การออกแบบและพัฒนาระบบค้นหาเว็บไซต์แบบเจาะจง

FOCUS WEB SEARCH: DESIGN AND IMPLEMENTATION

โดย

วงศกร ตั้งทรงจิตร

WONGSAKORN THUNGSONGJIT

อาจารย์ที่ปรึกษา

รศ. ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาการศึกษาอิสระ
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2552

FOCUS WEB SEARCH: DESIGN AND IMPLEMENTATION

WONGSAKORN THUNGSONGJIT

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS OF THE COURSE

INDEPENDENT STUDY

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2 / 2009

ใบรับรองการศึกษาอิสระ (Independent Study)

เรื่อง

การออกแบบและพัฒนาระบบค้นหาเว็บแบบเจาะจง

Focus Web Search: Design and Implementation

นายวงศกร ตั้งทรงจิตร

รหัสประจำตัว 51066405

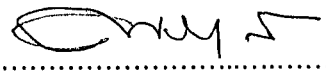
ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ
การศึกษาวิชาการศึกษาอิสระ หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)
ภาคเรียนที่ 2 ปีการศึกษา 2552

.....**อาจารย์ที่ปรึกษา**

(รศ.ดร. วรพจน์ กรีสุระเดช)

.....**กรรมการสอบ**

(รศ.ดร. อาริต ธรรมโน)



.....**กรรมการสอบ**

(ผศ.ดร. พรฤดี เนติโสภากุล)

หัวข้อ	การออกแบบและพัฒนาระบบค้นหาเว็บแบบเจาะจง
นักศึกษา	นายวงศกร ตั้งทรงจิตร
รหัสนักศึกษา	51066405
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2552
อาจารย์ที่ปรึกษา	รศ. ดร. วรพจน์ กรีสุระเดช

บทคัดย่อ

ในปัจจุบันนี้ เว็บไซต์กลายเป็นแหล่งข้อมูลที่แรกๆ ที่พวกเราทุกคนมักจะมองหากัน เนื่องจาก ในเว็บไซต์มีข้อมูลที่หลากหลาย ทั้งตัวอักษร ภาพ และเสียง การค้นหาเว็บไซต์ที่เราต้องการเราจำเป็นที่จะต้องใช้เครื่องมือที่เรียกว่า Search Engine ซึ่งก็มีอยู่หลายตัวให้เลือกใช้กันในปัจจุบัน แต่ว่าผลการค้นหาของ Search Engine นั้น ให้ผลลัพธ์ที่ตรงความต้องการของผู้ใช้กลับมาเพียงน้อยนิดเมื่อเทียบกับผลลัพธ์ทั้งหมดที่ค้นคืนได้กลับมา ด้วยเหตุนี้เราจึงมีแนวคิดที่จะพัฒนาระบบที่สามารถค้นหาเว็บไซต์ที่ตรงความต้องการยิ่งขึ้น โดยอาศัยข้อมูลผลลัพธ์จาก Search Engine มาใช้ประโยชน์ต่อยอด เพื่อให้ได้เว็บไซต์ที่ตรงความต้องการผู้ใช้มากขึ้น ซึ่งหวังว่าผลลัพธ์ที่ได้เพิ่มขึ้นมานี้จะเป็นประโยชน์ต่อผู้ใช้มากขึ้น

Title	Focus Web Search: Design and Implementation
Student	Mr. Wongsakorn Thungsongjit
Student ID.	51066405
Degree	Master of Science
Program	Information Technology
Major	Information Technology Management
Academic Year	2009
Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej

ABSTRACT

Today, website is a first source of information that we seek for. Website has many type of information such as text, audio and video. If we want to find information from web site, we need tool that called Search Engine. Today, there are many Search Engine is available on internet, but the recall information of Search Engine is too many if we compared it with information that user want. For this reason, we have ideas to develop system that have ability to find websites that meet the user needs better. Based on result from Search Engine, we hope result from our system that enhances power of Search Engine will be benefit to user who uses system.

ใบรับรองโครงการพัฒนาระบบงาน (System Development Project)

หรือ โครงการศึกษากรณีพิเศษ (Special Study Project)

หรือ การศึกษาอิสระ (Independent Study)

เรื่อง

การออกแบบและพัฒนาระบบค้นหาเว็บแบบเจาะจง

Focus Web Search: Design and Implementation

นายวงศกร ตั้งทรงจิตร

รหัสประจำตัว 51066405

ขอรับรองว่ารายงานฉบับนี้ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ
การศึกษาวิชาการศึกษาอิสระ หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)
ภาคเรียนที่ 2 ปีการศึกษา 2552

.....อาจารย์ที่ปรึกษา

(รศ.ดร.วรพจน์ กรีสุระเดช)

.....กรรมการสอบ

(รศ.ดร.อาริต ธรรมโน)

.....กรรมการสอบ

(ผศ.ดร.พรฤดี เนติโสภากุล)

กิตติกรรมประกาศ

โครงการพัฒนาระบบงานนี้ ไม่มีทางสำเร็จออกมาสมบูรณ์หากปราศจากความร่วมมือของบุคคลดังต่อไปนี้

พ่อกับแม่ที่คอยเป็นกำลังใจและคอยช่วยเหลือทุกปัญหาตลอดเวลา หากโครงการนี้เป็นประโยชน์ต่อใครในอนาคตขอให้ผลบุญอันนี้จงบัลดาลให้แม่หายหรือบรรเทาจากอาการป่วยที่เป็นอยู่

รศ. ดร. วรพจน์ กรีสุระเดช ที่สละเวลาให้คำปรึกษาทั้งเรื่องที่เกี่ยวกับโครงการนี้และไม่เกี่ยว ประสบการณ์ แนวคิด ที่อาจารย์แบ่งปันและเสนอแนะให้มันเป็นประโยชน์ทั้งทางตรงและทางอ้อม ทั้งในปัจจุบันและในอนาคต

เพื่อนๆ ทั้งหลายที่อุตสาห์พาไปเที่ยว ดูหนัง กินของอร่อยๆ กันในเวลาที่เราเซ็งๆ

วงศกร ตั้งทรงจิตร

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญตาราง

สารบัญตาราง

สารบัญรูป

รูปที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

มนุษย์เราในปัจจุบันนี้เริ่มตระหนักแล้วว่าข้อมูลนั้นมีความสำคัญต่อชีวิตของตัวเอง เพราะว่าข้อมูลเหล่านั้นส่งผลกระทบต่อชีวิตของมนุษย์ทุกคน แล้วในปัจจุบันแหล่งข้อมูลที่มนุษย์ทุกคนมองหากันเป็นอันดับแรกก็คือเว็บไซต์บนอินเตอร์เน็ต ไม่ว่าจะเป็นบทความ ภาพ เสียง หรือไฟล์วีดีโอ ข้อมูลเหล่านี้พร้อมให้บริการอยู่ในเว็บไซต์เหล่านั้นแทบตลอดเวลา

เมื่อนานวันเข้าเว็บไซต์เริ่มเติบโตและทวีจำนวนขึ้นเรื่อยๆ จนต้องมีการทำสารบัญเว็บเพื่อการค้นหาเว็บที่มีเนื้อหาตรงกับความต้องการของผู้ใช้ โดยการใช้ความรู้ทางด้าน Information Retrieval และพัฒนาจนกลายเป็น Search Engine ซึ่งสามารถช่วยเราหาเว็บไซต์ที่ตรงกับความต้องการของเราได้อย่างน่าประทับใจ

แต่ว่าการค้นหาโดยใช้ Search Engine นั้น ใช้การใส่ คีย์เวิร์ด แล้วให้ Search Engine นำไปค้นหา ซึ่งคีย์เวิร์ดดังกล่าวนี้เมื่อนำไปประมวลผล ผลลัพธ์ที่กลับมาจะมีปริมาณมากมาย แต่ผลลัพธ์ที่ตรงกับความต้องการของผู้ค้นหากลับมีอยู่เพียงน้อยนิดเท่านั้น

สาเหตุที่สำคัญที่ทำให้ผลลัพธ์ที่ได้จากการค้นหานำไปใช้งานได้จริงมีน้อยนั่นก็เพราะคีย์เวิร์ดที่ใส่เข้ามามักมีความกำกวมเสมอ เช่น คำว่า "กา" ซึ่งอาจจะหมายถึง "นกกา" หรือ "กาน้ำ" ก็เป็นไปได้ทั้งคู่ ซึ่งความกำกวมนี้เป็นธรรมชาติของภาษามนุษย์ที่เกิดขึ้นได้ ซึ่งคนส่วนใหญ่แก้ปัญหาด้วยการเพิ่ม หรือเปลี่ยนคีย์เวิร์ดที่ใช้ค้นหา จนได้ผลลัพธ์ที่ต้องการ แต่ก็ยังได้ผลลัพธ์ที่ไม่เกี่ยวข้องกลับมามากมายอยู่ดี

จากปัญหาดังกล่าว เราจึงต้องการที่จะศึกษาและพัฒนาระบบที่สามารถค้นหาเว็บไซต์ที่ตรงความต้องการของผู้ใช้เพิ่มขึ้นจากผลลัพธ์เดิมที่ได้จากการค้นหาของ Search Engine เพื่อให้ผู้ใช้นั้นได้รับข้อมูลมากขึ้น มีทางเลือกมากขึ้น ซึ่งสามารถที่จะเปรียบเทียบข้อมูลที่ได้จากหลายๆ เว็บไซต์ที่มีข้อมูลคล้ายกัน และนำไปใช้ประโยชน์ต่อยอดในด้านอื่นๆ ได้

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

จุดประสงค์ที่เราพยายามที่จะศึกษาและพัฒนาระบบงานนี้ออกมาใช้งานจริงก็เพื่อที่จะเพิ่มความสามารถ และนำผลลัพธ์จากการทำงานของ Search Engine มาใช้ประโยชน์ต่อยอด ซึ่งสามารถแจกแจงได้ดังนี้

1. หาแนวทางที่จะนำผลลัพธ์จากการทำงานของ Search Engine มาใช้ประโยชน์ต่อยอดในด้านต่างๆ
2. เพิ่มเติมความสามารถของ Search Engine ให้สามารถค้นหาเว็บไซต์ที่ผู้ใช้ต้องการได้มากขึ้น
3. นำความรู้ทางด้าน Data Mining และ Web Crawler มาประยุกต์ใช้งานให้เกิดประโยชน์

1.3 ทฤษฎีหรือแนวคิดที่ใช้ในการศึกษา

แนวคิดที่เราใช้ในการเริ่มต้นศึกษาคือการทำ Web Mining ในแบบ Web Classification ซึ่งเป็นกระบวนการที่พยายามจะจัดประเภทเว็บไซต์ให้เป็นกลุ่มๆ ตามที่เรากำหนด ซึ่งเราสามารถกำหนดลักษณะของกลุ่มที่เราต้องการได้ โดยมีการใช้ความรู้ในด้าน Data Mining ในเรื่อง Predictive Modeling เพื่อช่วยให้ระบบสามารถคำนวณหาความคล้ายคลึงของเว็บไซต์และจัดกลุ่มของเว็บไซต์เหล่านั้นได้

และบทความของ Mr.Soumen Chakrabarti เรื่อง Focused Crawler ซึ่งกล่าวถึง Crawler ที่สามารถประเมินได้ว่าเว็บไซต์ใดที่มีความน่าจะเป็นที่จะเป็นเว็บที่ตรงกับประเภทเว็บที่ระบบกำหนดไว้ ซึ่งจะทำให้ Crawler ประเภทนี้ไม่ต้องดึงข้อมูลของเว็บไซต์ทุกเว็บไซต์ที่มันมี URL ลดการดึงเว็บที่ไม่เกี่ยวข้องลง และทำให้ระบบสามารถนำทรัพยากรที่เหลือไปทำงานอย่างอื่นได้

1.4 ขอบเขตของการศึกษา

ระบบที่พัฒนาขึ้นนี้มีวัตถุประสงค์ในการนำเอาผลลัพธ์ของ Search Engine ไปพัฒนาต่อเพื่อให้ผู้ใช้ได้รับข้อมูลที่เป็นประโยชน์และตรงความต้องการมากขึ้น ซึ่งแบ่งได้เป็น 2 ส่วนใหญ่ๆ ดังจะมีขอบเขตของโครงการดังนี้

1. ระบบนำเสนอและรับข้อมูลจากผู้ใช้
    - ค้นหาเว็บไซต์ที่ต้องการจาก Search Engine
    - เลือกเว็บไซต์ที่ต้องการจากการค้นหาของ Search Engine
    - แสดงเว็บไซต์ที่มีความคล้ายคลึงกับเว็บไซต์ที่ผู้ใช้เลือก

2. ระบบค้นหาเว็บไซต์ที่คล้ายและเกี่ยวข้องกัน
    - Topic-Specific Web Crawler
    - เตรียมข้อมูลเว็บไซต์
    - หาตัวแทนเว็บไซต์
    - TF-IDF
    - K-Nearest Neighbors Algorithm

ข้อจำกัดของระบบมีดังต่อไปนี้

- Crawler ของระบบไม่สามารถกำหนดการเก็บข้อมูลเว็บไซต์แบบจำกัดประเภทของภาษาได้
- การเปรียบเทียบเว็บไซต์สามารถทำงานได้กับภาษาอังกฤษเท่านั้น ไม่รองรับการทำงานกับภาษาอื่น รวมทั้งภาษาไทย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ประโยชน์ต่อผู้พัฒนาระบบ
    - เป็นการศึกษาและพัฒนาความสามารถในการนำ Search Engine ไปใช้ประโยชน์เพิ่มเติม
    - เป็นการนำความรู้ในด้าน Data Mining มาทำงานและใช้ประโยชน์จริง
    - เป็นการฝึกฝนการออกแบบและพัฒนาซอฟแวร์แบบเว็บแอพพลิเคชั่น รวมถึงการพัฒนา Web Crawler เพื่อประยุกต์ใช้งานจริง
2. ประโยชน์ต่อผู้ใช้ระบบ
    - สามารถที่ค้นหาเว็บไซต์ที่คล้ายคลึงกับเว็บไซต์ที่ต้องการได้มากขึ้นจากผลลัพธ์เดิมของ Search Engine
    - สามารถนำข้อมูลที่ได้จากหลายๆ เว็บไซต์มาเปรียบเทียบ วิเคราะห์ และสรุปผลด้วยตนเองเพื่อนำไปใช้ ประโยชน์ต่อยอดในด้านต่างๆ ได้

1.6 ขั้นตอนของการศึกษา

ขั้นตอนในการออกแบบและพัฒนาโครงการประกอบด้วยขั้นตอนดังนี้

1. ศึกษาข้อมูลเบื้องต้นเกี่ยวกับการทำงานของ Search Engine ข้อมูลเกี่ยวกับ Web Crawler และข้อมูลใน เรื่อง Web Mining ในแบบ Web Classification

2. รวบรวมข้อมูลดังกล่าวข้างต้นมาวิเคราะห์และออกแบบระบบเพื่อแก้ปัญหาที่กล่าวมาข้างต้น
3. ออกแบบ UML Diagram ต่างๆ ที่จำเป็นต่อการพัฒนาระบบ
4. ออกแบบการเก็บข้อมูลใน Session
5. ศึกษาการเขียนโปรแกรมภาษา PHP ในส่วนที่ต้องใช้ และเทคโนโลยีที่เกี่ยวข้อง
6. พัฒนาระบบขึ้นตามที่ได้ออกแบบและศึกษาข้อมูลที่เกี่ยวข้อง
7. ทดสอบการทำงานของระบบ ตรวจหาข้อผิดพลาดและแก้ไขปรับปรุงให้เรียบร้อย
8. สรุปผลจากการพัฒนาระบบและทดสอบระบบ
9. จัดทำเอกสารที่เกี่ยวข้องกับการพัฒนาโครงการ

