รายงานโครงการวิจัยประจำปีงบประมาณ 2546

ศึกษาแนวทางการพัฒนาอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์คให้สามารถรับข้อมูลและประมวลผลข้อมูลประเภทข้อความเพื่อใช้ในงานแบ่งกลุ่มเอกสาร

A Study Developing A Text Adaptive Resonance Theory Neural Network Useing Processing Text Data For Document Clustering Task

โดย

ผศ.ดร. วรพจน์ กรีสุระเดช

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง
ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

รายงานโครงการวิจัยประจำปีงบประมาณ 2546

ศึกษาแนวทางการพัฒนาอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์คให้สามารถรับข้อมูลและประมวลผลข้อมูลประเภทข้อความเพื่อใช้ในงานแบ่งกลุ่มเอกสาร

A Study D eloping A Text Adaptive Resonance Theor Neural Network Usei Processing Text Data For Document Clu ering Task

00986
.b11648028
.f127392 2 7

โดย

ผศ.ดร. วรพจน์ กรีสุระเดช

วัน เดือน ปี 0 6 มิ.ย. 2550
เลขทะเบียน H000986
เลขเรียกหนังสือ QA76.87 .G225 ฉ.2
"ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล."

คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520
พ.ศ. 2546

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค

บทคัดย่อ

งานวิจัยนี้จะนำเสนออัลกอริทึมการจัดกลุ่มเอกสารโดยใช้ เท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค ซึ่งมีวิธีการที่แตกต่างจากอัลกอริทึมการแบ่งกลุ่มข้อมูลทั่วไป วิธีการของเท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์คจะกระทำกับข้อมูลที่เป็นข้อความโดยตรง และไม่มีการแปลงข้อความเป็นตัวเลข งานวิจัยนี้มีจุดประสงค์คือการปรับแต่ง ART Neural Network ให้สามารถจัดกลุ่มข้อมูลที่มีคุณสมบัติมีค่าเป็นข้อความ อัลกอริทึมนี้มีความสามารถในการเรียนรู้ด้วยตัวเอง และอาศัยหลักการวัดความคล้ายคลึงสำหรับซิมโบลลิกออบเจ็คต์ ซึ่งหลักการวัดความคล้ายคลึงสำหรับซิมโบลลิกออบเจ็คต์สามารถใช้กับข้อมูลที่ไม่ใช่ตัวเลขได้ โครงข่ายนิวรอลใหม่ที่ได้รับการปรับแต่งนี้สามารถรับข้อมูลที่เป็นข้อความได้โดยตรงและทำการจัดกลุ่มข้อมูลที่มีคุณสมบัติมีค่าเป็นข้อความได้เป็นอย่างดี

A Text Adaptive Resonance Theory Neural Network

ABSTRACT

This research proposes a Text Adaptive Resonance Theory Neural Network for document clustering. Unlike the conventional clustering algorithms, a Text Adaptive Resonance Theory Neural Network works directly on textual information without transforming text data into a numerical value. The main contribution of this paper is to show how to adapt the concepts of ART clustering on a data set, which has a qualitative feature values. The Text Adaptive Resonance Theory Neural Network utilizes the concept of similarity measure for symbolic objects, which is different from the conventional similarity measure for objects whose feature values are numerical values.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1. ความเป็นมา ความสำคัญของปัญหาและจุดประสงค์ของงานวิจัย

ปัจจุบันปริมาณข้อมูลที่เป็นข้อความหรือเอกสารมีปริมาณเพิ่มมากขึ้นอย่างรวดเร็ว วิธีการและอัลกอริทึมต่างๆของ Text Data Mining ได้ถูกนำมาประยุกต์เพื่อใช้กับข้อมูลที่เป็นข้อความ เช่น Text Clustering สำหรับ Text Clustering นั้นคือวิธีการจัดแบ่งกลุ่มข้อความโดยอาศัยค่าความแตกต่างหรือค่าความคล้ายคลึงของวัตถุหรือข้อมูลที่นำมาเปรียบเทียบกัน ข้อมูลที่มีความคล้ายคลึงกันมากจะถูกกำหนดให้อยู่กลุ่มเดียวกันและข้อมูลที่มีความคล้ายคลึงกันน้อยจะถูกกำหนดให้อยู่คนละกลุ่มกัน ซึ่งวิธีการต่างๆนั้นจำเป็นต้องการอัลกอริทึมที่สามารถทำงานได้ดีกับข้อมูลที่สนใจหรือต้องการที่จะนำมาเป็นข้อมูลอินพุตของอัลกอริทึม เพื่อให้กระบวนการทำงานของการจัดแบ่งกลุ่มข้อความมีประสิทธิภาพดีและได้ผลลัพธ์ที่ถูกต้องในระดับที่ยอมรับได้ วิธีการและอัลกอริทึมที่นำเสนอในงานวิจัยนี้จึงมุ่งเน้นเกี่ยวกับ อัลกอริทึมสำหรับการจัดแบ่งกลุ่ม (Clustering Algorithm) เพื่อปรับปรุงและออกแบบอัลกอริทึมใหม่สำหรับการจัดแบ่งกลุ่มที่มีความเหมาะสมกับข้อมูลที่เป็นข้อความหรือเอกสารประเภทข่าวต่างๆ

การจัดกลุ่มแบบคลัสเตอริ่ง (Clustering) คือ กระบวนการจัดกลุ่มของออบเจ็คต์เป็นกลุ่มต่างๆที่มีความคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน และกลุ่มที่มีความแตกต่างกันให้อยู่ในกลุ่มที่แตกต่างกัน และคลัสเตอร์คือกลุ่มข้อมูลที่รวบรวมให้อยู่กลุ่มเดียวกันโดยใช้หลักการวัดความคล้ายคลึง ซึ่งการจัดกลุ่ม (Clustering) มีลักษณะการทำงานแบบเรียนรู้ด้วยตัวเอง (Unsupervised Learning) แตกต่างจากการจำแนกกลุ่ม (Classification) ที่มีลักษณะการทำงานแบบเรียนรู้โดยมีสอน (Supervised Learning) การแบ่งกลุ่มสามารถแบ่งออกเป็น 2 วิธีการหลัก คือ Hierarchical clustering และ Partitioning clustering [1]

