน P กับชุดของสตริง S มีค่ามากกว่าหรือเท่ากับค่า affinity threshold ε ชุดของสตริง P ถูกจดจำก็จะถูกจำกัดออกและถ้าสตริง P ไม่ถูกจดจำก็จะนำไปกำหนดในชุด A



รูปที่ 2.14 แสดงอัลกอริทึมของการเลือกทางลบ

2.2.5 แบบจำลองเครือข่ายภูมิคุ้มกัน (Immune Network Model)

ทฤษฎีเครือข่ายภูมิคุ้มกันอธิบายว่าระบบมีภูมิคุ้มกันมีพฤติกรรมการปรับเครือข่ายให้เกิดความสมดุลตลอดเวลาซึ่งการปรับนี้ไม่ได้เกิดจากแอนติเจน รวมถึงระบบภูมิคุ้มกันกับพฤติกรรมที่แท้จริงของเครือข่ายการติดต่อสื่อสารระหว่างเซลล์ตัวรับ ซึ่งจะแตกต่างจากโคลนอลซีเล็คชันและการคัดเลือกทางลบซึ่งเป็นความสามารถของเซลล์บีในการจดจำซึ่งกันและกัน

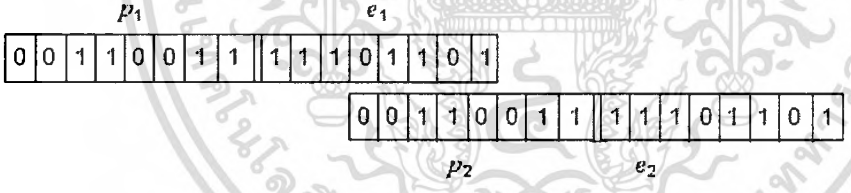
ทฤษฎีของภูมิคุ้มกันสนใจในการสร้างแบบจำลองของระบบเครือข่ายภูมิคุ้มกัน เพื่อหาวิธีการใหม่ในการอธิบายวิธีการทำงานของระบบภูมิคุ้มกันเทียมนักวิจัยทางด้านปัญญาเชิงคำนวณ (Computational Intelligence) ได้พยายามพัฒนาการทำงานนี้ ความสนใจในรูปแบบของ

ระบบภูมิคุ้มกัน ได้ถูกนำมาแก้ปัญหาในทางคอมพิวเตอร์ งานวิศวกรรมและในสาขาอื่น ๆ ช่วงแรกของเครือข่ายภูมิคุ้มกันจะอยู่ในแบบของสมการดิฟเฟอเรนเชียลและตัวแปรที่แทนขนาดของประชากรของแอนติบอดีและเซลล์บี ซึ่งเราเรียกการทำงานแบบนี้ว่าเครือข่ายภูมิคุ้มกันแบบต่อเนื่อง (continuous immune network models) ใช้อย่างแพร่หลายในกลุ่มของระบบภูมิคุ้มกันเทียม เช่น งานหุ่นยนต์ (robotics) การหาคำตอบดีที่สุด (optimization) และการควบคุม (control) ส่วนเครือข่ายภูมิคุ้มกันที่ใช้การเรียนรู้ของเครือข่ายเป็นหลัก และการใช้ในการวิเคราะห์ข้อมูล ในแบบหลังนี้เราจัดเป็นเครือข่ายภูมิคุ้มกันแบบไม่ต่อเนื่อง (discrete immune network models) ซึ่งไม่อยู่บนพื้นฐานสมการดิฟเฟอเรนเชียลแต่ขึ้นอยู่กับลำดับขั้นตอนของการปรับค่าหรือสมการที่แตกต่างไป

ต่อไปนี้จะพิจารณารูปแบบของเครือข่ายภูมิคุ้มกันแบบต่อเนื่องและเครือข่ายภูมิคุ้มกันแบบต่อไม่เนื่องที่ได้ใช้อย่างกว้างขวางในระบบภูมิคุ้มกันเทียม

2.2.5.1 เครือข่ายภูมิคุ้มกันแบบต่อเนื่อง (A Continuous Immune Network Models)

การแทนเซลล์ภูมิคุ้มกันและโมเลกุลเป็นบิตสตริงในแบบของ Hamming shape-space, แสดงดังในรูปที่ 2.15 โมเลกุลของแอนติบอดีแสดงเป็นสองส่วนโดยส่วนหนึ่งชื่อ epitope (e) และอีกส่วนชื่อ paratope(p) epitope เป็นส่วนของโมเลกุลแอนติบอดีที่สามารถจดจำ paratopes ของแอนติบอดีอื่น ๆ



รูปที่ 2.15 แสดงบิตสตริงในการแทน epitope และ paratope ของโมเลกุลของแอนติบอดีทั้งสอง

สตริงถูกจับคู่โดยส่วนประกอบที่ตรงกันในแนวทางต่างๆที่เป็นไปได้ รูปแบบของทั้งสองโมเลกุลนั้นจะแสดงปฏิกิริยาโต้ตอบได้มากกว่าหนึ่งทาง สมการที่ (2.7) แสดงเมทริกซ์การจับคู่ m_{ij} ที่สัมพันธ์กันระหว่างระดับของการจับคู่ของแต่ละส่วนในระบบภูมิคุ้มกันเทียม

$$m_{ij} = \sum_{k} G\left(\sum e_i(n+k) \wedge p_j(n) - \varepsilon + 1 \right) \qquad (2.7)$$

เมื่อ $e_i(n)$ เป็น n-th บิตของ i-th ของ epitope , pj(n) เป็น n-th บิตของ paratope , $\wedge$ ความสัมพันธ์ของระยะทางแฮมมิ่ง ระหว่าง ei(.) และ pj(.) และ ε เป็น affinity threshold พารามิเตอร์ k สัมพันธ์กับการแนวที่ตรงกันระหว่าง paratope และ epitope ถ้าการจับคู่ตรงกัน

มากกว่าหนึ่งตำแหน่งความแข็งแรงจะถูกรวมกับความยาวของสตริงที่แตกต่าง ฟังก์ชัน G จะวัดความแข็งแรงของปฏิกิริยาที่เป็นไปได้ระหว่าง epitope และ paratope แสดงดังสมการที่ (2.8)

$$G(x) = \begin{cases} x & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

ปริมาณการเปลี่ยนแปลงของเครือข่าย คือ ถ้า N เป็นแอนติบอดีโดย $\{c_1,\ldots,c_N\}$ และ M แอนติเจน โดย $\{y_1,\ldots,y_M\}$ อัตราการเปลี่ยนแปลงของแอนติบอดีเขียนได้ดังสมการที่ (2.9)

$$\frac{dc_i}{dt} = k_1\left[\sum_{j=1}^{N} m_{i,j}c_i c_j - k_2\sum_{j=1}^{N} m_{i,j}c_i c_j + \sum_{j=1}^{M} m_{j,i}c_i y_j\right] - k_3 c_i \tag{2.9}$$

เทอมแรกคือการจำลองของ paratope แอนติบอดีชนิด i โดย epitope ของแอนติบอดีชนิด j เทอมที่สองแทนการกำจัดของแอนติบอดีชนิด i เมื่อ epitope ถูกจดจำโดย paratope ชนิด j พารามิเตอร์ k_1 คือค่าคงที่ขึ้นอยู่กับจำนวนหน่วยการชนกันต่อหน่วยของเวลาและอัตราการกระตุ้นการผลิตแอนติบอดี ค่าคงที่ k_2 แสดงความไม่เท่ากันระหว่างกระตุ้น และการระงับ เทอมที่สามจำลองความเข้มข้นของแอนติเจน และเทอมล่าสุดจำลองความโน้มเอียงของเซลล์ที่จะตาย

สมการที่ควบคุมการเคลื่อนไหวของเครือข่ายในความหมายนี้ก็คือการจดจำแอนติเจนหรือการขยายจำนวนแอนติบอดีซึ่งจะไม่ถูกกำจัด การสร้างแอนติบอดีเพื่อการจัดระบบจะรวมถึงการต่อสู้กับแอนติเจนที่ไม่คาดหวัง

2.2.5.2 เครือข่ายภูมิคุ้มกันแบบไม่ต่อเนื่อง (Discrete Immune Network Models)

เครือข่ายภูมิคุ้มกันแบบไม่ต่อเนื่องต่างกับเครือข่ายภูมิคุ้มกันแบบต่อเนื่อง ในแง่ของการปรับตัว ไม่ขึ้นอยู่กับสมการดิฟเฟอเรนเชียลแต่เป็นกระบวนการทำซ้ำๆ เพื่อปรับให้เหมาะสม ใช้สำหรับพัฒนาการรู้จำ การจัดกลุ่มข้อมูลและการลดขนาดข้อมูล (data compression) อย่างไรก็ตามการเรียนรู้ของอัลกอริทึมเหล่านี้ยังสามารถประยุกต์ใช้ในสาขาอื่นๆ ดังเช่น การหาคำตอบที่ดีที่สุด การควบคุม และหุ่นยนต์ แต่การเรียนรู้ของอัลกอริทึมสามารถใช้โครงสร้างเครือข่ายภูมิคุ้มกันเทียมที่สามารถสกัดข้อมูลจากชุดของอินพุตที่สัมพันธ์กับสารกระตุ้น ทั้งเซลล์บีและแอนติบอดี จะเป็นส่วนหลักของอัลกอริทึมเครือข่ายภูมิคุ้มกันและแอนติเจนจะเป็นรูปแบบอินพุต

สำหรับอัลกอริทึมที่สำคัญในเครือข่ายแบบนี้จะกล่าวถึงเพียงแบบเดียวที่ใช้ในงานวิจัยนี้ คือ aiNet (Artificial Immune Network)[4]ได้ถูกนำเสนอโดย de Castro และ Von Zuben ในปี 2000 เครือข่ายเริ่มต้นเกิดจากกลุ่มขององค์ประกอบย่อยๆ ที่มีความสัมพันธ์คล้ายกับโมเลกุลของแอนติบอดี เช่น แอตตริบิวต์สตริงที่ถูกแทนในรูปของยูคลิดเชป-สเปซ

การแทนของรูปแบบของแอนติเจนและแอนติบอดีในเครือข่าย ทำการหาค่าแอฟฟินิตีโดยสมการในหัวข้อ (2.2.3) จำนวนของแอนติบอดีที่มีแอฟฟินิตีสูงจะถูกเลือกและทำการเพิ่ม

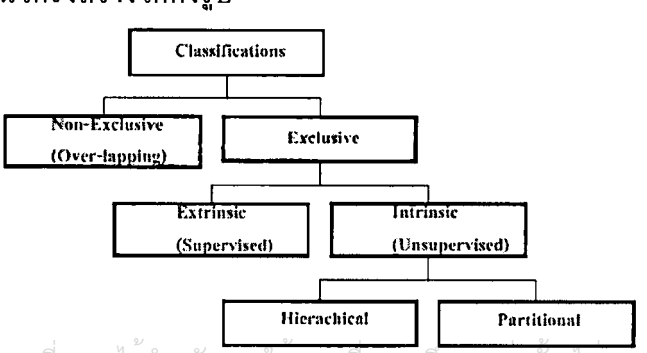
ขยายจำนวน (Clonal expansion) ตามค่าแอฟฟินิตี ถ้ามีค่าแอฟฟินิตีสูงจะมีจำนวนการเพิ่มขยายสูง และอัตราการเปลี่ยนแปลงจะเป็นสัดส่วนโดยผกผันกับค่าแอฟฟินิตีของแอนติเจน โดยที่ค่าแอฟฟินิตีสูงจะมีอัตราการเปลี่ยนแปลงของแอนติบอดีที่ทำการโคลนต่ำ จำนวนของแอนติบอดีที่มีค่าแอฟฟินิตีสูงกว่าจะถูกเลือกเข้าไปเป็นส่วนของโคลนอลเมมโมรี (Clonal memory) แอนติบอดีต่างๆ ที่เหลือจากการโคลน ถ้าแอนติบอดีใดมีค่าแอฟฟินิตีต่ำกว่าค่าขีดแบ่ง (Threshold) จะถูกกำจัดออก (Clonal suppression) และแอนติบอดีทั้งหมดภายในเครือข่ายที่มีค่าแอฟฟินิตีระหว่างแอนติบอดีในเครือข่ายด้วยกันต่ำกว่าค่าขีดแบ่งจะถูกกำจัดออกจากเครือข่ายด้วยเช่นกัน และจะมีการสร้างแอนติบอดีใหม่โดยการสุ่มมารวมเข้ากับเครือข่ายอีก (Metadynamics) สำหรับในการทำงานของอัลกอริทึม aiNet โดยละเอียดจะกล่าวถึงในบทที่ 3

2.3 การจัดกลุ่มข้อมูล (Clustering Methods)

การจัดกลุ่มข้อมูลคือ การแบ่งข้อมูลออกเป็นกลุ่มของข้อมูลที่เหมือนกันซึ่งแต่ละกลุ่มจะเรียกว่า คลัสเตอร์ (Cluster) ข้อมูลที่อยู่ในคลัสเตอร์เดียวกันจะมีความคล้ายคลึงกัน ข้อมูลที่อยู่ต่างคลัสเตอร์จะมีความแตกต่างกัน การจัดกลุ่มข้อมูลอัตโนมัติเป็นงานที่ต้องใช้เทคนิคที่ซับซ้อนเพื่อค้นหากลุ่มข้อมูลได้อย่างแม่นยำ ซึ่งกระบวนการจำเป็นต้องใช้ทั้งเวลาในการประมวลผลและสิ้นเปลืองหน่วยความจำจำนวนมาก อีกทั้งแต่ละเทคนิคยังมีข้อจำกัดที่แตกต่างกันออกไป ดังนั้นจึงมีผู้สนใจศึกษาและพัฒนาเทคนิคการจัดกลุ่มข้อมูลอย่างกว้างขวาง เนื่องจากอัลกอริทึมจัดกลุ่มข้อมูลมีอยู่มากมาย การเลือกใช้อัลกอริทึมที่เหมาะสมกับวัตถุประสงค์ของงานนั้นจึงเป็นสิ่งสำคัญ

การจัดกลุ่มข้อมูลเป็นประเภทหนึ่งของการจำแนกประเภท (Classification) ซึ่งเรียกว่าเป็นการจำแนกประเภทแบบไม่มีผู้สอน (Unsupervised Classification) ซึ่งไม่มีข้อมูลใดๆ ของกลุ่มที่ถูกจัดไว้ก่อนแล้ว ต่างจากการจำแนกประเภทแบบมีผู้สอน (Supervised Classification) ซึ่งกลุ่มของข้อมูลจะถูกกำหนดไว้ก่อนล่วงหน้า จากนั้นข้อมูลจะถูกจัดลงในกลุ่มที่ถูกกำหนดไว้แล้วซึ่งในหัวข้อนี้จะกล่าวถึงเทคนิคที่นิยมใช้ในการจัดกลุ่มข้อมูล

ใน[15,16] ได้กล่าวถึงความสัมพันธ์ของการจำแนกประเภทกับการจัดกลุ่มไว้โดยสามารถแสดงเป็นโครงสร้างได้ดังรูป



รูปที่ 2.16 แสดงโครงสร้างประเภทของการจำแนกข้อมูล

1. Exclusive VS. nonexclusive

   การจำแนกประเภทข้อมูลแบบ exclusive คือการแบ่งเซตข้อมูลออกเป็นกลุ่มโดยที่ข้อมูลแต่ละตัวจะสามารถอยู่ได้เพียงกลุ่มเดียวเท่านั้น ส่วนการจำแนกประเภทข้อมูลแบบ nonexclusive ข้อมูลแต่ละตัวจะสามารถปรากฏได้ในกลุ่มหลายกลุ่ม

2. Intrinsic VS. extrinsic

   การจำแนกประเภทข้อมูลแบบ Intrinsic จะใช้เพียงเมทริกซ์ของความคล้ายเป็นตัวจำแนกประเภทข้อมูล กล่าวคือ เราสามารถเรียกการจำแนกประเภทข้อมูลแบบ Intrinsic ได้ว่าเป็นการจำแนกประเภทข้อมูลแบบไม่มีผู้สอน เนื่องจากไม่มีการกำหนดการจัดกลุ่มใดๆ เริ่มต้น ต่างจากการจำแนกประเภทข้อมูลแบบ extrinsic ซึ่งจะมีการกำหนดกลุ่มของข้อมูลก่อน

3. Hierarchical VS. partial

   ในการจำแนกประเภทข้อมูลแบบ intrinsic นั้นแบ่งออกเป็นสองประเภทหลักคือ แบบ hierarchical ข้อมูลที่ถูกแบ่งจะเป็นลำดับชั้นเชื่อมโยงกัน แต่แบบ partitional ข้อมูลจะถูกแบ่งเป็นกลุ่มๆ แยกจากกัน ซึ่งจะยกตัวอย่างอัลกอริทึมทั้งสองแบบ

เราจะใช้คำว่าการจัดกลุ่ม(Clustering) สำหรับการจำแนกประเภทแบบ intrinsic hierarchical และแบบ intrinsic partitional เท่านั้น ส่วน extrinsic เราจะใช้คำว่า classification) แทน

2.4 อัลกอริทึมการจัดกลุ่มข้อมูล

อัลกอริทึมจัดกลุ่มข้อมูลสามารถแบ่งออกเป็นประเภทต่างๆ ได้ดังนี้

2.4.1 Partitioning Methods

บนฐานข้อมูลจำนวน n เรคคอร์ด การจัดกลุ่มข้อมูลประเภทนี้จะทำการสร้าง k พาร์-ทิชัน โดยแต่ละพาร์ทิชันจะแสดงถึงข้อมูลที่ถูกแบ่งออกเป็นกลุ่ม (โดยที่ $k \leq n$) ในแต่ละกลุ่มจะประกอบไปด้วยข้อมูลอย่างน้อยที่สุด 1 เรคคอร์ด และข้อมูลแต่ละเรคคอร์ดจะต้องถูกจัดให้อยู่ในกลุ่มข้อมูลใดเพียงกลุ่มเดียวเท่านั้น (สำหรับบางเทคนิคอาจอนุโลมให้เรคคอร์ดใดๆ สามารถถูกจัดให้อยู่ในกลุ่มข้อมูลได้มากกว่า 1 กลุ่ม)

การจัดกลุ่มข้อมูลประเภทนี้จะต้องระบุค่า k หรือจำนวนพาร์ทิชันที่ต้องการ โดยกระบวนการจัดกลุ่มเริ่มจากการสร้างพาร์ทิชันตั้งต้น จากนั้นอัลกอริทึมจะทำการวนซ้ำเพื่อปรับพาร์ทิชันให้เหมาะสมโดยการย้ายเรคคอร์ดหรืออ็อบเจกต์จากกลุ่มหนึ่งไปยังอีกกลุ่มหนึ่งที่มีความเหมาะสมกว่า โดยที่พาร์ทิชันที่ดีจะต้องสามารถแบ่งให้ข้อมูลที่มีความใกล้ชิดกันหรือมีความสัมพันธ์กันอยู่ในพาร์ทิชันเดียวกัน ส่วนข้อมูลที่มีความแตกต่างกันจะต้องถูกจัดให้อยู่ในพาร์-ทิชันที่ไกลออกไป ซึ่งแต่ละอัลกอริทึมจะใช้เทคนิคในการพิจารณาความคล้ายคลึงกันของข้อมูลของแต่ละคลัสเตอร์แตกต่างกันออกไป เทคนิคที่เป็นที่รู้จักได้แก่

K-means algorithm

ใช้ค่าเฉลี่ยของอ็อบเจกต์ที่ถูกจัดให้อยู่ในกลุ่มเดียวกันเป็นตัวแทนของทุกอ็อบเจกต์ในกลุ่มนั้น อัลกอริทึมเริ่มต้นจากการรับค่าพารามิเตอร์ k ซึ่งคือจำนวนคลัสเตอร์ที่ต้องการค้นหา จากนั้นอัลกอริทึมจะทำการสุ่มเลือกอ็อบ-เจกต์เริ่มต้นจำนวน k อ็อบเจกต์ ซึ่งแต่ละอ็อบเจกต์แสดงถึงตัวแทนของแต่ละคลัสเตอร์(ค่าเฉลี่ยหรือจุดศูนย์กลางของคลัสเตอร์หรือเซนทรอยด์: centroid) จากนั้นจะทำการจัดกลุ่มให้กับอ็อบ-เจกต์ที่เหลืออ็อบเจกต์จะถูกจัดให้อยู่ในคลัสเตอร์เดียวกันเมื่ออ็อบเจกต์นั้นมีความคล้ายกับตัวแทนของคลัสเตอร์นั้นมากที่สุด จากนั้นจึงทำการคำนวณหาค่าเฉลี่ยของคลัสเตอร์ใหม่ และดำเนินกระบวนการเดียวกันกับอ็อบเจกต์ที่เหลือต่อไปจนกระทั่งทุกอ็อบเจกต์ถูกจัดกลุ่มอย่างสมบูรณ์และอ็อบเจกต์ไม่มีการเปลี่ยนกลุ่มอีกต่อไป

โดยขั้นตอนของอัลกอริทึม K-means[24] จะทำการค้นหา k คลัสเตอร์โดยพยายามทำให้ค่า squared-error มีค่าน้อยที่สุด ซึ่งอัลกอริทึมจะทำงานได้ดีเมื่อคลัสเตอร์มีการกระจายตัวออกเป็นกลุ่มอย่างชัดเจนอีกทั้งยังสามารถรองรับกับข้อมูลที่มีขนาดใหญ่ได้อย่างมีประสิทธิภาพ เนื่องจากมีความซับซ้อนของอัลกอริทึม (computational complexity) เพียง $O(nkt)$ เมื่อ n คือจำนวนอ็อบเจกต์ทั้งหมด k คือจำนวนคลัสเตอร์ และ t คือจำนวนรอบของการวนซ้ำ ซึ่งโดยปกติค่า $k<<n$ และ $t<<n$ ดังนั้นเวลาส่วนใหญ่จึงขึ้นตรงกับขนาดของชุดข้อมูลเป็นสำคัญ อัลกอริทึมแบบ K-means แสดงได้ดังนี้

Algorithm: *k*-means. The *k*-means algorithm for partitioning based on the mean value of the object in the cluster.