บทที่ 2

เทคนิคที่จำเป็นต่อการพัฒนาโครงการ

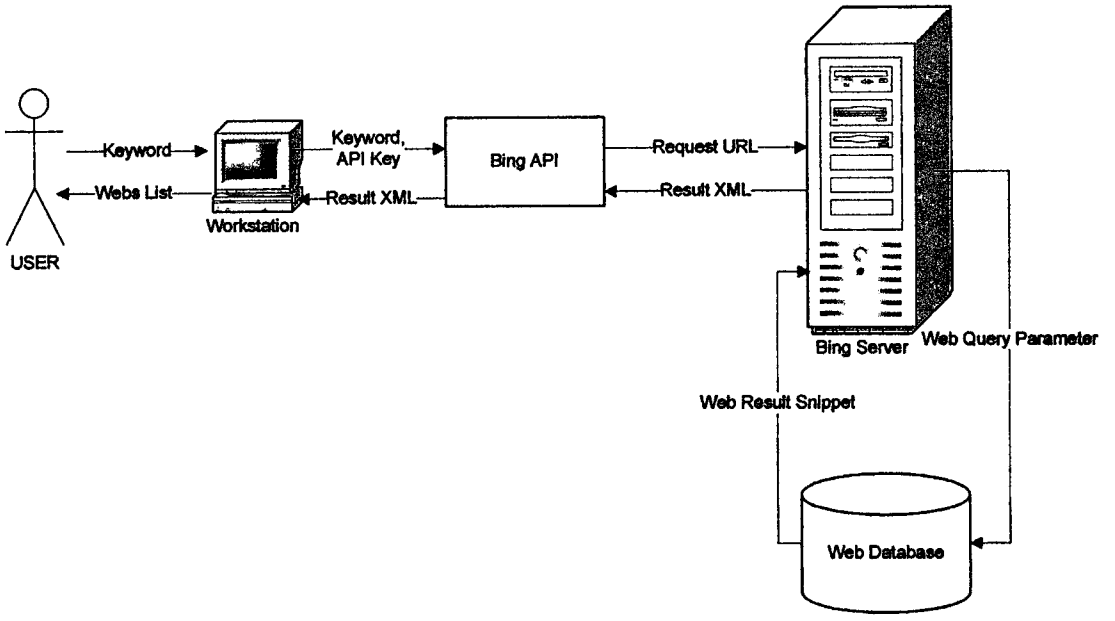
ในบทนี้จะกล่าวถึงเทคนิคต่างๆ ที่นำมาใช้พัฒนาโครงการจนเป็นรูปเป็นร่างและประสบผลสำเร็จลงได้ ซึ่งรายละเอียดของแต่ละเทคนิคมีดังต่อไปนี้

1. Bing API
2. Web Crawler : Topic-Specific Web Crawler
3. Web Document Representation
4. TF-IDF
5. K-Nearest Neighbor Algorithm

2.1 Bing API

ในการที่ระบบของเราจะทำงานต่อยอดจาก Search Engine ได้นั้น เราจำเป็นต้องได้รับข้อมูลการค้นหาจาก Search Engine ซึ่ง ตัว Search Engine ที่เราเลือกมาทำงานร่วมกันกับโครงการของเราคือ Bing เป็น Search Engine จาก Microsoft ซึ่ง Search Engine ตัวนี้เปิดบริการ API ให้เราใช้งานในลักษณะของ Web Service ฟรี

โดยจุดประสงค์ของ Microsoft ที่เปิดบริการให้ใช้บริการ API นี้ฟรีก็เพื่อให้นักพัฒนาและผู้ที่เกี่ยวข้องอื่นๆ นำการค้นหาของ Search Engine ไปใช้ประโยชน์ต่อยอด ซึ่งตรงกับความต้องการของเราพอดี โดย API ตัวนี้จะมีลักษณะเป็นการส่ง Request ไปยัง URL หนึ่งๆ แล้วทาง Server ก็จะตอบกลับมาเป็น XML ซึ่งเป็นรูปแบบที่เราสามารถจัดการได้ง่าย โดยแต่ละการร้องขอนั้นเราจำเป็นต้องบรรจุ API Key ลงไปใน Request URL ไปด้วย



รูปที่ 2.1 ขั้นตอนการทำงานของ Bing API

นอกจากการให้บริการค้นหาเว็บแล้ว Bing API ยังมีบริการอื่นๆ ให้เราเลือกใช้ในทางด้านการศึกษาและทางด้านธุรกิจให้เราอีก เช่น Bing MAP API, Phone Book Search, Image Search ซึ่งส่วนที่เรานำมาใช้มีเพียงแค่ Web Search เท่านั้น โปรโตคอลที่รองรับในการติดต่อนั้นได้แก่ JSON XML และ SOAP ในโครงการนี้เราเลือก XML มาใช้งาน

นอกจาก Bing API แล้วยังมีตัวเลือกอื่นให้ผู้ที่สนใจงานเกี่ยวกับการนำผลลัพธ์จาก Search Engine ไปใช้งานต่อยอดอีก เช่น Yahoo BOSS

2.2 Web Crawler และ Topic-Specific Web Crawler

Web Crawler เป็นคอมพิวเตอร์โปรแกรมแบบหนึ่งที่สามารถเยี่ยมชมเว็บไซต์และเก็บข้อมูลเว็บไซต์ต่างๆ ได้อัตโนมัติ ซึ่งอาจจะรู้จักกันในชื่ออื่นเช่น Web Spiders, Web Robot, ants ซึ่ง Web Crawler นี้เป็นโปรแกรมที่ Search Engine ต่างๆ ใช้งานในการเยี่ยมชมเว็บไซต์ปริมาณมหาศาลแล้วเก็บข้อมูลเหล่านั้นมาเพื่อประมวลผลและจัดเรียงตามแต่ละวิธีของ Search Engine แต่ละตัว ซึ่ง Web Crawler ที่มีประสิทธิภาพที่ดีนั้นจะต้องเก็บข้อมูลเว็บได้ปริมาณมากในเวลาที่น้อยที่สุด แต่ว่าแม้จะเป็น Web Crawler ที่ทำงานอยู่บนคอมพิวเตอร์ที่มีประสิทธิภาพสูงเพียงใดก็ตามก็ไม่สามารถที่จะเก็บข้อมูลเว็บไซต์ได้อย่างสมบูรณ์เพราะว่า

- เว็บไซต์ที่มีอยู่ในปัจจุบันที่เชื่อมโยงกันนั้นมีปริมาณมากมายมหาศาล
- เว็บไซต์ต่างๆ เหล่านั้นมีการเปลี่ยนแปลงเนื้อหาข้อมูลตลอดเวลา

ดังนั้นการให้ Web Crawler เยี่ยมชมเว็บไซต์ทุกเว็บไซต์เพื่อเก็บข้อมูลนั้นเป็นเรื่องที่แทบจะเป็นไปไม่ได้เลย แต่ถึงจะเก็บมาได้ข้อมูลบางเว็บไซต์นั้นก็ได้เปลี่ยนแปลงเพิ่มเติมไปแล้วทำให้ข้อมูลที่ Web Crawler เก็บมานั้นไม่เป็นข้อมูลที่ล่าสุดแล้ว ด้วยปัญหาเหล่านี้จึงได้มีการริเริ่มแนวคิดเกี่ยวกับ Web Crawler ประเภทใหม่ที่เรียกว่า Topic-Specific Web Crawler ขึ้นมา

Topic-Specific Web Crawler ต่างจาก Web Crawler แบบปกติตรงที่ Crawler ประเภทนี้จะไม่เยี่ยมชมเว็บไซต์ทุกเว็บไซต์แต่ว่ามันจะเลือกเยี่ยมชมเว็บไซต์ที่มีความเป็นไปได้ที่เว็บไซต์นั้นมีข้อมูลที่ Crawler นั้นถูกกำหนดมาให้มันสนใจ โดยความค่าความเป็นไปได้ที่เว็บไซต์ดังกล่าวจะมีข้อมูลที่มันสนใจเชื่อมต่อไปนั้นจำเป็นต้องใช้ความรู้ด้าน Data Mining มาช่วยในการทำนายและตัดสินใจ ด้วยการที่ Topic-Specific Web Crawler อาศัยความรู้และเทคนิคจากศาสตร์ด้านอื่นมาช่วยทำงาน จึงทำให้ Crawler ประเภทนี้สามารถลดปริมาณเว็บไซต์ที่มันต้องไปเก็บข้อมูลได้มากและข้อมูลเว็บไซต์ที่มันไปเก็บมาได้ก็ยังเกี่ยวข้องกับหัวข้อที่มันสนใจและเนื่องจากเว็บไซต์ที่เก็บมามีปริมาณลดลงทำให้มันสามารถเข้าไปเก็บข้อมูลใหม่ให้เป็นข้อมูลล่าสุดได้ในเวลารวดเร็วกว่า Web Crawler แบบปกติมาก

2.3 Web Document Preparation

ในการเตรียมข้อมูลเอกสารเว็บนั้นคล้ายกับการเตรียมข้อมูลเอกสารปกติ แต่ว่ามีบางส่วนที่เพิ่มเติมขึ้นมาเพื่อให้เอกสารเว็บนั้นสามารถถูกจัดการและหาตัวแทนเหมือนเอกสารปกติได้

1. ระบุข้อมูลที่อยู่ใน Tag HTML ตัวต่างๆ เพราะว่าข้อมูลที่อยู่ใน Tag HTML ต่างๆ นั้นสามารถสื่อถึงความหมายของเอกสารได้ไม่เท่ากัน ส่วนใหญ่ข้อมูลที่อยู่ใน Tag TITLE มักจะสื่อความหมายของเอกสารได้ดี (ไม่ได้นำมาใช้ในโครงการนี้)
2. ระบุลิงค์เชื่อมโยงด้วยแนวความคิดที่ว่าเว็บไซต์มักจะมีการเชื่อมโยงไปยังเว็บไซต์ที่มีเนื้อหาใกล้เคียงและเกี่ยวข้องกัน เราจึงต้องเก็บลิงค์เชื่อมโยงต่างๆ ของเว็บไซต์นั้นไว้ด้วย
3. เมื่อเก็บข้อมูลข้างต้นหมดแล้วให้นำ Tag HTML ออกให้หมด เพราะมันไม่มีประโยชน์ในการใช้งานในขั้นตอนต่อๆ ไปแล้ว

นอกจากการเตรียมข้อมูลเว็บดังที่กล่าวมาข้างต้นแล้วยังมีขั้นตอนเพิ่มเติมเพื่อที่จะทำให้ข้อมูลที่ได้จากเอกสารเว็บนั้นมีประสิทธิสูงขึ้นได้อีกคือ

- การระบุส่วนเนื้อหาของเว็บไซต์ เพราะเว็บไซต์นั้นมีส่วนที่เป็นเนื้อหาอยู่ส่วนใหญ่เราจะใช้งานแต่ส่วนนี้ ส่วนอื่นๆที่เป็นเมนู Header และ Footer นั้นเรามักไม่ได้ใช้ การนำส่วนที่ไม่ได้ใช้งานออกไปทำให้เอกสารนั้นสื่อความหมายมากขึ้นและลดการทำงานที่ไม่จำเป็นของโปรแกรมลง (ไม่ได้นำมาใช้งานในโครงการนี้)
- การตรวจสอบหน้าที่ซ้ำกันของเว็บไซต์ ปกติเว็บไซต์นั้นมีโอกาสที่จะซ้ำกันได้อยู่แล้วทั้งโดยเจตนาดีเช่น การทำ Mirror Page เพื่อกระจาย Bandwidth ไปหลายๆ ที่ และเจตนาร้ายเช่นการขโมยเนื้อหาไปใช้ในเว็บไซต์อื่น การระบุเนื้อหาหรือหน้าที่ซ้ำกันของเว็บไซต์จะทำให้เราสามารถลดตัวแทนของเว็บที่เกินออกมาได้ เพราะถ้าเราไม่หาหน้าเว็บไซต์ที่ซ้ำกันจะทำให้คำที่เป็นตัวแทนของเว็บนั้นมีมากกว่าความเป็นจริง (เพราะมีเอกสารที่เหมือนกันเยอะ) ทำให้ไม่สามารถหาคำที่สื่อความหมายได้ถูกต้องเท่าที่ควร (ไม่ได้นำมาใช้ในโครงการนี้)

จากขั้นตอนดังกล่าวข้างต้นเราจะได้เอกสารที่เป็น Plain Text มา จากนี้เราจะนำเอกสารเหล่านี้มาเข้าขั้นตอนเตรียมเอกสารแบบปกติซึ่งมีขั้นตอนดังนี้

1. การทำให้คำอยู่ในรูปเดียวกันหรือก็คือถ้าเป็นตัวพิพม์ใหญ่ก็ต้องเป็นพิมพ์ใหญ่ทั้งหมด ถ้าเป็นพิมพ์เล็กก็ต้องพิมพ์เล็กทั้งหมด
2. การนำคำธรรมดาออกจากเอกสาร เช่น a, an, the, for, in เป็นต้น เพราะคำเหล่านี้ไม่ได้บอกว่าเอกสารเหล่านี้มีเนื้อหาเกี่ยวกับเรื่องอะไร ไม่สามารถเป็นตัวแทนของเอกสารได้
3. การทำให้คำที่อยู่ในเอกสารอยู่ในรูปของรากศัพท์ เนื่องจากต่างๆ ในเอกสารนั้นคำที่มีความหมายคล้ายกันมักมาจากรากศัพท์เดียวกันเช่น computer, computing, computation คำเหล่านี้มีความหมายคล้ายกันและมาจากรากศัพท์คำเดียวกันคือ compute การทำให้คำอยู่ในรูปรากศัพท์จะทำให้หาตัวแทนเอกสารได้ชัดเจนยิ่งขึ้น

เมื่อผ่านการเตรียมข้อมูลเรียบร้อยแล้ว ข้อมูลเหล่านี้จะอยู่ในรูปที่สามารถสื่อความถึงเอกสารเว็บได้ดีกว่าข้อมูลที่เราเก็บในตอนแรก ซึ่งการจะหาตัวแทนเอกสารที่ดีและมีประสิทธิภาพได้เราจำเป็นต้องอาศัยข้อมูลเหล่านี้

2.4 TF-IDF

TF เป็นการหาความสำคัญของคำหนึ่งๆ ในเอกสารเหล่านั้น โดยมีแนวคิดว่าคำที่มีความสำคัญในเอกสารนั้นมากคือคำที่ปรากฏในเอกสารนั้นบ่อยๆ โดยปกติแล้วมักจะนำความถี่ของคำที่เราสนใจหารด้วยคำทั้งหมดในเอกสารนั้น ซึ่งค่าที่เราได้ออกมาจะอยู่ระหว่าง 0 – 1 ยิ่งค่าเข้าใกล้ 1 มากยิ่งแสดงว่าคำนั้นมีความสำคัญต่อเอกสารนั้นมาก ซึ่งสามารถแสดงออกมาได้เป็นดังสมการ 2.1 ด้านล่าง

$$TF(d,w) = \frac{Total\ Word\ [w]\ in\ Document\ [d]}{Total\ word\ in\ Document\ [d]} \tag{2.1}$$