อัลกอริทึมที่ใช้หลักการของ Hierarchical clustering มีสองอัลกอริทึมคือ Agglomerative และ Divisive clustering อัลกอริทึมที่ใช้หลักการของ Partitioning clustering ได้แก่ K-Means, Competitive Learning Neural Network, Kohonen Neural Network และ Adaptive Resonance Theory เป็นต้น [1] ซึ่งอัลกอริทึมเหล่านี้จะรับข้อมูลเข้าเป็นตัวเลข (Numerical Feature Value) ดังนั้น การจัดกลุ่มข้อมูลที่ไม่ใช่ตัวเลข จำเป็นต้องทำการแปลงคุณสมบัติข้อมูล (Feature Type) ให้อยู่ในลักษณะของตัวเลข เพื่อให้สามารถนำข้อมูลเข้าอัลกอริทึมได้ ขั้นตอนการแปลงคุณสมบัติของออปเจ็คต์ที่เป็นเอกสารมาเป็นข้อมูลตัวเลข เรียกว่า การแทนเอกสาร (Document Representation) ซึ่งจะต้องแปลงข้อมูลที่เป็นเอกสารข้อความให้อยู่ในรูปแบบที่สามารถประมวลผลด้วยอัลกอริทึมทั่วไปได้ วิธีการหนึ่งที่ได้แพร่หลายให้การแทนเอกสารข้อความคือ วิธีการ Vector Space Model นำเสนอโดย G. Salton [2] มีลักษณะข้อมูลเป็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Matrix Array 2 Dimensions ประกอบด้วย TERM x Document แต่การทำ Document Representation โดยวิธีการ Vector Space Model อาจสร้างจำนวน Term Frequency จำนานมากทำให้เกิดปัญหา High Dimensions ของ Vector Space Model มีผลกับการคำนวณที่ใช้เวลามากขึ้นของอัลกอริทึม และอาจทำให้ข้อมูลสูญเสียความหมายในตัวมันเองเนื่องจากการแปลงเอกสารข้อความมาเป็นอาเรย์ 2 มิติ ที่มี Feature type เป็นตัวเลข [6] ดังนั้นในปัจจุบันปัญหาการจัดกลุ่มได้ขยายปัญหาในการออกแบบอัลกอริทึมสำหรับการจัดกลุ่มข้อมูลที่เป็นตัวเลขมาเป็นปัญหาในการออกแบบอัลกอริทึมสำหรับการจัดกลุ่มที่มีลักษณะข้อมูลเป็นเอกสารข้อความหรือลักษณะข้อมูลที่ไม่ใช่ตัวเลข

ในช่วงเวลาที่ผ่านมามีโครงข่ายนิวรอลที่ได้รับการพัฒนาเพื่อให้สามารถรับข้อมูลที่เป็นข้อความได้ คือ Kohonen Neural Network พัฒนาเป็นโครงข่ายนิวรอลใหม่ชื่อ Text Processing Kohonen Neural Network [7] และ Competitive Learning Neural Network พัฒนาเป็นโครงข่ายนิวรอลใหม่ชื่อ Text Processing Competitive Learning Neural Network [8] ซึ่งโครงข่ายนิวรอลใหม่นี้สามารถจัดแบ่งกลุ่มข้อความได้โดยตรง โดยไม่มีแปลงรูปแบบข้อความนั้นมาเป็นตัวเลข ทำให้โครงข่ายนิวรอลใหม่นี้สามารถรับข้อมูลที่เป็นข้อความได้ การวัดความคล้ายคลึงสามารถทำได้สองแนวทางคือแนวทางแรกเป็นการวัดความแตกต่าง ถ้าเปรียบเทียบกับทุกๆโหนดแล้วโหนดใดมีความแตกต่างกันน้อยที่สุดโหนดนั้นคือโหนดที่ชนะ และแนวทางที่สองเป็นการวัดความคล้ายคลึง ถ้าเปรียบเทียบกับทุกๆโหนดแล้วโหนดใดมีความคล้ายคลึงมากที่สุด โหนดนั้นคือโหนดที่ชนะ [3]

Text Adaptive Resonance Theory Neural Network เป็นโครงข่ายนิวรอลที่ได้รับการปรับแต่งการทำงานของอัลกอริทึมขึ้นเพื่อขยายความสามารถของ ART Neural Network ให้สามารถรับข้อมูลเข้าเป็นข้อความได้โดยตรง โดยไม่ต้องผ่านขั้นตอนการแปลงรูปแบบข้อมูลให้เป็นตัวเลข และยังคงขั้นตอนการทำรีเซตเพื่อใช้ในการควบคุมระดับค่าคล้ายคลึงของ Pattern ที่ทำการ Clustering ได้ ด้วยหลักการทำงานแบบเรียนรู้ด้วยตัวเองและหลักการวัดความคล้ายคลึงของซิมโบลิคออบเจ็ต[5] โครงข่ายนิวรอลที่ได้ปรับแต่งอัลกอริทึมนี้สามารถจัดกลุ่มข้อมูลที่มีคุณสมบัติเป็นข้อความได้เป็นอย่างดี

1.1 จุดประสงค์ของงานวิจัย

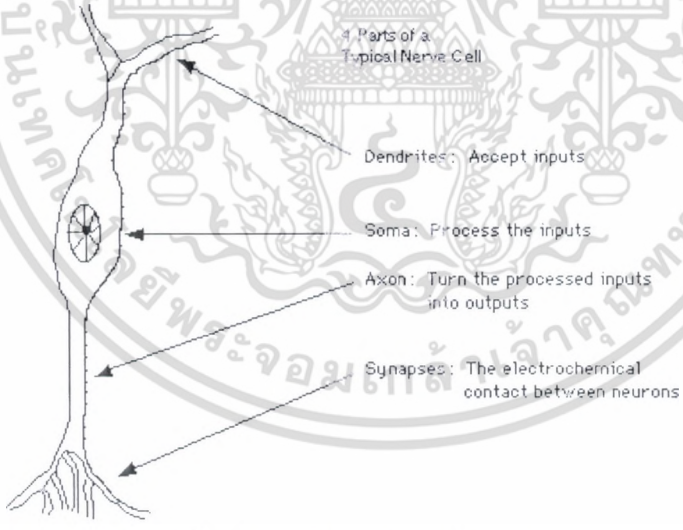
งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาวิธีการจัดแบ่งกลุ่มเอกสารด้วยโครงข่ายนิวรอล ที่สามารถรับค่าข้อมูลเข้าที่เป็น Qualitative Value ได้โดยตรงโดยไม่ต้องแปลงค่าข้อมูลให้อยู่ในรูปแบบตัวเลข เพื่อใช้เป็นแนวทางในการพัฒนาและปรับปรุงอัลกอริทึม ในการจัดกลุ่มเอกสารที่เป็นข้อความ เช่น ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวกีฬา เป็นต้น ให้มีประสิทธิภาพดีสามารถทำงานกับข้อมูลที่เป็นข้อความได้ดีและมีความถูกต้องในระดับที่ยอมรับได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.หลักการและทฤษฎีที่เกี่ยวข้อง

วิธีการจัดแบ่งกลุ่มโดยใช้โครงข่ายนิวรอล เป็นวิธีการเพื่อปรับปรุงความเหมาะหรือความสอดคล้องของกลุ่มข้อมูลที่ทำการวิเคราะห์และหลักการที่เกี่ยวกับคณิตศาสตร์ (Mathematical Model) วิธีการของโครงข่ายนิวรอลสำหรับการแบ่งกลุ่มที่มีอยู่ในขณะนี้ ส่วนใหญ่ใช้ Winner-Take-All Learning Rule ในธรรมชาติเซลล์ประสาทของสิ่งมีชีวิตมีหลายประเภทแล้วแต่หน้าที่ของมัน เซลล์ประสาทในตัวของคนเราก็เช่นกัน มีอยู่หลายประเภทตามตำแหน่งและหน้าที่ของมัน เช่น เซลล์ประสาทกล้ามเนื้อ เซลล์ประสาทในสมอง เซลล์ประสาทที่ลิ้นและจมูก เป็นต้น [3]

