การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่

นนท์ บุญนิธิประเสริฐ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา วิศวกรรมคอมพิวเตอร์และโทรคมนาคม บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2552

Short Message Service Filtering for Thai & English Language on Mobile Phone Network

NONT BOONNITIPRASERT

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

2009

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยความกรุณาเป็นอย่างยิ่งจาก อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่คอยให้คำแนะนำ ตลอดจนเปิดโลกทรรศน์ ในการค้นคว้าข้อมูลให้แก่ผู้วิจัย ขอขอบพระคุณอาจารย์ ดร.วราพร จิระพันธุ์ทอง กรรมการสอบวิทยานิพนธ์ ซึ่งสละเวลามาเป็นกรรมการสอบวิทยานิพนธ์ และได้ให้ข้อคิดเห็นที่เป็นประโยชน์ต่องานวิจัย และโดยเฉพาะอย่างยิ่ง ผู้วิจัยขอขอบพระคุณ ดร. พีรเดช ณ น่าน กรรมการสอบวิทยานิพนธ์ ผู้ช่วยฝ่ายวิจัยและพัฒนา บริษัท กสท โทรคมนาคม จำกัด (มหาชน) ที่ได้ให้โอกาสการเรียนในครั้งนี้แก่ผู้วิจัย รวมทั้งให้ความรู้อันเป็นประโยชน์แก่ผู้วิจัยมาโดยตลอด นอกจากนี้ผู้วิจัยขอขอบพระคุณคณาจารย์ทุกๆ ท่านในภาควิชา วิศวกรรมศาสตร์ ที่ได้ถ่ายทอดความรู้แก่ผู้วิจัยตลอดระยะเวลาการศึกษา

ผู้วิจัยขอขอบพระคุณ เจ้าหน้าที่ที่เกี่ยวข้องทุกท่าน ในภาควิชา วิศวกรรมศาสตร์ ที่คอยให้ความช่วยเหลือ ตลอดจนแนะนำกระบวนการในการทำงานให้แก่ผู้วิจัยด้วยดีเสมอมา

ผู้วิจัยขอขอบพระคุณ คุณ พีรวัสน์ พงษ์โพนทอง ผู้จัดการส่วนพัฒนาบริการเสริม บริษัท กสท. โทรคมนาคม จำกัด (มหาชน) ที่ให้คำแนะนำ ช่วยเหลือและส่งเสริมในการทำวิจัยครั้งนี้จนสำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอขอบคุณเพื่อนๆ ร่วมรุ่น 1 ทุกคน ที่ช่วยเหลือและให้กำลังใจกันเสมอมาตลอดระยะเวลาการศึกษา

ท้ายสุดนี้ ผู้วิจัยขอกราบขอบพระคุณ คุณแม่ และครอบครัว รวมทั้ง นางสาว ศันสนีย์ วาระนุช ที่คอยเป็นกำลังใจและให้การสนับสนุนผู้วิจัยในทุกๆ ด้านเสมอมาจวบจนสำเร็จการศึกษา

นนท์ บุญนิธิประเสริฐ

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

ประมวลคำศัพท์ และคำย่อ

BTS	Base Trans-receiver Stations
CDMA	Code Division Multiple Access
CDR	Call Detail Records
DB	Database
DMC	Dynamic Markov Compression
HLR	Home Location Register
HTML	Hyper Text Markup Language
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
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
STP	Signaling Transfer Point
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TD	Training Data
TN	Text Normalization
URL	Uniform Resource Locator
VLR	Visitor Location Register
WWW	World Wide Web
XML	Extensible Markup Language

หัวข้อวิทยานิพนธ์	การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่
ผู้เขียน	นนท์ บุญนิธิประเสริฐ
อาจารย์ที่ปรึกษาวิทยานิพนธ์	อาจารย์ ดร. ชัยพร เขมะภาตะพันธ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์และโทรคมนาคม
ปีการศึกษา	2552

บทคัดย่อ

การกรองข้อความสแปมที่เป็นภาษาไทยและภาษาอังกฤษในระบบส่งข้อความสั้นหรือ SMS ยังไม่มีการศึกษาและพัฒนาอย่างจริงจัง ซึ่งปัญหาการมีข้อความสแปมในระบบส่งข้อความสั้นของผู้ให้บริการโทรศัพท์เคลื่อนที่กำลังมีความรุนแรงเพิ่มขึ้น งานวิจัยนี้ได้ทำการศึกษาและเสนอวิธีการกรองข้อความสแปมที่ศูนย์กลางการส่งข้อความหรือ SMS Center ด้วยอัลกอริทึมแบบ Support Vector Machine (SVM) และ Naïve Bayesian (NB) โดยนำเสนอวิธีการ 2 แบบได้แก่ การกรองข้อความที่นำเสนอแบบที่ 1 ซึ่งปรับปรุงจากวิธีการกรองข้อความในภาษาต่างประเทศ ให้สามารถกรองข้อความภาษาไทยได้ และ การกรองข้อความที่นำเสนอแบบที่ 2 ที่ปรับปรุงวิธีการ Text Normalization การตัดคำแบบผสม และเพิ่มการแก้ไขคำผิดจากการกรองข้อความที่นำเสนอแบบที่ 1 แล้วทดสอบประสิทธิภาพการทำงานด้วยการกรองข้อความ SMS ทั้งภาษาไทย ภาษาอังกฤษ และภาษาไทยปนภาษาอังกฤษ

ผลการศึกษาพบว่า การกรองข้อความที่นำเสนอแบบที่ 2 มีประสิทธิภาพทั้งในด้านความถูกต้องและการใช้ระยะเวลาการประมวลผลดีกว่าการกรองข้อความที่นำเสนอแบบที่ 1 นอกจากนี้ ผลการทดสอบด้วยอัลกอริทึมทั้ง 2 แบบ พบว่า อัลกอริทึมแบบ SVM มีความถูกต้องในการกรองข้อความสูงกว่าอัลกอริทึมแบบ NB แต่อัลกอริทึมแบบ NB ใช้เวลาในการประมวลผลน้อยกว่า SVM ประมาณ 2.5 เท่า

Thesis Title	Short Message Service Filtering for Thai & English Language on Mobile Phone Network
Author	Nont Boonnitipasert
Thesis Advisor	Chaiyaporn Khemapatapan, Ph.D.
Department	Computer and Telecommunication Engineering
Academic Year	2009

ABSTRACT

Nowadays, all mobile operators in Thailand have faced with the severe problem from Short Message Service (SMS) spam. However, SMS spam filter system for Thai/English language has not seriously been studied and developed in order to solve this problem. This research aims to study in SMS spam filtering by using Support Vector Machine (SVM) and Naïve Bayesian (NB) algorithm. This research proposes 2 SMS filtering methods, the propose 1 uses current English filtering and upgrading for Thai Language support. The propose 2 modified the first one by upgrading text normalization, applying word segmentation processes, and correcting the missing words for Thai language. Finally, performance evaluations of SMS filtering in the form of Thai language, English language, and combination of Thai and English languages are studied.

The result show that the propose 2 has higher efficient result in filtering and time processing than the other one. Moreover, SMS algorithm has higher accuracy than NB algorithm. On the other hand, NB algorithm spends less time on filtering process than SVM algorithm about 2.5 times.

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

บริการเสริมที่มีผู้ใช้เป็นจำนวนมากที่สุดบริการหนึ่งของระบบโทรศัพท์เคลื่อนที่ในปัจจุบันคือ บริการส่งข้อความสั้นหรือ Short message service (SMS) ในขณะที่บริการดังกล่าวได้รับความนิยมมากขึ้น จึงเริ่มมีการใช้ SMS เป็นสื่อโฆษณาประชาสัมพันธ์ ข้อความประเภทนี้บางกลุ่มจัดเป็นข้อความขยะ (Spam SMS) ซึ่งนอกจากจะรบกวนการใช้งานของผู้ใช้โทรศัพท์แล้ว ยังส่งผลกระทบต่อการทำงานของระบบเครือข่าย SMS รายงานการวิจัยหลายฉบับได้ทำการวิเคราะห์เพื่อปรับปรุงบริการบนระบบโทรศัพท์เคลื่อนที่ต่างๆหลายบริการ แต่กลับมีงานวิจัยที่มุ่งเน้นการแก้ไขปัญหาข้อความสแปมในระบบ SMS น้อยมาก

SMS ถูกคิดค้นขึ้นและพัฒนาใช้งานในยุค 1980 บนระบบเครือข่ายโทรศัพท์เคลื่อนที่ GSM ซึ่งจัดเป็น Data Service บนระบบโทรศัพท์เคลื่อนที่ที่ประสบความสำเร็จสูงที่สุด และได้รับความนิยมเพิ่มขึ้นอย่างต่อเนื่อง มีอัตราส่วนแบ่งทางการตลาดเฉพาะยุโรปตะวันตกร้อยละ 80 เมื่อเทียบกับ Data Service แบบอื่นๆ โดยมีการส่งข้อความ SMS จำนวน 200,000 ล้านครั้งต่อเดือน[1] บริษัท China Unicom ในประเทศจีนรายงานว่า มีการส่ง SMS ในปี 2005 จำนวน 304,140 ล้านครั้งต่อเดือน[2] ดีแทค ซึ่งเป็นผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทย รายงานสถิติการใช้ SMS อวยพรปีใหม่ระหว่างวันที่ 31 ธันวาคม 2550 ถึงวันที่ 1 มกราคม 2551 จำนวน 38 ล้านข้อความ เพิ่มขึ้นจากช่วงเดียวกันของปีที่ผ่านมาร้อยละ 32 โดยมีการส่งสูงสุด 4 ล้านข้อความต่อชั่วโมง ในช่วงรอยต่อระหว่างวันที่ 31 ธ.ค. 2550 กับ วันที่ 1 ม.ค. 2551[3] ข้อมูลจากผู้จัดการออนไลน์ ฉบับวันที่ 2 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.

---

[1] S. Dixit, S. Gupta, and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam.** P. 1.

[2] Petros Zerfos, Xiaoqiao Meng, Starsky H.Y. Wong, Vidyut Samanta and Songwu Lu. (2006). **A Study of the Short Message Service of a Nationwide Cellular Network.** P. 1.

[3] สำนักข่าว ผู้จัดการออนไลน์ (2551). ดีแทค เผยสถิติส่งข้อความส่งความสุขปีใหม่ (SMS) 38 ล้านข้อความ (MMS) 6.8 แสนข้อความ.

ถึงเวลา 01.00 น. และบริษัทแอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) มีอัตราการส่งข้อความ SMS เพิ่มขึ้นจากช่วงเดียวกันของปี 2551 ร้อยละ 25[1]

เนื่องจากข้อความ SMS สามารถเข้าถึงผู้ใช้งานได้รวดเร็ว การนำบริการดังกล่าวมาเป็นเครื่องมือทางการประชาสัมพันธ์จึงมีประสิทธิภาพและถูกนำมาใช้อย่างกว้างขวาง ข้อความประชาสัมพันธ์ที่จัดว่าเป็นข้อความสแปมจึงเกิดมากขึ้นตามไปด้วย ประเทศเกาหลีใต้และญี่ปุ่นมีจำนวนข้อความสแปมในระบบ SMS มากกว่าครึ่งของการใช้งาน[2] ทำให้การกรองข้อความสแปมออกก่อนมีการส่งถึงผู้รับมีความจำเป็นอย่างยิ่ง เพราะนอกจากจะลดปัญหาความรำคาญให้กับผู้ใช้งานแล้ว ยังช่วยเพิ่มประสิทธิภาพของระบบ SMS ให้แก่ผู้ให้บริการอีกด้วย

ความหมายของคำว่าข้อความสแปมคือ การส่งข้อความไปยังผู้รับ โดยผู้รับมิได้ร้องขอ ซึ่งก่อความรำคาญแก่ผู้รับ และมีจุดประสงค์บางประการของผู้ส่งต่อผู้รับ ช่องทางของการส่งข้อความสแปมที่พบในชีวิตประจำวันได้แก่ E-Mail, ข้อความแสดงความคิดเห็นใน Forum บน Web Site ต่างๆ และ SMS ซึ่งมีองค์ประกอบดังนี้

- ไม่มีการร้องขอ : ผู้รับมิได้ร้องขอข้อมูล และไม่ทราบข้อมูลของผู้ส่ง
- ส่งครั้งละหลายข้อความ : ผู้ส่งทำการส่งข้อความจำนวนมากติดต่อกัน
- จุดประสงค์ : เช่นการชักจูงผู้รับให้ทำกิจกรรมบางอย่าง เพื่อให้ผู้ส่งได้รับประโยชน์

---

[1] สำนักข่าว ผู้จัดการออนไลน์ (2551). ดีแทค เผยสถิติส่งข้อความส่งความสุขปีใหม่ (SMS) 38 ล้านข้อความ (MMS) 6.8 แสนข้อความ.

[2] S. Dixit, S. Gupta, and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam.** P. 1-2.

ตารางที่ 1.1 ตัวอย่างข้อความ SMS แบบปรกติและแบบข้อความสแปม

SPAM	NORMAL
Get lots of xxx pictures in your e-mail!	Contact Me Privately (jackparkinson12@live.com)
Get free xxx account passwords! Press *4555	Hello, my friend!
ดาวน์โหลดริงโทน!! พิมพ์ ok ส่งที่ *123456	โหลดเอกสาร word ได้ที่ www.123.com/zip.zip
มันส์กับเกมEuro2008แล้วลุ้นPSP!	จองหนังให้แล้วนะ รีบมาด้วย

เมื่อข้อความสแปมในระบบ SMS เริ่มก่อปัญหามากขึ้น ดังเช่นกรณีการส่ง SMS ของบริษัท ทรู คอร์ปอเรชั่น ในการส่งเสริมการขาย SIM Card สำหรับโหวตรายการวาไรตี้โชว์ ไปยังเครื่องโทรศัพท์ลูกข่ายของระบบดีแทคจำนวนมากติดต่อกัน จนทำให้ต้องมีการปิดรับ SMS ระหว่าง 2 เครือข่ายเป็นเวลา 16 ชั่วโมง[1] และกรณีการส่ง SMS อวยพรปีใหม่ของผู้ใช้บริการโทรศัพท์เคลื่อนที่ในประเทศอินเดียที่มากกว่า 4 แสนข้อความ ต่อนาที จนทำให้ระบบหยุดทำงาน[2] จึงจำเป็นต้องมีการปรับปรุงระบบส่งข้อความสั้นเพื่อลดปัญหาดังกล่าว

การแก้ปัญหาข้อความสแปมสามารถทำได้หลายรูปแบบเช่น การใช้ Application กรองข้อความสแปมติดตั้งลงบนเครื่องโทรศัพท์เคลื่อนที่ปลายทาง ซึ่งช่วยลดปริมาณข้อความ SMS ที่ผู้รับไม่ต้องการรับลงได้ แต่ไม่สามารถลดปริมาณข้อความ SMS แบบสแปมในระบบส่งข้อความสั้นของผู้ให้บริการได้ ตัวอย่าง Application ที่ใช้กรองข้อความอาทิเช่น SMS Spam Manager[3]

นอกจากนี้ ผู้ให้บริการโทรศัพท์เคลื่อนที่บางราย ได้มีการเพิ่มมาตรการระบุตัวตนผู้ส่งด้วยการเพิ่มการทำ Authentication ก่อนการอนุญาตให้ส่งข้อความ SMS ระหว่างผู้ให้บริการโทรศัพท์เคลื่อนที่และผู้ให้บริการ Content ซึ่งมาตรการนี้มีวัตถุประสงค์เพื่อให้สามารถระบุผู้ส่ง

---

[1] ผู้จัดการออนไลน์ และ ASTV. (2549). ดีแทคชนทรูมูฟแก้ปัญหา SPAM SMS บล็อกไม่รับข้อความขายชิมกว่า 16 ชม.

[2] Petros Zerfos, Xiaoqiao Meng, Starsky H.Y. Wong, Vidyut Samanta and Songwu Lu. (2006). **A Study of the Short Message Service of a Nationwide Cellular Network.** P. 1.

[3] WebGate JSC (2007). SMS Spam Manager.

ข้อความและใช้เป็นข้อมูลอ้างอิงทางกฎหมาย โดยไม่มีการกรองข้อความสแปมออกจากระบบส่งข้อความ

ปัจจุบันสำนักงานคุ้มครองผู้บริโภค หรือ สคบ เตรียมออกระบบ Anti Spam SMS โดยเป็นกฎหมายที่จะมีผลบังคับใช้ในปี พ.ศ. 2553 และใช้ระบบ Call Center ในการรับเรื่องร้องเรียนข้อความ SMS ที่ไม่ต้องการได้รับ ซึ่งระบบดังกล่าวจะใช้การปิดกันข้อความจากการร้องขอของผู้ใช้บริการ ทำให้มีความล่าช้าและสิ้นเปลืองทรัพยากรของระบบได้แก่ วงจรเชื่อมต่อสำหรับรับการร้องเรียน และบุคลากรในการกำหนดลักษณะข้อความที่ไม่ต้องการส่งยังผู้ใช้บริการ[1]

แนวทางการแก้ปัญหาหนึ่งที่ถูกนำมาพิจารณาในระดับสากลคือ การกรองข้อความสแปมออกจากระบบที่ SMSC เพื่อลดความคับคั่งของข้อมูลด้วยการใช้การกรองข้อความสแปม (Filtering) โดยให้ Filter ทำการกรองข้อความที่ SMSC ก่อนการส่ง ในปัจจุบัน Filter ที่ใช้งานเพื่อกรองข้อความสแปม เช่น DMC, SVM, LR ล้วนแต่มีการพัฒนาต่อเนื่องจาก E-Mail Spam Filter ทั้งสิ้น[2] โดยได้ปรับปรุงหลักการกรองให้เหมาะสมกับ SMS ที่มีข้อมูลต่อหนึ่งข้อความที่สั้นกว่า E-Mail

การส่งข้อความในระบบ SMS มีรูปแบบการส่ง 2 วิธีได้แก่

1) การส่งข้อความแบบ Signaling (SS7)

2) การส่งข้อความผ่าน TCP Protocol

Spammer จะใช้ Robot Software เข้ามาช่วยให้ทำการส่งข้อความสแปมที่มีลักษณะส่งครั้งละหลายข้อความและหลายปลายทาง Software ประเภทนี้สามารถทำงานร่วมกับ TCP Protocol ได้สะดวกรวดเร็วกว่าการทำงานผ่าน Signaling ที่ต้องอาศัย Software สั่งการทำงานไปยังเครื่องโทรศัพท์เคลื่อนที่ผ่าน AT Command นอกจากนี้ระบบโทรศัพท์เคลื่อนที่ในยุคต่อไป จะทำงานบนพื้นฐาน TCP/IP เพียงอย่างเดียว ทำให้การกรองข้อความบน TCP Protocol ที่มีความซับซ้อนน้อยกว่า และสามารถครอบคลุมการใช้งานที่จะเกิดขึ้นในอนาคต

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 กำหนดความหมายของข้อความสแปมในระบบส่งข้อความสั้น SMS

---

[1] ข่าวทั่วไป MCOT. (2552). ผู้ให้บริการโทรมือถือขานรับ สคบ. เปิดให้ประชาชนโทรขอยกเลิก SMS ขายของ.

[2] S. Dixit, S. Gupta and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam.** P. 1.

1.2.2 ปรับปรุงการเตรียมข้อมูล Text ก่อนการประมวลผลเพื่อเพิ่มประสิทธิภาพในการกรองข้อความ

1.2.3 พัฒนาวิธีการกรองข้อความสแปมภาษาไทย สำหรับบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่ เพื่อลดปริมาณข้อความสแปมในระบบส่งข้อความสั้นให้สามารถรองรับการทำงานได้มากขึ้น

1.2.4 จำลองสถานการณ์รับส่งข้อความสั้นผ่านระบบกรองข้อความสแปมภาษาไทย เพื่อวัดประสิทธิภาพด้านความถูกต้องและความเร็วในการกรองข้อความ และทดสอบเปรียบเทียบวิธีการตัดคำแบบต่างๆที่มีความเหมาะสมในการนำมาพัฒนาระบบกรองข้อความต่อไป

1.3 สมมติฐานของการวิจัย

1.3.1 ศึกษาความหมายของข้อความสแปมในประเทศไทย เพื่อกำหนดขอบเขตการตัดสินข้อความสั้น SMS ให้มีความชัดเจน

1.3.2 ออกแบบและพัฒนาวิธีการการกรองข้อความสั้น SMS ให้สามารถใช้งานกับข้อความ SMS ภาษาไทย ภาษาอังกฤษ และภาษาไทยปนภาษาอังกฤษได้

1.3.3 ปรับปรุงประสิทธิภาพการทำ TN ก่อนการกรองข้อความ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถกรองข้อความสแปมภาษาไทยออกจากบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่ ด้วยขั้นตอนวิธีการกรองข้อความสแปม แบบใหม่ เพื่อเพิ่มประสิทธิภาพการทำงานของบริการส่งข้อความสั้น SMS ให้สามารถรองรับการส่งข้อความในปริมาณที่เพิ่มขึ้น ลดอัตราเสี่ยงที่ระบบส่งข้อความสั้นจะเกิดการ Overload จนไม่สามารถทำงานต่อไปได้ ลดปริมาณการรับข้อความสั้นของผู้ใช้บริการโทรศัพท์เคลื่อนที่และเพิ่มอัตราการประหยัดพลังงานของอุปกรณ์ที่ใช้รับข้อความ ด้วยการคัดแยกข้อความสแปมออกจากระบบส่งข้อความก่อนเกิดขั้นตอนการส่ง ผลการศึกษายังสามารถใช้เป็นพื้นฐานเพื่อพัฒนาระบบกรองข้อความสแปมที่จะนำไปใช้งานเชิงพาณิชย์ สำหรับผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยต่อไป

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 องค์ประกอบของ SMS

2.1.1 Message (SMS)

โครงสร้างของข้อความในระบบ SMS ประกอบด้วยตัวอักษรภาษาอังกฤษแบบ ASCII ขนาด 7 bit จำนวน 160 ตัวต่อข้อความ โดยอักษรแต่ละตัวมีความแตกต่างกัน 126 แบบ (ซึ่งมีทั้งตัวอักษรภาษาอังกฤษตัวพิมพ์เล็ก, ตัวพิมพ์ใหญ่, ตัวเลข, สัญลักษณ์พิเศษต่างๆ) และจำนวน 70 ตัวอักษรต่อข้อความในแบบ Unicode ขนาด 16 bit (เช่น ภาษาไทย เป็นต้น)

2.1.2 Sender

แบ่งออกเป็น 2 ประเภท หลักๆ ได้แก่

- ผู้ส่ง ที่ส่งข้อความจากโทรศัพท์เคลื่อนที่ หรือ Mobile Device อื่นๆ

ปัจจุบัน การส่งข้อความจากอุปกรณ์เหล่านี้จะไม่เป็นข้อความสแปมเนื่องจากมีข้อจำกัดหลายอย่าง เช่น ความเร็วการประมวล, หน่วยความจำ และแหล่งพลังงานของอุปกรณ์ อย่างไรก็ตามระบบโทรศัพท์ในยุคต่อไปที่กำลังจะมาถึง อาจทำให้เกิดข้อความสแปมจากอุปกรณ์เหล่านี้ได้

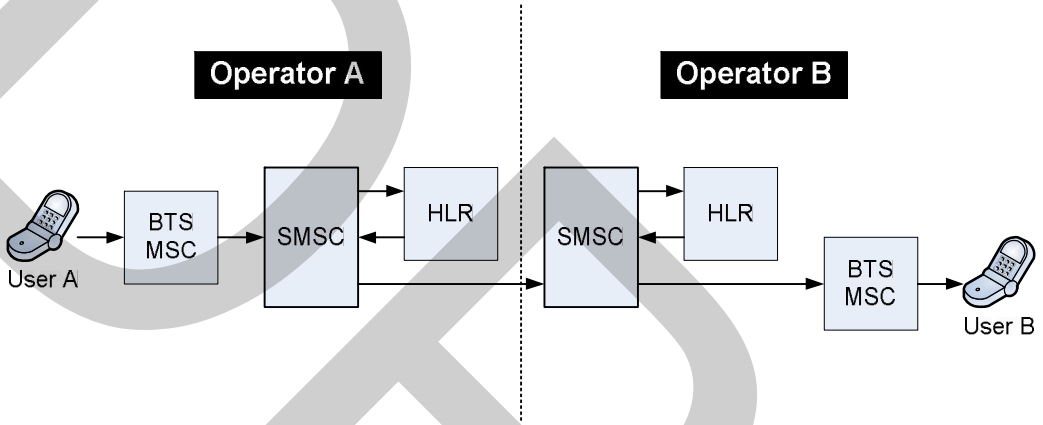
- ผู้ส่ง ที่ส่งข้อความผ่านเครือข่ายการสื่อสารข้อมูลสาธารณะ (Internet)

ผู้ให้บริการโทรศัพท์เคลื่อนที่ (Operator) อาจเปิดให้บริการส่งข้อความผ่าน TCP/IP Protocol โดยการให้ Download Program สำหรับส่ง SMS เช่น การส่ง SMS จากโรงพยาบาล เพื่อเตือนให้คนไข้พบแพทย์ตามกำหนดเวลา หรือผู้ให้บริการข่าวสารผ่าน SMS (Content Provider) เป็นต้น ผู้ส่งกลุ่มนี้มีความยืดหยุ่นในการส่งข้อความสูง ส่งได้ครั้งละหลายข้อความ และตลอดเวลา ทำให้ผู้ส่งกลุ่มนี้ มีโอกาสส่งข้อความสแปมได้ (Spammer)

นอกจากการส่ง SMS ด้วย เครื่องโทรศัพท์เคลื่อนที่และ Applications ที่ผู้ให้บริการโทรศัพท์เคลื่อนที่จัดเตรียมไว้ให้แล้ว การส่ง SMS จาก Applications อื่นๆ ก็สามารถทำได้ เช่น จาก E-Mail Server หรือ ระบบตอบรับด้วยเสียงอัตโนมัติ (IVR) โดยแต่ละระบบจะเชื่อมต่อกับ SMSC ด้วย SMPP ผ่าน Adapter แต่ผู้ส่งประเภทนี้จะควบคุมได้ง่ายกว่า และมีโอกาสเกิดข้อความสแปมน้อยมาก

2.1.3 SMS Network

โครงสร้างพื้นฐานของระบบ SMS จะมีการเชื่อมต่อกันหลายส่วน เช่น การเชื่อมต่อกับ Wireless Network เพื่อส่ง SMS ไปยังโทรศัพท์เคลื่อนที่ เชื่อมต่อกับ Internet รองรับบริการลักษณะ Web SMS หรือเชื่อมต่อกับ SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่รายอื่น เพื่อให้การส่ง SMS ระหว่างผู้ให้บริการโทรศัพท์เคลื่อนที่สามารถทำได้ 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 Protocol

มาตรฐานการสื่อสารของระบบ SMS บน TCP/IP ระหว่างผู้ส่งและผู้รับ แบ่งออกเป็น

- การสื่อสารภายในเครือข่ายและการสื่อสารระหว่างเครือข่าย โดยใช้ Short Message Peer-to-Peer protocol (SMPP)[1]

- การสื่อสารระหว่างเครือข่ายกับผู้ส่งทั่วไป มีหลายแบบขึ้นอยู่กับระบบของเครือข่ายและผู้ให้บริการ เช่น Socket ผ่าน Application ที่ผู้ให้บริการเปิดให้ Download หรือการเชื่อมต่อด้วย SOAP XML Web Service เป็นต้น

[1] Smsforum.net. (1999). **Short Message Peer to Peer Protocol Specification v3.4.** P. 36 - 163

2.1.5 Receiver

ผู้รับ คือเครื่องโทรศัพท์เคลื่อนที่ปลายทาง เมื่อผู้รับเปิดเครื่องรับ โทรศัพท์จะลงทะเบียนเพื่อแจ้งที่อยู่ของตนที่ HLR ว่ากำลังเชื่อมต่อกับ MSC ใด บริการต่างๆที่เกิดขึ้นก่อนที่ผู้รับเปิดเครื่อง เช่น SMS หรือ MMS จะทราบตำแหน่งของเครื่องผู้รับ และดำเนินการส่งให้กับผู้รับอย่างถูกต้อง

2.2 การส่งข้อมูล SMS

SMPP[1] เป็น Protocol มาตรฐานในการส่งข้อมูล SMS MMS หรือ PUSH Message ภายในโครงข่ายโทรศัพท์เคลื่อนที่ โดยประกอบด้วย 2 ส่วนสำคัญคือ ส่วนที่ 1 PDU Header ที่ใช้ในการระบุ ความยาว ชนิด และลำดับของข้อความ ส่วนที่ 2 PDU Body ใช้บรรจุข้อมูลที่ต้องการส่งผ่าน เช่น ข้อความภายใน SMS หรือ Link สำหรับ PUSH Message เป็นต้น

ตารางที่ 2.1 โครงสร้างของ SMPP PDU

PDU Header (mandatory)				**PDU Body (Optional)**
command length	*command id*	*command status*	*sequence number*	*PDU Body*
4 octets	Length = (Command Length value - 4) octets			

ที่มา: www.smsfourm.net

ตัวอย่าง Data-stream ของ SMPP PDU แบบ Hex format

```
00 00 00 2F 00 00 00 02 00 00 00 00 00 00 00 01 53 4D 50 50 33 54 45 53 54 00
73 65 63 72 65 74 30 38 00 53 55 42 4D 49 54 31 00 00 01 01 00
```

[1] Smsforum.net. (1999). **Short Message Peer to Peer Protocol Specification v3.4.** P. 36 - 163

โดยมีส่วน Header ประกอบไปด้วย

00 00 00 2F Command Length 0x0000002F (ความยาวของ PDU)
00 00 00 02 Command ID 0x00000002 (คำสั่งสำหรับ bind transmitter)
00 00 00 00 Command Status 0x00000000 (สถานะของข้อความ)
00 00 00 01 Sequence Number 0x00000001 (ลำดับของ PDU Message)

2.3 ระบบกรองข้อความ

งานวิจัย Spam filtering for short messages[1] การเปรียบเทียบระบบการกรองข้อความสแปมหรือ SMS Spam Filtering ที่พัฒนาต่อจากระบบ E-Mail Spam Filter ลักษณะเด่นของ Filter แต่ละตัวมีดังต่อไปนี้

- Bogofilter

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

- OSBF-Lua

เป็น Open source spam filter ที่พัฒนาขึ้นจากภาษา C ชนิดหนึ่ง (ภาษา C แบบ Lua) ที่ใช้เทคนิค orthogonal sparse bi-grams[2] ซึ่งเป็นวิธีการตรวจจับของ E-Mail โดยเน้นที่การทำงานแบบ pairs of collocated words

- DMC (Dynamic Markov Compression)

เน้นการประมวลผลกับข้อมูลที่ถูกบีบอัด โดยใช้หลักการพิจารณาข้อความเป็น String เพื่อตรวจสอบและทำนายว่าข้อความนั้นจัดเป็นข้อความสแปมหรือไม่

---

[1] Gordon V. Cormack, Jose Maria Gomez Hidalgo and Enrique Puertas Sanz. (2007). **Spam filtering for short messages.** P. 1-8.

[2] Fidelis Assis. (2006). **OSBF-Lua, Text classification module for the Lua Programming Language and a production class anti-spam in Lua using the module.** P. 1.

