การแบ่งกลุ่มข้อความ SMS ตามลักษณะการให้บริการ

ภูรดา นนทวาสี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2554

Service-Oriented Classifying of SMS Message

BHURADA NONTHAWASEE

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Computer and Telecommunication Engineering

Graduate School, Dhurakij Pundit University

2011

หัวข้อวิทยานิพนธ์	การแบ่งกลุ่มข้อความ SMS ตามลักษณะการให้บริการ
ชื่อผู้เขียน	ภูรดา นนทวาสี
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ดร.ชัยพร เขมะภาตะพันธ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์และโทรคมนาคม
ปีการศึกษา	2554

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการแบ่งประเภทของข้อความ SMS โดยใช้วิธีการ Naïve Bayesian โดยพิจารณาจากเนื้อหาของข้อความ SMS เพื่อแก้ปัญหาความคับคั่งของ SMSC เมื่อมีปริมาณผู้ส่งเป็นจำนวนมาก โดยมีประโยชน์ต่อระบบ SMSC ในการสร้างลำดับการส่งใหม่ตามระดับความสำคัญของข้อความ SMS ที่กำหนดขึ้น ทำให้ลดอัตราเสี่ยงที่ระบบจะเกิดการ overload และการไม่สามารถให้บริการได้ ผลการศึกษายังสามารถใช้เป็นพื้นฐานเพื่อพัฒนาระบบคัดแยกระดับความสำคัญข้อความ SMS ที่จะนำไปใช้งานเชิงพาณิชย์ สำหรับผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยต่อไปในอนาคต จากผลการทดสอบแสดงให้เห็นว่า วิธีการคัดแบ่งประเภทของข้อความที่นำเสนอใช้เวลาในการทำงานน้อยกว่าการคัดกรองแบบเดิมถึง 6% นอกจากนี้ยังมีความถูกต้องในการทำงานสูงกว่า 13.59%

Thesis Title	Service-Oriented Classifying of SMS Message
Author	Bhurada Nonthawasee
Thesis Advisor	Chaiyaporn Khemapatapan, Ph.D
Department	Computer and Telecommunication Engineering
Academic Year	2011

ABSTRACT

This research proposed a technique for classification of SMS messages by using Naïve Bayesian algorithm and the behavior of the content in an SMS message. The proposed system will be useful for SMSC by rearranging the order of sending an SMS according to the priority of an SMS type that is defined. This aims to solve the congestion at the SMSC when bulk subscribers are trying to send their SMSs. The system reduces the risks of overloading and denial of service. The studied results can also be used as a basis to develop the commercial system that is used to classify type of SMS messages for mobile service operators in Thailand in the future. The results show that the proposed system performs 6% faster and 13.59% higher accuracy than the conventional one.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยความกรุณาเป็นอย่างยิ่งจาก อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่คอยให้คำแนะนำ ตลอดจนเปิดโลกทรรศน์ ในการค้นคว้าข้อมูลให้แก่ข้าพเจ้า ขอขอบพระคุณท่านอาจารย์กรรมการสอบวิทยานิพนธ์ ซึ่งสละเวลามาเป็นกรรมการสอบวิทยานิพนธ์และได้ให้ข้อคิดเห็นที่เป็นประโยชน์ต่องานวิจัยในครั้งนี้แก่ข้าพเจ้า รวมทั้งให้ความรู้อันเป็นประโยชน์แก่ข้าพเจ้ามาโดยตลอด นอกจากนี้

ผู้วิจัยขอขอบพระคุณ คณาจารย์ทุกๆ ท่านในคณะวิศวกรรมศาสตร์ที่ได้ถ่ายทอดความรู้แก่ผู้วิจัยตลอดระยะเวลาการศึกษา

ผู้วิจัยขอขอบพระคุณ เจ้าหน้าที่ที่เกี่ยวข้องทุกท่าน ในคณะวิศวกรรมศาสตร์ที่คอยให้ความช่วยเหลือ ตลอดจนแนะนำกระบวนการในการทำงานให้แก่ผู้วิจัยด้วยดีเสมอมา

ผู้วิจัยขอขอบพระคุณ บริษัท กสท. โทรคมนาคม จำกัด ที่ให้ข้อมูลและข้อแนะนำช่วยเหลือและส่งเสริมในการทำวิจัยครั้งนี้จนสำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอขอบพระคุณ คุณ นนท์ บุญนิธิประเสริฐ บริษัท กสท. โทรคมนาคม จำกัด ที่ให้คำแนะนำ ช่วยเหลือและส่งเสริมในการทำวิจัยครั้งนี้จนสำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอขอบพระคุณ คุณ เปรมฤดี ผลชอบ บริษัท กสท. โทรคมนาคม จำกัด ที่ให้ความช่วยเหลือและส่งเสริมในการทำวิจัยครั้งนี้จนสำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอขอบคุณเพื่อนๆ ร่วมรุ่นและรุ่นพี่ทุกคน ที่ช่วยเหลือและให้กำลังใจกันเสมอมาตลอดระยะเวลาการศึกษา

ท้ายสุดนี้ ผู้วิจัยขอกราบขอบพระคุณ คุณแม่และครอบครัว ที่คอยเป็นกำลังใจและให้การสนับสนุนผู้วิจัยในทุกๆ ด้านเสมอมาจวบจนสำเร็จการศึกษา

ภูรดา นนทวาสี

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญภาพ

หน้า

ภาพที่

รายการคำย่อ

BTS	Base Trans-receiver Stations
CDMA	Code Division Multiple Access
CDR	Call Detail Records
DB	Database
HLR	Home Location Register
HTTP	Hyper-Text Transfer Protocol
IP	Internet Protocol
MSC	Mobile Switching Center
NB	Naïve Bayesian
ND	New Data
PDU	Protocol Data Unit
PHP	Personal Home Page
SMPP	Short Message Peer to Peer Protocol
SMS	Short Message Service
SMSC	Short Message Service Centre
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TD	Training Data
TN	Text Normalization

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

บริการเสริมที่มีผู้ใช้เป็นจำนวนมากที่สุดบริการหนึ่งของระบบโทรศัพท์เคลื่อนที่ในปัจจุบันคือ บริการส่งข้อความสั้นหรือ Short message service (SMS) ซึ่งเป็นบริการที่สามารถเข้าถึงผู้ใช้งานได้รวดเร็ว จึงมีการนำบริการดังกล่าวมาเป็นเครื่องมือในการแจ้งรายงานหรือประชาสัมพันธ์ในธุรกิจ ซึ่งเป็นช่องทางหนึ่งที่มีประสิทธิภาพและถูกนำมาใช้อย่างกว้างขวาง โดยมีอัตราส่วนแบ่งทางการตลาดเฉพาะยุโรปตะวันตกร้อยละ 80 เมื่อเทียบกับ Data Service แบบอื่นๆ พบว่ามีการส่งข้อความ SMS จำนวน 200,000 ล้านครั้งต่อเดือน[1] บริษัท China Unicom ในประเทศจีนรายงานว่า มีการส่ง SMS ในปี 2005 จำนวน 304,140 ล้านครั้งต่อเดือน[2] ดีแทคซึ่งเป็นผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทย รายงานสถิติการใช้ SMS อวยพรปีใหม่ระหว่างวันที่ 31 ธันวาคม 2550 ถึงวันที่ 1 มกราคม 2551[3] จำนวน 38 ล้านข้อความ เพิ่มขึ้นจากช่วงเดียวกันของปีที่ผ่านมาร้อยละ 32 โดยมีการส่งสูงสุด 4 ล้านข้อความต่อชั่วโมง ในช่วงรอยต่อระหว่างวันที่ 31 ธ.ค. 2550 กับ วันที่ 1 ม.ค. 2551 ข้อมูลจากหนังสือพิมพ์ไทยรัฐ ฉบับวันที่ 7 มกราคม 2552 แสดงอัตราการส่งข้อความสั้นของบริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) ในช่วงเทศกาลปีใหม่ 2552 อยู่ที่ 49.95 ล้านข้อความ เพิ่มขึ้นจากช่วงเดียวกันของปี 2551 ร้อยละ 31 โดยมีการส่ง SMS สูงสุด 8.9 ล้านข้อความต่อชั่วโมง ในช่วงระหว่างเวลา 00.00 น. ถึงเวลา 01.00 น. และบริษัทแอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) มีอัตราการส่งข้อความ SMS เพิ่มขึ้นจากช่วงเดียวกันของปี 2551 ร้อยละ 25[4]

เนื้อหาของข้อความมีหลายประเภท เช่น งานด้านการเงิน,การศึกษา รวมถึงเป็นสื่อโฆษณาประชาสัมพันธ์ในงานด้านต่างๆ ซึ่งข้อความบางประเภทมีความสำคัญต่อธุรกิจ แต่บางประเภทกลับจัดเป็นข้อความขยะ (Spam SMS) บางครั้งการส่งข้อความในปริมาณมากๆ นั้นส่ง

---

[1] Siddharth Dixit, Sandeep Gupta and Chinya V.Ravishankar. (November 14-16, 2005). "LOHIT: An Online Detection & Control System for Cellular SMS Spam." **Proceeding of the IASTED International Conference on Communication, Network, and Information Security.** p. 48-54. Phoenix, AZ, USA.

[2] Petros Zerfos, XiaoqiaoMeng, Starsky H.Y Wong, VidyutSamanta, Songwu Lu. (October 25-27, 2006) "A Study of the Short Message Service of a Nationwide Cellular Network". **Proceedings of the Internet Measurement Conference: IMC 2006**. p. 263-268. Rio de Janeiro, Brazil.

[3] ไทยซ่าส์, (2551). สถิติการใช้บริการส่งข้อความอวยพรปีใหม่. สืบค้นเมื่อ 27 พฤษภาคม 2552, จาก **http://phone.thaiza.com/A1_1212_83502_1212_.html.**

[4] ไทยรัฐ, (2552). ข่าวด้านเทคโนโลยี.สืบค้นเมื่อ 27 พฤษภาคม 2552. จาก **http://thairath.co.th/news.php?section=technology03b&content=118050**

ผลกระทบหลายๆด้าน ยกตัวอย่างเช่นกรณีการส่ง SMS ของบริษัท ทรู คอร์ปอเรชั่น ส่งเสริมการขาย SIM Card สำหรับโหวตรายการวาไรตี้โชว์ ไปยังเครื่องโทรศัพท์ลูกข่ายของระบบดีแทคจำนวนมากติดต่อกัน จนทำให้เกิดการปิดรับ SMS ระหว่าง 2 เครือข่ายเป็นเวลา 16 ชั่วโมง[5] และกรณีการส่ง SMS อวยพรปีใหม่ของผู้ใช้บริการโทรศัพท์เคลื่อนที่ในประเทศอินเดียที่มากกว่า 4 แสนข้อความ ต่อนาที จนทำให้ระบบหยุดทำงาน[6] เป็นต้น จะเห็นได้ว่านอกจากจะรบกวนการใช้งานของผู้ใช้โทรศัพท์แล้ว ยังส่งผลกระทบต่อการทำงานของระบบเครือข่าย SMS ทำให้ผู้ให้บริการหลายรายต้องการที่จะแก้ไขปัญหาด้านข้อความ SMS ที่จัดว่าอยู่ในกลุ่มที่สร้างความรบกวนให้กับผู้ใช้บริการของตนเอง แต่ในขณะเดียวกันผู้ให้บริการก็เป็นผู้ขายบริการการส่ง SMS เหล่านั้นนั้นเอง เมื่อผู้ให้บริการคิดที่จะทำการกรอง SMS ที่ตนคิดว่ารบกวนผู้ใช้บริการ ก็กลายเป็นว่าผู้ให้บริการลดความน่าเชื่อถือในระบบของตนลง ส่งผลถึงรายได้ที่ตามมา และเมื่อมีผู้ใช้บริการบางรายต้องการที่จะรับ SMS เหล่านั้น ก็จะส่งผลกระทบต่อความน่าเชื่อถือในระบบของผู้ให้บริการอีกต่อหนึ่ง จึงจำเป็นต้องมีการปรับปรุงระบบส่งข้อความสั้นเพื่อลดปัญหาดังกล่าว แต่รายงานการวิจัยหลายฉบับมุ่งเน้นที่จะแก้ไขปัญหาข้อความ Spam ในระบบมากกว่าที่จะคัดแยกระดับความสำคัญของ SMS

ดังนั้นจากปริมาณข้อความที่มีจำนวนมากเหล่านี้แต่ละข้อความก็จะมีระดับของความสำคัญที่ไม่เท่ากันขึ้นอยู่กับรูปแบบการใช้งานของธุรกิจนั้นๆ เช่น ธนาคารทหารไทย กำลังจะเปิดให้บริการจ่ายค่าสาธารณูปโภคผ่านทางธนาคาร ซึ่งลักษณะของการให้บริการคือ เมื่อครบรอบของการให้บริการระบบจะส่ง SMS มายังลูกค้าเพื่อให้ลูกค้ายืนยันการจ่ายเงิน และเมื่อจ่ายเงินเรียบร้อยแล้วระบบก็จะส่ง SMS แจ้งอีกครั้งหนึ่ง ดังนั้นข้อความเหล่านี้จะต้องส่งถึงลูกค้าโดยทันทีจะสูญหายหรือส่งล้าช้าไม่ได้ ระบบก็ไม่สามารถตัดเงินเพื่อจ่ายค่าสาธารณูปโภคต่างๆ ได้ เพราะไม่มี SMS มายืนยัน ส่งผลกระทบต่อลูกค้าเป็นอย่างมากเนื่องจากไม่สามารถจ่ายค่าสาราณูปโภคได้ ยิ่งไปกว่ายังส่งผลต่อความน่าเชื่อของผู้ให้บริการ และอาจจะทำให้ลูกค้าไม่ต้องการใช้บริการกับเราอีก ดังนั้นจึงต้องมีระบบในบริหารจัดการการส่งข้อความที่มีประสิทธิภาพ โดยเฉพาะในช่วงที่มีปริมาณการส่งSMS จำนวนมาก จะส่งผลกระทบต่อประสิทธิภาพการส่งข้อความ และอาจจะทำให้ข้อความบางส่วนสูญหายได้ จึงต้องมีระบบในการคัดแยกข้อความ ซึ่งทำหน้าที่ในการคัดแยกข้อความตามระดับความสำคัญ ซึ่งอาจจะพิจารณาจากหลายๆองค์ประกอบด้วยกัน ไม่ว่าจะเป็นผู้ส่ง

---

[5] ผู้จัดการ, (2549). แก้ปัญหา SPAM SMS บล็อก.สืบค้นเมื่อ 27 พฤษภาคม 2552. จาก **http://www.manager.co.th/Cyberbiz/ViewNews.aspx?NewsID=9490000090269**

[6] Petros Zerfos, XiaoqiaoMeng, Starsky H.Y Wong, VidyutSamanta, Songwu Lu. (October 25-27, 2006) "A Study of the Short Message Service of a Nationwide Cellular Network". **Proceedings of the Internet Measurement Conference: IMC 2006**. p. 263-268. Rio de Janeiro, Brazil.

เนื้อหาของข้อความเป็นต้น ทั้งนี้เพราะมีผู้ซื้อบริการ SMS จำนวนมากที่ซื้อบริการจากผู้ให้บริการโทรศัพท์มือถือ เพื่อนำไปสร้างบริการอีกต่อหนึ่ง

ผู้ใช้บริการ SMS นั้นมีปริมาณมากหลากหลายกลุ่ม โดยสามารถจำแนกกลุ่มของผู้ใช้บริการออกเป็น 4 ประเภทดังนี้

(1) ผู้ให้บริการต่างๆ เช่น ผู้ให้บริการโทรศัพท์มือถือ

(2) ธุรกิจ บริษัท เช่น บริษัทขายตรง

(3) องค์กร ภาครัฐ เช่น โรงพยาบาล หน่วยงานของรัฐ

(4) บุคคลทั่วไป

ซึ่งแต่ละกลุ่มผู้ใช้บริการก็จะมีลักษณะของการใช้บริการรับ-ส่งข้อความแตกต่างกันออกไปตามวัตถุประสงค์ของการใช้งาน ซึ่งสามารถแบ่งประเภทของข้อความออกเป็นประเภท ได้ดังนี้

(1) ข้อความโฆษณาประชาสัมพันธ์ เป็นข้อความเสนอขายสินค้าหรือบริการต่างๆ

(2) ข้อความทั่วไป เป็นข้อความทั่วไปที่มีการส่งในชีวิตประจำวัน

(3) ข้อความด้านการบริการ ส่วนใหญ่เป็นข้อความทางด้านการแพทย์,การศึกษา,การใช้งานระบบต่างๆ

(4) ข้อความแจ้งเตือน หรือข้อความที่ต้องรับรู้ ที่อาจส่งผลต่อการดำเนินธุรกิจหรือการใช้ชีวิตประจำวันต่างๆ เช่น การเงิน แจ้งเตือนภัยต่างๆ

ตารางที่ 1. 1 ตัวอย่างประเภทของข้อความ SMS

ประเภท	**ข้อความ**
ข้อความแจ้งเตือน	โอนเงินเข้า KBank 2767XXXX ผ่าน K-ATM 1,800 บ.
ข้อความด้านการบริการ	2-5 missed call from 0867367583@ 30/07/2010 17:44
ข้อความทั่วไป	จองหนังให้แล้วนะ รีบมาด้วย
ข้อความโฆษณาประชาสัมพันธ์	ดาวน์โหลดริงโทน พิมพ์ ok ส่งที่ *123456

การแก้ปัญหาดังกล่าวสามารถทำได้หลายรูปแบบ เช่น การใช้ Application กรองข้อความ Spam ซึ่งช่วยลดปริมาณข้อความ SMS ที่ลงได้ แต่ไม่สามารถแก้ไขปัญหาการลดทอนความน่าเชื่อถือในตัวระบบของผู้ให้บริการ และปริมาณข้อความ SMS ในระบบส่งข้อความสั้นของผู้ให้บริการได้ นอกจากนี้ผู้ให้บริการโทรศัพท์เคลื่อนที่บางรายได้มีการเพิ่มมาตรการระบุตัวตนผู้

ส่ง ด้วยการทำ Authentication ก่อนการอนุญาตให้ส่งข้อความ SMS ระหว่างผู้ให้บริการโทรศัพท์เคลื่อนที่และผู้ให้บริการ Content ซึ่งมาตรการนี้มีวัตถุประสงค์เพื่อให้สามารถระบุผู้ส่งข้อความและใช้เป็นข้อมูลอ้างอิงทางกฎหมาย โดยไม่มีการกรองข้อความ Spam ออกจากระบบส่งข้อความแต่ก็ไม่สามารถลดปริมาณข้อความที่ถูกส่งในช่วงเวลาที่มีการใช้งานระบบคับคั่งได้ แนวทางการแก้ปัญหาหนึ่งที่ถูกนำมาพิจารณาในระดับสากลคือ การกรองข้อความ Spam ออกจากระบบที่ SMSC เพื่อลดความคับคั่งของข้อมูลด้วยการใช้การกรองข้อความ Spam (Filtering) โดยให้ Filter ทำการกรองข้อความที่ SMSC ก่อนการส่งข้อความ

การส่งข้อความในระบบ SMS มีรูปแบบการส่ง 2 วิธีได้แก่

(1) การส่งข้อความแบบ Signaling (SS7)

(2) การส่งข้อความผ่าน TCP Protocol

โดยทั่วไปผู้ซื้อบริการ SMS จะใช้ Robot Software เข้ามาช่วยให้ทำการส่งข้อความ ที่มีลักษณะส่งครั้งละหลายข้อความและหลายปลายทาง Software ประเภทนี้สามารถทำงานร่วมกับ TCP Protocol ได้สะดวกรวดเร็วกว่าการทำงานผ่าน Signaling ที่ต้องอาศัย Software สั่งการทำงานไปยังเครื่องโทรศัพท์เคลื่อนที่ผ่าน AT Command นอกจากนี้ระบบโทรศัพท์เคลื่อนที่ในยุคต่อไปจะทำงานบนพื้นฐาน TCP/IP เพียงอย่างเดียว ทำให้การคัดแยกข้อความบน TCP Protocol ที่มีความซับซ้อนน้อยกว่า สามารถครอบคลุมการใช้งานที่จะเกิดขึ้นในอนาคต

จากความเป็นมาข้างต้นผู้วิจัยจึงนำเสนอวิธีการแก้ปัญหาความคับคั่งของการส่งข้อความและสร้างความน่าเชื่อถือให้กับผู้ให้บริการ โดยนำเสนอวิธีการคัดแยกข้อความตามระดับและประเภทของความสำคัญแล้วทยอยส่งหาลูกค้า ซึ่งหลักการทำงานของงานวิจัยนี้คือ ก่อนที่ระบบจะส่งข้อความออกไปยัง SMS Gateway นั้น ต้องนำข้อความเหล่ามาผ่านกระบวนการ normalize ข้อความหรือการเตรียมข้อความก่อน หลังเตรียมข้อความเรียบร้อยแล้ว จะนำข้อความมาผ่านกระบวนการคัดแยกโดยใช้เทคนิค Naïve Bayes เพื่อจัดประเภทของข้อความออกเป็นกลุ่มๆ หลัก 4 กลุ่มแล้วจึงส่งข้อความออกไปหาลูกค้าตามลำดับความสำคัญต่อไป

1.2 วัตถุประสงค์ของการวิจัย

1. จัดประเภทของข้อความภาษาไทย สำหรับบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่ เพื่อลดภาระในระบบส่งข้อความสั้นให้สามารถรองรับการทำงานได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

2. พัฒนาวิธีการคัดแยกประเภทของข้อความภาษาไทยโดยให้ความสำคัญกับเนื้อหา (Content) ของข้อความเป็นหลัก

3. จำลองสถานการณ์รับส่งข้อความสั้นผ่านระบบการจัดประเภทของข้อความเพื่อวัดความถูกต้องในการจัดระดับความสำคัญของข้อความที่มีได้

4. จำลองสถานการณ์รับส่งข้อความสั้นผ่านระบบการจัดประเภทของข้อความ เพื่อวัดประสิทธิภาพการเพิ่มความเร็วในการส่งข้อความสำคัญ และหน่วงเวลาในการส่งข้อความที่มีระดับความสำคัญต่ำเมื่อมีปริมาณการใช้งานของระบบมากขึ้น

1.3 สมมติฐานของการวิจัย

1. ออกแบบและพัฒนาวิธีการคัดแยกประเภทของข้อความสั้น SMS ให้สามารถใช้งานกับข้อความ SMS ภาษาไทย ภาษาอังกฤษ และภาษาไทยปนอังกฤษได้

