เหมืองข้อความและการประยุกต์ใช้

TEXT MINING AND ITS APPLICATIONS

วรรณวิภา วงศ์วิไลสกุล[1]

บทคัดย่อ

เหมืองข้อความ เป็นเทคนิคการค้นหาความรู้ใหม่จากข้อมูลประเภทข้อความที่มีปริมาณมากโดยอัตโนมัติ โดยการสกัดคำ ค้นหารูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความเอกสาร เพื่อให้เกิดความหมายและสามารถนำมาใช้ประโยชน์ได้ โดยสถาปัตยกรรมระบบของเหมืองข้อความมีความคล้ายคลึงกับสถาปัตยกรรมระบบในการค้นหาความรู้ในฐานข้อมูลประเภทข้อความ ซึ่งประกอบด้วย ขั้นตอนการเตรียมข้อมูล ขั้นตอนการค้นหากฎความสัมพันธ์ และขั้นตอนการนำเสนอผลลัพธ์ที่ชัดเจน ส่วนประโยชน์ของเหมืองข้อความนั้น คือการทำให้สารสนเทศในรูปแบบข้อความสามารถเข้าถึง วิเคราะห์ และประมวลผลภายในเวลาอันรวดเร็ว โดยผู้ใช้สามารถเข้าใจและใช้ประโยชน์จากเอกสารที่จัดเก็บไว้เป็นจำนวนมากได้อย่างมีประสิทธิภาพ แนวทางสำคัญในการประยุกต์ใช้เหมืองข้อความ คือ การประยุกต์ใช้ตามหน้าที่ และตามกลุ่มการใช้งาน และขณะนี้ได้มีการพัฒนาเหมืองข้อความขึ้นเป็นเหมืองข้อความแสดงความคิดเห็น ซึ่งผสานเทคนิคของการสืบค้นข้อมูลเข้ากับการประมวลผลภาษาธรรมชาติ ทำให้สามารถสรุปความคิดเห็นที่หลากหลายบนเครือข่ายสังคมออนไลน์ให้เข้าใจง่ายยิ่งขึ้น

คำสำคัญ : เหมืองข้อความ เหมืองข้อมูล การค้นหาความรู้ ฐานข้อมูลประเภทข้อความ

Abstract

Text Mining is a new technique of a search engine done by automatically going through large quantity of information. This is done by extracting an information, finding pattern, as well as hidden meaning and relation within the set documents. This makes it meaningful and very beneficial. The system architecture of text mining is similar to the one of Knowledge Discovery in Textual Databases. This is including Text Preprocessing phase, Association Rule Mining phase and Visualization phase. The benefit of text mining is to help improve information technology in the form of text to be easier accessible in an analytical way. The users will be able to understand and benefit from this efficiently. This will be done with the use of large number of collected documents. The main approach to text mining is to apply it according to function and usability. Nowadays, text mining has been developed into an opinion mining. This combines the mean of retrieving information with Natural Language Processing (NLP). Therefore, this will be easier especially with different opinion from an online social network.

Keywords : Text Mining, Data Mining, Knowledge Discovery, Textual Databases

[1] หัวหน้าสาขาวิชาเทคโนโลยีสารสนเทศ คณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์
E-mail: *wanvipawon@pim.ac.th*

บทนำ

ปัจจุบันการเติบโตของสารสนเทศในรูปแบบข้อความ (Textual Information) มีอัตราที่สูงขึ้น ทั้งจากเครือข่ายอินเทอร์เน็ตและจากเอกสารที่องค์กรจัดเก็บไว้ แม้ว่าเทคโนโลยีการสืบค้น (Search Engine) จะมีขีดความสามารถที่สูงขึ้น โดยสามารถค้นหาด้วยคำสำคัญ (Keywords) และให้ผลลัพธ์อย่างรวดเร็ว แต่วิธีดังกล่าวก็ยังไม่สามารถสรุปความ ประมวลความหมาย หรือความสัมพันธ์ของคำได้อย่างตรงประเด็น ทำให้การค้นคืนเอกสารไม่ตรงกับความต้องการของผู้ใช้ ดังนั้น จึงได้มีการพัฒนาเทคนิคเหมืองข้อความ เพื่อช่วยจัดการกับปัญหาทั้งในส่วนการตรวจสอบหัวข้อ การติดตาม และการแสดงแนวโน้ม ทำให้สารสนเทศในรูปแบบข้อความสามารถเข้าถึง วิเคราะห์ และประมวลผลภายในเวลาอันรวดเร็ว ผู้ใช้จึงสามารถเข้าใจหรือใช้ประโยชน์จากข้อความเอกสารที่จัดเก็บไว้เป็นจำนวนมากได้อย่างมีประสิทธิภาพ

บทความนี้จัดทำขึ้นเพื่ออธิบายแนวคิดของการทำเหมืองข้อความในการจัดการข้อความเอกสาร ประกอบด้วย ความหมายของเหมืองข้อความ สถาปัตยกรรมระบบของเหมืองข้อความ ประโยชน์ของเหมืองข้อความ การประยุกต์ใช้เหมืองข้อความโดยการจัดแบ่งตามหน้าที่และตามกลุ่มการใช้งาน บทสรุปและแนวโน้มในอนาคต ดังรายละเอียดต่อไปนี้