แต่ว่า TF ไม่เพียงพอต่อการตีค่าความสำคัญของคำที่มีต่อเอกสารจึงมีการใช้ IDF เข้ามาด้วยแนวความคิดที่ว่าคำที่ปรากฏอยู่ทุกๆ เอกสารน่าจะไม่ใช่คำที่สื่อถึงความหมายของเอกสารได้ดีพอและทำให้เอกสารนั้นสื่อความหมายของเอกสารไม่ชัดเจน IDF จึงจะลดความสำคัญของคำเหล่านี้ลง โดยค่า IDF ของแต่ละคำคิดจาก การนำจำนวนเอกสารทั้งหมดหารด้วยเอกสารที่ปรากฏคำที่เราสนใจ แล้วนำผลลัพธ์นำมาเข้าฟังชั่น Logarithm โดยจะเป็น log ฐานใดก็ได้เพราะไม่ได้กำหนดไว้แต่ในโครงนี้เราใช้ log ฐานธรรมชาติ ซึ่งการหาค่า IDF นี้สามารถแสดงออกมาเป็นดังสมการ 2.2 ด้านล่าง

$$IDF(w) = \ln\frac{Number\ of\ All\ Document}{Number\ of\ Document\ Contain\ Word\ [w]} \tag{2.2}$$

เมื่อได้ค่า IDF แล้วเราก็จะนำมาคูณกับ TF ของแต่ละคำของแต่ละเอกสารเพื่อหาค่า TF-IDF เหมือนกันสมการ 2.3 ด้านล่าง ซึ่งค่า TF-IDF นี่เองจะเป็นตัวแปรที่แสดงถึงความสำคัญของคำดังกล่าวของเอกสารนั้นๆที่เราเลือกใช้ในโครงการนี้

$$TF - IDF(d,w) = TF(d,w) \times IDF(w) \tag{2.3}$$

2.5 K-Nearest Neighbors Algorithm

โครงการนี้ได้นำศาสตร์ทางด้าน Data Mining มาช่วยในการทำงานด้วย ซึ่งส่วนที่เรานำมาใช้คือแบบจำลองในแบบ Predictive Modeling ซึ่งแบบจำลองประเภทนี้จะช่วยเราในการทำนายหรือตัดสินใจในด้านต่างๆ ซึ่งแบบจำลองประเภทนี้ที่รู้กันดีได้แก่ Decision Tree, Neuron Network, Support Vector Machine แต่ว่าแบบจำลองที่เราเลือกมาใช้ในโครงการของเราคือ K-Nearest Neighbors สาเหตุที่เราเลือกแบบจำลองนี้เพราะว่าแบบจำลองนี้ไม่จำเป็นต้อง Training เพื่อสร้างแบบจำลอง ทำให้แทบไม่เสียเวลาสำหรับการสร้างแบบจำลองเลย และข้อที่เรามีซึ่งเป็นค่า TF-IDF ค่านี้เหมาะมากสำหรับแบบจำลองประเภทนี้เพราะเราสามารถหาค่าความคล้ายกันของเว็บใดๆ โดยการนำค่า TF-IDF นี้มาหาระยะห่างด้วยสมการ Euclidian Distance ได้

$$D(d1,d2) = \sqrt{\sum_{t=1}^{n}(X(t,d1) - X(t,d2))^2} \tag{2.4}$$

จากสมการ 2.4 เป็นการหาค่าระยะห่าง D ของเอกสาร d1 และ d2 โดยนำค่า TF-IDF ซึ่งก็คือค่า X ของแต่ละคำ (t) ในแต่ละเอกสาร (d1, d2) มาลบกันแล้วยกกำลังสองเพื่อกำจัดค่าลบ แล้วนำค่าที่ได้แต่ละคำมาบวกกันทั้งหมดแล้วถอดรูทออก

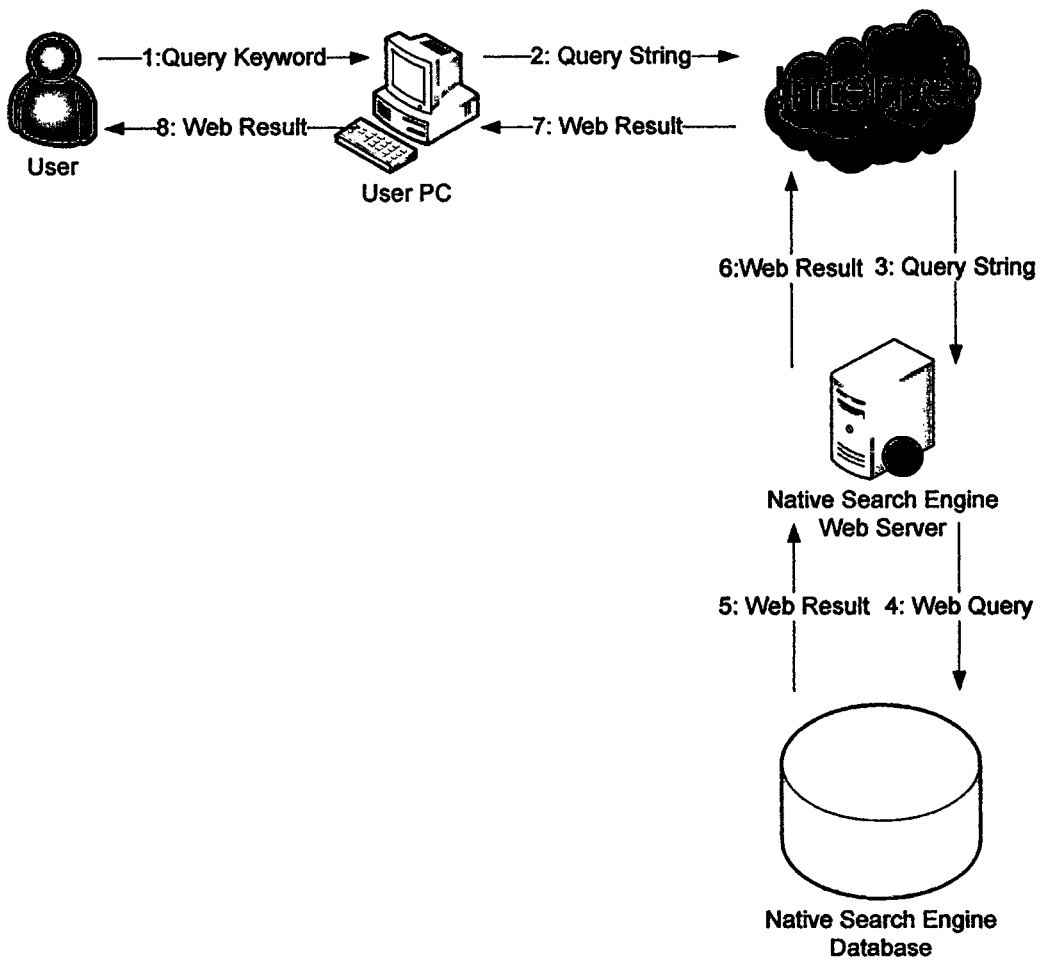
ค่าที่ได้ออกมาจะอยู่ระหว่าง 0-1 ยิ่งเข้าใกล้ 0 มากแสดงว่ายิ่งเอกสารสองฉบับนี้มีความเหมือนกันมาก ในโครงการของนำแบบจำลองนี้มาใช้สองจุด จุดแรกคือการตัดสินใจของ Crawler ที่จะเลือกเว็บไซต์ดังกล่าวมาหาข้อมูลต่อหรือไม่และอีกที่คือตรงการออกรายงานเพื่อเรียงเอกสารจากความที่ความเหมือนกันกับเอกสารต้นแบบสูงที่สุดและลดลงมาตามลำดับ

บทที่ 3

การทำงานของ Search Engine และระบบค้นหาเว็บแบบเจาะจง

ระบบค้นหาเว็บแบบเจาะจงของเรานั้นและการพัฒนาระบบที่นำผลการค้นหาของ Search Engine ไปทำงานต่อยอด เพื่อให้ผลลัพธ์ที่ผู้ใช้งานพอใจกับผลลัพธ์มากยิ่งขึ้น ดังนั้นเราจึงจำเป็นที่จะศึกษาการทำงานของ Search Engine แบบปกติ และนำการทำงานนั้นมาทำงานร่วมกับระบบค้นหาเว็บแบบเจาะจงของเรา และสุดท้ายเราจะแสดงโครงสร้างสถาปัตยกรรมของระบบค้นหาเว็บแบบเจาะจงโดยรวมของเรา

3.1 การค้นหาโดยใช้ Search Engine แบบปกติ



รูปที่ 3.1 การทำงานปกติของ Search Engine

ในการค้นหาเว็บแบบปกติเราจะต้องเข้าไปที่เว็บไซต์ที่ Search Engine นั้นให้บริการ จากนั้นเราจะต้องพิมพ์คีย์เวิร์ดที่ใช้ในการค้นหา เมื่อระบบค้นหาเสร็จเรียบร้อยระบบก็จะแสดงเว็บที่เกี่ยวกับข้องกับคีย์เวิร์ดนั้นกลับมาให้ผู้ใช้ ซึ่งผู้ใช้สามารถปรับเปลี่ยนคีย์เวิร์ดเพื่อค้นหาใหม่ได้อีก จนกว่าผู้ใช้จะพอใจในผลลัพธ์ เหมือนดั่งในรูปที่ 3.1

จากที่เรากล่าวไว้ในข้างต้น ผลลัพธ์ที่ได้จากการค้นหาของ Search Engine นั้น ที่เรามักใช้และเป็นประโยชน์นั้นมักอยู่ในหน้าแรกๆ เท่านั้น หน้าที่เหลือเราก็ไม่ได้ใช้ และถ้าหน้าแรกๆ ไม่มีผลลัพธ์ที่ตรงความต้องการของเรา เราก็จะปรับเปลี่ยนคีย์เวิร์ดเพื่อค้นหาใหม่ จนกว่าจะเจอเว็บที่เราพอใจ

จะเห็นได้ว่าเว็บที่ค้นคืนขึ้นมาได้นั้นมีเพียงเล็กน้อยที่ตรงความต้องการของผู้ใช้ เราได้เล็งเห็นถึงข้อด้วยในจุดนี้ จึงมีความต้องการที่จะพัฒนาระบบที่จะสามารถค้นหาเว็บที่ตรงความต้องการของผู้ใช้เพิ่มเติมจากการค้นหาเดิมของ Search Engine

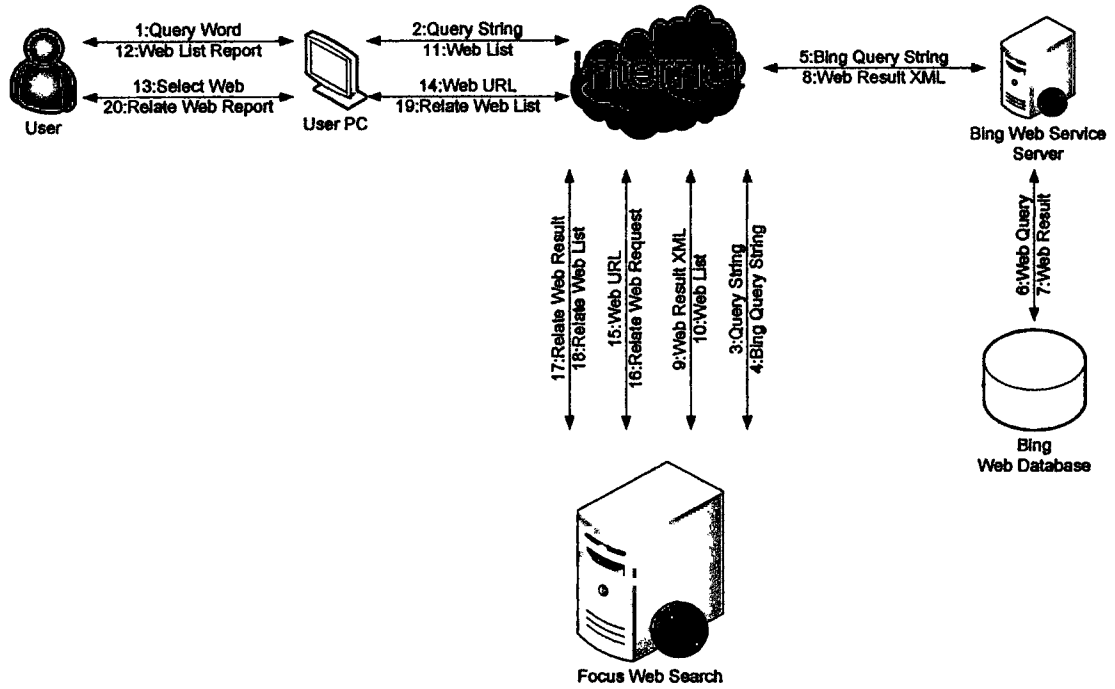
3.2 การทำงานของระบบค้นหาเว็บแบบเจาะจง

ระบบค้นหาเว็บแบบเจาะจง เป็นระบบที่นำมาเสริมการทำงานของ Search Engine ไม่ใช่ลักษณะการทำงานเพื่อแข่งว่าใครดีกว่ากัน แต่เป็นการประสานการทำงานร่วมกันเพื่อนำผลลัพธ์ของระบบหนึ่งมาใช้งานต่อเพื่อให้ได้ผลลัพธ์ที่ตรงความต้องการของผู้ใช้งานมากขึ้นกว่าเดิม

ตัว Search Engine ที่เรานำมาใช้งานร่วมกับระบบของเราคือ Bing เป็น Search Engine ของ Microsoft ซึ่ง Search Engine ตัวนี้ได้ให้บริการทั้งในแบบ Web Search Engine และแบบ Web Service API ซึ่งเราสามารถนำ Web Service API มาทำงานร่วมกับระบบค้นหาเว็บแบบเจาะจงของเราได้ ซึ่งจะดีกว่าการไปเก็บจากหน้าผลลัพธ์การค้นหาบนเว็บโดยตรงมาก ทั้งไม้ต้องเสียเวลาจัดเตรียมข้อมูล ความเร็วที่ได้เร็วกว่า และปริมาณแบนวิธที่ใช้ก็น้อยกว่า

ระบบของเรามีลักษณะเป็น Web Application พัฒนาขึ้นด้วยภาษา PHP ในแบบ OOP โดยใช้ Framework ของ Codeigniter เพื่อสร้างและกำหนดระเบียบในการเขียนโปรแกรม และมีการใช้ฟังชั่นพิเศษของ PHP ซึ่งก็คือ CURL ไว้ใช้ในการติดต่อเว็บไซต์อื่นเพื่อค้นหาข้อมูลเว็บไซต์และนำไปสร้างเป็น Crawler ของตัวระบบ

นอกจากนี้ระบบของเรายังประกอบด้วยส่วนที่สำคัญอื่นๆ อีก เช่น ระบบเปรียบเทียบเว็บไซต์โดยใช้แบบจำลอง K-NN, ระบบหาตัวแทนเว็บไซต์, ระบบเตรียมข้อมูลเว็บไซต์, ระบบหาค่าความถี่ของคำในเว็บ ซึ่งจะแสดงให้เห็นในหัวข้อต่อไป

