โปรแกรมคัดแยกประเภทหน้าเว็บด้วยแนวคิดแบบ TF.IDF

WEB PAGE CATEGORIZATION PROGRAM WITH TF.IDF METHODOLOGY

โดย

ณรงค์พันธ์ ปาการเสรี

NARONGPHAN PAKARNSEREE

อาจารย์ที่ปรึกษา

ผศ.ดร.พรฤดี เนติโสภากุล

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 1 ปีการศึกษา 2551

WEB PAGE CATEGORIZATION PROGRAM

WITH TF.IDF METHODOLOGY

NARONGPHAN PAKARNSEREE

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

1/2008

หัวข้อ	โปรแกรมการคัดแยกประเภทหน้าเว็บ
นักศึกษา	นายณรงค์พันธ์ ปาการเสรี
รหัสนักศึกษา	47066647
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2551
อาจารย์ที่ปรึกษา	ผศ.ดร.พรฤดี เนติโสภากุล

บทคัดย่อ

รายงานฉบับนี้เกิดขึ้นจากการศึกษา, ค้นคว้าและ วิเคราะห์ในเรื่องการคัดแยกประเภทหน้าเว็บด้วยวิธีการแบบ TF.IDF ซึ่งทำงานร่วมกับทฤษฎีต่างๆเช่น เวคเตอร์สเปซโมเดลและการหาค่ามุมความสัมพันธ์โคไซน์ฯลฯ เพื่อนำความรู้ที่ได้ทำการค้นคว้านี้ไปพัฒนาโปรแกรมการคัดแยกประเภทหน้าเว็บ (Web categorization) ซึ่ง TF.IDF เป็นแนวคิดที่สามารถค้นหาคำที่เป็นคำสำคัญของเอกสารนั้นๆได้อย่างยอดเยี่ยมทั้งยังสามารถบอกถึงน้ำหนักความสำคัญของทุกๆคำในเอกสารนั้นๆได้ หากแต่แนวคิดแบบ TF.IDF นี้ ไม่สามารถวิเคราะห์จากเอกสารเพียงเอกสารเดียวได้ซึ่งต้องวิเคราะห์จากกองเอกสาร จึงจำเป็นที่จะต้องหาวิธีในการเก็บข้อมูลเว็บตัวอย่างซึ่งมีอยู่มากมายหลายวิธี ค่า TF.IDF นี้ยังนำไปใช้ในการคำนวณเพื่อหาค่าความสัมพันธ์ระหว่างเว็บแต่ละเว็บในข้อมูลเว็บตัวอย่างโดยดูจากค่ามุมความสัมพันธ์โคไซน์ ซึ่งได้มีความพยายามนำความรู้เหล่านี้ไปพัฒนาเป็นโปรแกรมการคัดแยกประเภทหน้าเว็บ หลังการพัฒนานั้นทำให้ได้เข้าใจในสิ่งที่ศึกษามามากยิ่งขึ้นเนื่องจากได้ทำการทดลองจริง และหวังว่าโปรแกรมที่พัฒนาขึ้นมานี้อาจเป็นส่วนหนึ่งให้ผู้ที่สนใจสามารถนำไปทดลองใช้เพื่อการศึกษาและช่วยนำไปพัฒนาให้ดียิ่งๆขึ้น

Title	Web Page Categorization Program
Student	Mr.Narongphan Pakarnseree
Student ID.	47066647
Degree	Master of Science
Programme	Information Science
Academic Year	2008
Advisor	Asst.Prof. Dr.Ponrudee Netisopakul

ABSTRACT

This document is created based on the study, research and analysis of the Web Categorization using TF.IDF combined with other theories such as Vector Space Model and Cosine Similarity. To bring this findings into development of Web categorization program which TF.IDF is used for searching important word list of any text together with weight of each word. TF.IDF weight can be used for calculation of the Web site relationship by using the angle of cosine. Fortunately, I have brought this research to develop the Web categorization program and find out more understandings due to the experiment. I hope that this developped program and report is useful for people who interested in and would like to develop programs to have more functions and more stable.

กิตติกรรมประกาศ

โปรเจ็คการพัฒนาโปรแกรมการคัดแยกประเภทหน้าเว็บนี้สำเร็จลุล่วงไปได้ด้วยดีต้องขอขอบพระคุณมารดา,คุณน้า,คุณลุง,คุณป้า,พี่ๆและญาติทุกท่านที่คอยให้ความช่วยเหลืออนุเคราะห์, สนับสนุนและให้กำลังใจข้าพเจ้าอยู่เสมอมา ขอขอบพระคุณอาจารย์ ผศ.ดร.พรฤดี เนติโสภากุล ซึ่งเป็นอาจารย์ที่ปรึกษาโปรเจ็ค, วิชาสัมมนา1และสัมมนา2 ที่อาจารย์คอยให้คำปรึกษาและให้คำแนะนำอยู่เสมอ ทั้งนี้ขอขอบพระคุณอาจารย์ทุกท่านที่ให้ความรู้ในด้านต่างๆ และจะขาดไปเสียไม่ได้ที่จะขอบคุณเป็นอย่างมากสำหรับเพื่อนของข้าพเจ้า นายกิตจา เพ็งส้ม ซึ่งคอยช่วยเหลือและให้คำปรึกษาในด้านการพัฒนาโปรแกรมมาโดยตลอดและสุดท้ายนี้ขอขอบคุณเพื่อนๆที่เรียนมาด้วยกันทุกท่านด้วย

ณรงค์พันธ์ ปาการเสรี

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

รูปที่ หน้า

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญในการพัฒนาโปรแกรมการคัดแยกประเภทหน้าเว็บ

การพัฒนาโปรแกรมการคัดแยกประเภทหน้าเว็บเกิดขึ้นเนื่องจาก เว็บไซต์ในปัจจุบัน (Website) เป็นแหล่งข้อมูลขนาดใหญ่ที่ใช้งานและเข้าถึงข้อมูลได้ง่ายกว่าห้องสมุดเป็นอย่างมาก หากแต่การค้นหาข้อมูลด้วยผู้ให้บริการสืบค้นข้อมูล (Search engine) ที่ยังไม่ได้ใช้วิธีการคัดแยกประเภทหน้าเว็บ ผู้ใช้อาจจะไม่ได้ข้อมูลที่ตรงกับคำถามคิวรี(Query) ตามความเข้าใจของมนุษย์ที่ได้ทำการป้อนเข้าไปเช่น ค้นหาคำว่า "car company" โดยผู้ใช้คาดว่าจะได้ผลลัพธ์เป็นรายชื่อผู้ผลิตรถยนต์เช่น โตโยต้า, ฮอนด้า, นิสสัน เป็นต้น แต่ผลลัพธ์ที่ได้กลับกลายเป็น เว็บไซต์ประเภทอื่นๆที่มีคำว่า "car company" อยู่ในเว็บไซต์แทน ซึ่งเว็บไซต์อย่างโตโยต้าอาจมีคำว่า "car" แต่ไม่มีคำว่า "company" ภายในเว็บไซต์ซึ่งทำให้ผลลัพธ์อาจไปปรากฏอยู่ท้ายๆหรือไม่ปรากฏเลยก็ได้ ดังนั้นหากเราได้ทำการศึกษาและพัฒนาโปรแกรมการคัดแยกประเภทของหน้าเว็บให้เป็นหมวดหมู่จะช่วยให้เกิดความสะดวกสบายในการสืบค้นและ ช่วยให้คอมพิวเตอร์ทำงานได้ใกล้เคียงกับความเข้าใจของมนุษย์มากยิ่งขึ้น

การคัดแยกประเภทหน้าเว็บ(Web page Categorization) นั้นยากยิ่งกว่าการคัดแยกประเภทข้อความล้วนๆ (Pure text categorization : Pure TC) เป็นอย่างมาก ด้วยเหตุที่หน้าเว็บ)Web page(นั้นมีความหลากหลายในลักษณะรูปแบบการใช้งานของข้อมูลอยู่หลายรูปแบบด้วยกัน ซึ่งข้อมูลแต่ละแบบนั้นส่วนใหญ่ล้วนสามารถนำมาใช้เป็นเกณฑ์ในการคัดแยกประเภทหน้าเว็บได้ แต่ก็ขึ้นอยู่กับอัลกอริธึม(Algorithm)ที่จะนำมาใช้ร่วมกันกับข้อมูลในแบบนั้นๆด้วยว่ามีความเหมาะสมกันมากน้อยขนาดไหน และสุดท้ายภายในโปรแกรมจะต้องมีส่วนรายงานการวัดประสิทธิภาพผลลัพธ์การคัดแยกประเภทของเว็บระหว่างข้อมูลกับแนวคิดทั้ง TF-.IDF และ VSM ที่นำมาใช้ร่วมกันว่าจะทำให้เกิดความถูกต้องเพียงใดเพื่อนำไปประยุกต์ใช้ให้เกิดประสิทธิภาพสูงสุด

1.2 วัตถุประสงค์ของการพัฒนาโปรแกรมคัดแยกประเภทหน้าเว็บมีดังต่อไปนี้

1.2.1 หาความสัมพันธ์ของเว็บแต่ละคู่ว่ามีความสัมพันธ์กันมากน้อยเพียงไร โดยใช้วิธีการ Cosine similarity ในการวัด

1.2.2 โปรแกรม Web page Categorization พัฒนาขึ้นมาด้วยแนวคิดแบบ TF.IDF เพื่อทำการทดสอบทฤษฎีที่ได้เรียนรู้และศึกษามา นอกจากนั้นยังช่วยให้รู้จักประยุกต์ทฤษฎีมาพัฒนาโปรแกรม

1.2.3 เป็นตัวอย่างหนึ่งสำหรับผู้ที่สนใจการคัดแยกประเภทหน้าเว็บให้สามารถนำแนวคิด TF.IDF, VSM จากโปรแกรมนี้ไปพัฒนาการทำงานของโปรแกรมคัดแยกประเภทหน้าเว็บอื่นต่อไปได้ เพื่อให้ผู้ใช้งานเว็บสามารถสืบค้นและเข้าถึงข้อมูลที่ได้ทำการค้นหาผ่านโปรแกรมที่พัฒนาได้อย่างถูกต้องแม่นยำและรวดเร็วยิ่งขึ้น

1.2.4 ได้ทักษะและความรู้ด้านการคัดแยกตัวอักษร,คำและประโยคเพื่อสามารถนำไปพัฒนาประยุกต์ใช้กับการคัดแยกประเภทหน้าเว็บซึ่งมีข้อมูลที่ซับซ้อนมากยิ่งขึ้นได้

1.3 ขอบเขตของโปรแกรมคัดแยกประเภทหน้าเว็บ

1.3.1 โปรแกรมทำการคัดแยกประเภทหน้าเว็บจะคัดแยกเฉพาะหน้าเว็บที่เป็นภาษาอังกฤษเท่านั้น

1.3.2 การเก็บข้อมูลของประเภทหน้าเว็บจะจัดเก็บข้อมูลของทั้งเว็บโดยใช้โปรแกรมจัดเก็บเว็บตัวอย่างที่ทำงานร่วมกับโปรแกรมแชร์แวร์ชื่อ Httrack โดยข้อมูลเว็บตัวอย่างทั้งหมดจะถูกเก็บอยู่ในรูปโฟลเดอร์

1.3.3 โปรแกรมจะทำการดาวน์โหลด (Download) กี่ครั้งก็ได้ครั้งละ 1 คำที่สนใจ (Query) แต่ในรายงานฉบับนี้การทดลองจะค้นหาคำที่สนใจ 10 คำ และจะเก็บจำนวนเว็บตัวอย่างทั้งหมดเท่ากับ 500เว็บ (50 เว็บต่อหนึ่งคำที่สนใจ)

1.3.4 นำข้อมูลเว็บตัวอย่างที่เก็บมาได้มาผ่านกระบวนการคำนวณในเว็บกลุ่มตัวอย่างเพื่อให้ได้ค่าน้ำหนักของแต่ละเว็บและแต่ละคำที่สนใจเพื่อนำไปใช้ในการคำนวณต่อไป

1.3.5 ค่าน้ำหนักที่คำนวณได้จะนำไปใช้เพื่อคำนวณในสูตรเพื่อหาความสัมพันธ์ระหว่างเว็บเป็นค่ามุม ø ต่อไป

1.3.6 ค่าของมุม ø ที่ได้ออกมานี้ สามารถบอกความสัมพันธ์ระหว่างเว็บสองเว็บว่ามีความหมายใกล้เคียงกันมากน้อยเพียงไรซึ่งรายงายสรุปจะแสดงให้เห็นว่ากลุ่มของเว็บใดมีความใกล้เคียงกันเช่นกลุ่มของเว็บตัวอย่างที่เก้บมาจากคำว่า Cook กับ Nutrition จะมีความสัมพันธ์กันมากกว่า Cook กับ Dog หรือ Cook กับ Guitar เป็นต้น

1.3.7 ส่วนรายงานจะทำการสรุปรายงานในรูปของตาราง และนำตารางผลลัพธ์ที่ได้ไปเขียนรายงานการทดลองต่อไป

1.4 เครื่องมือที่ใช้พัฒนาโปรแกรม

โปรแกรมการคัดแยกประเภทหน้าเว็บใช้เครื่องมือ (Tools) ในการพัฒนาโปรแกรมในหลายๆส่วนดังต่อไปนี้

ตารางที่1.1 แสดงเครื่องมือที่ใช้ในการพัฒนาโปรแกรมในแต่ละส่วน

ส่วนการทำงานในโปรแกรม	เครื่องมือที่ใช้ในการพัฒนาโปรแกรม
- ส่วนเก็บข้อมูลเว็บตัวอย่าง	- HTTrack, Internet Explorer7.0, Google.com
- ส่วนโปรแกรมคำนวณ	- Visual Studio2005 (C#.NET)

1.5 เครื่องมือที่ใช้ในการจัดทำเอกสารฉบับนี้

ตารางที่1.2 แสดงเครื่องมือที่ใช้ในการจัดทำรายงาน

ส่วนการทำงาน	เครื่องมือที่ใช้ในการพัฒนาโปรแกรม
- เอกสาร	- Microsoft Office Word 2007
- โมเดลต่างๆ	- Microsoft Office Visio 2007
- ส่วนรายงาน	- Microsoft Office Excel 2007

บทที่ 2
ความรู้และแนวคิดที่เกี่ยวข้องกับการพัฒนาการคัดแยกประเภทหน้าเว็บ

2.1 ข้อมูลในหน้าเว็บที่สามารถนำข้อมูลเหล่านี้มาเป็นเกณฑ์ในการคัดแยกประเภท

หน้าเว็บนั้นมีความซับซ้อนมากกว่าชุดข้อมูลตัวหนังสือธรรมดา เนื่องจากหน้าเว็บประกอบไปด้วยข้อมูลหลากหลายชนิด ทำหน้าที่ต่างกันออกไปแต่ละส่วนนั้นสามารถนำมาใช้เป็นเกณฑ์ในการคัดแยกประเภทหน้าเว็บได้ ซึ่งสามารถแบ่งออกเป็นหลักๆที่สำคัญได้ดังต่อไปนี้

- URLs text
- Meta data text
- Title text
- Contents (Body text)
- Link (Anchor text)
- Document

2.1.1 URLs text

URLs (Uniform Resource Locators) text คือชุดตัวอักษรที่ทำหน้าที่บอกแหล่งที่อยู่ของหน้าเว็บมีโครงสร้างดังต่อไปนี้

scheme://host/path-elements/document.extension

-scheme แทนโครงสร้างโปรโตคอลที่ใช้ เช่น http หรือ ftp เป็นต้น

-host บอกถึงชื่อที่ได้ทำการจดทะเบียนโดเมนเนมเช่น www.hotmail.com เป็นต้น

-path-element ใช้อ้างอิงถึงโฟลเดอร์ที่อยู่ภายใน host server นั้นๆ เช่นโฟลเดอร์ mails

-document อ้างถึงชื่อของเอกสารเช่น default

- extension อ้างถึงนามสกุลของไฟล์เอกสารเช่น htm, html, asp, aspx, php เป็นต้น

2.1.2 Meta data text

Meta data text เป็นข้อความที่อยู่ภายใน tag meta <meta> ... </meta> เดิมทีใช้เพื่อช่วยให้ผู้ทำเว็บบอกถึงรายละเอียดที่มีอยู่ภายในเว็บไซต์ ช่วยให้การค้นหาข้อมูลทำได้ง่ายยิ่งขึ้นแต่ ผู้ทำเว็บส่วนใหญ่มักใส่คำที่คนนิยมค้นหาเข้าไปด้วยทั้งที่เว็บตนเองไม่มีเนื้อหาในส่วนนั้นด้วยเช่น คำว่า

"sex" ทำให้ Metadata มีความน่าเชื่อถือในการสืบค้นต่ำ tag meta ต้องอยู่ภายใน tag head มีโครงสร้างดังต่อไปนี้

<meta name="meta type" content="desc01, desc02">

meta type แบ่งออกเป็น Description, Keyword เป็นต้น

2.1.3 Title text

Title text คือหัวเรื่องในหน้าเว็บนั้นๆ tag title อยู่ภายใน tag head เช่นเดียวกับ tag meta มีโครงสร้างดังต่อไปนี้เช่น

<title>New York Times</title>

2.1.4 Contents (Body text)

Content (Body text) คือข้อความที่เป็นเนื้อหา ส่วนที่บอกถึงรายละเอียดต่างๆ ภายในเว็บไซต์เนื่องจากเป็นส่วนที่มักมีข้อมูลขนาดใหญ่แต่มีคำที่ไม่จำเป็นจะนำมาใช้ในการคัดแยกเป็นจำนวนมาก จึงควรทำการกลั่นกรอง (Refine) ข้อความก่อนจะนำไปใช้คัดแยกประเภทหน้าเว็บ

2.1.5 Link (Anchor text)

Link (Anchor text) เป็นชุดข้อความที่ใช้เพื่อเชื่อมโยงไปยังเว็บที่ต้องการได้ Link มีโครงสร้างดังต่อไปนี้

add this book to cart

2.1.6 Document

Document เป็นการพิจารณาเอกสารทั้งเอกสารไม่ได้แบ่งเนื้อหาออกเป็นส่วนๆ การพิจารณาทั้งเอกสารนี้หากมองโดยปกติอาจเห็นได้ว่าน่าจะให้ผลลัพธ์การวิเคราะห์ด้อยกว่าการพิจารณาบางส่วนที่สำคัญในกรณีของการวิเคราะห์ข้อมูลชนิดหน้าเว็บเนื่องจาก หน้าเว็บมักจะประกอบไปด้วย html tag และ สคริปต์ข้อความขยะต่างๆจำนวนมาก แต่แท้ที่จริงแล้วการพิจารณาทั้งเอกสารสามารถทำได้และมีประสิทธิภาพเช่นเดียวกันหากนำไปใช้ร่วมกับแนวคิด, วิธีการที่เหมาะสม ตัวอย่างเช่นวิธีการแบบ TF.IDF เนื่องจาก TF.IDF สามารถกรองคำที่ไม่มีความสำคัญในการพิจารณาออกไปได้ หากลองพิจารณาค้นหาคำสำคัญในเว็บไซต์หลายๆหน้าแล้ว แน่นอนว่าถ้าพิจารณาทั้งเอกสารจะต้องมีโอกาสในการพบ html tag จำนวนมากแต่ด้วยความสามารถของเทอม IDF ซึ่งให้ค่าการพบเทอมนั้นๆหลายๆครั้ง มีค่าเข้าใกล้ศูนย์ทำให้ html tag ไม่มีผลนำมาใช้ในการพิจารณา (อ่านรายละเอียดเพิ่มเติมการพิจารณาค่า IDF ในหัวข้อถัดไป 2.2 แนวคิดและวิธีการคัดแยกประเภทหน้าเว็บ)

2.2 แนวคิดและวิธีการคัดแยกประเภทหน้าเว็บ

แนวคิดและวิธีการคัดแยกประเภทหน้าเว็บนั้นมีผู้คิดและทำการวิจัยออกมาเป็นจำนวนมาก จะเห็นว่าส่วนใหญ่มีอัลกอริธึมที่ไม่เหมือนกันต่างก็พยายามคิดวิธีการและอัลกอริธึมที่ช่วยให้ได้รายชื่อประเภทของหน้าเว็บผลลัพธ์ที่ถูกต้องและแม่นยำที่สุด โดยเอกสารฉบับนี้จะยกตัวอย่างแนวคิดที่นำมาประยุกต์ใช้ในโปรแกรมที่ทำการพัฒนาดังต่อไปนี้

2.2.1 การคัดแยกประเภทหน้าเว็บโดยใช้แนวคิดแบบ TF.IDF (Term Frequency Inverse Document Frequency) Salton, G. and Buckley, C. (1988 : 513–523)

การคำนวณค่าน้ำหนักของ TF.IDF นั้นเป็นที่นิยมกันอย่างมากใช้ในเรื่องการค้นคืนข้อมูล (Information Retrieval) และเหมืองข้อความ (Text Minings) น้ำหนักดังกล่าวเป็นตัววัดทางสถิติที่ช่วยในการประเมินความสำคัญของคำนั้นๆในเอกสารจำนวนหนึ่ง สัดส่วนของคำที่ปรากฏในเอกสารเป็นส่วนสำคัญในการคำนวณ ค่าน้ำหนักของ TF.IDF นี้มักใช้เป็นเครื่องมือหลักสำหรับเว็บผู้ให้บริการด้านสืบค้นข้อมูล (Search Engine) เป็นอย่างมากทั้งการให้คะแนน (Scoring) และจัดอันดับ (Ranking)

2.2.1.1 ค่าความถี่ของคำ (Term Frequency)

ค่าความถี่ของคำ(Term Frequency) ก่อนอื่นเราต้องเข้าใจความหมายของคำว่าเทอม เทอมนั้นหมายถึงคำที่เราสนใจและต้องการค้นหาในเอกสารเช่น หากเราสนใจที่จะหาคำว่า "bank" ดังนั้น "bank" จึงถือว่าเป็นเทอมๆหนึ่งหรือที่จะเรียกในเอกสารนี้ว่าเทอมที่สนใจแทนค่าด้วย t (term) และหากมีหลายๆเทอมที่สนใจมากกว่า 1 ขึ้นไปเราก็จะเรียกเป็น $t_1, t_2, ...,t_n$ และจะแทนค่าคำ (เทอม)ทั้งหมดที่มีในเอกสารด้วย T (All terms in document)

เอกสาร(Document) ในที่นี่เอกสารที่จะนำมาพิจารณาในเรื่อง TF.IDF จะต้องเป็นเอกสารประเภทเอกสารตัวอักษร(Text Document) ซึ่งอาจมีสิ่งที่ไม่ใช่ตัวอักษรอยู่ด้วยก็ได้เช่นภาพทั่วไป หรือภาพที่แสดงเป็นตัวอักษรเพียงแต่เราไม่สามารถนำสิ่งอื่นที่ไม่ใช่ตัวอักษรเข้ามาพิจารณาร่วมด้วย ในเอกสารฉบับนี้เราจะแทนค่าเอกสารด้วย d หรือ document นั้นเองซึ่งถ้ามีหลายเอกสารจะแสดงเป็นหรือ $d_1, d_2, ...,d_n$ และถ้าพูดถึงจำนวนเอกสารทั้งหมดจะแทนค่าด้วย D (All documents)

การที่จะได้มาของจำนวนครั้งของเทอมที่สนใจ ที่ปรากฏอยู่ในเอกสารหนึ่งๆ ค่าความถี่ (Frequency) ที่นับมาได้นี้จะถูกนำเข้าสู่การนอร์มัลไลซ์ (Normalized) เพื่อป้องกันการถ่วงน้ำหนักในเอกสารนั้นจากสูตร

$$tf = t / T \qquad (2.1)$$

จากสมการข้างต้นหมายถึงค่าความถี่ของเทอม TF (Term Frequency) เท่ากับ t จำนวนครั้งที่พบเทอมที่สนใจในเอกสารd หารด้วย T ซึ่งหมายถึงจำนวนคำ(เทอม)ทั้งหมดที่อยู่ในเอกสาร d

ดังนั้นถ้าต้องการพิจารณาหลายๆเทอมในหลายๆเอกสารก็จะได้เป็นสูตรดังต่อไปนี้

กำหนดให้ t_i เป็นที่สนใจที่พบในเอกสาร d_j โดย $n_{i,j}$ เป็นจำนวนของคำที่พบในเอกสารนั้นๆและนำมาหารด้วยจำนวนของคำทั้งหมดในเอกสาร d_j

$$tf_{i,j} = n_{i,j} / \sum_k n_{kj} \tag{2.2}$$

2.2.1.2 ค่าความถี่ผกผันของเอกสาร(Inverse Document Frequency)

ค่าความถี่ผกผันของเอกสาร(Inverse Document Frequency) เป็นตัววัดความสำคัญของทุกเทอมที่ปรากฏอยู่ในเอกสาร โดยการพิจารณาจากจำนวนเอกสารมากกว่าหนึ่งขึ้นไป โดยหาค่าได้เป็นค่าความถี่ผกผันซึ่งจะช่วยบอกความสำคัญของคำที่พบในเอกสารทั้งหมดได้เป็นอย่างมาก ว่าเป็น Term(คำที่สนใจ), Normal word(คำทั่วไป), หรือ Stop word (คำหยุด) สามารถคำนวณโดยการเอาค่าจำนวนเอกสารทั้งหมดหารด้วยจำนวนเอกสารที่พบคำที่สนใจแล้วคูณด้วยค่า log ซึ่งส่วนใหญ่จะใช้ค่า ln

$$idf_i = \log (|D| / |\{d_j : t_i \in d_j\}|) \tag{2.3}$$

โดยที่ $|D|$ เป็นจำนวนของเอกสารทั้งหมด

$|\{d_j : t_i \in d_j\}|$ เป็นจำนวนของเอกสารที่มีเทอมที่สนใจ t_i ในเอกสารนั้น(เมื่อ $n_{i,j} \neq 0$)

ดังนั้น

$$Tfidf_{i,j} = tf_{i,j} * idf_i \tag{2.4}$$

ความรู้เพิ่มเติมเกี่ยวกับ TF-IDF หลังจากที่ได้ศึกษานิยามและสูตรทั้งหมดของค่า TF และค่า IDF ไปแล้วจะทำให้เข้าใจได้ว่า

ค่า TF นั้นเป็นค่าน้ำหนักเฉพาะพื้นที่ (Local Weight) เท่านั้นเพราะเป็นค่าที่ได้มาจากการนับและคำนวณในเอกสารเพียงเอกสารเดียวดังนั้นค่าที่ได้จึงไม่สามารถนำไปอ้างอิงหรือใช้ในเอกสารอื่นได้เนื่องจากเป็นเพียงค่าในเอกสารนั้นๆ

ส่วนค่า IDF นับเป็นค่าน้ำหนักรวม (Global Weight) เนื่องจากค่า IDF คำนวณได้มาจากจำนวนเอกสารทั้งหมดที่เราได้ทำการพิจารณาหารด้วยจำนวนเอกสารที่พบเทอมที่สนใจในเอกสาร

นั้นๆและนำผลที่ได้ไปติด log ค่าที่ได้มีส่วนสำคัญอย่างมากในการคัดแยกคำในเอกสารว่าเป็นคำที่เป็นสาระสำคัญในบทความหรือไม่ โดยดูได้จากตัวอย่างด้านล่าง

ตัวอย่าง ถ้าเรามีเอกสารทั้งหมดในการพิจารณาจำนวน 1000 เอกสารแต่เราพบเอกสารที่มีคำที่เราสนใจเช่น "bird" อยู่ 10 เอกสารดังนั้นค่า IDF จะเท่ากับ log(1000/10) ซึ่งมีค่าเท่ากับ 2 แต่ถ้ามีเอกสารที่พบเทอมที่สนใจอยู่ถึง 1000 เอกสารก็จะได้ค่า IDF เท่ากับ log(1000/1000) มีค่าเท่ากับ 0 จะเห็นได้ว่าถ้าพบคำที่สนใจแทบจะทุกเอกสารหรือค่า d ที่นับได้เข้าใกล้ค่าจำนวนเอกสารทั้งหมด (D) ค่า IDF จะเข้าใกล้ศูนย์เมื่อนำไปคูณกับค่า TF ก็จะได้ค่าน้ำหนักเข้าใกล้ศูนย์เช่นกัน ด้วยเหตุนี้เนื่องจากคำที่มีโอกาสพบในทุกเอกสารมักจะเป็นคำที่ไม่มีความสำคัญเช่น คำหยุด(Stop Word) is, am ,are ,a ,an ,the ฯลฯ เป็นต้น และแน่นอนว่าโอกาสที่พบคำที่สนใจได้ในเอกสารจำนวนน้อยมากๆก็ไม่มีความน่าสนใจเช่นกัน จะเห็นได้ว่าค่าที่พบบ่อยมากๆกับพบได้น้อยเกินไปเราจะไม่สนใจทั้งคำนั้นทั้งคู่ ดังนั้นจึงถือได้ว่าค่า IDF เป็นค่าแปรผกผันที่ช่วยคัดแยกความสำคัญของคำได้อย่างมาก