เซลล์ประสาทประกอบด้วยส่วนใหญ่ๆ 4 ส่วน [3] คือตัวเซลล์ประสาทหรือนิวรอน ซึ่งมีนิวเคลียสอยู่ตรงกลาง รอบๆ ตัวเซลล์ประสาทมีสิ่งที่ยื่นออกไปเพื่อรับและส่งสัญญาณจากเซลล์ประสาทอื่นๆ สิ่งดังกล่าวเรียกว่า แอกซอน (Axon) ที่ปลายกิ่งจะแตกออกเป็นก้านย่อยๆ เรียกว่า เดนไดรท์ (Dendrite) รอยต่อระหว่างก้านของเซลล์ประสาทที่ต่างกันเรียกว่า ไซแนปส์ (Synapse) ซึ่งสามารถเปลี่ยนค่าความต้านทานได้ตามสัญญาณที่ส่งระหว่างกันของเซลล์ประสาท การส่งสัญญาณระหว่างเซลล์ประสาททำได้โดยการถ่ายเทสารประกอบโซเดียมและโพแทสเซียม [3]



รูปที่ 1 โครงสร้างของเซลล์ประสาท

ฮอดกิ้น (Hodkin) และฮักลีย์ (Huxley) [3] ซึ่งได้รับรางวัลโนเบลทางชีววิทยาได้ค้นพบว่าการไหลของสารประกอบโซเดียมและโพแทสเซียมของเซลล์ประสาทของปลาหมึกได้ทำให้เกิดความต่างศักย์จะอยู่ระหว่าง 50 ถึง 70 มิลลิโวลต์ จากผลการศึกษาดังกล่าวทำให้เราสามารถจำลองการทำงานของเซลล์ประสาทโดยอาศัยวงจรอิเลคทรอนิกส์ได้ โครงสร้างของโครงข่ายนิวรอลมีลักษณะที่คล้ายคลึงกับ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สมองและระบบประสาทซึ่งประกอบด้วยส่วนของการประมวลผลต่างๆ ที่เรียกว่านิวรอน (Neuron) ทุกๆ นิวรอนสามารถมีอินพุทได้หลายอินพุท แต่มีเอาต์พุทเพียงเอาต์พุทเดียว และทุกๆ เอาต์พุทจะแยกไปยังอินพุทของนิวรอนอื่นๆ ภายในโครงข่าย การติดต่อกันภายในระหว่างนิวรอนไม่ใช่ลักษณะของการต่อแบบธรรมดา ทุกๆ อินพุทจะมีค่าถ่วงน้ำหนักเป็นตัวกำหนดกำลังของการติดต่อภายในและช่วยในการตัดสินใจ การทำงานของนิวรอนในบางโครงข่ายจะถูกกำหนดไว้ตายตัว แต่บางโครงข่ายสามารถที่จะปรับแต่งได้ ซึ่งอาจจะเป็นการปรับแต่งจากภายนอกโครงข่ายหรือนิวรอนสามารถปรับตัวได้ด้วยตัวของมันเอง[4] ในจุดนี้เองแสดงถึงความสามารถในการเรียนรู้และจดจำของโครงข่ายนิวรอล มีโครงข่ายนิวรอลหลายรูปแบบ สำหรับโครงข่ายนิวรอลที่ใช้ในการ Clustering ส่วนใหญ่จะนำวิธีการของ Winner-Take-All มาใช้ เช่น Competitive Learning, Kohonen self-organizing maps และ Adaptive Resonance Theory การใช้ Winner-Take-All

2.1 โครงข่ายนิวรอลแบบ Adaptive Resonance Theory

Adaptive Resonance Theory (ART) พัฒนาโดย Carpenter และ Grossberg ในปี 1987 เป้าหมายของอัลกอริทึมนี้คือ การจัดแบ่งกลุ่มข้อมูลที่เป็นตัวเลข ซึ่งมีอัลกอริทึมของ ART1 สำหรับจัดแบ่งกลุ่มข้อมูลตัวเลขที่เป็นไบนารี และอัลกอริทึมของ ART2 สำหรับการจัดแบ่งกลุ่มข้อมูลตัวเลขที่เป็นเลขจำนวนจริง ลักษณะการทำงานของนิวรอลนี้เป็นแบบเรียนรู้ด้วยตัวเอง (Unsupervised Learning) [4]

Adaptive Resonance Theory (ART) ถูกออกแบบเพื่อให้ผู้ใช้สามารถควบคุมระดับความคล้ายคลึงของข้อมูล (Pattern) ในการจัดแบ่งกลุ่มได้ โครงสร้างของ ART แบ่งออกเป็น 3 ระดับ คือ

1. Input Layer ($F_{1(a)}$ layer)
2. Interface Layer ($F_{1(b)}$ layer)
3. Output Layer (F_2 layer)

วิธีการควบคุมระดับความคล้ายคลึงของข้อมูลทำได้โดยการรีเซตเมื่อค่าความคล้ายคลึงน้อยกว่าค่า Vigilance ที่กำหนดโดยผู้ใช้ ดังนั้นการทำงานของ ART จึงแบ่งออกได้ 2 ส่วน คือ Competition และ Decision [4]

1. Competition คือส่วนที่ทำการเปรียบเทียบความเหมือนกันระหว่างข้อมูลเข้ากับโหนดต่างๆ ในนิวรอล

2. Decision คือส่วนที่ตัดสินการจัดแบ่งกลุ่มที่เกิดจากส่วน Competition ขึ้นอยู่กับเงื่อนไขในการรีเซต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนการทำงานอัลกอริทึมของ ART1

ขั้นตอนที่ 0: กำหนดค่าเริ่มต้นให้กับ bottom-up weight และ top-down weight ในแต่ละโหนดของโครงข่ายนิวรอล์ ค่าเริ่มต้นเหล่านี้อาจได้จากการสุ่มเลือกจากข้อมูลที่นำมาใช้ในกระบวนการเรียนรู้

และกำหนดค่าVigilance parameter, $\rho_0 = [0, 1]$ และ L (Learning Rate)