ความหมายของเหมืองข้อความ

เหมืองข้อความ (Text Mining) ถือเป็นสาขาวิจัยใหม่ที่น่าสนใจซึ่งพยายามแก้ปัญหาภาวะท่วมท้นของสารสนเทศ (Information Overload) โดยใช้แขนงความรู้ในด้านต่างๆ อาทิ การทำเหมืองข้อมูล (Data Mining) การเรียนรู้เครื่องจักร (Machine Learning) การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การค้นคืนสารสนเทศ (Information Retrieval) และการจัดการความรู้ (Knowledge Management) โดยอาศัยกระบวนการรวบรวมเอกสาร การวิเคราะห์ข้อความเอกสาร และการนำเสนอผลลัพธ์ให้มีความชัดเจนมากยิ่งขึ้น (Feldman, Ronen & Sanger, James, 2007)

เหมืองข้อความ ถือเป็นรูปแบบหนึ่งของการทำเหมืองข้อมูลและเป็นส่วนหนึ่งขององค์ความรู้ในด้านการประมวลผลภาษาธรรมชาติ โดยศึกษาในด้านการค้นคืนเอกสาร การเรียนรู้เครื่องจักร สถิติ และการประมวลผลทางภาษา โดยเน้นการทำงานด้านการจัดประเภทเอกสาร (Text Categorization) การจัดกลุ่มเอกสาร (Text Clustering) การสกัดความรู้ (Extraction) การวิเคราะห์ความรู้สึก (Sentiment Analysis) การสรุปเอกสาร (Document Summarization) และการหาความสัมพันธ์ (Entity Relation Modeling) ซึ่งการประมวลผลภาษาธรรมชาติถือเป็นแขนงหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence: AI) โดยมุ่งศึกษาปัญหาและทำความเข้าใจกับภาษามนุษย์ โดยใช้องค์ความรู้ด้านภาษา หลักไวยากรณ์ภาษา และหลักสถิติ (Saša Petrovic, 2007)

เหมืองข้อความ อาจเรียกว่า เหมืองข้อมูลประเภทข้อความ (Text Data Mining) หรือการค้นหาความรู้ในฐานข้อมูลประเภทข้อความ (Knowledge Discovery in Textual Databases: KDT) ซึ่งหมายถึง กระบวนการสกัดสิ่งที่น่าสนใจและสิ่งที่ไม่เคยทราบมาก่อนจากคลังข้อความเอกสาร เพื่อให้ได้ความรู้ใหม่ที่เป็นประโยชน์ (Mahesh, T.R, 2009)

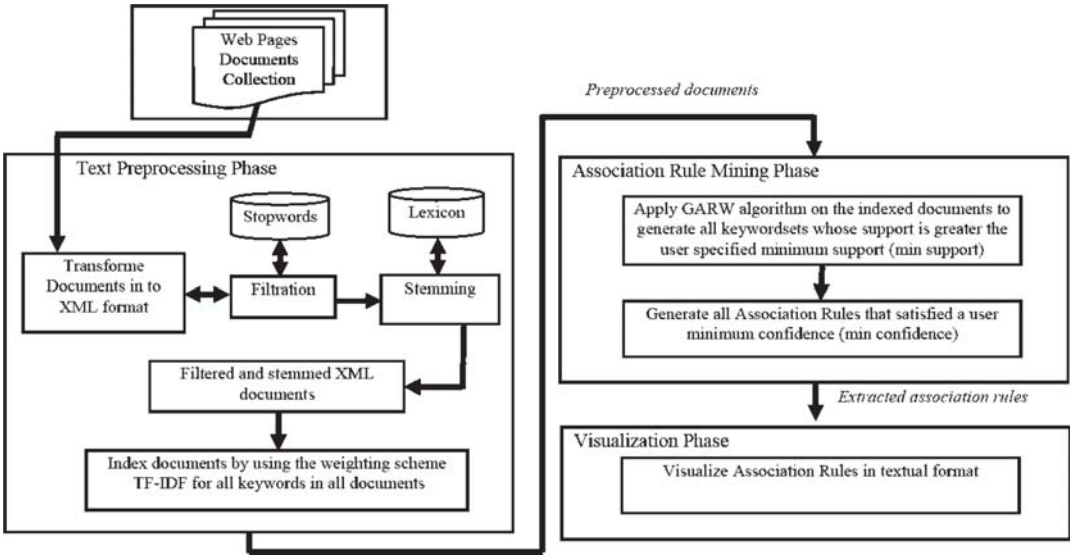
เหมืองข้อความ หมายถึง วิธีการที่ใช้สกัดสารสนเทศที่มีคุณค่าและการค้นหากฎความสัมพันธ์ของข้อความจำนวนมาก โดยการวิเคราะห์ข้อมูลจากชุดข้อมูลที่ไม่มีโครงสร้างหรือกึ่งโครงสร้าง ให้สามารถนำมาใช้ในการทำนายเหตุการณ์หรือจัดกลุ่มของข้อมูลได้อย่างมีประสิทธิภาพ (Vishwadeepak Singh Baghela, 2012)

เหมืองข้อความ เป็นระบบที่สามารถทำงานอย่างอัตโนมัติในการระบุประเด็นที่สนใจจากคลังข้อความขนาดใหญ่ ซึ่งจะช่วยให้ผู้ใช้งานเกิดความเข้าใจและสามารถใช้ประโยชน์จากเอกสารที่จัดเก็บไว้ (Ashok Srivastava, 2009)

จากนิยามข้างต้นสรุปได้ว่า เหมืองข้อความ คือ เทคนิคการค้นหาความรู้ใหม่จากข้อมูลประเภทข้อความที่มีปริมาณมากโดยอัตโนมัติ โดยใช้การสกัดคำ ค้นหารูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความเอกสาร เพื่อให้เกิดความหมายและสามารถนำมาใช้ประโยชน์ได้