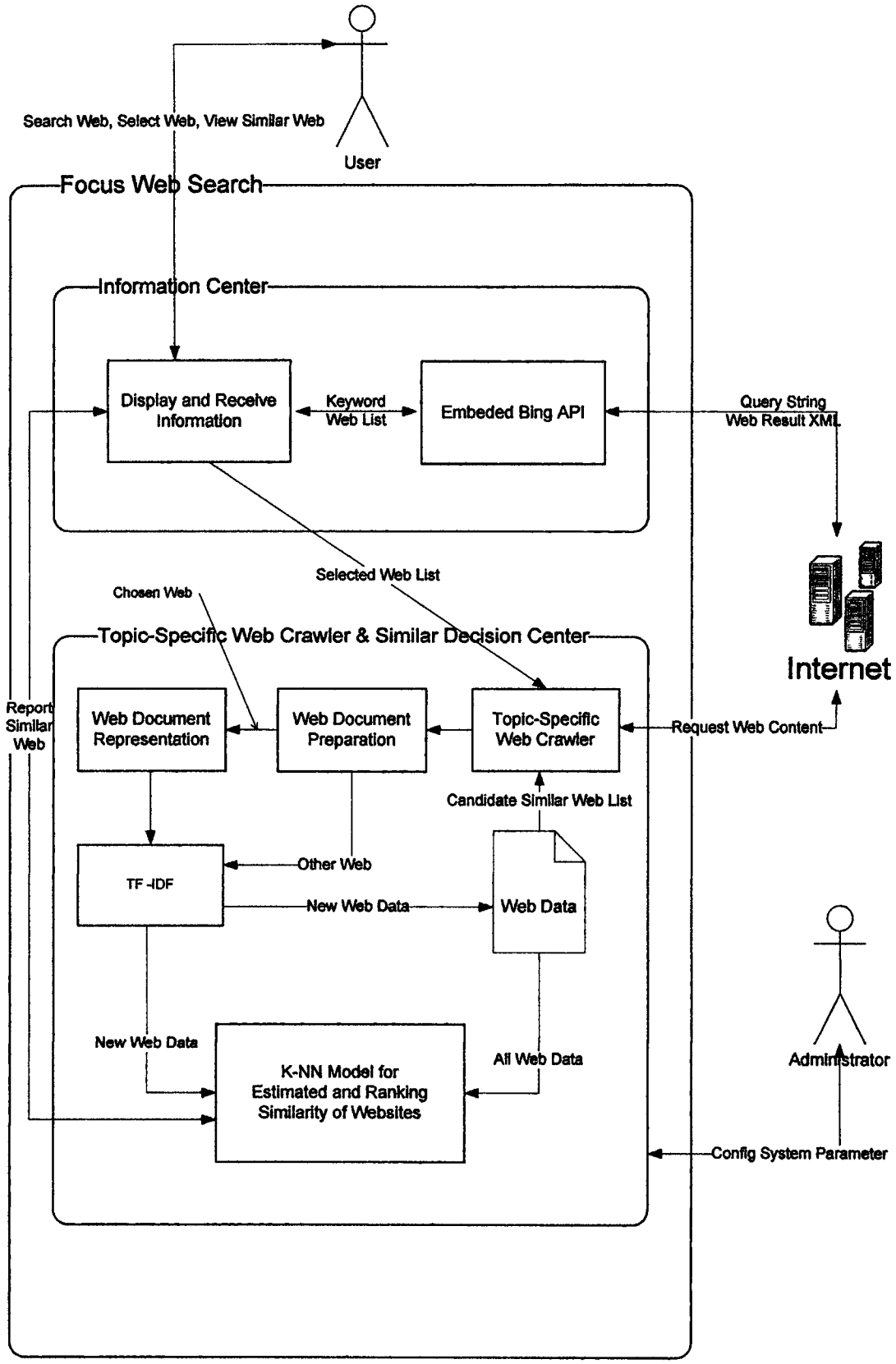


รูปที่ 3.2 การทำงานของระบบค้นหาเว็บแบบเจาะจง

การทำงานของระบบค้นหาเว็บแบบเจาะจงโดยรวมก็จะเหมือนกับรูปที่ 3.2 ซึ่งแสดงการทำงานของระบบร่วมกับผู้ใช้งานและ Bing Web Service Server ซึ่งเราหวังว่าระบบของเราจะสามารถช่วยค้นหาเว็บไซต์ที่ใกล้เคียงหรือเกี่ยวกับข้องกับเว็บไซต์ที่ผู้ใช้ต้องการได้ ซึ่งเว็บไซต์ที่ระบบของเราหาได้นั้นอาจจะใช้เวลาน้อยกว่าในการค้นหาโดย Search Engine ซึ่งต้องปรับเปลี่ยนคีย์เวิร์ดและค้นหาใหม่ไปเรื่อยๆ ซึ่งถ้าสามารถทำได้ผู้ใช้ก็ได้ประโยชน์จากเว็บไซต์ที่ตรงความต้องการที่มีมากขึ้นและ และทาง Search Engine ก็จะรับภาระการทำงานน้อยลง

3.3 สถาปัตยกรรมของระบบค้นหาเว็บแบบเจาะจง

โครงสร้างสถาปัตยกรรมของระบบค้นหาเว็บแบบเจาะจงที่เราจะพัฒนาขึ้นนั้นสามารถแบ่งออกไปได้เป็นสองส่วนใหญ่ๆ คือ ส่วนแรกเป็นระบบนำเสนอและรับข้อมูลต่อผู้ใช้ ส่วนนี้มีหน้าที่รับและแสดงข้อมูลให้กับผู้ใช้รวมถึงติดต่อและทำงานร่วมกับ Bing API และอีกส่วนหนึ่งคือระบบค้นหาเว็บไซต์ที่คล้ายหรือเกี่ยวข้องกัน ส่วนนี้ทำหน้าที่ในการค้นหาและคัดเลือกเว็บไซต์ที่คาดว่าจะมีความคล้ายคลึงและเกี่ยวข้องกันกับเว็บไซต์ที่ผู้ใช้เลือก



รูปที่ 3.3 สถาปัตยกรรมระบบค้นหาเว็บแบบเจาะจง

ส่วนระบบนำเสนอและรับข้อมูลต่อผู้ใช้ ส่วนนี้จะทำหน้าที่ในการติดต่อกับผู้ใช้เป็นหลัก ผู้ใช้สามารถค้นหาเว็บจาก Search Engine API ที่เรานำมาทำงานร่วมกับระบบของเราและสามารถเลือกเว็บไซต์ที่ได้จากการค้นหาปกติเหล่านั้นมาแล้วเลือกว่าเป็นเว็บไซต์ที่ตรงความต้องการได้จากส่วนนี้ ซึ่งเว็บไซต์ที่เลือกนี้จะนำไปใช้ค้นหาเว็บที่คล้ายหรือเกี่ยวข้องกันต่อไป นอกจากนั้นส่วนนี้ยังทำหน้าที่ได้การแสดงผลลัพธ์จากการทำงานของค้นหาเว็บไซต์แบบเจาะจงโดยจะแสดงเว็บไซต์ที่มีความคล้ายหรือเกี่ยวข้องกับเว็บไซต์ที่ผู้ใช้เลือกกลับคืนมา โดยส่วนของระบบนี้อยู่ด้านบนของรูป

อีกส่วนคือระบบค้นหาเว็บไซต์ที่คล้ายหรือเกี่ยวข้องกัน ส่วนนี้จะนำเว็บที่ผู้ใช้เลือกไปทำการ ทำความสะอาดข้อมูลเว็บและทำการหาข้อมูลที่เป็นตัวแทนเว็บ โดยใช้ความถี่ของคำที่ปรากฏเป็นตัวแทนในเอกสารเว็บเหล่านั้น และตัว Crawler ก็จะนำ URL ที่อยู่ในเว็บเหล่านี้ไปหาข้อมูลเว็บอื่นๆ ต่อ เมื่อตัวระบบได้ข้อมูลเว็บอื่นเข้ามาแล้ว ระบบก็จะนำมาทำความสะอาดข้อมูลเว็บและความถี่ของคำที่เหมือนกับกับตัวแทนเอกสารเว็บที่ผู้ใช้เลือก

จากนั้นก็จะส่งข้อมูลมาให้อีกส่วนที่สำคัญและเป็นหัวใจของระบบค้นหาเว็บแบบเจาะจงก็คือ ส่วนที่ใช้ในการวัดความคล้ายคลึงกันของเว็บไซต์ ซึ่งส่วนนี้จะสามารถบอกได้ว่าเว็บแต่ละเว็บที่ได้มานั้นมีความคล้ายคลึงกับเว็บที่ผู้ใช้เลือกมากน้อยเท่าใด ซึ่งวิธีการและกระบวนการที่เราใช้ก็คือการใช้แบบจำลองชนิด K-Nearest Neighbors โดยใช้เนื้อหาของเว็บเป็นหลัก (Content-Based Filtering)

3.4 ลำดับการทำงานของระบบค้นหาเว็บแบบเจาะจง

1. ผู้ใช้ค้นหาเว็บไซต์โดยใช้ Search Engine API ที่ระบบเตรียมมาให้ หลังจากนั้น ผู้ใช้ดังกล่าวก็เลือกเว็บไซต์ที่ตรงความต้องการ เว็บไซต์ที่เลือก ระบบจะทำเก็บเว็บนั้นไว้เพื่อนำไปใช้ในขั้นตอนต่อไป
2. เว็บไซต์ที่ผู้ใช้เลือกจะถูกทำให้อยู่ในรูปที่ใช้งานได้ ซึ่งก็คือการเตรียมและหาตัวแทนเอกสารเว็บ (Web Preparation) ซึ่งมีรายละเอียดในบทที่ 2 จากนั้นก็เก็บข้อมูลซึ่งแบบจำลองของ K-Nearest Neighbors สามารถนำไปใช้งานได้ และตัว Crawler จะนำข้อมูลของเว็บที่ยังไม่ได้ทำความสะอาดไปหา URL เพื่อที่จะไปหาข้อมูลต่อ

3. เมื่อ Crawler ได้ข้อมูลเว็บใหม่มาจะมาทำการเปรียบเทียบกับเว็บที่ผู้ใช้เลือกโดยใช้แบบจำลอง K-Nearest Neighbors เพื่อคิดหาค่าความเหมือนกันของเว็บไซต์ หลังจากนั้นก็ใช้ข้อมูลเว็บนี้ไปหาข้อมูลเว็บอื่นต่อ แต่ถ้าตัวเว็บที่หามาเมื่อคิดหาค่าความเหมือนแล้วมีค่าความเหมือนกันกับเว็บไซต์ที่ผู้ใช้เลือกต่ำกว่าค่าที่กำหนด ตัว Crawler ก็จะไม่เลือกเว็บนี้มาไปหาข้อมูลต่อ ซึ่ง Crawler ประเภทนี้มีรายละเอียดในบทที่ 2
4. เมื่อระบบได้เว็บไซต์พร้อมกับค่าความเหมือนของเว็บไซต์เมื่อเปรียบเทียบกับเว็บที่ผู้ใช้เลือกครั้งแรกมาแล้ว ระบบจะทำการทำนายว่าเว็บไซต์ใดในระบบที่ผู้ใช้น่าจะเห็นว่าเป็นประโยชน์หรือตรงความต้องการที่สุด โดยดูจากระยะความห่างของเว็บไซต์ ระบบจะเลือกเว็บที่มีระยะความห่างจากเว็บที่ผู้ใช้เลือกน้อยที่สุดออกมา
5. ระบบนำเสนอเว็บไซต์ที่ทำนายต่อผู้ใช้

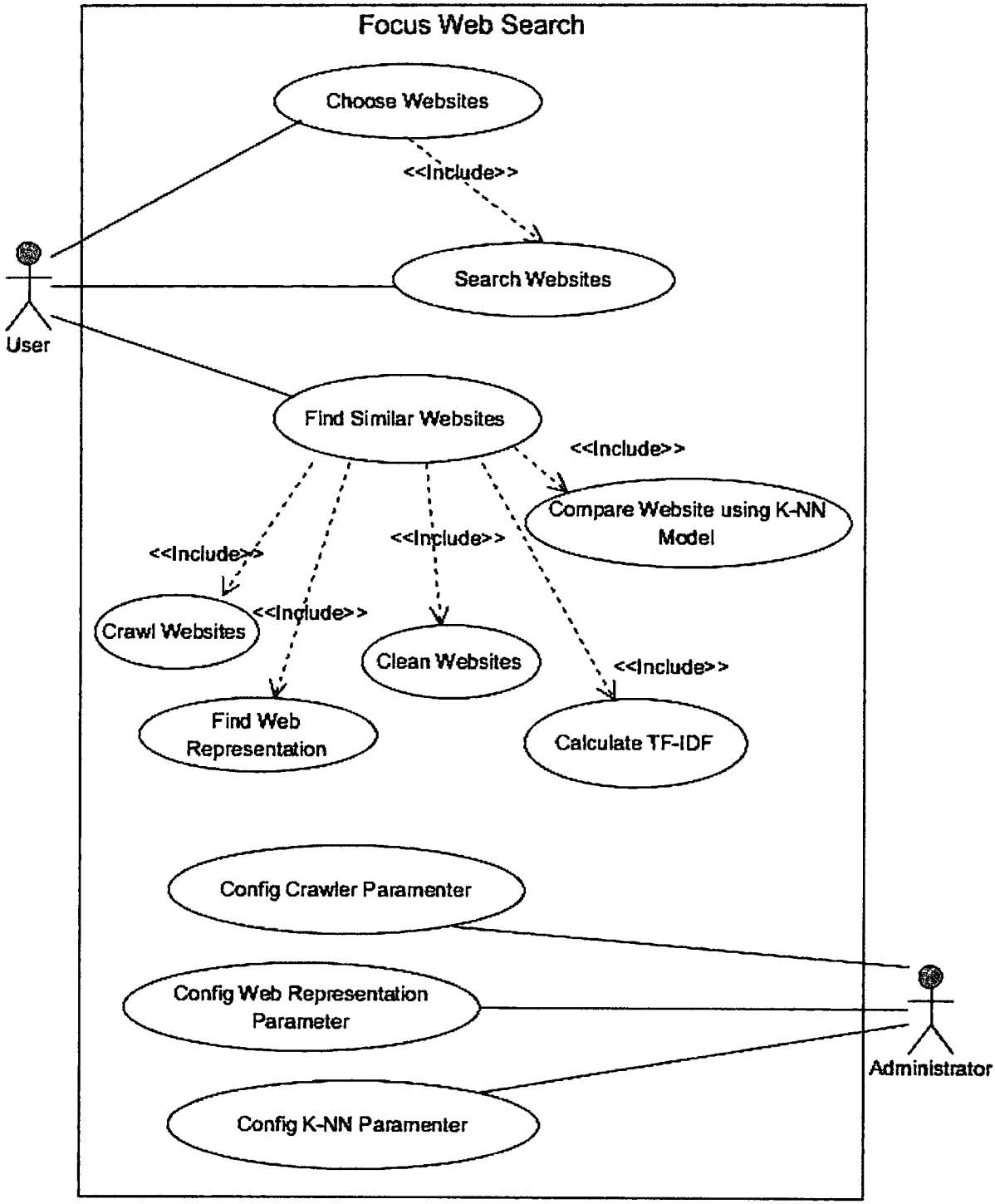
บทที่ 4

การออกแบบระบบค้นหาเว็บแบบเจาะจง

หลังจากที่เราได้ทราบถึงรายละเอียดของโครงการที่จะพัฒนาแล้ว ในบทนี้เราจึงได้เริ่มในส่วนของการออกแบบระบบ ซึ่งการออกแบบระบบของเราจะแสดงออกมาในรูปแบบของยูเอ็มแอลไดอะแกรม (UML Diagram) ซึ่งไดอะแกรมประเภทนี้เป็นที่ใช้กันอย่างแพร่หลายในวงการพัฒนาซอฟแวร์ เราจะนำไดอะแกรมที่อยู่ในยูเอ็มแอลบางส่วนมาใช้ในการออกแบบระบบงานของเราซึ่งประกอบด้วย ยูสเคสไดอะแกรมที่แสดงภาพรวมของระบบและมีการอธิบายเพิ่มเติมด้วย ไดอะแกรมลำดับกิจกรรมการทำงาน (Activity Diagram) กับไดอะแกรมลำดับการทำงาน (Sequence Diagram) และสุดท้ายเป็นคลาสไดแกรมในระดับข้อมูลของระบบ (Data Model Class Diagram)