Input: The number of cluster *k* and a database containing *n* objects.

Output: A set of *k* clusters that minimizes the squared-error criterion.

Method:

(1) arbitrarily choose *k* objects as the initial cluster center;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

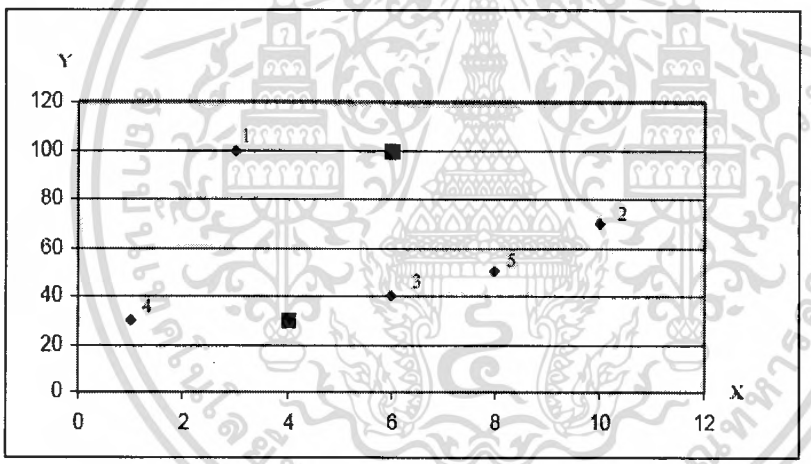
(5) until no change;

k-means สามารถใช้งานได้ดีตราบใดที่สามารถหาค่าเฉลี่ยของคลัสเตอร์ได้แต่ในบางกรณีซึ่งข้อมูลประกอบด้วยข้อมูลเชิงอักขระ จึงอาจต้องปรับปรุงวิธีการคำนวณหาค่าเฉลี่ยและวิธีการพิจารณาความคล้ายคลึงกันของอ็อบเจกต์เพื่อความเหมาะสม การที่ต้องระบุค่า k ให้กับ

อัลกอริทึมอาจทำได้ยากเมื่อผู้ใช้ไม่มีความรู้พื้นฐานเกี่ยวกับข้อมูลที่ใช้ การระบุค่า k ที่ไม่เหมาะสมอาจทำให้กระบวนการจัดกลุ่มข้อมูลต้องใช้เวลานานขึ้น อีกทั้งยังอาจได้คลัสเตอร์ที่ไม่มีคุณภาพอีกด้วยและเนื่องจากการพิจารณาความใกล้ชิดกันของอ็อบเจกต์กับคลัสเตอร์โดยการวัดระยะห่างระหว่างอ็อบเจกต์กับตัวแทนคลัสเตอร์ทำให้คลัสเตอร์ที่พบมีรูปทรงกลม (spherical-shape) ซึ่งเทคนิคนี้อาจไม่เหมาะสำหรับการค้นหาคลัสเตอร์ที่มีรูปทรงอื่นๆ (arbitrary-shape, non-convex shape) อีกทั้งยังมีความอ่อนไหวต่อข้อมูลรบกวน เนื่องจากข้อมูลรบกวนเป็นข้อมูลที่มีลักษณะต่างจากข้อมูลอื่น (มีค่าสูงหรือต่ำกว่าข้อมูลปกติ) ซึ่งอาจส่งผลให้การคำนวณค่าเฉลี่ยของอ็อบเจกต์ในคลัสเตอร์ผิดเพี้ยนได้

ตัวอย่างที่ 2.1 การจัดกลุ่มข้อมูลแบบ K-means

สมมุติมีข้อมูลดังนี้ (3,100),(10,70),(6,40),(1,30),(8,50) กำหนดให้ K=2 นั้นคือเราจะทำการแบ่งกลุ่มข้อมูลทั้งหมดเป็น 2 กลุ่ม ค่าเริ่มต้นของทั้ง 2 กลุ่มคือ (4,30) และ (6,100) ดังรูปที่ 2.17



รูปที่ 2.17 แสดงกราฟ XY ของข้อมูล 1-5 และค่าเริ่มต้นรูปสี่เหลี่ยม

คำนวณระยะห่างระหว่างข้อมูลกับค่าเริ่มต้นของกลุ่มซึ่งสามารถแสดงได้ดังตารางที่ 2.1

ตารางที่ 2.1 แสดงการจัดกลุ่มของอัลกอริทึมแบบ K-means

	กลุ่มที่ 1 (4,30)	กลุ่มที่ 2 (6,100)
1	70.0	3.0
2	40.4	30.3
3	10.2	60.0
4	3	70.2
5	20.4	50.0

จะได้ว่าในกลุ่มที่ 1 มีสมาชิกคือ ข้อมูลที่ 3 และ 4 ในกลุ่มที่ 2 มีสมาชิกคือข้อมูลที่ 1 ,2 และ 5 จากนั้นทำการคำนวณค่าตัวแทนของกลุ่มใหม่เป็น (3,5.30) ตัวแทนกลุ่มที่ 2 ใหม่เป็น (7,73.3)

หลังจากนั้นทำการหาสมาชิกกลุ่มใหม่และหาตัวแทนกลุ่มใหม่ ทำไปจนกระทั่งสมาชิกในกลุ่มไม่เปลี่ยนแปลงหรือตัวแทนของกลุ่มไม่เปลี่ยนแปลง

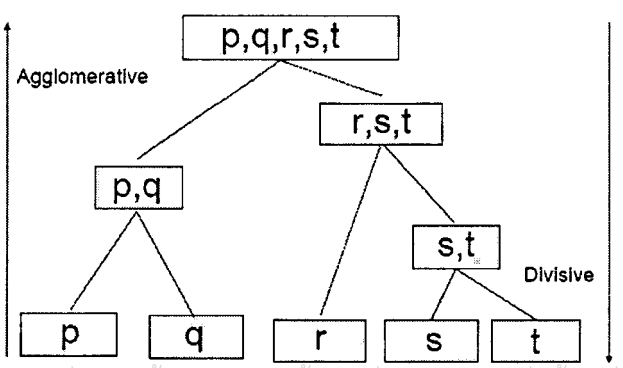
2.4.2 Hierarchical Methods

เป็นวิธีการจัดกลุ่มข้อมูลประเภทหนึ่งที่อาศัยหลักการแบ่งข้อมูลออกเป็นลำดับขั้นคล้ายกับต้นไม้ ซึ่งวิธีการแบ่งกลุ่มข้อมูลแบบนี้สามารถแบ่งออกเป็น 2 แนวทางตามลักษณะการสร้างลำดับขั้นคือ agglomerative approach กับ divisive approach

Agglomerative approach หรือ bottom-up approach เริ่มต้นโดยการให้แต่ละอ็อบเจกต์ อยู่ในคลัสเตอร์ที่ต่างกัน จากนั้นจึงทำการวนซ้ำเพื่อรวมคลัสเตอร์ที่ใกล้ชิดกันเข้าด้วยกัน จนกระทั่งทุกคลัสเตอร์ถูกรวมเข้าเป็นคลัสเตอร์เดียว หรือเมื่อเข้าเงื่อนไขการสิ้นสุดการค้นหา อัลกอริทึมจัดกลุ่มข้อมูลแบบลำดับขั้นส่วนใหญ่จะใช้แนวทางนี้ในการจัดกลุ่มข้อมูล แต่จะต่างกันที่เทคนิคและวิธีการคำนวณค่าความคล้ายคลึงกันระหว่างคลัสเตอร์ (intercluster similarity)

Divisive approach หรือ top-down approach เริ่มต้นโดยให้ทุกอ็อบเจกต์อยู่ในคลัสเตอร์เดียวกัน จากนั้นจึงค่อยๆ แบ่งคลัสเตอร์ออกเป็นคลัสเตอร์ที่เล็กลงมาเรื่อยๆ จนกระทั่งทุกอ็อบเจกต์ถูกแยกออกจากกันทั้งหมดหรือเมื่อเข้าเงื่อนไขการสิ้นสุดการค้นหา เช่นจำนวนของคลัสเตอร์ที่ได้ เป็นต้น

ในการแบ่งหรือรวมคลัสเตอร์แต่ละครั้งจะอาศัยการพิจารณาบนพื้นฐานของสิ่งที่ได้เรียนรู้ขณะปัจจุบันเป็นสำคัญ โดยเมื่อตัดสินใจที่จะรวมหรือแบ่งคลัสเตอร์ใดแล้วจะไม่สามารถย้อนกลับมาแก้ไขได้อีก ทำให้การตัดสินใจในแต่ละรอบใช้การคำนวณเพียงเล็กน้อยเนื่องจากไม่จำเป็นต้องพิจารณาทุกทางเลือก เพียงแต่เลือกตัดสินใจในสิ่งที่ดีที่สุด ณ ขณะนั้นเป็นสำคัญแต่อย่างไรก็ตามเนื่องจากการไม่สามารถย้อนกลับมาแก้ไขการตัดสินใจที่ผิดพลาดได้ การตัดสินใจในแต่ละครั้งจึงต้องพิจารณาอย่างรอบคอบที่สุด



รูปที่ 2.18 แสดงแผนภาพเดนโดแกรมของการจัดกลุ่มแบบ Hierarchical

เทคนิคของ agglomerative จะได้รับความนิยมมากกว่า เนื่องจากสามารถทำได้สะดวกกว่าวิธีของ divisive เริ่มต้นจากการกำหนดฟังก์ชันการวัดระยะห่างแบบยูคลิด เป็นตัววัดความคล้าย อัลกอริทึมของ agglomerative hierarchical สามารถแสดงได้ดังนี้

1. สร้างเมทริกซ์ความแตกต่าง(dissimilarity matrix) NxN ของเอกสารทั้งหมด (เริ่มต้นแต่ละกลุ่มจะมีสมาชิกเป็นหนึ่งเอกสาร)
2. รวมสองกลุ่มที่เหมือนกันที่สุด(ในกรณีที่ใช้ฟังก์ชั่นยูคลิดจะคิดสองกลุ่มที่มีค่าระยะห่างน้อยที่สุด)
3. ทำการปรับปรุงเมทริกซ์ความไม่เหมือน โดยคิดรวมสองกลุ่มก่อนหน้าที่เหมือนกันมากที่สุดเป็นกลุ่มเดียวกัน
4. ทำซ้ำกระบวนการ 2 และ 3 จนกระทั้งเหลือกลุ่มเดียว

ตัวอย่างที่ 2.2 การจัดกลุ่มแบบ agglomerative hierarchical

ให้ A,B,C,D,E เป็นเอกสารที่มีความแตกต่างดังนี้

ตารางที่ 2.2 แสดงเมทริกซ์ความต่าง

	A	B	C	D	E
A	-	9	3	6	11
B	9	-	7	5	10
C	3	7	-	9	2
D	6	5	9	-	8
E	11	10	2	8	-

กลุ่มที่มีความเหมือนกันมากที่สุดคือ C-E=2 ดังนั้นทำการรวมกลุ่มเป็น CE จะได้ตารางความเหมือนใหม่ดังนี้

ตารางที่ 2.3 แสดงเมทริกซ์ความไม่เหมือนหลังจากรวมกลุ่ม C และ E

	CE	A	B	D
CE	-	11	10	9
A	11	-	9	6
B	10	9	-	5
D	9	6	5	-

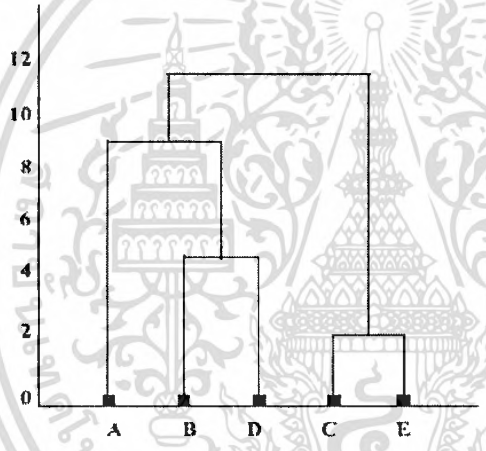
ในการคำนวณความเหมือนระหว่าง CE และ A จะสามารถหาได้จาก

ในกรณีถ้าเป็นแบบ single link เราจะเลือกค่าน้อยที่สุดระหว่าง Sim(A,C) และ Sim(A,E) ซึ่งมีค่าดังนี้ Sim(A,C)=3 และ Sim(A,E)=11 ดังนั้นถ้าเป็นแบบ single link ค่า Sim(CE,A)=3

ในกรณีถ้าเป็นแบบ complete link เราจะเลือกค่ามากที่สุดระหว่าง Sim(A,C) และ Sim(A,E) ซึ่งมีค่าดังนี้ Sim(A,C)=3 และ Sim(A,E)=11 คือ Sim(CE,A)=11 ดังนั้นถ้าเป็นแบบ complete link ค่า Sim(CE,A)=11

ในกรณีถ้าเป็นแบบ group average เราจะใช้ค่าเฉลี่ยระหว่าง Sim(A,C) และ Sim(A,E) ซึ่งมีค่าดังนี้ Sim(A,C)=3 และ Sim(A,E)=11 คือ Sim(CE,A)=7 ดังนั้นถ้าเป็นแบบ group average ค่า Sim(CE,A)=7

จากตัวอย่างตารางที่ 2.3 เราใช้การคำนวณแบบ complete link หลังจากนั้นทำการคำนวณความต่างของกลุ่มใหม่กับกลุ่มเก่าทุกกลุ่ม ทำต่อไปเรื่อยๆ จนกระทั่งเหลือกลุ่มเดียวจะได้ดังรูป 2.19



รูปที่ 2.19 แสดงแผนภาพเดนโดแกรมของตัวอย่างที่ 2.2 แกนแสดงลำดับก่อนหลังการรวมตัว

2.5 การวัดความคล้ายของข้อมูล

-การวัดระยะทางแบบยูคลิด(Euclidean distance)

การวัดค่าความคล้ายคลึงของข้อมูล โดยการใช้ฟังก์ชันการวัดค่าระยะทางแบบยูคลิดเป็นการวัดระยะห่างระหว่างข้อมูลที่ต้องการเปรียบเทียบความคล้ายคลึงกัน โดยระยะห่างระหว่างข้อมูล จะแปรผันตรงกับความคล้ายคลึงกันของข้อมูลโดยผลลัพธ์ที่ได้จากการวัดที่มีค่ามากกว่าแสดงว่าความคล้ายคลึงกันของข้อมูลมีน้อยกว่า หรือค่าที่วัดได้น้อยกว่าจะมีความคล้ายคลึงกันของข้อมูลมากกว่า การวัดระยะทางแบบยูคลิด (Euclidean distance) ซึ่งแสดงดังสมการ(2.10)

$$d(D_i,D_j)=\sqrt{\sum_{k=1}^{n}(w_{dik}-w_{djk})^2} \qquad (2.10)$$

เมื่อ $D_i=(wd_{i1},wd_{i2},\ldots,wd_{it})$ และ $D_j==(wd_{j1},wd_{j2},\ldots,wd_{jt})$ เป็นเวกเตอร์ตัวแทนเอกสาร

-การวัดความเหมือนแบบโคไซน์ (Cosine Similarity)

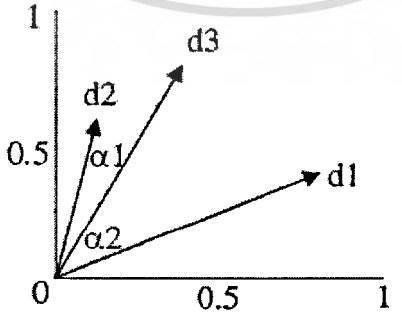
การวัดความคล้ายคลึงอีกแบบหนึ่งที่นิยมใช้ในการวัดความคล้ายคลึงของเอกสารที่เรียกว่าการวัดความเหมือนแบบโคไซน์ (Cosine Similarity) เป็นวิธีการเปรียบเทียบความคล้ายของเอกสารสองเอกสาร โดยแต่ละเอกสารจะถูกแทนด้วย N-dimensional vector ซึ่งเก็บค่าน้ำหนักของคำแต่ละคำในเอกสารนั้น (N-dimensional vector in term space) การเปรียบเทียบความคล้ายกัน ของเอกสารจะเปรียบเทียบโดยดูจากความคล้ายของมุมระหว่าง 2 document vector หากเอกสารคล้ายกันมาก เวกเตอร์จะเกือบทับกันสนิท มุมจึงมีค่าน้อย ค่าโคไซน์ที่ได้จึงมีค่ามาก แสดงดังสมการ (2.11)

$$Sim(D_i,D_j)=\frac{\sum_{k=1}^{t} w_{dik} * w_{djk}}{\sqrt{\sum_{k=1}^{t}(w_{dik})^2 * \sum_{k=1}^{t}(w_{djk})^2}} \qquad (2.11)$$

เมื่อ $D_i=(wd_{i1},wd_{i2},\ldots,wd_{it})$ และ $D_j=(wd_{j1},wd_{j2},\ldots,wd_{jt})$ เป็นเวกเตอร์ตัวแทนเอกสาร

ตัวอย่างที่ 2.3 การคำนวณการวัดความเหมือนแบบโคไซน์

กำหนดให้ d1=(0.4,0.8), d2=(0.8,0.3), d3=(0.2,0.7) เป็นเวกเตอร์ของเอกสาร เราสามารถหาค่าความความเหมือนได้ดังนี้



รูปที่ 2.20 แสดงเวกเตอร์ของเอกสารใน 2 มิติ

$$Sim(d_1,d_2) = \frac{(0.4\times0.8)+(0.8\times0.3)}{\sqrt{[(0.4)^2+(0.8)^2]\times[(0.8)^2+(0.3)^2]}} = \frac{0.56}{\sqrt{0.58}} = 0.74$$

$$Sim(d_1,d_3) = \frac{(0.4\times0.2)+(0.8\times0.7)}{\sqrt{[(0.4)^2+(0.2)^2]\times[(0.8)^2+(0.7)^2]}} = \frac{0.64}{\sqrt{0.42}} = 0.98$$

$$Sim(d_2,d_3) = \frac{(0.8\times0.2)+(0.3\times0.7)}{\sqrt{[(0.8)^2+(0.2)^2]\times[(0.3)^2+(0.7)^2]}} = \frac{0.37}{\sqrt{0.63}} = 0.59$$

2.6 การประเมินผลคุณภาพของการจัดกลุ่ม (Evaluation of Clustering Quality)

ในการวัดคุณภาพของการจัดกลุ่มเอกสารสามารถประเมินผลได้ [17] ได้เป็นสองประเภทหลักคือ

1. การประเมินผลโดยไม่อาศัยข้อมูลภายนอก (Internal Quality Measure) ซึ่งเรียกว่าเป็นการคุณภาพแบบภายใน กล่าวคือ เป็นการวัดค่าเฉลี่ยของค่าความเหมือนของเอกสารทั้งหมด การวัดคุณภาพโดยวิธีนี้จะวัดจากค่าความเหมือนของเอกสารภายในกลุ่ม (Intercluster similarity) ซึ่งควรจะมีค่าสูง และ ค่าความเหมือนระหว่างกลุ่มเอกสาร (Intracluster similarity) ควรมีค่าต่ำ [18] กรรมวิธีในการประเมินแบบไม่อาศัยข้อมูลภายนอกแบบหนึ่งคือใช้วัดค่าความเหมือนรวม เช่น การวัดค่า mean squared error ซึ่งเป็นที่นิยมอย่างมาก