2.2.2 เวคเตอร์สเปซโมเดล (Vector Space Model) (Wikipedia. 2007)

เวคเตอร์สเปซโมเดล (Vector Space Model) หรือ เทอมเวคเตอร์โมเดล(Term Vector Model) ถือเป็นโมเดลทางพีชคณิตที่ใช้ในการคำนวณเพื่อช่วยอธิบายเรื่อง เอกสารประเภทตัวอักษร (Text Document) โดยจะมองแต่ละเทอมในเอกสารเป็นแต่ละเวคเตอร์ (Vector) วิธีการทำเวคเตอร์สเปซโมเดลนี้ได้ถูกนำไปใช้คำนวณในเรื่องต่างๆอีกมากมายเช่น การกรองข้อมูล(Information Filtering) , การสืบค้นข้อมูล (Information Retrieval), การทำดัชนี (Indexing), การจัดอันดับความสัมพันธ์ (Relevancy Ranking)

เวคเตอร์สเปซโมเดลสามารถแบ่งออกได้เป็น 2 กรณีได้แก่

- เวคเตอร์สเปซโมเดลแบบไบนารี (Binary Vector Space Model)
  เวคเตอร์สเปซโมเดลแบบไบนารี คือเวคเตอร์สเปซโมเดลที่พิจารณาค่าเวคเตอร์เพียง 0 กับ 1 เท่านั้นโดยเวคเตอร์ใดๆจะมีค่า 0 เมื่อไม่พบเทอมที่สนใจในเอกสารและเวคเตอร์จะมีค่าเท่ากับ 1 เมื่อพบสิ่งที่สนใจในเอกสารเท่านั้น
- เวคเตอร์สเปซโมเดลแบบเชิงตัวเลข (Non-Binary Vector Space Model)
  เวคเตอร์สเปซโมเดลแบบเชิงตัวเลข คือเวคเตอร์สเปซโมเดลที่พิจารณาค่าเวคเตอร์จากการนับจำนวนครั้งในการพบและไม่พบเทอมที่สนใจในเอกสาร 0, 1, 2, 3,... ตามจริง

ค่า TF.IDF ของแต่ละเทอมที่คำนวณออกมาได้จากเอกสารทั้งหมด เราจะถือว่าเป็นค่าน้ำหนักแต่ละเทอมของแต่ละเอกสารนั้นๆ(Weight : W) จึงได้สูตรว่า

$$W_{t,d} = tf_t * \log (|D| / |\{t \in d\}|) \quad (2.5)$$

โดยที่ tf_t เท่ากับความถี่ของเทอม หรือ เทอม t ในเอกสาร d (ค่า local)

$\log (|D| / |\{t \in d\}|)$ เท่ากับค่าความถี่ผกผันของเอกสาร (ค่า global), $|D|$ มีค่าเท่ากับจำนวนเอกสารทั้งหมด, $|\{t \in d\}|$ คือค่าจำนวนของเอกสารที่พบเทอมที่สนใจ

เพราะฉะนั้นค่าน้ำหนักของเทอมอย่างง่ายที่สุดที่พบจากเอกสารเพียงเอกสารเดียวจึงมีค่าเท่ากับ

$$W_{t,d} = tf_t \quad (2.6)$$

จากการศึกษาจะพบว่าค่าน้ำหนักของเทอมจะเพิ่มมากขึ้น เมื่อ tf_t มีค่ามากขึ้นหรือเรียกได้ว่าพบจำนวนคำที่สนใจในเอกสารเป็นจำนวนมากในเอกสารนั้น

ค่าน้ำหนักของทุกๆเทอมในเอกสารหนึ่งๆนั้นรวมกันจะเรียกว่าเป็นเวคเตอร์ของเอกสารนั้นๆจากสูตร

$$V_d = [W_{1,d}, W_{2,d}, \ldots, W_{N,d}]^T \quad (2.7)$$

หากเปรียบเทียบเอกสารหนึ่งเป็นเว็บหนึ่งเว็บเช่น google.com ดังนั้น $V_{google.com}$ ก็จะเท่ากับค่าน้ำหนักของแต่ละเทอม (เช่น W_{earth}, W_{map}, $W_{calendar}$, ...) ทั้งหมดประกอบเข้าด้วยกัน รูปแบบค่าน้ำหนักของ TF.IDF นั้นมักจะใช้ในเรื่องเวคเตอร์สเปซโมเดล ร่วมกับ Cosine similarity เพื่อพิจารณาหาความเหมือน (Similarity) ของทั้งสองเว็บ หรือสองเอกสาร

2.2.3 ทฤษฎีมุมความสัมพันธ์โคไซน์ (Cosine Similarity) Garcia, E. (2006)

โดยทฤษฎีเอกสารที่ใกล้เคียงกันนั้นจะทำการคำนวณหาค่าความสัมพันธ์ของสองเอกสารระหว่างค่าเวคเตอร์หลายมิติจากทั้งสองเอกสาร โดยมีสูตรดังต่อไปนี้

$$\cos \theta = (V1 * V2) / (|V1| * |V2|) \quad (2.8)$$

โดยที่ผลลัพธ์ค่ามุม θ ที่คำนวณออกมาได้นั้นจะมีขอบเขตอยู่ในช่วง [0,90] ซึ่งผลลัพธ์ค่ามุม θ ที่ได้ออกมาแต่ค่านั้นมีผลความสัมพันธ์ระหว่างสองเว็บดังต่อไปนี้

ถ้า $\theta = 90$,เว็บทั้งสองมีเนื้อหาที่แตกต่างกัน

$0 <= \theta <= 90$,เว็บทั้งสองที่ส่วนสัมพันธ์กันตามมุม θ ที่เกิดขึ้นนี้

$ø = 0$,เว็บทั้งสองมีเนื้อหาเหมือนกัน

ถ้าค่า Cosine เท่ากับ 0 หมายถึงคำที่สนใจหรือเทอมๆนั้นในเอกสารทั้งสองไม่สัมพันธ์กัน

ตัวอย่าง การคำนวณค่าเวคเตอร์ของสองเว็บเพื่อหาค่าความสัมพันธ์ของเว็บทั้งสอง ว่าจะนำค่าเวคเตอร์ของน้ำหนัก TF.IDF แต่ละเทอมที่สนใจไปคำนวณในสูตรที่ (8) ทฤษฎีเอกสารที่ใกล้เคียงกันได้อย่างไร

สมมุติว่า $V_{d1} = <1,1,2>$

$V_{d2} = <0,3,0>$

หาค่า Cosine Measure ได้จากสูตร $COS ø = (V_1 * V_2) / (|V_1| |V_2|)$

$V_1 * V_2 = (1)(0)+(1)(3)+(2)(0) = 3$

$|V_1| = (1^2 + 1^2 + 2^2)^{1/2} = (6)^{1/2}$

$|V_2| = (0^2 + 3.^2 + 0^2)^{1/2} = 3$

$COS ø = 3 / ((6)^{1/2} * 3)$

ทำการถอดค่า Cosine ก็จะได้มุมความสัมพันธ์ระหว่างเว็บ d_1 กับเว็บ d_2

2.3 แนวคิดและวิธีการของเว็บผู้ให้บริการสืบค้นข้อมูล

2.3.1 Northern Light

Northern Light ใช้วิธีการที่เรียกว่า "Custom Search Folder" แล้วเปลี่ยนมาเป็น "Category profile extraction" ซึ่งมีความต้องการที่จะเก็บผลลัพธ์ข้อมูลการค้นหาออกเป็นกลุ่มๆเดียวกันในหลายมุมมองเช่น มุมมองด้านหัวเรื่อง, มุมมองด้านแหล่งอ้างอิง, มุมมองด้านชนิดของข้อมูล เป็นต้น

2.3.2 Google

Google อาศัยวิธีการที่เรียกว่า "Citation Analysis" ร่วมกับ "link-based search " และมีการให้คะแนนแก่หน้าที่มีคำที่ค้นหาโดยเรียกว่า "PageRanks" โดยที่พัฒนามาจากแนวคิดของ "HITS" และ "Hubs and Authorities"

2.3.3 Infoseek

Infoseek ได้พัฒนาโปรแกรมที่ชื่อว่า CCE (Content Classification Engine) ซึ่งสามารถรวบรวมข้อมูลแต่นำมาแยกประเภทได้โดยอัตโนมัติโดย นำข้อมูลเข้ามาและนำไปเปรียบเทียบกับ

โครงสร้างของข้อมูลประเภทต่างๆบนเซิร์ฟเวอร์ โดยสามารถเรียนรู้และคาดเดาว่าจะคัดแยกหน้าเว็บได้อย่างไรได้

และ Infoseek ยังมีแนวคิดที่จะจัดกลุ่มของผลลัพธ์ในการค้นหาแต่ละครั้ง และมีการเก็บข้อมูลที่เหมือนกันแต่ต่างที่กันเป็นอันเดียวกันได้ทำให้ไม่เสียเวลาในการดูข้อมูลเดิมซ้ำๆกันหลายๆครั้ง

2.3.4 Lycos

Lycos มีการสร้างรายชื่อการคัดแยกเว็บแบบอัตโนมัติแต่รายชื่อที่ได้เป็นเพียงรายชื่อลิงค์ไปยังหน้าเว็บเท่านั้น ไม่มีการสรุปผลหรือจัดกลุ่มของข้อมูลอีกที

2.3.5 Yahoo

Yahoo เป็นมีการเก็บข้อมูลแบบ “Web directories“ ซึ่งเป็นการสร้างการคัดแยกหน้าเว็บโดยใช้มนุษย์

2.3.6 Altavista

Altavistaใช้วิธีการที่เรียกว่า ”Catalogue-based service“ และมีส่วนที่ทำหน้าที่กลั่นกรองข้อมูลให้เที่ยงตรงยิ่งขึ้น และมีการแนะนำผู้ใช้ในการกรอกส่วนการสืบค้นเพื่อเพิ่มความแม่นยำในการค้นหาผลลัพธ์ตามความต้องการของผู้ใช้

จะเห็นได้ว่าแต่ละผู้ให้บริการสืบค้นข้อมูลกลุ่มที่เกี่ยวข้องกับการสืบค้นข้อมูล (Information Retrieval : IR) ต่างก็มีวิธีการเป็นของตัวเอง แต่วิธีการส่วนใหญ่ที่ได้จัดอันดับเอกสารโดยแยกออกเป็นเทอมเวคเตอร์ของแต่ละเอกสารนั้นๆล้วนได้พัฒนามาจากสูตรที่ (2.5) หัวข้อที่ 2.2.2 เรื่องแนวคิดแบบเวคเตอร์สเปซโมเดล

บทที่ 3
การวิเคราะห์ระบบร่วมกับทฤษฎี

3.1 การวิเคราะห์ระบบร่วมกับทฤษฎีเพื่อออกแบบโปรแกรม

จากการศึกษาเรื่อง TF.IDF, เวคเตอร์สเปซโมเดล (Vector Space Model) และทฤษฎีมุมความสัมพันธ์โคไซน์ (Cosine Similarity) จากบทที่2 มาแล้ว ในบทนี้จะทำการวิเคราะห์และประยุกต์ใช้ทฤษฎีดังกล่าวเพื่อนำไปพัฒนาเป็นโปรแกรม Web Categorization ขึ้นมาโดยพิจารณา

- วิธีการจัดเก็บข้อมูลหน้าเว็บตัวอย่าง (Sampling Web page)
- การนำข้อมูลหน้าเว็บข้อมูลตัวอย่างมาวิเคราะห์เพื่อหาค่า TF
- การนำข้อมูลหน้าเว็บข้อมูลตัวอย่างมาวิเคราะห์เพื่อหาค่า IDF
- การหาค่าน้ำหนัก TF.IDF ของแต่ละเทอมในแต่ละเว็บนำน้ำหนักที่ได้ไปพิจารณาตามหลักการเวคเตอร์สเปซโมเดล
- การคำนวณหามุมความสัมพันธ์โคไซน์ (θ)
- การรายงานตารางผลการทดลอง

3.1.1วิธีการจัดเก็บข้อมูลหน้าเว็บข้อมูลตัวอย่าง (Sampling Web page)

3.1.1.1 การค้นหารายชื่อเว็บข้อมูลตัวอย่าง

- พิจารณาว่าจะใช้หลักการใดในการจัดเก็บข้อมูลหน้าเว็บตัวอย่าง เริ่มจากการตัดสินใจเลือกเว็บ เสิร์ชเอ็นจิ้นกูเกิ้ล (Google) เป็นหลักในการได้มาซึ่งรายชื่อของกลุ่มข้อมูลหน้าเว็บตัวอย่าง การใช้วิธีนี้ถือว่าเป็นการเก็บข้อมูลตัวอย่างแบบกึ่งอัตโนมัติ (Semi-Automatic) โปรแกรม Web Categorization สามารถเซ็ตค่าจำนวนหน้าในกูเกิ้ลได้เช่นถ้าเซ็ตค่าจำนวนหน้า เท่ากับ1 โปรแกรมก็จะเก็บรายชื่อหน้าเว็บตัวอย่างมาได้เท่ากับ 10 เว็บเป็นต้น เริ่มการค้นหารายชื่อโดยที่ผู้ใช้ทำการป้อนคำที่ต้องการค้นหา (Query) ในเว็บกูเกิ้ลเซ็ตค่าจำนวนหน้าในกูเกิ้ลไว้ที่ 5 เพื่อให้ได้รายชื่อเว็บประมาณ 50 เว็บจากการค้นหาด้วย คำที่ต้องการค้นหา 1ครั้งเพราะฉะนั้นหากทำการค้นหาด้วย คำที่ต้องการค้นหา 10 ครั้ง ก็จะได้รายชื่อเว็บตัวอย่างน้อยกว่าหรือเท่ากับ 500 เว็บ (< = 500) สาเหตุเนื่องจากถ้ารายชื่อเว็บจากการค้นหาซ้ำกันในการค้นหาด้วยคำเดียวกัน โปรแกรมจะทำการตัดรายชื่อที่ 2, 3, ..., n ออกทันที จะเก็บเพียงเว็บอันแรกเท่านั้น แต่ถ้ามีรายชื่อเว็บจากการค้นหาซ้ำกันในการค้นหาครั้งถัดไปโปรแกรมตัดรายชื่อเว็บเดิมทั้งหมดเนื่องจากการคำนวณหาค่าความสัมพันธ์ระหว่างเว็บโดยหลักการเวคเตอร์สเปซโมเดล (VSM) จะต้องนำค่า น้ำหนักที่ได้จากแต่ละเทอมในเว็บๆนั้นมาเปรียบเทียบซึ่งไม่ควรเป็นเว็บเดียวกัน

3.1.1.2 วิธีการดาวน์โหลดข้อมูลจากรายชื่อเว็บข้อมูลตัวอย่าง

- พิจารณาใช้โปรแกรมแชร์แวร์ (Shareware) ชื่อ HTTrack ในการดาวน์โหลดเนื่องจากเป็นโปรแกรมที่สามารถเปิดได้หลายเซสชั่น (Session) ในเวลาเดียวกันได้ซึ่งจะได้ทำการดาวน์โหลดรายชื่อเว็บตัวอย่างทีละ 50 เว็บพร้อมๆกันโดยกำหนดการดาวน์โหลดแต่ละครั้งในการค้นหารายชื่อเป็นเวลา 1วัน และโปรแกรม Web Categorization ควรจะมีส่วนติดต่อโดยตรงไปยัง โปรแกรมแชร์แวร์ HTTrack เพื่อตั้งค่าไฟล์ที่ต้องการดาวน์โหลดได้เช่น htm, html, shtml, xml, pdf ฯลฯ เป็นต้น เพื่อที่ตัวโปรแกรม Httrack ได้ไม่เก็บไฟล์อื่นๆที่ไม่จำเป็นเช่น bmp, jpg, gif, js, ฯลฯ

3.1.1.3 การจัดเก็บเว็บข้อมูลตัวอย่าง

- ทำการเก็บเว็บข้อมูลตัวอย่างในรูปของไฟล์และโฟลเดอร์ ไม่ทำการจัดเก็บในรูปแบบของฐานข้อมูล (Database) เนื่องจากเป็นรูปแบบที่โปรแกรมแชร์แวร์ HTTrack ได้ทำการจัดเก็บอยู่แล้วและทั้งยังสะดวกในการอ่านค่าออกมาวิเคราะห์และคำนวณผ่านโปรแกรม Visual Studio .NET C#
- โฟลเดอร์ที่ทำการเก็บข้อมูลเว็บจะมีชื่อตาม วันเดือนปีและเวลา เช่น "30-08-2551_104723" เพื่อทราบว่าข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลเมื่อไหร่ พร้อมทั้งเก็บค่าหน้าเว็บที่ใช้กูเกิ้ลค้นหาเจอมาในรูปไฟล์ชื่อ ตย. "cake_1.html, cake_2.html, cake_3.html, cake_4.html, cake_5.html เป็นต้น

3.1.2 การนำข้อมูลหน้าเว็บข้อมูลตัวอย่างมาวิเคราะห์เพื่อหาค่า TF

การนำข้อมูลหน้าเว็บข้อมูลตัวอย่างมาวิเคราะห์เพื่อหาค่า TF ได้นั้นต้องทำการวางแผนเพื่อให้ได้มาซึ่งค่า จำนวนครั้งในการพบเทอมที่สนใจ (t) และค่าจำนวนคำทั้งหมดที่ปรากฏในเอกสาร(T) จึงใช้การเขียนโปรแกรมเพื่อเข้าไปอ่านค่าทีละเว็บและทีละหน้าเว็บโดยตัวโปรแกรม Web Categorization จะนำรายชื่อเทอมที่เคยค้นหาทั้งหมดมาเปรียบเทียบทีละอันเพื่อให้ได้ ค่าจำนวนครั้งที่พบเทอมที่สนใจในเว็บนั้นมาเก็บไว้

ส่วนค่าจำนวนคำทั้งหมดที่ปรากฏในเอกสาร (T) ได้มาจากการที่โปรแกรมนับ (Count) แต่ละคำโดยอาศัยช่องว่าง (Space) ในการแยกเป็นคำหนึ่งคำ

นำผลลัพธ์ที่ได้ทั้ง ค่า t และ T มาเข้าสูตร t / T ก็ได้ค่า TF ออกมา

3.1.3 การนำข้อมูลหน้าเว็บข้อมูลตัวอย่างมาวิเคราะห์เพื่อหาค่า IDF

การหาค่า IDF ในโปรแกรม ค่าจำนวนของเอกสารทั้งหมด (D) หาได้หลังจากหัวข้อที่ 3.1.1.1 เรื่องการค้นหารายชื่อเว็บข้อมูลตัวอย่างว่าได้ตัดเว็บที่ซ้ำกันออกไปเท่าไหร่ ก็จะได้ค่าจำนวนเว็บทั้งหมดที่นำมาใช้เป็นกลุ่มเว็บข้อมูลตัวอย่าง

ส่วนค่าจำนวนของเอกสารที่พบเทอมที่สนใจ (d) หาได้จากการคำนวณจากตารางจำนวนครั้งที่พบเทอม (t) โดยกำหนดให้ เว็บที่พบเทอม (t) ในเอกสารหรือกล่าวได้ว่า เทอม (t) มีค่า

มากกว่าเท่ากับ 1 (t >= 1) จะถือว่าได้เกิดการพบเทอมที่สนใจในเว็บนั้นแล้วจึงให้ค่า d = 1 แต่ถ้าไม่พบเทอมที่สนใจในเว็บนั้นๆเลย เพราะฉะนั้น เทอม t มีค่าเท่ากับ 0 ซึ่งโปรแกรมก็จะให้ค่า d = 0 ด้วย เมื่อให้ค่า d กับเอกสารเว็บทั้งหมดแล้ว ให้รวมผลค่า d ออกมาก็จะได้เป็นค่าจำนวนของเอกสารที่พบเทอมที่สนใจ (d)

นำผลลัพธ์ที่ได้ทั้ง ค่า D และ ก มาเข้าสูตร log (|D| / |d|)ก็ได้ค่า IDF ออกมา เป็นที่น่าสนใจว่าค่า IDF นั้นจะเท่ากันทั้งหมดสำหรับเทอมๆหนึ่งในเว็บนั้นๆ

3.1.4 การหาค่าน้ำหนัก TF.IDF ของแต่ละเทอมในแต่ละเว็บนำน้ำหนักที่ได้ไปพิจารณาตามหลักการเวคเตอร์สเปซโมเดล

นำค่า TF และ IDF จากข้อ 3.1.2 และ 3.1.3 มาคำนวณหาค่า TF.IDF ตามสูตรคือนำค่าความถี่ของเทอม (TF) มาคูณกับ ค่าความถี่แปรผกผันของเอกสาร (IDF)

ค่า TF.IDF ที่คำนวณได้ออกมาเป็นน้ำหนักของแต่ละเทอมในเว็บนั้น ดังนั้นเมื่อกำหนดให้เทอมที่สนใจ (t) มีจำนวนเท่ากับ i , เว็บไซต์ (Web site) หรือเอกสาร (Document) d มีจำนวนอยู่เท่ากับ n จะสามารถเขียนผลลัพธ์ออกมาให้ดูได้ดังต่อไปนี้

$$d_n = TF.IDF_{t1,dn}, TF.IDF_{t2,dn}, TF.IDF_{t3,dn}, \ldots TF.IDF_{(ti,dn)} \quad (3.1)$$

เปรียบเทียบเป็นเวคเตอร์สเปซโมเดลจะได้เป็น

$$V_{dn} = [\, W_{t1,dn}, W_{t2,dn}, W_{t3,dn}, \ldots W_{ti,dn} \,] \quad (3.2)$$

เพื่อพร้อมในการแทนค่าในสูตรการหามุมความสัมพันธ์โคไซน์ต่อไป

3.1.5 การคำนวณหามุมความสัมพันธ์โคไซน์ (ø)

จากที่ได้ค่าคำนวณ TF.IDF ของทุกเทอมที่สนใจในทุกๆเว็บ และนำไปเปรียบเทียบเป็นเวคเตอร์ของเว็บนั้นๆได้แล้ว นำผลลัพธ์ค่าน้ำหนัก ($W_{ti,dn}$) ไปคำนวณในการหาค่าทฤษฎีมุมความสัมพันธ์โคไซน์ ตามสูตร COS ø = (V1 * V2) / (|| V1 || * || V2 ||) นำค่ามุมความสัมพันธ์ระหว่างเว็บ (ø) เก็บไว้เพื่อนำไปแสดงเป็นรายงาน

3.1.6 การรายงานตารางผลการทดลอง

นำค่ามุมความสัมพันธ์โคไซน์ (ø) ไปแสดงเป็นรายงานในรูปตารางขนาด n*n ดังต่อไปนี้

ตารางที่ 3.1 แสดงวิธีการรายงานตารางผลลัพธ์มุมความสัมพันธ์โคไซน์ (ø) ระหว่างแต่ละเว็บ

Cosine Similarity (ø)	Web site1	Web site2	...	Web site N
Web site1	0	ø	ø	ø
Web site2	ø	0	ø	ø
...	ø	ø	0	ø
Web site N	ø	ø	ø	0

จากตารางรายงานความสัมพันธ์ระหว่างเว็บค่า 0 ในช่องเปรียบเทียบ เป็นเว็บเดียวกันเปรียบเทียบกันเองจึงมีค่าความสัมพันธ์เท่ากับ 0 สุดท้ายนำผลตารางการคำนวณนี้ไปสรุปเป็นเอกสารผลการวิเคราะห์

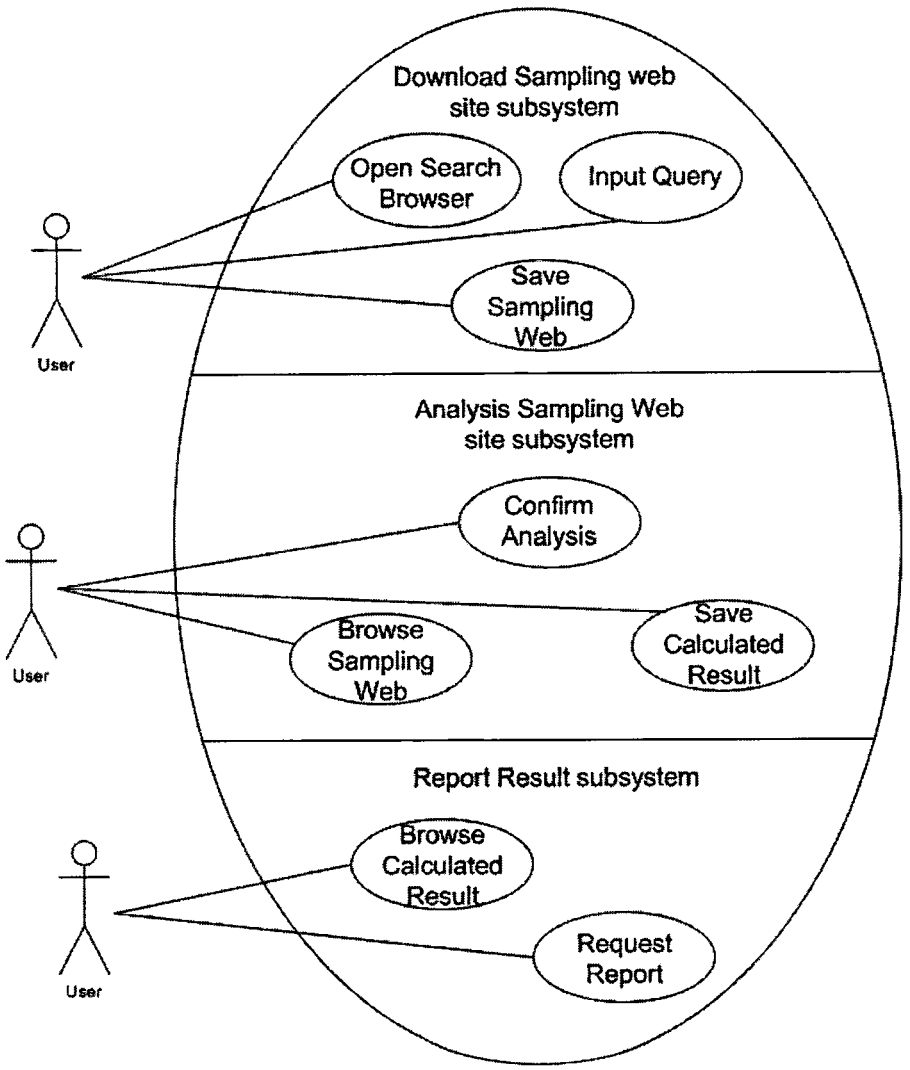
บทที่ 4

ออกแบบระบบในการพัฒนาโปรแกรม

4.1 แบบจำลองแสดงการทำงานของโปรแกรมคัดแยกประเภทหน้าเว็บ

4.1.1 แบบจำลอง Use-case

4.1.1.1 แบบจำลอง Use-case Diagram ของโปรแกรม



รูปที่ 4.1 Use-case Diagram แสดงความสัมพันธ์ระหว่างผู้ใช้กับฟังก์ชันระบบย่อยการทำงานภายในโปรแกรม

4.1.1.2 แบบจำลอง Use-case Description ของโปรแกรม

4.1.1.2.1 แบบจำลอง Use-case Description ในส่วนของ Download Sampling Web site subsystem