4.1 ยูสเคสไดอะแกรม

ยูสเคสไดอะแกรมจะแสดงถึงภาพรวมทั้งหมดของระบบ โดยแสดงบทบาทต่างๆ ของผู้ใช้งานระบบ และสิ่งที่ระบบให้บริการแก่ผู้ใช้แต่ละคนที่มีบทบาทเหล่านั้น ซึ่งจะเห็นได้จากรูปที่ 4.1 ในหน้าถัดไป

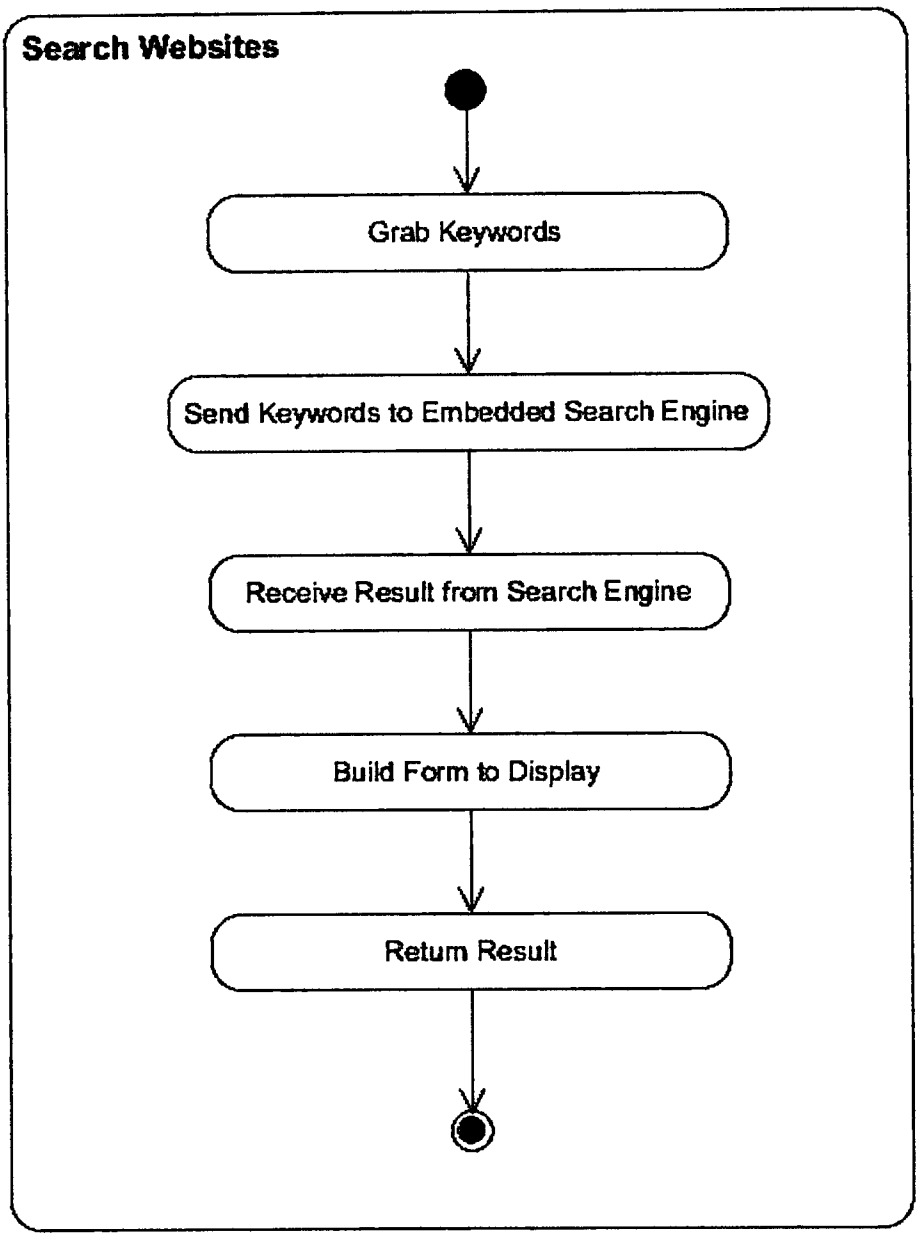


รูปที่ 4.1 ยูสเคสไดอะแกรมระบบค้นหาเว็บแบบเจาะจง

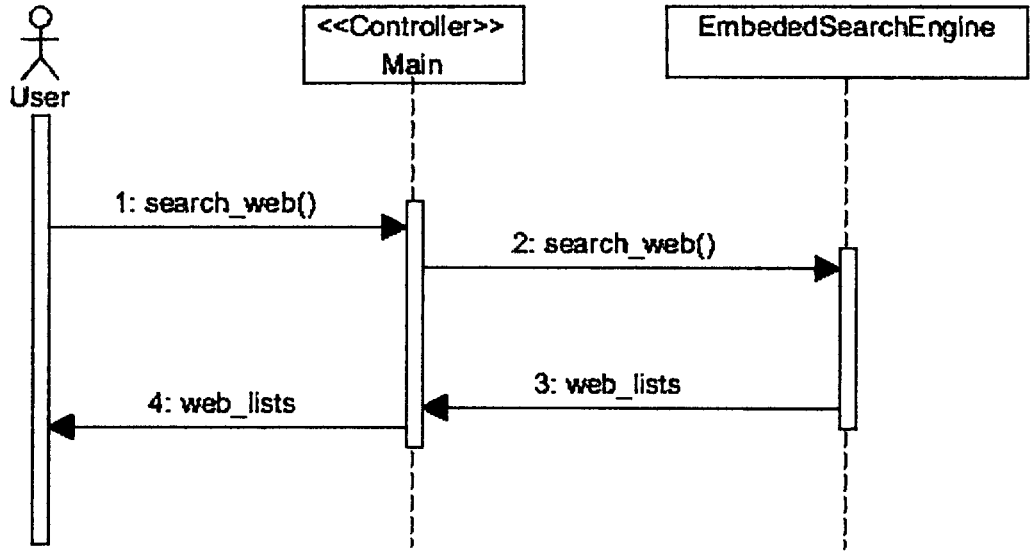
4.1.1 รายละเอียดยูสเคส Search Websites

ตารางที่ 4.1 คำอธิบายยูสเคส Search Websites

คำอธิบายลำดับกิจกรรมของยูสเคส Search Websites
ขั้นตอนหลัก : 1. รับคีย์เวิร์ดจากผู้ใช้ 2. ส่งคีย์เวิร์ดดังกล่าวไปให้ Bing API 3. รับผลลัพธ์กลับมาจาก Bing API 4. จัดรูปแบบข้อมูลที่มาในแบบ XML ให้เหมาะสมกับการแสดงผล 5. ส่งผลลัพธ์กลับไปยังผู้ใช้
เงื่อนไขข้อยกเว้น : 1. ถ้าไม่มีผลลัพธ์จากการค้นหาด้วยคีย์เวิร์ดดังกล่าว a. ระบบแจ้งต่อผู้ใช้ว่าคีย์เวิร์ดนี้ไม่มีผลลัพธ์ที่สามารถแสดงได้



รูปที่ 4.2 ไดอะแกรมลำดับกิจกรรมการทำงานของยูสเคส Search Websites

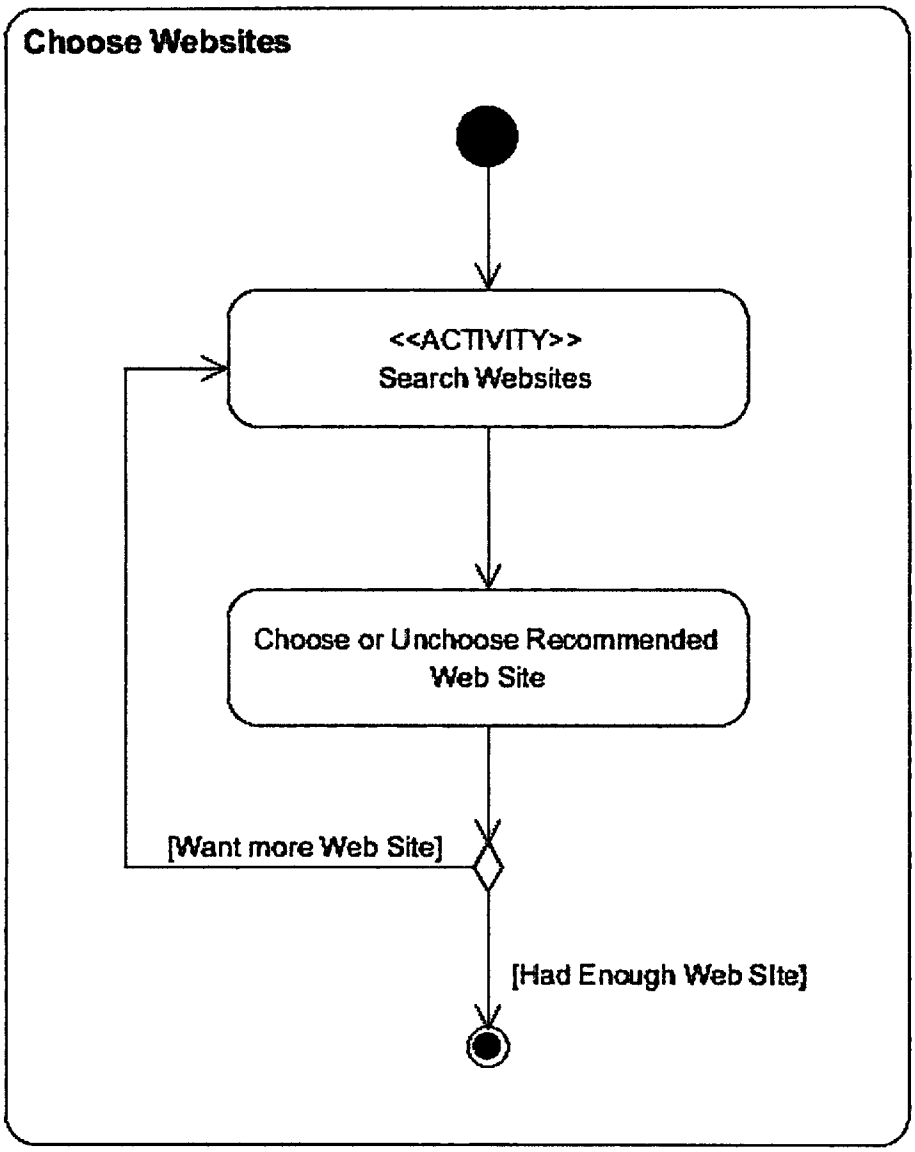


รูปที่ 4.3 ไดอะแกรมลำดับการทำงานของยูสเคส Search Websites

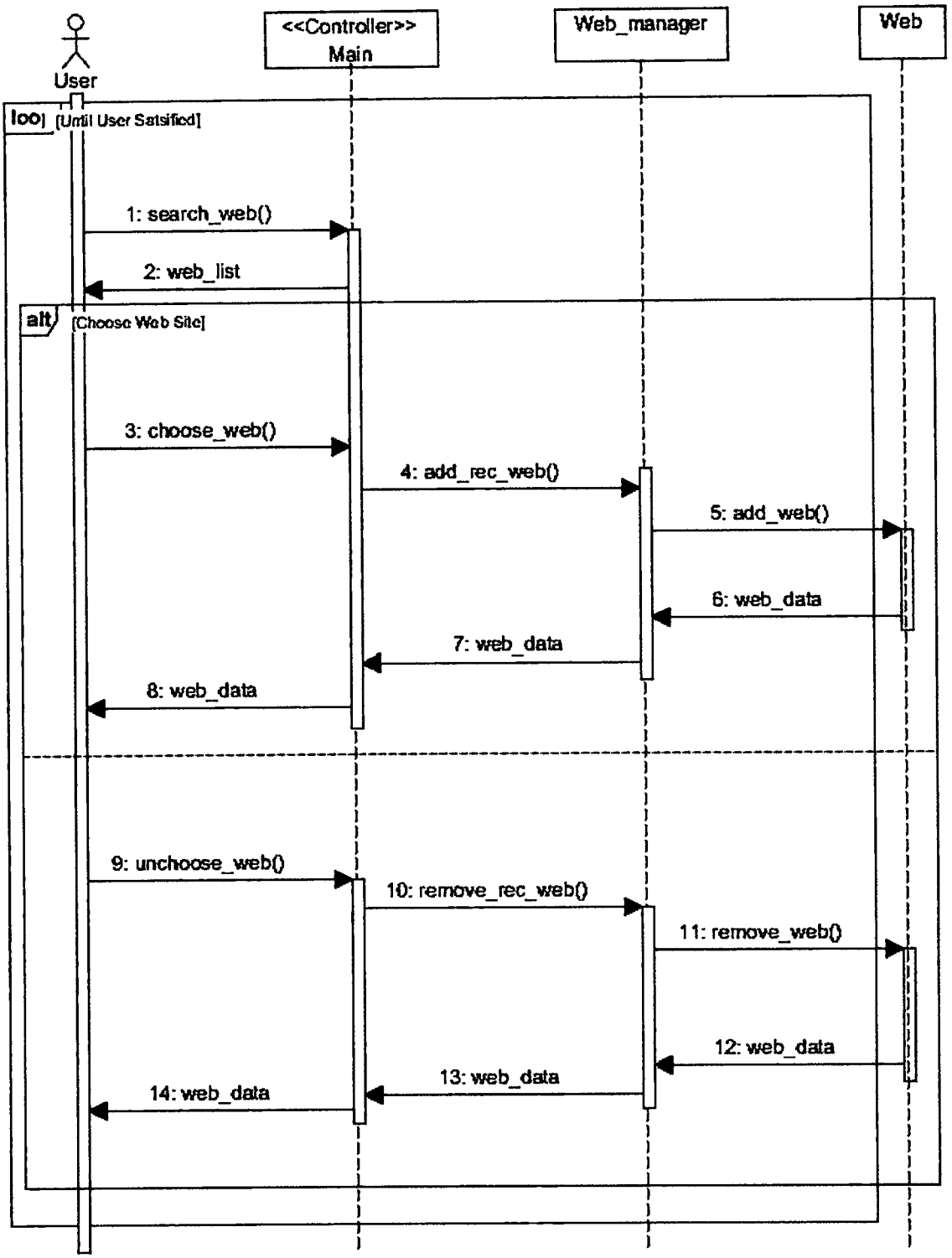
4.1.2 รายละเอียดยูสเคส Choose Websites

ตารางที่ 4.2 คำอธิบายยูสเคส Choose Websites

คำอธิบายลำดับกิจกรรมของยูสเคส Choose Websites
ขั้นตอนหลัก : 1. ค้นหาเว็บไซต์โดยยูสเคส Search Websites 2. เลือกเว็บไซต์ที่ตรงความต้องการหรือถอนเว็บไซต์เว็บไซต์ที่เลือกไว้แล้ว 3. วนกลับไปทำข้อ 1 จนกว่าผู้ใช้จะพอใจ
เงื่อนไขข้อยกเว้น : 1. ถ้าไม่มีผลลัพธ์จากการค้นหาเว็บไซต์ a. ระบบจะแนะนำให้ผู้ใช้ใส่คีย์เวิร์ดใหม่