2. สามารถคัดประเภทของข้อความได้อย่างถูกต้อง

3. สามารถเพิ่มประสิทธิภาพของการส่งข้อความที่มีความสำคัญให้กับลูกค้าได้อย่างรวดเร็ว

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถคัดแยกประเภทของข้อความภาษาไทยได้อย่างถูกต้องและเพิ่มประสิทธิภาพการทำงานของบริการส่งข้อความสั้น SMS ให้สามารถรองรับการส่งข้อความในปริมาณที่เพิ่มขึ้น

2. ลดอัตราเสี่ยงที่ระบบส่งข้อความสั้นจะเกิดการส่งข้อความมากเกินไป (overload) จนไม่สามารถทำงานต่อไปได้

3. ผลการศึกษายังสามารถใช้เป็นพื้นฐานเพื่อพัฒนาระบบคัดแยกระดับความสำคัญข้อความที่จะนำไปใช้งานเชิงพาณิชย์ สำหรับผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยต่อไป

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงองค์ประกอบของการส่ง SMS หลักการ ทฤษฎีและขั้นตอนวิธีการที่นำมาประยุกต์ใช้ในกระบวนการคัดแยกประเภทของข้อความเพื่อนำมาจัดลำดับตามความสำคัญก่อนส่ง

2.1 องค์ประกอบของการใช้บริการ SMS

2.1.1 เนื้อหาหรือข้อความ (SMS)

SMS หรือ การส่งข้อความสั้น โดยลักษณะของการส่งข้อความสั้นจะมีลักษณะคล้ายกับการส่งข้อความไปยังเพจเจอร์ คือ ผู้ใช้สามารถส่งข้อความไปยังผู้รับ โดยที่ผู้รับสามารถกดอ่านได้จากเครื่องโทรศัพท์มือถือได้ทันที ข้อดีของ SMS ที่ทำให้ต่างกับเพจเจอร์ก็คือ ผู้ใช้หรือผู้ที่ต้องการส่งข้อความสามารถพิมพ์ข้อความได้เองจากโทรศัพท์มือถือและสามารถส่งไปยังโทรศัพท์มือถือของผู้รับได้ทันที

SMS สามารถส่งได้ในรูปแบบของ ตัวเลข, ตัวอักษรและสัญลักษณ์ต่างๆ SMS ได้ถูกสร้างขึ้นมาครั้งแรกให้ทำงานร่วมกับโทรศัพท์เคลื่อนที่แบบดิจิตอล ระบบ GSM โดยข้อความแรกได้ถูกส่งในเดือนธันวาคม 1992 จากเครื่องคอมพิวเตอร์ส่วนบุคคลไปสู่เครื่องโทรศัพท์เคลื่อนที่บนโครงข่ายระบบ GSM ของ Vodafone ในประเทศอังกฤษ ปัจจุบันบริการ SMS สนับสนุนโครงข่าย GSM, CDMA และ TDMA สำหรับการส่ง SMSภาษาไทยจะส่งได้ 70 ตัวอักษร ภาษาอังกฤษส่งได้ 160 ตัวอักษร (ซึ่งมีทั้งตัวอักษรภาษาอังกฤษตัวพิมพ์เล็ก, ตัวพิมพ์ใหญ่, ตัวเลขและสัญลักษณ์พิเศษต่างๆ)

2.1.2 ผู้ส่ง (Sender)

เป็นอุปกรณ์ที่ใช้ในการส่งข่าวสาร (Message) เป็นต้นทางของการสื่อสารข้อมูลมีหน้าที่เตรียมสร้างข้อมูลสามารถ สำหรับการส่งข้อความนั้น Sender จะแบ่งออกเป็น 2 ประเภทหลักๆ ได้แก่ [1]

---

[1] นนท์ บุญนิธิประเสริฐ. (2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่.** วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

1) ผู้ส่งส่งข้อความจากโทรศัพท์เคลื่อนที่ หรือ Mobile Device อื่นๆ

เป็นการส่งข้อความที่ผู้ส่งส่งจากอุปกรณ์โดยตรง ซึ่งเป็นวิธีการส่งข้อความตั้งแต่เริ่มมีให้บริการ ซึ่งข้อความแบบไม่เป็นข้อความ Spam เนื่องจากมีข้อจำกัดหลายอย่าง เช่น ความเร็วการประมวล, หน่วยความจำ และแหล่งพลังงานของอุปกรณ์ อย่างไรก็ตาม ระบบโทรศัพท์ในยุคต่อไปที่กำลังจะมาถึง อาจทำให้เกิดข้อความ Spam จากอุปกรณ์เหล่านี้ได้

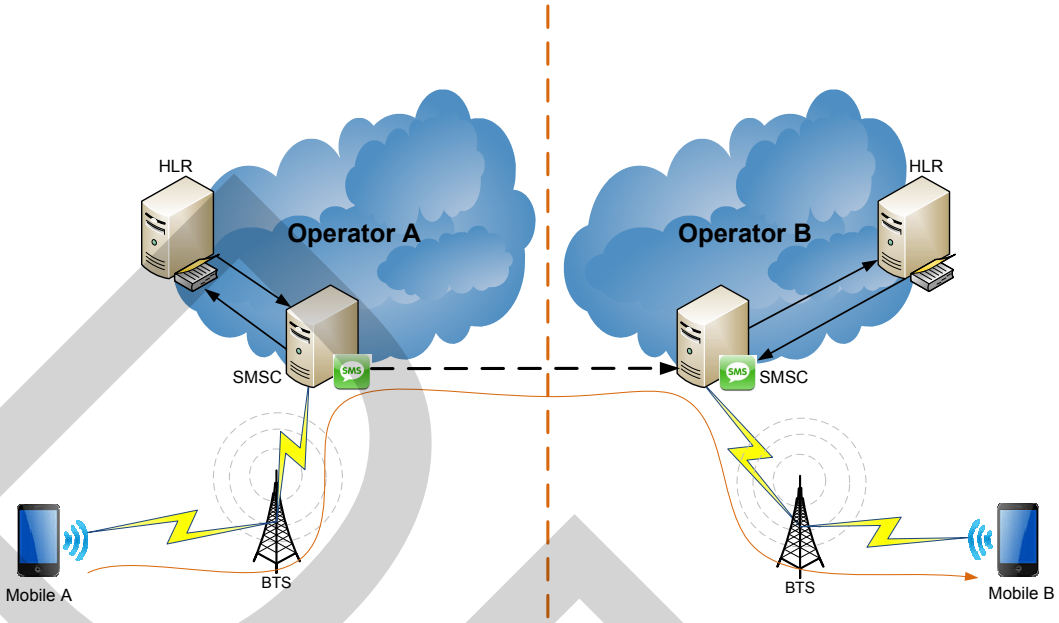
2) ผู้ส่งส่งข้อความผ่านระบบแอพพลิเคชั่น

เป็นวิวัฒนาการของการส่งข้อความที่เพิ่มความสามารถในการส่งข้อความมากขึ้น คือสามารถส่งข้อความได้ครั้งละมากๆ ซึ่งผู้ให้บริการโทรศัพท์เคลื่อนที่ (Operator) อาจจะมีการเตรียมแอพพลิเคชั่นไว้ โดยผู้ใช้บริการอาจจะทำการให้ Download แอพพลิเคชั่นสำหรับส่งข้อความเหล่านั้นมาใช้งานได้ เช่น การส่ง SMS เพื่อทำการประชาสัมพันธ์สินค้า เป็นต้น ยิ่งไปกว่านั้นผู้ใช้บริการเองก็สามารถพัฒนาระบบแอพพลิเคชั่นสำหรับการส่งข้อความขึ้นเองได้เช่นกัน ผู้ส่งกลุ่มนี้มีความยืดหยุ่นในการส่งข้อความสูง ส่งได้ครั้งละหลายข้อความ และตลอดเวลา ทำให้ผู้ส่งกลุ่มนี้ มีโอกาสส่งข้อความ Spam ได้ (Spammer)

นอกจากการส่ง SMS ด้วยโทรศัพท์เคลื่อนที่และแอพพลิเคชั่นที่ผู้ให้บริการโทรศัพท์เคลื่อนที่ได้จัดเตรียมไว้ให้แล้ว การส่ง SMS จากแอพพลิเคชั่นอื่นๆ ก็สามารถทำได้ ยกตัวอย่างเช่น E-Mail Server หรือ ระบบตอบรับด้วยเสียงอัตโนมัติ (IVR) โดยแต่ละระบบจะเชื่อมต่อกับ SMSC ด้วย SMPP ผ่าน Adapter แต่ผู้ส่งประเภทนี้จะควบคุมได้ง่ายกว่าและมีโอกาสเกิดข้อความ Spam น้อยมาก

2.1.3 เครือข่าย (SMS Network)

SMS Network เปรียบเสมือนช่องทางหรือตัวกลางนำข้อมูลจากต้นทางไปยังปลายทาง ซึ่งเป็นเส้นทางการสื่อสารเพื่อพื้นฐานของระบบ SMS จะมีการเชื่อมต่อกันหลายส่วน ทั้งภายในเครือข่ายตนเองและเครือข่ายของผู้บริการรายอื่นด้วยง ซึ่งการรับ-ส่ง SMS เป็นเทคนิคการสื่อสารที่ไม่จำเป็นต้องใช้การสร้างวงจรสนทนา (Call Set-up) จึงทำให้สามารถรับหรือส่งข้อความได้ในขณะที่กำลังสนทนาอยู่ หรือในขณะที่เปิดเครื่องทิ้งไว้บริการ SMS ไม่ใช่บริการแบบ Realtime เนื่องจากการส่งข้อความต้องส่งผ่าน Platform กลาง คือ Short Message Service Center หรือ SMS-C ซึ่งเป็นอุปกรณ์ที่ผู้ให้บริการเครือข่ายโทรศัพท์เคลื่อนที่ติดตั้งไว้เพื่อให้บริการรับ-ส่งข้อความได้ ดังแสดงในภาพที่ 2.1



ภาพที่ 2.1 ลำดับการส่งข้อความ SMS ระหว่าง Operator A และ B

จากภาพที่ 2.1 มีลำดับการส่งข้อความเริ่มต้นจากเครื่องโทรศัพท์เคลื่อนที่ของผู้ส่ง (User A) ผ่านเสา รับ – ส่ง สัญญาณโทรศัพท์ (Base Transceiver Station) และชุมสาย (Mobile Switching Center MSC) ไปยัง SMS Center (SMSC) ซึ่งจะทำหน้าที่ค้นหาผู้รับ (User B) จาก Home Location Register (HLR) แล้วดำเนินการจัดส่งข้อความ

2.1.4 โปรโตคอล (Protocol)

วิธีการหรือกฎระเบียบที่ใช้ในการสื่อสารข้อมูลเพื่อให้ผู้รับและผู้ส่งสามารถเข้าใจกันหรือคุยกันรู้เรื่อง โดยทั้งสองฝั่งทั้งผู้รับและผู้ส่งได้ตกลงกันไว้ก่อนล่วงหน้าแล้ว ซึ่งมาตรฐานการสื่อสารของระบบ SMS บน TCP/IP ที่ใช้ในธุรกิจโทรคมนาคม สำหรับแลกเปลี่ยน SMS Message ระหว่างกันของ SMSC ของแต่ละผู้ให้บริการ คือ Short Message Peer-to-Peer protocol (SMPP)[2] ซึ่ง SMPP สามารถนำไปใช้ในการทำ value-added service providers เพื่อใช้ในการให้บริการส่ง SMS ได้ เช่น บริการข่าว, บริการ SMS Bulk เป็นต้น

2.1.5 ผู้รับ (Receiver)

ผู้รับคือเครื่องโทรศัพท์เคลื่อนที่ปลายทาง หากเครื่องปลายได้ทำการเปิดใช้งานปกติก็จะได้รับข้อความทันทีที่มีการส่งแต่ถ้าหากมีการส่ง SMS พร้อมกันครั้งละมากๆ ผู้รับก็อาจจะไม่ได้

[2] SMS Forum, (1999-2007). SMPP Protocol. Retrieved January 25, 2010. From **http://smsforum.net/**

รับในทันได้ ต้องรอลำดับในการส่ง SMS ของ SMSC แต่ถ้าหากปิดเครื่องอยู่เมื่อผู้รับเปิดเครื่องรับโทรศัพท์จะลงทะเบียนเพื่อแจ้งที่อยู่ของตนที่ HLR ว่ากำลังเชื่อมต่อกับ MSC ใด บริการต่างๆที่เกิดขึ้นก่อนที่ผู้รับเปิดเครื่อง เช่น SMS หรือ MMS จะทราบตำแหน่งของเครื่องผู้รับ และดำเนินการส่งให้กับผู้รับอย่างถูกต้อง

2.2 ระบบการคัดแยกข้อความ

เป็นงานวิจัยเกี่ยวกับนำทฤษฎีต่างๆ มาพัฒนาเพื่อแก้ปัญหา Spam Mail หรือ Spam SMS โดยจะแบ่งออกเป็น 2 ประเภทคือ

2.2.1 โปรแกรม

โปรแกรม Spam Mail Killer[3] เป็นโปรแกรมที่ช่วยกำจัดอีเมล์ขยะ (Spam Mail Killer) ตัวกำจัดอีเมล์ขยะที่ดีจะต้องสามารถจัดการอีเมล์ขยะได้อย่างถูกต้องและมีประสิทธิภาพ สำหรับโครงงานนี้ได้นำวิธีของ Bayesian ซึ่งเป็นวิธีการคำนวณข้อความทางสถิติมาประยุกต์ใช้ในการสร้างโปรแกรมกำจัดอีเมล์ขยะ

2.2.2 วิธีการ

2.2.2.1 Content Based SMS Spam Filtering[4]

เป็นงานวิจัยเกี่ยวกับการแก้ปัญหาของ SMS spam โดยนำเอาเทคนิคของ Bayesian Filtering ที่ใช้สำหรับกรอง Spam Mail มาประยุกต์ใช้ในการตรวจจับ SMS Spam เพื่อดำเนินการบล็อกข้อความเหล่านั้นและช่วยให้ลดปัญหา SMS Spam ได้

2.2.2.2 Bogofilter[5]

เป็น Open source spam filter ใช้เทคนิคการตรวจจับด้วยวิธีการ Naïve Bayesian (NB) และการเทียบคำที่กำหนดไว้ (Keyword matching) เพื่อแยกข้อความ Spam ออกจากข้อความทั่วไป โดยคำนึงถึงบริเวณที่มีการเทียบ เช่น ความหมายของคำว่า “platypus” ที่อยู่ส่วนของ Subject กับที่อยู่ใน from จะมีคุณลักษณะในการเปรียบเทียบต่างกัน

---

[3] ชัยณรงค์ รุจิเสถียรทรัพย์และณธัช เลี่ยวชาลิต. (2547). **ตัวกำจัดเมล์ขยะ Spam Mail killer** (รายงานวิชาโครงงาน). กรุงเทพฯ: สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[4] José María Gómez Hidalgo, Guillermo CajigasBringas and Enrique PuertasSánz. (October 10 - 13, 2006). “Content Based SMS Spam Filtering.” **Proceedings of ACM Symposium on Document Engineering**: **ACM2006.** p. 107-114. Amsterdam, Netherlands.

[5] Gordon V. Cormack, José María Gómez Hidalgo and Enrique Puertas Sánz. (November 6-9, 2007). “Spam Filtering for Short Messages.” **Proceedings of the ACM sixteenth conference on information and knowledge management: CIKM 2007.** p. 313-319. Lisboa, Portugal.

2.2.2.3 DMC (Dynamic Markov Compression)[6 - 7]

เน้นการประมวลผลกับข้อมูลที่ถูกบีบอัด โดยใช้หลักการพิจารณาข้อความเป็น String เพื่อตรวจสอบและทำนายว่าข้อความนั้นจัดเป็นข้อความ Spam หรือไม่

2.2.2.4 LOHIT Algorithm[8]

An Online Detection & Control System for Cellular SMS Spam เป็นงานวิจัยที่กล่าวถึงการพัฒนา LOHIT Filter สำหรับกรองข้อความ Spam ในระบบ SMS ที่ต่างจาก E-Mail Spam Filter โดยใช้สมการทางคณิตศาสตร์ เพื่อระบุความน่าจะเป็นของข้อความ Spam และทำการจำลองการส่งข้อมูล SMS เพื่อทดสอบประสิทธิภาพ โดยแสดงผลออกมาในรูป 3D subspace ผลการจำลองการส่ง SMS สามารถให้ประสิทธิภาพดีกว่า Filter ที่พัฒนาจาก E-Mail Spam Filter อื่นๆ

2.2.2.5 การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่ [9]

เป็นงานวิจัยที่ศึกษาวิธีการกรองแบบ Support Vector Machine (SVM) และ Naive Bayesian (NB) การทำ Text Normalization และการตัดคำแบบต่างๆที่มีการใช้งานอยู่ในปัจจุบัน โดยได้ปรับปรุงการทำ Text Normalization และการใช้วิธีตัดคำแบบผสมด้วยการกรองข้อความ SMS ทั้ง ภาษาไทย ภาษาอังกฤษและภาษาไทยปนอังกฤษ พบว่าการปรับปรุงขั้นตอนการทำ Text Normalization สามารถลดปริมาณคำที่ไม่ถูกต้องในพจนานุกรมฐานข้อมูลลงได้ ผลการทดสอบด้วยวิธีการกรอง แบบ SVM มีความถูกต้องในการกรองข้อความสูงกว่าวิธีการแบบ NB แต่วิธีการกรองแบบ NB ใช้เวลาในการประมวลผลน้อยกว่าและการใช้วิธีตัดคำแบบผสมส่งผลให้การกรองข้อความ SMS ที่มีเนื้อความเกินกว่า 1 ข้อความมีความถูกต้อง มากกว่าการตัดคำด้วยวิธีการแบบใดแบบหนึ่งเพียงแบบเดียวซึ่งต้องใช้ระยะเวลาในการประมวลผลเพิ่มขึ้น

---

[6] Gordon V. Cormack, José María Gómez Hidalgo and Enrique Puertas Sánz. (November 6-9, 2007). "Spam Filtering for Short Messages." **Proceedings of the ACM sixteenth conference on information and knowledge management: CIKM 2007.** p. 313-319. Lisboa, Portugal.

[7] Andrej Bratko, Gordon V.Cormack, Bogdan Filipic, Thomas R. Lynam and Blaz Zupan. (December, 2006). "Spam Filtering Using Statistical Data Compression Models." **Journal of Machine Learning Research,7**. p. 2673-2678.

[8] Siddharth Dixit, Sandeep Gupta and Chinya V.Ravishankar. (November 14-16, 2005). "LOHIT: An Online Detection & Control System for Cellular SMS Spam." **Proceeding of the IASTED International Conference on Communication, Network, and Information Security.** p. 48-54. Phoenix, AZ, USA.

[9] นนท์ บุญนิธิประเสริฐ และ ดร. ชัยพร เขมะภาตะพันธ์. (22-23 พฤษภาคม 2552). "Short Message Service Filtering for Thai & English Language on Mobile Phone Network" **Proceeding of the 5th National Conference on Computing and Information Technology; NCCIT2009.** p.34-39. กรุงเทพมหานคร.

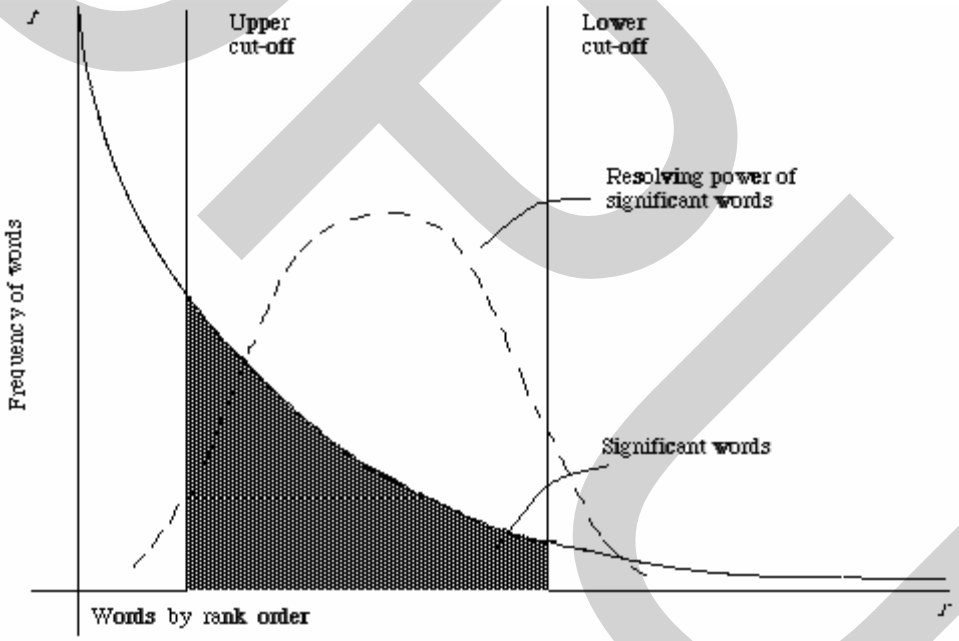
2.3 เทคนิคการคัดแยกข้อความ

2.3.1 Text Operation[10]

คือเป็นขั้นตอนการเตรียมข้อมูลเบื้องต้น เป็นกระบวนการนำข้อความที่รวบรวมได้เข้าสู่กระบวนการเตรียมข้อความ ซึ่งจะมีการแยกคำ การตัดคำ การเลือกกลุ่มคำที่เป็นคำหลัก (Keyword) และลบสัญลักษณ์พิเศษ เช่น **$** | **#** | **@** | **?** | **!** หรือตัวเลขที่ไม่ต้องการเพื่อกำจัดข้อมูลส่วนเกินออก ใช้ในการประมวลผลข้อมูลที่อยู่ในรูปแบบ text หรือ string เมื่อต้องการนำข้อมูลเหล่านั้นไปคำนวณค่า โดยการใช้ฐานความรู้ในพจนานุ กรม (Dictionary) สำหรับภาษาไทย

2.3.2 Stop Words[11-12]

ในการจัดหมวดหมู่ข้อมูลประเภท text หรือการกรองข้อความ จะต้องค้นหาลักษณะแทนข้อมูล (representation data) จากความถี่ของคำทั้งหมดในเอกสารดังภาพที่ 2.2



ภาพที่ 2.2 แสดงความถี่ของคำที่ใช้แทนลักษณะของเอกสาร

---

[10] István Pilászy. (November 18-19, 2005). "Text Categorization and Support Vector Machines." **6th International Symposium of Hungarian Researchers on Computational Intelligence.** Paper list No.064. Budapest, Hungary.