- LR (Logistic Regression)

เป็น Open source spam filter ที่ใช้หลักการคำนวณทางตรรกศาสตร์ เพื่อหาค่าสัมประสิทธิ์ของฟังก์ชั่นเชิงเส้น แล้วพิจารณาความเป็นไปได้ที่จะเป็นข้อความสแปมคล้ายกับ SVM แต่มีระดับการคำนวณต่ำกว่า

- SVM (Support Vector Machine)

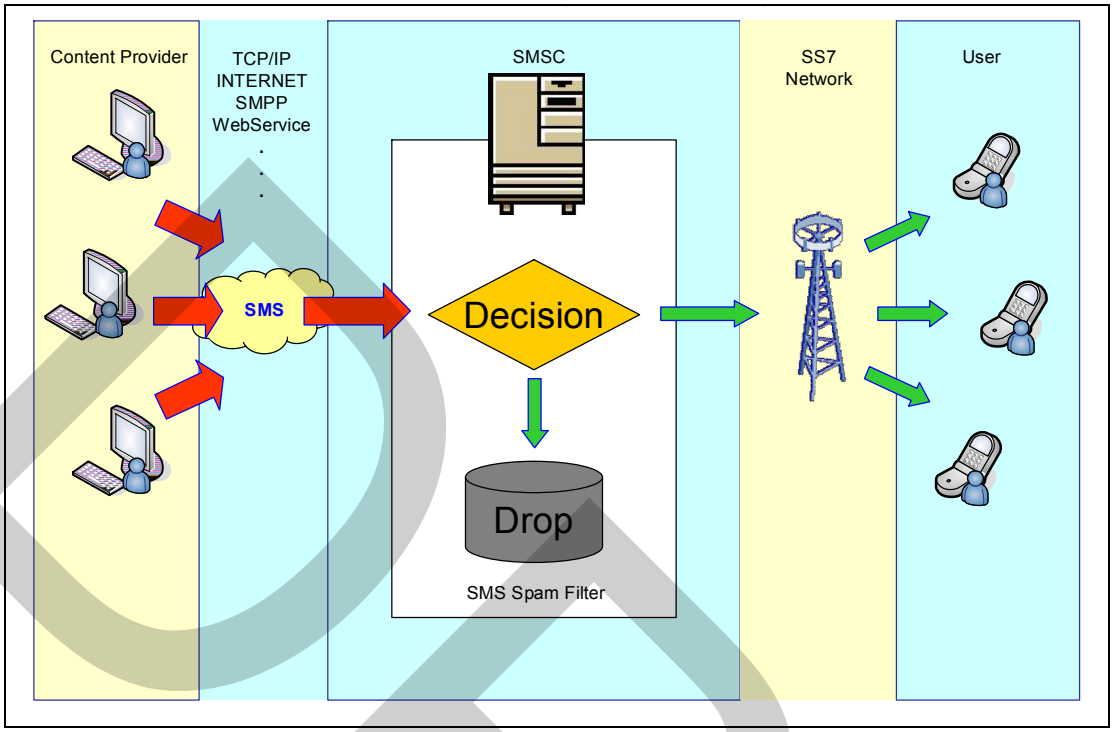
มีชื่อเสียงในการค้นหาข้อความสแปมด้วยหลักการหาค่าสัมประสิทธิ์ของฟังก์ชั่นเชิงเส้น เพื่อแยกข้อความสแปมด้วยระยะห่างบนกราฟ และมีคุณลักษณะในการตรวจสอบข้อมูล Alphanumeric characters อีกด้วย ซึ่งให้ผลการกรองได้ดีที่สุดในการทดสอบ

- LOHIT Algorithm

งานวิจัยที่ใช้อัลกอริทึมการกรองข้อความแบบ LOHIT[1] กล่าวถึงการพัฒนา LOHIT Filter สำหรับกรองข้อความสแปมในระบบ SMS ที่ต่างจาก E-Mail Spam Filter โดยใช้ Probability เพื่อระบุความน่าจะเป็นของข้อความสแปมและทำการจำลองการส่งข้อความ SMS เพื่อทดสอบประสิทธิภาพ โดยแสดงผลออกมาในรูป 3D subspace ผลการจำลองการส่ง SMS สามารถให้ประสิทธิภาพดีกว่า Filter ที่พัฒนาจาก E-Mail Spam Filter อื่นๆ

โครงสร้างของระบบส่งข้อความ SMS ที่มีการกรองข้อความ SMS ซึ่งมีใช้งานในปัจจุบัน โดยหลักการทำงานคือ เมื่อข้อความ SMS จากผู้ให้บริการข่าวสารหรือ Content Provider ถูกส่งด้วย TCP Interface ข้อความ SMS จะถูกแปลงให้อยู่ในรูป SMPP Message แล้วทำการส่งไปยัง SMSC เพื่อทำการส่งให้แก่สมาชิกผู้ร้องขอข่าวสารนั้นๆ การตรวจสอบข้อความจะทำได้โดยระบบกรองข้อความที่ถูกติดตั้งใน SMSC ซึ่งจะทำหน้าที่คัดแยกข้อความสแปมออกก่อนแล้วทำการส่งข้อความปรกติไปยังผู้รับ ดังรายละเอียดตามภาพที่ 2.2

[1] S. Dixit, S. Gupta, and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam.** P. 1.



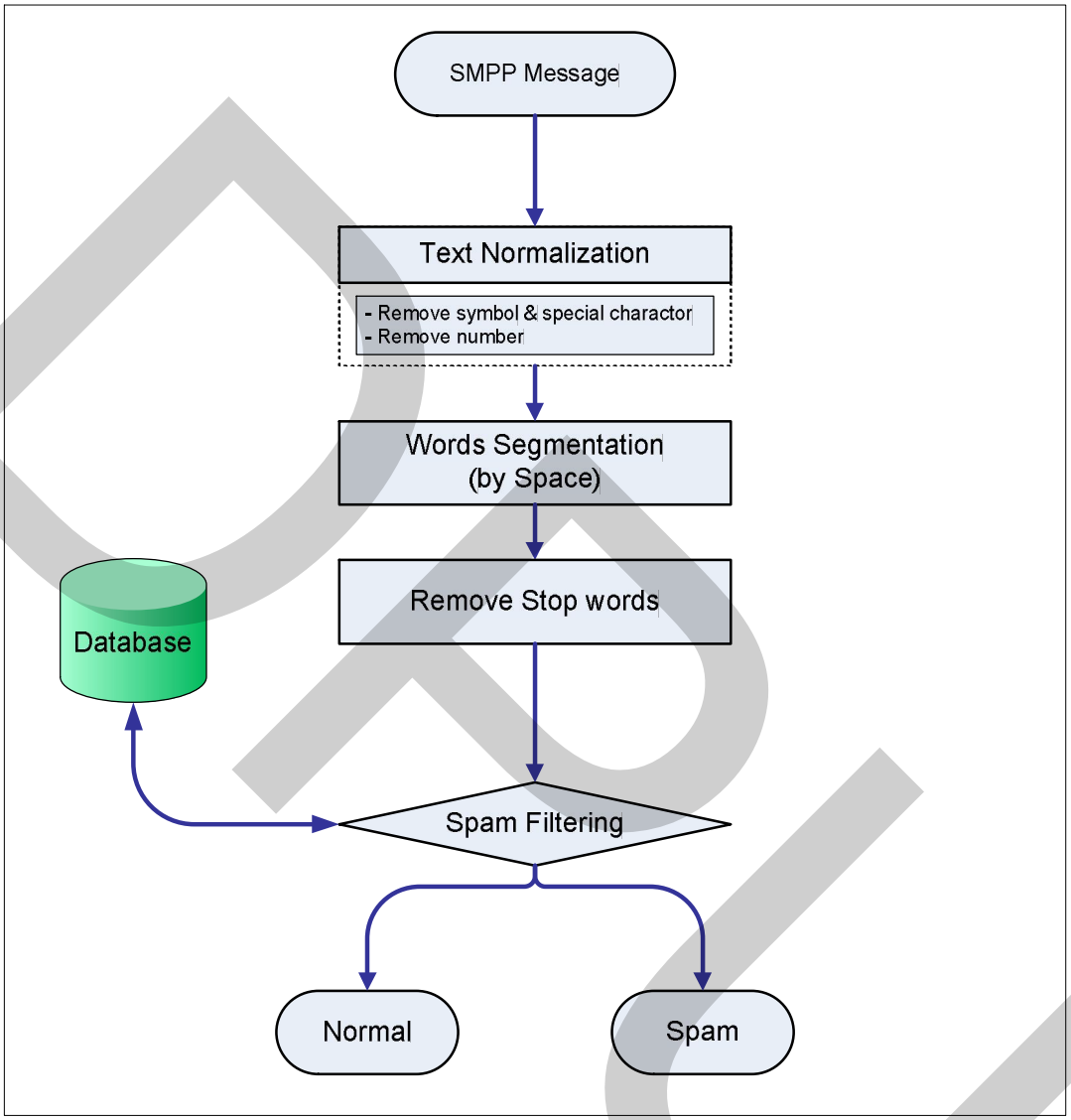
ภาพที่ 2.2 โครงสร้างระบบกรองข้อความ SMS

ที่มา: LOHIT: An Online Detection & Control System for Cellular SMS Spam.

การศึกษางานวิจัยที่เกี่ยวข้องกับการกรองข้อความ SMS จากบทความทางวิชาการและงานวิจัยต่างประเทศ[1] สามารถอธิบายการกรองข้อความของ Filter ได้ คือ เมื่อ Filter รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message ให้อยู่ในรูปของ text แล้ว จะทำกระบวนการ TN เพื่อลบ Character ที่ไม่สามารถตัดเป็นคำได้ออกไป เช่น @ # ! รวมถึงตัวเลข เป็นต้น เมื่อได้ข้อความที่มีความพร้อมแล้ว จะทำตัดคำภาษาอังกฤษด้วยการตรวจสอบการเว้นวรรค จากนั้นจะลบคำที่จัดอยู่ในประเภท Stop words ออกไป และทำการ Mapping คำเข้ากับค่า TFIDF หรือค่า Spam Rate ตามอัลกอริทึมการกรองเพื่อสรุปผลของข้อความว่าจัดเป็นข้อความปรกติ หรือข้อความสแปม เมื่อได้ผลการกรองเป็นข้อความปรกติ ระบบจะทำการส่งข้อความไปยังผู้รับ หรือหากไม่ผ่านการกรองข้อความ ระบบจะละทิ้งข้อความนั้นโดยไม่ส่ง ซึ่งมีรายละเอียดการทำงานตามตามภาพที่ 2.3

---

[1] Petros Zerfos, Xiaoqiao Meng, Starsky H.Y. Wong, Vidyut Samanta and Songwu Lu. (2006). **A Study of the Short Message Service of a Nationwide Cellular Network.** P. 1.



ภาพที่ 2.3 ขั้นตอนการทำงานของ SMS Spam filter

ที่มา: LOHIT: An Online Detection & Control System for Cellular SMS Spam.

ซึ่งวิธีการกรองข้อความดังกล่าวไม่สามารถนำมาใช้กรองข้อความภาษาไทย และภาษาไทยปนภาษาอังกฤษได้ เพราะในขั้นตอน Text Normalize และ ขั้นตอน Word Segmentation ไม่สามารถรองรับกับข้อความภาษาไทย และภาษาไทยปนภาษาอังกฤษ

2.4 อัลกอริทึมสำหรับกรองข้อความ

2.4.1 Rule Base

เป็นวิธีการตรวจจับข้อความด้วยกฎและเงื่อนไขการใช้ Keywords Matching ไม่มีความซับซ้อนในการคำนวณทางคณิตศาสตร์ สามารถได้ทำงานรวดเร็ว แต่มีความถูกต้องน้อยกว่าวิธีการอื่น โดยถูกนำมาใช้งานในการกรองข้อความ SMS ในรูปแบบของ Software ขนาดเล็กที่ติดตั้งบนโทรศัพท์เคลื่อนที่[1] การทำงานของวิธีการนี้ ผู้ใช้งานต้องเป็นผู้กำหนดกฎและเงื่อนไขขึ้น เพื่อให้สามารถกรองข้อความได้ถูกต้อง เช่น

- การใช้กฎ Black/white list ซึ่งจะตรวจสอบผู้ส่งข้อความว่าได้รับอนุญาตหรือไม่
- การใช้ Keywords Matching ที่จะตรวจสอบคำในข้อความ หากพบคำที่ผู้ใช้งานกำหนดให้เป็นข้อความสแปมจะดำเนินการคัดแยกข้อความออกไป

2.4.2 TFIDF[2]

เป็นวิธีการค้นหาลักษณะเด่นของเอกสาร (Document) ให้อยู่ในรูปของกลุ่มข้อมูล (Feature Vector) โดยอ้างอิงจากชุดตัวอักษรหรือคำ (Term) ในเอกสาร และจำนวนเอกสารทั้งหมดที่ถูกกำหนดให้เป็นข้อมูลฝึกสอน ดังสมการต่อไปนี้

$$TFIDF(i,j) = TF(i,j) \cdot IDF(i) \qquad \cdots\cdots (1)$$

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

TF คือ ความถี่ของ Term นี้ ที่ปรากฏใน Document

DF คือ ความถี่ของ Document ที่มี Term นี้

IDF คือ ค่าแทน Discrimination power ของ DF

จากนั้นจะทำการ Normalization ให้ vector เอกสารให้มีขนาด 1 หน่วยโดยใช้สมการ

$$w_{ik} = \frac{tf_{ik} \bullet \log\left(\frac{N}{df_k}\right)}{\sqrt{\sum_{j=1}^{t}\left(tf_{ij}\right)^2 \bullet \left(\log\left(\frac{N}{df_j}\right)\right)^2}} \qquad \cdots\cdots (3)$$

tf_{ik} คือ ความถี่ของคำ k ในเอกสาร i

N คือ คือจำ นวนของเอกสารในชุดเอกสาร

df_k จำ นวนของเอกสารในชุด เอกสารซึ่งบรรจุคำ k

---

[1] WebGate JSC (2007). SMS Spam Manager.

[2] อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย และ ชุลีรัตน์ จรัสกุลชัย. (2544). **Document Summarization.** หน้า 4-6

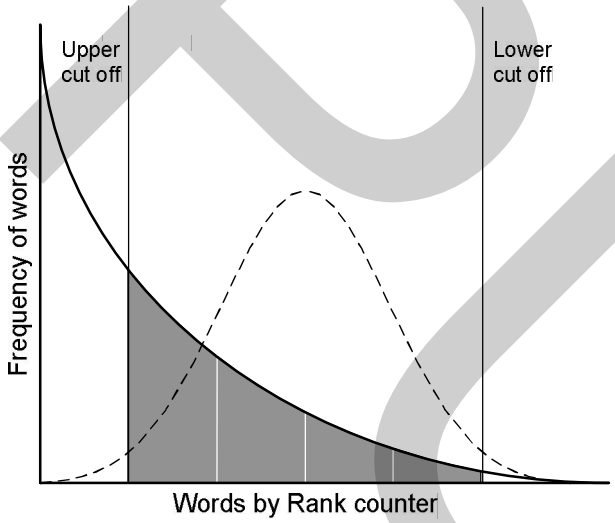
ในการจัดกลุ่มเอกสาร หรือการกรองข้อความแบบต่างๆที่ใช้การคำนวณทางคณิตศาสตร์เชิงเส้น จะใช้วิธีการ TFIDF ในการแปลงเอกสารที่ต้องการนำไปคำนวณ ให้เป็นชุดข้อมูล vector เพื่อนำไปคำนวณต่อไป

2.4.3 Text Normalization[1]

คือขั้นตอนการลบ สัญลักษณ์พิเศษ เช่น **$ | # | @ | ? | !** หรือตัวเลขที่ไม่ต้องการเพื่อกำจัดข้อมูลส่วนเกินออก ใช้ในการประมวลผลข้อมูลที่อยู่ในรูปแบบ text หรือ string เมื่อต้องการนำข้อมูลเหล่านั้นไปคำนวณค่า หรือเข้าสู่กระบวนการตัดคำ

2.4.4 Stop Words[2]

ในการจัดหมวดหมู่ข้อมูลประเภท text หรือการกรองข้อความ จะต้องค้นหาลักษณะแทนข้อมูล[3] (representation data) โดยการลบคำประเภท Stop words ที่ไม่มีความหมายใดๆในข้อมูลชุดนั้น โดยดูจากความถี่ของคำทั้งหมดเทียบกับปริมาณเอกสารดังภาพที่ 2.4



ภาพที่ 2.4 แสดงความถี่ของคำที่ใช้แทนลักษณะของเอกสาร

ที่มา: อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย, ชุลีรัตน์ จรัสกุลชัย. Document Summarization.

---

[1] István Pilászy. (2005). **Text Categorization and Support Vector Machines.** P. 2 – 3.

[2] ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. ข้อมูลพื้นฐาน ภาษาไทย, คำศัพท์ที่พบบ่อยในฐานข้อมูล.

[3] Mark Sanderson. (1999). Stop words.

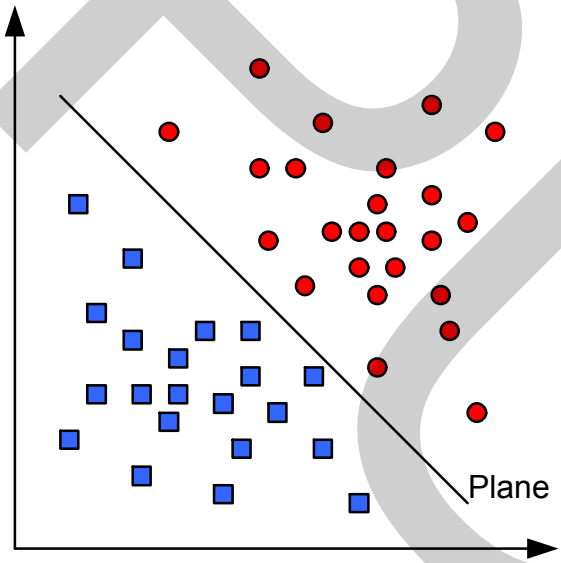
โดยในส่วนของคำที่มีความถี่สูงเกินไป (ทางซ้ายของเส้น Upper cut-off) และส่วนของคำที่มีความถี่น้อยเกินไป (ทางขวาของเส้น Lower cut-off) เป็นคำที่ไม่แสดงลักษณะแทนข้อมูลนั้นๆ จึงต้องทำการลบคำเหล่านั้นออก เพื่อจะได้ขอบเขตของคำที่แสดงลักษณะแทนข้อมูลโดย ซึ่งคำประเภทนี้จะมีองค์ประกอบคือ

- คำที่พบเป็นจำนวนมากในข้อความทุกข้อความ
- มีลักษณะเป็นคำขยาย หรือคำที่ไม่แสดงความหมาย

ตัวอย่างคำที่มีลักษณะเป็น Stop words ได้แก่ ฉัน เธอ นาย ที่ นี่ และ ไป เป็นต้น

2.4.5 Support Vector Machine (SVM)[1]

แนวความคิดของ Support Vector Machine เกิดขึ้นจากการนำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) จากนั้นจึงคำนวณหาเส้นตรงที่ใช้แบ่งข้อมูลทั้งสอง (Plane) ออกจากกันดังภาพที่ 2.5



ภาพที่ 2.5 ตำแหน่งข้อมูลสองกลุ่มในฟีเจอร์สเปซ (Feature Space)

ที่มา: István Pilászy. (2005). Text Categorization and Support Vector Machines.

[1] wikipedia.org (2006). Support vector machine.

และเพื่อให้ทราบว่าเส้นตรงที่แบ่งทั้งสองกลุ่มออกจากกันนั้น เส้นตรงใดเป็นเส้นที่ดีที่สุด จะใช้การฝึกสอนด้วยการกำหนดชุดข้อมูลตั้งต้นลงใน Feature Space และใช้สมการดังต่อไปนี้ในการสร้าง Model สำหรับ SVM ในการ Classify ข้อมูล

$$a = w^T \cdot x + b \qquad \cdots\cdots (4)$$

$$\text{new weight} = w + \eta(t-a)x^T \qquad \cdots\cdots (5)$$

$$\text{new bias} = b + \eta(t-a) \qquad \cdots\cdots (6)$$

w คือ ค่า vector น้ำหนักของสมการ SVM

x คือ vector ที่บรรจุ feature parameter ของข้อความ

b คือ ค่าคงที่สำหรับกำหนดความเบี่ยงเบน

t คือ ค่าที่ vector_x ควรจะเป็น (ปรกติ = 1 หรือ Spam = -1)

η คือ ค่าการเรียนรู้ของสมการ SVM

อธิบายกระบวนการสร้าง Model ของ SVM ตามสมการที่ 4 ถึง 6 ได้ดังนี้

กำหนดให้ Vector 1 (v1) มีค่า [1 1 1] , และมีค่า t = 1 (เป็นข้อความปรกติ)

กำหนดค่า vector น้ำหนักตั้งต้นเป็น [2 0 0] และค่าการเรียนรู้ของสมการเป็น 0.1

$$w = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}$$

$$\eta = 0.1$$

$$b = 0$$

$$a = f(n) = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases}$$

$$n = w^t \bullet x + b$$

$$n = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} \bullet \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + 0$$

$$n = 2, a = 1;$$

$$w_{new} = w + \eta(t-a)x^t = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix} + (0.1)(1-1)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}$$

$$b = b + \eta(t-a) = 0 + (0.1)(1-1) = 0$$

$$a = f(n) = a$$

$$w_{new} = w + \eta(t-a)x^t = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix} + (0.1)(1-2)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.9 & -0.1 & -0.1 \end{bmatrix}$$

$$b = b + \eta(t-a) = 0 + (0.1)(1-2) = -0.1$$

จากการคำนวณนี้จะได้ vector น้ำหนักใหม่เป็น [1.9 -0.1 -0.1] และค่าคงที่สำหรับกำหนดความเบี่ยงเบนเป็น -0.1 เพื่อนำค่าดังกล่าวไปคำนวณหาค่า Model SVM กับตัวอย่างฝึกสอนต่อไป สำหรับการฝึกสอน vector ที่ใช้ในการฝึกสอนต้องผ่านการกำหนดชนิดของข้อมูลเรียบร้อยแล้วได้แก่ ข้อความสแปม(-1) หรือ ข้อความปรกติ (1)

โดยเขียนแทนกระบวนการสร้าง Model ของ SVM ที่มีความสมบูรณ์ได้เป็นชุดสมการต่อไปนี้[1]

$$\min(w,b) = \frac{1}{2} \cdot w^T \cdot w + C \cdot \sum_{i=1}^{n} \xi_i \qquad \cdots\cdots (7)$$

$$\text{subject to} : \forall_{i=1}^{n} : y_i \left[w^T \cdot x_i + b \right] \geq \xi_i \cdots\cdots (8)$$

2.4.6 Naïve Bayesian (NB)[2]

ในทฤษฎีความน่าจะเป็น สถิติ การอนุมาน และ ปัญญาประดิษฐ์ บางครั้งจะพบคำว่า แบบเบย์ (Bayesian) มาขยายชื่อทฤษฎีหรือโมเดลต่างๆ โดยทุกครั้งที่พบคำขยายนี้หมายความว่า ได้มีการนำปรัชญาหรือหลักการของ ทฤษฎีความน่าจะเป็นแบบเบย์ (การอนุมานแบบเบย์ หรือ สถิติแบบเบย์) มาใช้กับสาขาความรู้นั้นๆชื่อเรียก "แบบเบย์" เริ่มต้นใช้ในช่วงปี ค.ศ. 1950 โดยมีต้นกำเนิดมาจากชื่อของ โทมัส เบย์ ผู้ซึ่งเสนอทฤษฎีบทของเบย์เป็นคนแรก (เท่าที่มีบันทึกในประวัติศาสตร์) ในเวลาถัดมาปีแยร์ ซิมง ลาปลาซ ได้เสนอทฤษฎีบทของเบย์เช่นกัน โดยในขณะนั้นลาปลาซไม่ทราบว่ามีงานของเบย์อยู่ ทฤษฎีบทของเบย์ในแบบของลาปลาซถูกนำไปใช้งานอย่างกว้างขวาง ทั้งนี้เนื่องจากการแปลความหมาย ความน่าจะเป็น ของลาปลาซนั้นกว้างมาก โดยลาปลาซได้นำไปในประยุกต์ใช้ในปัญหาของกลศาสตร์, ดาราศาสตร์, สถิติการแพทย์ (medical statistics) หรือแม้แต่ นิติศาสตร์

ลาปลาซได้ใช้ทฤษฎีความน่าจะเป็น (แบบเบย์) ในการทำนายมวลของดาวเสาร์โดยใช้ข้อมูลของวงโคจรดาวเสาร์ที่มีอยู่ในขณะนั้น โดยการคำนวณของลาปลาซมีความผิดพลาดไปเพียงร้อยละ 0.63 เท่านั้น ซึ่งในช่วงเวลานั้น ปัญหานี้ไม่สามารถสร้างการทดลองเชิงแนวคิดในลักษณะทดลองสร้างดาวเสาร์มา N ครั้ง มี M ครั้งที่มวลของดาวเสาร์เท่ากับ X ได้อย่างสมเหตุสมผล

ผู้บุกเบิกทฤษฎีความน่าจะเป็นแบบเบย์ที่มีชื่อเสียงคนอื่นๆ คือ จอห์น เมย์นาร์ด เคนส์, เลโอนาร์ด ซาเวจ, แฟรงค์ แรมซีย์, รูดอร์ฟ คาร์นาพ โดยนักทฤษฎีความน่าจะเป็นแบบเบย์ที่โด่งดังในช่วงปี ค.ศ. 1930 ถึง 1960 ที่ยังมีชีวิตอยู่ในปัจจุบันก็คือ เดนนิส ลินด์ลีย์ เจนส์ให้ข้อสังเกตไว้ว่า

---

[1] István Pilászy (2005). **Text Categorization and Support Vector Machines.** P. 2 – 3.

[2] wikipedia.org (2006). Naive Bayes classifier.

ผู้สนับสนุนทฤษฎีแบบเบย์ที่มีชื่อเสียงมักเป็นบุคคลจากสาขาอื่นที่ไม่ใช่คณิตศาสตร์ ไม่ว่าจะเป็น ตนเอง และคอกซ์ นักฟิสิกส์ เซอร์ แฮโรลด์ เจฟฟรีย์ นักธรณีวิทยา เคนส์ นักเศรษฐศาสตร์ หรือ คาร์นาพ นักปรัชญาวิทยาศาสตร์ ทั้งนี้อาจเป็นเพราะบุคคลเหล่านี้ต้องการนำทฤษฎีความน่าจะเป็นไปใช้งานจริง และต่างก็พบว่าทฤษฎีความน่าจะเป็นเชิงความถี่ไม่กว้างขวางพอที่จะเอาไปใช้จริงได้ อีกทั้งสถิติเชิงความถี่ยังไม่มีความน่าเชื่อถือ และสมเหตุสมผลพอ บุคคลเหล่านี้จึงต้องพัฒนาทฤษฎีความน่าจะเป็นที่สามารถนำไปใช้ได้จริงขึ้นมา และต่างก็ค้นพบแนวทางเดียวกันซึ่งก็คือสิ่งที่ลาปลาซได้แสดงไว้แล้วเมื่อราวต้นคริสต์ศตวรรษที่ 19

Naïve Bayesian เป็นทฤษฎีการใช้ Probability ในการแก้ปัญหาต่างๆที่ไม่สามารถใช้หลักสถิติกับชุดข้อมูลโดยตรงได้ การนำ NB มาใช้กับการกรองข้อความสแปมจะใช้การหา Probability ของคำในข้อความที่มีโอกาสเป็นสแปมเปรียบเทียบกับ Probability ของคำในข้อความที่มีโอกาสเป็นข้อความปรกติ ซึ่งหากเปรียบเทียบกันโดยใช้ ln แล้วมีค่ามากกว่า 0 แสดงว่า ข้อความดังกล่าวน่าจะเป็นสแปม[1] อธิบายได้จากสมการดังต่อไปนี้

$$p(Spam|D) = \frac{p(Spam)}{p(D)} \prod_i p(w_i|Spam) \quad \cdots\cdots (9)$$

$$p(Norm|D) = \frac{p(Norm)}{p(D)} \prod_i p(w_i|Norm) \quad \cdots\cdots (10)$$

$$\text{Spam} = \ln \frac{p(Spam|D)}{p(Norm|D)} > 0, \text{ otherwise Norm} \cdot (11)$$

P คือ ค่าความน่าจะเป็น

D คือ Document หรือ ข้อความ SMS

w_i คือ ลำดับของคำ ในข้อความ SMS

การตรวจสอบข้อความ จะใช้ความน่าจะเป็นข้อความสแปมหรือ P (Spam|D) เทียบกับความน่าจะเป็นข้อความปรกติ หรือ P (Norm|D) หากผลการเปรียบเทียบมีค่ากกว่า 0 แสดงว่าข้อความดังกล่าวเป็นข้อความสแปม

[1] Rafael Pinto. (2005). SpamFilter 1.1.

อธิบายการทำงานของ NB ได้จากตัวอย่างดังต่อไปนี้

ตารางที่ 2.2 แสดงตัวอย่างข้อมูลสำหรับคำนวณด้วยอัลกอริทึมแบบ NB

SMS	
Spam SMS	Norm SMS
A B E	A B C
A B C	A E F
B E E F	A C D F
D E F	A B C
A B E	B D F
	C D E

Words	A	B	C	D	E	F
Spam	3	4	1	1	5	2
Norm	4	3	4	3	2	3
$P(W_{spam})$	3/5	4/5	1/5	1/5	5/5	2/5
$P(W_{norm})$	4/6	3/6	4/6	3/6	2/6	3/6

P(Spam)	5/11
P(Norm)	6/11

กำหนดให้อักษร A B C D E F แทนคำในข้อความ SMS

กำหนดให้ SMS1 มีคำดังนี้ [A C D E] จะสามารถคำนวณค่าการเป็นสแปมได้ดังนี้

P(Spam|SMS) = (5/11) x (3/5) x (1/5) x (1/5) x (5/5) = 0.0109090909

P(Norm|SMS) = (6/11) x (4/6) x (4/6) x (3/6) x (2/6) = 0.0404040404

P(Norm|SMS) > P(Spam|SMS) → ข้อความ SMS1 เป็น ข้อความปรกติ

กำหนดให้ SMS2 มีคำดังนี้ [A B D E] จะสามารถคำนวณหาค่าการเป็นสแปมได้ดังนี้

P(Spam|SMS) = (5/11) x (3/5) x (4/5) x (1/5) x (5/5) = 0.0436363636

P(Norm|SMS) = (6/11) x (4/6) x (3/6) x (3/6) x (2/6) = 0.0303030303

P(Norm|SMS) < P(Spam|SMS) → ข้อความ SMS2 เป็น ข้อความสแปม

2.5 การตัดคำภาษาไทย

การกรองข้อความโดยทั่วไป เช่น ระบบกรองคำไม่สุภาพในเอกสารหน้า Web page หรือ E-Mail Spam Filter จะวิเคราะห์ข้อความจากคำ ซึ่งในภาษาอังกฤษใช้การเว้นวรรคเพื่อตัดคำ (word segmentation)[1] ในขณะที่ข้อความภาษาไทยไม่สามารถทำได้ เพราะใช้หลักการเขียนคำต่อกันเป็นประโยค ทำให้ต้องใช้ อัลกอริทึมในการตัดแยกคำ อีกทั้งข้อจำกัดของระบบ SMS ทำให้พฤติกรรมการส่งข้อความ มีลักษณะของ คำย่อ, คำทับศัพท์, หรือคำภาษาอังกฤษ ปะปนกันอย่างไม่เป็นระเบียบ จึงจำเป็นต้องปรับปรุงระบบการตัดคำภาษาไทยให้รองรับข้อความดังกล่าวได้

[1] วิรัช ศรเลิศล้ำวานิช, National Electronics and Computer Technology Center (NECTEC).(2543). โปรแกรมตัดคำภาษาไทย.

