การสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์

สุริยศักดิ์ เลิศสกุลสมบูรณ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมเว็บคณะเทคโนโลยีสารสนเทศมหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2556

Stuff Relation Extraction from Scientific Research Paper

Suriyasak Lertsakunsomboon

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Web Engineering

Faculty of Information Technology, Dhurakij Pundit University

2013

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก รศ.ดร.ฉวีวรรณ เพ็ชรศิริ ผู้จัดทำวิทยานิพนธ์ใคร่ขอกราบขอบพระคุณที่ท่านอาจารย์กรุณาให้คำปรึกษา ให้ข้อคิดเห็น และสละเวลาให้กับผู้จัดทำวิทยานิพนธ์ตลอดมา ตลอดจนช่วยตรวจสอบต้นฉบับและแก้ไขข้อบกพร่องของงานวิจัยเพื่อให้วิทยานิพนธ์ฉบับนี้เสร็จสิ้นไปได้ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ คณะเทคโนโลยีสารสนเทศ สาขาวิศวกรรมเว็บ มหาวิทยาลัยธุรกิจบัณฑิตย์ด้วยความเคารพอย่างสูง ผู้วิจัยรู้สึกซาบซึ้งยิ่งนัก จึงขอขอบพระคุณเป็นอย่างสูงมา ณ โอกาสนี้

ตลอดระยะเวลาในการจัดทำวิทยานิพนธ์เล่มนี้ขอกราบขอบพระคุณบิดา มารดาผู้ซึ่งให้ความรักความเมตตาความห่วงใยและเป็นกำลังใจให้กับผู้วิจัยจนสำเร็จและขอขอบพระคุณพี่ๆน้ องๆ รวมทั้งเพื่อนๆ ทุกคนที่ให้กำลังใจผู้วิจัยโครงการรู้สึกซาบซึ้งในพระคุณอย่างสูง

สุริยศักดิ์ เลิศสกุลมบูรณ์

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

หัวข้อวิทยานิพนธ์	การสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์
ชื่อผู้เขียน	สุริยศักดิ์ เลิศสกุลสมบูรณ์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร.ฉวีวรรณ เพ็ชรศิริ
สาขาวิชา	วิศวกรรมเว็บ
ปีการศึกษา	2555

บทคัดย่อ

งานวิจัยนี้มีเป้าหมายเพื่อสกัดความสัมพันธ์แบบสตัฟฟ์ (Stuff Relation) ซึ่งเป็นหนึ่งในประเภทของความสัมพันธ์แบบพาธ- โฮล (Part-Whole Relation) (ในงานวิจัยนี้ความสัมพันธ์แบบสตัฟฟ์หมายถึงความสัมพันธ์ระหว่าง สารเคมี(สตัฟฟ์)กับพืช(วัตถุ)) จากข้อมูลที่ไม่เป็นโครงสร้าง ความสัมพันธ์แบบสตัฟฟ์จำเป็นสำหรับการสร้างออนโทโลยีของสารผลิตภัณฑ์ธรรมชาติเพื่อช่วยอุตสาหกรรมโดยเฉพาะอุตสาหกรรมการผลิตยา วิทยานิพนธ์นี้นำเสนอการสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยจากเว็บ ซึ่งมีปัญหาสำคัญ 3 ปัญหาในการสกัดความสัมพันธ์แบบสตัฟฟ์ 1) ปัญหาการระบุความสัมพันธ์แบบสตัฟฟ์โดยไม่มีการระบุชนิดของคำ 2) การระบุนิพจน์ระบุนามทางวิทยาศาสตร์ของพืช และ 3) การระบุนิพจน์ระบุนามของ สารเคมี โดยที่วิทยานิพนธ์นี้ขอเสนอการใช้เทคนิคการเรียนรู้แบบเนอีฟ-เบย์ในการเรียนรู้และสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์โดยไม่มีการระบุชนิดของคำ และใช้ฐานข้อมูล NBCI-pubchem และ NBCI-taxonomy เพื่อระบุชื่อสารเคมีและชื่อพืชตามลำดับ ซึ่งใช้คุณลักษณะ (feature) ในการเรียนรู้ประกอบไปด้วย แนวคิดของสารเคมี , แนวคิดของพืช และ คำทั้งหมดที่อยู่ในหน้าต่างซึ่งปรากฏระหว่างแนวคิดของสารเคมีและแนวคิดของพืช(เมื่อขนาดของหน้าต่างมีขนาดตั้งแต่ 3 คำถึง 5 คำ)

ในการประเมินประสิทธิภาพของแบบจำลองในงานวิจัยนี้พบว่าได้ค่าระลึกสูงถึง 98.5% และค่าความถูกต้องเป็น 35.51% โดยใช้ขนาดหน้าต่างเป็น 3 คำ

Thesis Title	Stuff Relation Extraction from Scientific Research Paper
Author	Suriyasak Lertsakunsomboon
Thesis Advisor	Assoc. Prof. Dr. Chaveevan Pechsiri
Academic Program	Web Engineering
Academic Year	2012

ABSTRACT

This research aims to extract Part-Whole relations, especially the stuff relation (where stuff relation in this research is relation between the stuff, i.e. chemical compound, and the object, i.e. scientific name of plant), from unstructured textual data is the challenging work. The Stuff relation is necessary for constructing the natural product Ontology used to represent all natural product knowledge which benefits to the industries, especially the pharmaceutical industry. This thesis presents how to extract the stuff relation from scientific research paper on the Web for supporting chemical industries. There are three problems of extracting the stuff relation: a) the stuff relation identification of problem without POS (Part-of-Speech) annotation, b) the scientific name entity identification of plant and c) the chemical name entity identification problems. Therefore this thesis proposes using machine learning technique, Naïve Bayes, to learn and extract the stuff relation from the scientific research paper without applying POS annotation. The research also applies NCBI-pubchem and NCBI-taxonomy to identify chemical name and scientific name of plant, respectively. The features used in learning the stuff relation consist of the chemical name concept (the chemical name entity), the plant name concept (the scientific name of plan) and all words within one window existing between the chemical name concept and the plant name concept (where the window size is vary from 3 words to 5 words).

The evaluation of this research model shows the highest recall 98.5% and 35.51% precision at the window size is 3 words

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของงาน

ผลิตภัณฑ์สมุนไพรและผลิตภัณฑ์จากธรรมชาติได้รับความนิยมจากผู้บริโภคเป็นอย่างมากในปัจจุบัน ผู้บริโภค ทั้งหลายต่างก็เลือกใช้ผลิตภัณฑ์เหล่านี้ ซึ่งมีสาเหตุมาจากการที่ผู้บริโภคในปัจจุบันได้คำนึงถึงสุขภาพของตนเองมากขึ้น ต่างก็หันมาใส่ใจเรื่องสุขภาพในวิถีทางธรรมชาติกันมากขึ้น กล่าวคือกลับไปสู่แบบดั้งเดิมอันเป็นภูมิปัญญาแต่โบราณผลิตภัณฑ์เหล่านี้ถูกผลิตออกมาเพื่อตอบสนองความต้องการของผู้บริโภคอย่างไม่สิ้ นสุด เมื่อมีการค้นพบว่าผลิตภัณฑ์จากธรรมชาติชนิดไหนมีสรรพคุณหลากหลายก็จะถูกบอกต่อ และขยายไปสู่ผลิตภัณฑ์อื่นๆ อาทิ อาหาร, ยารักษาโรค, ผลิตภัณฑ์บำรุงผิวพรรณ , หรือบรรดาผลิตภัณฑ์เสริมอาหารต่างๆ รวมไปถึงศาสตร์ของการแพทย์ทางเลือก ที่เป็นศาสตร์ของการปรับสมดุลของร่างกายเช่น การนวด ดัด ดึง เพื่อการบำบัดและทำให้ร่างกายผ่อนคลายในวัฒนธรรมต่างๆ เป็น การจัดระเบียบร่างกายรูปแบบหนึ่งซึ่งจะมีการใช้น้ำมันหอมระเหยมาร่วมด้วย หรือ การบำบัดด้วยกลิ่น หรือการล้างพิษ เป็นต้น การใช้การ บำบัดทางธรรมชาติจะเป็นวิธีค่อยเป็นค่อยไปซึ่งแตกต่างจากการรักษาโรค เพราะการรักษาโรคจะปล่อยทิ้งไว้นานไม่ได้ต้องเข้าไปควบคุมเชื้อโรคไม่ให้ขยายหรือแพร่กระจายไปยังอวัยวะส่วนอื่นๆหรือบุคคลอื่น ในสมัยก่อนการรักษาโรคจะใช้ยาสมุนไพร ซึ่งส่วนใหญ่จะใช้การนำเข้าสู่ร่างกายทางปาก ก่อนที่จะนำเข้าสู่ร่างกายก็จะนำไปต้ม ตัวยาในสมุนไพรก็จะละลายออกมาอยู่ในน้ำ แล้วดื่มน้ำที่ใช้ต้มสมุนไพร วิธีการรักษาแบบนี้จะใช้เวลานาน น้ำต้มสมุนไพรก็มีรสชาติเฝื่อน ขม ทำให้ดื่มได้ยาก แต่ถ้าเราทราบว่าสารเคมีที่อยู่ในสมุนไพรที่ใช้รักษาโรคเป็นสารเคมี ชนิดไหน จะทำให้นำสารเคมีตัวนั้นมาใช้รักษาโรคได้ซึ่งยาประเภทนี้ก็คือยาแผนปัจจุบัน เมื่อเทียบปริมาณของยาที่ต้องนำเข้าสู่ร่างกายแต่ละครั้ง พบว่ายาแผนปัจจุบันจะนำเข้าสู่ร่างกายได้ง่ายว่ายาสมุนไพรและปริมาณการใช้แต่ละครั้งก็น้อยกว่าด้วย เนื่องจากในยาแผนปัจจุบัน จะนำเอาสารเคมีที่เป็นยาโดยเฉพาะมาใช้ แต่ในยาสมุนไพรจะมีสารเคมีส่วนอื่นที่ไม่ใช่ยาผสมอยู่ด้วย แต่ก็ ไม่ส่งผลอันตรายแก่ร่างกายผู้ป่วยซึ่งสมุนไพรบางตัวมีสารเคมีที่ใช้รักษาโรคปริมาณน้อยมาก จำเป็นต้องใช้ยาสมุนไพรจำนวนมาก หรือใช้เวลาในการรักษายาวนาน การที่เราจะทร าบถึงชนิดของสารเคมีที่ใช้รักษาโรคในยาสมุนไพร จะต้องผ่านกระบวนการการสกัดเพื่อแยกเอาสารเคมีออกมาแต่ละตัว และนำไปวิเคราะห์หาโครงสร้างทางเคมี ชนิดหรือประเภทของสารเคมี และผลการออกฤทธิ์ทางชีวภาพ

กระบวนการเหล่านี้จะเป็นงานของกลุ่มงานวิจัยทางด้านสารผลิตภัณฑ์ธรรมชาติ

งานวิจัยทางด้านสารผลิตภัณฑ์ธรรมชาติคืองานวิจัยเกี่ยวกับการสกัด, วิเคราะห์สารเคมีที่อยู่ในสิ่งมีชีวิต ทั้งที่ถูกผลิตขึ้นมาเองและที่ได้รับจากสภาพแวดล้อมบริเวณถิ่นที่อยู่ รวมถึงการวิเคราะห์การออกฤทธิ์ทางชีวภาพของสารเคมีนั้นๆ โดยข้อมูลที่ได้จากงานวิจัยเหล่านี้เป็นที่ต้องการของอุตสาหกรรมด้านการผลิตยา, เครื่องสำอาง, ผลิตภัณฑ์สปา, อาหารเสริมเป็นต้นแต่รายงานการวิจัยทางด้านนี้มีปริมาณมากและหลากหลาย รวมทั้งชื่อและประเภทของสารเคมีก็มีปริมาณมาก ทำให้ใช้เวลาในการศึกษาเรียนรู้มากตามไปด้วย ฉะนั้นหากมีเครื่อ งมืออัตโนมัติช่วยสร้างองค์ความรู้ต่างๆจากเอกสารงานวิจัยด้าน สารผลิตภัณฑ์จากธรรมชาติ โดยเฉพาะอย่างยิ่งองค์ความรู้เกี่ยวกับออนโทโลยีของ สารผลิตภัณฑ์จากธรรมชาติซึ่งเป็นองค์ความรู้ที่แสดงแนวความคิด (Concept) ความสัมพันธ์ (Relation) ประเภทต่างๆเช่น is-a-relation, part-of-relation, property-relation เป็นต้นของสารเคมีต่างๆที่เป็นผลิตภัณฑ์จากธรรมชาติ จะช่วยทำให้ผู้ที่สนใจ เช่น นักอุตสาหกรรม นักวิจัย เป็นต้น สามารถเรียนรู้และติดตามเพื่อนำไปพัฒนาอุตสาหกรรมที่เกี่ยวข้องกับ สารผลิตภัณฑ์ธรรมชาติดังนั้นงานวิจัยนี้มีจุดมุ่งหมายที่จะสกัดความสัมพันธ์แบบสตัฟฟ์ (stuff relation) ซึ่งเป็นประเภทหนึ่งของ part-of-relation จากเอกสารงานวิจัยด้านผลิตภัณฑ์จากธรรมชาติ เพื่อนำไปสู่การการสร้างองค์ความรู้เกี่ยวกับออนโทโลยีของผลิตภัณฑ์จากธรรมชาติในการสกัดความสัมพันธ์แบบ สตัฟฟ์ อย่างอัตโนมัตินี้ มีปัญหาหลัก 2 ปัญหาคือ ปัญหาการระบุความสัมพันธ์แบบสตัฟฟ์ ในเอกสาร/วารสารทางวิชาการภาษาอังกฤษ ที่เป็นไฟล์ รูปแบบ PDF และปัญหาการระบุคำนามที่มี แนวความคิด เป็นสารเคมีต่างๆที่เป็นผลิต ภัณฑ์จากธรรมชาติ และพืชในระดับ ดิวิชัน (Division) ฉะนั้นงานวิจัยนี้จึงของเสนอการเรียนรู้ความสัมพันธ์แบบ สตัฟฟ์ จากคู่คำนามระหว่างคำนามที่มีแนวความคิด เป็นสารเคมีต่างๆที่เป็นผลิตภัณฑ์จากธรรมชาติ และพืช ในระดับดิวิชันด้วยการเรียนรู้ของเครื่อง (Machine Learning) ที่เป็นตัวจำแนก Naïve Bayes ทั้งนี้เพื่อใช้ในการสกัดความสัมพันธ์แบบ สตัฟฟ์ จากเอกสาร และในขณะเดียวกัน ใช้ NBIC-PubChem และ NBIC-Taxonomy เพื่อหาแนวความคิด เป็นสารเคมีต่างๆที่เป็นผลิตภัณฑ์จากธรรมชาติ และพืชใน ระดับดิวิชัน ตามลำดับ นอกจากนี้งานวิจัยนี้ ยังได้หาความส หสัมพันธ์ระหว่างกลุ่มสารเคมีกับกลุ่ม พืชในระดับดิวิชันว่ามีความสัมพันธ์กันอย่างไรด้วยค่าสหสัมพันธ์ทั้งนี้เพื่อช่วยชี้นำว่าสารเคมีต่างๆที่เป็นผลิตภัณฑ์จากธรรม ชาติมีความสัมพันธ์กับพืชในระดับ ดิวิชัน มากน้อยอย่าง ไรกับโดเมนวารสารทางวิชาการของสารผลิตภัณฑ์จากธรรมชาติ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและ สกัดความสัมพันธ์แบบ สตัฟฟ์ (stuff relation) จากเอกสาร งานวิจัยทางวิทยาศาสตร์เพื่อนำไปสร้างออนโทโลยีที่มีโดเมนเป็นสารผลิตภัณฑ์ธรรมชาติ

1.3 ประโยชน์และผลที่คาดว่าจะได้รับ

1. ได้ข้อมูลที่มีความสัมพันธ์แบบสตัฟฟ์ (stuff relation) เพื่อนำไปสร้างออนโทโลยี
2. ได้คลังข้อมูลของสารผลิตภัณฑ์ธรรมชาติ
3. ทำให้นักอุตสาหกรรมหรือผู้ที่สนใจสามารถคัดกรองพืชที่มีสารที่ต้องการในเบื้องต้นได้

1.4 ขอบเขตของการวิจัย

1. ข้อมูลของงานวิจัยนี้เป็นเอกสารที่เป็น รูปแบบ PDF ซึ่งดาวน์โหลดมาจาก วารสารทางวิชาการชื่อ Journal of natural products
2. ข้อมูลจะเกี่ยวข้องกับสารผลิตภัณฑ์ธรรมชาติที่สกัดได้จากพืช

1.5 คำนิยามศัพท์

ในวิทยานิพนธ์ฉบับนี้ได้นิยามศัพท์ที่ใช้ในการวิจัยไว้ ดังนี้

สารผลิตภัณฑ์ธรรมชาติ (Natural Product) หมายถึง สารเคมีที่สกัดมาจากสิ่งมีชีวิต เช่น พืช สิ่งมีชีวิตในท้องทะเล จุลินทรีย์ซึ่งสารเคมีเหล่านี้เกิดขึ้นจากกระบวนการทางชีวภาพของสิ่งมีชีวิตนั้นในการดำรงชีวิตหรือสาร เคมีที่ถูกเปลี่ยนแปลงไปเนื่องจากนำเข้าร่างกายของสิ่งมีชีวิตนั้นๆ สารผลิตภัณฑ์ธรรมชาติไม่ใช่สมุนไพร เพราะสมุนไพรคือยาที่ได้มาจากพืชหรือสัตว์โดยที่ไม่มีการเปลี่ยนแปลงสภาพ

การออกฤทธิ์ทางชีวภาพ (Biological Activity) หมายถึง ผลกระทบของสารเคมีหรือตัวยาที่มีผลต่อ เซลล์หรือเนื้อเยื่อเป้าหมาย

การเรียนรู้ ของเครื่อง (Machine learning) หมายถึง การทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างหรือจากสภาพแวดล้อม จุดมุ่งหมายคือการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้นเมื่อเรียนรู้แล้วความรู้ที่เรียนได้จะเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่งเช่น กฎ ฟังก์ชัน ฯลฯ

รูปแบบโครงสร้างทางไวยากรณ์ (Lexico-syntactic pattern) หมายถึงรูปแบบที่บ่งบอกและจำแนกสารสนเทศ ออกมาจากข้อความทั่วไปโดยใช้รูปแบบโครงสร้างทางไวยากรณ์ตามหลักของภาษา

ชนิดของคำ (Part-of-speech, POS) หมายถึงหน้าที่ของคำนั้นๆในประโยคมี 8 ประเภทคือ คำนาม (Nouns) คำสรรพนาม (Pronouns) คำกริยา (Verbs) คำคุณศัพท์ (Adjectives) คำกริยาวิเศษณ์ (Adverbs) คำบุพบท (Prepositions) คำสันธาน (Conjunctions) และคำอุทาน (interjections)