2. การประเมินผลโดยอาศัยข้อมูลภายนอก (External Quality Measure) การวัดเปรียบเทียบสัดส่วนของเอกสารที่มีการจัดกลุ่มโดยขั้นตอนวิธี กับข้อมูลที่มีการจัดกลุ่มไว้แล้ว ซึ่งอาจให้ผู้เชี่ยวชาญในเรื่องดังกล่าวอ่านแต่ละเอกสาร แล้วจัดเป็นหมวดหมู่เหมือนงานบรรณารักษ์ เช่น การวัดค่าพลังงาน(entropy) และการวัดค่าเอฟ-เมเชอร์ (F-measure) ซึ่งเป็นค่าผสมระหว่างค่าความแม่นยำ และ ค่าความระลึก ไว้ในสูตรเดียวกัน

- การวัดค่าแบบ Sum Squared Error

คือ ผลรวมของความผิดพลาดของข้อมูลแต่ละอันกับจุดศูนย์กลางกลุ่ม การวัดความผิดพลาดสามารถวัดได้จากการพิจารณาว่าข้อมูลนั้นอยู่ห่างจากจุดศูนย์กลางกลุ่มเท่าไร ซึ่งสามารถแสดงสมการได้

$$SSE = \sum_{i=1}^{k} \sum_{x\in C} \|x - c_i\|^2 \qquad (2.12)$$

โดยที่ k คือ จำนวนกลุ่มทั้งหมด

C_i คือ กลุ่มแต่ละกลุ่ม, X คือสมาชิกของกลุ่มนั้น

$\|x - c_i\|$ คือระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม

-การวัดค่าเอฟ-เมเชอร์ (F-measure)

การประเมินประสิทธิภาพการจัดกลุ่มคำหลักในการวิจัยนี้ จะใช้การวัดค่าความถูกต้อง (Precision) ค่าความครบถ้วน (Recall) และค่าเอฟ-เมเชอร์เพื่อทดสอบความถูกต้องจากการจัดกลุ่มเอกสาร

ค่าเอฟ-เมเชอร์ คือ ค่าการวัดประสิทธิภาพพื้นฐานในการจัดกลุ่มเอกสาร ซึ่งเกิดจากการรวมเอาค่าการวัดความถูกต้อง (Precision) และค่าความครบถ้วน (Recall) มาคำนวณ โดยค่าRecall คือ ค่าที่บ่งบอกถึงอัตราผลลัพธ์ที่ถูกต้องจากการจัดกลุ่มเอกสาร และค่าความถูกต้อง คือค่าที่บ่งบอกถึงอัตราผลลัพธ์ที่ไม่ถูกต้องจากการจัดกลุ่มเอกสาร สามารถคำนวณได้จากสมการ ดังนี้

$$\text{F-measure}=\frac{2RP}{R+P} \tag{2.13}$$

$$P=\frac{A}{A+B} \tag{2.14}$$

$$R=\frac{A}{A+C} \tag{2.15}$$

เมื่อ P คือ ค่าความถูกต้อง ,R คือ ค่าความครบถ้วน , A คือ จำนวนเอกสารที่สามารถจัดกลุ่มได้ถูกต้อง B คือจำนวนเอกสารที่จัดกลุ่มไม่ถูกต้อง และ C คือจำนวนเอกสารที่ต้องการ

-การวัดความถูกต้องของการจัดกลุ่ม

การวัดผลของการจัดกลุ่มเอกสาร โดยการวัดประสิทธิภาพของความถูกต้องข้อมูล โดยวัดค่าความถูกต้องของการจัดกลุ่มเอกสารดังนี้

$$\text{ความถูกต้อง} = \frac{\text{จำนวนสมาชิกที่ถูกต้องทั้งหมด}}{\text{จำนวนสมาชิกทั้งหมดในกลุ่ม}} \tag{2.16}$$

บทนี้ได้นำเสนอพื้นฐานการทำงานของระบบภูมิคุ้มกันและความเป็นมาของระบบภูมิคุ้มกันเทียม และได้กล่าวถึงการจัดกลุ่มเอกสาร อัลกอริทึม Agglomerative hierarchical และ K-means ซึ่งเป็นอัลกอริทึมที่นิยมใช้ในการจัดกลุ่มเอกสารและแสดงตัวอย่างให้เห็นและวิธีการในการวัดคุณภาพของการจัดกลุ่มข้อมูล

ในบทถัดไปจะเป็นการนำเสนออัลกอริทึม aiNet การประยุกต์ใช้งานอัลกอริทึม aiNet และการปรับวิธีการวัดค่าแอฟฟินิตีของอัลกอริทึม aiNet ต่อไป

บทที่ 3
อัลกอริทึม aiNet และการประยุกต์ใช้

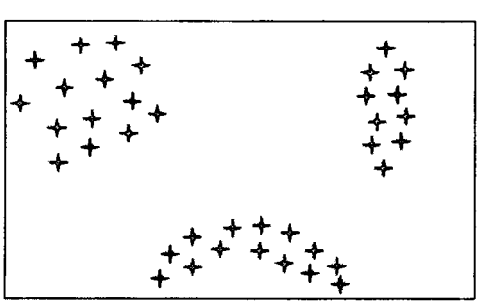
บทนี้จะกล่าวถึงอัลกอริทึม aiNet การประยุกต์ใช้งานและในหัวข้อสุดท้ายจะกล่าวถึง aiNet ที่มีการปรับวิธีการคำนวณค่าแอฟฟินิตีที่ใช้ในงานวิจัยนี้

3.1 อัลกอริทึม aiNet

aiNet (Artificial Immune Network) เสนอขึ้นในปี 2000 โดย de Castra and Von Zuben การกำหนดปริมาณการจดจำของระบบภูมิคุ้มกันเทียม จะพิจารณากระบวนการที่เกิดขึ้นใน shape-space S ซึ่งวัดหลายมิติแต่ละหลักจะเป็นคุณสมบัติทางเคมีของรูปร่างโมเลกุลเพื่อใช้ในการวัด การกำหนดปัญหาขึ้นอยู่กับขนาดของกลุ่ม L ที่เป็นของรูปร่างโมเลกุลดังเช่น จุดของ $s \in S$ ดังนั้นจุดที่ปรากฏในมิติของ L (L-dimensional) เราเรียกว่า เชป-สเปซ ซึ่งเป็นกลุ่มของลักษณะเด่นที่ใช้กำหนดปฏิกิริยาระหว่างแอนติบอดีกับแอนติบอดีหรือปฏิกิริยาระหว่างแอนติเจนกับแอนติบอดี ในทางคณิตศาสตร์รูปแบบนี้จะกำหนดให้อยู่ในรูป L-dimension ของสตริงหรือของเวกเตอร์ ปฏิกิริยาที่เกิดภายใน aiNet จะแทนในรูปของที่มีการเชื่อมต่อของกราฟ เครือข่ายภูมิคุ้มกันเทียมอธิบายได้ดังนี้

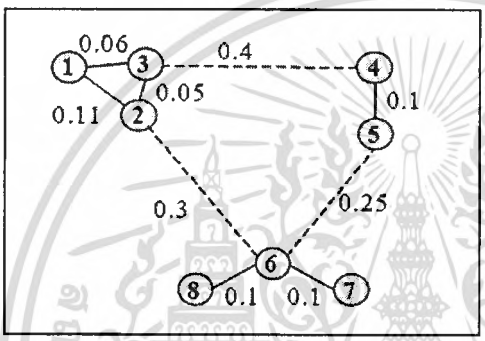
นิยาม aiNet เป็นกราฟที่มีเส้นเชื่อมที่มีน้ำหนักระหว่างโหนด แต่ไม่ต้องเชื่อมต่ออย่างครบถ้วน ประกอบด้วยกลุ่มของโหนดเรียกว่า แอนติบอดี และเส้นเชื่อมระหว่างโหนดเรียกว่าเอดจ์ (edges) มีหมายเลขกำหนดเรียกว่า น้ำหนัก (weight) หรือ ความแรงการเชื่อมต่อ (connection strength) ที่มีความสัมพันธ์กันแต่ละเอดจ์

ภาพคลัสเตอร์ภายใน aiNet จะแสดงถึงวางตัวของคลัสเตอร์ที่มีอยู่ในชุดข้อมูลไปยังเครือข่ายของคลัสเตอร์ โดยสมมุติชุดของมูลประกอบด้วยข้อมูลที่มีความหนาแน่นสูง 3 กลุ่ม ดังรูปที่ 3.1



รูปที่ 3.1 แสดงกลุ่มข้อมูลที่มีความหนาแน่น 3 กลุ่ม

เครือข่ายที่เกิดจากการเรียนรู้ของอัลกอริทึมแสดงได้ดังรูปที่ 3.2 จำนวนตัวเลขภายในเป็นป้ายแสดงอยู่ในแต่ละโหนด (จำนวนของตัวเลขที่สร้างจะมากกว่าจำนวนคลัสเตอร์และน้อยกว่าจำนวนของตัวอย่าง) ตัวเลขถัดไปเป็นการแสดงถึงความแรงในการเชื่อมต่อ และเส้นประแสดงถึงการยกเลิกการเชื่อมต่อเพื่อตรวจสอบคลัสเตอร์และกำหนดโครงสร้างเครือข่ายผลลัพธ์ สังเกตที่คลัสเตอร์ของแอนติบอดีที่แตกต่างกันทั้งสามกลุ่ม โดยแต่ละคลัสเตอร์มีจำนวนแอนติบอดีที่เชื่อมต่อและความแรงแตกต่างกันแผนภาพคลัสเตอร์จะมีลักษณะเหมือนแผนภาพของชุดข้อมูลเริ่มต้น จำนวนของแอนติบอดีในเครือข่ายจะมีจำนวนน้อยกว่าจำนวนข้อมูลตัวอย่าง ซึ่งเป็นคุณสมบัติที่เหมาะสมสำหรับการลดขนาดข้อมูล รูปร่างของแอนติบอดีจะกระจายตามรูปร่างการกระจายของข้อมูลทดสอบ



รูปที่ 3.2 แสดงเครือข่ายผลลัพธ์จากการจัดกลุ่มข้อมูลทดสอบ 3 กลุ่ม

ทฤษฎีเบื้องต้นของเครือข่ายภูมิคุ้มกันที่นำเสนอนั้น เซลล์ที่มีอยู่จะรับรู้สารกระตุ้นและทำให้เกิดการกระตุ้นในเครือข่ายเพื่อเพิ่มจำนวนเซลล์ ขณะที่เซลล์ใดล้มเหลวจะถูกกำจัดออก นอกจากนี้การจดจำระหว่างแอนติบอดีกับแอนติบอดีจะถูกระงับภายในเครือข่ายด้วย การกำจัดจะกระทำด้วยตัวแอนติบอดีเอง เมื่อขีดแบ่งการกำจัด (suppression threshold) คือ σ_s ทุกคู่ของ Ag_j-Ab_i ที่ j=1,…,M, i=1,…,N จะสัมพันธ์ซึ่งกันและกันภายในเชป-สเปซ S ด้วย ค่าแอฟฟินิตี $d_{i,j}$ ปฏิกิริยาจะสะท้อนถึงคุณสมบัติของปฏิกิริยาการเริ่มตอบสนองในการโคลน(clone) และค่าแอฟฟินิตี $s_{i,j}$ จะสะท้อนถึงปฏิกิริยาของ Ab_j-Ab_i เมื่อ i,j=1,…,N

สัญลักษณ์ที่นำมาใช้ในการอธิบายขั้นตอน aiNet[22] มีดังต่อไปนี้

- Ab : เป็นแอนติบอดีในเครือข่าย ($Ab \in S^{Nxl}$, $Ab = b_d \cup Ab_m$);
- Ab_m : กลุ่มของแอนติบอดีเมมโมรี่ ($Ab_m \in S^{mxl}$, $m \leq N$);
- Ab_d : แอนติบอดีที่เกิดขึ้นใหม่และจะรวมกับแอนติบอดีในเครือข่าย ($Ab_d \in S^{dxL}$);
- Ag : ประชากรของแอนติเจน ($Ag \in S^{Mxl}$);
- f_j : เวกเตอร์ที่เก็บค่าแอฟฟินิตีระหว่างแอนติบอดี Ab_i กับแอนติเจน Ag_j เมื่อ i,j=1,…,N
- S : เมทริกซ์ความคล้ายระหว่างแต่ละคู่ของ Ab_i-Ab_j , ที่มีสมาชิก $s_{i,j}(i, j = 1,..,N)$;
- C : ประชากรของการโคลนที่เกิดจากแอนติบอดี Ab($C \in S^{Ncxl}$);

- C* : ประชากรของ C ที่ได้ หลังจากกระบวนการแอฟฟินิตีมัทรูเรชัน;
- d_j : เวกเตอร์ที่เก็บค่าแอฟฟินิตีระหว่างทุกๆ สมาชิกใน C* กับแอนติเจน Ag_j ;
- ζ : เปอร์เซ็นต์ของการเลือกแอนติบอดีที่ดี (re-select 20 %);
- M_j : โคลนอลเมมโมรี่ของแอนติเจน Ag_j ซึ่งเหลือจากกระบวนการกำจัด;
- M_j*: ผลลัพธ์ของโคลนอลเมมโมรี่สำหรับแอนติเจน Ag_j;
- σ_d : เป็นค่าขีดแบ่งในการกำจัดแอนติบอดีที่ต่ำกว่าค่าแอฟฟินิตีหลังจากที่ได้ทำการเลือกแล้ว (re-selection)
- σ_s : ค่าขีดแบ่งในการกำจัดแอนติบอดีที่มีความซ้ำซ้อนออกจากเครือข่าย

การเรียนรู้ของ aiNet มีจุดประสงค์ในการสร้างชุดของหน่วยความจำในการจดจำและการแทนข้อมูลโครงสร้างของเครือข่าย แอนติบอดีที่จำเพาะส่วนมากจะมีความสัมพันธ์ภายในเครือข่ายต่ำ ขณะที่แอนติบอดีทั่วไปส่วนมากจะมีความสัมพันธ์ภายในเครือข่ายมากกว่า ขีดแบ่งการกำจัดจะควบคุมเฉพาะระดับของแอนติบอดีที่เกี่ยวกับ ความถูกต้อง, การจัดกลุ่ม และเครือข่ายที่ยืดหยุ่น ผู้ใช้จะต้องมีแนวทางในการกำหนดค่าพารามิเตอร์ของ aiNet การวิเคราะห์ความเคลื่อนไหวของอัลกอริทึมและความสัมพันธ์ของตัวแปรสำคัญที่ผู้ใช้กำหนดเอง

การเรียนรู้ของอัลกอริทึม aiNet[22] อธิบายได้ดังนี้

1. เริ่มทำแต่ละกระบวนการ:

1.1. เริ่มทำสำหรับแอนติเจนแต่ละตัว $Ag_j, j = 1,...,M, (Ag_j \in \mathbf{Ag})$, :

1.1.1. วัดค่าแอฟฟินิตี $f_{i,j}, i = 1,...,N$, ระหว่างแอนติบอดี Ab_i ทุกตัวกับแอนติ Ag_j โดยที่ $f_{i,j} = 1/D_{i,j}, i = 1,...,N$ และ$D_{i,j} \|\mathrm{Ab_i}\text{-}\mathrm{Ag_j}\|, \mathrm{i}= 1,...,\mathrm{N}$

1.1.2. ทำการเลือกแอนติบอดี $\mathbf{Ab}_{(n)}$ ที่มีค่าแอฟฟินิตีสูงที่สุดจำนวน n ตัวซึ่ง n ได้มาจากสมการที่ (3.1);

1.1.3. ทำการสร้างกลุ่มโคลน C ที่สัมพันธ์กับแอฟฟินิตีของแอนติเจน

1.1.4. กลุ่มของ C จะเกิดกระบวนการแอฟฟินิตีมัทรูเรชันด้วยอัตรา α_k โดยแปรผกผันกับแอฟฟินิตี $f_{i,j}$ ตามสมการ$C_{k*} = C_k + \alpha_k (Ag_j - C_k)$; $\alpha_k \propto 1/f_{i,j}$; $k = 1,..., N_c$, $i = 1,..., N.$;

1.1.5. วัดค่าแอฟฟินิตีระหว่าง **C*** และแอนติเจน $\mathbf{Ag}_j$ $d_{k,j} = 1/D_{k,j}$ โดยที่ $D_{k,j} = \| \mathrm{C^*_k} \text{-Ag}\|, k= 1,...,N$

1.1.6. จาก **C*** ที่ได้จากการโคลนเลือกแอนติบอดีที่มีแอฟฟินิตีสูงโดยเลือกจำนวน ζ% ของแอนติบอดีและใส่ไปยังโคลนอลเมมโมรี่ $\mathbf{M}_j$;

1.1.7. ทำการกำจัดโคลนอลเมมโมรี่ตัวที่มีแอฟฟินิตี $D_{k,j}$ มากกว่าค่าขีดแบ่ง σ_d:

1.1.8. โคลนอลเมมโมรี่ที่เหลือทำการหาค่าแอฟฟินิตี $s_{i,k}$ ระหว่างกันทุกตัว

$s_{i,k} = \| \mathbf{M}_{j,i} - \mathbf{M}_{j,k} \|, \forall_{i,k}$

1.1.9. ทำการกำจัดโคลนอลเมมโมรี่ที่มีค่าต่ำกว่าσ_s $s_{i,k} < \sigma_s$:

1.1.10. ผลลัพธ์โคลนอลเมมโมรี่ที่เหลือจากการกำจัดนำมาเชื่อมต่อกับเมมโมรี่แอนติบอดีเมทริกซ์ $\mathbf{M}_j^*$ for Ag_j: $\mathbf{Ab}_{\{m\}} \leftarrow [\mathbf{Ab}_{\{m\}};\mathbf{M}_j^*]$;

1.2. วัดค่าแอฟฟินิตีของแอนติบอดีเมมโมรี่ทุกๆตัว $\mathbf{Ab}_{\{m\}}$:

$s_{i,k} = \| Ab^i_{\{m\}} - Ab^k_{\{m\}} \|, \forall_{i,k}$

1.3. ทำการกำจัดแอนติบอดีที่มีแอฟฟินิตีต่ำกว่า $s_{i,k} < \sigma_s$:

1.4. ทำการรวมแอนติบอดีเมมโมรี่ $\mathbf{Ab}_{\{m\}}$ เข้ากับแอนติบอดีใหม่ที่สร้าง $\mathbf{Ab} \leftarrow [\mathbf{Ab}_{\{m\}};\mathbf{Ab}_{\{d\}}]$

2. ทำจนครบเงื่อนไข.