[11] ข้อมูลพื้นฐานภาษาไทย, (2552). คำที่พบบ่อย (Stop word th). สืบค้นเมื่อ 2 สิงหาคม 2553. จาก **http://thailang.nectec.or.th/thaichar/word_thai.php?page=1&n_p_page=100.**

[12] University of Glasgow, (2009). Stop Word EN. Retrieved August 2, 2009. From **http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words.**

โดยในส่วนของคำที่มีความถี่สูงเกินไป (ทางซ้ายของเส้น Upper cut-off) และส่วนของคำที่มีความถี่น้อยเกินไป (ทางขวาของเส้น Lower cut-off) เป็นคำที่ไม่แสดงลักษณะแทนข้อมูลนั้นๆ จึงต้องทำการลบคำเหล่านั้นออก เพื่อจะได้ขอบเขตของคำที่แสดงลักษณะแทนข้อมูลโดย ซึ่งคำประเภทนี้จะมีองค์ประกอบคือ

1) คำที่พบเป็นจำนวนมากในข้อความทุกข้อความ

2) มีลักษณะเป็นคำขยาย หรือคำที่ไม่แสดงความหมาย

ตัวอย่างคำที่มีลักษณะเป็น Stop words ได้แก่ ฉัน, เธอ, นาย, ที่, ซึ่ง ไป เป็นต้น

2.3.3 TFIDF[13]

เป็นวิธีการค้นหาลักษณะเด่นของเอกสาร (Document) ให้อยู่ในรูปของกลุ่มข้อมูล (Feature Vector) โดยอ้างอิงจากชุดตัวอักษรหรือคำ (Term) ในเอกสาร และจำนวนเอกสารทั้งหมดที่ถูกกำหนดให้เป็นข้อมูลฝึกสอน ซึ่งวิธการนี้มีการทำงานโดยการคำนวณน้ำหนักของคำซึ่งพิจารณาเอกสารที่เกี่ยวข้องทั้งหมดประกอบ โดยใช้หลักการความถี่ของคำและความถี่ผกผันของเอกสาร (Term Frequency/Inverse Document Frequency—TF/IDF) น้ำหนักของคำที่อยู่ในเอกสารเป็นหลักการที่ใช้ค้นหาน้ำหนักของคำที่อยู่ในเอกสารทั้งหมดว่ามีค่าน้ำหนักมากหรือน้อย การเปรียบเทียบค่าน้ำหนักของคำหนึ่งๆ นำคำที่ถูกตัดจากการสืบค้นไปเปรียบเทียบกับเอกสารที่มีคำนี้อยู่ แล้วคำนวณหาค่าน้ำหนักใหม่ออกมา ซึ่งมีสมการการคำนวณคือ

$$TFIDF(i, j) = TF(i, j) \cdot IDF(i) \qquad (2\text{-}1)$$

$$IDF(i) = \log \frac{N}{DF(i)} \qquad (2\text{-}2)$$

TF คือ ความถี่ของ Term นี้ ที่ปรากฏใน Document

DF คือ ความถี่ของ Document ที่มี Term นี้

IDF คือ ค่าแทน Discrimination power ของ DF

ในการจัดกลุ่มเอกสาร หรือการกรองข้อความแบบต่างๆที่ใช้การคำนวณทางคณิตศาสตร์ จะใช้วิธีการ TFIDF ในการแปลงเอกสารที่ต้องการนำไปคำนวณ ให้เป็นชุดข้อมูลเพื่อนำไปคำนวณต่อไป

---

[13] อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย, ชุลีรัตน์ จรัสกุลชัย. (2547). **การสรุปใจความสำคัญของเอกสาร.** วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน.

2.3.4 Naïve Bayesian (NB)[14-15]

Naïve Bayes เป็นเทคนิคที่ถูกตั้งชื่อตาม Thomas Bayes (ค.ศ. 1702-1761) ซึ่งใช้วิธีการเรียนรู้ที่อาศัยหลักการทางสถิติและความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes'Theorem ซึ่งผลลัพธ์ที่ให้ความน่าจะเป็นสูงสุดเป็นคำตอบของการตัดสินใจ

Bayes'Rule

Bayes's rule: $$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)} \quad (2\text{-}3)$$

จากสมการแนวคิดพื้นฐานของ Bayes's Rule คือผลลัพธ์ของสมมติฐานหรือเหตุการณ์ (H) ที่สนใจสามารถทำนายได้จากหลักฐาน (E) ที่สามารถสังเกตได้ จากกฎของ Bayes จะได้ว่า

1) Priori คือความน่าจะเป็นของ H หรือ P(H) นี่เป็นความน่าจะเป็นของเหตุการณ์ก่อนหลักฐานถูกสังเกต

2) Posterior คือความน่าจะเป็นของ H หรือ P(H|E) นี่เป็นความน่าจะเป็นของเหตุการณ์หลังจากหลักฐานถูกสังเกตได้

ตัวอย่างการใช้ Bayes's rule

เป็นตัวอย่างการทำนายโอกาสหรือความน่าจะเป็นของการเกิดฝนตก เราใช้หลักฐานบางอย่างสำหรับการทำนาย เช่น จำนวนก้อนเมฆในอากาศ

กำหนดให้ H คือเหตุการณ์ของการเกิดฝนตกและ E ปืนหลักฐานของการเกิดเมฆดำ (dark cloud) ดังสมการที่ 2-4

$$P(rainning \mid darkcloud) = \frac{P(darkcloud \mid rainning) \times P(raining)}{P(darkcloud)} \quad (2\text{-}4)$$

$P(darkcloud \mid rainning)$ คือ ความน่าจะเป็นที่เมื่อมีเมฆดำแล้ว ซึ่งเมฆดำ(Dark cloud) เกิดขึ้นได้หลายกรณีเช่น วันที่ท้องฟ้ามืดครึ้มหรือไฟไหม้ป่าแต่ในที่นี้สนใจเฉพาะกรณีที่เมฆดำแล้วฝนตก และความน่าจะเป็นนี้สามารถหาได้จากข้อมูลในอดีตโดยนักนักอุตุนิยมวิทยา

$P(raining)$ คือ priori ความน่าจะเป็นของการเกิดฝนตก ซึ่งความน่าจะเป็นนี้สามารถหาได้จากการเก็บสถิติ ตัวอย่างเช่นจำนวนวันของการฝนตกในปีนั้นๆ

---

[14] Wikipedia. (2009). Naive Bayes classifier**.** Retrieved August 2, 2009. From **http://en.wikipedia.org/wiki/Naive_bayes.**

[15] Wikipedia. (2009). Bayesian network. Retrieved August 2, 2009. From **http://en.wikipedia.org/wiki/Bayesian_network.**

$P(darkcloud)$ เป็นความน่าจะเป็นของหลักฐานที่เราใช้คือความน่าจะเป็นของการเมฆดำซึ่งได้มาจากการเห็นสถิติ

จากสมการของ Bayes's rule จะเห็นว่าหลักฐานที่ใช้สนับสนุนสำหรับการคัดแยกกลุ่มข้อความมีเพียงอย่างเดียวเท่านั้น ซึ่งในความเป็นจริงแล้วการที่มีหลักฐานสนับสนุนในการตัดสินใจที่มากกว่าหนึ่งอย่างนั้น จะทำให้การคัดแยกข้อความมีความถูกต้องและแม่นยำมากยิ่งขึ้น Naïve Bay เป็นเทคนิคในการแก้ปัญหาแบบ classification ที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างสิ่งที่ต้องการ classification แต่ละตัวกับหลักฐานที่ใช้ประกอบการตัดสินใจเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ในทางทฤษฎีแล้วการทำนายผลของ Naïve-Bayes จะถูกต้องถ้าหลักฐานทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับหลักฐานตัวใดตัวหนึ่ง ดังสมการที่ 2-5

$$P(H \mid E_1, E_2, ..., E_n) = \frac{P(E_1 \mid H) \times P(E_2 \mid H) \times ... P(E_n \mid H) \times P(H)}{P(E_1, E_2, ..., E_n)} \qquad (2\text{-}5)$$

ดังนั้นเพื่อให้เข้าใจการสร้าง NB Model มายิ่งขึ้นนั้นโดยยกตัวอย่างของ data set ของอากาศประกอบไปด้วย outlook, temperature, humidity และ windy สำหรับการทำนายเงื่อนไขของการเล่นเทนนิส ดังตารางที่ 2.1

ตารางที่ 2.1 แสดงชุดข้อมูลของสภาพอากาศ

No.	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes

ตารางที่ 2.1 แสดงชุดข้อมูลของสภาพอากาศ (ต่อ)

No.	outlook	temperature	humidity	windy	play
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

จากนั้นสามารถสร้าง NB Model จากข้อมูลที่ให้มาข้างต้น ซึ่งผลลัพธ์ได้ตามตารางที่ 2.2

ตารางที่ 2.2 แสดงผลการสร้าง NB model

outlook			**temperature**			**humidity**			**windy**			**Play**	
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
sunny	2	3	hot	2	2	high	3	4	FALSE	6	2	9	5
overcat	4	0	mild	4	2	normal	6	1	TRUE	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/9	hot	2/9	2/5	high	3/9	4/5	FALSE	6/9	2/5	9/14	5/14
overcat	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	TRUE	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

จากตารางจะแสดงให้เห็นความถี่ที่เกิดขึ้นของหลักฐานแต่ละชิ้นที่ใช้การทำนายการจะเล่นเทนนิสหรือไม่เล่นเทนนิส ยกตัวอย่างเช่นจาก data set ทั้งหมดแสดง outlook=sunny เมื่อ play =yes ดังนี้

$$P(outlook = sunny \mid play = yes) = \frac{2}{9}$$

$$P(play = yes) = \frac{9}{14}$$

เมื่อได้ NB model มาแล้ว เราสามารถทำนายเหตุการณ์ของการเล่น “play” โดยทำนายจากหลักฐานต่างๆ ที่ได้เตรียมไว้แล้วนั้น จากตัวอย่างจะสังเกตได้ว่า outlook=sunny, temperature=cool, humidity=high และ windy=ture แล้วสามารถประมาณค่า posterior probability ได้ดังนี้

outlook	temperature	humidity	windy	play
sunny	cool	high	TRUE	?

← Evidence E

$$\Pr[yes|E] = \Pr[outlook = sunny|yes] \times \Pr[temperature = cool|yes] \times \Pr[humidity = high|yes] \times \Pr[windy = ture|yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class “yes”

$$= \frac{\frac{2}{9}\times\frac{3}{9}\times\frac{3}{9}\times\frac{3}{9}\times\frac{9}{14}}{\Pr[E]}$$

ในที่นี้สามารถมองข้าม Pr(E) เพราะเราต้องการเพียงแต่หา ความสัมพัทธ์(relatively) เพื่อเปรียมเทียบค่าระหว่าง 2 class ซึ่งผลที่ได้ดังนี้

Likelihood of the two classe

For “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For “no” = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P(“yes”) = 0.0053/ (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206/ (0.0053 + 0.0206) = 0.795

สังเกตได้จาก NB model ข้างต้นจะเห็นได้ว่า P(outlook=overcast|play=no)=0/5 ซึ่งจะทำให้เกิดปัญหาเมื่อคำนวณหาค่า P(“no”) เนื่องจากผลลัพธ์ที่ได้จะเท่ากับ “0” แก้ปัญหาได้โดยใช้เทคนิคที่เรียกว่า “smoothing techigue ซึ่งวิธีการหนึ่งของการ smoothing technique คือ Laplace estimation

$$P(outlook = sunny \mid play = no) = \frac{3 + \mu p_1}{5 + \mu}$$

$$P(outlook = overcast \mid play = no) = \frac{0 + \mu p_2}{5 + \mu}$$

$$P(outlook = overcast \mid play = no) = \frac{1 + \mu p_3}{5 + \mu}$$

โดยที่ $(p_1 + p_2 + p_3) = 1.0$

ตามที่ได้มีการกำหนดค่าให้หลักฐานทั้งหมดที่ถูกกระจายข้อมูลเท่าๆ กัน ตามข้างต้น p1=p2=p3=1/3

$$P(outlook = sunny \mid play = no) = \frac{3 + \mu/3}{5 + \mu} = \frac{3 + 3/3}{5 + 3} = \frac{4}{8}$$

$$P(outlook = overcast \mid play = no) = \frac{0 + \mu/3}{5 + \mu} = \frac{0 + 3/3}{5 + 3} = \frac{1}{8}$$

$$P(outlook = overcast \mid play = no) = \frac{1 + \mu/3}{5 + \mu} = \frac{2 + 3/3}{5 + 3} = \frac{3}{8}$$

Naïve Bayes ถูกใช้อย่างกว้างขว้างในหลายๆ แอพพลิเคชั่น ซึ่งแอพพลิเคชั่นหนึ่งที่ได้รับความสนใจคือ text classification และ information filtering (เช่น spam filtering) เหตุผลหลักคือ NB model ทำงานได้ดีสำหรับ text domain เพราะหลักฐานที่ใช้ในการแบ่งกลุ่มเป็น “vocabularies” หรือ “word” รูปแบบของ text ที่จะนำมาใช้สำหรับเป็นหลักฐานในการประกอบการแบ่งประเภทของข้อความนั้นมีรูปแบบแน่นอนที่ปรากฏในกลุ่มของข้อความที่มีจำนวนเป็นพันๆ หรือมากกว่านั้น ขนาดของหลักฐานที่ใหญ่อย่างนี้ทำให้ NB model ทำงานได้ดีสำหรับการแก้ปัญหา text classification

สำหรับ text โดเมนนั้น ก็สามารถสร้าง NB model ที่คล้ายกับข้อมูลประเภทอื่นเช่นกัน สมมติถ้าหากมี data set ของdocument 6 document คือ D0…D5 เป็น data set ที่เอาไว้เป็นข้อมูล training และพิจารณา vocabularies เพียง 6 คำจากข้อมูลทั้งหมดและในที่นี่เราจะแบ่ง class ของ document ออกเป็น 2 Class คือ terrorism และ entertainment โดยจะทำการเตรียม document ก่อนนำไปคำนวณดังตารางที่ 2.3 ซึ่งจะแสดงจำนวนความถี่ของคำนั้นๆ ใน document ตัวอย่างเช่นคำว่า “kill” มีใน D0 จำนวน 2 ครั้ง

ตารางที่ 2.3 การเตรียมข้อมูลโดยการนับจำนวนความถี่ของคำในแต่ละ document เพื่อสร้าง NB model

Training Doc	kill	bomb	kidnap	music	movie	TV	C
D0	2	1	3	0	0	1	Terrorism
D1	1	1	1	0	0	0	Terrorism
D2	1	1	2	0	1	0	Terrorism

ตารางที่ 2.3 การเตรียมข้อมูลโดยการนับจำนวนความถี่ของคำในแต่ละ document เพื่อสร้าง NB model (ต่อ)

Training Doc	kill	bomb	kidnap	music	movie	TV	C
D3	0	1	0	2	1	1	Entertainment
D4	0	0	1	1	1	0	Entertainment
D5	0	0	0	2	2	0	Entertainment

ขั้นตอนที่ 1

สร้าง NB Model เมื่อทำการเตรียมข้อมูลเรียบร้อยแล้วสร้าง NB model ดังตารางที่ 2.4

ตารางที่ 2.4 การสร้าง NB model

|V|	C	P(Ci)	ni	P(kill|Ci)	P(bomb|Ci)	P(kidnap|Ci)	P(music|Ci)	P(movie|Ci)	P(TV|Ci)
6	Terrorism	0.5	15	0.2380952	0.1904762	0.3333333	0.047619	0.0952381	0.0952381
	Entertainment	0.5	12	0.0555556	0.1111111	0.1111111	0.3333333	0.2777778	0.1111111

|V| คือ จำนวนของ keyword หรือ vocaburaries

$P(c_i)$ คือ priori probability ของแต่ละ class โดยหาได้จาก จำนวนของ document ใน class / จำนวนของ document ทั้งหมด

P (Terrorism) = 3/6 = 0.5

P (Entertainment) = 3/6 = 0.5

ni คือ ผลรวมของความถี่ของคำแต่ละ class

$n_{terrorism}$ = 2+1+3+1+1+1+1+1+1+2+1 = 15

$n_{Entertainment}$ =1+2+1+1+1+1+1+2+2 = 12

$P(w_i|c_i)$ คือเงื่อนไขของความน่าจะเป็นของคำที่เกิดขึ้นที่มีใน class เช่น

P(kill|Terrorism) = (2+1+1)/15 = 4/15

P(kill|Entertainment) = (0+0+0)/12

ในที่นี่หลีกเลี่ยง “zero frequence” โดยการประยุกต์ใช้ Laplace estimation โดยการสมมติว่ามีการกระจายของข้อมูลที่เท่ากันดังนี้

P(kill|Terrorism) = (2+1+1+1) / (15+|V|) = 5/12 = 0.2380

P(kill|Entertainment) = (0+0+0+0+1) / (12+|V|) = 1/18 = 0.0555

ขั้นตอนที่ 2

จัดกลุ่ม (Classifying) ข้อมูลทดสอบ

เมื่อสร้าง NB model เรียบร้อยแล้ว นำข้อมูลทดสอบมาทำการทดสอบเพื่อแบ่งกลุ่มข้อมูล

ตารางที่ 2.5 การทดสอบข้อมูล

Test Doc	kill	bomb	kidnap	music	movie	TV	C
Dt	2	1	2	0	0	1	?

$$P(c_i \mid W) = P(ci) \times \prod_{j=1}^{v} P(w_j \mid c_i) \qquad (2\text{-}6)$$

P(Terrorism | W) = P(Terrorism) x P(kill | Terrorism) x P(bomb | Terrorism) x P(kidnap | Terrorism) x P(music | Terrorism) x P(movie | Terrorism) x P(TV | Terrorism)

= $0.5 \times 0.2380^{2} \times 0.1904^{1} \times 0.3333^{2} \times 0.0476^{0} \times 0.0952^{0} \times 0.0952^{1}$

= 0.5 x 0.0566 x 0.1904 x 0.1110 x 1 x 1 x 0.0952

= 5.7×10^{-5}

P(Entertainment | W) = P(Entertainment) x P(kill | Entertainment) x P(bomb | Entertainment) x P(kidnap | Entertainment) x P(music | Entertainment) x P(movie | Entertainment) x P(TV | Terrorism)

= $0.5 \times 0.0555^{2} \times 0.1111^{1} \times 0.1111^{2} \times 0.3333^{0} \times 0.2777^{0} \times 0.1111^{1}$

= 0.5 x 0.0030 x 0.1111 x 0.0123 x 1 x 1 x 0.1111

= 2.27×10^{-7}

ซึ่งผลที่ได้จากกรคำนวณหาค่าความน่าจะเป็นมีค่าน้อยมาก จำนวนของเงื่อนไขความน่าจะเป็นในขอบเขตของคำเป็นพันหรือมากกว่า จะมีค่าต่ำมากสำหรับการจัดการของ CPU นี่เป็น

ปัญหาที่ถูกอ้างถึง ภายใต้ปัญหานี้ การแก้ปัญหานี้ เราสามารถ take logarithm บนความน่าจะเป็นนี้ ดังสมการ 2-7

$$P(Ci \mid W) = \log(P(Ci) \times \prod_{j=1}^{v} P(wj \mid Ci)) \qquad (2\text{-}7)$$

P(Terrorism | W) $= \log(0.5 \text{ x } 0.2380^2 \text{ x } 0.1904^1 \text{ x } 0.3333^2 \text{ x } 0.0476^0 \text{ x } 0.0952^0 \text{ x } 0.0952^1)$

$= \log(0.5) + 2 \log(0.2380) + 1 \log(0.1904) + 2 \log(0.3333) + 0 \log(0.0476) +$

$0 \log(0.0952) + 1 \log (0.0952)$

$= -0.3010 - 1.2468 - 0.7203 - 0.9543 + 0 + 0 - 1.0213$

$= -4.2437$

P(Entertainment | W) = log(0.5 x 0.05552 x 0.11111 x 0.11112 x 0.33330 x 0.27770 x 0.11111)

= log(0.5) + 2 log(0.0555) + 1 log(0.1111) + 2 log(0.1111) + 0 log(0.3333) +0 log(0.2777) + 1 log (0.1111)

= – 0.3010 – 2.511 – 0.9542 – 1.9085 + 0 + 0 – 0.9542

= – 6.6289

หลังจากผ่านการคำนวณหาค่าความน่าจะเป็นของข้อความจาก Naïve Bayes model แล้ว ต่อมาก็ตัดสินใจว่าข้อความนั้นจัดอยู่ในประเภทไหน ดังนั้นจึงนำ Decision Rule มาใช้สำหรับการตัดสินใจด้วย ซึ่ง Rule ที่ใช้กันโดยปกติทั่วไปคือ การพิจารณาว่าค่าความน่าของ Class ไหนมีค่าสูงที่สุด ก็จัดว่าเป็นประเภทนั้น ดังสมการ 2-8

$$classify(c_1,..,c_n) = \arg\max\, p(W = w) \prod_{i=1}^{n} p(C_i = C_i \mid W = w) \qquad (2\text{-}8)$$

2.4 การตัดคำภาษาไทย

การกรองข้อความทั่วไป เช่น ระบบกรองข้อความในหน้า Web page หรือ E-Mail Spam Filter จะวิเคราะห์ข้อความจากคำ ซึ่งในภาษาอังกฤษใช้การเว้นวรรคเพื่อตัดคำ (word segmentation) ในขณะที่ข้อความภาษาไทยไม่สามารถทำได้ เพราะใช้หลักการเขียนคำต่อกันเป็นประโยค ทำให้ต้องใช้ Algorithm ในการตัดแยกคำ อีกทั้งข้อจำกัดของระบบ SMS ทำให้พฤติกรรมการส่งข้อความ มีลักษณะของ คำย่อ, คำทับศัพท์, หรือคำภาษาอังกฤษ ปะปนกันอย่างไม่เป็นระเบียบ จึงจำเป็นต้องปรับปรุงระบบการตัดคำภาษาไทยให้รองรับข้อความดังกล่าวได้

เทคนิคการตัดคำภาษาไทยแบ่งออกเป็น 4 แบบใหญ่ๆ[16] ดังนี้

---

[16] **สารานุกรมไทยสำหรับเยาวชนฯ**. (2544). เล่มที่ 25. กรุงเทพมหานคร: โครงการสารานุกรมไทยสำหรับเยาวชน โดยพระราชประสงค์ในพระบาทสมเด็จพระเจ้าอยู่หัว.

