ระบบจำแนกและค้นคืนข้อมูลเว็บกระทู้ข่าว ด้วยโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น

A Web News Information Classification and Retrieval System using Multilayer Perceptron Neural Network

สุภะ จันทา (Supa Chanta)* และ นลินภัสร์ ปรวัฒน์ปรียกร (Nalinpat Porrawatpreyakorn)*

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบจำแนกและค้นคืนข้อมูลเว็บกระทู้ข่าว โดยใช้โครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น (Multilayer Perceptron) ซึ่งใช้ 3 เครื่องมือหลักในการจำแนกข้อมูลกระทู้ข่าวจากเว็บไซต์พันทิป ได้แก่ 1) Rapidminer ใช้ในการพัฒนาโมเดลของโครงข่ายประสาทเทียม 2) Javascript และ jQuery ใช้ในการพัฒนาระบบเก็บรวบรวมข้อมูล (Crawler) จากเว็บไซต์พันทิป และ 3) คลาสไลบรารีเล็กซ์โต (Thai Lexeme Tokenizer : LexTo) ใช้ในการตัดคำภาษาไทยและคำนวณหาค่าน้ำหนัก (Weight) ของคำนั้น ๆ เพื่อใช้เป็นชุดข้อมูลสำหรับเรียนรู้ของโครงข่ายประสาทเทียมในการจำแนกข้อมูล และใช้ 4 เครื่องมือหลักในการค้นคืนข้อมูล ได้แก่ 1) Vector Space Model (VSM) ใช้ในการค้นคืนข้อมูลเพื่อเปรียบเทียบความคล้ายของคำค้นกับเอกสาร 2) Apache Solr ใช้ในการสร้างดัชนีข้อมูล (Index) ของเอกสารเพื่อใช้ในการค้นคืนข้อมูลอย่างมีประสิทธิภาพ 3) N-Gram ใช้ในการแนะนำชุดคำถามที่ถูกต้องแบบอัตโนมัติ และ 4) LexTo ใช้ในการตัดคำเพื่อขยายชุดคำถาม (Query Expansion) ให้ได้ผลลัพธ์ที่ตรงตามความต้องการมากที่สุด พร้อมทั้งตัดคำหยุดหรือคำที่ไม่มีความหมาย (Stop-Word) เพื่อให้ได้ผลลัพธ์ที่ตรงกับความต้องการของผู้สืบค้นมากที่สุด จากการทดสอบประสิทธิภาพของการจำแนกข้อมูล และการค้นคืนข้อมูลได้ค่าความแม่นยำ (Precision) เท่ากับ 74.51% และ 86.30% และค่าความระลึก (Recall) เท่ากับ 75.36% และ 100% ตามลำดับ ซึ่งเป็นค่าที่น่าพอใจ จึงสรุปได้ว่างานวิจัยนี้สามารถจำแนกและค้นคืนข้อมูลได้ในระดับที่ดีมาก

คำสำคัญ: การค้นคืนสารสนเทศ โครงข่ายประสาทเทียม การจำแนกข้อมูล เปอร์เซ็ปตรอนแบบหลายชั้น

Abstract

This paper proposes a web news information retrieval and classification system, using multilayer perceptron neural network. In the part of web news information classification, Rapidminer was used to model an artificial neural network (ANN). Javascript and jQuery was used to develop web Crawler for downloading data from www.pantip.com; while LexTo was used to cut stop-words and calculate word weights. The results of this serve as learning data for the ANN model. In the part of web news information retrieval, a vector space model was used to compare word similarity between words in query and documents. Apache Solr was used to create indexes in documents for improving the retrieval performance. N-Gram was also used for automatic suggestion on a set of queries; while LexTo was used for query expansion in order to get the most accurate results. The testing results of this system reveal the precision values of information classification and retrieval which are 74.51% and 86.30% respectively, and the recall values of information classification and retrieval which are 75.36% and 100% respectively. This shows that the system can be used effectively.

Keyword: Information Retrieval, Artificial Neural Network, Classification, Multilayer Perceptron.