รูปที่ 4.4 ไดอะแกรมลำดับกิจกรรมการทำงานของยูสเคส Choose Websites

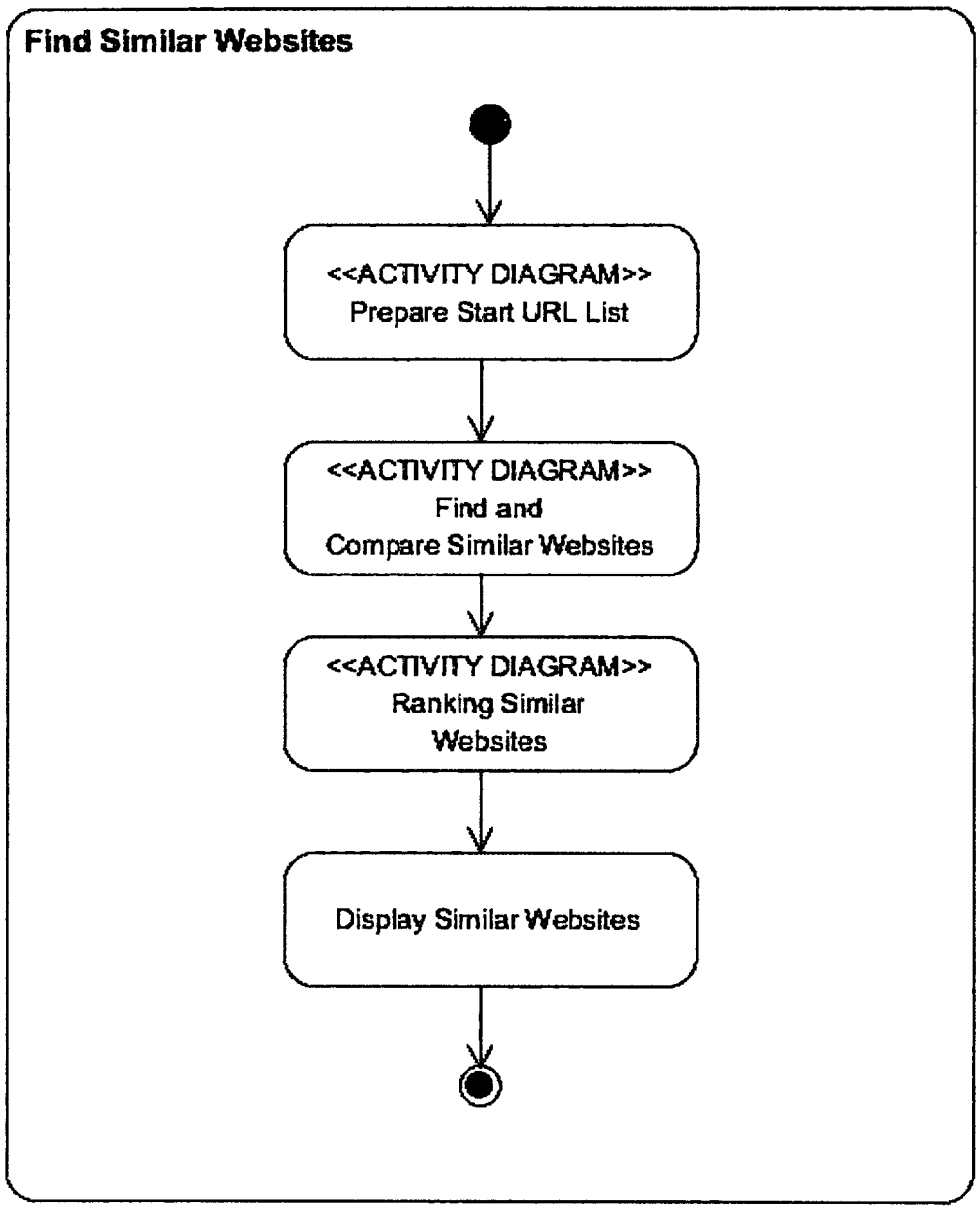


รูปที่ 4.5 ไดอะแกรมลำดับการทำงานของยูสเคส Choose Websites

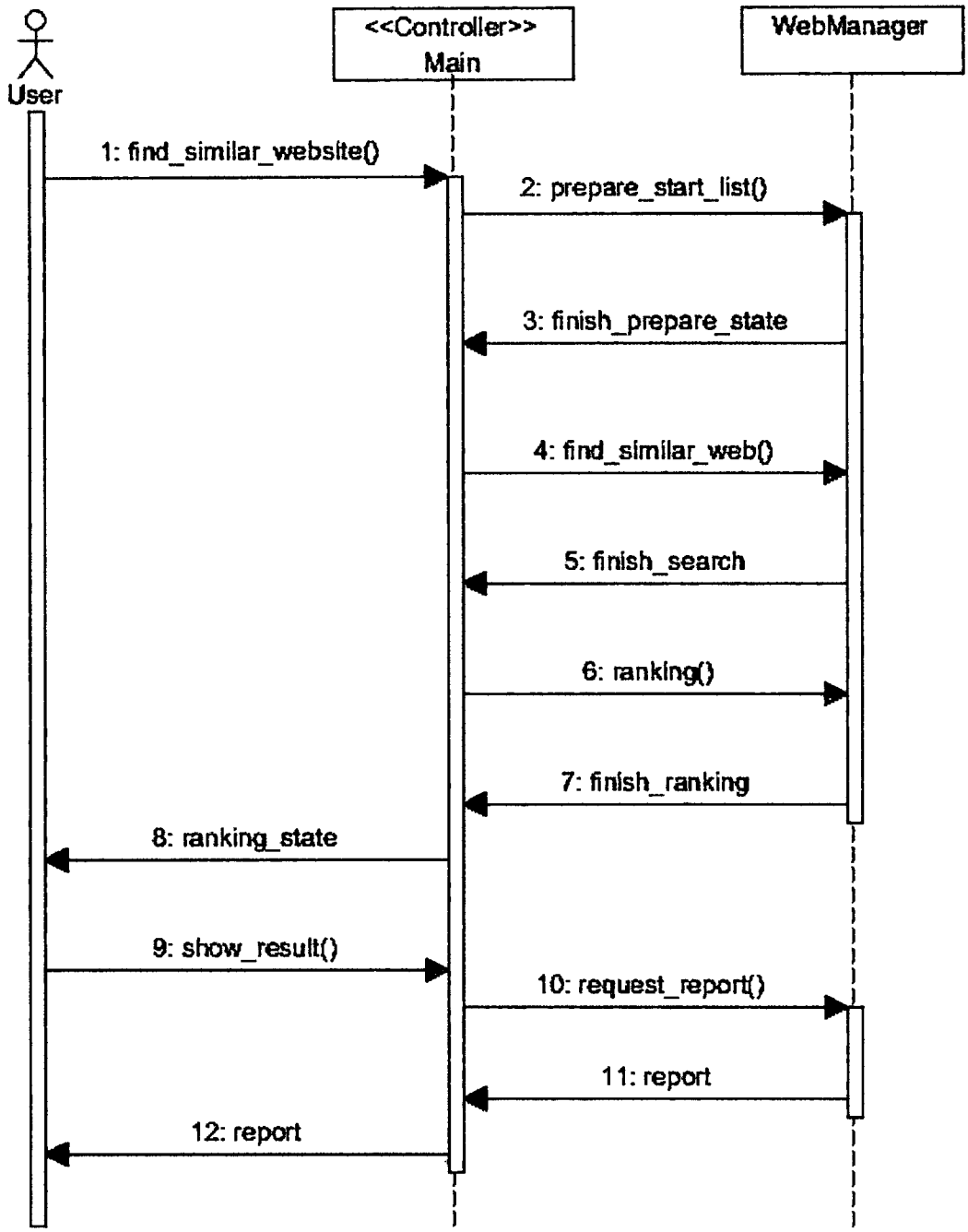
4.1.3 รายละเอียดยูสเคส Find Similar Websites

ตารางที่ 4.3 คำอธิบายยูสเคส Find Similar Websites

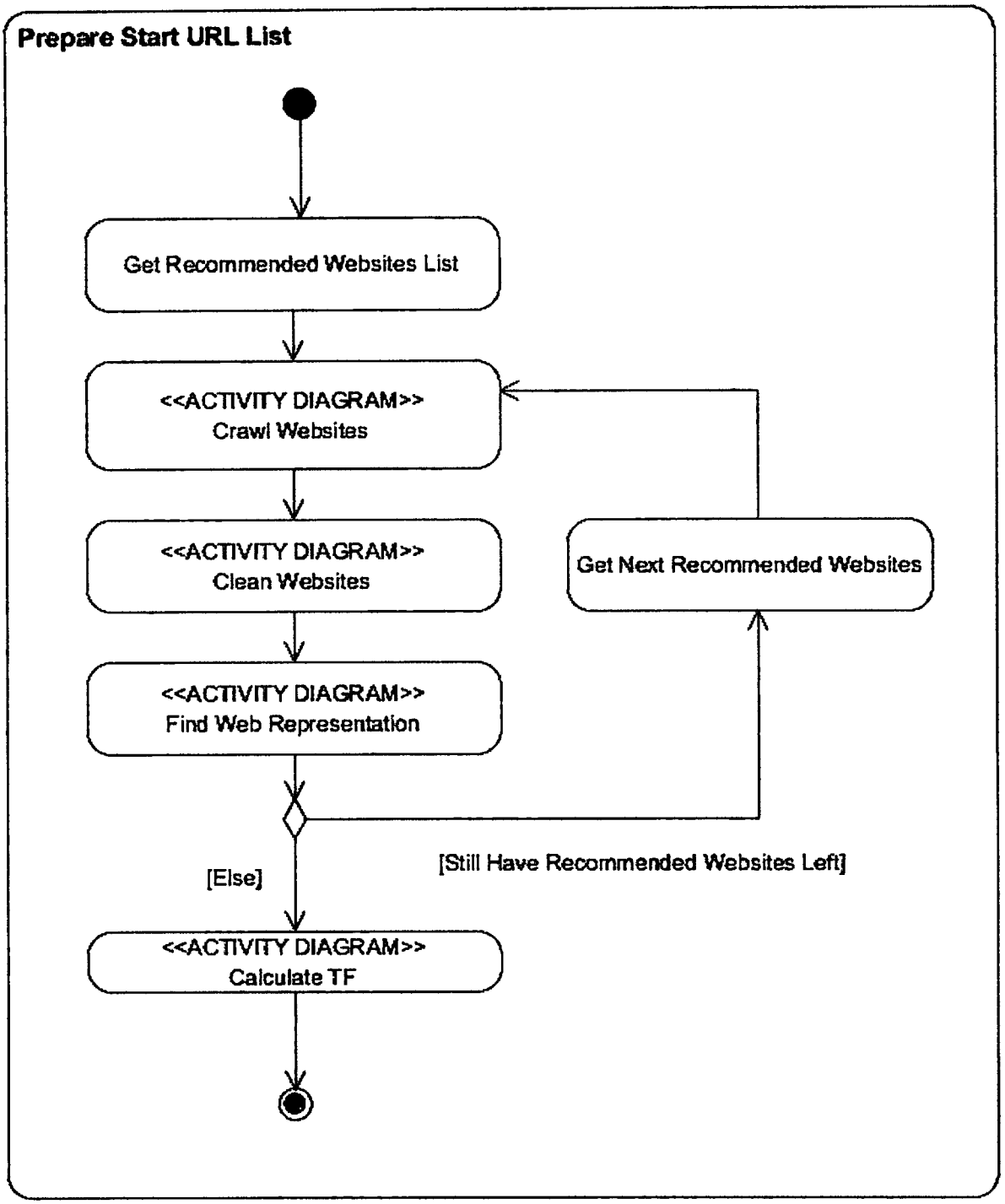
คำอธิบายลำดับกิจกรรมของยูสเคส Find Similar Websites
ขั้นตอนหลัก : 1. เตรียมรายชื่อของ URL ที่จะให้ Crawler ทำงาน (ไดอะแกรมลำดับกิจกรรมการทำงาน รูปที่ 4.8) 2. ค้นหาเว็บไซต์จาก URL ที่ได้มา เมื่อได้เนื้อหาเว็บไซต์ของ URL นั้นๆ แล้วก็นำไปเปรียบเทียบความเหมือนระหว่างเว็บไซต์ที่หามาได้กับเว็บไซต์ที่ผู้ใช้เลือกจากนั้นจึงคัดเลือกเว็บไซต์ที่ความเหมือนกับเว็บไซต์ที่ผู้ใช้เลือกไปทำงานต่อ (ไดอะแกรมลำดับกิจกรรมการทำงาน รูปที่ 4.10) 3. ทำการจัดลำดับเว็บไซต์ที่ผ่านการคัดเลือกทั้งหมด โดยการใช้แบบจำลอง K-NN เปรียบเทียบค่าความเหมือนกันของเว็บไซต์โดยใช้ค่า TF-IDF ของแต่ละคำในแต่ละเว็บไปทำการหาค่า Euclidean distance กับเว็บที่ผู้ใช้เลือก (ไดอะแกรมลำดับกิจกรรมการทำงาน รูปที่ 4.12) 4. แสดงผลลัพธ์
เงื่อนไขข้อยกเว้น : 1. ถ้าไม่มีผลลัพธ์จากการค้นหาเว็บไซต์จากเว็บที่ผู้เลือกเลย a. ระบบจะแจ้งว่าไม่เว็บไซต์ที่คล้ายหรือเกี่ยวข้องกับเว็บไซต์ที่ผู้ใช้เลือก



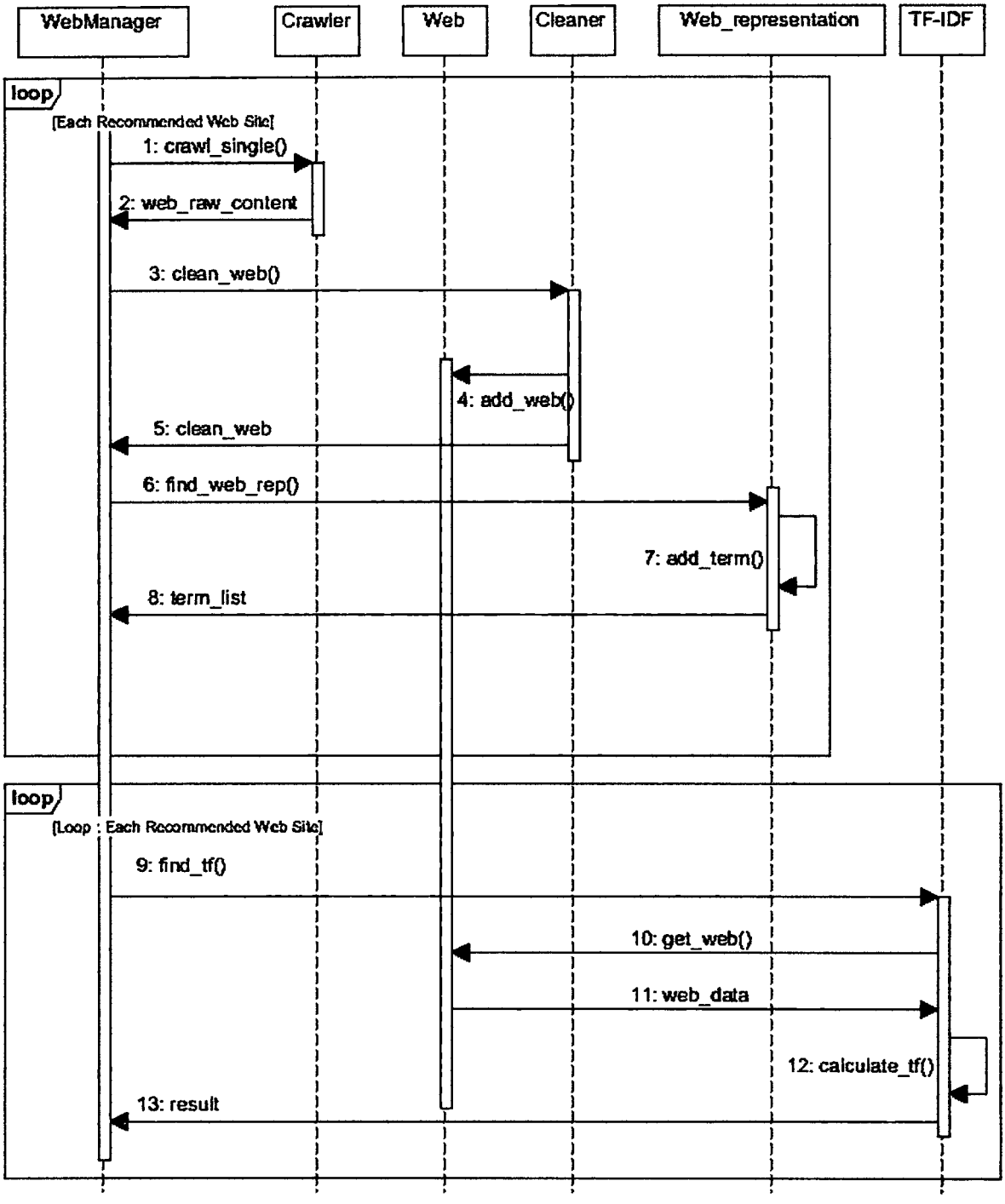
รูปที่ 4.6 ไดอะแกรมลำดับกิจกรรมการทำงานของยูสเคส Find Similar Websites