จำนวนของการโคลน N_c ของแอนติบอดีสำหรับแต่ละแอนติเจน M แสดงดังสมการที่ (3.1)

$$N_c = \sum_{i=1}^{n} \text{round}(N - D_{i,j}N) \qquad (3.1)$$

เมื่อ

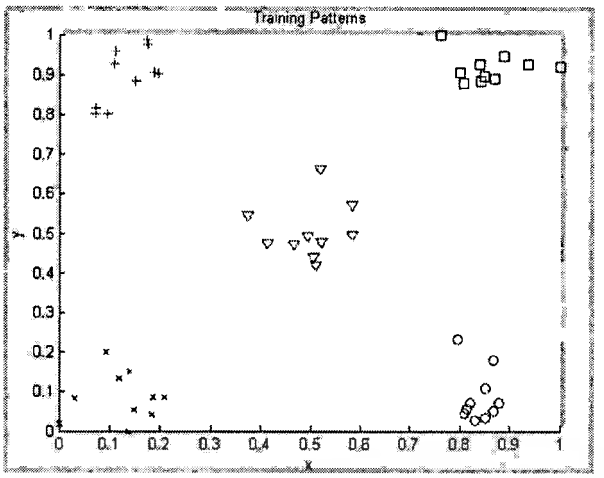
N เป็นจำนวนของแอนติบอดีใน Ab ,

D_{ij} เป็นระยะทางระหว่างแอนติบอดี Ab_i ที่เลือกโดยแอนติเจน Ag_j

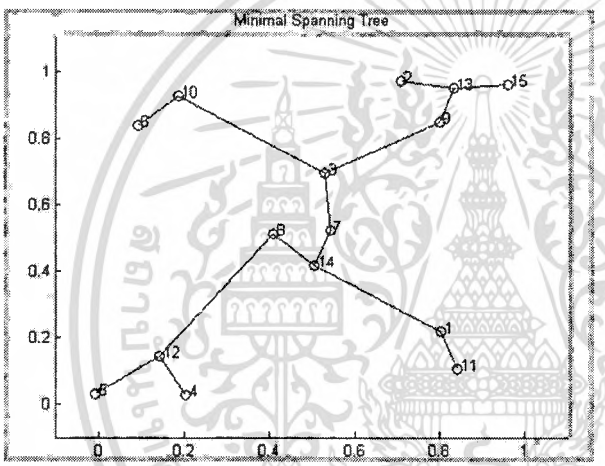
จากอัลกอริทึมข้างบนขั้นตอนที่ 1.1.1 ถึง 1.1.7 จะเป็นขั้นตอนของโคลนอลซีเล็คชันและแอฟฟินิตีมัทรูเรชัน ขั้นตอนที่ 1.1.8 ถึง 1.3 จะเป็นการจำลองกิจกรรมของเครือข่าย ในขั้นตอนของ 1.1.9 จะเรียก Clonal suppression และขั้นตอนที่ 1.3 จะเรียกว่า Network suppression สำหรับ Clonal suppression จะเป็นการกำจัดการเชื่อมต่อภายในของตัวแอนติบอดีเอง เครือข่ายผลลัพธ์จะอยู่ในรูปของแอนติบอดีเมมโมรี่เมทริกซ์

ตัวอย่างที่ 3.1 การจัดกลุ่มข้อมูลโดย aiNet

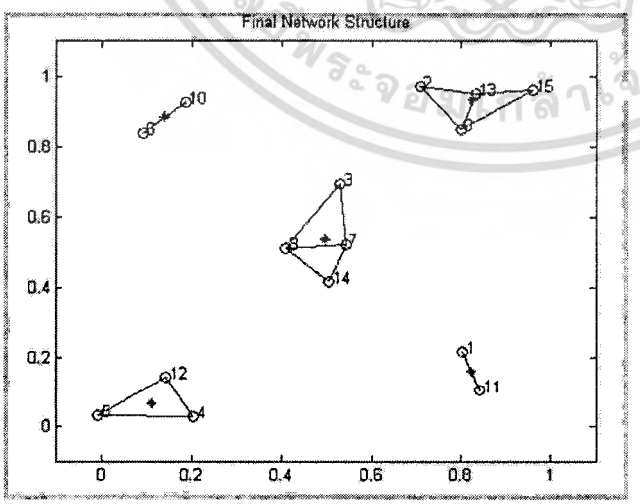
ตัวอย่างนี้เป็นการอธิบายปัญหาในการจัดกลุ่มข้อมูลโดยแสดงดังรูปที่ 3.3 มีตัวอย่างข้อมูลจำนวน 50 ตัวอย่างโดยที่แบ่งเป็น 5 คลัสเตอร์ (ซึ่งคลาสไม่มีการซ้อนทับกัน)โดยมีข้อมูลกลุ่มละ 10 ตัวอย่าง รูปที่ 3.5 แสดงให้เห็นภาพสเหมือนในการสร้างเซลล์เครือข่ายโดยอัตโนมัติและการพิจารณาการเรียนรู้ใน aiNet โดยกำหนดค่าพารามิเตอร์ในการเรียนรู้ดังนี้ n=4 , ζ%=0.2 , σ_d =1.0 σ_s =0.1 หยุดเมื่อเงื่อนไขเมื่อจำนวน generation N=10 เครือข่ายผลลัพธ์จะมี 15 โหนด มีขนาดลดลง 30 % มีอัตราการลดขนาด 70% ซึ่งเครือข่ายผลลัพธ์ที่ได้สามารถแสดงในลักษณะต่างๆ เช่น ในรูปที่ 3.4 จะแสดงเป็น minimal spanning tree ในรูปที่ 3.5 แสดงเป็นกราฟของเครือข่ายโดยมีจุด + เป็นจุดศูนย์กลางของกลุ่ม และในรูปที่ 3.6 เป็นแผนภาพเดนโดแกรมแสดงลำดับขั้นในการแบ่งกลุ่มของเครือข่ายผลลัพธ์ตามค่าแอฟฟินิตีของแต่ละโหนด



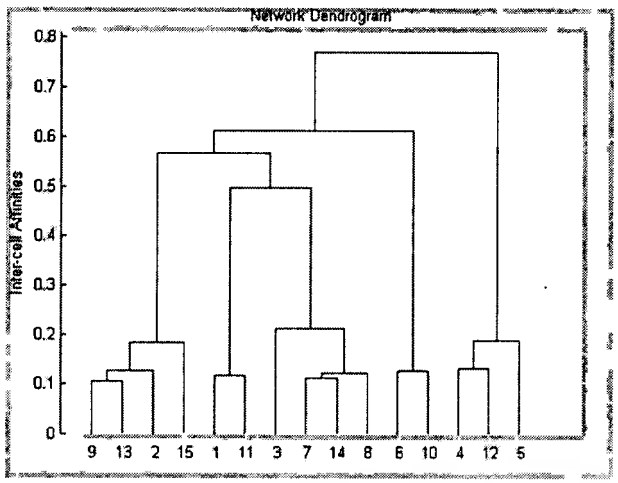
รูปที่ 3.3 ข้อมูลที่ใช้ในการเรียนรู้ตัวในตัวอย่างที่ 3.1



รูปที่ 3.4 แสดง minimal spanning tree ในตัวอย่างที่ 3.1



รูปที่ 3.5 แสดงเครือข่ายผลลัพธ์ประกอบด้วยกราฟย่อยจำนวน 5 กลุ่มในตัวอย่างที่ 3.1

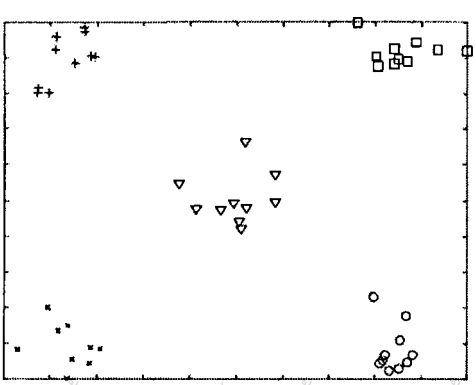


รูปที่ 3.6 แสดงเดนโดแกรม (dendogram) ของ aiNet ในตัวอย่างที่ 3.1

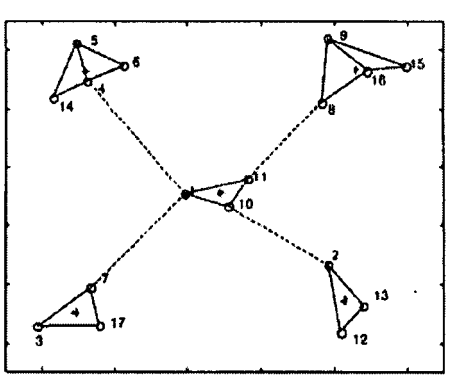
3.2 การประยุกต์ใช้ aiNet ในงานต่างๆ

ในงานวิจัยของ Xiaoshu Hang and Honghua Dai [3] ได้ประยุกต์ใช้ aiNet มาใช้ในการจัดกลุ่มเอกสารที่เป็นเว็บเพจ(web page) มีจำนวน 314 เว็บเพจ ตัวแทนเอกสารที่ใช้แต่ละเอกสารจะเป็นไบนารีเวกเตอร์ ผลการทดลองแสดงถึงผลกระทบที่เกิดจากการปรับค่าพารามิเตอร์ซึ่งให้ผลลัพธ์เครือข่ายสุดท้ายแตกต่างกัน และยังแสดงถึงคุณสมบัติในการลดขนาดข้อมูลของเครือข่ายภูมิคุ้มกันและยังแสดงถึงการเปรียบเทียบกับการจัดกลุ่มแบบต่างๆ โดยวัดคุณภาพการจัดกลุ่มด้วย เอฟ-เมชเชอร์ และ เอนโทรพี

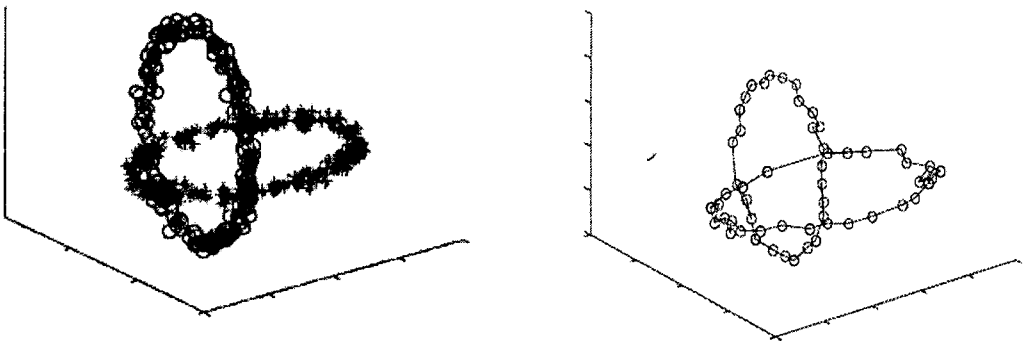
ในงานวิจัยของ Leandro Nunes de Castro and Fernado J. Von Zuben[4] ได้ใช้ aiNet จัดกลุ่มข้อมูลโดยทำการทดลองเพื่ออธิบายเกี่ยวกับคุณสมบัติของ aiNet โดยปัญหาแรกเป็นการอธิบายการจัดกลุ่มข้อมูลตัวอย่างจำนวน 50 ตัวอย่างเพื่อจะจัดกลุ่มออกเป็น 5 กลุ่ม และอีกปัญหาหนึ่งก็คือ two-donut problem ผลของปัญหาแรกจะแสดงถึงคุณสมบัติในการลดขนาดของข้อมูลและปัญหาที่ 2 แสดงถึงการลดขนาดข้อมูลและความถูกต้องของรูปร่างของคลัสเตอร์ที่ได้รับซึ่งข้อมูลทดสอบและผลลัพธ์ของเครือข่ายแสดงดังรูปที่ 3.7



ก. ข้อมูลทดสอบ 50 ตัวอย่าง



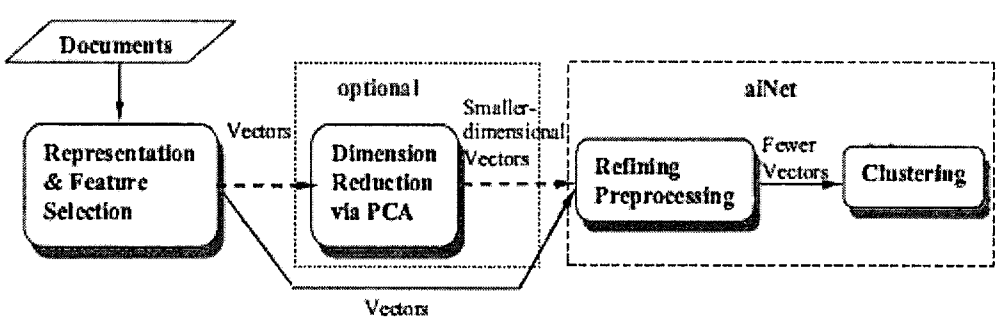
ข. เครือข่ายผลลัพธ์



ค.ข้อมูลทดสอบ two-donut problem ง.เครือข่ายผลลัพธ์ two-donut problem

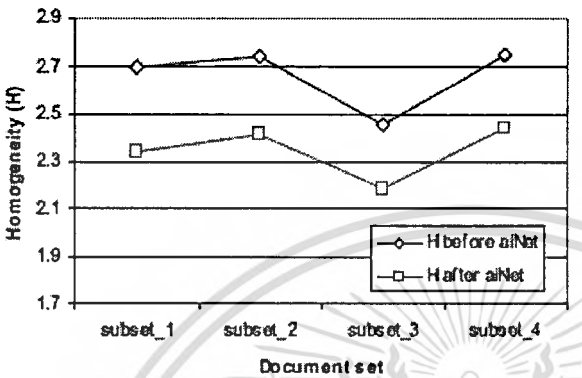
รูปที่ 3.7 แสดงผลการทดสอบเพื่ออธิบายคุณสมบัติ aiNet

ในงานวิจัยของ Natang and V.Rao Vemuri [14] ได้ใช้ aiNet ในการจัดกลุ่มเอกสารข่าว (20 Newgroups)[19] ซึ่งเอกสารข่าวได้ถูกจัดกลุ่มไว้เรียบร้อยแล้วโดยมีเอกสารจำนวน 4 ตัวอย่าง โดยแต่ละตัวอย่างมีเอกสาร 2 กลุ่มและแต่ละกลุ่มมีเอกสารข่าวจำนวน 150 ข่าว โดยตัวแทนข้อเอกสารแต่ละเอกสารแทนในรูปของไบนารีเวกเตอร์และใช้เทคนิคการลดขนาดข้อมูลด้วย Feature Selection[8] และอินพุตที่ใช้ PCA (Principal Component Analysis) [20]และไม่ใช้ PCA เพื่อลดขนาดของอินพุตเวกเตอร์ แล้วใช้ aiNet ที่ปรับปรุงโดยใช้ K-means มาช่วยจัดกลุ่มแอนติบอดีเมมโมรี่ และ aiNet ที่ปรับปรุงโดยใช้ HAC มาช่วยจัดกลุ่มแอนติบอดีเมมโมรี่โดยเปรียบเทียบกับการจัดกลุ่มกับอัลกอริทึม aiNet, K-means และ HAC โดยขอบข่ายในการทดลองแสดงดังรูปที่ 3.8 คือ จากเอกสารทำการหาตัวแทนเอกสารในรูปของเวกเตอร์แล้วนำมาผ่านกระบวนการ Feature Selection เมื่อได้เอกสารชุดนี้แล้วจะนำไปทดลอง 2 วิธี วิธีที่หนึ่ง ผ่านกระบวนการ PCA แล้วนำไปจัดกลุ่มโดย aiNet ,K-means และ HAC และวิธีที่สองนำเอกสารที่ได้ไปจัดกลุ่มโดย aiNet, K-means และ HAC เลยโดยไม่ผ่านกระบวนการ PCA

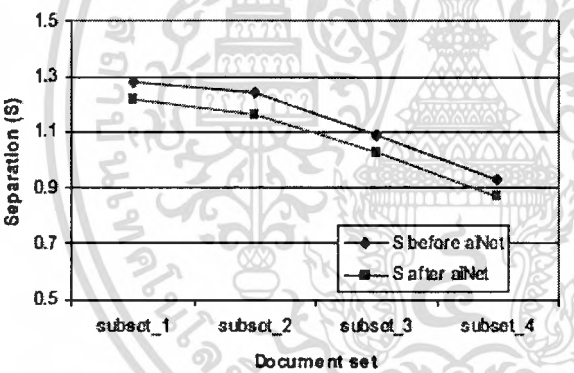


รูปที่ 3.8 แสดงขอบข่ายในการทดลองของ Natang and V.Rao Vemuri

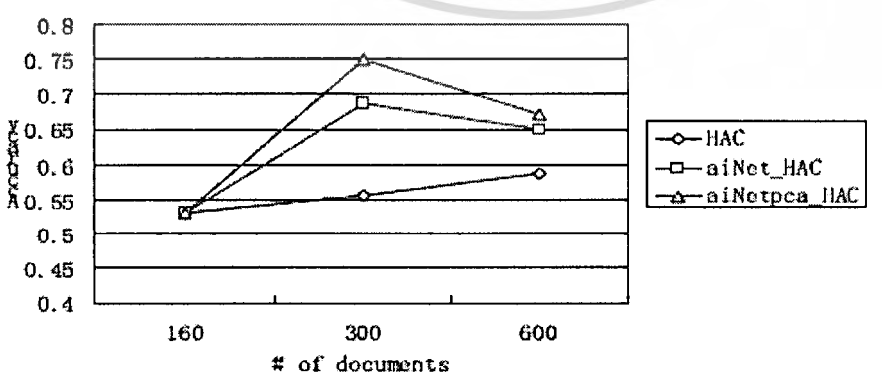
ผลการทดลองที่ได้แสดงถึงคุณภาพของของการจัดกลุ่มด้วยอัลกอริทึมที่แตกต่างกันและการใช้ PCA กับไม่ใช้ PCA ในการลดขนาดของอินพุตและคุณภาพในการจัดกลุ่มก่อนใช้และหลังการใช้ aiNet ผลลัพธ์ที่ได้แสดงให้เห็นว่า aiNet มีความถูกต้องสูงกว่าและให้ผลในการจัดกลุ่มดีกว่า ซึ่งกราฟในรูปที่ 3.9 ได้แสดงการเปรียบเทียบให้เห็นประเด็นต่างๆ เช่น ลักษณะที่คล้ายคลึงกัน การแยกกลุ่ม ความถูกต้องในการจัดกลุ่ม



ก. แสดงลักษณะที่คล้ายคลึงกัน (Homogeneity)



ข. แสดงการแยกกลุ่ม (Separation)



ค. แสดงความถูกต้อง (Accuracy)

รูปที่ 3.9 แสดงผลการเปรียบเทียบของการจัดกลุ่มโดย aiNet ของ Natang and V.Rao Vemuri

3.3 aiNet ที่ปรับปรุงการหาค่าแอฟฟินิตี

อัลกอริทึม aiNet ที่อธิบายในหัวข้อ (3.1) ซึ่งการหาค่าแอฟฟินิตีระหว่างแอนติบอดีกับแอนติเจนและระหว่างแอนติบอดีกับแอนติบอดีภายในเครือข่ายนั้น เป็นกลไกที่ทำให้ aiNet มีการเลือกแอนติบอดีที่มีค่าแอฟฟินิตีกับแอนติเจนสูงกว่าขีดแบ่งเข้ามายังเครือข่ายและรักษาความสมดุลของแอนติบอดีภายในเครือข่ายโดยกำจัดแอนติบอดีที่มีค่าแอฟฟินิตีกับแอนติบอดีด้วยกันต่ำกว่าขีดแบ่งออกไป ซึ่งการหาค่าแอฟฟินิตีโดยปกติกับอินพุตชนิด real valued หรือ real valued เชป-สเปซ นั้นโดยทั่วไป aiNet ใช้การวัดระยะทางแบบยูคลิดแต่สำหรับในงานวิจัยนี้ใช้การหาค่าแอฟฟินิตีโดยใช้การวัดความเหมือนแบบโคไซน์มาแทนเนื่องจากปัญหาการใช้ การวัดระยะทางแบบยูคลิดกับอินพุตหรือแอนติเจนที่มีมิติสูงให้ผลคลาดเคลื่อน

งานวิจัยนี้จึงได้ทำการปรับปรุงในส่วนของการคำนวณค่าแอฟฟินิตีโดยใช้การวัดความเหมือนแบบโคไซน์แทนการวัดระยะทางแบบยูคลิดซึ่งค่าความความเหมือนเชิงมุมมีค่ามากแสดงว่ามีค่าแอฟฟินิตีสูงซึ่งจะตรงกันข้ามกับการวัดระยะทางแบบยูคลิด คือ ระยะทางน้อยมีค่าแอฟฟินิตีสูง สำหรับอัลกอริทึม aiNet ที่ใช้ในงานวิจัยแสดงดังรูปที่ 3.10 โดยคำนวณค่าแอฟฟินิตีระหว่างแอนติบดีและแอนติเจนในแบบของการวัดความเหมือนแบบโคไซน์ ดังสมการที่ (3.3)

$$\text{Affinity}(Ab_i, Ag_j) = \frac{\sum_{k=1}^{L} Ab_{ik} * Ag_{jk}}{\sqrt{\sum_{k=1}^{L} (Ab_{ik})^2 * \sum_{k=1}^{L} (Ag_{jk})^2}} \qquad (3.3)$$

เมื่อ $Ab_i=<Ab_{i1},Ab_{i2},\ldots Ab_{iL}>$ เป็นแอนติบอดี , $Ag_j=<Ag_{j1},Ag_{j2},\ldots Ag_{jL}>$ เป็นแอนติเจน