สถาปัตยกรรมระบบของเหมืองข้อความ

กระบวนการทำเหมืองข้อความประกอบด้วยหลายขั้นตอน ตั้งแต่การนำข้อความปริมาณมากมาจัดกลุ่มและกำหนดคุณลักษณะของข้อความ จากนั้นทำการตัดส่วนที่ไม่สำคัญออกจากชุดข้อความและแยกข้อความออกจากกัน แล้วจึงนำมาวิเคราะห์โดยใช้เทคนิคต่างๆ ดังตัวอย่างที่ Hany Mahgoub (2008) ได้อธิบายถึงการทำเหมืองข้อความโดยใช้เทคนิคกฎความสัมพันธ์ (Association Rules) ซึ่งประกอบด้วย 3 ขั้นตอน ดังภาพที่ 1



ภาพที่ 1 : สถาปัตยกรรมระบบของเหมืองข้อความ (Hany Mahgoub, 2008)

จากภาพจะเห็นว่าระบบเหมืองข้อความเริ่มต้นจากการเลือกเอกสารที่ต้องการจากเว็บไซต์ หรือ จากระบบไฟล์ข้อมูลในองค์กร จากนั้นเข้าสู่กระบวนการทำงาน 3 ขั้นตอน ได้แก่

1) ขั้นตอนการเตรียมข้อมูล (Text Preprocessing phase)

1.1 การแปลงข้อมูล (Transformation) เป็นการเปลี่ยนรูปแบบเอกสารที่มีไฟล์นามสกุลอยู่หลากหลายให้กลายเป็นรูปแบบมาตรฐาน เช่น XML เพื่อให้เอกสารอยู่ในรูปแบบเดียวกัน

1.2 การกรองข้อมูล (Filtration) เป็นการสกัดคุณลักษณะของข้อความ (Text Extraction) โดยคัดเลือกเฉพาะส่วนที่ต้องการด้วยคำเดี่ยว พยางค์ วลี กลุ่มคำ หรือประโยค มาเป็นตัวแทนของเอกสาร และกำจัดคำหยุด (Stopword) โดยตัดคำที่ไม่มีนัยสำคัญต่อเอกสารและไม่ทำให้ใจความของเอกสารเปลี่ยน เช่น คำสรรพนาม คำเชื่อม เป็นต้น

1.3 การหารากศัพท์ (Stemming) เป็นการจัดกลุ่มคำหลายคำที่มีความหมายคล้ายคลึงกันหรือใกล้เคียงกันให้อยู่ในกลุ่มเดียวกัน

1.4 การสร้างดัชนี (Indexing) เป็นการนำเอกสารในรูปแบบมาตรฐาน เช่น XML ที่ได้คัดเลือกคุณสมบัติแล้วมาสร้างดัชนีแบบถ่วงน้ำหนัก โดยสามารถดำเนินการสร้างโดยผู้ใช้เองหรือให้ระบบดำเนินการให้โดยอัตโนมัติ

2) การใช้อัลกอริธึม (Algorithm) ในการค้นหาสารสนเทศที่ต้องการจากดัชนีของเอกสารทั้งหมดด้วยเทคนิคต่างๆ จากภาพเป็นขั้นตอนการค้นหากฎความสัมพันธ์ (Association Rule Mining Phase) ซึ่งเป็นวิธีที่สามารถหาความสัมพันธ์ของข้อความให้เป็นไปตามเกณฑ์ของค่าความเชื่อมั่นขั้นต่ำที่ผู้ใช้กำหนด

3) ขั้นตอนการนำเสนอผลลัพธ์ที่ชัดเจน (Visualization Phase) เป็นการสร้างภาพคอมพิวเตอร์กราฟิกที่สามารถนำเสนอข้อมูลที่ต้องการได้อย่างครบถ้วนแทนการใช้ข้อความ ซึ่งจะช่วยให้ผู้ใช้สามารถตีความสารสนเทศที่ได้รับอย่างมีประสิทธิภาพมากยิ่งขึ้น

ประโยชน์ของเหมืองข้อความ

Diane McDonald (2012) ได้อธิบายถึงประโยชน์ของการทำเหมืองข้อความไว้ดังนี้

1) *ช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล* เนื่องจากการทำเหมืองข้อมูลอาศัยองค์ความรู้ในหลายด้านมาช่วยในการวิเคราะห์ เช่น ด้านคณิตศาสตร์ สถิติ ความน่าจะเป็น ปัญญาประดิษฐ์ การค้นคืนเอกสารและฐานข้อมูล เป็นต้น อีกทั้งหลายปีที่ผ่านมามีงานวิจัยจำนวนมากที่ทำการค้นหาอัลกอริธึมและการนำเสนอข้อมูลที่เหมาะสม จึงทำให้ผลการวิเคราะห์ข้อมูลมีประสิทธิภาพมากยิ่งขึ้น ผู้ใช้จึงสามารถวิเคราะห์เนื้อหาจากข้อความเอกสารที่ถูกรวบรวมไว้ได้อย่างมีประสิทธิภาพ

2) *ปลดล็อกข้อมูลที่ถูกซ่อนอยู่ในฐานข้อมูลขนาดใหญ่และพัฒนาให้เกิดเป็นความรู้ใหม่* โดยผู้ใช้สามารถเข้าถึงเนื้อหาเอกสารที่เป็นภาษาเขียนได้มากยิ่งขึ้น เช่น ข่าวในหน้าหนังสือพิมพ์ บทความในนิตยสาร อีเมล บล็อก และข้อความออนไลน์ เป็นต้น เนื่องจากข้อมูลเหล่านี้ นับวันจะมีปริมาณมากยิ่งขึ้น การใช้เหมืองข้อความจะช่วยสกัดคุณลักษณะข้อความเอกสารให้ตรงตามความสนใจของผู้ใช้ และสามารถแสดงผลทางหน้าจอในรูปแบบกราฟิก จากนั้นผลลัพธ์จึงถูกนำมาตีความโดยผู้ใช้ในภายหลัง