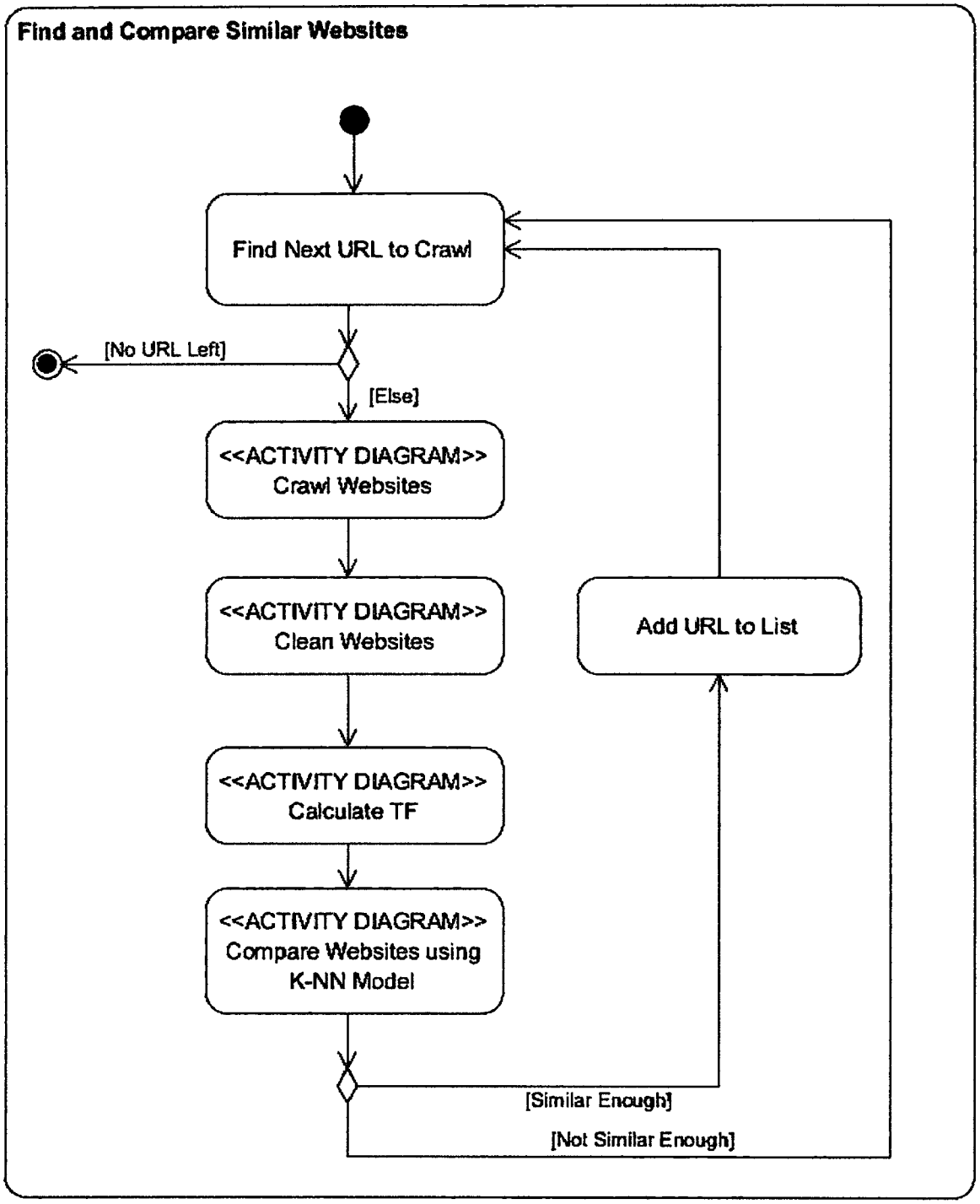
รูปที่ 4.7 ไดอะแกรมลำดับการทำงานของยูสเคส Find Similar Websites



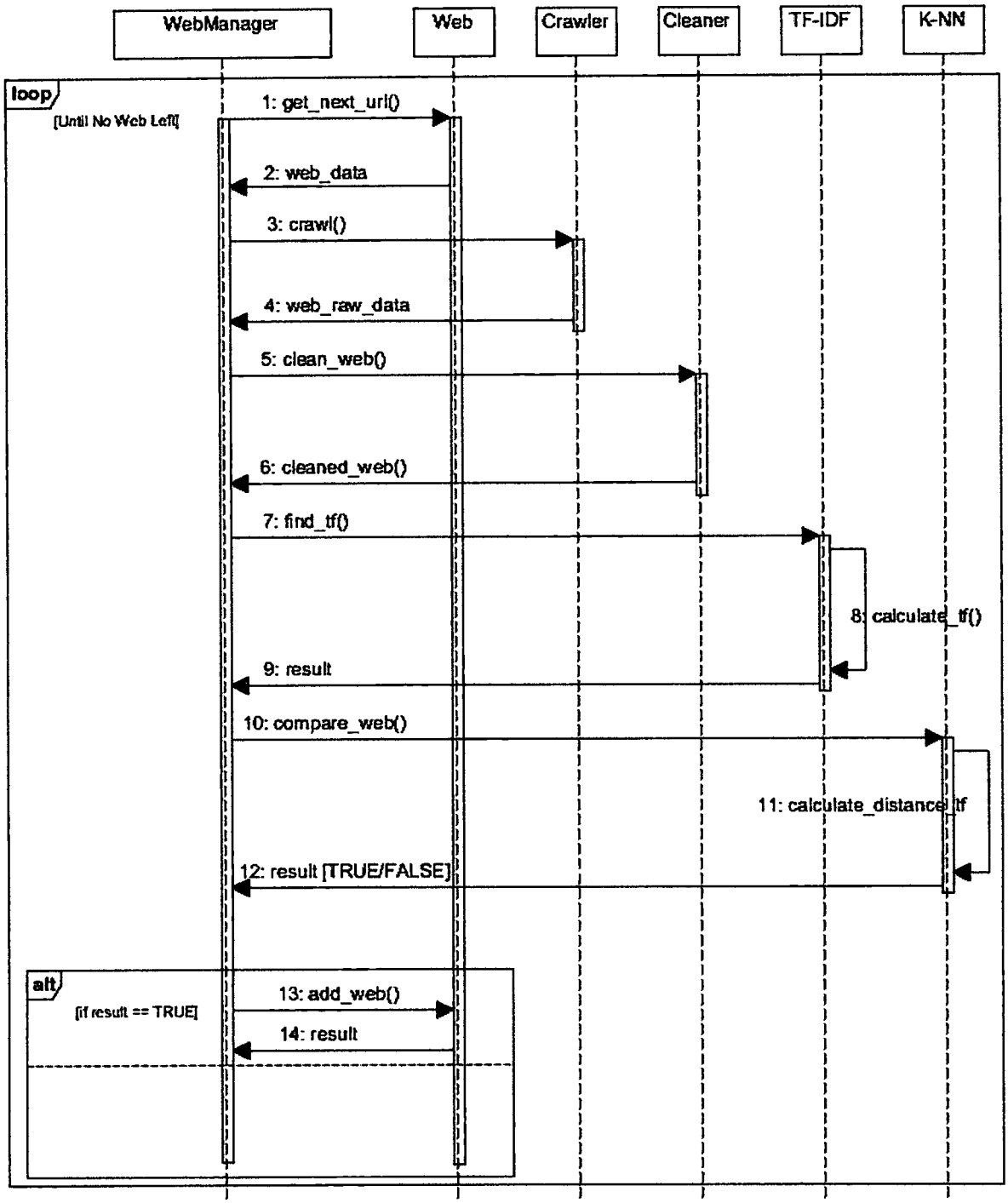
รูปที่ 4.8 ไดอะแกรมลำดับกิจกรรมการทำงานย่อย Prepare Start URL List ของไดอะแกรมลำดับกิจกรรมการงาน Find Similar Websites



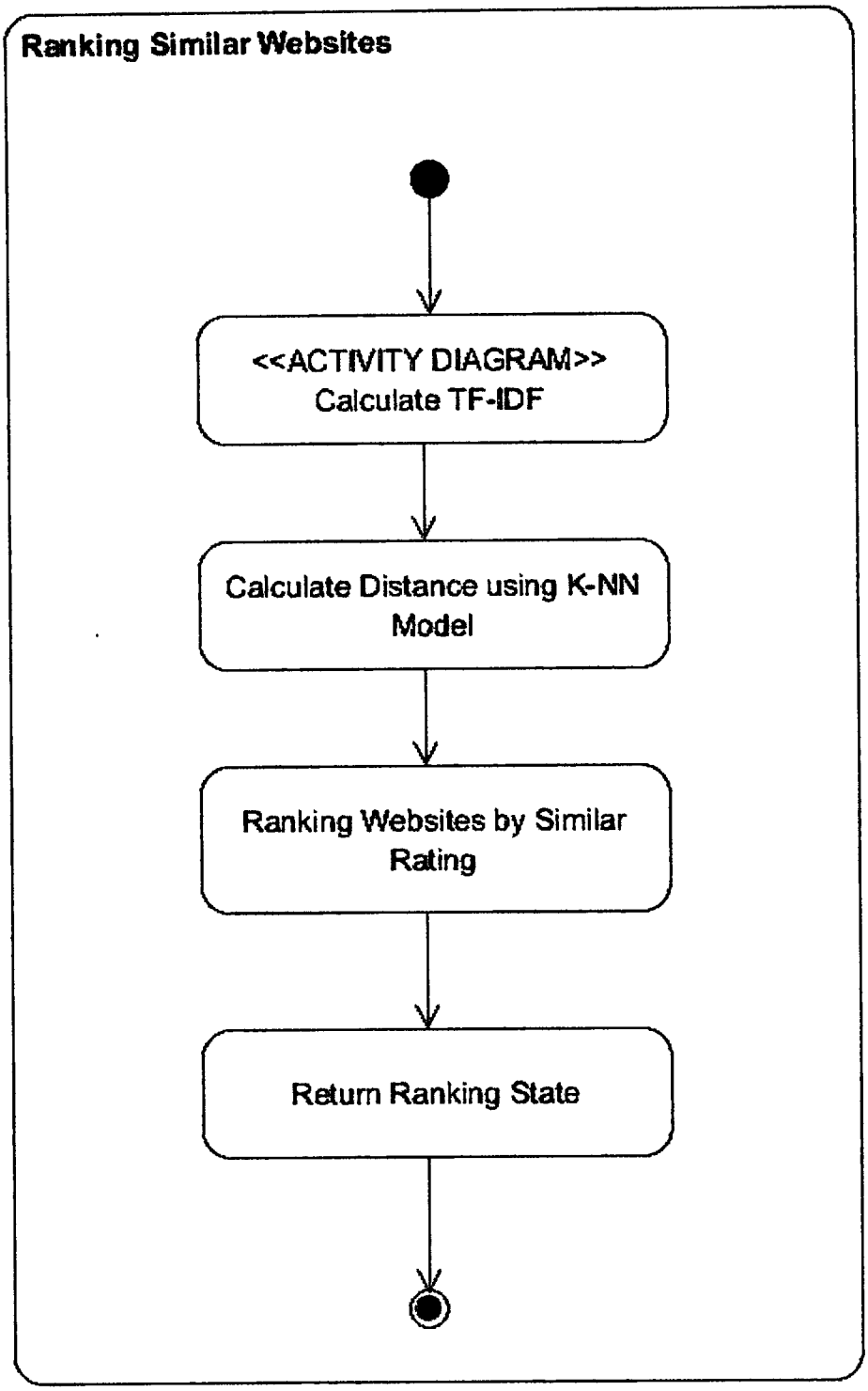
รูปที่ 4.9 ไดอะแกรมลำดับการทำงานของกิจกรรมย่อย Prepare Start URL List



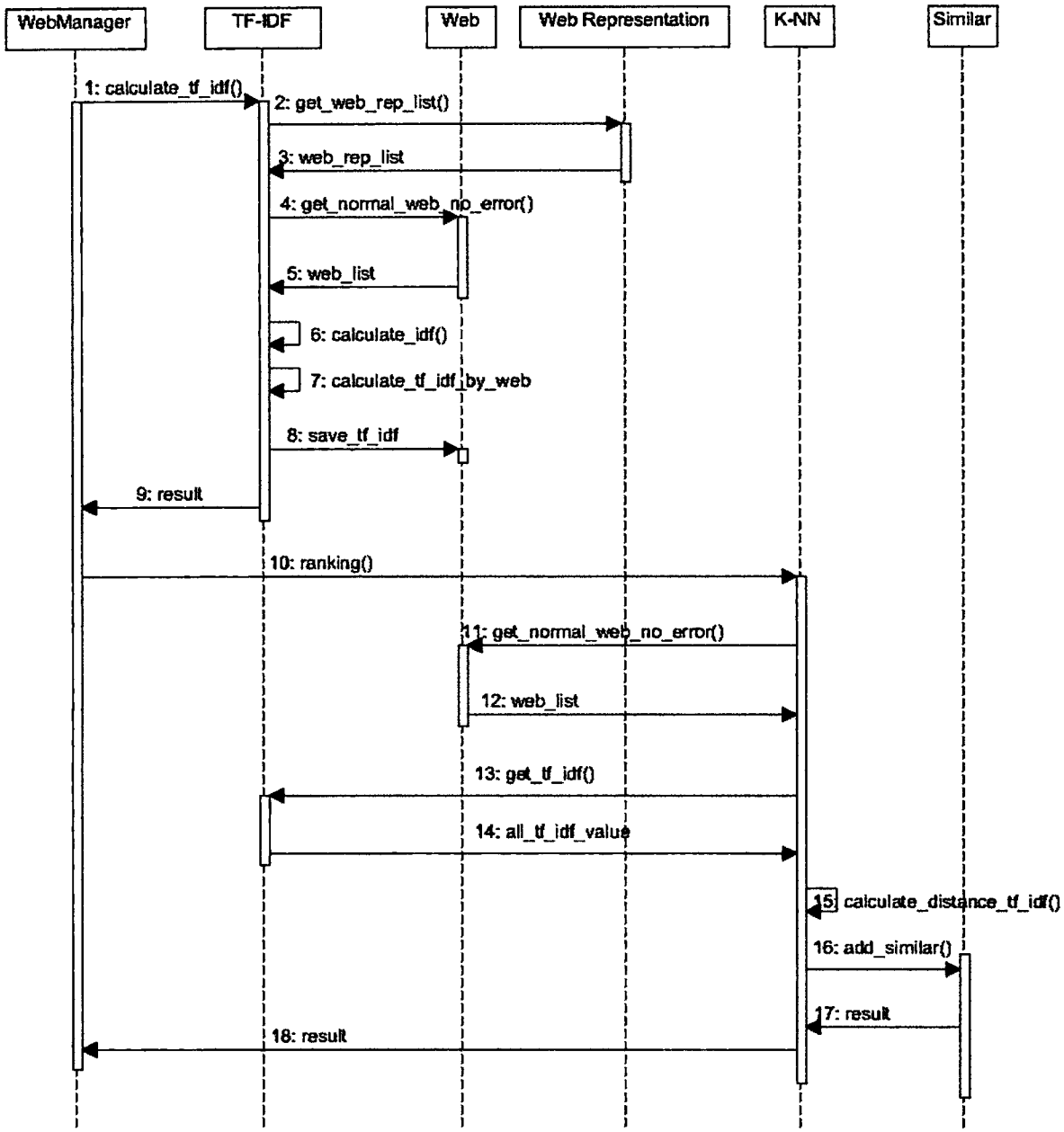
รูปที่ 4.10 ไดอะแกรมลำดับกิจกรรมการทำงานย่อย Find and Compare Similar Websites ของไดอะแกรมลำดับกิจกรรมการทำงาน Find Similar Websites



