การหาค่าความสัมพันธ์ระหว่างคำศัพท์กับหมวดหมู่

RELATION TEST BETWEEN
ONLINE DICTIONARY VOCABULARY AND CATEGORY

โดย

เอมอร สิริศุภางค์

AIMON SIRISUPHANG

อาจารย์ที่ปรึกษา

ผศ.ดร. พรฤดี เนติโสภากุล

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2549

RELATION TEST BETWEEN
ONLINE DICTIONARY VOCABULARY AND CATEGORY

AIMON SIRISUPHANG

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2006

หัวข้อ	การหาค่าความสัมพันธ์ระหว่างคำศัพท์กับหมวดหมู่
นักศึกษา	นางสาวเอมอร สิริศุภางค์
รหัสนักศึกษา	48066539
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	ผศ.ดร. พรฤดี เนติโสภากุล

บทคัดย่อ

ในปัจจุบัน มี Dictionary ต่างๆมากมาย อยู่หลายรูปแบบ ทั้งแบบหนังสือ และแบบ dictionary online ซึ่งมีคำศัพท์ใหม่ๆ เกิดขึ้นอยู่เสมอ ดังนั้น การค้นหาคำศัพท์จึงอาจจะไม่พบความหมายของศัพท์ที่ต้องการ และขณะนี้ได้มี website ต่างๆเกิดขึ้นมากมาย ซึ่งบนหน้า website มักมีคำศัพท์ใหม่ๆเสมอ และบางครั้งก็อาจพบคำศัพท์ที่เป็นความหมายของกันและกัน และสามารถนำมาใช้ประโยชน์ได้ ในรายงานฉบับนี้จึงได้วิเคราะห์วิธีการจัดทำ ดิกชั่นเนรี่อังกฤษ-ไทยจากเวปไซต์ ด้วยทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web โดยค้นหาคำศัพท์จากหน้า website และนำมาเป็นเก็บรวบรวมไว้ และหาความสัมพันธ์ของหมวดหมู่ที่จัดเตรียมไว้และคำศัพท์ที่เก็บรวบรวมได้ เพื่อใช้เป็นประโยชน์ในการทำ dictionary online ซึ่งจะได้คำศัพท์ใหม่ๆมากมายที่รวบรวมได้จาก website เหล่านี้ และใช้ประโยชน์จากการทราบถึงความสัมพันธ์ของคำศัพท์ต่างๆ เพื่อเป็นประโยชน์ในการสืบค้นหาความหมาย และเพื่อพัฒนาระบบงานอื่นๆ เช่น machine translation ต่อไปในอนาคต

Title	Relation test between online dictionary vocabulary and category
Student	Miss Aimon Sirisuphang
Student ID.	48066539
Degree	Master of Science
Programme	Information Science
Academic Year	2006
Advisor	Asst.Prof. Dr.Ponrudee Netisopakul

ABSTRACT

Nowadays, there are many dictionaries in many different format, e.g. book, online website. New words appear everyday, so finding the desire word's meaning maybe difficult. However, the online websites always contain new vocabularies, and they are also the meaning of the other words. These websites are very useful. This project analyzes the method of creating Dictionary English - Thai from website with Automatic Collection of Vocabulary and Related Term from the Web. This method searches vocabularies from websites and stores on the database, finds the relation between prepared categories and the stored vocabularies. The result is dictionary online which contains many new vocabularies from websites. Moreover, the relation values of each vocabulary can be used for other searching its meaning, and for other further research, such as, machine translation.

กิตติกรรมประกาศ

การจัดทำโครงการพัฒนาระบบงานในหัวข้อเรื่องการหาค่าความสัมพันธ์ระหว่างคำศัพท์กับหมวดหมู่นี้สำเร็จได้นั้นด้วยความกรุณาจากอาจารย์ที่ปรึกษา ผศ.ดร. พรฤดี เนติโสภากุล ที่กรุณาให้คำปรึกษาแนะนำ และช่วยตรวจสอบแก้ไขข้อบกพร่องของโครงการนี้ ตลอดจนให้ความรู้และข้อคิดเห็นที่เป็นประโยชน์อย่างยิ่งต่อโครงการ

ขอขอบพระคุณ คุณชัยยงค์ รักขิตเวชสกุล และ คุณชัชพงศ์ สุธีสุขสถาพร Software engineer บริษัท Microsoft (Thailand) ที่ช่วยสนับสนุนการทำโครงการ ช่วยให้คำแนะนำแนวทางและข้อเสนอแนะที่เป็นประโยชน์ยิ่งต่อทำโครงการ

และขอขอบคุณเพื่อนๆ ทุกคนที่ให้กำลังใจซึ่งกันและกัน แลกเปลี่ยนประสบการณ์ซึ่งกันและกัน จนทำให้การพัฒนาโครงการนี้สำเร็จลุล่วงไปด้วยดี

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่ร่วมเป็นกำลังใจ และให้การสนับสนุนในทุกๆ เรื่อง ด้วยดีเสมอมา

สำหรับคุณงานความดีและประโยชน์อันพึงมาจากโครงการฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และถ่านทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

เอมอร สิริศุภางค์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันความก้าวหน้าทางเทคโนโลยีด้านสารสนเทศได้รับการพัฒนาขึ้นอย่างมาก ทำให้ข้อมูลต่างๆจำเป็นต้องถูกรวบรวมไว้ในหน่วยความจำของคอมพิวเตอร์ และนำมาใช้งานได้ตรงความต้องการของผู้ใช้งานที่สุด หนึ่งในข้อมูลที่สำคัญคือคำศัพท์และคำแปล ซึ่งเป็นสิ่งที่เราสามารถนำมาใช้ประโยชน์ต่างๆได้ โดยเฉพาะการทำ Online Dictionaryในขณะนี้มี Dictionary ต่างๆ มากมาย อยู่หลายรูปแบบ ทั้งแบบหนังสือ และแบบ Online Dictionary ซึ่งมีคำศัพท์ใหม่ๆ เกิดขึ้นอยู่เสมอ ดังนั้น การค้นหาคำศัพท์จึงอาจจะไม่พบความหมายของศัพท์ที่ต้องการ Dictionary คือสิ่งที่รวบรวมการแปลภาษาจาก ภาษาหนึ่งไปยังอีกภาษาหนึ่ง เพื่อช่วยผู้ที่ต้องการแปลภาษาที่ตนเองไม่ถนัด เพื่อใช้ประโยชน์ต่างๆกัน ในปัจจุบันมีการทำอยู่ในหลายรูปแบบเช่น การทำเป็นหนังสือ, เครื่อง talking dictionary หรือ ที่นิยมกันมาในปัจจุบัน คือ Online Dictionary ซึ่งมีการจัดทำผ่านระบบ website เพราะผู้ใช้จะสะดวกต่อการใช้งาน เพียงติดต่อระบบ internet เท่านั้น และขณะนี้ได้มี website ต่างๆเกิดขึ้นมากมาย ซึ่งบนหน้า website มักมีคำศัพท์ใหม่ๆเสมอ และบางครั้งก็อาจพบคำศัพท์ที่เป็นความหมายของกันและกัน และสามารถนำมาใช้ประโยชน์ได้ ในรายงานฉบับนี้จึงได้วิเคราะห์วิธีการจัดทำ dictionary อังกฤษ-ไทย โดยการรวบรวมคำศัพท์จาก website ต่างๆที่มีคำศัพท์และคำแปล นำมาเป็นเก็บรวบรวมไว้ โดยที่มีการจัดทำหมวดหมู่ต่างๆของคำศัพท์ เพื่อหาความเกี่ยวข้องกันของคำศัพท์ต่างๆ และบ่งบอกว่าเป็นคำศัพท์ที่เกี่ยวข้องกับเรื่องใด โดยที่มีการจัดเรียงคำศัพท์ที่ได้ตามหมวดหมู่ และเก็บไว้เรียงตามหมวดหมู่โดยที่จะมีค่าความเป็นไปได้ในการที่คำศัพท์แต่ละตัวจะอยู่หมวดต่างๆอีกด้วย เพื่อใช้ให้เป็นประโยชน์ต่างๆ โดยเฉพาะการทำ Online Dictionary ซึ่งจะได้คำศัพท์ใหม่ๆมากมายที่รวบรวมได้จาก website เหล่านี้ เพื่อเป็นประโยชน์ในการสืบค้นหาความหมายต่อไปในอนาคต

การพัฒนาโปรแกรมรวบรวมคำศัพท์จากหน้าเวปไซต์ มีการทำงานโดยใช้ Automatic Term Recognition เพื่อทดสอบความเกี่ยวเนื่องกันของคำศัพท์และหมวดหมู่การทำงาน และมีการวัดผลค่าที่ได้ว่าเป็นความเกี่ยวเนื่องกันแค่ไหน โดยใช้การทำ Relation test และเก็บผลของการทำงาน โดยใช้หลักการวิเคราะห์ความแม่นยำของผลที่ได้จากการทำ Relation test ว่าต้องปรับปรุงแก้ไขอย่างไรเมื่อนำมาใช้จริง และการวิเคราะห์ผลลัพท์ รวมถึงการเพิ่มประสิทธิภาพในการค้นหาศัพท์ และความถูกต้องแม่นยำของศัพท์ที่ได้ เพื่อนำมาใช้ประโยชน์ในการทำ Online Dictionary และการใช้ประโยชน์เพิ่มเติมด้านอื่นๆในอนาคต

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

การศึกษาโครงงานนี้มีวัตุประสงค์คือ

1. เพื่อศึกษาการทำการเก็บรวบรวมคำศัพท์เป็นจากหน้าเวปไซต์ ศึกษาวิธีการหาคำศัพท์จากเวปไซต์พร้อมทั้งคำแปลและเก็บรวบรวม
2. เพื่อพัฒนาโปรแกรมรวบรวมคำศัพท์จากหน้าเวปไซต์ด้วยวิธีการ Compiling Corpus , Automatic Term Recognition และหาความสัมพันธ์ของคำศัพท์และหมวดหมู่ต่างๆ Relation test (Filtering)
3. เพื่อศึกษาความถูกต้องและความแม่นยำของวิธีการทำ Relation test (Filtering) ว่าค่าความน่าจะเป็นที่ได้มีความถูกต้องมากเพียงใดเพื่อสามารถนำไปใช้ในการหาความสัมพันธ์ของข้อมูลอื่นๆได้ หรือควรปรับปรุงต่อไปในอนาคต
4. เพื่อประยุกต์คำศัพท์ที่ได้มาใช้ในการทำประโยชน์ด้านต่างๆ โดยที่โครงงานนี้จะประยุกต์ใช้ในการทำ Online Dictionary
5. เพื่อให้คำศัพท์ที่รวบรวมได้นำไปทำโครงงานอื่นได้ในอนาคตเช่น การทำ machine translation

1.3 สมมติฐานของการศึกษา

ข้อด้อยของการหาค้นหาคำศัพท์จากเวปไซต์และความสัมพันธ์ของคำศัพท์และหมวดหมู่ คือ ค่าความน่าจะเป็นที่คำศัพท์จะเกี่ยวข้องกับหมวดหมู่ที่เราหาได้จากสูตร อาจจะไม่ได้ถูกต้อง 100% ซึ่งด้วยสมมุติฐานของผลลัพท์จากวิธีการหา related term นี้ คำศัพท์ที่ได้มากกว่า 80% จะมีความเกี่ยวข้องกับหมวดหมู่นั้น แต่อาจจะเป็นส่วนนึงของหมวดหมู่นั้นๆหรือไม่ก็ได้ การดึงคำศัพท์จากหน้าเวปไซต์ และการหาค่าความสัมพันธ์ใช้เวลานานต่อศัพท์ 1 คำ และทั้งนั้นเรามีศัพท์ทั้งหมดที่จะรวบรวมและนำมาหาความสัมพันธ์มากกว่า 100,000 ตัว โดยที่มีหมวดหมู่หลักและหมวดหมู่ย่อยรวมกันมากกว่า 50 หมวดหมู่ และเราต้องหาความสัมพันธ์ของคำศัพท์ทั้งหมดกับหมวดหมู่ที่เราต้องการทราบ จะเห็นได้ว่าเป็นปริมาณจำนวนมากที่คอมพิวเตอร์จะต้องนำมาประมวลผล รวมถึงการที่ต้องมีการทำงานกับฐานข้อมูลทั้งการอ่านและการเขียนจำนวนมากกว่า 100000 ครั้ง ดังนั้นทำให้สมรรถนะของเครื่องคอมพิวเตอร์ต่ำลงซึ่งเป็นผลทำให้การทำงานของโปรแกรมช้าลงไปตามระยะเวลา ทำใช้เวลานานมากในการประมวลผลทั้งการหาคำศัพท์และการหาความสัมพันธ์ ดังนั้นจึงต้องหาแนวทางในการแก้ปัญหาในจุดนี้ เมื่อลองแล้วว่าการทำงานมีปัญหาตรงจุดนี้จริง เราจึงต้องทำการดึงคำศัพท์จากหน้าเวปไซต์จะต้องทำการตั้งเวลาให้หยุดการทำงานและเริ่มใหม่ที่จุดเดิมเป็นช่วงๆเพื่อทำให้การทำงานมีประสิทธิภาพดีขึ้น

1.4 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

เราใช้เทคนิคการทำ Automatic Collection of Vocabulary and Related Term from the Web ที่ประกอบไปด้วยการทำ 3 เทคนิค คือ Compiling Corpus, Automatic Term Recognition, Relation test (Filtering) จากแนวคิดที่ต้องการจัดเรียงคำศัพท์เป็นหมวดหมู่แทนการจัดเรียงคำศัพท์ตามตัวอักษร และความสัมพันธ์ของคำศัพท์กับคำศัพท์และหมวดหมู่ต่างๆ โดยแนวคิดนี้จะทำให้เราสามารถดึงประโยชน์จากข้อมูลบนเวปไซต์และ search engine website ให้เป็นประโยชน์ และเป็นแนวคิดรูปแบบใหม่ในการรวบรวมและจัดเก็บคำศัพท์ เพื่อให้สามารถนำคำศัพท์ที่เราได้มาใช้ให้เกิดประโยชน์ในแนวทางใหม่ๆ และรวบรวมคำศัพท์ที่ได้มาพร้อมค่าความสัมพันธ์มาจัดเป็นฐานข้อมูล คดยจะมีการจัดทำ online dictionary ที่มีรูปแบบใหม่ ซึ่งสามารถจะค้นหาความหมายของคำศัพท์ได้ในแบบเดิมๆและในแบบใหม่ คือคำศัพท์ทั้งหมดที่มีความเกี่ยวข้องกัน หรือเกี่ยวข้องกับหมวดหมู่ โดยที่จะมีการจัดทำฐานข้อมูลใหม่ที่มีการจัดเรียนตามหมวดหมู่ เพื่อสามารถนำไปประยุกต์ใช้ในโครงงานอื่นๆในอนาคตต่อไปได้อีกด้วย เช่นการทำ machine translator เป็นต้น

1.5 การเปรียบเทียบระหว่างวิธีการที่นำเสนอกับวิธีการแบบพื้นฐาน

จากวิธีการเก็บรวบรวมคำศัพท์แบบพื้นฐานที่เรียงตามอักษร ซึ่งการค้นหาคำศัพท์มาใช้งาน เช่นการทำ online dictionary ก็จะหาได้เฉพาะความหมายของคำนั้นๆ แต่ในกรณีนี้ วิธีการทำ Automatic Collection of Vocabulary and Related Term from the Web ที่เสนอในรายงานฉบับนี้ จะได้ให้ได้ฐานข้อมูลที่มีการจัดลำดับความสัมพันธ์ของคำศัพท์ไว้กับหมวดหมู่ต่างๆ ว่าศัพท์คำใดมีความเกี่ยวข้องกับหมวดหมู่ใดมากน้อยเพียงใด ทำให้ได้ค่าของความน่าจะเป็นในการเกี่ยวข้องกันของคำศัพท์และคำศัพท์ หรือของคำศัพท์และหมวดหมู่ ทำให้เมื่อมีการค้นหาคำศัพท์ผ่านทาง online dictionary สามารถค้นหาคำศัพท์ในหมวดหมู่เดียวกันกับคำศัพท์ที่เราต้องการค้นหาด้วย การทำ dictionary โดยใช้คำศัพท์จาก website ก่อให้เกิดประโยชน์ในการรวบรวมและค้นหาคำศัพท์ใหม่ๆ ที่เกิดขึ้นอยู่ตลอดเวลา ซึ่ง dictionary แบบเก่าที่มีมา มักไม่มีการรวบรวมคำศัพท์ใหม่ๆ หรือคำศัพท์เฉพาะด้านต่างๆเอาไว้ ทำให้การค้นหาของผู้ใช้งานเป็นไปได้อย่างไม่ครบถ้วน ดังนั้น รายงานฉบับนี้จึงกล่าวถึงการจัดทำ website Online Dictionary ซึ่งเป็นรูปแบบใหม่ของการค้นหาศัพท์ โดยวิธีการที่ได้วิเคราะห์ในรายงานฉบับนี้ ทั้งหารหาคำศัพท์ การหา website การเก็บรวบรวม จนถึงการนำมาใช้งาน สามารถนำไปใช้ได้จริง ทำได้ไม่ยาก และมีประสิทธิภาพ โดยที่สามารถทำให้ระบบกลับมาค้นหาคำศัพท์อีกครั้งได้ใหม่ ตามความถี่ที่เราตั้งไว้ ซึ่งจะทำให้ได้คำศัพท์ใหม่ๆ เสมอ และจะเป็นประโยชน์กับผู้ใช้งาน ในการค้นหาคำศัพท์ โดยเฉพาะในการใช้คำศัพท์จาก website Online Dictionary ซึ่งจะมีคำศัพท์ใหม่ๆ และคำศัพท์แตกต่างกันไปในแต่ละ website ทำให้เราได้ข้อมูลมากมายที่รวบรวมมาจากหลายๆ website และเมื่อทำให้เกิดการ update ตามระยะเวลาที่ตั้งไว้ จะทำ

ให้เมื่อ website ที่เรารวบรวมข้อมูลมามีการเพิ่มคำศัพท์ เราก็จะได้คำศัพท์เหล่านั้นมาด้วย ทำให้ได้คำศัพท์จำนวนมากมาย และนำมาใช้ประโยชน์ได้หลายอย่าง ซึ่งเป็นประโยชน์ต่อผู้ที่ต้องการค้นหาคำศัพท์เป็นอย่างยิ่ง

1.6 ขอบเขตการวิจัย

ผลของการศึกษาทำให้ทราบว่ามีกระบวนการการทำงานที่สามารถรวบรวมคำศัพท์และหาความสัมพันธ์ของคำศัพท์และหมวดหมู่ต่างๆได้ โดยรายงานชุดนี้ จะศึกษาการแปลภาษาอังกฤษ ไปเป็นภาษาไทย และภาษาไทยเป็นภาษาอังกฤษ โดยที่จะมีความแตกต่างกับ Dictionary ทั่วไป คือเป็นคำศัพท์ที่หามาจากหน้าของ website โดยการเปรียบเทียบศัพท์ อังกฤษ-ไทยที่ปรากฏบนหน้า website ซึ่งจะทำให้ได้คำศัพท์ใหม่ๆที่เกิดขึ้น เนื่องจาก website มีการเกิดใหม่ทุกวัน และมักมีศัพท์ใหม่ๆนำมาใช้งาน ดังนั้น ผู้ที่ค้นหาคำศัพท์จาก dictionary มาตรฐานที่มีอยู่เดิม แล้วไม่พบนั้นสามารถลองค้นหาจาก Online Dictionary ที่จัดทำจาก ศัพท์ในหน้า website ได้ ทำให้มีโอกาสที่จะพบความหมายของคำที่ต้องการเพิ่มขึ้น

ในเอกสารสัมมนาฉบับนี้ค้นหาวิธีการดึงข้อมูลจากหน้า Website และวิธีการเปรียบเทียบว่าคำศัพท์ใด เป็นความหมายของคำศัพท์ใด โดยใช้หลักการจากการรวบรวมข้อมูลว่ามีเงื่อนไขใดบ้างที่จะเป็นสิ่งบ่งบอกว่า คำศัพท์ใด เป็นความหมายของคำศัพท์ใด และนำคำศัพท์ที่ได้เก็บรวบรวมและหาความสัมพันธ์ของคำศัพท์และหมวดหมู่ต่างๆ โดยจัดทำเป็น Online Dictionary บน website ต่อไปในอนาคต และใช้ ทฤษฎีการทำ Automatic Collection of Vocabulary and Related Term from the Web

โดยมีขอบเขตหลักๆที่จะศึกษาต่อไปนี้

- การทำ Compiling Corpus
- การทำ Automatic Term Recognition
- การทำ Filtering
- การทำ Relation test

1.7 ขั้นตอนของการศึกษา

วิทยานิพนธ์ฉบับนี้ได้แบ่งเนื้อหาออกเป็น 5 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความเป็นมาของงานวิจัย ความมุ่งหมายและวัตถุประสงค์ สมมติฐาน ทฤษฎีที่ใช้ ขอบเขตของการวิจัย และขั้นตอนการศึกษา

บทที่ 2 กล่าวถึงวิธีการทำ Automatic Collection of Vocabulary and Related Term from the Web และวิธีการทำ Compiling Corpus Automatic, วิธีการทำ Term Recognition, วิธีการทำ Filtering, วิธีการทำ Relation test ทฤษฎีและแนวคิดการประเมินของผลที่ได้รับจากโครงงาน สำหรับแต่ละขั้นตอนของการดำเนินการ

บทที่ 3 กล่าวถึงวิธีการดำเนินการทำโครงงาน เริ่มตั้งแต่การเตรียมข้อมูลเพื่อใช้ค้นหาคำศัพท์จากเวปไซต์ การหาคำศัพท์และความหมายจากเวปไซต์ การเก็บรวบรวมคำศัพท์ การคัดเลือกหมวดหมู่เพื่อนำมาค้นหาความสัมพันธ์กับคำศัพท์ การค้นหาความสัมพันธ์ของคำศัพท์และหมวดหมู่โดยการเก็บค่าความน่าจะของการเกี่ยวข้องกัน การนำคำศัพท์มาเรียงลงฐานข้อมูลโดยเรียงตามความสัมพันธ์ การสร้าง online dictionary เพื่อนำข้อมูลที่ได้มามาใช้ให้เกิดประโยชน์สูงสุดเพื่อใช้ในการค้นหคำศัพท์และความหมายของคำศัพท์ต่างๆ

บทที่ 4 กล่าวถึงผลการทดลอง โดยสรุปผลของการทำการใช้ทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web ในการทำโครงงาน ว่าได้ผลถูกต้องแม่นยำเพียงใด ควรมีการปรับปรุงด้านใด และเหมาะนำไปใช้ประโยชน์ หรือพัฒนาโครงงานใดต่อไปในอนาคต

บทที่ 5 เป็นบทสรุปผลการวิจัยและข้อเสนอแนะ

บทที่2
ทฤษฎีที่เกี่ยวข้อง

2.1 ทฤษฎีการทำ Automatic Collection of Vocabulary and Related Term from the Web (Sato and Y.Sasaki. 2003. "Automatic collection of related terms from the web." 121-124. InProc.)