2.4.1 วิธีการตัดคำแบบยาวที่สุด (Longest Matching)

วีธีการเทียบคำที่ยาวที่สุดเป็นวิธีการตัดคำทางวิทยาการศึกษา (heuristic) วิธีหนึ่งที่ตัดคำด้วยการค้นหาคำเริ่มจากตัวอักษรซ้ายสุดของข้อความนั้นไปยังตัวอักษรถัดไปจนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรมที่นำมาใช้ในการช่วยรู้จำคำไทย โดยจะทำการเปรียบเทียบสายอักขระดังกล่าวเป็นหนึ่งคำหรือไม่หากไม่พบว่าสายอักขระดังกล่าวสามารถเทียบเป็นคำได้ในพจนานุกรม ก็จะทำการลดควายาวของสายอักขระลงทีละตัว จนกว่าสายอักขระที่ตรวจสอบจะสามารถเทียบเป็นคำในพจนานุกรมได้ ก็จะทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ จากนั้นก็จะเริ่มทำงานจากจุดย้อนกลับนั้นเพื่อตรวจสอบสายอักขระที่เหลือว่าจะสามารถตัดสายอักขระใดต่อไปให้เป็นคำได้ หากตัวเลือกในตอนแรกนี้สามารถทำให้ค้นหาคำที่เหลือได้ ตัวเลือกนี้ก็จะเป็นคำแรกของข้อความได้จริง ไม่เช่นนั้นขั้นตอนวิธีก็จะกลับไปยังจุดย้อนกลับ หากยังไม่สามารถเทียบสายอักขระกับคำในพจนานุกรมได้ ก็จะทำการลดตัวอักษรลงที่ละตัวจนกว่าจะเทียบคำในพจนานุกรมได้ และทำงานรูปแบบนี้ต่อไปจนจบข้อความ ตัวอย่างเช่น ถ้าป้อนคำว่า “ความก้าวหน้าทางวิทยาศาสตร์มีบทบาทสำคัญ” เข้าไป ข้อความนี้เมื่อไม่สามารถเทียบคำกับพจนานุกรมได้ ก็ลดลงเหลือ --> “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคั” --> “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำ” จนได้สายอักขระ “ความก้าวหน้า” ซึ่งสามารถเทียบคำในพจนานุกรมได้ จึงตัดเป็นคำแรกของข้อความ และทำเครื่องหมายไว้เป็นจุดย้อนกลับ

4.2.2 วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching)

เป็นวิธีการตัดคำทาง heuristic อีกวิธีหนึ่งที่ใช้พจนนุกรมช่วยรู้จำคำภาษาไทย ใช้การตัดคำที่สามารถจะเป็นไปได้ทั้งหมด แล้วเลือกข้อความที่ตัดได้จำนวนคำน้อยที่สุดมาใช้งาน เพื่อแก้ปัญหาที่ปรากฏในวิธีการเทียบคำที่ยาวที่สุด เริ่มจากการเลือกของรูปแบบการตัดคำทั้งหมดที่เป็นไปได้เสียก่อน โดยทำการย้อนกลับ (backtracing) ทีละคำหลังจากได้คำตอบจากวิธีการเทียบคำที่ยาวที่สุดแล้วจึงเลือกทางเลือกที่มีจำนวนคำน้อยที่สุด suprapant Meknavin และ บุญเสริม กิจศิริกุล กล่าวว่า การค้นหาทุกทางเลือกที่เป็นไปได้นี้ทำให้ต้องเสียเวลาในการคำนวณมาก แต่ก็สามารถลดเวลาลงได้โดยใช้โปรแกรมแบบพลวัติ (dynamic programming) ตัวอย่างเช่น หากป้อนข้อความ “ไปหามเหสี” เข้าไปขั้นตอนวิธีนี้จะหาทางเลือกทั้งหมดของรูปแบบการตัดคำที่เป็นไปได้ เป็นต้น

4.2.3 วิธีการตัดคำแบบคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Model)

วิธีการนี้นำเอาค่าสถิติการเกิดของคำและลำดับหน้าที่ของคำ (Part of speech) เข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อที่จะใช้เลือกแบบที่มีโอกาสการเกิดมากที่สุด ซึ่งสามารถจะตัดคำได้ดีกว่า 2 แบบแรก แต่ข้อจำกัดของวิธีการนี้คือ จะต้องมีฐานข้อมูลที่มีการตัดคำที่ถูกต้องและมีการกำหนดหน้าที่ของคำให้ เพื่อนำไปใช้ในการสร้างสถิติ

4.2.4 วิธีการตัดคำแบบใช้คุณลักษณะ (Feature - Based Approach)

วิธีการนี้จะพิจารณาจากบริบท (context words) และการเกิดร่วมกันของคำ หรือ หน้าที่ของคำ (collocation) เข้ามาช่วยในการตัดคำ ซึ่งตัวอย่างเช่น “ตากลม” ถ้าพบคำว่า “แป๋ว” ในบริบทก็จะสามารถตัดคำได้ว่า “ตา” “กลม” “มากว่า” ถ้าในบริบทที่ตามมาเป็นตัวเลขก็สามารถตัดคำได้ว่า “มา” “กว่า” ซึ่งแนวคิดของวิธีการนี้คือ พยายามเรียนรู้คุณลักษณะต่างๆ จากคลังข้อมูลที่บ่งชี้ลักษณะของบริบทที่คำหนึ่งๆ สามารถจะปรากฏได้ ซึ่งทำให้สามารถใช้คุณลักษณะต่างๆ นั้นร่วมกันเพื่อแก้ปัญหาความกำกวมในการตัดคำ

บทที่ 3
ระเบียบวิธีวิจัย

3.1 แนวทางการวิจัยและพัฒนา

งานวิจัยนี้มีวัตถุประสงค์ในการออกแบบวิธีการคัดแยกข้อความตามระดับความสำคัญ เพื่อจัดลำดับการส่ง อีกทั้งยังลดความคับคั่งในการส่งข้อความ SMS ในประเทศไทย ที่มีการใช้ภาษาไทย ภาษาอังกฤษ และภาษาไทยปนภาษาอังกฤษ ในการส่งข้อความ โดยแบ่งขั้นตอนการวิจัยออกเป็น 2 ส่วนดังนี้

3.1.1 ศึกษาและเปรียบเทียบหาวิธีการคัดแยกที่เหมาะสม

เนื่องจากยังไม่มีงานวิจัยที่ศึกษาการการคัดแยกข้อความตามลำดับความสำคัญสำหรับบริการส่งข้อความ SMS ในประเทศไทยอย่างจริงจัง จึงต้องใช้การศึกษาวิธีคัดแยกข้อความที่มีใช้ในงานอื่นๆ เป็นพื้นฐานอ้างอิง โดยวิธีการที่ถูกนำมาใช้อย่างแพร่หลายในงานคัดแยกข้อความได้แก่ Naive Bayesian ซึ่งจะทำการศึกษาและปรับปรุงการทำงานบางส่วนให้สามารถใช้งานร่วมกับภาษาไทยได้ เพื่อวิจัยเปรียบเทียบในด้านประสิทธิภาพความถูกต้องและระยะเวลาในการประมวลผล จากข้อความ SMS ที่มีการตรวจสอบลักษณะข้อความด้วยมนุษย์ โดยนำวิธีการคัดแยกที่มีประสิทธิภาพ มาปรับปรุงให้สอดคล้องกับข้อความ SMS ในประเทศไทยต่อไป

3.1.2 วิเคราะห์ปัญหาการคัดแยกข้อความ SMS ของประเทศไทย

ในขั้นตอนการวิจัยส่วนที่ 1 จะทราบหลักการทำงานและข้อบกพร่องในการคัดแยกข้อความ SMS ซึ่งจะนำผลการทดสอบมาทำการวิเคราะห์เพื่อแก้ไขและปรับปรุงวิธีการคัดแยกให้มีความสอดคล้องกับข้อความ SMS ในประเทศไทยด้วยการดำเนินงานในส่วนที่ 2 และทำการทดสอบการคัดแยกข้อความด้วยข้อความ SMS ชุดเดิมอีกครั้ง เพื่อหาส่วนต่างของประสิทธิภาพที่เพิ่มขึ้น แล้วสรุปงานวิจัยเพื่อนำไปพัฒนาวิธีการคัดแยกให้สามารถใช้งานในระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่

3.2 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้ในงานวิจัยครั้งนี้ คือ เครื่อง Desktop Computer หรือ Laptop สำหรับติดตั้ง Software ในการทดสอบคัดแยกข้อความ จำนวน 1 เครื่อง ซึ่งมีคุณสมบัติดังต่อไปนี้

1) CPU Intel Core 2 Duo Processor 2.40 GHz
2) Mainboard
3) RAM 3 GB DDRII 800 MHz
4) HARD DISK 320 GB SATA II
5) DVD - RW
6) 10/100/1000 LAN Built-In
7) USB Mouse & Keyboard
8) Mini Tower Case
9) LCD 13.1 Inch Display
10) Genuine Windows Vista® Business

3.3 แผนการดำเนินงาน

ตารางที่ 3.1 แผนการดำเนินงาน

รายการดำเนินงาน	ระยะเวลา (เดือน)							
	1	2	3	4	5	6	7	8
รวบรวมข้อมูลจากผู้ใช้งานโทรศัพท์เคลื่อนที่								
รวบรวมข้อมูล(CDR)								
ศึกษา Classify ที่มีการใช้งานในปัจจุบัน								
ออกแบบและพัฒนาระบบการคัดแยกข้อความ								
ทดสอบการประมวลผลของระบบขั้นต้น								
ทดสอบเปรียบเทียบเพื่อหาประสิทธิภาพ								
สรุปผลการเปรียบเทียบและประโยชน์								

3.4 ขั้นตอนการดำเนินงานวิจัย

3.4.1 รวบรวมและวิเคราะห์ข้อความ SMS ในประเทศไทย

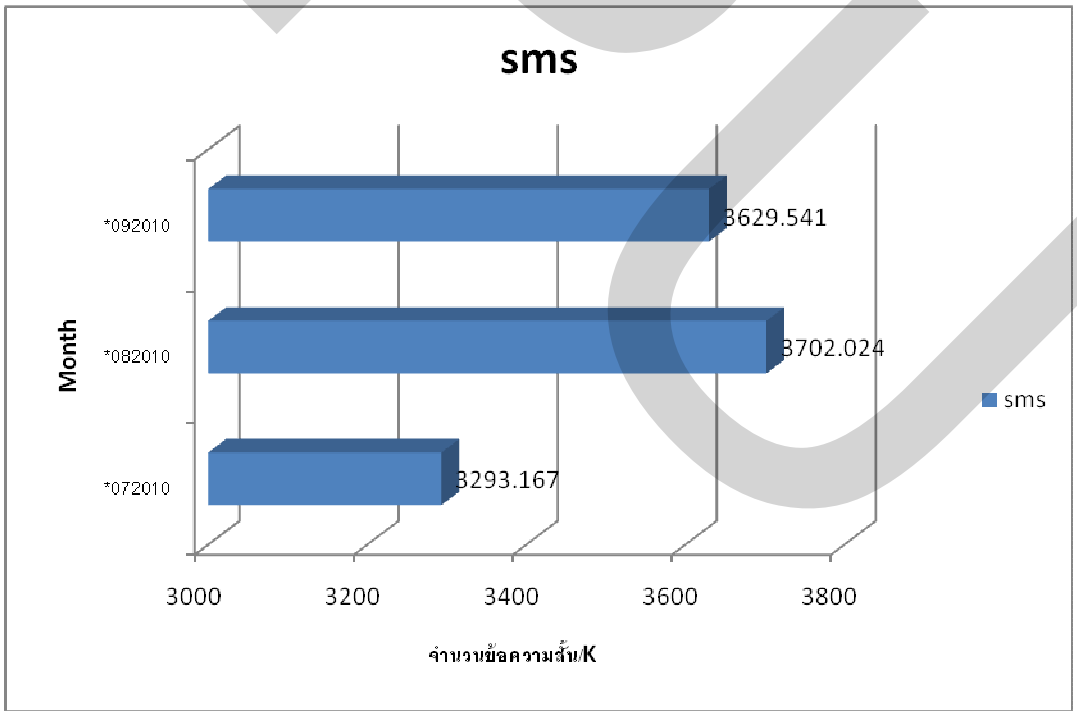
ก่อนทำการออกแบบเพื่อสร้างระบบคัดแยกนั้น จำเป็นต้องมีการวิเคราะห์พฤติกรรมของระบบ SMS เพื่อเป็นแนวทางในการพัฒนา โดยจะทำการรวบรวมข้อมูลการใช้งานบริการส่งข้อความ SMS จากระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยรายหนึ่ง เป็นระยะเวลา 3 เดือน เพื่อใช้ในการศึกษาลักษณะและแนวโน้มทางสถิติของข้อความในระบบส่งข้อความ โดยมีวิธีดำเนินการดังนี้

1) ติดต่อผู้ให้บริการโทรศัพท์เคลื่อนที่ เพื่อขอควาอนุเคราะห์ข้อมูลการ รับ – ส่งข้อความสั้น SMS เป็นระยะเวลา 3 เดือน

2) ทำการเก็บรวบรวม CDR ของบริการ SMS ที่เครื่อง SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่

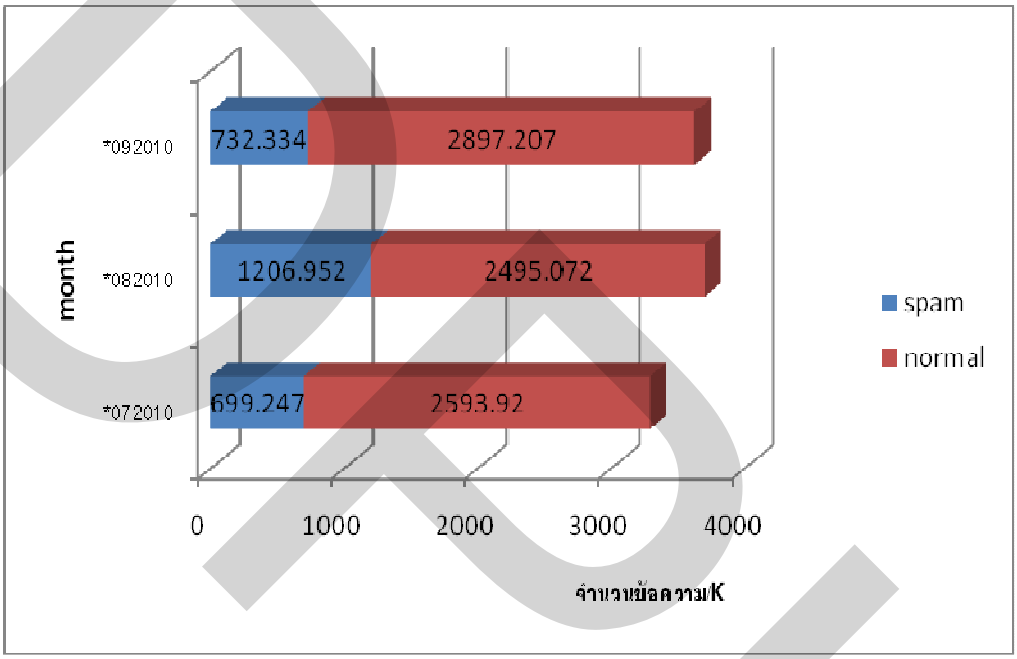
3) นำข้อความ SMS ที่ได้รับมาจัดเตรียมลงสู่ฐานข้อมูล เพื่อใช้ในการเรียกดึงข้อมูลได้อย่างเป็นระบบและมีความรวดเร็ว

4) สร้างกราฟแสดงข้อมูลทางสถิติซึ่งมีรายละเอียดดังนี้



ภาพที่ 3.1 ข้อมูลการใช้บริการ SMS ทั้งหมด ของ CAT CDMA เป็นระยะเวลา 3 เดือน ตั้งแต่เดือนกรกฎคม – กันยายน พ.ศ. 2553

จากภาพที่ 3.1 ซึ่งแสดงให้เห็นแนวโน้มการใช้บริการที่เพิ่มสูงขึ้นในแต่ละเดือน และจะมีเพิ่มสูงมากขึ้นในเดือนที่มีเทศกาลพิเศษ เช่น เดือนสิงหาคม ที่มีเทศกาลวันแม่แห่งชาติ พบว่ามีอัตราการส่งข้อความโฆษณาประชาสัมพันธ์ ที่เกี่ยวข้องหรือส่งเสริมเรื่องของครอบครัวมากขึ้น อีกทั้งงานด้านบริการด้านต่างๆ ที่ใช้ช่องทาง SMS ในการส่งข้อมูลให้กับลูกค้าที่เพิ่มมากขึ้นทุกวัน



ภาพที่ 3.2 ข้อมูลการใช้บริการโฆษณาประชาสัมพันธ์ทาง SMS ของ CAT CDMA เป็นระยะเวลา 3 เดือน ตั้งแต่เดือน กรกฎคม – กันยายน พ.ศ. 2553

จากภาพที่ 3.2 ซึ่งแสดงให้เห็นแนวโน้มการใช้บริการโฆษณาประชาสัมพันธ์ทาง SMS ที่มีการเพิ่มสูงขึ้นในแต่ละเดือน โดยเฉพาะการส่งข้อความโฆษณาประชาสัมพันธ์ในเดือนที่มีเทศกาลพิเศษ เช่น เทศกาลวันแม่แห่งชาติ จะมีสถิติการใช้งานที่สูงขึ้นเป็นพิเศษ โดยเมื่อนำข้อมูลเหล่านี้มาหาสัดส่วนของการส่งข้อความ พบว่าจำนวนข้อความโฆษณาประชาสัมพันธ์เหล่านี้ต่อเดือน ถึง 20 – 25 % เมื่อเทียบกับจำนวนของข้อความในระบบทั้งหมด และขณะที่อัตราการการส่งข้อความจากบุคคลทั่วไปหรือจากบริการด้านอื่นๆยังคงมีปริมาณเท่าเดิมหรืออาจเพิ่มมากขึ้นตามช่วงเทศกาลต่างๆ เมื่อมีปริมาณข้อความในระบบมากขึ้นอัตราการส่งข้อความปกติหรือข้อความที่มีความสำคัญก็อาจถูกส่งให้ผู้รับช้าลงกว่าเดิม

3.4.2 การกำหนดประเภทของข้อความ SMS

การคัดแยกข้อความจากระบบส่งข้อความ SMS นั้น จำเป็นต้องกำหนดความหมายของประเภทข้อความแต่ละประเภทให้ชัดเจน ในขณะที่ผู้ใช้งานโทรศัทพ์เคลื่อนที่แต่ละคนอาจมีทัศนคติในการตัดสินว่าข้อความใดเป็นข้อความ Spam หรือข้อความปรกติที่แตกต่างกัน แต่การแยกประเภทของข้อความบางประเภทอาจมีความชัดเจนกว่า เช่น ข้อความบริการ ข้อความด้านการเงินที่ผู้ใช้งานโทรศัพท์เคลื่อนที่จำเป็นต้องได้รับเพื่อให้การดำเนินกิจกรรมต่างๆเป็นไปได้อย่างต่อเนื่องและมีความถูกต้อง

จากการเก็บข้อมูลตัวอย่างใน SMSC เป็นระยะเวลา 3 เดือน ตั้งแต่กรกฎคม – กันยายน พ.ศ. 2553 ตามภาพที่ 3.2 พบว่าอัตราการส่งข้อความมีแนวโน้มที่เพิ่มขึ้นตลอดเวลาโดยเฉพาะในช่วงเดือนที่มีเทศกาลหรือวันสำคัญต่างๆ เช่น วันที่ 12 สิงหาคม ที่คนส่วนใหญ่นิยมส่งข้อความอวยพรหรือบอกรักแม่ ทำให้ปริมาณการใช้งาน SMS สูงขึ้น แต่ข้อความที่จัดอยู่ในกลุ่มของโฆษณาก็สูงขึ้นด้วยเช่นกัน เมื่อเทียบแล้วมีปริมาณถึงร้อยละ 25 ของข้อความทั้งหมด และเมื่อแยกข้อความที่ไม่ใช่โฆษณาออกแล้วก็ยังมีข้อความบริการอื่นๆอีก เช่น การนัดตรวจสุขภาพของโรงพยาบาล, การแจ้งผลการใช้ยอดบัตรเครดิต, การส่งรายงานการใช้งานโทรศัพท์มือถือและอื่นๆ ซึ่งข้อความเหล่านี้จัดเป็นข้อความดีที่มีความสำคัญสูงกว่าข้อความทั่วไป ความจำเป็นด้านข้อมูลดังตัวอย่างข้างต้น ทำให้การกำหนดประเภทของความหมายในข้อความ SMS มีความจำเป็นมากขึ้นและเพื่อใช้เป็นข้อมูลฝึกสอนสำหรับออกแบบวิธีกรองข้อความ ให้สามารถกำหนดระดับความสำคัญของข้อความได้ถูกต้องมากขึ้น โดยในปัจจุบันกลุ่มของผู้ใช้บริการ SMS มีหลายประเภทและแต่ละกลุ่มผู้ใช้บริการก็จะมีลักษณะของการใช้บริการรับ-ส่งข้อความแตกต่างกันออกไปตามวัตถุประสงค์ของการใช้งาน เราสามารถแยกกลุ่มผู้ใช้งาน SMS ได้ดังนี้

1) ผู้ให้บริการต่างๆ เช่น ผู้ให้บริการโทรศัพท์มือถือ
2) ธุรกิจ, บริษัท เช่น บริษัทขายตรง
3) องค์กร, ภาครัฐ เช่น โรงพยาบาล, หน่วยงานของรัฐ
4) บุคคลทั่วไป

โดยกลุ่มผู้ส่งแต่ละกลุ่มมีวัตถุประสงค์ในการส่งข้อความแตกต่างกัน ทำให้เราสามารถกำหนดประเภทของข้อความได้ง่ายขึ้นโดยพิจารณาจากกลุ่มผู้ส่งและเนื้อหาของข้อความเป็นหลัก ดังนั้นสามารถแบ่งประเภทของข้อความออกเป็นกลุ่มๆ ได้ดังนี้

1) ข้อความโฆษณาประชาสัมพันธ์ เป็นข้อความเสนอขายสินค้าหรือบริการต่างๆ

2) ข้อความด้านการบริการ ส่วนใหญ่เป็นข้อความทางด้านการแพทย์,การศึกษา,การใช้งานระบบต่างๆ

3) ข้อความแจ้งเตือน หรือข้อความที่ต้องรับรู้ ที่อาจส่งผลต่อการดำเนินธุรกิจหรือการใช้ชีวิตประจำวันต่างๆ เช่น การเงิน แจ้งเตือนภัยต่างๆ

