การจัดกลุ่มเอกสารโดยใช้เครือข่ายภูมิคุ้มกันเทียม
Document Clustering Using Artificial Immune Network

บัณฑิต ปุญญวัฒนะ บุญวัฒน์ อัตชู

สาขาวิชาวิศวกรรมคอมพิวเตอร์ 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

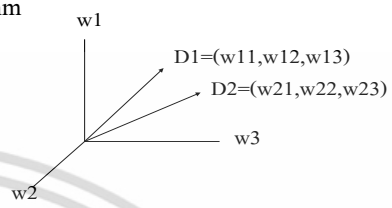
2. ทฤษฎีที่เกี่ยวข้อง

2.1 ความรู้เบื้องต้นระบบค้นคืนเอกสาร

การจำลองแบบเชิงแนวคิดของระบบค้นคืนสารสนเทศสามารถจำแนกได้เป็น 3 แบบ คือ แบบจำลองทางบูลีน (Boolean Model) แบบจำลองทางสถิติ (Statistic Model) และแบบจำลองทางเวคเตอร์ (Vector Model) บทความฉบับนี้จะกล่าวเฉพาะแบบจำลองแบบเวคเตอร์ซึ่งเป็นแบบจำลองที่นิยมใช้ในการจัดกลุ่มเอกสารเพราะแบบจำลองดังกล่าวแทนเอกสารแต่ละฉบับโดยแต่ละมิติของเวคเตอร์จะแทนคำที่ปรากฏในเอกสาร 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 คือจำนวนเอกสารในชุดเอกสารทั้งหมดและ df (document frequency) คือจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ วิธีให้น้ำหนักของคำใน[7] มีการ normalization ทำให้เวคเตอร์เอกสารมีขนาด 1 หน่วยมีสูตรดังนี้

$$W_{i,j} = \frac{tf_{ik} * \log(N / df_k)}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 * (\log(N / df_j))^2}} \qquad (1)$$

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

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

ตัวอย่าง เช่น ในชุดเอกสารหนึ่งประกอบด้วยเอกสาร D1, D2 , D3 นำเอกสารแต่ละฉบับมาตัดคำ (word segmentation) ดึงคำหยุดออกไปและหารากศัพท์ก็จะได้เอกสารตามรูปที่ 2 จากนั้นหาความถี่ของคำที่ไม่

ซ้ำกันในเอกสารแต่ละฉบับจะได้ดังตารางที่ 1 แล้วทำการหาค่า df และ idf ให้แต่ละคำจะได้ดังตารางที่ 2 และทำการหาเวคเตอร์เอกสารทั้งหมดโดยที่แถวของเมทริกซ์คือเอกสารทั้งหมด และสดมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร ถ้าคำในสดมภ์ปรากฏอยู่ในเวคเตอร์เอกสารฉบับหนึ่งๆ สามารถหาค่าน้ำหนักได้ตามสมการ (1) แต่ถ้าไม่ปรากฏคำนั้นในเอกสารที่กำลังพิจารณาอยู่ก็ให้คำนั้นมีค่าน้ำหนักเป็น 0 จากเอกสารที่ไม่มีโครงสร้างก็จะถูกแทนเป็นระบบด้วยเวคเตอร์ ซึ่งอยู่ในรูปของเมทริกซ์เอกสาร-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)

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

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 เพื่อที่ว่าในเวลาต่อมาถ้าร่างกาย

ได้รับแอนติเจนตัวเดิมอีก B-cell ตัวนั้นก็จะแบ่งตัวเพิ่มจำนวนมากขึ้นอย่างรวดเร็ว 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

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

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

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

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

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

และจำนวนของ การ clone ของแอนติบอดี้สำหรับแต่ละแอนติเจนได้แสดงดังสมการ (6)

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

Ab : available antibody repertoire ($Ab \in S^{Nxl}, Ab = b_d \cup Ab_m$);
Ab_m : total memory antibodyset($Ab_m \in S^{mxl}, m \leq N$);
Ab_d : d new antibodies to be inserted in Ab($Ab_d \in S^{dxL}$);
Ag : population of antigens($Ag \in S^{Mxl}$);
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 $s_{i,j}$(i, j = 1, ..,N);
C : population of clones generated from Ab($C \in S^{Ncxl}$);
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 $f_{i,j}$,i=1,...,N to all
    $Abi,fi,j = D_{i,j}$, $i = 1, ...N, D_{i,j}$ = 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 $d_{k,j} = D_{k,j}$ among Ag_j and all the elements of C_*,$D_{k,j}$ = affinity (C^*_k, Ag_j) $k = 1, ...,Nc$
    Reselect a subset ζ% of the antibodies with highest $d_{k,j}$ and put them into M_j as memory clones;
    Remove the meomory clones from M_j whose $D_{k,j} > \sigma d$
    Determine $s_{i,k}$ among the memory clones:$s_{i,k}$ =affinity ($M_{j,i}$, $M_{j,k}$), $\forall i, k$
    Eliminate those memory clone whose $s_{i,k} < \sigma s$
    Concatenate the total antibody memory matrix with resultant clonal memory M^*_j :$Abm \leftarrow [Abm; M^*_j]$
  end
  Calculate $s_{i,k}$ = affinity (Ab^i_m, Ab^k_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.politics.mideast,talk.politics.misc	150, 150	300

3.2 ขั้นตอนวิธีการทดลองจัดกลุ่มโดยใช้ aiNet

นำข้อมูลแต่ละกลุ่มมาทำการตัดคำตามหัวข้อที่ 2.1,2.2 โดยรูปแบบเอกสารแต่ละฉบับจะอยู่ในรูปของเวกเตอร์ที่มีมิติเป็น l มิติต่างๆ ก็คือเซตของคำภายในเอกสาร เมื่อมีเอกสารจำนวน N เอกสารก็จะเป็นเมตริกซ์ N x l โดยค่าที่อยู่ในเมตริกซ์ก็คือค่าของน้ำหนักของคำที่ปรากฏในแต่ละเอกสาร โดยจะเป็นค่าแบบ real value และเวกเตอร์ของเอกสารก็คือ กลุ่มของแอนติเจนที่ใช้เรียนรู้ภายในอัลกอริธึมของ aiNet โดยทำการปรับค่าพารามิเตอร์$(\sigma_d, \sigma_s, \varsigma\%, iteration)$ ของ aiNet ที่วัดค่า affinity แบบEuclidean และแบบ Cosine เพื่อจัดกลุ่มข้อมูลโดยข้อมูลชุดที่ 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] ซึ่งเป็นค่าที่รวมเอาค่าความแม่นยำ (Precision: P) และค่าความระลึก (Recall: R) ไว้ในค่าเดียว เรากำหนดให้ชนิดของข้อความที่มีจำนวนมากที่สุดในกลุ่มใดๆ เป็นหัวข้อเรื่อง (topic) มีสูตรดังนี้

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

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

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

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	0.5113	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 สามารถแก้ปัญหาของระยะห่างระหว่างคู่ของเวกเตอร์มีค่ามากเนื่องจากเวคเตอร์มีขนาดใหญ่ที่ส่งผลกระทบในการจัดกลุ่มข้อมูล 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 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 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 Fermando 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 Janeiiro/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.