เทคนิคการตัดคำภาษาไทยแบ่งออกเป็น 3 แบบใหญ่ๆดังนี้[1]

1) วิธีการตัดคำแบบยาวที่สุด (Longest Matching)

ตัดคำด้วยการค้นหาคำเริ่มจากตัวอักษรซ้ายสุดของข้อความนั้นไปยังตัวอักษรถัดไป จนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรม

2) วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching)

ใช้การตัดคำที่สามารถจะเป็นไปได้ทั้งหมด แล้วเลือกข้อความที่ตัดได้จำนวนคำน้อยที่สุดมาใช้งาน

3) วิธีการตัดคำแบบคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Model)

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

4) วิธีการตัดคำแบบใช้คุณลักษณะ (Feature - Based Approach)

วิธีการนี้จะพิจารณาจากบริบท (context words) และการเกิดร่วมกันของคำ หรือหน้าที่ของคำ (collocation) เข้ามาช่วยในการตัดคำตัวอย่างเช่น “ตากลม” ถ้าพบคำว่า “แป๋ว” ในบริบทก็จะสามารถตัดคำได้ว่า “ตา” “กลม” “มากว่า” ถ้าในบริบทที่ตามมาเป็นตัวเลขก็สามารถตัดคำได้ว่า “มา” “กว่า”

2.6 ภาษา PHP

ประวัติความเป็นมาของภาษา PHP ถูกคิดค้นขึ้นในปี 1994 โดย ราสมัส เลอร์ดอร์ฟ (Rasmus Lerdorf) แต่ใน Version ที่ไม่เป็นทางการหรือกำลังทดสอบเท่านั้น โดย Rasmus ได้ทำการทดสอบกับWeb Page ของตนเองโดยใช้การตรวจสอบติดตามเก็บสถิติข้อมูล ผู้ที่เข้ามาเยี่ยมชมประวัติส่วนตัวบนWeb Page เท่านั้น

ต่อมา PHP Version แรกได้ถูกพัฒนา และเผยแพร่ให้กับผู้อื่นที่ต้องการใช้ศึกษาในปี ค.ศ. 1995 โดยเรียกว่า Person Home Page Tool ซึ่งเป็นที่มาของคำว่า PHP แต่ในระยะเวลานั้น PHP ยังไม่มีความสามารถอะไรที่โดดเด่นมากนัก จนกระทั่งเมื่อประมาณกลางปี 1995 Rasmus ได้คิดค้นและพัฒนาให้ PHP/FI หรือ PHP Version 2 ให้มีความสามารถจัดการเกี่ยวกับแบบรูปแบบข้อมูลที่ถูกสร้างมาจากภาษา HTML และสนับสนุนกับโปรแกรมจัดการฐานข้อมูล เอ็มเอสคิวแอล

[1] กำธน สินธวานนท์, พล.อ.ต. (2544). โปรแกรมตัดคำภาษาไทย (Thai Word Segmentation).

(mSQL) จึงทำให้ PHP เริ่มถูกใช้มากขึ้นอย่างรวดเร็ว และเริ่มมีผู้สนับสนุนการใช้งาน PHP มากขึ้น โดยในปลายปี ค.ศ. 1996 PHP ถูกนำไปใช้ประมาณ 15,000 เว็บไซต์ทั่วโลก และเพิ่มจำนวนขึ้นเรื่อย ๆ เป็น 50,000 เว็บไซต์

นอกจากนี้ ในราวกลางปี ค.ศ. 1997 PHP ได้มีการเปลี่ยนแปลงและถูกพัฒนาจาก Rasmus ซึ่งเป็นนักพัฒนาเพียงผู้เดียว มาเป็นทีมงาน โดยมี Zeev Suraski และ Andi Gutmans ทำการวิเคราะห์พื้นฐานของ PHP/FI และได้นำ Code มาพัฒนาใหม่เป็น PHP Version 3 ซึ่งมีความสมบูรณ์มากขึ้น

ในกลางปี ค.ศ. 1999 PHP Version 3 หรือ PHP 3 สามารถทำงานร่วมกับ C2's StrongHold Web Server และ Red Hat Linux ได้ ในปัจจุบัน PHP ถูกนำไปใช้ในเว็บไซต์ต่าง ๆ ทั่วโลกมากกว่า 150,000 เว็บไซต์และคาดว่าในอนาคต PHP รุ่นต่อไปจะถูกพัฒนาใหม่ให้มีประสิทธิภาพสูงขึ้น และสามารถที่จะทำงานภายใต้เว็บเซิร์ฟเวอร์ตัวอื่นได้นอกเหนือจากอะปาเชเว็บเซิร์ฟเวอร์ (Apache Web Server) ที่ใช้อยู่ในปัจจุบัน

ลักษณะสำคัญของ PHP ในการวิจัยและพัฒนา คือ

- Open source

การเป็น Software แบบ Open Source ทำให้ไม่มีข้อจำกัดด้านต้นทุนการพัฒนาโปรแกรมเมอร์สามารถนำ PHP ไปใช้งานใดๆ โดยสามารถพัฒนาต่อยอดให้เหมาะสมกับงานได้ ทำให้ ภาษา PHP มีพัฒนาการที่รวดเร็ว และการใช้งานที่แพร่หลาย

- Speed

PHP นำข้อดีของภาษาสคริปต์ที่เคยมีในภาษา C, Perl และ Java ร่วมกับความเร็วของ CGI มาใช้พัฒนาเป็นหลัก

- Process

ลักษณะการทำงานของ PHP นอกจากจะสามารถทำงานในรูปแบบ Web Server แล้ว PHP ยังสามารถทำงานในแบบ Console Mode และ Background Mode ซึ่งสามารถประยุกต์การ Process ข้อมูลได้หลายรูปแบบ

- Crossable Platform

ลักษณะเด่นของ PHP ที่สำคัญอีกประการ คือความเข้ากันได้กับระบบปฏิบัติการหลายชนิด เช่น Windows, Linux, UNIX เป็นต้น โดยไม่จำเป็นต้องเปลี่ยน Code คำสั่ง

- Database Access

PHP สามารถติดต่อกับฐานข้อมูลอย่าง dBase, Access, SQL Server, Oracle, Sybase, Informix, PostgreSQL, MySQL, Empress, FilePro, mSQL, PostgreSQL ได้อย่างมีประสิทธิภาพ

- String Handle

คำสั่งการจัดการ String ของภาษา PHP มีความยืดหยุ่นสูง รองรับการทำงานได้หลายรูปแบบ อีกทั้งยังมีความรวดเร็วในการประมวลผล

- Protocol Support

การสนับสนุนโปรโตคอลหลายแบบ ทั้ง IMAP, SNMP, NNTP, POP3, HTTP, SSL, Socket และยังสามารถพัฒนาให้รองรับการทำงานของ Protocol อื่นๆที่ใช้งาน TCP ได้ เช่น SMPP และ SOAP เป็นต้น

2.7 ระบบฐานข้อมูล

ฐานข้อมูล[1] (Database) คือ การจัดเก็บข้อมูลที่มีความสัมพันธ์กันไว้ด้วยกัน เพื่อลดปัญหาความซ้ำซ้อนของข้อมูลและการที่ไม่สามารถใช้ข้อมูลร่วมกันได้ เพื่อให้เป็นข้อมูลที่ใช้สนับสนุนการดำเนินงานอย่างใดอย่างหนึ่งขององค์กรเช่น ระบบฐานข้อมูลเงินเดือนที่จัดเก็บข้อมูลต่าง ๆที่สนับสนุนการคำนวณเงินเดือน เป็นต้น

ระบบการจัดการฐานข้อมูล (DBMS: Database Management System) คือ ซอฟต์แวร์ระบบที่ใช้ในการจัดการฐานข้อมูล ทำหน้าที่เป็นตัวกลางในการติดต่อระหว่างผู้ใช้กับฐานข้อมูล โดยมีหน้าที่สำคัญที่ต้องกระทำ ได้แก่ การจัดการพจนานุกรมข้อมูลการจัดเก็บข้อมูล การควบคุมการเข้าถึงข้อมูลจากผู้ใช้หลายคน การสำรองและการกู้คืนข้อมูล และภาษาที่ใช้ในการเข้าถึงฐานข้อมูลและการเชื่อมต่อกับโปรแกรมประยุกต์

2.8 ภาษา SQL

ภาษา SQL[2] (สามารถอ่านออกเสียงได้ 2 แบบ คือ “เอสคิวแอล” (SQL) หรือ “ซีเควล” (Sequel) ย่อมาจาก Structured Query Language หรือภาษาในการสอบถามข้อมูล เป็นภาษาทางด้านฐานข้อมูล ที่สามารถสร้างและปฏิบัติการกับฐานข้อมูลแบบสัมพันธ์ (relational database) โดยเฉพาะ และ เป็นภาษาที่มีลักษณะคล้ายกับภาษาอังกฤษ ภาษา SQLถูกพัฒนาขึ้นจากแนวคิดของ relational calculus และ relational algebra เป็นหลัก ภาษา SQL เริ่มพัฒนาครั้งแรกโดย almaden research center ของบริษัท IBM โดยมีชื่อเริ่มแรกว่า “ซีเควล” (Sequel) ต่อมาได้เปลี่ยนชื่อเป็น “เอสคิวแอล” (SQL) หลังจากนั้นภาษา SQLได้ถูกนำมาพัฒนาโดยผู้ผลิตซอฟแวร์ด้านระบบจัดการ

---

[1] วิเชียร เปรมชัยสวัสดิ์. (2546). ระบบฐานข้อมูล.

[2] ศิรินุช เทียนรุ่งโรจน์. (2552). ระบบฐานข้อมูล, มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร.

ฐานข้อมูลเชิงสัมพันธ์จนเป็นที่นิยมกันอย่างแพร่หลายในปัจจุบัน โดยผู้ผลิตแต่ละรายก็พยายามที่จะพัฒนาระบบจัดการฐานข้อมูลของตนให้มีลักษณะเด่นเฉพาะขึ้นมา ทำให้รูปแบบการใช้คำสั่ง SQL มีรูปแบบที่แตกต่างกันไปบ้าง เช่น ORACLE ACCESS SQL Base ของ Sybase INGRES หรือ SQL Server ของ Microsoft เป็นต้น ดังนั้นในปี ค.ศ. 1986 ทางด้าน American National Standards Institute (ANSI) จึงได้กำหนดมาตรฐานของ SQL ขึ้น อย่างไรก็ดี โปรแกรมฐานข้อมูลที่ขายในท้องตลาด ได้ขยาย SQL ออกไปจนเกินข้อกำหนดของ ANSI โดยเพิ่มคุณสมบัติอื่นๆ ที่คิดว่าเป็นประโยชน์เพิ่มเติม แต่โดยหลักทั่วไปแล้วก็ยังปฏิบัติตามมาตรฐานของ ANSI ในการอธิบายคำสั่งต่างๆของภาษา SQL

1. ประเภทของคำสั่งของภาษา SQL

ภาษา SQL เป็นภาษาที่ใช้งานได้ตั้งแต่ระดับเครื่อง Computer ส่วนบุคคล (Personal Computer หรือ PC) ไปจนถึงระดับ Mainframe ประเภทของคำสั่งในภาษา SQL (The subdivision of sql) แบ่งออกเป็น 3 ประเภท คือ

1.1 ภาษาสำหรับการนิยามข้อมูล (Data Definition Language: DDL)

ประกอบด้วยคำสั่งที่ใช้ในการกำหนดโครงสร้างข้อมูลว่ามีคอลัมน์อะไร แต่ละคอลัมน์เก็บข้อมูลประเภทใด รวมถึงการเพิ่มคอลัมน์ การกำหนดดัชนี การกำหนดวิวหรือตารางเสมือนของผู้ใช้ เป็นต้น

1.2 ภาษาสำหรับการจัดการข้อมูล (Data Manipulation Language: DML)

ประกอบด้วยคำสั่งที่ใช้ในการเรียกใช้ข้อมูล การเปลี่ยนแปลงข้อมูล การเพิ่มหรือลบข้อมูล เป็นต้น

1.3 ภาษาควบคุม (Data Control Language: DCL)

ประกอบด้วยคำสั่งที่ใช้ในการควบคุม การเกิดภาวะพร้อมกัน หรือการป้องกันการเกิดเหตุการณ์ที่ผู้ใช้หลายคนเรียกใช้ข้อมูลพร้อมกัน และคำสั่งที่เกี่ยวข้องกับการควบคุมความปลอดภัยของข้อมูลด้วยการกำหนดสิทธิของผู้ใช้ที่แตกต่างกัน เป็นต้น

2. ชนิดของข้อมูลที่ใช้ในภาษา SQL

ในภาษา SQL การบรรจุข้อมูลลงในคอลัมน์ต่าง ๆ ของตารางจะต้องกำหนดชนิดของข้อมูล (data type) ให้แต่ละคอลัมน์ ชนิดของข้อมูลนี้จะแสดงชนิดของค่าที่อยู่ในคอลัมน์ ค่าทุกค่าในคอลัมน์ที่กำหนดจะต้องเป็นชนิดเดียวกัน เช่น ในตารางลูกค้าคอลัมน์ที่เป็นรายชื่อลูกค้าจะต้องเป็นตัวหนังสือ ในขณะที่คอลัมน์จำนวนเงินที่ลูกค้าซื้อสินค้าเป็นตัวเลข

ชนิดของข้อมูลของแต่ละคอลัมน์จะขึ้นกับลักษณะของข้อมูลแต่ละคอลัมน์ ซึ่งแบ่งได้ดังนี้

2.1 ตัวหนังสือ (Character) ในภาษา SQL มีดังนี้

- ตัวหนังสือแบบความยาวคงที่ (fixed-length character) จะใช้ char (n) หรือ character (n) แทนประเภทของข้อมูลที่เป็นตัวหนังสือใดๆที่มีความยาวของข้อมูลคงที่โดยมีความยาว n ตัวหนังสือประเภทนี้จะมีการจองเนื้อที่ตามความยาวที่คงที่ตามที่กำหนดไว้ ชนิดของข้อมูลประเภทนี้จะเก็บความยาวของข้อมูลได้มากที่สุดได้ 255 ตัวอักษร

- ตัวหนังสือแบบความยาวไม่คงที่ (variable-length character) จะใช้ varchar (n) แทนประเภทของข้อมูลที่เป็นตัวหนังสือใดๆที่มีความยาวของข้อมูลไม่คงที่ โดยมีความยาว n ตัวหนังสือประเภทนี้จะมีการจองเนื้อที่ตามความยาวของข้อมูล ชนิดของข้อมูลประเภทนี้จะเก็บความยาวของข้อมูลได้มากที่สุดได้ 4000 ตัวอักษร

2.2 จำนวนเลข (Numeric)

- จำนวนเลขที่มีจุดทศนิยม (Decimal) ในภาษา SQL จะใช้ dec (m,n) หรือ decimal (m,n) เป็นประเภทข้อมูลที่เป็นจำนวนเลขที่มีจุดทศนิยมโดย m คือจำนวนตัวเลขทั้งหมด (รวมจุดทศนิยม) และ n คือจำนวนตัวเลขหลังจุดทศนิยม

- จำนวนเลขที่ไม่มีจุดทศนิยมในภาษา SQL จะใช้ int หรือ integer เป็นเลขจำนวนเต็มบวกหรือลบขนาดใหญ่ เป็นตัวเลข 10 หลัก ที่มีค่าตั้งแต่ –2,147,483,648 ถึง +2,147,483,647 และในภาษา SQL จะใช้ smallint เป็นประเภทข้อมูลที่เป็นเลขจำนวนเต็มบวกหรือลบขนาดเล็ก เป็นตัวเลข 5 หลัก ที่มีค่าตั้งแต่ –32,768 ถึง +32,767 ตัวเลขจำนวนเต็มประเภทนี้จะมีการจองเนื้อที่น้อยกว่าแบบ integer

- เลขจำนวนจริง ในภาษา SQL อาจใช้ number (n) แทนจำนวนเลขที่ไม่มีจุดทศนิยมและจำนวนเลขที่มีจุดทศนิยม

2.3 ข้อมูลในลักษณะอื่นๆ

วันที่และเวลา (Date/Time) เป็นชนิดวันที่หรือเวลาในภาษา SQL จะใช้ date เป็นข้อมูลวันที่ ซึ่งจะมีหลากหลายรูปแบบให้เลือกใช้ เช่น yyyy-mm-dd (1999-10-31) dd.mm.yyyy (31.10.1999) หรือ dd/mm/yyyy (31/10/1999)

บทที่ 3
ระเบียบวิธีวิจัย

3.1 แนวทางการวิจัยและพัฒนา

งานวิจัยนี้มีวัตถุประสงค์ในการออกแบบวิธีการกรองข้อความ SMS ในประเทศไทย ที่มีการใช้ภาษาไทย ภาษาอังกฤษ และภาษาไทยปนภาษาอังกฤษ ในการส่งข้อความ โดยแบ่งขั้นตอนการวิจัยออกเป็น 2 ส่วนดังนี้

3.1.1 ศึกษาและเปรียบเทียบหาวิธีการกรองที่เหมาะสม

เนื่องจากยังไม่มีงานวิจัยที่ศึกษาการกรองข้อความจากบริการส่งข้อความ SMS ในประเทศไทยอย่างจริงจัง จึงต้องใช้การศึกษาวิธีกรองข้อความที่มีใช้งานในต่างประเทศเป็นพื้นฐานอ้างอิง โดยวิธีการที่ถูกใช้งานอย่างแพร่หลายมีด้วยกัน 2 วิธีได้แก่ SVM และ NB ซึ่งจะทำการศึกษาและปรับปรุงการทำงานบางส่วนให้สามารถใช้งานร่วมกับภาษาไทยได้ เพื่อวิจัยเปรียบเทียบในด้านประสิทธิภาพความถูกต้องและระยะเวลาในการประมวลผล จากข้อความ SMS ที่มีการตรวจสอบลักษณะข้อความสแปมด้วยมนุษย์ โดยนำวิธีการกรองที่มีประสิทธิภาพ มาปรับปรุงให้สอดคล้องกับข้อความ SMS ในประเทศไทยต่อไป

3.1.2 วิเคราะห์ปัญหาการกรองข้อความ SMS ของประเทศไทย

ในขั้นตอนการวิจัยส่วนที่ 1 จะทำให้ทราบหลักการทำงานและข้อบกพร่องในการกรองข้อความ SMS ซึ่งจะนำผลการทดสอบมาทำการวิเคราะห์ เพื่อนำไปแก้ไขและปรับปรุงวิธีการกรองให้มีความสอดคล้องกับข้อความ SMS ในประเทศไทย ด้วยการดำเนินงานในส่วนที่ 2 และทำการทดสอบการกรองข้อความด้วยข้อความ SMS ชุดเดิมอีกครั้ง เพื่อหาส่วนต่างของประสิทธิภาพที่เพิ่มขึ้น แล้วสรุปงานวิจัยเพื่อนำไปพัฒนาวิธีการกรองให้สามารถใช้งานในระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่

3.2 เครื่องมือที่ใช้ในงานวิจัย

3.2.1 เครื่อง Desktop Computer หรือ Laptop สำหรับติดตั้ง Software ในการทดสอบกรองข้อความสแปมจำนวน 1 เครื่อง

CPU Intel Core 2 Duo Processor 2.26 GHz

Mainboard

RAM 3 GB DDRII 800 MHz

HARD DISK 250 GB SATA II

DVD - RW

10/100/1000 LAN Built-In

USB Mouse & Keyboard

LCD 14.1 Inch Display

3.2.2 Software สำหรับใช้ในการสร้างโปรแกรม Spam Filter

PHP

SWATH

SVM Light

MsSQL

Genuine Windows Vista® Ultimate

Windows 2003 Server

Vmware

3.2.3 แบบสอบถาม สำหรับนิยามความหมายของข้อความสแปม

เอกสารแบบสอบถามจำนวน 500 ชุด

แบบสอบถาม Online สำหรับกรองข้อมูลผ่าน Internet

3.3 แผนการดำเนินงาน

ตารางที่ 3.1 แผนการดำเนินงาน

รายการดำเนินงาน	ระยะเวลา (เดือน)									
	1	2	3	4	5	6	7	8	9	10
รวบรวมข้อมูลจากผู้ใช้งานโทรศัพท์เคลื่อนที่	■									
รวบรวมข้อมูล (CDR)	■	■	■							
ศึกษา Filter ที่มีการใช้งานในปัจจุบัน	■	■	■							
ออกแบบและพัฒนาระบบการกรองข้อความ		■	■	■	■					
ทดสอบการประมวลผลของระบบขั้นต้น			■	■	■					
ทดสอบเปรียบเทียบเพื่อหาประสิทธิภาพ						■	■			
สรุปผลการเปรียบเทียบและประโยชน์							■	■		

3.4 ขั้นตอนการดำเนินงานวิจัย

3.4.1 รวบรวมและวิเคราะห์ข้อวาม SMS ในประเทศไทย

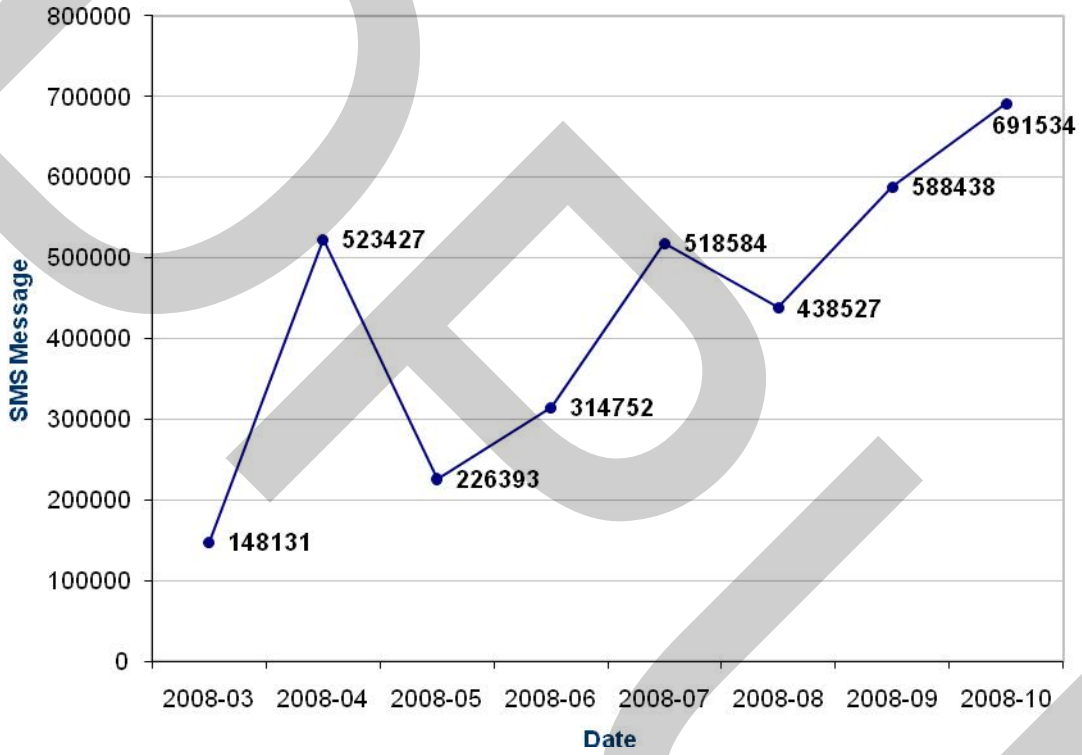
ก่อนทำการออกแบบ Filter นั้น จำเป็นต้องมีการวิเคราะห์พฤติกรรมของข้อความ SMS เพื่อเป็นแนวทางในการพัฒนา โดยจะทำการรวบรวมข้อมูลการใช้งานบริการส่งข้อความ SMS จากระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่ในประเทศไทยรายหนึ่ง เป็นระยะเวลา 8 เดือน เพื่อใช้ในการศึกษาลักษณะและแนวโน้มทางสถิติของข้อความในระบบส่งข้อความ โดยมีวิธีดำเนินการดังนี้

1) ติดต่อผู้ให้บริการโทรศัพท์เคลื่อนที่ เพื่อขอความอนุเคราะห์ข้อมูลการ รับ – ส่งข้อความสั้น SMS เป็นระยะเวลา 8 เดือน

2) ทำการเก็บรวบรวมข้อมูล CDR ของบริการ SMS จากเครื่อง SMSC ของผู้ให้บริการโทรศัพท์เคลื่อนที่

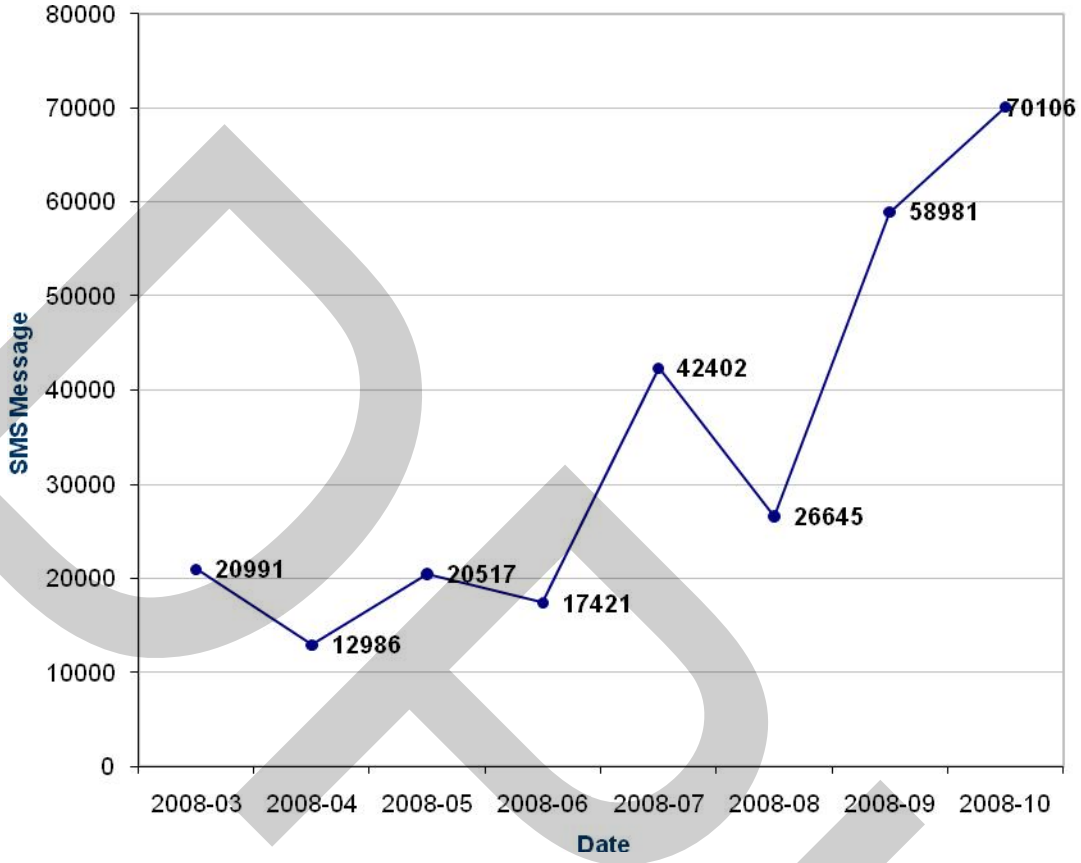
3) นำข้อความ SMS ที่ได้รับมาจัดเตรียมลงสู่ฐานข้อมูล เพื่อใช้ในการเรียกดึงข้อมูลได้อย่างเป็นระบบและมีความรวดเร็ว

4) สร้างกราฟแสดงข้อมูลทางสถิติซึ่งมีรายละเอียดดังตารางที่ 3.1



ภาพที่ 3.1 ข้อมูลการใช้บริการ SMS ทั้งหมด ของ CAT CDMA ตั้งแต่ 03/2008 ถึง 10/2008

จากภาพที่ 3.1 ซึ่งแสดงให้เห็นแนวโน้มการใช้บริการที่มีการเพิ่มสูงขึ้นในแต่ละเดือน สำหรับในเดือนที่มีเทศกาลพิเศษ เช่น เทศกาลสงกรานต์ จะมีสถิติการใช้งานสูงกว่าการใช้งานในเดือนอื่นๆ เนื่องจากเป็นช่วงวันหยุดพักผ่อน ซึ่งผู้ใช้งานเดินทางไปท่องเที่ยวในต่างจังหวัด ทำให้มีการสื่อสารถึงบุคคลอื่นสูงกว่าปรกติ



ภาพที่ 3.2 ข้อมูลการใช้บริการ SMS ผ่าน TCP/IP ของ CAT4SMS ตั้งแต่ 03/2008 ถึง 10/2008

จากภาพที่ 3.2 ซึ่งแสดงให้เห็นแนวโน้มการใช้บริการส่งข้อความ SMS ผ่าน TCP/IP ที่มีการเพิ่มสูงขึ้นในแต่ละเดือนเช่นเดียวกับการส่งข้อความจากโทรศัพท์เคลื่อนที่ สำหรับในเดือนที่มีเทศกาลพิเศษ เช่น เทศกาลสงกรานต์ จะมีสถิติการใช้งานต่ำกว่าการใช้งานในเดือนอื่นๆ เนื่องจากเป็นช่วงวันหยุดพักผ่อน ซึ่งผู้ใช้งานเดินทางไปท่องเที่ยวในต่างจังหวัด ทำให้การใช้งานบริการส่งข้อความผ่าน TCP/IP จากบุคคลทั่วไปลดลง นอกจากนี้ การบันทึกข้อมูลจากระบบตรวจสอบความถี่ของข้อมูลที่ SMS Gateway แสดงรูปแบบความถี่ในการส่งข้อความในแต่ละวันดังภาพที่ 3.3

ภาพที่ 3.3 ตัวอย่างบันทึกการใช้งานบริการ CAT4SMS (บริการส่ง SMS ผ่านเว็บ) ตั้งแต่วันที่ 12/05/2551 ถึง 15/05/2551

จากภาพที่ 3.3 แสดงให้เห็นความถี่ในการส่งข้อความตามช่วงเวลาต่างๆ ของบริการ CAT4SMS ซึ่งมีความถี่ในการส่งหนาแน่นในช่วงเวลา 09:00 น. ถึง 16:00 น. ในแต่ละวัน ลักษณะของข้อความจะเป็นการส่งข้อความ 1 ข้อความ ไปยังผู้รับมากกว่า 1 ราย

3.4.2 การกำหนดความหมายของข้อความสแปมและข้อความปรกติ

การกรองข้อความสแปมออกจากระบบส่งข้อความ SMS นั้น จำเป็นต้องกำหนดความหมายของข้อความสแปมให้มีความชัดเจน ขณะที่ผู้ใช้งานโทรศัพท์เคลื่อนที่แต่ละคนอาจมีทัศนคติในการตัดสินว่าข้อความใดเป็นข้อความสแปมหรือข้อความปรกติที่แตกต่างกัน ตัวอย่างความแตกต่างทางความคิดในการตัดสินข้อความสแปมเช่น ผู้ใช้งานโทรศัพท์เคลื่อนที่ที่มีรูปร่างอ้วนจะมีความสนใจในข้อความ SMS ที่เกี่ยวข้องกับการลดความอ้วน ได้แก่ ข้อความส่งเสริมการขายยาลดความอ้วน อาหารเสริม หรือสถานออกกำลังกายต่างๆเพื่อลดความอ้วน ในขณะที่ผู้ใช้งานโทรศัพท์เคลื่อนที่ที่รูปร่างปรกติไม่ต้องการรับข้อมูลข่าวสารดังกล่าว

ความแตกต่างทางความคิดดังตัวอย่างข้างต้น ทำให้การตัดสินความหมายของข้อความ SMS เพื่อใช้เป็นข้อมูลฝึกสอนสำหรับออกแบบวิธีกรองข้อความ ไม่สามารถทำได้ถูกต้องทุกข้อความ การทำแบบสำรวจความคิดเห็น เพื่อใช้กำหนดความหมายของข้อความสแปมกับผู้ใช้งานโทรศัพท์เคลื่อนที่ ทำให้สามารถเข้าใจทัศนคติของกลุ่มผู้ใช้งานที่มีต่อข้อความสแปมได้ชัดเจน โดยใช้แบบสำรวจปลายปิด เพื่อให้กลุ่มตัวอย่างสามารถแสดงความคิดเห็นได้อย่างสะดวกรวดเร็ว