ขั้นตอนที่ 1: เมื่อยังไม่ตรงเงื่อนไขในการหยุด ให้ทำขั้นตอนที่ 2-9

ขั้นตอนที่ 2: เลือกข้อมูลเข้าทั้งหมด หาค่า Norm

$$\|S\| = \sum_i S_i \text{ , ทำขั้นตอนที่ 3-8 ต่อไป} \quad (1)$$

ขั้นตอนที่ 3: กำหนดข้อมูลเข้า D จาก $F_{1(a)}$ ให้กับตัวแปร X ใน $F_{1(b)}$

$$X_i = S_i \quad (2)$$

ขั้นตอนที่ 4: คำนวณเปรียบเทียบข้อมูลเข้า X กับแต่ละโหนดของโครงข่ายนิวรอลจนครบทุกโหนด

$$Y_j = \sum_i b_{ij} x_i \quad (3)$$

ขั้นตอนที่ 5: เปรียบเทียบหาโหนด J ของโครงข่ายนิวรอลที่มีค่ามากที่สุด

ขั้นตอนที่ 6: เปรียบเทียบข้อมูลเข้า X กับ top-down weight ที่เชื่อมต่อกับ winning node J

$$x_i = s_i t_{ji} \quad (4)$$

คำนวณหาค่า Norm ของ X

$$\|x\| = \sum_i x_i \quad (5)$$

ขั้นตอนที่ 7: ทดสอบเงื่อนไขการ reset mechanism

ถ้า $\frac{\|x\|}{\|s\|} < \rho_0$ ให้ $Y_J = -1$ (ยับยั้งโหนด J), ทำขั้นตอนที่ 5 อีกครั้ง

ถ้า $\frac{\|x\|}{\|s\|} \geq \rho_0$ ให้ทำขั้นตอนที่ 8

ขั้นตอนที่ 8: ปรับปรุงค่า bottom-up weight และ top-down weight ของโหนด J

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$b_{ij}(new) = \frac{Lx_i}{L-1+\|x\|} \tag{6}$$

$$t_{ji}(new) = x_i \tag{7}$$

ขั้นตอนที่ 9: ทดสอบเงื่อนไขการหยุด เงื่อนไขการหยุดสามารถกำหนดโดย:

9.1Weight ไม่มีการเปลี่ยนแปลง

9.2ไม่มีการรีเซต

9.3 จำนวนรอบของการเรียนรู้

ในงานวิจัยนี้มีขอบเขตการทำโดยจะทำการปรับแต่งอัลกอริทึมของ ART1 เพื่อให้ ART สามารถรับข้อมูลที่เป็นข้อความได้

2.2 การเปรียบเทียบความคล้ายคลึงของเอกสาร (Similarity Measure)

ขั้นตอนก่อนการเปรียบเทียบความคล้ายคลึงของเอกสารคือ ขั้นตอนการแทนเอกสารต่างๆ (Document Representation) ให้อยู่รูปแบบที่สามารถนำมาคำนวณได้ เอกสารต่างๆ ซึ่งประกอบด้วย ชื่อเรื่อง คำสำคัญ สามารถแทนอยู่ในรูปแบบของ Cartesian Product ได้ดังนี้ [5]

$$Doc = D_1 \times D_2 \times D_3 \times \ldots \times D_d \tag{8}$$

d คือจำนวนคุณสมบัติของเอกสาร และคุณสมบัติของเอกสารมีลักษณะข้อมูลเป็นข้อความ ตัวอย่างเช่น

$$Doc = Title \times Keyword \tag{9}$$

เมื่อ Title Feature คือคำที่ใช้อธิบายชื่อเรื่องของเอกสาร

Keyword Feature คือคำที่ใช้อธิบายคำที่พบบ่อยในเอกสาร หรือคำสำคัญของเอกสาร

การเปรียบเทียบความคล้ายคลึงของเอกสาร ระหว่างเอกสาร A และเอกสาร B ตามแนวคิดของ El-Sonbaty สามารถเขียนในสมการได้ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$S(A,B) = \sum_{k=1}^{d} S(A_k, B_k) \tag{10}$$

ในการเปรียบเทียบความคล้ายคลึงของ $S(A_k, B_k)$ มีการเปรียบเทียบ 2 ส่วนย่อยคือ ส่วนที่เป็นขนาดของ Feature เรียกว่า Span, $S_s(A,B)$ และส่วนที่เป็นเนื้อหาของ Feature เรียกว่า Content, $S_c(A,B)$ ซึ่งนิยามสมการไว้ดังนี้ [5]

ส่วนการเปรียบเทียบความคล้ายคลึงของ Span มีนิยามเป็น

$$S_s(A,B) = \frac{(l_a + l_b)}{2xl_s} \tag{11}$$

ส่วนการเปรียบเทียบความคล้ายคลึงของ Content มีนิยามเป็น

$$S_c(A,B) = \frac{inters}{l_s} \tag{12}$$

นิยามของสัญลักษณ์ในสมการมีดังนี้

l_a = จำนวนสมาชิกทั้งหมดใน Feature A

l_b = จำนวนสมาชิกทั้งหมดใน Feature B

$inters$ = จำนวนของสมาชิกทั้งหมดที่ Intersection กันระหว่าง Feature A และ Feature B

$l_s = l_a + l_b - inters$ = จำนวนสมาชิกทั้งหมดของ Feature A และ Feature B รวมกันลบด้วยจำนวนของสมาชิกทั้งหมดที่ Intersection กันระหว่าง Feature A และ Feature B

เมื่อเปรียบเทียบความคล้ายคลึงครบทั้งสองส่วนย่อยแล้ว จึงทำการรวมค่าผลลัพธ์จากสองย่อยเข้าด้วยดังสมการ Net Similarity นี้

$$S(A_k, B_k) = S_s(A_k, B_k) + S_c(A_k, B_k) \tag{13}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่าความคล้ายคลึงของ Net Similarity นี้สามารถมีค่าได้ระหว่าง [0.5,2] ดังตัวอย่างในตารางที่ 1

ตารางที่ 1 ข้อมูลเอกสาร

DOCUMENT	*TITLE*	*KEYWORD*
Doc1	JAVA Handbook	JAVA Object Oriented
Doc2	C++ Handbook	C++ Object Oriented
Doc3	MatLab Handbook	MatLab matrix laboratory

ตัวอย่างที่ 1 วัดความคล้ายคลึงระหว่าง Doc1 กับ Doc2

S(Doc1,Doc2) = S({JAVA Handbook}, {C++ Handbook}) + S({JAVA Object Oriented}, {C++ Object Oriented})

S({JAVA HandBook},{C++ Handbook})

= [(1+1) / 2.(1+1-0)]+[0/(1+1-0)]

= 0.5+0.0

= 0.5

ค่า Net Similarity ของ Title feature เท่ากับ 0.5

S({JAVA Object Oriented} ,{ C++ Object Oriented})

= [(1+1)/2.(1+1-1)]+ [0/(1+1-1)]