4.1.1.2.1.1 แบบจำลอง Use-case Description ในส่วน Open Search Browser

ตารางที่ 4.1 แสดงแบบจำลอง Use-case Description ในส่วน Open Search Browser

Use Case Name:	Open Search Browser	
Scenario:	เปิดกูเกิลเบราเซอร์	
Triggering Event:	ผู้ใช้งานเลือก File > Open หรือคลิกไอคอน Open	
Brief Description:	เปิดกูเกิลเบราเซอร์เพื่อเริ่มต้นการดาวน์โหลดเว็บตัวอย่าง	
Actors:	ผู้ใช้งาน	
Related Use Cases:	-	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	-	
Postconditions:	ผลลัพธ์รายชื่อเว็บที่ต้องการดาวน์โหลด	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือก File > Open หรือคลิกไอคอน Open	1.1 ระบบแสดงหน้าจอกูเกิลเบราเซอร์
Exception Conditions:	-	

4.1.1.2.1.2 แบบจำลอง Use-case Description ในส่วน Input query

ตารางที่ 4.2 แสดงแบบจำลอง Use-case Description ในส่วน Input query

Use Case Name:	Input query and Search	
Scenario:	ใส่ค่าคำที่ต้องการค้นหาลงในกูเกิลเบราเซอร์และคลิกปุ่ม Search	
Triggering Event:	ผู้ใช้งานใส่ค่าคำที่ต้องการค้นหาลงในกูเกิลเบราเซอร์	
Brief Description:	ผู้ใช้งานใส่ค่าคำที่ต้องการค้นหาเพื่อเตรียมค้นหารายชื่อเว็บไซต์	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Open Search Browser	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ทำการเปิดกูเกิลเบราเซอร์	
Postconditions:	ส่งค่าคำที่ต้องการค้นหาให้เว็บกูเกิลหาผลลัพธ์รายชื่อเว็บไซต์ต่อไป	
Flow of Events:	Actor	System
	1. ใส่ค่าคำที่ต้องการค้นหาลงในกูเกิลเบราเซอร์และคลิกปุ่ม Search	1.1 ทำการค้นหาผลลัพธ์รายชื่อเว็บไซต์ตามคำที่ต้องการค้นหาส่งมา
Exception Conditions:	-	

4.1.1.2.1.3 แบบจำลอง Use-case Description ในส่วน Save Sampling Web

ตารางที่ 4.3 แสดงแบบจำลอง Use-case Description ในส่วน Save Sampling Web

Use Case Name:	Save Sampling Web	
Scenario:	เก็บผลลัพธ์เว็บตัวอย่าง	
Triggering Event:	ผู้ใช้งานเลือก Tools > Download หรือคลิกไอคอน Download และสั่ง Download	
Brief Description:	ผู้ใช้งานทำการดูและเลือกรายชื่อผลลัพธ์เว็บตัวอย่างที่ต้องการจัดเก็บและเมื่อเลือกรายชื่อเรียบร้อยกด Download เพื่อสั่งให้เริ่มการดาวดหลด	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Open Search Browser, Input query	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องได้ผลลัพธ์รายชื่อเว็บไซต์จากการค้นหามาก่อน	
Postconditions:	เก็บผลลัพธ์เว็บตัวอย่างไว้ใน c:\Program Files\Web Categorization\CategoryData	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือก File > Download หรือคลิกไอคอน Download 2. เลือกรายชื่อเว็บไซต์ที่ต้องการดาวน์โหลด 3. กดคำสั่งให้ทำการ Download	1.1 ระบบเปิดหน้าต่าง Analyst ขึ้นมา 2.1 ระบบแสดงรายชื่อเว็บที่ผู้ใช้เลือก 3.1 ระบบทำการเปิดโปรแกรมแชร์แวร์ HTTrack จำนวนเว็บไซต์ที่ต้องการดาวน์โหลด 3.2 โปรแกรม HTTrack 1 session ทำการดาวน์โหลด 1 เว็บไซต์ 3.3 แสดงแถบสถานะการดาวน์โหลดสำหรับแต่ละเว็บว่ายังดาวน์โหลดอยู่หรือว่าดาวน์โหลดเสร็จแล้ว

ตารางที่ 4.3 (ต่อ)

Exception Conditions:	ผู้ใช้งานต้องทำการเลือกชื่อเว็บเพื่อทำการดาวน์โหลดอย่างน้อย 1 เว็บ (ถ้าผู้ใช้งานไม่ทำการเลือกรายชื่อเว็บเพื่อดาวน์โหลดเลยแล้วกดคำสั่ง Downloadโปรแกรมไม่ทำการดาวน์โหลดและกลับมายังหน้าจอเลือกรายชื่อเว็บไซต์

4.1.1.2.2 แบบจำลอง Use-case Description ในส่วนของ Analysis Sampling Web site subsystem

4.1.1.2.2.1 แบบจำลอง Use-case Description ในส่วน Browse Sampling Web site

ตารางที่ 4.4 แสดงแบบจำลอง Use-case Description ในส่วน Browse Sampling Web site

Use Case Name:	Browse Sampling Web site	
Scenario:	เลือกเว็บไซต์ตัวอย่างมาใช้ในการวิเคราะห์	
Triggering Event:	ผู้ใช้งานเลือก Tools > Analyst หรือคลิกไอคอน Analyst และสั่ง Browse	
Brief Description:	ผู้ใช้งานทำการเลือกโฟลเดอร์ที่เก็บเว็บตัวอย่างที่จะนำมาทำการวิเคราะห์	
Actors:	ผู้ใช้งาน	
Related Use Cases:	-	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีโฟลเดอร์ที่ได้ทำการจัดเก็บเว็บตัวอย่างเอาไว้ก่อนหน้านี้	
Postconditions:	แสดงรายชื่อโฟลเดอร์เว็บไซต์ตัวอย่างที่ต้องการทำการวิเคราะห์	
Flow of Events:	Actor	System
	1. ผู้ใช้งานเลือก File > Analyst หรือคลิกไอคอน Analyst 2. กดคำสั่ง Browse 3. เลือกรายชื่อเว็บไซต์ที่ต้องการ 4. กดคำสั่ง Load data	1.1 ระบบเปิดหน้าต่างส่วน Analyst ขึ้นมาแสดง 2.1 ระบบเปิดหน้าต่างย่อยสำหรับการ Browse ไฟล์ 3.1 ระบบแสดงรายชื่อเว็บไซต์ที่ต้องการ 4.1 ระบบทำการเขียนตารางด้านล่างโดยแถวแสดงค่าเว็บทั้งหมด ส่วนคอลัมภ์แสดงเทอมของคำที่ผู้ใช้ได้ใส่ไว้ให้ค้นหาเมื่อตอนดาวน์โหลด
Exception Conditions:	ต้องเลือกอย่างน้อย 2 โฟลเดอร์และต้องเป็นโฟลเดอร์ที่ถูกสร้างจากการดาวน์โหลดด้วยโปรแกรม HTTrack ผ่านโปรแกรมหลักอีกทีเท่านั้น	

4.1.1.2.2.2 แบบจำลอง Use-case Description ในส่วน Confirm Analysis

ตารางที่ 4.5 แสดงแบบจำลอง Use-case Description ในส่วน Confirm Analysis

Use Case Name:	Confirm Analysis	
Scenario:	ทำการยืนยันให้โปรแกรมเริ่มทำการวิเคราะห์	
Triggering Event:	ผู้ใช้งานกดคำสั่ง Analyst data	
Brief Description:	ผู้ใช้งานทำการยืนยันว่าให้คำการวิเคราะห์ข้อมูล	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Browse Sampling Web site	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีการ Load data เข้ามาเพื่อให้ระบบเขียนตารางความสัมพันธ์ระหว่างเว็บกับเทอมที่สนใจก่อน	
Postconditions:	ได้ค่ามุม ø และ ค่าน้ำหนัก TF.IDF สำหรับทุกเทอมในแต่ละเว็บออกมาทั้งหมด	
Flow of Events:	Actor	System
	1. ผู้ใช้งานทำการกดคำสั่ง Analyst data เพื่อยืนยันว่าต้องการให้ทำการวิเคราะห์ข้อมูล	1.1 ระบบทำการโยนค่าเทอมทีละค่าเข้าไปในเว็บเพื่อนับจำนวนครั้งที่พบเทอมที่สนใจ (t) ในเว็บนั้นๆ ซึ่งขณะเดียวกันโปรแกรมก็ได้ทำการนับจำนวนคำทั้งหมดที่มีในเอกสารทำให้ได้ค่า (T) 1.2 ระบบทำทีละเว็บจนครบจำนวนเว็บทั้งหมดซึ่งในช่วงการทำงานดังกล่าวระบบจะทำการนับจำนวนเว็บทั้งหมดไปด้วยทำให้ได้ค่า (D) และในขณะที่พบเทอมที่สนใจในเอกสารใดๆระบบจะทำการนับค่าเอกสารที่พบเทอมที่สนใจเป็น 1 สำหรับเทอมนั้นทันทีแต่ถ้าไม่พบเทอมที่สนใจใน

ตารางที่ 4.5 (ต่อ)

Flow of Events:	Actor	System
		เอกสารนั้นเลยระบบจะให้ค่าเป็น 0 ทำให้ได้ค่า (d) ออกมา 1.3 ระบบนำค่า t, T, D, d ไป คำนวณหาค่า TF.IDF มาแสดงในตารางโดยอัตโนมัติและนำค่าน้ำหนัก TF.IDF ที่คำนวณออกมาได้ทุกเทอมของเว็บทีละคู่ไปคำนวณหาค่ามุม ø ออกมาโดยคำนวณเป็นจำนวนครั้งเท่ากับจำนวนเว็บทั้งหมด Factorial (D!) นำค่าทั้งหมดที่ได้เก็บไว้ในเครื่องและแสดงผลลัพธ์บนหน้าจอ
Exception Conditions:	-	

4.1.1.2.2.3 แบบจำลอง Use-case Description ในส่วน Save Calculated Result

ตารางที่ 4.6 แสดงแบบจำลอง Use-case Description ในส่วน Save Calculated Result

Use Case Name:	Save Calculated Result	
Scenario:	ทำการเก็บบันทึกผลลัพธ์การคำนวณ	
Triggering Event:	ผู้ใช้งานกดคำสั่ง Save data	
Brief Description:	หลังจากโปรแกรมทำการคำนวณเสร็จแล้วสามารถเก็บบันทึกผลลัพธ์การคำนวณเอาไว้ได้	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Confirm Analysis	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีหน้าจอแสดงผลลัพธ์การคำนวน	
Postconditions:	ต้องการทำการเก็บผลลัพธ์การคำนวณเอาไว้ในเครื่องหรืออุปกรณ์ที่มีหน่วยความจำสามารถจัดเก็บข้อมูลเอาไว้ได้	
Flow of Events:	Actor	System
	1. ผู้ใช้งานกดคำสั่ง Save data 2. เลือกสถานที่โฟลเดอร์ภายในเครื่องที่ต้องการเก็บผลลัพธ์การคำนวณนี้ 3. กดยืนยันคำสั่ง Save	1.1 ระบบแสดงหน้าจอย่อยโฟลเดอร์ภายในเครื่อง 1.2 ระบบแสดงสถานที่โฟลเดอร์ที่ผู้ใช้ต้องการเก็บภายในหน้าต่างย่อยนี้ 1.3 ระบบทำการเก้บบันทึกข้อมูลผลลัพธ์
Exception Conditions:	-	

4.1.1.2.3 แบบจำลอง Use-case Description ในส่วนของ Analysis Sampling Web site subsystem

4.1.1.2.3.1 แบบจำลอง Use-case Description ในส่วน Browse Calculated Result

ตารางที่ 4.7 แสดงแบบจำลอง Use-case Description ในส่วน Browse Calculated Result

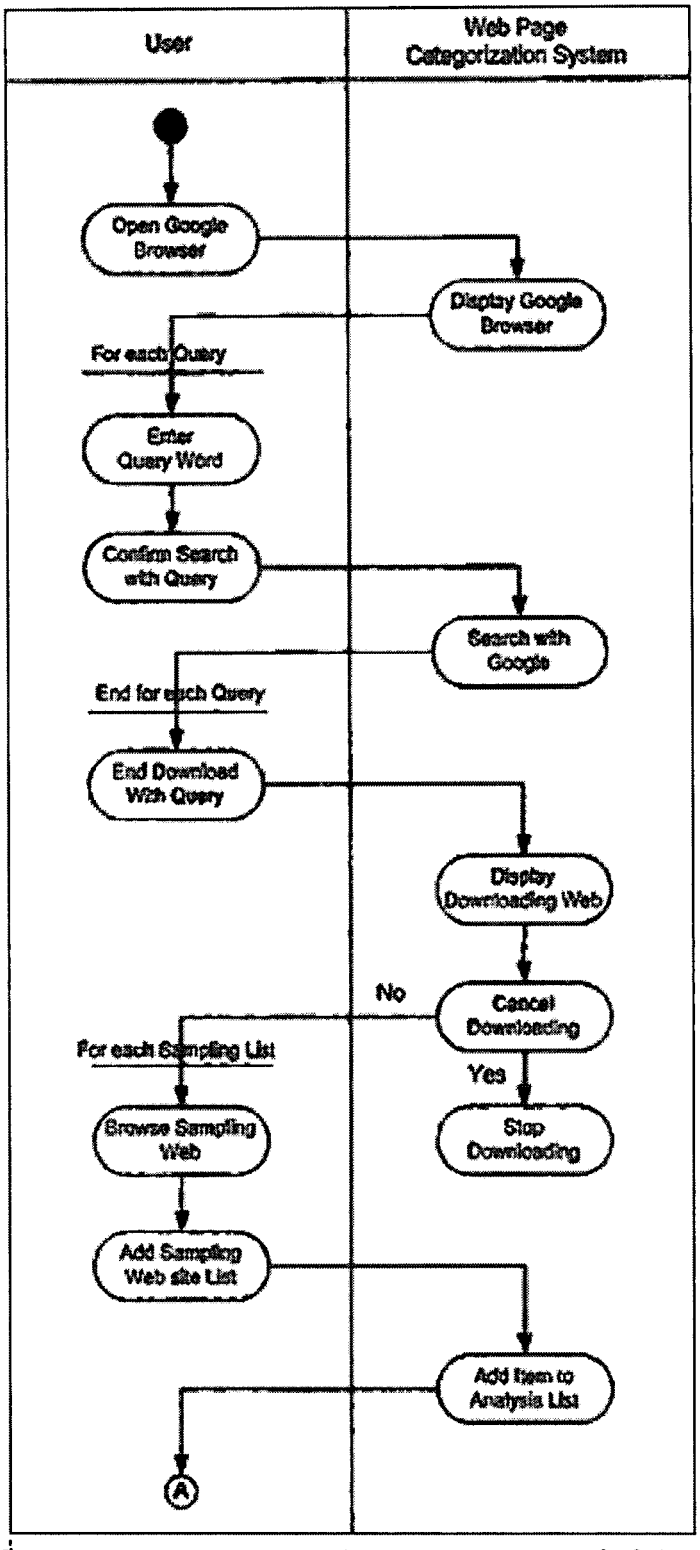
Use Case Name:	Browse Calculated Result	
Scenario:	เลือกไฟล์ที่เก็บผลลัพธ์การคำนวณ	
Triggering Event:	ผู้ใช้งานทำการเลือก Tools > History หรือเลือกไอคอน History	
Brief Description:	เลือกไฟล์ที่เก็บผลลัพธ์การคำนวณเพื่อที่จะนำมาแสดงเป็นรายงาน	
Actors:	ผู้ใช้งาน	
Related Use Cases:	-	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	ต้องมีไฟล์ผลลัพธ์เก็บอยู่ในอุปกรณ์ที่มีหน่วยความจำสามารถจัดเก็บข้อมูลเอาไว้ได้หรือไฟล์ผลลัพธ์ภายในเครื่อง	
Postconditions:	-	
Flow of Events:	Actor	System
	1. ผู้ใช้งานทำการเลือก Tools > History หรือเลือกไอคอน History 2. กดคำสั่ง Browse เพื่อเลือกสถานที่ที่เก็บไฟล์ผลลัพธ์ไว้ 3. กดคำสั่ง Open	1.1 ระบบแสดงหน้าต่างส่วน History 2.1 ระบบทำการแสดงหน้าต่างย่อย Browse 3.1 ระบบนำผลลัพธ์จากไฟล์ที่เลือกออกมา
Exception Conditions:	ต้องเลือก 1 ไฟล์ผลลัพธ์ที่ถูกจัดเก็บไว้	

4.1.1.2.3.2 แบบจำลอง Use-case Description ในส่วน Request Report

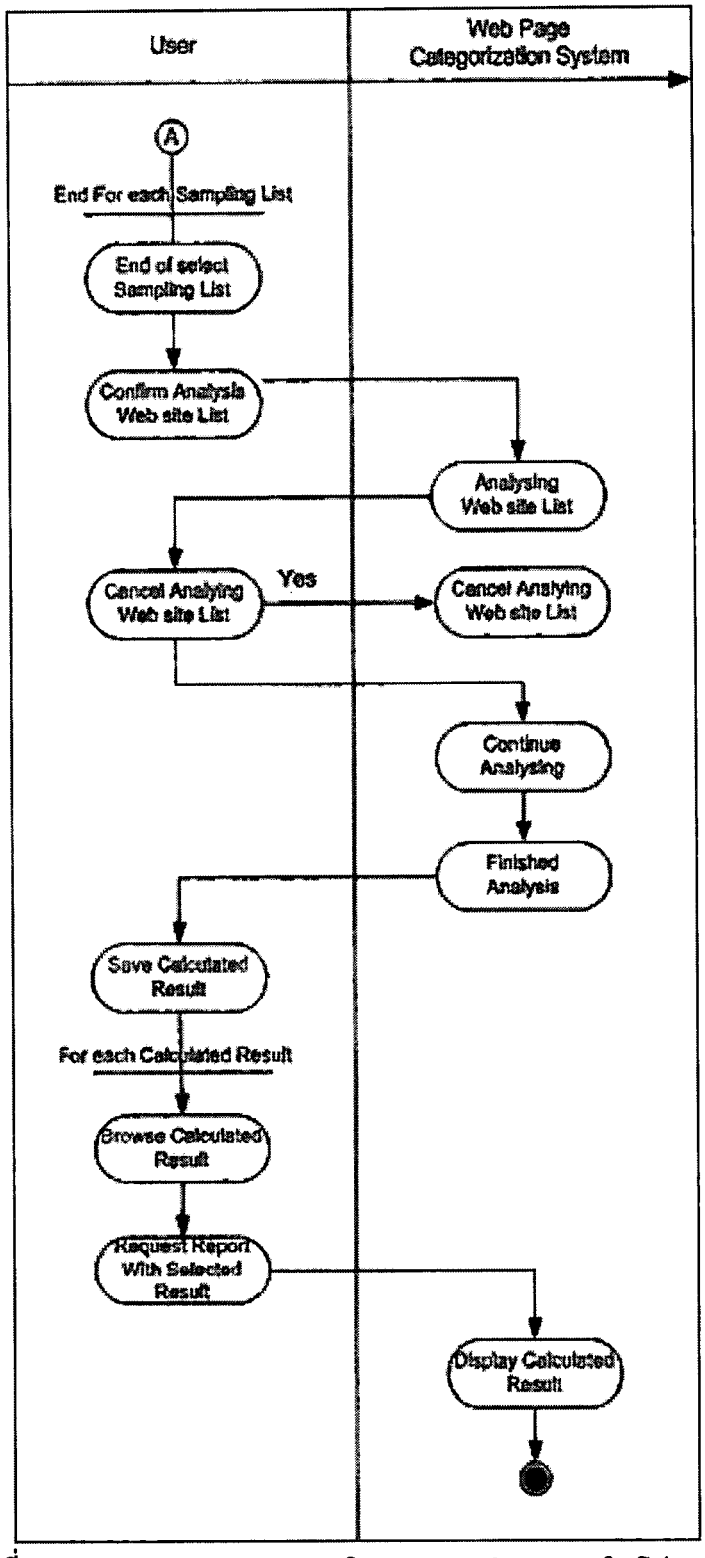
ตารางที่ 4.8 แสดงแบบจำลอง Use-case Description ในส่วน Request Report

Use Case Name:	Request Report	
Scenario:	แสดงรายงานผลลัพธ์การคำนวณ	
Triggering Event:	ผู้ใช้ทำการเปิดไฟล์ผลลัพธ์ขึ้นมา	
Brief Description:	เป็นส่วนรายงานการแสดงผลลัพธ์ที่ได้เก็บเอาไว้ในไฟล์	
Actors:	ผู้ใช้งาน	
Related Use Cases:	Browse Calculated Result	
Stakeholders:	ผู้ใช้งาน	
Preconditions:	เลือกสถานที่ที่เก็บไฟล์ผลลัพธ์ไว้	
Postconditions:	-	
Flow of Events:	Actor	System
	1. ผู้ใช้งานดูรายงานผลลัพธ์การคำนวณ	1.1 ระบบทำการแสดงผลลัพธ์การคำนวณ
Exception Conditions:	-	

4.1.2 แบบจำลอง Activity Diagram ของโปรแกรม



รูปที่ 4.2 Activity Diagram แสดงกิจกรรมการทำงานภายในโปรแกรม



รูปที่ 4.3 Activity Diagram แสดงกิจกรรมการทำงานภายในโปรแกรม

จากการวิเคราะทฤษฎีในบทที่3 และการออกแบบโมเดลทั้งหมดในบทที่ 4 นี้ช่วยให้เห็นว่าในช่วงที่พัฒนาโปรแกรมต้องมีการทำงานอย่างไรเพื่อให้ได้ค่าที่ต้องการเพื่อที่จะได้นำมาคำนวณต่อไปทั้งยังทำให้เห็นภาพรวมของฟังก์ชันฟการทำงาน, การไหลของข้อมูล, ขั้นตอนคำสั่งที่ผู้ใช้

โปรแกรมสามารถเข้าถึงและส่งความต้องการออกไปได้ภาพรวมทั้งหมดนี้ช่วยให้การพัฒนาและออกแบบโปรแกรมเป็นไปอย่างมีขั้นตอนและไม่ผิดวัตถุประสงค์

บทที่ 5

โปรแกรม Web Categorization

5.1 การติดตั้งและใช้งานโปรแกรม Web categorization

5.1.1 ขั้นตอนการติดตั้งโปรแกรม Web Categorization

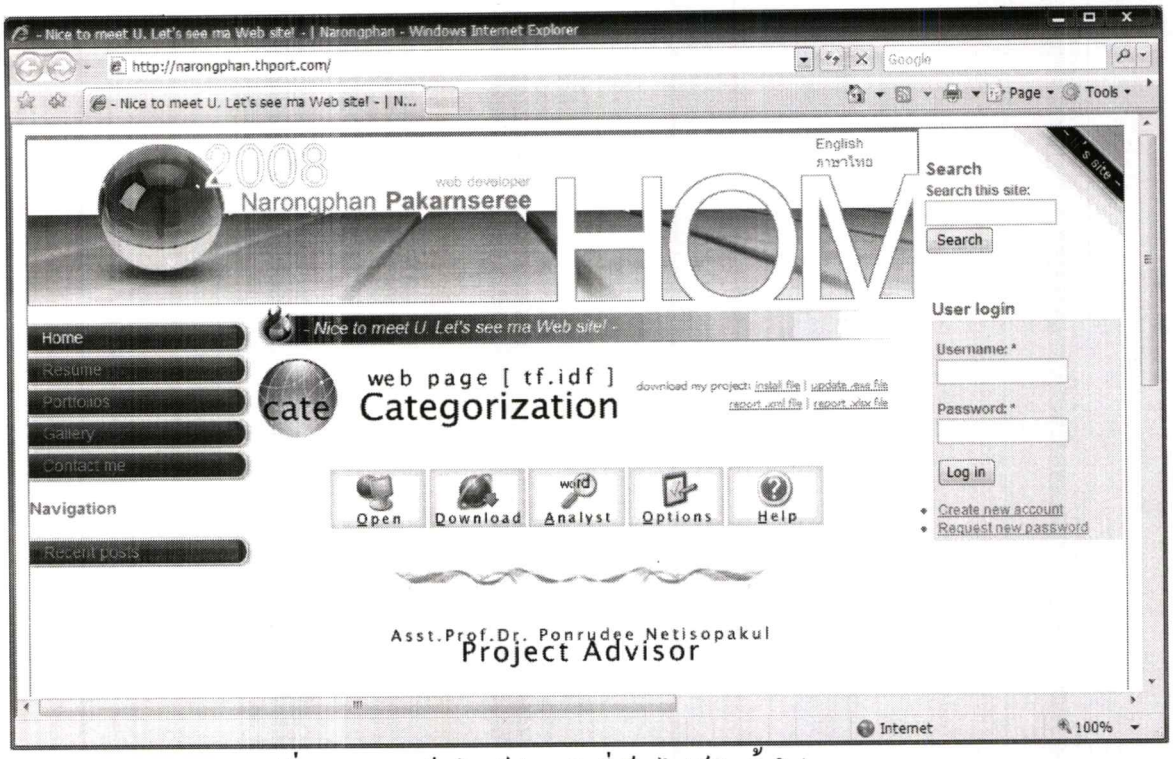
5.1.1.1 ขั้นตอนการดาวน์โหลดโปรแกรม Web Categorization

ทำการดาวน์โหลดโปรแกรม Web Categorization เวอร์ชันล่าสุดได้ที่เว็บไซต์ส่วนตัวที่

URL: *http://narongphan.thport.com/files/Release.zip*

และดาวน์โหลด update ไฟล์ execute ที่

URL: *http://narongphan.thport.com/files/Web%20Categorization.zip*

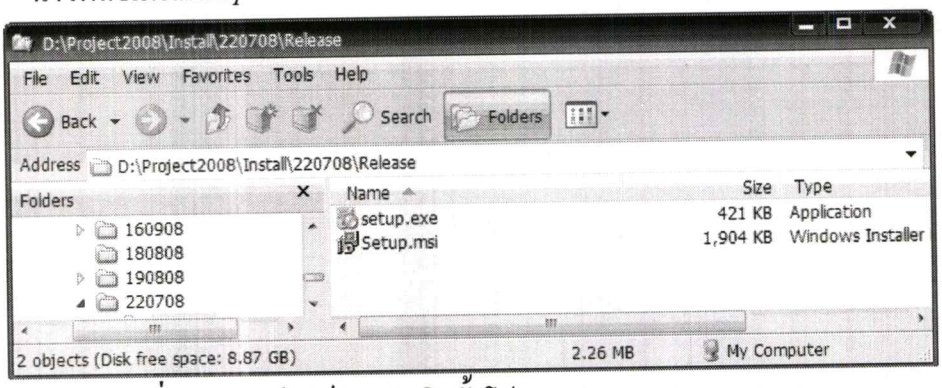


รูปที่ 5.1 แสดงเว็บไซต์ส่วนตัวที่เก็บไฟล์ติดตั้งโปรแกรม

ด้านขวามือของเว็บไซต์จะพบตัวหนังสือตัวเล็กๆว่า download my project: install file | update .exe file ให้คลิ้กที่ลิ้งค์นี้เพื่อทำการดาวน์โหลด

5.1.1.2 ขั้นตอนการแตกไฟล์ติดตั้ง Release.zip

นำไฟล์ *Release.zip* มาแตกออกจะได้ไฟล์สำหรับติดตั้งดังภาพ



รูปที่ 5.2 แสดงไฟล์สำหรับติดตั้งโปรแกรม Web Categorization

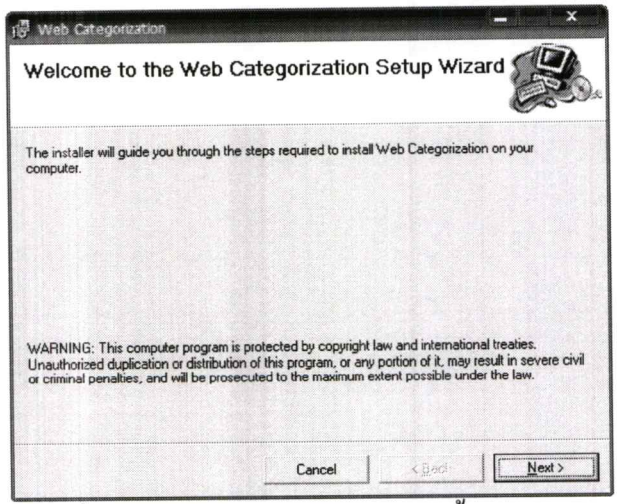
ไฟล์สำหรับติดตั้งโปรแกรม Web Categorization ประกอบไปด้วย :