4) ข้อความทั่วไป เป็นข้อความทั่วไปที่มีการส่งในชีวิตประจำวัน

ตารางที่ 3.2 ตัวอย่างประเภทของข้อความ SMS

ประเภท	**ข้อความ**
ข้อความแจ้งเตือน	โอนเงินเข้า KBank 2767XXXX ผ่าน K-ATM 1,800 บ.
ข้อความด้านการบริการ	2-5 missed call from 0867367583@ 30/07/2010 17:44
ข้อความทั่วไป	จองหนังให้แล้วนะ รีบมาด้วย
ข้อความโฆษณาประชาสัมพันธ์	ดาวน์โหลดริงโทน พิมพ์ ok ส่งที่ *123456

3.4.3 ศึกษาวิธีการคัดแยกข้อความ SMS ในปัจจุบัน

โดยค้นคว้างานวิจัยประเภทต่างๆดังต่อไปนี้

1) งานวิจัยที่เกี่ยวข้องกับระบบส่งข้อความ SMS

2) งานวิจัยที่เกี่ยวข้องกับการจัดหมวดหมู่เอกสารภาษาไทย

3) งานวิจัยที่เกี่ยวข้องกับการตรวจสอบข้อความ Spam ทั้งจากระบบส่งข้อความสั้น SMS และระบบรับ-ส่ง E-Mail

4) งานวิจัยที่เกี่ยวข้องกับ Algorithm ในการจัดกลุ่มข้อมูล

5) งานวิจัยที่เกี่ยวข้องกับการตัดคำภาษาไทย

พัฒนาโปรแกรมสำหรับคัดแยกข้อความโดยวิธีกรองแบบ NB ด้วยภาษา PHP โดยมีขั้นตอนดังนี้

1) จัดเตรียม Module การตรวจสอบ Stop words โดยการพัฒนาจากภาษา PHP

2) จัดเตรียม Module การตัดคำภาษาไทย ที่สามารถตัดคำด้วยวิธีการดังต่อไปนี้

2.1) Longest Matching

2.2) Maximal Matching

2.3) Probabilistic Model

โดยใช้โปรแกรม Swath[1] ของ NECTEC มี license แบบ GNU GPL และพัฒนาวิธีการเชื่อมต่อข้อมูลระหว่าง PHP กับ Swath เพิ่มเติมเพื่อให้ PHP เข้าใจและรับรู้การตัดคำของ Swath ได้ถูกต้อง

3) จัดเตรียม Module การแปลงข้อมูล text เป็น feature vector ตามวิธีการ TFIDF โดยพัฒนาด้วย PHP

4) จัดเตรียม Module การประมวลผลด้วย NB โดยการพัฒนาจากภาษา PHP ตามทฤษฎี[2-3]

จากสมการของ Bayes's rule จะเห็นว่าหลักฐานที่ใช้สนับสนุนสำหรับการคัดแยกกลุ่มข้อความมีเพียงอย่างเดียวเท่านั้น ซึ่งในความเป็นจริงแล้วการที่มีหลักฐานสนับสนุนในการตัดสินใจที่มากกว่าหนึ่งอย่างนั้น จะทำให้การคัดแยกข้อความมีความถูกต้องและแม่นยำมากยิ่งขึ้น Naïve Bay เป็นเทคนิคในการแก้ปัญหาแบบ classification ที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างสิ่งที่ต้องการ classification แต่ละตัวกับหลักฐานที่ใช้ประกอบการตัดสินใจเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ ในทางทฤษฎีแล้วการทำนายผลของ Naïve-Bayes จะถูกต้องถ้าหลักฐานทั้งหมดเป็นอิสระต่อกัน ไม่ขึ้นกับหลักฐานตัวใดตัวหนึ่ง

Naïve Bayes ถูกใช้อย่างกว้างขว้างในหลายๆ แอพพลิเคชั่น ซึ่งแอพพลิเคชั่นหนึ่งที่ได้รับความสนใจคือ text classification และ information filtering (เช่น spam filtering) เหตุผลหลักคือ NB model ทำงานได้ดีสำหรับ text domain เพราะหลักฐานที่ใช้ในการแบ่งกลุ่มเป็น "vocabularies" หรือ "word" รูปแบบของ text ที่จะนำมาใช้สำหรับเป็นหลักฐานในการประกอบการแบ่งประเภทของข้อความนั้นมีรูปแบบแน่นอนที่ปรากฏในกลุ่มของข้อความที่มีจำนวนเป็นพันๆ หรือมากกว่านั้น ขนาดของหลักฐานที่ใหญ่อย่างนี้ทำให้ NB model ทำงานได้ดีสำหรับการแก้ปัญหา text classification ดังนั้นในงานวิจัยนี้จึงได้ประยุกต์เทคนิค Naïve Bayes สำหรับแบ่งกลุ่มของข้อความ โดยในปัจจุบันกลุ่มของผู้ใช้บริการ SMS มีหลายประเภทและแต่ละกลุ่มผู้ใช้บริการก็จะมีลักษณะของการใช้บริการรับ-ส่งข้อความแตกต่างกันออกไปตามวัตถุประสงค์ของการใช้งาน

---

[1] ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, (2537). โปรแกรมตัดคำภาษาไทย. สืบค้นเมื่อ 25 มกราคม 2553. จาก **http://www.hlt.nectec.or.th/products/swath.php**

[2] Wikipedia. (2009). Naive Bayes classifier. Retrieved August 2, 2009. From **http://en.wikipedia.org/wiki/Naive_bayes**

[3] Wikipedia. (2009). Bayesian network. Retrieved August 2, 2009. From **http://en.wikipedia.org/wiki/Bayesian_network**

จากประเภทของข้อความผู้วิจัยนำเอาข้อความจำนวนหนึ่งมาฝึกเพื่อสร้าง NB Model จากตารางที่ 3.2 เป็นตัวอย่างของการนำเอาข้อความ S0–Sn เป็นข้อความที่นำมาฝึก Model และกำหนดคีย์เวิร์ดที่ผู้วิจัยกำหนดในแต่ละข้อความ โดยระบบจะนับจำนวนความถี่ที่พบคีย์เวิร์ดนั้นๆ จากนั้นกำหนดหมวดหมู่ข้อความที่นำมาฝึกโมเดล ในที่นี้มี 4 ประเภท ตามตารางที่ 3.2 และกำหนดให้แต่ละประเภทของข้อความแทนด้วยตัวแปรดังนี้

ข้อความแจ้งเตือน คือ Class A

ข้อความด้านการบริการ คือ Class B

ข้อความทั่วไป คือ Class C

ข้อความโฆษณาประชาสัมพันธ์ คือ Class D

ตารางที่ 3.3 แสดงตัวอย่างการนำทฤษฎีนาอีฟ เบย์เซียนมาใช้ในการจัดหมวดหมู่ข้อความ

Training SMS	**K1**	**K2**	**K3**	**K4**	**…**	**Kn**	**Type**
S0	NF	NF	NF	NF	NF	NF	A
S1	NF	NF	NF	NF	NF	NF	A
S2	NF	NF	NF	NF	NF	NF	B
S3	NF	NF	NF	NF	NF	NF	B
S4	NF	NF	NF	NF	NF	NF	C
S5	NF	NF	NF	NF	NF	NF	C
….	…	…	…	…	…	…	D
Sn	NF	NF	NF	NF	NF	NF	D

โดยที่

Sn คือ ข้อความที่นำมาเป็นข้อมูลฝึกฝน

Kn คือ คีเวิร์ดที่นำมาเป็นหลักฐานในการจัดประเภทข้อความ

NF คือ จำนวนความถี่ของคำที่เกิดขึ้นในข้อความนั้นๆ

จากค่าต่างๆ ในตารางที่ 3.3 นำมาคำนวณหาค่าความน่าจะเป็นของเอกสารแต่ละประเภทที่พบคีย์เวิร์ดต่างๆ โดยระบบจะทำการคำนวณค่าความน่าจะเป็นทุกๆ คีย์เวิร์ดในเอกสารแต่ละประเภท ดังจะได้ค่าดังตารางที่ 3.4

ตารางที่ 3.4 แสดงค่าต่าง ๆ ที่ได้จากการคำนวณตามทฤษฎี Naïve Bayes

$\|V\|$	C	$P(C_i)$	n_i	P(W1|Ci)	P(W2|Ci)	P(W3|Ci)	P(W4|Ci)	P(W5|Ci)	P(Wj|Ci)
n	A	$P(C_A)$	n_i	$P(W_1 \mid C_A)$	$P(W_2 \mid C_A)$	$P(W_3 \mid C_A)$	$P(W_4 \mid C_A)$	$P(W_5 \mid C_A)$	$P(W_j \mid C_A)$
	B	$P(C_B)$	n_i	$P(W_1 \mid C_B)$	$P(W_2 \mid C_B)$	$P(W_3 \mid C_B)$	$P(W_4 \mid C_B)$	$P(W_5 \mid C_B)$	$P(W_j \mid C_B)$
	C	$P(C_C)$	n_i	$P(W_1 \mid C_C)$	$P(W_2 \mid C_C)$	$P(W_3 \mid C_C)$	$P(W_4 \mid C_C)$	$P(W_5 \mid C_C)$	$P(W_j \mid C_C)$
	D	$P(C_D)$	n_i	$P(W_1 \mid C_D)$	$P(W_2 \mid C_D)$	$P(W_3 \mid C_D)$	$P(W_4 \mid C_D)$	$P(W_5 \mid C_D)$	$P(W_j \mid C_D)$

โดยที่

$|V|$ คือ จำนวนของ keyword

n_i = จำนวนครั้งที่พบคีย์เวิร์ดทั้งหมดในข้อความแต่ละหมวดหมู่

$P(C_i)$ ความน่าจะเป็นของแต่ละ class

$$P(W_j \mid C_i) = \frac{Number\ of\ SMS\ in\ Class}{All\ of\ SMS} \qquad (3\text{-}1)$$

ดังนั้นการคำนวณหาค่าความน่าจะเป็นของข้อความแต่ละประเภทที่พบคีเวิร์ดต่างๆ ที่เกิดขึ้นในแต่ละ Class และเพื่อเป็นการหลีกเลี่ยงปัญหาความถี่ 0 (zero frequency) เราประยุกต์สมการของการคำนวณโดยการสมมติให้รูปแบบการกระจายทั้งหมดของคีเวิร์ดทั้งหมดนั้นเท่ากัน ดังสมการที่ 3-2

$$P(W_j \mid C_i) = \frac{n_{ij} + 1}{n_i + |V_i|} \qquad (3\text{-}2)$$

จากค่าความน่าจะเป็นของทุก w_i หรือทุกคีย์เวิร์ด ในตารางที่ 3.5 นำข้อความที่ต้องการจัดหมวดหมู่หรือTest Data นั้นไปคำนวณหาค่าความน่าจะเป็นเพื่อหาหมวดหมู่ที่เหมาะสมกับข้อความดังกล่าวโดยเลือกค่าความน่าจะเป็นของประเภทที่มากที่สุด ตามสมการ Argmax P(Ci)* P(wi| Ci)Dt ดังนี้

ตารางที่ 3.5 แสดงข้อมูลทดสอบ

Test Data	K1	K2	K3	K4	…	Kn	Type
S0	NF	NF	NF	NF	NF	NF	?

$$P(C_i \mid W) = (P(C_i) \times \prod_{j=1}^{v} P(w_j \mid C_i) \quad (3\text{-}3)$$

ซึ่งผลที่ได้จากการคำนวณหาค่าความน่าจะเป็นมีค่าน้อยมาก จำนวนของเงื่อนไขความน่าจะเป็นในขอบเขตของคำเป็นพันหรือมากกว่า จะมีค่าต่ำมากสำหรับการจัดการของ CPU นี่เป็นปัญหาที่ถูกอ้างถึง ภายใต้ปัญหานี้ การแก้ปัญหานี้ เราสามารถ take logarithm บนความน่าจะเป็นนี้ ดังสมการที่ 3-4

$$P(C_i \mid W) = \log(P(C_i) \times \prod_{j=1}^{v} P(w_j \mid C_i)) \quad (3\text{-}4)$$

หลังจากผ่านการคำนวณหาค่าความน่าจะเป็นของข้อความจาก Naïve Bayes model แล้ว ต่อมาก็ตัดสินใจว่าข้อความนั้นจัดอยู่ในประเภทไหน ดังนั้นจึงนำ Decision Rule มาใช้สำหรับการตัดสินใจด้วย ซึ่ง Rule ที่ใช้กันโดยปกติทั่วไปคือ การพิจารณาว่าค่าความน่าของ Class ไหนมีค่าสูงที่สุด ก็จัดว่าเป็นประเภทนั้น ดังสมการที่ 3-5

$$classify(c_1,..,c_n) = \arg\max\ p(W = w)\prod_{i=1}^{n} p(C_i = C_i \mid W = w) \quad (3\text{-}5)$$

3.4.4 ออกแบบการทดสอบประสิทธิภาพวิธีการคัดแยกข้อความ

การทดสอบประสิทธิภาพ ต้องจัดเตรียมชุดข้อมูลสำหรับทำการทดสอบ โดยนำข้อความ SMS ที่ผ่านการคัดแยกด้วยมนุษย์มาแบ่งออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลฝึกสอน (Training data หรือ TD) และชุดข้อมูลใหม่สำหรับทำการทดสอบในการทำงานจริงที่ชุดข้อมูลตรวจสอบอาจไม่มีในชุดข้อมูลฝึกสอน (New data หรือ ND) โดยวิธีการกรองแต่ละวิธีจะทำการเรียนรู้จากชุดข้อมูลฝึกสอนเพื่อให้เข้าใจความหมายของข้อความแต่ละประเภทและสามารถ คัดแยกได้ แล้วจึงวัดประสิทธิภาพด้วยการกรองชุดข้อมูลฝึกสอนอีกครั้งร่วมกับข้อมูลใหม่

จัดเตรียมเครื่อง Computer หรือ Laptop สำหรับทำการทดสอบ ซึ่งประกอบด้วย CPU Core 2 Duo Processor ความเร็ว 2.40 GHz หน่วยความจำขนาด 3 GB ระบบปฏิบัติการ Windows Vista Business โดยแบ่งการทดสอบออกเป็น

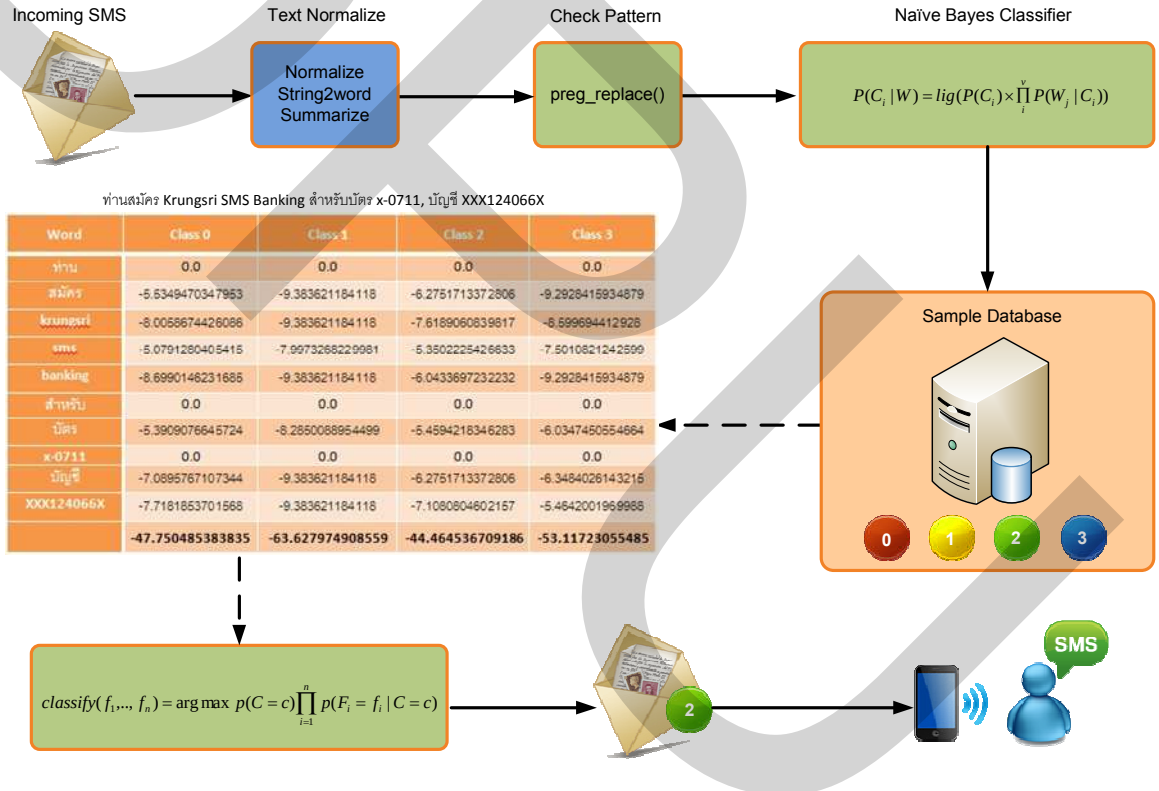
1) ทำการทดสอบด้วยวิธีการตัดคำและทำ Text Normalize ภาษาไทย

2) ทำการทดสอบการจัดกลุ่มรูปแบบคำภาษาไทยปนอังกฤษเพื่อแยกประเภทกลุ่มคำ

3) ทำการทดสอบการตัดคำที่พบบ่อยและไม่มีผลกับการคัดแยกประเภทข้อความออกก่อนทำการคัดแยกข้อความแบบ NB

4) ทำการทดสอบการ Training ข้อมูล SMS แบบ NB

5) ทำการทดสอบโดยใช้วิธีคัดแยกข้อความแบบ NB



ภาพที่ 3.3 การคัดแยกข้อความแต่ละประเภท

จากภาพที่ 3.3 แสดงให้เห็นว่าเมื่อมีการส่งข้อความ “ท่านสมัคร Krungsri SMS Banking สำหรับบัตร X-0711, บัญชี XXX124066X” ระบบจะทำ Text Normalize ข้อความเพื่อให้ข้อความมีความสมบูรณ์ก่อนทำการ Check Pattern คำในข้อความเพื่อจัดกลุ่มคำก่อนทำการตัดคำและคัดกรองประเภทของข้อความต่อไป

Word	Class 0	Class 1	Class 2	Class 3
ท่าน	0.0	0.0	0.0	0.0
สมัคร	-5.5349470347953	-9.383621184118	-6.2751713372806	-9.2928415934879
krungsri	-8.0058674426086	-9.383621184118	-7.6189060839817	-8.599694412928
sms	-5.0791280405415	-7.9973268229981	-5.3502225426633	-7.5010821242599
banking	-8.6990146231685	-9.383621184118	-6.0433697232232	-9.2928415934879
สำหรับ	0.0	0.0	0.0	0.0
บัตร	-5.3909076645724	-8.2850088954499	-5.4594218346283	-6.0347450554664
x-0711	0.0	0.0	0.0	0.0
บัญชี	-7.0895767107344	-9.383621184118	-6.2751713372806	-6.3484026143215
XXX124066X	-7.7181853701568	-9.383621184118	-7.1080804602157	-5.4642001969968
	-47.750485383835	**-63.627974908559**	**-44.464536709186**	**-53.11723055485**

ภาพที่ 3.4 ตารางคำที่ใช้ในการคัดแยกประเภทของข้อความ

จากภาพที่ 3.4 จะเห็นว่าตารางการคำนวณนั้นได้ตัดคำที่ไม่มีผลต่อการคำนวณออกให้เหลือเฉพาะคำที่คาดว่าจะมีผลต่อการแยกประเภทข้อความเท่านั้น และคำที่ระบบได้ตัดออกไปจะเป็นคำที่พบว่าใช้บ่อยในข้อความทั่วไป เช่น คำว่า “ฉัน”, “เธอ”, “ท่าน”

เมื่อการทดสอบเสร็จสิ้น ดำเนินการเก็บข้อมูลการทดสอบเพื่อใช้ในการศึกษาปัญหาการกรองข้อความในภาษาไทยต่อไป

3.4.5 ศึกษาปัญหาและการแก้ไขปัญหาการกรองข้อความ SMS

ศึกษาความเป็นไปได้และปัญหาที่จะเกิดขึ้นในการกรองภาษาไทย ของวิธีกรองข้อความจากการทดสอบประสิทธิภาพ ได้แก่

1) การตรวจสอบความผิดปรกติต่างๆของคำในข้อความ

ทั้งข้อความภาษาไทย ภาษาอังกฤษและภาษาไทยปนภาษาอังกฤษ ที่ถูกสร้างจากโทรศัพท์เคลื่อนที่ การพิมพ์ข้อความที่ผิดพลาด การพิมพ์ตัวอักษรเดียวกันมากกว่า 1 ครั้ง และคำที่ไม่สมบูรณ์เนื่องจากมีการตัดแบ่งข้อความที่ไม่ถูกต้อง และออกแบบวิธีการแก้ปัญหาดังกล่าว

2) การตรวจสอบการตัดคำ

วิธีการตัดคำในปัจจุบัน เป็นวิธีการที่ออกแบบมาเพื่อตัดคำที่มีความถูกต้องตามหลักภาษาศาสตร์ ซึ่งไม่สามารถตัดคำจากข้อความ SMS ได้อย่างถูกต้อง การตรวจสอบวิธีตัดคำ จะทำ

ให้เข้าใจความผิดพลาดที่เกิดขึ้น และสามารถปรับปรุงแก้ไขให้การตัดคำในข้อความ SMS มีความถูกต้องมากขึ้น

3) การปรับปรุงในส่วนอื่นๆ

นอกจากความผิดปรกติของคำ และการตัดคำที่ไม่ถูกต้องแล้ว อาจมีองค์ประกอบอื่นๆ ที่สามารถปรับปรุงและส่งผลกระทบต่อการกรองข้อความให้มีประสิทธิภาพสูงขึ้นได้

3.4.6 การทดสอบประสิทธิภาพวิธีการคัดแยกที่พัฒนาเสร็จสิ้น

ทำการทดสอบเช่นเดียวกับการทดสอบก่อนหน้า โดยเปรียบเทียบระหว่างวิธีการคัดแยกที่ยังไม่ผ่านการปรับปรุงขั้นตอนต่างๆกับวิธีการคัดแยกข้อความที่ผ่านการปรับปรุงการแก้ไขข้อผิดพลาด