3) *ปรับปรุงกระบวนการวิจัยให้มีคุณภาพ* เนื่องจากมีการใช้โปรแกรมคอมพิวเตอร์ในการอ่านข้อความเอกสารและการประมวลผลสามารถกระทำได้หลายรูปแบบ ผู้ใช้สามารถกำหนดอัลกอริธึมที่เหมาะสมและมีประสิทธิภาพมากที่สุด ทำให้ประมวลผลในเวลาอันรวดเร็ว และได้ผลลัพธ์ที่ต้องการ

4) *เกิดผลประโยชน์ทางเศรษฐกิจที่กว้างขึ้น* ผู้ใช้สามารถประหยัดค่าใช้จ่ายในการผลิตและสร้างกำไรจากการพัฒนานวัตกรรมบริการใหม่ รวมถึงก่อให้เกิดรูปแบบธุรกิจใหม่ที่สร้างมูลค่าเพิ่มให้แก่องค์กร

โดยสรุปแล้วเหมืองข้อความทำให้สารสนเทศในรูปแบบข้อความสามารถเข้าถึง วิเคราะห์ และประมวลผลภายในเวลาอันรวดเร็ว โดยผู้ใช้สามารถเข้าใจและใช้ประโยชน์จากเอกสารที่จัดเก็บไว้เป็นจำนวนมากได้อย่างมีประสิทธิภาพ

การประยุกต์ใช้เหมืองข้อความ

ในยุคแรกๆ เหมืองข้อความได้ถูกนำมาใช้เป็นเครื่องมือในการสืบค้น โดยผู้ใช้สามารถพิมพ์คำหรือข้อความที่ต้องการค้นหาก็จะได้เอกสารที่เกี่ยวข้องออกมา ต่อมาได้มีการนำวิธีการนี้ไปใช้ในการลงโค้ดข้อมูลแบบสอบถามที่เป็นคำถามปลายเปิดได้โดยอัตโนมัติ หรือนำมาใช้วิเคราะห์สแปมเมล โดยการวิเคราะห์จากชื่อเรื่องและเนื้อหา จนถึงปัจจุบันมีการใช้เหมืองข้อความในองค์กรต่างๆ กันแพร่หลายมากยิ่งขึ้น

Sergio Bolasco (2002) ได้ศึกษาแนวทางในการประยุกต์ใช้เหมืองข้อความ โดยแบ่งเป็น 2 ลักษณะ คือ การประยุกต์ใช้ตามหน้าที่ และตามกลุ่มการใช้งาน รายละเอียดมีดังนี้

1) การประยุกต์ใช้เหมืองข้อความตามหน้าที่ มี 4 ประเภท ได้แก่

1.1 การจัดการความรู้ (KM) และทรัพยากรมนุษย์ (HR)

- CI-Competitive Intelligence ใช้ในการรวบรวมข้อมูลขององค์กร ตลาด และคู่แข่งขัน และจัดการข้อมูลจำนวนมากเพื่อใช้ในการวิเคราะห์และ

วางแผน โดยการเลือกอ่านข้อมูลที่เกี่ยวข้องโดยอัตโนมัติ มีการจัดกลุ่มและพัฒนาสู่ฐานข้อมูลเพื่อสร้างกลยุทธ์ให้องค์กร อาทิ การสอบถามข้อมูลเกี่ยวกับผลิตภัณฑ์คู่แข่งขันและคู่ค้าในตลาด ดัชนีทางการเงินที่เกี่ยวข้อง รายชื่อพนักงาน เป็นต้น ตัวอย่างเช่น บริษัท IBM ใช้ในการวิเคราะห์ข้อมูลออนไลน์ของฝ่ายขาย โดยข้อมูลของ Call Center ทั้งหมดจะถูกแบ่งเป็น 30 กลุ่มและทำตารางสรุปตามคำถาม ชื่อ และเวลา ตลอดจนข้อมูลการบริการที่เกี่ยวข้อง หรือบริษัท GlaxoSmithKline ที่ใช้ SAS Text Miner ในการแบ่งกลุ่มผลิตภัณฑ์ของคู่แข่ง นอกจากนี้ ยังมี University of Louisville (Kentucky, USA) ที่วินิจฉัยใบสั่งยาของแพทย์จากรหัสยา ชื่อยา และชื่อของแพทย์ โดยใช้ในกรณีทั่วไปและกรณีที่ผิดปกติ ทำให้ทราบล่วงหน้าถึงสิ่งที่จำเป็นต้องตรวจสอบความสัมพันธ์ระหว่างการวินิจฉัยและยา

• ETL-Extraction Transformation Loading เป็นการแปลงข้อความที่ไม่มีโครงสร้างให้เป็นข้อความที่มีโครงสร้างและสามารถนำเข้าสู่ฐานข้อมูลได้ ทำให้สามารถค้นคืนเอกสาร และค้นหาขั้นสูงได้ อาทิ เอกสารทางกฎหมาย และทางการแพทย์ เป็นต้น ตัวอย่างเช่น NHS Medical Centre ที่ผู้ป่วยสามารถแบ่งปันประสบการณ์ที่แตกต่างกันในการผ่าตัด และนำมาบูรณาการ ทำให้แพทย์สามารถเข้าถึงข้อมูลในฐานข้อมูลได้อย่างทั่วถึง