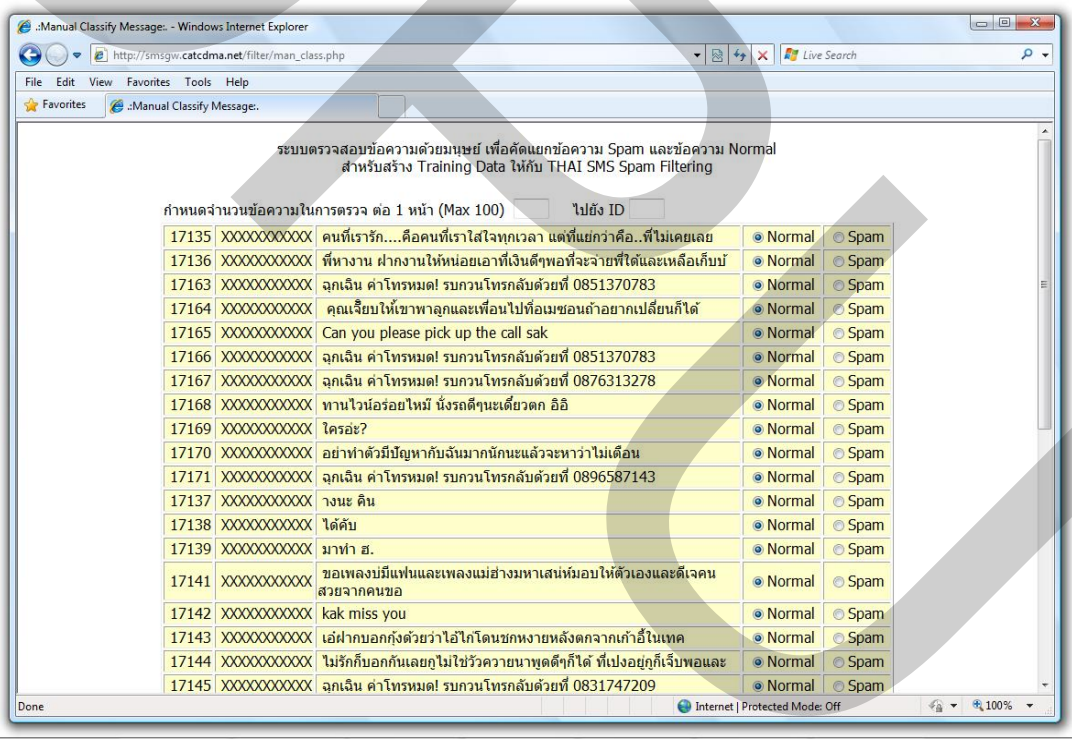
และจัดทำแบบสำรวจเป็น 2 ช่องทาง ได้แก่ แบบสำรวจที่จัดทำขึ้นเป็นเอกสารแบบ 2 หน้ากระดาษ A4 และรูปแบบ Online ในรูปของ HTML เชื่อมต่อกับ ฐานข้อมูล Microsoft SQL และการนำข้อมูลดังกล่าวมาใช้กำหนดทิศทางของการออกแบบวิธีกรองข้อความ จะช่วยเพิ่มความถูกต้องในการกรองได้อีกด้วย โดยมีวิธีดำเนินการดังนี้

1) สร้างแบบสำรวจความคิดเห็นโดยให้ความสำคัญในการนิยามความหมายของคำว่า “ข้อความสแปม”

2) จัดทำแบบสำรวจในรูปสิ่งพิมพ์ และ HTML เพื่อเป็นช่องทางการรวบรวมข้อมูล

3) เก็บรวบรวมความคิดเห็น และจัดเตรียมข้อมูลลงสู่ฐานข้อมูลเพื่อใช้ในการสรุปผล

4) พัฒนาเครื่องมือในการคัดแยกข้อความสแปมแบบ Web Application เพื่อให้การคัดแยกข้อความด้วยมนุษย์จากข้อสรุปนิยามของ “ข้อความสแปม” ตามภาพที่ 3.4



ภาพที่ 3.4 ระบบจำแนกข้อความด้วยมนุษย์ ผ่าน Web Application

จากภาพที่ 3.4 แสดงหน้าจอการทำงานของเครื่องมือการคัดแยกข้อความสแปมแบบ Online ผ่าน Web Application ซึ่งพัฒนาด้วยภาษา HTML และ PHP โดยเชื่อมต่อกับฐานข้อมูล Microsoft SQL ที่เก็บรวบรวข้อมูล SMS

3.4.3 ศึกษาวิธีการกรองข้อความ SMS ในปัจจุบัน

โดยค้นคว้างานวิจัยประเภทต่างๆดังต่อไปนี้

1) งานวิจัยที่เกี่ยวข้องกับระบบส่งข้อความ SMS

2) งานวิจัยที่เกี่ยวข้องกับการจัดหมวดหมู่เอกสารภาษาไทย

3) งานวิจัยที่เกี่ยวข้องกับการตรวจสอบข้อความสแปมทั้งจากระบบส่งข้อความสั้น SMS และระบบรับ-ส่ง E-Mail

4) งานวิจัยที่เกี่ยวข้องกับอัลกอริทึมในการจัดกลุ่มข้อมูล

5) งานวิจัยที่เกี่ยวข้องกับการตัดคำภาษาไทย

พัฒนาโปรแกรมสำหรับกรองข้อความโดยอัลกอริทึมแบบ SVM และ NB ด้วยภาษา PHP โดยมีขั้นตอนดังนี้

1) จัดเตรียม Module การตรวจสอบ Stop words โดยการพัฒนาจากภาษา PHP

2) จัดเตรียม Module การตัดคำภาษาไทย ที่สามารถตัดคำด้วยวิธีการดังต่อไปนี้

- Longest Matching
- Maximal Matching
- Probabilistic Model

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

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

4) จัดเตรียม Module การประมวลผลด้วย SVM โดยใช้โปรแกรม SVM-Light[2] ซึ่งเป็นโปรแกรมในการคัดแยกข้อมูลที่ได้รับความนิยมสูง และพัฒนาวิธีการเชื่อมต่อข้อมูลระหว่าง PHP กับ SVM-Light เพิ่มเติมเพื่อให้ PHP เข้าใจและรับรู้ผลการคัดแยกข้อมูลแบบ SVM

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

---

[1] วิรัช ศรเลิศล้ำวานิช. National Electronics and Computer Technology Center (NECTEC).(2543). โปรแกรมตัดคำภาษาไทย.

[2] Thorsten Joachims. (2008). SVM Light, Support Vector Machine.

[3] wikipedia.org (2006). Naive Bayes classifier.

[4] Rafael Pinto. (2005). SpamFilter 1.1.

3.4.4 ออกแบบการทดสอบประสิทธิภาพวิธีการกรองข้อความ

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

จัดเตรียมเครื่อง Computer หรือ Laptop สำหรับทำการทดสอบ ซึ่งประกอบด้วย CPU Core 2 Duo Processor ความเร็ว 2.26 GHz หน่วยความจำขนาด 3 GB ระบบปฏิบัติการ Windows 2003 Server ใน VMWare บน Windows Vista Ultimate โดยแบ่งการทดสอบออกเป็น

1) การทดสอบด้วยวิธีการตัดคำแบบต่างๆโดยใช้วิธีกรองข้อความแบบ NB

2) การทดสอบด้วยวิธีการตัดคำแบบต่างๆโดยใช้วิธีกรองข้อความแบบ SVM

เมื่อการทดสอบเสร็จสิ้น ดำเนินการเก็บข้อมูลการทดสอบเพื่อใช้ในการศึกษาปัญหาการกรองข้อความในภาษาไทยต่อไป

3.4.5 ศึกษาปัญหาและการแก้ไขปัญหาการกรองข้อความ SMS

ศึกษาความเป็นไปได้และปัญหาที่จะเกิดขึ้นในการกรองข้อความภาษาไทย ของวิธีกรองข้อความจากการทดสอบประสิทธิภาพ ได้แก่

1) การตรวจสอบความผิดปรกติต่างๆของคำในข้อความ

ทั้งข้อความภาษาไทย ภาษาอังกฤษและภาษาไทยปนภาษาอังกฤษ ที่ถูกสร้างจากโทรศัพท์เคลื่อนที่ การพิมพ์ข้อความที่ผิดพลาด การพิมพ์ตัวอักษรเดียวกันมากกว่า 1 ครั้ง และคำที่ไม่สมบูรณ์เนื่องจากมีการตัดแบ่งข้อความที่ไม่ถูกต้อง และออกแบบวิธีการแก้ปัญหาดังกล่าว

2) การตรวจสอบการตัดคำ

วิธีการตัดคำในปัจจุบัน เป็นวิธีการที่ออกแบบมาเพื่อตัดคำที่มีความถูกต้องตามหลักภาษาศาสตร์ ซึ่งไม่สามารถตัดคำจากข้อความ SMS ได้อย่างถูกต้อง การตรวจสอบวิธีการตัดคำ จะทำให้เข้าใจถึงความผิดพลาดที่เกิดขึ้น และสามารถนำไปปรับปรุงแก้ไขให้การตัดคำในข้อความ SMS มีความถูกต้องมากขึ้น

3) การปรับปรุงในส่วนอื่นๆ

นอกจากความผิดปรกติของคำ และการตัดคำที่ไม่ถูกต้องแล้ว อาจมีองค์ประกอบอื่นๆ ที่สามารถปรับปรุงและส่งผลกระทบต่อการกรองข้อความให้มีประสิทธิภาพสูงขึ้นได้

3.4.6 การทดสอบประสิทธิภาพวิธีการกรองที่พัฒนาเสร็จสิ้น

ทำการทดสอบเช่นเดียวกับการทดสอบก่อนหน้า โดยเปรียบเทียบระหว่างวิธีการกรองที่ยังไม่ผ่านการปรับปรุงขั้นตอนต่างๆกับวิธีการกรองข้อความที่ผ่านการปรับปรุงการแก้ไขข้อผิดพลาด

3.4.7 รายงานผลการวิจัยและสรุปข้อเสนอแนะ

แสดงผลการวิจัยและสรุปผลการดำเนินงาน พร้อมทั้งข้อเสนอแนะต่างๆ จัดทำรายงานการวิจัย และนำเสนอผลงาน

บทที่ 4
การกรองข้อความ SMS ภาษาไทย

4.1 การนิยามข้อความสแปม

4.1.1 การสำรวจความคิดเห็น

การกรองข้อความสแปมออกจากระบบส่งข้อความสั้น SMS นั้น จำเป็นต้องกำหนดความหมายของข้อความสแปมให้มีความชัดเจน เพราะผู้ใช้บริการแต่ละคนอาจมีทัศนคติในการตัดสินว่าข้อความใดเป็นข้อความสแปมหรือข้อความปรกติที่แตกต่างกัน ซึ่งจากการจัดทำแบบสำรวจความคิดเห็น (รายละเอียดของแบบสำรวจความคิดเห็นในเอกสารภาคผนวก ข.) ของผู้ใช้งานโทรศัพท์เคลื่อนที่ในประเทศไทย ซึ่งมีสาระสำคัญของแบบสอบถามดังนี้

1) ความหมายของข้อความสแปมซึ่งจะนำมาใช้กำหนดชุดข้อมูลฝึกสอน เพื่อให้การกรองข้อความ SMS สามารถทำได้ตรงกับกลุ่มผู้ใช้งานในประเทศไทยมากที่สุด

2) ตัวอย่างคำและลักษณะเฉพาะของข้อความที่พบได้ในข้อความสแปม เพื่อใช้กำหนดเงื่อนไขเพิ่มเติม จากระบบการกรองที่มีอยู่ในปัจจุบัน ให้มีความถูกต้องในการกรองสูงขึ้น

3) ผลกระทบและการแก้ไขปัญหาข้อความสแปมใช้ในการอ้างอิงถึงความรุนแรงของปัญหาข้อความสแปมที่เกิดขึ้นในประเทศไทย

การสำรวจความคิดเห็นจากกลุ่มตัวอย่างผู้ที่ใช้งานโทรศัพท์เคลื่อนที่จำนวน 468 ตัวอย่าง มีรายละเอียดดังตารางที่ 4.1

ตารางที่ 4.1 แสดงผลการตอบแบบสำรวจส่วนที่ 1 ข้อมูลทั่วไป

คำถาม/คำตอบ	จำนวนผู้ตอบ (คน)
เพศ	
ชาย	265
หญิง	203
อายุ	
ต่ำกว่า 20 ปี	56
21 – 30 ปี	307
31 – 40 ปี	75
40 ปีขึ้นไป	30
สถานะ	
โสด	393
สมรส	71
หม้ายหรือหย่า	4
อาชีพ	
นักเรียน / นักศึกษา	239
รับราชการ / รัฐวิสาหกิจ	139
พนักงานบริษัท / ธุรกิจส่วนตัว	90
จำนวนโทรศัพท์เคลื่อนที่ที่ใช้งานพร้อมกัน	
1 เครื่อง	294
2 เครื่อง	161
มากกว่า 2 เครื่อง	13
จำนวนข้อความ SMS ที่ได้รับโดยเฉลี่ยใน 1 วัน	
ไม่มี	39
น้อยกว่า 3 ข้อความ	276
น้อยกว่า 5 ข้อความ	112
มากกว่า 5 ข้อความ	41
จำนวนข้อความ SMS ที่ส่งไปยังหมายเลขอื่นโดยเฉลี่ยใน 1 วัน	
ไม่มี	183
น้อยกว่า 3 ข้อความ	236
น้อยกว่า 5 ข้อความ	37
มากกว่า 5 ข้อความ	12

ตารางที่ 4.2 แสดงผลการตอบแบบสำรวจส่วนที่ 2 ข้อมูลลักษณะสแปม

คำถาม/คำตอบ	จำนวนผู้ตอบ (คน)
ความหมายของข้อความสแปมทาง SMS (ตอบได้มากกว่า 1 ข้อ)	
ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ	361
ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่	275
ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้	153
ข้อความหยาบคาย หรือข้อความที่ไม่มีสาระสำคัญ	65
ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง	133
ความถี่ของข้อความสแปมที่ท่านได้รับในแต่ละวัน	
ไม่มี	86
น้อยกว่า 3 ข้อความ	289
น้อยกว่า 5 ข้อความ	70
มากกว่า 5 ข้อความ	21
ภาษาของข้อความสแปมที่ท่านได้รับ	
ภาษาอังกฤษเพียงอย่างเดียว	21
ภาษาไทยหรือภาษาไทยปนภาษาอังกฤษ	429
คำใดบ้างที่ท่านคิดว่าจะพบในข้อความสแปม	ตารางที่ 4.3
ช่วงเวลาที่ท่านได้รับข้อความ สแปม	
06:01 น. – 20:00 น.	266
20:01 น. – 06:00 น.	37
ตลอดทั้งวัน ไม่แน่นอน	165
ผลกระทบของข้อความสแปม (ตอบได้มากกว่า 1 ข้อ)	
ทำให้แบตเตอร์รี่หมดเร็วขึ้น	101
ก่อความรำคาญและทำให้ใช้งานไม่สะดวก	357
ถูกละเมิดสิทธิส่วนบุคคล	97
เสียค่าบริการจากโฆษณาในข้อความสแปมเพิ่มขึ้น	53
เสียพื้นที่ในการเก็บข้อความที่จำเป็น (Inbox เต็ม)	66
วิธีการแก้ปัญหาข้อความสแปม (ตอบได้มากกว่า 1 ข้อ)	
ไม่ดำเนินการใดๆ	86
รับข้อความและเปิดอ่านตามปกติ	181
ลบข้อความทิ้งทันที	297
ปิดเครื่องโทรศัพท์ทันที	470
แจ้งผู้ให้บริการโทรศัพท์เคลื่อนที่เพื่อให้ดำเนินการแก้ไข	54
เลือกใช้โทรศัพท์เคลื่อนที่ที่สามารถกรองข้อความสแปมได้	27

ตารางที่ 4.3 แสดงคำที่ผู้ตอบแบบสำรวจพบในข้อความสแปม

ลำดับ	คำ	ลำดับ	คำ	ลำดับ	คำ	ลำดับ	คำ
1	bonus	21	ซื้อ	41	ลักษณ์	61	เชิญชวน
2	download	22	ดวง	42	ลุ้น	62	เซียมซี
3	duty	23	ดารา	43	สนุก	63	เด็ด
4	free	24	ดาวน์	44	สมัคร	64	เติม
5	mail	25	ดาวน์โหลด	45	สลาก	65	เบอร์
6	mms	26	ดูดวง	4	สอบถาม	66	เพลง
7	promotion	27	ดูหมอ	47	สิทธิพิเศษ	67	เพิ่ม
8	push	28	ด่วน	48	สินค้า	68	เพิ่มเติม
9	ringtone	29	บริการ	49	สุขภาพ	69	เพียง
10	sms	30	พยากรณ์	50	ส่ง	70	เวลา
11	vote	31	พิเศษ	51	ส่งเสริม	71	เสียง
12	www	32	ฟรี	52	ส่วนลด	72	แม่น
13	xxx	33	ฟุตบอล	53	หาคู่	73	โฆษณา
14	กด	34	ราคา	54	ห้างสรรพสินค้า	74	โชคดี
15	ขาย	35	รางวัล	55	อ้วน	75	โทน
16	คลิป	36	รายการ	56	ฮิต	76	โทรศัพท์
17	คอร์ส	37	รายละเอียด	57	เกมส์	77	โบนัส
18	คูปอง	38	รูปภาพ	58	เครดิต	78	โหลด
19	ค่าบริการ	39	ร้านค้า	59	เงิน	79	ใหม่
20	ชิง	40	ลด	60	เงินพิเศษ	-	-

จากตารางที่ 4.3 สามารถสรุปสาระสำคัญจากการสำรวจความคิดเห็นได้ดังต่อไปนี้

1. ความหมายของข้อความสแปม ซึ่งสามารถแบ่งออกได้เป็น 4 ประเภทเรียงตามลำดับคะแนนดังนี้

1.1. ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ จำนวน 361 คะแนน

1.2. ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่ จำนวน 275 คะแนน

1.3. ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้ จำนวน 153 คะแนน

1.4. ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง จำนวน 133 คะแนน

2. ผลกระทบและการแก้ไขปัญหาข้อความสแปมเรียงตามลำดับคะแนนดังนี้

2.1. ก่อความรำคาญและทำให้ใช้งานไม่สะดวก จำนวน 357 คะแนน

2.2. ทำให้แบตเตอร์รี่หมดเร็วขึ้น จำนวน 101 คะแนน

2.3. ถูกละเมิดสิทธิส่วนบุคคล จำนวน 97 คะแนน

2.4. เสียพื้นที่ในการเก็บข้อความที่จำเป็น (Inbox เต็ม) จำนวน 66 คะแนน

4.1.2 สรุปนิยามของข้อความสแปม

จากการทำแบบสำรวจความคิดเห็นพบว่า ผู้ใช้งานโทรศัพท์เคลื่อนที่ในประเทศไทยส่วนใหญ่ ตัดสินข้อความในลักษณะโฆษณาขายสินค้า ข่าวสารทั่วไปที่ผู้ให้บริการโทรศัพท์เคลื่อนที่แจ้งเตือน ข้อความที่ไม่ทราบที่มา และข้อความที่มีความหมายใกล้เคียงกันที่ส่งหลายครั้งว่าเป็นข้อความสแปมซึ่งแตกต่างจากนิยามที่กำหนดไว้ในงานวิจัย LOHIT[1] ที่มุ่งเน้นเฉพาะข้อความที่โฆษณาขายสินค้าหรือการชักชวนให้ใช้บริการพิเศษซึ่งมีอัตราค่าบริการสูงกว่าปรกติ และใช้ข้อมูลดังกล่าวอ้างอิงการคัดแยกข้อความที่รวบรวมจาก SMSC

นิยามของข้อความสแปมมีดังนี้

1) ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ

---

[1] S. Dixit, S. Gupta, and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam.** P. 1.

2) ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่บางข้อความ

3) ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้

นอกจากนิยามของข้อความสแปมที่กล่าวถึงข้างต้นแล้ว ลักษณะเฉพาะของข้อความสแปมในประเทศไทยอีกประการคือ ข้อความจะมีลักษณะการเขียนที่ถูกต้องตามหลักภาษาศาสตร์เป็นส่วนใหญ่ ซึ่งแตกต่างจากลักษณะของข้อความภาษาอังกฤษ ในงานวิจัยต่างประเทศที่ใช้หลักการพ้องรูปและพ้องเสียงในการเขียนข้อความเพื่อหลีกเลี่ยงระบบกรองข้อความอีกด้วย

ข้อความตัวอย่างที่ผ่านการคัดแยกด้วยมนุษย์แสดงรายละเอียดดังตารางที่ 4.4

ตารางที่ 4.4 แสดงข้อความปรกติและข้อความสแปมที่ผ่านการคัดแยกจากนิยาม

ข้อความปรกติ	ข้อความสแปม
ถ้าไปทำได้ ตอนนี้อยากหายตัวไปนอนข้างๆ เธอในคืนนี้ เหงาจัง	2,350บเครื่อง+SimCATCDMAคุยทั้งวันทั้งปีทุกเครือข่าย149บ/ดที่mshopศก
มองโลกในแง่ดี ซิคะ,	www.pec9.comเปิดคอร์สสอนสดติวเข้มชั้น ป.4-ม.6 เริ่มเรียน 1 มิ.ย. 51
สุขสันต์วันเกิด ขอให้มีความสุขมากๆนะ นึกอะไรก็ขอให้ได้ดั่งใจนึก	ลุ้นดูดวงฟรีกับปู โลกเบี้ยวแค่รับดวงรายวันกด*298*193#(4บ/วัน)022076888
ฉุกเฉิน ค่าโทรหมด! รบกวนโทรกลับด้วยที่ 084XXXXXXX	สุดคุ้ม!ใช้ฟรี15วัน โมโนลูกทุ่งฮิต แถมผลหวย โทร* 45223111110/027302424
ไม่เคยกลัวคำขู่	ด่วนเซนต์เทเรซารับสมัครน.ศ.พยาบาลอีก10คนสุดท้ายโทร 0867227493/037395313
เอากระโปรงผู้หญิงไปใส่ซะไป	ส่งความรักด้วยเสียงผ่านVoiceSMSฟรี ถึง31มี.ค.51 โทร50100(เฉพาะภูมิภาค)
เทอเป็นคนเดียวที่ทำร้ายจิตใจฉันมากที่สุด ถ้าฉันตายเทอคงดีจัยนะ	เชิญแข่งลดน้ำหนักชิงทองและครอสไม่จำกัดจำนวนยูนิเซ็นตปิ่นเกล้า028848692
ไม่คิดถึงเค้าแล้วหรอ?	สมัครสมาชิกทรูมูฟซูปเปอร์สตาร์ภายในวันที่ 20 พ.ค. นี้ รับฟรี!
การที่เรารักใครสักคนบางครั้งเราก็ไม่ต้องการให้เขารักเราตอบขอแค่ห่วงใย	ฟรีกระเป๋ามูลค่า350บ.ที่บูธTMBMoneyExpo 8-11พค.51_โชว์บัตร Ready Cash
มองในแง่ดีเสมอแต่พระเจ้าซิ!กำลังไม่ให้โอกาส!	ร่วมเล่นเกมส์ค้นหาสมบัติ พร้อลุ้นรับตุ๊กตา Mickey & Minnie Mouse Big S
ขี้เกียดทะเลาะกับผู้หญิงปัญญาอ่อน	ดูภาพมันส์ๆจาก100 Rock Uncensored BKK และลุ้นรับ CD พร้อมลายเซ็นต์จ
กำลังจะนอนค่ะ	ดวงคุณเดือนพฤษภาจะมีโชคลาภหรือไม่?ปูโลกเบี้ยวมีคำตอบกด*4988แม่นมากๆ
ไหงงั้น?	คุณมากกว่าใคร ดอกเบี้ย 0% 1 ปี ฟรีโอน+จดจำนอง โทร 1375
เค้าคนนั้นก็พอใจแล้วฝันดีนะคับหมูขาวของชัย	ด่วน!มีจำกัดโมบายดูทีวี-ใส่Sim2ระบบลูกเล่นครบเพียง8,900ที่M-Shopศก.
Sorry. Good night.	เจาะลึก!แม่นยำ!ดวงคุณสัปดาห์นี้จะเป็นอย่างไรอ.ลักษณ์ฟันธงโทร1900190065

4.2 ลักษณะของข้อความ SMS ในประเทศไทย

ลักษณะข้อความ SMS ของประเทศไทยหลังจากการเก็บข้อมูลเป็นระยะเวลา 8 เดือน พบว่า มีลักษณะของข้อความที่ไม่เป็นไปตามหลักภาษาศาสตร์ ทั้งภาษาไทยและภาษาอังกฤษจำนวนมาก เนื่องจาก SMS เป็นการสื่อสารที่ไม่จำเป็นต้องใช้ภาษาอย่างเป็นทางการ อีกทั้งข้อจำกัดของจำนวนตัวอักษรที่พิมพ์ได้ในข้อความ นอกจากนี้ยังพบคำที่พิมพ์ไม่สมบูรณ์ในตอนต้นและท้ายข้อความ ซึ่งเกิดจากการส่งข้อความที่มีเนื้อความยาวเกินกว่าขนาดของ SMS ผ่านเครื่องโทรศัพท์เคลื่อนที่หรือโปรแกรมส่งข้อความที่ตัดข้อความไม่ถูกต้อง ดังตัวอย่างจากตารางที่ 4.5

ตารางที่ 4.5 แสดงลักษณะข้อความที่พบในประเทศไทย

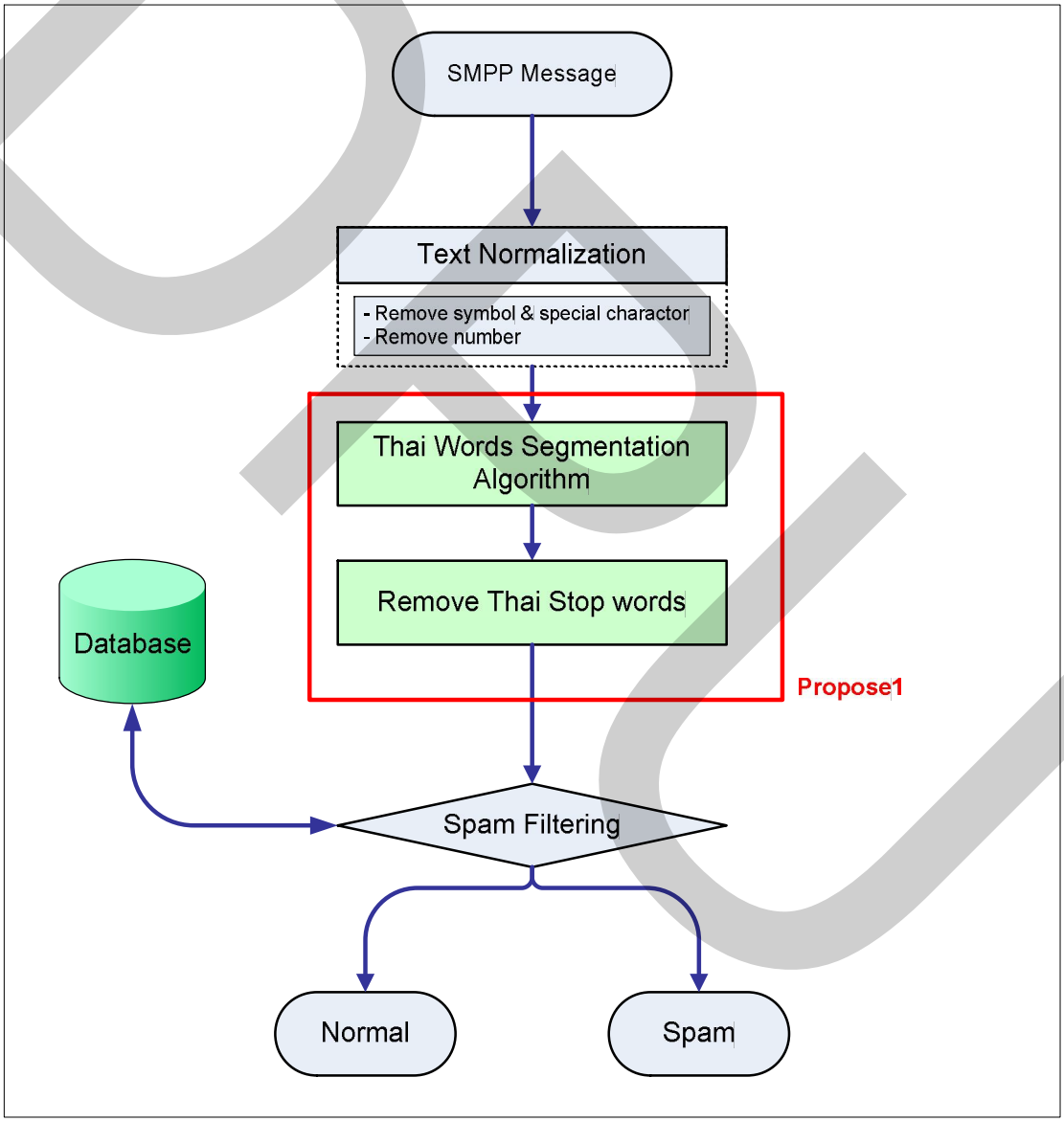
ข้อความที่ส่งทั้งหมด	ลำดับข้อความ	ข้อความที่ถูกแบ่ง
ช่วงนี้อากาศเปลี่ยนแปลงบ่อย ดูแลสุขภาพให้ดีนะ ด้วยความปรารถนาดีจากทิฟฟี่แผงสีเขียว ฮิๆ	1	ช่วงนี้อากาศเปลี่ยนแปลงบ่อย ดูแลสุขภาพให้ดีนะ ด้วยความปรารถนาดีจากท
	2	ิฟฟี่แผงสีเขียว ฮิๆ
คิดถึงเป็นคำสั้นๆดูเหมือนไม่มีความหมายแต่ก็ทำให้ผีบ้าอ่านแล้วยิ้มได้ก็แล้วกัน	1	คิดถึงเป็นคำสั้นๆดูเหมือนไม่มีความหมายแต่ก็ทำให้ผีบ้าอ่านแล้วยิ้มได
	2	้ก็แล้วกัน

4.3 การกรองข้อความที่นำเสนอแบบที่ 1 (Propose1)

เป็นวิธีการกรองข้อความ SMS ที่รองรับภาษาไทย ซึ่งทำการปรับเปลี่ยนเพิ่มเติมจากภาพที่ 2.3 ซึ่งรองรับการกรองภาษาอังกฤษเท่านั้น โดยใช้การทำ TN ที่ปรับปรุงใหม่ให้สามารถทำกระบวนการ TN และลบ Stop words กับภาษาไทย แล้วเพิ่มกระบวนการตัดคำให้สามารถรองรับข้อความ SMS ภาษาไทย และภาษาไทยปนภาษาอังกฤษ

ขั้นตอนการทำงานของการกรองข้อความที่นำเสนอแบบที่ 1 คือ รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message ให้อยู่ในรูปของ text จากนั้นทำกระบวนการ TN เพื่อลบ Character ที่ไม่สามารถตัดเป็นคำได้ออกไป เช่น @ # ! รวมถึงตัวเลข เป็นต้น เมื่อได้ข้อความที่มีความพร้อมแล้ว จะทำการส่งข้อความไปตัดคำภาษาไทยและภาษาอังกฤษ โดยจะลบคำ

ที่จัดอยู่ในประเภท Stop words ออกไป จากนั้นจะทำการ Mapping คำเข้ากับค่า TFIDF หรือค่า Spam Rate ตามอัลกอริทึมการกรองที่ใช้งานเพื่อสรุปผลของข้อความว่าจัดเป็นข้อความปรกติ หรือข้อความสแปม เมื่อได้ผลการกรองเป็นข้อความปรกติ ระบบจะทำการส่งข้อความไปยังผู้รับ หรือหากไม่ผ่านการกรองข้อความ ระบบจะละทิ้งข้อความนั้นโดยไม่ส่ง ซึ่งมีรายละเอียดการทำงานตามตามภาพที่ 4.1



ภาพที่ 4.1 ขั้นตอนการทำงานของ การกรองข้อความที่นำเสนอแบบที่ 1 ที่รองรับข้อความภาษาไทย

ในการประมวลผลข้อความ SMS เพื่อกรองข้อความสแปมจำเป็นต้องมีการทำ TN และการตัดคำ เพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผล โดยในการทดสอบที่ผ่านมา ใช้วิธีการทำ TN และการตัดคำแบบที่มีในงานวิจัย Document Summarization[1] และโปรแกรมตัดคำ[2] กล่าวถึงวิธีการตัดคำภาษาไทยและภาษาไทยปนภาษาอังกฤษที่ผ่านมา ใช้การตัดคำกับเอกสารประเภท หนังสือพิมพ์ วรรณกรรม หรือ เว็บไซต์ ที่ข้อความมีความถูกต้องตามหลักภาษาศาสตร์

เมื่อมีคำที่พิมพ์ผิดและคำที่พิมพ์ไม่สมบูรณ์เป็นจำนวนมาก การตัดคำและการหาความหมายของข้อความที่คลาดเคลื่อน ซึ่งส่งผลโดยตรงต่อการกรองข้อความ จึงจำเป็นต้องมีการปรับปรุงการทำ TN และการตัดคำ เพื่อให้รองรับกับลักษณะของข้อความ SMS

4.4 การกรองข้อความที่นำเสนอแบบที่ 2 (Propose2)

วิธีการที่นำเสนอนี้ สามารถประยุกต์ใช้กับการตัดคำและวิธีการกรองได้หลายแบบ ซึ่งมีรายละเอียดดังนี้

4.4.1 การแก้ไขตำแหน่งและการพิมพ์เกินของ สระ วรรณยุกต์ (vowel) และตัวอักษร

เพิ่มการแก้ไขตำแหน่งและการพิมพ์เกินของ สระ วรรณยุกต์ และตัวอักษรลงในขั้นตอนการทำ TN ซึ่งเป็นขั้นตอนการลบ สัญลักษณ์พิเศษ เช่น $ # @ ? ! หรือตัวเลขที่ไม่ต้องการเพื่อกำจัดข้อมูลส่วนเกินออกในระบบ Filter แต่เนื่องจากข้อความ SMS ที่สร้างขึ้นด้วยการพิมพ์จากเครื่องโทรศัพท์เคลื่อนที่ ทำให้พบการพิมพ์สระ วรรณยุกต์ และตัวอักษรตัวเดียวกันมากกว่า 1 ครั้ง หรือการพิมพ์สระและวรรณยุกต์สลับที่เป็นปริมาณมาก เช่น การพิมพ์คำว่า “อยู่” จะต้องพิมพ์ สระอู ก่อนพิมพ์ ไม้เอก เสมอ ซึ่งหลักการพิมพ์ภาษาไทยที่ถูกต้อง จะพิมพ์สระก่อนวรรณยุกต์ดัง ตารางที่ 4.6

ตารางที่ 4.6 สระที่ต้องพิมพ์ก่อนหน้าวรรณยุกต์

สระ	วรรณยุกต์
ั ิ ี ึ ื ุ ู ็ ฺ	่ ้ ๊ ๋

[1] อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย และ ชุลีรัตน์ จรัสกุลชัย. (2544). **Document Summarization.** หน้า 4-6

[2] วิรัช ศรเลิศล้ำวานิช. National Electronics and Computer Technology Center (NECTEC).(2543). โปรแกรมตัดคำภาษาไทย.