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

ในบทที่ 2 นี้จะกล่าวถึงทฤษฎีแนวคิดและองค์ความรู้อื่นที่เกี่ยวข้องในการดำเนินการวิจัย ซึ่งจะแบ่งออกเป็น 2 ส่วน คือส่วนที่เป็นทฤษฎี กับส่วนที่เป็นงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังต่อไปนี้

2.1 ทฤษฎี

2.1.1 ความสัมพันธ์

Meronym (Cruse, Alan, Meaning in Language: An Introduction to Semantics and Pragmatics, Oxford University Press, Oxford, 2000)คือแนวความคิดเกี่ยวกับความสัมพันธ์แบบมีความหมายใช้ในด้านภาษาศาสตร์โดยที่ meronymจะหมายถึงการเป็นส่วนประกอบของบางสิ่งบางอย่าง (Part-of) หรือการเป็นสมาชิกของบางสิ่งบางอย่าง (Member-of) ยกตัวอย่างเช่น ‘นิ้วมือ’ เป็น meronymของ ‘มือ’ เนื่องจากนิ้วมือเป็นส่วนหนึ่งของมือ หรือ ‘ล้อ’ เป็น meronymของ ‘รถจักรยาน’ เนื่องจากล้อเป็นองค์ประกอบของรถจักรยาน

ในบางครั้งการพูดคุยกันหรือการถามตอบจะไม่มีวลี ‘เป็นส่วนหนึ่ง’ หรือ‘เป็นองค์ประกอบ’(‘part’)ในการสนทนาแต่จะมีแนวความคิด ‘part’(Morton E. Winston et al.,1987) เช่น

1) “Bicycles have wheels.” // “รถจักรยานมีล้อ”

2) “Bicycles are made of aluminum.” // “รถจักรยานทำด้วยอลูมิเนียม”

ซึ่งถ้าเปลี่ยนรูปประโยคให้มี ‘part’จะได้

3) “Wheels are parts of bicycles.” // “ล้อเป็นส่วนประกอบของจักรยาน”

4) “Bicycles are partly aluminum.” // “จักรยานมีส่วนหนึ่งเป็นอลูมิเนียม”

จากตัวอย่างข้างต้นจะเห็นว่า ‘part’ในประโยคแรกจะหมายถึงการเป็นส่วนประกอบ (component-integral object) และ ‘part’ในประโยคที่ 2 จะหมายถึงการเป็นองค์ประกอบ (stuff-object) จึงได้มีการแบ่ง ‘Part-Whole relation’เป็น 6 หมวด ดังต่อไปนี้

2.1.1.1 Component-Integral Object

ความสัมพันธ์หมวดนี้จะหมายถึงการนำวัตถุแต่ละชิ้นมาประกอบรวมกันเป็นวัตถุขนาดใหญ่ขึ้นมีคุณสมบัติมากขึ้นใช้งานได้หลากหลายขึ้นเช่น

1) “Wheels are parts of cars.” // “ล้อเป็นส่วนประกอบของรถยนต์”

2) “Chapters are parts of books.” // “บทเป็นส่วนประกอบของหนังสือ”

3) “The refrigerator is part of the kitchen.” // “ตู้เย็นเป็นส่วนประกอบของห้องครัว”

4) “Belgium is part of NATO.” // “ประเทศเบลเยี่ยมเป็นเป็นส่วนหนึ่งขององค์การสนธิสัญญาป้องกันแอตแลนติกเหนือ”

1) “Phonology is part of linguistics.” // “สัทวิทยาเป็นสาขาของภาษาศาสตร์”

ส่วนประกอบแต่ละส่วนของตัวอย่างข้างต้นต่างก็เป็นวัตถุที่มีคุณสมบัติของตัวเองอยู่ แต่เมื่อนำมาประกอบกันจะได้เป็นกลุ่มใหม่, องค์กรใหม่, หรือวัตถุชิ้นใหม่ ซึ่งสามารถแยกวัตถุแต่ละชิ้นออกจากกันได้ ซึ่งเมื่อแยกออกจากกันแล้วก็ยังได้วัตถุชิ้นเก่าก่อนนำมาประกอบรวมกัน

2.1.1.2 Member-Collection

ความสัมพันธ์ประเภทนี้จะแตกต่างจากความสัมพันธ์แบบ Component-Integral Object ข้างต้นแม้จะเกิดจากการประกอบกันขึ้นมาก็ตาม เนื่องจากความสัมพันธ์แบบ Member-Collection ไม่ได้อาศัยคุณลักษณะของแต่ละองค์ประกอบเช่น

1) “A tree is part of forest.” //“ต้นไม้เป็นสมาชิกของป่า”

2) “A juror is part of a jury.” // “ตุลาการเป็นสมาชิกของคณะตุลาการ”

3) “This ship is part of a fleet.” // “เรือลำนี้เป็นสมาชิกของกองทัพเรือ”

แต่ความสัมพันธ์แบบ Member-Collection ก็แตกต่างจากคลาสเช่นกัน เพราะความสัมพันธ์ในคลาสไม่ใช่ความสัมพันธ์แบบMeronymเพราะใช้ ‘part’แทนไม่ได้เช่น

1) “The nile is a river.” // “ไนล์คือแม่น้ำ”

2) “Fido is a dog.” //“ฟิโด้คือสุนัข”

จะเห็นว่าความสัมพันธ์ในคลาสจะเป็นแบบความคล้ายคลึงกันของแต่ละสมาชิก กล่าวคือ มีคุณสมบัติพื้นฐานเหมือนกัน ในขณะที่ความสัมพันธ์แบบ Member-Collection จะเป็นแบบการอยู่ร่วมกัน, อยู่ใกล้ชิดกัน หรือเป็นการติดต่อเชื่อมโยงกันทางสังคม ดังจะเห็นว่า ‘สมาชิกของป่าคือต้นไม้’ ต้นไม้ในที่นี้จะหมายถึงต้นไม้ที่อยู่ใกล้หรืออยู่ติดกับต้นไม้อื่นๆ ซึ่งอีกความหมายหนึ่งของความสัมพันธ์ประเภทนี้ก็คือการรวมกลุ่ม (groups)

2.1.1.3 Portion-Mass

ความสัมพันธ์ของบางส่วนของวัตถุหรือมิติทางกายภาย จะแตกต่างจาก 2 ความสัมพันธ์ข้างต้นนั่นคือแต่ละองค์ประกอบที่ประกอบกันด้วยความสัมพันธ์ประเภทนี้จะเหมือนกันเช่น

1) “This slice is part of a pie.” // “พายชิ้นนี้เป็นส่วนหนึ่งของก้อนพาย”

2) “A yard is part of a mile.” //“หลาเป็นส่วนประกอบของไมล์”

3) “This hunk is part of my clay.” // “ดินก้อนนี้เป็นส่วนหนึ่งของดินของฉัน”

จากตัวอย่างพบว่า ทุกๆชิ้นส่วนของแต่ละองค์ประกอบจะต้องเหมือนกันเมื่อนำมาประกอบกันก็จะได้ชิ้นที่ใหญ่กว่าเดิม โดยลักษณะของPortion-Mass มักจะเป็นหน่วยการวัดพื้นฐานเช่น นิ้ว, เมตร, ปอนด์, ชั่วโมง โดยความสัมพันธ์แบบ Portion-Mass มักจะใช้ตัวดำเนินการทางคณิตศาสตร์ (บวก, ลบ, คูณ, หาร)

2.1.1.4 Stuff-Object

ความสัมพันธ์ประเภทนี้จะเป็นความสัมพันธ์ขององค์ประกอบที่เป็นส่วนหนึ่ง(‘is partly’)ของวัตถุ แต่ไม่เหมือนความสัมพันธ์แบบ Component-Integral Object ดังตัวอย่าง

1) “A martini is partly alcohol.” // “มาร์ตินี่มีส่วนหนึ่งเป็นแอลกอฮอล์”

2) “The bike is partly steel.” // “จักรยานมีส่วนหนึ่งเป็นเหล็กกล้า”

3) “Water is partly hydrogen.” // “น้ำมีส่วนหนึ่งเป็นไฮโดรเจน”

จะเห็นว่าลักษณะของความสัมพันธ์ที่ใช้‘is partly’จะไม่สามารถแยกส่วนแอลกอฮอล์ออกจากมาร์ตินี่ หรือ แยกส่วนเหล็กกล้าออกจากจักรยาน หรือ แยกส่วนไฮโดรเจนออกจากน้ำ เหมือนกับการแยกวัตถุในความสัมพันธ์แบบ Component-Integral Object

‘is partly’จะหมายถึง‘ทำด้วย’,‘ผลิตด้วย’ หรือ ‘made of’ โดยที่ลักษณะของ ‘made of’ เป็นการสร้างวัตถุ, ผลิตวัตถุ เมื่อเราต้องการแยกองค์ประกอบของวัตถุจะทำไม่ได้ด้วยวิธีทางกายภาพเหมือนกับการแยกวัตถุในความสัมพันธ์แบบ Component-Integral Object

พบว่าลักษณะของ ‘Stuff-Object’ เมื่อพิจารณาด้วยตาเปล่าจะเป็นวัตถุที่เป็นเนื้อเดียวกัน เช่น เหล็กกล้า, เครื่องดื่มแอลกอฮอล์เป็นต้น แต่เมื่อพิจารณาถึงระดับโมเลกุลหรืออะตอมวัตถุเกือบทุกชนิดก็จะเป็น‘Stuff-Object’ได้

ในบางครั้งการพิจารณาวัตถุที่ซับซ้อนว่าเป็นวัตถุชนิดไหน หรือมีความสัมพันธ์ขององค์ประกอบเป็นอะไร เช่น ‘จักรยาน’ จะเห็นว่าจักรยานมีองค์ประกอบหลายชนิดเมื่อแยกล้อออกจากจักรยาน จักยานนั้นก็ยังเรียกว่าจักยาน หรือ ‘สลัด’ เมื่อเอามะเขือเทศออกจากสลัด สลัดก็ยังคงเป็นสลัด ลักษณะของวัตถุแบบนี้จะไม่ใช่ ‘Stuff-Object’แต่ถ้าวัตถุนั้นขาดองค์ประกอบบางอย่าง

1) “Alcohol is a constituent of wine.” // “แอลกอฮอล์เป็นส่วนผสมของไวน์”

2) “Tomato is an ingredient of salad.” // “มะเขือเทศเป็นส่วนผสมของสลัด”

จะเห็นว่าทั้ง แอลกอฮอล์ และ มะเขือเทศ ต่างก็เป็นเป็นส่วนผสมของวัตถุ โดย มะเขือเทศต้องผ่าน ‘การจัดเตรียม’ ก่อนนำไปปรุงเป็นสลัด แต่แอลกอฮอล์ไม่ต้องผ่านขั้นตอน‘การจัดเตรียม’ซึ่งใน‘Part-Whole relation’ ทั้ง 6 หมวดนี้ ‘ingredient’ จะหมายถึง ‘component’นั่นคือแอลกอฮอล์เป็น Stuff-Object และมะเขือเทศเป็น Component-Integral Object

2.1.1.5 Feature-Activity

ความสัมพันธ์แบบนี้จะเป็นการใช้ ‘part’ ในคุณลักษณะของกิจกรรมและการดำเนินการ เช่น

1) “Paying is part of shopping.” // “การจ่ายเงินเป็นส่วนหนึ่งของการซื้อสินค้า”

2) “Ovulation is part of the menstrual cycle.” // “การตกไข่เป็นส่วนหนึ่งของรอบประจำเดือน”

3) “Bidding is part of playing bridge.” // “การบิดเป็นส่วนหนึ่งของเกมบริดจ์”

แต่ความสัมพันธ์แบบ Feature-Activity จะไม่สามารถอยู่ในรูป ‘X has Y’ได้

2.1.1.6 Place-Area

ความสัมพันธ์แบบนี้จะเป็นความสัมพันธ์ของพื้นที่(area) กับสถานที่(place) เช่น

1) “The Everglades are part of Florida.” // “Everglades เป็นส่วนหนึ่งของรัฐฟลอริดา”

2) “An oasis is a part of a desert.” // “โอเอซิสเป็นส่วนหนึ่งของทะเลทราย”

3) “The baseline is a part of a tennis court.” // “เส้นหลังเป็นส่วนหนึ่งของสนามเทนนิส”

Place-Area จะมีลักษณะคล้าย Member-Collectionตรงที่ไม่ได้อาศัยคุณลักษณะของแต่ละองค์ประกอบเพื่อมาประกอบกันแต่จะเป็นตำแหน่งของสถานที่ (place) ในพื้นที่ (area) โดยที่ไม่สามารถแบ่งแยก สถานที่ (place) ออกจากพื้นที่ (area)ได้

2.1.2 อาณาจักรพืช(Kingdom Plantae)

พืชเป็นสิ่งมีชีวิตที่มีกำเนิดขึ้นมาแล้วไม่ต่ำกว่า 400ล้านปี มีหลักฐานหลายอย่าง ที่ทำให้เชื่อว่า พืชมีวิวัฒนาการมาจากสาหร่ายสีเขียว กลุ่ม Charophytesโดยมีการปรับตัวจากสภาพที่เคยอยู่

ในน้ำ ขึ้นมาอยู่บนบก ด้วยการสร้างคุณสมบัติต่างๆ ที่เหมาะสมขึ้นมา เช่น มีการสร้างคิวติน(cutin) ขึ้นมาปกคลุมผิว ของลำต้นและใบเรียกว่า คิวทิเคิล (cuticle) เพื่อป้องกันการสูญเสียน้ำ และการเกิดสโทมาตา (stomata) เพื่อทำหน้าที่ระบายน้ำ และแลกเปลี่ยนก๊าซ เป็นต้น

พืชมีโครงสร้างที่ประกอบขึ้นด้วยหลายเซลล์ที่มารวมกลุ่มกันเป็นเนื้อเยื่อที่ทำหน้าที่เฉพาะ อย่างเซลล์ของพืชมีผนังเซลล์ที่มีสารประกอบ เซลลูโลส (cellulose) เป็นองค์ประกอบ ที่พบเป็นส่วนใหญ่ พืชทุกชนิดที่คุณสมบัติที่สามารถสร้างอาหารได้เอง จากกระบวนการ สังเคราะห์ด้วยแสง โดยบทบาทของรงควัตถุ คลอโรฟิลล์(chlorophyll a & b) ที่อยู่ในคลอโรพลาสต์เป็นสำคัญ รงควัตถุหลักที่พบได้ในเซลล์พืช จะเหมือนกับที่พบในเซลล์ ของสาหร่ายสีเขียว ได้แก่ คลอโรฟิลล์เอ คลอโรฟิลล์บี และแคโรทีนอยด์ นอกจากนี้ พืชยังสะสมอาหารในรูปของแป้ง (starch)

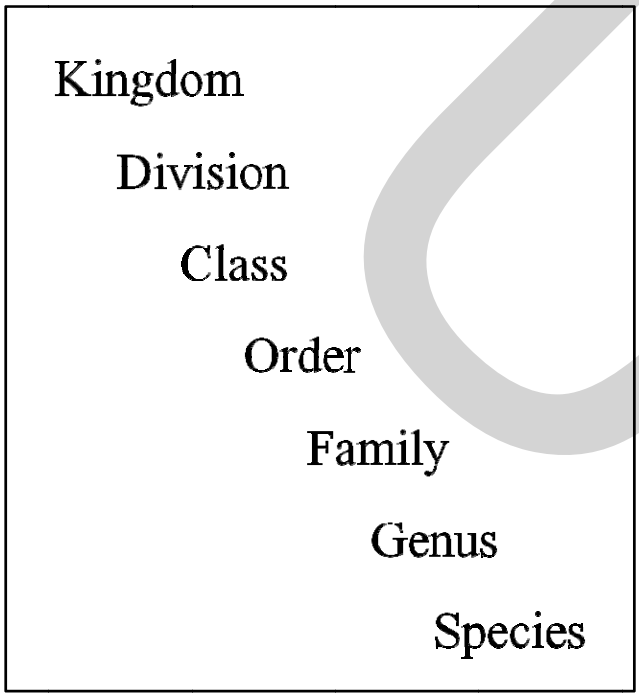
คุณสมบัติพื้นฐานที่แสดงว่าสิ่งมีชีวิตประเภทนั้นคือพืช มี 3 ประการ คือ

1) พืชต้องมีคลอโรฟิลล์ (chlorophyll)

2) พืชต้องมีผนังเซลล์ที่มีสารประกอบ เซลลูโลส (cellulose) เป็นองค์ประกอบ

3) พืชไม่สามารถเคลื่อนที่ได้เอง

การจำแนกสิ่งมีชีวิตในอาณาจักรพืช ทางชีววิทยาจะเรียกว่า อนุกรมวิธานของพืช (taxonomy) ซึ่งจะมีลำดับดังต่อไปนี้



ภาพที่ 2.1 แสดงอนุกรมวิธานพืช

ในทางชีววิทยา(แบบเรียนชีววิทยาของ สสวท.)ได้จำแนกสิ่งมีชีวิตในอาณาจักรพืช ออกเป็นกลุ่มใหญ่ที่เรียกว่า ดิวิชัน(division) จำนวน 9 ดิวิชัน ดังต่อไปนี้

2.1.2.1 ดิวิชันไบรโอไฟตา

เรียกโดยทั่วไปว่า ไบรโอไฟต์ (bryophyte) มีทั้งสิ้นประมาณ 16,000 ชนิด พืชในดิวิชันนี้มีขนาดเล็ก มีโครงสร้างง่าย ๆ ยังไม่มีราก ลำต้นและใบที่แท้จริง ชอบอาศัยอยู่ตามที่ชุ่มชื้น การสืบพันธุ์แบบอาศัยเพศยังต้องอาศัยน้ำสำหรับให้สเปิร์มที่มีแฟลกเจลลา (flagella) ว่ายไปผสมกับไข่ ต้นที่พบเห็นโดยทั่วไปคือแกมีโทไฟต์ (มีแกมีโทไฟต์เด่น) รูปร่างลักษณะมีทั้งที่เป็นแผ่นหรือแทลลัส (thallus) และคล้ายลำต้นและใบของพืชชั้นสูง (leafy form) มีไรซอยด์ (rhizoid) สำหรับยึดต้นให้ติดกับดินและช่วยดูดน้ำและแร่ธาตุ มีส่วนคล้ายใบ เรียก phylloidและส่วนคล้ายลำต้นเรียกว่า cauloidแกมีโทไฟต์ของไบรโอไฟต์มีสีเขียวเพราะมีคลอโรฟิลล์สามารถสร้างอาหารได้เอง ทำให้อยู่ได้อย่างอิสระ เมื่อแกมีโทไฟต์เจริญเต็มที่จะสร้างเซลล์สืบพันธุ์คือสเปิร์มและไข่ต่อไป ภายหลังการปฏิสนธิของสเปิร์มและไข่จะได้ไซโกตซึ่งแบ่งตัวเจริญต่อไปเป็นเอ็มบริโอและสปอร์โรไฟต์ตามลำดับ สปอโรไฟต์ของ ไบรโอไฟต์มีรูปร่างลักษณะง่าย ๆ ไม่สามารถอยู่ได้อย่างอิสระจะต้องอาศัยอยู่บนแกมีโทไฟต์ตลอดชีวิต พืชในดิวิชันนี้สร้างสปอร์เพียงชนิดเดียว ตัวอย่างของพืชในกลุ่มนี้ได้แก่ ลิเวอร์เวิร์ต (liverwort), ฮอร์นเวิร์ต (hornwort)และ มอส(moss)