• HR-Human Resources ใช้ในการสร้างแรงจูงใจแก่พนักงาน การวิเคราะห์ประวัติผู้สมัคร และการวิเคราะห์ความคิดเห็นของพนักงาน อาทิ ความเครียด ความเชื่อถือในองค์กร เป็นต้น ตัวอย่างเช่น บริษัท Temis Enabled Conoco ใช้ติดตามระดับความพึงพอใจในการทำงานจากข้อความเอกสาร เช่น อีเมล การสำรวจความคิดเห็นภายใน และการสนทนาออนไลน์ เพื่อนำเสียงสะท้อนของพนักงานมาปรับใช้ให้สอดคล้องกับวัฒนธรรมองค์กร รวมถึงการอ่านและจัดเก็บประวัติผู้สมัครเพื่อคัดเลือกพนักงานใหม่ให้มีความรวดเร็วมากยิ่งขึ้น

1.2 การบริหารลูกค้าสัมพันธ์ (CRM) และการวิเคราะห์ตลาด (MA)

• CRM-Customer Care and CRM เป็นการบริหารเนื้อหาจากข้อความของลูกค้า แล้วนำมาวิเคราะห์ความต้องการด้านบริการ เพื่อให้การตอบสนองบริการหรือตอบคำถามได้อย่างเหมาะสม เช่น การให้บริการของพนักงาน Call Center ในการตอบคำถามจากสายลูกค้าที่โทรศัพท์เข้ามา

• MA-Market Analysis ใช้ในการวิเคราะห์คู่แข่ง และติดตามความคิดเห็นของลูกค้าเพื่อกำหนดลูกค้าที่มีศักยภาพ วิเคราะห์แหล่งสื่อที่มีผลต่อธุรกิจและนำเก็บในฐานข้อมูล เช่น หลายธุรกิจใช้อีเมลเป็นแหล่งในการสร้างลูกค้าใหม่ เป็นต้น

1.3 การติดตามเทคโนโลยี (TW) เพื่อนำมาวิเคราะห์สิทธิบัตร

• TW-Patents, Scientific Abstracts and Financial News เป็นการวิเคราะห์คุณลักษณะของเทคโนโลยีที่มีอยู่ในปัจจุบัน เช่น ใช้ในการอ้างบรรณานุกรมของแต่ละเอกสารที่มีหัวเรื่องเดียวกัน และจัดเก็บข้อความเอกสารจากหลายภาษาให้อยู่ในฐานข้อมูลเดียวกันเพื่อให้เรียกใช้ได้ง่ายขึ้น และสามารถนำมาวิเคราะห์สิทธิบัตรเพื่อคุ้มครองความเป็นเจ้าของผลงานการประดิษฐ์คิดค้นหรือการออกแบบผลิตภัณฑ์

1.4 การประมวลผลภาษา ธรรมชาติ (NLP) ประกอบด้วย

• ORT-Dictionary and Spelling Correction ด้านพจนานุกรมและการสะกดคำ

• QNL-Questioning in Natural Language and Search Engines ด้านการถามคำถามในภาษาธรรมชาติ และการใช้เครื่องมือค้นหา

• ML–Multi-lingual Writing and Teaching Languages ด้านการเขียนและการสอนหลายภาษา

• VR-Voice Recognition ด้านการรู้จำเสียง

ตัวอย่างเช่น การนำมาใช้ในการวิเคราะห์เว็บเพจที่มาจากหลายภาษาโดยวิเคราะห์จากความหมายสามารถสืบค้นจากบางส่วนของประโยค ถือเป็นการส่งเสริมการการตลาดระหว่างประเทศ

2) การประยุกต์ใช้เหมืองข้อความตามกลุ่มการใช้งาน

2.1 กลุ่ม BIF (Banks, Insurance and Financial Markets) ใช้ในการวิเคราะห์ข่าวทางการเงิน ศูนย์บริการตอบรับทางโทรศัพท์ หรือ Call Center

2.2 กลุ่ม INF (Informatics, Internet, Linguistic Resources On Line) ใช้ในการค้นคืนสารสนเทศ ข้อความตามการใช้งานของภาษา และการแปลภาษาอัตโนมัติ

2.3 กลุ่ม PH (Pharmaceutical and Healthcare) ใช้ในการสกัดข้อมูลอัตโนมัติจากเนื้อหาทางชีวการแพทย์และบริหารข้อมูลคลีนิค ด้วยการติดตามข่าวจากหน้าหนังสือพิมพ์และสื่อต่างๆ เพื่อแจ้งเตือนข่าวสารเกี่ยวกับโรคติดต่อที่สำคัญ เช่น SARS หรือการที่ Health Care ทำการวิเคราะห์และตรวจจับการฉ้อโกงใบเสร็จที่พบจากการการเคลมสิทธิ์ต้องสงสัย เช่น การใช้ชื่อย่อ ชื่อเต็ม ในการเคลมประกันสุขภาพด้วยนามแฝงที่แตกต่างกันของบุคคลเดียวกัน

2.4 กลุ่ม PM (Publishing and Media) ใช้ในการสร้างสารบัญอัตโนมัติของสำนักพิมพ์และสื่อกระจายเสียง

2.5 กลุ่ม POL (Political Institutions, Public Administration and Legal Documents) ใช้ในการวิเคราะห์เอกสารอัตโนมัติ และวิเคราะห์ข้อมูลจากเครื่องมือค้นหา