และกำหนดอัตราการมิวเตชันในการโคลนของแอนติบอดีที่มีค่าแอฟฟินิตีสูง ดังสมการที่ (3.4)

$$C^*_k = C_k + \alpha_k(Ag_j - C_k) \qquad (3.4)$$

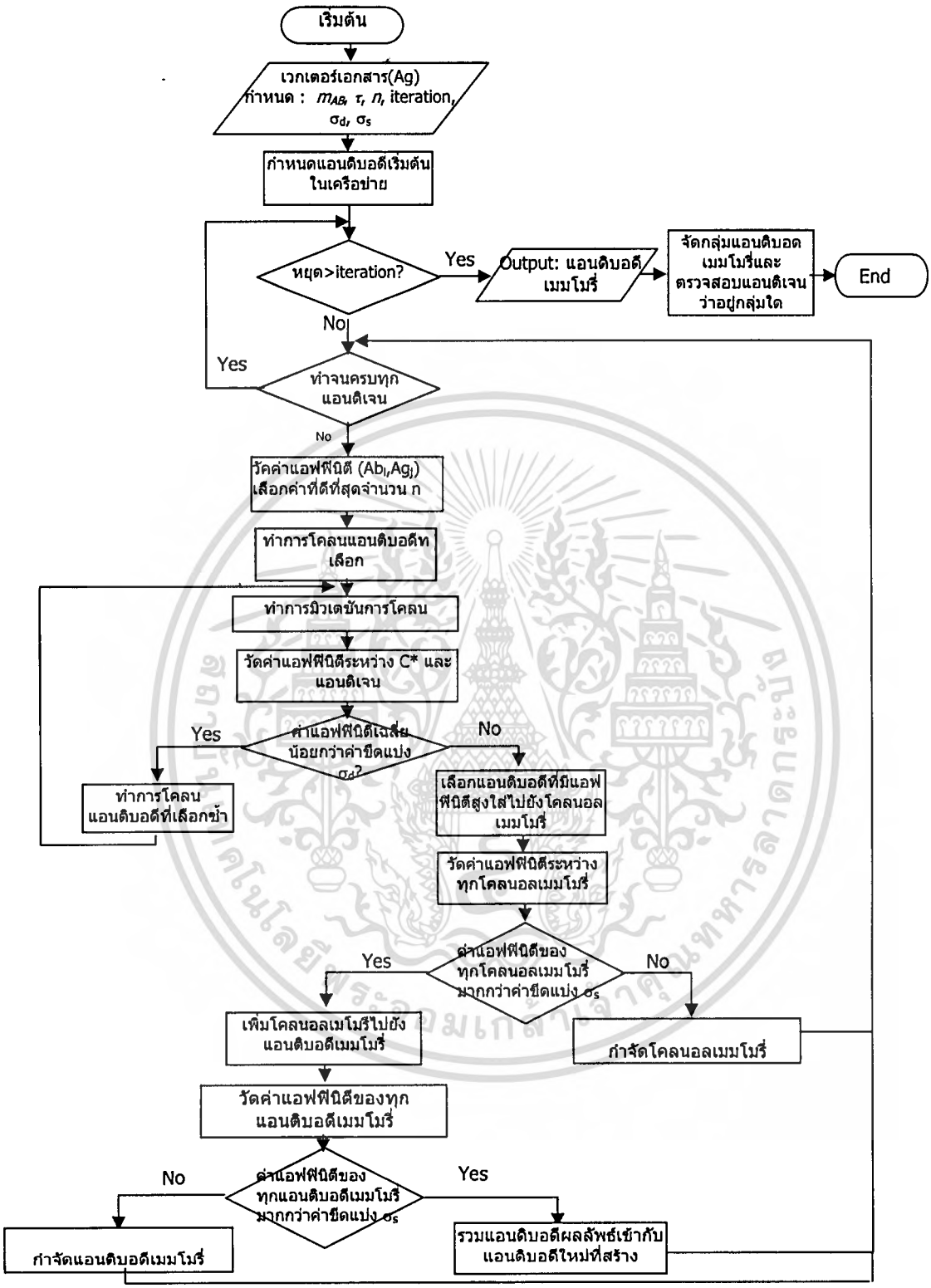
เมื่อ $\alpha_k \propto 1/f_{i,j}$ $k = 1,\ldots,Nc,\ i = 1,\ldots,N$

และจำนวนของการโคลน N_c ของแอนติบอดีสำหรับแต่ละแอนติเจน M แสดงดังสมการที่ (3.5)

$$N_c = \sum_{i=1}^{n} \text{round}(N - D_{i,j}N) \qquad (3.5)$$

เมื่อ N เป็นจำนวนของแอนติบอดีในเครือข่าย

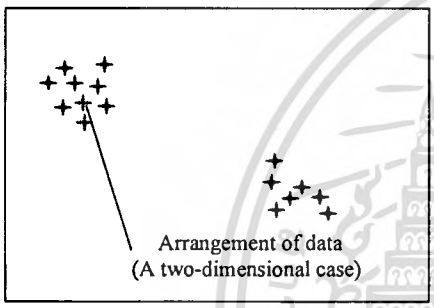
$D_{i,j}$ เป็นแอฟฟินิตีแอนติบอดีที่เลือกโดยแอนติเจน



รูปที่ 3.10 แสดงโฟวชาร์ทอัลกอริทึม aiNet ที่ใช้ในงานวิจัย

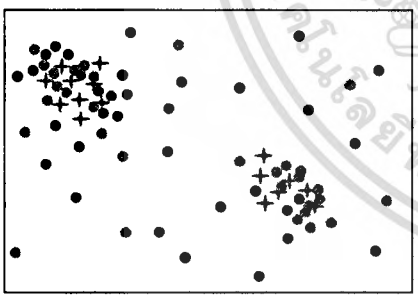
จากโฟวชาร์ทอัลกอริทึม aiNet ในรูปที่ 3.10 เมื่อกำหนดค่าพารามิเตอร์และแอนติบอดีเริ่มต้นในเครือข่ายแล้ว ทำการวัดค่าแอฟฟินิตี(Ab_i,Ag_j)ตามสมการที่ (3.3) ระหว่างแอนติบอดี (•)

ที่อยู่ในเครือข่าย แสดงดังรูปที่ 3.11(ข) และแอนติเจน(+)แสดงดังรูปที่ 3.11(ก) เลือกแอฟฟินิตีที่ดีที่สุดจำนวนหนึ่ง ตามสมการที่ (3.5) เพื่อทำการโคลนมีการมิวเตชั่นตามสมการ (3.4) จะเกิดแอนติบอดีจำนวนมากแสดงดังรูปที่ 3.11(ค) ทำการคัดเลือกเซลล์ที่มีค่าแอฟฟินิตีสูงไปยังโคลนอลเมมโมรี่และทำการวัดค่าแอฟฟินิตีระหว่างทุกๆ โคลนอลเมมโมรี่และทำการคัดเลือกโคลนอลเมมโมรี่ที่มีค่าแอฟฟินิตีมากกว่าค่าขีดแบ่งและ ส่วนที่น้อยกว่าจะถูกกำจัดทิ้งแสดงดังรูปที่ 3.11(ง) และส่วนที่เหลือจะเพิ่มโคลนอลเมมโมรี่เข้าไปเป็นแอนติบอดีเมมโมรี่แสดงดังรูปที่ 3.11(จ) และวัดค่าแอฟฟินิตีของทุกๆ แอนติบอดีเมมโมรี่และกำจัดแอนติบอดีเมมโมรี่ที่มีค่าแอฟฟินิตีมากกว่าค่าขีดแบ่งออกไปซึ่งแอนติบอดีเมมโมรี่ที่เหลือเป็นผลลัพธ์ของเครือข่ายดังรูปที่ 3.11(ฉ) ซึ่งจะไปรวมกับแอนติบอดีที่สร้างใหม่ในเครือข่ายต่อไปและทำกระบวนซ้ำต่อไปจนครบเงื่อนไขของอัลกอริทึม

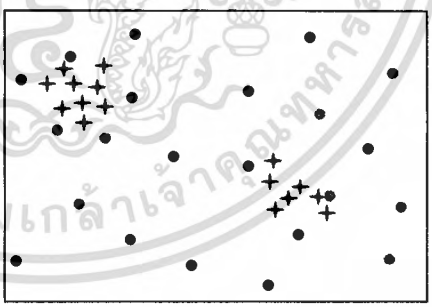


ก. แสดงแอนติเจนหรืออินพุต aiNet

ข. แสดงแอนติบอดีและแอนติเจน

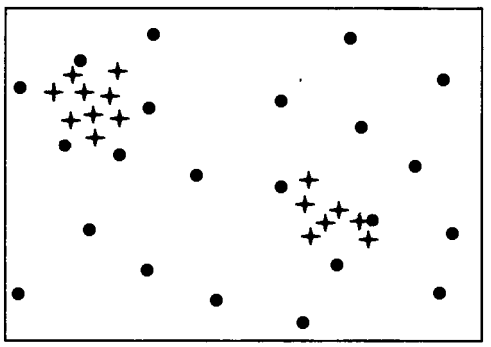


ค. แสดงโคลนอลเมมโมรี่ที่เกิดจากการโคลน

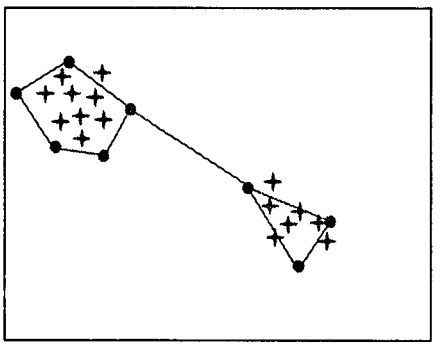


ง. แสดงโคลนอลเมมโมรี่หลังจากการกำจัด

รูปที่ 3.11 แสดงแอนติบอดีและแอนติเจนของอัลกอริทึม aiNet



จ. แสดงโคลนอลเมมโมรี่ที่รวมเป็นแอติบอดีเมมโมรี่



ฉ. แสดงแอนติบอดีเมมโมรี่หลังจากการกำจัด

รูปที่ 3.11 แสดงแอนติบอดีและแอนติเจนของอัลกอริทึม aiNet

ในรูปที่ 3.11 แสดงถึงแอนติบอดีและแอนติเจนภายในเครือข่ายแต่ละช่วงเวลาของอัลกอริทึม aiNet รูปที่ 3.11(ก) เป็นแอนติเจนหรืออินพุตของ aiNet ส่วนรูปที่ 3.11(ข) จุดวงกลมจะเป็นแอนติบอดีภายในเครือข่าย และจะมีกระบวนการวัดค่าแอฟฟินิตีระหว่างแอนติเจนกับแอนติบอดี แอนติบอดีใดมีค่าแอฟฟินิตีสูงก็จะทำการโคลนและเพิ่มจำนวนซึ่งจะได้ดังรูปที่ 3.11(ค) และมีการวัดค่าแอฟฟินิตีระหว่างโคลนอลเมมโมรี่ทุกๆ โคลนอลเมมโมรี่และกำจัดโคลนอลเมมโมรี่ที่มีค่าแอฟฟินิตีมากกว่าค่าขีดแบ่งออกไปซึ่งเป็นกระบวนการ Clonal suppression จะได้โคลนอลเมมโมรี่และแอนติบอดีดังรูปที่ 3.11(ง) สำหรับรูปที่ 3.11(จ) จะเป็นกระบวนการเพิ่มโคลนอลเมมโมรี่ที่ได้จากกระบวนการ Clonal suppression เข้าไปยังแอนติบอดีเมมโมรี่ และวัดค่าแอฟฟินิตีระหว่างแอนติบอดีเมมโมรี่ทุกๆ แอนติบอดีเมมโมรี่และกำจัดแอนติบอดีเมมโมรี่ที่มีค่าแอฟฟินิตีมากกว่าค่าขีดแบ่งออกไป(Network suppression)ซึ่งจะได้เครือข่ายผลลัพธ์สำหรับแอนติเจนดังรูปที่ 3.11(ฉ)

สำหรับงานวิจัยนี้ได้ใช้อัลกอริทึม K-means ในการจัดกลุ่มแอนติบอดีเมมโมรี่(ผลลัพธ์ของ aiNet) เพื่อช่วยให้การตรวจสอบนั้นทำได้ง่ายขึ้น โดยทำให้ทราบว่าแอนติเจนหรืออินพุตนั้นอยู่ในกลุ่มใดของแอนติบอดีเมมโมรี่

ในบทนี้ได้นำเสนอการวิธีการปรับการคำนวณค่าแอฟฟินิตีของ aiNet ด้วยวิธีการวัดความเหมือนแบบโคไซน์เพื่อเพิ่มคุณภาพและเพิ่มความถูกต้องในการจัดกลุ่มเอกสาร

ในบทถัดไปจะทำการทดลองเปรียบเทียบคุณภาพของการจัดกลุ่มข้อมูลของโมเดลเดิมกับโมเดลที่ปรับวิธีการวัดค่าแอฟฟินิตี พร้อมทั้งวัดคุณภาพในการจัดกลุ่มเอกสารที่ได้จากโมเดลทั้งสอง

บทที่ 4
การทดลองในการจัดกลุ่มเอกสาร

ในบทนี้จะกล่าวถึงขั้นตอนในการจัดเตรียมเอกสารและการกำหนดค่าพารามิเตอร์ที่ใช้ในการทดลองและผลที่ได้จากการทดลอง

4.1 การทดลองที่ 1 การจัดกลุ่มเอกสารข่าวจำนวนเอกสารที่แตกต่างกัน

4.1.1 จุดประสงค์ในการทดลอง

เพื่อทดสอบว่าการจัดกลุ่มโดยใช้กลุ่มของเอกสารที่เหมือนกันแต่จำนวนของเอกสารที่ใช้ในการจัดกลุ่มมีจำนวนต่างกันนั้น จะส่งผลต่อคุณภาพในการจัดกลุ่มของเอกสารอย่างไร

4.1.2 ข้อมูลในการทดลอง

เอกสารที่ใช้ในการทดลองคือ 20 Newgroup data set [19] ซึ่งนำมาจาก http://people.csail.mit.edu/jrennie/20Newsgroups ประกอบด้วยเอกสารจำนวน 20,000 เอกสารที่มีการจัดกลุ่มหัวข้อที่ต่างกัน 20 หัวข้อ

ตารางที่ 4.1 แสดงหัวข้อข่าว

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

ข้อมูลทดสอบในการทดลองนี้จะเลือกเอกสารจาก 2 หัวข้อคือ sci.crypt และ sci.electronics โดย Subset A เลือกเอกสารโดยการสุ่มมากลุ่มละ 80 เอกสาร, Subset B เลือกโดยการสุ่มมากลุ่มละ 150 เอกสารและ Subset C เลือกโดยการสุ่มมากลุ่มละ 300 เอกสาร

4.1.3 ขั้นตอนในการจัดเตรียมเอกสารในการจัดกลุ่ม

การจำลองแบบเชิงแนวคิดเพื่อหาตัวแทนเอกสารสามารถจำแนกได้เป็น 3 แบบ คือ แบบจำลองทางบูลีน (Boolean Model) แบบจำลองทางสถิติ (Statistic Model) และแบบจำลอง

ทางเวกเตอร์ (Vector Model) งานวิจัยนี้จะกล่าวเฉพาะแบบจำลองแบบเวกเตอร์ ซึ่งเป็นแบบจำลองที่นิยมใช้ในการจัดกลุ่มเอกสารเพราะแบบจำลองดังกล่าวแทนเอกสารแต่ละฉบับโดยแต่ละมิติของเวกเตอร์จะแทนน้ำหนักของคำที่ปรากฏในเอกสารแสดงดังรูปที่ 4.1 กรรมวิธีในการเลือกคำที่จะมาเป็นตัวแทนของเอกสาร โดยมีหลักเบื้องต้นดังนี้

1.การหาคำหยุด (Stopwords) คำหยุดเป็นคำที่เกิดในเอกสารทุกฉบับและเกิดเป็นปริมาณมากๆทำให้ไม่สามารถใช้เป็นคำในการจำแนกเอกสารได้ต้องกำจัดออก

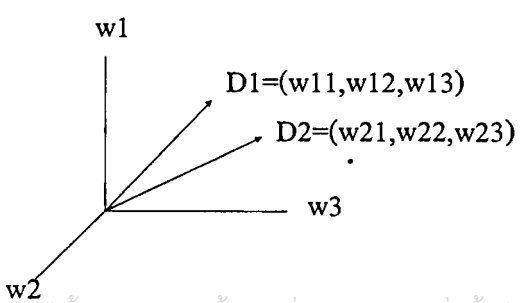
ตารางที่ 4.2 แสดงตัวอย่างคำหยุด (Stoplist Word)

a	been	get	least	our
about	before	getting	left	ourselves
after	being	go	less	out
again	between	goes	let	over
ago	but	going	like	per
all	by	gone	make	put
also	can	gotten	may	same

2.การหารากศัพท์ (Stemming) เป็นการหารูปเดิมของคำหรือหาคำที่มีความหมายคล้ายกันเพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ สำหรับบทความฉบับนี้ใช้วิธีการหารากศัพท์ด้วย Porter Algorithm[23]

ตารางที่ 4.3 แสดงตัวอย่างคำที่ผ่านการหารากศัพท์ด้วยอัลกอริทึม Porter

คำต้นฉบับ	คำที่ผ่านการหารากศัพท์
Andy	andi
Murray	murrai
Run	run
Running	run



รูปที่ 4.1 แสดงการเก็บเอกสารด้วยแบบจำลองแบบเวกเตอร์ (Vector Model) ใน 3-มิติ

วิธีการให้น้ำหนักของคำที่ใช้กันอย่างมากในการสืบค้นข้อมูล โดยคิดน้ำหนักคำจากค่าผลคูณของ tf(term frequency) ซึ่งเป็นความถี่ของคำที่ปรากฏในเอกสารและค่า idf(inverse document frequency) คำนวณจากค่า log(N/df) ซึ่ง N คือจำนวนเอกสารในชุดเอกสารทั้งหมด และ df คือจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ วิธีให้น้ำหนักของคำใน [7] มีการ normalization ทำให้เวกเตอร์เอกสารมีขนาด 1 หน่วยมีสูตรดังนี้

$$W_{i,k} = \frac{tfik * \log(N/dfk)}{\sqrt{\sum_{j=1}^{i} (tfij)^2 * (\log(N/df))^2}} \qquad (4.1)$$

โดยที่ tf_{ik} คือความถี่ของคำในเอกสาร i, N คือจำนวนของเอกสารในชุดเอกสาร, df_k คือจำนวนเอกสารในชุดเอกสารซึ่งบรรจุคำ k เมื่อผ่านกระบวนการทั้งหมดแล้วจะได้เอกสารที่ถูกแทนอยู่ในรูปของ Document Word Matrix

ตารางที่ 4.4 แสดง Document Word Matrix

เอกสาร	ค่า TF-IDF			
	W_1	W_2	...	W_m
Doc_1	W_{11}	W_{12}	...	W_{1m}
Doc_2	W_{21}	W_{22}	...	W_{2m}
...	...	...	...	...
Doc_n	W_{n1}	W_{n2}	...	W_{nm}

ตัวอย่างที่ 4.1 แสดงการหาความถี่ของคำ

ในชุดเอกสารหนึ่งประกอบด้วยเอกสาร D_1, D_2 , D_3 นำเอกสารแต่ละฉบับมาตัดคำ (word segmentation) ดึงคำหยุดออกไปและหารากศัพท์ก็จะได้เอกสารตามรูปที่ 4. 2 จากนั้นหาความถี่ของคำที่ไม่ซ้ำกันในเอกสารแต่ละฉบับจะได้ดังตารางที่ 4.5 แล้วทำการหาค่า df และ idf ให้แต่ละคำจะได้ดังตารางที่ 4.6 และทำการหาเวคเตอร์เอกสารทั้งหมดโดยที่แถวของเมทริกซ์คือเอกสารทั้งหมด และสดมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร ถ้าคำในสดมภ์ปรากฏอยู่ในเวคเตอร์เอกสารฉบับหนึ่งๆ สามารถหาค่าน้ำหนักได้ตามสมการ (4.1) แต่ถ้าไม่ปรากฏคำนั้นในเอกสารที่กำลังพิจารณาอยู่ก็ให้คำนั้นมีค่าน้ำหนักเป็น 0 จากเอกสารที่ไม่มีโครงสร้างก็จะถูกแทนเป็นระบบด้วยเวกเตอร์ซึ่งอยู่ในรูปของเมทริกซ์เอกสาร-คำดังรูปที่ 4.3 ซึ่งใช้เป็นอินพุตในขั้นตอนการจัดกลุ่มเอกสารต่อไป

D_1: computer information computer computer

D_2: internet computer internet data

D_3: system internet

รูปที่ 4.2 แสดงชุดเอกสารในการหาความถี่

ตารางที่ 4.5 แสดงความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	tf
D_1	Computer	3
D_1	Information	1
D_2	Internet	2
D_2	Computer	1
D_2	Data	1
D_3	System	1
D_3	Internet	1

ตารางที่ 4.6 แสดงค่า idf ของคำในชุดเอกสาร

คำ	df	idf
T1: computer	2	0.18
T2: information	1	0.48
T3: internet	2	0.18
T4: system	1	0.48
T5: data	1	0.48

	T1	T2	T3	T4	T5
D_1	0.74	0.67	0	0	0
D_2	0.57	0	0.29	0	0.77
D_3	0	0	0.35	0.94	0

รูปที่ 4.3 แสดงเมทริกซ์เอกสาร-คำ ที่ได้โดยการคำนวณน้ำหนักของคำจากสมการ (4.1)