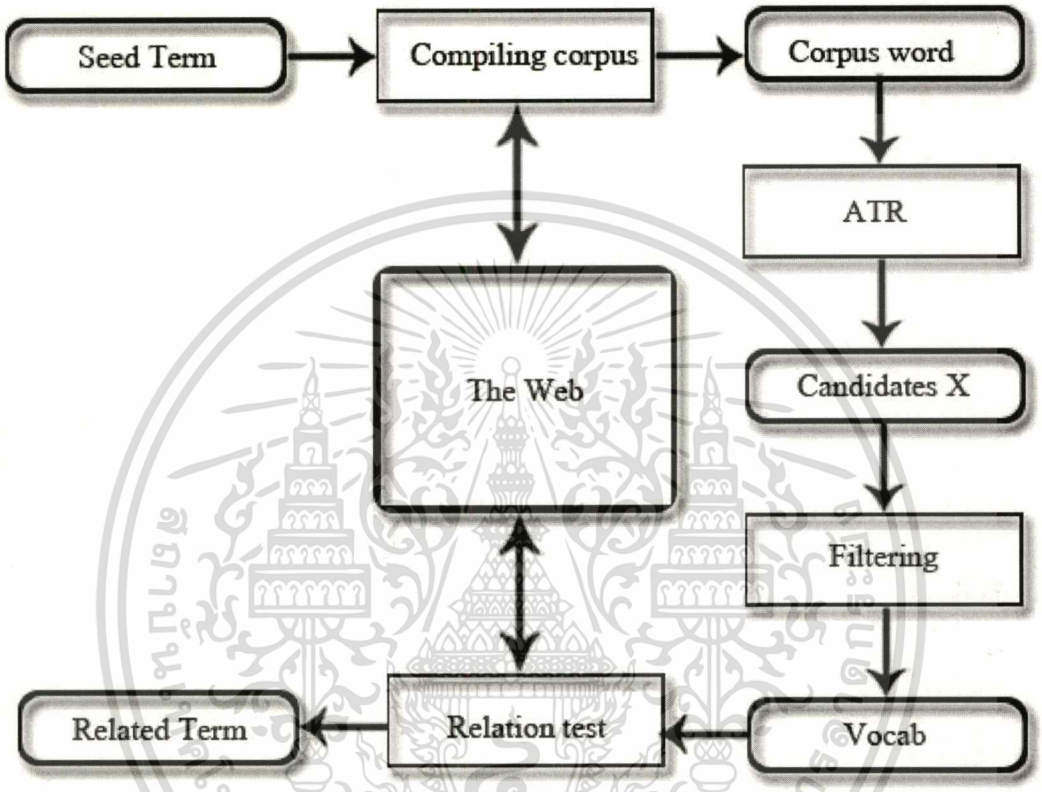
Automatic Collection of Vocabulary and Related Term from the Web คือกระบวนการในการค้นหาคำศัพท์และคำแปลจากหน้าเวปไซต์แบบออโตเมติก และหาคำศัพท์ที่เกี่ยวข้องกันโดยแบ่งตามหมวดหมู่ และหาความน่าจะเป็นในหมวดหมู่นั้นๆแล้วรวบรวมเก็บไว้เป็น Small Corpus ทำให้ได้ข้อมูลที่เป็นคำศัพท์ใหม่ๆจากหน้าเวปไซต์เพื่อนำมาสนับสนุนการทำประโยชน์ในด้านต่างๆเช่นการทำ Online Dictionary กระบวนการนี้เป็นขั้นตอนในการดึงเอาคำศัพท์และคำแปลมาจากเวปไซต์ทำให้ได้ความคำศัพท์ใหม่ๆที่ไม่เคยรู้มาก่อน หรือได้คำศัพท์ที่มีความถูกต้อง หรือคำศัพท์ที่น่าจะนำไปใช้ประโยชน์ได้ Automatic Collection of Vocabulary and Related Term from the Web เป็นส่วนหนึ่งของกระบวนการการค้นพบฐานความรู้จากเวปไซต์ การนำแนวโน้มของคำศข้อมูลต่างๆที่ซ่อนอยู่ในหน้าเวปไซต์ออกมาใช้งานเป็นสิ่งสำคัญ เพราะถ้าไม่รู้จักใช้ประโยชน์จากข้อมูลเหล่านั้นก็สูญเปล่า ดังนั้น กระบวนการนี้เป็นเครื่องมือที่สำคัญในการค้นหาคำศัพท์จากเวปไซต์ที่มีขนาดใหญ่ทำให้ได้คำศัพท์ที่มีประโยชน์ซึ่งซ่อนอยู่ในเวปไซต์ เป็นการเพิ่มคุณค่าให้กับข้อมูลบนเวปไซต์ที่เรามีอยู่

นอกจากนั้น Automatic Collection of Vocabulary and Related Term from the Web ยังเป็นกระบวนการที่ช่วยแบ่งคำศัพท์ออกตามหมวดหมู่ และศึกษาค่าความน่าจะเป็นในการเกี่ยวเนื่องกันของคำศัพท์และหมวดหมู่ต่างๆโดยเก็บข้อมูลที่ได้ไว้ในฐานข้อมูลโดยเรียงตามหมวดหมู่ของคำศัพท์ ดังนั้นกระบวนการนี้ช่วยให้เกิดศักยภาพในการใช้ข้อมูลในเวปไซต์ให้มากยิ่งขึ้น

2.2 กระบวนการทำ Automatic Collection of Vocabulary and Related Term from the Web (Sato and Y.Sasaki. 2003. "Automatic collection of related terms from the web." 121-124. InProc.)

Automatic Collection of Vocabulary and Related Term from the Web มีการรวบรวมเทคนิคต่างๆ เช่น การทำ Compiling Corpus โดยแบ่งเป็น Web page collection และ Sentence extraction, การทำ Automatic term recognition , การทำ Filtering , การทำ Relation Test และสุดท้ายคือการวิเคราะห์ Experiments และ Discussion ขั้นตอนทั้งหมดนี้ได้ถูกทำเพื่อหาคำศัพท์ที่สามารถนำไปใช้งานได้จริง และหาความสัมพันธ์ของคำศัพท์นั้นๆเพื่อใช้ให้เกิดประโยชน์สูงสุด

โดยทั่วไปเมื่อกล่าวถึงการทำ Automatic Collection of Vocabulary and Related Term from the Web ส่วนใหญ่จะคำนึงถึงและให้ความสำคัญกับขั้นตอนการทำ Filtering และ การทำ Relation Test เนื่องจากเป็น 2 ขั้นตอนที่ทำให้เกิดกระบวนการหาคำศัพท์ที่เกี่ยวข้องกันของ และเป็นกระบวนการที่ช่วยค้นพบข้อมูลที่เป็นความรู้จากหน้าเวปไซต์โดยแบ่งขั้นตอนของการทำ Automatic Collection of Vocabulary and Related Term from the Web ได้ดังนี้



รูปที่2.1 Automatic Collection of Vocabulary and Related Term from the Web โดยสรุป

การทำ Automatic Collection of Vocabulary and Related Term from the Web เป็นการทำการค้นหาและตรวจสอบความเกี่ยวข้องกันคำศัพท์สำหรับแต่ละหมวดหมู่ของคำศัพท์ ซึ่งจะประกอบด้วยหลายขั้นตอนที่บางขั้นตอนมีการทำซ้ำๆ หรือต้องมีการวนกลับมาทำซ้ำใหม่ ซึ่งจะประกอบด้วยการทำงาน 4 กระบวนการ ได้แก่

- Compiling corpus
- Automatic Term Recognition
- Filtering
- Relation test

2.3 งานหลักของการทำ Automatic Collection of Vocabulary and Related Term from the Web แบ่งออกเป็น 4 ส่วน ดังนี้

- Compiling corpus รวบรวมหน้าเวปไซต์ที่มีคำศัพท์และคำแปลมาจัดเก็บเป็น Corpus
- Automatic Term Recognition กำหนดเงื่อนไขในการค้นหาคำศัพท์และตัดประโยคที่มีคำศัพท์พร้อมคำแปลมาจัดเก็บไว้เป็น Corpus ที่มีขนาดเล็กลง
- Filtering ตัดคำศัพท์และคำแปลเฉพาะที่ต้องการจริงๆและนำมาเก็บลงฐานข้อมูล
- Relation test หาความสัมพันธ์ของคำศัพท์และหมวดหมู่ต่างๆและนำมาเก็บลงฐานข้อมูล โดยเรียงคำศัพท์ตามหมวดหมู่และบอกค่าความน่าจะเป็นในการอยู่ในหมวดหมู่นั้นๆ

Compiling corpus

Compiling corpus เป็นกระบวนการที่เริ่มจากการส่งค่า seed term เข้าไปใน compiling corpus โดยในที่นี้จะใช้ seed term คือ URL ของ Website dictionary และกระบวนการทำ Compiling corpus จะทำการเลือก Web page ทั้งหมดที่มีเก็บค่า seed term เอาไว้ ในที่นี้เราใช้ The Web สำหรับการดึงหน้าเวปไซต์ที่ต้องการ และหน้านำหน้าเวปไซต์ที่เราเก็บได้จาก Compiling corpus นำมาตัดเฉพาะ passages ที่เกี่ยวข้องกับ seed term โดยการผ่านกระบวนการทั้งหมด 2 กระบวนการดังนี้

1. Web page collection คือการเก็บรวบรวมหน้า web page โดยการเอา URL ที่ได้มาทำการเปิดหน้า web page เพื่อรวบรวมคำศัพท์ และทำ compiling corpus คือเอา web page ทั้งหมดที่ได้จากการเปิด URL มาเป็น corpus ของคำศัพท์
2. Sentence extraction คือการตัดเอาแต่ละประโยคที่เราต้องการ มาจาก corpus ที่เราเก็บไว้ จะได้ออกมาเป็นประโยคที่มีคำศัพท์และคำแปลอยู่ข้างใน

Automatic Term Recognition

Automatic term recognition เป็นกระบวนการการสร้างเงื่อนไขในการตัดเอาเฉพาะคำศัพท์และคำแปลที่เราต้องการออกมาจาก corpus สามารถทำได้โดยการสร้าง term list ในการดึงคำศัพท์ออกจากหน้า website แบบ automatic เริ่มจากการสร้าง term list เป็นเงื่อนไขในการดึงคำศัพท์จากแต่ละหน้าเวปไซต์และนำมาใช้ในการดึงคำศัพท์และคำแปล สิ่งที่ได้จากกระบวนการนี้จะได้เป็นชุดของคำที่มีคำศัพท์และคำแปลอยู่ข้างใน

Filtering

Filtering เป็นกระบวนการในการตัดคำที่ต้องการ ทั้งคำศัพท์และคำแปล เพื่อนำมาเก็บรวบรวมลงฐานข้อมูล โดยผลจากกระบวนการนี้คือการได้คำศัพท์มาเก็บลงฐานข้อมูล โดยการสร้าง

เงื่อนไขของการตัดคำศัพท์และนำมาใช้ในการตัดคำศัพท์และคำแปลแบบ automatic ซึ่งจะมีรูปแบบสำหรับแต่ละ Corpus แบ่งตาม web page ที่ได้รวบรวมมา เมื่อได้คำศัพท์และคำแปล ก็จะนำมาเก็บรวบรวมลงฐานข้อมูล เพื่อใช้ในการทำ Relation test ต่อไป

Relation test

Relation test เป็นกระบวนการในการหาความสัมพันธ์ของ หมวดหมู่ที่เราต้องการกับคำศัพท์ในฐานข้อมูลของเรา การแบ่งหมวดหมู่ของคำศัพท์ในที่นี้เมื่อเราได้คำศัพท์และคำแปลมาเก็บไว้ในฐานข้อมูลแล้ว เราจะทำการแบ่งคำศัพท์ออกมาเป็นหมวดหมู่ และสร้างฐานข้อมูลใหม่โดยการเรียงคำศัพท์ตามหมวดหมู่ และระบุค่าความน่าจะเป็นของการเกี่ยวข้องกันของคำศัพท์แต่ละตัวกับหมวดหมู่หลักและหมวดหมู่ย่อย

การหาความสัมพันธ์ของหมวดหมู่ที่เรากำหนดไว้และคำศัพท์ในฐานข้อมูล เราจะตรวจสอบว่าทั้ง 2 คำนี้เกี่ยวข้องกันจริงไหม โดยการเกี่ยวข้องกัน จะต้องประกอบด้วย 2 อย่าง

1. หมวดหมู่ต้องมีความหมายแคบหรือกว้างกว่าคำศัพท์
2. ความสัมพันธ์ของหมวดหมู่กับคำศัพท์ต้องสูงมากพอเกินกว่าค่าหนึ่ง โดยที่ Term ทั้งหมดที่ได้จะถูกแบ่งออกเป็น 4 ชนิด

Type 0: หมวดหมู่และคำศัพท์เป็นคำๆเดียวกัน

Type 1: คำศัพท์มีความเกี่ยวข้องกับหมวดหมู่หมวดหมู่

Type 2: คำศัพท์อาจจะเกี่ยวข้องหรือไม่เกี่ยวข้องกับหมวดหมู่

Type 3: คำศัพท์และหมวดหมู่ไม่มีความเกี่ยวข้องกันเลย

เหตุผลที่เราต้องกำหนดชนิดของความสัมพันธ์เพราะว่าการหาว่า Relation degree มีมากแค่ไหน ในที่นี้เราจะใช้วิธีการ search เพื่อการหาความน่าจะเป็น และค่าความน่าจะเป็นที่เราจะได้สามารถนำมาจำแนกความสัมพันธ์ตามหมวดหมู่ทั้ง 4 เพื่อดูประสิทธิภาพของวิธีการทำ Relation test นี้ด้วย โดยวิธีการหาค่าความน่าจะเป็นสามารถหาได้ด้วยวิธีดังต่อไปนี้

เราสามารถหาค่า Probability จากการทำ Relation test จากสูตรต่อไปนี้

$$P(s|x) = H(s\hat{}x)/H(x) \qquad \textbf{(2.1)}$$

โดยที่

s คือ หมวดหมู่ย่อย

x คือ คำศัพท์ภาษาอังกฤษ

P(s|x) คือ ค่าความน่าจะเป็นที่ s และ x จะมีความเกี่ยวข้องกัน

H(x) คือ จำนวน hits ที่มี x ปรากฏอยู่

H(s^x) คือ จำนวน hits ที่มีทั้ง s และ x ปรากฏอยู่

สมมุติฐานของการกำหนดค่าความน่าจะเป็นเพื่อให้เกี่ยวข้องกับหมวดหมู่คือ ถ้า P(s|x)>=0.05 หรือ P(x|s)>=0.05 แล้วคำศัพท์จะเกี่ยวข้องกับหมวดหมู่นี้อย่างใกล้ชิด สมมุติฐานเพิ่มเติมอื่นๆจะได้หลังจากการทำโครงงานนี้เสร็จสิ้นและสรุปผลของการทำ relation test นี้

บทที่3

วิธีการดำเนินการศึกษา

ในการศึกษาโครงงานนี้ เป็นการนำทฎษฎีการทำ Automatic Collection of Vocabulary and Related Term from the Web มาพัฒนาเป็นโปรแกรมคอมพิวเตอร์ เพื่อประยุกต์ใช้ในการทำ online dictionary รูปแบบใหม่ที่สามารถหาความเกี่ยวเนื่องกันของคำศัพท์ได้ด้วย โดยจะแบ่งการดำเนินการศึกษาออกเป็น 4 ส่วน ดังนี้

- โครงสร้างการทำงานตามทฤษฎี
- ขั้นตอนการดำเนินงาน
- การออกแบบโปรแกรมการทำงาน
- อัลกอลิทึมในการทำงานของโปรแกรมตามทฤษฎี

3.1 โครงสร้างการทำงานตามทฎษฎี

โครงสร้างการทำงานตามทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web ที่ใช้ในโครงงานนี้ จะประกอบด้วยโครงสร้างการทำงานที่แบ่งเป็นจำนวนขั้นตอนการทำงานทั้งหมด 4 ขั้นตอน ดังนี้

1. การทำ Compiling Corpus
2. การทำ Automatic Term Recognition
3. การทำ Filtering
4. การทำ Relation test

3.2 ขั้นตอนการดำเนินงาน

โดยเริ่มจากกระบวนการของการวิเคราะห์การค้นหาคำศัพท์ จนกระทั่งถึงการหาความสัมพันธ์ของคำศัพท์และจัดเก็บลงฐานข้อมูล ซึ่งประกอบการทำงานได้เป็น 9 ขั้นตอนหลักๆ ดังนี้

1. กำหนดวัตถุประสงค์ (Objectives Determination) มีการศึกษาความต้องการในการวิเคราะห์ คือกำหนดปัญหาและวัตถุประสงค์การทำให้ชัดเจน ต้องเข้าใจถึงปัญหาและความต้องการของผู้ใช้งาน Online Dictionary ประกอบด้วยการวิเคราะห์เวปไซต์เบื้องต้นว่ามีข้อมูลอะไรบ้างบนหน้าเวปไซต์และเราต้องการคำศัพท์อะไรบ้างจากหน้าเวปไซต์และเป็นคำศัพท์ที่จะนำมาใช้งานเมื่อใด
2. การจัดเตรียมรูปแบบของคำศัพท์ (Vocabulary preparation)ที่เราจะทำการค้นหาจากหน้าเวปไซต์ว่าต้องการคำศัพท์ประเภทใดบ้าง ต้องการคำแปลแบบใดและจะสามารถค้นหาได้

จากเวปไซต์ใดบ้าง มีเวปไซต์ใดที่เหมาะสมและอยู่ในประเด็นที่ต้องการประกอบด้วย 2 ขั้นตอนย่อยคือ

2.1 การคัดเลือกคำศัพท์ที่ต้องการ (Vocabulary Selection) เป็นการกำหนดรูปแบบของคำศัพท์ว่าจะเอาคำศัพท์ประเภทใดบ้าง เป็นการเริ่มต้นของการค้นหาคำศัพท์ การเลือกคำศัพท์ขึ้นอยู่กับจุดประสงค์ของการใช้งานของเรา จึงเลือกคำศัพท์ทุกรูปแบบที่เจอและมีความหมายที่ถูกต้องแม่นยำ คำศัพท์ที่จะนำมาเก็บคือคำศัพท์ที่มีความหมายประกอบด้วยเท่านั้น

2.2 การตรวจสอบความถูกต้องของคำศัพท์และความหมายของคำศัพท์ที่ได้ (Vocabulary Preprocess) เป็นวิธีที่นำไว้ตรวจสอบคุณภาพของคำศัพท์และความหมายของคำศัพท์ที่ได้โดยใช้วิธีการหลากหลายอาทิเช่น การนำคำศัพท์ไปค้นหาความหมายอีกครั้งว่าตรงกันกับความหมายที่ได้ในตอนแรกหรือไม่

3. การจัดเตรียมรูปแบบของการค้นหาคำศัพท์ (Vocabulary Platform) การวิเคราะห์วิธีการเลือกคำศัพท์ที่ เหมาะสมกับแต่ ละเวปไซต์ขั้นตอนนี้มีความสัมพันธ์กับการวิเคราะห์ข้อมูลในขั้นตอนที่ ผ่านมา โดยอาจจะไปย้อนทำในขั้นตอนที่ 2 ใหม่ เนื่องจากถ้าวิธีในการค้นหาคำศัพท์ได้ผลลัพท์ไม่ตรงกับที่ต้องการก็ย้อนกลับไปทำใหม่จนกว่าจะได้คำศัพท์และความหมายที่ถูกต้อง
4. วิเคราะห์คำศัพท์ที่ได้ (Analysis of Result) การตรวจสอบคำศัพท์และคำแปลว่ามีความถูกต้องแม่นยำแค่ไหน เพื่อดูว่ารูปแบบของการค้นหาคำศัพท์ที่ได้มีประสิทธิภาพมากเพียงใด และประเมินผลลัพท์ที่ได้จากขั้นตอนที่ 3
5. นำคำศัพท์ที่ได้มารวบรวมและจัดเก็บเข้าสู่ฐานข้อมูลโดยเรียงตามตัวอักษร A-Z
6. จัดทำหมวดหมู่ของคำศัพท์ โดยในรายงานชุดนี้จะมีการจัดทำหมวดหมู่ของคำศัพท์แบ่งเป็น 6 หมวดหมู่ใหญ่
7. หาความสัมพันธ์ของคำศัพท์และหมวดหมู่ที่จัดทำขึ้น ว่าคำศัพท์มีความสัมพันธ์ลับหมวดหมู่ใด โดยมีความน่าจะเป็นของการอยู่ในหมวดหมู่นี้แค่ไหน เพื่อใช้ในการทำเป็นฐานความรู้ต่อไป
8. จัดเก็บคำศัพท์และคำแปลโดยเรียงตามหมวดหมู่ลงในฐานข้อมูล พร้อมเก็บค่าความน่าจะเป็นของความสัมพันธ์ของคำศัพท์ในแต่ละหมวดหมู่
9. วิเคราะห์ค่าความน่าจะเป็นได้ที่ได้ว่ามีความถูกต้องแม่นยำแค่ไหน เพื่อให้ในการปรับปรุงต่อไปในอนาคต
10. จัดทำเวปไซต์ เพื่อให้บริการ Online Dictionary โดยสามารถค้นหาคำศัพท์แบบมาตราฐานและการค้นหาคำศัพท์ที่ขึ้นต้นด้วยตัวอักษรที่ต้องการ, การค้นหาคำศัพท์ที่ประกอบด้วย

ตัวอักษรที่ต้องการ, การค้นหาคำศัพท์เรียงตามหมวดหมู่ , การค้นหาคำศัพท์ที่ต้องการ คำศัพท์ที่อยู่ในหมวดหมู่เดียวกัน

การทำงานในแต่ละขั้นตอนจะใช้ระยะเวลาแตกต่างกันออกไป โดยจะใช้เวลาในการทำขั้นตอนต่างๆดังนี้

-การ กำหนดวัตถุประสงค์ 5%

-การจัดเตรียมรูปแบบของคำศัพท์ 5%

-การจัดเตรียมรูปแบบของการค้นหาคำศัพท์ 5%

- วิเคราะห์คำศัพท์ที่ได้ 5%

-นำคำศัพท์ที่ได้มารวบรวมและจัดเก็บเข้าสู่ฐานข้อมูล 20%

-จัดทำหมวดหมู่ของคำศัพท์ 10%

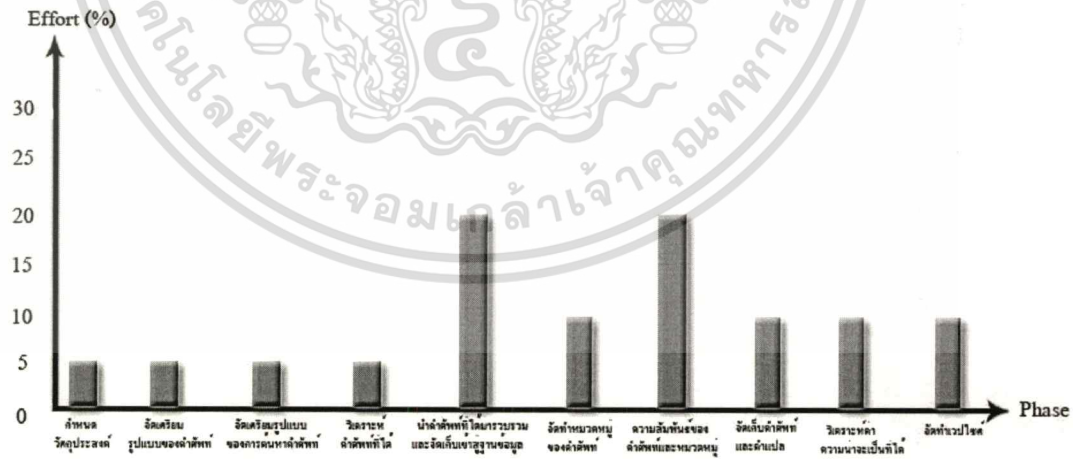
-หาความสัมพันธ์ของคำศัพท์และหมวดหมู่ที่จัดทำขึ้น 20%