= 0.5+0.0

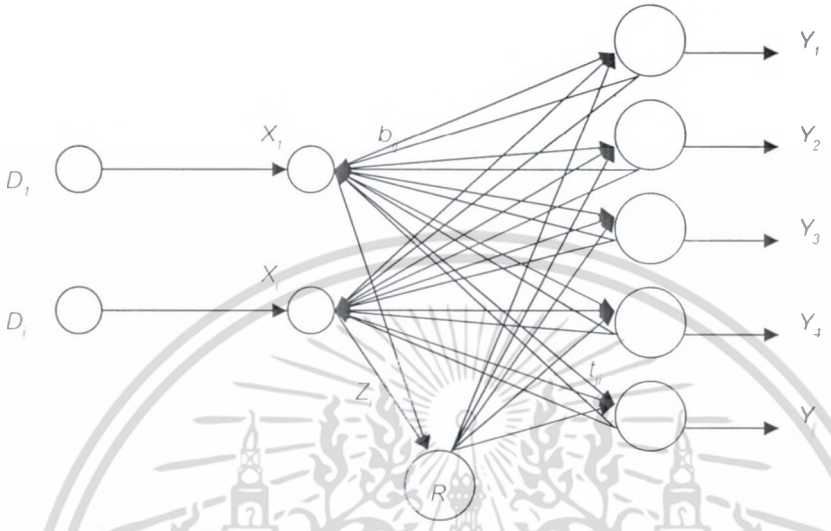
= 0.5

ค่า Net Similarity ของ Keyword feature เท่ากับ 0.67

ดังนั้นค่า Net Similarity ระหว่าง Doc1 และ Doc2 จากการเปรียบเทียบด้วย 2 Feature เท่ากับ 0.5 + 0.5 = 1.0 หลักการของ Similarity Measure นี้จะได้ถูกนำมาใช้ในการจัดกลุ่มข้อมูลของอัลกอริทึม Text Adaptive Resonance Theory Neural Network ที่จะได้นำเสนอนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3 เท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค (A Text Adaptive Resonance Theory Neural Network)



รูปที่ 2 โครงสร้างของ Text Adaptive Resonance Theory Neural

โครงข่ายนิวรอลที่นำเสนอนี้มีความแตกต่างจากโครงข่ายนิวรอลทั่วไปคือ ในส่วนของการรับข้อมูลเข้า โครงข่ายนิวรอลนี้ถูกออกแบบมาให้รับข้อมูลที่เป็นข้อความ การเรียนรู้ข้อมูลที่เข้ามาเป็นแบบเรียนรู้ด้วยตัวเอง (Unsupervised Learning) โครงสร้างของโครงข่ายนิวรอลนี้ประกอบด้วย 3 ชั้น คือ

4. Input Layer ($F_{1(a)}$ layer)
5. Interface Layer ($F_{1(b)}$ layer)
6. Output Layer (F_2 layer)

Text Adaptive Resonance Theory Neural Network มี R unit เพื่อใช้ในการรีเซตการจัดกลุ่มข้อมูล เรียกว่า วิธีการ Reset mechanism ซึ่งจะทำการรีเซตโดยเงื่อนไขที่มีค่า Vigilance Parameter โดยค่า Vigilance Parameter เป็นค่าที่ใช้ควบคุมระดับความคล้ายคลึงของข้อมูลเข้า และ Weights ในโครงข่ายนิวรอล

การทำงานของ Text Adaptive Resonance Theory Neural Network มีการทำงานแบบ Feed Back เพื่อให้โครงข่ายนิวรอลนี้มีความสามารถในการควบคุมระดับความคล้ายคลึงของการจัดกลุ่มโดยใช้วิธีการรีเซต [4]ดังนั้นโครงข่ายประสาทนี้จึงมี 2 Weights คือ Bottom-up weight $,b_{ij}$,ซึ่งจะเชื่อมต่อและส่งข้อมูลระหว่างโหนดที่ i ใน $F_{1(b)}$ layer ไปยังโหนดที่ j ใน F_2 layer และ Top-down weight. t_{ji} ,ซึ่งจะเชื่อมต่อและส่งข้อมูลระหว่างโหนดที่ j ใน F_2 layer และโหนดที่ i ใน $F_{1(b)}$ layer

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เพื่อที่จะทำให้โครงข่ายนิวรอลนี้สามารถหาความคล้ายคลึงของข้อมูลเข้าที่เป็น Qualitative value ในที่นี้คือ Text ได้ ดังนั้นทั้ง Bottom-up weight และ Top-down weight จะเก็บค่าใน Weight เป็น Qualitative value และค่าแสดงความเป็นสมาชิกของ Weight

$$b_{ij} = \{(A_{1ij}, e_{1ij}), (A_{2ij}, e_{2ij}), ..., (A_{pij}, e_{pij})\} \quad (14)$$

กำหนดให้ A_{pij} คือ ค่า qualitative value ของ Bottom-up Weightและ e_{pij} คือค่า Degree แสดงความเป็นสมาชิกของ A_{pij}

ค่า e_{pij} ของ A_{pij} ในโครงข่ายนิวรอลนี้มีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 ซึ่งค่า Degree นี้เป็นค่าที่ให้ระดับความเป็นสมาชิกของ A_{pij} กับ ข้อมูลเข้า ถ้า Degree ของ A_{pij} มีค่าเท่ากับ 0 ให้ความหมายว่า Qualitative value ของ A_{pij} ไม่ได้เป็นส่วนหนึ่งของข้อมูลเข้า i ถ้า Degree ของ A_{pij} มีค่าเท่ากับ 1 ให้ความหมายว่า Qualitative value ของ A_{pij} เป็นสมาชิกของข้อมูลเข้า i

$$t_{ji} = \{(B_{1ji}, e_{1ji}), (B_{2ji}, e_{2ji}), ..., (B_{pji}, e_{pji})\} \quad (15)$$

กำหนดให้ B_{pji} คือ ค่า qualitative value ของ Top-down weight และ e_{pji} คือค่า Degree แสดงความเป็นสมาชิกของ B_{pji}

ค่า e_{pji} ของ B_{pji} ในโครงข่ายนิวรอลนี้มีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 ซึ่งค่า Degree นี้เป็นค่าที่ให้ระดับความเป็นสมาชิกของ B_{pji} กับ ข้อมูลเข้า ถ้า Degree ของ B_{pji} มีค่าเท่ากับ 0 ให้ความหมายว่า Qualitative value ของ B_{pji} ไม่ได้เป็นส่วนหนึ่งของข้อมูลเข้า j ถ้า Degree ของ B_{pji} มีค่าเท่ากับ 1 ให้ความหมายว่า Qualitative value ของ B_{pji} เป็นสมาชิกของข้อมูลเข้า j

2.4 กระบวนการเรียนรู้ของอัลกอริทึม (Learning Algorithm)

Text Adaptive Resonance Theory Neural Network เป็นการขยายความสามารถของ ART1 Neural Network โดยเพิ่มเติมแนวคิดเกี่ยวกับคุณสมบัติที่มีค่าของข้อมูลเป็นแบบข้อมูลเชิงคุณภาพ Qualitative value ซึ่งทำให้อัลกอริทึมนี้สามารถจัดการกับข้อมูลแบบเชิงคุณภาพได้โดยตรงโดยไม่ต้องผ่านกระบวนการแปลงค่า Qualitative Value เป็น Numerical Value ขั้นตอนของการเรียนรู้ของอัลกอริทึมมีดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 0: กำหนดค่าเริ่มต้นให้กับ bottom-up weight และ top-down weight ในแต่ละโหนดของโครงข่ายนิวรอล ค่าเริ่มต้นเหล่านี้อาจได้จากการสุ่มเลือกจากข้อมูลที่นำมาใช้ในกระบวนการเรียนรู้