* *ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

1. บทนำ

เว็บไซต์พันทิป (www.pantip.com) เป็นเว็บกระทู้ข่าวที่ได้รับความนิยมเป็นอย่างมาก โดยเว็บไซต์ทรูฮิต (www.truehits.net) [1] ได้เก็บสถิติผู้เข้าชมเว็บไซต์ภายในประเทศไทยประจำเดือนกุมภาพันธ์ พ.ศ. 2556 มีผู้เข้าชมเว็บไซต์พันทิปรวม 129,678 ไอพีแอดเอรส (IP Address) 200,573 เซสชั่น (Session) และจำนวนเข้าชมหน้าเว็บไซต์ทั้งหมด 688,025 ครั้ง เว็บไซต์พันทิปมีส่วนร่วมในการให้ผู้ใช้เข้ามาร่วมแสดงความคิดเห็น และมีบทบาทสำคัญในการทำหน้าที่เป็นสื่อกลางในการแสดงออกถึงความคิดเห็นเกี่ยวกับประเด็นต่างๆ ทางสังคม ปริมาณข้อมูลที่ผู้ใช้งานสร้างมีจำนวนเพิ่มมากขึ้นเรื่อยๆ และเป็นอุปสรรคสำหรับการค้นคืนข้อมูลเป็นอย่างมาก และจำเป็นต้องรู้วิธีการกรองข้อมูลที่เหมาะสมจึงจะสามารถค้นคืนข้อมูลได้รวดเร็วและแม่นยำ ข้อมูลที่ได้จากการค้นคืนจะไม่มีการจำแนกประเภทจึงเป็นอุปสรรคสำหรับการเข้าถึงข้อมูลของผู้ใช้งาน

งานวิจัยนี้จึงคิดค้นวิธีในการแก้ไขปัญหาโดยใช้โครงข่ายประสาทเทียม (Artificial Neural Network) ซึ่งเป็นกระบวนการที่คอมพิวเตอร์จำลองการทำงานโดยเลียนแบบการทำงานของสมองมนุษย์ โดยใช้วิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) เพื่อจำแนกกลุ่มข้อมูล (Classification) กระทู้ข่าวของเว็บไซต์พันทิป (www.pantip.com) เพื่อบันทึกลงฐานข้อมูลสำหรับการค้นคืนข้อมูลในขั้นตอนต่อไป โดยใช้โครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น (Multilayer Perceptron) ซึ่งเป็นรูปแบบที่เหมาะสำหรับการประมวลผลกับงานที่มีความซับซ้อนได้เป็นอย่างดี [2] ข้อมูลที่ได้จะถูกจำแนกดังนี้ สังคมและการเมือง กีฬา ท่องเที่ยว บันเทิง เทคโนโลยี การศึกษา และอื่นๆ

ผู้ใช้จะต้องกรอกข้อมูลคำถาม (Query) เข้าสู่กระบวนการค้นคืนข้อมูลสารสนเทศ (Information Retrieval) เพื่อการค้นคืนข้อมูลให้ตรงตามความต้องการของผู้ใช้งาน โดยใช้วิธีการค้นคืนแบบ (Vector Space Model : VSM) ซึ่งเป็นการแทนค่าข้อมูลคำค้นและข้อมูลเอกสารเป็น Vector โดยใช้มุมองศาเพื่อเปรียบเทียบความคล้ายกันของเอกสาร Vector เอกสารที่มีมุมองศาที่น้อยสุดจะเป็นเอกสารที่มีความคล้ายหรือใกล้เคียงกับคำค้นมากที่สุดและเรียงไปตามลำดับ [3]

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 การสร้างดัชนี (Indexing)

เป็นกระบวนการสร้างตัวแทนเอกสารเพื่อใช้สำหรับการค้นคืนข้อมูล โดยการสร้างดัชนีแบบแฟ้มผกผัน (Inverted File Index) มีสองขั้นตอนดังนี้ ขั้นแรกทำการวิเคราะห์และแบ่งคำเป็นรายการคำ (Index Term) โดยตัดข้อความที่ไม่มีความหมายหรือคำหยุดออก (Stop Word) และแปลงคำเป็นรากศัพท์ (Stemming) และขั้นที่สองนำข้อความที่ได้จากขั้นแรกมาจัดเก็บเป็นแฟ้มดัชนีพร้อมทั้งเก็บจำนวนเอกสารที่คำนั้นๆ ปรากฏและระบุหมายเลขเอกสารด้วย [3]