4.1.4 ตัวแทนเอกสารในการจัดกลุ่ม

เมื่อได้เมทริกซ์เอกสาร-คำตาม(4.1.3)แล้ว ให้ L คือจำนวนของคำที่ปรากฏในเอกสาร L มีมิติขนาดใหญ่ การทำให้ L มีขนาดเล็กลงเพื่อให้มีมิติ *l* ที่ดีสำหรับการจัดกลุ่มเอกสาร โดย *l* << L เพื่อลดเวลาในการคำนวณและให้การใช้ทรัพยากรน้อย โดยใช้วิธีที่เรียกว่า Feature Selection [8] ซึ่งมีสมการดัง(4.2)

$$q(w)= \sum_{i=1}^{N} f_i^2 - \frac{1}{N}[\sum_{i=1}^{N} f_i]^2 \qquad (4.2)$$

เมื่อ f_i ความถี่ของคำ w ในเอกสาร d_i N เป็นจำนวนทั้งหมดของเอกสาร ซึ่งวิธีนี้เราเลือกมา 15% ของคำในเอกสารที่ใช้ทดลอง ซึ่งยังให้ผลการจัดกลุ่มเหมือนเดิม หลังจาก Feature Selection แล้วใช้ Principal Component Analysis (PCA) [20] เพื่อลดมิติของ weight vector ที่ได้จาก Feature Selection โดยลดมิติให้น้อยลง และใช้เพียง 20 dimensional เป็นอินพุตของ aiNet อัลกอริทึม

4.1.5 ขั้นตอนวิธีการทดลองจัดกลุ่มโดยใช้ aiNet

นำข้อมูลที่ได้จากหัวข้อที่ 4.1.3-4.1.5 โดยรูปแบบเอกสารแต่ละฉบับจะอยู่ในรูปของเวกเตอร์ที่มีมิติเป็น *l* มิติต่างๆ ก็คือเซตของคำภายในเอกสาร เมื่อมีเอกสารจำนวน n เอกสารก็จะเป็นเมทริกซ์ n x *l* โดยค่าที่อยู่ในเมทริกซ์ก็คือค่าของน้ำหนักของคำที่ปรากฏในแต่ละเอกสาร โดยจะเป็นค่าแบบ real value

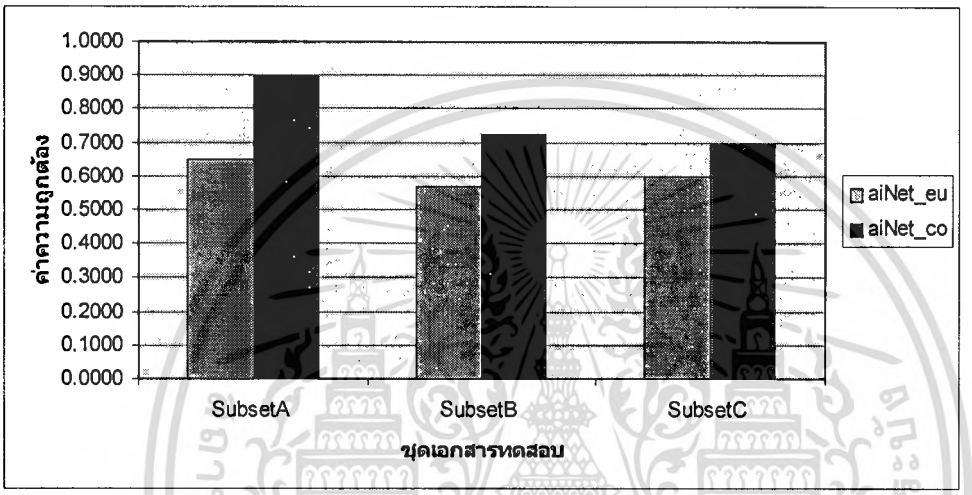
เวกเตอร์ของเอกสารก็คือ กลุ่มของแอนติเจนที่จะใช้เรียนรู้ภายใน aiNet อัลกอริทึม โดยทำการปรับค่าพารามิเตอร์*(σd, σs, ς%, iteration)* ของอัลกอริทึม aiNet ที่วัดค่าแอฟฟินิตี แบบระยะทางแบบยูคลิด(aiNet_eu)และแบบการวัดความเหมือนแบบโคไซน์(aiNet_co) เพื่อจัดกลุ่มข้อมูล กับข้อมูลทดสอบแสดงดังตารางที่ 4.7

ตารางที่ 4.7 แสดงค่าพารามิเตอร์ต่างๆ ที่ใช้ในการทดลอง

Subset	Algorithms	ค่าพารามิเตอร์			
		σd	*σs*	*ς%*	*iteration*
Subset A	aiNet_eu	0.5110	0.0014	0.2	10
	aiNet_co	0.9985	0.6424	0.2	10
Subset B	aiNet_eu	0.5115	0.0015	0.2	10
	aiNet_co	0.9980	0.6225	0.2	10
Subset C	aiNet_eu	0.5110	0.0016	0.2	10
	aiNet_co	0.9985	0.6515	0.2	10

เมื่อสิ้นสุดการทำงานของ aiNet อัลกอริทึมกับข้อมูลแล้วจะได้กลุ่มของแอนติบอดีกลุ่มหนึ่งในแอนติบอดีเมมโมรี่เซลล์แล้วนำมาจัดกลุ่มด้วย K-means และนำแอนติเจนมาตรวจสอบว่าอยู่ในกลุ่มใด แล้ววัดผลของการจัดกลุ่มเอกสารโดยการวัดประสิทธิภาพของความถูกต้องข้อมูลโดยวัดค่าความถูกต้องของการจัดกลุ่มเอกสารดังสมการ (4.3)นี้

$$\text{ความถูกต้อง} = \frac{\text{จำนวนสมาชิกที่ถูกต้องทั้งหมด}}{\text{จำนวนสมาชิกทั้งหมดในกลุ่ม}} \qquad (4.3)$$



รูปที่ 4.4 แสดงการเปรียบเทียบผลการทดลองวัดความถูกต้องของการทดลองที่ 4.1

ผลการทดลองการจัดกลุ่มเอกสารโดยการใช้อัลกอริทึม aiNet โดยวิธีการประยุกต์การใช้การวัดความเหมือนแบบโคไซน์ระหว่างคู่ของเวกเตอร์เอกสารใดๆ แทนการใช้การวัดระยะทางยูคลิด แสดงได้ดังกราฟรูปที่ 4.4 โดยอัลกอริทึม aiNet ที่ใช้การวัดความเหมือนแบบโคไซน์ (aiNet_co) ให้ค่าความถูกต้องดีกว่าอัลกอริทึม aiNet ที่ใช้การวัดระยะทางแบบยูคลิด (aiNet_eu) ในทุกชุดเอกสารที่ใช้ทดสอบ

4.2 การทดลองที่ 2 การจัดกลุ่มโดยใช้เอกสารจากหลายกลุ่มที่มีจำนวนเอกสารเท่ากัน

4.2.1 จุดประสงค์ในการทดลอง

เพื่อทดสอบการจัดกลุ่มเอกสาร ที่ใช้เอกสารหลายกลุ่มโดยมีจำนวนเอกสารเท่ากันแล้วเปรียบเทียบผลที่ความถูกต้องและค่าเอฟ-เมเชอร์ ว่าข้อมูลในแต่ละตัวอย่าง ส่งผลการจัดกลุ่มอย่างไร มีข้อมูล 4 กลุ่มนำมาโดยวิธีการสุ่มมาจากเอกสารแต่ละหัวข้อที่แตกต่างกันแสดงรายละเอียดดังตารางที่ 4.8

4.2.2 ข้อมูลสำหรับการทดลอง

ข้อมูลที่ใช้ในการทดลองจะเป็นข้อมูลเดียวกับหัวข้อ 4.1.2 แต่จะเลือกในหัวข้อที่แตกต่างกัน แสดงดังตารางที่ 4.8

ตารางที่ 4.8 แสดงข้อมูลทดสอบในการทดลองที่ 2 โดยเลือกแต่ละ Topic

Dataset	Topic	จำนวนเอกสารแต่ละกลุ่ม	Total #docs
subset 1	sci.crypt, sci.space	150, 150	300
subset 2	sci.crypt, sci.electronics	150, 150	300
subset 3	sci.space, rec.sports.basketball	150, 150	300
subset 4	talk.politics.mideast, talk.politics.misc	150, 150	300

4.2.3 ขั้นตอนในการจัดเตรียมเอกสารในการจัดกลุ่ม

ใช้แบบจำลองแบบเวกเตอร์โมเดล ซึ่งเป็นแบบจำลองที่นิยมใช้ในการจัดกลุ่มเอกสาร เพราะแบบจำลองดังกล่าวแทนเอกสารแต่ละฉบับโดยแต่ละมิติของเวกเตอร์จะแทนน้ำหนักของคำที่ปรากฏในเอกสาร สำหรับวิธีการหาได้กล่าวแล้วในหัวข้อที่ 4.1.2

4.2.4 ตัวแทนเอกสารในการจัดกลุ่ม

สำหรับตัวแทนเอกสารนั้นจะต้องผ่านกระบวนการ Feature Selection และ Principal Component Analysis (PCA) เพื่อลดมิติของ weight vector เพื่อลดเวลาในการประมวลผล

4.2.5 ขั้นตอนวิธีการทดลองจัดกลุ่มโดยใช้ aiNet

ค่าพารามิเตอร์ที่ใช้ในการทดลองของอัลกอริทึม aiNet ในการทดลองที่ 4.2 เป็นการตรวจสอบความถูกต้องและประสิทธิผลของเอกสารในหัวข้อเรื่อง (topic) ที่หลากหลายและทำปรับค่าพารามิเตอร์ในการทดลองจนให้ผลที่ดีที่สุด แสดงดังตารางที่ 4.9

ตารางที่ 4.9 แสดงค่าพารามิเตอร์ต่างๆ ที่ใช้ในการทดลอง

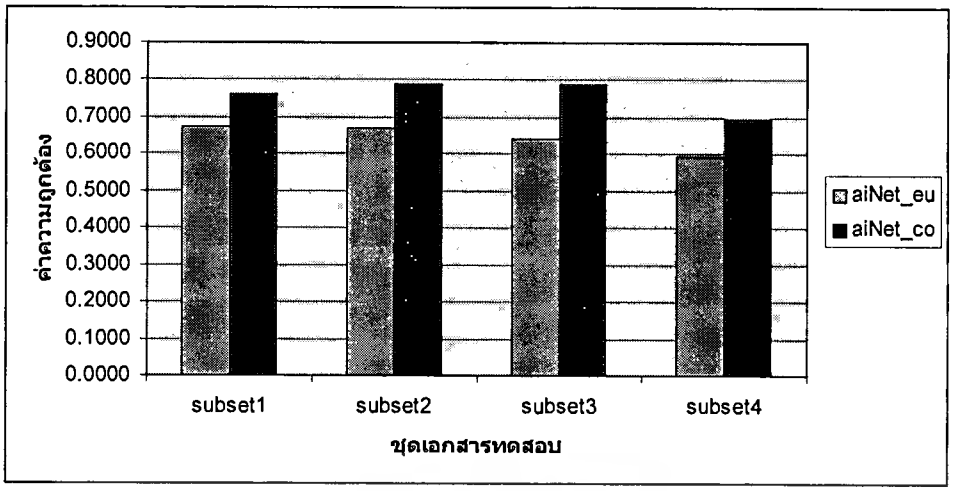
Subset	Algorithms	ค่าพารามิเตอร์			
		σd	σs	$\zeta\%$	*iteration*
Subset 1	aiNet_eu	0.5100	0.0014	0.2	10
	aiNet_co	0.9980	0.6229	0.2	10
Subset 2	aiNet_eu	0.5120	0.0016	0.2	10
	aiNet_co	0.9957	0.7812	0.2	10
Subset 3	aiNet_eu	0.5515	0.0015	0.2	10
	aiNet_co	0.9985	0.6424	0.2	10
Subset 4	aiNet_eu	0.5100	0.0015	0.2	10
	aiNet_co	0.9985	0.6223	0.2	10

เมื่อสิ้นสุดการทำงานของอัลกอริทึม aiNet กับข้อมูลแล้วจะได้กลุ่มของแอนติบอดีกลุ่มหนึ่งในแอนติบอดีเมมโมรี่เซลล์ แล้วนำมาจัดกลุ่มและนำแอนติเจนมาตรวจสอบว่าอยู่ในกลุ่มใด วัดผลของการจัดกลุ่มเอกสาร โดยการวัดประสิทธิภาพของความถูกต้องข้อมูลโดยวัดค่าความถูกต้องของการจัดกลุ่มเอกสารดังสมการ(4.3)และวัดค่าเอฟ-เมเชอร์ [21] ซึ่งเป็นค่าที่รวมเอาค่าความถูกต้อง P และ ค่าความครบถ้วน R ไว้ในค่าเดียว โดยกำหนดให้ชนิดของข้อความที่มีจำนวนมากที่สุดในกลุ่มใดๆ เป็นหัวข้อเรื่อง แสดงดังสมการที่(4.6)

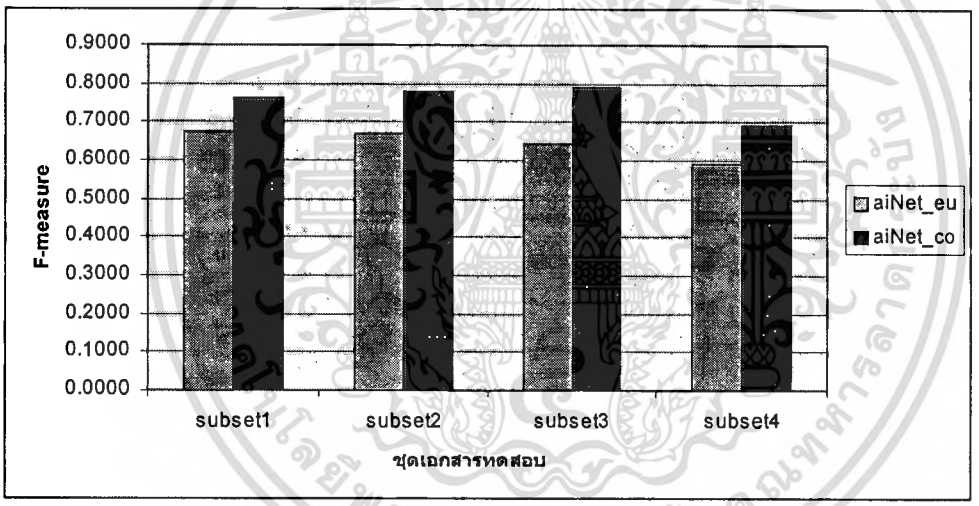
$$P_{i,t} = \frac{\text{จำนวนเอกสารที่เป็นหัวข้อเรื่อง t ในกลุ่ม i}}{\text{จำนวนเอกสารในกลุ่ม i}} \tag{4.4}$$

$$R_{i,t} = \frac{\text{จำนวนเอกสารที่เป็นหัวข้อเรื่อง t ในกลุ่ม i}}{\text{จำนวนเอกสารหัวข้อเรื่อง t ในเอกสาร}} \tag{4.5}$$

$$F_{i,t} = \frac{2(P_{i,t}R_{i,t})}{P_{i,t} + R_{i,t}} \tag{4.6}$$



รูปที่ 4.5 แสดงการเปรียบเทียบผลการทดลองวัดความถูกต้องของการทดลองที่ 4.2



รูปที่ 4.6 แสดงการเปรียบเทียบผลการทดลองวัดค่าเอฟ-เมเชอร์ ของการทดลองที่ 4.2

จากการทดลองจัดกลุ่มข้อมูลโดยนำเอกสารมาจากหลายกลุ่มแต่จำนวนเอกสารในแต่ละกลุ่มเท่ากันผลการทดลองแสดงให้เห็นว่าอัลกอริทึม aiNet ที่วัดค่าแอฟฟินิตีด้วยวิธีการวัดความเหมือนแบบโคไซน์ให้ผลในการจัดกลุ่มดีกว่า aiNet ที่ใช้การคำนวณโดยการวัดระยะทางยูคลิดโดยวัดที่ความถูกต้องแสดงดังรูปที่ 4.5 และวัดด้วยค่าเอฟ-เมเชอร์แสดงดังรูปที่ 4.6 และกราฟยังแสดงถึงผลการจัดกลุ่มของแต่ละชุดข้อมูล ที่ให้ผลต่างกันดังเช่น ชุด subset4 จะได้ค่าความถูกต้องและค่าเอฟ-เมเชอร์ที่ต่ำกว่าทุกชุดข้อมูลทดสอบ ซึ่งแสดงว่าข้อมูลชุดนี้มีความคล้ายคลึงกันมากจึงส่งผลให้การจัดกลุ่มมีคุณภาพต่ำ

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

aiNet เป็นอัลกอริทึมแบบหนึ่งที่มีประสิทธิภาพที่ใช้จัดกลุ่มข้อมูลซึ่งจากการศึกษางานวิจัยต่างๆ ที่ได้ทำการทดลองเปรียบเทียบ aiNet กับอัลกอริทึมแบบอื่นๆ ที่นิยมใช้กันในการจัดกลุ่มข้อมูลผลการเปรียบเทียบของงานวิจัยแสดงถึงความสามารถของ aiNet ว่ามีประสิทธิภาพในการจัดกลุ่มข้อมูลได้ดีกว่า และ aiNet ยังมีคุณสมบัติในการลดขนาดข้อมูลซึ่งจะเห็นได้ว่าแอนติบอดีเมมโมรี่ซึ่งเป็นผลลัพธ์ของ aiNet มีจำนวนแอนติบอดีลดลงไปจำนวนมากเมื่อเทียบกับข้อมูลอินพุทหรือแอนติเจน ซึ่งปัจจัยหลักของอัลกอริทึม aiNet ส่วนหนึ่งก็คือการกำหนดรูปแบบของปัญหาคือ การกำหนดเชป-สเปซ ว่าจะกำหนดเป็นชนิดใดและวิธีการในการคำนวณค่าแอฟฟินิตี ซึ่งเป็นสิ่งสำคัญสำหรับการแยกกลุ่มของเอกสาร โดยที่ค่าแอฟฟินิตีที่มีความแม่นยำสูงในการวัดความคล้ายคลึงเอกสารจะส่งผลให้คุณภาพของการจัดกลุ่มดีขึ้น

ในงานวิจัยนี้ผู้วิจัยได้ประยุกต์ใช้การวัดความคล้ายคลึงโดยใช้การวัดความเหมือนแบบโคไซน์แทนการวัดระยะทางแบบยูคลิดในการคำนวณ ค่าแอฟฟินิตีและใช้ตัวแทนเอกสารในรูปของ real value เนื่องจากการวัดระยะทางแบบยูคลิดของอินพุตที่มีมิติสูงจะทำให้ระยะทางมีค่ามากส่งผลให้มีข้อผิดพลาดเกิดขึ้นจึงเปลี่ยนเป็นการวัดเชิงมุมแทนเพื่อแก้ปัญหาดังกล่าว

จากการทดลองแสดงให้เห็นว่าการจัดกลุ่มโดยอัลกอริทึม aiNet ที่วัดค่าแอฟฟินิตีด้วยวิธีการวัดความเหมือนแบบโคไซน์ให้ผลของความถูกต้องในการจัดกลุ่มดีกว่า และการทดลองได้ทำการเปรียบเทียบให้เห็นว่าในกรณีที่เอกสารหัวข้อเดียวกันแต่มีจำนวนเอกสารในแต่ละกลุ่มไม่เท่ากัน พบว่าการจัดกลุ่มโดยอัลกอริทึม aiNet ที่วัดค่าแอฟฟินิตีด้วยวิธีการวัดความเหมือนแบบโคไซน์ให้ผลดีกว่าในทุกชุดข้อมูลทดสอบ แสดงให้เห็นว่าจำนวนของเอกสารที่ไม่เท่ากันไม่มีผลกับการจัดกลุ่มเอกสารด้วยวิธีนี้ที่จะทำให้คุณภาพการจัดกลุ่มด้อยลงกว่า aiNet ที่วัดค่าแอฟฟินิตีด้วยการวัดระยะทางแบบยูคลิด