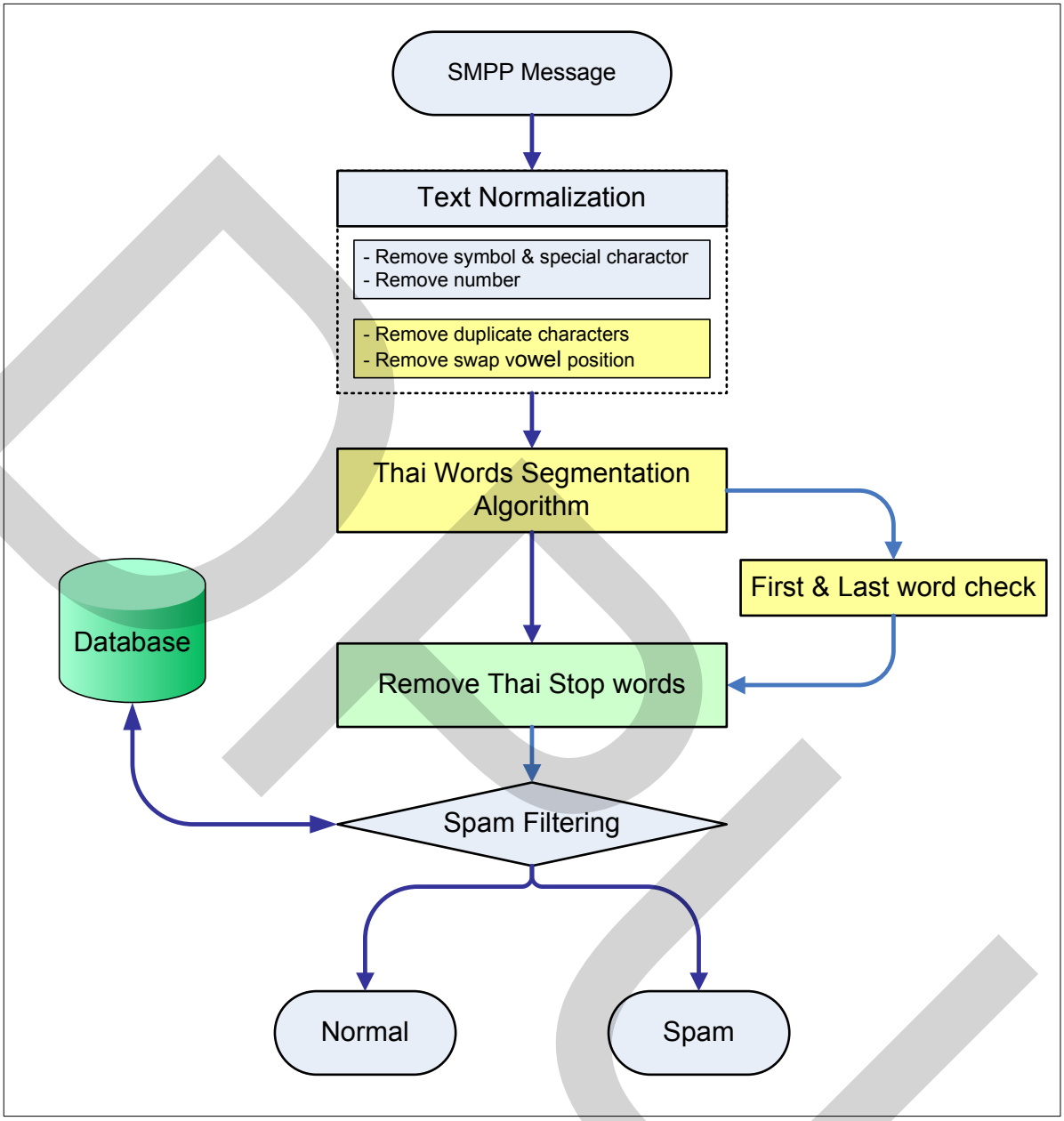
3.4.7 รายงานผลการวิจัยและสรุปข้อเสนอแนะ

แสดงผลการวิจัยและสรุปผลการดำเนินงาน พร้อมทั้งข้อเสนอแนะต่างๆ จัดทำรายงานการวิจัย และนำเสนอผลงาน

3.5 Model ที่นำเสนอในการทดสอบระบบคัดกรองข้อความสั้น

จากงานวิจัยที่ผ่านมามีการนำเสนอระบบวิธีการคัดกรองข้อความในแบบต่างๆกัน เช่น การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่[4] ที่ใช้ทฤษฎี NB เข้ามาคัดแยกข้อความ และมีการทำ Normalize ข้อความก่อนที่จะนำมาคัดกรอง ตามภาพที่ 3.5

[4] นนท์ บุญนิธิประเสริฐ. (2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่.** วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.



ภาพที่ 3.5 ขั้นตอนการทำงานของ SMS Spam filter

ที่มา: นนท์ บุญนิธิประเสริฐ. (2552). การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

จากภาพที่ 3.5 สามารถสรุปลำดับการกรองข้อความ SMS ที่รองรับภาษาไทย ได้ดังนี้

1) รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message

2) ทำกระบวนการ Text Normalization โดยการ

2.1) ลบสัญลักษณ์พิเศษต่างๆ

2.2) ลบตัวเลข

2.3) ลบ Character ที่เกินกว่า 1 ตัว

2.4) สลับตำแหน่งสระและวรรณยุกต์ให้ถูกต้อง

3) ตัดคำด้วยวิธีการที่รองรับภาษาไทย โดยนำคำแรกและคำสุดท้ายของข้อความไปผ่านกระบวนการตัดคำอีกครั้ง เพื่อตรวจสอบคำที่ไม่สมบูรณ์

4) ลบคำประเภท Stop words ทั้งภาษาอังกฤษ และภาษาไทย

5) นำคำที่ผ่านขั้นตอนทั้งหมดเข้าสู่กระบวนการกรองข้อความ

6) จัดส่งข้อความที่ผ่านการตรวจสอบให้กับผู้รับ และบันทึกข้อความที่ไม่ผ่านการตรวจสอบลงในฐานข้อมูล เพื่อใช้ในการเรียนรู้ต่อไป

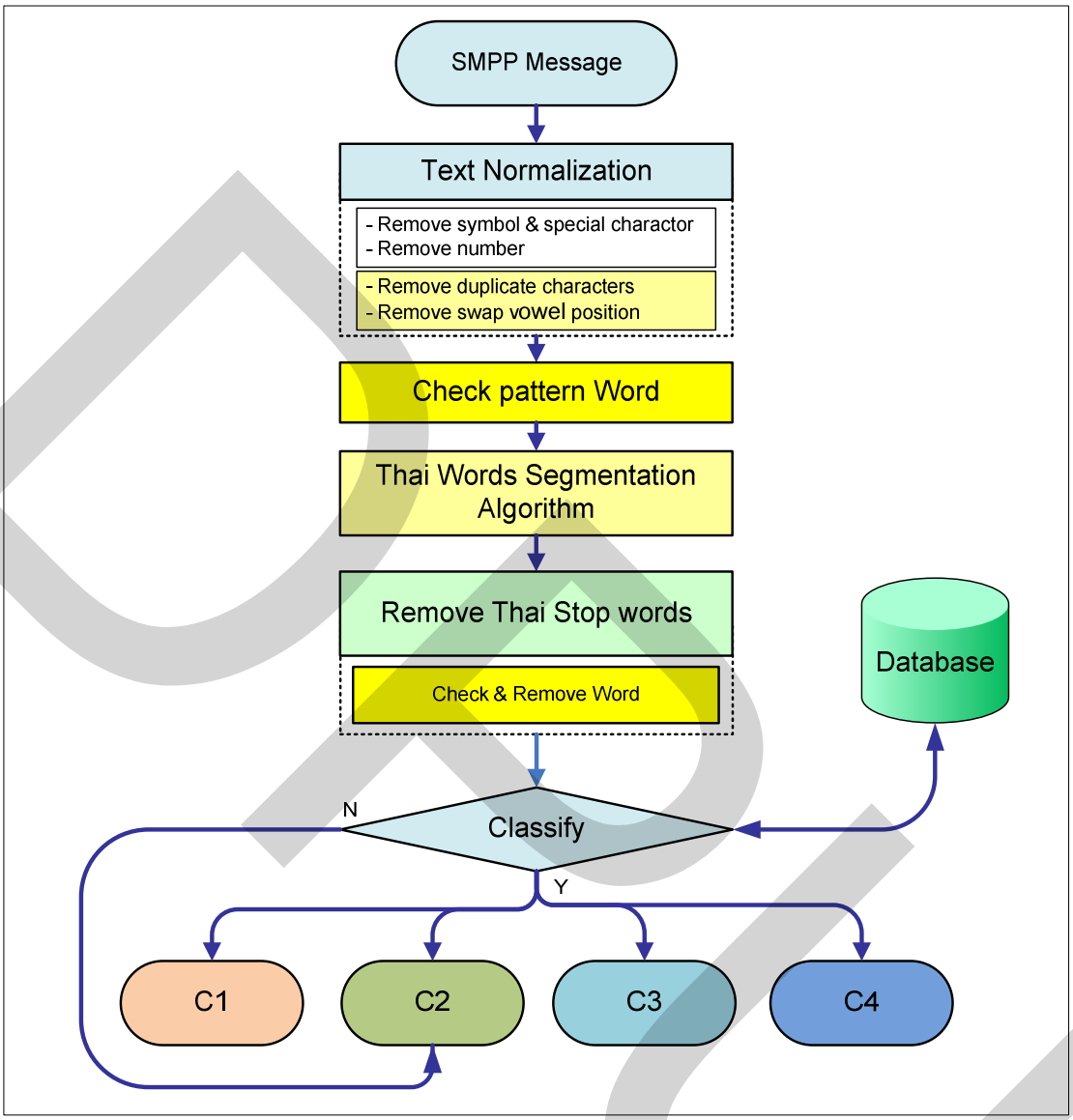
ตารางที่ 3.6 ผลการทดสอบการกรองข้อความแบบ NB

วิธีการกรอง	วิธีการตัดคำ	ข้อมูล	แบบที่ไม่มีการทำ Normalize		แบบที่มีการทำ Normalize		แบบที่มีการทำ Normalize ดีกว่า	
			ความถูกต้อง (%)	เวลา/SMS (millisec)	ความถูกต้อง (%)	เวลา/SMS (millisec)	ความถูกต้อง (%)	เวลา/SMS (%)
NB	ยาวที่สุด	TD	89.76	39.1673	89.84	39.6484	0.08	-1.2282
		ND	82.7095	42.4202	83.3832	45.0292	0.673652695	-6.1503
	ค่าเฉลี่ย		**86.1175**	**40.7814**	**86.5042**	**42.3280**	**0.3866**	**-4.2585**
	สอดคล้องที่สุด	TD	90.08	42.2466	89.92	45.1168	-0.16	-6.7938
		ND	82.5598	46.5521	82.9341	51.1245	0.3742	-9.8221
	ค่าเฉลี่ย		**86.1948**	**44.3771**	**86.3109**	**48.0991**	**0.1160**	**-8.5330**

ที่มา: นนท์ บุญนิธิประเสริฐ. (2552). การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

จากตารางที่ 3.6 แสดงให้เห็นว่า วิธีการกรองข้อความ Spam ด้วยทฤษฎี NB สามารถกรองข้อความได้อย่างรวดเร็ว แต่มีความถูกต้องประมาณร้อยละ 86.15 และวิธีการกรองที่มีการทำ Normalize ข้อความก่อนที่นำมาคัดกรอง สามารถเพิ่มความถูกต้องในการการกรองได้ประมาณร้อยละ 0.25 แต่ต้องใช้เวลาในการประมวลผลต่อข้อความสูงขึ้นประมาณร้อยละ 4.25 ทำให้วิธีการที่นำเสนอนี้ไม่มีความเหมาะสมในการกรองข้อความด้วยวิธีการแบบ NB

ในงานวิจัยนี้ได้นำเสนอทฤษฎี NB มาปรับปรุงให้สามารถทำงานได้เต็มประสิทธิภาพมากขึ้น โดยทำการจัดรูปแบบของคำเพื่อกำหนดกลุ่มคำของข้อความก่อนที่นำเข้าสู่การคัดกรองเพื่อเพิ่มความถูกต้องและปรับปรุงวิธีการคัดกรอง โดยตัดคำที่พบว่ามีการใช้บ่อยครั้งและไม่มีผลต่อการคำนวณเพื่อแยกประเภทของข้อความเพื่อเพิ่มความเร็วในการคำนวณ ทั้งนี้เพื่อให้เหมาะสมกับรูปแบบของข้อความในปัจจุบัน ตามภาพที่ 3.6



ภาพที่ 3. 6 ขั้นตอนการคัดแยกประเภทข้อความที่นำเสนอ

จากภาพที่ 3.6 สามารถสรุปลำดับการกรองข้อความ SMS แบบที่นำเสนอ ได้ดังนี้

1) รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message

2) ทำกระบวนการ Text Normalization

2.1) ลบสัญลักษณ์พิเศษต่างๆ เช่น ^_^ , +_+

2.2) ลบตัวเลข ที่ไม่มีผลในการคัดกรอง คือ ตัวเลขที่ไม่อยู่ในรูปแบบใดๆ เช่น ไม่เป็นทั้งเบอร์โทรศัพท์ หรือ ราคาสินค้า ฯลฯ

2.3) ลบ Character ที่เกินกว่า 1 ตัว เช่น สระ หรือ วรรณยุกต์ ต่างๆ

2.4) สลับตำแหน่งสระและวรรณยุกต์ให้ถูกต้อง

3) ตรวจสอบและแยกรูปแบบของคำเพื่อจัดกลุ่มคำ เช่น Email, Phone number, Account number

4) ลบคำประเภท Stop words ทั้งภาษาอังกฤษ และภาษาไทย

5) นำคำที่ผ่านขั้นตอนทั้งหมดเข้าสู่กระบวนการกรองข้อความแบบที่นำเสนอ

6) ข้อความที่ไม่สามารถระบุกลุ่มของข้อความได้ให้จัดอยู่ในกลุ่มของข้อความทั่วไป หรือ C2

7) จัดส่งข้อความที่ผ่านการตรวจสอบให้กับผู้รับ และบันทึกข้อความลงในฐานข้อมูล เพื่อใช้ในการเรียนรู้ต่อไป

บทที่ 4
การคัดแยกข้อความ SMS

4.1 ลักษณะของข้อความ SMS ในประเทศไทย

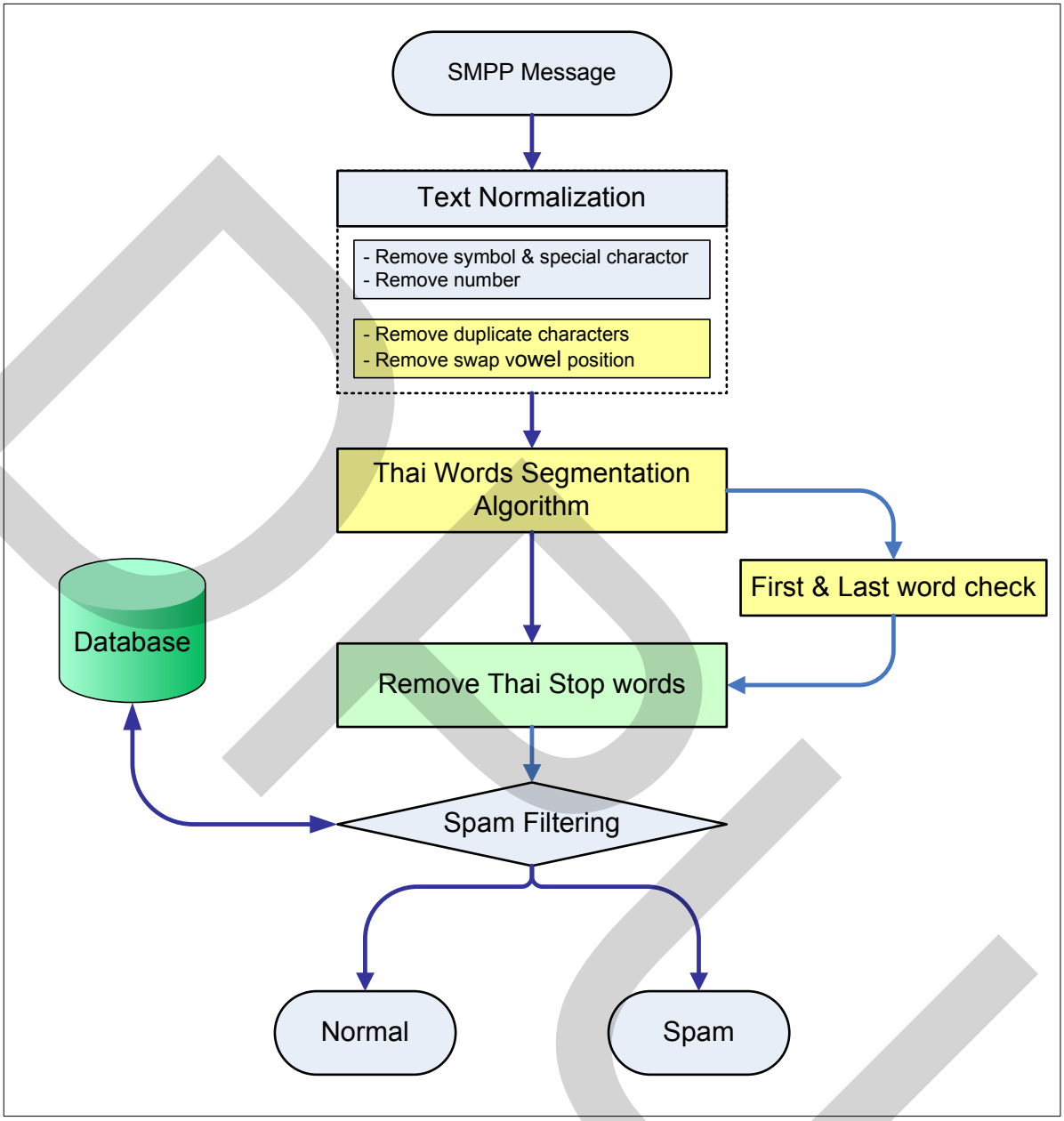
ลักษณะข้อความ SMS ของประเทศไทยหลังจากการเก็บข้อมูลเป็นระยะเวลา 3 เดือน พบว่า มีลักษณะของข้อความที่ใช้ไม่เป็นไปตามหลักภาษาศาสตร์ มีทั้งคำภาษาไทยและคำภาษาอังกฤษปะปนกันเป็นจำนวนมากและไม่มีรูปแบบการใช้งานที่แน่นอนของคำ ขึ้นอยู่กับวัตถุประสงค์การใช้งานของข้อความนั้นๆและเนื่องจาก SMS เป็นการสื่อสารที่ไม่จำเป็นต้องใช้ภาษาอย่างเป็นทางการ อีกทั้งข้อจำกัดของจำนวนตัวอักษรที่พิมพ์ได้ในข้อความ ทำให้ลักษณะการใช้งานส่วนใหญ่นิยมที่จะพิมพ์ข้อความให้ได้ใจความสำคัญครบทั่วในข้อความเดียว และจากการเก็บตัวอย่างข้อความ SMS สามารถแบ่งประเภทข้อความออกเป็น 4 ประเภทเพื่อให้เหมาะสมกับลักษณะการให้บริการข้อความ SMS ดังตัวอย่างจากตารางที่ 4.1

ตารางที่ 4.1 แสดงลักษณะข้อความที่พบในประเทศไทย

ประเภท	ข้อความ
ข้อความแจ้งเตือน	โอนเงินเข้า KBank 2767XXXX ผ่าน K-ATM 1,800 บ.
ข้อความด้านการบริการ	2-5 missed call from 0867367583@ 30/07/2010 17:44
ข้อความทั่วไป	จองหนังให้แล้วนะ รีบมาด้วย
ข้อความโฆษณาประชาสัมพันธ์	ดาวน์โหลดริงโทน พิมพ์ ok ส่งที่ *123456

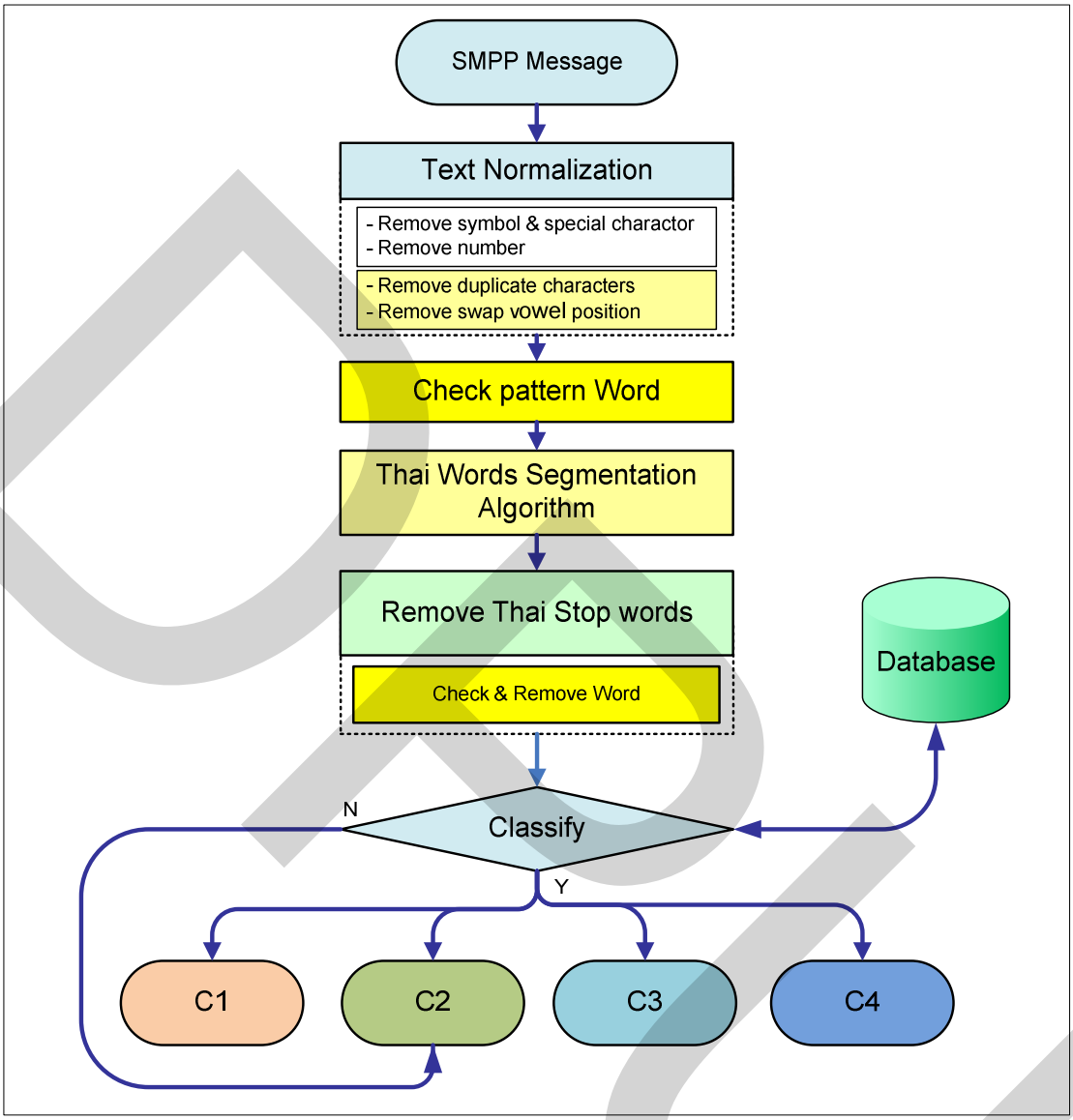
4.2 การคัดกรองข้อความแบบที่นำเสนอ

เป็นวิธีการคัดแยกประเภทข้อความ SMS ที่ทำการปรับเปลี่ยนเพิ่มเติมจากภาพที่ 4.1 ซึ่งรองรับภาษาไทยปนภาษาอังกฤษ โดยใช้การทำ TN ที่ปรับปรุงใหม่ให้สามารถทำกระบวนการ TN และลบ Stop words กับภาษาไทย แล้วลดกระบวนการตัดคำบางขั้นตอนลงเพื่อเพิ่มประสิทธิภาพทางเวลาให้เพิ่มขึ้น มีขั้นตอนการทำงานใหม่ตามภาพที่ 4.2



ภาพที่ 4.1 ขั้นตอนการกรองข้อความแบบเดิม

ที่มา: นนท์ บุญนิธิประเสริฐ. (2552). การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.