2.6 กลุ่ม TES (Telecommunications, Energy and Other Services Industries) ใช้ในการบริหารศูนย์บริการตอบรับทางโทรศัพท์ หรือ Call Center ในการสกัดรายชื่อของลูกค้าและคู่แข่งขัน รวมถึงการวางแผนเชิงกลยุทธ์

2.7 กลุ่มอื่นๆ

Louise A. Fransis (2006) ได้อธิบายถึงการนำเหมืองข้อความมาใช้ในธุรกิจประกันภัย เนื่องจากข้อมูลส่วนใหญ่ในฐานข้อมูลขององค์กรเป็นข้อมูลแบบไม่มีโครงสร้าง เช่น คำอธิบายการเคลมสิทธิ์ประกันภัย เนื้อหาของอีเมล ผลการประเมินตามกรมธรรม์ และการสำรวจความพึงพอใจของลูกค้าแบบคำถามปลายเปิด ข้อมูลดังกล่าวเมื่อถูกแปลงเป็นสารสนเทศที่ต้องการแล้วสามารถใช้ในการวิเคราะห์และพยากรณ์ได้ในอนาคต

ปัจจุบันได้มีการนำเทคนิคเหมืองข้อความมาใช้ในการพัฒนางานวิจัยด้านต่างๆ มากยิ่งขึ้น ตัวอย่างเช่น การนำเทคนิคเหมืองข้อความมาใช้กับฐานข้อมูลบทความวิจัยด้านการเกษตรที่เผยแพร่และตีพิมพ์ของศูนย์สนเทศทางการเกษตรแห่งชาติ สำนักหอสมุดมหาวิทยาลัยเกษตรศาสตร์ ซึ่งเป็นข้อมูลเชิงบรรณานุกรม จำนวน 2,580 บทความ โดยใช้คำสำคัญมาทำการสกัดคุณลักษณะ (Feature Extraction) เพื่อดึงคุณลักษณะของคำมาเป็นตัวแทนของเอกสารภาษาไทย โดยมีการตัดคำหยุดซึ่งเป็นคำที่ไม่มีความหมายสำคัญต่อเอกสารและไม่ทำให้ใจความของเอกสารเปลี่ยน แล้วนำคุณลักษณะของคำที่สำคัญมาใช้ในการจำแนกหมวดหมู่ เพื่อสร้างแบบจำลองการจำแนกหมวดหมู่ข้อมูล และใช้ในการพยากรณ์หมวดหมู่ของข้อมูล ทั้งนี้ ช่วยให้เกิดประโยชน์ต่อการอ้างอิงและใช้ประโยชน์ในการศึกษาค้นคว้าวิจัยจากสารสนเทศด้านการเกษตรของไทย และสามารถประยุกต์ใช้กับระบบการสืบค้นผู้เชี่ยวชาญด้านการเกษตรต่อไปได้ (Pilapan Phonarin, 2011) รวมถึงมีการพัฒนาระบบการสกัดสารสนเทศทางการแพทย์จากบทคัดย่อภาษาอังกฤษเกี่ยวกับโรคไข้หวัดใหญ่ H1N1 โดยใช้ข้อมูลจากฐานข้อมูล MEDLINE จำนวน 50 เอกสาร ประกอบด้วยชื่อเรื่องและเนื้อหา นำมาตัดคำที่ไม่ใช่คำหลักออกเพื่อให้ได้คำศัพท์ที่เกี่ยวข้องกับเนื้อหาในเอกสารมากที่สุด จากนั้นทำการสร้างกลุ่มคำที่ใช้แทนโรคไข้หวัดใหญ่ H1N1 โดยการเปรียบเทียบกับพจนานุกรมที่รวบรวมไว้ก่อนหน้า แล้วจึงนำมากำหนดดัชนีสำหรับระบุชนิดของกลุ่มคำ เพื่อหาคำตอบเกี่ยวกับสถานที่และ

เวลาในการแพร่กระจายของโรค ช่วยให้ผู้ใช้สามารถค้นหาข้อมูลเกี่ยวกับโรคดังกล่าวได้สะดวกและสามารถรับทราบข้อมูลที่ถูกต้องจากเอกสารงานวิจัยโดยตรง (Ekamorn Meesomsak, 2011)

นอกจากนี้แล้ว วิชุดา โชติรัตน์ (2554) ได้พัฒนาระบบวิเคราะห์ข่าวออนไลน์ในรูปแบบเว็บ แอพพลิเคชันกรณีศึกษาข่าวในพื้นที่ 5 จังหวัดชายแดนภาคใต้ประเทศไทย เพื่อค้นหาข้อมูลที่มีประโยชน์ เช่น ข้อมูลความสัมพันธ์ของบุคคล เหตุการณ์ สถานที่ เวลา และอาวุธ เป็นต้น โดยใช้ฐานความรู้ออนโทโลจี (Ontology Knowledge Based) เป็นโครงสร้างในการจัดเก็บคำสำคัญที่มีผลต่อการวิเคราะห์เนื้อหาข่าวออนไลน์ภาษาไทย และใช้เทคนิคการทำเหมืองข้อความในการแสวงหากฎความสัมพันธ์ (Association Rule) ของคำสำคัญที่ซ่อนอยู่ในเนื้อหาข่าวออนไลน์ ทำให้การวิเคราะห์ข้อเท็จจริงจากเนื้อหาข่าวออนไลน์ที่มีจำนวนมากเกิดความสะดวกรวดเร็ว ผู้ใช้สามารถวิเคราะห์สภาพปัญหา สถานการณ์ความรุนแรง และแนวทางแก้ไขปัญหาในพื้นที่ดังกล่าวได้อย่างมีประสิทธิภาพมากยิ่งขึ้น เช่นเดียวกันนี้ได้มีงานวิจัยที่นำเหมืองข้อความมาช่วยวิเคราะห์ข่าวสถานการณ์ในพื้นที่ 3 จังหวัดชายแดนภาคใต้ของประเทศไทย เพื่อทำนายพฤติกรรมของผู้ก่อการร้าย และใช้สนับสนุนการตัดสินใจในการป้องกันวินาศภัยที่อาจเกิดขึ้น ซึ่งได้มีการรวบรวมข่าวที่เกี่ยวข้องจากแหล่งต่างๆ มาจัดเก็บไว้ในฐานข้อมูล แล้วนำมาสร้างโมเดลหาความสัมพันธ์และจัดโครงสร้างของรูปแบบความสัมพันธ์ของเหตุการณ์ที่เกิดขึ้น (Inyaem, Uraiwan, 2010)