สำหรับ สระอำ (ำ) ต้องพิมพ์ตามหลังวรรณยุกต์ ได้แก่ ไม้เอก (่) ไม้โท (้) ไม้จัตวา (๊) เท่านั้น จึงต้องปรับปรุงการทำ TN ให้สามารถแก้ไขการพิมพ์สระหรือวรรณยุกต์ตัวเดียวกันมากกว่า 1 ครั้ง หรือสระและวรรณยุกต์ที่พิมพ์สลับลำดับทั้งข้อความ เพื่อลดปริมาณคำผิดและเพิ่มประสิทธิภาพในการตัดคำ

4.4.2 การตรวจสอบคำแรกและคำสุดท้ายของข้อความ

วิธีการตัดคำแบบการคำนวณเชิงสถิติ แม้จะมีประสิทธิภาพสูงเมื่อใช้งานกับเอกสารที่มีความถูกต้อง แต่เมื่อนำมาใช้ร่วมกับข้อความ SMS พบว่า มีความผิดพลาดมากกว่า การตัดคำแบบยาวที่สุด และการตัดคำแบบสอดคล้องมากที่สุด

สำหรับการตัดคำแบบสอดคล้องมากที่สุด จะตัดคำผิดพลาดเมื่อพบคำที่พิมพ์ไม่สมบูรณ์ โดยจะนำคำที่พิมพ์ไม่สมบูรณ์รวมเข้ากับคำถัดไปแล้วทำการตัด ทำให้คำถัดไปที่เป็นคำที่ถูกต้อง มีความหมายผิดไปจากเดิม ปัญหาดังกล่าวนี้ สามารถแก้ไขได้ด้วยวิธีการนำคำที่พิมพ์ไม่สมบูรณ์และคำถูก ที่การตัดคำแบบสอดคล้องไม่สามารถตัดได้ มาผ่านกระบวนการตัดคำอีกครั้งด้วยวิธีการคำนวณเชิงสถิติ และเนื่องจากคำที่พิมพ์ไม่สมบูรณ์ สามารถตรวจพบได้ที่คำแรกสุดของข้อความที่เป็นข้อความต่อเนื่อง ตามตัวอย่างในตารางที่ 5.1 ข้อความที่ 2 จึงสามารถตรวจสอบคำผิดตามลักษณะดังกล่าวได้ง่าย โดยแสดงผลการตัดคำแบบผสมดังตารางที่ 4.7

ตารางที่ 4.7 เปรียบเทียบวิธีตัดคำ

แบบสอดคล้อง	**แบบผสม**
้ ก็ | แล้ว | กัน	**ก็** | แล้ว | กัน

4.4.3 การตรวจสอบรูปแบบของเลขหมายพิเศษ

นอกจากการตรวจสอบคำผิดและการปรับปรุงโดยใช้การตัดคำแบบผสมแล้ว ยังมีการตรวจพบข้อความสแปมที่สามารถผ่านการกรองโดยมีลักษณะข้อความที่แสดงข้อมูลเลขหมายพิเศษที่มีอัตราค่าบริการสูงกว่าปรกติ (Premium rate Number) ซึ่งเกิดขึ้นจากกรณีที่ข้อความสแปมมีเนื้อความยาวเกินกว่า 1 ข้อความ โดยเมื่อข้อความถูกแบ่งเนื้อความออกเป็น 2 ข้อวาม ใจความในส่วนของการแสดงข้อมูลเลขหมายพิเศษของข้อความโฆษณาเชิญชวน อาจถูกแยกจากเนื้อความที่แสดงความหมายเป็นสแปมทำให้เนื้อความส่วนนี้ ไม่มีคำใดๆที่แสดงความหมายเป็นข้อความสแปมและสามารถผ่านการกรอง ซึ่งจากการเก็บข้อมูล SMS ทำให้ทราบว่า เลขหมายพิเศษ จะมี

เครื่องหมาย # หรือ * ประกอบในตัวเลข ทั้งก่อนหน้าชุดตัวเลข ในระหว่างชุดตัวเลข และต่อท้ายชุดตัวเลข และเลขหมายพิเศษที่ขึ้นต้นด้วย 1900 ดังมีตัวอย่างตามตารางที่ 4.4 เป็นต้น

การออกแบบวิธีการกรองข้อความภาษาไทยในงานวิจัยนี้ จึงได้เพิ่มการตรวจสอบรูปแบบของเลขหมายพิเศษ เพื่อให้การกรองข้อความ SMS มีความถูกต้องมากขึ้น ด้วยเงื่อนไขของการมีเครื่องหมาย # หรือ * ประกอบในตัวเลข ทั้งก่อนหน้าชุดตัวเลข ในระหว่างชุดตัวเลข และต่อท้ายชุดตัวเลข และเลขหมายพิเศษที่ขึ้นต้นด้วย 1900 ดังนี้

1) ตรวจพบคำว่า “โทร” หรือ “กด” และมีเครื่องหมาย # หรือ * แล้วตามด้วย ชุดตัวเลข

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

3) ตรวจพบหมายเลข 1900 แล้วตามด้วยชุดตัวเลข 6 หลัก

4.4.4 ปรับปรุงลำดับการกรองภาษาไทย

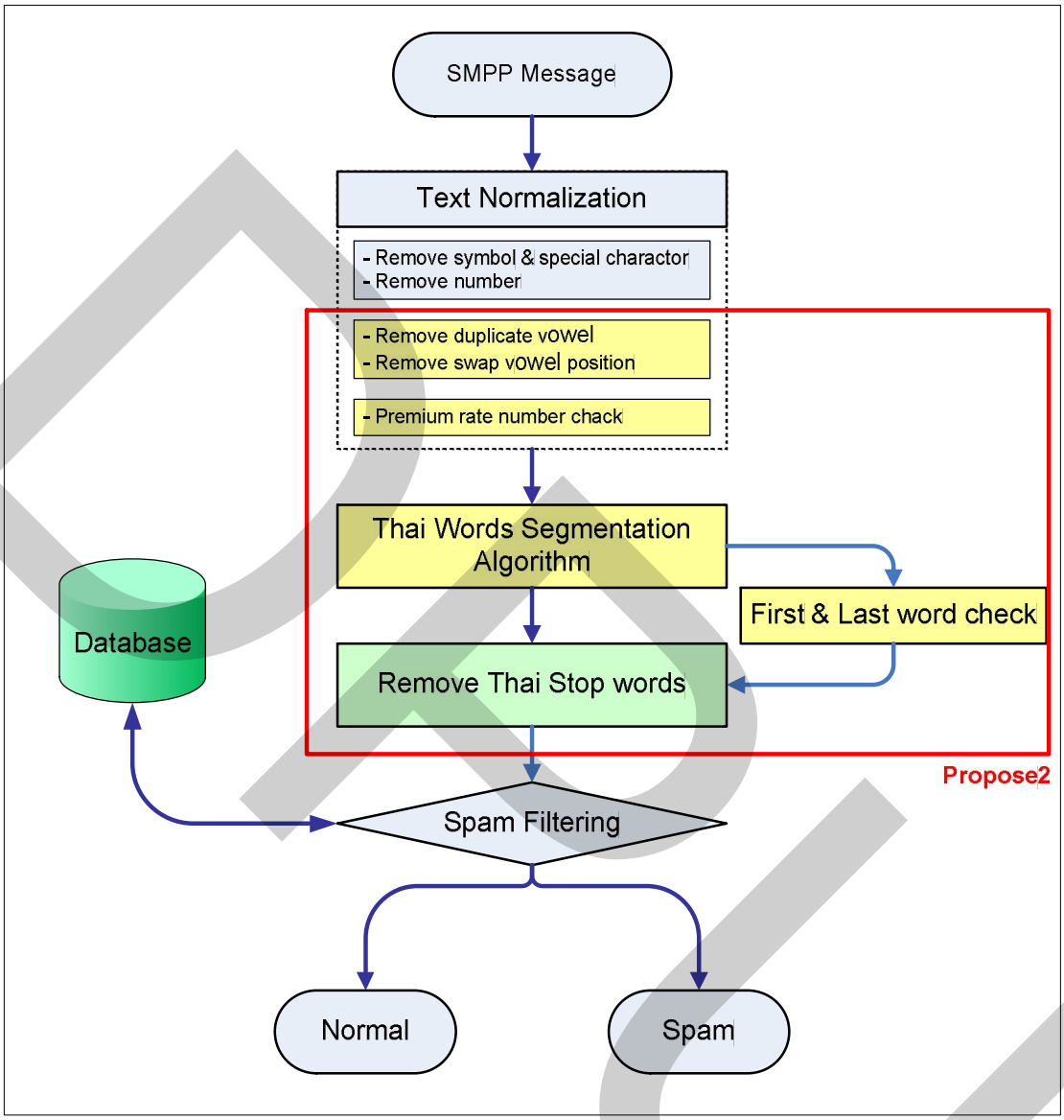
ทำการปรับเปลี่ยนเพิ่มเติมจากภาพที่ 4.1 ดังนี้

4.4.4.1 ปรับรุงการทำ TN จากเดิมที่ทำหน้าที่ลบเครื่องหมายพิเศษต่างๆออกเพียงอย่างเดียว ให้สามารถแก้ไขการพิมพ์ภาษาไทยที่ไม่ตรงตามหลักภาษาศาสตร์ให้ถูกต้อง

4.4.4.2 เพิ่มขั้นตอนการตรวจสอบหมายเลขพิเศษที่มีค่าบริการสูงกว่าปรกติซึ่งมีโอกาสเป็นข้อความสแปมในการทำ TN

4.4.4.3 ปรับปรุงการตัดคำ โดยเพิ่มเติมการตรวจสอบคำแรกและคำสุดท้าย เพื่อค้นหาคำที่พิมพ์ไม่สมบูรณ์

ขั้นตอนการทำงานของการกรองข้อความที่นำเสนอแบบที่ 2 คือ รับข้อความจากชุมสาย (BTS/MSC) และ ถอดข้อความจาก SMPP Message ให้อยู่ในรูปของ text จากนั้นทำกระบวนการ TN เพื่อลบ Character ที่ไม่สามารถตัดเป็นคำได้ออกไป เช่น @ # ! ตัวเลข การลบ Character ที่มีการซ้ำกันมากกว่า 3 ตัวขึ้นไป การพิมพ์สระและวรรณยุกต์สลับตำแหน่งกัน และการตรวจสอบหมายเลขพิเศษที่มีการคิดค่าบริการสูงกว่าปรกติ เมื่อได้ข้อความที่มีความพร้อมแล้ว จะทำการส่งข้อความไปตัดคำภาษาไทยและภาษาอังกฤษ โดยจะลบคำที่จัดอยู่ในประเภท Stop words ออกไป และทำการตัดคำครั้งที่ 2 กับคำแรกและคำสุดท้ายของข้อความเพื่อเพิ่มความถูกต้องในการตัดคำที่ไม่สมบูรณ์ จากนั้นจะทำการ Mapping คำเข้ากับค่า TFIDF หรือค่า Spam Rate ตามอัลกอริทึมการกรองเพื่อสรุปผลของข้อความว่าจัดเป็นข้อความปรกติ หรือข้อความสแปม เมื่อได้ผลการกรองเป็นข้อความปรกติ ระบบจะทำการส่งข้อความไปยังผู้รับ หรือหากไม่ผ่านการกรองข้อความ ระบบจะละทิ้งข้อความนั้นโดยไม่ส่ง ซึ่งมีรายละเอียดการทำงานตามตามภาพที่ 4.2



ภาพที่ 4.2 ขั้นตอนการทำงานของ การกรองข้อความที่นำเสนอแบบที่ 2

4.5 การเตรียมข้อมูลทดสอบ

จากการวิเคราะห์ข้อมูล SMS และการสำรวจความคิดเห็นในประเทศไทย สามารถแบ่งการกำหนดชุดข้อมูลออกได้เป็น 2 ชุด ได้แก่ ชุดข้อมูลสำหรับฝึกสอน (TD) จำนวน 1250 ข้อความ ประกอบด้วยข้อความปรกติจำนวน 652 ข้อความ ข้อความสแปมจำนวน 598 ข้อความ และชุดข้อมูลใหม่สำหรับทดสอบ (ND) จำนวน 1336 ข้อความประกอบด้วยข้อความปรกติจำนวน 1055 ข้อความ ข้อความสแปมจำนวน 281 ข้อความ ซึ่งข้อความ SMS ชุดข้อมูล TD จะนำมาใช้ในการ

ฝึกสอน Filter ให้สามารถคัดแยกข้อความ SMS ในประเทศไทยได้อย่างถูกต้อง อีกทั้งข้อความ SMS ชุดข้อมูล TD จะนำมาใช้ทดสอบเพื่อวัดประสิทธิภาพในการกรองข้อความ

กระบวนการฝึกสอนกระทำโดยการป้อนชุดข้อมูล TD เข้าสู่ Filter โดยระบุความหมายของข้อมูลแต่ละตัว เพื่อให้ Filter จดจำรูปแบบของข้อความ SMS แบบสแปมและข้อความ SMS แบบปรกติ แล้วทำการทดสอบด้วยชุดข้อมูล TD เดิม และชุดข้อมูล ND เพื่อวัดประสิทธิภาพทั้งทางเวลาและความถูกต้อง

4.6 การเขียนโปรแกรมจำลอง

4.6.1 การคำนวณหาลักษณะแทนข้อความด้วยวิธีการ TFIDF

จากสมการที่ 2 ใช้การคำนวณค่าน้ำหนักของข้อมูล TD จำเป็นต้องทราบค่า TF และค่า DF โดยมี Pseudocode ที่ใช้ในการหาค่าดังต่อไปนี้

```
$id = 0;
for($i=0;$i<count($TD_data);$i++) {
        $sms_ary = word_segment(normalize($TD_data[$i]),$algorithm); // text normalization and word segmentation
        $data_DB[$i] = $sms_ary;
        for($i=0;$j<count($sms_ary);$j++) {
                $word = $sms_ary[$j];
                if ($word != $stop_words) { // if not stop words
                        if ($word not in $db_ary) { // if new data
                                $db_ary[$word][0] = $id; // id of data
                                $db_ary[$word][1] = 1; // TF (count)
                                $db_ary[$word][2] = 0; // DF (count)
                                $id++;
                        }else{ // if existing data
                                $db_ary[$word][1] += 1; // +TF
                        }
                }
        }
```

```
        // add DF
        $unique = array_unique($sms_ary); // remove duplicate words
        foreach ($unique as $word) {
                $db_ary[$word][2] += 1; // +DF (count)
        }
}
// function
normalize($data){
        return string_replace(array(symbol, number), "", $data);
}
word_segment($text,$algorithm,$spliter="|") {
        return system('cmd swath.exe -b "'.$spliter.'" '."-m ".$algorithm." -d data < ".$text." >");
}
```

ซึ่ง Psudocode ดังกล่าวจะคำนวณข้อมูลจากตัวแปร $TD_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 $data_DB โดยจะจัดเก็บเป็น Array 2 ชั้น รูปแบบ SMS_1{$word_1${id, TF, DF}, $word_2${id, TF, DF},..., $word_n${id, TF, DF}} และ SMS_1{...}, SMS_2{...},...,SMS_n{...} ในตัวแปร $db_ary และนำไปคำนวณหาค่า Feature Vector จาก Psudocode ดังต่อไปนี้

```
for($i=0;$i<count($data_DB);$i++) {
        $mod = 0;
        for($i=0;$j<count($data_DB[$i]);$j++) {
                $word = $data_DB[$i][$j];
                $TF = $db_ary[$word][1]
                $DF = $db_ary[$word][2];
                $IDF = log10($N/$DF);
                $mod += ($TF*$TF)*($IDF*$IDF);
```

```
        }
        $mod = sqrt($mod);
        for($i=0;$i<count($data_DB[$i]);$i++) {
                $word = $data_DB[$i][$j];
                $TF = $db_ary[$word][1]
                $DF = $db_ary[$word][2];
                $IDF = log10($N/$DF);
                $feature_vector = ($TF*$IDF)/$mod;
                $db_ary[$word][3] = $feature_vector; // feature vector (FV)
        }
}
```

ข้อมูลที่ได้จาก $db_ary ที่มีรูปแบบ SMS_1{word_1{id, TF, DF, FV}, word_2{id, TF, DF, FV},..., word_n{id, TF, DF, FV}} และ SMS_1{...}, SMS_2{...},...,SMS_n{...} จะสามารถระบุ Feature Vector ของคำแต่ละคำเพื่อนำไปใช้ประมวลผลกับอัลกอริทึมกรองข้อความแบบ SVM และ NB ต่อไป

4.6.2 การเขียนโปรแกรมกรองข้อความที่นำเสนอแบบที่ 1 (Propose1 Programing)

การกรองข้อความที่นำเสนอแบบที่ 1 มีการเขียนโปรแกรมดัง Psudocode ต่อไปนี้

```
$SMS_vector = "";
$sms_ary = word_segment(normalize($SMS_data),$algorithm); // text normalization and word segmentation
for($i=0;$i<count($sms_ary);$i++) {
        $word = $sms_ary[$i];
        if ($word != $stop_words) { // if not stop words
          $SMS_vector .= $db_ary[$word][0] . ":" . $db_ary[$word][3] + " "; // Insert feature vector
        }
}
$result = spam_filter_algorithm($SMS_vector); // NB or SVM
```

```
normalize($data){
        return string_replace(array(symbol, number), "", $data);
}

word_segment($text,$algorithm,$spliter="|") {
        return system('cmd swath.exe -b "'.$spliter.'" '."-m ".$algorithm." -d data < ".$text." >");
}
```

ข้อความ SMS จะถูกแทนด้วยตัวแปร $SMS_data และผ่านการทำ TN ก่อนการตัดคำ จากนั้นจึงเข้าสู่กระบวนการกรองข้อความด้วย อัลกอริทึม SVM หรือ NB เพื่อระบุลักษณะของข้อความต่อไป

4.6.3 การเขียนโปรแกรมกรองข้อความที่นำเสนอแบบที่ 2 (Propose2 Programing)

การกรองข้อความที่นำเสนอแบบที่ 2 มีการเขียนโปรแกรมดัง Psudocode ต่อไปนี้

```
$SMS_vector = "";
$sms_ary = new_word_segment(new_normalize($SMS_data)); // text normalization and word segmentation
for($i=0;$i<count($sms_ary);$i++) {
        $word = $sms_ary[$i];
        if ($word != $stop_words) { // if not stop words
                $SMS_vector .= $db_ary[$word][0] . ":" . $db_ary[$word][3] + " "; // Insert feature vector
        }
}
$result = spam_filter_algorithm($SMS_vector); // NB or SVM
// function
new_normalize($data){
        $new_data = string_remove_dup(dup_char, char, $data); // remove duplicate characters
```

```
        $new_data = string_swap(array(vowelB, vowelA), $data); // correct vowel
        $new_data = string_convert_to_lower($data); // Convert Eng character to lowercase
        $new_data = string_mapping(phone_number, uppercase , $data); // Convert phone number to uppercase
        $new_data = string_replace(array(symbol, number), "", $data); // remove symbol and number
        return $newdata;
}
word_segment($text,$spliter="|") {
        $words = system('cmd swath.exe -b "'.$spliter.'" '."-m ".$algorithm." -d data < ".$text." >");
        $words[0] = system('cmd swath.exe -b "'.$spliter.'" '."-m max -d data < ".$words[0]." >");
        $words[last] = system('cmd swath.exe -b "'.$spliter.'" '."-m max -d data < ".$words[last]." >");
        return $words;
}
```

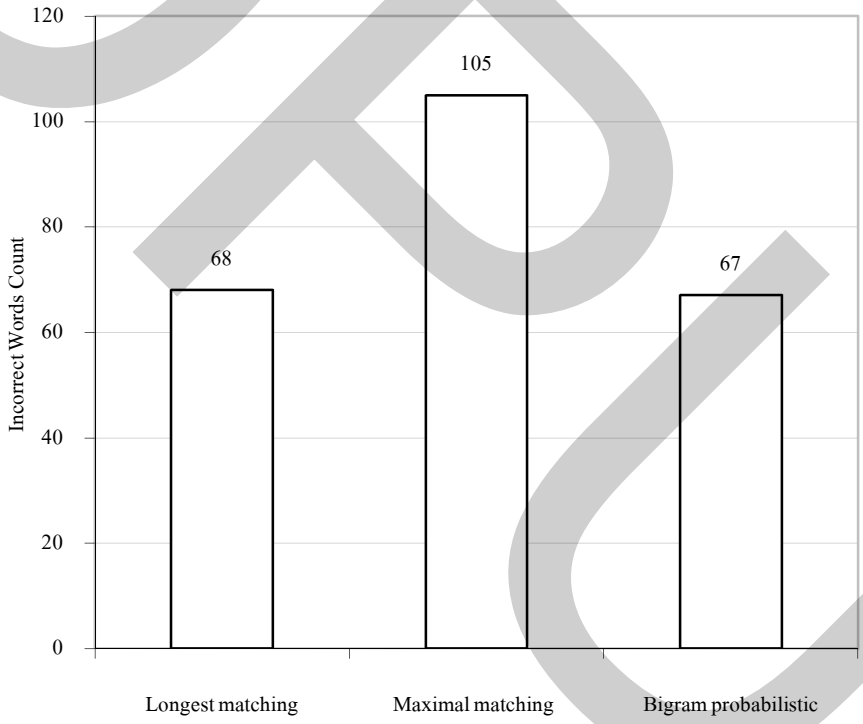
การกรองข้อความที่นำเสนอแบบที่ 2 นี้ จะเพิ่มขั้นตอนในส่วนของการทำ TN การแก้ไขสระและวรรณยุกต์ที่ผิดตำแหน่ง การลดปริมาณตัวอักษรที่พิมพ์เกินกว่า 1 ครั้ง การตรวจสอบหมายเลขโทรศัพท์พิเศษ และการตัดคำแรกและคำสุดท้ายของข้อความ เพื่อเพิ่มความถูกต้องในการนำคำไประบุลักษณะต่อไป

บทที่ 5
ผลการวิจัย

5.1 การวัดประสิทธิภาพ

5.1.1 ผลการลดจำนวนคำที่ไม่ถูกต้องในการตัดคำ

จากการปรับปรุงการทำ TN และเพิ่มเติมการตรวจสอบคำที่ไม่สมบูรณ์ ในการกรองข้อความที่นำเสนอแบบที่ 2 สามารถลดจำนวนคำที่ไม่ถูกต้องในการตัดคำลงได้เมื่อเทียบกับวิธีการที่ 1 โดยแสดงผลการทดสอบตัดคำกับข้อความ SMS จำนวน 18145 ข้อความ ดังภาพที่ 5.1



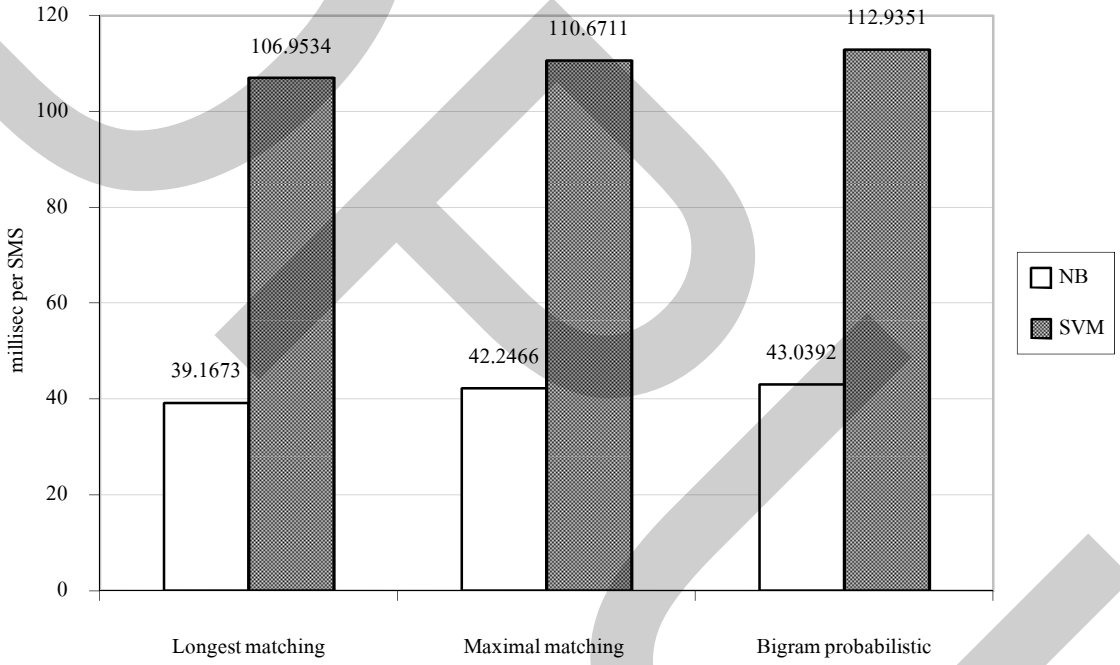
ภาพที่ 5.1 การตรวจสอบคำผิดในขั้นตอน TN ระหว่างการตัดคำแบบต่างๆ ในการกรองข้อความที่นำเสนอแบบที่ 2

จากภาพที่ 5.1 แสดงผลการทำ TN และการตัดคำแบบผสมกับชุดข้อมูลฝึกสอน ซึ่งการกรองข้อความที่นำเสนอแบบที่ 2 สามารถตรวจจับคำที่ไม่ถูกต้องในการตัดแบบสอดคล้องมากที่สุดลงได้ 105 คำ เมื่อเปรียบเทียบกับการกรองข้อความที่นำเสนอแบบที่ 1 เพราะได้รับผลจากการตัดคำ

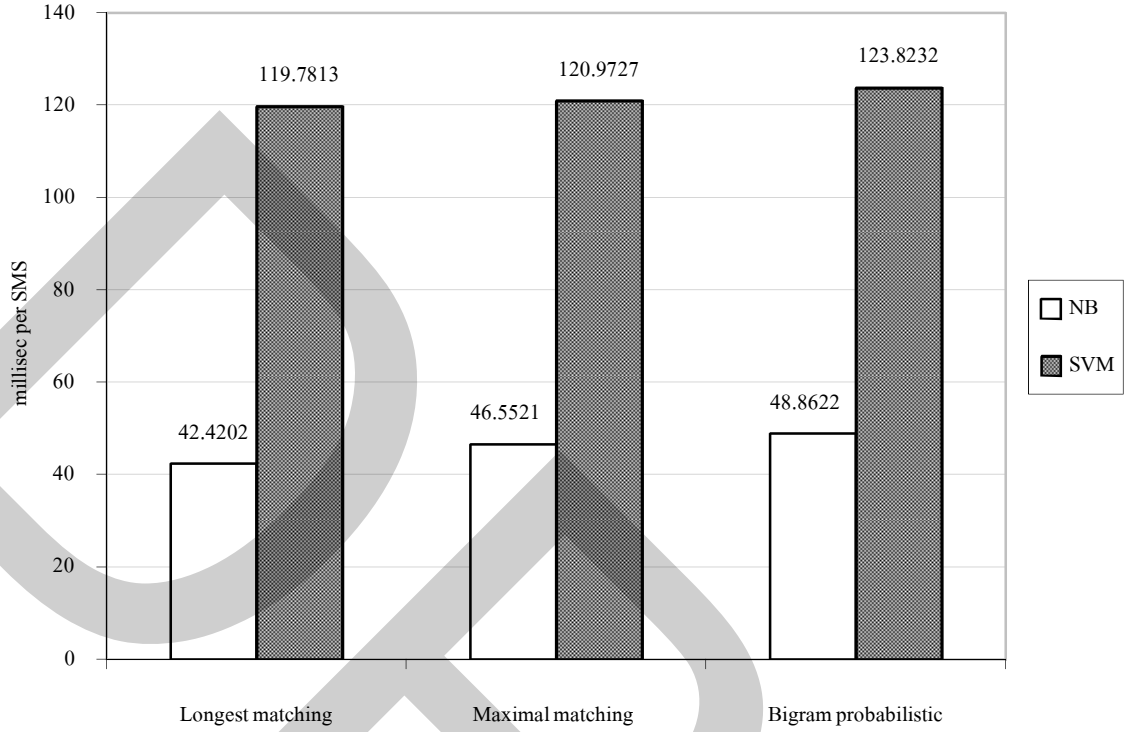
แบบผสมกับคำแรกและคำสุดท้ายของข้อความที่การตัดคำแบบนี้ มีการรวมคำที่พิมพ์ไม่ครบเข้ากับคำถูก ในขณะที่การตัดคำแบบอื่นจะได้รับผลจากการแก้ไขลำดับการพิมพ์สระและวรรณยุกต์ผิดหรือการพิมพ์สระและวรรณยุกต์มากกว่า 1 ครั้ง โดยจำนวนคำที่ได้ตรวจพบนี้เป็นคำผิดที่จะส่งผลต่อความเร็วการกรองข้อความในการทดสอบหัวข้อถัดไป