-จัดเก็บคำศัพท์และคำแปลโดยเรียงตามหมวดหมู่ลงในฐานข้อมูล 10%

-วิเคราะห์ค่าความน่าจะเป็นได้ที่ได้ว่ามีความถูกต้องแม่นยำแค่ไหน 10%

-จัดทำเวปไซต์ เพื่อให้บริการ Online Dictionary 10%

โดยสามารถแสดงเป็นกราฟดังภาพ 3.1 ดังนี้



รูปที่3.1 เปอร์เซนต์ของเวลาที่ใช้แต่ละขั้นตอน

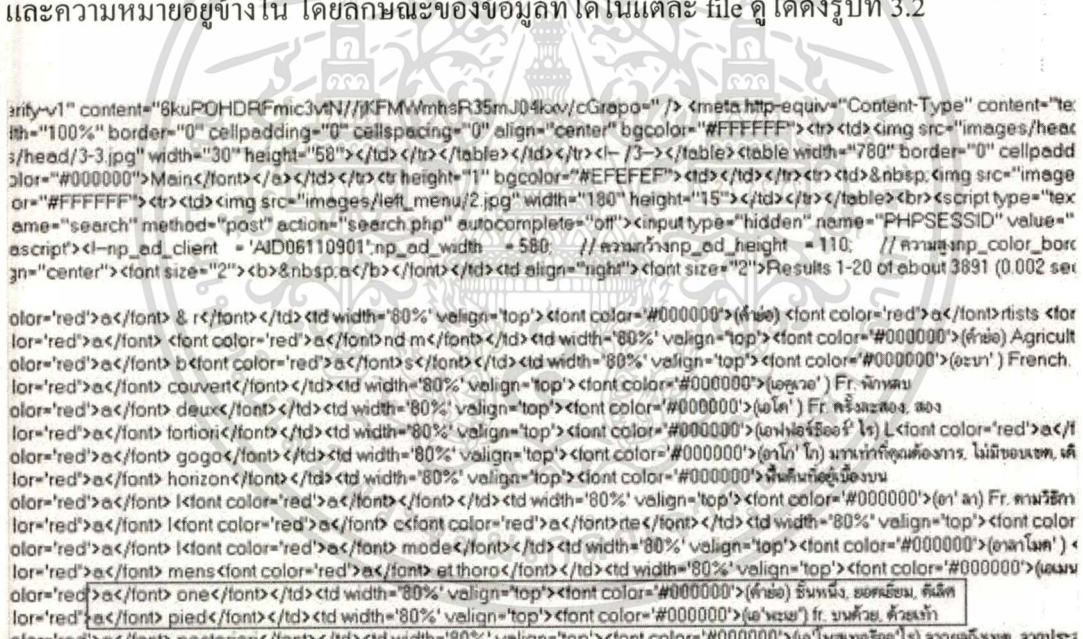
การศึกษาโครงงานนี้จะดำเนินการศึกษา Automatic Collection of Vocabulary and Related Term from the Web และทำการพัฒนาระบบสร้างเป็นโปรแกรมคอมพิวเตอร์เพื่อประยุกต์ใช้ในการสร้าง online dictionary ที่สามารถค้นหาคำศัพท์ที่มีความเกี่ยวข้องกัน และคำศัพท์ที่เกี่ยวข้องกับ

หมวดหมู่ได้ โดยใช้โปรแกรมไมโครซอฟท์วิชวลสตูดิโอดอทเน็ต โดยมีขั้นตอนการดำเนินงานดังนี้

1. ขั้นตอนการจัดเตรียมข้อมูล โดยใช้เวปไซต์ Online dictionary ที่มีอยู่ทั่วไป โดยคัดแยกมา 3 เวปไซต์ซึ่งมีการรวบรวมคำศัพท์ที่แตกต่างกัน โดยเวปไซต์ที่ได้ทำการคัดเลือกมาสำหรับโครงงานนี้คือ http://www.siamdic.com/, http://truehits.net/dic.php /, http://www.pixiart.com/dict/ มาทำการค้นหาคำศัพท์ และคำแปล เพื่อนำมาจัดเก็บลงฐานข้อมูลที่เตรียมไว้

2. ขั้นตอนการรวบรวม source จากเวปไซต์ทั้ง 3 โดยใช้วิธีการทำ compiling corpus โดยที่วิธีการทำ compiling corpus จะประกอบด้วย 2 ขั้นตอน ได้แก่ การทำ Web page collection และการทำ Sentence extraction

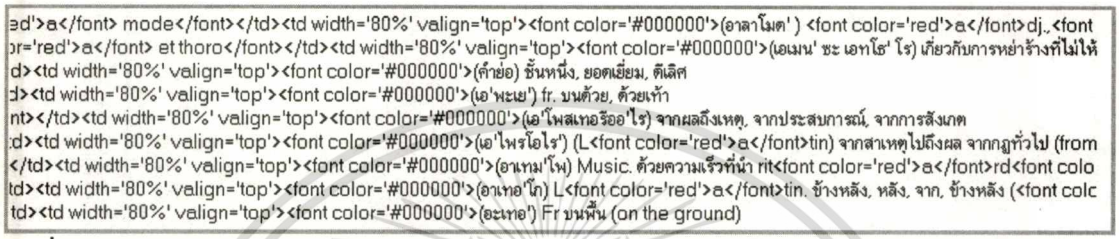
- การทำ Web page collection คือการทำการรวบรวม source ของเวปไซต์ทั้ง 3 เวปไซต์ นำมาทำการเก็บรวบรวมโดยการใช้การทำงาน 3 ขั้นตอน ได้แก่ การทำ CheckWebResponse การกำหนดค่า WebClient และการทำการ DownloadFile จะได้ corpus ของ file จำนวนมากที่มีคำศัพท์และความหมายอยู่ข้างใน โดยลักษณะของข้อมูลที่ได้ในแต่ละ file ดูได้ดังรูปที่ 3.2



รูปที่3.2 ลักษณะของข้อมูลของแต่ละ file ใน corpus ที่รวบรวมได้จากการทำ Web page collection

จากภาพจะเห็นได้ว่าข้อมูลใน Corpus ที่เรารวบรวมได้ มีคำศัพท์และคำแปลอยู่ข้างใน แต่เนื่องจากว่ารูปแบบของ source ที่ได้นี้ยังมีส่วนที่เราไม่ได้นำมาใช้งานเป็นจำนวนมาก เราจึงเป็นต้องตัดเอาเฉพาะประโยคที่ต้องการใช้งานด้วยการทำขั้นตอนต่อไปคือการทำ Sentence extraction

-การทำ Sentence extraction คือการทำการตัดเฉพาะประโยคที่เราต้องการจะใช้งานออกมาจาก file แต่ละ file ใน corpus โดยนำมาเก็บรวบรวมเป็น corpus ตัวใหม่ที่มีขนาดเล็กลงเนื่องจากเก็บรวบรวมเฉพาะประโยคที่ต้องการใช้งานจริงๆ เราจะรวบรวมเอาเฉพาะประโยคที่มีคำศัพท์และคำแปลอยู่ข้างใน โดยจะรวบรวมทุกประโยคที่เรามีใน corpus เพื่อให้ได้คำศัพท์จำนวนมากที่สุดและครอบคลุมต่อการดำเนินการทดลองในโครงงานนี้ และโครงงานถัดๆไป ผลที่ได้จากการทำ Sentence extraction คือ ประโยคดังรูปที่ 3.3

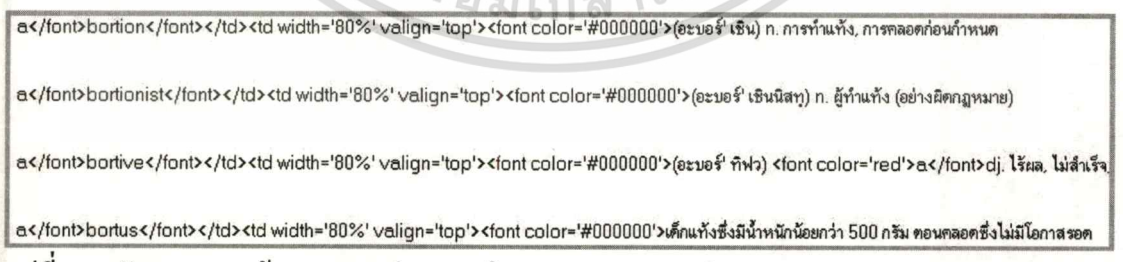


รูปที่3.3 ลักษณะของข้อมูลของแต่ละ file ใน corpus ที่รวบรวมได้จากการทำ Sentence extraction

3. ขั้นตอนการทำ Automatic term recognition โดยการนำ corpus ที่ได้จากการทำ Sentence extraction มาทำการทำงาน 2 ขั้นตอน ได้แก่การทำ Generate term format และการทำ Selection

- การทำ Generate term format คือกาทำการสร้างรูปแบบของ term ที่เราต้องการจะดึงออกมาจาก source file และจัดเก็บรูปแบบไว้สำหรับการทำการ Selection เพื่อให้ได้ term ที่ต้องการเท่านั้น ซึ่งในที่นี้คือ term ที่เริ่มด้วยคำศัพท์และจบด้วยคำแปลของคำศัพท์แต่ละคำ

- การทำ Selection คือการนำ format ที่ได้จากการทำ Generate term format มาทำการตัดเอาเฉพาะ term ที่เราต้องการออกมา และจัดเก็บเป็น corpus ที่มีขนาดเล็กลงและตรงตามการใช้งานมากขึ้น ซึ่งจะได้ผลลัพท์ดังรูปที่ 3.4



รูปที่ 3.4 ลักษณะของข้อมูลของแต่ละ file ใน corpus จากการทำ Automatic term recognition

4. การทำ Filtering คือขั้นตอนการทำการตัดคำศัพท์เฉพาะคำศัพท์และคำแปลที่เราต้องการออกมาจาก corpus ที่เป็นผลลัพท์จากการทำ Automatic term recognition ซึ่งเราจะนำ file ที่อยู่ใน corpus ทั้งหมด มาทำการตัดเฉพาะ term ที่เราต้องการเพื่อนำมาเก็บรวบรวม โดยจะใช้วิธีการตัดคำศัพท์และคำแปลแบบ automatic เพื่อความสะดวกและรวดเร็ว ในโครงงานการนี้ การทำการตัดคำศัพท์

และคำแปลแบบ automatic เพื่อความสะดวกและรวดเร็ว ในโครงงานการนี้ การทำการตัดคำศัพท์โดยการพัฒนาโปรแกรมขึ้นมา และมีการแยกโปรแกรมไว้เป็นส่วนๆเพื่อทำขั้นตอนต่างๆตามทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web

5. การออกแบบฐานข้อมูล และ จัดเก็บคำศัพท์และคำแปลที่ได้ลงฐานข้อมูล ในโครงงานนี้ เราจะมีการออกแบบฐานข้อมูล โดยจัดเก็บทั้งหมด 5 ตาราง ได้แก่ ตาราง Dictionary, ตาราง Directory, ตาราง Subdirectory, ตาราง Relation, ตาราง Link โดยที่แต่ละตารางมีการจัดเก็บข้อมูลดังต่อไปนี้

- ตาราง Dictionary จัดเก็บข้อมูลทั้งหมด 4 ข้อมูลได้แก่ ID สำหรับเก็บรหัสของคำศัพท์, English สำหรับเก็บคำศัพท์, type สำหรับเก็บชนิดของคำศัพท์ และ Thai สำหรับเก็บคำแปลของคำศัพท์

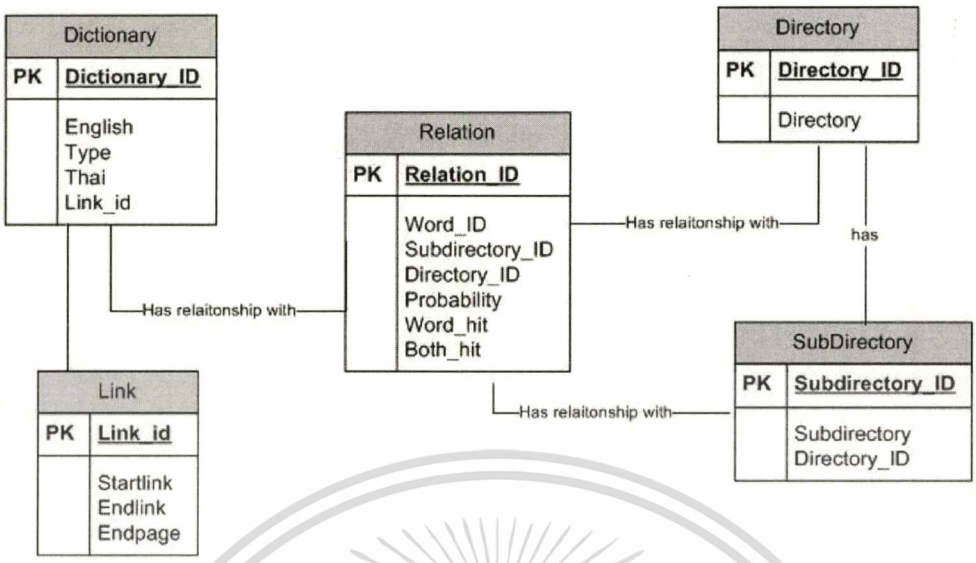
- ตาราง Directory จัดเก็บข้อมูลทั้งหมด 2 ข้อมูลได้แก่ Directory_ID สำหรับเก็บรหัสของหมวดหมู่, Directory สำหรับเก็บชื่อหมวดหมู่

- ตาราง Subdirectory จัดเก็บข้อมูลทั้งหมด 3 ข้อมูลได้แก่ Subdirectory_ID สำหรับเก็บรหัสของหมวดหมู่ย่อย, Subdirectory สำหรับเก็บชื่อหมวดหมู่ย่อย, Directory_ID สำหรับเก็บรหัสหมวดหมู่

- ตาราง Relation จัดเก็บข้อมูลทั้งหมด 7 ข้อมูลได้แก่ Relation_ID สำหรับเก็บรหัสของความสัมพันธ์, Word_ID สำหรับเก็บรหัสของคำศัพท์ Subdirectory สำหรับเก็บชื่อหมวดหมู่ย่อย, Directory_ID สำหรับเก็บรหัสหมวดหมู่, Subdirectory_ID สำหรับเก็บรหัสของหมวดหมู่ย่อย, Directory_ID สำหรับเก็บรหัสของหมวดหมู่, Word_hit เก็บจำนวนที่เจอคำศัพท์, Both_hit เก็บจำนวนที่เจอทั้งคำศัพท์และหมวดหมู่, Probability เก็บค่าความน่าจะเป็นที่คำศัพท์และหมวดหมู่จะเกี่ยวข้องกัน

- ตาราง Link จัดเก็บข้อมูลทั้งหมด 3 ข้อมูลได้แก่ Start_link , End_link และ End_page เพื่อเก็บ URL ที่นำมาเปิดหา source file เพื่อใช้ในการค้นหาคำศัพท์ และความหมายของคำศัพท์

รูปแบบของฐานข้อมูล (Database diagram) สามารถดังแสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 แสดงรูปแบบของฐานข้อมูล (Database diagram)

6. การคัดเลือกหมวดหมู่และหมวดหมู่ย่อย มาใช้งานในการจัดหมวดหมู่ของคำศัพท์ โดยในที่นี้ เราจะกำหนดให้มีหมวดหมู่ทั้งหมด 5 หมวดหมู่ ดังนี้

- Business & Finance คือ หมวดหมู่ที่เกี่ยวข้องกับ ธุรกิจและการเงิน ที่จะประกอบไปด้วย 20 หมวดย่อย

- Computers & Internet คือ หมวดหมู่ที่เกี่ยวข้องกับ คอมพิวเตอร์และอินเตอร์เน็ต ที่จะประกอบไปด้วย 15 หมวดย่อย

- Schools & Education คือ หมวดหมู่ที่เกี่ยวข้องกับโรงเรียนและการศึกษา ที่จะประกอบไปด้วย 24 หมวดย่อย

- Entertainment & Arts คือ หมวดหมู่ที่เกี่ยวข้องกับบันเทิงและศิลปะ ที่จะประกอบไปด้วย 28 หมวดย่อย

- Health & Wellness คือ หมวดหมู่ที่เกี่ยวข้องกับ ธุรกิจและการเงิน ที่จะประกอบไปด้วย 19 หมวดย่อย

- Recreation & Sport คือ หมวดหมู่ที่เกี่ยวข้องกับ การพักผ่อนหย่อนใจและกีฬา ที่จะประกอบไปด้วย 9 หมวดย่อย

ในที่นี้เราจะระบุหมวดหมู่ในการทำในโครงงานนี้ออกเป็น 6 หมวดหมู่ใหญ่ ดังต่อไปนี้

Business & Finance - Business Schools	Entertainment & Arts - Actors

- Companies
- Consumer
- Employment
- Work
- Financial Professionals
- Free Stuff
- Home Business
- Industry Associations
- Investments
- Labor
- Marketing
- Advertising
- News
- Media
- Public Relations
- Real Estate
- Small Business
- Trade
- Transportation

Computers & Internet

- Communications
- Networking
- Cyber culture
- Data Formats
- Desktop Publishing
- Education
- Hardware
- Internet
- Multimedia
- Programming Languages
- Security
- Software
- Technical Support
- Collecting
- Desktop Customization

Schools & Education

- Bilingual
- Subject
- Classmates
- Colleges
- Universities
- Courses
- Distance Learning
- Exchange Students

- Actresses
- Amusement
- Theme Parks
- Audio
- Visual Equipment
- Celebrities
- Comics
- Animation
- Contests
- Sweepstakes
- Cool Clubs
- Fashion
- Fine Arts
- Genres
- Humanities
- Humor
- Magic
- Movies
- Official Groups
- Clubs
- Performing Arts
- Quotations
- Radio
- Television
- Trivia
- Dancing
- Music

Health & Wellness

- Advice
- Alternative Medicine
- Beauty
- Children's
- Drugs
- First Aid
- Fitness
- Health Care
- Men's
- Professional
- Reproductive
- Seniors
- Stress Management
- Support
- Teens
- Women's
- Pet Health
- Nutrition

· Fund Raising	· Medications
· Graduating Classes	
· Homework Help	Recreation & Sports
· Issues	
· Preschools	· Automotive
· Religious Schools	· Aviation
· Special Education	· Gambling
· Student Organizations	· Outdoors
· Study Abroad	· Sports
· Teaching	· Toys
· Research	· Travel
· Testing	· Disabled
· Theory	· Games
· Methods	
· University Life	
· Dental Schools	

โดยจะนำมาจัดเก็บในฐานข้อมูลโดยแบ่งเป็น 2 ตาราง ได้แก่ตาราง Directory และ ตาราง Subdirectory โดยที่จะมีการให้ค่า Directory_id ไว้ในตาราง Subdirectory เพื่อบ่งบอกว่าเป็นหมวดหมู่ย่อยของหมวดหมู่ใด ดังตัวอย่างรูป 3.6

Directory : Table

Directory_id	Directory
1	Business & Finance
2	Computers & Internet
3	Schools & Education
4	Entertainment & Arts
5	Health & Wellness
6	Recreation & Sports

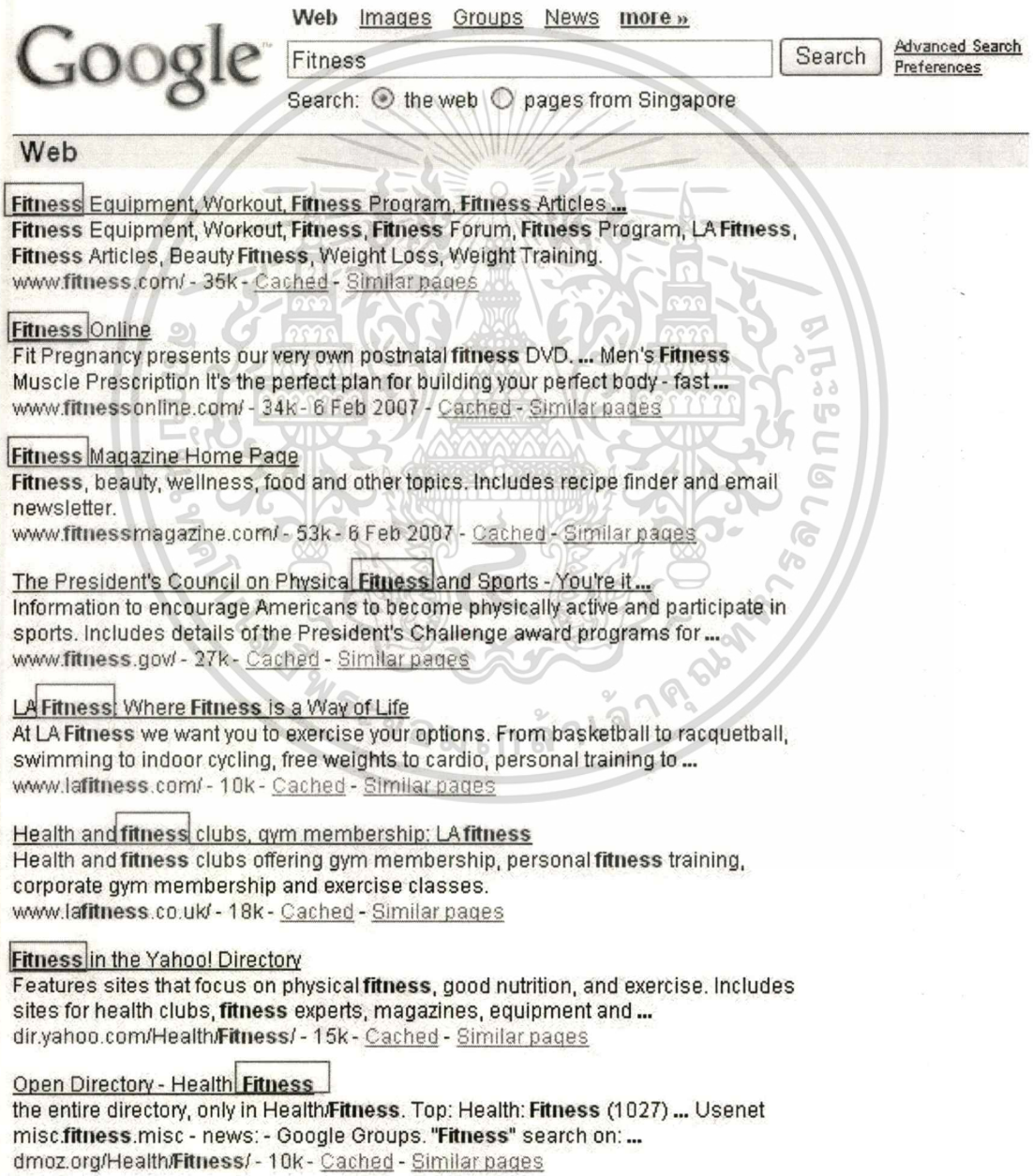
Subdirectory : Table

Subdirectory_id	Subdirectory	Directory_id
1	Communications	2
2	Cyber culture	2
3	Data Formats	2
4	Desktop Publishing	2
5	Education	2
6	Hardware	2
7	Internet	2
8	Multimedia	2
9	Programming Languages	2
10	Security	2
11	Software	2
12	Technical Support	2
13	Collecting	2
14	Desktop Customization	2
15	Bilingual	3
16	By Subject	3
17	Classmates	3
18	Colleges	3
19	Courses	3
20	Distance Learning	3
21	Exchange Students	3
22	Fund Raising	3
23	Graduating Classes	3

รูปที่3.6 แสดงวิธีการเก็บหมวดหมู่และหมวดหมู่ย่อยลงในฐานข้อมูล

7. การทำ Relation test และการเก็บผลลัพท์ที่ได้ เราจะทำการหาค่าทั้งหมด 3 ค่า คือค่า Word hits และค่า both hits เพื่อนำมาหาค่า Probability ที่เราต้องการ โดยเราสามารถคำนวณหาแต่ละค่าได้ด้วยวิธีการดังต่อไปนี้