ภาพที่ 4.2 ขั้นตอนการคัดแยกประเภทข้อความแบบที่นำเสนอ

ในการประมวลผลข้อความ SMS เพื่อคัดกรองประเภทของข้อความจำเป็นต้องมีการทำ TN และการตัดคำ เพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผล โดยในการทดสอบที่ผ่านมาใช้วิธีการทำ TN และการตัดคำแบบที่มีในงานวิจัย[1] และงานวิจัยที่กล่าวถึงวิธีการตัดคำภาษาไทย

[1] นนท์ บุญนิธิประเสริฐ. (2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่.** วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

และภาษาไทยปนภาษาอังกฤษที่ผ่านมา[2] ใช้การตัดคำกับเอกสารประเภท หนังสือพิมพ์ วรรณกรรม หรือเว็บไซต์ ที่ข้อความมีความถูกต้องตามหลักภาษาศาสตร์

เมื่อมีคำที่พิมพ์ผิดและคำที่พิมพ์ไม่สมบูรณ์เป็นจำนวนมาก การตัดคำและการหาความหมายของข้อความที่คลาดเคลื่อน ซึ่งส่งผลโดยตรงต่อการคัดกรองประเภทของข้อความ จึงจำเป็นต้องมีการปรับปรุงการทำ TN และการตัดคำ เพื่อให้รองรับกับลักษณะของข้อความ SMS

เพิ่มการตรวจสอบรูปแบบของกลุ่มคำเฉพาะเช่น เลขหมายหรือเบอร์พิเศษ อีเมล์ หมายเลขบัญชี เพื่อนำมาใช้ในการระบุประเภทของคำให้ชัดเจนมากยิ่งขึ้นและเพื่อให้การคัดแยกข้อความ SMS มีความถูกต้องมากขึ้น ด้วยเงื่อนไขของการมีเครื่องหมายหรือคำที่เป็นส่วนประกอบในข้อความ ซึ่งการตรวจสอบกลุ่มคำเฉพาะที่ใช้และหมายเลขพิเศษเหล่านี้ได้ช่วยลดระยะเวลาและเพิ่มประสิทธิภาพในการคัดแบ่งประเภทข้อความให้มีประสิทธิภาพมากขึ้น

4.2.1 ปรับปรุงลำดับการคัดแยกประเภทของข้อความ

ทำการปรับเปลี่ยนเพิ่มเติมจากภาพที่ 4.1 ดังนี้

1) เพิ่มขั้นตอนการตรวจสอบลักษณะของกลุ่มคำเฉพาะที่มีการใช้งานในข้อความประเภทต่างๆ ซึ่งทำให้การระบุประเภทของข้อความมีความถูกต้องและประหยัดเวลามากขึ้น

2) ลดขั้นตอนการตัดคำของการการตรวจสอบคำแรกและคำสุดท้าย เพื่อเพิ่มประสิทธิภาพทางเวลา

3) ลบคำประเภท Stop words ทั้งภาษาอังกฤษ และภาษาไทย

จากภาพที่ 4.2 แสดงลำดับการกรองข้อความ SMS ที่นำเสนอดังนี้

1) รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message

2) ทำกระบวนการ Text Normalization

2.1) ลบสัญลักษณ์พิเศษต่างๆ เช่น ^_^ , +_+

2.2) ลบตัวเลข ที่ไม่มีผลในการคัดกรอง คือ ตัวเลขที่ไม่อยู่ในรูปแบบใดๆ เช่น ไม่เป็นทั้งเบอร์โทรศัพท์ หรือราคาสินค้า ฯลฯ

2.3) ลบ Character ที่เกินกว่า 1 ตัว เช่น สระ หรือ วรรณยุกต์ ต่างๆ

2.4) สลับตำแหน่งสระและวรรณยุกต์ให้ถูกต้อง

3) ตรวจสอบและแยกรูปแบบของคำเพื่อจัดกลุ่มคำ เช่น Email, Phone number, Account number

---

[2] ปโยธร อุราธรรมกุลและผศ. ดร. กานดา รุณนะพงศา. (2549). **การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่** . วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. กรุงเทพฯ: มหาวิทยาลัยขอนแก่น.

4) ลบคำประเภท Stop words ทั้งภาษาอังกฤษ และภาษาไทย (ตัดคำที่พบบ่อยและไม่มีผลต่อการคำนวณผลของข้อความออก)

5) นำคำที่ผ่านขั้นตอนทั้งหมดเข้าสู่กระบวนการกรองข้อความแบบที่นำเสนอ

6) ข้อความที่ไม่สามารถระบุกลุ่มของข้อความได้ให้จัดอยู่ในกลุ่มของข้อความทั่วไป หรือ C2

7) จัดส่งข้อความที่ผ่านการตรวจสอบให้กับผู้รับ และบันทึกข้อความลงในฐานข้อมูลเพื่อใช้ในการเรียนรู้ต่อไป

4.2.2 การตรวจสอบรูปแบบของกลุ่มคำเฉพาะ

การออกแบบวิธีการคัดแยกประเภทข้อความในงานวิจัยนี้ นอกจากการตรวจสอบคำผิดและการปรับปรุงโดยใช้การตัดคำแบบผสมแล้ว ยังมีการตรวจตามลักษณะการใช้งานในรูปแบบต่างๆ เช่น Mail, ราคา, เบอร์โทรศัพท์, หมายเลขบัญชีธนาคาร

จากการเก็บข้อมูล SMS เบื้องต้นนั้นพบว่าลักษณะการใช้งาน SMS นั้นค่อนข้างแยกกันอย่างชัดเจนตามวัตถุประสงค์ของการใช้งาน แต่ทั้งนี้ขึ้นอยู่กับคำที่ถูกใช้ในข้อความต่างๆ เช่น ข้อความที่แสดงข้อมูลเลขหมายพิเศษที่มีอัตราค่าบริการสูงกว่าปรกติ (Premium rate Number) จะมีเครื่องหมาย # หรือ * ประกอบในตัวเลข ทั้งก่อนหน้าชุดตัวเลข ในระหว่างชุดตัวเลข และต่อท้ายชุดตัวเลข และเลขหมายพิเศษที่ขึ้นต้นด้วย 1900 หรือคำที่เป็นรูปแบบของ E-Mail Address ซึ่งถ้าทำการตัดคำตามปกติจะไม่สามารถแยกคำนั้นๆได้ว่าเป็น E-Mail หรือไม่ แต่จะได้เป็นคำที่ไม่มีความหมายใดๆแทน ซึ่งยากต่อการจำแนกประเภทของข้อความดังมีตัวอย่างตามตารางที่ 4.2 เป็นต้น

ตารางที่ 4.2 รูปแบบการตรวจสอบตัวอักษรพิเศษ

กลุ่มคำ	**รูปแบบการตรวจสอบ**
เลขหมายพิเศษ	- ตรวจคำว่า “โทร” หรือ “กด” และมีเครื่องหมาย # หรือ * แล้วตามด้วยชุดตัวเลข - ตรวจข้อความ “1900” แล้วตามด้วยชุดตัวเลขหลายหลัก - ตรวจคำว่า “โทร” หรือ “กด” และมีชุดตัวเลข แล้วตามด้วย เครื่องหมาย # หรือ เครื่องหมาย *
เบอร์โทรศัพท์	ตรวจคำว่า “โทร” หรือมีชุดตัวเลขที่มีรูปแบบเข้าข่าย เช่น 02-1044828

ตารางที่ 4.2 รูปแบบการตรวจสอบตัวอักษรพิเศษ(ต่อ)

กลุ่มคำ	รูปแบบการตรวจสอบ
อีเมล์	ตรวจกลุ่มตัวอักษรและกลุ่มตัวเลขที่มีรูปแบบเป็น E-mail Address เช่น knot_099505411@hotmail.com
หมายเลขบัญชี	ตรวจกลุ่มตัวอักษร X ที่มีชุดตัวเลขผสมอยู่ เช่น 2767XXXX
URL	ตรวจกลุ่มคำที่มีรูปแบบเป็น URL เช่น www.sanook.com
DateTime	ตรวจกลุ่มคำที่มีรูปแบบวันเวลา เช่น DD/MM/YYYY
User & Password	ตรวจคำว่า “username | password ” หรือมีชุดตัวเลข หรือตัวอักษร ไม่เกิน 4-8 หลัก

4.2.2.1 การตรวจสอบรูปแบบเลขหมายพิเศษ

การตรวจสอบรูปแบบของเลขหมายพิเศษ เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการมีเครื่องหมาย # หรือ * ประกอบในตัวเลขทั้งก่อนหน้าชุดตัวเลขในระหว่างชุดตัวเลขและต่อท้ายชุดตัวเลข และเลขหมายพิเศษที่ขึ้นต้นด้วย 1900 ดังนี้

1) ตรวจพบคำว่า “โทร” หรือ “กด” และมีเครื่องหมาย # หรือ * แล้วตามด้วยชุดตัวเลข

2) ตรวจพบคำว่า “โทร” หรือ “กด” และมีชุดตัวเลข แล้วตามด้วย เครื่องหมาย # หรือ เครื่องหมาย *

3) ตรวจพบหมายเลข 1900 แล้วตามด้วยชุดตัวเลข 6 หลัก

4.2.2.2 การตรวจสอบรูปแบบหมายเลขโทรศัพท์

การตรวจสอบรูปแบบของหมายเลขโทรศัพท์ เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุตำแหน่ง หรือมีเครื่องหมาย - หรือ () ประกอบในตัวเลข ทั้งก่อนหน้าชุดตัวเลข ในระหว่างชุดตัวเลข และต่อท้ายชุดตัวเลข ดังนี้

1) ตรวจพบชุดตัวเลขที่มีรูปแบบ ขึ้นต้นด้วย (08), (668), (02) และมีชุดตัวเลขตามจำนวนที่กำหนด

2) ตรวจพบชุดตัวเลขที่มีรูปแบบตามที่กำหนด เช่น (0)(\d{2})(\-*|\s*)(\d{6}) : 042-240444

4.2.2.3 การตรวจสอบรูปแบบ E-Mail Address

การตรวจสอบรูปแบบของ E-Mail Address เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุรูปแบบ E-Mail หรือมีเครื่องหมาย @ และ (.) ประกอบในคำที่ตรวจพบ

4.2.2.4 การตรวจสอบรูปแบบหมายเลขบัญชี

การตรวจสอบรูปแบบของหมายเลขบัญชี เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุรูปแบบที่พบว่ามีการใช้มากในข้อมูล SMS ที่พบในประเทศไทย เช่น มีตัวอักษรขึ้นต้นหรือลงท้ายด้วย XXX และมีชุดตัวเลขประกอบตามจำนวนที่ระบุ

4.2.2.5 การตรวจสอบรูปแบบ URL

การตรวจสอบรูปแบบของ URL เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุรูปแบบที่ขึ้นต้นด้วย http หรือ www และตามด้วยกลุ่มตัวอักษรหรือกลุ่มตัวเลข

4.2.2.6 การตรวจสอบรูปแบบวันเวลา

การตรวจสอบรูปแบบของวันเวลา เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุรูปแบบวันเวลาที่มีการใช้งาน เช่น กลุ่มรูปแบบของวันเดือนปี YYYY/MM/DD

4.2.2.7 การตรวจสอบรูปแบบรหัสผ่าน

การตรวจสอบรูปแบบของรหัสผ่านต่างๆ เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยกฎการระบุรูปแบบรหัสผ่านที่มักประกอบด้วยกลุ่มตัวอักษรหรือกลุ่มตัวเลขรวมกันแบบไม่มีความหมายตามจำนวนที่กำหนด

4.3 การเตรียมข้อมูลทดสอบ

จากการวิเคราะห์ข้อมูล SMS และการสำรวจความคิดเห็นในประเทศไทย สามารถแบ่งการกำหนดชุดข้อมูลออกได้เป็น 2 ชุด คือ ข้อมูลสำหรับฝึกสอน (TD) จำนวน 5,198 ข้อความ และชุดข้อมูลใหม่สำหรับทดสอบ (ND) จำนวน 1,369 ข้อความ ที่ไม่เหมือนกับชุดข้อมูลสำหรับฝึกสอน (TD) ซึ่งประกอบด้วยข้อความดังตารางที่ 4.3

ตารางที่ 4.3 จำนวนข้อมูล TD และ ND

ประเภท SMS	TD	ND
ข้อความปรกติ	1,247	359
ข้อความบริการ	1,369	328
ข้อความแจ้งเตือน	1,067	256
ข้อความโฆษณาประชาสัมพันธ์	1,515	426
	5,198	1,369

ซึ่งข้อความ SMS TD ดังกล่าวจะนำมาใช้ในการฝึกสอน Classify ให้สามารถคัดแยกประเภทของข้อความ SMS ในประเทศไทยได้อย่างถูกต้อง อีกทั้งยังนำ SMS ND มาใช้ทดสอบเพื่อวัดประสิทธิภาพในการคัดกรองข้อความ

กระบวนการฝึกสอนกระทำโดยการป้อนชุดข้อมูล TD เข้าสู่ Classify โดยระบุความหมายของข้อมูลแต่ละตัว เพื่อให้ Classify จดจำรูปแบบของข้อความ แล้วทำการสอน Classify ด้วยชุดข้อมูล TD และทดสอบด้วยชุดข้อมูล ND เพื่อวัดประสิทธิภาพทั้งทางเวลาและความถูกต้อง

4.4 การเขียนโปรแกรม

4.4.1 การคำนวณหาลักษณะแทนข้อความด้วยวิธีการ TFIDF

การคำนวณค่าน้ำหนักของข้อมูล TD จำเป็นต้องทราบค่า TF และค่า DF โดยมี Pseudocode ที่ใช้ในการหาค่าดังต่อไปนี้

```
for($ji=0;$ji<count($sms_data);$ji++) {
        $sms_text = string2word(normalize($sms_data [$ji]),"long");
        $data_DB[0][$ji] = $sms_ary;
        for($i=0;$i<count($sms_ary);$i++) {
                $word = $sms_ary[$i];
                // ignor stop word
                if ((!array_search($word, $sw_ary))&&(strlen($word)!=1)) {
                        if (array_key_exists($word, $db_ary)) {
                                $db_ary[$word][0] = $ij; // index
                                If $db_ary[$word][1] += 1; // TF (count) in priority 0 Doc
                                If $db_ary[$word][2] += 1; // TF (count) in priority 1 Doc
                                If $db_ary[$word][3] += 1; // TF (count) in priority 2 Doc
                                If $db_ary[$word][4] += 1; // TF (count) in priority 3 Doc
                                $ij++;
                        } else {
                                New $db_ary[$word][1] = 1; // TF (count) in priority 0 Doc
                                New $db_ary[$word][2] = 1; // TF (count) in priority 1 Doc
                                New $db_ary[$word][3] = 1; // TF (count) in priority 2 Doc
                                New $db_ary[$word][4] = 1; // TF (count) in priority 3 Doc
                        }
                }
        }
        // add DF
        $unique = array_unique($sms_ary);
        foreach ($unique as $word) {
                if (array_key_exists($word, $db_ary)) {
                        $db_ary[$word][5] += 1; // DF (count)
                }
        }
}
```

ซึ่ง Psudocode ดังกล่าวจะคำนวณข้อมูลจากตัวแปร $priority0_data ซึ่งเป็นชุดข้อมูล TD ในรูปแบบ Array ของ ข้อความ SMS แล้วใช้ Function จัดการข้อมูล TN และ words segmentation เพื่อให้ได้ Normalization ข้อมูลและตัดคำเพื่อเก็บลงในตัวแปร $data_DB

ทำการรวบรวมค่า TF และค่า DF จากคำใน $data_DB โดยจะจัดเก็บเป็น Array 4 ชั้น รูปแบบ SMS_1{$word_1${id, TF_1, TF_2, TF_3, TF_4, DF}, $word_2${ id, TF_1, TF_2, TF_3, TF_4, DF },..., $word_n${ id, TF_1, TF_2, TF_3, TF_4, DF }} และ SMS_1{...}, SMS_2{...},...,SMS_n{...} ในตัวแปร $db_ary และนำไปคำนวณหาค่า Feature Vector จาก Psudocode ดังต่อไปนี้

```
for($i=0;$i<$n_row;$i++) {
                //cut mgs
                 $ncm = strlen($sms_row[$i]);
                 $sms_text = string2word(normalize($sms_row[$i]),"long"); //<<<
                 $sms_ary = explode("|",$sms_text);
                // ได้คำทั้งประโยค
                for($j=0;$j<count($sms_ary);$j++) {
                  $word = $sms_ary[$j];
                        if (array_key_exists($word, $db_ary)) {
                        P(C_i | W) = log(P(C_i) × ∏_{j=1}^{v} P(w_j | C_i))
                        }
                } // for sms ทั้งประโยค
                if (max_key($pop)) {
                   classify(c_1,..,c_n) = arg max p(W = w) ∏_{i=1}^{n} p(C_i = C_i | W = w)
                }
}               // count sms_row
```

4.4.2 Funtion การทำ normalize

การทำ Normalize ข้อความ คือการทำข้อความเพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผลต่อไปได้ ซึ่งเป็นขั้นตอนการลบ สัญลักษณ์พิเศษ เช่น **$ | # | @ | ? | !** หรือตัวเลขที่ไม่ต้องการเพื่อกำจัดข้อมูลส่วนเกินออกและแก้ไขการพิมพ์สระหรือวรรณยุกต์ตัวเดียวกันมากกว่า 1 ครั้ง หรือสระและวรรณยุกต์ที่พิมพ์สลับลำดับทั้งข้อความ เพื่อลดปริมาณคำผิดและเพิ่มประสิทธิภาพในการตัดคำ มีการเขียนโปรแกรมดัง Psudocode ต่อไปนี้

```
function normalize($text) {
                for($i=65;$i<91;$i++) $text = str_replace(chr($i),strtolower(chr($i)),$text);
                $text = Check word in SMS ( array('$word in SMS'          ),
                  " URL ",
                  " MAIL ",
                  " SCN ",
                  " PRN ",
                  " ACC ",
                  " NPN ",
                  " SPN ",
                  " DMY ",
                  " PWD ",
                   $text
                  );
               // del spacial symbol
               // del symbol
               // del symbol
               // del math symbol and number
               // del spacial symbol
               // del thai symbol
               // del space
               // del space
               // rule 2 swap
               // rule 3 same dup
return trim($text);
}
```

4.4.3 Funtion การทำ string2word

การทำ String2word ข้อความ คือการทำข้อความเพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผลต่อไปได้ ซึ่งเป็นขั้นตอนการตัดคำ และในภาษาอังกฤษใช้การเว้นวรรคเพื่อตัดคำ (word segmentation) ในขณะที่ข้อความภาษาไทยไม่สามารถทำได้ เพราะใช้หลักการเขียนคำต่อกันเป็นประโยค ทำให้ต้องใช้ อัลกอริทึมในการตัดแยกคำ อีกทั้งข้อจำกัดของระบบ SMS ทำให้พฤติกรรมการส่งข้อความ มีลักษณะของ คำย่อ, คำทับศัพท์, หรือคำภาษาอังกฤษ ปะปนกันอย่างไม่เป็นระเบียบ จึงจำเป็นต้องปรับปรุงระบบการตัดคำภาษาไทยให้รองรับข้อความดังกล่าวได้ โดยใช้โปรแกรม Swath [20] ของ NECTEC ซึ่งมี license แบบ GNU GPL และพัฒนาวิธีการเชื่อมต่อข้อมูลระว่าง PHP กับ Swath เพิ่มเติมเพื่อให้ PHP เข้าใจและรับรู้การตัดคำของ Swath ได้ถูกต้อง มีการเขียนโปรแกรมดัง Psudocode ต่อไปนี้

```
function string2word($text,$algorithm = "max",$spliter="|") {
        $self_array = explode(chr(92),$_SERVER["PHP_SELF"]);
        $self_file = $self_array[count($self_array)-1];
        $self_path = str_replace($self_file,"",implode("/", $self_array));
        fwrite($handle, 'swath -b "'.$spliter.'" '."-m $algorithm -d data < $self_path$input_tmp > $self_path$output_tmp\n");
        pclose($handle);
        return str_replace($spliter." ".$spliter,$spliter,trim(file_get_contents($self_path.$output_tmp)));
}
```

ข้อความ SMS จะถูกแทนด้วยตัวแปร $SMS_data และผ่านการทำ TN ก่อนการตัดคำ จากนั้นจึงเข้าสู่กระบวนการกรองข้อความด้วย อัลกอริทึมแบบNB เพื่อระบุประเภทของข้อความต่อไป