ปัจจุบันอินเทอร์เน็ตกลายเป็นเครื่องมือสำคัญของผู้ใช้งานทั่วโลกในการเข้าถึงข้อมูล ผลกระทบในเชิงลบที่ตามมาคือการก่ออาชญากรรมทางอินเทอร์เน็ตมีจำนวนมากขึ้น ทั้งนี้ เทคนิคเหมืองข้อความสามารถนำมาใช้ในการแก้ปัญหาอาชญากรรมบนโลกไซเบอร์ทั้งในส่วนการล่อลวงทางอินเทอร์เน็ตและการข่มขู่บนโลกไซเบอร์ โดยการดักจับข้อมูลจากโปรแกรมส่งข้อความทางอินเทอร์เน็ต (Internet Messager : IM) และโปรแกรมสนทนาออนไลน์ (Internet Relay Chat : IRC) จากห้องสนทนาออนไลน์และสื่อสังคมออนไลน์ต่างๆ เพื่อนำมาวิเคราะห์ข้อความสนทนาโต้ตอบ และหาบทสนทนาระหว่างผู้ล่อลวงและผู้เยาว์ จากนั้นนำมาแยกข้อมูลการสนทนาที่ถูกบันทึก เพื่อวิเคราะห์ข้อความคิดเห็น การแสดงความรู้สึกทั้งขั้วบวกและขั้วลบ ตลอดจนการตรวจสอบพฤติกรรมที่ไม่เหมาะสม ทำให้พบที่อยู่ของกิจกรรมที่ไม่เหมาะสม ที่อาจมีการตามมาราวีหรือมีพฤติกรรมที่น่ารังเกียจ ตลอดจนมีคลังคำศัพท์ที่สามารถใช้สถิติในการวิเคราะห์และการจัดกลุ่ม ส่งผลให้การตรวจสอบผู้ข่มขู่ในโลกไซเบอร์ทำได้ง่ายขึ้น ซึ่งจะช่วยปกป้องไม่ให้ผู้เยาว์ตกเป็นเหยื่อของผู้ไม่หวังดี ซึ่งในปัจจุบันได้ถูกพัฒนาเป็นซอฟต์แวร์ที่ใช้ติดตามการสนทนาที่สามารถส่งข้อความเตือนผู้ปกครองเมื่อโปรแกรมตรวจพบว่าคอมพิวเตอร์ที่บุตรหลานใช้งานอยู่มีข้อความการสนทนา อีเมล หรือการเข้าใช้งานเว็บไซต์ที่เข้าข่ายการล่อลวงหรือการข่มขู่เกิดขึ้น ตลอดจนสามารถบันทึกทุกตัวอักษรทางแป้นพิมพ์ที่ใช้งานผ่านสื่อสังคมออนไลน์และจับภาพหน้าจอเพื่อนำมาวิเคราะห์ เพื่อดูแลให้บุตรหลานห่างไกลจากภัยอาชญากรรมที่อาจเกิดขึ้น (Michael W. Berry & Jacob Kogan, 2010)

บทสรุปและแนวโน้มในอนาคต

องค์กรส่วนใหญ่ที่มีการจัดเก็บข้อมูลประเภทข้อความจากหลายแหล่งในปริมาณมากและเป็นระยะเวลายาวนาน มักประสบปัญหาในการนำข้อมูลมาใช้ให้เกิดประโยชน์สูงสุด แต่หากองค์กรใดนำเทคนิคเหมืองข้อความมาใช้ในการวิเคราะห์ข้อมูลก็จะช่วยให้ค้นพบความรู้ใหม่ที่สามารถสร้างมูลค่าเพิ่มให้แก่องค์กรได้ ทั้งนี้ เนื่องจากการทำเหมืองข้อความจะช่วยหารูปแบบความสัมพันธ์ของข้อมูลประเภทข้อความที่ซับซ้อนและมีปริมาณมากซึ่งไม่สามารถกระทำได้จากการทำงานของมนุษย์เพียงอย่างเดียว

แม้ว่าเทคโนโลยีการสืบค้นจะสามารถค้นหาด้วยคำสำคัญให้ได้ผลลัพธ์อย่างรวดเร็วแต่การค้นข้อความแสดง