2.1.2.2 ดิวิชันไซโลไฟตา

พืชในดิวิชันนี้ที่พบได้ในประเทศไทย ได้แก่ Psilotumรู้จักกันในชื่อไทยว่า หวายทะนอยสปอโรไฟต์ของพืชนี้ มีรูปร่างลักษณะง่ายๆ คือมีแต่ลำต้นยังไม่มีรากและใบ ลำต้นมีลักษณะเป็นไม้เนื้ออ่อนขนาดสูงประมาณ 20 –30 เซนติเมตร ขึ้นอยู่ตามพื้นดิน (tcrrestrial) หรือเกาะติดกับต้นไม้อื่น (epiphyte) ลำต้นแบ่งออกเป็น 2 ส่วน คือ ส่วนที่อยู่ใต้ดินเป็นลำต้นชนิดไรโซม (rhizome) มีสีน้ำตาล และมีไรซอยด์ทำหน้าที่ดูดน้ำและแร่ธาตุ ลำต้นส่วนที่อยู่เหนือพื้นดิน (acrial stem) มีสีเขียว มีลักษณะเป็นเหลี่ยม ลำต้นส่วนนี้ทำหน้าที่สังเคราะห์แสง ทั้งลำต้นใต้ดินและลำต้นเหนือพื้นดิน แตกกิ่งเป็น 2 แฉก (dichotomous branching) ที่ส่วนของลำต้นเหนือพื้นดินมีระยางค์เล็กๆ (appcndage) ยื่นออกมาเห็นได้ทั่วไป สปอโรไฟต์ที่เจริญต้นที่จะสร้างอับสปอร์ที่มีรูปร่างเป็น 3 พู ที่ซอกของระยางค์บนลำต้นเหนือพื้นดิน อับสปอร์สร้างสปอร์ชนิดเดียว แกมีโทไฟต์มีขนาดเล็ก สีน้ำตาลไม่มีคลอโรฟิลล์ รูปร่างเป็นแท่งทรงกระบอก แตกแขนงได้

2.1.2.3 ดิวิชันไลโคไฟตา

สปอโรไฟต์ของพืชดิวิชันนี้มีราก ลำต้น และใบครบทุกส่วน มีลักษณะเป็นไม้เนื้ออ่อนที่มีขนาดไม่ใหญ่มากนัก พวกที่เจริญอยู่บนพื้นดิน อาจมีลำต้นตั้งตรงหรือทอดนอน บางชนิดอาศัยเกาะบนต้นไม้อื่น ลำต้นแตกกิ่งเป็น 2 แฉก ใบมีขนาดเล็ก เป็นใบแบบไมโครฟิลล์ (microphyll) คือ

Lycopodiumรู้จักในชื่อไทยว่า ช้องนางคลี่ สร้อยสุกรม สามร้อยยอด และหางสิงห์เป็นต้น ที่พบในปัจจุบันมีประมาณ 200 ชนิด ใบในขนาดเท่า ๆ กันเรียงตัวเป็นเกลียวโดยรอบลำต้นและกิ่ง เป็นพืชที่สร้างสปอร์ชนิดเดียว แกมีโทไฟต์มีขนาดเล็ก บางชนิดมีคลอโรฟิลล์เจริญอยู่บนพื้นดิน บางชนิดไม่มีคลอโรฟิลล์เจริญอยู่ใต้ดิน

2.1.2.4 ดิวิชันสฟีโนไฟตา

ดิวิชันสฟีโนไฟตา (Division Sphenophyta) พืชที่มีท่อลำเลียงในดิวิชันนี้มีเพียง วงศ์เดียว คือ Equisetaceaeแกมีโตไฟต์มีขนาดเล็ก เจริญอยู่ใต้ดิน สปอโรไฟต์มีขนาดใหญ่ อายุยืน มีซิลิกา ลำต้นเป็นข้อปล้องชัดเจน ปล้องเป็นร่องและสัน ข้อมีใบแบบไมโครฟิลล์อยู่รอบข้อเรียงแบบ whorl เป็น homosporous plant โดยสปอแรงเจียมเจริญอยู่บนโครงสร้างที่เรียกว่าสปอแรงจิโอฟอร์ (sporangiophore) ตัวอย่างของพืชในกลุ่มนี้ได้แก่ หญ้าถอดปล้อง (equisetum)

2.1.2.5 ดิวิชันเทอโรไฟตา

พืชดิวิชันนี้มีชื่อทั่วไปว่า เฟิร์น (fern) มีจำนวนมากที่สุดในบรรดาพโรไฟต์ของเฟิร์นมีราก ลำต้นและใบเจริญดี เฟิร์นส่วนใหญ่มีลำต้นใต้ดิน ใบของเฟิร์นเรียกว่า ฟรอนด์ (frond) เป็นส่วนที่เห็นเด่นชัด มีขนาดใหญ่เป็นใบแบบเมกะฟิลล์ (megaphyll) มีรูปร่างลักษณะเป็นหลายแบบ มีทั้งที่เป็นใบเดี่ยว (simple leaf) และใบประกอบ (compound leaf) ใบอ่อนของเฟิร์นมีลักษณะพิเศษ คือ จะม้วนเป็นวง (circinatevenantion) สปอโรไฟต์ที่เจริญเต็มที่จะสร้างอับสปอร์ ซึ่งมารวมกลุ่มอยู่ที่ด้านใต้ใบ แต่ละกลุ่มของอับสปอร์เรียกว่า ซอรัส (sorus) เฟิร์นส่วนใหญ่สร้างสปอร์ชนิดเดียว ยกเว้นเฟิร์นบางชนิดที่อยู่ในน้ำ และที่ชื้นแฉะ ได้แก่ จอกหูหนู แหนแดง และผักแว่นมีการสร้างสปอร์ 2 ชนิด

แกมีโทไฟต์ของเฟิร์นที่สร้างสปอร์ชนิดเดียว มีลักษณะเป็นแผ่นแบนบางสีเขียว (มีคลอโรฟิลล์) ด้านล่างมีไรซอยด์ ส่วนใหญ่มักมีรูปร่างคล้ายรูปหัวใจ (prothallus)

2.1.2.6 ดิวิชันโคนิเฟอโรไฟตา

เป็นจิมโนสเปิร์มที่มีจำนวนมากที่สุด มีหลายสกุลด้วยกัน ที่รู้จักกันดีคือ Pinusได้แก่ สนสองใบ และสนสามใบ เป็นต้น สปอโรไฟต์ของ Pinusมีลักษณะเป็นไม้ยืนต้นขนาดค่อนข้างใหญ่ และแตกกิ่งก้านสาขาจำนวนมาก ใบมีขนาดเล็ก รูปร่างคล้ายเข็ม อยู่รวมกันเป็นกลุ่ม สปอโรไฟต์ที่เจริญเต็มที่จะสร้างโคนเพศผู้ที่มีขนาดเล็กและโคนเพศเมียที่มีขนาดใหญ่บนต้นเดียวกัน

2.1.2.7 ดิวิชันไซแคโดไฟตา

พืชดิวิชันนี้มีอยู่ประมาณ 60 ชนิด ตัวอย่างที่รู้จักกันดีคือ พวกปรง (Cycas) สปอโรไฟต์มีลำต้นอวบ เตี้ย และมักไม่แตกแขนง มีใบเป็นใบประกอบแบบขนนกขนาดใหญ่ เกิดเป็นกระจุกที่บริเวณยอดของลำต้น ใบย่อยมีรูปร่างเรียวยาว และแข็งสปอโรไฟต์ที่เจริญเต็มที่จะสร้างโคนเพศผู้และโคนเพศเมีย แยกตัวกัน

2.1.2.8 ดิวิชันกิงโกไฟตา

ปัจจุบันมีเพียงชนิดเดียวคือ Ginkgo bilobaหรือแปะก๊วย เป็นพืชที่ขึ้นอยู่ในเขตอบอุ่น เช่น ในประเทศจีน สปอโรไฟต์มีลักษณะเป็นไม้ยืนต้นขนาดสูงใหญ่ แตกกิ่งก้านสาขาเป็นจำนวนมาก ใบมีรูปร่างคล้ายพัด สปอโรไฟต์ที่เจริญเติบโตเต็มที่จะสร้างโคนเพศผู้และโคนเพศเมียแยกต้นกัน

2.1.2.9 ดิวิชันแอนโทไฟตา

พืชในกลุ่มนี้จะเป็นกลุ่มพืชชั้นสูง หรือพืชที่มีดอก ผล และสืบพันธ์ด้วยเมล็ดแบ่งออกได้เป็น 2 คลาส คือ

1.คลาสไดคอทีเลโดเนส (Class Dicotyledones) ได้แก่ พืชใบเลี้ยงคู่ทั้งหมด มีอยู่ประมาณ 170,000 ชนิด ลักษณะทั่วไปคือ มีใบเลี้ยง 2 ใบ เส้นใบเป็นร่างแห รากเป็นระบบรากแก้ว และส่วนประกอบของดอก (เช่น กลีบเลี้ยง กลีบดอก) มีจำนวนเป็น 4–5 หรือ ทวีคูณของ 4–5

2.คลาสมอโนคอทีเลโดเนส (Class Monocotyledones) ได้แก่ พืชใบเลี้ยงเดี่ยวทั้งหมดมีอยู่ประมาณ 60,000 ชนิด ลักษณะทั่วไป คือ มีใบเลี้ยงใบเดียว ใบมีเส้นใบเรียงตัวแบบขนาน รากเป็นระบบรากฝอย ส่วนประกอบของดอกมีจำนวนเป็น 3 หรือทวีคูณของ 3

จากทั้ง 9ดิวิชัน จะพบว่า มี 3 ดิวิชันที่พืชในกลุ่มนี้มีจำนวนน้อยมากจนใกล้สูญพันธ์คือ ดิวิชันไซโลไฟตา ดิวิชันสฟีโนไฟตาและ ดิวิชันกิงโกไฟตา ซึ่งในการวิจัยนี้จะไม่นำมาร่วมด้วย

2.1.3 สารประกอบทุติยภูมิ (Secondary Metabolites)

ตัวตั้งแต่อดีตกาลมนุษย์มีการใช้สารเคมีจากพืช ในขั้นต้นได้ใช้เพื่อการล่าและการทำลายชีวิต สารดังกล่าวได้แก่ ทูโบคูแรรีน (tubocurarine) ที่ส่วนใหญ่ได้จากพืชเถา Chondrodendrontomentosumวงศ์ Menispermaceaeในป่าแถบลุ่มน้ำอะเมซอน สารดังกล่าวหยุดการส่งผ่านของกระแสประสาทไปยังกล้ามเนื้อ เมื่อใช้อาบปลายลูกธนูจะทำให้สัตว์ที่ถูกยิงด้วยลูกธนูดังกล่าวเกิดเป็นอัมพาตเคลื่อนไหวไม่ได้ ในการปลิดชีวิตมีการใช้เฮมล็อก (Conium maculatum วงศ์ Umbelliferae) ซึ่งทำให้เกิดอัมพาตกล้ามเนื้อตามด้วยการชักและจบชีวิตลงด้วยอัมพาตของระบบทางเดินหายใจ เฮมล็อกมีพิษทั้งต้นเพราะมีอัลคาลอยด์มากอัลคาลอยด์ที่สำคัญคือโคนิอีน (coniine)

สารประกอบทุติยภูมิ (Secondary Metabolites) คือกลุ่มของสารเคมีที่สร้างโดยพืช สัตว์ ราหรือแบคทีเรีย ที่ไม่มีความจำเป็นในขั้นวิกฤติต่อสิ่งมีชีวิตผู้ผลิต หากแต่ถูกสร้างโดยขบวนการทางชีวเคมีของผู้ผลิตเป็นสารจำเพาะต่อผู้ผลิตนั้นๆ เป็นสารที่ให้กลิ่น สี หรือสรรพคุณจำเพาะของพืชที่พบในอาหาร ยาและสารพิษต่างๆ จากพืชและมีการกระจายตัวอย่างจำกัด

การใช้สมุนไพรและตัวยาจากธรรมชาติมีประวัติการใช้งานอันยาวนาน การใช้พืชทั้งต้นหรือการเตรียมยาอย่างหยาบๆ มีข้อด้อยหลายประการดังนี้

1. ความไม่แน่นอนของปริมาณสารที่ออกฤทธิ์จากแหล่งเพาะปลูกที่ต่างกัน หรือจากฤดูหนึ่งๆ ไปยังฤดูอื่น รวมถึงส่วนของพืชที่ต่างกันและลักษณะทางกายวิภาคที่ต่างกันด้วย ตัวอย่างได้แก่ เปล้าน้อยพืชที่ปลูกในจังหวัดปราจีนบุรีมีสารออกฤทธิ์น้อยกว่าพืชที่ปลูกในจังหวัดประจวบคีรีขันธ์

2. การสะสมของสารที่ไม่ต้องการซึ่งอาจมีส่วนก่อให้เกิดความเปลี่ยนแปลงต่อสารออกฤทธิ์เช่น มีฤทธิ์ต่างกันหรือมีฤทธิ์แรงขึ้น เช่น ลิวโรซีน (leurosine) เป็น indole alkaloid คล้ายอัลคาลอยด์จากต้นพังพวยฝรั่ง สารสกัดหยาบอัลคาลอยด์ไม่มีฤทธิ์ชีวภาพในเซลล์มะเร็งเม็ดเลือดขาวแต่สารบริสุทธิ์มีพิษต่อเซลล์มะเร็งนี้ในหลอดทดลองอย่างชัดเจน

3. การสูญเสียสารออกฤทธิ์เนื่องจากความแตกต่างกันของวิธีเก็บเกี่ยวเก็บรักษาและการเตรียมวัตถุดิบเพื่อการใช้ในการรักษา ตัวอย่างได้แก่ต้นฟีเวอร์ฟิวจะรักษาสารสำคัญไว้ได้มากที่สุดถ้าอบพืชแห้งทันทีที่ 60 องศาเซลเซียสในระยะเวลาอันสั้น ถ้าตากในร่มที่อุณหภูมิห้องต้องใช้เวลาตาก 3-7 วันและจะได้ปริมาณพาร์ทีโนไลด์ไม่เต็มที่

ข้อดีของการสกัดแยกสารผลิตภัณฑ์ธรรมชาติให้ได้สารออกฤทธิ์มีดังนี้

1. สารออกฤทธิ์ในรูปของสารบริสุทธิ์สามารถนำจ่ายได้ในปริมาณที่ถูกต้องและสามารถวัดให้เท่ากันได้อีก

2. การรู้จักองค์ประกอบทางเคมีและโครงสร้างสารบริสุทธิ์สามารถนำไปสู่การพัฒนาวิธีทดสอบทางชีวภาพหาสารอื่นที่มีฤทธิ์เดียวกันจากสิ่งมีชีวิตในกลุ่มเดียวกันหรือกลุ่มที่ใกล้เคียงกันได้

3. การทราบโครงสร้างที่แน่นอนของสารออกฤทธิ์อาจเป็นการชี้นำให้เกิดการสังเคราะห์สารดังกล่าวหรือการเปลี่ยนแปลงองค์ประกอบโมเลกุลเล็กน้อยให้ได้สารที่มีฤทธิ์มากขึ้นหรือมีพิษต่ำลง ตัวอย่างได้แก่ สารโพโดฟิลโลท็อกซิน (podophyllotoxin) ซึ่งเป็นลิกแนนในยางของ Podophyllumpeltatumมีฤทธิ์ใช้รักษามะเร็งได้แต่มีผลข้างเคียงมาก ขณะที่ยาสังเคราะห์อีโทโปไซด์ (etoposide) และเทนิโปไซด์ (teniposide) มีสูตรโครงสร้างหลักเหมือนกันต่างกันที่อนุพันธ์มีฤทธิ์แรงกว่าและมีผลข้างเคียงน้อยกว่าสารที่ได้จากธรรมชาติมาก

4. สามารถศึกษาถึงเส้นทางชีวสังเคราะห์ของสารนั้นๆ ทำให้รู้จักสารตั้งต้นของมัน อาจมีผลทำให้ใช้ส่วนของพืชที่มีสารสำคัญน้อยลงโดยใช้สารอื่นในเส้นทางชีวสังเคราะห์ทดแทน หรือนำสารในเส้นทางเมตาบอลิสมมาเป็นสารตั้งต้นในการสังเคราะห์ตัวยาที่ต้องการ ตัวอย่าง ได้แก่ ยาพาซิทาเซลหรือแท็กซอล เป็นต้น

ซึ่งในงานวิจัยนี้ได้แบ่งประเภทของสารประกอบทุติยภูมิออกเป็น 5 กลุ่มใหญ่ตามโครงสร้างหลักได้แก่

1. อัลคาลอยด์ (Alkaloids)
2. เทอร์พีนอยด์ (Terpenoid)
3. ไกลโคไซด์ (Glycoside)
4. ฟีนอล (Phenols)
5. โพลีคีไทด์(Polyketides)

2.1.4 ตัวจำแนกประเภทเนอีฟเบย์ (Naïve Bayes Classifier)

ตัวจัดประเภทเนอีฟเบย์ (Naïve Bayes classifier, NB) (Mitchell 1997) หรือ ตัวเรียนรู้ NB เป็นวิธีการเรียนรู้ที่นิยมใช้กันมาก และเป็นการเรียนรู้ที่อยู่บนพื้นฐานของความน่าจะเป็น (Probability) กับข้อมูลที่สังเกต (Observed Data) ตามที่ Mitchell T.M., (1997) ได้กล่าวว่าตัวจัดประเภท NB สามารถประยุกต์ใช้กับงานเรียนรู้ที่ซึ่งแต่ละตัวอย่าง x (Instance x) ได้ถูกอธิบายโดยการเชื่อมโยงค่าแอททริบิวท์ (Attribute Values) ต่างๆ และที่ซึ่งฟังก์ชันเป้าหมาย (Target Function, f(x)) สามารถแสดงค่าคลาส (Class Value, v) จาก คลาสไฟไนท์เซท (Class Finite Set, V) ดังนั้นเซทของตัวอย่างการเรียนรู้ของฟังก์ชันเป้าหมายได้ถูกกำหนดไว้ให้ และเมื่อมีตัวอย่างใหม่เกิดขึ้นก็สามารถอธิบายได้ คือบอกค่าคลาสได้ด้วยทูปเพิล (Tuple) ของค่าแอททริบิวท์$<a_1, a_2, \ldots, a_n>$นั่นคือตัวเรียนรู้ทำนายค่าเป้าหมายหรือการจัดแบ่งประเภทสำหรับตัวอย่างใหม่ที่เข้ามา