- ไฟล์ setup.exe
- ไฟล์ Setup.msi

ทำการติดตั้งโดยการดับเบิ้ลคลิกไฟล์ setup.exe หรือ ไฟล์ Setup.msi ก็ได้จะพบกับหน้าจอเริ่มต้นการติดตั้ง

5.1.1.3 ขั้นตอนหน้าจอเริ่มต้นการติดตั้งโปรแกรม Web categorization

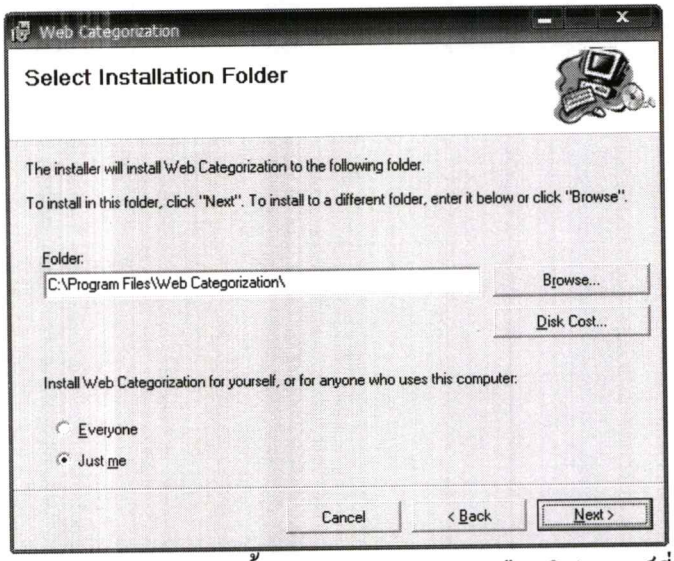
เข้าสู่หน้าจอเริ่มต้นการติดตั้งโปรแกรม Web categorization ให้คลิกปุ่ม Next > เพื่อเข้าสู่ขั้นตอนการติดตั้งขั้นต่อไป



รูปที่ 5.3 แสดงหน้าจอแรกในการติดตั้งโปรแกรม

5.1.1.4 ขั้นตอนการเลือกโฟลเดอร์ที่จะติดตั้งโปรแกรม Web categorization

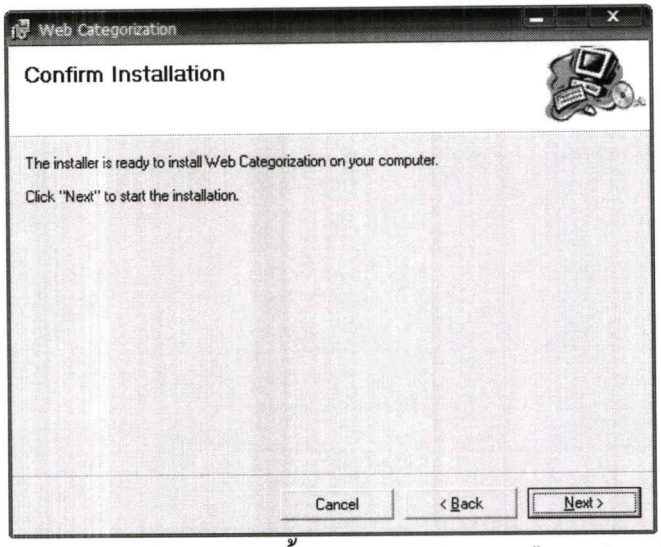
เข้าสู่ขั้นตอนการเลือกโฟลเดอร์ที่จะติดตั้งโปรแกรม Web categorization ค่าดีฟอลต์โฟลเดอร์ (Default Folder) ที่โปรแกรมได้ตั้งไว้คือที่ C:\Program Files\Web categorization\ โดยที่ผู้ใช้สามารถเปลี่ยนโฟลเดอร์ติดตั้งเป็นที่โฟลเดอร์อื่นได้



รูปที่ 5.4 แสดงหน้าจอการติดตั้งโปรแกรม ส่วนการเลือกโฟลเดอร์ที่จะติดตั้ง

5.1.1.5 ขั้นตอนการยืนยันการติดตั้งโปรแกรม Web categorization

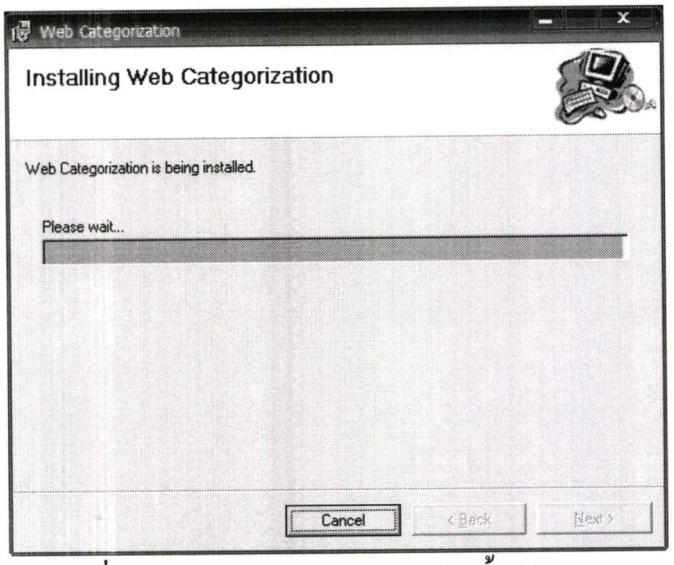
ทำการยืนยันการติดตั้งโปรแกรม Web categorization โดยคลิกปุ่ม Next > เพื่อเข้าสู่ขั้นตอนการติดตั้งขั้นต่อไป



รูปที่ 5.5 แสดงหน้าจอการติดตั้งโปรแกรม ส่วนการยืนยันว่าจะติดตั้ง

5.1.1.6 ขั้นตอนขณะระบบทำการติดตั้งโปรแกรม Web categorization

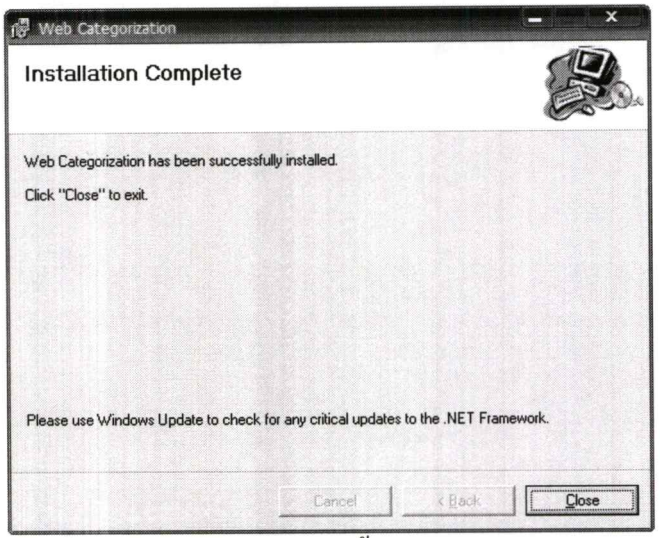
ขั้นตอนนี้ระบบทำการติดตั้งโปรแกรม Web categorization บนโฟลเดอร์ที่ผู้ใช้ได้ระบุก่อนหน้า หากผู้ใช้ไม่ได้ทำการเปลี่ยนค่าดีฟอลต์โฟลเดอร์ ระบบจะติดตั้งโปรแกรมไปยัง C:\Program Files\Web categorization\ เมื่อการติดตั้งเสร็จ ให้คลิกที่ปุ่ม Next > เพื่อเข้าสู่ขั้นตอนการติดตั้งขั้นต่อไป



รูปที่ 5.6 แสดงหน้าจอขณะทำการติดตั้งโปรแกรม

5.1.1.7 ขั้นตอนการติดตั้งโปรแกรมสมบรูณ์

หากระบบทำการติดตั้งโปรแกรม Web categorization เสร็จจะเข้าสู่ขั้นตอนนี้ให้คลิกปุ่ม Close เพื่อออกจากส่วนการติดตั้งโปรแกรม

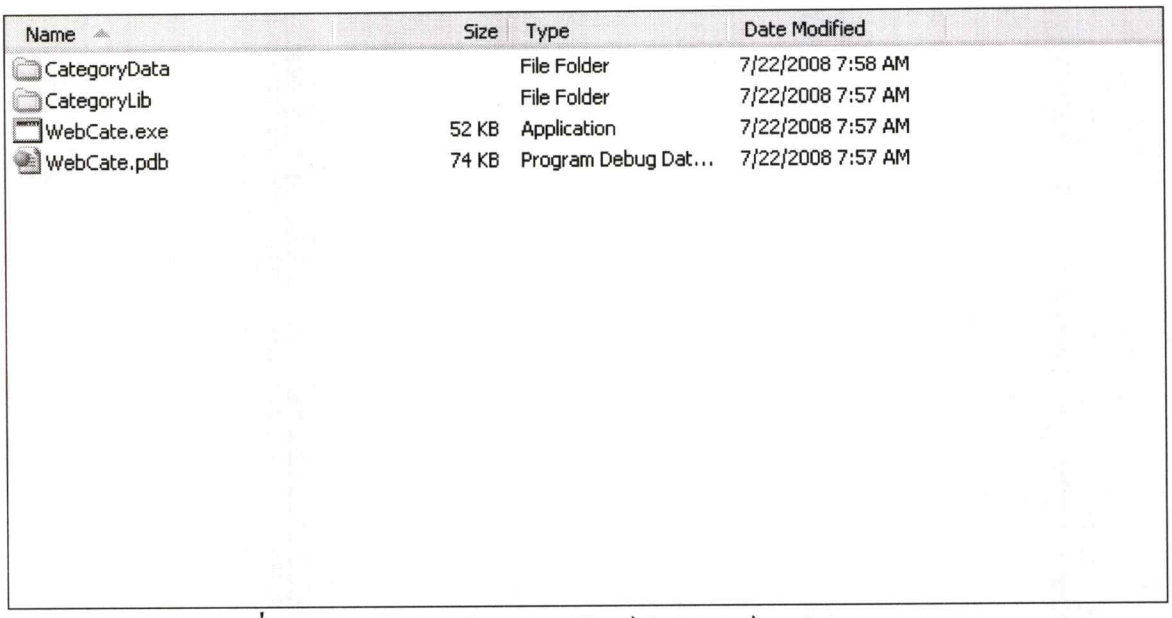


รูปที่ 5.7 แสดงหน้าจอการติดตั้งโปรแกรมเสร็จสมบรูณ์

5.1.2 ขั้นตอนการใช้งานโปรแกรม Web categorization

5.1.2.1 ขั้นตอนการตรวจสอบไฟล์ที่ติดตั้งและทำการอัพเดตไฟล์

หลังจากทำการติดตั้งโปรแกรม Web Categorization เรียบร้อยให้เข้าไปยังโฟลเดอร์ที่ได้ติดตั้งจะพบว่าภายในโปรแกรมมี โฟลเดอร์และไฟล์ดังต่อไปนี้



รูปที่ 5.8 แสดงภาพหน้าจอแสดงไฟล์, โฟลเดอร์ของโปรแกรม

- CategoryData เป็นโฟลเดอร์ที่ไว้เก็บกลุ่มเว็บตัวอย่างจากการดาวน์โหลด
- CategroyLib เป็นโฟลเดอร์เก็บไฟล์ของโปรแกรมแชร์แวร์ HTTrack
- WebCate.exe คือไฟล์ Execute ของโปรแกรม ให้ทำการอัพเดตไฟล์WebCate.exe ล่าสุดที่ได้ดาวน์โหลดมาจากURL: *http://narongphan.thport.com/files/Web%20Categorization.zip* (ในหัวข้อที่ 5.1.1.1 เรื่องขั้นตอนการดาวน์โหลดโปรแกรม Web Categorization) โดยการคัดลอก (copy) ไฟล์ที่ดาวน์โหลดมานี้ไปวาง (paste) ทับในโฟลเดอร์ที่ได้ทำการติดตั้งโปรแกรมระบบจะขึ้นหน้าต่างย่อยถามว่า ไฟล์นี้ได้มีในโฟลเดอร์อยู่แล้วต้องการจะเขียนไฟล์ทับหรือไม่? ให้คลิกปุ่ม “Over write” เพื่อทำการเขียนทับ
- WebCate.pdb คือไฟล์ Program Debug Database ไฟล์นี้จะถูกเขียนขึ้นหลังจากได้ทำการรัน (Run)โปรแกรมครั้งแรก (แสดงว่าในหน้าจอที่โชว์ด้านบนนี้ได้ทำการรันโปรแกรม Web Categorization มาแล้วอย่างน้อยหนึ่งครั้ง)

5.1.2.2 ขั้นตอนตรวจสอบไฟล์ภายในโฟลเดอร์ CategoryLib

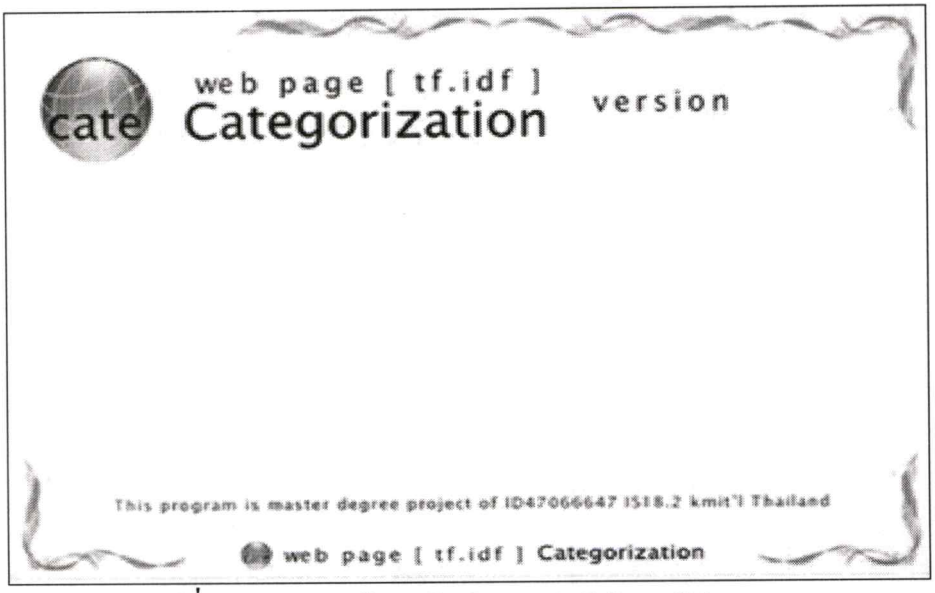
ภายในโฟลเดอร์ CategoryLib จะประกอบไปด้วยไฟล์ดังภาพ

Name	Size	Type	Date Modified
htsjava.dll	56 KB	Application Extension	6/15/2007 8:27 PM
htsswf.dll	72 KB	Application Extension	6/15/2007 8:27 PM
httrack.exe	15 KB	Application	6/15/2007 8:27 PM
httrack-doc.html	1 KB	HTML Document	11/1/2002 4:50 PM
install	1 KB	File	6/9/2003 1:50 PM
lang.def	22 KB	Export Definition File	7/24/2005 10:17 PM
libeay32.dll	668 KB	Application Extension	8/15/2002 2:03 PM
libhttrack.dll	636 KB	Application Extension	6/15/2007 8:27 PM
license.txt	2 KB	Text Document	11/28/2002 7:25 PM
MFC71.dll	1,036 KB	Application Extension	3/19/2003 6:20 AM
msvcr71.dll	340 KB	Application Extension	2/21/2003 1:42 PM
readme	2 KB	File	11/1/2002 4:54 PM
ssleay32.dll	152 KB	Application Extension	8/15/2002 2:03 PM
zlib1.dll	55 KB	Application Extension	11/18/2003 1:29 AM

รูปที่ 5.9 แสดงไฟล์ภายในโฟลเดอร์ชื่อ CategoryLib

5.1.2.3 ขั้นตอนการรันโปรแกรม Web Categorization

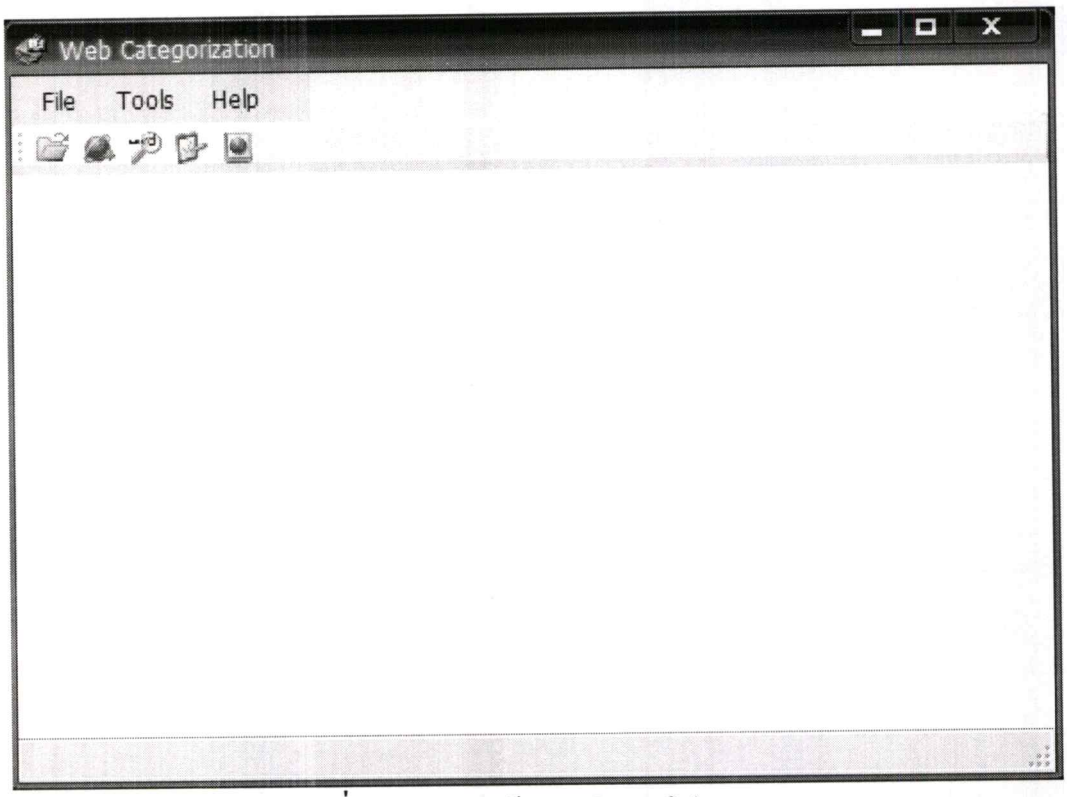
ทำการรันโปรแกรม Categorization ด้วยการดับเบิ้ลคลิกไฟล์ WebCate.exe ที่โฟลเดอร์ที่ได้ทำการติดตั้งโปรแกรมจะได้เห็นหน้าจอเปิดตัวโปรแกรมดังรูป



รูปที่ 5.10 แสดงหน้าจอเปิดตัวขณะกำลังโหลดโปรแกรม

5.1.2.4 เข้าสู่หน้าจอหลักโปรแกรม Web Categorization

โปรแกรม Web Categorization มีหน้าจอหลักดังต่อไปนี้



รูปที่ 5.11 แสดงหน้าจอหลักของโปรแกรม

ภายในหน้าจอหลักประกอบไปด้วยเมนูบาร์ :

File ประกอบไปด้วยคำสั่ง

- Open ใช้เพื่อเข้าสู่หน้าจอกูเกิลเบราเซอร์
- Exit ใช้เพื่อปิดโปรแกรม

Tools ประกอบไปด้วยคำสั่ง

- Download ใช้เพื่อเข้าสู่หน้าจอย่อย Download ทำการดาวน์โหลดเว็บตัวอย่าง
- Analysis ใช้เพื่อเข้าสู่หน้าจอย่อย Analysis ทำการวิเคราะห์ข้อมูล
- Options ใช้เพื่อเข้าสู่หน้าจอย่อย Options ทำการเซ็ตค่าของโปรแกรม
- History ใช้เพื่อเข้าสู่หน้าจอย่อย History ทำการแสดงรายงานจากไฟล์ xml

Help ประกอบไปด้วยคำสั่ง

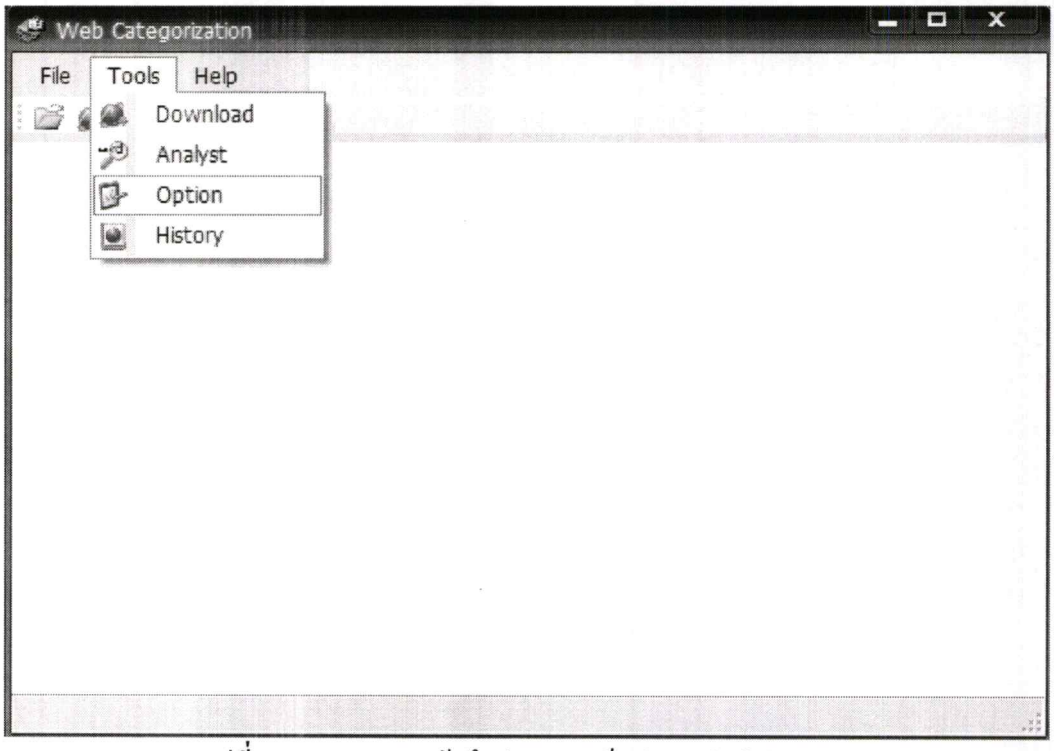
- About ใช้เพื่อเข้าสู่หน้าจอย่อย About เพื่อแสดงเวอร์ชันของโปรแกรม

หน้าจอหลักในส่วนไอคอนบาร์ :

- Open ใช้เป็นทางลัดเพื่อเข้าสู่หน้าจอกูเกิลเบราเซอร์
- Download ใช้เป็นทางลัดเพื่อเข้าสู่หน้าจอย่อย Download ทำการดาวน์โหลดเว็บตัวอย่าง
- Analysis ใช้เป็นทางลัดเพื่อเข้าสู่หน้าจอย่อย Analysis ทำการวิเคราะห์ข้อมูล
- Options ใช้เป็นทางลัดเพื่อเข้าสู่หน้าจอย่อย Options ทำการเซ็ตค่าของโปรแกรม
- History ใช้เป็นทางลัดเพื่อเข้าสู่หน้าจอย่อย History ทำการแสดงรายงานจากไฟล์ xml

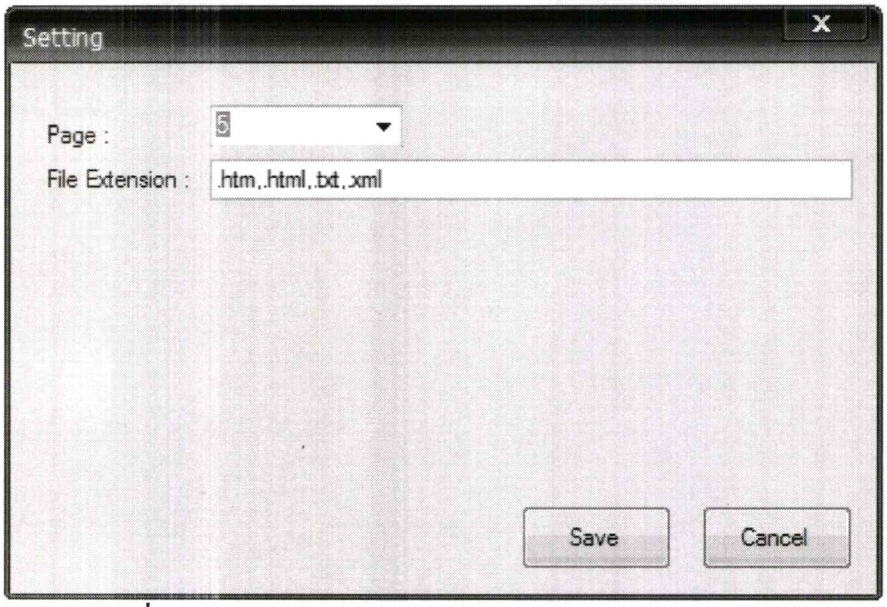
5.1.2.5 เข้าสู่การเซ็ตค่าในโปรแกรม

เข้าสู่หน้าการเซ็ตค่าในโปรแกรมด้วยคำสั่ง Tools > Options



รูปที่ 5.12 แสดงการเข้าถึงส่วนการเซ็ตค่าภายในโปรแกรม

เข้าสู่ส่วนหน้าต่างย่อย Options เพื่อทำการเซ็ตค่า Page และ File Extension ดังมีรายละเอียดดังต่อไปนี้



รูปที่ 5.13 แสดงหน้าต่างย่อยส่วนการเซ็ตค่าภายในโปรแกรม

เซ็ตค่า Page เพื่อกำหนดจำนวนหน้าของผลลัพธ์ในการค้นหาด้วยกูเกิลโดยที่แต่ละหน้าจะมีรายชื่อเว็บอยู่ 10 เว็บ หากเซ็ตค่า Page ไว้ที่ 5 หมายความว่าโปรแกรมจะทำการเก็บรายชื่อเว็บไว้ใช้ในขั้นตอนดาวน์โหลดจำนวน 50 รายชื่อเว็บ

เซ็ตค่า File Entension เพื่อกำหนดไฟล์ที่โปรแกรม Web categorization ส่วน Analysis จะเข้าไปทำการวิเคราะห์ซึ่งหมายความว่าโปรแกรมจะไม่พิจารณาไฟล์ที่ไม่ได้กำหนดไว้ในค่านี้

จบขั้นตอนการใช้งานโปรแกรมขั้นพื้นฐานในหัวข้อ 5.1 เรื่องการติดตั้งและใช้งานโปรแกรม Web categorization ในหัวข้อถัดไปจะทำการทดลองไปพร้อมกับเรียนรู้การใช้งานโปรแกรมไปพร้อมๆกัน

5.2 ขั้นตอนการทดลองหาค่าน้ำหนัก TF.IDF ของแต่ละเทอมในแต่ละเว็บ

โปรแกรม Web Categorization นี้สามารถแบ่งออกเป็น 3 ส่วนหลักๆได้แก่

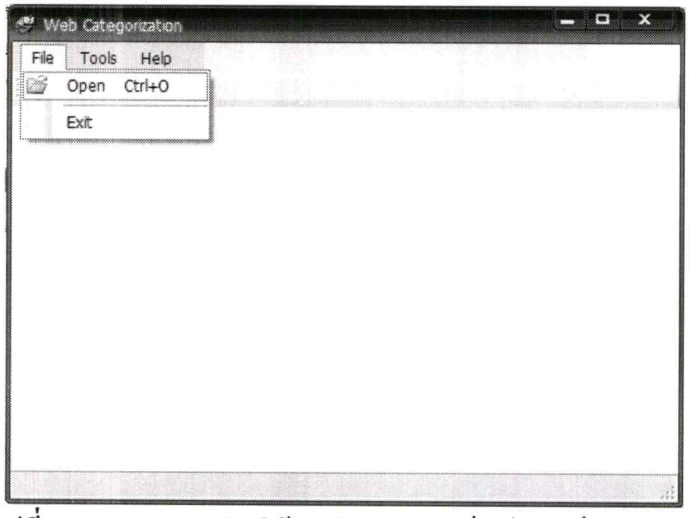
- ส่วนค้นหาและจัดเก็บข้อมูลเว็บตัวอย่าง (Search and Download Sampling Web Part)
- ส่วนคำนวณหาผลลัพธ์ (Calculate Part)
- ส่วนแสดงรายงาน (Report Part)