- ค่า Word hits คือการนำคำศัพท์ที่เราต้องการ ไปหาค่า hitsโดยใช้ search engine hits โดยที่เราจะนำคำศัพท์ไป search ในเวป search engine และนับจำนวน hits ที่เจอโดยเลือกเฉพาะ 30 ลิงค์แรกที่เจอ และนับเฉพาะ ลิงค์ที่มีคำศัพท์ที่เราต้องการปรากฏอยู่ในส่วนของลิงค์เท่านั้น เราจะไม่นับถ้าไม่เจอในส่วนของลิงค์ ถึงแม้ว่าจะเจอในส่วนของ content ก็ตาม ซึ่งในโครงงานนี้ เราจะพัฒนาโปรแกรมอันโนมัติขึ้นมาทำการคำนวณค่า Word hits ตามวิธีดังกล่าว ดังตัวอย่างรูป 3.7



รูปที่3.7 แสดงวิธีการนับจำนวน word hits จาก search engine

- ค่า Both hits คือการนำคำศัพท์ และหมวดหมู่ที่เราต้องการ ไปหาค่า hitsโดยใช้ search engine hits โดยที่เราจะนำคำศัพท์และหมวดหมู่ไป search ในเวป search engine และนับจำนวน hits ที่เจอโดยเลือกเฉพาะ 30 ลิงค์แรกที่เจอ และนับเฉพาะ ลิงค์ที่มีคำศัพท์และหมวดหมู่ที่เราต้องการปรากฏอยู่ในส่วนของลิงค์เท่านั้น เราจะไม่นับถ้าไม่เจอในส่วนของลิงค์ ถึงแม้ว่าจะเจอในส่วนของ content ก็ตาม โดยที่รูปแบบของการค้นหาคือ “คำศัพท์” “หมวดหมู่” ซึ่งในโครงงานนี้ เราจะพัฒนาโปรแกรมอันโนมัติขึ้นมาทำการคำนวณค่า Word hits ตามวิธีดังกล่าว ดังตัวอย่างรูป3.8

รูปที่3.8 แสดงวิธีการนับจำนวน both hits จาก search engine

-ค่า Probability คือค่าความน่าจะเป็นที่คำศพท์และหมวดหมู่จะมีความเกี่ยวข้องกันโดยที่เราจะแยกประเภทของความสัมพันธ์ออกเป็น 4 ประเภท โดยเราจะกำหนดสมมุติฐานของการทำลองในโครงงานนี้ โดยแบ่งตามค่า probability ดังนี้

Probability =1

อยู่ใน Type 0: หมวดหมู่และคำศัพท์เป็นคำๆเดียวกัน หรือมีความเกี่ยวข้องกันมาก

0.8=< Probability <1

อยู่ใน Type 1: คำศัพท์มีความเกี่ยวข้องกับหมวดหมู่หมวดหมู่

0.5=<Probability <0.8

อยู่ใน Type 2: คำศัพท์อาจจะเกี่ยวข้องหรือไม่เกี่ยวข้องกับหมวดหมู่

Probability <0.5

อยู่ใน Type 3: คำศัพท์และหมวดหมู่ไม่มีความเกี่ยวข้องกันเลย

โดยที่เราจะแบ่งการพิจารณาออกจามชนิดของคำศัพท์ คือคำนาม คำกริยา และคำอื่นๆ เนื่องจากคำที่จะถูกต้องมากที่สุดในการนำมาทำ Relation test ตามสมมุติฐานของโครงงานนี้คือคำนาม ดังนั้นเราจะพิจารณาเฉพาะคำนามเป็นหลัก แต่จะมีคำชนิดอื่นๆในการวิเคราะห์ผลลัพท์ในโครงงานนี้ด้วย ซึ่งในส่วนของผลการทดลองครั้งนี้จะแบ่งผลการทำ Relation ออกเป็นคำนาม กับคำชนิดอื่นๆ เพื่อใช้แสดงในการค้นหาผ่านเวปไซต์ online dictionary และเพื่อใช้ในการวิเคราะห์ความถูกต้องของทฤษฎีการทำ Automatic Collection of Vocabulary and Related Term from the Web

เราสามารถหาค่า Probability จากการทำ Relation test จากสูตรต่อไปนี้

$$P(s|x) = H(s^x)/H(x) \quad (3.1)$$

โดยที่

s คือ หมวดหมู่ย่อย

x คือ คำศัพท์ภาษาอังกฤษ

P(s|x) คือ ค่าความน่าจะเป็นที่ s และ x จะมีความเกี่ยวข้องกัน

H(s) คือ จำนวน hits ที่มี s ปรากฏอยู่

H(x) คือ จำนวน hits ที่มี x ปรากฏอยู่

H(s^x) คือ จำนวน hits ที่มีทั้ง s และ x ปรากฏอยู่

โดยจะนำค่าทั้งหมดที่ได้ ไปเก็บลงฐานข้อมูลดังตัวอย่างรูป 3.9

Relation : Table

relation_id	Word_id	Subdirectory_id	Directory_id	probability	Word_hit	Both_hit
20780	20228	45	1	0.03703703703	27	1
20781	20228	46	1	0	27	0
20782	20228	47	1	0	27	0
20783	20228	48	1	0	27	0
20784	20228	49	1	0	27	0
20785	20228	50	1	0.07407407407	27	2
20786	20228	51	1	0.07407407407	27	2
20787	20228	52	1	0	27	0
20788	20229	36	1	0	27	0
20789	20229	37	1	0.59259259259	27	16
20790	20229	38	1	0.62962962962	27	17
20791	20229	39	1	0.03703703703	27	1
20792	20229	40	1	0.40740740740	27	11
20793	20229	41	1	0.03703703703	27	1
20794	20229	42	1	0.14814814814	27	4
20795	20229	43	1	0.03703703703	27	1
20796	20229	44	1	0.44444444444	27	12
20797	20229	45	1	0.59259259259	27	16
20798	20229	46	1	0	27	0
20799	20229	47	1	0.22222222222	27	6
20800	20229	48	1	0.29629629629	27	8
20801	20229	49	1	0.29629629629	27	8
20802	20229	50	1	0.11111111111	27	3
20803	20229	51	1	0.62962962962	27	17
20804	20229	52	1	0.59259259259	27	16

รูปที่3.9 แสดงการเก็บค่าต่างๆที่ได้จากการทำ Relation test

จากรูป 3.9 จะเห็นได้ว่าเราจะได้ค่า Probability ต่างๆกันสำหรับคำศัพท์และหมวดหมู่ต่างๆ ซึ่งจะทำค่าเหล่านี้ไปวิเคราะห์ในส่วนของผลการทำลอง เพื่อใช้ในการสรุปสมมุติฐานของค่า Probability สำหรับพัฒนาการทำ Relation test ต่อไปในอนาคต

7. การวิเคราะห์ผลลัพท์ที่ได้จากการทำ Relation test เราจะนำฐานข้อมูลที่ได้มาทำการวิเคราะห์ผลด้วยการวิเคราะห์แบบ manual check ว่าคำศัพท์และหมวดหมู่ที่ได้ มีผลตามสมมุติฐานที่เราตั้งไว้หรือไม่ และควรปรับสมมุติฐานอย่างไร เพื่อให้เหมาะสมและได้ผลลัพท์ที่ถูกต้องที่สุด พร้อมทั้งยังหาค่าของความถูกต้องของวิธีการทำ Relation test นี้ว่ามีความถูกต้องมากน้อยเพียงใดโดยการนำค่า Probability ที่ได้จากแต่ละแถว มาพิจารณาว่าคำศัพท์ และ หมวดหมู่ย่อย มีความสัมพันธ์กันอย่างไรในแต่ละช่วงค่าของ Probability ที่ได้

8. การสร้าง Online dictionary เพื่อนำฐานข้อมูลที่ได้มาใช้ให้เกิดประโยชน์สูงสุด โดยจะจัดทำ online dictionary ที่สามารถค้นหาคำศัพท์ได้ 3 ลักษณะ ได้แก่

- แบบมาตราฐาน ได้แก่การค้นหาจากส่วนเริ่มต้นของคำ, การค้นหาจากส่วนประกอบของคำ. การค้นหาจากส่วนท้ายสุดของคำ, ค้นหาเฉพาะเจาะจงคำนั้นๆเท่านั้น

- แบบการค้นหาคำศัพท์ที่ relation กัน โดยเราจะค้นหาคำศัพท์ในหมวดหมู่เดียวกันขึ้นมาแสดงทั้งหมด เพื่อช่วยให้ผู้ใช้งานนำคำศัพท์เหล่านี้ไปใช้งานในการเขียนหรือการอ่านได้ง่ายดายยิ่งขึ้น

- แบบการค้นหาคำศัพท์ที่เกี่ยวข้องกับหมวดหมู่ โดยเลือกหมวดหมู่ที่มีจัดเตรียมไว้ให้เพื่อดูคำศัพท์ที่มีความเกี่ยวข้องกัน เพื่อใช้ประโยชน์ในด้านต่างๆ

ดังแสดงขั้นตอนการดำเนินงานของโปรแกรมได้ดังตัวอย่างรูป 3.10

รูปที่3.10 แสดงขั้นตอนการดำเนินงาน

3.3 การพัฒนาโครงงานตามลำดับขั้น

การพัฒนาโครงงานนี้ได้แบ่งการทำงานเป็น 3 ลำดับขั้น ซึ่งพัฒนาตามผลการทดลองของโครงงานเพื่อให้ผลลัพธ์ที่มีความถูกต้องสูงขึ้น

- การพัฒนาโครงงานในลำดับที่ 1 คือการพัฒนาโครงงานตามทฤษฎี Automatic collection of related terms from the web โดยมี algorithm การทำงานดังนี้

1. คัดเลือก online dictionary สำหรับการค้นหาคำศัพท์และคำแปลโดยการทำแบบ manual โดยการเลือกจาเวปไซต์ที่มีคำศัพท์ปริมาณมากที่สุด และมีการแสดงคำศัพท์และคำแปลใน source ของหน้า website ที่สามารถเรียกขึ้นมาดูได้ และมีรูปแบบการแสดงผลลัพท์ที่แน่นอน

2. รวบรวมคำศัพท์และเก็บรวบรวมเป็น corpus โดยใช้ฟังก์ชั่นการทำงาน wc.DownloadFile(URL)

3. ตัดประโยคที่มีคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้ โดยพิจารณาจากsourceในcorpus ที่เรารวบรวมได้ดังตัวอย่าง

<table width='100%' border='0' cellpadding='5' cellspacing='0' align='center' bgcolor='#EEEBFA'>
<tr><td width='20%' valign='top'>   a gogo</td>
<td width='80%' valign='top'>(อาโก' โก) มากเท่าที่คุณต้องการ, ไม่มีขอบเขต, เดินเพ่นพ่าน, จังหวะเต้นรำอะโกโก้, เวทีเต้นรำจังหวะอะโกโก. -S... a Go go, a go-go, au gogo </td></tr></table>

จาก Code ของ website นี้ เราจะทำการตัดตั้งแต่ <table width='100%' border='0' cellpadding='5' cellspacing='0' align='center' bgcolor='#EEEBFA'>
<tr><td width='20%' valign='top'>   จนถึง
จะทำให้ได้ผลของการทำ Sentence extraction คือ a gogo</td>
<td width='80%' valign='top'>(อาโก' โก) มากเท่าที่คุณต้องการ, ไม่มีขอบเขต, เดินเพ่นพ่าน, จังหวะเต้นรำอะโกโก้, เวทีเต้นรำจังหวะอะโกโก. -S...
และนำมาจัดเก็บลง corpus ใหม่ที่มีขนาดเล็กลง เพื่อพร้อมสำหรับการทำ Filtering ในขั้นตอนต่อไป

4. ตัดคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้เช่นจากตัวอย่างในข้อ 3 จาก corpus ที่มีข้อมูล a gogo</td><td width='80%' valign='top'>(อาโก' โก) มาก

เท่าที่คุณต้องการ, ไม่มีขอบเขต, เดินเพ่นพ่าน, จังหวะเต้นรำอะโกโก้, เวทีเต้นรำจังหวะอะโกโก. - S...

จะทำการตัดคำศัพท์โดยตัดคำที่อยู่ด้านหน้า มาเป็นคำศัพท์ และ หลังจาก <font color='#000000'> มาเป็นคำแปลของคำศัพท์แต่ละคำ

5. เก็บรวบรวมคำศัพท์และคำแปลลงฐานข้อมูล

6. เลือกหมวดหมู่หลักมา 6 หมวดหมู่ และหมวดหมู่ย่อยมาจำนวน 115 หมวดหมู่โดยนำมาจากหมวดหมู่ของ www.yahoo.com

7. ทำ Relation test ตามสูตร P(s|x) = H(s^x)/H(x) โดยใช้ค่า hits ที่ได้จากการค้นหาจาก www.google.com

8. เก็บผลลัพท์จากการทำ Relation test ของทุกๆคำศัพท์กับทุกๆหมวดหมู่ย่อยที่เรามี

9. วิเคราะห์ความถูกต้องของผลลัพท์โดยการทำ manual test โดยพิจารณาจากการใช้ search engine เข้าช่วยในการพิจารณาว่าค่า Relation ระหว่างผลลัพท์และหมวดหมู่มีความเกี่ยวข้องกันมากน้อยเพียงใด โดยนำคำศัพท์และหมวดหมู่มาทำการค้นหาใน www.google.com โดยใส่เงื่อนไขในการค้นหาคือ "หมวดหมู่"+"คำศัพท์" และดูว่าพบมากน้อยเพียงใด และรวมถึงการตรวจสอบจากการพิจารณาแบบ manual test

จากการทำงานตามวิธีการข้างต้นจะสามารถสรุปผลลัพท์ได้ดังนี้

1. ในเวลา 11 นาทีสามารถหาความสัมพันธ์ของคำศัพท์ได้ 1 คำ กับ หมวดหมู่ย่อย 115 หมวดหมู่

2. ผลลัพท์ที่ได้สำหรับคำศัพท์ชนิดต่างๆมีความแตกต่างกัน สำหรับคำศัพท์ที่เป็นคำนามจะมีความแม่นยำมากสุด คำสรรพนาม คำกริยา และคำชนิดอื่นๆจะมีความแม่นยำน้อยมาก

3. หมวดหมู่ย่อยจำนวนมากที่มีประโยชน์ต่อการใช้งานน้อย และไม่ชี้เฉพาะเจาะจงเพียงพอสำหรับการนำมาหาค่าความสัมพันธ์

4. การทำ Relation ทุกคำศัพท์ที่เก็บไว้กับทุกหมวดหมู่ใช้เวลามากและได้ผลลัพท์ที่มีความแม่นยำน้อย

- การพัฒนาโครงงานในลำดับที่ 2 คือการพัฒนาโครงงานจากการพัฒนาโครงงานในลำดับที่ 1 ให้มีความถูกต้องของผลลัพท์มากขึ้น โดยมีการทำงานดังนี้

1. คัดเลือก online dictionary สำหรับการค้นหาคำศัพท์และคำแปลโดยการทำแบบ manual โดยการเลือกจาเวปไซต์ที่มีคำศัพท์ปริมาณมากที่สุด และมีการแสดงคำศัพท์และคำแปลใน source ของหน้า website ที่สามารถเรียกขึ้นมาดูได้ และมีรูปแบบการแสดงผลลัพท์ที่แน่นอน

2. รวบรวมคำศัพท์และเก็บรวบรวมเป็น corpus โดยใช้ฟังก์ชั่นการทำงาน wc.DownloadFile(URL)

3. ตัดประโยคที่มีคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้ โดยพิจารณาจากsourceในcorpus ที่เรารวบรวมได้ดังเช่นเดียวกับการพัฒนาโครงงานในลำดับที่1

4. ตัดคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้ดังเช่นเดียวกับการพัฒนาโครงงานในลำดับที่1

5. เก็บรวบรวมคำศัพท์และคำแปลลงฐานข้อมูล

6. เลือกหมวดหมู่หลักมา 6 หมวดหมู่ และหมวดหมู่ย่อยมาจำนวน 50 หมวดหมู่จากเดิม 115 หมวดหมู่ โดยเลือกเฉพาะหมวดหมู่ที่มีความน่าสนใจและสามารถนำมาใช้ประโยชน์ได้

7. ทำการคัดกรองคำศัพท์เฉพาะคำนาม โดยดูจากคำแปลจะมีอักษร n. หรือ (n) เพื่อบอกว่าเป็นชนิดคำนาม โดยคัดเลือกมาทั้งหมด 1326 คำ โดยวิธี Random เพื่อให้ได้คำนามที่มีความหลากหลายและต่างชนิดกัน

8. ทำ Relation test ตามสูตร P(s|x) = H(s^x)/H(x) โดยใช้ค่า hits ที่ได้จากการค้นหาจาก www.google.com

9. เก็บผลลัพท์จากการทำ Relation test ของทุกๆคำศัพท์กับทุกๆหมวดหมู่ย่อยที่เรามี ได้ผลลัพท์ทั้งหมด 60000 Relation

10. วิเคราะห์ความถูกต้องของผลลัพท์โดยการทำ manual test โดยพิจารณาจากการใช้ search engine เข้าช่วยในการพิจารณาว่าค่า Relation ระหว่างผลลัพท์และหมวดหมู่มีความเกี่ยวข้องกันมากน้อยเพียงใด โดยนำคำศัพท์และหมวดหมู่มาทำการค้นหาใน www.google.com โดยใส่เงื่อนไขในการค้นหาคือ “หมวดหมู่”+”คำศัพท์” และดูว่าพบมากน้อยเพียงใด และรวมถึงการตรวจสอบจากการพิจารณาแบบ manual test

จากการทำงานตามวิธีการข้างต้นจะสามารถสรุปผลลัพท์ได้ดังนี้

1. ในเวลา 3 นาทีสามารถหาความสัมพันธ์ของคำศัพท์ได้ 1 คำ กับ หมวดหมู่ย่อย 50 หมวดหมู่

2. ผลลัพท์ที่ได้สำหรับคำศัพท์ชนิดคำนามมีความถูกต้องแม่นยำมากกว่าการพัฒนาในลำดับขั้นตอนแรก แต่ยังไม่ได้ผลลัพท์ที่น่าพอใจ

3. การคัดเลือกหมวดหมู่ย่อยเพื่อนำมาใช้งานทำให้ช่วยประหยัดเวลาและได้ผลลัพท์ที่แม่นยำขึ้น

4. การพัฒนาโครงงานในลำดับที่2 ได้ผลลัพท์ที่แม่นยำกว่าการพัฒนาโครงงานในลำดับที่1 แต่ยังไม่มีความแม่นยำเท่าที่ควร โดยความแม่นยำที่ควรจะได้คือ 70% ขึ้นไป

- การพัฒนาโครงงานในลำดับที่ 3 คือการพัฒนาโครงงานจากการพัฒนาโครงงานในลำดับที่ 2 ให้มีความถูกต้องของผลลัพท์มากขึ้น โดยมีการทำงานดังนี้

1. คัดเลือก online dictionary สำหรับการค้นหาคำศัพท์และคำแปลโดยการทำแบบ manual โดยการเลือกจาเวปไซต์ที่มีคำศัพท์ปริมาณมากที่สุด และมีการแสดงคำศัพท์และคำแปลใน source ของหน้า website ที่สามารถเรียกขึ้นมาดูได้ และมีรูปแบบการแสดงผลลัพท์ที่แน่นอน

2. รวบรวมคำศัพท์และเก็บรวบรวมเป็น corpus โดยใช้ฟังก์ชั่นการทำงาน wc.DownloadFile(URL)

3. ตัดประโยคที่มีคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้ โดยพิจารณาจากsourceในcorpus ที่เรารวบรวมได้ดังเช่นเดียวกับการพัฒนาโครงงานในลำดับที่1และ2

4. ตัดคำศัพท์และคำแปลโดยใช้การทำ sentence extraction โดยดูจากเงื่อนไขของภาษา html ตามแต่ละ online dictionary ที่เรานำมาใช้ดังเช่นเดียวกับการพัฒนาโครงงานในลำดับที่1และ2

5. เก็บรวบรวมคำศัพท์และคำแปลลงฐานข้อมูล

6. เลือกหมวดหมู่หลักมา 6 หมวดหมู่ และหมวดหมู่ย่อยมาจำนวน 50 หมวดหมู่จากเดิม 115 หมวดหมู่

7. ทำการคัดกรองคำศัพท์เฉพาะคำนาม โดยดูจากคำแปลจะมีอักษร n. หรือ (n) เพื่อบอกว่าเป็นชนิดคำนาม โดยคัดเลือกมาทั้งหมด 1326 คำ โดยวิธี Random เพื่อให้ได้คำนามที่มีความหลากหลายและต่างชนิดกัน

8. ทำ Relation test ตามสูตร $P(s|x) = H(s^x)/H(x)$ โดยเปลี่ยนแปลงจากการใช้ค่า hits เป็นการพิจารณาเฉพาะ 30 ผลลัพท์แรกที่ได้ และใช้วิธีการนับจำนวนที่เจอคำศัพท์และหมวดหมู่ย่อยในหัวข้อของผลลัพท์เป็นค่า $H(s^x)$ และ ใช้วิธีการนับจำนวนที่เจอหมวดหมู่ย่อยในหัวข้อของผลลัพท์เป็นค่า $H(s)$

9. เก็บผลลัพท์จากการทำ Relation test ของทุกๆคำศัพท์กับทุกๆหมวดหมู่ย่อยที่เรามี ได้ผลลัพท์ทั้งหมด 60000 Relation

10. วิเคราะห์ความถูกต้องของผลลัพท์โดยการทำ manual test โดยพิจารณาจากการใช้ search engine เข้าช่วยในการพิจารณาว่าค่า Relation ระหว่างผลลัพท์และหมวดหมู่มีความเกี่ยวข้องกันมากน้อยเพียงใด โดยนำคำศัพท์และหมวดหมู่มาทำการค้นหาใน www.google.com โดยใส่เงื่อนไขในการค้นหาคือ “หมวดหมู่”+”คำศัพท์” และดูว่าพบมากน้อยเพียงใด และรวมถึงการตรวจสอบจากการพิจารณาแบบ manual test

จากการทำงานตามวิธีการข้างต้นจะสามารถสรุปผลลัพท์ได้ดังนี้

1. ในเวลา 3 นาทีสามารถหาความสัมพันธ์ของคำศัพท์ได้ 1 คำ กับ หมวดหมู่ย่อย 50 หมวดหมู่

2. ผลลัพท์ที่ได้สำหรับคำศัพท์ชนิดคำนามมีความถูกต้องแม่นยำมากกว่าการพัฒนาในลำดับขั้นตอนที่ 2 และจาก 10 คำที่มีค่า Relation มากกว่า 8 มีได้ผลลัพท์ที่แม่นยำ 70%

3. การพัฒนาโดยการเปลี่ยนจากการหาค่า hits มาเป็นการหาค่าที่พิจารณาเฉพาะ 30 ผลลัพท์แรกที่ได้ ทำให้ได้ผลลัพท์ที่แม่นยำและถูกต้องมากขึ้น

4. การพัฒนาโครงงานในลำดับที่3 ได้ผลลัพท์ที่แม่นยำกว่าการพัฒนาโครงงานในลำดับที่1 และการพัฒนาโครงงานในลำดับที่2 และแม่นยำเท่าที่ควรจึงพร้อมสำหรับนำไปทำ online dictionary เพื่อแสดงผลลัพท์ของโครงงานนี้