และกำหนดค่าVigilance parameter, ρ

ขั้นตอนที่ 1: เมื่อยังไม่ตรงเงื่อนไขในการหยุด ให้ทำขั้นตอนที่ 2-9

ขั้นตอนที่ 2: เลือกข้อมูลเข้าทั้งหมด แล้วทำการ Transpose

$$Doc = (D_1, D_2, D_3, \ldots, D_d)^t \text{ , ทำขั้นตอนที่ 3-8 ต่อไป} \quad (16)$$

ขั้นตอนที่ 3: กำหนดข้อมูลเข้า D จาก $F_{1(a)}$ ให้กับตัวแปร X ใน $F_{1(b)}$

$$X_i = D_i \quad (17)$$

ขั้นตอนที่ 4: คำนวณเปรียบเทียบข้อมูลเข้า X กับแต่ละโหนดของโครงข่ายนิวรอลจนครบทุกโหนด

$$Y_j = \sum_{i=1}^{d} \sum_{k=1}^{p} S(X_k, A_{nkj}).e_{nkj} \quad (18)$$

กำหนดให้

p คือจำนวนของ bottom-up weight

d คือจำนวนของ feature values

ขั้นตอนที่ 5: เปรียบเทียบหาโหนด J ของโครงข่ายนิวรอลที่มีค่ามากที่สุด

ขั้นตอนที่ 6: เปรียบเทียบข้อมูลเข้า X กับ top-down weight ที่เชื่อมต่อกับ winning node J

$$Z = S(X_i, t_{Ji}).e_{Ji} \quad (19)$$

ขั้นตอนที่ 7: ทดสอบเงื่อนไขการ reset mechanism

ถ้า $Z < \rho$, ให้ $Y_J = -1$ (ยับยั้งโหนด J), ทำขั้นตอนที่ 5 อีกครั้ง

ถ้า $Z \geq \rho$, ให้ทำขั้นตอนที่ 8

ขั้นตอนที่ 8: ปรับปรุงค่า bottom-up weight และ top-down weight ของโหนด J

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$b_{i,J}^{(new)} = b_{i,J}^{(old)} \cup X$$

$$e_{ni,J}^{(new)} = \begin{cases} f(e_{ni,J}^{(old)} + \eta) \text{ if } A_{nij} \in b_{ij} \cap X, \\ f(e_{ni,J}^{(old)} - \eta) \text{ if } A_{nij} \notin b_{ij} \cap X \end{cases} \quad (20)$$

$$t_{Ji}^{(new)} = X$$

$$e_{nJi}^{(new)} = \begin{cases} f(e_{nJi}^{(old)} + \eta) \text{ if } B_{nji} \in t_{Ji} \cap X, \\ f(e_{nJi}^{(old)} - \eta) \text{ if } B_{nji} \notin t_{Ji} \cap X \end{cases} \quad (21)$$

กำหนดให้ $f(.)$ นิยามดังนี้

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x < 0 \\ 1 & \text{if } x > 1 \end{cases} \quad (22)$$

ขั้นตอนที่ 9: ทดสอบเงื่อนไขการหยุด เงื่อนไขการหยุดสามารถกำหนดโดย:

9.1Weight ไม่มีการเปลี่ยนแปลง

9.2ไม่มีการรีเซต

9.3 จำนวนรอบของการเรียนรู้

ตัวอย่างที่ 2 อัลกอริทึมการทำงานของ Text Adaptive Resonance Theory Neural Network Neural Network

ข้อความที่ทำการจัดแบ่งกลุ่มมี 4 ชุด คือ (tv,vy),(rstv,uwy),(hmn,mnp),(n,mp) ให้โครงข่ายนิวรอลแบ่งเป็น 3 กลุ่ม

ขั้นตอนที่ 0 กำหนดค่าเริ่มต้น

Learning Rate = 0.01

Vigilance = 0.5

สุ่มค่า Weight จากข้อมูลอินพุต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$b_{11} = \{tv\}$

$b_{12} = \{n\}$

$b_{13} = \{hmn\}$

$b_{21} = \{mnp\}$

$b_{22} = \{mnp\}$

$b_{23} = \{mp\}$

$t_{11} = \{tv\}$

$t_{21} = \{uwy\}$

$t_{31} = \{tv\}$

$t_{12} = \{rstv\}$

$t_{22} = \{vy\}$

$t_{32} = \{uwy\}$

Degree ของ bottom-up weight เท่ากับ 0.25

Degree ของ top-down weight เท่ากับ 1

ขั้นตอนที่ 1: เมื่อยังไม่ตรงเงื่อนไขในการหยุด ให้ทำขั้นตอนที่ 2-9

ขั้นตอนที่ 2: เลือกข้อมูลเข้าทั้งหมด แล้วทำการ Transpose

$Doc = (D_1, D_2)'$, ทำขั้นตอนที่ 3-8 ต่อไป

ขั้นตอนที่ 3: กำหนดข้อมูลเข้า D จาก $F_{1(a)}$ ให้กับตัวแปร X ใน $I_{(a)}$

$X_i = D_i$

ขั้นตอนที่ 4: คำนวณเปรียบเทียบข้อมูลเข้า X กับแต่ละโหนดของโครงข่ายนิวรอลจนครบทุกโหนด

$$Y_j = \sum_{k=1}^{d} \sum_{n=1}^{p} S(X_k, A_{nkj}).e_{nkj}$$

$Y_1 = S(X_1, A_{11})e_{11}$ // Output Neuron 1, d = 1 (TITLE FEATURE)

$S(X_1, A_{11}) = S_s(X_1, A_{11}) + S_c(X_1, A_{11})$

$S_s(X_1, A_{11}) = S_s(tv, tv)$

$S_s(tv, tv) = [(1+1)/(2x1+1-1)] = 1.0$

$S_c(X_1, A_{11}) = S_c(tv, tv)$

$S_c(tv, tv) = [1/1+1-1] = 1.0$

$S(tv, tv) = 1.0 + 1.0 = 2$

$S(tv, tv).e = 2x0.25 = 0.5$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$Y_1 = S(X_2, A_{21})e_{21}$ // Output Neuron 1, d=2 (KEYWORD FEATURE)

$S(X_2, A_{21}) = S_s(X_2, A_{21}) + S_c(X_2, A_{21})$

$S_s(X_2, A_{21}) = S_s(vy, mnp)$

$S_s(vy, mnp) = [(1+1)/(2x1+1-0)] = 0.5$

$S_c(X_2, A_{21}) = S_c(vy, mnp)$

$S_c(vy, mnp) = [0/(1+1-0)] = 0$

$S(vy, mnp) = 0.5 + 0 = 0.5$

$S(vy, mnp).e = 0.5x0.25 = 0.125$

$Y_1 = 0.5 + 0.125 = 0.625$ // S(TITLE)+S(KEYWORD)

$Y_2 = S(X_1, A_{12})e_{12}$ // Output Neuron 2, d = 1 (TITLE FEATURE)