5.2.1 ส่วนค้นหาและจัดเก็บข้อมูลเว็บตัวอย่าง (Search and Download Sampling Web Part)

5.2.1.1 เริ่มขั้นตอนการค้นหาข้อมูลเว็บตัวอย่าง (Search Sampling Web site)

5.2.1.1.1 ขั้นตอนไปยังส่วนค้นหาข้อมูลเว็บตัวอย่าง

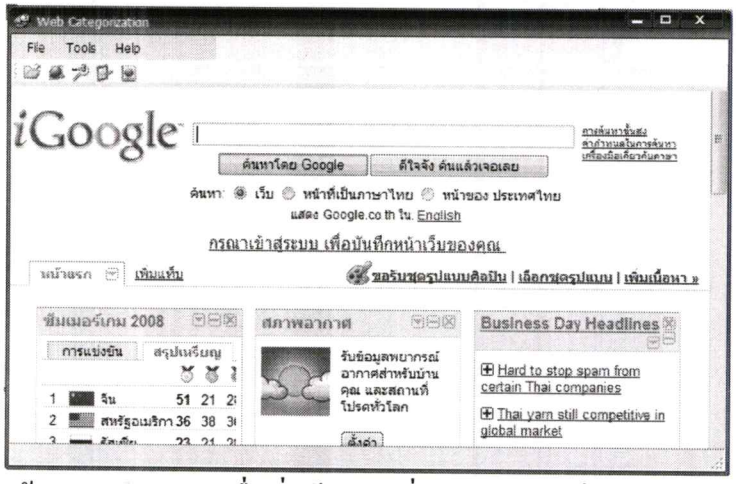
ที่หน้าจอโปรแกรมหลักไปที่ File > Open เพื่อเข้าสู่หน้าจอย่อยส่วนการค้นหาข้อมูลเว็บตัวอย่างดังรูป



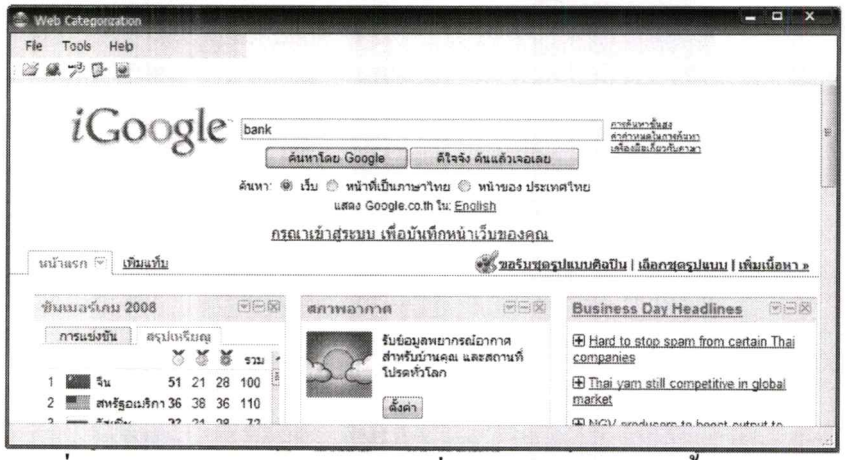
รูปที่ 5.14 แสดงการเปิดใช้งานในส่วนจัดเก็บข้อมูลเว็บตัวอย่าง

การทำการทดลองครั้งนี้ ได้ทำการดาวน์โหลดเว็บตัวอย่างจากคำที่ต้องการค้นหาทั้งหมด 10 คำได้แก่คำว่า Bank, Finance, flash, game, coke, pepsi, Thailand, Bangkok, adobe และ golf

5.2.1.1.2 เข้าสู่หน้าจอกูเกิลเบราเซอร์

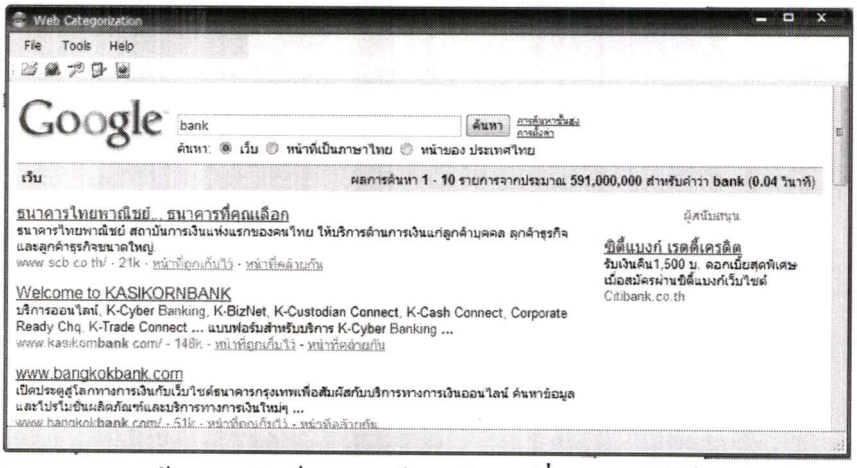


รูปที่ 5.15 แสดงหน้าจอของ Google เพื่อเริ่มป้อนคำเพื่อหากลุ่มของเว็บตัวอย่างทำการใส่คำสนใจที่ต้องการค้นหาลงบนกูเกิลเบราเซอร์แล้วคลิกปุ่ม “ค้นหาโดย Google”

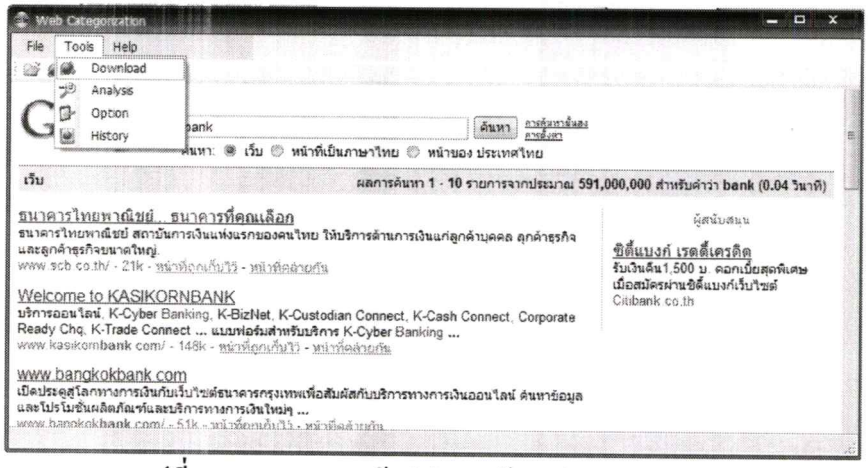


รูปที่ 5.16 แสดงตัวอย่างการใส่ค่าคำที่สนใจ (Query) บนกูเกิ้ลเบราเซอร์

5.2.1.1.3 หน้าจอผลลัพธ์จากการค้นหาคำที่สนใจในกูเกิล



รูปที่ 5.17 แสดงหน้าจอผลลัพธ์จากการค้นหาด้วยคำที่จะนำมาใช้เป็นกลุ่มเว็บตัวอย่าง

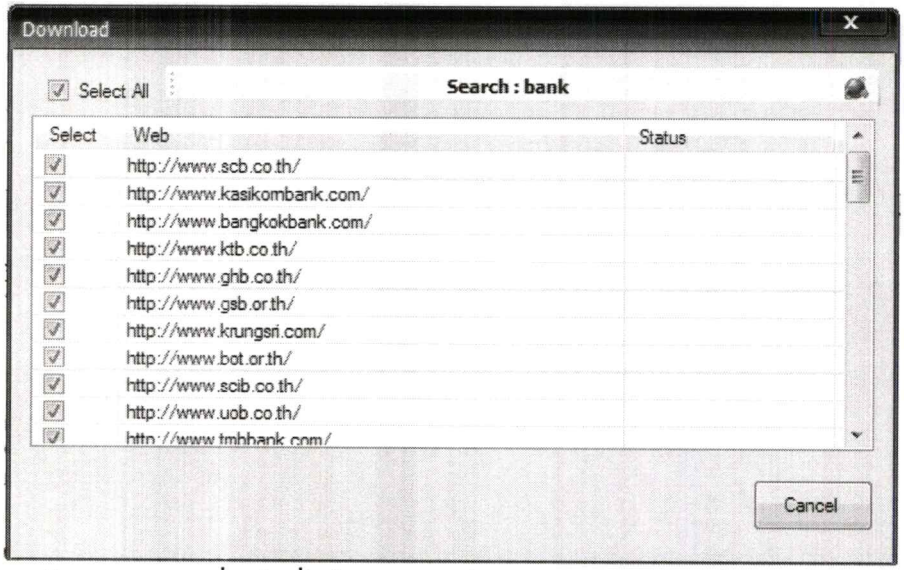


รูปที่ 5.18 แสดงการเข้าสู่ส่วนหน้าจอย่อย Download

5.2.1.1.4 ส่วนหน้าจอย่อย Download

เมื่อเข้ามาสู่ส่วนหน้าจอย่อย Download จะเห็นรายชื่อเว็บไซต์ที่ได้ทำการค้นหาไว้ในหัวข้อที่ 5.1.2.3 เรื่องหน้าจอผลลัพธ์จากการค้นหาคำที่สนใจในกูเกิล ปรากฏอยู่

ให้ทำการเลือกรายชื่อเว็บตัวอย่างที่ต้องการดาวน์โหลดด้วยการ เช็คกล่องเช็คบล็อค (Check box)ให้เป็นเครื่องหมายถูกแต่ถ้าต้องการเลือกรายชื่อเว็บทั้งหมดก็สามารถที่จะทำได้โดยเช็คกล่องเช็คบล็อคชื่อ “Select All” ให้เป็นเครื่องหมายถูกจะเห็นว่ารายชื่อเว็บตัวอย่างทั้งหมดจะถูกเลือกทุกอันในทันที



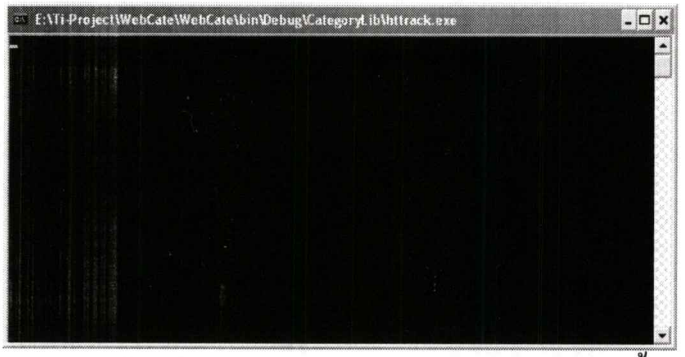
รูปที่ 5.19 หน้าจอแสดงรายชื่อเว็บที่ได้มาจากการค้นหาพร้อมเช็กค์บ็อกซ์สำหรับเลือกดาวน์โหลด

ทำการคลิกไอคอนรูป (Download) ด้านขวาบนเพื่อเข้าสู่ในขั้นตอนถัดไป (ดูรูปด้านบนประกอบ)

5.2.1.2 เริ่มขั้นตอนการดาวน์โหลดข้อมูลเว็บตัวอย่าง (Search Sampling Web site)

5.2.1.2.1 ขั้นตอนการดาวน์โหลดข้อมูลเว็บตัวอย่างด้วย HTTrack

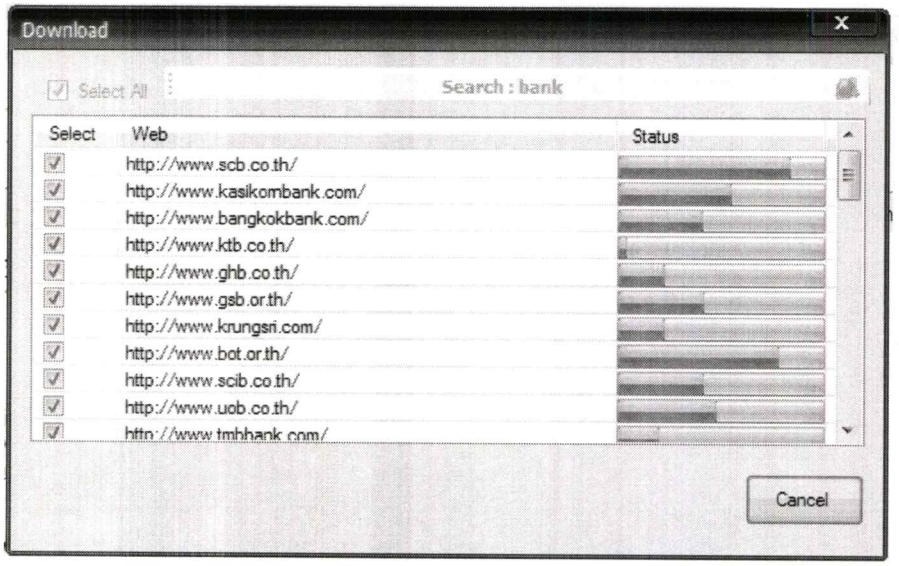
หลังจากขั้นตอนที่แล้วที่ได้กดไอคอนรูป (Download) เพื่อทำการดาวน์โหลดข้อมูลเว็บตัวอย่าง โปรแกรม Web categorization จะทำการเปิดโปรแกรมย่อยแชร์แวร์ชื่อ HTTrack ทั้งหมด 50 เซสชั่น (Sessions) เพื่อทำการดาวน์โหลดเว็บหนึ่งต่อโปรแกรมแชร์แวร์ HTTrack 1 เซสชั่น



รูปที่ 5.20 แสดงหน้าจอโปรแกรมหลักเรียกโปรแกรม HTTrack ขึ้นมาทำงาน

ส่วนการเก็บข้อมูลเว็บตัวอย่างจะเก็บโดยทำการค้นหาทั้งหมด 10 ครั้ง (10 คำที่สนใจ) แต่ละครั้งใช้คำที่ไม่ซ้ำกัน ผลลัพธ์ในแต่ละครั้งโปรแกรมจะทำการเก็บครั้งละ 50 เว็บ ดังนั้นจะได้กลุ่มเว็บตัวอย่างเป็นจำนวนทั้งหมดเท่ากับ 500 เว็บ

5.2.1.2.2 ขั้นตอนขณะทำการดาวน์โหลด



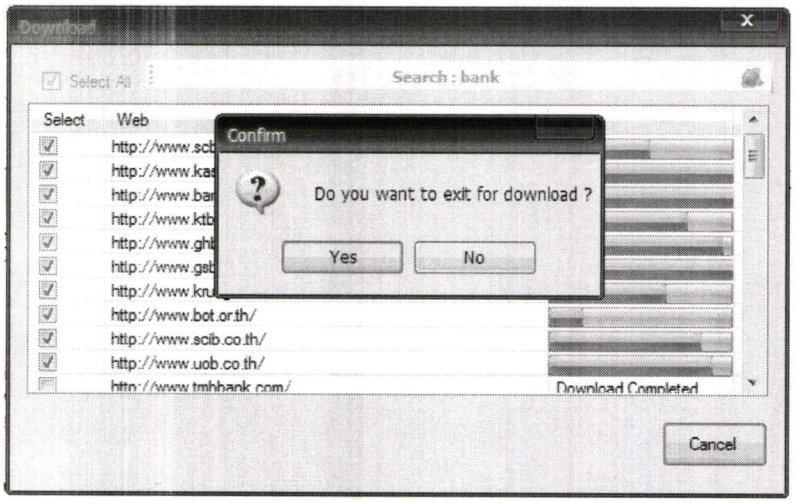
รูปที่ 5.21 แสดงหน้าจอย่อย Download แสดงการทำงานขณะดาวน์โหลดเว็บข้อมูลตัวอย่าง

ในขั้นตอนการดาวน์โหลดเว็บข้อมูลตัวอย่างนี้โปรแกรมมีแถบสถานะสีเขียวแสดงเพื่อบอกว่าเว็บดังกล่าวยังดาวน์โหลดอยู่ และถ้าหากเว็บใดมีการดาวน์โหลดเสร็จเรียบร้อยแล้วแถบสีเขียวจะหายไปและปรากฏคำว่า “Download Completed” เข้ามาแทนที่

5.2.1.2.3 ขั้นตอนการยกเลิกคำสั่งดาวน์โหลด

สามารถยกเลิกคำสั่งดาวน์โหลดได้ทุกเมื่อ โดยการคลิกที่ปุ่ม Cancel ด้านขวาล่างดังรูป โปรแกรมจะถามอีกครั้งเพื่อยืนยันว่าผู้ใช้ต้องการยกเลิกคำสั่งดาวน์โหลดหรือไม่ คลิกปุ่ม Yes เพื่อยืนยันการยกเลิกการดาวน์โหลด หรือคลิกปุ่ม No เพื่อดาวน์โหลดข้อมูลเว็บตัวอย่างต่อไป

หลังจากทำการยกเลิกการดาวน์โหลด ข้อมูลที่ดาวน์โหลดไปก่อนหน้านี้ทั้งหมดจะไม่หายไปแต่จะถูกเก็บอยู่ในโฟลเดอร์ CategoryData โดยอัตโนมัติ (เนื่องจากการโหลดเว็บทั้งหมด 50 เว็บบางครั้งกินเวลานานมากเกินไป)

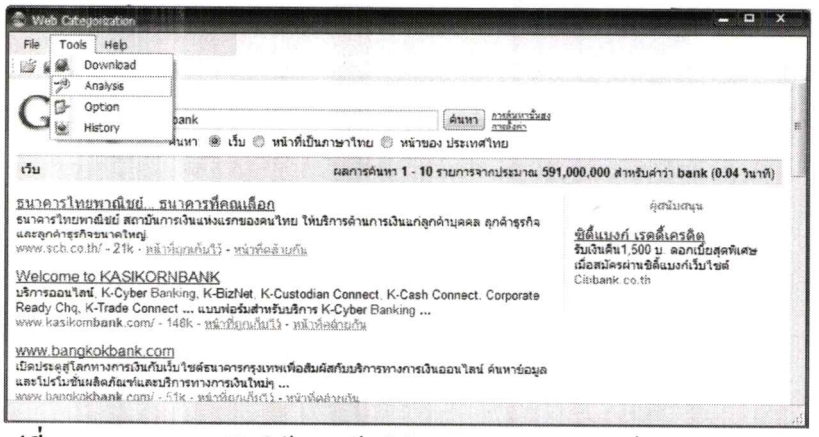


รูปที่ 5.22 แสดงหน้าจอย่อย Download แสดงการทำงานขณะยกเลิกการดาวน์โหลดเว็บข้อมูลตัวอย่าง

5.2.2 ส่วนคำนวณหาผลลัพธ์ (Calculate Part)

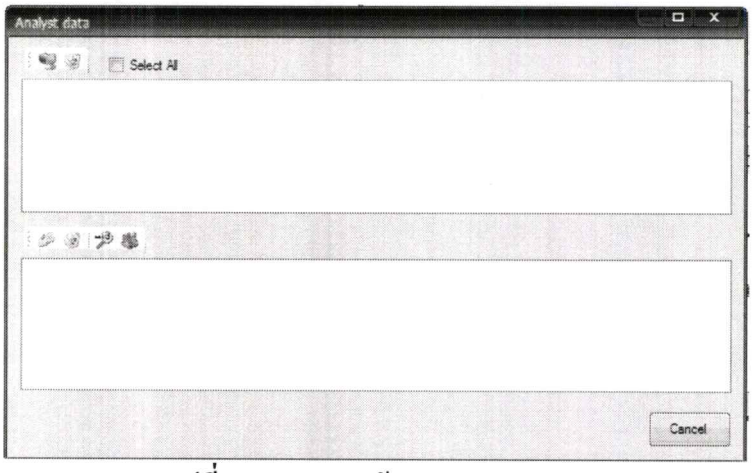
5.2.2.1 เริ่มขั้นตอนการคำนวณหาผลลัพธ์จากข้อมูลเว็บตัวอย่าง (Calculate Sampling Web site)

ที่หน้าจอโปรแกรมหลักไปที่ Tools > Analyst เพื่อเข้าสู่หน้าจอย่อยส่วนการค้นหาข้อมูลเว็บตัวอย่างดังรูป



รูปที่ 5.23 แสดงการเปิดใช้งานเข้าสู่ส่วนการกรองคำศัพท์และประมวลผล

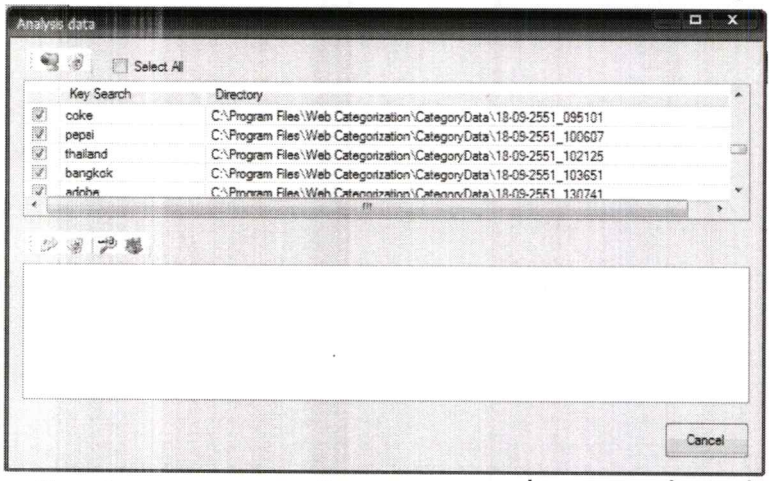
5.2.2.2 เข้าสู่ส่วนหน้าจอย่อย Analysis data



รูปที่ 5.24 แสดงหน้าจอย่อย Analysis

5.2.2.2.1 ส่วนการแสดงรายชื่อกลุ่มเว็บตัวอย่างในโฟลเดอร์ Category

คลิกไอคอนรูป (Browse) เพื่อให้โปรแกรมทำการโหลดรายชื่อกลุ่มเว็บตัวอย่างในโฟลเดอร์ที่ได้ทำการติดตั้งโปรแกรมโฟลเดอร์ย่อยชื่อ CategoryData มาแสดงดังรูป



รูปที่ 5.25 แสดงหน้าจอย่อย Analysis และเข้าสู่ส่วน browse เพื่อแสดงรายชื่อเว็บที่ต้องการวิเคราะห์

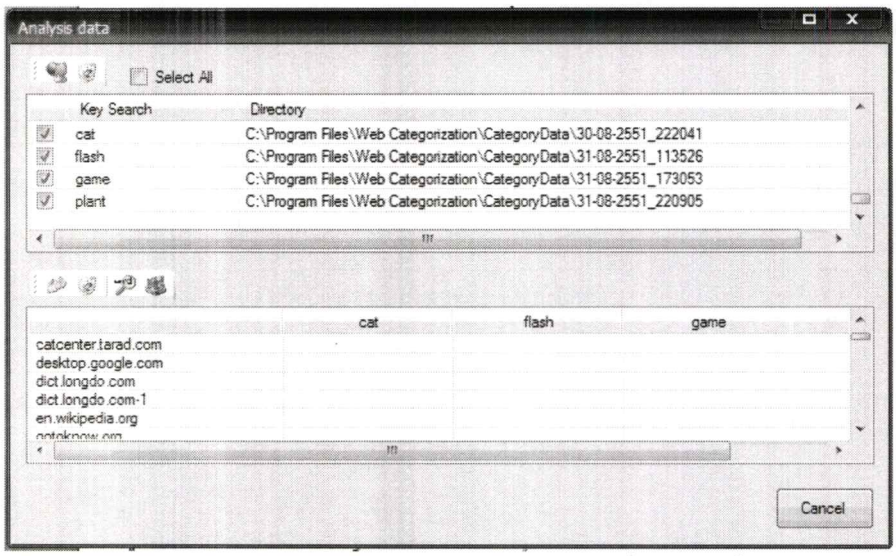
รายชื่อกลุ่มเว็บตัวอย่างนี้จะเห็นว่าประกอบไปด้วยคำที่ผู้ใช้ได้ป้อนเพื่อทำการค้นหา (Key Search or Query) กับโฟลเดอร์กลุ่มเว็บตัวอย่างที่ถูกเก็บไว้ในเครื่อง (Directory)

ให้ทำการเลือกกลุ่มเว็บตัวอย่างที่ต้องการคำนวณด้วยการ เช็คกล่องเช็คบล็อค (Check box) ให้เป็นเครื่องหมายถูกแต่ถ้าต้องการเลือกกลุ่มเว็บตัวอย่างทั้งหมดก็สามารถที่จะทำได้โดย เช็คกล่องเช็คบล็อคชื่อ “Select All” ด้านบนให้เป็นเครื่องหมายถูกจะเห็นว่ารายชื่อเว็บตัวอย่างทั้งหมดจะถูกเลือกทุกอันในทันที

5.2.2.2.2 ส่วนการโหลดกลุ่มเว็บตัวอย่างเข้าสู่ตารางเตรียมการคำนวณ

คลิกไอคอนรูป (Load Data) เพื่อนำรายชื่อกลุ่มของเว็บไซต์ที่ได้เลือกไว้ เข้าไปอยู่ในตารางเตรียมการคำนวณโดยที่โปรแกรมจะทำการเรียงรายชื่อเว็บในโฟลเดอร์ของกลุ่มข้อมูลเว็บตัวอย่างในแถวของตารางและเรียงรายชื่อคำที่ผู้ใช้ได้ทำการค้นไว้ (Query) ซึ่งตอนนี้เราจะเรียกว่าเทอมที่สนใจ (term : t) ในคอลัมภ์ของตารางดังรูปด้านล่าง

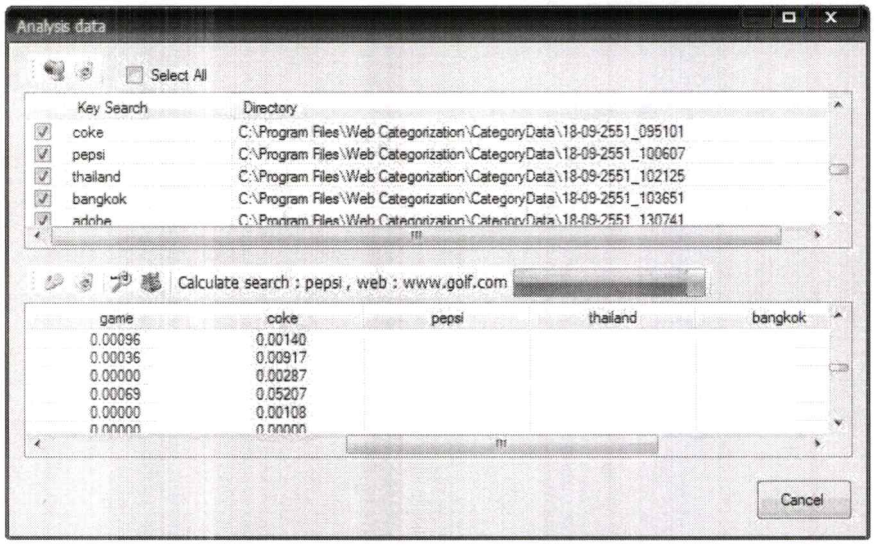
ถ้าโหลดกลุ่มของเว็บตัวอย่างมาในตารางเพียงกลุ่มเดียวโปรแกรมจะไม่สามารถทำการคำนวณโดยกดปุ่ม (Analysis) ได้เนื่องจากจะไม่มีน้ำหนัก TF.IDF มากกว่า 1 เพื่อนำไปคำนวณในการหามุมความสัมพันธ์โคไซน์ในสูตรได้



รูปที่ 5.26 แสดงหน้าจอย่อย Analysis และเข้าสู่ส่วน Load Data เพื่อนำรายชื่อเว็บและเทอมที่สนใจเตรียมพร้อมเข้าสู่การวิเคราะห์ในตาราง