5.1.2 ผลการเปรียบเทียบระหว่างอัลกอริทึม NB และ SVM

เป็นการวัดประสิทธิภาพความถูกต้องโดยใช้การกรองข้อความที่นำเสนอแบบที่ 1 ร่วมกับอัลกอริทึมแบบ NB เปรียบเทียบกับอัลกอริทึมแบบ SVM

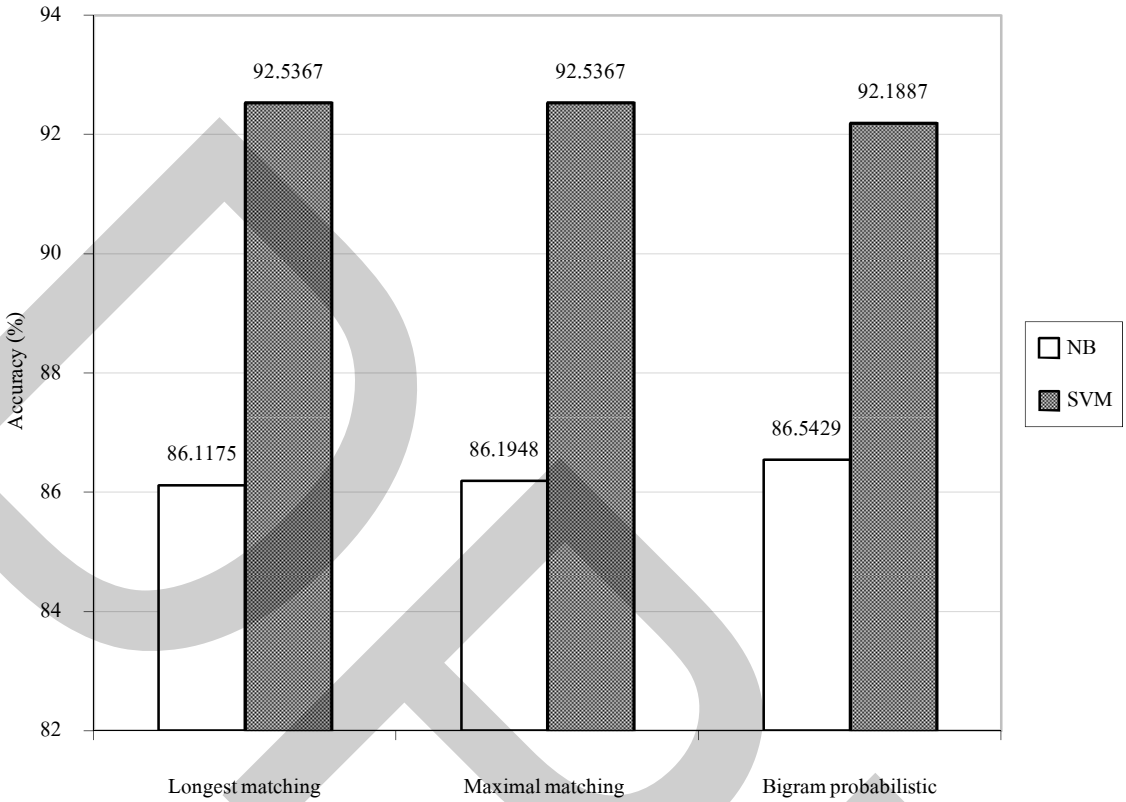


ภาพที่ 5.2 เวลาประมวลผลระหว่างอัลกอริทึม NB และ SVM การกรองข้อความที่นำเสนอแบบที่ 1 จากชุดข้อมูลที่ใช้ฝึกสอน (TD)



ภาพที่ 5.3 เวลาประมวลผลระหว่างอัลกอริทึม NB และ SVM การกรองข้อความที่นำเสนอแบบที่ 1 จากชุดข้อมูลใหม่ (ND)

จากภาพที่ 5.2 และ 5.3 การทดสอบกรองข้อความด้วยวิธีการตัดคำภาษาไทยทั้ง 3 แบบ แสดงให้เห็นว่า ประสิทธิภาพทางเวลาของอัลกอริทึมแบบ NB จะใช้เวลาในการประมวลผลน้อยกว่าอัลกอริทึมแบบ SVM ในทุกการทดสอบ



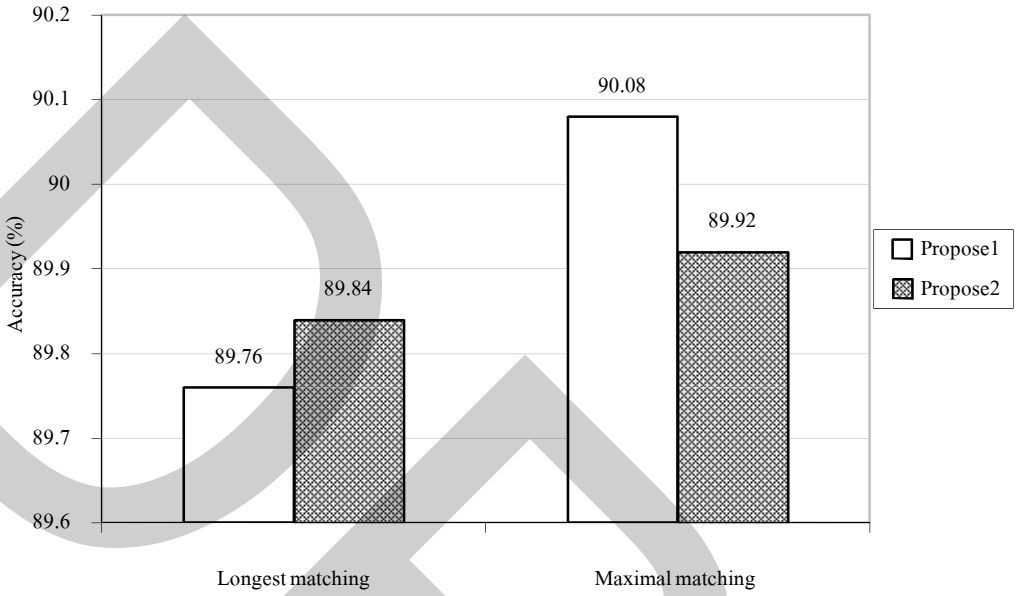
ภาพที่ 5.4 ประสิทธิภาพความถูกต้องระหว่าง NB และ SVM การกรองข้อความที่นำเสนอแบบที่ 1

จากภาพที่ 5.4 ผลการทดสอบความถูกต้องในการกรองข้อความด้วยวิธีการตัดคำภาษาไทยทั้ง 3 แบบแสดงให้เห็นว่า ประสิทธิภาพของวิธีการกรองแบบ SVM มีความถูกต้องเฉลี่ยที่ร้อยละ 92.42 ในขณะที่การกรองแบบ NB มีความถูกต้องเฉลี่ยที่ร้อยละ 86.28 ซึ่งวิธีการแบบ SVM ให้ผลการกรองที่ดีกว่าวิธีการแบบ NB อยู่ประมาณร้อยละ 6.14

นอกจากนี้วิธีการตัดคำแบบต่างๆยังส่งผลต่อความถูกต้องของการกรอง โดยวิธีการตัดคำแบบยาวที่สุด และวิธีการตัดคำแบบสอดคล้องมากที่สุด มีความถูกต้องมากกว่าวิธีการแบบสถิติในการกรองด้วยอัลกอริทึม SVM

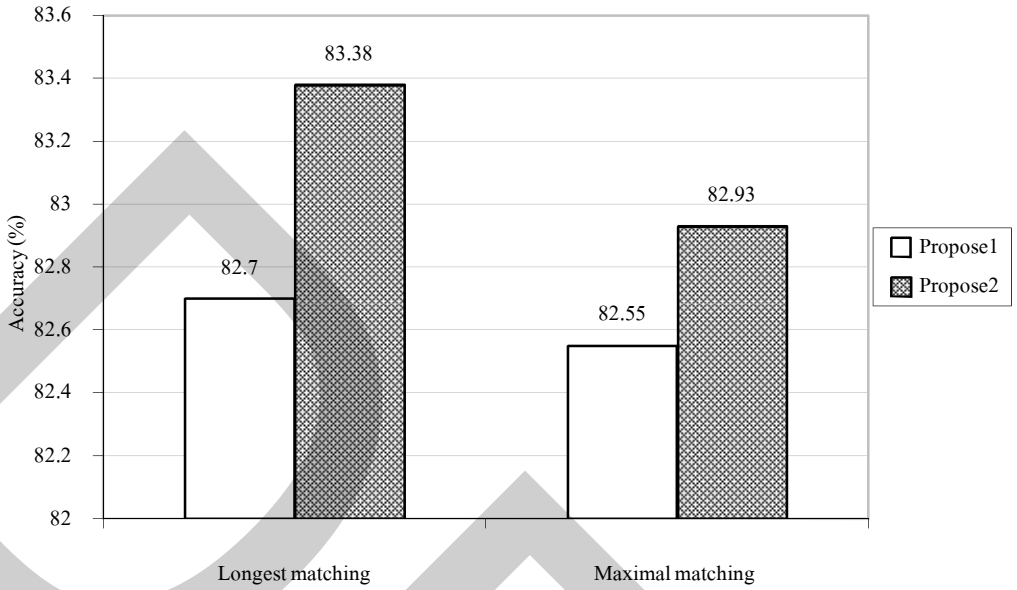
5.1.3 ผลการเปรียบเทียบอัลกอริทึม NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

การวัดประสิทธิภาพความถูกต้องในการกรองข้อความระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 โดยทดสอบด้วยอัลกอริทึม NB และการตัดคำ 2 แบบได้แก่ การตัดคำแบบยาวที่สุด และแบบสอดคล้องที่สุด ตามภาพที่ 5.5 และ 5.6



ภาพที่ 5.5 ความถูกต้องของอัลกอริทึม NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลที่ใช้ฝึกสอน (TD)

จากภาพที่ 5.5 แสดงให้เห็นว่าการกรองข้อความที่นำเสนอแบบที่ 2 ไม่สามารถเพิ่มความถูกต้องของอัลกอริทึม NB ในการกรองข้อมูล TD ได้สูงขึ้น โดยในการตัดคำแบบสอดคล้องที่สุดมีความถูกต้องลดลงเล็กน้อย

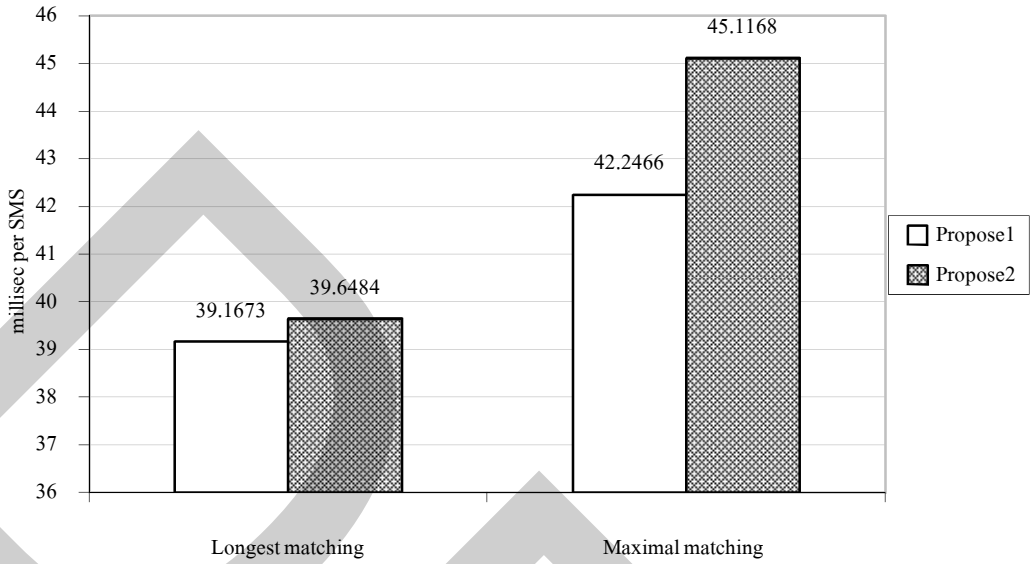


ภาพที่ 5.6 ความถูกต้องของอัลกอริทึม NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลใหม่ (ND)

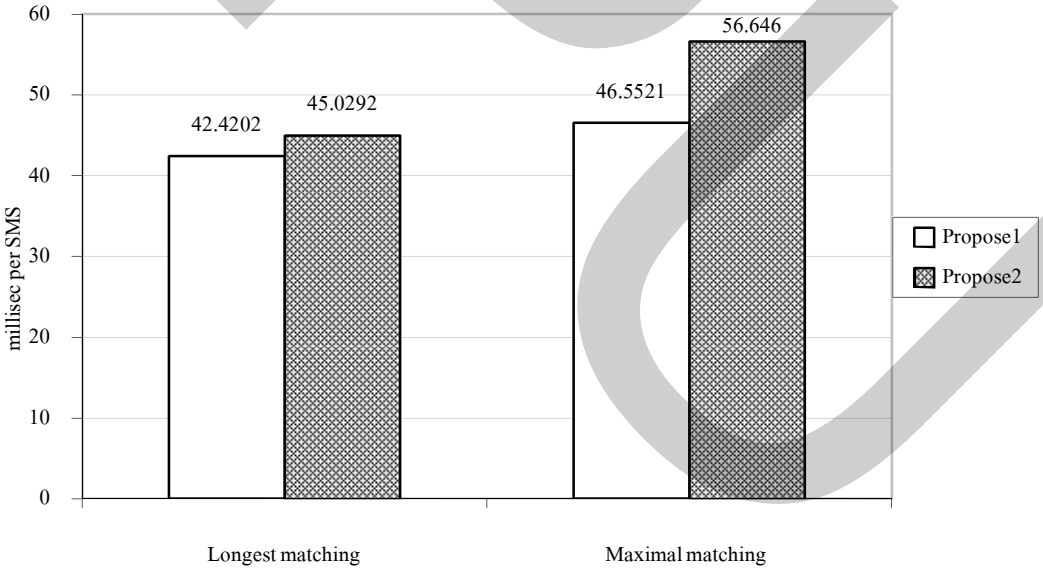
จากภาพที่ 5.6 แสดงให้เห็นว่าการกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มความถูกต้องของอัลกอริทึม NB ในการกรองข้อมูล ND ได้สูงขึ้นเล็กน้อย และความถูกต้องในการกรองข้อความของข้อมูลชุด TD และ ND โดยรวมสูงขึ้นน้อยมาก

จากผลการเปรียบเทียบข้างต้นแสดงให้เห็นว่า การกรองข้อความที่นำเสนอแบบที่ 2 ไม่สามารถเพิ่มความถูกต้องในการกรองข้อความด้วยอัลกอริทึมแบบ NB ได้

การวัดประสิทธิภาพทางเวลาในการกรองข้อความระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 โดยทดสอบด้วยอัลกอริทึม NB และการตัดคำ 2 แบบได้แก่ การตัดคำแบบยาวที่สุด และแบบสอดคล้องที่สุด ตามภาพที่ 5.7 และ 5.8



ภาพที่ 5.7 เวลาในการกรองของอัลกอริทึม NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลที่ใช้ฝึกสอน (TD)

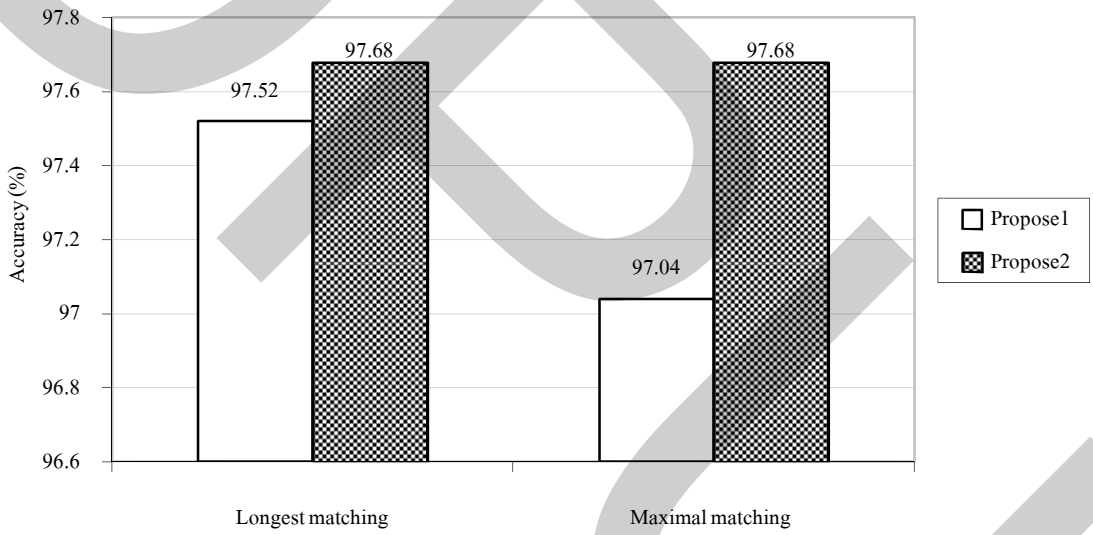


ภาพที่ 5.8 เวลาในการกรองของอัลกอริทึม NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลใหม่ (ND)

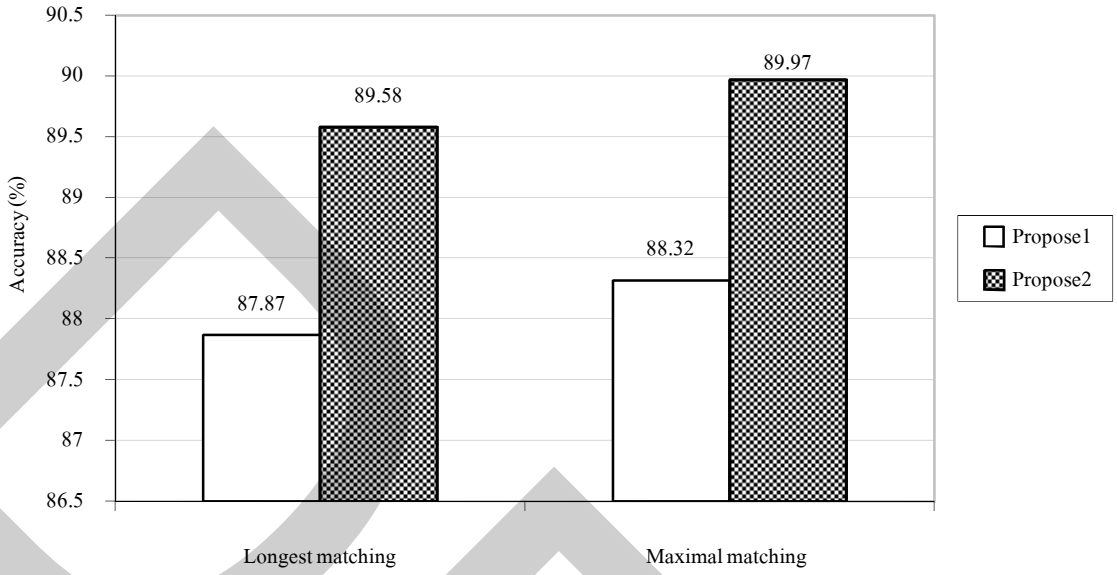
จากภาพที่ 5.7 และ 5.8 แสดงให้เห็นว่าการกรองข้อความที่นำเสนอแบบที่ 2 ใช้เวลาในการกรองมากกว่าการกรองข้อความที่นำเสนอแบบที่ 1 ทำให้สิ้นเปลืองเวลาในการประมวลข้อมูลมากขึ้น อีกทั้งความถูกต้องที่เพิ่มขึ้นเล็กน้อยทำให้การกรองข้อความที่นำเสนอแบบที่ 2 ไม่มีความเหมาะสมในการกรองข้อความด้วยอัลกอริทึมแบบ NB

5.1.4 ผลการเปรียบเทียบอัลกอริทึม SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

การวัดประสิทธิภาพความถูกต้องในการกรองข้อความระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 โดยทดสอบด้วยอัลกอริทึม SVM และการตัดคำ 2 แบบได้แก่ การตัดคำแบบยาวที่สุด และแบบสอดคล้องที่สุด ตามภาพที่ 5.9 และ 5.10



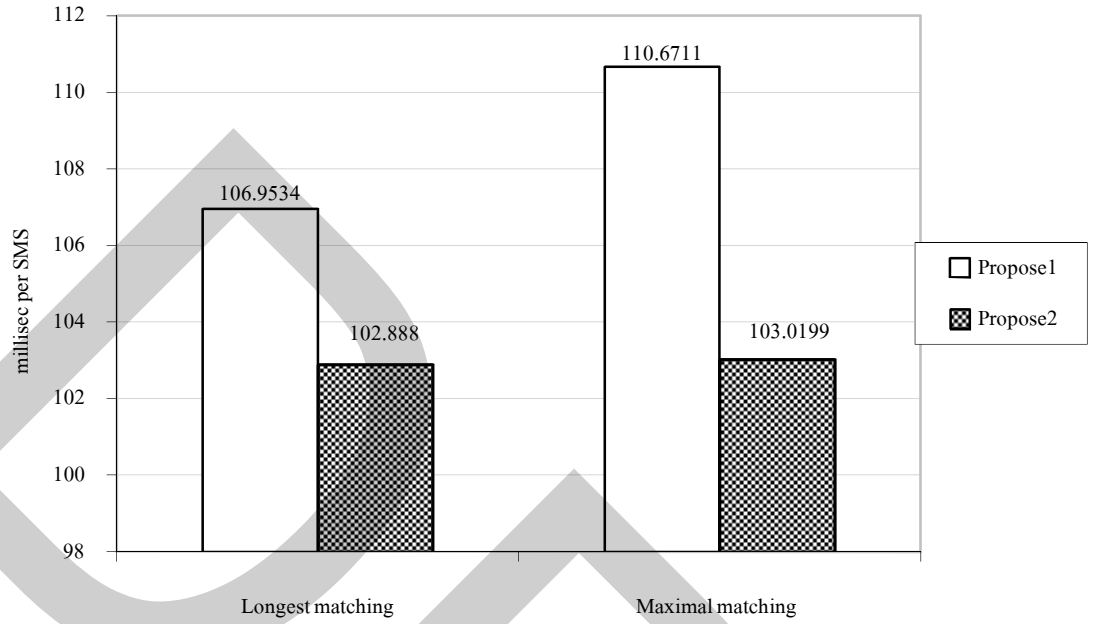
ภาพที่ 5.9 ความถูกต้องของอัลกอริทึม SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลที่ใช้ฝึกสอน (TD)



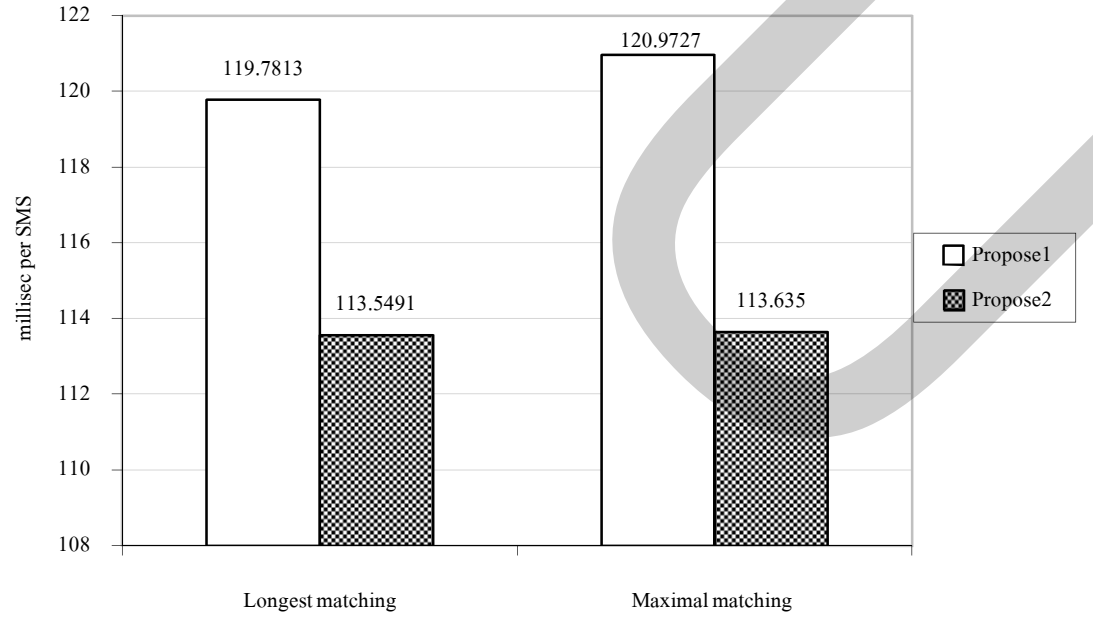
ภาพที่ 5.10 ความถูกต้องของอัลกอริทึม SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลใหม่ (ND)

จากภาพที่ 5.9 และ 5.10 แสดงให้เห็นว่าการกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มความถูกต้องของอัลกอริทึม SVM ในการกรองข้อมูล ND ได้สูงขึ้น

การวัดประสิทธิภาพทางเวลาในการกรองข้อความระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 โดยทดสอบด้วยอัลกอริทึม SVM



ภาพที่ 5.11 เวลาในการกรองของอัลกอริทึม SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลที่ใช้ฝึกสอน (TD)



ภาพที่ 5.12 เวลาในการกรองของอัลกอริทึม SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 จากชุดข้อมูลใหม่ (ND)

จากภาพที่ 5.11 และ 5.12 แสดงให้เห็นว่าการกรองข้อความที่นำเสนอแบบที่ 2 ใช้เวลาในการกรองน้อยกว่าการกรองข้อความที่นำเสนอแบบที่ 1 อีกทั้งความถูกต้องที่เพิ่มขึ้น ทำให้การกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มประสิทธิภาพในการกรองข้อความให้กับอัลกอริทึมแบบ SVM ได้ดี

5.2 สรุปผลการเปรียบเทียบระหว่าง NB SVM การกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

ผลการทดสอบประสิทธิภาพระหว่างอัลกอริทึมแบบ NB และ SVM

ตารางที่ 5.1 ผลการทดสอบประสิทธิภาพระหว่างอัลกอริทึมแบบ NB และ SVM ในการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

เปรียบเทียบความถูกต้องระหว่าง NB และ SVM		
อัลกอริทึม	การกรองข้อความที่นำเสนอแบบที่ 1 (%)	การกรองข้อความที่นำเสนอแบบที่ 2 (%)
NB	86.15622583	86.40757927
SVM	92.53673627	93.5808198
เปรียบเทียบการเวลาประมวลผลข้อความระหว่าง NB และ SVM		
อัลกอริทึม	การกรองข้อความที่นำเสนอแบบที่ 1 (millisec)	การกรองข้อความที่นำเสนอแบบที่ 2 (millisec)
NB	42.57928731	45.21358868
SVM	114.4988508	108.2246556

จากการทดสอบที่ผ่านสามารถสรุปผลการทำงานของ SMS Spam Filter ที่ปรับปรุงวิธีการกรองตามแบบที่นำเสนอในบทที่ 4 ได้ดังตารางที่ 5.2 และ 5.3

ตารางที่ 5.2 ผลการทดสอบการอัลกอริทึมแบบ NB ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

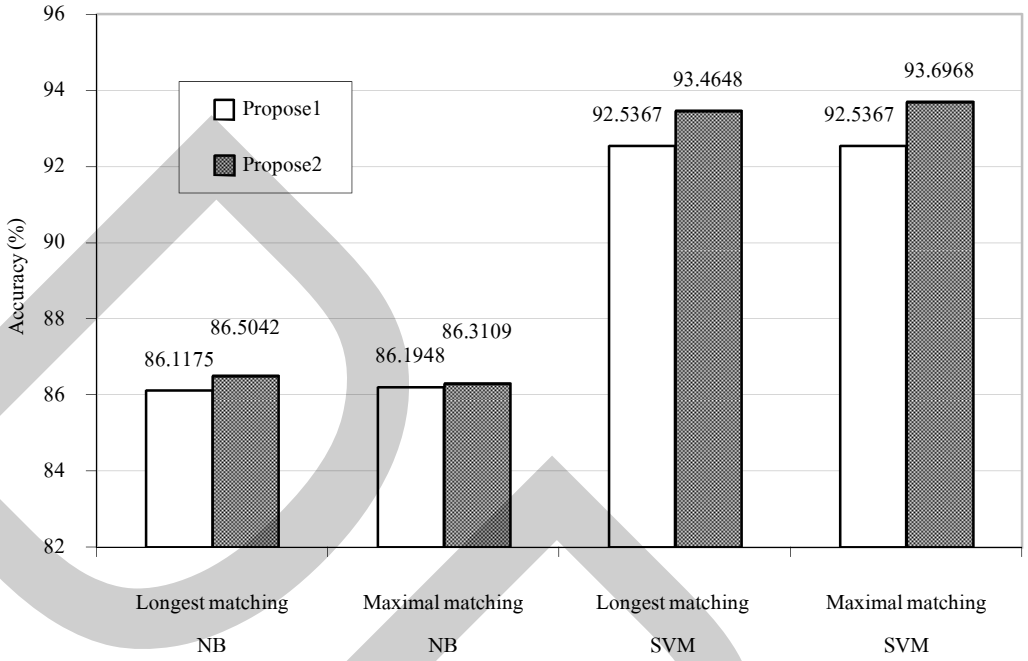
อัลกอริทึม	วิธีการตัดคำ	ชุดข้อมูล	การกรองข้อความที่นำเสนอแบบที่ 1		การกรองข้อความที่นำเสนอแบบที่ 2	
			ความถูกต้อง (%)	เวลา/SMS (millisec)	ความถูกต้อง (%)	เวลา/SMS (millisec)
NB	ยาวที่สุด	TD	89.76	39.1673	89.84	39.6484
		ND	82.7095	42.4202	83.3832	45.0292
	ค่าเฉลี่ย		**86.1175**	**40.7814**	**86.5042**	**42.3280**
	สอดคล้องที่สุด	TD	90.08	42.2466	89.92	45.1168
		ND	82.5598	46.5521	82.9341	51.1245
	ค่าเฉลี่ย		**86.1948**	**44.3771**	**86.3109**	**48.0991**

จากตารางที่ 5.2 แสดงการเปรียบเทียบระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 ด้วยอัลกอริทึม NB มีความถูกต้องระหว่างร้อยละ 86.1 ถึงร้อยละ 86.5 ซึ่งการกรองข้อความที่นำเสนอแบบที่ 2 นี้สามารถเพิ่มความถูกต้องในการการกรองได้ประมาณร้อยละ 0.25 แต่ต้องใช้เวลาในการประมวลผลต่อข้อความสูงขึ้นประมาณร้อยละ 6.39 ทำให้การกรองข้อความที่นำเสนอแบบที่ 2 ไม่มีความเหมาะสมในการกรองข้อความด้วยวิธีการแบบ NB

ตารางที่ 5.3 ผลการทดสอบการอัลกอริทึมแบบ SVM ระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