สำหรับการทดลองที่เปรียบเทียบเอกสารโดยนำมาจากหลายกลุ่มแต่จำนวนเอกสารในแต่ละกลุ่มเท่ากันผลการทดลองแสดงให้เห็นว่าอัลกอริทึม aiNet ที่วัดค่าแอฟฟินิตีด้วยวิธีการวัดความเหมือนแบบโคไซน์ให้ผลในการจัดกลุ่มดีกว่า โดยการวัดคุณภาพการจัดกลุ่มด้วย ค่าเอฟ-เมเชอร์ และค่าความถูกต้อง และจะสังเกตได้ว่าชุดของเอกสารที่นำมาจัดทดลองในแต่ละชุดมีผลในการวัดคุณภาพต่างกัน เนื่องมาจากแต่ละชุดของเอกสารที่นำมาทดสอบนั้นมีเนื้อหาที่แตกต่างกันหรือบางชุดของเอกสารที่นำมาจัดกลุ่มอาจมีเนื้อหาของข้อความที่ใกล้เคียงกันจึงส่งผลให้ข้อมูลทดสอบแต่ละชุดมีคุณภาพในการจัดกลุ่มต่างกัน ส่งผลให้ค่าเอฟ-เมเชอร์ และความถูกต้องมีค่าแตกต่างกันไป แต่อย่างไรก็ตามวิธีของ aiNet ที่วัดค่าแอฟฟินิตีด้วยการวัดความเหมือนแบบ

โคไซน์ยังให้คุณภาพในการจัดกลุ่มดีกว่า aiNet ที่วัดค่าแอฟฟินิตีด้วยการวัดระยะทางแบบยูคลิด กับทุกชุดของเอกสารที่ใช้ทดสอบ

ซึ่งสรุปได้ว่าในการจัดกลุ่มเอกสารด้วยวิธีที่ใช้ aiNet ที่วัดค่าแอฟฟินิตีด้วยการวัดความเหมือนแบบโคไซน์มีประสิทธิภาพในการจัดกลุ่มเอกสารดีกว่า aiNet ที่วัดค่าแอฟฟินิตีด้วยการวัดระยะทางแบบยูคลิด

5.2 ข้อเสนอแนะ

ในงานวิจัยนี้ผู้วิจัยได้ทำการทดลอง aiNet กับข้อมูลที่เป็นเอกสารเท่านั้นและไม่ได้นำข้อมูลทั้งหมดมาทดสอบโดยสุ่มมาทดสอบบางเอกสารตามจำนวนที่ต้องการเท่านั้น และผลการทดสอบเป็นการนำค่าที่ดีที่สุดของแต่ละวิธีมาซึ่งได้จากการทดสอบหลายๆ ครั้ง ซึ่งควรทดสอบให้มากกว่านี้และทำการหาค่าเฉลี่ยของผลการทดสอบมาใช้ สำหรับข้อมูลในลักษณะอื่นๆ หรือชุดข้อมูลทดสอบอื่นๆ ที่มีการรวบรวมไว้อย่างมีมาตรฐานที่นิยมใช้ในการทดสอบ ผู้วิจัยยังมิได้นำมาศึกษาเปรียบเทียบว่า aiNet ทั้งสองแบบจะให้ผลในการทดสอบอย่างไร ซึ่งในโอกาสต่อไปผู้ทำวิจัยจะได้ทำการศึกษาเพิ่มเติม

และสำหรับปัญหาของการปรับค่าพารามิเตอร์ของ aiNet ในการจัดกลุ่มให้ได้ประสิทธิภาพที่ดีที่สุดนั้นเป็นสิ่งที่ค่อนค้างยากในกรณีที่ข้อมูลมีมิติที่สูงและการเรียนรู้นั้นค่อนข้างใช้เวลานาน ซึ่งในโอกาสต่อไปผู้วิจัยจะทำการศึกษาวิธีการปรับค่าพารามิเตอร์แบบอัตโนมัติเพื่อเพิ่มประสิทธิภาพของ aiNet และปรับปรุงระยะเวลาในการเรียนรู้ของ aiNet ให้เร็วขึ้นอีก และแผนภาพเครือข่ายผลลัพธ์ของ aiNet สำหรับข้อมูลที่มีมิติจำนวนมากนั้นเป็นสิ่งที่ค่อนข้างยากในการแสดงได้อย่างถูกต้องซึ่งผู้ทำวิจัยจะหาแนวทางในการแสดงแผนภาพเครือข่ายผลลัพธ์ของ aiNet สำหรับข้อมูลที่มีมิติจำนวนมากในโอกาสต่อไป

บรรณานุกรม

[1] Yingfeng Cben and Lianying Zhou. "An Innovative IDS immune System Model." IEEE International Conference on Systems, Man and Cybemetics 2004 vol. 5 pp. 4810-4814.

[2] Andraw Watkins and Lois Boggess. "A new classifier based on resource limited artificial immune systems." Proceedings of Congress on Evolutionary Computation, Honolulu, HI, USA, vol. 2 IEEE, May 2002, pp. 1546-1551.

[3] Xiaoshu Hang and Honghua Dai. "An Immune Network Approach for Web Document Clustering." IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), pp. 278-284.

[4] Leandro Nunes de Castro Fermando J. von Zuben. "An Evolutionary Immune Network for data Clustering." IEEE computer society press, SBRN'00 (Brazilian Symposium on neural network), Rio de Janeiiro/RJ, vol. 1, 2000. pp. 84-89,

[5] Lifang Xu, Hongwei Mo, Kejun Wang, and Na Tang. "Document Clustering Based on Modified Artificial Immune Network." Springer-Verlag Berlin Heidelberg. vol. 4062, 2006. pp. 516-524.

[6] Yates Baeza and Neto Ribeiro. **Modern Information Retrieval.** New York:Addison-Wesley, 1999

[7] Salton, G. and J. Allen. "Selective Text Utilization and Text Traversal." Proc. of Hypertext '93' pp. 131-144.

[8] I. Dhillon, J. Korgan, and C. Nicholas. "Feature selection and document clustering." Survey of Text Mining, Springer-Verlag, 2003. pp. 73-100.

[9] Burnet, F.H. **The Clonal Selection Theory of Acquired Immunity.** Cambridge University Press, 1959

[10] Jerne, N.K. **Towards a Network Theory of the Immune System.** Ann. Immunol (inst. Pasteur) 125C, pp.373-389.

[11] de Castro, L. N. and Timmis,J. **Artificial Immune System: A New Computational Intelligence Approach.** Springer-Verlag, 2002.

[12] de Castro, L. N. and Von Zuben, F.J. "The Clonal Selection Algorithm with Engineering Applications." Proc.of GECCO'00, pp.36-37.

[13] Forrest, S. , A. Perelson, Allen, L. and Cherukurl, R. "Self-Nonself Discrimination in a Computer." Proc. of the IEEE Symposium on Research in Security and Privacy, pp. 202-212.

[14] Na Tang and V. Rao Vemuri. "An Artificial Immune System Approach to Document Clustering." Proc. of the Twentieth ACM Sysposium on Applied Computing. Sata Fe, New Mexico, USA 2005. pp.918-922.

[15] Anil K. Jain and Richard C. Dubes. Algorithms for Clustering Data. New Jersey: Prentice Hall. 1988.

[16] Michael Steinbach, George Karypis and Vipin Kunar . "A comparison of Document Clustering Techniques." Technical Report, Department of Computer Science and Engineering University of Minnesota, 2000.

[17] Steinback M., Karypis G., and Kumar V., A Comparison of Document Clustering Techniques, Technical Reprot #00-034, Department of Computer Science and Engineering, University of Minnesota .

[18] Karypis G., Han E., and Kumar V., CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling, Technical Report #99-007, Department of Computer Science and Engineering, University of Minisesota.

[19] "20 newsgroup data set." [Online]. Available : http://people.csail.mit.edu/jrennie/20newsgroups.

[20] I. T. Jolliffe. "Principal Component Analysis." Springer- Verlag, second edition, 2002.

[21] Larsen, B. and C. Anoe. "Fast and Effective Text Mining Using Linear-time Document Clustering." KDD-99, SanDiego, California, 1999. pp. 16-22 .

[22] L.N. De Castro, F. J. Von Zuben. "AiNet: an Artificial Immune Network for Data Analysis." International Journal of Computation Intelligence and Application (IJCIA), vol.1 (3). 2001.

[23] M.F. Porter. "An algorithm for suffix stripping" Program 14(3) 1980. pp. 130-137

[24] Han, J. and M. Kamber . "Data mining: Concepts and techniques." San Diego: Academic Press. 2001

ภาคผนวก

ภาคผนวก ก.

บทความและผลงานวิจัยที่ได้รับการตีพิมพ์

1. บัณฑิต ปุญญวัฒนะ และบุญวัฒน์ อัตชู. "การจัดกลุ่มเอกสารโดยใช้เครือข่ายภูมิคุ้มกันเทียม." วิศวสารลาดกระบัง, ปีที่ 27, ฉบับที่ 1, 2553.
2. บัณฑิต ปุญญวัฒนะ และบุญวัฒน์ อัตชู. "การเพิ่มความเร็ว SOM ในการจัดกลุ่มข้อมูลด้วยเทคนิคการสุ่มแบบเบี่ยงเบนตามความหนาแน่นและความเป็นปึกแผ่นของข้อมูล." วิศวสารลาดกระบัง, ปีที่ 27, ฉบับที่ 2, 2553.

การจัดกลุ่มเอกสารโดยใช้เครือข่ายภูมิคุ้มกันเทียม
Document Clustering Using Artificial Immune Network

บัณฑิต ปุญญวัฒนะ บุญวัฒน์ อัตชู

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

ในปัจจุบันมีการนำการทำงานของระบบภูมิคุ้มกันมาประยุกต์ใช้แก้ปัญหาต่างๆ ในด้านการเรียนรู้ด้วยคอมพิวเตอร์ บทความนี้ได้นำเสนอการใช้ aiNet (Artificial Immune Network) ซึ่งเป็นอัลกอริธึมแบบหนึ่งของการทำงานของระบบภูมิคุ้มกันเพื่อใช้จัดกลุ่มเอกสาร การทำงานของ aiNet จะมีการคำนวณค่า affinity โดยทั่วไปใช้การวัดระยะทางแบบ Euclidean distance กับข้อมูลที่เป็นค่าแบบ real value สำหรับบทความนี้ได้มีการปรับปรุงโดยนำวิธีการวัดความคล้ายคลึงของเอกสารโดยใช้ค่าสัมประสิทธิ์โคไซน์มาคำนวณค่า affinity ของ aiNet แทนการใช้การวัดระยะทางแบบ Euclidean distance โดยทดลองกับเอกสารที่ถูกจัดกลุ่มไว้แล้ว ซึ่งผลที่ได้แสดงให้เห็นว่าวิธีการที่นำเสนอมีประสิทธิภาพในการจัดกลุ่มเอกสารดีกว่า

คำสำคัญ: เครือข่ายภูมิคุ้มกันเทียม, การจัดกลุ่มเอกสาร, ค่าสัมประสิทธิ์โคไซน์

Abstract

It has recently been shown that Artificial Immune Network (aiNet) provides inspiration for solving a wide range of machine learning problems. In this paper we propose the application of aiNet for document clustering. Traditional aiNet algorithm determines the affinity of real value data set by using Euclidean Distance. Cosine Similarity is used to determine the affinity instead of Euclidean Distance in this paper. The experiment results show that our proposed technique gets better results.

Key words: Artificial Immune Network, Document Clustering, Cosine Similarity

1.บทนำ

ในปัจจุบันเทคโนโลยีสารสนเทศมีข้อมูลข่าวสารปริมาณมากขึ้น ดังเช่น ในอินเตอร์เน็ตมีเอกสารมากมายในรูปแบบต่างๆ เทคนิคในการจัดลำดับความสำคัญของเอกสาร (Ranking) ไม่เพียงพอในการที่จะเพิ่มประสิทธิภาพของการค้นคืน การจัดกลุ่มเอกสาร (Document Clustering) การกรองสารสนเทศ (Information Filtering) หรือการกลั่นกรองเอกสาร (Information Extraction) เข้ามามีบทบาทในการช่วยค้นคืนสารสนเทศให้กับผู้ใช้มากขึ้น การจัดกลุ่มเอกสารมีจุดประสงค์คือแยกเอกสารออกเป็นกลุ่มตามความคล้ายคลึงและความสัมพันธ์กันซึ่งขึ้นอยู่กับข้อความที่ปรากฏในเอกสารแต่ละฉบับ และพยายามคิดค้นวิธีที่จะจัดกลุ่มเอกสารปริมาณมากๆ แบบอัตโนมัติโดยมีงานวิจัยพัฒนาขั้นตอนและวิธีการในการจัดกลุ่มเอกสารออกมาอย่างต่อเนื่อง

ในช่วงไม่กี่ปีมานี้ระบบภูมิคุ้มกันทำให้เกิดแนวความคิดมากมายสำหรับวิธีการแก้ปัญหาที่

เปลี่ยนแปลงไป ระบบภูมิคุ้มกันได้ถูกนำมาประยุกต์ใช้ในด้านต่างๆ เช่น Anomaly Detection[1], Pattern Recognition[2]. Web Document Classification[3],Data Clustering[4,5] เป็นต้น โดย aiNet เป็นอัลกอริธึมหนึ่งที่มีการทำงานคล้ายกับ Immune Network ซึ่งนำมาประยุกต์ใช้ในการลดข้อมูลที่ซ้ำซ้อนและจัดกลุ่มเอกสาร หลักสำคัญในการจัดกลุ่มเอกสารโดย aiNet คือ การคำนวณค่า affinity ระหว่าง system unit และ input data โดยปกติ aiNet นั้นใช้ Euclidean distance ในการคำนวณค่า affinity กับข้อมูลแบบ real value ปัญหาก็คือเมื่อ input ซึ่งเป็นเวคเตอร์มีขนาดใหญ่ทำให้ระยะห่างระหว่างเวคเตอร์ทั้งสองมีค่ามากส่งผลให้การจัดกลุ่มเอกสารเกิดข้อผิดพลาดมาก

ดังนั้นบทความนี้ได้นำเสนอการปรับปรุง aiNet โดยใช้การคำนวณค่า affinity ด้วยค่าสัมประสิทธิ์โคไซน์ (Cosine Similarity) [6]ซึ่งเป็นวิธีที่นิยมใช้ในการวัดความคล้ายคลึงของเอกสารโดยวัดมุมระหว่างเวคเตอร์แทนเพื่อลดข้อผิดพลาด และได้ทำการทดลองจัดกลุ่มข้อมูลโดยเปรียบเทียบผลของทั้งสองวิธีโดยวัดที่ความถูกต้องและค่า F-measure

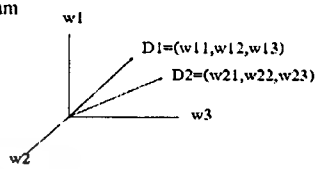
2. ทฤษฎีที่เกี่ยวข้อง

2.1 ความรู้เบื้องต้นระบบค้นคืนเอกสาร

การจำลองแบบเชิงแนวคิดของระบบค้นคืนสารสนเทศสามารถจำแนกได้เป็น 3 แบบ คือ แบบจำลองทางบูลีน (Boolean Model) แบบจำลองทางสถิติ (Statistic Model) และแบบจำลองทางเวคเตอร์ (Vector Model) บทความฉบับนี้จะกล่าวเฉพาะแบบจำลองแบบเวคเตอร์ ซึ่งเป็นแบบจำลองที่นิยมใช้ในการจัดกลุ่มเอกสารเพราะแบบจำลองดังกล่าวแทนเอกสารแต่ละฉบับโดยแต่ละมิติของเวคเตอร์จะแทนคำที่ปรากฏในเอกสาร กรรมวิธีในการเลือกคำที่จะมาเป็นตัวแทนของเอกสาร โดยมีหลักเบื้องต้นดังนี้

1.การหาคำหยุด (Stopwords) คำหยุดเป็นคำที่เกิดในเอกสารทุกฉบับและเกิดเป็นปริมาณมากๆทำให้ไม่สามารถใช้เป็นคำในการจำแนกเอกสารได้ต้องกำจัดออก

2.การหารากศัพท์ (Stemming) เป็นการหารูปเดิมของคำหรือหาคำที่มีความหมายคล้ายกันเพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ สำหรับบทความฉบับนี้ใช้วิธีการหารากศัพท์ด้วย Porter Algorithm



รูปที่ 1 แสดงการเก็บเอกสารด้วยแบบจำลองแบบเวคเตอร์ (Vector Model) ใน 3-มิติ

วิธีการให้น้ำหนักของคำที่ใช้กันอย่างมากในการสืบค้นข้อมูล โดยคิดน้ำหนักคำจากค่าผลคูณของ tf(term frequency) ซึ่งเป็นความถี่ของคำที่ปรากฏในเอกสารและค่า idf(inverse document frequency) คำนวณจากค่า log(N/df) ซึ่ง N คือจำนวนเอกสารในชุดเอกสารทั้งหมด และ df (document frequency) คือจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ วิธีให้น้ำหนักของคำใน[7] มีการ normalization ทำให้เวคเตอร์เอกสารมีขนาด 1 หน่วยมีสูตรดังนี้

$$W_{ij}=\frac{tfik*\log(N/dfk)}{\sqrt{\sum_{j=1}^{t}(tfij)^2*(\log(N/df_j))^2}} \quad (1)$$

โดยที่ tf_{ik} คือความถี่ของคำในเอกสาร i, N คือจำนวนของเอกสารในชุดเอกสาร, df_k คือจำนวนเอกสารในชุดเอกสารซึ่งบรรจุคำ k เมื่อผ่านกระบวนการทั้งหมดแล้วจะได้เอกสารที่ถูกแทนอยู่ในรูปของ

$D_j=\{w_{1j},w_{2j},w_{3j},\ldots,w_{tj}\}$ โดยที่ $w_{ij}\geq 0$

ตัวอย่าง เช่น ในชุดเอกสารหนึ่งประกอบด้วยเอกสาร D1, D2 , D3 นำเอกสารแต่ละฉบับมาตัดคำ (word segmentation) ตัดคำหยุดออกไปและหารากศัพท์ก็จะได้เอกสารตามรูปที่ 2 จากนั้นหาความถี่ของคำที่ไม่ซ้ำกันในเอกสารแต่ละฉบับจะได้ดังตารางที่ 1 แล้วทำการหา

ค่า df และ idf ให้แต่ละคำจะได้ดังตารางที่ 2 และทำการหาเวคเตอร์เอกสารทั้งหมดโดยที่แถวของเมทริกซ์คือเอกสารทั้งหมด และสดมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร ถ้าคำในสดมภ์ปรากฏอยู่ในเวคเตอร์เอกสารฉบับหนึ่งๆ สามารถหาค่าน้ำหนักได้ตามสมการ (1) แต่ถ้าไม่ปรากฏคำนั้นในเอกสารที่กำลังพิจารณาอยู่ก็ให้คำนั้นมีค่าน้ำหนักเป็น 0 จากเอกสารที่ไม่มีโครงสร้างก็จะถูกแทนเป็นระบบด้วยเวคเตอร์ซึ่งอยู่ในรูปของเมทริกซ์เอกสาร-คำดังรูปที่ 3 ซึ่งใช้เป็น input ในขั้นตอนการจัดกลุ่มเอกสารต่อไป

D1: computer information computer computer

D2: internet computer internet data

D3: system internet

รูปที่ 2 ชุดเอกสาร

ตารางที่ 1 ความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	freq
D1	computer	3
D1	information	1
D2	Internet	2
D2	computer	1
D2	data	1
D3	system	1
D3	Internet	1

ตารางที่ 2 ค่า idf ของคำในชุดเอกสาร

คำ	Df	idf
T1: computer	2	0.18
T2: information	1	0.48
T3: internet	2	0.18
T4: system	1	0.48
T5: data	1	0.48

	T1	T2	T3	T4	T5
D1	0.74	0.67	0	0	0
D2	0.57	0	0.29	0	0.77
D3	0	0	0.35	0.94	0

รูปที่ 3 เมทริกซ์เอกสาร-คำ

2.2 เอกสารในการจัดกลุ่ม

เมื่อได้เมทริกซ์เอกสาร-คำตาม(2.1)แล้ว ให้ L คือจำนวนของคำที่ปรากฏในเอกสาร L มีมิติขนาดใหญ่ การทำให้ L มีขนาดเล็กลงเพื่อให้มีมิติ *l* ที่ดีสำหรับการจัดกลุ่มเอกสาร โดย *l* << L เพื่อลดเวลาในการคำนวณและให้การใช้ทรัพยากรน้อย โดยใช้วิธีที่เรียกว่า Feature Selection[8] ซึ่งมีสมการดัง(2)

$$q(w)=\sum_{i=1}^{N} f_i^2 - \frac{1}{N}\left[\sum_{i=1}^{N} f_i\right]^2 \qquad (2)$$

เมื่อ f_i ความถี่ของคำ w ในเอกสาร d_i N เป็นจำนวนทั้งหมดของเอกสาร ซึ่งวิธีนี้เราเลือกมา 15% ของคำในเอกสารที่ใช้ทดลอง ซึ่งยังให้ผลการจัดกลุ่มเหมือนเดิมหลังจาก Feature Selection แล้วใช้ Principal Component (PCA) [9] เพื่อลดมิติของ weight vector ที่ได้จาก Feature Selection โดยลดมิติให้น้อยลง และใช้เพียง 20 dimensional เป็น Input ของ aiNet