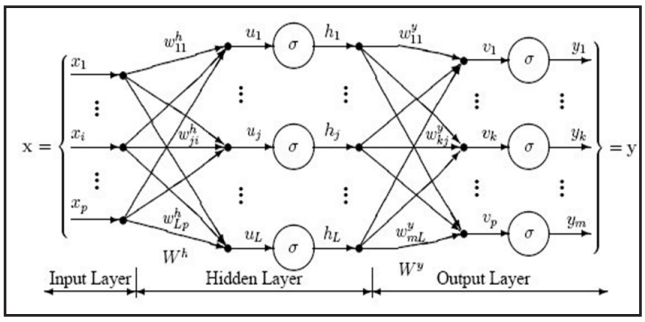
2.2 Vector Space Model

เป็นการค้นคืนเอกสารจากคลังเอกสารที่มีจำนวนมาก ซึ่งเป็นการแทนค่าข้อมูลคำค้นและข้อมูลเอกสารเป็น Vector จากนั้นทำการหาความคล้ายกันระหว่าง Vector ที่มีความคล้ายกันกับข้อมูลคำค้นมากที่สุดโดยการแทนขนาดแต่ละมิติ (Dimension) ของ Vector ด้วยค่าน้ำหนัก (Weight) แล้วเปรียบเทียบความคล้ายของ Vector (Similarity Measurement) ด้วยการทำ Inner Product หรือ Cosine Product โดยที่ Inner Product คือการเปรียบเทียบความคล้ายระหว่าง Vector ด้วยการนำ Vector มาคูณกัน ผลลัพธ์ที่ได้มีค่ามากแสดงว่ามีความคล้ายกันมาก ส่วน Cosine Product คือการเปรียบเทียบระหว่าง Vector ด้วยมุมที่มีองศาน้อยที่สุด แสดงว่ามีความคล้ายกันมากที่สุด [3]

2.3 โครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น

เป็นกระบวนการเรียนรู้ข้อมูลของเครื่อง โดยลักษณะของโมเดลจะเป็นแบบหลายชั้น มีทั้งหมด 3 Layer ได้แก่ 1) Input Layer คือชั้นของการรับข้อมูลเข้าเพื่อทำการเรียนรู้ ในชั้นนี้จะมีการกำหนดจำนวนข้อมูลนำเข้า (Input) และรูปแบบของข้อมูลตามที่ผู้ออกแบบได้ออกแบบไว้ 2) Hidden Layer เป็นชั้นของการเรียนรู้ข้อมูลโดย Hidden Layer ชั้นที่ 1 จะสุ่มค่าน้ำหนัก (Weight) ของข้อมูลนั้นๆ เพื่อเป็นการให้ความสำคัญของแต่ละข้อมูลที่ไม่เท่ากัน และสุ่มค่าความโน้มเอียง (Bias) เพื่อเป็นตัวกำหนดทิศทางการเรียนรู้ของโมเดล ส่วน Hidden Layer ชั้นที่เหลือจะมีการปรับค่าความโน้มเอียง (Bias) และค่าน้ำหนัก (Weight) ในขณะเรียนรู้ให้ได้ค่าที่

เหมาะสม ผู้ออกแบบสามารถกำหนดจำนวน Hidden Layer และจำนวนโครงข่ายประสาทเทียม (Neural Network) ในการเรียนรู้ได้ตามความเหมาะสม และ 3) Output Layer เป็นชั้นของข้อมูลผลลัพธ์จากการเรียนรู้ของโมเดล โดยข้อมูลนำเข้า (Input) ของชั้นนี้คือข้อมูลผลลัพธ์จาก Hidden Layer ซึ่งจะมีการปรับค่าความโน้มเอียง (Bias) และค่าน้ำหนัก (Weight) ในขณะเรียนรู้เช่นเดียวกับชั้น Hidden Layer ซึ่งผลลัพธ์ที่ได้นั้นจะอยู่ในลักษณะของคลาส (Class) อาจมีได้มากกว่าสองคลาส [4] ลักษณะของโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น (MLP) สามารถแสดงได้ดังภาพที่ 1