ความคิดเห็นที่มีอยู่เป็นจำนวนมากในกระทู้ เว็บบอร์ด หรือบล็อกต่างๆ ยังไม่สามารถใช้วิธีดังกล่าวได้ผลนัก หากองค์กรต้องการสืบค้นความคิดเห็นของลูกค้าที่มีต่อสินค้าหรือบริการหรือเรื่องใดเรื่องหนึ่ง จะพบว่าข้อมูลความคิดเห็นบนอินเทอร์เน็ตและสื่อสังคมออนไลน์มีปริมาณมหาศาลและมาจากหลายแหล่ง ทำให้ใช้เวลาในการอ่าน ตีความ และสรุปข้อความคิดเห็นเหล่านั้นเป็นเวลานาน นักวิจัยจึงได้พัฒนาเทคนิคเหมืองข้อความแสดงความคิดเห็น (Opinion Mining) ซึ่งอาศัยเทคนิคของการสืบค้นข้อมูลและการประมวลผลทางภาษา ทำให้ผู้ใช้สามารถสรุปความคิดเห็นที่หลากหลายบนเครือข่ายสังคมออนไลน์ให้เข้าใจง่ายยิ่งขึ้น สามารถประเมินความพึงพอใจของลูกค้าและนำความคิดเห็นต่างๆ มาใช้ปรับปรุงรูปแบบสินค้าและบริการได้อย่างเหมาะสม (Kongthon, 2010)

นอกจากเทคนิคเหมืองข้อความแสดงความคิดเห็นแล้ว แนวโน้มในอนาคตจะมีการนำเทคนิคเหมืองข้อมูลมาใช้กับข้อมูลสื่อประสมที่มีรูปแบบหลากหลายมากยิ่งขึ้น อาทิ การทำเหมืองข้อมูลบนเว็บ (Web Mining) เหมืองข้อมูลวิดีโอ (Video Mining) และเหมืองข้อมูลรูปภาพ (Image Mining) เป็นต้น ทั้งนี้ ตราบใดที่ข้อมูลในทุกรูปแบบยังมีอยู่จำนวนมาก และผู้ใช้งานยังคงต้องการแสวงหาความรู้ใหม่ๆ จากข้อมูลเหล่านั้น การทำเหมือง (Mining) ก็จะทวีคุณค่าและมีความจำเป็นมากยิ่งขึ้น ซึ่งสามารถพิสูจน์ได้จากการนำเทคนิคอันทรงพลังนี้ไปประยุกต์ใช้งานในวงการอุตสาหกรรมและด้านวิชาการมากยิ่งขึ้น

บรรณานุกรม

วิชุดา โชติรัตน์. (2554). *ระบบวิเคราะห์ข่าวออนไลน์โดยใช้ฐานความรู้ออนโทโลยี และการทำเหมืองข้อความ กรณีศึกษาข่าวในพื้นที่ 5 จังหวัดชายแดนภาคใต้ ประเทศไทย*, วิทยานิพนธ์หลักสูตรวิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

Ashok Srivastava, Mehran Sahami. (2009). *Text mining : classification, clustering, and applications*. Chapman & Hall/CRC.

Diane McDonald. (2012). *Value and benefits of text mining*. Retrieved 28 January 2013, from http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

Ekamorn Meesomsak, Natthapong Sriphadung, Phatthanaphong Wanchanthuek, Panida Songram, Kaveepoj Bunluewong and Umaporn. (2011). *An Information Extraction System for Medical Informatics using Text Mining Techniques: A Case Study of the H1N1 influenza*. The 7th National Conference on Computing and Information Technology: NCCIT2011.

Feldman, Ronen & Sanger, James. (2007). *The text mining handbook advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey. (2008). A Text Mining Technique Using Association Rules Extraction. *International Journal of Information and Mathematical Sciences* 4(1).

Inyaem, Uraiwan, Meesad, Phayung, Haruechaiyasak, Choochart and Tran, Dat. (2010). Terrorism event classification using fuzzy inference systems, *International Journal of Computer Science and Information Security*, 7(3), 247-256.

Kongthon, Alisa, Kongyoung, Sarawoot Sangkeettrakarn, Chatchawal Haruechaiyasak, Choochart. (2010). *Thailand's Tourism Information Service Based on Semantic Search and Opinion Mining*, Proceedings of ITC-CSCC 2010.

Louise A. Francis. (2006). *Taming Text: An Introduction to Text Mining*, Casualty Actuarial Society Forum, Winter 2006.

Mahesh T R, Suresh M B, M Vinayababu. (2009). *Text Mining : Advancements, Challenges and Future Directions*. *International Journal of Reviews in Computing* © 2009-2010 IJRIC & LLS.

Michael W. Berry and Jacob Kogan. (2010). *Text Mining Applications and Theory*. United Kingdom: John Wiley & Sons.

Pilapan Phonarin, Supot Nitsuwat and Choochart Haruechaiyasak. (2011). *Text Categorization using Feature Selection Techinque for Agricuture Bibliographic Data*. The 7th National Conference on Computing and Information Technology: NCCIT2011.

Saša Petrovic. (2007). *Collocation extraction measures for text mining applications*. University of Zagreb Faculty, of Electrical Engineering and Computing.

Sergio Bolasco, Alessio Canzonetti, Federico M. Capo. Francesca Della Ratta-Rinaldi, Bhupesh K. Singh. (2002). Understanding Text Mining: a Pragmatic Approach. Roma Italy: U. La Sopienza.

Vishwadeepak Singh Baghela, Dr. S.P.Tripathi. (2012). *Text Mining Approaches to Extract Interesting Association Rules from Text Documents*, *IJCSI International Journal of Computer Science*, 9(3).

Wanvipa Wongvilaisakul received her Master of Science in Information Technology Degree, and Bachelor of Business Administration (Business Computer) from Sripatum University. She is Chairperson, Department of Information Technology in Faculty of Engineering and Technology, Panyapiwat Institute of Management. She main interests are in Management Information System and Decision Support System.