อัลกอริทึม	วิธีการตัดคำ	ชุดข้อมูล	การกรองข้อความที่นำเสนอแบบที่ 1		การกรองข้อความที่นำเสนอแบบที่ 2	
			ความถูกต้อง (%)	เวลา/SMS (millisec)	ความถูกต้อง (%)	เวลา/SMS (millisec)
SVM	ยาวที่สุด	TD	97.52	106.9534	97.68	102.8880
		ND	87.8742	119.7813	89.5209	113.5491
	ค่าเฉลี่ย		**92.5367**	**113.2468**	**93.4648**	**108.1634**
	สอดคล้องที่สุด	TD	97.04	110.6711	97.68	103.0199
		ND	88.3233	120.9727	89.9700	113.6350
	ค่าเฉลี่ย		**92.5367**	**115.7508**	**93.6968**	**108.2858**

จากตารางที่ 5.3 แสดงการเปรียบเทียบระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2 ด้วยอัลกอริทึม SVM สามารถกรองข้อความได้ความถูกต้องคิดเป็นร้อยละ 92.53 ซึ่งมีประสิทธิภาพมากกว่าอัลกอัลกอริทึมแบบ NB อยู่ที่ประมาณร้อยละ 6.38 และการกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มความถูกต้องในการกรองได้ประมาณร้อยละ 1.160 อีกทั้งยังใช้เวลาในการประมวลผลต่อข้อความลดลงถึงประมาณ 6.3 millisec คิดเป็นร้อยละ 5.27 ด้วยวิธีการตัดคำแบบสอดคล้องมากที่สุด ทำให้การกรองข้อความที่นำเสนอแบบที่ 2 สามารถช่วยเพิ่มประสิทธิภาพการกรองข้อความสแปมภายในประเทศไทยได้เป็นอย่างดี โดยแสดงการเปรียบเทียบประสิทธิภาพการกรองตามภาพที่ 5.13



ภาพที่ 5.13 ประสิทธิภาพความถูกต้องระหว่างการกรองข้อความที่นำเสนอแบบที่ 1 และแบบที่ 2

5.3 ข้อจำกัดของการกรองข้อความที่นำเสนอแบบที่ 2

ในการทดสอบกรองข้อความด้วยวิธีการตัดคำและวิธีการกรองแบบต่างๆ พบข้อจำกัดบางประการของการกรองข้อความที่นำเสนอแบบที่ 2 โดยมีตัวอย่างข้อความที่ผ่านการกรองแสดงตามตารางที่ 5.4

ตารางที่ 5.4 ตัวอย่างข้อความที่ผ่านการกรองและผลของการกรอง

SVM with Maximal Matching word segmentation		
SMS	**Classify**	**Correction**
นั่นโรคฮิตเลยนี่.	NORMAL	NORMAL
พระเจ้า พลาด?	**SPAM**	**NORMAL**
โอกาสสุดท้ายจาก Fitness First ถึงอาทิตย์ที่ 17 กุมภานี้ สำหรับสมาชิ	**NORMAL**	**SPAM**
โชว์SMSนี้ภายใน31มี.ค. ให้คุณ&เพื่อน3คนรับสมาชิก1เดือนเพียง399บ พร้	SPAM	SPAM
NB with Maximal Matching word segmentation		
SMS	**Classify**	**Correction**
พักแล้วนะ!บาย!	NORMAL	NORMAL
คุณคือผู้โชคดี ได้รับบัตรอาบน้ำฟรี ที่ฟาร์มจรเข้..... จร้าฮิฮิฮิ	**SPAM**	**NORMAL**
วิลด์ อัลไลแอนซ์0819477355,026454455	**NORMAL**	**SPAM**
แลกของรางวัลที่กรุงศรี ATM ถึง 31 มี.ค.51 สอบถามโทร. 1572	SPAM	SPAM

จากตารางที่ 5.4 แสดงตัวอย่างความผิดพลาดของการกรองข้อความที่นำเสนอแบบที่ 2 โดยแสดงผลของการกรองในช่อง classify เปรียบเทียบกับชนิดของข้อความที่แท้จริงจากการตัดสินด้วยมนุษย์ในช่อง correction ข้อความที่กรองผิดพลาดส่วนใหญ่คือข้อความที่ไม่ใช่สแปมแต่มีจำนวนคำน้อย ซึ่งคำส่วนใหญ่เป็นคำที่เกิดขึ้นแบบสแปมมากกว่าข้อความปรกติจากชุดข้อมูลฝึกสอนทั้งหมด และ ข้อความสแปมที่มีเนื้อความมากกว่า 1 ข้อความ ซึ่งเนื้อความที่ไม่ครบถ้วนนี้ทำให้ไม่สามารถกรองข้อความได้อย่างถูกต้อง

บทที่ 6
สรุปผลการวิจัย

6.1 สรุปผลการศึกษาและวิจัย

จากผลการศึกษาและเปรียบเทียบในบทที่ 5 แสดงให้เห็นถึงความแตกต่างของวิธีการกรองและวิธีการตัดคำภาษาไทยได้ดังนี้

6.1.1 วิธีการกรองแบบ SVM มีความถูกต้องมากกว่าวิธีการกรองแบบ NB แต่ในทางกลับกันวิธีการแบบ NB ต้องการเวลาในการประมวลผลน้อยกว่าวิธีการกรองแบบ SVM ประมาณ 2.5 เท่า

6.1.2 การปรับปรุงวิธีทำ TN แบบใหม่กับข้อความทั้งข้อความ สามารถลดปริมาณคำผิดในฐานข้อมูลลงได้ โดยเฉพาะการพิมพ์ตัวอักษร สระ และวรรณยุกต์มากกว่า 1 ตัวและการพิมพ์ สระและวรรณยุกต์ สลับตำแหน่งที่ก่อให้เกิดคำที่ไม่สามารถนำมาคำนวณได้เป็นจำนวนมาก

6.1.3 การใช้วิธีตัดคำแบบผสมในการตรวจสอบคำแรกและคำสุดท้ายของข้อความ สามารถลดปริมาณคำผิดที่เกิดจากการตัดคำไม่สมบูรณ์ลงได้ ทำให้วิธีการกรองแบบ SVM ที่ต้องใช้การคำนวณทางคณิตศาสตร์ปริมาณมาก สามารถลดระยะเวลาการประมวลผลต่อข้อความลงได้

6.1.4 การใช้วิธีตัดคำแบบผสมในการกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มความถูกต้องในการกรองได้เพราะลักษณะของข้อความสแปมในประเทศไทย มีการเขียนอย่างถูกต้องตามหลักภาษาศาสตร์เป็นส่วนใหญ่

6.1.5 การตรวจสอบเลขหมายพิเศษภายในข้อความ สามารถตรวจสอบข้อความสแปมที่มีเนื้อความเกินกว่า 1 ข้อความ โดยแต่ละข้อความขาดใจความที่แสดงถึงการเป็นข้อความสแปมได้ดี

6.1.6 ข้อความบางชนิด ไม่สามารถกรองได้อย่างถูกต้อง ได้แก่ ข้อความสแปมที่มีเนื้อความยาวเกินกว่า 1 ข้อความ และการตัดแบ่งข้อความทำให้ความหมายของแต่ละข้อความไม่มีความต่อเนื่อง จนไม่สามารถตรวจสอบได้ว่า ข้อความนี้เป็นสแปมหรือไม่ เช่น

“ร้านอาหารลีลาวดี ขอนแก่น บริการจัดบุฟเฟ่ต์ อาหารกล่อง คอฟฟี่เบรค โต๊ะจีนราคาเริ่มต้น 700.- สอบถามโทร. 0815455450 www.leelawadee.tht.in” จะถูกแบ่งออกเป็น 2 ข้อความดังนี้

1) ร้านอาหารลีลาวดี ขอนแก่น บริการจัดบุฟเฟ่ต์ อาหารกล่อง คอฟฟี่เบรค โต

2) ๊ะจีนราคาเริ่มต้น 700.- สอบถามโทร. 0815455450 www.leelawadee.tht.in

ซึ่งไม่สามารถตัดสินข้อความแรกให้เป็นข้อความสแปมได้เนื่องจากเนื้อหาในข้อความไม่มีคำที่มีอัตราการเป็นข้อความสแปมสูง จึงถูกตัดสินให้เป็นข้อความปรกติ

6.1.7 ไม่สามารถตรวจสอบข้อความปรกติที่มีเนื้อความคล้ายคลึงกับข้อความสแปมได้ เนื่องจากคำที่บรรจุในข้อความเป็นคำที่พบในชุดข้อมูลฝึกสอนแบบสแปมทั้งสิ้น เช่น “คุณคือผู้โชคดี ได้รับบัตรอาบน้ำฟรี ที่ฟาร์มจรเข้..... จร้าฮิฮิฮิ” เป็นต้น

6.1.8 การกรองข้อความที่นำเสนอแบบที่ 2 สามารถเพิ่มความถูกต้องในการกรองข้อความ SMS ในประเทศไทย เมื่อเปรียบเทียบกับการกรองข้อความที่นำเสนอแบบที่ 1 ซึ่งใช้การดัดแปลงจากกระบวนการกรองข้อความภาษาอังกฤษลงได้ร้อยละ 1.16 ซึ่งมีความถูกต้องของการกรองข้อความอยู่ที่ร้อยละ 93.69 ในการใช้อัลกอริทึมแบบ SVM

6.1.9 การกรองข้อความที่นำเสนอแบบที่ 2 สามารถลดระยะเวลาการประมวลผลต่อข้อความจากการกรองข้อความที่นำเสนอแบบที่ 1 ลงได้ร้อยละ 5.27 ซึ่งทำให้ระบบกรองข้อความ สามารถประมวลผลข้อความจำนวนมากได้อย่างมีประสิทธิภาพ

6.2 ข้อเสนอแนะและงานวิจัยในอนาคต

การตรวจสอบคำจากข้อความ SMS ที่ผ่านกระบวนการตัดคำด้วยวิธีการแบบผสม ยังมีความผิดพลาด จำเป็นต้องมีการวิจัยเพื่อหาวิธีการตัดคำที่สามารถรองรับการเขียนข้อความที่มีคำผิดจำนวนมากได้ เช่น การวิจัยเรื่องการหารากศัพท์ในภาษาไทยที่ยังไม่มีงานวิจัยที่มุ่งศึกษาหัวข้อดังกล่าวอย่างจริงจัง

ในการวิเคราะห์ข้อความที่ไม่สามารถคัดกรองได้อย่างถูกต้องพบว่า ข้อความส่วนใหญ่เป็นข้อความสแปมที่มีเนื้อความยาวเกินกว่า 1 ข้อความ และการตัดแบ่งข้อความทำให้ความหมายของแต่ละข้อความไม่มีความต่อเนื่อง จนไม่สามารถตรวจสอบได้ว่า ข้อความนี้เป็นสแปมหรือไม่ หากสามารถตรวจสอบความหมายของข้อความที่มีความต่อเนื่องนี้ได้อย่างถูกต้อง จะสามารถเพิ่มอัตราการกรองที่ถูกต้องได้สูงยิ่งขึ้น และในกรณีที่ข้อความที่มีคำน้อยเกินไป และคำแต่ละคำเกิดขึ้นในฐานข้อมูลแบบสแปมจำเป็นต้องมีวิธีการตรวจสอบเพิ่มเติม

งานวิจัยนี้ ผู้วิจัยต้องการศึกษาวิธีการทำ TN และวิธีการตัดคำในปัจจุบัน โดยปรับปรุงขั้นตอนดังกล่าวเพื่อประยุกต์เข้ากับการกรองข้อความด้วยวิธีการกรองแบบ SVM และ NB สำหรับข้อความ SMS ที่ใช้งานในประเทศไทย โดยคาดหวังว่าองค์ความรู้นี้จะเป็นพื้นฐานในการพัฒนาระบบการกรองข้อความ SMS ให้สามารถนำมาแก้ไขปัญหาข้อความสแปมในประเทศไทยต่อไป

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

หนังสือ

วิเชียร เปรมชัยสวัสดิ์. (2546). **ระบบฐานข้อมูล**. กรุงเทพฯ : สมาคมส่งเสริมเทคโนโลยี (ไทย-ญี่ปุ่น).

รายงานการวิจัย

กานดา รุณนะพงศา และ ปโยธร อุราธรรมกุล. (2549). **การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่**. ขอนแก่น : ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น.

บทความ

กิตติ ภักดีวัฒนะกุล และ ทวีศักดิ์ กาญจนสุวรรณ. (2544). **สร้างระบบสารสนเทศบนเว็บด้วย FrontPage 2002**. หน้า 2-5.

อติชาต ขานทอง, วัลลภา ตันติประสงค์ชัย และ ชุลีรัตน์ จรัสกุลชัย. (2544). **Document Summarization**. หน้า 1-3.

สารสนเทศจากสื่ออิเล็กทรอนิกส์

กำธน สินธวานนท์. (2544). โปรแกรมตัดคำภาษาไทย (Thai Word Segmentation). สืบค้นเมื่อ 14 กันยายน 2550. จาก **http://guru.sanook.com/encyclopedia/%E0%B9%82%E0%B8%9B%E0%B8%A3%E0%B9%81%E0%B8%81%E0%B8%A3%E0%B8%A1%E0%B8**

%95%E0%B8%B1%E0%B8%94%E0%B8%84%E0%B8%B3%E0%B8%A0%E0%B8%B2%E0%B8%A9%E0%B8%B2%E0%B9%84%E0%B8%97%E0%B8%A2_(Thai_Word__Segmentation)/

ข่าวทั่วไป. (2552). ผู้ให้บริการโทรมือถือขานรับ สคบ. เปิดให้ประชาชนโทรขอยกเลิก SMS ขายของ. MCOT. สืบค้นเมื่อ 1 สิงหาคม 2552. จาก **http://news.mcot.net/social/inside.php?value=bmlkPTEwNjk1MiZudHlwZT10ZXh0**

นิศารัตน์ วิเชียรศรี. (2552). DTAC เผยยอด SMS และ MMS ฉลองปีใหม่เพิ่มขึ้น 31% และ 43% จากปี 51. สำนักข่าวอินโฟเควสท์ (IQ). สืบค้นเมื่อ 13 เมษายน 2552. จาก **http://www.ryt9.com/s/iq05/498137**

ผู้จัดการออนไลน์. (2549). ดีแทคชนทรูมูฟแก้ปัญหา SPAM SMS บล็อกไม่รับข้อความขายซิมกว่า 16 ชม. ASTV. สืบค้นเมื่อ 14 เมษายน 2552. จาก **http://www.manager.co.th/Cyberbiz/ViewNews.aspx?NewsID=9490000090269**

ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. ข้อมูลพื้นฐาน ภาษาไทย, คำศัพท์ที่พบบ่อยในฐานข้อมูล. สืบค้นเมื่อ 15 เมษายน 2552. จาก **http://thailang.nectec.or.th/thaichar/word_thai.php?page=1&n_p_page=100**

วิรัช ศรเลิศล้ำวานิช. (2543). โปรแกรมตัดคำภาษาไทย. National Electronics and Computer Technology Center (NECTEC). สืบค้นเมื่อ 12 สิงหาคม 2551. จาก **http://www.hlt.nectec.or.th/products/swath.php**

ศิรินุช เทียนรุ่งโรจน์. (2552). ระบบฐานข้อมูล, มหาวิทยาลัยศรีนครินทรวิโรฒ ประสานมิตร. สืบค้นเมื่อ 1 มิถุนายน 2552. จาก **http://sot.swu.ac.th/CP342/lesson06/ms1t1.htm**

สำนักข่าว ผู้จัดการออนไลน์. (2551). ดีแทค เผยสถิติส่งข้อความส่งความสุขปีใหม่ (SMS) 38 ล้านข้อความ (MMS) 6.8 แสนข้อความ. สืบค้นเมื่อ 29 มกราคม 2551. จาก **http://phone.thaiza.com/%E0%B8%94%E0%B8%B5%E0%B9%81%E0%B8%97%E0%B8%84%20%E0%B9%80%E0%B8%9C%E0%B8%A2%E0%B8%AA%E0%B8%96%E0%B8%B4%E0%B8%95%E0%B8%B4%E0%B8%AA%E0%B9%88%E0%B8%87%E0%B8%82%E0%B9%89%E0%B8%**

B8%AD%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1%E0%B8%AA%E0%B9%88%E0%B8%87%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1%E0%B8%AA%E0%B8%B8%E0%B8%82%E0%B8%9B%E0%B8%B5%E0%B9%83%E0%B8%AB%E0%B8%A1%E0%B9%88%202551%20SMS%2038%20%E0%B8%A5%E0%B9%89%E0%B8%B2%E0%B8%99%E0%B8%82%E0%B9%89%E0%B8%AD%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1%20%20MMS%2068%20%E0%B9%81%E0%B8%AA%E0%B8%99%E0%B8%82%E0%B9%89%E0%B8%AD%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1_1212_83502_1212_.html

ภาษาต่างประเทศ

ARTICLES

Giovanni Camponovo, Davide Cerutti. (2004). **The spam issue in mobile business a comparative regulatory overview.** P. 1-2.

Gordon V. Cormack, Jose Maria Gomez Hidalgo, Enrique Puertas Sanz. (2007). **Spam filtering for short messages.** P. 1-4.

Petros Zerfos, Xiaoqiao Meng, Starsky H.Y. Wong, Vidyut Samanta, Songwu Lu. (2006). **A Study of the Short Message Service of a Nationwide Cellular Network.** P. 1-6.

S. Dixit, S. Gupta, and C.V. Ravishankar. (2005). **LOHIT: An Online Detection & Control System for Cellular SMS Spam**. P. 2-8.

István Pilászy (2005) **Text Categorization and Support Vector Machines.** P. 1-9.

ELECTRONIC SOURCES

Fidelis Assis. (2006). OSBF-Lua, Text classification module for the Lua Programming Language and a production class anti-spam in Lua using the module. Retrieved April 14, 2007, from **http://osbf-lua.luaforge.net/**

Justin Mason. (2008). The Apache SpamAssassin Project, The Powerful #1 Open-Source Spam Filter. Retrieved March 18, 2009, from **http://spamassassin.apache.org/**

Mark Sanderson. (1999). Stop words. Retrieved April 15, 2009, from **http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words**

Rafael Pinto. (2005). SpamFilter 1.1. Retrieved April 13, 2007, from **http://www.phpclasses.org/browse/package/2275.html**

Smsforum.net. (1999). Short Message Peer to Peer Protocol Specification v3.4. Retrieved October 12, 2007, from **http://smsforum.net/SMPP_v3_4_Issue1_2.zip**

Thorsten Joachims. (2008). SVM Light, Support Vector Machine. Retrieved April 15, 2009, from **http://svmlight.joachims.org/**

WebGate JSC. (2007). SMS Spam Manager. Retrieved December 18, 2007, from **http://www.webgate.bg/products/ssm/**

wikipedia.org. (2006). Support vector machine. Retrieved January 29, 2009, from **http://en.wikipedia.org/wiki/Support_vector_machine**

wikipedia.org. (2006). Naive Bayes classifier. Retrieved April 29, 2009, from **http://en.wikipedia.org/wiki/Naive_Bayes_classifier**

ภาคผนวก

ภาคผนวก ก.

บทความทางวิชาการ เรื่องการกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่

การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้น บนเครือข่ายโทรศัพท์เคลื่อนที่

Short Message Service Filtering for Thai & English Language on Mobile Phone Network

ชัยพร เขมะภาตะพันธ์ *นนท์ บุญนิธิประเสริฐ*

chaiyaporn.k@eng.dpu.ac.th *nontwork@yahoo.com*

ภาควิชาวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม มหาวิทยาลัยธุรกิจบัณฑิตย์

บทคัดย่อ

การกรองข้อความ Spam ในระบบส่งข้อความสั้น Short Message Service (SMS) ของประเทศไทยนั้น ยังไม่มีการศึกษาและพัฒนาอย่างจริงจัง ซึ่งปัญหาการมีข้อความ Spam ในระบบส่งข้อความของผู้ให้บริการโทรศัพท์เคลื่อนที่กำลังมีความรุนแรงเพิ่มขึ้น งานวิจัยนี้ได้ทำการศึกษาการกรองข้อความ Spam ด้วย วิธีการกรองแบบ Support Vector Machine (SVM) และ Naive Bayesian (NB) การทำ Text Normalization และการตัดคำแบบต่างๆที่มีการใช้งานอยู่ในปัจจุบัน โดยได้ปรับปรุงการทำ Text Normalization และการใช้วิธีตัดคำแบบผสม แล้วทดสอบประสิทธิภาพการทำงานด้วยการกรองข้อความ SMS ทั้งภาษาไทย ภาษาอังกฤษ และภาษาไทยปนอังกฤษ

ผลการทดสอบพบว่า การปรับปรุงขั้นตอนการทำ Text Normalization สามารถลดปริมาณคำที่ไม่ถูกต้องในพจนานุกรมฐานข้อมูลลงได้ ผลการทดสอบด้วยวิธีการกรองแบบ SVM มีความถูกต้องในการกรองข้อความสูงกว่าวิธีการแบบ NB แต่วิธีการกรองแบบ NB ใช้เวลาในการประมวลผลน้อยกว่า และการใช้วิธีตัดคำแบบผสมส่งผลให้การกรองข้อความ SMS ที่มีเนื้อความเกินกว่า 1 ข้อความมีความถูกต้องมากกว่าการตัดคำด้วยวิธีการแบบใดแบบหนึ่งเพียงแบบเดียว ซึ่งต้องใช้ระยะเวลาในการประมวลผลเพิ่มขึ้น

คำสำคัญ: ข้อความสั้น, ข้อความขยะ, การกรองข้อความ, โทรศัพท์เคลื่อนที่

Abstract

Nowadays, all mobile operators in Thailand have faced with the severe problem in Short Message Service (SMS) spam. However, SMS spam filter system has not seriously been studied and developed seriously in order to solve this problem. This research aims to study in SMS spam filtering by using Support Vector Machine (SVM) and Naive Bayesian (NB) algorithm by using text normalization and several words segmentations processes. This research had shown the process of the text normalization and mixing words segmentation algorithm and then tests SMS filtering in the form of Thai Language, English Language, and Combination of Languages between Thai and English.

The result shows that using the text normalization process to filter SMS can decrease wrong words in dictionary's database. It also reveals SMS spam filtering by using SVM has higher efficient result in correct filtering than NB. On the other hand, NB processing spends less time on testing process than SVM. Moreover, the result of SMS filtering by using mixing words segmentation to separate length SMS has more efficiently than filtering only one words segmentation technique at a time, it therefore spends longer time for filtering than others.

Keyword: SMS, Spam SMS, SMS Filtering, Mobile Phone

1. บทนำ

บริการเสริมที่มีผู้ใช้งานเป็นจำนวนมากที่สุดบริการหนึ่งของระบบโทรศัพท์เคลื่อนที่ในปัจจุบันคือ บริการส่งข้อความสั้นหรือ Short Message Service (SMS) ถูกคิดค้นและพัฒนาประมาณปี 1980 บนระบบโทรศัพท์เคลื่อนที่ Global Systems for Mobile communications (GSM) ซึ่งจัดเป็นบริการข้อมูล หรือ Data Service บนระบบโทรศัพท์เคลื่อนที่ที่ประสบความสำเร็จสูงที่สุด ในขณะที่บริการดังกล่าวได้รับความนิยมมากขึ้น จึงเริ่มมีการใช้ SMS เป็นสื่อโฆษณาประชาสัมพันธ์ ข้อความประเภทนี้บางกลุ่มจัดเป็นข้อความ Spam (Spam SMS) ประเทศเกาหลีใต้และญี่ปุ่นมีจำนวนข้อความ Spam ในระบบ SMS มากกว่าร้อยละ 50 ของการใช้งาน [1] ซึ่งนอกจากจะรบกวนการใช้งานของผู้ใช้โทรศัพท์แล้ว ยังส่งผลกระทบต่อการทำงานของระบบเครือข่าย SMS (SMS Network) ทำให้การกรองข้อความ Spam ก่อนมีการส่งถึงผู้รับมีความจำเป็นอย่างยิ่ง เพราะนอกจากจะลดปัญหาการรบกวนต่อผู้ใช้งานแล้ว ยังช่วยเพิ่มประสิทธิภาพของระบบ SMS ให้แก่ผู้ให้บริการอีกด้วย

รายงานการวิจัยหลายฉบับได้ทำการวิเคราะห์เพื่อปรับปรุงบริการบนระบบโทรศัพท์เคลื่อนที่ต่างๆหลายบริการ แต่งานวิจัยที่มุ่งเน้นการแก้ไขปัญหาข้อความ Spam ในระบบ SMS ยังมีไม่มาก อาจเนื่องมาจากปัญหาเกิดขึ้นกับผู้ใช้บริการโทรศัพท์เคลื่อนที่มากกว่าผู้ให้บริการ ซึ่งการแก้ปัญหาดังกล่าวมีองค์ประกอบหลายประการขึ้นอยู่กับผู้ให้บริการ จึงทำให้การศึกษาวิจัยเพื่อแก้ไขปัญหาทำได้ลำบาก

ความหมายของคำว่า "ข้อความ Spam" คือ การส่งข้อความ ซึ่งก่อความรำคาญแก่ผู้รับ ซึ่งข้อความ Spam ที่พบในชีวิตประจำวันได้แก่ E-Mail, กระทู้แสดงความคิดเห็นใน Forum บน Web Site ต่างๆ ข้อความโฆษณาขายสินค้าผ่านเครื่องโทรศัพท์เคลื่อนที่ งานวิจัย [2] ให้นิยามความหมายของข้อความ Spam ไว้ดังนี้

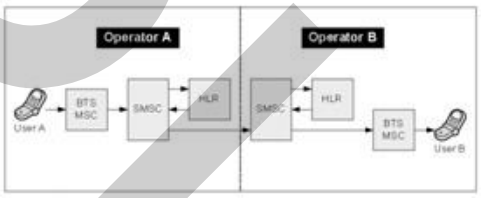
- ไม่มีการร้องขอ : ผู้รับมิได้ร้องขอข้อมูล และไม่ทราบข้อมูลของผู้ส่ง
- ส่งครั้งละหลายข้อความ : ผู้ส่งทำการส่งข้อความจำนวนมากติดต่อกัน
- จุดประสงค์ : เช่น การชักจูงผู้รับให้ทำกิจกรรมบางประการเพื่อให้ผู้ส่งได้รับประโยชน์

การแก้ปัญหาข้อความ Spam สามารถทำได้หลายรูปแบบ เช่น การใช้ Software กรองข้อความ Spam ติดตั้งลงบนเครื่องโทรศัพท์เคลื่อนที่ ซึ่งช่วยลดปริมาณการรับข้อความ Spam ที่ผู้รับไม่ต้องการลงได้ แต่ยังคงมีปริมาณข้อความ Spam ในระบบส่งข้อความสั้นของผู้ให้บริการอยู่ นอกจากนี้ ผู้ให้บริการโทรศัพท์เคลื่อนที่บางราย ได้มีการเพิ่มมาตรการระบุตัวตนผู้ส่ง (Authentication) เพื่อให้สิทธิ์การส่งข้อความ SMS กับผู้ให้บริการ Content ซึ่งมาตรการนี้มีวัตถุประสงค์เพื่อเป็นข้อมูลอ้างอิงทางกฎหมาย โดยไม่สามารถลดปริมาณข้อความ Spam ออกจากระบบส่งข้อความ SMS ลงได้

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ระบบส่งข้อความสั้น SMS

ระบบส่งข้อความสั้นมีองค์ประกอบสำคัญได้แก่ ผู้ส่ง, ผู้รับ, ข้อความ, SMS Network และ Protocol



ภาพที่ 1: การส่งข้อความ SMS ระหว่าง User A และ B

จากภาพที่ 1 มีลำดับการส่งข้อความเริ่มต้นจากเครื่องโทรศัพท์เคลื่อนที่ของผู้ส่ง (User A) ผ่านเสา รับ – ส่งสัญญาณโทรศัพท์ (Base Transceiver Station) และชุมสาย (Mobile Switching Center) ไปยัง SMS Center (SMSC) ซึ่งจะทำหน้าที่ค้นหาตำแหน่งผู้รับ (User B) จาก Home Location Register (HLR) แล้วดำเนินการจัดส่งข้อความ

การกรองข้อความ จะทำการตรวจสอบที่มาของ ผู้ส่งและความหมายของข้อความที่ SMSC ซึ่งเป็นศูนย์กลางในการรับส่งข้อความ ด้วยการถอดข้อความจาก Short Message Peer-to-Peer Protocol (SMPP Protocol) แล้วตรวจสอบเนื้อหา (Content) ที่อยู่ภายใน

SMPP ซึ่งเป็น Protocol มาตรฐานในการส่งข้อมูล SMS MMS หรือ PUSH Message ภายในระบบโทรศัพท์เคลื่อนที่ ประกอบด้วย 2 ส่วนสำคัญดังตารางที่

2 คือ ส่วนที่ 1 PDU Header ที่ใช้ในการระบุ ความยาว ชนิด และลำดับของข้อความ ส่วนที่ 2 PDU Body ใช้บรรจุข้อมูลที่ต้องการส่งผ่าน เช่น ข้อความภายใน SMS หรือ Link สำหรับ PUSH Message เป็นต้น

ตารางที่ 1: รูปแบบของ SMPP PDU [3]

SMPP PDU				
PDU Header (mandatory)				PDU Body (Optional)
command length	command id	command status	sequence number	PDU Body
4 octets	Length = (Command Length value - 4) octets			

2.2 Spam Filtering for Short Messages [4]

เป็นงานวิจัยที่ได้ทำการเปรียบเทียบระบบกรองข้อความ Spam หรือ SMS Spam Filter ดังต่อไปนี้

Bogofilter ใช้เทคนิคการตรวจจับด้วยทฤษฎี Bayesian หรือการหาความน่าจะเป็น

OSBF-Lua พัฒนาขึ้นจากภาษา C ชนิดหนึ่ง ที่ใช้เทคนิค orthogonal sparse bigrams [5]

Dynamic Markov Compression (DMC) เน้นการประมวลผลกับข้อมูลที่ถูกบีบอัด

Logistic Regression (LR) เป็น Open source spam filter ที่ใช้หลักการคำนวณทางตรรกศาสตร์ คล้ายกับ SVM แต่มีระดับการคำนวณต่ำกว่า

Support Vector Machine (SVM) มีชื่อเสียงในการค้นหาข้อความ Spam ด้วยหลักการหาค่าสัมประสิทธิ์ของฟังก์ชันเชิงเส้น ซึ่งมีประสิทธิภาพความถูกต้องสูงที่สุดในการทดสอบ

2.3 Filter Algorithm

Rule Base เป็นวิธีการกรองข้อความด้วยกฎและเงื่อนไขการใช้ Keywords Matching ไม่มีความซับซ้อน สามารถได้ทำงานรวดเร็ว แต่ความถูกต้องน้อยกว่าวิธีการอื่น และขาดความยืดหยุ่นในการใช้งาน

Support Vector Machine (SVM) [6] ใช้การหาลักษณะแทนข้อความด้วย วิธี TFIDF [7] ดังสมการที่ 1 และ 2

$$TFIDF(i,j) = TF(i,j) \cdot IDF(i) \quad \cdots\cdots (1)$$

$$IDF(i) = \log \frac{N}{DF(i)} \quad \cdots\cdots (2)$$

โดยที่ TF คือ ความถี่ของ Term หรือ คำ ที่ปรากฏใน Document หรือข้อความ SMS

DF คือ ความถี่ของ Document ที่มี Term นี้

IDF คือ ค่าแทน Discrimination power ของ DF

และใช้การฝึกสอนด้วยชุดข้อมูลตั้งต้น เพื่อสร้าง Model สำหรับ SVM ในการ Classify ข้อมูล จากสมการพื้นฐานที่ 3 ถึง 5 ตามลำดับ

$$a = w^T \cdot x + b \quad \cdots\cdots (3)$$

$$\text{new weight} = w + \eta(t-a)x^T \quad \cdots\cdots (4)$$

$$\text{new bias} = b + \eta(t-a) \quad \cdots\cdots (5)$$

โดยที่ w คือ ค่า vector น้ำหนักของสมการ SVM

x คือ vector feature ของข้อความ

b คือ ค่าคงที่สำหรับกำหนดความเบี่ยงเบน

t คือ ค่าที่ vector x ควรจะเป็น (1 หรือ -1)