$S(X_1, A_{12}) = S_s(X_1, A_{12}) + S_c(X_1, A_{12})$

$S_s(X_1, A_{12}) = S_s(tv, n)$

$S_s(tv, n) = [(1+1)/(2x1+1-0)] = 0.5$

$S_c(X_1, A_{12}) = S_c(tv, n)$

$S_c(tv, n) = [0/2] = 0$

$S(tv, rstv) = 0.5 + 0 = 0.5$

$S(tv, rstv).e = 0.5x0.25 = 0.125$

$Y_2 = S(X_2, A_{22})e_{22}$ // Output Neuron 2, d=2 (KEYWORD FEATURE)

$S(X_2, A_{22}) = S_s(X_2, A_{22}) + S_c(X_2, A_{22})$

$S_s(X_2, A_{22}) = S_s(vy, mnp)$

$S_s(vy, mnp) = [(1+1)/(2x2)] = 0.5$

$S_c(X_2, A_{22}) = S_c(vy, mnp)$

$S_c(vy, mnp) = [0/2] = 0$

$S(vy, mnp) = 0.5 + 0 = 0.5$

$S(vy, mnp).e = 0.5x0.25 = 0.125$

$Y_2 = 0.125 + 0.125 = 0.250$ // S(TITLE)+S(KEYWORD)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$Y_3 = S(X_1, A_{13})e_{13}$ // Output Neuron 3, d = 1 (TITLE FEATURE)

$S(X_1, A_{13}) = S_s(X_1, A_{13}) + S_c(X_1, A_{13})$

$S_s(X_1, A_{13}) = S_s(tv, hmn)$

$S_s(tv, hmn) = [(1+1)/(2x2)] = 0.5$

$S_c(X_1, A_{13}) = S_c(tv, hmn)$

$S_c(tv, hmn) = [0/2] = 0$

$S(tv, hmn) = 0.5 + 0 = 0.5$

$S(tv, hmn).e = 0.5x0.25 = 0.125$

$Y_3 = S(X_2, A_{23})e_{23}$ // Output Neuron 3, d=2 (KEYWORD FEATURE)

$S(X_2, A_{23}) = S_s(X_2, A_{23}) + S_c(X_2, A_{23})$

$S_s(X_2, A_{23}) = S_s(vy, mp)$

$S_s(vy, mp) = [(1+1)/(2x2)] = 0.5$

$S_c(X_2, A_{23}) = S_c(vy, mp)$

$S_c(vy, mp) = [0/2] = 0$

$S(vy, mp) = 0.5 + 0 = 0.5$

$S(vy, mp).e = 0.5x0.25 = 0.125$

$Y_3 = 0.125 + 0.125 = 0.250$ // S(TITLE)+S(KEYWORD)

ขั้นตอนที่ 5: เปรียบเทียบหาโหนด J ของโครงข่ายนิวรอลที่มีค่ามากที่สุด

Y_1 มีค่ามากที่สุด ดังนั้นโหนดที่ 1 เป็นโหนดที่ชนะ

ขั้นตอนที่ 6: เปรียบเทียบข้อมูลเข้า X กับ top-down weight ที่เชื่อมต่อกับ winning node J

$Z = S(X_i, t_{ji})e_{ji}$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$S(X_1, t_{11}).e_{11} = S(tv, tv).(1)$

$S(tv, tv) = S_s(tv, tv) + S_c(tv, tv)$

$S_s(tv, tv) = [(1+1)/(2x1)] = 1.0$

$S_c(tv, tv) = (1/1) = 1$

$S(tv, tv).(1) = (1+1).(1) = 1$

$S(X_2, t_{12}).e_{12} = S(vy, rstv).(1)$

$S(vy, rstv) = S_s(vy, rstv) + S_c(vy, rstv)$

$S_s(vy, rstv) = [(1+1)/(2x2)] = 0.5$

$S_c(vy, rstv) = (0/2) = 0$

$S(vy, rstv).(1) = (0.5+0).(1) = 0.5$

$Z = 1 + 0.5 = 1.5$

ขั้นตอนที่ 7: ทดสอบเงื่อนไขกับ reset mechanism

ถ้า $Z < \rho_*$ ให้ $Y_J = -1$ (ยับยั้งโหนด J), ทำขั้นตอนที่ 5 อีกครั้ง

ถ้า $Z > \rho_*$ ให้ทำขั้นตอนที่ 8

ทดสอบเงื่อนไขแล้ว ผ่านเงื่อนไข ดังนั้นทำขั้นตอนที่ 8 เพื่อปรับค่า Weight

ขั้นตอนที่ 8: ปรับปรุงค่า bottom-up weight และ top-down weight ของโหนด J

$$b_{iJ}^{(new)} = b_{iJ}^{(old)} \cap X$$

$$e_{miJ}^{(new)} = \begin{cases} f(e_{miJ}^{(old)} + \eta) \text{ if } A_{mij} \in b_{ij} \cap X, \\ f(e_{miJ}^{(old)} - \eta) \text{ if } A_{mij} \notin b_{ij} \cap X \end{cases}$$

$b_{11}^{(new)} = \{tv\} \cap \{tv\} = \{tv\}$

$e_{11}^{(new)} = 0.25 + 0.01 = 0.26$

$b_{21}^{(new)} = \{mnp\} \cap \{vy\} = \{mnp, vy\}$

$e_{21}^{(new)} = 0.25 - 0.01 = 0.24$

$$t_{Ji}^{(new)} = X$$

$$e_{nJi}^{(new)} = \begin{cases} f(e_{nJi}^{(old)} + \eta) \text{ if } B_{nji} \in t_{Ji} \cap X, \\ f(e_{nJi}^{(old)} - \eta) \text{ if } B_{nji} \notin t_{Ji} \cap X \end{cases}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

$t_{11}^{(new)} = \{tv\}$

$e_{11}^{(new)} = 1 + 0.01 = 1.01 = 1$

$t_{12}^{(new)} = \{vy\}$

$e_{12}^{(new)} = 1 - 0.01 = 0.99$

กำหนดให้ $f(.)$ นิยามดังนี้

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x < 0 \\ 1 & \text{if } x > 1 \end{cases}$$

ขั้นตอนที่ 9: ทดสอบเงื่อนไขการหยุด เงื่อนไขการหยุดสามารถกำหนดโดย:

9.1Weight ไม่มีการเปลี่ยนแปลง

9.2ไม่มีการรีเซต

9.3 จำนวนรอบของการเรียนรู้

เมื่อสิ้นสุดการนำข้อมูล 1 ชุดมาจัดแบ่งกลุ่ม หลังจากนั้นนำข้อมูลชุดที่เหลือนำเข้าอัลกอริทึม จบครบทุกชุดและครบจำนวนรอบการเรียนรู้