แนวทางเบย์ที่จะจัดประเภทให้กับตัวอย่างใหม่ที่เข้ามานั้นเป็นการกำหนดค่าเป้าหมายที่มีโอกาสเป็นไปได้มากสุด หรือที่เรียกว่า $v_{maximum}$ a posterior (v_{MAP}) เมื่อกำหนดค่าแอททริบิวท์ต่างๆให้ $<a_1, a_2, \ldots, a_n>$ที่ใช้อธิบายตัวอย่าง ดังแสดงในสมการ(2) และ (3)

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, \ldots, a_n) \quad (1)$$

$$v_{\text{MAP}} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, \ldots, a_n | v_j) P(v_j) \quad (2)$$

ตัวจัดประเภท NBดำเนินงานบนพื้นฐานของข้อสมมติฐานแบบง่ายๆที่มีเงื่อนไขว่าค่าแอททริบิวท์แต่ละแอททริบิวท์ จะต้องเป็นอิสระต่อกันเมื่อกำหนดค่าเป้าหมายไว้ให้ กล่าวคือข้อสมมติฐานเป็นการกำหนดค่าเป้าหมายของตัวอย่าง (คือคลาสของตัวอย่าง) ฉะนั้นความน่าจะเป็นของการสังเกตการเชื่อมโยงกันของ $a_1, a_2, \ldots, a_n$ คือผลคูณของค่าความน่าจะเป็นของแอททริบิวท์ต่างๆ ดังนั้นตัวจัดประเภท NB (v_{NB}) สามารถแสดงได้ดังต่อไปนี้

$$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j) \quad (3)$$

สำหรับงานวิจัยนี้ เราได้ประยุกต์ใช้ตัวจัดประเภท NB ที่เป็นสมการ (4) สำหรับการเรียนรู้แยกประเภทของแนวความคิดของสารเคมีและแนวความคิดของพืชในแต่ละประโยคดังต่อไปนี้

$$StuffRelationForNaturalProduct$$
$$= \operatorname{argmax}_{class \in Class} P(class | c_1, c_2, .., c_{max1}, s_1, s_2, .., s_{max2}, a_1, a_2, .., a_{max3})$$
$$= \underset{class \in Class}{\operatorname{argmax}} P(class) \prod_{num=1}^{max1} P(c_{num}) \prod_{num=1}^{max2} P(s_{num}) \prod_{num=1}^{max3} P(a_{num}) \quad (4)$$

เมื่อ Class ={“yes”, “no”}

r = ขนาดกรอบหน้าต่างของคำระหว่างแนวคิดสารเคมีกับแนวคิดของพืช (ดูหัวข้อ 4.1) โดยr เป็นเซตของตัวเลข {1,2,3,4,5}ซึ่งเป็นค่า max3

r = set of {1,2,3,4,5}

สำหรับงานวิจัยนี้ $a_1, a_2, \ldots, a_n \in A$, A = เซตของคำที่มีความถี่สูงซึ่งอยู่ระหว่างแนวคิดของสารเคมี(c_i)กับแนวคิดของพืช(s_j)

$c_1, c_2, \ldots, c_i \in C$, C = เซตของแนวคิดของสารเคมี

$s_1, s_2, \ldots, s_j \in S$, S = เซตของแนวคิดของพืช

i = {1, 2, …, m}

j = {1,2, …, k}

จากการศึกษาข้อมูล max1= 16 max2 = 14 max3 = r หรือ n (nคือจำนวนคำที่อยู่ระหว่างคำที่เป็นแนวคิดของสารเคมี และคำที่เป็นแนวคิดของพืช) โดย r มีขนาด 3,4, และ5 และ n มีค่ามากที่สุดคือ 10

เมื่อตัวแปร “Class” เป็นไฟไนท์เซท (Finite Set) ของประเภทความสัมพันธ์แบบสตัฟฟ์สำหรับ แนวคิดของสารเคมีและความคิดของพืช (StuffRelationForNaturalProduct) โดยใช้ฟีเจอร์ (Feature) ต่างๆที่ประกอบด้วย 3 กลุ่มดังนี้ ฟีเจอร์กลุ่มแรกคือคำหรือสมาชิก (Element) ของเซต C ซึ่งเป็น เซตของแนวคิดของสารเคมี $<c_1, c_2, \ldots, c_i>$กลุ่มที่ 2 คือคำหรือสมาชิก ของเซต S ซึ่งเป็นเซตของแนวคิดของพืชที่เป็น Natural Sourceในระดับ Genus $<s_1, s_2, \ldots, s_j>$และกลุ่มที่ 3 ซึ่งแบ่งออกเป็น 2 แบบคือ แบบที่1เป็นคำหรือสมาชิกต่างๆใน เซตA ซึ่งเป็นเซตของคำที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช โดยจะใช้คำทั้งหมดเหล่านี้ $< a_1, a_2, .., a_r>$ที่มีความถี่มากที่สุด ภายใต้ขนาดกรอบหน้าต่าง r ของคำต่างๆระหว่างแนวคิดสารเคมีกับแนวคิดของพืช แบบที่2 เป็นคำทั้งหมด (n) ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช และเป็นคำใน เซตA $<a_1, a_2, .., a_n>$

2.2 งานวิจัยที่เกี่ยวข้อง

R. Girjuใช้เทคนิคการเรียนรู้ของเครื่องชนิด ID3(C4.5)โดยใช้เงื่อนไขในรูปแบบโครงสร้างทางไวยากรณ์ (Lexico-syntactic pattern) เพื่อหาความสัมพันธ์แบบ Part-Whole จากบทความ LA Times ของคลังข้อความ TREC-9โดยแบ่งความสัมพันธ์แบบ Part-Whole เป็น 3 ประเภทตาม WordNetได้แก่ Member-of(เช่น UK IS-MEMBER-OF NATO), Stuff-of (เช่น carbon IS-STUFF-OF coal) และ Part-of (เช่น leg IS-PART-OF table) จาก 10,000 ประโยคตัวอย่างพบว่าได้ความถูกต้องเป็น ความแม่นยำ(precision) 83% และ การเรียกคืน(recall) 98%

P. Pantelและ M. Pennacchiottiเสนอ Espresso algorithm ซึ่งเป็นอัลกอริทึ่มการเรียนรู้แบบมีการสอนแบบอ่อนๆ(weakly-supervised algorithm) เพื่อใช้สกัดความสัมพันธ์แบบ Is-a(Is-a relation) และความสัมพันธ์แบบ Part-of(Par-of relation) Espresso algorithm สามารถรองรับข้อมูลปริมาณมหาศาลที่ดาวน์โหลดมาจากเว็บด้วยการใช้รูปแบบทั่วไปที่น่าเชื่อถือในการเก็บข้อมูลซึ่งโดยทั่วไปผลที่ได้จากรูปแบบที่น่าเชื่อนี้จะให้ precision ที่สูงแต่ recallจะต่ำมากเช่น “Xconsists of Y” ในกรณีของความสัมพันธ์แบบ Part-of (Part-of relation)

โดยที่ P. Pantelและ M. Pennacchiotti ได้ทำการทดลองโดยใช้ข้อมูล 2 กลุ่ม คือ คลังข้อความ CHEM และ คลังข้อความ TREC-9 ซึ่งพบว่าจากคลังข้อความ CHEM ได้ความถูกต้องของความสัมพันธ์แบบ Part-of (Par-of relation) เป็น precision 51% และ relative recall เป็น 46 ในขณะ

อเป็นเครื่องมือในการสกัดความสัมพันธ์จากข้อความในส่วนของบทคัดย่อของคลังข้อความ MEDLINE ซึ่งเป็นคลังข้อความที่มีเนื้อหาเกี่ยวกับชีวการแพทย์ โดย RelExเป็นโมเดลที่ใช้ตรวจจับเหตุการณ์ร่วม (co-occurrences) ของนิพจน์ระบุนามภายในประโยคหรือบทคัดย่อและใช้ เซทของกฎอย่างง่าย, ใช้การเครื่องมือในการกำหนดประเภทของคำ (part-of-speech-tagging), noun-phrase-chunking และการขึ้นต่อกันเพื่อหาความสัมพันธ์ของนิพจน์ระบุนามของยีนกับนิพจน์ระบุนามของโปรตีนซึ่งจะใช้กฎทางภาษาศาสตร์ที่พบมากในภาษาอังกฤษดังนี้

(1) effector-relation-effectee (‘α activates β’)

(2) relation-of-effectee-by-effector (‘Activation of α by β’)

(3) relation-between-effector-and-effectee(‘Interaction between α and β’).

โดยได้ความถูกต้อง precision เป็น 80 % และ recall 80 %

G. I. Brownนำเสนอระบบการสกัดความสัมพันธ์โดยใช้ตัวจำแนก support vector machine (SVM)โดยใช้คลังข้อความที่เป็นบทวิจารณ์จาก J.D. Power และ Associates Sentiment 3 ประเภท ดังนี้ 1) ผู้เชี่ยวชาญเป็นผู้เขียน 2) บล็อกเกอร์เป็นผู้เขียน และ3) ผู้ใช้ทั่วไป โดยคุณลักษณะ (features) ที่ใช้ใน SVM ประกอบไปด้วย ‘คำ’ที่เกี่ยวข้องกับคำนามที่ปรากฏในวลี, ประเภทของนิพจน์ระบุนาม และคลาส token ซึ่งงานนี้ได้เน้นการสกัดระหว่างความแตกต่างของประเภทบทวิจารณ์ โดยได้ความถูกต้องของการสกัดแบบ Part-of เป็น precision 46% โดยเฉลี่ย และ recall เป็น 33% โดยเฉลี่ย

จะเห็นว่างานวิจัยเหล่านี้ดำเนินการบนพื้นฐานที่มีการกำหนดชนิดของคำ(part of speech) และโดยส่วนใหญ่จะสนใจเฉพาะคำนามอย่างเดียว แต่สำหรับงานวิจัยนี้ไม่สามารถกำหนดชนิดของคำได้อย่างอัตโนมัติ เนื่องจากความซับซ้อนของคำที่เป็นชื่อสารเคมี ต่อไปนี้จะแสดงให้เห็นถึงความซับซ้อนของชื่อสารเคมี เนื่องจากว่าชื่อเรียกของสารเคมีมีหลากหลายชื่อ เช่น ชื่อทางการค้า, ชื่อสามัญ, ชื่อIUPAC, ชื่อย่อ, ชื่อที่ใช้สูตรอย่างย่อ ดังตารางที่2.3จะเห็นว่าชื่อของสารเคมีจะมีสัญลักษณ์ต่างๆเช่น นขลิขิต “( ), [ ], { }”, ยัติภังค์ “-”, อะพอสทรอฟี “′”, จุลภาค “,”, มหัพภาค “.” รวมถึงช่องว่าง เป็นต้น บางครั้งก็จะมีอักษรภาษาอังกฤษที่เป็นตัวโดดและเป็นตัวพิมพ์ใหญ่

ตารางที่ 2.1 แสดงตัวอย่างของชื่อของสารเคมี

ชื่อของสารเคมี
1-(3-hydroxyphenyl)-3-(4-hydroxy-2,5-dimethoxyphenyl)propane
5-bromo-3-(3′-hydroxy-3′-methylpent-4′-enylidene)-2,4,4-trimethylcyclohexanone
isoliquiritigenin
methyl carnosate
4R-hydroxy-18-normanoyl oxide
N-pentacosanoyl-4,5-dihydroxy-tryptamine
alanine
2′-hydroxy- 6,4′,6″,4‴-tetramethoxy-[7-O-7″]-bisisoflavone
(7S,8R,1′S,5′S,6′R)-′2′,8′-3′,6′-dihydroxy-5′-methoxy-3,4-methylenedioxy-4′-oxo-8.1′,7.5′-neolignan

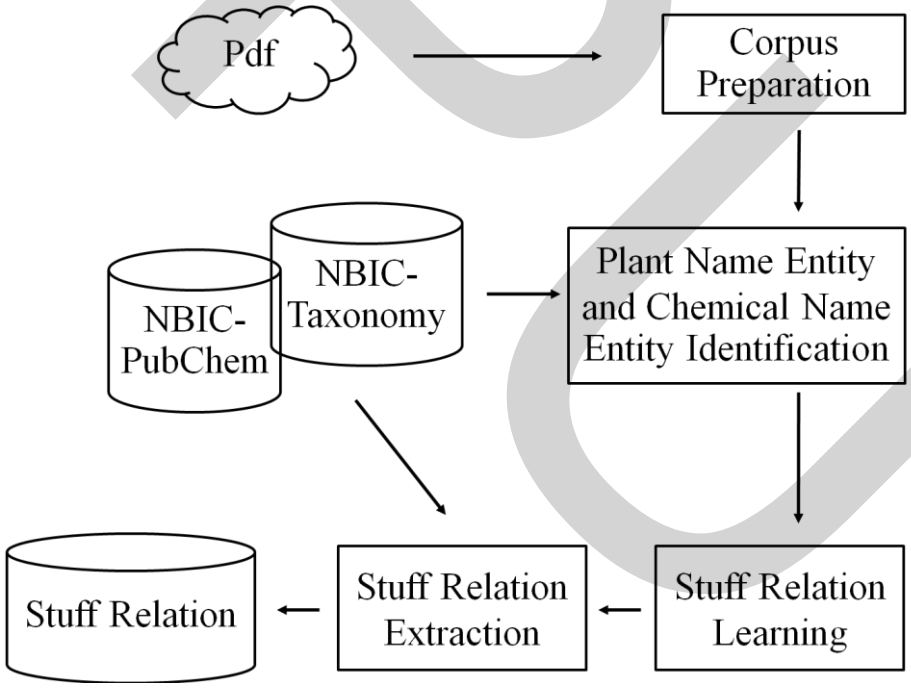
การที่จะใช้เครื่องมือในการระบุชนิดของคำโดยเฉพาะส่วนที่เป็นชื่อสารเคมีอย่างอัตโนมัตินั้นทำไม่ได้ เนื่องมาจากสัญลักษณ์ต่างๆ ที่ปรากฏอยู่ในชื่อของสารเคมี ดังนั้น งานวิจัยนี้จึงเสนอการสกัดความสัมพันธ์แบบสตัฟฟ์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ที่เป็นเนอีฟเบย์(Naïve Bayes) ด้วยการเปรียบเทียบคำต่างๆ ในเอกสารงานวิจัยทางวิทยาศาสตร์กับฐานข้อมูล NBCI-PubChemและ NBCI-taxonomy

บทที่ 3
วิธีการดำเนินการวิจัยและเครื่องมือ

ในบทที่ 3 นี้จะเป็นบทที่อธิบายถึง วิธีดำเนินการ วิจัยและเครื่องมือ โดยจะแบ่งออกเป็นสองส่วนคือ ส่วนของวิธีการดำเนินการวิจัย เป็นการแสดงลำดับขั้นตอนของ การสกัดความสัมพันธ์แบบสตัฟฟ์และส่วนของเครื่องมือที่ใช้ โดยมีรายละเอียดขั้นตอนดังต่อไปนี้

3.1 วิธีดำเนินการวิจัย

ขั้นตอนของการสกัดความสัมพันธ์แบบสตัฟฟ์ประกอบไปด้วย 4ขั้นตอน ดังนี้คือ 1.การเตรียมคลังข้อมูล 2.การระบุชื่อของพืชและสารเคมี 3.การเรียนรู้ความสัมพันธ์แบบสตัฟฟ์และ 4. การสกัดความสัมพันธ์แบบสตัฟฟ์ดังแสดงในภาพที่3.1



ภาพที่ 3.1 แสดงภาพรวมของระบบการสกัดความสัมพันธ์แบบสตัฟฟ์

1. การเตรียมคลังข้อมูล (Corpus Preparation)

เอกสารที่ใช้ในงานวิจัยนี้เป็นเอกสารทางวิชาการด้านวิทยาศาสตร์ในโดเมนสารเคมีที่เป็นผลิตภัณฑ์จากธรรมชาติ ซึ่งดาวน์โหลดมาจากสำนักพิมพ์ ACS เป็นไฟล์รูปแบบ PDF ซึ่งต้องนำมาแปลงให้อยู่ในรูปของข้อความ text โดยใช้ PDFTextStream (http://snowtide.com) เอกสารที่ผ่านการแปลงแล้วจะนำมาทำเหมืองข้อมูลจำนวน20,000 ประโยค โดยวิธี 10 folds cross-validation

ในการศึกษาพฤติกรรมทางภาษาทำให้ได้เซตคำศัพท์ที่อยู่ระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช(ตาราง 3.1) และจะทำการกำกับประโยคที่มีความสัมพันธ์แบบสตัฟฟ์ เป็นClass = “Yes” ด้วย ตัวกำกับสตัฟฟ์ (Stuff Tag) ดังแสดงในภาพที่ 3.2แล้วจะถูกแยกเอา Stop word Set ออก

ตารางที่ 3.1 แสดงความถี่ของคำศัพท์ที่น่าสนใจ

คำศัพท์ที่อยู่ระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช	ความถี่
isolated	939
extract	418
isolation	228
parts	170
leaves	167
aerial	153
roots	149
species	88
bark	81
seeds	58
obtained	55

<stuff_relation class="yes">Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii* by following their potential to inhibit in vitro the growth of *Plasmodium falciparum*.</ stuff_relation>
<stuff_relation class="no">Of these, the ethyl acetate extract obtained from the heartwood of *Dalbergia louvelii* R. Viguier (Fabaceae).</stuff_relation>
<stuff_relation class="yes">Although several isoflavonoids have been obtained from roots of *P. floribundum*, none of the abovementioned compounds have been isolated previously from this species.</stuff_relation>
<stuff_relation class="no">The cytotoxicity of the isolates obtained herein from *P. floribundum* has been evaluated against a small panel of cancer cell lines.</stuff_relation>

ภาพที่ 3.2 แสดงการกำกับ stuff relation class "YES/NO" แต่ละประโยค

2. การระบุชื่อของพืชและสารเคมี (Plant name entity and chemical name entity identification)

Four new sesquiterpenes, (8R*)-8-bromo-10-epi-â-snyderol (1), (8S*)-8-bromo-α-snyderol (2), 5-bromo-3-(3′-hydroxy-3′-methylpent-4′-enylidene)-2,4,4-trimethylcyclohexanone (3), and the epoxide (4), have been isolated from the chloroform-methanol extract of Laurencia obtusa, together with the three known compounds R-snyderol (5), R-snyderol acetate (6), and stigmasterol.