3.4 รูปแบบของผลลัพท์ตามขั้นตอนการทำงาน

ในส่วนของ Application

- การทำ Compiling Corpus ผลลัพท์ที่ได้คือชุดของข้อมูลทั้งหมดที่ได้จากการรวบรวมหน้าเวปไซต์
- การทำ Automatic Term Recognition ผลลัพท์ที่ได้คือชุดของข้อมูลทที่มีขนาดเล็กลงที่ได้จากการตัดเฉพาะส่วนของข้อมูลที่มีคำศัพท์และคำแปลอยู่ข้างใน
- การทำ filtering ผลลัพท์ที่ได้คือคำศัพท์และคำแปลสำหรับการเก็บลงฐานข้อมูล
- การทำ Relation test ผลลัพท์ที่ได้คือค่า Relation และเก็บรวบรวมลงฐานข้อมูล

ในส่วนของ Website

- การทำ online dictionary เพื่อแสดงผลลัพท์ในส่วนของ online dictionary มาตราฐาน คือ การค้นหาความหมายของคำศัพท์
- การทำ online dictionary เพื่อแสดงผลลัพท์ในส่วนของ relation คือการแสดงว่าศัพท์แต่ละคำมีความเกี่ยวข้องกับหมวดหมู่ใดบ้าง

บทที่ 4

การทำงานของโครงงานและผลการทำงาน

จากการทดลอง สามารถสรุปผลการทดลองการทำงานจากการพัฒนาโปรแกรม Automatic Collection of Vocabulary and Related Term from the Web โดยแบ่งการอธิบายออกเป็นส่วนต่างๆ ดังนี้

1. ส่วนของการเตรียมข้อมูล
2. ส่วนการทำงานของโปรแกรมในส่วนของ back office
3. ส่วนของการทำงานโปรแกรมในส่วนของเวปไซต์
4. ส่วนของการแสดงผลลัพท์

4.1 การเตรียมข้อมูล

ในการทดลองนี้ได้มีการเตรียมข้อมูลที่จะใช้ในการทดลองเพื่อการทำ Automatic Collection of Vocabulary and Related Term from the Web โดยมีการแบ่งการเตรียมข้อมูลเป็น 2 ส่วน คือส่วนแรกคือข้อมูลที่จัดเตรียมเพื่อมาทำเป็น online dictionary และส่วนที่สองคือข้อมูลที่เราจัดเตรียมเพื่อใช้ในการหา related term from website ของคำศัพท์และหมวดหมู่ที่เรากำหนด สิ่งที่เราจะต้องจัดเตรียมในขั้นตอนนี้คือการจัดเตรียมข้อมูลในส่วนของคำศัพท์และหมวดหมู่ ในที่นี้เรามีคำศัพท์ในฐานข้อมูลเป็นจำนวนมากกว่า 70,000 คำ และมีหมวดหมู่ย่อยทั้งหมด 118 หมวดหมู่ จากผลการทดลองการหาค่าความสัมพันธ์ทำให้ทราบว่าจากทฤษฎีนี้สามารถหาค่าความสัมพันธ์ของหมวดหมู่และคำศัพท์ได้ถูกต้องในระดับที่ดี โดยการเตรียมข้อมูลสำหรับใช้ในการทดลองนี้เป็นสิ่งสำคัญต่อการทำงานของการหาค่า Relation test ดังนั้คำศัพท์และหมวดหมู่ที่นำมาหาค่าความสัมพันธ์จะผ่านขั้นตอนในการจัดเตรียมให้เป็นข้อมูลที่พร้อมในการใช้งาน จากนั้นต้องนำไปผ่านกระบวนการคัดเลือกข้อมูล เพื่อให้ข้อมูลพร้อมใช้สำหรับโปรแกรม Automatic Collection of Vocabulary and Related Term from the Web ในที่นี้การเตรียมข้อมูลสำหรับในการทำ Relation test เราจัดคัดเลือกคำศัพท์มาทำ เป็นจำนวน 1300 ตัวเฉพาะที่เป็นชนิดคำนาม และ หมวดหมู่หลัก 6 หมวดหมู่ และ หมวดหมู่ย่อย 46 หมวดหมู่ นำมาทำ relation test ซึ่งจะได้ข้อมูลทั้งหมดคือ 1300*20 คือ 59800 relation เพื่อนำมาใช้ในการเปรียบเทียบดูผลการทดลอง

ตารางที่4.1 แสดงข้อมูลที่ใช้ในการทดลองของ Directory

Directory : Table

Directory_id	Directory
1	Business & Finance
2	Computers & Internet
3	Schools & Education
4	Entertainment & Arts
5	Health & Wellness
6	Recreation & Sports

ตารางที่4.2 แสดงข้อมูลที่ใช้ในการทดลองของ Subdirectory

Hardware	First Aid
Internet	Fitness
Programming	Health Care
Software	Pet Health
Technical	Automotive
Computer	Aviation
Subject	Gambling
Courses	Outdoor
Student	Sport
Graduate	Toy
Schools	Travel
Education	Game
University	Business
Actors	Company
Audio	Consumer
Comics	Employment
Fashion	Financial
Movies	Business
Radio	Industry
Television	Marketing
Dancing	Drug
Music	First Aid
Advice	Fitness
Beauty	Health Care
Children	Pet Health

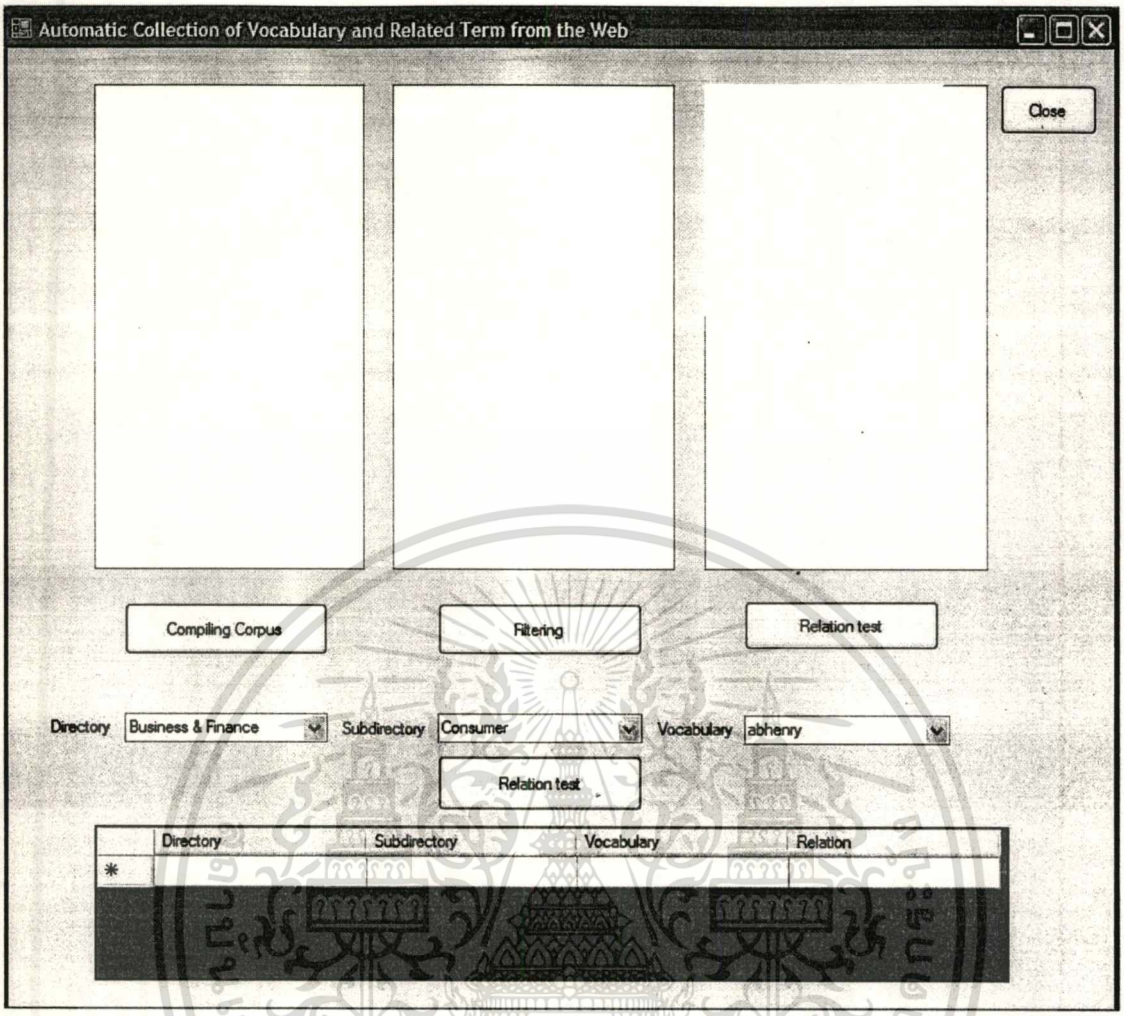
ตารางที่4.3 แสดงข้อมูลที่ใช้ในการทดลองของ Dictionary

English	Thai
abhenry	(แอบเฮน' รี) n. หน่วยการนำไฟฟ้าที่เป็นเซ็นติเมตร-กรัม-วินาที และมีค่าเท่ากบ 10-9
ablegate	(แอบ' ลิเคท) n. ผู้แทนของโป๊ป
accusation	(แอคคิวเซ่' เชิน) n. การกล่าวหา, คำประนาม, การตำหนิ, การกล่าวโทษ, การใส่ควา
achievement	(อะชีฟว' เมินท์) n. การบรรลุผล, ความสำเร็จ, ความสัมฤทธิ์,ผลสัมฤทธิ์, สัมฤทธิ์ภาพ
acolyte	(แอค' โคไลท์) n. เด็กผู้ช่วยขาดหลวงทำพิธี, พระในนิกายตรมันคาทอลิก
actionon	(แอค' ทินอน) n. ธาตุเฉื่อยที่เป็นแก๊สและมีสัญลักษณ์ An; at. no. : 86; at. wt. : 2
actor	(แอค' เทอะ) n. นักแสดงชาย, ผู้กระทำ, ผู้ดำเนินการ (คำที่มีความหมายเหมือนกัน
acuity	(อะคิว' อิที) n. ความหลักแหลม, ความคม, ความรุนแรง, ความคมกริบ, ความชัดเจน
addend	(แอด' เดนด์) n. เลขหรือจำนวนที่บวกเข้าด้วยกัน
adit	(แอด' ดิท) n. ทางตามแนวนอนเข้าสู่เหมือง, การเข้าหา, ทางเข้า
admittance	(แอดมิท' เทินซ์) n. การอนุญาติให้เข้า, การรับเข้า, การสารภาพ,การวัดการนำไฟฟ้า
advocacy	(แอด' โวคะซี) n. การเป็นทนาย, ทนาย, การสนับสนุน, ผู้สนับสนุน, การแก้ต่าง
aerodrome	(แอ' โรโดรม) n. สนามบิน
afflux	(แอฟ' ฟลัคซ์) n. สิ่งที่ไหลไปทางจุดหนึ่ง, การไหลไปทาง (that which flows)
afreet	(แอฟ' รีท , อะฟรีท') n. ปีศาจร้ายในนิยายอาหรับ (monster, afrit)
ageism	(เอจ' จิสซึม) n. การเลือกที่รักมักที่ชังโดยถืออายุเป็นเกณฑ์และโดยเฉพาะกับคนแก
agger	(แอจ' เจอะ) n. กระแสน้ำต่ำที่น้ำลดแล้วค่อยๆ ขึ้นแล้วลดอีก. -S... double tide
agrapha	(แอก' ราฟา) n. คำสอนของพระเยซูคริสต์ที่บันทึกโดยชาวคริสเตียนใน New Testa
agrimony	(แอก' กริโมนี) n. หญ้าพันธุ์มังกรจำพวก Agrimonia
agueweed	(เอ' กิววีด) n. พืชจำพวก Eupatorium ในอเมริกา gentian (Gentiana quinquefolia
aisle	('ไอล) n. ทางเดินระหว่างที่นั่ง (ในรถ, โรงหนัง, เครื่องบิน, โบสถ์), ที่นั่งฟังธรรมที่แบ่
albania	(แอลเบ' เนีย) n. ประเทศอัลบาเนีย
album	(แอล' บัม) n. สมุดหน้าเปล่าสำหรับเก็บภาพแสตมป์หรืออื่นๆ,แผ่นเสียงขนาดใหญ่, ช
alcoholicity	(แอลกะฮอลลิซ' ซิที) n. คุณภาพหรือความเข้มข้นของเครื่องดื่มแอลกอฮอล
alexander	(แอลลิกซาน' เดอะ) n. ชื่อเหล้า Cocktail ชนิดหนึ่ง
alleyway	(แอล' ลีเว) n. ทางที่แคบ, ตรอก, ซอย
allium	(แอล' เลียม) n. พืชจำพวก Allium (ต้นหอม, กระเทียม) (onion, leek, shallot)
allograph	(แอล' โลกราฟ) n. การเขียนหรือการเซ็นชื่อคดยคนหนึ่งเพื่ออีกคนหนึ่ง

4.2 ส่วนของการทำงานของโปรแกรม

การทำงานจะแบ่งเป็น 2 ส่วน คือส่วนที่เป็น Application ในการค้นหาคำศัพท์และทำ Relation test และส่วนที่เป็น website สำหรับ online dictionary เพื่อค้นหาความหมายของคำศัพท์และเพื่อค้นหาการทำ relation test

-ในส่วนของ Application เมื่อเริ่มต้นโปรแกรมการทำงานจะปรากฏหน้าจอในรูป 4.1 โดยแบ่งเมนูการทำงานออกเป็นส่วนๆคือส่วนของการเก็บรวบรวม Corpus มาจากเวปไซต์ ส่วนของการ filtering คำศัพท์มากจาก Corpus ที่เรารวบรวมได้ และส่วนของการทำ Relation test ดังแสดงในรูป 4.1

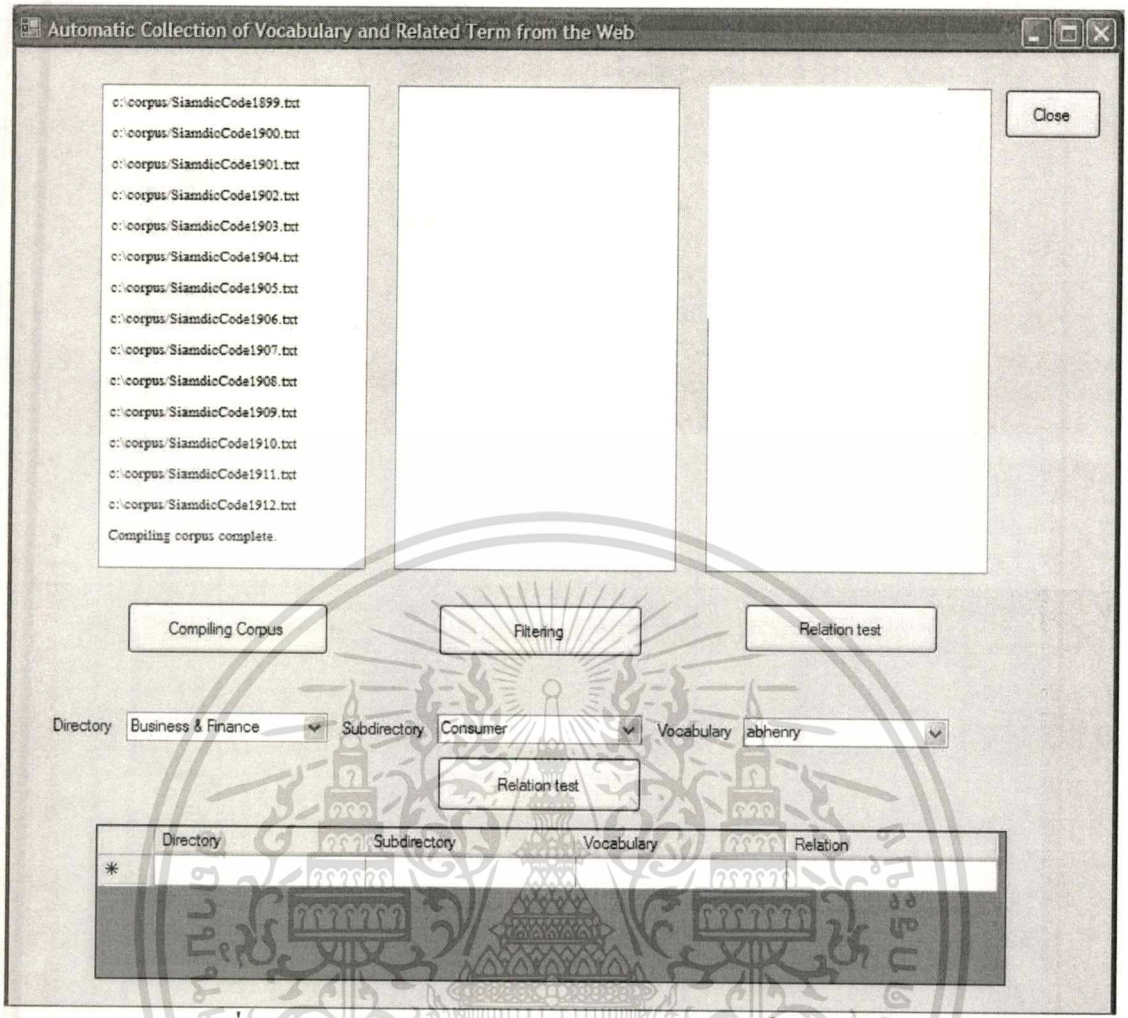


รูปที่4.1 หน้าจอหลักของ Application

1. ส่วนของการเก็บรวบรวมข้อมูลจาก Corpus จะเป็นการทำงานเมื่อกดปุ่ม Compiling Corpus ซึ่งเมื่อทำการกดแล้ว การทำงานคือการเก็บรวบรวมข้อมูลจากเวปไซต์โดยใช้เทคโนโลยี search engine มาช่วย ซึ่งในส่วนของการทำงานนี้ input ของโปรแกรมส่วนนี้ได้แก่

- URL ของเวปไซต์ของ dictionary ที่เราคัดเลือกมา
- จำนวนหน้าเวปไซต์ของเวปไซต์ dictionary นั้นๆ
- รูปแบบการเก็บ source file ของเวปไซต์นั้นๆ
- รูปแบบการเก็บคำศัพท์ภาษาอังกฤษของเวปไซต์นั้นๆ
- รูปแบบการเก็บคำศัพท์ภาษาไทยของเวปไซต์นั้นๆ

เมื่อทำงานแล้ว ผลของ Corpus ที่ถูกสร้างขึ้นจะโชว์ขึ้นมาให้ดูที่ช่องแสดงผลตามรูป 4.2 โดยที่ถ้าการทำงานเสร็จสมบูรณ์จะขึ้นคำว่า Compiling corpus complete ขึ้นมา ดังรูป 4.2

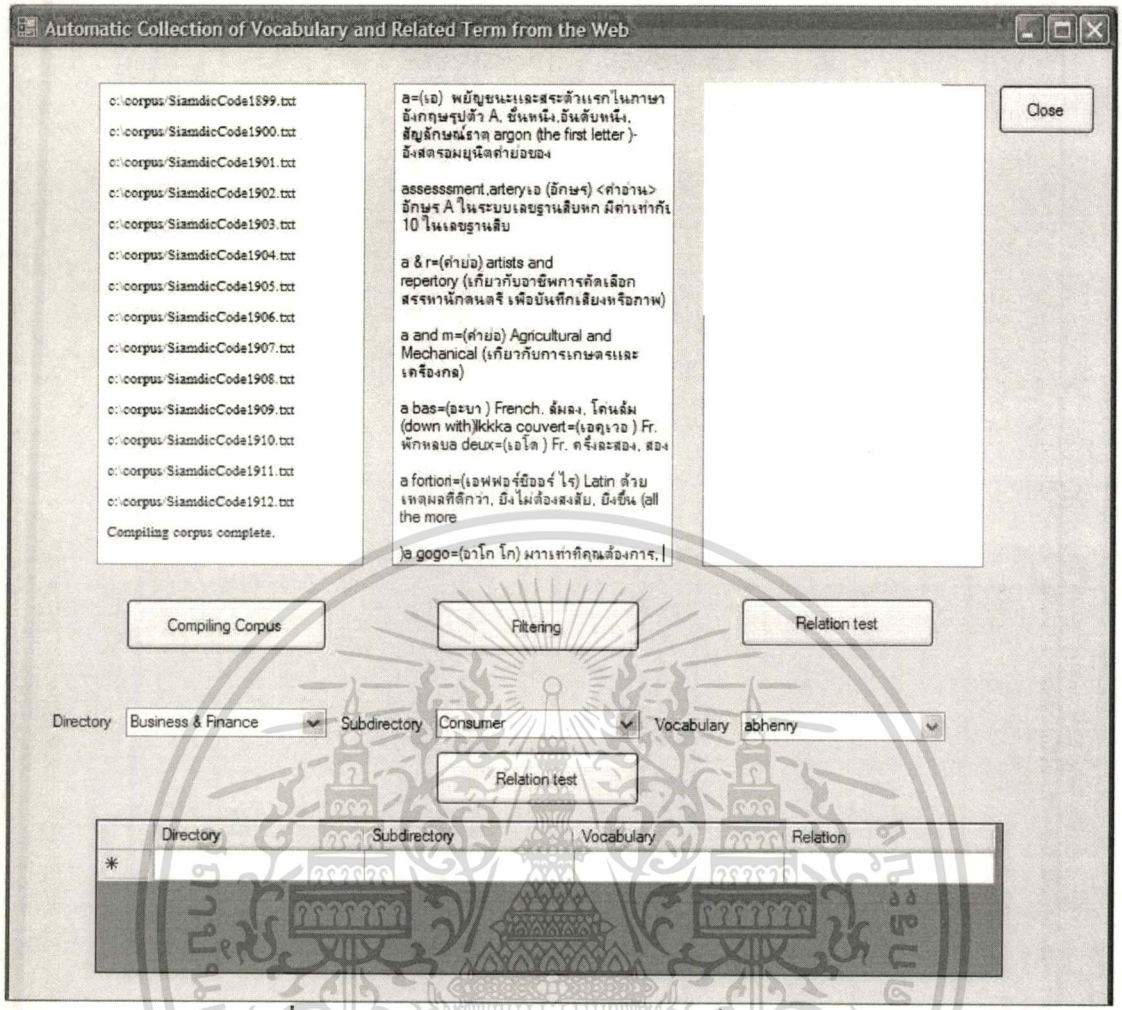


รูปที่4.2 ผลลัพท์จากการกดปุ่ม Compiling corpus ที่ปรากฏบน application

2. ส่วนของการทำการ Filtering เพื่อดึงคำศัพท์และคำแปลออกมาจาก Corpus ที่เรารวบรวมได้จากการทำ compiling corpus โดยการทำงาน 2 ขั้นตอนได้แก่การทำ Web page collection และการทำ Sentence extraction ซึ่งในส่วนของการทำงานนี้ input ของโปรแกรมส่วนนี้ได้แก่