ภาพที่ 1 *โครงข่ายประสาทเทียมแบบ Multilayer Perceptron [4]*

2.2 Apache Solr

โซล่าร์ (Solr หรือ Apache Solr) [5] เป็นเครื่องมือหรือซอฟต์แวร์ตัวกลาง (Midden Ware) สำหรับค้นคืนข้อมูลโดยทำการสืบค้นข้อมูลจากดัชนีของข้อมูล (Index) [6] ที่ถูกพัฒนาจากไลบารีของระบบสืบค้นลูซีน (Lucene) [7] ใช้เอกสาร XML สำหรับการนำเข้าและส่งออกข้อมูลเพื่อแสดงผลลัพธ์การค้นคืนข้อมูล ทำให้ประสิทธิภาพการค้นคืนมีความรวดเร็วมากยิ่งขึ้น ซึ่งสามารถใช้งานข้อมูลผ่านโพรโทคอล HTTP ให้สามารถเรียกใช้งานในรูปแบบโปรแกรมงานประยุกต์ (Application Programming Interface) ได้

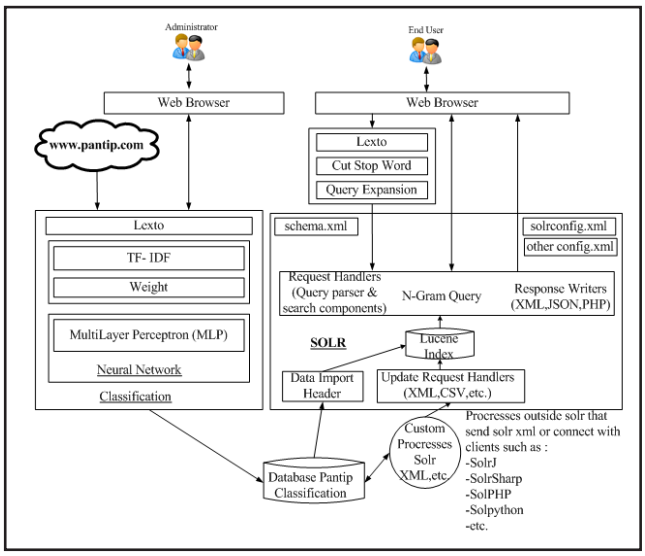
3. วิธีการดำเนินงานวิจัย

3.1 ศึกษาเครื่องมือและวิเคราะห์ปัญหาที่เกี่ยวข้อง

ศึกษาระบบค้นคืนของเว็บกระทู้ข่าวพันทิป และวิเคราะห์ปัญหาที่เกี่ยวข้องได้แก่ ผลจากการค้นคืนไม่สามารถจำแนกหมวดหมู่ และไม่สามารถค้นหาด้วยคำคล้ายได้ จากนั้นจึงได้นำมาคิดหาวิธีการแก้ไขปัญหาที่เหมาะสมและมีประสิทธิภาพมากที่สุด โดยศึกษาเครื่องมือการสร้างดัชนีของข้อมูล (Index) ในเอกสารโดยใช้ Apache Solr เพื่อใช้สำหรับกระบวนการค้นคืนข้อมูลให้มีประสิทธิภาพ ซึ่งเป็นเครื่องมือที่ได้รับความนิยมและมีประสิทธิภาพสูงอีกทั้งยังเป็น Open Source อีกด้วย ศึกษาเครื่องมือในการเรียนรู้ข้อมูลของโครงข่ายประสาทเทียมสำหรับจำแนกหมวดหมู่ข้อมูลเพื่อให้ได้โมเดลที่แม่นยำมากที่สุด โดยใช้ Rapidminer และศึกษาคลาสไลบรารีสำหรับใช้ในการตัดคำภาษาไทยอย่างมีประสิทธิภาพโดยใช้คลาสไลบรารีเล็กซ์โต (Thai Lexeme Tokenizer : LexTo) ซึ่งเป็นไลบรารีที่ถูกพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) จากการศึกษาเครื่องมือที่ใช้ในการแก้ปัญหาทั้งหมด ผู้วิจัยได้นำมาประยุกต์ใช้งานด้วยกัน พบว่าสามารถทำงานร่วมกันได้อย่างมีประสิทธิภาพ

3.2 วิเคราะห์และออกแบบระบบ

ทำการวิเคราะห์ข้อมูลเพื่อใช้ในการออกแบบระบบจำแนกและค้นคืนข้อมูลเว็บกระทู้ข่าวด้วยโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น มีรายละเอียดดังภาพที่ 2



ภาพที่ 2 *กระบวนการทำงานของระบบจำแนกและค้นคืนข้อมูลเว็บกระทู้ข่าวด้วยโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น*