W_1 W_2 W_3 W_4 W_5 W_6 W_7 W_8 W_9 W_{10} W_{11} W_{12} W_{13} W_{14} W_{15} W_{16} W_{17} W_{18} W_{19} W_{20}

ภาพที่ 3.3 แสดงประโยคตัวอย่างของการระบุชื่อของสารเคมี

โดยเริ่มจาก

1. ระบุตำแหน่งของคำที่เป็นแนวคิดของพืชที่อยู่ในแต่ละประโยคกับ NCBI-Taxonomy

2. ระบุคำที่เป็นแนวคิดของสารเคมีโดยใช้คำที่อยู่ในกรอบหน้าต่างซึ่งมีขนาดตั้งแต่ 1, 2,.., n (nคือจำนวนคำที่อยู่ด้านหน้าของตำแหน่งของคำที่เป็นแนวคิดของพืช) แล้วทำการเปรียบเทียบคำที่อยู่ในกรอบหน้าต่างกับ NBCI-PubChemที่ระดับกรอบหน้าต่างขนาดต่างๆ โดยเลื่อนกรอบหน้าต่างด้วยระยะทางครั้งละ 1 คำ

3. ทำเช่นเดียวกับข้อ 2 แต่ n คือจำนวนคำที่อยู่ด้านหลังของตำแหน่ งของคำที่เป็นแนวคิดของพืช

3. การเรียนรู้ความสัมพันธ์แบบสตัฟฟ์ (Stuff Relation Learning)

ขั้นตอนการเรียนรู้ของเครื่องด้วย Naïve Bayes Classifier โดย ใช้ โดยเครื่องมือวีก้า (Weka Tool) (Mark Hall, 2009) หา Conditional Probabilities ของฟีเจอร์ (Feature) ต่างๆที่แบ่งออกเป็น 3 กรณีศึกษาดังนี้

กรณีศึกษาที่ 1 ฟีเจอร์ที่ประกอบด้วย 3 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} จาก หัวข้อ 2.1.4 ฟีเจอร์กลุ่มแรกคือคำหรือสมาชิก (Element) ของเซต C ซึ่งเป็น เซตของแนวคิดของสาร เคมี กลุ่มที่ 2 คือคำหรือสมาชิก ของเซต S ซึ่งเป็นเซตของแนวคิดของพืชที่เป็น Natural Sourceในระดับดิวิชัน และกลุ่มที่ 3 คือ คำหรือสมาชิกต่างๆใน เซต A ซึ่งเป็นเซตของคำที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช โดยจะใช้คำทั้งหมดเหล่านี้ a_1, a_2,.., a_rที่มีความถี่มากที่สุด ภายใต้ขนาดกรอบหน้าต่าง r ของคำต่างๆ ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., r) ภายใต้ขนาดกรอบหน้าต่าง r ดังตารางที่ 3.2 ถึงตารางที่ 3.6 สำหรับ r = 5 ตารางที่ 3.7 ถึงตารางที่ 3.10 สำหรับ r = 4 ตารางที่ 3.11ถึงตารางที่ 3.13 สำหรับ r = 3

ตารางที่ 3.2 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00136519	0.00074627
isolated	0.51262799	0.08208955
extract	0.05119454	0.04104478
addition	0.00204778	0.00074627
obtained	0.00204778	0.00298507
…	…	…

ตารางที่ 3.3 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.11083123	0.00956938
roots	0.02518892	0.00683527
parts	0.02078086	0.01025290
leaves	0.01700252	0.00751880
seeds	0.00881612	0.00410116
…	…	…

ตารางที่ 3.4 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03947368	0.00452196
aerial	0.02033493	0.00968992
parts	0.01614833	0.00129199
leaves	0.01555024	0.00129199
roots	0.01435407	0.00129199
…	…	…

ตารางที่ 3.5 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01375358	0.00123381
parts	0.01088825	0.00061690
fruits	0.00916905	0.00123381
leaves	0.00744986	0.00061690
stems	0.00401146	0.00061690
…	…	…

ตารางที่ 3.6 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_5	Class = ‘YES’	Class = ‘NO’
parts	0.01528014	0.0006079
aerial	0.01075269	0.0006079
leaves	0.00962083	0.0006079
stems	0.00509338	0.0006079
bark	0.00509338	0.0006079
…	…	…

ตารางที่ 3.7 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
leaves	0.01012146	0.01920236
isolated	0.50067476	0.07754801
extract	0.04453441	0.03545052
investigated	0.00067476	0.00295421
roots	0.01079622	0.00590842
…	…	…

ตารางที่ 3.8 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.11407316	0.00810263
bark	0.00557967	0.00540176
parts	0.01673900	0.00810263
leaves	0.01487911	0.00607698
stems	0.00247985	0.00202566
…	…	…

ตารางที่ 3.9 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03256365	0.00448430
aerial	0.01539372	0.00768738
obtained	0.00296033	0.00064061
leaves	0.01480166	0.00128123
roots	0.01835406	0.00128123
…	…	…

ตารางที่ 3.10 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01534963	0.00122624
parts	0.01421262	0.00061312
roots	0.00397953	0.00674433
leaves	0.01193860	0.00061312
stems	0.00397953	0.00061312
…	…	…

ตารางที่ 3.11 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00133869	0.00074349
isolated	0.48728246	0.08104089
extract	0.04350736	0.04163569
addition	0.00200803	0.00074349
obtained	0.00200803	0.00371747
…	…	…

ตารางที่ 3.12 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.09588199	0.00949153
stems	0.00184388	0.00203390
pigment	0.00184388	0.00067797
leaves	0.01536570	0.00610169
fruit	0.00491703	0.00067797
…	…	…

ตารางที่ 3.13 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03627853	0.00448430
heartwood	0.00292569	0.00064061
addition	0.00175541	0.00064061
aerial	0.01872440	0.01089045
flowers	0.00526624	0.00064061
…	…	…

กรณีศึกษาที่ 2 ฟีเจอร์ที่ประกอบด้วย 3 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มแรกคือคำหรือสมาชิก (Element) ของเซต C กลุ่มที่ 2 คือคำหรือสมาชิก ของเซต S และกลุ่มที่ 3 คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้ทุกคำทั้งหมดที่อยู่ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i=1,2,..,nและnคือจำนวนคำที่อยู่ระหว่างคำที่เป็นแนวคิดของสารเคมีและคำที่เป็นแนวคิดของพืช) ดังตารางที่3.14ถึงตารางที่3.23สำหรับ n=10

ตารางที่ 3.14 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00128866	0.00079177
isolated	0.50193299	0.06650831
extract	0.04317010	0.04275534
addition	0.00193299	0.00079177
obtained	0.00193299	0.00395883
…	…	…

ตารางที่ 3.15 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10159480	0.00999286
obtained	0.00177200	0.00356888
heartwood	0.00590667	0.00071378
leaves	0.01476669	0.00499643
parts	0.02303603	0.01213419
…	…	…

ตารางที่ 3.16 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03943662	0.00470746
aerial	0.02140845	0.01143241
flowers	0.00507042	0.00067249
leaves	0.01464789	0.00067249
roots	0.01746479	0.00134499
…	…	…

ตารางที่ 3.17 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01302225	0.00128783
heartwood	0.00325556	0.00064392
fruits	0.00868150	0.00128783
leaves	0.01302225	0.00064392
identified	0.00217037	0.00064392
…	…	…

ตารางที่ 3.18 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_5	Class = ‘YES’	Class = ‘NO’
strain	0.00053591	0.00189873
aerial	0.01339764	0.00063291
growth	0.00160772	0.00126582
stems	0.00643087	0.00063291
colors	0.00053591	0.00126582
…	…	…

ตารางที่ 3.19 แสดงความน่าจะเป็นของคำที่ 6 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_6	Class = ‘YES’	Class = ‘NO’
inhibit	0.00163755	0.00064851
investigation	0.00054585	0.00129702
leaves	0.00545852	0.00064851
regions	0.00054585	0.00129702
stems	0.00109170	0.00064851
…	…	…

ตารางที่ 3.20 แสดงความน่าจะเป็นของคำที่ 7 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_7	Class = ‘YES’	Class = ‘NO’
potential	0.00167038	0.00066181
inhibit	0.00167038	0.00066181
agents	0.00055679	0.00529451
leaves	0.00111359	0.00066181
resulted	0.00111359	0.00132363
…	…	…

ตารางที่ 3.21 แสดงความน่าจะเป็นของคำที่ 8 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_8	Class = ‘YES’	Class = ‘NO’
potential	0.00172018	0.00068540
stems	0.00114679	0.00068540
natural	0.00057339	0.00274160
trunk	0.00057339	0.00137080
seeds	0.00516055	0.00068540
…	…	…

ตารางที่ 3.22 แสดงความน่าจะเป็นของคำที่ 9 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_9	Class = ‘YES’	Class = ‘NO’
bacterial	0.00058617	0.00140647
combined	0.00058617	0.00140647
fractionated	0.00058617	0.00210970
leaves	0.00234467	0.00070323
stems	0.00117233	0.00070323
…	…	…

ตารางที่ 3.23 แสดงความน่าจะเป็นของคำที่ 10 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_{10}	Class = ‘YES’	Class = ‘NO’
assay	0.00059809	0.00144404
stems	0.00299043	0.00072202
fractionated	0.00059809	0.00216606
seeds	0.00179426	0.00072202
inhibition	0.00059809	0.00216606
…	…	…

กรณีศึกษาที่ 3 ฟีเจอร์ที่ประกอบด้วย 1กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มนี้คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้คำทั้งหมดเหล่านี้ a_1, a_2, .., a_r ที่มีความถี่มากที่สุด ภายใต้ขนาดกรอบหน้าต่าง r ของคำต่างๆระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., r)ภายใต้ขนาดกรอบหน้าต่าง r ดังตารางที่ 3.24 ถึงตารางที่ 3.28 สำหรับ r = 5 ตารางที่ 3.29 ถึงตารางที่ 3.32 สำหรับ r = 4 ตารางที่ 3.33 ถึงตารางที่ 3.35 สำหรับ r = 3

ตารางที่ 3.24 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00136893	0.00073314
isolated	0.45995893	0.07331378
extract	0.05065024	0.04325513
addition	0.00205339	0.00073314
obtained	0.00068446	0.00366569
…	…	…

ตารางที่ 3.25 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.09699625	0.00801068
roots	0.02252816	0.00667557
parts	0.02440551	0.01134846
leaves	0.01564456	0.00734312
seeds	0.00750939	0.00333778
…	…	…

ตารางที่ 3.26 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.04045211	0.00441640
aerial	0.02201071	0.01072555
parts	0.01963117	0.00126183
leaves	0.01011303	0.00126183
roots	0.0184414	0.00126183
…	…	…

ตารางที่ 3.27 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01600000	0.00120919
parts	0.01428571	0.00060459
fruits	0.00914286	0.00120919
leaves	0.01371429	0.00060459
stems	0.00400000	0.00060459
…	…	…

ตารางที่ 3.28 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_5	Class = ‘YES’	Class = ‘NO’
parts	0.01634724	0.00059524
aerial	0.01409245	0.00059524
leaves	0.00958286	0.00059524
stems	0.00676437	0.00059524
bark	0.00507328	0.00059524
…	…	…

ตารางที่ 3.29 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
leaves	0.00986842	0.02148887
isolated	0.49736842	0.07828089
extract	0.05263158	0.02916347
investigated	0.00065789	0.00460476
roots	0.00921053	0.00613968
…	…	…

ตารางที่ 3.30 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10935601	0.00981767
bark	0.00546780	0.00280505
parts	0.02308627	0.00981767
leaves	0.01761847	0.00631136
stems	0.00303767	0.00210379
…	…	…

ตารางที่ 3.31 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.02537486	0.00329381
aerial	0.02133795	0.00922266
obtained	0.00288351	0.00329381
leaves	0.01384083	0.00131752
roots	0.01557093	0.00131752
…	…	…

ตารางที่ 3.32 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01551247	0.00062933
parts	0.01385042	0.00062933
roots	0.00387812	0.00692259
leaves	0.01218837	0.00062933
stems	0.00387812	0.00062933
…	…	…

ตารางที่ 3.33 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00135962	0.00074516
isolated	0.49558124	0.06333830
extract	0.05234534	0.04470939
addition	0.00203943	0.00074516
obtained	0.00203943	0.00372578
…	…	…

ตารางที่ 3.34 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10627719	0.00405954
stems	0.00310752	0.00202977
pigment	0.00186451	0.00067659
leaves	0.01429459	0.00744249
fruit	0.00435053	0.00067659
…	…	…

ตารางที่ 3.35 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03775811	0.00191449
heartwood	0.00294985	0.00063816
addition	0.00176991	0.00063816
aerial	0.02005900	0.01084876
flowers	0.00530973	0.00063816
…	…	…

กรณีศึกษาที่ 4 ฟีเจอร์ที่ประกอบด้วย 1 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มนี้คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้ทุกคำทั้งหมดที่อยู่ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., n และ n คือจำนวนคำที่อยู่ระหว่างคำที่เป็นแนวคิดของสารเคมี และคำที่เป็นแนวคิดของพืช) ดังตารางที่ 3.36 ถึงตารางที่ 3.45 สำหรับ n = 10

ตารางที่ 3.36 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00134771	0.00074906
isolated	0.51347709	0.08164794
extract	0.05256065	0.04344569
addition	0.00202156	0.00074906
obtained	0.00202156	0.00374532
…	…	…

ตารางที่ 3.37 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10864198	0.00885559
obtained	0.00123457	0.00885559
heartwood	0.00246914	0.00068120
leaves	0.01666667	0.00749319
parts	0.02160494	0.01158038
…	…	…

ตารางที่ 3.38 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.04110393	0.00450161
aerial	0.01996477	0.01093248
flowers	0.00528479	0.00064309
leaves	0.01409278	0.00128617
roots	0.01761597	0.00064309
…	…	…

ตารางที่ 3.39 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01351351	0.00122850
heartwood	0.00337838	0.00061425
fruits	0.00394144	0.00122850
leaves	0.01295045	0.00061425
identified	0.00225225	0.00061425
…	…	…

ตารางที่ 3.40 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_5	Class = ‘YES’	Class = ‘NO’
strain	0.00055772	0.00182149
aerial	0.00836587	0.00060716
growth	0.00167317	0.00121433
stems	0.00669269	0.00060716
colors	0.00055772	0.00121433
…	…	…

ตารางที่ 3.41 แสดงความน่าจะเป็นของคำที่ 6 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_6	Class = ‘YES’	Class = ‘NO’
inhibit	0.00171429	0.0018797
investigation	0.00057143	0.00125313
leaves	0.00514286	0.00062657
regions	0.00057143	0.00125313
stems	0.00285714	0.00062657
…	…	…

ตารางที่ 3.42 แสดงความน่าจะเป็นของคำที่ 7 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_7	Class = ‘YES’	Class = ‘NO’
potential	0.00175336	0.00063939
inhibit	0.00175336	0.00063939
agents	0.00058445	0.00511509
leaves	0.00116891	0.00063939
resulted	0.00058445	0.00255754
…	…	…

ตารางที่ 3.43 แสดงความน่าจะเป็นของคำที่ 8 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_8	Class = ‘YES’	Class = ‘NO’
potential	0.00180941	0.00066269
stems	0.00120627	0.00066269
natural	0.00060314	0.00265076
trunk	0.00060314	0.00132538
seeds	0.00542823	0.00066269
…	…	…

ตารางที่ 3.44 แสดงความน่าจะเป็นของคำที่ 9 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_9	Class = ‘YES’	Class = ‘NO’
bacterial	0.00061690	0.00135777
combined	0.00061690	0.00135777
fractionated	0.00061690	0.00203666
leaves	0.00246761	0.00067889
stems	0.00123381	0.00067889
…	…	…

ตารางที่ 3.45 แสดงความน่าจะเป็นของคำที่ 10 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_{10}	Class = ‘YES’	Class = ‘NO’
assay	0.00063052	0.00139373
stems	0.00315259	0.00069686
fractionated	0.00063052	0.00209059
seeds	0.00189155	0.00069686
inhibition	0.00063052	0.00209059
…	…	…

4. การสกัดความสัมพันธ์แบบสตัฟฟ์ (Stuff Relation Extraction)

ขั้นตอนการสกัดความสัมพันธ์แบบ สตัฟฟ์ เป็นการค้นหาและสกัดประโยคที่มีความสัมพันธ์แบบ สตัฟฟ์จากคลังข้อมูลสำหรับใช้ทดสอบ ด้วยสมการ (4) ร่วมกับค่าความน่าจะเป็นของฟีเจอร์ต่างๆตามกรณีศึกษา1-4แสดงในอัลกอริทึมการสกัดความสัมพันธ์แบบสตัฟฟ์ (ภาพที่ 3.4) และอัลกอริทึม การสกัดฟีเจอร์จากประโยค (ภาพที่ 3.5) โดย R คือความสัมพันธ์แบบสตัฟฟ์

L is a list of sentence.

i is a index of sentence list.

W is a set of word in each sentence.

C is natural-product-compounds concept set.

S is the natural-sources concept set.