- Corpus ที่เราเก็บรวบรวมได้
- Automatic Term Recognition

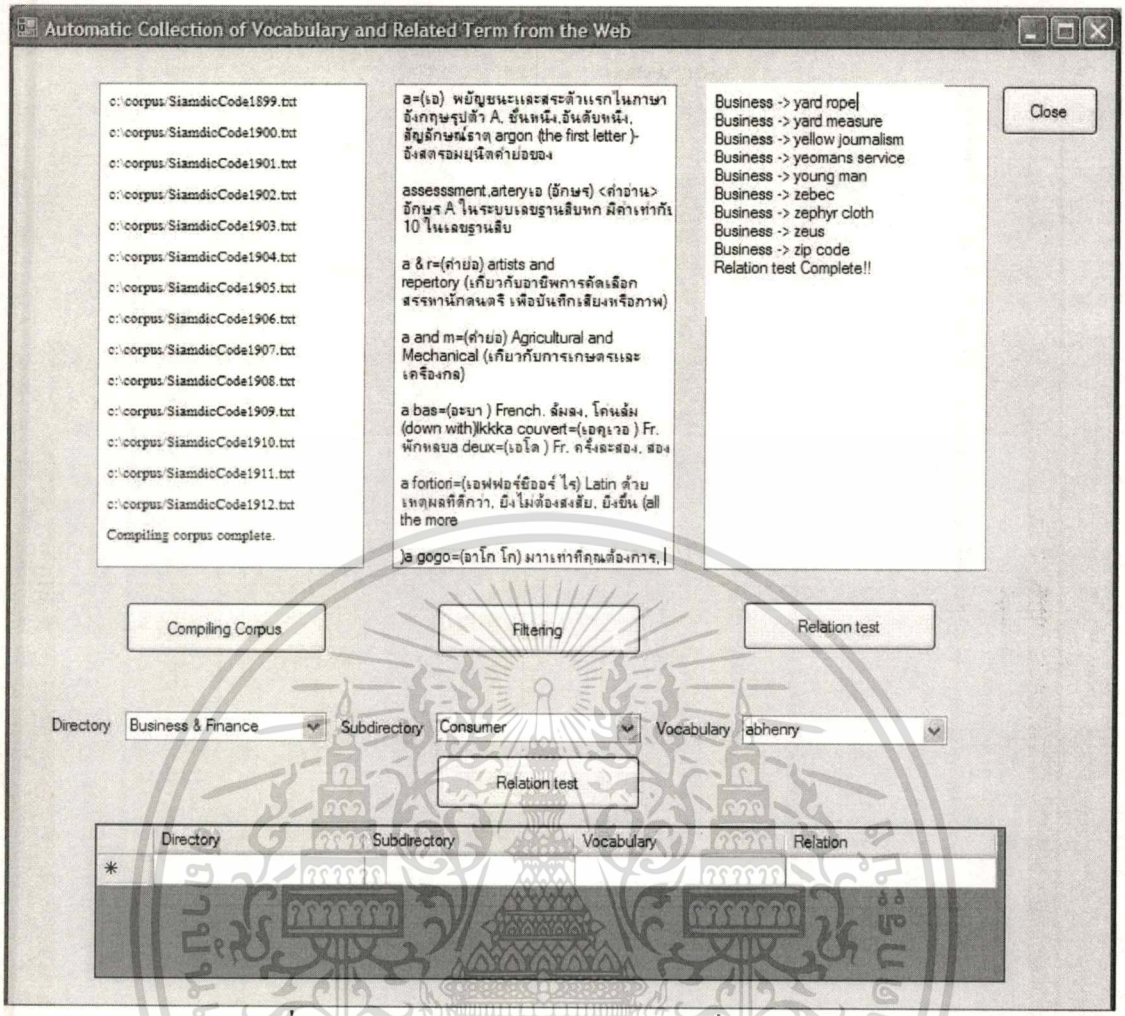
เมื่อทำงานแล้ว ผลของ คำศัพท์และคำแปล ที่ถูกทำ Filtering มาได้จะโชว์ขึ้นมาให้ดูที่ช่องแสดงผลตามรูป 4.3 โดยที่ถ้าการทำงานเสร็จสมบูรณ์จะขึ้นคำว่า Filtering complete ขึ้นมา ดังรูป 4.3 และจะทำการเก็บคำศัพท์และความหมายลงในฐานข้อมูล ในตาราง Dictionary



รูปที่4.3 ผลลัพท์จากการกดปุ่ม Filtering ที่ปรากฏบน application

3. ส่วนของการทำการ Relation test เพื่อทำการหาความสัมพันธ์ระหว่างคำศัพท์และหมวดหมู่ โดยจะเริ่มทำงานเมื่อมีการกดปุ่ม Relation test ตัวโปรแกรมจะนำหมวดหมู่ที่เราต้องการ และคำศัพท์ที่เราจัดไว้ มาทำการหาค่า Relation โดยได้ผลเป็นค่าความน่าจะเป็นที่คำศัพท์และหมวดหมู่จะมีความเกี่ยวข้องกัน

เมื่อทำงานแล้ว ผลของความสัมพันธ์ระหว่างคำศัพท์และหมวดหมู่ จากการทำ Relation test จะแสดงขึ้นมาให้ดูที่ช่องแสดงผลตามรูป 4.4 โดยที่ถ้าการทำงานเสร็จสมบูรณ์จะขึ้นคำว่า Relation complete ขึ้นมา ดังรูป4.4



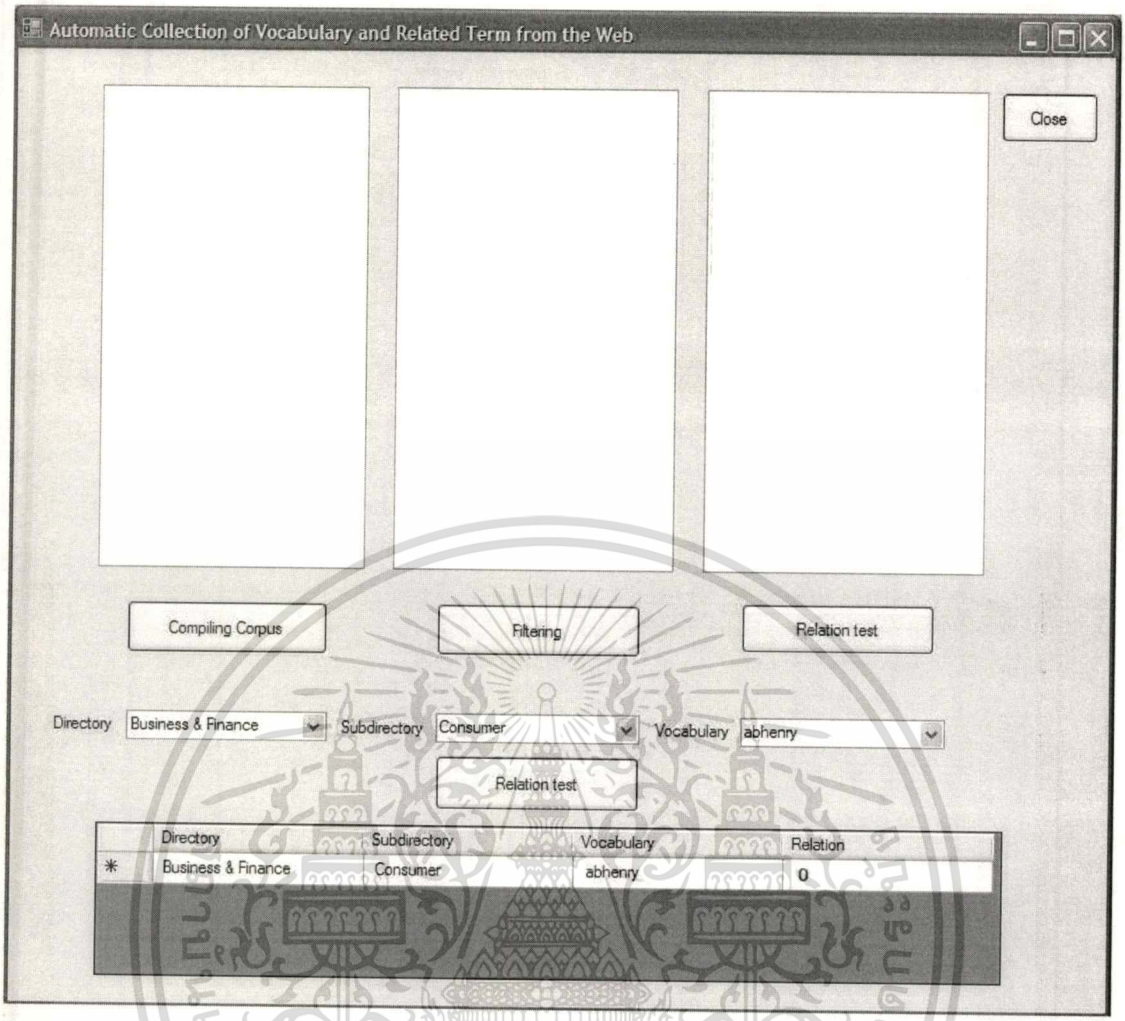
รูปที่4.4 ผลลัพท์จากการกดปุ่ม Relation ที่ปรากฏบน application

และผลของการทำ Relation test จะได้ผลดังตาราง 4.4

ตารางที่4.4 แสดงผลที่ได้จากการทดลองของ Relation test

relation_id	Word_id	Subdirectory_id	Directory_id	probability	Category_hit	Word_hit	Both_hit
210301	12541	9	2	0	136000000	18	0
210302	12541	11	2	0	1150000000	18	0
210303	12541	12	2	0.111111111111111	474000000	18	2
210304	12541	14	2	0.0555555555555556	6450000	18	1
210305	12541	16	3	0.0555555555555556	975000000	18	1
210306	12541	19	3	0	209000000	18	0
210307	12541	21	3	0.333333333333333	149000000	18	6
210308	12541	23	3	0	1150000	18	0
210309	12541	27	3	0.0555555555555556	35000000	18	1
210310	12541	28	3	0.222222222222222	863000000	18	4
210311	12541	34	3	0.111111111111111	1010000000	18	2
210312	12541	36	1	0.444444444444444	223000000	18	8
210313	12541	37	1	0.222222222222222	530000000	18	4
210314	12541	38	1	0.0555555555555556	287000000	18	1
210315	12541	39	1	0	265000000	18	0
210316	12541	40	1	0.0555555555555556	83800000	18	1
210317	12541	42	1	0.444444444444444	2460000000	18	8
210318	12541	43	1	0.166666666666667	60500000	18	3
210319	12541	46	1	0.111111111111111	248000000	18	2
210320	12541	53	4	0	1380000	18	0
210321	12541	55	4	0.111111111111111	9010000	18	2
210322	12541	57	4	0	5340000	18	0
210323	12541	60	4	0	295000000	18	0
210324	12541	66	4	0.0555555555555556	516000000	18	1

4. ส่วนของการเลือกทดสอบ Relation test แบบเจาะจงคำศัพท์หรือตามกลุ่มของตัวอักษร เพื่อหาความสัมพันธ์ระหว่างหมวดหมู่ที่เลือก หรือทุกหมวดหมู่ กับคำศัพท์ที่เรากำหนดไว้ โดยหลักของการทำงานนี้ ใช้เพื่อตรวจสอบความสัมพันธ์ของคำศัพท์เฉพาะเจาะจงที่ต้องการ โดยการเลือกหมวดหมู่ที่ต้องการ เลือกหมวดหมู่ย่อยที่ต้องการ เลือกคำศัพท์ที่ต้องการ หรือเลือกคำขึ้นต้นที่ต้องการ เมื่อเลือกทั้งหมดเสร็จสิ้น การทำงานจะเริ่มเมื่อทำการกดปุ่ม Relation test ผลลัพท์ของการทำ Relation test จะแสดงให้เห็นในช่องแสดงผลบน application ดังรูป 4.5



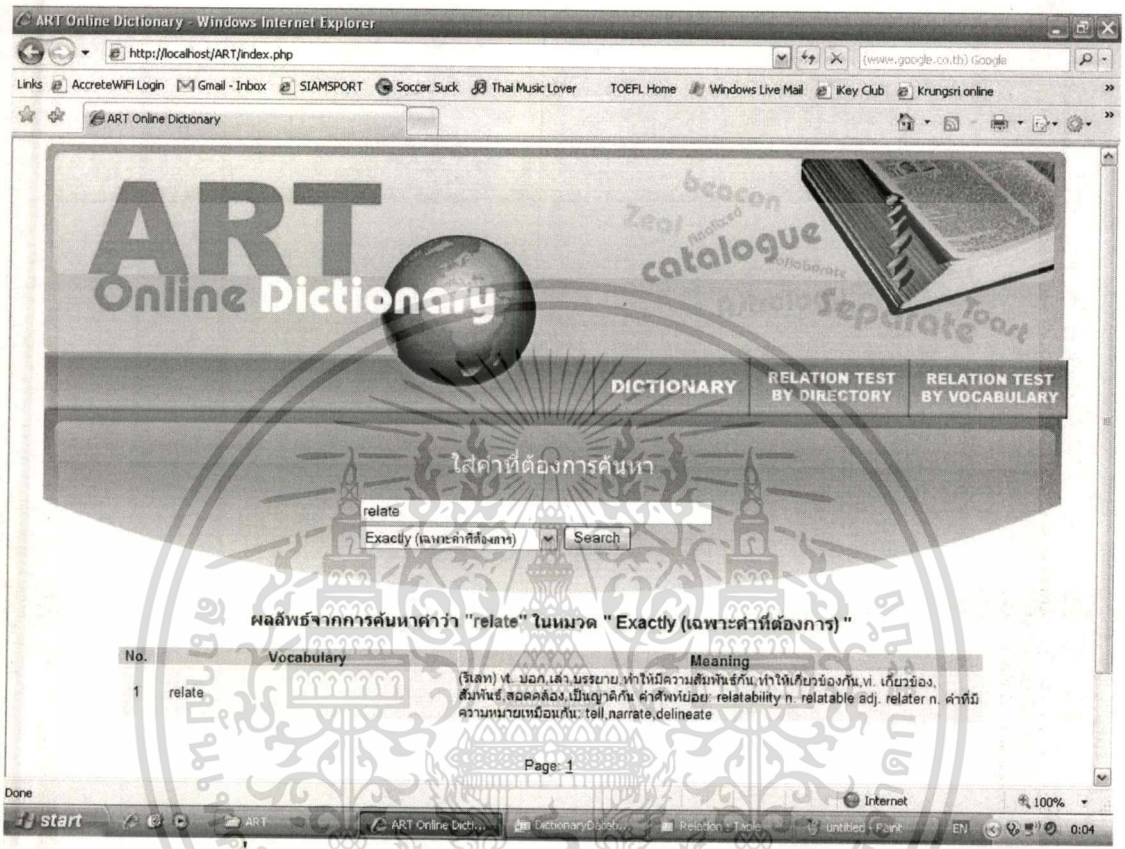
รูปที่4.5 แสดงผลที่ได้จากการทำ Relation test แบบเจาะจงคำศัพท์

-ในส่วนของ online dictionary เราจะแสดงผลของการทำงานของโปรแกรมตามทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web ด้วยการใช้ website ในที่นี้เราจัดทำในรูปแบบของ online dictionary โดยใช้ชื่อว่า ART Dictionary ซึ่งจะมีการทำงาน 3 รูปแบบ ได้แก่

1. การทำงาน Online Dictionary ทั่วไป
2. การทำงานของการตรวจสอบ Relation test ตามหมวดหมู่
3. การเลือกคำศัพท์ที่ต้องการเพื่อตรวจผลของ Relation test กับทุกๆหมวดหมู่ย่อย

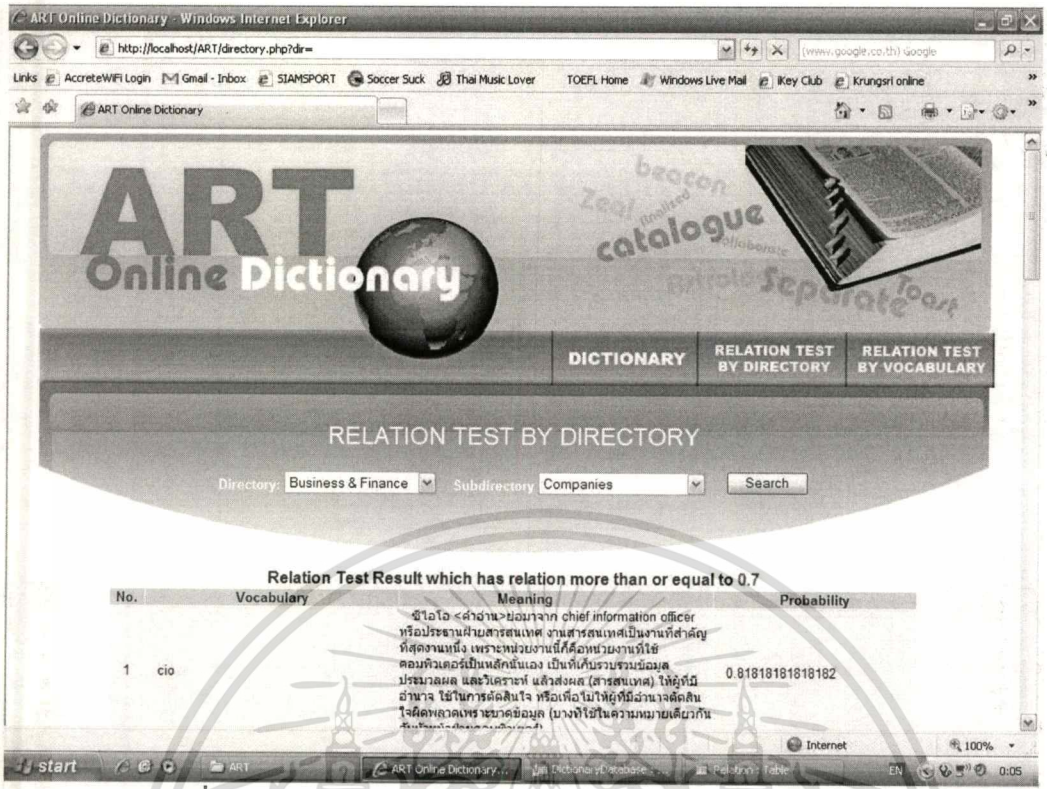
-การทำงานของ Dictionary ทั่วไปโดยหน้าแรกของ online dictionary สามารถดูได้ดังรูป 4.6 ซึ่งเป็นความสามารถในหาค้นหาความหมายของคำศัพท์ในรูปแบบทั่วไป โดยแบ่งการทำงานในส่วนนี้ออกเป็น 3 ส่วนย่อยได้แก่

1. การค้นหาคำศัพท์เฉพาะคำศัพท์ที่เราต้องการ
2. การค้นหาคำศัพท์ที่มีคำศัพท์ที่เราต้องการเป็นตัวประกอบ
3. การค้นหาคำศัพท์ที่ขึ้นต้นด้วยคำศัพท์ที่เราต้องการ

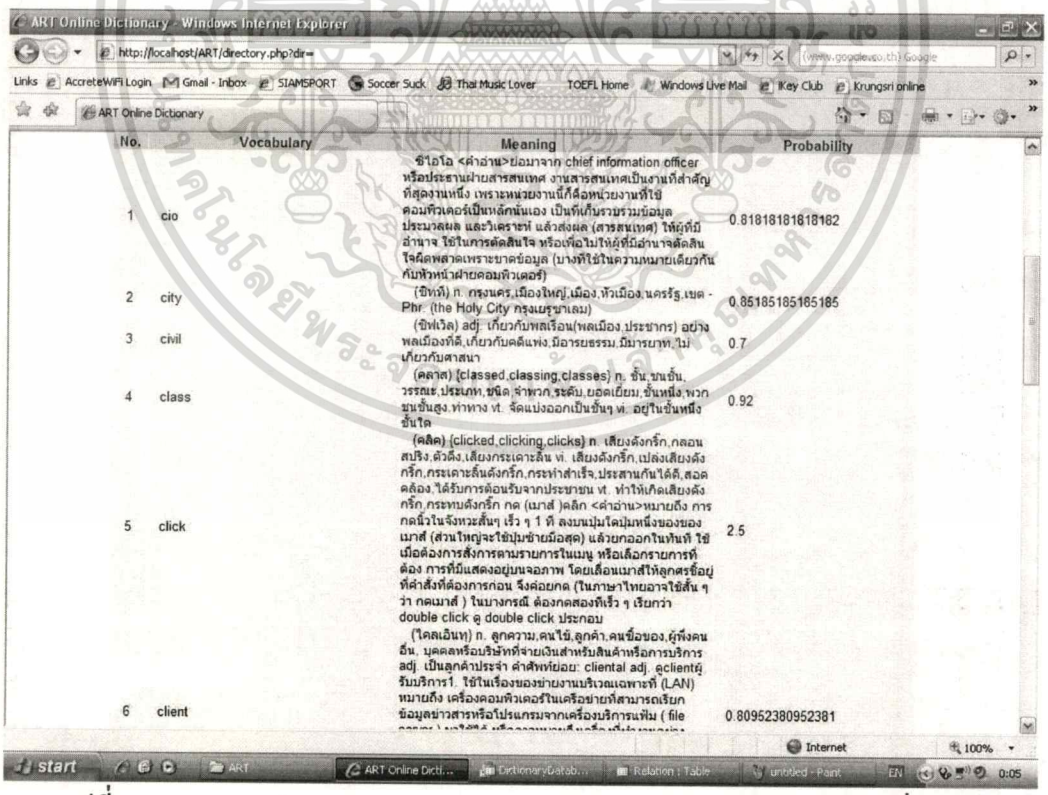


รูปที่4.6 แสดงรูปแบบหน้าการทำงานหลักของ online dictionary

-การทำงานของการตรวจสอบ Relation test ตามหมวดหมู่ จะเป็นการเลือกดูค่า Relation test โดยเลือกจากหมวดหมู่หลัก และหมวดหมู่ย่อย และ online dictionary จะแสดงผลของคำศัพท์ต่างๆที่มีค่า Relation test กับหมวดหมู่ย่อยที่ผู้ใช้งานเลือก โดยที่ผลลัพท์ที่จะมีจำนวนเท่าใด หรือตรอบคลุมการคำศัพท์ทุกคำหรือไม่ ขึ้นกับการทำงานของส่วน application ซึ่งสามารถจะหาค่า Relation test ของคำศัพท์ทุกคำ หรือแค่บางส่วนได้ โดยผลของการทำงานสามารถดูได้ดังรูป 4.7 และรูป 4.8 ซึ่งเป็นความสามารถในการตรวจสอบ Relation test ตามหมวดหมู่