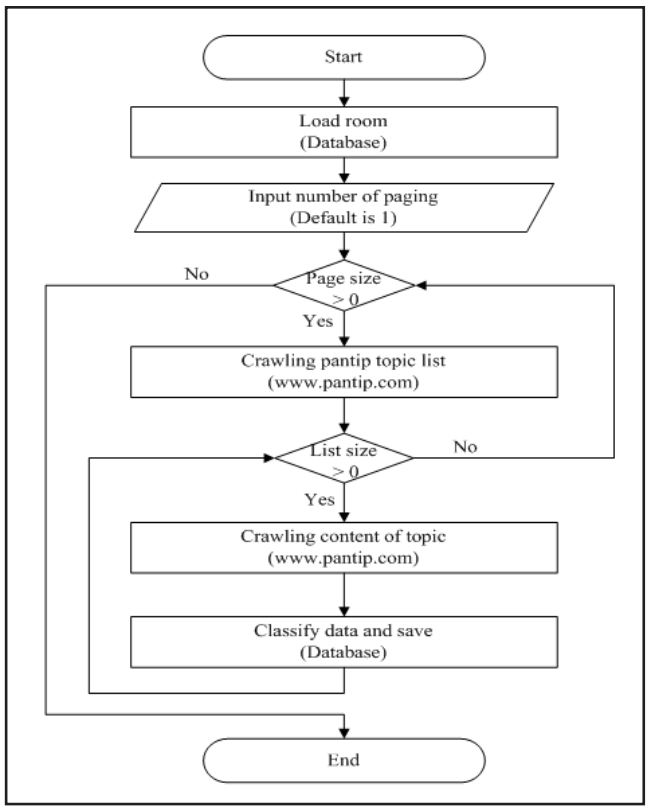
จากภาพที่ 2 ขั้นตอนการทำงานเริ่มจากการเก็บรวบรวมข้อมูลกระทู้ข่าวจากเว็บไซต์พันทิปโดยผู้ดูแลระบบ จากนั้นระบบจะตัดคำเพื่อคำนวนค่า TF ค่า IDF และค่า Weight สำหรับเป็นชุดข้อมูลในการเรียนรู้ของโครงข่ายประสาทเทียมเพื่อจำแนกลุ่มข้อมูลและเก็บลงในฐานข้อมูลต่อไป ขั้นต่อมาผู้ใช้จะกรอกข้อมูลชุดคำถาม ซึ่งระบบจะแนะนำชุดคำถามแบบอัตโนมัติให้ โดยผู้ใช้จะเลือกหรือไม่ก็ได้ จากนั้นระบบจะตัดคำหยุด (Stop Word) ออก และทำการขยายชุด

คำถาม (Query Expansion) เพื่อให้ได้ข้อมูลชุดคำถามที่เหมาะสม จากนั้นจะนำชุดข้อคำถามนี้ไปค้นหาใน Lucene ที่ทำการสร้างดัชนีข้อมูลไว้ เพื่อให้ได้ผลลัพธ์ที่แม่นยำและตรงตามความต้องการมากที่สุด

3.3 การพัฒนาระบบ

3.3.1 กระบวนการรวบรวมและจำแนกข้อมูล

การรวบรวมข้อมูลใช้ Javascript และ jQuery เป็นตัวสร้างเครื่องมือสำหรับเก็บรวบรวมข้อมูลจากเว็บไซต์พันทิป เนื่องจาก Javascript มีหลักการทำงานแบบ Asynchronous [8] จึงจำเป็นต้องใช้วิธีการเขียนโปรแกรมแบบ Recursive [9] เพื่อให้กระบวนการเก็บรวบรวมข้อมูลมีประสิทธิภาพสูงสุด โดยมีขั้นตอนการทำงานดังภาพที่ 3