5.2.2.2.3 ส่วนการคำนวณหาค่า TF.IDF จากกลุ่มของเว็บตัวอย่างมากกว่า 2 กลุ่มขึ้นไป

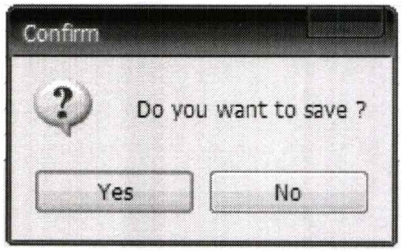
ในส่วนการคำนวณหาค่า TF.IDF จากกลุ่มของเว็บตัวอย่างมากกว่า 2 กลุ่มขึ้นไปนี้ โปรแกรมจะทำการคำนวณหาค่า t, T, D และ d เพื่อนำค่าทั้งหมดที่ได้นี้ไปคำนวณต่อในสูตร TF.IDF = (t/T)/(log(D/d)) และจะนำผลลัพธ์การคำนวณค่า TF.IDF ออกมาแสดงในตารางดังรูปต่อไปนี้



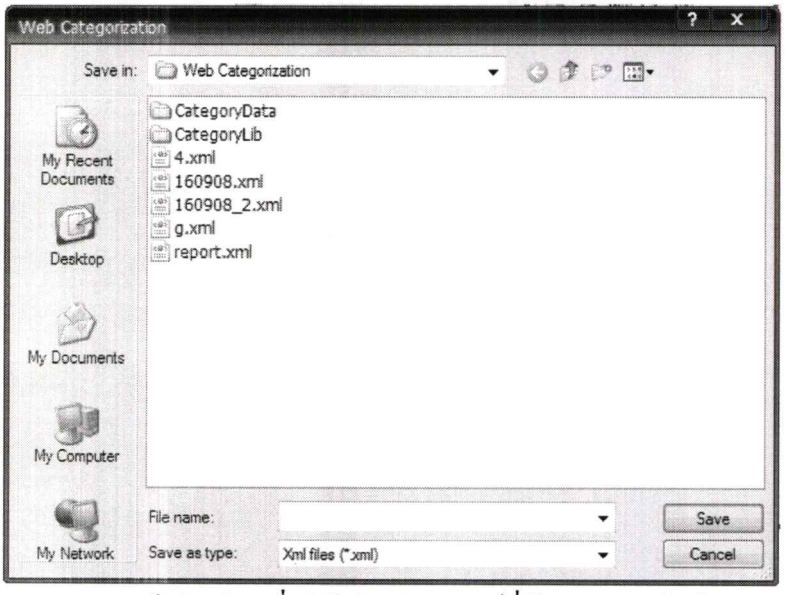
รูปที่ 5.27 แสดงหน้าจอย่อย Analysis และเข้าสู่ส่วน Analysis data ทำการประมวลผลข้อมูล

5.2.2.2.4 ส่วนการเก็บบันทึกข้อมูลผลลัพธ์จากการคำนวณไว้ในไฟล์ XML

คลิกไอคอนรูป (Save data) เพื่อการบันทึกข้อมูลผลลัพธ์จากการคำนวณไว้ในไฟล์ XML โดยที่ค่าที่ถูกเก็บทั้งหมดก็จะมี เทอมที่สนใจ, รายชื่อเว็บทุกเว็บ, ค่า t, T, D, d และ TF.IDF เพื่อที่จะนำไปแสดงเป็นตารางรายงานค่ามุมความสัมพันธ์โคไซน์ในหัวข้อถัดไป ทั้งยังสามารถทำให้เก็บค่าผลลัพธ์ไว้ไปแสดงอีกกี่ครั้งก็ได้ ไม่ต้องทำการคำนวณใหม่ทุกครั้ง



รูปที่ 5.28 แสดงหน้าต่างย่อย Confirm ยืนยันว่าต้องการบันทึกข้อมูลผลลัพธ์



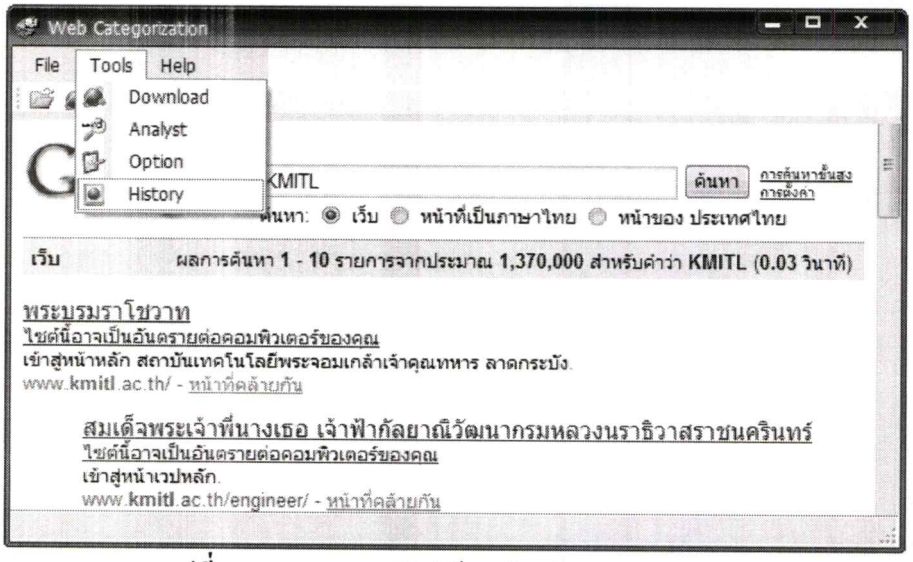
รูปที่ 5.29 แสดงหน้าต่างย่อยเพื่อให้เลือกโฟลเดอร์ที่ต้องการบันทึกข้อมูลผลลัพธ์

ทำการเลือกโฟลเดอร์และตั้งชื่อไฟล์ที่ต้องการบันทึกในรูปเอกสาร XML แล้วคลิกปุ่ม Save

5.2.3 ส่วนแสดงรายงาน (Report Part)

5.2.3.1 เริ่มขั้นตอนการแสดงรายงาน

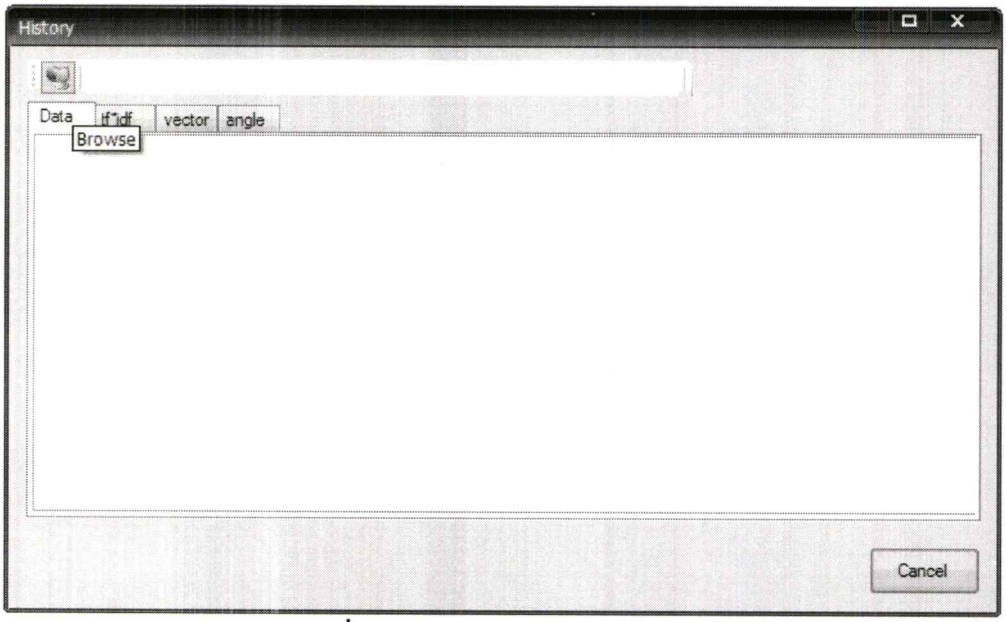
ที่หน้าจอโปรแกรมหลักไปที่ Tools > History เพื่อเข้าสู่หน้าจอย่อยส่วนแสดงรายงาน ดังรูป



รูปที่ 5.30 แสดงการเปิดใช้งานในส่วนแสดงรายงาน

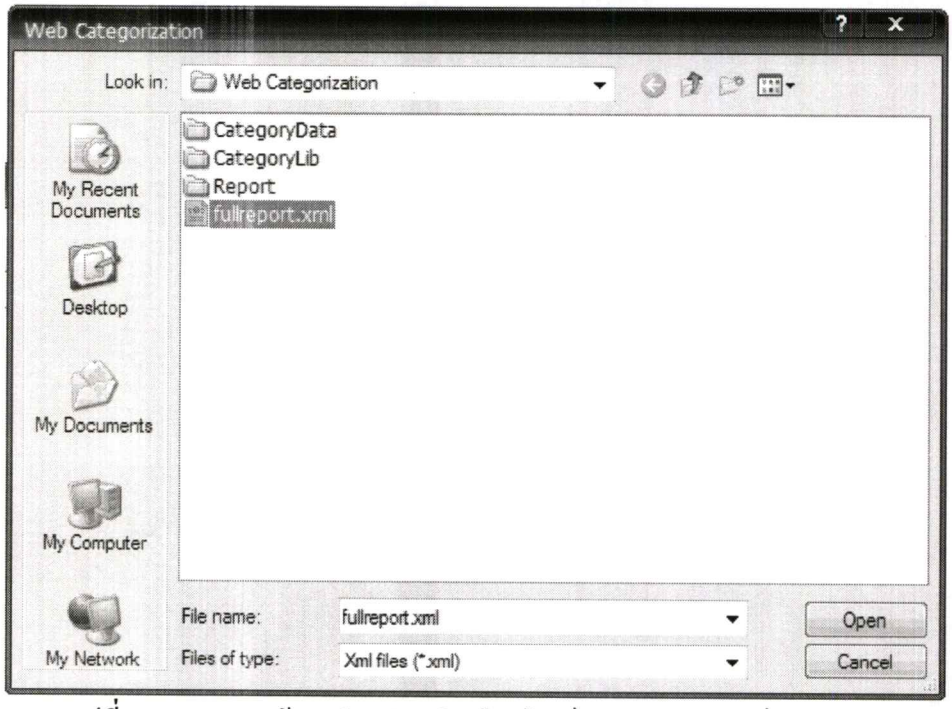
5.2.3.2 เข้าสู่ส่วนหน้าจอย่อย History

คลิกไอคอนรูป (Browse) เพื่อแสดงหน้าจอย่อย



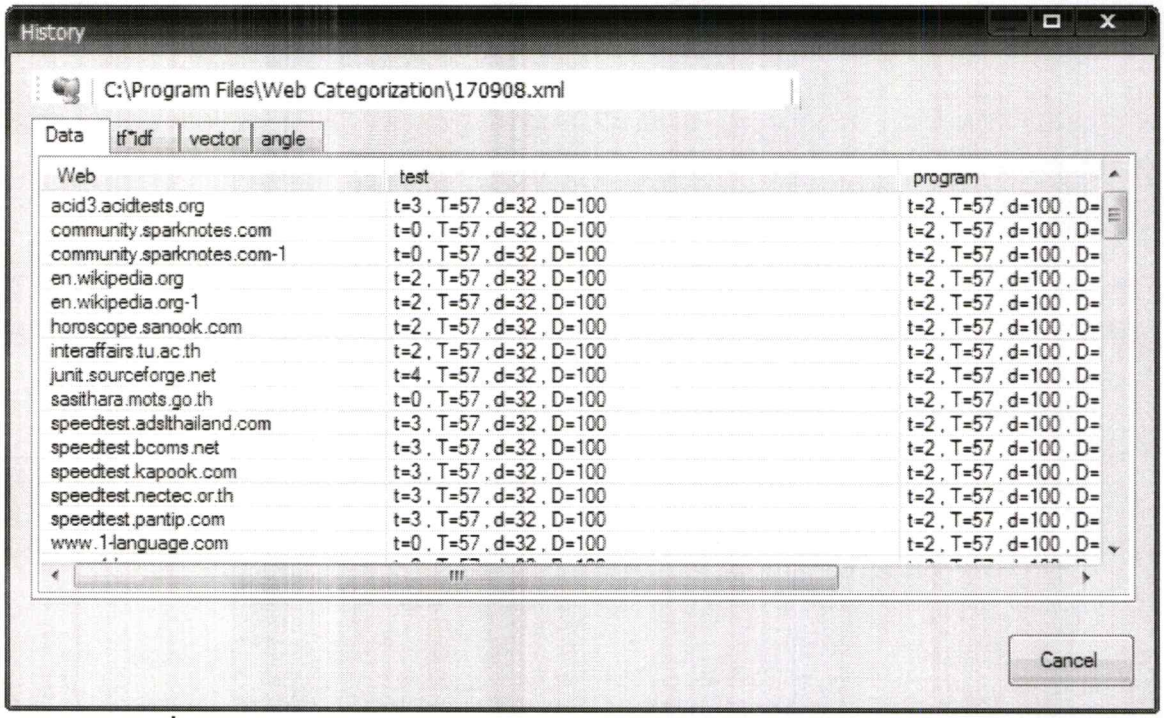
รูปที่ 5.31 แสดงหน้าจอย่อย History

5.2.3.3 ส่วนหน้าต่างย่อยสำหรับเลือกไฟล์ XML มาแสดงเป็นรายงาน



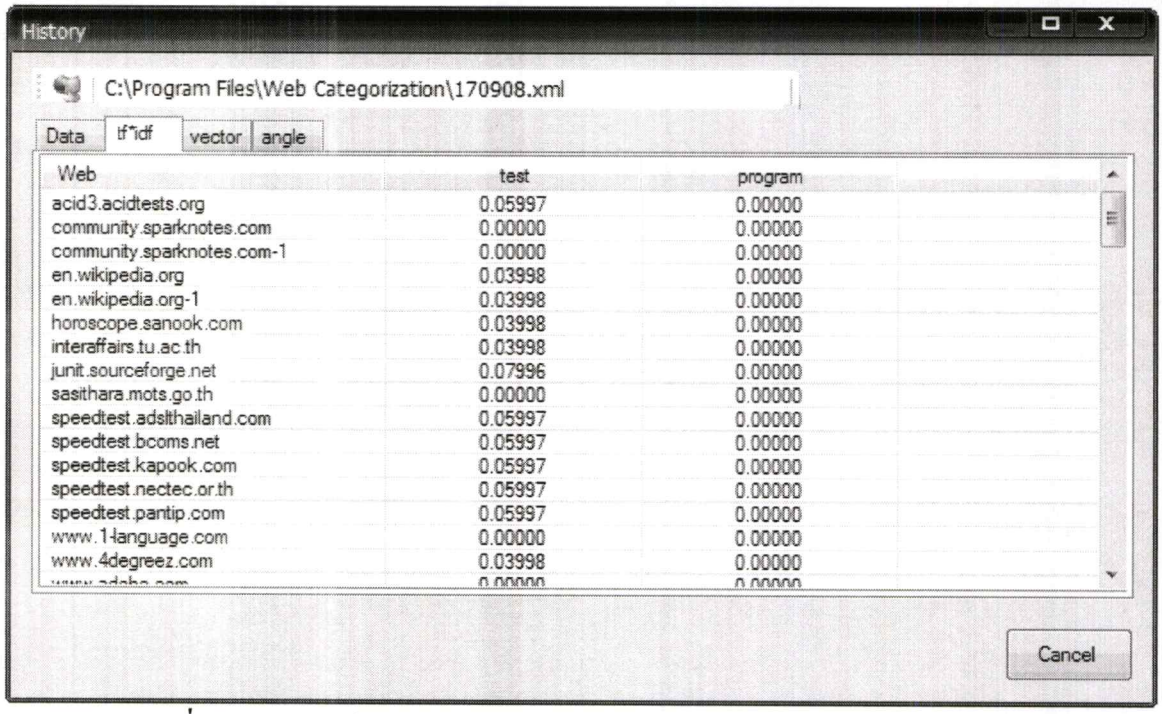
รูปที่ 5.32 แสดงหน้าจอย่อยสำหรับเลือกไฟล์ XML มาแสดงเป็นรายงาน

5.2.3.4 ส่วนแท็บชีทแสดงค่า Data ทั่วไปของแต่ละเทอมในแต่ละเว็บไซต์



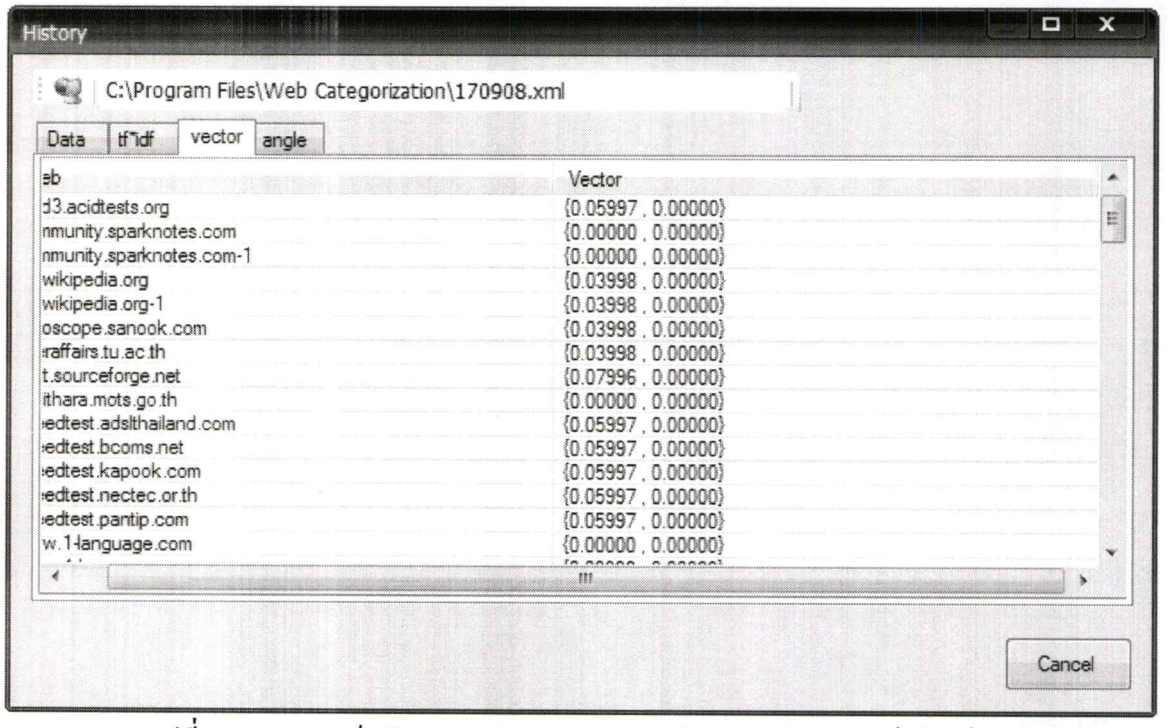
รูปที่ 5.33 แสดงแท็บชีทแสดงค่า Data ทั่วไปของแต่ละเทอมในแต่ละเว็บไซต์

5.2.3.5 ส่วนแท็บชีทแสดงค่าน้ำหนัก TF.IDF ของแต่ละเทอมในแต่ละเว็บไซต์



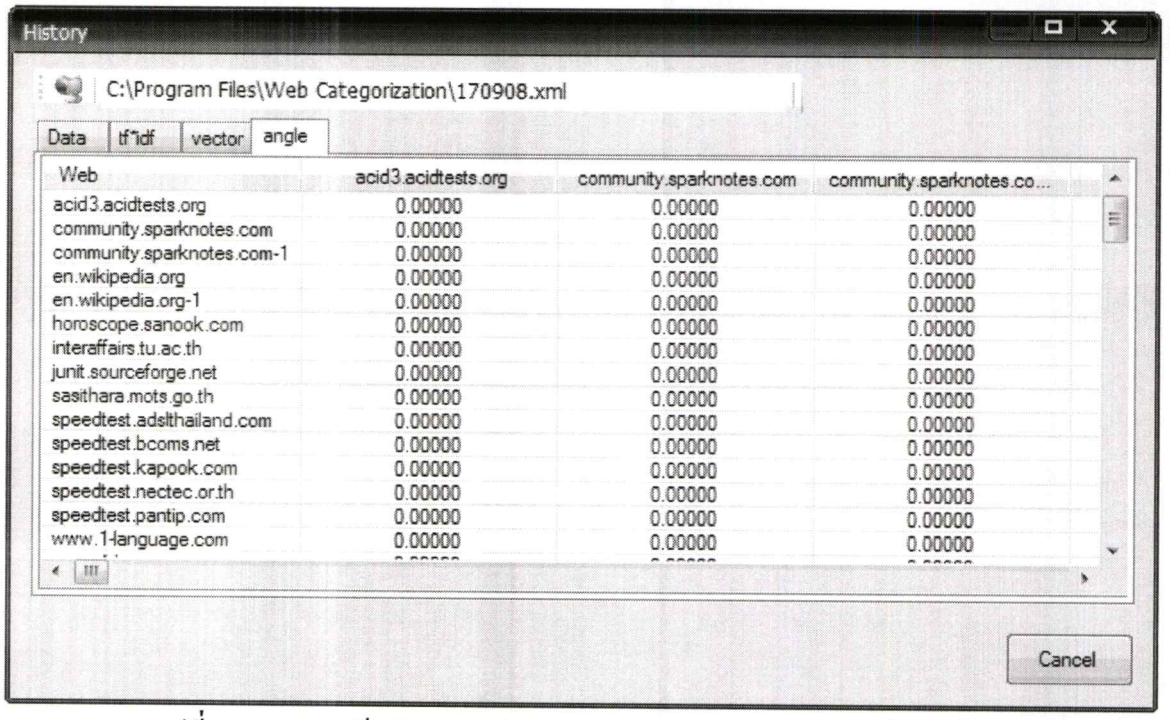
รูปที่ 5.34 แสดงแท็บชีทแสดงค่า TF.IDF ของแต่ละเทอมในแต่ละเว็บไซต์

5.2.3.6 ส่วนแท็บชีทแสดงค่า Vector ของแต่ละเทอมในแต่ละเว็บไซต์



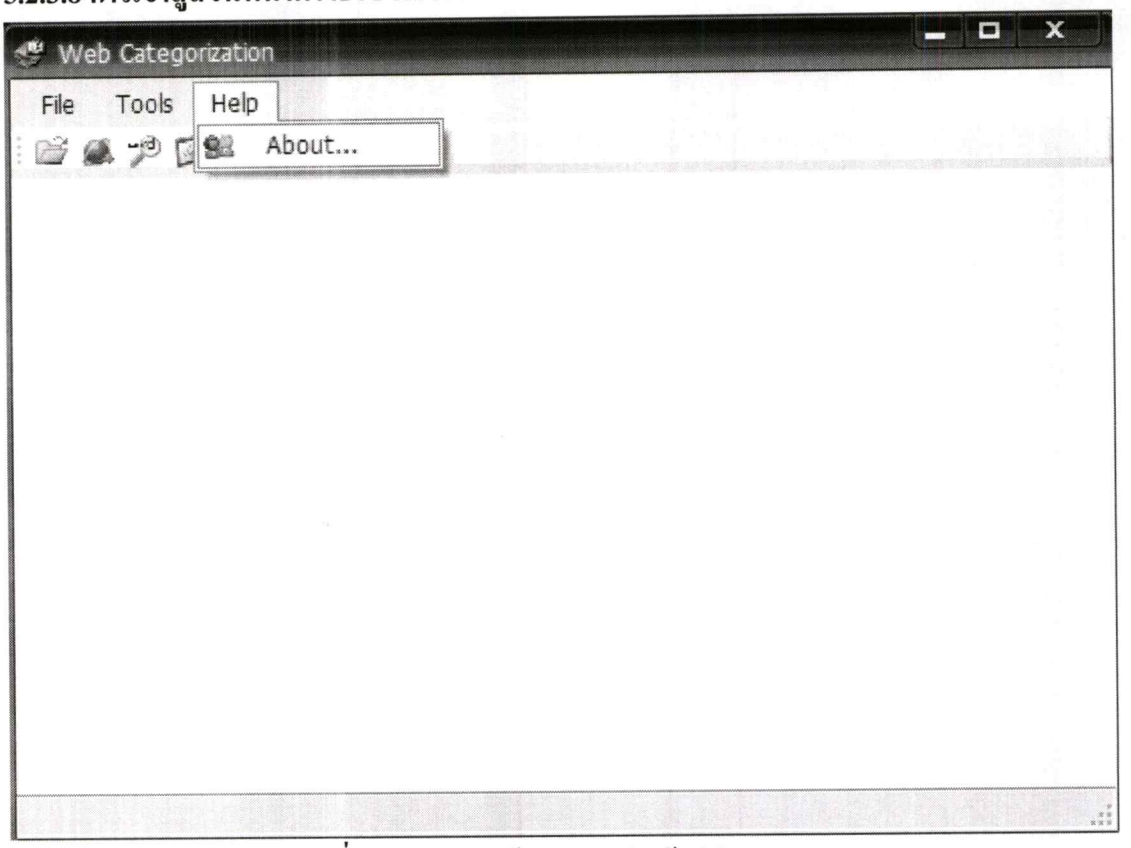
รูปที่ 5.35 แสดงแท็บชีทแสดงค่า Vector ของแต่ละเทอมในแต่ละเว็บไซต์

5.2.3.7 ส่วนแท็บชีทแสดงค่า Angle ของแต่ละเทอมในแต่ละเว็บไซต์

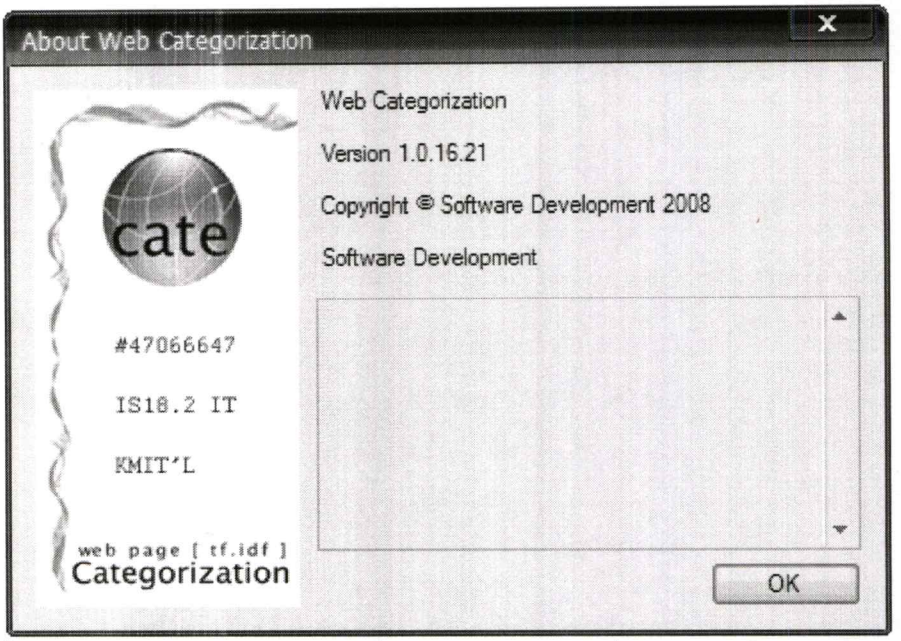


รูปที่ 5.36 แสดงแท็บชีทแสดงค่า Angleของแต่ละเทอมในแต่ละเว็บไซต์

5.2.3.8 การเข้าสู่ส่วนหน้าต่างย่อย About



รูปที่ 5.37 แสดงหน้าจอการเปิดเข้าสู่ส่วน About



รูปที่ 5.38 แสดงหน้าจอย่อย About เพื่อแสดงรายละเอียดเกี่ยวกับเวอร์ชั่นของโปรแกรม

บทที่ 6

วิเคราะห์ผลลัพธ์การคำนวนในโปรแกรม Web Categorization

6.1 ผลลัพธ์การคำนวณจากโปรแกรม Web Categorization

ผลลัพธ์การคำนวณจากโปรแกรม Web Categorization ได้ถูกนำมาแสดงในตารางที่ 6.1 – ตารางที่ 6.7 ดังต่อไปนี้