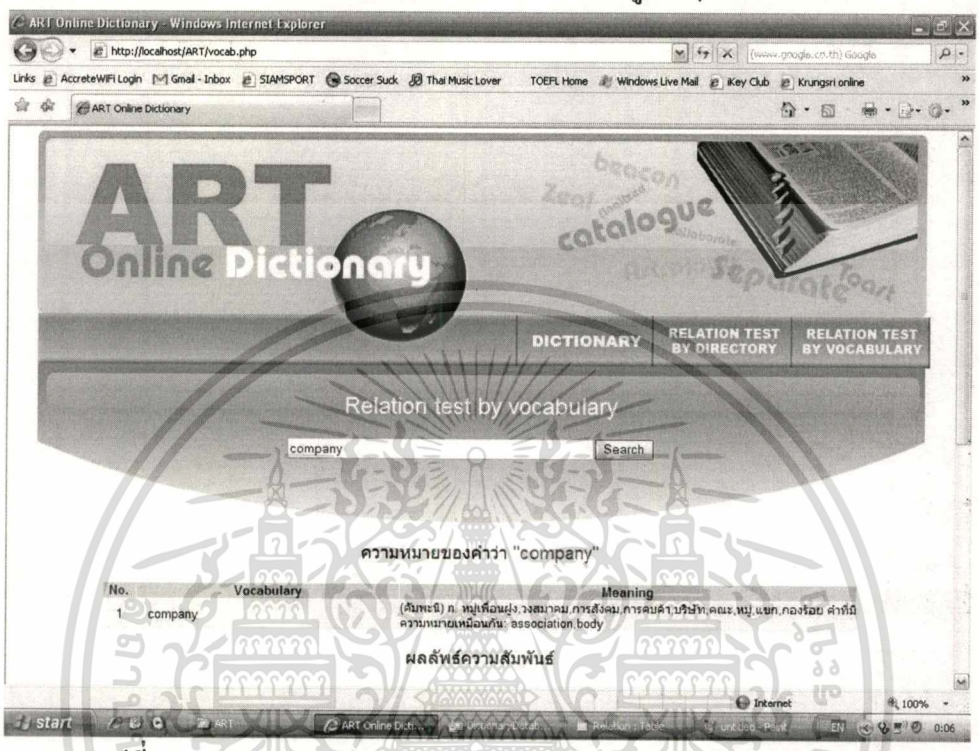


รูปที่4.7 แสดงการทำงานของการตรวจสอบ Relation test ตามหมวดหมู่

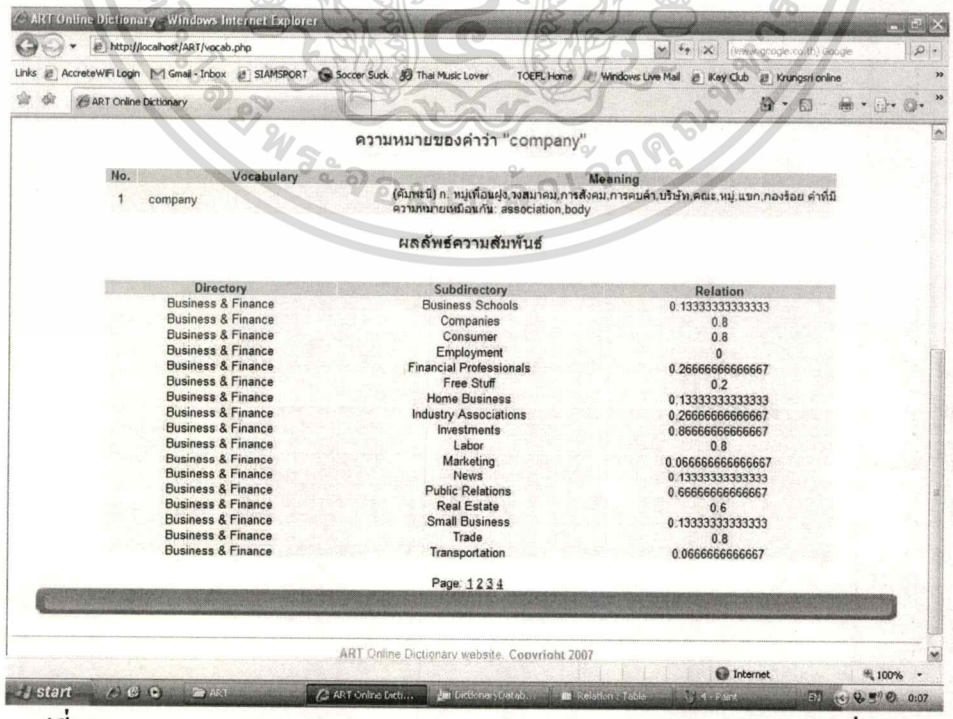


รูปที่4.8 แสดงการทำงานของการตรวจสอบ Relation test ตามหมวดหมู่ (เพิ่มเติม)

- การเลือกคำศัพท์ที่ต้องการเพื่อตรวจผลของ Relation test กับทุกๆหมวดหมู่ย่อย โดยผลของการทำงานสามารถดูได้ดังรูป 4.9 และรูป 4.10 ซึ่งเป็นความสามารถในการตรวจสอบ Relation test ของทุกๆหมวดหมู่ตามคำศัพท์ที่เราต้องการ โดยที่ผลลัพท์ที่ได้จะมีการแสดงความหมายของคำศัพท์นั้นๆก่อน และมีการแสดงค่า Relation test กับหมวดหมู่ต่างๆต่อเนื่องลงมา



รูปที่4.9 แสดงการทำงานของการตรวจสอบ Relation test ตามคำศัพท์



รูปที่4.10 แสดงการทำงานของการตรวจสอบ Relation test ตามคำศัพท์ (เพิ่มเติม)

จากผลการทำลองที่ได้ทำในโปรแกรมนี้ เราจะแบ่งผลการทำลองออกเป็นส่วนๆ ตามค่าของ Relation test ที่ได้เป็น 4 ชนิดดังนี้

-ชนิดที่ 1 ค่า Probability =1 จะแบ่งได้ 2 ประเภท คือ

Type 0: หมวดหมู่และคำศัพท์อาจจะเป็นคำๆเดียวกัน

Type 1: คำศัพท์มีความเกี่ยวข้องกับหมวดหมู่หมวดหมู่มาก ดังตัวอย่างรูป 4.11

ความหมายของคำว่า "Company"

Subdirectory	Relation
Labor	1.0
Trade	1.0

รูปที่4.11 แสดงผลของ Relation test สำหรับคำศัพท์มีความเกี่ยวข้องกับหมวดหมู่หมวดหมู่มาก

-ชนิดที่ 2 ค่า Probability <1 และค่า Probability >=0.8

Type 2: คำศัพท์มีความเกี่ยวข้องกับหมวดหมู่ ดังตัวอย่างรูป 4.12

ความหมายของคำว่า "Company"

Subdirectory	Relation
Companies	0.8
Consumer	0.8

รูปที่4.12 แสดงผลของ Relation test สำหรับ คำศัพท์มีความเกี่ยวข้องกับหมวดหมู่

-ชนิดที่ 3 ค่า Probability <0.8 และค่า Probability >=0.5

Type 3: คำศัพท์และหมวดหมู่อาจเกี่ยวข้องหรือไม่เกี่ยวข้องกัน ดังตัวอย่างรูป 4.13

ความหมายของคำว่า "connection"

Subdirectory	Relation
Companies	0.5
Consumer	0.75

รูปที่4.13 แสดงผลของ Relation test สำหรับคำศัพท์และหมวดหมู่อาจเกี่ยวข้องหรือไม่เกี่ยวข้องกัน

-ชนิดที่ 4ค่า Probability <0.5

Type 4: คำศัพท์และหมวดหมู่ไม่มีความเกี่ยวข้องกันเลย ดังตัวอย่างรูป 4.14

ความหมายของคำว่า "connection"

Subdirectory	Relation
Financial Professionals	0.041666666666667
Free Stuff	0.166666666666667

รูปที่4.14 แสดงผลของ Relation test สำหรับคำศัพท์และหมวดหมู่ไม่มีความเกี่ยวข้องกันเลย

บทที่ 5
สรุปผลและข้อเสนอแนะ

จากการศึกษาการทำงานของการเก็บรวบรวมคำศัพท์และคำแปลจากหน้าเวปไซต์และการค้นหาความสัมพันธ์ของคำศัพท์และหมวดหมู่ต่าง โดยใช้ทฤษฎี Automatic Collection of Vocabulary and Related Term from the Web ในโครงการนี้ สามารถสรุปผลกรดำเนินงานและสรุปผลการทดลอง รวมถึงขอเสนอแนะได้ดังต่อไปนี้

5.1 สรุปผลการทำงาน

จากการศึกษาในโครงการนี้สามารถสรุปผลการดำเนินงานในการทำ Automatic Collection of Vocabulary and Related Term from the Web ได้ดังนี้

- ในการเก็บรวบรวมข้อมูลคำศัพท์และคำแปลมาจากเวปไซต์ สามารถทำได้ 2 รูปแบบ คือ การเก็บรวบรวมคำศัพท์จากเวปไซต์ทั่วไป โดยใช้เทคโนโลยี search engine โดยที่การเก็บรวบรวมคำศัพท์จากเวปไซต์ทั่วไป จะได้คำศัพท์ประมาณ 3 คำจากหน้าเวปไซต์ประมาณ 3000 เวปไซต์ ซึ่งเป็นจำนวนน้อยมากและใช้เวลามากในการรวบรวมคำศัพท์ สำหรับกรณีที่ต้องการสามารถให้ program ทำงานโดยใช้เวลานานได้ แต่ในกรณีที่ต้องการคำศัพท์จำนวนมากในเวลาที่จำกัด อีกทางเลือกคือการค้นหาคำศัพท์จาก website online dictionary ที่มีอยู่ทั่วไป เนื่องจากในเวปไซต์เหล่านี้มีคำศัพท์และคำแปลจำนวนมาก ทำให้เราสามารถรวบรวมข้อมูลได้ แต่ในกรณีที่มีเวลามากในการทำงาน ควรจะทำการรวบรวมคำศัพท์จากเวปไซต์ทั่วไปเพิ่มเติมจากคำศัพท์ที่ได้จาก online dictionary ปรกติเพื่อการได้คำศัพท์ที่มีความหลากหลายมากขึ้น

- ในการทำ Compiling Corpus ใช้เวลามากในการรวบรวมคำศัพท์ เนื่องจากมีคำศัพท์ในเวปไซต์จำนวนมาก ทำให้เครื่องคอมพิวเตอร์ที่จะใช้ในการทำงานนี้ต้องมีประสิทธิภาพสูงและสามารถทำงานต่อเนื่องได้เป็นเวลานาน จึงจะสามารถทำการรวบรวมคำศัพท์ได้ โดยควรแบ่งการรวบรวมคำศัพท์ออกเป็นส่วนๆ เช่นการแบ่งตามตัวอักษร เพื่อให้เครื่องคอมพิวเตอร์ไม่ต้องทำงานมากจนเกิดความสามารถและแบ่งการทำงานได้

- ในการทำการ Filtering คำศัพท์ ด้วยเทคนิค Web page collection และการทำ Sentence extraction ใช้เวลานานในการคัดแยกคำศัพท์ออกมาจาก Corpus ที่รวบรวมไว้ ซึ่งใช้เวลามากกว่าการทำ Compiling Corpus ดังนั้นก็เช่นกับการทำ Compiling Corpus คือควรมีการแบ่งตามตัวอักษรเพื่อการทำงานที่มีประสิทธิภาพ หรือต้องใช้เครื่องคอมพิวเตอร์ที่มีประสิทธิภาพเพียงพอในการทำงาน

- ในการทำ Relation test เป็นส่วนของการทำงานที่ใช้เทคโนโลนี search engine เข้าช่วย ดังนั้นเครื่องคอมพิวเตอร์ที่ใช้ในการคำนวณต้องมีการต่ออินเตอร์เน็ตตลอดเวลา และด้วยการที่เรามีคำศัพท์และหมวดหมู่ย่อยจำนวนมากทำให้เราจะเป็นต้องใช้อินเตอร์เน็ตความเร็วสูง

- ผลที่ได้จากการทำ Relation test ยังคงต้องใช้การทำ manual checking ในการดูว่าผลที่ได้นั้นควรจะสรุปผลยังไงซึ่งผลได้ที่ปริมาณมากทำให้ต้องใช้เวลาจำนวนมากในการวิเคราะห์ผล

- เมื่อมีการเรียนรู้จากผลของการทำ Relation test ทำให้ต้องมีการปรับปรุงการทำงาน Relation test ให้มีประสิทธิภาพยิ่งขึ้น โดยใช้เทคนิคต่างๆของ Automatic Collection of Vocabulary and Related Term มาปรับปรุงเช่นการคัดเลือกเฉพาะคำศัพท์ที่เป็นชนิดคำนามในการหา Relation test เนื่องจากจะช่วยเพิ่มความแม่นยำของผลลัพท์ และมีการปรับเปลี่ยนสมมุติฐานที่เราจะได้ในตอนแรกและข้อสรุปที่ได้หลังจากได้ผลลัพท์ออกมาแล้ว

5.2 สรุปผลการทดลอง

1. ค่าที่ได้จากการทำ Relation test ขึ้นกับคำนาม ชนิดของคำนาม และหมวดหมู่ที่จัดเตรียมไว้ คำนามนั้นจะมีความเกี่ยวข้องกับหมวดหมู่มากน้อยแค่ไหน หรือตรงตามที่ต้องการหรือไม่ขึ้นอยู่กับกาจัดเตรียมคำศัพท์และการหมวดหมู่ต่างๆนั่นเอง

2. ค่า Relation test ที่ได้จะมีประโยชน์ต่อการหาคำศัพท์ที่มีความเกี่ยวข้องกัน หรือคำศัพท์กับหมวดหมู่ที่มีความเกี่ยวข้องกัน เพื่อใช้ประโยชน์ต่างๆมากมาย เช่นการทำวิจัย การเขียนโครงงานภาษาอังกฤษ หรือการนำไปทำ machine translation ต่อไปได้

3. การหาค่า Relation test เพื่อให้ได้ผลที่ถูกต้องแม่นยำยิ่งขึ้น ต้องมีการพัฒนาปรับปรุงวิธีการทำ Relation test และต้องมีการคัดเลือกคำศัพท์และหมวดหมู่ให้ตรงกับความต้องการในด้านการใช้งานให้มากที่สุด และผลลัพท์ที่ได้จะตรงตามความต้องการของผู้ใช้งานมากยิ่งขึ้น

5.3 ข้อเสนอแนะ

เพื่อปรับปรุงโปรแกรมประยุกต์นี้ในอนาคต ผู้ศึกษามีความเห็นว่า เพื่อเพิ่มความเร็วให้กับการทำงานของ Automatic Collection of Vocabulary and Related Term from the Web คือการเตรียม Hardware และสภาวะแวดล้อมที่เหมาะสมกับการทำงาน และถ้ามีเวลานานสำหรับการทำ Relation test ควรจะทำ Relation test กับคำศัพท์ทุกตัว เพื่อนำมาใช้ประโยชน์อื่นๆและเพื่อวิเคราะห์และปรับปรุงการทำ Relation test ให้ได้ผลลัพท์ดียิ่งขึ้น

โปรแกรมประยุกต์นี้อาจจะนำไปใช้พัฒนาต่อเนื่องเพื่อการทำ Relation test ที่มีประสิทธิภาพมากขึ้น หรือการนำไปใช้ในการทำงานอื่นๆเช่น machine translation เป็นต้น

บรรณานุกรม

ไพศาล โมลิสกุลมงคล. 2545. **Microsoft Visual C#.net.** กรุงเทพฯ : หจก. ไทยเจริญการพิมพ์.

ศุภชัย สมพานิช. 2546. **คู่มือการเขียนโปรแกรม Visual C#.NET ฉบับโปรแกรมเมอร์.** นนทบุรี อินโฟเพรส.

สิทธิศักดิ์ คล่องดี. 2542. **สร้างฐานข้อมูลด้วย Microsoft Access2000 อย่างมืออาชีพ.** กรุงเทพฯ : เยลโล่การพิมพ์(1988).

Sato and Y.Sasaki. 2003. "Automatic collection of related terms from the web." 121-124. In Proc

ภาคผนวก

ภาคผนวก ก.

คู่มือการติดตั้งระบบ

ภาคผนวก ข.

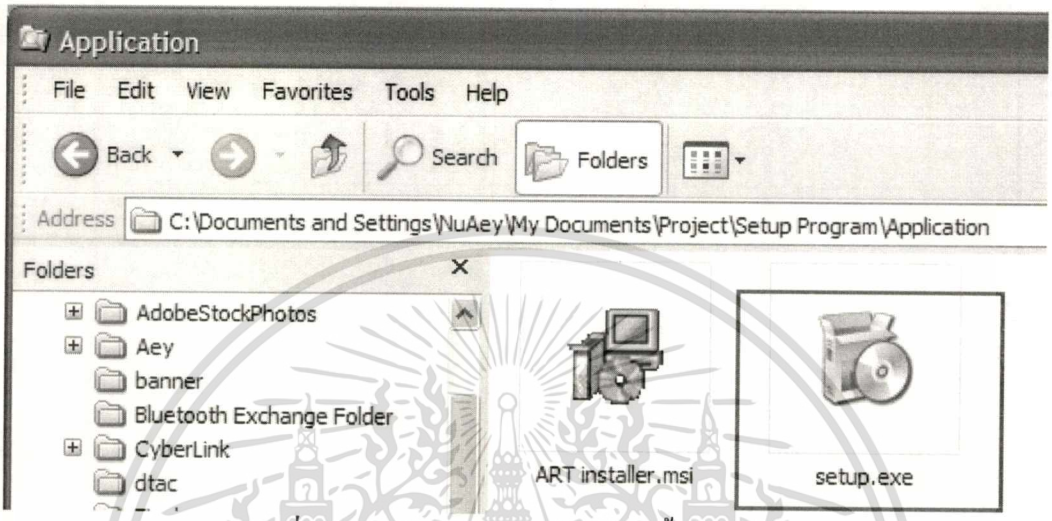
คู่มือการใช้งานระบบ

คู่มือการติดตั้งระบบการหาค่าความสัมพันธ์ระหว่างคำศัพท์กับหมวดหมู่

1. การติดตั้งโปรแกรมส่วนวินโดวส์แอพลิเคชั่น

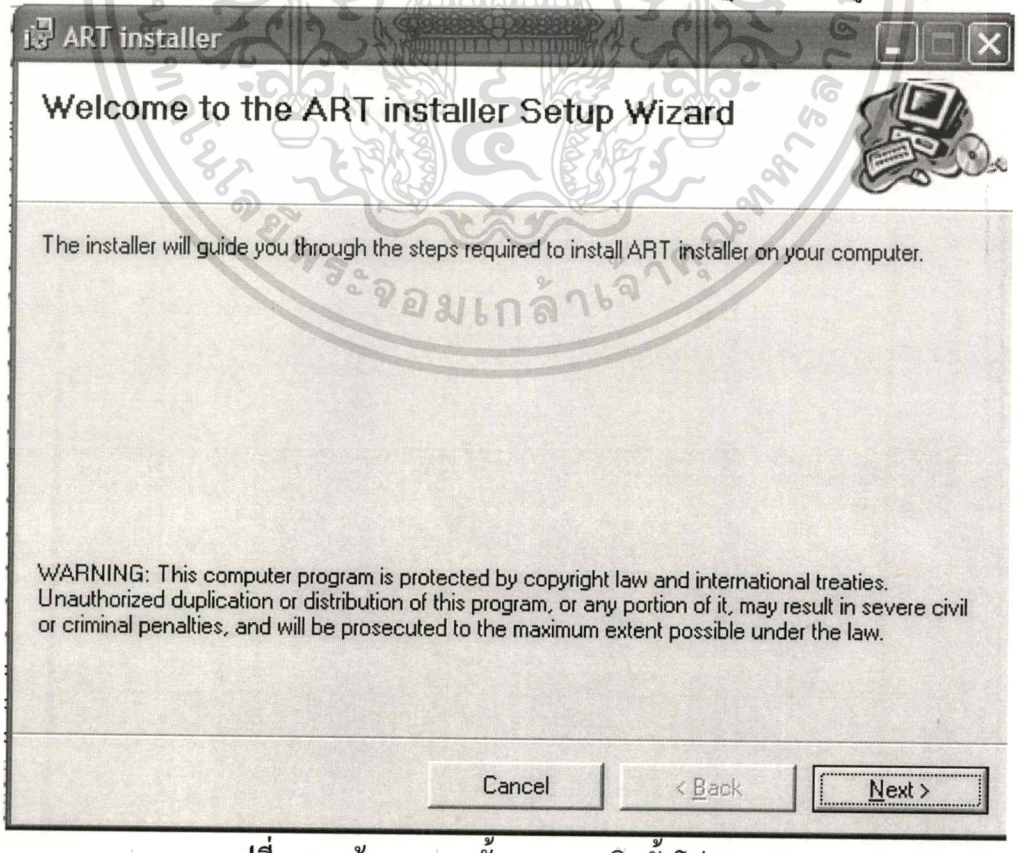
โดยมีขั้นตอนเป็นไปตาม Step by Step ดังต่อไปนี้

1.1 ดับเบิ้ลคลิกที่ตัวไอคอน setup.exe ตามรูปที่ ก.1 ต่อไปนี้



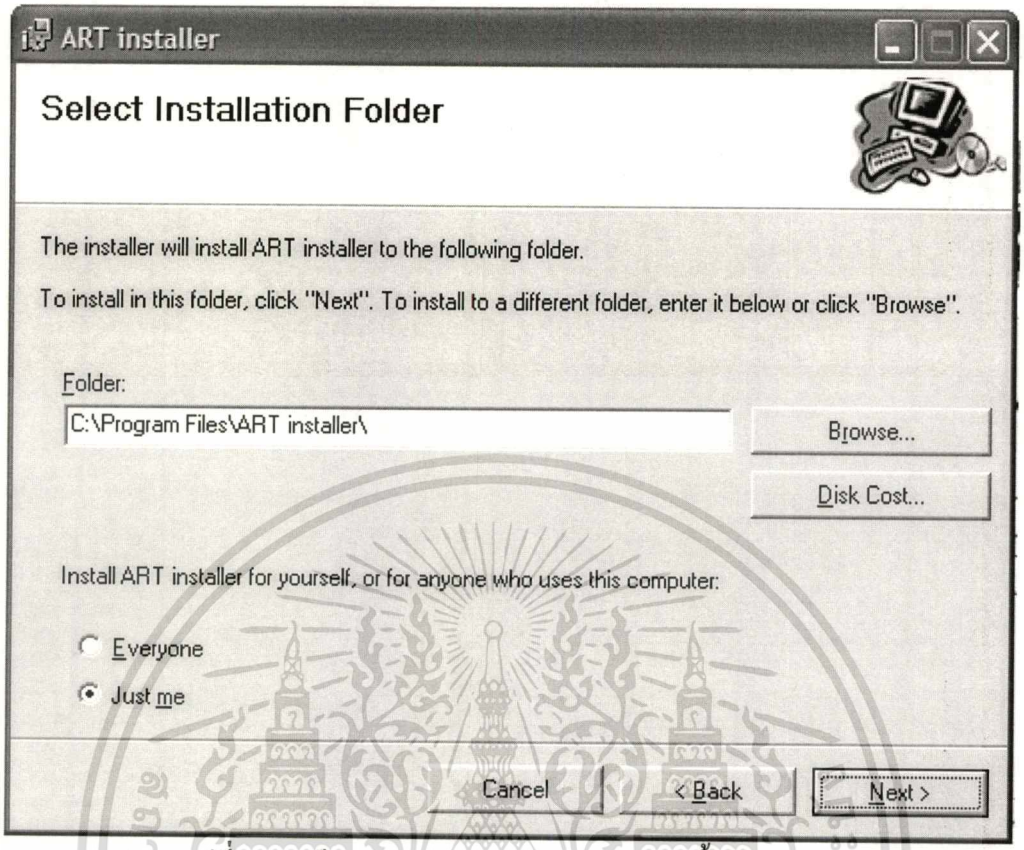
รูปที่ ก.1 หน้าจอแสดงไอคอนตัวติดตั้งโปรแกรม

1.2 เมื่อดับเบิ้ลคลิกที่ไอคอนตัวติดตั้งเรียบร้อยแล้ว จะปรากฏหน้าจอดังรูปที่ ก.2 นี้



รูปที่ ก.2 หน้าจอแสดงขั้นตอนการติดตั้งโปรแกรม

1.3 หลังจากนั้นให้คลิก Next จะปรากฏหน้าจอดังรูปที่ ก.3 นี้



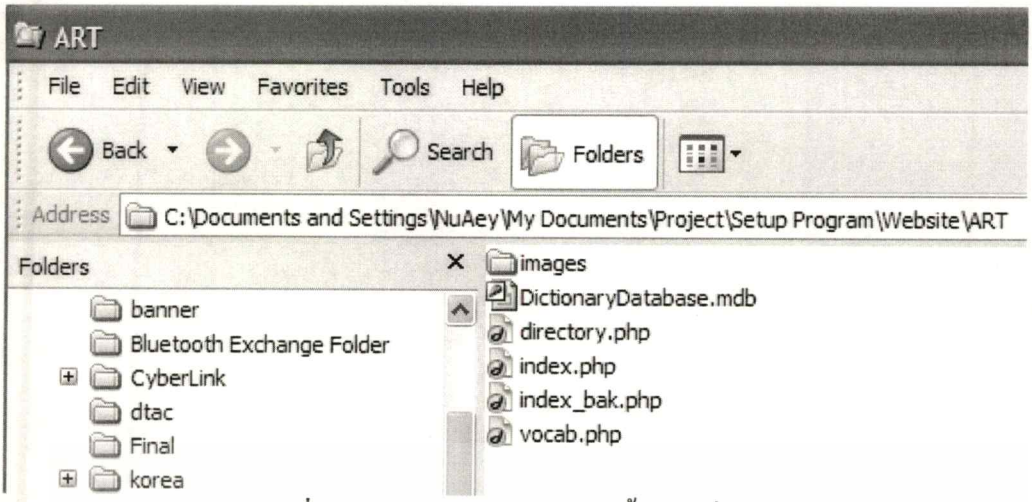
รูปที่ ก.3 หน้าจอแสดงการกำหนดพาธการติดตั้งโปรแกรม