***ภาพที่ 3** ขั้นตอนการทำงานของระบบเก็บรวบรวมข้อมูล*

จากภาพที่ 3 เป็นการอธิบายขั้นตอนการทำงานของระบบเก็บรวบรวมข้อมูลกระทู้ข่าวจากเว็บไซต์พันทิป ขั้นแรกโหลดข้อมูลห้องข่าวจากฐานข้อมูลที่ได้รวบรวมไว้ โดยมีทั้งหมด 28 ห้องข่าว ขั้นที่ 2 กำหนดจำนวน Paging ที่ต้องการดาวน์โหลดซึ่งแต่ละ Paging จะมีค่าเท่ากับ 50 เนื้อหา ขั้นที่ 3 ทำการเก็บรวบรวมข้อมูลหัวข้อกระทู้ข่าวจากเว็บไซต์พันทิปโดยกำหนดให้ทำงานแบบคู่ขนาน (Parallel) เพื่อให้สามารถทำงานได้อย่างมีประสิทธิภาพไม่ต้องรอห้องใดห้องหนึ่งทำเสร็จก่อน ขั้นที่ 4 เก็บรวบรวมข้อมูลเนื้อหาของแต่ละหัวข้อข่าวที่ถูกส่งมาจากขั้นที่ 3 เพื่อใช้เป็นชุดข้อมูลในการจำแนกกลุ่มต่อไป

การจำแนกข้อมูลได้นำกระบวนการเรียนรู้ข้อมูลของเครื่อง (Machine Learning) แบบมีผู้สอน (Supervised Learning) [4] ในการเรียนรู้ข้อมูลเพื่อจำแนกกลุ่ม โดยใช้ Rapidminer เป็นเครื่องมือสร้างโมเดลโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น (MLP) สำหรับจำแนกกลุ่มของกระทู้ข่าวที่ได้จากขั้นตอนที่ 4 เพื่อบันทึกลงในฐานข้อมูล มีข้อมูลนำเข้า (Input Layer) เท่ากับ 1,007 ข้อมูล และใช้โครงข่ายประสาทเทียมแบบ 17-17-17-7 นั่นคือมี Hidden Layer ทั้งหมด 3 ชั้น โดย Hidden Layer ชั้นที่ 1 ชั้นที่ 2 และชั้นที่ 3 แต่ละชั้นมีจำนวน 17 Nodes และชั้นผลลัพธ์ (Output Layer) มีจำนวน 7 Nodes ด้วยกันโดยที่ทุกชั้นของโครงข่ายประสาทเทียมจะใช้ฟังก์ชันการกระตุ้นแบบซิกมอยด์ฟังก์ชัน (Sigmoid Function) ซึ่งได้ค่าประสิทธิภาพของโครงข่ายประสาทเทียมดังตารางที่ 1

***ตารางที่ 1** ตารางการประเมินประสิทธิภาพการค้นคืนข้อมูล*

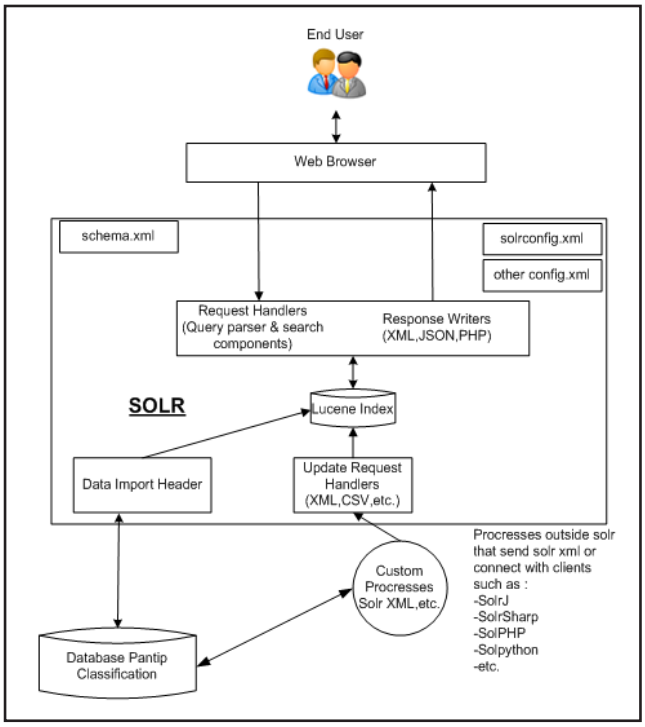
	การเมือง	กีฬา	ท่องเที่ยว	บันเทิง	เทคโนโลยี	การศึกษา	อื่น ๆ
การเมือง	102	5	3	4	1	3	11
กีฬา	3	101	3	2	2	1	6
ท่องเที่ยว	4	7	81	0	6	2	13
บันเทิง	1	0	2	154	5	1	2
เทคโนโลยี	2	1	2	6	74	0	6
การศึกษา	6	0	2	9	2	51	4
อื่น ๆ	7	5	11	2	5	5	25