- ตารางที่6.1 แสดงผลการทดลองตารางชื่อเว็บไซต์ทั้งหมด
- ตารางที่6.2 แสดงผลการทดลองค่าเทอมที่สนใจที่พบในเอกสาร (t) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด
- ตารางที่6.3 แสดงผลการทดลองค่าจำนวนคำทั้งหมดที่พบในเอกสาร (T) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด
- ตารางที่6.4 แสดงผลการทดลองค่าเอกสารทั้งหมด (D) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด
- ตารางที่6.5 แสดงผลการทดลองค่าเอกสารที่พบเทอมที่สนใจ (d) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด
- ตารางที่6.6 แสดงผลการทดลองค่าน้ำหนัก TF.IDF ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด
- ตารางที่6.7 แสดงผลการทดลองค่ามุมความสัมพันธ์ (ø) ระหว่างเว็บไซต์แต่ละคู่

เนื่องจากชื่อเว็บไซต์ที่นำมาทำการวิเคราะห์นี้ บางชื่อมีความยาวพอสมควร จึงได้ทำตารางที่ 6.1 เพื่อแสดงชื่อเว็บไซต์ทั้งหมดซึ่งจะใช้ตัวแปร D (Document or Web site) ร่วมกับตัวเลขเพื่อแทนชื่อเว็บไซต์นั้นๆกรุณาใช้ตารางที่ 6.1 ในการประกอบการพิจารณาผลลัพธ์ในตารางที่ 6.2 – ตารางที่ 6.7

ตารางที่6.1 แสดงผลการทดลองตารางชื่อเว็บไซต์ทั้งหมด

D1	ebank.kasikornbank.com	D252	en.wikipedia.org
D2	en.wikipedia.org	D253	forum.nkc.kku.ac.th
D3	en.wikipedia.org-1	D254	game.mthai.com
D4	kbank.nfe.go.th	D255	hijack.smfforfree3.com
D5	www.aclbank.com	D256	music.mercigod.com
D6	www.asbbank.co.nz	D257	pepsistuff.amazon.com
D7	www.baac.or.th	D258	play.kapook.com
D8	www.bangkokbank.com	D259	starwars.igetweb.com
D9	www.bangkokbank.com-1	D260	th.wordpress.com
D10	www.bank-art.com	D261	video.gigchat.com
D11	www.bankofamerica.com	D262	video.google.com
D12	www.bankthai.co.th	D263	video.mthai.com
D13	www.bmb.co.th	D264	video.sanook.com
D14	www.bot.or.th	D265	winkkk.com
D15	www.bot.or.th-1	D266	www.adintrend.com
D16	www.cb.ktb.co.th	D267	www.bangkokcity.com
D17	www.citi.com	D268	www.bloggang.com
D18	www.citibank.co.th	D269	www.blognone.com
D19	www.comerica.com	D270	www.clipmass.com
D20	www.exim.go.th	D271	www.dek-d.com
D21	www.ghb.co.th	D272	www.eguide.co.th
D22	www.gsb.or.th	D273	www.eguide.co.th-1
D23	www.hsbc.co.th	D274	www.expert2you.com
D24	www.hsbc.com	D275	www.exteen.com
D25	www.isbt.co.th	D276	www.fwdder.com
D26	www.kasikornbank.com	D277	www.hunsa.com
D27	www.kasikornbank.com-1	D278	www.isriya.com
D28	www.kiatnakin.co.th	D279	www.jedineko.com
D29	www.krungsri.com	D280	www.kosanathai.com
D30	www.krungsri.com-1	D281	www.kosanathai.com-1
D31	www.ktb.co.th	D282	www.oppapers.com
D32	www.neopets.com	D283	www.pbg.com
D33	www.royalbank.com	D284	www.pepsi.ca
D34	www.scb.co.th	D285	www.pepsi.co.uk

ตารางที่6.1 (ต่อ)

D35	www.scb.co.th-1	D286	www.pepsi.com
D36	www.scib.co.th	D287	www.pepsi.nl
D37	www.smebank.co.th	D288	www.pepsiamericas.com
D38	www.smebank.co.th-1	D289	www.pepsico.com
D39	www.standardchartered.co.th	D290	www.pepsithai.com
D40	www.suntrust.com	D291	www.pepsithai.com-1
D41	www.thanachart.com	D292	www.pepsiworld.com
D42	www.thanachartbank.co.th	D293	www.pollpub.com
D43	www.tisco.co.th	D294	www.rssthai.com
D44	www.tmbbank.com	D295	www.ryt9.com
D45	www.tmbbank.com-1	D296	www.ryt9.com-1
D46	www.uob.co.th	D297	www.vcharkarn.com
D47	www.wellsfargo.com	D298	www.video.articlesphere.com
D48	www.worldbank.org	D299	www.yopi.co.th
D49	www.worldbank.org-1	D300	www.youtube.com
D50	www.youtube.com	D301	en.wikipedia.org
D51	e-learning.tu.ac.th	D302	en.wikipedia.org-1
D52	en.wikipedia.org	D303	kanchanapisek.or.th
D53	en.wikipedia.org-1	D304	maps.google.com
D54	finance.google.com	D305	th.jobsdb.com
D55	finance.sympatico.msn.ca	D306	th.msn.com
D56	finance.truehits.net	D307	th.yahoo.com
D57	finance.yahoo.ca	D308	thai.tourismthailand.org
D58	finance.yahoo.ca-1	D309	wikitravel.org
D59	finance.yahoo.com	D310	www.02aflower.com
D60	finance.yahoo.com-1	D311	www.afsthailand.org
D61	finance.yahoo.com.au	D312	www.agoda.com
D62	finmin.nic.in	D313	www.airportthai.co.th
D63	finmin.nic.in-1	D314	www.alexa.com
D64	gotoknow.org	D315	www.amnesty.or.th
D65	in.finance.yahoo.com	D316	www.athailand.com
D66	money.aol.com	D317	www.bot.or.th
D67	uk.finance.yahoo.com	D318	www.britishcouncil.org
D68	www.branchorientation.com	D319	www.cia.gov

ตารางที่6.1 (ต่อ)

D69	www.bus.tu.ac.th	D320	www.ethnologue.com
D70	www.careers-in-finance.com	D321	www.ftpi.or.th
D71	www.cob.ohio-state.edu	D322	www.fulbrightthai.org
D72	www.eguide.co.th	D323	www.google.co.th
D73	www.eguide.co.th-1	D324	www.gothailand.com
D74	www.fin.gc.ca	D325	www.inet.co.th
D75	www.fin.gov.on.ca	D326	www.ipthailand.org
D76	www.fin24.co.za	D327	www.lonelyplanet.com
D77	www.finance.army.mil	D328	www.microsoft.com
D78	www.finance.gov.ab.ca	D329	www.mtvthailand.com
D79	www.finance.gov.au	D330	www.nectec.or.th
D80	www.finance.gov.ie	D331	www.nlt.go.th
D81	www.finance.gov.pk	D332	www.nod32th.com
D82	www.finance.gov.sk.ca	D333	www.nso.go.th
D83	www.finance.ku.ac.th	D334	www.parliament.go.th
D84	www.gnb.ca	D335	www.railway.co.th
D85	www.gov.mb.ca	D336	www.school.net.th
D86	www.gov.ns.ca	D337	www.set.or.th
D87	www.hfathai.com	D338	www.sriwittayapaknam.ac.th
D88	www.jobthaiweb.com	D339	www.tei.or.th
D89	www.mint.com	D340	www.thailand.com
D90	www.mof.go.th	D341	www.thailand.idp.com
D91	www.mof.go.th-1	D342	www.thailand.net
D92	www.news.com.au	D343	www.thailandmaps.net
D93	www.news.com.au-1	D344	www.thailandpost.com
D94	www.nyc.gov	D345	www.thaimet.tmd.go.th
D95	www.nyc.gov-1	D346	www.tourismthailand.org
D96	www.senate.gov	D347	www.toyota.co.th
D97	www.tisco.co.th	D348	www.tu.ac.th
D98	www.youtube.com	D349	www.worldbank.org
D99	www.youtube.com-1	D350	www.yellowpages.co.th
D100	www2.mof.go.th	D351	www.youtube.com
D101	a4esl.org	D352	bangkok.sawadee.com
D102	alistapart.com	D353	bangkok2night.com

ตารางที่6.1 (ต่อ)

D103	api.rubyonrails.org	D354	commons.wikimedia.org
D104	en.wikipedia.org	D355	en.wikipedia.org
D105	en.wikipedia.org-1	D356	eo.wikipedia.org
D106	flash.desy.de	D357	lonelyplanet.com
D107	flash.thaimisc.com	D358	maps.google.com
D108	flash.uchicago.edu	D359	wikitravel.org
D109	get.adobe.com	D360	www.absoluteyogabangkok.com
D110	livedocs.adobe.com	D361	www.agoda.com
D111	multimedia.journalism.berkeley.edu	D362	www.airportthai.co.th
D112	th.wikipedia.org	D363	www.at-bangkok.com
D113	www.adobe.com	D364	www.bangkok-maps.com
D114	www.adobe.com-1	D365	www.bangkok-today.com
D115	www.arip.co.th	D366	www.bangkok.com
D116	www.bcoms.net	D367	www.bangkok.go.th
D117	www.bestflashanimationsite.com	D368	www.bangkokairportonline.com
D118	www.blognone.com	D369	www.bangkokbank.com
D119	www.eas.asu.edu	D370	www.bangkokbiznews.com
D120	www.echoecho.com	D371	www.bangkokcentrehotel.com
D121	www.flash.gr	D372	www.bangkokfightclub.com
D122	www.flash.org	D373	www.bangkokhearthospital.com
D123	www.flashearth.com	D374	www.bangkokmetro.co.th
D124	www.flashkit.com	D375	www.bangkokpost.com
D125	www.flashkit.com-1	D376	www.bangkoksite.com
D126	www.flashloaded.com	D377	www.bangkoktourist.com
D127	www.gigchat.com	D378	www.barcampbangkok.org
D128	www.grandmasterflash.com	D379	www.bigth.com
D129	www.hotscripts.com	D380	www.bki.co.th
D130	www.hyperborea.org	D381	www.bkkflights.com
D131	www.hyperborea.org-1	D382	www.brazilembassy.or.th
D132	www.imdb.com	D383	www.bts.co.th
D133	www.macromedia.com	D384	www.bu.ac.th
D134	www.marmoon.com	D385	www.bu.ac.th-1
D135	www.meganova.com	D386	www.chula.ac.th
D136	www.nectec.or.th	D387	www.emeraldhotel.com

ตารางที่6.1 (ต่อ)

D137	www.rsc.org	D388	www.hotelclub.net
D138	www.shockwave.com	D389	www.ileabangkok.com
D139	www.swishzone.com	D390	www.jalcargobkk.com
D140	www.thaiall.com	D391	www.jfbkk.or.th
D141	www.thaiflashdev.com	D392	www.officebangkok.com
D142	www.tutorialized.com	D393	www.pridefestival.org
D143	www.upscale.utoronto.ca	D394	www.samuihospital.com
D144	www.useit.com	D395	www.thegrandhotelgroup.com
D145	www.w3schools.com	D396	www.timeanddate.com
D146	www.webaim.org	D397	www.tourismthailand.org
D147	www.webthaidd.com	D398	www.worldfilmbkk.com
D148	www.webthaidd.com-1	D399	www.wunderground.com
D149	www.youtube.com	D400	www.youtube.com
D150	www.zalim-code.com	D401	www.youtube.com-1
D151	audition.playpark.com	D402	www.zabzaa.com
D152	flashgame.tarad.com	D403	adobe.elementk.com
D153	game.deedeejang.com	D404	adobe.istreamplanet.com
D154	game.deedeejang.com-1	D405	adobemax2007.com
D155	game.giggog.com	D406	download.siamhrm.com
D156	game.hunsa.com	D407	en.wikipedia.org
D157	game.hunsa.com-1	D408	en.wikipedia.org-1
D158	game.kapook.com	D409	feeds.adobe.com
D159	game.meemodel.com	D410	finance.yahoo.com
D160	game.mthai.com	D411	flex.org
D161	game.popcornfor2.com	D412	get.adobe.com
D162	game.sanook.com	D413	graphicssoft.about.com
D163	game.sanook.com-1	D414	guru.google.co.th
D164	game.siamha.com	D415	kb.adobe.com
D165	game.siamhrm.com	D416	labs.adobe.com
D166	game.siamhrm.com-1	D417	labs.adobe.com-1
D167	game.siamza.com	D418	opensource.adobe.com
D168	game.thaihealth.net	D419	photoshopnews.com
D169	game.unseencar.com	D420	searchpdf.adobe.com
D170	gamecenter.kapook.com	D421	www.abobe.com

ตารางที่6.1 (ต่อ)

D171	games.narak.com	D422	www.acrobat.com
D172	happy.teenee.com	D423	www.adobe.com
D173	mobilemagic.sanook.com	D424	www.adobe.com-1
D174	play.kapook.com	D425	www.adobebuilder.com
D175	postjung.com	D426	www.adobeevangelists.com
D176	www.bkkonline.com	D427	www.adobepress.com
D177	www.compgamer.com	D428	www.adobetutorialz.com
D178	www.empireinteractive.com	D429	www.adobetutorialz.com-1
D179	www.funwhan.com	D430	www.amanwithapencil.com
D180	www.game1188.com	D431	www.aptana.com
D181	www.gamehothit.com	D432	www.bu.ac.th
D182	www.hitsplay.com	D433	www.c4lpt.co.uk
D183	www.icygang.com	D434	www.cs.cmu.edu
D184	www.manager.co.th	D435	www.download.com
D185	www.marmoon.com	D436	www.downloadadobe.net
D186	www.mediathai.net	D437	www.easyhome.in.th
D187	www.narak.com	D438	www.epsea.org
D188	www.online-station.net	D439	www.filehippo.com
D189	www.pantip.com	D440	www.it-guides.com
D190	www.siam2.com	D441	www.macromedia.com
D191	www.siamcomic.com	D442	www.nectec.or.th
D192	www.siamcomic.com-1	D443	www.officesnapshots.com
D193	www.siamzone.com	D444	www.pdf-format.com
D194	www.siamzone.com-1	D445	www.smashingmagazine.com
D195	www.suansanook.com	D446	www.smethai.com
D196	www.thaifreegame.com	D447	www.thaiadobeuser.com
D197	www.tlcthai.com	D448	www.thaiall.com
D198	www.yenta4.com	D449	www.thaiware.com
D199	www.yenta4.com-1	D450	www.thaiware.com-1
D200	www.zabzaa.com	D451	www.tucows.com
D201	coke-sleeping.hi5.com	D452	www.vue.com
D202	en.wikipedia.org	D453	en.wikipedia.org
D203	en.wikipedia.org-1	D454	en.wikipedia.org-1
D204	finance.yahoo.com	D455	game.siamhrm.com

ตารางที่6.1 (ต่อ)

D205	guru.google.co.th	D456	golf.ru.ac.th
D206	news.cnet.com	D457	guru.google.co.th
D207	shopping.sanook.com	D458	sports.espn.go.com
D208	video.google.com	D459	th.thaigolfer.com
D209	video.google.com-1	D460	th.thaigolfer.com-1
D210	www-cse.ucsd.edu	D461	thai.tourismthailand.org
D211	www.adintrend.com	D462	webindex.sanook.com
D212	www.adrants.com	D463	wordpress.com
D213	www.ajc.com	D464	www.108golfer.com
D214	www.balajicoke.com	D465	www.alpinegolfclub.com
D215	www.co.coke.tx.us	D466	www.ethaimusic.com
D216	www.coca-cola.com	D467	www.gassangolf.com
D217	www.coke.co.nz	D468	www.golf.com
D218	www.coke.dk	D469	www.golf2thailand.com
D219	www.coke.dk-1	D470	www.golfasian.com
D220	www.coke.hu	D471	www.golfdigest.com
D221	www.coke.net	D472	www.golfinthailand.com
D222	www.cokecce.co.uk	D473	www.golfmikethailand.com
D223	www.cokedares.com	D474	www.golforient.com
D224	www.cokefacts.com	D475	www.golfvariety.com
D225	www.cokemachineglow.com	D476	www.kirimaya.com
D226	www.cokesideoflifethai.com	D477	www.lochpalm.com
D227	www.cokezerogame.com	D478	www.maejogolfclub.com
D228	www.dietcoke.com	D479	www.muangkaewgolf.com
D229	www.eepybird.com	D480	www.palmhills-golf.com
D230	www.eepybird.com-1	D481	www.panyagolf.com
D231	www.guardian.co.uk	D482	www.pga.com
D232	www.gujaratnre.com	D483	www.royalchiangmai.com
D233	www.icoke.ca	D484	www.royalgolfclubs.com
D234	www.killercoke.org	D485	www.royalhillsresort.com
D235	www.metacafe.com	D486	www.santiburi.com
D236	www.msnbc.msn.com	D487	www.sawangresortgolf.com
D237	www.music.coca-cola.com	D488	www.sportsline.com
D238	www.mycoke.com	D489	www.springfieldresort.com

ตารางที่6.1 (ต่อ)

D239	www.mycokemusic.com	D490	www.summitwindmillgolfclub.com
D240	www.mycokerewards.com	D491	www.suwangolf.com
D241	www.pl8s.com	D492	www.tga.or.th
D242	www.revver.com	D493	www.thaigolfer.com
D243	www.rootsweb.ancestry.com	D494	www.thaigolfer.com-1
D244	www.snopes.com	D495	www.thaigolfguide.com
D245	www.thecoca-colacompany.com	D496	www.thailandgolfcourse.com
D246	www.tlcthai.com	D497	www.thegolfchannel.com
D247	www.vcharkarn.com	D498	www.thepinegolf.com
D248	www.woccatlanta.com	D499	www.tjgc.org
D249	www.youtube.com	D500	www.wingolfer.com
D250	www.youtube.com-1	D501	www.youtube.com
D251	chiangmaischnauzer.tarad.com	D502	www.youtube.com-1

ตารางที่6.2 แสดงผลการทดลองค่าเทอมที่สนใจที่พบในเอกสาร (t) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D1	261	0	0	0	0	0	0	0	0	0
D2	5	0	0	0	0	0	5	0	0	0
D3	5	0	0	0	0	0	0	0	0	0
D4	23	0	0	0	0	0	0	0	0	0
D5	646	1	26	0	0	0	0	0	5	0
D6	443	0	0	0	0	0	0	0	0	0
D7	2	1	55	0	0	0	0	0	5	0
D8	329	0	0	0	0	0	0	309	0	0
D9	1410	0	0	0	0	0	0	1233	1	0
D10	471	0	6	0	0	0	0	0	0	0
D11	102	0	0	0	0	0	0	0	0	0
D12	16	0	0	0	0	0	0	0	0	0
D13	0	0	8	0	0	0	17	0	0	0
D14	2	0	0	0	0	0	0	2	0	0
D15	4	0	0	0	0	0	0	0	0	0
D16	7	0	0	0	0	0	1	0	0	0

ตารางที่6.2 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D17	13	0	0	0	0	0	0	0	0	0
D18	321	0	0	0	0	0	0	0	0	0
D19	1	0	62	0	0	0	0	0	2	0
D20	0	0	0	0	0	0	0	0	0	0
D21	638	1	24	0	0	0	2	0	4	0
D22	1	0	1	0	0	0	0	0	0	0
D23	12	0	0	0	0	0	48	0	0	0
D24	135	0	0	0	0	0	6	0	0	0
D25	2	0	9	0	0	0	2	0	0	0
D26	477	0	0	0	0	0	0	0	0	0
D27	25031	0	0	0	0	0	0	0	0	0
D28	5	0	64	0	0	0	0	0	5	0
D29	9	0	47	0	0	0	0	1	0	0
D30	20	0	45	0	0	0	3	4	0	0
D31	6	0	0	0	0	0	1	0	0	0
D32	49	0	6	22	0	0	0	0	0	0
D33	710	0	25	0	0	0	0	0	0	0
D34	4	0	9	0	0	0	0	0	3	0
D35	6	0	0	0	0	0	1	0	0	0
D36	6	0	0	0	0	0	0	0	0	0
D37	538	0	17	0	0	0	0	0	0	0
D38	546	0	22	0	0	0	0	0	0	0
D39	144	3	0	0	0	0	5	3	0	0
D40	0	0	0	0	0	0	0	0	0	0
D41	52	29	54	0	0	0	14	0	0	0
D42	16	0	0	0	0	0	0	0	0	0
D43	99	22	8	0	0	0	0	1	26	0
D44	41	0	0	0	0	0	0	0	0	0
D45	675	2	20	0	0	0	1	0	0	0
D46	22	3	13	0	0	0	1	13	0	10
D47	0	0	0	0	0	0	0	0	0	0

ตารางที่6.2 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D48	231	0	6	0	0	0	62	0	0	0
D49	322	0	4	0	0	0	0	0	0	0
D50	112	0	14	23	0	4	0	0	5	0
D51	0	3	0	0	0	0	0	0	0	0
D52	0	0	0	0	0	0	0	0	0	0
D53	0	5	0	0	0	0	0	0	0	0
D54	0	465	0	0	0	0	0	0	0	0
D55	20	688	11	0	0	0	0	0	4	0
D56	10	237	2	0	0	0	9	1	0	0
D57	0	318	0	0	0	0	0	0	0	0
D58	0	300	0	0	0	0	0	0	0	0
D59	8	360	0	0	0	0	0	0	0	0
D60	4	154	0	0	0	0	0	0	0	0
D61	0	228	0	0	0	0	0	0	0	0
D62	1	22	5	0	0	0	0	0	0	0
D63	1	17	4	0	0	0	0	0	0	0
D64	0	6	2	0	0	0	2	0	0	0
D65	2	211	0	0	0	0	0	0	0	0
D66	16	84	42	0	0	0	0	0	0	0
D67	0	125	0	0	0	0	0	0	0	0
D68	1	489	28	0	0	0	0	0	4	0
D69	2	40	0	0	0	0	4	1	0	0
D70	31	165	0	0	0	0	0	0	0	0
D71	25	35	0	0	0	0	0	0	0	0
D72	4	46	0	0	0	0	0	4	0	0
D73	3	66	0	0	0	0	0	6	0	0
D74	2	27	0	0	0	0	0	0	0	0
D75	0	28	13	0	0	0	0	0	0	0
D76	0	15	0	0	0	0	0	0	0	0
D77	0	153	0	0	0	0	0	0	0	0
D78	0	266	2	0	0	0	0	0	0	0
D79	0	200	0	0	0	0	0	0	0	0

ตารางที่6.3 แสดงผลการทดลองค่าจำนวนคำทั้งหมดที่พบในเอกสาร (T) ระหว่างเทอมที่สนใจกับเว็บไซต์

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D1	646	646	646	646	646	646	646	646	646	646
D2	131	131	131	131	131	131	131	131	131	131
D3	131	131	131	131	131	131	131	131	131	131
D4	862	862	862	862	862	862	862	862	862	862
D5	26828	26828	26828	26828	26828	26828	26828	26828	26828	26828
D6	965	965	965	965	965	965	965	965	965	965
D7	75742	75742	75742	75742	75742	75742	75742	75742	75742	75742
D8	828	828	828	828	828	828	828	828	828	828
D9	32898	32898	32898	32898	32898	32898	32898	32898	32898	32898
D10	5164	5164	5164	5164	5164	5164	5164	5164	5164	5164
D11	1684	1684	1684	1684	1684	1684	1684	1684	1684	1684
D12	770	770	770	770	770	770	770	770	770	770
D13	5625	5625	5625	5625	5625	5625	5625	5625	5625	5625
D14	814	814	814	814	814	814	814	814	814	814
D15	125	125	125	125	125	125	125	125	125	125
D16	6832	6832	6832	6832	6832	6832	6832	6832	6832	6832
D17	1028	1028	1028	1028	1028	1028	1028	1028	1028	1028
D18	904	904	904	904	904	904	904	904	904	904
D19	1332	1332	1332	1332	1332	1332	1332	1332	1332	1332
D20	958	958	958	958	958	958	958	958	958	958
D21	15587	15587	15587	15587	15587	15587	15587	15587	15587	15587
D22	789	789	789	789	789	789	789	789	789	789
D23	856	856	856	856	856	856	856	856	856	856
D24	223118	223118	223118	223118	223118	223118	223118	223118	223118	223118
D25	1827	1827	1827	1827	1827	1827	1827	1827	1827	1827
D26	984	984	984	984	984	984	984	984	984	984
D27	188270	188270	188270	188270	188270	188270	188270	188270	188270	188270
D28	19981	19981	19981	19981	19981	19981	19981	19981	19981	19981
D29	16549	16549	16549	16549	16549	16549	16549	16549	16549	16549
D30	16634	16634	16634	16634	16634	16634	16634	16634	16634	16634
D31	1366	1366	1366	1366	1366	1366	1366	1366	1366	1366
D32	947	947	947	947	947	947	947	947	947	947

ตารางที่6.3 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D33	9155	9155	9155	9155	9155	9155	9155	9155	9155	9155
D34	1763	1763	1763	1763	1763	1763	1763	1763	1763	1763
D35	1880	1880	1880	1880	1880	1880	1880	1880	1880	1880
D36	955	955	955	955	955	955	955	955	955	955
D37	19401	19401	19401	19401	19401	19401	19401	19401	19401	19401
D38	52584	52584	52584	52584	52584	52584	52584	52584	52584	52584
D39	10524	10524	10524	10524	10524	10524	10524	10524	10524	10524
D40	777	777	777	777	777	777	777	777	777	777
D41	22986	22986	22986	22986	22986	22986	22986	22986	22986	22986
D42	758	758	758	758	758	758	758	758	758	758
D43	18217	18217	18217	18217	18217	18217	18217	18217	18217	18217
D44	839	839	839	839	839	839	839	839	839	839
D45	20534	20534	20534	20534	20534	20534	20534	20534	20534	20534
D46	16814	16814	16814	16814	16814	16814	16814	16814	16814	16814
D47	850	850	850	850	850	850	850	850	850	850
D48	779	779	779	779	779	779	779	779	779	779
D49	934	934	934	934	934	934	934	934	934	934
D50	65734	65734	65734	65734	65734	65734	65734	65734	65734	65734
D51	1668	1668	1668	1668	1668	1668	1668	1668	1668	1668
D52	1670	1670	1670	1670	1670	1670	1670	1670	1670	1670
D53	1670	1670	1670	1670	1670	1670	1670	1670	1670	1670
D54	1021	1021	1021	1021	1021	1021	1021	1021	1021	1021
D55	3865	3865	3865	3865	3865	3865	3865	3865	3865	3865
D56	1437	1437	1437	1437	1437	1437	1437	1437	1437	1437
D57	946	946	946	946	946	946	946	946	946	946
D58	936	936	936	936	936	936	936	936	936	936
D59	2986	2986	2986	2986	2986	2986	2986	2986	2986	2986
D60	685	685	685	685	685	685	685	685	685	685
D61	922	922	922	922	922	922	922	922	922	922
D62	15772	15772	15772	15772	15772	15772	15772	15772	15772	15772
D63	11447	11447	11447	11447	11447	11447	11447	11447	11447	11447
D64	10178	10178	10178	10178	10178	10178	10178	10178	10178	10178

ตารางที่6.3 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D65	4354	4354	4354	4354	4354	4354	4354	4354	4354	4354
D66	7342	7342	7342	7342	7342	7342	7342	7342	7342	7342
D67	684	684	684	684	684	684	684	684	684	684
D68	29205	29205	29205	29205	29205	29205	29205	29205	29205	29205
D69	17252	17252	17252	17252	17252	17252	17252	17252	17252	17252
D70	2213	2213	2213	2213	2213	2213	2213	2213	2213	2213
D71	2406	2406	2406	2406	2406	2406	2406	2406	2406	2406
D72	700	700	700	700	700	700	700	700	700	700
D73	732	732	732	732	732	732	732	732	732	732
D74	4968	4968	4968	4968	4968	4968	4968	4968	4968	4968
D75	5013	5013	5013	5013	5013	5013	5013	5013	5013	5013
D76	849	849	849	849	849	849	849	849	849	849
D77	9389	9389	9389	9389	9389	9389	9389	9389	9389	9389
D78	920	920	920	920	920	920	920	920	920	920
D79	1216	1216	1216	1216	1216	1216	1216	1216	1216	1216
D80	906	906	906	906	906	906	906	906	906	906
D81	1682	1682	1682	1682	1682	1682	1682	1682	1682.	1682
D82	5188	5188	5188	5188	5188	5188	5188	5188	5188	5188
D83	608	608	608	608	608	608	608	608	608	608
D84	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216
D85	738	738	738	738	738	738	738	738	738	738
D86	3015	3015	3015	3015	3015	3015	3015	3015	3015	3015
D87	6539	6539	6539	6539	6539	6539	6539	6539	6539	6539
D88	2598	2598	2598	2598	2598	2598	2598	2598	2598	2598
D89	2426	2426	2426	2426	2426	2426	2426	2426	2426	2426
D90	1057	1057	1057	1057	1057	1057	1057	1057	1057	1057
D91	4767	4767	4767	4767	4767	4767	4767	4767	4767	4767
D92	1219	1219	1219	1219	1219	1219	1219	1219	1219	1219
D93	1215	1215	1215	1215	1215	1215	1215	1215	1215	1215
D94	984	984	984	984	984	984	984	984	984	984
D95	3305	3305	3305	3305	3305	3305	3305	3305	3305	3305
D96	2806	2806	2806	2806	2806	2806	2806	2806	2806	2806