2.5 ผลการทดลองเบื้องต้น

ในการทดลองเพื่อทดสอบความถูกต้องเบื้องต้นของอัลกอริทึม โดยใช้ชุดข้อมูลซึ่งสร้างขึ้นเพื่อใช้ในการฝึกฝนจำนวน 100 ชุด ซึ่งสร้างจากกลุ่มตัวอักษรจำนวน 3 กลุ่ม กลุ่มตัวอักษรนี้ใช้เป็นตัวแทนของข้อมูลที่มีค่าของคุณสมบัติเป็นข้อความ ข้อมูลแต่ละตัวประกอบด้วยคุณสมบัติ 2 คุณสมบัติ คือ

Title และ Keyword

รูปที่ 3 a ชุดตัวอักษรที่ใช้สร้างข้อมูลใน Title และรูปที่ 3 b ชุดตัวอักษรที่ใช้สร้างข้อมูลใน Keyword

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2. แสดงข้อมูลบางส่วนที่ใช้ในการฝึกฝน

DATA NO.	*TITLE*	*KEYWORD*
1	tv	vy
2	rstv	uwy
3	hmn	mnp
4	n	mp
5	stv	vwxy

ตารางที่ 3. แสดงข้อมูลบางส่วนที่ใช้ในการทดสอบ

DATA NO.	*TITLE*	*KEYWORD*
1	gn	py
2	hnv	ijmp
3	adh	fj
4	vuv	wxy
5	vdgh	ydij

ตัวอย่างของข้อมูลที่สร้างขึ้นจากชุดของตัวอักษรในภาพที่ 3 ได้แสดงในตารางที่ 2 และตารางที่ 3 ตามลำดับ โดยข้อมูลที่นำมาใช้ในการฝึกฝนประกอบด้วยข้อมูลที่เป็นสมาชิกของ Cluster 1 จำนวน 34 ข้อมูล ข้อมูลที่เป็นสมาชิกของ Cluster 2 จำนวน 33 ข้อมูล ข้อมูลที่เป็นสมาชิกของ Cluster 3 จำนวน 33 ข้อมูล และโครงข่ายนิวรอลที่ใช้ในการฝึกฝนนี้ประกอบด้วยจำนวนเอาพุตนิวรอลจำนวน 4 โหนด แต่ละโหนดประกอบด้วย Bottom-Up weight และ Top-down weight ซึ่งเป็นตัวแทนของคุณสมบัติ Title และ Keyword หลังจากที่ทำการฝึกฝน เราจะได้ค่า Weight ของโครงข่ายนิวรอลซึ่งเป็นตัวแทนคุณสมบัติของ Title และ Keyword ของข้อมูลที่นำมาฝึกฝน เพื่อวัดค่าความถูกต้องของอัลกอริทึมที่นำเสนอ ค่าความถูกต้องของการจัดแบ่งกลุ่ม r นิยามได้ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$r = \left[1 - \left(\frac{\sum_{i=1} doc}{n} \right) \right] \times 100 \qquad (23)$$

เมื่อ *doc* คือจำนวนของเอกสารที่ผิดพลาดในการจัดแบ่งกลุ่ม

C คือจำนวนของการจัดแบ่งกลุ่มที่แท้จริงของข้อมูล

ซึ่งการทดลองจะทำการฝึกฝนและทดสอบโครงข่ายนิวรอลที่นำเสนอนี้จำนวน 2 ครั้ง ผลลัพธ์ของการทดลองคือค่าเฉลี่ยจากการทดสอบ 2 ครั้ง

ในการฝึกฝนครั้งที่ 1 นิวรอลลำดับที่ 4 เป็นตัวแทนของ Cluster 1 นิวรอลลำดับที่ 1 เป็นตัวแทนของ Cluster 2 และ นิวรอลลำดับที่ 5 เป็นตัวแทนของ Cluster 3 ในการทดสอบกับข้อมูลที่ใช้ทดสอบครั้งที่ 1 ได้ทำการทดสอบกับข้อมูล 100 ตัว เปอร์เซนต์ความถูกต้องได้เท่ากับ 98.0 เปอร์เซ็นต์

ในการฝึกฝนครั้งที่ 2 นิวรอลลำดับที่ 2 เป็นตัวแทนของ Cluster 1 นิวรอลลำดับที่ 3 เป็นตัวแทนของ Cluster 2 และ นิวรอลลำดับที่ 5 เป็นตัวแทนของ Cluster 3 ในการทดสอบกับข้อมูลที่ใช้ทดสอบครั้งที่ 2 ได้ทำการทดสอบกับข้อมูล 100 ตัว เปอร์เซนต์ความถูกต้องได้เท่ากับ 100.0 เปอร์เซ็นต์ ดังนั้นเมื่อหาค่าเฉลี่ยค่าถูกต้องของการจัดแบ่งกลุ่มของโครงข่ายนิวรอลที่นำเสนอนี้ โครงข่ายนิวรอล Text Adaptive Resonance Theory Neural Network นี้สามารถให้ค่าเฉลี่ยความถูกต้องเท่ากับ 99 เปอร์เซ็นต์ ซึ่งแสดงให้เห็นว่าโครงข่ายนิวรอลที่ออกแบบสร้างขึ้นนี้สามารถทำงานในการจัดแบ่งกลุ่มข้อมูลที่เป็นข้อความได้เป็นอย่างดี

3. เอกสารอ้างอิง

[1] Han Jiawei and Kamber Micheline. **Data Mining: Concepts and Technique.** Margan Kaufmann Publishers, New York, 2001.

[2] Salton Gerard. **Automatic Text Processing: the transformation, analysis, and retrieval of information by computer.** Addison Wesley Publishing Company, New York, 1989.

[3] Jacek M. Zurada. **Introduction to Artificial Neural Systems.** West Publishing Company, New York, 1992.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

[4] Fausett Lanrene. **Fundamentals of Neural Networks Architecture, Algorithms and Application**. Prentice Hall International, New Jersey, 1994.

[5] K.C. Gowda and E.Diday. "Symbolic Clustering Using a New Similarity Measure." IEEE Trans. On Syst., Man, Cybern., vol. 22, 1992. no. 2, pp. 368-378.

[6] El-SonBaty YA and Ismail MA. "Fuzzy Clustering for Symbolic Data." IEEE Trans. On Fuzzy Systems, vol. 6, no. 2, May, 1998. pp. 195-204

[7] ทรงผล ชุติพงศ์พัฒนกุล. "การจัดกลุ่มเอกสารโดยใช้เท็กโปรเซสซิ่งโคโฮเนนนิวรอลเน็ตเวิร์ค." วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2544

[8] สมคิด แสนเสนาะ. "การแบ่งกลุ่มเอกสารโดยใช้เทคนิคการประมวลผลข้อความด้วยโครงข่ายประสาทเทียมที่เรียนรู้แบบหาผู้ชนะ." วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2544

[9] T.V. Ravi and K.C. Gowda. "Clustering of Symbolic Objects Using Gravitational Approach." IEEE Trans. On Syst., Man, Cybern., vol. 29, 1999. no. 6, pp. 888-894.

[10] Hsin-Chang Yang and Chung-Hong Lee. "Automatic Category Generation for Text Documents by Self-Organizing Maps." IEEE Trans., 2000. pp. 581-586

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้