จากตารางที่ 1 ประสิทธิภาพของโครงข่ายประสาทเทียมแบบ 17-17-17-7 มีค่าความถูกต้อง (Accuracy) เท่ากับ 74.51% ค่าความระลึก (Recall) เท่ากับ 75.36% และค่าเฉลี่ยผิดพลาดกำลังสอง (Mean Squared Error: MSE) เท่ากับ 0.442

3.3.2 กระบวนการค้นคืนข้อมูลแบบ (VSM)

การค้นคืนข้อมูลใช้หลักการ Vector Space Model ใน

การค้นคืนข้อมูลโดยใช้ Apache Solr ในการสร้างดัชนีของข้อมูล (Index) สำหรับใช้ในการค้นคืนข้อมูลให้มีประสิทธิภาพสูงสุดซึ่งมีกระบวนการทำงานดังภาพที่ 4



ภาพที่ 4 *กระบวนการทำงานของ Apache Solr*

จากภาพที่ 4 แสดงถึงโครงสร้างการทำงานของ Apache Solr โดยเริ่มจากผู้ใช้ป้อนข้อมูลชุดคำถามผ่านเว็บบราวเซอร์ (Web Browser) โดยระบบจะทำการแปลงชุดคำถามโดยผ่าน Query Parser เพื่อตรวจสอบประเภทและฟิวส์ของข้อมูลที่ตั้งค่าไว้ใน Apache Solr จากนั้นนำชุดข้อคำถามนี้ไปค้นหาใน Lucene ที่ทำการสร้างดัชนีข้อมูลไว้ และส่งผลลัพธ์กลับมายังผู้ใช้โดยจะแปลงข้อมูลให้เป็นประเภทตามที่ตั้งค่าไว้

4. ผลการดำเนินงาน

งานวิจัยนี้มีการทดสอบประสิทธิภาพด้วยค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่า F-Measure [10] ซึ่งเป็นที่ยอมรับ และนิยมใช้กันอย่างแพร่หลายเพื่อทดสอบความถูกต้องของข้อมูล

4.1 ผลการประเมินประสิทธิภาพการค้นคืน

ประสิทธิภาพการค้นคืนข้อมูลทดสอบจากข้อมูลชุดคำถาม 12 ชุดคำถามได้ผลการทดสอบดังตารางที่ 2

ตารางที่ 2 *ตารางการประเมินประสิทธิภาพการค้นคืน*

Keyword	Precision	Recall
เลือกตั้งผู้ว่ากทม	0.93	1
ประชุมสภา	0.72	1
ศึกแดงเดือด	0.95	1
นักกีฬาดีเด่น	0.89	1
เที่ยวเชียงใหม่	0.84	1
น้ำตกสวย	0.91	1
อ่านการ์ตูน	0.82	1
ละครไทยทำไมต้องน้ำเน่า	0.83	1
อยากได้มือถือใหม่	0.88	1
ราคา iphone	0.89	1
โรงเรียนน่าอยู่	0.87	1
นักเรียนนักศึกษา	0.94	1
ค่าเฉลี่ยรวม	**0.863**	**1**

จากการทดสอบพบว่าได้ค่า Precision เท่ากับ 86.30% ค่า Recall เท่ากับ 100% และค่า F-Measure เท่ากับ 80%

4.2 ผลการประเมินประสิทธิภาพการจำแนกข้อมูล

ประสิทธิภาพของโมเดลจากการเรียนรู้ได้ค่า Precision เท่ากับ 74.51% ค่า Recall เท่ากับ 75.36% และค่า Mean Square Error (MSE) เท่ากับ 0.44 ส่วนประสิทธิภาพของโมเดลเมื่อทดสอบกับข้อมูลจริงจำนวน 1,000 ข้อมูลได้ค่า Precision เท่ากับ 75.41 % ค่า Recall เท่ากับ 75.38 % จึงสรุปได้ว่าประสิทธิภาพการจำแนกข้อมูลอยู่ในระดับที่ดี