A is a set of the high frequency words existing between $\langle c_1, c_2, .., c_i \}$ and $\langle s_1, s_2, .., s_j \rangle$

cf is a array of natural-product-compound occurrence features from $\langle c_1, c_2, .., c_i \rangle$

sf is a array of natural-source occurrence features from $\langle s_1, s_2, .., s_j \}$

af is a array of the high-frequency-word occurrence features existing between $\langle c_1, c_2, .., c_i \rangle$ and $\langle s_1, s_2, .., s_j \rangle$ varying to window sizes r (r = 3, 4, 5) and n (where n is the total number of words existing between the natural-product-compound concept words and the natural-source concept words)

STUFF_RELATION_EXTRACTION

1. { i ← 0, R ← ϕ
2. Array [max1] cf, Array [max2] sf, Array [max3] af
3. initialize each element of cf[], sf[], and af[] with " "
4. while (i ≤ length[L]) do
5. { ExtFea(cf[], sf[], af[])
6. Case 1 where max3 = r
7. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max1} P(c_{num}) \prod_{num=1}^{max2} P(s_{num}) \prod_{num=1}^{max3} P(a_{num})$
8. Case 2 where max3 = n
9. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max1} P(c_{num}) \prod_{num=1}^{max2} P(s_{num}) \prod_{num=1}^{max3} P(a_{num})$
10. Case 3 where max3 = r
11. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max3} P(a_{num})$
12. Case 4 where max3 = n
13. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max3} P(a_{num})$
14. if (StuffRelationForNaturalProduct == "yes") then
15. R = R ∪ {i} ;
16. }
17. return R
18. }

ภาพที่ 3.4 แสดงอัลกอริทึมของการสกัดความสัมพันธ์แบบสตัฟฟ์

```
ExtFeat(Array [max1] cf,  Array [max2] sf, Array [max3] af)
1.    { j ⟵ 0; k ⟵ 0
2.    initialize each element of cf[], sf[], and af[] with " "
3.      while(FindPositionOfS(w_j)) do
4.            j++;
5.      sf[]=FindElementsOfSfromPlantSourceBase
6.      while(FindPositionOfC (w_k)) do
7.            k++;
8.      cf[]=FindElementOfCfromChemNet
9.      Array [max3]  temp
10.     if( j > k ) then
11.       for(m ⟵ k to j)
12.           temp[m] = w_m
13.       Next m
14.     else if( k > j ) then
15.           for(m ⟵ j to k)
16.               temp[m] = w_m
17.           Next m
18.     Array featureA[] = SortingWord(temp[])
19.     af[]= featureA[]
20.   }
```

ภาพที่ 3.5 แสดงอัลกอริทึมของการสกัดฟีเจอร์สำหรับความสัมพันธ์แบบสตัฟฟ์

การวัด การวัดประสิทธิภาพของระบบ การสกัดความสัมพันธ์แบบสตัฟฟ์ จะอ้างอิงความถูกต้องจากเอกสารที่ผ่านการกำ กับจากผู้เชี่ยวชาญ ซึ่งการวัดประสิทธิภาพ จะวัดโดยใช้ค่าความถูกต้อง (precision) ค่าความระลึก (recall) และค่า F-measure ซึ่งทั้ง 3 ค่า สามารถคำ นวณได้ดังนี้

$$\text{Precision(P)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (7)$$

$$\text{Recall(R)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNagative}} \quad (8)$$

$$\text{F} - \text{measure} = \frac{(\beta^2+1)\text{PR}}{\beta^2\text{P}+\text{R}} \quad (9)$$

โดย βคือค่าพารามิเตอร์ที่แสดงสัดส่วนความสำคัญระหว่างค่าความถูกต้องและค่าความระลึก โดยทั่วไป βจะมีค่าเท่ากับ 1

3.2 เครื่องมือที่ใช้

ในการวิจัยครั้งนี้ใช้เครื่องมือที่ประกอบไปด้วยส่วนของฮาร์ดแวร์ (Hardware) และซอฟต์แวร์ (Software) สรุปได้ดังต่อไปนี้

3.2.1ฮาร์ดแวร์(Hardware)

1.เครื่องแมคบุ๊กโปร(Macbook Pro) โดยมีรายละเอียดดังนี้

Model: Early 2011

CPU: 2.3GHz (2410M) Intel Core i5 พร้อม 3MB on-chip L3cache

RAM: 4GB 1333 MHz DDR3

VGA: Intel HD Graphics 3000 พร้อม 384 MB DDR3 SDRAM shared พร้อม main memory

Hard disk: 320 GB

Monitor: LED Display 13.3" @Ressolution 1,280 × 800

2.เครื่องคอมพิวเตอร์ตั้งโต๊ะ (Desktop Computer) โดยมีรายละเอียดดังนี้

CPU:3.40 GHz Intel Core i7-2600

RAM: 4GB

VGA: AMD Radeon HD 6450

Hard disk: 1TB

Monitor: Lenovo LED Display 21" @Ressolution 1,600 × 900

3.เครื่องบริการ (Server Computer) โดยมีรายละเอียดดังนี้

CPU:2.66GHzIntel Core i7-920พร้อม 8MB on-chip L3cache

RAM: 12GB 1333 MHz DDR3

VGA: AMD Radeon HD 6450

Hard disk: 2x500 GB

3.3.2ซอฟต์แวร์(Software)

1.ซอฟต์แวร์ระบบปฏิบัติการ (Operating System Software)โดยมีรายละเอียดดังนี้

Microsoft® Windows® 7 Enterprise

Mac OS X Lion 10.7.5

Linux: CentOS

2.ซอฟต์แวร์ปฏิบัติการประยุกต์ (Application Software) โดยมีรายละเอียดดังนี้

Adobe®DreamWeaver® CS5

Apache Web Server

PHP: Hypertext Preprocessor

MySQL RDBMS

Eclipse Classic (INDIGO)

บทที่ 4
ผลการดำเนินงานวิจัย

จากการดำเนินการ วิจัยการสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์ ได้ผลดังต่อไปนี้

หลังจากการศึกษาพฤติกรรมของภาษาพบว่าค่าทางสถิติของคำที่อยู่ระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชดังตารางต่อนี้

ตารางที่ 4.1 แสดงค่าทางสถิติของคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช

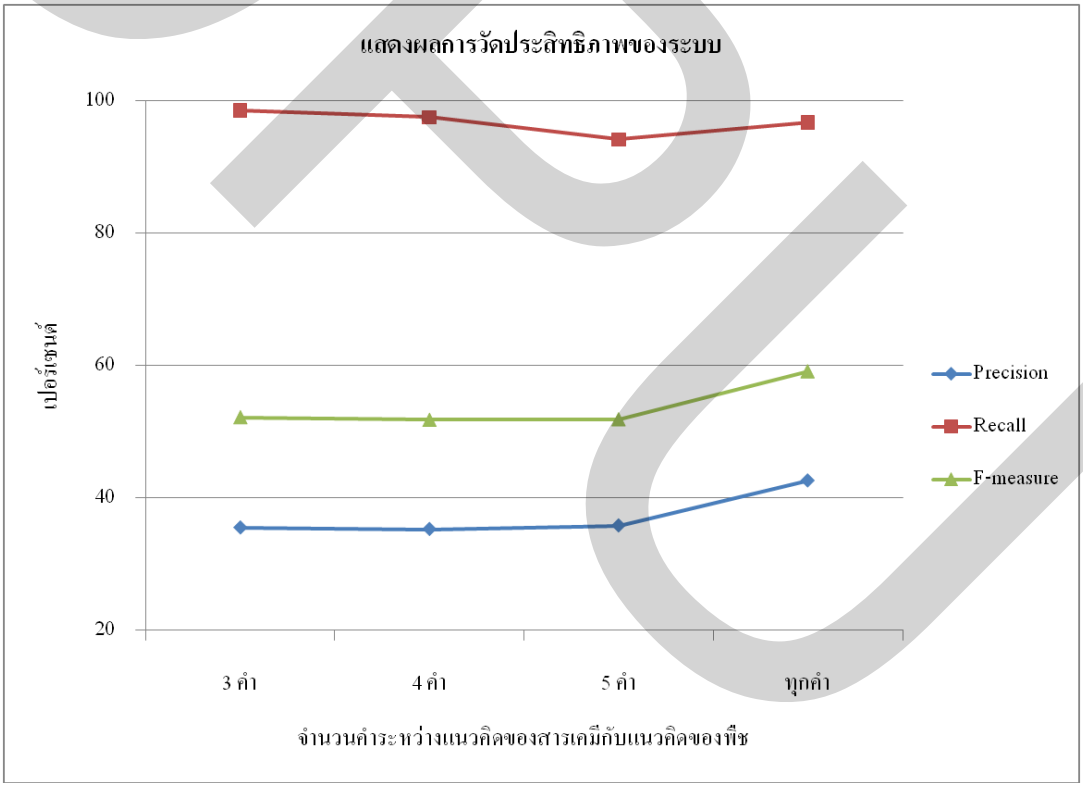
ค่าทางสถิติของคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช	จำนวน
ค่ามากที่สุด	10
ค่าน้อยที่สุด	1
ค่ามัธยฐาน	4
ค่าฐานนิยม	4
ค่าเฉลี่ยเลขคณิต	5
ค่าเบี่ยงเบนมาตรฐาน	2.55

จากตารางที่ 4.1 จึงใช้จำนวนคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช เป็น 4 ตามค่าฐานนิยมแต่ในการทดลองจะใช้ ขนาดกรอบหน้าต่างของคำระหว่างแนวคิดสารเคมีกับแนวคิดของพืช (r) เป็น 3, 4 และ 5 คำ

ในการวัดประสิทธิภาพของระบบ การสกัดความสัมพันธ์แบบสตัฟฟ์ โดยใช้ประโยคทั้งหมด 20,000 ประโยค ในการทดสอบซึ่งจะทำการทดลองโดยใช้ 10-folds cross-validation แล้ววัดค่าความถูกต้อง (precision) ค่าความระลึก (recall) และค่า F-measure ดังสมการ(7), สมการ(8) และ สมการ(9) ตามลำดับ ซึ่งได้ผลการทดลองดังนี้

ตารางที่ 4.2 แสดงผลการทดสอบการสกัดความสัมพันธ์แบบสตัฟฟ์โดยใช้ฟีเจอร์เป็น แนวคิดของสารเคมี แนวคิดของพืช และคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชที่เป็น 3 คำ 4 คำ 5 คำ และทุกคำ

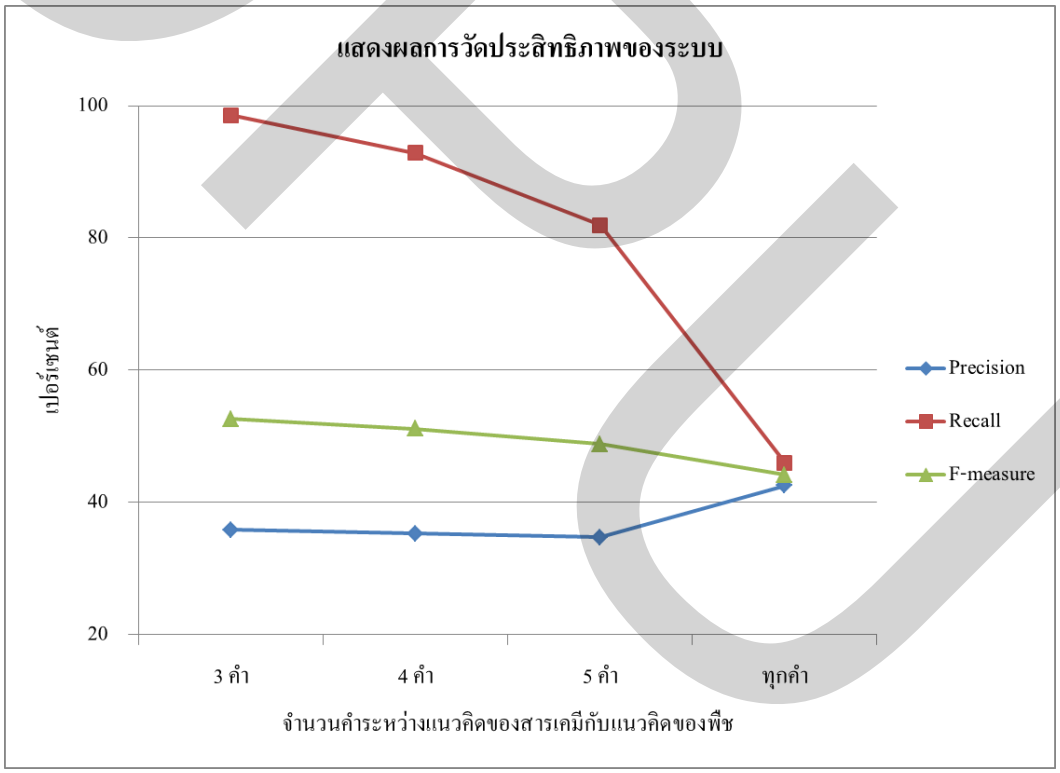
คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช	จำนวนความสัมพันธ์แบบสตัฟฟ์	ค่าความถูกต้อง (%)	ค่าความระลึก (%)	F-measure
3 คำ	23	35.51	98.5	52.20
4 คำ	23	35.28	97.5	51.81
5 คำ	22	35.82	94.14	51.89
ทุกคำ	18	42.62	96.71	59.17



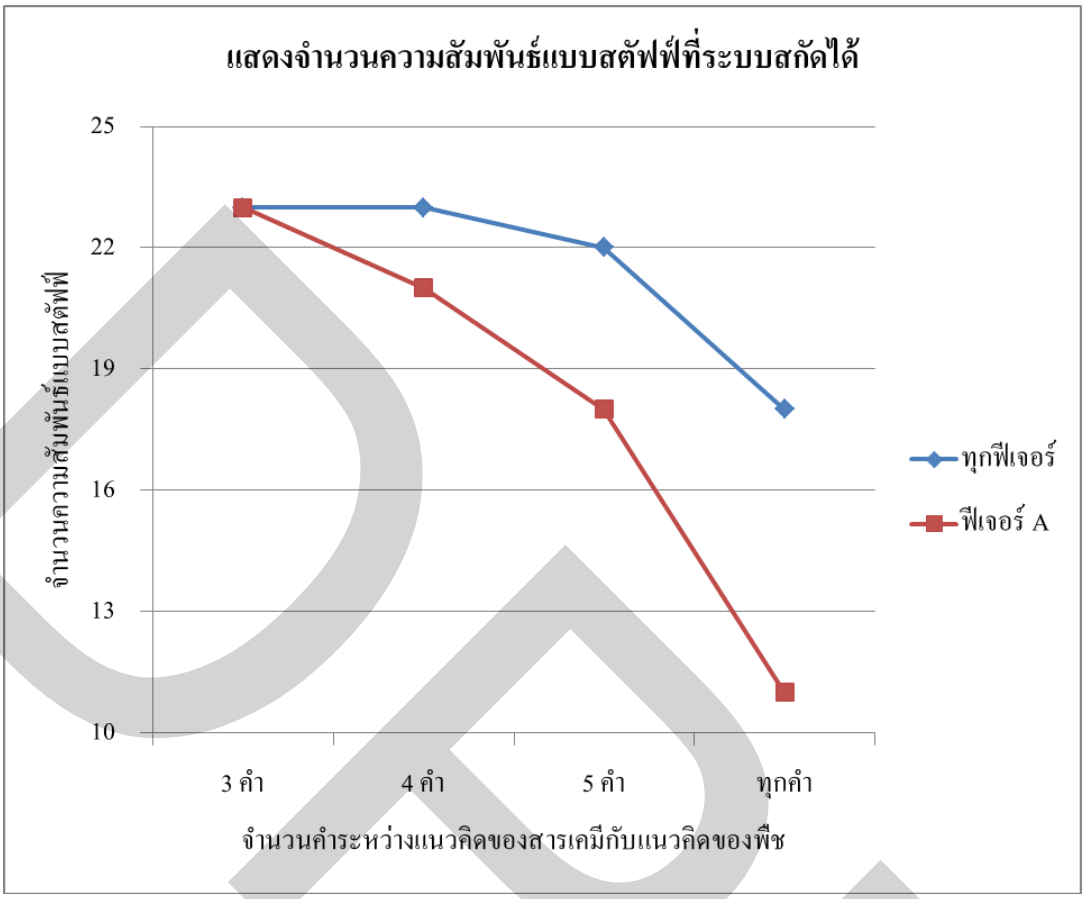
ภาพที่ 4.1 แสดงกราฟความสัมพันธ์ระหว่างประสิทธิภาพของระบบการสกัดความสัมพันธ์แบบสตัฟฟ์โดยใช้ฟีเจอร์เป็น แนวคิดของสารเคมี แนวคิดของพืช และคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชที่เป็น 3 คำ 4 คำ 5 คำ และทุกคำ

ตารางที่ 4.3 แสดงผลการทดสอบการสกัดความสัมพันธ์แบบสตัฟฟ์โดยใช้ฟีเจอร์เป็นคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชที่เป็น 3 คำ 4 คำ 5 คำ และทุกคำ

คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช	จำนวนความสัมพันธ์แบบสตัฟฟ์	ค่าความถูกต้อง (%)	ค่าความระลึก (%)	F-measure
3 คำ	23	35.87	98.57	52.60
4 คำ	21	35.26	92.86	51.11
5 คำ	18	34.73	81.95	48.78
ทุกคำ	11	42.51	45.97	44.17



ภาพที่ 4.2 แสดงกราฟความสัมพันธ์ระหว่างประสิทธิภาพของระบบการสกัดความสัมพันธ์แบบสตัฟฟ์โดยใช้ฟีเจอร์เป็น คำระหว่าง แนวคิดของสารเคมีกับแนวคิดของพืช ที่เป็น 3 คำ 4 คำ 5 คำ และทุกคำ



ภาพที่ 4.3 แสดงกราฟเปรียบเทียบ จำนวนความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้โดยระบบกับ ขนาดกรอบหน้าต่างของ คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช ที่เป็น 3 คำ 4 คำ 5 คำ และทุกคำ

ในการวัดผลระบบการสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์ แสดงในตารางที่ 4.2 กับตารางที่ 4.3 และภาพที่ 4.1 กับภาพที่ 4.2 โดยใช้ค่าความถูกต้อง (precision) เมื่อใช้ทุกฟีเจอร์และใช้เฉพาะฟีเจอร์ A พบว่าเมื่อเพิ่มจำนวน คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชค่าความถูกต้องจะเพิ่มขึ้น แต่ก็อยู่ในระดับค่าที่ไม่สูงมากนัก สำหรับค่าความระลึก (recall) เมื่อใช้เฉพาะฟีเจอร์ A พบว่าเมื่อเพิ่มจำนวน คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช ค่าความระลึกจะลดลงอย่างมากในขณะที่ใช้ทุกฟีเจอร์ค่าความระลึกจะไม่ค่อยเปลี่ยนแปลงเมื่อเพิ่มหรือลดจำนวนคำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช สุดท้ายเมื่อวัดประสิทธิภาพของระบบด้วย F-measure พบว่าเมื่อใช้ทุกฟีเจอร์ ค่า F-measure จะมีแนวโน้มเพิ่มขึ้นตามจำนวน คำระหว่าง แนวคิดของสารเคมีกับแนวคิดของพืช แต่เมื่อใช้เฉพาะฟีเจอร์ A ค่า F-

measure จะมีแนวโน้มลดลงสวนทางกับการเพิ่มของจำนวน คำระหว่าง แนวคิดของสารเคมีกับแนวคิดของพืช

เมื่อเปรียบเทียบระหว่างจำนวนความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้กับฟี เจอร์ที่ใช้ (ใช้ทุกฟีเจอร์และใช้เฉพาะฟีเจอร์ A) พบว่าเมื่อเพิ่มจำนวน คำระหว่างแนวคิดของสารเคมีกับแนวคิดของพืชแนวโน้มของความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้ก็จะลดลง แต่ใช้ทุกฟีเจอร์จะลดลงไม่มาก นักจาก 23 ความสัมพันธ์แบบสตัฟฟ์เหลือ 18 ความสัมพันธ์แบบสตัฟฟ์ถ้าใช้เฉพาะฟีเจอร์ A จะลดลงอย่างมากจาก 23 ความสัมพันธ์แบบสตัฟฟ์เหลือ 11 ความสัมพันธ์แบบสตัฟฟ์ ดังภาพที่ 4.3

บทที่ 5
สรุปอภิปรายผลการศึกษาและข้อเสนอแนะ

ในบทนี้จะกล่าวถึง ข้อสรุปจากการดำเนินงานวิจัยปัญหาและอุปสรรคระหว่างการ วิจัย รวมทั้งข้อเสนอแนะ โดยมีรายละเอียดดังต่อไปนี้