ตารางที่6.4 แสดงผลการทดลองค่าเอกสารทั้งหมด (D) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D1	502	502	502	502	502	502	502	502	502	502
D2	502	502	502	502	502	502	502	502	502	502
...	...	...	...	...	...	...	...	...	...	...
D502	502	502	502	502	502	502	502	502	502	502

ตารางที่6.5 แสดงผลการทดลองค่าเอกสารที่พบเทอมที่สนใจ (d) ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D1	118	114	290	161	55	58	160	120	157	60
D2	118	114	290	161	55	58	160	120	157	60
...	...	...	...	...	...	...	...	...	...	...
D4	118	114	290	161	55	58	160	120	157	60

ตารางที่6.6 แสดงผลการทดลองค่าน้ำหนัก TF.IDF ระหว่างเทอมที่สนใจกับเว็บไซต์ทั้งหมด

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D1	0.58499	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D2	0.05526	0.00000	0.00000	0.00000	0.00000	0.00000	0.04364	0.00000	0.00000	0.00000
D3	0.05526	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D4	0.03863	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D5	0.03486	0.00006	0.00053	0.00000	0.00000	0.00000	0.00000	0.00000	0.00022	0.00000
D6	0.66469	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D7	0.00004	0.00002	0.00040	0.00000	0.00000	0.00000	0.00000	0.00000	0.00008	0.00000
D8	0.57532	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.53407	0.00000	0.00000
D9	0.06206	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.05364	0.00004	0.00000
D10	0.13206	0.00000	0.00064	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D11	0.08770	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D12	0.03009	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D13	0.00000	0.00000	0.00078	0.00000	0.00000	0.00000	0.00346	0.00000	0.00000	0.00000
D14	0.00356	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00352	0.00000	0.00000
D15	0.04633	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D16	0.00148	0.00000	0.00000	0.00000	0.00000	0.00000	0.00017	0.00000	0.00000	0.00000
D17	0.01831	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

ตารางที่6.6 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D18	0.51414	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D19	0.00109	0.00000	0.02554	0.00000	0.00000	0.00000	0.00000	0.00000	0.00175	0.00000
D20	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D21	0.05927	0.00010	0.00084	0.00000	0.00000	0.00000	0.00015	0.00000	0.00030	0.00000
D22	0.00184	0.00000	0.00070	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D23	0.02030	0.00000	0.00000	0.00000	0.00000	0.00000	0.06412	0.00000	0.00000	0.00000
D24	0.00088	0.00000	0.00000	0.00000	0.00000	0.00000	0.00003	0.00000	0.00000	0.00000
D25	0.00159	0.00000	0.00270	0.00000	0.00000	0.00000	0.00125	0.00000	0.00000	0.00000
D26	0.70189	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D27	0.19250	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D28	0.00036	0.00000	0.00176	0.00000	0.00000	0.00000	0.00000	0.00000	0.00029	0.00000
D29	0.00079	0.00000	0.00156	0.00000	0.00000	0.00000	0.00000	0.00009	0.00000	0.00000
D30	0.00174	0.00000	0.00148	0.00000	0.00000	0.00000	0.00021	0.00034	0.00000	0.00000
D31	0.00636	0.00000	0.00000	0.00000	0.00000	0.00000	0.00084	0.00000	0.00000	0.00000
D32	0.07492	0.00000	0.00348	0.02642	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D33	0.11229	0.00000	0.00150	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D34	0.00329	0.00000	0.00280	0.00000	0.00000	0.00000	0.00000	0.00000	0.00198	0.00000
D35	0.00462	0.00000	0.00000	0.00000	0.00000	0.00000	0.00061	0.00000	0.00000	0.00000
D36	0.00910	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D37	0.04015	0.00000	0.00048	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D38	0.01503	0.00000	0.00023	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D39	0.01981	0.00042	0.00000	0.00000	0.00000	0.00000	0.00054	0.00041	0.00000	0.00000
D40	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D41	0.00328	0.00187	0.00129	0.00000	0.00000	0.00000	0.00070	0.00000	0.00000	0.00000
D42	0.03056	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D43	0.00787	0.00179	0.00024	0.00000	0.00000	0.00000	0.00000	0.00008	0.00166	0.00000
D44	0.07076	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D45	0.04760	0.00014	0.00053	0.00000	0.00000	0.00000	0.00006	0.00000	0.00000	0.00000
D46	0.00189	0.00026	0.00042	0.00000	0.00000	0.00000	0.00007	0.00111	0.00000	0.00126
D47	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D48	0.42936	0.00000	0.00423	0.00000	0.00000	0.00000	0.09100	0.00000	0.00000	0.00000
D49	0.49917	0.00000	0.00235	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D50	0.00247	0.00000	0.00012	0.00040	0.00000	0.00013	0.00000	0.00000	0.00009	0.00000

ตารางที่6.6 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D51	0.00000	0.00267	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D52	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D53	0.00000	0.00444	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D54	0.00000	0.67514	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D55	0.00749	0.26388	0.00156	0.00000	0.00000	0.00000	0.00000	0.00000	0.00120	0.00000
D56	0.01008	0.24449	0.00076	0.00000	0.00000	0.00000	0.00716	0.00100	0.00000	0.00000
D57	0.00000	0.49831	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D58	0.00000	0.47513	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D59	0.00388	0.17872	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D60	0.00845	0.33327	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D61	0.00000	0.36658	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D62	0.00009	0.00207	0.00017	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D63	0.00013	0.00220	0.00019	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D64	0.00000	0.00087	0.00011	0.00000	0.00000	0.00000	0.00022	0.00000	0.00000	0.00000
D65	0.00067	0.07184	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D66	0.00316	0.01696	0.00314	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D67	0.00000	0.27091	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D68	0.00005	0.02482	0.00053	0.00000	0.00000	0.00000	0.00000	0.00000	0.00016	0.00000
D69	0.00017	0.00344	0.00000	0.00000	0.00000	0.00000	0.00027	0.00008	0.00000	0.00000
D70	0.02028	0.11053	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D71	0.01504	0.02156	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D72	0.00827	0.09741	0.00000	0.00000	0.00000	0.00000	0.00000	0.00818	0.00000	0.00000
D73	0.00593	0.13366	0.00000	0.00000	0.00000	0.00000	0.00000	0.01173	0.00000	0.00000
D74	0.00058	0.00806	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D75	0.00000	0.00828	0.00142	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D76	0.00000	0.02619	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D77	0.00000	0.02416	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D78	0.00000	0.42861	0.00119	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D79	0.00000	0.24382	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D80	0.00000	0.42050	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D81	0.00000	0.02468	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D82	0.00000	0.11944	0.00000	0.00022	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D83	0.00000	0.17799	0.00090	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

ตารางที่6.6 (ต่อ)

	bank	finance	flash	game	coke	pepsi	thailand	bangkok	adobe	golf
D84	0.00000	0.00669	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00420	0.00000
D85	0.00000	0.24707	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00315	0.00000
D86	0.00000	0.13914	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D87	0.00022	0.00861	0.00210	0.00000	0.00000	0.00000	0.00087	0.00175	0.00000	0.00000
D88	0.00000	0.00799	0.00000	0.00000	0.00000	0.00000	0.00000	0.00110	0.00000	0.00000
D89	0.00239	0.02444	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D90	0.00411	0.00280	0.00052	0.00000	0.00000	0.00000	0.00433	0.00000	0.00000	0.00000
D91	0.00182	0.00187	0.00000	0.00000	0.00000	0.00000	0.00312	0.00030	0.00073	0.00000
D92	0.00950	0.04256	0.00315	0.00000	0.00000	0.00000	0.00000	0.00000	0.00191	0.00000
D93	0.00715	0.04026	0.00226	0.00000	0.00000	0.00000	0.00000	0.00000	0.00383	0.00000
D94	0.00000	0.00603	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00354	0.00000
D95	0.00000	0.00179	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00106	0.00000
D96	0.00000	0.08875	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00207	0.00000
D97	0.00637	0.00224	0.00028	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D98	0.00000	0.03050	0.00030	0.00062	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D99	0.00000	0.03060	0.00030	0.00062	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D100	0.00630	0.00967	0.00083	0.00057	0.00000	0.00000	0.00688	0.00168	0.00000	0.00000
D101	0.00000	0.00115	0.00319	0.00000	0.00000	0.00000	0.00000	0.00000	0.10845	0.00000
D102	0.00000	0.00000	0.01272	0.00028	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D103	0.00000	0.00000	0.02345	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D104	0.00000	0.00000	0.02063	0.00000	0.00000	0.00000	0.00000	0.00000	0.04370	0.00000
D105	0.00000	0.00000	0.02094	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D106	0.00000	0.00000	0.00661	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D107	0.00000	0.00000	0.04287	0.00000	0.00000	0.00000	0.00000	0.00000	0.07404	0.00000
D108	0.00000	0.00000	0.06294	0.00000	0.00000	0.00000	0.00000	0.00000	0.00131	0.00000
D109	0.00000	0.00123	0.00410	0.00000	0.00000	0.00000	0.00000	0.00000	0.39276	0.00000
D110	0.00000	0.00254	0.00469	0.00000	0.00000	0.00000	0.00000	0.00000	0.58466	0.00000
D111	0.00000	0.00000	0.00258	0.00637	0.00000	0.00000	0.00000	0.00000	0.00027	0.00000
D112	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
D113	0.00000	0.00120	0.01842	0.00000	0.00000	0.00000	0.00000	0.00000	0.39827	0.00000
D114	0.00000	0.00131	0.01573	0.00000	0.00000	0.00000	0.00000	0.00000	0.26496	0.00000
D115	0.00000	0.00000	0.00187	0.00038	0.00000	0.00000	0.00034	0.00000	0.00010	0.00000
D116	0.00000	0.00000	0.01544	0.00000	0.00000	0.00000	0.00027	0.00000	0.00000	0.00000

ตารางที่6.7 แสดงผลการทดลองค่ามุมความสัมพันธ์ (ø) ระหว่างเว็บไซต์แต่ละคู่

	D143	D144	D145	D146	D147	D148	D149	D150	D151	D152
D130	0.01593	0.00000	0.11066	0.10498	0.00000	0.14741	0.31516	0.14358	1.12005	1.12799
D131	0.01593	0.00000	0.11066	0.10498	0.00000	0.14741	0.31516	0.14358	1.12005	1.12799
D132	0.77063	0.77631	0.78252	0.78190	0.77631	0.78729	0.54942	0.66687	0.34374	0.35168
D133	1.53088	1.54572	1.43508	1.45415	1.54572	1.54599	1.44481	1.45879	1.55987	1.56005
D134	1.56500	1.57080	1.57080	1.57080	1.57080	1.57080	1.31914	1.45694	0.45075	0.44280
D135	0.74040	0.74607	0.75266	0.75201	0.74607	0.75773	0.52123	0.63694	0.37398	0.38192
D136	0.01593	0.00000	0.11066	0.10498	0.00000	0.14741	0.31516	0.14358	1.12005	1.12799
D137	1.53930	1.55414	1.44351	1.46139	1.55414	1.55432	1.45181	1.46716	1.56354	1.56366

ตาราง6.7 ที่นำมาแสดงนี้เป็นเพียงส่วนหนึ่งของตารางทั้งหมดเนื่องจากตาราง6.7 นี้มีขนาด 502*502 ซึ่งหากนำตารางมาแสดงในรายงานจะไม่สะดวกเป็นอย่างมากในการพิจารณาและใช้งาน ทั้งยังทำให้จำนวนหน้าของรายงานเพิ่มขึ้นโดยที่ไม่จำเป็น

เพื่อการพิจารณาผลลัพธ์การทดลองครั้งนี้มีประสิทธิภาพมากขึ้น ผู้อ่านสามารถดาวน์โหลดผลลัพธ์ในรูปไฟล์Excelได้ที่URL: *http://narongphan.thport.com/files/ReportExcel2007.zip* แต่เนื่องจากผลลัพธ์ค่ามุมความสัมพันธ์ในตารางที่ 6.7ในรูปแบบไฟล์ Excel นั้นอยู่ใน MS.Office เวอร์ชัน 2007 (.xlsx) เหตุเพราะตาราง 6.7 เป็นตารางขนาด 502*502 (เปรียบเทียบความสัมพันธ์ระหว่างเว็บทุกเว็บ) จึงไม่สามารถเก็บเป็นไฟล์ .xls สำหรับเวอร์ชันธรรมดาได้ (เวอร์ชันที่ไฟล์เป็น .xls สามารถเก็บคอลัมภ์ได้เพียง 256 คอลัมภ์เท่านั้น) จึงจำเป็นที่ผู้ใช้ต้องมี MS.Office เวอร์ชัน 2007 หรือเวอร์ชันที่สามารถเปิดไฟล์ .xlsx ได้เท่านั้น

หรือทำการดาวน์โหลดไฟล์รายงานผลลัพธ์ในรูปของไฟล์ XML ในการทดลองครั้งนี้ที่ URL: *http://narongphan.thport.com/files/XMLfullreport.zip* (ไฟล์ XML) เพื่อนำไฟล์นี้ไปเปิดดูตารางผลลัพธ์ทั้งหมดได้ในโปรแกรม Web Categorization ส่วนการทำงาน History อ่านวิธีการเปิดไฟล์ด้วยโปรแกรม Web Categorization ได้ที่หัวข้อ 5.2.3 ส่วนแสดงรายงาน (Report Part)

6.2 วิเคราะห์ผลลัพธ์การคำนวณจากโปรแกรม Web Categorization

จะเห็นว่าจากตารางที่6.7ที่แสดงค่ามุมความสัมพันธ์ (ø) ระหว่างเว็บไซต์แต่ละคู่ จะมีค่ามุม øอยู่ในช่วง 0 - 1.57080 เนื่องจากค่าดังกล่าวเป็นค่าเรเดียน (Radian) จากตารางเปรียบเทียบค่ามุมเป็นดีกรี(Degree) กับค่ามุมเป็นเรเดียน (Radian)

ตารางที่6.8 ตารางเปรียบเทียบค่ามุมเป็นดีกรี(Degree) กับค่ามุมเป็นเรเดียน (Radian)

Degrees	0°	30°	45°	60°	90°	180°	270°	360°
Radians	0	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$	π	$(3\pi)/2$	2π

จะเห็นว่าค่าดีกรี 90° จะได้ค่าเรเดียนเท่ากับ $\pi/2$ ซึ่ง π มีค่าเท่ากับ 22/7 หรือประมาณ 3.1416 ดังนั้นค่า $\pi/2$ หรือ 90° จึงมีค่าเรเดียนเท่ากับ 3.1416 / 2 เท่ากับ 1.57080 นั่นเอง

ซึ่งถ้าค่ามุมความสัมพันธ์ระหว่างเว็บ มีมุมเรเดียนเท่ากับ 1.57080 หรือมุมดีกรีเท่ากับ 90 ก็หมายถึง สองเว็บดังกล่าวไม่สัมพันธ์กันโดยสิ้นเชิง และถ้าค่ามุมเรเดียนเท่ากับ 0 หรือมุมดีกรีเท่ากับ 0 ก็จะหมายถึงว่า สองเว็บดังกล่าวสัมพันธ์กันนั่นเอง

จากการที่การทดลองในครั้งนี้โปรแกรมมีวิธีในการเก็บข้อมูลเว็บตัวอย่างด้วยการค้นหารายชื่อจากกูเกิลทำให้รายชื่อเว็บที่ 1-50, 51-100, 101-150, ... นั้นเป็นเว็บที่อยู่ในกลุ่มคำเดียวกันซึ่งตรงกับผลการทดลองในตารางที่ 6.7 เรื่องค่ามุมสัมพันธ์ระหว่างเว็บ จะเห็นได้ว่าช่วงการเปรียบเทียบเว็บที่ 1-50 นั้นค่ามุม ø ที่ได้ออกมาจะมีค่าเข้าใกล้ 0 ซึ่งหมายความว่ามีความสัมพันธ์กันนั่นเองและมีความสัมพันธ์กันในหัวเรื่องที่ชื่อว่า "Bank" สามารถรู้ได้โดยดูค่า TF.IDF ที่มากที่สุดของเว็บ 1-50 คือค่า TF.IDF ของเทอมที่ชื่อว่า Bank (ในตารางที่ 6.6 ดูช่วงเว็บที่ 1-10) และรอบๆในช่วงกลุ่มเว็บนั้นๆจะมีค่ามุม ø เข้าใกล้ค่า 1.57080 ซึ่งหมายความว่าอยู่คนละกลุ่มเว็บกัน แต่ก็มีอยู่บางเว็บที่มีความสัมพันธ์กันมากทั้งที่อยู่คนละกลุ่มกันเช่น เว็บที่ D152 (flashgame.tarad.com) ซึ่งอยู่ในกลุ่มของเว็บ Game และมีค่าน้ำหนัก TF.IDF ของคำว่า Game สูงถึง 0.90641 และยังมีค่าน้ำหนัก TF.IDF ของคำว่า Flash ถึง 0.42983 ด้วย เว็บที่ D152 นี้มีค่ามุมความสัมพันธ์กับเว็บ D132 (www.imdb.com) เท่ากับ 0.35168 ซึ่งถือว่ามีความสัมพันธ์ใกล้เคียงกันทั้งที่เว็บ D132 อยู่ในกลุ่มของ Flash เราจึงทำการตรวจสอบค่า TF.IDF ของเว็บ D132 พบว่า เว็บ D132 มีค่าน้ำหนัก TF.IDF ของคำว่า Game เท่ากับ 0.00218 และมีค่าน้ำหนัก TF.IDF ของคำว่า Flash เท่ากับ 0.00222 จะเห็นได้ว่าทั้งเว็บ D152 กับ D132 อยู่คนละกลุ่มแต่มีค่าน้ำหนัก TF.IDF ของเทอมที่สนใจเหมือนกันด้วยเหตุนี้จึงทำให้ค่ามุมความสัมพันธ์ของเว็บทั้งสองมีค่ามุม ø เรเดียนเข้าใกล้ 0 หรือเรียกได้ว่าอยู่ในเว็บกลุ่มเดียวกันประเภท Flash game นั่นเอง

การทดลองในครั้งนี้ยังมีค่าความสัมพันธ์อีกมากมายระหว่างเว็บแต่ละคู่ ซึ่งการอ่านค่าดังกล่าวทำให้สามารถทราบได้ว่าเว็บไหนเป็นเว็บประเภทอะไรมีคำอะไรเป็นคำสำคัญในเว็บนั้นๆ และยังสามารถทราบได้ว่าเว็บดังกล่าวมีความสัมพันธ์เชิงเนื้อหาลักษณะเหมือนกันกับเว็บอื่นๆเว็บใดบ้าง

บทที่ 7
สรุปโปรแกรมการคัดแยกประเภทเว็บ เพื่อทำการพัฒนาและปรับปรุง

7.1 ส่วนพัฒนาและปรับปรุงโปรแกรม

โปรแกรมการคัดแยกประเภทเว็บนี้ยังสามารถปรับปรุงส่วนต่างๆอีกได้แก่

- ส่วนการค้นหาเทอมที่สนใจแบบกรอกเทอมที่สนใจด้วยตนเอง
- ส่วนการรองรับการคัดแยกหน้าเว็บด้วยภาษาไทย
- ส่วนการใช้คำที่สนใจในการค้นหารายชื่อเว็บตัวอย่าง
- ส่วนการค้นหาคำหน้าและคำหลังของคำที่สนใจมาช่วยประกอบการพิจารณาความหมาย

7.1.1 การพัฒนาและปรับปรุงโปรแกรมส่วนการค้นหาเทอมที่สนใจแบบกรอกเทอมที่สนใจด้วยตนเอง

โปรแกรมควรมีการปรับปรุงพัฒนาเพิ่มเติมส่วนการค้นหาเทอมที่สนใจโดยที่ผู้ใช้สามารถกรอกคำที่ต้องค้นหาเองในส่วนการวิเคราะห์ข้อมูล เนื่องจากโปรแกรมในปัจจุบันผู้ใช้สามารถใส่คำที่ต้องการได้เพียงในส่วนการค้นหาข้อมูลเท่านั้น แต่ในส่วนการวิเคราะห์ข้อมูลผู้ใช้ไม่สามารถใส่เทอมที่ต้องการค้นหาเองได้โปรแกรมจะนำคำเดิมที่ผู้ใช้ใส่ในส่วนการค้นหาข้อมูลมาทำการวิเคราะห์อีกทีซึ่งทำให้การทดลองไม่มีความยืดหยุ่นและผู้ใช้ไม่สามารถทดลองเทอมที่สนใจใหม่ๆ ในกลุ่มเว็บตัวอย่างเดิมได้ การปรับปรุงโปรแกรมส่วนติดต่อกับผู้ใช้ (User Interface) ทำได้โดยเพิ่มช่องกรอกเทอมที่ต้องการทำการวิเคราะห์เท่ากับจำนวนโฟลเดอร์กลุ่มเว็บตัวอย่างที่ได้เลือกเข้ามาทำการวิเคราะห์ เช่นหากผู้ใช้เลือก 3 โฟลเดอร์เว็บข้อมูลตัวอย่างหลังทำการดึงข้อมูลเข้ามาแสดงในส่วนการวิเคราะห์จะมีช่องสำหรับกรอกเทอมที่สนใจและต้องการค้นหาแสดงขึ้นมา 3 ช่องโดยอัตโนมัติ

7.1.2 การพัฒนาและปรับปรุงโปรแกรมส่วนการรองรับการคัดแยกหน้าเว็บด้วยภาษาไทย

สามารถพัฒนาโปรแกรมให้สามารถรองรับการคัดแยกหน้าเว็บที่เป็นภาษาไทย แต่ว่าการแยกคำในภาษาไทยนั้นทำได้ยากกว่าภาษาอังกฤษเนื่องจากภาษาไทยไม่มีช่องว่าง (space) ระหว่างคำดังเช่นภาษาอังกฤษจึงต้องใช้วิธีคิดในการแยกคำภาษาไทยเช่น วิธีการแบบ N-Grams โดยส่วนใหญ่มักนิยมใช้ 2-Grams, 3-Grams หรือ 4-Grams ในการคัดแยก ยกตัวอย่างเช่นการใช้วิธีการแบบ 3-Grams ในการคัดแยกคำว่า "คนไทย" จะเริ่มจัดกลุ่มตัวอักษรทีละ 3 ตัวอักษรนำไปเปรียบเทียบ

กับพจนานุกรมได้ดังนี้ “คนไ”, “นไท”, “ไทย” หลังจากทำการเปรียบเทียบเสร็จสิ้นโปรแกรมจะสามารถแยกคำว่า “ไทย” ออกมาได้เป็นต้น

7.1.3 การพัฒนาและปรับปรุงโปรแกรมส่วนการใช้คำที่สนใจในการค้นหารายชื่อเว็บตัวอย่าง

จากที่ได้ทำการทดลองทำให้ได้ทราบว่าโปรแกรมนี้จะนำคำที่สนใจที่ผู้ใช้ได้ป้อนลงบนกูเกิ้ลเบราเซอร์เพื่อค้นรายชื่อเว็บตัวอย่างโปรแกรมนี้ จะทำการจัดเก็บคำที่ผู้ใช้สนใจค้นหาและป้อนเข้าไปนี้ไว้ เพื่อไปใช้เป็นเทอมที่สนใจต่อไป

ทำให้ถ้าหากผู้ใช้โยนคำที่ไม่ใช่คำรากศัพท์ของคำนั้นๆ (Stem) เช่น Transportation ซึ่งมีคำรากศัพท์เป็นคำว่า Transport เป็นต้น จะเกิดข้อผิดพลาดเนื่องจากโปรแกรมจะทำการนับจำนวนที่พบเทอมที่สนใจ “Transportation” เท่านั้น ไม่รวมคำว่า “Transport” เข้าไปด้วยและค่าเทอมที่สนใจที่พบในเอกสาร (t) ก็จะคลาดเคลื่อนได้

วิธีการแก้และพัฒนาปรับปรุงโปรแกรมสามารถทำได้โดยเปรียบเทียบคำใน dictionary เพื่อหาว่าเทอมที่สนใจอยู่ในกลุ่มของคำรากศัพท์ใดและนำคำในกลุ่มนั้นทั้งหมดเข้าไปหาความถี่ของเทอมในเอกสารนั้นๆ

7.1.4 การพัฒนาและปรับปรุงโปรแกรมส่วนการค้นหาคำหน้าและคำหลังของคำที่สนใจมาช่วยประกอบการพิจารณาความหมาย

สามารถพัฒนาโปรแกรมเพิ่มเติมได้โดยนำคำหน้าและคำหลังของคำที่สนใจมาช่วยประกอบการพิจารณาความหมายเช่น โปรแกรมพิจารณาเทอมที่สนใจคำว่า cat และมีเว็บตัวอย่าง 2 เว็บที่มีประโยคดังต่อไปนี้ “This is the best cute cat race.” และ “Welcome to CAT Telecom.” เป็นต้น โปรแกรมจะทำการตัดคำหน้าและหลังของคำที่สนใจในเว็บที่1 และ 2 ได้ว่า

- เว็บที่1 คำหน้าเท่ากับ “cute cat” และ คำหลัง “cat race”
- เว็บที่2 คำหน้าเท่ากับ “to CAT” และ คำหลัง “CAT Telecom”

โปรแกรมอาจมีการเก็บความรู้ (Knowledge) ไว้ก่อนแล้วว่าหากพบคำหลังของ cat เป็นคำว่า telecom คำว่า “cat“ นั้นจะไม่ได้หมายถึงแมว แต่หมายถึงการสื่อสารแห่งประเทศไทย (กสท.) เป็นต้น ซึ่งการนำคำหน้าและคำหลังของคำที่สนใจมาช่วยประกอบการพิจารณาความหมายจะช่วยให้โปรแกรมสามารถคัดแยกประเภทของหน้าเว็บได้มีประสิทธิภาพมากยิ่งขึ้น

บรรณานุกรม

Garcia, E. 2006. **Cosine Similarity and Term Weight Tutorial.** [Online]. Available : http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html

Garcia, E. 2006. **Term Vector Theory and Keyword Weights.** [Online]. Available : http://www.miislita.com/term-vector/term-vector-1.html

Illinois Institute of Technology. 2006. **Term Vector Theory and Keyword Weights.** [Online]. Available : http://www.ir.iit.edu/~dagr/cs529/files/handouts/03VectorSpaceImplementation-6per.pdf

Mathematic of University Degli Studi Di Padova. 2006. **Automatic Web Page Categorization by Link and Context Analysis.** [Online]. Available : http://www.math.unipd.it/fabseb/Publication/THAI99.pdf

Salton, G. and Buckley, C. 1988. **Term-weighting approaches in automatic text retrieval.** Newyork : Cornell University.

Stanford.edu. 2008. **Tf-idf weighting.** [Online]. Available : http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

Wikipedia. 2007. **Vector space model.** [Online]. Available : http://en.wikipedia.org/wiki/Vector_space_model

ประวัติผู้เขียน

ชื่อผู้เขียน	นายณรงค์พันธ์ ปาการเสรี
วันเกิด	16 ตุลาคม 2521
สถานที่เกิด	รพ.เซนต์หลุยส์
วุฒิการศึกษาระดับปริญญาตรี	วท.บ. (คณิตศาสตร์ประยุกต์) คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2543
การทำงานปัจจุบัน	เว็บดีเวลลอปเปอร์และเว็บมาร์เก็ตติ้ง บริษัท อินเตอร์เนชันแนลโนเลจเน็ตเวิร์ค จำกัด