5. สรุปผลและข้อเสนอแนะของการวิจัย

งานวิจัยนี้ได้ถูกพัฒนาขึ้นโดยการประยุกต์ใช้ทฤษฎีการค้นคืนข้อมูลแบบ Vector Space Model โดยใช้ Apache Solr ในการสร้างดัชนีข้อมูลสำหรับใช้ในการค้นคืนข้อมูล ซึ่งใช้ Lexto ในการตัดคำของชุดคำถามเพื่อนำไปขยายข้อคำถาม (Query Expansion) ให้ได้ข้อมูลที่เกี่ยวข้องมากที่สุดผลจากการทดสอบประสิทธิภาพได้ค่า Precision เท่ากับ 86.30% ค่า Recall เท่ากับ 100% ส่วนโครงข่ายประสาทเทียมแบบเปอร์เซ็ปตรอนหลายชั้น ถูกนำมาใช้สำหรับการจำแนก

ข้อมูลโดยใช้ Rapidminer ในการสร้างโมเดลการจำแนกข้อมูลนี้ มีข้อมูลนำเข้า (Input Layer) เท่ากับ 1,007 ข้อมูล และใช้โครงข่ายประสาทเทียมแบบ 17-17-17-7 นั่นคือมี Hidden Layer ทั้งหมด 3 ชั้น โดย Hidden Layer ชั้นที่ 1 ชั้นที่ 2 และชั้นที่ 3 แต่ละชั้นมีจำนวน 17 Nodes และชั้นผลลัพธ์ (Output Layer) มีจำนวน 7 Nodes ด้วยกัน โดยที่ทุกชั้นของโครงข่ายประสาทเทียมจะใช้ฟังก์ชันการกระตุ้นแบบซิกมอยด์ฟังก์ชัน (Sigmoid Function) ซึ่งจากการทดสอบประสิทธิภาพของโมเดลจากข้อมูลจริงจำนวน 1,000 ข้อมูล ได้ค่า Precision เท่ากับ 75.41% ค่า Recall เท่ากับ 75.38% ทำให้ระบบมีประสิทธิภาพในการจำแนกและค้นคืนข้อมูลอยู่ในระดับที่ดีมาก

งานวิจัยนี้มีข้อจำกัดอยู่ 2 ประการดังนี้ ประการแรกคือ ระบบไม่สามารถค้นคืนข้อมูลแบบ Real Time ได้เนื่องจากข้อมูลหลักจะอยู่ที่เว็บไซต์พันทิปจำเป็นต้องทำการเก็บรวบรวมข้อมูลเป็นระยะๆ ประการที่ 2 ระบบไม่สามารถจำแนกคำที่เขียนผิดได้ อย่างไรก็ตามในอนาคตงานวิจัยนี้พัฒนาระบบเพิ่มเติม ให้สามารถรองรับกับการค้นคืนข้อมูลด้วยการประมวลผลภาษาธรรมชาติ และสามารถเรียนรู้ข้อความที่เขียนผิดได้

6. เอกสารอ้างอิง

[1] Page Of Truehits, Available online at http://truehits.net/script/201302/ rank0.php [Access: 20 March 2013]

[2] เอกรินทร์ แซ่เฮ้ง, “โครงข่ายประสาทเทียมกับการประยุกต์ใช้งาน (ตอนที่ 1 รู้จักกับโครงข่ายประสาทเทียม).” แผนกสารสนเทศ สำนักวิชาการวิทยาลัยนอร์ทกรุงเทพ, 2010.

[3] W. B. Frakes and R. Baeza-Yates, eds., *Information Retrievall: Data Structures & Algorithms*, Pretice Hall, 1992.

[4] พยุง มีสัจ. *ระบบฟัซซีและโครงข่ายประสาทเทียม*. คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2555.

[5] Main page of Solr, Available online at http://lucene.apache.org/solr/ [Access: 20 March 2013]

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieva*l, ACM Press, Addison Wesley, 1999.

[7] Main page of Apache Lucene, Available online at http://lucene.apache.org [Access: 20 March 2013]

[8] PageOfMicrosoft , Available online at http://blogs.msdn.com/b/ie/archive/2011 /09/11/asynchronous-programming-in-javascript-with-promises.aspx [Access: 20 March 2013]

[9] Page Of Wikipedia, Available online at http://en.wikipedia.org/wiki /Recursion_(computer_science) [Access: 20 March 2013]

[10] A. Kongthon, C. Haruechaiyasak, C. Sangkeettrakarn, P. Palingoon, and W. Wunnasri, “HotelOpinion: An opinion mining system on hotel reviews in Thailand.” *Technology Management in the Energy Smart World (PICMET), 2011 Proceedings of PICMET ‘11. Portland : Portland State University*, pp. 1-6, 2011.