5.1 สรุปผลการดำเนินงานวิจัย

ในการดำเนินงานวิจัยการสกัดความสัมพันธ์แบบสตัฟฟ์จากเอกสารงานวิจัยทางวิทยาศาสตร์ในวิทยานิพนธ์นี้เป็นการใช้หลักสถิติมาประยุกต์ใช้ในการสกัดความสัมพันธ์แบบสตัฟฟ์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ที่เป็นเนอีฟเบย์ (Naïve Bayes) ด้วยการเปรียบเทียบคำต่างๆ ในเอกสารงานวิจัยทางวิทยาศาสตร์กับฐานข้อมูลความรู้เฉพาะทาง ด้านวิทยาศาสตร์เคมีและชีววิทยา ซึ่งผลลัพธ์ของการสกัดความสัมพันธ์ที่ได้ขึ้นกับการเตรียมคลังข้อมูล (Corpus) ที่เหมาะสม ถูกต้องและ รวมไปถึงการเลือกคุณสมบัติ ต่างๆที่ใช้พิจารณาให้เหมาะสม จะส่งผลให้ได้ผลลัพธ์จากระบบได้อย่างถูกต้องแล้วทำการประเมินผลเปรียบเทียบกับการประเมินโดยใช้ผู้เชี่ยวชาญเฉพาะด้านซึ่งปรากฏว่าผลลัพธ์ในส่วนของค่าความถูกต้อง (precision)ยังไม่ค่อยดีนักนัก เนื่องจากงานวิจัยนี้ศึกษาบนแพทเทิร์นพื้นฐานแบบง่ายๆทางภาษาแบบเดียวคือ NP1 verb NP2 | NP2 verb NP1(เมื่อ NP1 เป็นนามวลีที่เป็น Chemical-Name Entity และ NP2 เป็นนามวลีที่เป็น Natural-Source-Name Entity) ดังนั้นในการสร้างคลังข้อมูลเพื่อฝึกฝนระบบจะแบ่ง คุณสมบัติต่างๆ เป็น3กลุ่มคือกลุ่มที่เป็นแนวคิดของสารเคมี กลุ่มที่เป็นแนวคิดของพืช และกลุ่มที่เป็นคำระหว่างแนวคิดของสารเคมีและแนวคิดของพืช โดยไม่ได้สนใจกลุ่มคำที่อยู่ด้านนอกแนวคิดของสารเคมีและแนวคิดของพืช กลุ่มคำที่อยู่ระหว่างแนวคิดของสารเคมีด้วยกัน และกลุ่มคำที่อยู่ระหว่างแนวคิดของพืชด้วยกัน

หากนำงานวิจัยนี้มา พัฒนาเพิ่มเติมจะสามารถนำผลลัพธ์ที่ได้ไปสร้างออนโทโลยี ของสาร ผลิตภัณฑ์จากธรรมชาติ ซึ่งจะเป็นประโยชน์ต่ออุตสาหกรรมการสกัดสารเคมีจากพืช อาทิ อุตสาหกรรมยา อุตสาหกรรมผลิตภัณฑ์เสริมอาหาร เป็นต้น

5.2 ปัญหาและอุปสรรคจากการดำเนินงานวิจัย

ในการเตรียมคลังข้อมูลเพื่อฝึกฝนระบบ เนื่องจากจึงจำเป็นต้องแปลงข้อมูลในรูปแบบ pdfให้อยู่ในรูปข้อความก่อน ซึ่งจะมีปัญหาเรื่องสัญลักษณ์บางตัวที่ผิดเพี้ยนไป จำเป็นต้องใช้เวลาและแรงงานในการสร้างคลังข้อมูลมาก

5.3 ข้อเสนอแนะ

1. ระบบที่พัฒนาขึ้นนี้ เป็นการใช้เพียงรูปแบบโครงสร้างทางไวยากรณ์เดียว(NP1 verb NP2) ในการตัดสินใจเท่านั้น หากมีการศึกษาใช้รูปแบบโครงสร้างทางไวยากรณ์ที่แตกต่างออกไปมาช่วยในการพิจารณาข้อมูลอาจทำให้ได้ประสิทธิภาพของระบบดีขึ้น
2. ในการสร้างคลังข้อมูลเพื่อฝึกฝนระบบควรนำคุณสมบัติอื่นมาพิจารณาร่วมด้วย เช่นกลุ่มคำที่อยู่ด้านนอกแนวคิดของสารเคมีและแนวคิดของพืช กลุ่มคำที่อยู่ระหว่างแนวคิดของสารเคมีด้วยกัน และกลุ่มคำที่อยู่ระหว่างแนวคิดของพืชด้วยกัน อาจช่วยเพิ่มประสิทธิภาพของระบบได้
3. ควรเพิ่มขนาดของคลังข้อมูลให้มากขึ้น

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

บุษกร สุคนธวงศาโรจน์. (2008). *อัลกอริทึมในการจำแนกบุคลากรในองค์กรสำหรับการสร้างแผนที่ความรู้*. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์.

ภาษาต่างประเทศ

Katrin Fundel. (2007). *Text Mining and Gene Expression Analysis Towards Combined Interpretation of High Throughput Data*. München.

Tom Mitchell. (1997). *Machine* Learning The McGraw Hill Companies Inc. and MIT Press, Singapore.

Cruse, Alan. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford.

Morton E. Winston, Roger Chaffin, Douglas Herrmann. (1987). *A taxonomy of part-whole relations*. vol. 11, pp. 417-444. Cognitive Science.

Roxana Girju, Adriana Badulescu, Dan Moldovan. (2003). *Learning semantic constraints for the automatic discovery of part-whole relations*. vol. 1, pp. 1-8. *In* Proceedings of HLT/NAACL-03.

Patrick Pantel, Marco Pennacchiotti. (2006). *Espresso: Leveraging generic patterns for automatically harvesting semantic relations*. pp. 113-120. *In* Proceedings of COLING/ACL-06.

Gregory Ichneumon Brown. (2011). *An Error Analysis of Relation Extraction in Social Media Documents*. pp. 64-68. *In* Proceedings of ACL-HLT-49.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer. 2009. *The WEKA Data Mining Software: An Update.* vol. 11, pp. 10-18. *In* SIGKDD Explorations.

Alistair Kennedy, Graeme Hirst. (2012). *Measuring Semantic Relatedness Across Languages*. Computational Linguistics Seminar. Canada.

ภาคผนวก

ภาคผนวก ก
ตัวอย่างการเตรียมคลังข้อมูล

1384 *J. Nat. Prod.* **2003,** *66,* 1384−1387

New Antioxidant *C*-Glucosylxanthones from the Stems of *Arrabidaea samydoides*

Patrícia Mendonça Pauletti,[†] Ian Castro-Gamboa,[†] Dulce Helena Siqueira Silva,[†] Maria Claudia Marx Young,[‡] Daniela Maria Tomazela,[§] Marcos Nogueira Eberlin,[§] and Vanderlan da Silva Bolzani*[,†]

NuBBE- Núcleo de Biossíntese, Bioensaios e Ecofisiologia de Produtos Naturais, Instituto de Química, Universidade Estadual Paulista, UNESP, CP 355, 14801-970, Araraquara, SP, Brazil, Secção de Fisiologia e Bioquímica de Plantas, Instituto de Botânica, CP 4009, 01061-970, São Paulo, Brazil, and Thompson Mass Spectroscopy Laboratory, Instituto de Química, Universidade Estadual de Campinas, UNICAMP, CP 6154, 13083-970, Campinas, SP, Brazil

Received March 7, 2003

Three new *C*-glucosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**2**), and 2-(2′-*O*-*trans*-coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**3**), were isolated from the stems of *Arrabidaea samydoides*, in addition to three known *C*-glucosylxanthones, mangiferin (**4**), 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**5**), and muraxanthone (**6**). Their chemical structures were assigned on the basis of MS and 1D and 2D NMR experiments. Xanthones **1**−**6** showed moderate free radical scavenging activity against 1,1-diphenyl-2-picrylhydrazyl (DPPH) as well as antioxidant activity evidenced by redox properties measured on ElCD-HPLC.

As part of our bioprospecting program Biota-FAPESP (The Virtual Institute of Biodiversity), whose main goal is to discover potential antitumoral, antifungal, and antioxidant agents from plants of the Cerrado and Atlantic forest, we have screened hundreds of plants collected in the State of São Paulo. Among these, *Arrabidaea samydoides* was chosen for detailed chemical investigation due to prior antioxidant activity revealed on a TLC autographic assay sprayed with *β*-carotene solution, and to our knowledge there are no previous reports on chemical and biological studies. This species belongs to the family Bignoniaceae, which contains about 120 genera and 800 species distributed throughout tropical regions of South America and Africa.[1] Species from the genus *Arrabidaea* have been used in traditional medicine for wound asepsis and treating intestinal disorders.[2] In northeast Brazil, *Arrabidaea chica* is used in tattoos by Indians due to the pigments carajurin and carajurone.[2,3] A literature review indicated that this genus is a source of anthocyanins, flavonoids, and tannins.[3−7] The ethanolic extract from the stems showed promising antioxidant activity and led to the isolation of three new *C*-glucopyranosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**2**), and 2-(2′-*O*-*trans*-coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**3**), and the known mangiferin (**4**),[8] 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**5**),[9] and muraxanthone (**6**).[10] In this paper, we report the isolation, structure elucidation, and antioxidant properties of these *C*-glucopyranosylxanthones.

Compounds **4** and **6** were identified by comparison with previously published NMR and other physical data.[8,10] Compound **5** was described previously as a mixture of three isomers from *Hymenophyllum recurvum*.[9] Only a few ^{13}C NMR data were analyzed and discussed. In this paper we describe the complete ^{1}H, ^{13}C NMR and ES-MS/MS data for this compound.

Compound **1** was shown to have the molecular formula $C_{28}H_{29}O_{14}$ $[M - H]^-$ *m/z* 583.1008, by analysis of the negative-ion HRESIMS. The IR spectrum showed bands at 3370, 1615, and 1474 cm^{-1} accounting for hydroxyl, conjugated carbonyl, and aromatic groups, respectively. The ^{13}C NMR spectrum showed six signals for hydroxymethine carbons, suggesting the presence of a sugar moiety, and 22 signals for sp^2 carbons, which could be assigned to three aromatic rings, and also two carbonyls and one additional olefinic function. In the ^{1}H NMR spectrum (Table 1) of **1**, a caffeoyl moiety was identified by signals at *δ* 6.79 (1H, d, *J* = 2.0 Hz, H-2″), 6.58 (1H, d, *J* = 8.0 Hz, H-5″), and 6.67 (1H, br d, *J* = 8.0 Hz, H-6″),

* Author to whom correspondence should be addressed. Tel: 55(16)-2016660. Fax: 55(16)2227932. E-mail: bolzaniv@iq.unesp.br.
† Instituto de Química, Universidade Estadual Paulista−UNESP.
‡ Seção de Fisiologia e Bioquímica de Plantas, Instituto de Botânica.
§ Instituto de Química, Universidade Estadual de Campinas−UNICAMP.

10.1021/np030100t CCC: $25.00
Published on Web 09/24/2003

ตัวอย่างข้อมูลที่ทำการ annotation และนำเอา stop word ออก

<stuff_relation class="yes">Three new *C*-glucosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone(**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone isolated stems *Arrabidaea samydoides C*-glucosylxanthones, mangiferin 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone muraxanthone</stuff_relation>

<stuff_relation class="no">chemical structures assigned MS 1D 2D NMR experiments</stuff_relation>

<stuff_relation class=" no">Xanthones **1**-**6** showed moderate radical scavenging 1,1-diphenyl-2-picrylhydrazyl (DPPH) antioxidant evidenced redox properties measured ElCD-HPLC</stuff_relation>

<stuff_relation class=" no"> bioprospecting program Biota-FAPESP Virtual Institute Biodiversity goal discover potential antitumoral antifungal antioxidant agents Cerrado Atlantic forest screened hundreds collected State Sa˜o Paulo</stuff_relation>

<stuff_relation class=" no"> *Arrabidaea samydoides* chosen detailed chemical investigation prior antioxidant revealed TLC autographic assay sprayed *â*-carotene solution knowledge reports chemical biological studies </stuff_relation>

<stuff_relation class=" no"> species belongs family Bignoniaceae contains 120 genera 800 species distributed tropical regions South America Africa</stuff_relation>

<stuff_relation class=" no">Species genus *Arrabidaea* traditional medicine wound asepsis treating intestinal disorders</stuff_relation>

<stuff_relation class=" no"> northeast Brazil *Arrabidaea chica* tattoos Indians pigments carajurin carajurone </stuff_relation>

<stuff_relation class=" no"> literature review indicated genus source anthocyanins flavonoids tannins. </stuff_relation>

<stuff_relation class=" no"> ethanolic extract stems showed promising antioxidant isolation *C*-glucopyranosylxanthones 2-(2′-*O*-*trans*-caffeoyl)-*C*-*â*-Dglucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*â*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*-coumaroyl)-*C*-*â*-Dglucopyranosyl-1,3,6,7-tetrahydroxyxanthone mangiferin 2-(2′-*O*-benzoyl)-*C*-*â*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone muraxanthone</stuff_relation>

<stuff_relation class=" no"> paper report isolation structure elucidation antioxidant properties *C*-glucopyranosylxanthones</stuff_relation>

<stuff_relation class=" no">Compounds **4 6** identified comparison published NMR physical data</stuff_relation>

<stuff_relation class=" no">Compound **5** described mixture isomers *Hymenophyllum recurvum*</stuff_relation>

<stuff_relation class=" no">^{13}C NMR data analyzed discussed</stuff_relation>

<stuff_relation class=" no"> paper describe complete ^{1}H, ^{13}C NMR ES-MS/MS data compound</stuff_relation>

<stuff_relation class=" no">Compound **1** shown molecular formula $C_{28}H_{23}O_{14}$ [M - H]$^-$ *m*/*z* 583.1008 analysis negative-ion HRESIMS</stuff_relation>

<stuff_relation class=" no">IR spectrum showed bands 3370 1615 1474 cm^{-1} accounting hydroxyl conjugated carbonyl aromatic groups</stuff_relation>

<stuff_relation class=" no">^{13}C NMR spectrum showed signals hydroxymethine carbons suggesting presence sugar moiety 22 signals sp^2 carbons assigned aromatic rings carbonyls additional olefinic function </stuff_relation>

ภาคผนวก ข
ตัวอย่างประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้

id	ตัวอย่างประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้
88	*C*-glucosylxanthones, 2-(2′-O-trans-caffeoyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-O-trans-cinnamoyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-O-trans-coumaroyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone isolated stems Arrabidaea samydoides
217	5-epi-vibsanin G 18-O-methylvibsanin G vibsanin M aldovibsanin C isolated acetone extract leaves flowers Viburnum odoratissimum
318	(8R*)-8-bromo-10-epi-β-snyderol (8S*)-8-bromo-β-snyderol 5-bromo-3-(3′-hydroxy-3′-methylpent-4′-enylidene)-2,4,4-trimethylcyclohexanone isolated chloroform-methanol extract Laurencia obtuse
811	jujubogenin 3-O-R-L-arabinofuranosyl(1f2)-[β-D-glucopyra-nosyl(1f6) *O* -D-glucopyranosyl(1f3)]-R-L-arabinopyranoside jujubogenin 3-O-R-L-arabinofuranosyl-(1f2)-{6-O-[3-hydroxy-3-methylglutaryl]- *O* -D-glucopyranosyl(1f3)}-R-L-arabinopyranoside *O*-hydroxylup-20(29)-en-27,28-dioic acid 28-O-β-D-glucopyranosyl(1f2)-[*O*-D-xylopyra-nosyl(1f3)]- *O*-D-xylopyranosyl(1f2)-β-D-glucopyranoside ester jujubogenin 3-O-R-L-arabinofuranosyl(1f2)-[β-D-glucopyranosyl(1f3)]-R-L-arabinopyranoside *O*-hydroxylup-20(29)-ene-27,28-dioic acid isolated methanol extract stems Anomospermum grandifolium.
878	5R,9R,10β,13R-tetraacetoxy-14β-O-(β-D-glucopyranosyl)taxa-4(20),11-diene 1β,2R,9R,10β-tetrahydroxy-5R-cinnamoyoxytaxa-4(20),11-dien-13-one (2), 2R,9R,10β-trihydroxy-5R-cin-namoyoxytaxa-4(20),11-dien-13-one 9R-acetoxy-2R,10β-dihydroxy-5R-cinnamoyoxytaxa-4(20),11-dien-13-one 2R,10β-diacetoxy-1β,9R-dihydroxy-5R-cinnamoyoxy-3,11-cyclotaxa-4(20)-dien-13-one identified Taxus baccata
1188	wilfordine alatamine wilforidine alatusinine euonine euonymine ebenifoline forrestine mayteine 4-hydroxy-7-epi-chuchuhuanine isolated leaves Maytenus chiapensis.
1364	Asparagus cochichinensis led isolation asparacoside asparacosins A 3″-methoxyasparenydiol 3′-hydroxy-4′-methoxy-4′-dehydroxynyasol asparenydiol nyasol 3″-methoxynyasol 1,3-bis-di-p-hydroxyphenyl-4-penten-1-one trans-coniferyl alcohol

1734	C-glycosides 2″-O-(2‴-methylbutyryl)isoswertisin 3″-O-(2‴-methylbutyryl)-isoswertisin 2″-O-(2‴-methylbutyryl)vitexin 2″-O-(2‴-methylbutyryl)orientin 2″-O-(3‴,4‴-dimethoxybenzoyl)vitexin 2″-O-(3‴,4‴-dimethoxybenzoyl)orientin isoswertisin isolated flowers Trollius ledebouri.
2163	2,3,6,8-tetrahydroxy-1-(3-methylbut-2-enyl)-5-(2-methylbut-3-en-2-yl)-9H-xanthen-9-one isolated root bark Cudrania
2205	5R,7R,10βH-3-patchoulen-2-one 5R,7R10βH-4(14)-patchoulen-2R-ol 9R,10β-dihydroxy-2β,4β-peroxy-1R,5β,7RH-guaiane isolated aerial parts Croton arboreous
2762	10,11-dimethoxynareline alstohentine alstomicine 16-hydroxyalstonisine 16-hydroxyalstonal 16-hydroxy-N(4)-demethylalstophyllal oxindole alstophyllal 6-oxoalstophylline 6-oxoalstophyllal ones obtained leaf extract Malayan Alstonia macrophylla.
2772	14β-benzoyloxybaccatin IV 14β-benzoyloxy-13-deacetylbaccatin IV 14β-benzoyloxy-2-deacetylbaccatin VI isolated leaves stems Taxus chinensis.
3846	4β-hydroxy-19-normanoyl oxide 4R-hydroxy-18-normanoyl oxide 18-O-R-L-arabinopyranosylmanoyl oxide jhanol 18-hydroxy-13-epi-manoyl oxide isolated constituents Argentine collection Grindelia scorzonerifolia

ภาคผนวก ค

บทความการประชุม 7th International Conference on Computer Sciences and Convergence Information Technology (ICCIT2012) ณ กรุงโซล สาธารณรัฐเกาหลีใต้

Automatic Stuff Relation Extraction from Scientific Documents for Natural Product Ontology Construction

Suriyasak Lertsakunsomboon
Search Engines and Intelligent Information Systems Research Laboratory
Graduate Program in Web Engineering
Faculty of Information Technology
Dhurakij Pundit University
Bangkok, Thailand
535159090014@mydpu.net

Chaveevan Pechsiri
Search Engines and Intelligent Information Systems Research Laboratory
Graduate Program in Web Engineering
Faculty of Information Technology
Dhurakij Pundit University
Bangkok, Thailand
chaveevan.pec@dpu.ac.th

***Abstract*— To extract Part-Whole relations, especially the stuff relation, from unstructured textual data is the challenging work. This paper presents how to automatically extract the stuff relation from technical documents on the Web for supporting chemical industries. The research extracts the stuff relation without applying POS (Part-of-Speech) annotation. There are three problems of extracting the stuff relation: a) the identification of stuff relation without POS annotation problem, b) the chemical-formula-embedded name entity determination problem and c) the genus-species name entity determination problem. We propose using Naive Bayes to learn the stuff relation. The results from our proposed methodology are 87% precision and 61% recall.**

Keywords-stuff relation; scientific name entity; chemical name entity

I. INTRODUCTION

Through out history, there are consistent amount of interests to extract chemicals of natural products for using in different areas of the pharmaceutical industry [1], the alternative energy research [2], and the commercial biorefinery system [3]. Therefore, several natural product researches work on analysis of natural substances of land and sea and of plants, microbes and animals where the research results bring significant benefits to the industries, especially the pharmaceutical industry. For example, many natural product researches focus on identifying novel bioactive constituents from mushrooms and medicinal plant. Recent studies identified antibacterial and cytotoxic mushroom species (where cytotoxic mushroom uses as anti-cancer medicines), as well as novel compounds from Bangladeshi medicinal plants. Since numerous scientific literatures exist on natural substances, the allocation of large amounts of time and resources are required by industries to seek source organism alternatives for a specific constituent. Therefore this research proposes to automatically extract the Part-Whole relation (i.e. "X is the component of Y") from the research documents to reveal the source organism substances of the natural product. According to M.E. Winston et. al. 1987, the Part-Whole relation can be classified into six types: Component-Integral object (wheel-car), Member-Collection (soldier-army), Portion-Mass (meter-kilometer), Stuff-Object (alcohol-wine), Feature-Activity (paying-shopping), and Place-Area (oasis-desert). Therefore, this research aims to determine and extract the Part-Whole relation, especially Stuff-Object type, from the documents. Unlike the Component-Integral relation, the Stuff-Object relation (or Stuff relation) refers to a relation where the stuff cannot be physically separated from the object without altering its identity [4]. The Stuff relation is required for automatically constructing the natural product Ontology used to represent all natural product knowledge.