1.4 หลังจากกำหนดพาธการติดตั้งโปรแกรมเรียบร้อยแล้ว คลิก Next ต่อไปจนเสร็จสิ้นขั้นตอนการติดตั้งโปรแกรม จะปรากฏไอคอนโปรแกรมบนหน้าจอเดสท็อบ ดังรูปที่ ก.4 นี้

รูปที่ ก.4 หน้าจอแสดงไอคอนของโปรแกรม

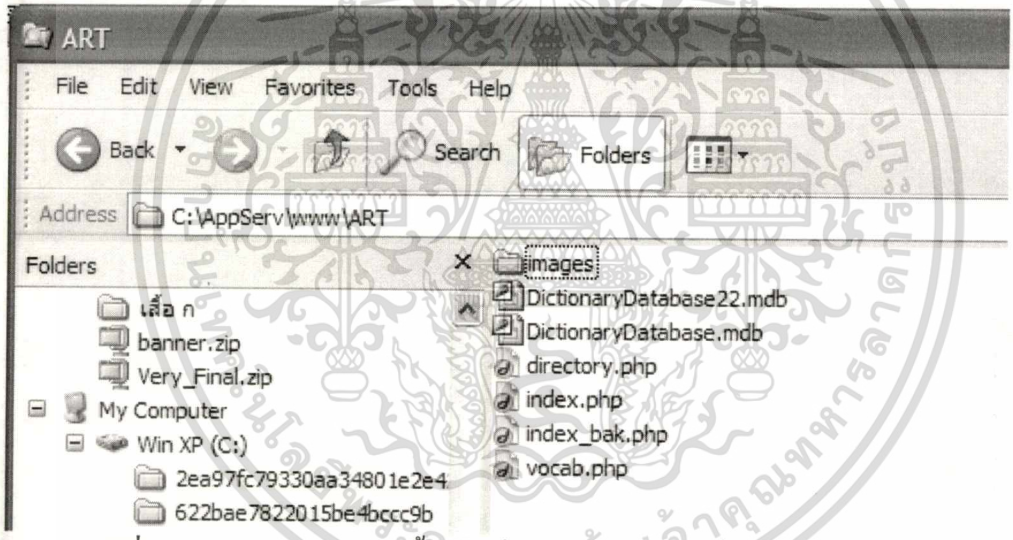
2. <u>การติดตั้งโปรแกรมส่วนเว็บแอพลิเคชั่น</u>

2.1 นำ file.php ทั้งหมดใน folder website ตามรูปที่ ก.5 ต่อไปนี้



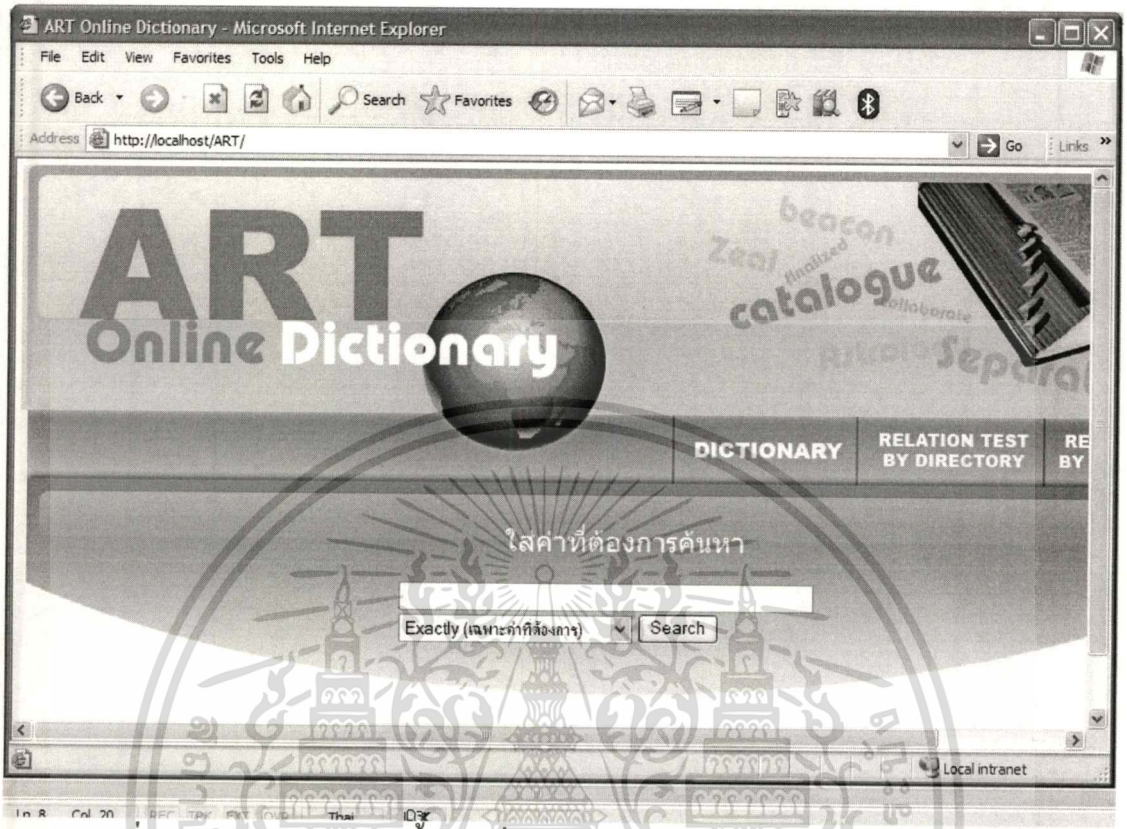
รูปที่ ก.5 หน้าจอแสดง file.php ทั้งหมดที่นำมาใช้งาน

2.2 ติดตั้งโปรแกรม Appserv ให้เสร็จสมบูรณ์ และนำ file ทั้งหมดวางใน Folder www ซึ่งอยู่ภายใต้ Folder AppServ ตามรูปที่ ก.6 ต่อไปนี้



รูปที่ ก.6 หน้าจอแสดง file ทั้งหมดที่นำมาวางใน path ที่เกิดจากโปรแกรม Appserv

2.3 เรียกดู website ด้วย URL “http://localhost/ART ตามรูปที่ ก.7 ต่อไปนี้

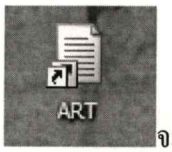


รูปที่ ก.7 หน้าจอแสดง file ทั้งหมดที่นำมาวางใน path ที่เกิดจากโปรแกรม Appserv

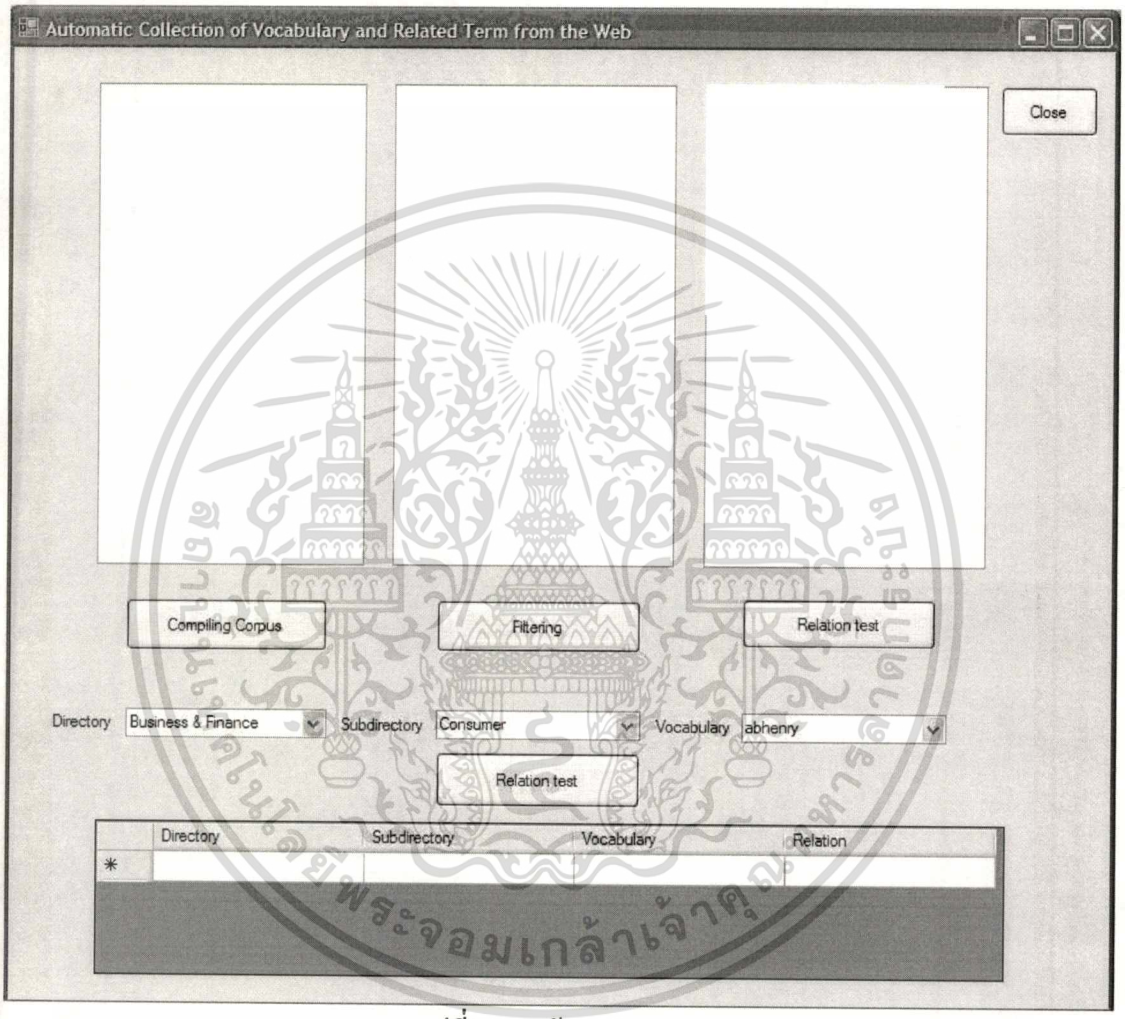
ภาคผนวก ข.

คู่มือการใช้งานระบบ

คู่มือการใช้งานระบบการหาค่าความสัมพันธ์ระหว่างคำศัพท์กับหมวดหมู่

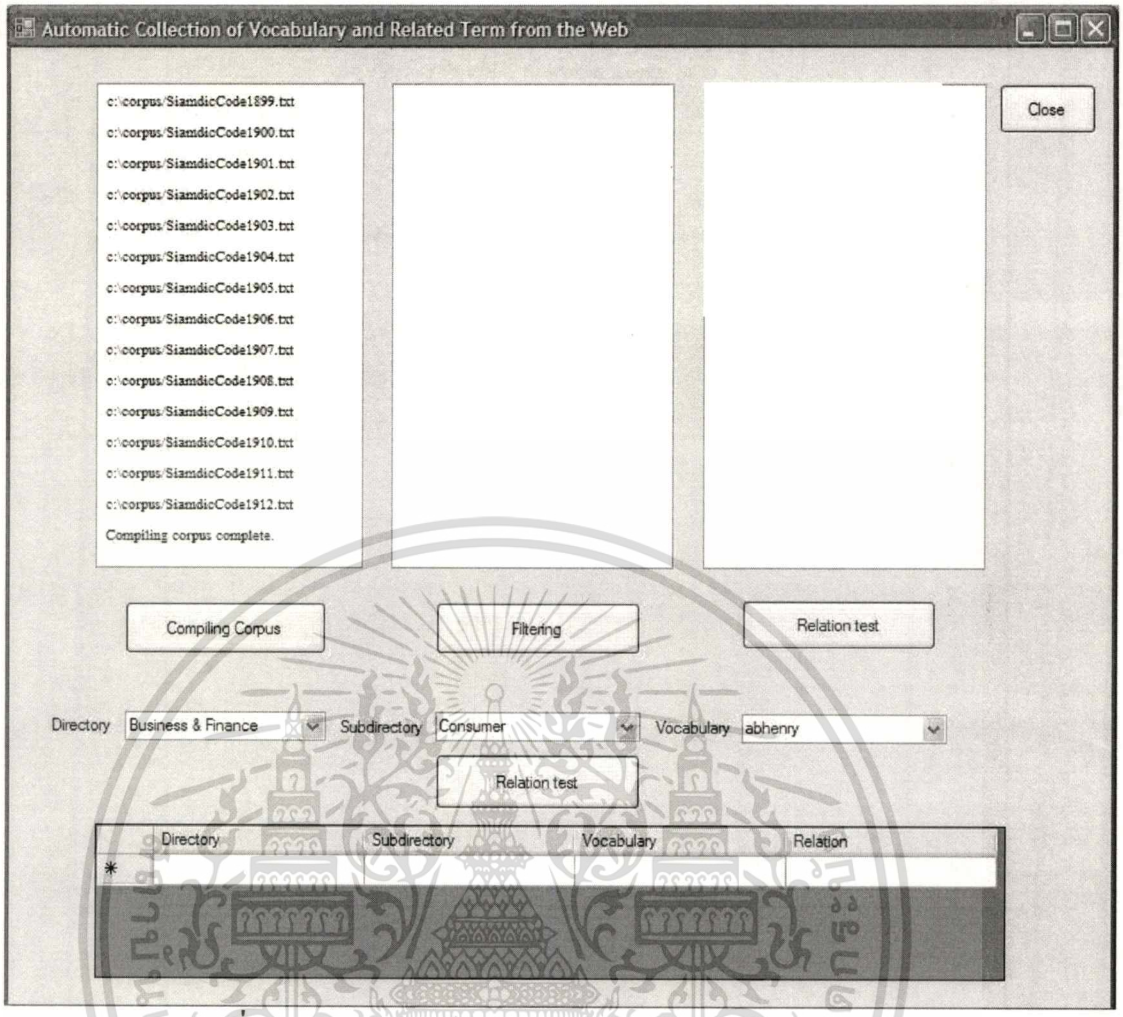
1. เข้าสู่โปรแกรมด้วยการดับเบิ้ลคลิกที่ไอคอนโปรแกรม 

จะปรากฏหน้าจอเมนูหลักดังรูป ข.8 นี้



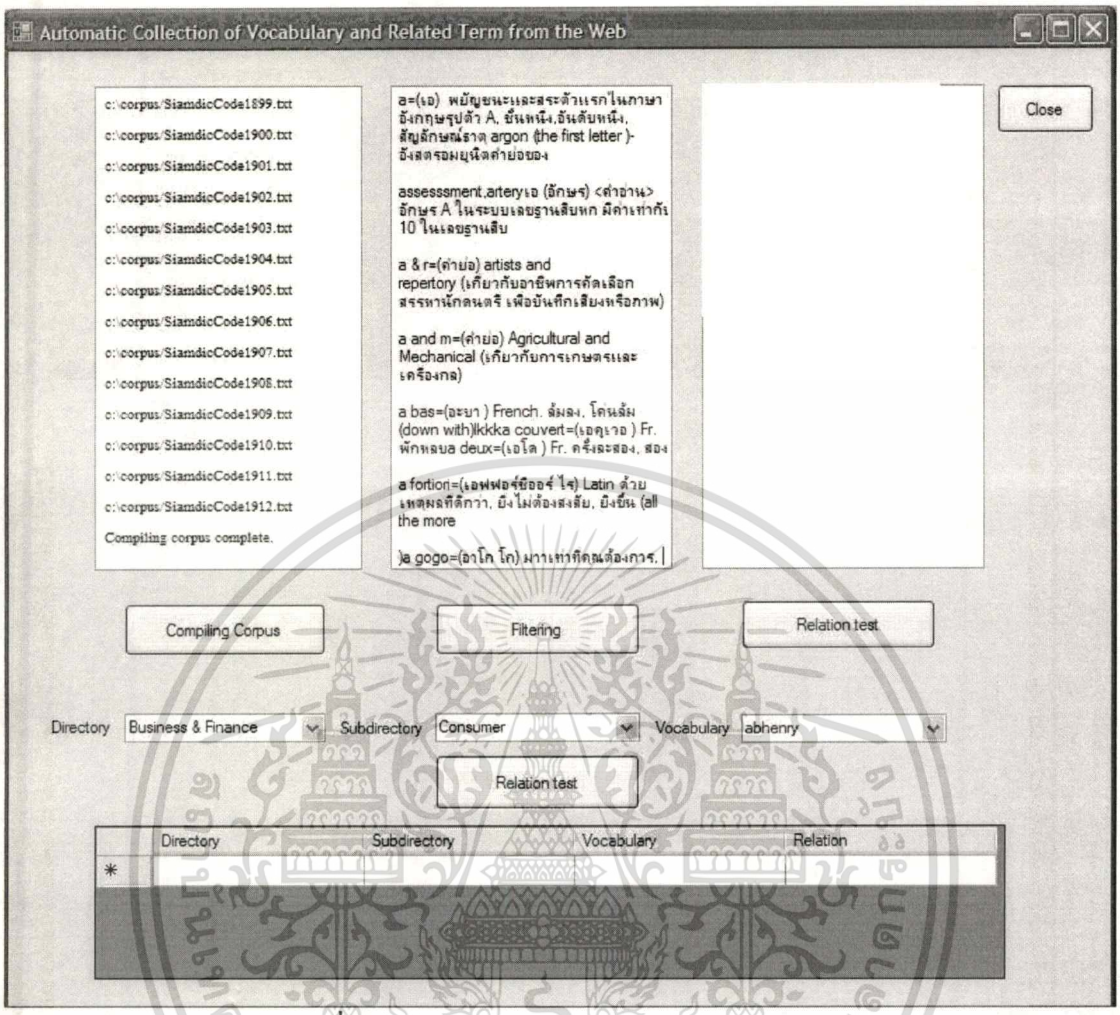
รูปที่ ข.8 หน้าจอเมนูหลัก

2. กดปุ่ม Compiling Corpus ซึ่งเมื่อทำการกดแล้ว การทำงานคือการเก็บรวบรวมข้อมูลจากเวปไซต์ เมื่อการทำงานเสร็จสมบูรณ์จะปรากฏหน้าจอเมนูหลักดังรูป ข.9 นี้



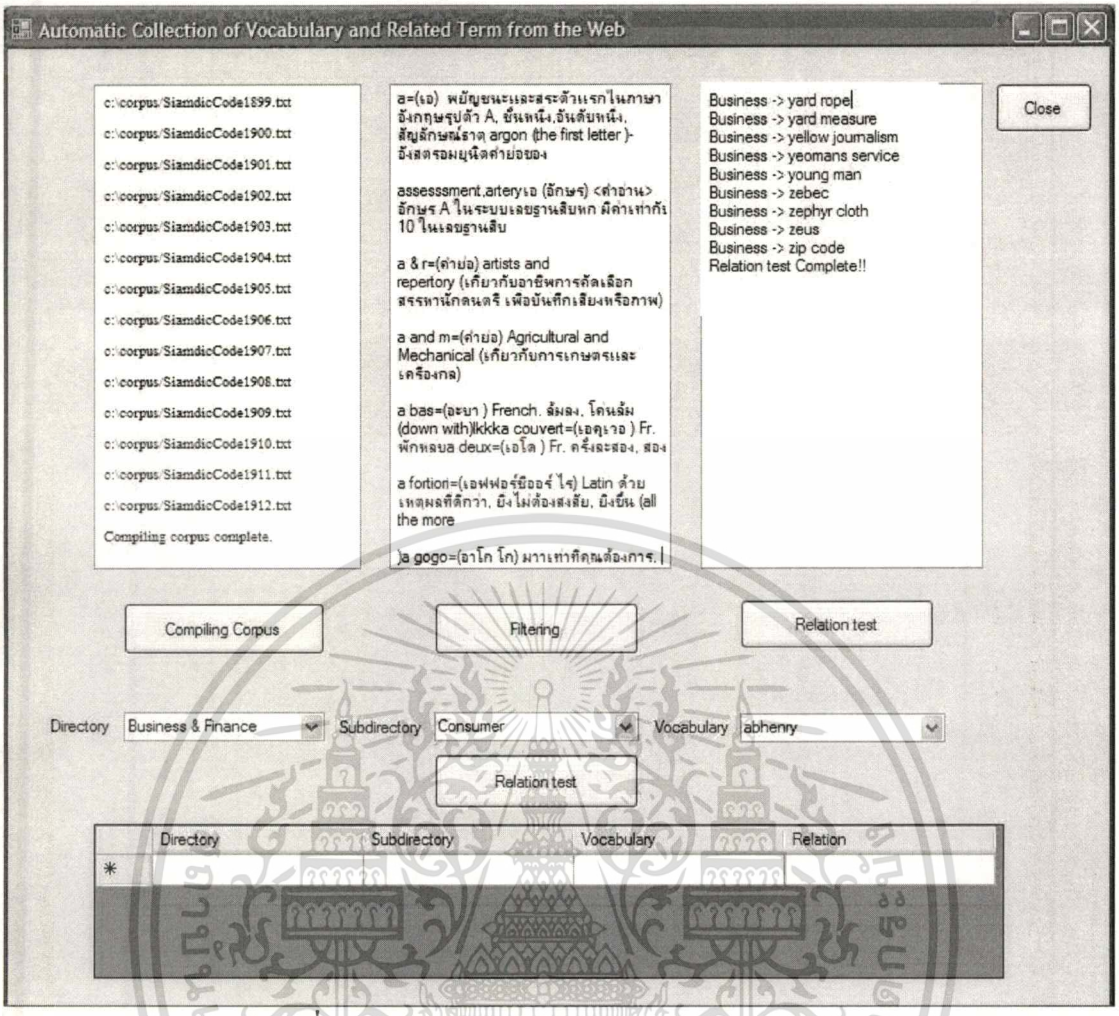
รูปที่ ข.9 แสดงผลลัพท์จากการกดปุ่ม Compiling Corpus

3. กดปุ่ม Filtering ซึ่งเมื่อทำการกดแล้ว การทำงาน Filtering เพื่อดึงคำศัพท์และคำแปลออกมาจาก Corpus ที่เรารวบรวมได้จากการทำ compiling corpus เมื่อการทำงานเสร็จสมบูรณ์จะปรากฏหน้าจอเมนูหลักดังรูป ข.10



รูปที่ ข.10 แสดงผลลัพท์จากการกดปุ่ม Filtering

4. การทำการกดปุ่ม Relation test เพื่อทำการหาความสัมพันธ์ระหว่างคำศัพท์และหมวดหมู่ โดยจะเริ่มทำงานเมื่อมีการกดปุ่ม Relation test เมื่อการทำงานเสร็จสมบูรณ์จะปรากฏหน้าจอเมนูหลักดังรูป ข.11



รูปที่ ข.11 แสดงผลลัพท์จากการกดปุ่ม Relation test

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวเอมอร สิริศุภางค์
สถานที่เกิด	จังหวัดกรุงเทพ
การศึกษา	ระดับปริญญาตรี
	วส.บ. (วิทยาการสารสนเทศบัณฑิต)
	สาขาวิชาวิทยาการคอมพิวเตอร์
	มหาวิทยาลัยธรรมศาสตร์
ประสบการณ์การทำงาน	Programmer
	Samart Info media
	Programmer
	Ictus
	Software testing engineer
	Microsoft(Thailand)