η คือ ค่าการเรียนรู้ของสมการ SVM

โดยเขียนแทนได้เป็นชุดสมการที่ 6 และ 7 [6]

$$\min(w,b) = \frac{1}{2} \cdot w^T \cdot w + C \cdot \sum_{i=1}^{n} \xi_i \quad \cdots\cdots (6)$$

$$\text{subject to} : \forall_{i=1}^{n} : y_i \left[w^T \cdot x_i + b \right] \geq \xi_i \quad \cdots\cdots (7)$$

Naive Bayesian (NB) [8] เป็นทฤษฎีการใช้ Probability ในการแก้ปัญหาที่ไม่สามารถใช้หลักสถิติได้ การนำ NB มาใช้กับการกรองข้อความ Spam จะใช้การหา Probability ของคำในข้อความที่มีโอกาสเป็น Spam หรือ P(D|Spam) เปรียบเทียบกับ Probability ของคำในข้อความที่มีโอกาสเป็น Normal หรือ P(D|Norm) ซึ่งหากเปรียบเทียบกันแล้วมีค่ามากกว่า 0 แสดงว่า ข้อความดังกล่าวน่าจะเป็น Spam [9] อธิบายได้จากสมการที่ 8 ถึง 10 ตามลำดับ

$$p(D \mid Spam) = \frac{p(Spam)}{p(D)} \prod_i p(w_i \mid Spam) \quad \cdots\cdots (8)$$

$$p(D \mid Norm) = \frac{p(Norm)}{p(D)} \prod_i p(w_i \mid Norm) \quad \cdots\cdots (9)$$

$$\text{Spam} = \ln \frac{p(D \mid Spam)}{p(D \mid Norm)} > 0, \text{ otherwise Norm} \cdot \quad (10)$$

โดยที่ P คือ ค่าความน่าจะเป็น

D คือ Document หรือ ข้อความ SMS

w_i คือ ลำดับของคำ ในข้อความ SMS

2.4 Thai Words Segmentation [10]

การตัดคำ เป็นขั้นตอนที่ต้องทำก่อนที่จะนำเอกสารแบบ Text ไปใช้คำนวณหาค่าใดๆ

การตัดคำภาษาไทยแบ่งออกเป็น 3 วิธีหลักๆได้ดังนี้

วิธีการตัดคำแบบยาวที่สุด (Longest Matching) จะตัดคำโดยตรวจสอบจากตัวอักษรแรกและตัวอักษรถัดไปตามลำดับจนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรม

วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) ใช้การตัดคำที่สามารถจะเป็นไปได้ทั้งหมดแล้วเลือกกลุ่มคำที่ตัดได้จำนวนคำน้อยที่สุดมาใช้งาน

วิธีการตัดคำแบบคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Model) เป็นการนำเอาค่าสถิติการเกิดของคำและลำดับหน้าที่ของคำเข้ามาช่วยในการตัดคำ

3. การดำเนินงานวิจัย

3.1 ศึกษาและวิเคราะห์ปัญหา

ในการประมวลผลข้อความ SMS เพื่อกรองข้อความ Spam จำเป็นต้องมีการทำ Text Normalization และการตัดคำ เพื่อให้ข้อความอยู่ในสถานะที่จะนำไปประมวลผลงานวิจัยที่กล่าวถึงวิธีการตัดคำภาษาไทยและภาษาไทยปนอังกฤษที่ผ่านมา [11] ใช้การตัดคำกับเอกสารที่ข้อความมีความถูกต้องตามหลักภาษาศาสตร์ ในขณะที่ลักษณะข้อความ SMS ของประเทศไทยที่เก็บข้อมูลเป็นระยะเวลา 3 เดือน พบว่า มีลักษณะของข้อความที่ไม่เป็นไปตามหลักภาษาศาสตร์ ทั้งภาษาไทยและภาษาอังกฤษจำนวนมาก เนื่องจาก SMS เป็นการสื่อสารที่ไม่จำเป็นต้องใช้ภาษาอย่างเป็นทางการ อีกทั้งข้อจำกัดของจำนวนตัวอักษรที่พิมพ์ได้ในข้อความ นอกจากนี้ยังพบคำที่พิมพ์ไม่สมบูรณ์ในตอนต้นและท้ายข้อความ ซึ่งเกิดจากการส่งข้อความที่มีเนื้อความยาวเกินกว่าขนาดของ SMS ผ่านเครื่องโทรศัพท์เคลื่อนที่หรือโปรแกรมส่งข้อความที่ตัดคำไม่ถูกต้อง ดังตัวอย่างจากตารางที่ 2

ตารางที่ 2: แสดงลักษณะข้อความที่พบใน SMSC

ข้อความที่ถูกต้อง	ลำดับ	ข้อความที่ตรวจพบ
ช่วงนี้อากาศเปลี่ยนแปลงบ่อย ดูแลสุขภาพให้ดีนะ ด้วยความปรารถนาดีจากทิฟฟี่แผงสีเขียว ฮิๆ	1	ช่วงนี้อากาศเปลี่ยนแปลงบ่อย ดูแลสุขภาพให้ดีนะ ด้วยความปรารถนาดีจากท
	2	ิฟฟี่แผงสีเขียว ฮิๆ
คิดถึงเป็นคำสั้นๆดูเหมือนไม่มีความหมายแต่ก็ทำให้ผีบ้าอ่านแล้วยิ้มได้ก็แล้วกัน	1	คิดถึงเป็นคำสั้นๆดูเหมือนไม่มีความหมายแต่ก็ทำให้ผีบ้าอ่านแล้วยิ้มโด
	2	้ก็แล้วกัน

เมื่อมีคำที่พิมพ์ผิดและคำที่พิมพ์ไม่สมบูรณ์เป็นจำนวนมาก การตัดคำและการหาความหมายของข้อความจึงผิดพลาด ซึ่งส่งผลโดยตรงต่อการกรองข้อความ ทำให้ต้องมีการปรับปรุงการทำ Text Normalization และการตัดคำ เพื่อให้รองรับกับลักษณะของข้อความ SMS

3.2 การออกแบบและแก้ไขปัญหา

ปรับปรุงการทำ Text Normalization คือขั้นตอนการลบ สัญลักษณ์พิเศษ เช่น **$ | # | @ | ? | !** หรือตัวเลขที่ไม่ต้องการเพื่อกำจัดข้อมูลส่วนเกินออก แต่เนื่องจากข้อความ SMS ที่สร้างขึ้นด้วยการพิมพ์จากเครื่องโทรศัพท์เคลื่อนที่ ทำให้พบการพิมพ์สระตัวเดียวกันมากกว่า 1 ครั้ง หรือการพิมพ์สระและวรรณยุกต์สลับที่เป็นปริมาณมาก เช่น การพิมพ์คำว่า “อยู่” จะต้องพิมพ์ **สระอู** ก่อนพิมพ์ **ไม้เอก** เสมอ ซึ่งหลักการพิมพ์ภาษาไทยที่ถูกต้อง จะพิมพ์สระก่อนวรรณยุกต์ดังตารางที่ 3

ตารางที่ 3: สระที่ต้องพิมพ์ก่อนหน้าวรรณยุกต์

สระ	วรรณยุกต์
◌ั ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู	◌่ ◌้ ◌๊ ◌๋

สำหรับ สระอำ (◌ำ) ต้องพิมพ์ตามหลังวรรณยุกต์ ได้แก่ ไม้เอก (◌่) ไม้โท (◌้) ไม้จัตวา (◌๊) จึงต้องปรับปรุงการทำ Text Normalization ให้สามารถแก้ไขปัญหาดังกล่าว เพื่อลดปริมาณคำผิดและเพิ่มประสิทธิภาพในการตัดคำ

การตัดคำ วิธีการตัดคำแบบคำนวณเชิงสถิติ แม้จะมีประสิทธิภาพสูงเมื่อใช้งานกับเอกสารที่มีความถูกต้อง [10] แต่เมื่อนำมาใช้กับข้อความ SMS พบว่า มีความผิดพลาดสูงกว่าการตัดคำแบบสอดคล้องมากที่สุด แต่การตัดคำแบบสอดคล้องเพียงวิธีเดียว จะตัดคำผิดพลาดเมื่อพบคำที่พิมพ์ไม่สมบูรณ์ โดยจะนำคำที่พิมพ์ไม่สมบูรณ์รวมเข้ากับคำถัดไปแล้วทำการตัด ทำให้คำถัดไปที่พิมพ์ถูก มีความหมายผิดไปจากเดิม

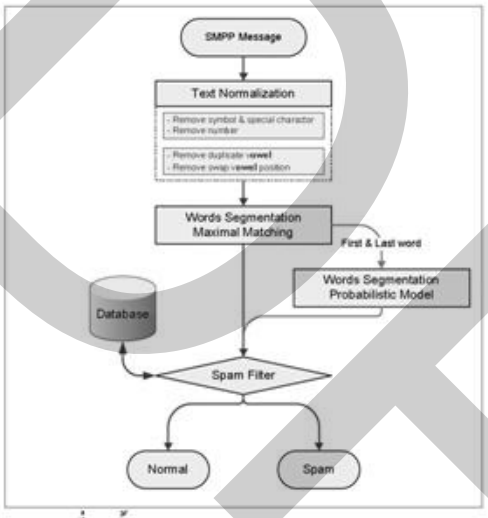
ปัญหาดังกล่าว สามารถแก้ไขด้วยวิธีการนำคำที่พิมพ์ไม่สมบูรณ์และคำถูก ที่การตัดคำแบบสอดคล้องไม่สามารถตัดได้ มาผ่านกระบวนการตัดคำอีกครั้งด้วยวิธีการคำนวณเชิงสถิติ และเนื่องจากคำที่พิมพ์ไม่สมบูรณ์สามารถตรวจพบได้ที่คำแรกสุดของข้อความที่เป็นข้อความต่อเนื่อง ตามตัวอย่างในตารางที่ 2 ข้อความที่ 2 จึงสามารถตรวจสอบคำผิดตามลักษณะดังกล่าวได้ง่าย โดยแสดงผลการตัดคำแบบผสมดังตารางที่ 4

ตารางที่ 4: เปรียบเทียบวิธีตัดคำ

แบบสอดคล้อง	แบบผสม
้ก็ | แล้ว | กัน	ก็ | แล้ว | กัน

3.3 SMS Filter

ระบบกรองข้อความ SMS จะถูกติดตั้งที่ SMSC หรือ SMS Gateway โดยมีขั้นตอนการทำงานตามภาพที่ 2



ภาพที่ 2: ขั้นตอนการทำงานของ SMS Spam filter

จากภาพที่ 2 มีรายละเอียดการทำงานตามลำดับดังนี้

- ถอดข้อความจาก SMPP Message
- ทำกระบวนการ Text Normalization โดยการ
  - ลบสัญลักษณ์พิเศษต่างๆ
  - ลบตัวเลข
  - แก้ไขการพิมพ์สระหรือวรรณยุกต์ตัวเดียวกันมากกว่า 1 ครั้ง
  - แก้ไขตำแหน่งของสระและวรรณยุกต์ให้มีความถูกต้อง
- ตัดคำด้วยวิธีการแบบสอดคล้องมากที่สุด
- นำคำแรกและคำสุดท้ายของข้อความ ไปตัดคำอีกครั้งด้วยวิธีการทางสถิติ และตรวจสอบความถูกต้อง
- นำคำทั้งหมดเข้าสู่กระบวนการกรองข้อความ
- จัดส่งข้อความที่ผ่านการตรวจสอบให้กับผู้รับ และบันทึกข้อความที่ไม่ผ่านการตรวจสอบลงในฐานข้อมูล เพื่อใช้ในการเรียนรู้ต่อไป

4. การวัดประสิทธิภาพและผลการทดสอบ

การทดสอบกระทำบนเครื่อง Computer Server ที่ติดตั้งระบบปฏิบัติการ Windows Server 2003 CPU Dual core 2.4 GHz Ram 3.0 GB กับชุดข้อความจำนวน 2 ชุด ได้แก่ ชุดข้อความที่ใช้ทำการฝึกสอน Training Data (TD) และ ข้อความ SMS ชุดใหม่ New Data (ND)

4.1 การปรับปรุงวิธีการทำ Text Normalization

จากการปรับปรุงการทำ Text Normalization ที่ใช้ในการแปลงข้อความ SMS จำนวน 18111 ข้อความ พบว่าวิธีการแบบปรกติ ตรวจพบคำจำนวน 10188 คำ และวิธีการที่ปรับปรุง ตรวจพบคำจำนวน 9219 คำ ช่วยลดจำนวนคำที่ไม่จำเป็นต่อการคำนวณได้ร้อยละ 9.51

4.2 การเปรียบเทียบระหว่างวิธี SVM และ NB

การทดสอบกรองข้อความระหว่างวิธีกรองข้อความ Spam ด้วยวิธีการแบบ SVM และ NB แสดงให้เห็นว่าวิธีการแบบ SVM มีความถูกต้องเฉลี่ยที่ร้อยละ 90 ในขณะที่การกรองแบบ NB มีความถูกต้องเฉลี่ยที่ร้อยละ 82.2 โดยมีรายละเอียดการทดสอบตามตารางที่ 5 และ 6

ตารางที่ 5: เปรียบเทียบการกรองระหว่าง SVM และ NB

Algorithm	Data	Maximal Matching word segmentation			
		SMS	correct	correct (%)	Averaged processing Time per SMS (millisecond)
SVM	TD	1204	1157	96.09634551	102.8
	ND	540	459	85	117.6
NB	TD	1204	991	82.3089701	38.3
	ND	540	444	82.22222222	38.2

ตารางที่ 6: ตัวอย่างผลของการกรองข้อความ

SMS	Classify	Correction
SVM with Maximal Matching word segmentation		
บันโรคฮิตเลยนี.	NORMAL	NORMAL
พระเจ้า พลาด?	SPAM	NORMAL
โอกาสสุดท้ายจาก Fitness First ถึงอาทิตย์ที่ 17 กุมภาฯนี้ สำหรับสมาชิ	NORMAL	SPAM
โชร์SMSนี้ภายใน31มี.ค. ให้คุณ&เพื่อน3คนรับสมาชิก1เดือนเพียง399บ พร้	SPAM	SPAM
NB with Maximal Matching word segmentation		
พักแล้วนะ!บาย!	NORMAL	NORMAL
เห็นข้อความละโทกับด่วน	SPAM	NORMAL
วิลด์ อัลไลแอนซ์ 0819477355,026454455	NORMAL	SPAM
แลกของรางวัลที่กรุงศรี ATM ถึง 31 มี.ค.51 สอบถามโทร. 1572	SPAM	SPAM

จากตารางที่ 6 แสดงผลของการกรองข้อความ SMS ในช่อง classify เปรียบเทียบกับชนิดของข้อความที่แท้จริงด้วยการตัดสินของมนุษย์ในช่อง correction ข้อความที่กรองผิดพลาดส่วนใหญ่คือข้อความที่ไม่ใช่

Spam แต่มีจำนวนคำน้อย ซึ่งคำส่วนใหญ่เป็นคำที่เกิดขึ้นแบบ Spam ในชุดข้อมูลฝึกสอนมากกว่าข้อความปรกติ และ ข้อความ Spam ที่มีเนื้อความมากกว่า 1 ข้อความ ซึ่งเนื้อความที่ไม่ครบถ้วนนี้ทำให้ไม่สามารถกรองข้อความได้อย่างถูกต้อง

4.3 การเปรียบเทียบระหว่าง SVM และ SVM ที่มีการตรวจสอบคำแรกของข้อความ

การทดสอบกรองข้อความจำนวน 2 ชุดข้อมูลระหว่างวิธีกรองข้อความ Spam แบบ SVM และ SVM ที่เพิ่มการตรวจสอบคำแรกของข้อความด้วยวิธีการตัดคำแบบผสม พบว่า การตัดคำแบบผสมช่วยเพิ่มความถูกต้องได้ เช่น การกรองข้อความ

ับฟรี LCD TV + Home Theatre 02-869-7788

ซึ่งการตัดคำแบบสอดคล้อง ตัดคำว่า **ฟรี** ผิดพลาด และส่งผลทำให้การกรองข้อความผิดพลาดตามไปด้วย ในขณะที่วิธีเพิ่มการตรวจสอบคำแรกของข้อความสามารถตัดคำได้ถูกต้อง ทำให้การกรองข้อความมีความถูกต้อง โดยมีรายละเอียดตามตารางที่ 7

ตารางที่ 7: เปรียบเทียบการกรองแบบ SVM ที่ปรับปรุง

Algorithm	Data	Maximal Matching words segmentation			
		SMS	correct	correct (%)	Averaged processing Time per SMS (millisecond)
SVM	TD	1204	1157	96.09634551	102.8
	ND	540	459	85	117.6
		Mixing words segmentation			
SVM	TD	1204	1162	96.51162791	130.5
	ND	540	459	85	150.3

5. สรุปผลการวิจัยและข้อเสนอแนะ

จากผลการทดสอบแสดงให้เห็นว่า วิธีการกรองแบบ SVM มีความแม่นยำมากกว่าวิธีการกรองแบบ NB แต่ในทางกลับกัน วิธีการแบบ NB ต้องการเวลาในการประมวลผลน้อยกว่าวิธีการกรองแบบ SVM การปรับปรุงวิธีทำ Text Normalization ช่วยลดคำผิดในฐานข้อมูลลงได้ และการใช้วิธีตัดคำแบบผสมสามารถเพิ่มความถูกต้องการตรวจสอบคำแรกและคำสุดท้ายได้ แต่ขั้นตอนดังกล่าวใช้เวลาในการคำนวณมากกว่า อีกทั้งผลการกรองยังมีความถูกต้องเพิ่มขึ้นไม่มากนัก จำเป็นต้องมีการศึกษาและพัฒนาเพิ่มเติม

6. เอกสารอ้างอิง

[1] S. Dixit, S. Gupta, and C.V. Ravishankar - LOHIT: An Online Detection & Control System for Cellular SMS Spam. - Proceeding (499) Communication, Network, and Information Security – 2005

[2] Giovanni Camponovo, Davide Cerutti - The spam issue in mobile business a comparative regulatory overview - Proceedings of the Third International Conference on Mobile Business - M-Business - 2004

[3] http://smsforum.net/ - SMPP Protocol pdf file

[4] Gordon V. Cormack, Jose Maria Gomez Hidalgo, Enrique Puertas Sanz - Spam filtering for short messages. - Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - 2007

[5] http://osbf-lua.luaforge.net/

[6] Istvan Pilaszy - Text Categorization and Support Vector - Machines - Department of Measurement and Information Systems - Budapest University of Technology and Economics - 2005

[7] Canasai Kruengkrai, Chuleerat Jaruskulchai - Thai Text Document Clustering using Parallel Spherical K-Means Algorithm on PIRUN Linux Cluster - Intelligent Information Retrieval and Database Laboratory Department of Computer Science, Faculty of Science Kasetsart University, Bangkok, 10900, Thailand - 2001

[8] http://wiki.nectec.or.th/ntl/Project/MobilityClassification - Classification: Naive Bayes Model - nectec

[9] http://en.wikipedia.org/wiki/Naive_Bayesian_classification#The_naive_Bayes_probabilistic_model - Naive Bayes classifier – Wikipedia

[10] http://guru.sanook.com/encyclopedia/%E2%BB%C3%E1%A1%C3%C1%B5%D1%B4%A4%D3%C0%D2%C9%D2%E4%B7%C2_(Thai_Word__Segmentation) - สารานุกรมไทยสำหรับเยาวชนฯ เล่มที่ 25

[11] กานดา รุณนะพงศา, ปโยธร อุราธรรมกุล - ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น - การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่ – 2006

ภาคผนวก ข.

เอกสาร แบบสอบถาม แสดงความคิดเห็นเกี่ยวกับข้อความ Spam ทาง SMS

แบบสัมภาษณ์แสดงความคิดเห็นเกี่ยวกับข้อความ Spam ทาง SMS

หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

คำชี้แจง โปรดทำเครื่องหมาย ✓ ลงใน ❑ ที่ตรงกับความเป็นจริงของท่านมากที่สุด หรือเติมข้อความลงในช่องว่าง

ส่วนที่ 1 ข้อมูลทั่วไป

1. เพศ / อายุ

❑ ชาย ❑ หญิง

❑ ต่ำกว่า 20 ปี ❑ 21-30 ปี

❑ 31-40 ปี ❑ 41 ปีขึ้นไป

2. สถานภาพ

❑ โสด ❑ สมรส ❑ หม้ายหรือหย่า

3. อาชีพ

❑ นักเรียน / นักศึกษา ❑ พนักงานบริษัท/ธุรกิจส่วนตัว

❑ รับราชการ / รัฐวิสาหกิจ ❑ อื่นๆโปรดระบุ.....

4. จำนวนโทรศัพท์เคลื่อนที่ที่ใช้งานพร้อมกัน

❑ 1 เครื่อง ❑ 2 เครื่อง ❑ มากกว่า 2 เครื่อง โปรดระบุ.....

5. จำนวนข้อความ SMS ที่ได้รับโดยเฉลี่ยใน 1 วัน

❑ ไม่มี ❑ น้อยกว่า 3 ข้อความ

❑ น้อยกว่า 5 ข้อความ ❑ มากกว่า 5 ข้อความ โปรดระบุ.....

6. จำนวนข้อความ SMS ที่ส่งไปยังหมายเลขอื่นโดยเฉลี่ยใน 1 วัน

❑ ไม่มี ❑ น้อยกว่า 3 ข้อความ

❑ น้อยกว่า 5 ข้อความ ❑ มากกว่า 5 ข้อความ โปรดระบุ.....

ส่วนที่ 2 ข้อมูลของข้อความ Spam ในระบบ SMS

1. ความหมายของข้อความ Spam ทาง SMS (ตอบได้มากกว่า 1 ข้อ)

❑ ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ

❑ ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่

❑ ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้

❑ ข้อความหยาบคาย หรือข้อความที่ไม่มีสาระสำคัญ

❑ ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง

❑ อื่นๆ โปรดระบุ.....

.....

2. ความถี่ของข้อความ Spam ที่ท่านได้รับในแต่ละวัน

- ❑ ไม่มี
- ❑ น้อยกว่า 3ข้อความ
- ❑ น้อยกว่า 5 ข้อความ
- ❑ มากกว่า 5 ข้อความ โปรดระบุ.....

3. ภาษาของข้อความ Spam ที่ท่านได้รับ

- ❑ ภาษาอังกฤษเพียงอย่างเดียว
- ❑ ภาษาไทยหรือภาษาไทยปนภาษาอังกฤษ

4. คำใดบ้างที่ท่านคิดว่าจะ**พบ**ในข้อความ Spam (เช่น ฟรี, ดาวน์โหลด เป็นต้น)

.....

.....

5. ช่วงเวลาที่ท่านได้รับข้อความ Spam

.....

6. ผลกระทบของข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)

- ❑ ทำให้แบตเตอรี่หมดเร็วขึ้น
- ❑ ก่อความรำคาญและทำให้ใช้งานไม่สะดวก
- ❑ ถูกละเมิดสิทธิส่วนบุคคล
- ❑ เสียค่าบริการจากโฆษณาในข้อความ Spam เพิ่มขึ้น
- ❑ อื่นๆ โปรดระบุ.....

7. วิธีการแก้ปัญหาข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)

- ❑ ไม่ดำเนินการใดๆ
- ❑ รับข้อความและเปิดอ่านตามปกติ
- ❑ ลบข้อความทิ้งทันที
- ❑ ปิดเครื่องโทรศัพท์ทันที
- ❑ แจ้งผู้ให้บริการโทรศัพท์เคลื่อนที่เพื่อให้ดำเนินการแก้ไข
- ❑ เลือกใช้โทรศัพท์เคลื่อนที่ที่สามารถกรองข้อความ Spam ได้
- ❑ อื่นๆ โปรดระบุ.....

8. ตัวอย่างข้อความ Spam ทาง SMS ที่พบในชีวิตประจำวัน

.....

.....

.....

9. ข้อเสนอแนะวิธีการจัดการกับปัญหาข้อความ Spam ในระบบ SMS

.....

.....

ขอขอบคุณที่ท่านให้ความร่วมมือในการตอบแบบสัมภาษณ์ด้วยดี

ภาคผนวก ค.

แบบสอบถาม แสดงความคิดเห็นเกี่ยวกับข้อความ Spam ทาง SMS Online

แบบสอบถามความคิดเห็นเกี่ยวกับข้อความ Spam ทาง SMS

Spam SMS คือข้อความที่ก่อความรำคาญแก่ผู้รับ

หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

คำชี้แจง โปรดทำเครื่องหมายลงในช่อง ☐ ○ หรือ เติมข้อความลงในช่องว่าง ที่ตรงกับความเป็นจริงมากที่สุด

ส่วนที่ 1 ข้อมูลทั่วไป

1. เพศ ○ชาย ○หญิง
2. อายุ ○ต่ำกว่า 20 ปี ○21-30 ปี ○31-40 ปี ○41 ปีขึ้นไป
3. สถานภาพ ○โสด ○สมรส ○หม้ายหรือหย่า
4. อาชีพ
   - ○นักเรียน / นักศึกษา
   - ○รับราชการ / รัฐวิสาหกิจ
   - ○พนักงานบริษัท / ธุรกิจส่วนตัว
   - ○อื่นๆ
5. จำนวนโทรศัพท์เคลื่อนที่ที่ใช้งานพร้อมกัน
   - ○1 เครื่อง
   - ○2 เครื่อง
   - ○มากกว่า 2 เครื่องโปรดระบุ
6. จำนวนข้อความ SMS ที่ได้รับโดยเฉลี่ยใน 1 วัน
   - ○ไม่มี
   - ○น้อยกว่า 3 ข้อความ
   - ○น้อยกว่า 5 ข้อความ
   - ○มากกว่า 5 ข้อความ โปรดระบุ
7. จำนวนข้อความ SMS ที่ส่งไปยังหมายเลขอื่นโดยเฉลี่ยใน 1 วัน
   - ○ไม่มี
   - ○น้อยกว่า 3 ข้อความ
   - ○น้อยกว่า 5 ข้อความ
   - ○มากกว่า 5 ข้อความ โปรดระบุ

ส่วนที่ 2 ข้อมูลของข้อความ Spam ในระบบ SMS

1. ความหมายของข้อความ Spam ทาง SMS (ตอบได้มากกว่า 1 ข้อ)
   - ☐ข้อความโฆษณาประชาสัมพันธ์ การขายสินค้า หรือมีการเสนอของรางวัลโดยมีเงื่อนไขต่างๆ
   - ☐ข่าวสารทั่วไป หรือข้อความแจ้งข้อมูลจากผู้ให้บริการโทรศัพท์เคลื่อนที่
   - ☐ข้อความจากผู้ที่ท่านไม่รู้จัก หรือไม่สามารถระบุที่มาของผู้ส่งได้
   - ☐ข้อความหยาบคาย หรือข้อความที่ไม่มีสาระสำคัญ
   - ☐ข้อความเดิมหรือข้อความที่มีความหมายใกล้เคียงกันที่ส่งซ้ำหลายครั้ง
   - ☐อื่นๆโปรดระบุ
2. ความถี่ของข้อความ Spam ที่ท่านได้รับในแต่ละวัน
   - ○ไม่มี
   - ○น้อยกว่า 3 ข้อความ
   - ○น้อยกว่า 5 ข้อความ
   - ○มากกว่า 5 ข้อความ โปรดระบุ
3. ภาษาของข้อความ Spam ที่ท่านได้รับ
   - ○ภาษาอังกฤษเพียงอย่างเดียว
   - ○ภาษาไทยหรือภาษาไทยปนภาษาอังกฤษ
4. คำใดบ้างที่ท่านคิดว่าจะ**พบ**ในข้อความ Spam (เช่น ฟรี, ดาวน์โหลด เป็นต้น)
5. ช่วงเวลาที่ท่านได้รับข้อความ Spam
6. ผลกระทบของข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)
   - ☐ทำให้แบตเตอรี่หมดเร็วขึ้น
   - ☐ก่อความรำคาญและทำให้ใช้งานไม่สะดวก
   - ☐ถูกละเมิดสิทธิส่วนบุคคล
   - ☐เสียค่าบริการจากโฆษณาในข้อความ Spam เพิ่มขึ้น
   - ☐เสียพื้นที่ในการเก็บข้อความที่จำเป็น (Inbox เต็ม)
   - ☐อื่นๆโปรดระบุ
7. วิธีการแก้ปัญหาข้อความ Spam (ตอบได้มากกว่า 1 ข้อ)
   - ☐ไม่ดำเนินการใดๆ

- ☐ รับข้อความและเปิดอ่านตามปกติ
- ☐ ลบข้อความทิ้งทันที
- ☐ ปิดเครื่องโทรศัพท์ทันที
- ☐ แจ้งผู้ให้บริการโทรศัพท์เคลื่อนที่เพื่อให้ดำเนินการแก้ไข
- ☐ เลือกใช้โทรศัพท์เคลื่อนที่ที่สามารถกรองข้อความ Spam ได้
- ☐ อื่นๆโปรดระบุ

8. ตัวอย่างข้อความ Spam ทาง SMS ที่พบในชีวิตประจำวัน

9. ข้อเสนอแนะวิธีการจัดการกับปัญหาข้อความ Spam ในระบบ SMS

ส่งความคิดเห็น

ภาคผนวก ฆ.

โปรแกรมกรองข้อความ SMS แบบ Online

Spam Message Checker

type some message th or en in the box (more than 5 character)

select Algorithm : support vector Check

Example Message From CAT SMSC (Human Classify)

#	Spam	Normal
1	รับฟรี!! วิธีอยู่แบบไม่มีหนี้ตลอดชีวิต สนใจ กด*48253320012 โทรออก	พรุ่งนี้เอาสายเกียร์ออโต้2เส้นไปให้พ่อด่วน
2	รับข้อเสนอMotorExpoก่อนใคร ถอย MazdaBT50ดอกเบี้ย0%ฟรีประกัน โทร026619880	ฝันดีนะเทอ!รักเทอมากนะ รักกัลบ้างป่าว?~รักกัลไปนานๆนะ*คิดถึงมากมาย
3	คอร์สพิเศษ!90วันเปลี่ยนชีวิตสู่ธุรกิจทำเงินที่มั่นคงโทร.0866825974 ฟรี	ส่งช้าไปมั้ย นอนยัง ?!? พึ่งท่องเสดค่ะ รักสุดขอบโลกเลย > < จุ๊บส์~!! *
4	คืนนี้ ฟรีคอนเสิร์ต * ตั๊กแตน ชลดา * @สภาดินโทร 0863676676	nuttcg2@gmail.com
5	หมอกฤษฎ์ คอนเฟิร์มแม่นรับดวงฟรี7วัน(3บ/SMS)โทร*299*001#สอบถาม025026767	Call me back now!!

ประวัติผู้เขียน

ชื่อ-นามสกุล	นาย นนท์ บุญนิธิประเสริฐ
วัน เดือน ปีเกิด	29 มกราคม 2522
ประวัติการศึกษา	สำเร็จการศึกษาระดับปริญญาตรี จากคณะศิลปกรรมศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ (ประสานมิตร) ปีการศึกษา 2544
ตำแหน่งและสถานที่ทำงานปัจจุบัน	ดูแลรับผิดชอบระบบบริการเสริมบนเครือข่าย โทรศัพท์เคลื่อนที่ภายใต้ชื่อผลิตภัณฑ์ CAT CDMA บริษัท กสท โทรคมนาคม จำกัด (มหาชน)