There are several researches work on the relation extraction [5][6][7][8], but literature on the Part-Whole (or Part-of) relation extraction is still lacking. Most of these researches worked on the text file involved with the linguistic patterns or the linguistic rule bases at the phrase level (e.g. "the oil and vinegar salad dressing") or the sentences level (e.g. "A vinaigrette-type salad dressing in United States and Canadian cuisine consists of water, vinegar or lemon juice, vegetable oil, chopped bell peppers corn syrup, and a blend of various herbs and spices."). Our research concerns only the sentence level since the Stuff relation expression on our corpus of the natural product research papers mostly occurs at the sentence level. In addition, our corpus contains several characteristics differ from other general corpora, especially the name entity (especially IUPAC nomenclature or chemical name entity) such as the chemical-formula-embedded name entity as shown the following with the underline.

"*Seven new lanostane-type triterpenes, hypo-crellols A−G (1−7), and six new hopane-type triterpenes, 7β,15α-dihydroxy-22**(29)**-hopene (8), 3β,7β-dihydroxy-22(29)-hopene (9), 3β-acetoxy-15α-hydroxy-22(29)-hopene (10), 3β,7β,15α,22-tetrahydroxyhopane (11),*

3β-acetoxy- 7β,15α,22-trihydroxyhopane (12), and 7β,15α,22-trihydroxy-hopane (13), were isolated from the scale insect pathogenic fungus Hypocrella sp. BCC 14524."

Where each number in the parenthesis except **(29)** stands for the identification number of the element in front of the parenthesis, for example:

"1−7" (1 to 7) stand for "*lanostane-type triterpenes, hypocrellols A−G*"

"8" stands for "*hopane-type triterpenes, 7β,15α-dihydroxy-22(29)-hopene*"

Another problem is the chemical name entity contains several commas whose functions differ from the general comma function separating word in a series (www.brighthubeducation.com). For example:

> "*… … …, 3,4-dihydroxy-9-methoxypterocarpan (vesticarpan) (5), 2′,4,4′-trihydroxychalcone (isoliquiritigenin), and 7,4′-dihydroxyflavanone (liquiritigenin) (6), were isolated from the heartwood of Platymiscium floribundum.*"

where "*2′,4,4′-trihydroxychalcone*" is one word. Whereas "*dogs*" / "*cats*" / "*mice*" is a one word in the following sentence.

"*Dogs, cats, and mice are mammals.*"

All of these characteristics are involved in the three main problems; first is how to identify the Stuff relation. Second is how to determine the chemical formula name entity. Third is how to determine the organism (species) name entity. From all of these problems, we propose applying the Naïve Bayes machine learning technique to learn the stuff relation from a sentence that contains the interesting word set {"*obtain*", "*isolate*", "*extract*", …}(from corpus behavior study) along with the stuff word set {"*7,4′-dihydroxy-3′-methoxyisoflavone*", "*isoliquiritigenin*", "*(R)-4″-methoxydalbergione*", …}(from NBIC-PubChem database) and the object word set { "*Dalbergia louvelii*", "*Dendrolobium lanceolatum*", … }(from NBIC-Taxonomy database).

In section II, related work is summarized. Problems in extracting the Stuff relation from the published research papers is described in section III and in section IV is purposed our framework for the Stuff relation extraction. In section V, we evaluate and conclude our proposed model.

II. Related Works

There are some previous researches including [5][6][7][8] working on the relation extracting from texts as described in the following.

R. Girju et. al.(2003)'s work [5] is to present a ID3 (C4.5) learning technique for learning semantic constraints based on the lexico syntactic pattern , NP1 verb NP2 where NP1 and NP2 are noun phrases, to detect Part-Whole relations (meronymy) from the LA Times articles of TREC 9 text collections. They also stated that the Part-Whole relations in WordNet were classified into three basic types: Member-of (e.g., UK IS-MEMBER-OF NATO), Stuff-of (e.g., carbon IS-STUFF-OF coal), and Part-of (e.g., leg IS-PART-OF table). According to their proposed model based on sentence level, the Part Whole relations are detected with an accuracy of 83% precision 98% recall from 10000 sentences.

In 2006, P. Pantel and M. Pennacchiotti's work [6] is to present the weakly-supervised algorithm, named Espresso, using the generic patterns to extract the semantic relations, especially the Is-a relation and the Part-of relation. The Espresso was applied to the very large corpora downloaded from webs with the generic pattern which was determined the pattern and instance reliability. The reliable patterns resulted in having high precision but often very low recall (e.g., "X consists of Y" for the Part-of relation). Their experimental results showed that their generic patterns substantially increased system recall with small effect on overall precision. The results of their system performance using the CHEM corpus with the Part-of relation is 51% precision and 46 relative recall whereas using TREC-9 with the Part-of relation is 70% precision 577 relative recall.

In 2007. K.Fundel et. al. [7] developed RelEx, an approach for relation extraction from free text, especially biomedical publications from million MEDLINE abstracts. RelEx was based on natural language preprocessing producing dependency parse trees dealing with gene and protein relations. The RelEx model involved the detection of co-occurrences of entities within sentences or abstracts and uses a small set of simple rules, applied for part-of-speech-tagging, noun-phrase-chunking and dependency. Their Relation Extraction is based on three linguistic rules frequently used in English language for describing relations where one rule is based on the sentence level and the other two rules are based on the phrase level as follow:

(1) effector-relation-effectee ('α activates β')

(2) relation-of-effectee-by-effector ('Activation of α by β')

(3) relation-between-effector-and-effectee('Interaction between α and β').

Finally, RelEx is estimated performance of both 80% precision and 80% recall.

G. I. Brown (2011) [8] presented an analysis of a relation extraction system using a support vector machine (SVM) classifier to the J.D. Power and Associates Sentiment Corpus separated into three style documents: professionally written reviews, blog reviews, and social networking reviews. The SVM features includes the word features involved the head noun of the phrase existing the least deep in the dependency parse tree, the entity types, and the token class. However, his research aims to study how the extraction system works on different styles of documents. The results of the Part-of relation extraction are 46%precision on average and 33% recall on average. Then, [8] concluded that the relation extraction task was being negatively impacted by the relation

classification itself and the poor tokenization or parsing of the documents.

However, all of the relation extraction techniques from [5][6][7][8] cannot be applied to our research. The complicated chemical-formula name entity and the organism name entity limit the use of POS standard tools (http://open.xerox.com/Services/fst-nlp-tools/Consume/181). Therefore, we apply Naïve Bayes to learn the Stuff relation from a sentence that contains the interesting-stem word set along with the Stuff word set and the object word set.

III. Problem of Stuff Relation Extraction

There are three main problems; how to identify the Stuff relation, how to determine the chemical-formula-embedded name boundary, and how to determine the scientific name boundary.

A. How to identify Stuff Relation without POS annotation

In order to identify the Stuff relation without the POS annotation, we propose using the A-B-C or C-B-A sequence (where A is the Stuff word set, B is the interesting word set, and C is the object word set) occurring within one sentence to identify the Stuff relation using Naïve Bayes. A and C are obtained from NBIC-PubChem and NBIC-Taxonomydatabase (http://pubchem.ncbi.nlm.nih.gov) respectively (see Fig. 1.).

> "… Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii*. ..."

Fig. 1. Example of the linguistic expression in scientific documents.

where A is "flavonoids (1-4)", B is "isolate", and C is "*Dalbergia lovelii*".

B. How to determine the chemical-formula-embedded name boundary

The chemical formula is very complex word (as shown in TABLE I.) where the chemical names are embedded within another chemical name (see section I). Therefore we used NBIC-PubChem to determine the chemical name.

TABLE I. List of Chemical Formula Example.

Chemical Formula
Example 1 : (R)-4″-methoxydalbergione
Example 2 : 3-(2,4-dihydroxy-5-methoxy)phenyl-7-hydroxycoumarin
Example 3 : (R)-4″-methoxydalbergione
Example 4 : (7S,8R,1′S,5′S,6′R)-$\Delta^{2',8'}$-5′,6′-dihydroxy-3′-methoxy-3,4-methylenedioxy-4′-oxo-8.1′,7.5′-neolignan
Example 5 : 2,4-dimethoxy- 5,6-methylenedioxy-1-(2-propenyl)benzene
Example 6 : 2′-hydroxy- 6,4′,6″,4‴-tetramethoxy-[7-O-7″]-bisisoflavone

Also, the chemical formula often contains "," as separating word (see section I). We propose using NBIC-PubChem database to determine the chemical formula name entity.

C. How to determine the scientific (species and genus) name boundary

Often the species and genus names of species are written in multiple word italic form (as shown in TABLE II.). However, when the surrounding texts are also in italics, the identification of the scientific species names become a challenging task.

TABLE II. List of Scientific Name Example.

scientific name
Example 1 : *Dalbergia louvelii*
Example 2 : *Dendrolobium lanceolatum*
Example 3 : *Hypericum perforatum*
Example 4 : *Platymiscium floribundum*
Example 5 : *Dalbergia candenatensis*

Therefore we propose using NBIC-Taxonomy database to solve this problem.

IV. Methodology

There are 4 major steps including corpus preparation, name entity determination, Stuff relation learning, and Stuff relation extraction as shown in Fig. 2.

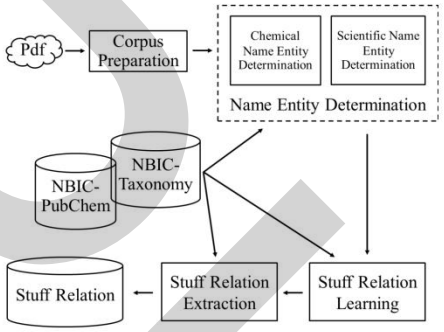


Fig. 2. System architecture.

A. Corpus Preparation

In the corpus preparation step, English scientific abstracts and introductions in chemical related areas are downloaded from online journals (20,000 sentences from 500 scientific documents in pdf format). The pdf documents are converted to text documents using PDFTextStream (http://snowtide.com). The corpus is separated into 2 parts; one part is 15,000 sentences for learning the stuff relation and the other part of 5,000 sentences for stuff relation extraction. The sentences with stuff relation are manually annotated as "yes" class and the others are "no" class (see Fig. 3.). Then the stop words are filtered out while all symbols , e.g. "," "-" … etc., still exist in the corpus.

<stuff_relation class="yes">Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii* by following their potential to inhibit in vitro the growth of *Plasmodium falciparum*.</ stuff_relation>
<stuff_relation class="no">Of these, the ethyl acetate extract obtained from the heartwood of *Dalbergia louvelii* R. Viguier (Fabaceae).</stuff_relation>
<stuff_relation class="yes">Although several isoflavonoids have been obtained from roots of *P. floribundum*, none of the abovementioned compounds have been isolated previously from this species.</stuff_relation>
<stuff_relation class="no">The cytotoxicity of the isolates obtained herein from *P. floribundum* has been evaluated against a small panel of cancer cell lines.</stuff_relation>

Fig. 3. Examples of sentences annotation.

B. Scientific Name Entity Determination

1) Chemical Name Entity Determination: There are 2 steps involved: the translation of numeric representation of chemicals and the identification of chemical name entities.

a) Translating Numeric Representation of chemicals: The following observed rule are used to translate the numeric representations of chemicals.

Rule:

if((Cname is the first occurrence)∧ (Cname(num))then
num is the numeric representation of Cname

Where Cname = chemical name on a scientific document
num = the integer

b) Chemical Name Entity Identification: The results from a) coupled with the adjacent surrounding word are compared to the NBIC-PubChem.

2) Genus-species Name Entity Determination: Using NBIC-Taxonomy database solves the Genus-species Name Entity.

C. Stuff Relation Learning

In the learning step, b (where b ∈ B and B is obtained from the corpus behavior studied) is used to anchor sentences. Sentences where the left hand side of the anchor contains a (where a ∈ A) or c (where c ∈ C) and the right hand side of the anchor contains c or a are collected. Then the frequency of a, and c with class "yes" and class "no" are determined for each sentences (see TABLE III.)

TABLE III. FREQUENCY OF A AND C WITH CLASS "YES"/"NO"

A	Class=yes	Class=no
3,10-dihydroxy-9-methoxypterocarpan	0.05882353	0.03571429
ethyl acetate	0.05882353	0.17857143
dibenzocycloheptene	0.3921569	0.03571429
…	…	…
C	**Class=yes**	**Class=no**
Dendrolobium	0.06976744	0.05
Platymiscium	0.34883721	0.25
Dalbergia	0.06976744	0.25
…	…	…

D. Stuff Relation Extraction

In order to start the Stuff relation extraction process, Naive Bayes Classifier[9] shown in equation (1) is applied. The class "yes" means Stuff relation, as shown in Fig. 4.

$$Stuff\ \mathrm{Re}lationClass = \arg\max_{class \in Class} P(class \mid a, c) = \arg\max_{class \in Class} P(a \mid class)P(c \mid class)P(class) \quad (1)$$

$$where\ \ Class = \{"yes","no"\}$$
$$a \in A \quad (A\ is\ a\ Stuff\ word\ set)$$
$$c \in C \quad (C\ is\ an\ object\ word\ set)$$

```
L is a list of sentence.
W is a set of word in each sentence.
WS is a sequence of n words where n = 1, 2, 3, … .
C is natural-product-compounds concept set.
S is the natural-sources concept set.
A is the marker set.
1   i ← 0, R ← ϕ
2   while (i ≤ length[L]) do
3   Begin
4     j ← 0, len ← length[Wi]
5     while(wij ∈ A) do
6       for (k ← 0 to j)
7         if (wsik ∈ C) then
8           for (m ← j+1 to len)
9             if ((wsim ∈ S) ∧ (StuffRelationClass
              == "yes")then
10              R = R ∪ {(wsik, wsim)}
11            end if
12          next
13        else if (wsik ∈ S) then
14          for (m ← j+1 to len)
15            if ((wsim ∈ C) ∧ (StuffRelationClass
              == "yes")) then
16              R = R ∪ {(wsim, wsik)}
17            end if
18          next
19        endif
20      next
21  End
22  return R
```

Fig. 4. Stuff Relation Extraction Algorithm.

V. EVALUATION AND CONCLUSION

The English corpora of the technical documents in chemistry domain are used to evaluate the proposed stuff relation extraction algorithm consisting of about 5,000 sentences. The evaluation of the Stuff Relation extraction performance in research is expressed in terms of the precision and the recall as shown below, where R is the stuff relation:

$$\text{Precision} = \frac{\#\text{ of samples correctly extracted as R}}{\#\text{ of all samples output as being R}} \quad (2)$$

$$\text{Recall} = \frac{\#\text{ of samples correctly extracted as R}}{\#\text{ of all samples holding the target relation R}} \quad (3)$$

The results of precision and recall are evaluated by three expert judgments with max win voting. The precision of the extracted Stuff relation is 87% and 61% recall. The reason of the recall limited to 61% is misplaced-B problem where B is before or after A and C. These problems are a subject of further studies. Finally, these extracted Stuff relations are beneficial for natural product chemical ontology construction (see Fig. 5.) which will bring significant benefits to the cosmetics, pharmaceuticals and other chemicals industries.

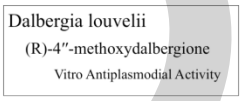


Fig. 5. Example of Natural Product Chemical Ontology.

REFERENCES

[1] Bernard Munos, "Lessons from 60 years of pharmaceutical innovation," Nature Reviews in Drug Discovery, vol. 8, pp. 959-968, December 2009.

[2] Yusuf Chisti, "Biodiesel from microalgae," Biotechnology Advances, vol. 25, pp. 294–306, February 2007.

[3] Gail Taylor, "Biofuels and the biorefinery concept," Energy Policy, vol. 36, pp. 4406–4409, 2008.

[4] Morton E. Winston, Roger Chaffin, Douglas Herrmann, "A taxonomy of part-whole relations," Cognitive Science, vol. 11, pp. 417-444, 1987.

[5] Roxana Girju, Adriana Badulescu, Dan Moldovan, "Learning semantic constraints for the automatic discovery of part-whole relations," In Proceedings of HLT/NAACL-03, vol. 1, pp. 1-8, 2003.

[6] Patrick Pantel, Marco Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," In Proceedings of COLING/ACL-06, pp. 113-120, 2006.

[7] Katrin Fundel, "Text Mining and Gene Expression Analysis Towards Combined Interpretation of High Throughput Data," München, 2007

[8] Gregory Ichneumon Brown, "An Error Analysis of Relation Extraction in Social Media Documents," In Proceedings of ACL-HLT-49, pp. 64-68, June 2011.

[9] Tom Mitchell, Machine Learning. Singapore: The McGraw Hill Companies Inc. and MIT Press, 1997.

ประวัติผู้เขียน

ชื่อ-นามสกุล	สุริยศักดิ์ เลิศสกุลสมบูรณ์
ประวัติการศึกษา	สำเร็จการศึกษาระดับปริญญาตรีสาขาวิชา วิทยาศาสตร์(เคมี) มหาวิทยาลัยมหิดล ปีการศึกษา 2550
ตำแหน่งและสถานที่ทำงานปัจจุบัน	ผู้ช่วยสอน (Teaching Assistant) สาขาวิสวกรรมเว็บ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์