4.5 การวัดประสิทธิภาพ

จากการวิเคราะห์ข้อมูล SMS และการสำรวจความคิดเห็นในประเทศไทย สามารถแบ่งประเภทของข้อความสั้นออกเป็น 4 ประเภทตามลักษณะการใช้งาน คือ ข้อความแจ้งเตือน, ข้อความด้านการบริการ, ข้อความทั่วไป, ข้อความโฆษณาประชาสัมพันธ์ นั้นทำให้สามารถกำหนดชุดข้อมูลสำหรับทดสอบเป็น 2 ชุด คือ ข้อมูลสำหรับฝึกสอน (TD) จำนวน 5,198 ข้อความ และชุดข้อมูลใหม่สำหรับทดสอบ (ND) จำนวน 1,369 ข้อความ ซึ่งประกอบด้วยข้อความดังตารางที่ 4.4

ตารางที่ 4.4 จำนวนข้อมูล TD และ ND ที่ใช้ทดสอบ

ประเภท SMS	TD	ND
ข้อความปรกติ	1,247	359
ข้อความบริการ	1,369	328
ข้อความแจ้งเตือน	1,067	256
ข้อความโฆษณาประชาสัมพันธ์	1,515	426
	5,198	1,369

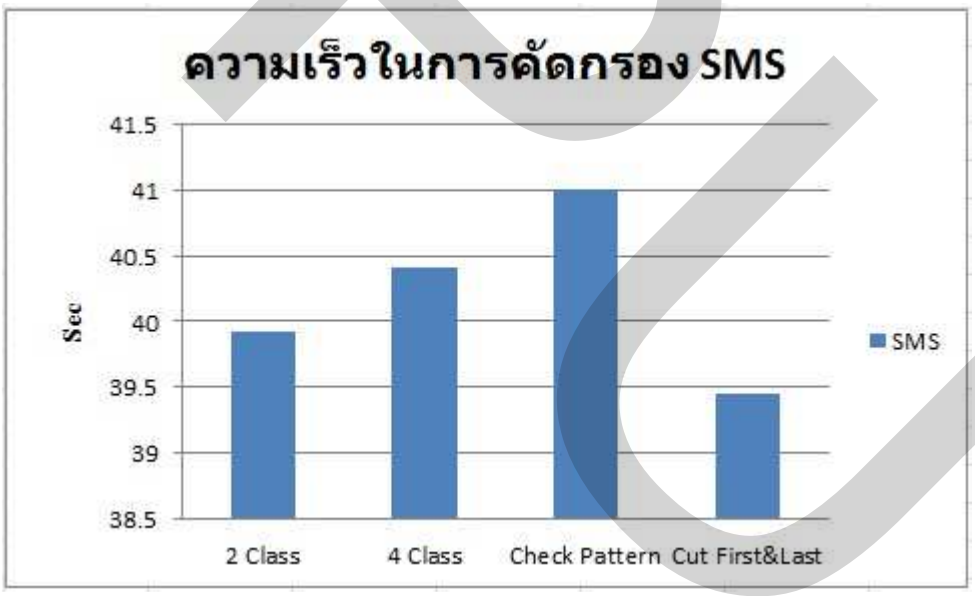
ซึ่งข้อความ SMS ดังกล่าวจะนำมาใช้ในการฝึกสอน Classify ให้สามารถคัดแยกประเภทของข้อความ SMS ในประเทศไทยได้อย่างถูกต้อง อีกทั้งยังนำมาใช้ทดสอบเพื่อวัดประสิทธิภาพในการคัดแยกประเภทข้อความต่างๆได้อีกด้วย

4.6 ผลการคัดแยกข้อความแบบที่นำเสนอในแต่ละขั้นตอน

การวัดประสิทธิภาพความถูกต้องและการวัดประสิทธิทางเวลาโดยวัดตามขั้นตอนการคัดแยกข้อความแบบที่นำเสนอในแต่ละขั้นตอน สามารถแยกได้ดังตารางที่ 4.5 และตารางที่ 4.6

ตารางที่ 4.5 ผลการเปรียบเทียบทางเวลาการคัดแยกแบบที่นำเสนอ

เปรียบเทียบเวลาที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	เวลาเฉลี่ยที่ใช้ในการประมวลผลต่อ 1,369 ข้อความ (sec)
การคัดกรอง 2 Class	39.9296
การคัดกรอง 4 Class	40.4084
ทำ Check Pattern	40.9982
ลดขั้นตอน First&Last	39.4559



ภาพที่ 4.3 ประสิทธิภาพทางเวลาของการคัดแยกข้อความแบบที่นำเสนอ

ตารางที่ 4.6 ผลการเปรียบเทียบความถูกต้องการคัดแยกแบบที่นำเสนอ

เปรียบเทียบความถูกต้องที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	ความถูกต้อง (%)
การคัดกรอง 2 Class	91.59970782
การคัดกรอง 4 Class	96.49379109
ทำ Check Pattern	98.17384953
ลดขั้นตอน First&Last	96.05551497

ภาพที่ 4.4 ประสิทธิภาพทางความถูกต้องของการคัดแยกข้อความแบบที่นำเสนอ

จากภาพที่ 4.3 และภาพที่ 4.4 ผลการทดสอบความถูกต้องและความเร็วในการคัดกรองข้อความด้วยวิธีการแบบที่นำเสนอ พบว่าการคัดแบ่งประเภทของข้อความด้วยวิธีการที่นำเสนอใช้เวลาในการแบ่งประเภทของข้อความน้อยกว่าการกรองข้อความแบบเดิมเล็กน้อย ทั้งนี้เพราะการทำงานตามวิธีการที่เสนอใหม่ลดความซับซ้อนของขั้นตอนลงในส่วนของกระบวนการการตรวจสอบคำแรกและคำสุดท้ายของข้อความ ซึ่งสามารถลดปริมาณงานที่ต้องทำงานได้ระดับหนึ่ง และเมื่อพิจารณาถึงความถูกต้องพบว่าวิธีการที่นำเสนอสามารถคัดแบ่งประเภทข้อความได้ถูกต้องมากขึ้น ทั้งนี้เนื่องจากการเพิ่มการตรวจสอบรูปแบบของคำเฉพาะที่เป็นเลขหมาย หรือข้อความพิเศษประเภทต่างๆ เข้าไป ซึ่งในการคัดกรองแบบเดิมไม่ได้นำมาใช้ในการประมวลผล

4.7 ข้อจำกัดของการคัดแยกประเภทของข้อความแบบที่นำเสนอ

ในการทดสอบการคัดแยกประเภทข้อความด้วยวิธีการตัดคำและวิธีการคัดกรองแบบต่างๆ พบข้อจำกัดบางประการของการคัดแยกประเภทข้อความที่นำเสนอแบบที่นำเสนอ โดยมีตัวอย่างข้อความที่ผ่านการกรองแสดงตามตารางที่ 4.7

ตารางที่ 4.7 ตัวอย่างข้อความที่ผ่านการกรองและผลของการกรอง

SMS	Classify	Correction
มุงดูตะเคียนอยู่ๆล้มต่อหน้าคนเป็นร้อยสำแดงเลขเด็ดโทร 1900888668 9บ.	ข้อความทั่วไป	ข้อความโฆษณาประชาสัมพันธ์
รับส่วนลดสูงสุด 15 % เมื่อใช้บัตรเครดิตธนาคารกรุงเทพซื้อบัตรเข้าชมม	ข้อความโฆษณาประชาสัมพันธ์	ข้อความโฆษณาประชาสัมพันธ์
I Miz U เจอกันชาติหน้านะคิดถึงมั่กๆเลย จุ๊ฟๆๆ B33R นี่ไม่ใช่เบอเค้า	ข้อความทั่วไป	ข้อความทั่วไป
0864459492;วันไหนอยู่ กทม ว่างๆ โทรบอกกันบ้าง ไปกินข้าวกัน นะ	ข้อความทั่วไป	ข้อความทั่วไป
บริการ P (50.00บ./สัปดาห์)ได้ต่ออายุสมาชิกถึง 06/08/10 [022612678]	ข้อความโฆษณาประชาสัมพันธ์	ข้อความด้านการบริการ
ถึงระยะเคลือบสีเต็มระบบครั้งที่2แล้วค่ะ.	ข้อความทั่วไป	ข้อความด้านการบริการ
ดต่อกลับ 024799500 ขออภัยหากชำระแล้ว	ข้อความทั่วไป	ข้อความแจ้งเตือน
กรุณาติดต่อกลับเพื่อหาข้อยุติเกี่ยวกับเงินกู้ที่ค้างชำระ 024799786	ข้อความด้านการบริการ	ข้อความแจ้งเตือน
6,251.05 บาท วันที่ 30/07	ข้อความโฆษณาประชาสัมพันธ์	ข้อความแจ้งเตือน

จากตารางที่ 4.7 แสดงตัวอย่างความผิดพลาดของการคัดกรองข้อความที่นำเสนอ โดยแสดงผลของการคัดกรองในช่อง classify เปรียบเทียบกับชนิดของข้อความที่แท้จริงจากการตัดสินด้วยมนุษย์ในช่อง correction ข้อความที่คัดแยกผิดพลาดส่วนใหญ่ คือ ข้อความที่มีคำไม่ครบทั่วและข้อความที่ไม่ใช่ประเภทที่กำหนดแต่มีจำนวนคำน้อย ซึ่งคำส่วนใหญ่เป็นคำที่เกิดขึ้นแบบอื่นมากกว่าข้อความจากชุดข้อมูลฝึกสอนทั้งหมด หรือ เมื่อมีข้อความที่มีการใช้คำใหม่ๆที่มีจำนวน

น้อยในชุดข้อมูลฝึกสอน จะทำให้การคัดแยกประเภทของข้อความมีความผิดเพี้ยนไปตามชุดข้อมูลฝึกสอน ดังนั้นเราจำเป็นต้องมีการกำหนดกลุ่มของชุดข้อมูลฝึกสอนให้ชัดเจนและสม่ำเสมอตามลักษณะการใช้งานในช่วงเวลานั้นๆ

และจากตารางที่ 4.7 จะเห็นว่าข้อความประเภททั่วไปนั้นมีความผิดเพี้ยนในการคัดแยกน้อย เนื่องจากคำที่ใช้ในกลุ่มนี้มีหลากหลายทำให้ค่าของคำมีค่ากระจัดกระจาย เมื่อเข้าสู่การคัดแยกระบบจึงจัดกลุ่มข้อความประเภทนี้ได้ง่าย และเมื่อระบบพบข้อความที่ไม่เคยพบมาก่อนหรือไม่อยู่ในชุดข้อมูลฝึกสอนระบบจะจัดกลุ่มข้อความนั้นให้อยู่ในประเภทของข้อความทั่วไปก่อน

บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผลการคัดแยกข้อความแบบที่นำเสนอ

ผลการทดสอบแสดงให้เห็นว่าการคัดแบ่งประเภทของข้อความด้วยวิธีการที่นำเสนอใช้เวลาในการแบ่งประเภทของข้อความน้อยกว่าการกรองข้อความแบบเดิม[1] เล็กน้อยกล่าวคือวิธีการที่นำเสนอใช้เวลาประมวลผลเฉลี่ยประมาณ 49.11 msec ในขณะที่วิธีการคัดกรองแบบเดิมนั้นใช้เวลาประมวลผลเฉลี่ย 52.10 msec ทั้งนี้เพราะการทำงานตามวิธีการที่เสนอใหม่ลดความซับซ้อนของขั้นตอนลงในส่วนของกระบวนการการตรวจสอบคำแรกและคำสุดท้ายของข้อความ ซึ่งสามารถลดปริมาณงานที่ต้องทำงานได้ระดับหนึ่ง ดังตารางที่ 5.1 และ ตารางที่ 5.2

ตารางที่ 5.1 ผลการเปรียบเทียบทางเวลาระหว่างการคัดกรองแบบเดิมและการคัดแยกแบบที่นำเสนอ

เปรียบเทียบเวลาที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	เวลาเฉลี่ยที่ใช้ในการประมวลผลต่อข้อความ (millisec)
การคัดกรองแบบเดิม	52.10
แบบที่นำเสนอ	49.11

ตารางที่ 5.2 ผลการเปรียบเทียบความถูกต้องระหว่างการคัดกรองแบบเดิมและการคัดแยกแบบที่นำเสนอ

เปรียบเทียบความถูกต้องที่ใช้ในการประมวลผลข้อความ	
อัลกอริทึม	ความถูกต้อง (%)
การคัดกรองแบบเดิม	84.2951
แบบที่นำเสนอ	97.8816

[1] นนท์ บุญนิธิประเสริฐ. (2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่.** วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

อย่างไรก็ตามเพื่อพิจารณาถึงความถูกต้องพบว่าวิธีการที่นำเสนอสามารถคัดแบ่งประเภทข้อความได้ถูกต้องมากถึง 97.88% ในขณะที่วิธีการคัดกรองแบบเดิม มีความถูกต้องเพียง 84.29% ทั้งนี้เนื่องจากการเพิ่มการตรวจสอบรูปแบบของคำเฉพาะที่เป็นเลขหมาย หรือข้อความพิเศษต่างๆเข้าไป ซึ่งในการคัดกรองแบบเดิมไม่ได้นำมาใช้ในการประมวลผล ซึ่งกลุ่มเลขหมายหรือข้อความพิเศษดังกล่าวพบมากในข้อความประเภท โฆษณาประชาสัมพันธ์, บริการและแจ้งเตือน ทำให้เมื่อประมวลผลพบว่าสามารถระบุประเภทของกลุ่มข้อความดังกล่าวได้ถูกต้องมากขึ้นและมีผลอย่างมากในการคัดแบ่งข้อความประเภทต่างๆ

จากผลการทดสอบแสดงให้เห็นว่า วิธีการคัดแยกประเภทข้อความแบบที่นำเสนอมีความแม่นยำมากกว่าวิธีการคัดกรองแบบเดิมและใช้เวลาในการประมวลผลน้อยกว่า ซึ่งเป็นผลมาจากการตัดขั้นตอนการตรวจสอบคำแรกและคำสุดท้ายสามารถลดเวลาในการทำงานลงได้เล็กน้อยประมาณ 6% เมื่อเทียบกับวิธีการคัดกรองแบบเดิม และการปรับปรุงวิธีตรวจสอบรูปแบบของคำเฉพาะ สามารถช่วยให้กำหนดกลุ่มของข้อความได้ง่ายขึ้น ลดคำผิดในฐานข้อมูลลงได้อย่างมาก และช่วยเพิ่มความถูกต้องให้กับวิธีการที่นำเสนอแบบใหม่ได้มากกว่าถึง 13.59% เมื่อเทียบกับวิธีการคัดกรองแบบเดิม

5.4 ข้อเสนอแนะการคัดแยกข้อความแบบที่นำเสนอ

ในการทดสอบการคัดแยกประเภทข้อความด้วยวิธีการตัดคำและวิธีการคัดกรองแบบต่างๆ พบข้อจำกัดบางประการของการคัดแยกประเภทข้อความแบบที่นำเสนอ เมื่อความถูกต้องในการคัดแยกข้อความขึ้นอยู่กับการกำหนดข้อมูลในการฝึกสอน(TD) เพื่อให้สามารถระบุกลุ่มของข้อความได้อย่างถูกต้อง และเนื่องจากรูปแบบการใช้งานข้อความสั้น (SMS) ในปัจจุบันไม่มีรูปแบบที่แน่นอนมีความเปลี่ยนแปลงอยู่ตลอดเวลาขึ้นอยู่กับปัจจัยทางสังคมของผู้ใช้ เช่น งานเทศกาลต่างๆ, แนวคิดทางการเมือง, ความนิยมในด้านต่างๆ และเมื่อมีข้อความที่มีการใช้คำใหม่ๆที่มีจำนวนน้อยในชุดข้อมูลฝึกสอน จะทำให้การคัดแยกประเภทของข้อความมีความผิดเพี้ยนไปตามชุดข้อมูลฝึกสอน ดังนั้นเราจำเป็นต้องมีการกำหนดกลุ่มของชุดข้อมูลฝึกสอนให้ชัดเจนและสม่ำเสมอตามลักษณะการใช้งานในช่วงเวลานั้นๆ

และจากการคัดแยกข้อความแบบที่นำเสนอพบว่าถ้าเราสามารถลดขั้นตอนการตัดคำของข้อความลงได้จะสามารถลดเวลาที่ใช้ได้มากขึ้นจากเดิมเนื่องจากการตัดคำที่ใช้ในงานที่นำเสนอเป็นการใช้งานโปรแกรมสำเร็จ Swath ของ NECTEC ที่มี license แบบ GNU GPL และพัฒนาเชื่อมต่อระหว่าง PHP กับ Swath ซึ่งต้องส่งข้อความเข้าไปตัดคำทั้งประโยคของข้อความโดยไม่สามารถกำหนดจำนวนคำที่ต้องการตัดได้ก่อนที่โปรแกรมจะทำการตัดคำได้ทั้งหมดและ

สามารถประสิทธิภาพทางความถูกต้องได้โดยเพิ่มการตรวจสอบความถูกต้องของคำในข้อความให้มากขึ้นแต่ก็อาจมีผลกระทบทางด้านเวลาที่เพิ่มมากขึ้น

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

สารานุกรม

สารานุกรมไทยสำหรับเยาวชนฯ. (2544). เล่มที่ 25. กรุงเทพมหานคร: โครงการสารานุกรมไทยสำหรับเยาวชน โดยพระราชประสงค์ในพระบาทสมเด็จพระเจ้าอยู่หัว.

วิทยานิพนธ์

นนท์ บุญนิธิประเสริฐ. (2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่.** วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย, ชุลีรัตน์ จรัสกุลชัย. (2547). **การสรุปใจความสำคัญของเอกสาร.** วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน.

ปโยธร อุราธรรมกุลและผศ. ดร. กานดา รุณนะพงศา. (2549). **การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่** . วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. กรุงเทพฯ: มหาวิทยาลัยขอนแก่น.

บทความ

นนท์ บุญนิธิประเสริฐ และ ดร. ชัยพร เขมะภาตะพันธ์. (22-23 พฤษภาคม 2552). “Short Message Service Filtering for Thai & English Language on Mobile Phone Network” **Proceeding of the 5th National Conference on Computing and Information Technology; NCCIT2009.** p.34-39. กรุงเทพมหานคร.

รายงานวิชาโครงงาน

ชัยณรงค์ รุจิเสถียรทรัพย์และณธัช เลี่ยวชวลิต. (2547). **ตัวกำจัดเมล์ขยะ Spam Mail killer** (รายงานวิชาโครงงาน). กรุงเทพฯ: สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

สารสนเทศจากสื่ออิเล็กทรอนิกส์

ไทยซ่าส์, (2551). สถิติการใช้บริการส่งข้อความอวยพรปีใหม่. สืบค้นเมื่อ 27 พฤษภาคม 2552, จาก **http://phone.thaiza.com/A1_1212_83502_1212_.html.**

ไทยรัฐ, (2552). ข่าวด้านเทคโนโลยี.สืบค้นเมื่อ 27 พฤษภาคม 2552. จาก **http://thairath.co.th/news.php?section=technology03b&content=118050**

ผู้จัดการ, (2549). แก้ปัญหา SPAM SMS บล็อก.สืบค้นเมื่อ 27 พฤษภาคม 2552. จาก **http://www.manager.co.th/Cyberbiz/ViewNews.aspx?NewsID=9490000090269**

ข้อมูลพื้นฐานภาษาไทย, (2552). คำที่พบบ่อย (Stop word th). สืบค้นเมื่อ 2 สิงหาคม 2553. จาก **http://thailang.nectec.or.th/thaichar/word_thai.php?page=1&n_p_page=100**

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, (2537). โปรแกรมตัดคำภาษาไทย. สืบค้นเมื่อ 25 มกราคม 2553. จาก **http://www.hlt.nectec.or.th/products/swath.php**

ภาษาอังกฤษ

ARTICLE

Andrej Bratko, Gordon V.Cormack, Bogdan Filipic, Thomas R. Lynam and Blaz Zupan. (December, 2006). “Spam Filtering Using Statistical Data Compression Models.” **Journal of Machine Learning Research,7**. p. 2673-2678.

PROCEEDINGS

Siddharth Dixit, Sandeep Gupta and Chinya V.Ravishankar. (November 14-16, 2005). "LOHIT: An Online Detection & Control System for Cellular SMS Spam." **Proceeding of the IASTED International Conference on Communication, Network, and Information Security.** p. 48-54. Phoenix, AZ, USA.

Petros Zerfos, XiaoqiaoMeng, Starsky H.Y Wong, VidyutSamanta, Songwu Lu. (October 25-27, 2006) "A Study of the Short Message Service of a Nationwide Cellular Network". **Proceedings of the Internet Measurement Conference: IMC 2006**. p. 263-268. Rio de Janeiro, Brazil.

José María Gómez Hidalgo, Guillermo CajigasBringas and Enrique PuertasSánz. (October 10 - 13, 2006). "Content Based SMS Spam Filtering." **Proceedings of ACM Symposium on Document Engineering**: **ACM2006.** p. 107-114. Amsterdam, Netherlands.

Gordon V. Cormack, José María Gómez Hidalgo and Enrique Puertas Sánz. (November 6-9, 2007). "Spam Filtering for Short Messages." **Proceedings of the ACM sixteenth conference on information and knowledge management: CIKM 2007.** p. 313-319. Lisboa, Portugal.

István Pilászy. (November 18-19, 2005). "Text Categorization and Support Vector Machines." **6th International Symposium of Hungarian Researchers on Computational Intelligence.** Paper list No.064. Budapest, Hungary.

ELECTRONIC SOURCES

University of Glasgow, (2009). Stop Word EN. Retrieved August 2, 2009. From **http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words**

Wikipedia. (2009). Naive Bayes classifier. **Retrieved August 2, 2009. From http://en.wikipedia.org/wiki/Naive_bayes**

Wikipedia. (2009). Bayesian network**. Retrieved August 2, 2009. From http://en.wikipedia.org/wiki/Bayesian_network**

SMS Forum, (1999-2007). SMPP Protocol. Retrieved January 25, 2010. From **http://smsforum.net/**

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวภูรดา นนทวาสี
ประวัติการศึกษา	สำเร็จการศึกษาระดับปริญญาตรี จากคณะวิทยาศาสตร์ สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยราชภัฏเชียงใหม่ ปีการศึกษา 2547
ตำแหน่งและสถานที่ทำงานปัจจุบัน	ดูแลระบบบริการเสริมบนเครือข่ายโทรศัพท์เคลื่อนที่ ภายใต้ชื่อผลิตภัณฑ์ CAT CDMA บริษัท กสท โทรคมนาคม จำกัด (มหาชน)