2.3 ระบบภูมิคุ้มกัน immune System

ระบบภูมิคุ้มกัน immune System เป็นระบบที่ซับซ้อนอันหนึ่งของเซลล์มีจุดประสงค์ในการป้องกันร่างกายโดยเกิดขึ้นเมื่อมีเชื้อโรคหรือสิ่งแปลกปลอม (antigen) เข้ามาจะมีเซลล์ไปทำความรู้จักกับเชื้อโรคแล้วบรรจุข้อมูลส่งไปให้เซลล์ที่มีหน้าที่สร้างสารต่อต้านคือ B-lymphocyte เป็นเซลล์ที่กำเนิดและเจริญที่ Bone marrow ร่างกายจะตอบสนองต่อแอนติเจน โดยสร้างสารภูมิคุ้มกัน(antibody) ขึ้นมาต่อต้านแบบจำเพาะเจาะจงต่อแอนติเจนนั้นๆ หรือเรียกว่า Humoral immune response ซึ่ง B-cell จะถูกกระตุ้นให้เปลี่ยนเป็น Plasma cell ทำหน้าที่สร้างแอนติบอดี้ที่อยู่ในเลือด หรือน้ำเหลือง ซึ่งเป็นการทำลายแอนติเจนที่อยู่ในเลือดหรือน้ำเหลือง ส่วนแอนติเจนที่อยู่นอกเซลล์หรือเรียกว่า Extra cellular pathogen B-cell จะมีความสามารถในการสร้างแอนติบอดี้ได้เพียงแบบเดียวอย่างจำเพาะเจาะจง โดยแอนติบอดี้ที่สร้างขึ้นจะปรากฏอยู่บนผิวเซลล์ ทำหน้าที่เป็น receptor สำหรับจับกับแอนติเจนที่จำเพาะ เมื่อ แอนติเจน เข้าสู่ร่างกาย B-cell ที่มีแอนติบอดี้ที่เหมาะสมหรือเข้ากันได้กับแอนติเจนจะกระตุ้นให้แอนติบอดี้แบ่งตัว(Clonal expansion) เป็นกลุ่ม plasma cell หรือ effector cell ที่จะสร้าง แอนติบอดี้ แบบเดียวกันกับแอนติเจนที่รุกรานนั้นและ B-cell บางส่วนจะกลายเป็น Memory B-cell เพื่อที่ว่าในเวลาต่อมาถ้าร่างกายได้รับแอนติเจนตัวเดิมอีก B-cell ตัวนั้นก็จะแบ่งตัวเพิ่ม

จำนวนมากขึ้นอย่างรวดเร็ว และผลิตแอนติบอดี้เพื่อต่อต้านแอนติเจนเดิมนั้นได้ทันท่วงที แอนติบอดี้แต่ละชนิดจะมีอายุไม่เท่ากัน บางชนิดก็อยู่ได้ไม่นาน บางชนิดก็อยู่ได้หลายปี บางชนิดก็อยู่ได้ตลอดชีวิต

2.4 aiNet Algorithm

aiNet (Artificail Immune Network)[10] เสนอขึ้นในปี 2000 โดย de Castra and Von Zuben ซึ่งเป็นอัลกอริธึมแบบหนึ่งที่ทำงานเลียนแบบกับระบบภูมิคุ้มกันของสิ่งมีชีวิต Network กำหนดขึ้นโดยการสุ่มซึ่งก็คือกลุ่มของแอนติบอดี้ จะสามารถอยู่ใน Network ได้โดยค่า affinity ระหว่างแอนติบอดี้กับแอนติเจน แอนติบอดี้ใดมีค่า affinity สูงแอนติบอดี้ก็จะถูกเลือกและเกิดการแบ่งตัวเพิ่มจำนวนมากขึ้น (Clonal selection) และทำการคำนวณค่า affinity ระหว่างแอนติบอดี้ที่เกิดใหม่กับแอนติเจน โดยแอนติบอดี้ใดที่มีค่า affinity ที่สูงจะถูกเลือกไปยัง Network เพื่อกำหนดเป็น Clonal memory แอนติบอดี้ที่เหลือนั้นจะถูกกำจัดถ้ามีค่า affinity มีค่าต่ำกว่า threshold (Clonal suppression) ที่กำหนด ซึ่งอัลกอริธึม aiNet แบบเดิมใช้การคำนวณค่า affinity โดยใช้ Euclidean distance สมการ(3) กับข้อมูลชนิด real value

$$d(D_i,D_j)=\sqrt{\sum_{k=1}^{n}(w_{d\,ik}-w_{d\,jk})^2} \qquad (3)$$

ส่วนในบทความนี้ได้ทำการปรับปรุงการคำนวณค่า affinity โดยใช้ค่าสัมประสิทธิ์โคไซน์ (Cosine Similarity) สมการ (4) แทน

$$Sim(D_i,D_j)=\frac{\sum_{k=1}^{l} w_{d\,ik} * w_{d\,jk}}{\sqrt{\sum_{k=1}^{l}(w_{d\,ik})^2 * \sum_{k=1}^{l}(w_{d\,jk})^2}} \qquad (4)$$

อัลกอริธึมของ aiNet ได้แสดงในรูปที่ 4 กำหนดอัตราการ mutation ในการ clone ดังสมการ (5)

$$C_k^* = C_k + \alpha_k(Ag_j - Ck) \;,\; \alpha_k \propto 1/f_{ij},\; k = 1,...,Nc,\; i = 1,...,N \qquad (5)$$

และจำนวนของ การ clone ของแอนติบอดี้สำหรับแต่ละแอนติเจนได้แสดงดังสมการ (6)

$$NC = \sum_{i=1}^{n} round(N - Di,jN) \qquad (6)$$

Ab : available antibody repertoire ($Ab \in S^{Ab}$,$Ab = b_d \cup Ab_m$);
Ab_m : total memory antibodyset($Ab_m \in S^{Abd}$,$m \leq N$);
Ab_d : d new antibodies to be inserted in $Ab(Ab_d \in S^{Abd})$;
Ag : population of antigens($Ag \in S^{Ab}$);
f_j : vectors containing the affinity of all the antibodies Ab_i with relation to antigen Ag_j.i, j = 1, ...,N;
S : similarity matrix between each pair Ab_i-Ab_j,with element s_{ij}(i, j = 1, ..,N);
C : population of clones generated from Ab($C \in S^{Ncu}$);
C* : population C after the affinity maturation process;
d_j : vector containing the affinity between every element from the set C* with Agj :
σ_s : the suppression threshold ,which defines the threshold to eliminate redundant Abs. ζ:the percentage of reselected Abs ;
σd : the death rate,which defined the threshold to remove the low-affinity Abs after the reselection.

Algorithm 1. Document Clustering by Modified aiNet
Input : Feature vectors of documents *Ag*
Output: Number of document clustering *N*.
Initialize *Ab* = []; Convert *n* Ags documents into *n* Ags via document representation and feature selection; Randomly generate *k* Abs and put them into *Ab*;
for *each iteration* do
  for *each* Ag_j, $j = 1, ...M$,$Ag_j \in Ag$ do
    Calculate f_{ij}, $i=1,....N$ to all
    $Abi,fi,j = D_{ij}$, $i = 1, ...N$,D_{ij} = affinity(Abi, Ag_j), $i = 1, ...N$;
    Select Ab_n composed of *n* highest affinity antibodies
    Clone the n selected antibodies according to (5),generating *C*
      C is submitted to process of affinity maturation process according to (6),generating *C**
    Calculate $d_{kj} = D_{ij}$ among Ag_j and all the elements of C^*,D_{kj} = affinity (C^*_k, Ag_j) $k = 1, ...,Nc$
    Reselect a subset ζ% of the antibodies with highest d_{kj} and put them into M_j as memory clones;
    Remove the memory clones from M_j whose $D_{kj} > \sigma d$
    Determine s_{ik} among the memory clones:s_{ik} =affinity ($M_{j,i}$,M_{jk}). $\forall i, k$
    Eliminate those memory clone whose $s_{ik} < \sigma s$
    Concatenate the total antibody memory matrix with resultant clonal memory M^*_j :$Abm \leftarrow [Abm;M^*_j]$
  end
  Calculate s_{ik} = affinity (Ab^i_m, Ab^i_m). $\forall i, k$;
  Eliminate all the antibodies whose $si,k < \sigma s$;
  $Ab \leftarrow [Abm; Abd]$;
end
Cluster *M* which contains *n Abs* via K-means;
Check the *Ags* of each *Ab* in *M* to obtain each Ag's cluster.

รูปที่ 4 แสดงอัลกอริธึม aiNet

3. การทดลอง

3.1 ข้อมูลสำหรับการทดลอง

เอกสารที่ใช้ในการทดลองคือ 20 Newgroup data set [11] ซึ่งนำมาจาก http://people.csail.mit.edu/jrennie/20Newsgroups ประกอบด้วยเอกสารจำนวน 20000 เอกสารที่มีการจัดกลุ่มหัวข้อที่ต่างกัน 20 หัวข้อ ข้อมูลทดสอบชุดแรกเป็นการทดสอบความถูกต้องโดยเลือกเอกสารจาก 2 หัวข้อคือ sci.crypt และ sci.electronics โดย Subset A เลือกเอกสารโดยการสุ่มมากลุ่มละ 80 เอกสาร Subset B เลือกมากลุ่มละ 150 เอกสาร Subset C เลือกมากลุ่มละ 300 เอกสาร ส่วนข้อมูลที่ใช้ทดสอบชุดที่ 2 เป็นการทดสอบความถูกต้องและ F-measure มีข้อมูล 4 กลุ่ม นำมาโดยวิธีการสุ่มมาจากเอกสารแต่ละหัวข้อดังแสดงในตารางที่ 3

ตารางที่ 3 แสดงข้อมูลทดสอบชุดที่ 2

Dataset	Topic	Included per Group #docs	Total #docs
subset 1	sci.crypt, sci.space	150, 150	300
subset 2	sci.crypt, sci.electronics	150, 150	300
subset 3	sci.space, rec.sports.basketball	150, 150	300
subset 4	talk.polities.mideast.talk .polities.misc	150, 150	300

3.2 ขั้นตอนวิธีการทดลองจัดกลุ่มโดยใช้ aiNet

นำข้อมูลแต่ละกลุ่มมาทำการตัดคำตามหัวข้อที่ 2.1,2.2 โดยรูปแบบเอกสารแต่ละฉบับจะอยู่ในรูปของเวกเตอร์ที่มีมิติเป็น l มิติต่างๆ ก็คือเซตของคำภายในเอกสาร เมื่อมีเอกสารจำนวน N เอกสารก็จะเป็นเมตริกซ์ $N \times l$ โดยค่าที่อยู่ในเมตริกซ์ก็คือค่าของน้ำหนักของคำที่ปรากฏในแต่ละเอกสาร โดยจะเป็นค่าแบบ real value และเวกเตอร์ของเอกสารก็คือ กลุ่มของแอนติเจนที่ใช้เรียนรู้ภายในอัลกอริธึมของ aiNet โดยทำการปรับค่าพารามิเตอร์ $(\sigma_d, \sigma_s, \zeta\%, iteration)$ ของ aiNet ที่วัดค่า affinity แบบEuclidean และแบบ Cosine เพื่อจัดกลุ่มข้อมูล โดยข้อมูลชุดที่ 1 SubsetA[0.4,0.12,0.2,10 (Euclidean) ,0.6,0.9,0.2,10(Cosine)] กับข้อมูลชุดอื่นๆเป็นดังนี้ SubsetB[0.32.0.1,0.2.10(Euclidean).0.6,0.65,0.2 .10(Cosine)] SubsetC[0.4,0.12,0.2,10(Euclidean),0.6.0.9, 0.2, 10(Cosine)] เมื่อสิ้นสุดการทำงานของ aiNet กับข้อมูลแล้วจะได้กลุ่มของแอนติบอดี้กลุ่มหนึ่งใน Memory cell แล้วนำมาจัดกลุ่มและนำแอนติเจนมาตรวจสอบว่าอยู่ในกลุ่มใดประเมินผลตามสมการ(7) สำหรับค่าพารามิเตอร์ที่ใช้ในการทดลองของ aiNet กับข้อมูลชุดที่ 2 ซึ่งเป็นการตรวจสอบความถูกต้องและประสิทธิผลโดยปรับค่าพารามิเตอร์ในการทดลองดังนี้ Subset1[0.4,0.12,0.2,10(Euclidean),0.6,0.9,0.2,10(Cosine)],Subset2[0.4,0.12,0.2,10(Euclidean),0.6,0.9,0.2,10(Cosine)],Subset3[0.4,0.12,0.2,10(Euclidean),0.6,0.9,0.2,10(Cosine)],Subset4[0.4,0.12,0.2,10(Euclidean),0.6,0.9,0.2,10(Cosine)] เมื่อได้ผลการทดลองแล้วก็ทำการประเมินผลการจัดกลุ่มเอกสารตามสมการ(7,10)

การวัดผลของการจัดกลุ่มเอกสารโดยการวัดประสิทธิภาพของความถูกต้องข้อมูลโดยวัดค่าความถูกต้องของการจัดกลุ่มเอกสารดังนี้

$$\text{ความถูกต้อง} = \frac{\text{จำนวนสมาชิกที่ถูกต้องทั้งหมด}}{\text{จำนวนสมาชิกทั้งหมดในกลุ่ม}} \quad (7)$$

และวัดค่าประสิทธิผล F-measure [12] ซึ่งเป็นค่าที่ รวมเอาค่าความแม่นยำ (Precision: P) และค่าความระลึก (Recall: R) ไว้ในค่าเดียว เรากำหนดให้ขนิดของข้อความที่มีจำนวนมากที่สุดในกลุ่มใดๆ เป็นหัวข้อเรื่อง (topic) มีสูตรดังนี้

$$P_{i,t} = \frac{\text{จำนวนเอกสารที่เป็นหัวข้อเรื่อง } t \text{ ในกลุ่ม } i}{\text{จำนวนเอกสารในกลุ่ม } i} \quad (8)$$

$$R_{i,t} = \frac{\text{จำนวนเอกสารที่เป็นหัวข้อเรื่อง } t \text{ ในกลุ่ม } i}{\text{จำนวนเอกสารหัวข้อเรื่อง } t \text{ ในเอกสาร}} \quad (9)$$

$$F_{i,t} = \frac{2(P_{i,t}R_{i,t})}{P_{i,t} + R_{i,t}} \quad (10)$$

3.3 ผลการทดลอง

ผลการทดลองการจัดกลุ่มเอกสารโดยการใช้ aiNet โดยวิธีการประยุกต์การใช้ค่าค่าสัมประสิทธิ์โคไซน์ระหว่างคู่ของเวกเตอร์เอกสารใดๆ แทนการใช้ Euclidean

distance แสดงได้ดัง ตารางที่ 4 เปรียบเทียบผลการทดลองวัดที่ความถูกต้องและจำนวนเอกสาร ตารางที่ 5 เปรียบเทียบผลการทดลองที่วัดความถูกต้องและประสิทธิผล(F-measure)

ตารางที่ 4 แสดงผลการทดลองวัดความถูกต้อง

Algorithms	Document		
	SubsetA 160	SubsetB 300	SubsetC 600
aiNet_eu	0.6000	0.5933	0.5833
aiNet_co	0.6937	0.6400	0.7166

ตารางที่ 5 แสดงผลการทดลองวัดความถูกต้องและประสิทธิผล (F-measure)

Algorithms	subset1		subset2		subset3		subset4	
	Acc.	F-mea	Acc.	F-mea	Acc.	F-mea	Acc.	F-mea
aiNet_eu	0.6100	0.5428	0.5200	0.5133	0.6966	0.6728	0.6100	0.5832
aiNet_co	0.7266	0.7265	0.7233	0.7210	0.7900	0.7953	0.7000	0.6986

ผลการทดลองของวิธีที่นำเสนอในการวัดความถูกต้องสำหรับข้อมูลชุด SubsetA ได้ค่า 0.6937, ข้อมูลชุด SubsetB ได้ค่า 0.6400 ข้อมูลชุด SubsetC ได้ค่า 0.7166 ซึ่งให้ค่าสูงกว่า ส่วนการวัดความถูกต้องและประสิทธิผล (F-measure) ข้อมูล Subset1 ได้ค่า 0.7266,0.7265 ข้อมูล Subset2 ได้ค่า 0.7233,0.7210 ข้อมูลSubset3 ได้ค่า 0.7900,0.7953 ข้อมูล Subset4 ได้ค่า 0.7000,0.6986 ซึ่งมีค่าความถูกต้องและค่า F-measure ดีกว่า aiNet ที่ใช้การคำนวณค่าโดย Euclidean distance

4. สรุป

การจัดกลุ่มเอกสารโดยใช้ aiNet อัลกอริทึมโดยใช้ค่าสัมประสิทธิ์โคไซน์ คำนวณค่า affinity ระหว่างคู่ของเวกเตอร์ใดๆ แทนการใช้ Euclidean distance สามารถแก้ปัญหาของระยะห่างระหว่างคู่ของเวกเตอร์มีค่ามาก เนื่องจากเวคเตอร์มีขนาดใหญ่ที่ส่งผลกระทบในการจัดกลุ่มข้อมูล โดยทดลองกับเอกสารที่มีเนื้อหาในด้านต่างๆ ซึ่งถูกจัดกลุ่มไว้เรียบร้อยแล้ว ผลการทดลองแสดงให้เห็นว่าความถูกต้องและประสิทธิผล F-measure ที่ได้จากการจัดกลุ่มเอกสารโดยใช้วิธีการของ aiNet ที่มีการปรับปรุงการคำนวณค่า affinity สามารถจัดกลุ่มเอกสารได้ผลลัพธ์ที่ดีกว่าทำให้การจัดกลุ่มเอกสารมีประสิทธิภาพมากขึ้น

5. เอกสารอ้างอิง

[1] Yingfeng Cben and Lianying Zhou, "An Innovative IDS immune System Model", Proceedings of the IEEE International Conference on Systems, Man and Cybernetics 2004 vol. 5 pp.4810-4814

[2] Andraw Watkins and Lois Boggess, "A new classifier based on resource limited artificial immune systems", Proceedings of Congress on Evolutionary Computation, Honolulu, HI, USA, vol.2 pp. 1546-1551., IEEE, May 2002.

[3] Xiaoshu Hang and Honghua Dai, "An Immune Network Approach for Web Document Clustering", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04),pp.278-284

[4] Leandro Nunes de Castro Fernando J. von Zuben, "An Evolutionary Immune Network for data Clustering", Proceedings of the IEEE computer society press, SBRN'00 (Brazilian Symposium on neural network), vol.1, pp. 84-89,Rio de Janeiro/RJ, 22-25,Nov., 2000.

[5] Lifang Xu, Hongwei Mo, Kejun Wang, and Na Tang, "Document Clustering Based on Modified Artificial Immune Network", Springer-Verlag Berlin Heidelberg, vol. 4062 pp. 516-524., 2006.

[6] Yates Baeza and Neto Ribeiro, Modern Information Retrieval, Addison-Wesley, 1999

[7] Salton, G. and J. Allen, "Selective Text Utilization and Text Traversal", Proceedings of Hypertext '93' pp. 131-144.

[8] I. Dhillon, J. Korgan, and C. Nicholas, "Feature selection and document clustering", Survey of Text Mining, Springer-Verlag, pp. 73-100, 2003.

[9] I. T. Jolliffe. "Principal Component Analysis", Springer- Verlag, second edition, 2002.

[10] De Castro, L.N and Von Zuben, F. (2001), "aiNet: An Artificial Immune Network for Data Analysis", in Data Mining: A Heuristic Approach. Abbas, H, Sarker, R and Newton. C(Eds). Idea Group Publishing. USA, Chapter XII , pp. 231-259

[11] 20 newsgroup data set. http://people.csail.mit.edu/jrennie/20newsgroups.

[12] Larsen, B. and C. Anoe, "Fast and Effective Text Mining Using Linear-time Document Clustering", KDD-99, SanDiego, California, pp. 16-22 ,1999.

ประวัติผู้เขียน

นายบัณฑิต ปุญญวัฒนะ เกิดเมื่อวันที่ 14 ตุลาคม พ.ศ.2513 ที่จังหวัดชัยนาท สำเร็จการศึกษาปริญญาตรีวิทยาศาสตร์บัณฑิต สาขาศาสตร์คอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง ในปีการศึกษา 2536 และเข้าศึกษาต่อในระดับปริญญาโท หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในปีการศึกษา 2548