รูปที่ 4.11 ไดอะแกรมลำดับการทำงานของกิจกรรมย่อย Find and Compare Similar Websites



รูปที่ 4.12 ไดอะแกรมลำดับกิจกรรมการทำงานย่อย Ranking Similar Websites ของไดอะแกรมแสดงลำดับกิจกรรมการทำงาน Find Similar Websites

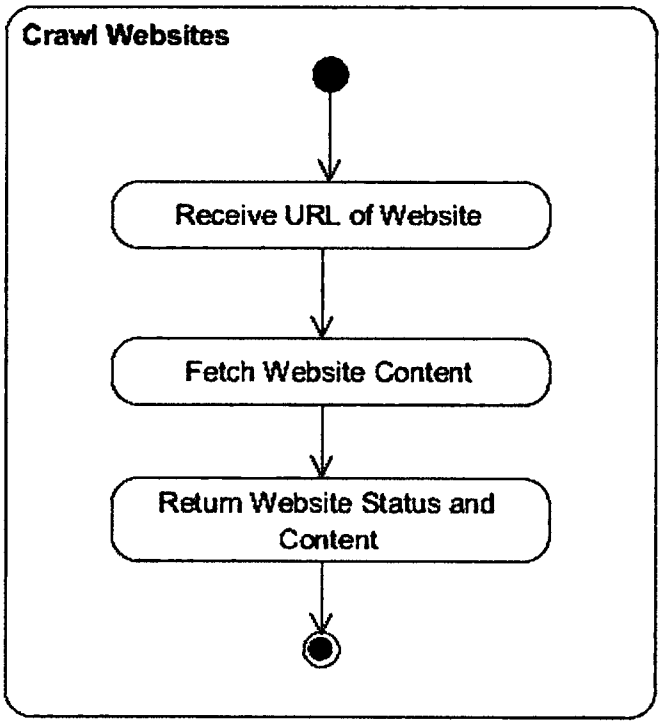


รูปที่ 4.13 ไดอะแกรมลำดับการทำงานของกิจกรรมย่อย Ranking Similar Websites

4.1.4 รายละเอียดยูสเคส Crawl Websites

ตารางที่ 4.4 คำอธิบายยูสเคส Crawl Websites

คำอธิบายลำดับกิจกรรมของยูสเคส Crawl Websites
ขั้นตอนหลัก : 1. รับ URL ของเว็บไซต์ที่ต้องการจะเก็บเนื้อหาเข้ามา 2. ดึงเนื้อหาจากเว็บไซต์นั้นขึ้นมาโดยใช้ Library CURL ของ PHP ซึ่งสามารถใช้ได้ตามเว็บเซอร์เวอร์ทั่วไปในปัจจุบัน 3. ส่งคืนเนื้อหาคืนให้กับผู้เรียกใช้
เงื่อนไขข้อยกเว้น : 1. ถ้า URL ผิด หรือ ไม่อยู่ในรูปแบบที่ถูกต้อง หรือเว็บไซต์ดังกล่าวไม่มีอยู่จริง a. ระบบจะทำการบันทึกไว้ว่าเว็บดังกล่าวไม่สามารถเข้าถึงเนื้อหาได้ และจะไม่ถูกนำมาคิดคำนวณหาค่าความเหมือนและเกี่ยวข้องกันของเว็บไซต์ 2. ถ้าการดึงเนื้อหาของเว็บใช้เวลามากเกิน 30 วินาที (ต่อ 1 เว็บไซต์) a. ระบบจะทำการบันทึกไว้ว่าเว็บดังกล่าวไม่สามารถเข้าถึงเนื้อหาได้ และจะไม่ถูกนำมาคิดคำนวณหาค่าความเหมือนและเกี่ยวข้องกันของเว็บไซต์

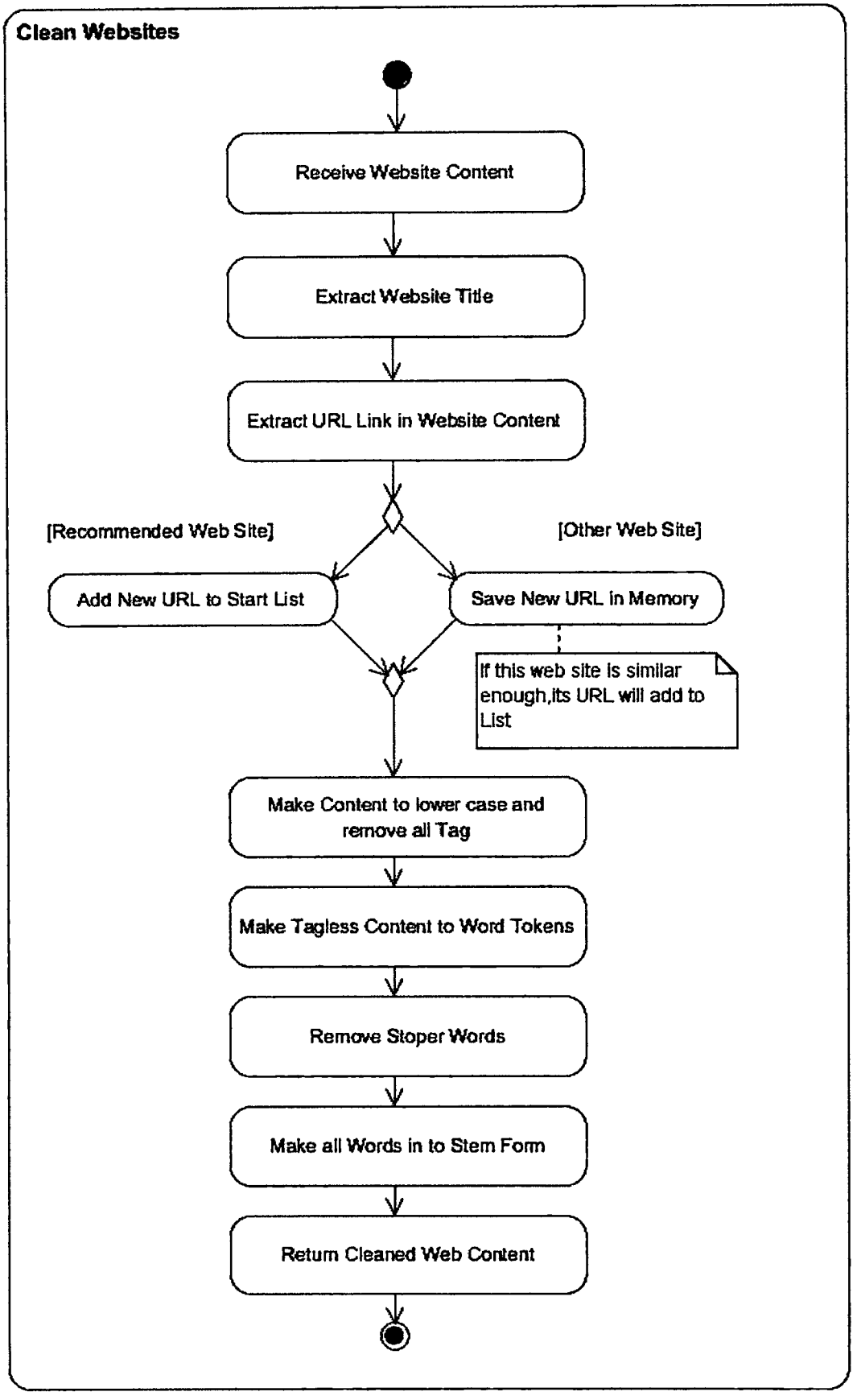


รูปที่ 4.14 ไดอะแกรมลำดับการทำงานของกิจกรรมของยูสเคส Crawl Websites

4.1.5 รายละเอียดยูสเคส Clean Websites

ตารางที่ 4.5 คำอธิบายยูสเคส Clean Websites

คำอธิบายลำดับกิจกรรมของยูสเคส Clean Websites
ขั้นตอนหลัก : 1. รับเนื้อหาของเว็บเข้ามา 2. ดึงหัวข้อเว็บไซต์ออกมาเพื่อนำไปใช้เป็นชื่อเว็บไซต์ 3. ดึง URL ในเนื้อหาเว็บออกมาให้หมด 4. ตรวจสอบประเภทของเว็บไซต์ปัจจุบัน 4.1. ถ้าเป็นเว็บที่ผู้ใช้เลือกให้บันทึก URL ทั้งหมดไว้ใน List ของ Crawler เลย 4.2. ถ้าไม่ใช่ให้เก็บไว้ชั่วคราว ถ้าเว็บไซต์ปัจจุบันมีความเหมือนเพียงพอจะทำการบันทึกเข้า List 5. เอา Tag Html ออกทั้งหมด และทำให้ตัวอักษรเป็นตัวเล็กทั้งหมด 6. ทำเนื้อหาที่เป็น String ต่อกันยาวๆ ให้กลายเป็น array of words 7. กำจัดคำที่อยู่ในข่ายของ Stop Word ออก โดยเราจะนำ array ของคำที่เป็น stop word ที่เราเตรียมไว้ไปลบกับ array ของคำทั้งหมดในเอกสารนั้น 8. ทำให้คำทั้งหมดอยู่ในรูปของรากศัพท์ ซึ่งทางผู้พัฒนาได้ไปโหลด Library จาก chuggnutt.com ซึ่งสามารถโหลดมาใช้งานได้ฟรี Library ตัวนี้รับค่าเป็น array of word แล้วจะส่งค่าคืนเป็น array of word ที่ทำ stemming แล้วกลับคืนมา 9. ส่งข้อมูลคำทั้งหมดกลับ
เงื่อนไขข้อยกเว้น : 1. ถ้าไม่มี URL อยู่ในเนื้อหาเลย a. ระบบจะทำงานเพียงการทำความสะอาดเนื้อหาเว็บเท่านั้น 2. ถ้าไม่มีเนื้อหาในเว็บเลย a. ระบบจะส่งคืน array of word ที่เป็นค่า ว่างกลับไป ในขั้นตอนต่อๆ ไป ค่าตรงนี้ก็จะทำให้เว็บดังกล่าวมีค่าความถี่ทุกคำเป็น 0 หมด สุดท้ายเว็บนี้ก็จะถูกพิจารณาว่าไม่ใช่เว็บที่มีความคล้ายคลึงกับเว็บที่ผู้ใช้เลือก แต่ถ้าหากเว็บนี้เป็นเว็บไซต์ที่ผู้ใช้เลือกระบบจะแจ้งว่าเว็บดังกล่าวไม่สามารถหาเว็บที่มีความคล้ายคลึงหรือเกี่ยวข้องกันได้

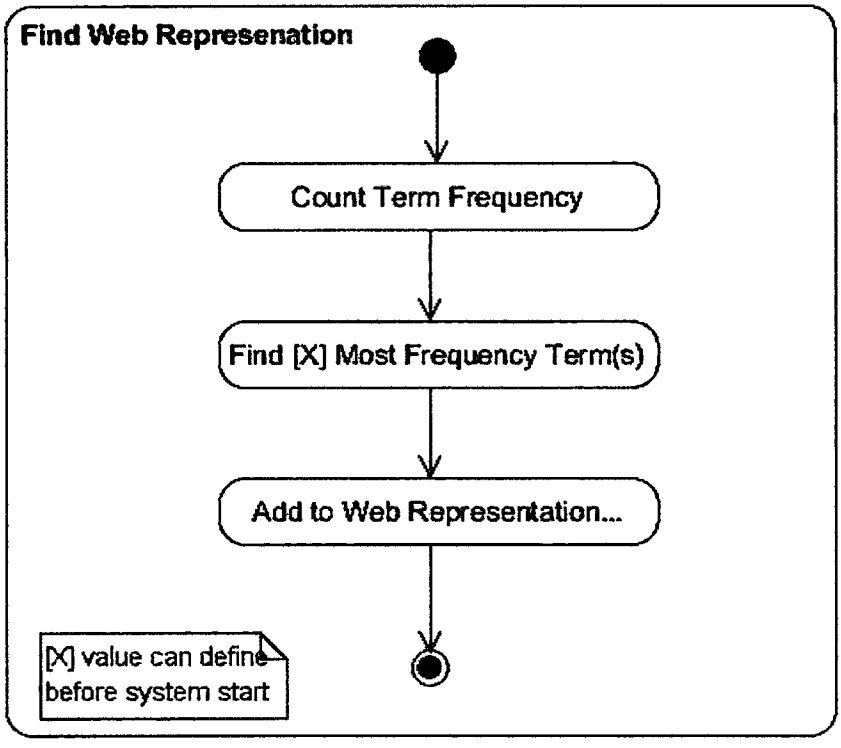


รูปที่ 4.15 ไดอะแกรมลำดับการทำงานของกิจกรรมของยูสเคส Clean Websites

4.1.6 รายละเอียดยูสเคส Find Web Representation

ตารางที่ 4.6 คำอธิบายยูสเคส Find Web Representation

คำอธิบายลำดับกิจกรรมของยูสเคส Find Web Representation
ขั้นตอนหลัก : 1. นับความถี่ของแต่ละคำในเนื้อหาของเว็บไซต์ดังกล่าว 2. นำคำที่ทั้งหมดที่นับความถี่แล้วมาเรียงจากคำที่มีความถี่มากที่สุด ไปยังคำที่มีความถี่น้อยที่สุด จากนั้นก็เลือกคำที่มีความถี่สูงสุดออกมา 10 คำ แต่ถ้ามีน้อยกว่า 10 คำ ก็เอาเท่าที่มี 3. เพิ่มคำที่เลือกดังกล่าวไปยังรายชื่อคำที่เป็นตัวแทนเอกสารเว็บ
เงื่อนไขข้อยกเว้น : 1. ถ้าเว็บไซต์ดังกล่าวไม่มีคำใดๆ อยู่เลย a. ระบบก็จะไม่เพิ่มคำที่เป็นตัวแทนเอกสารเว็บจากเนื้อหาเว็บไซต์ดังกล่าว 2. ถ้าความถี่ของคำมีค่าต่ำกว่าค่าที่กำหนด a. ระบบจะเพิ่มคำที่เหลือที่มีความถี่มากที่สุดตามเข้ามาจนครบตามจำนวนที่ระบบกำหนด 3. ถ้ามีคำที่เป็นตัวแทนเว็บไม่พอกับที่ระบบกำหนด a. ระบบจะทำการบันทึกคำที่เป็นตัวแทนเว็บเท่าที่มี เท่านั้น

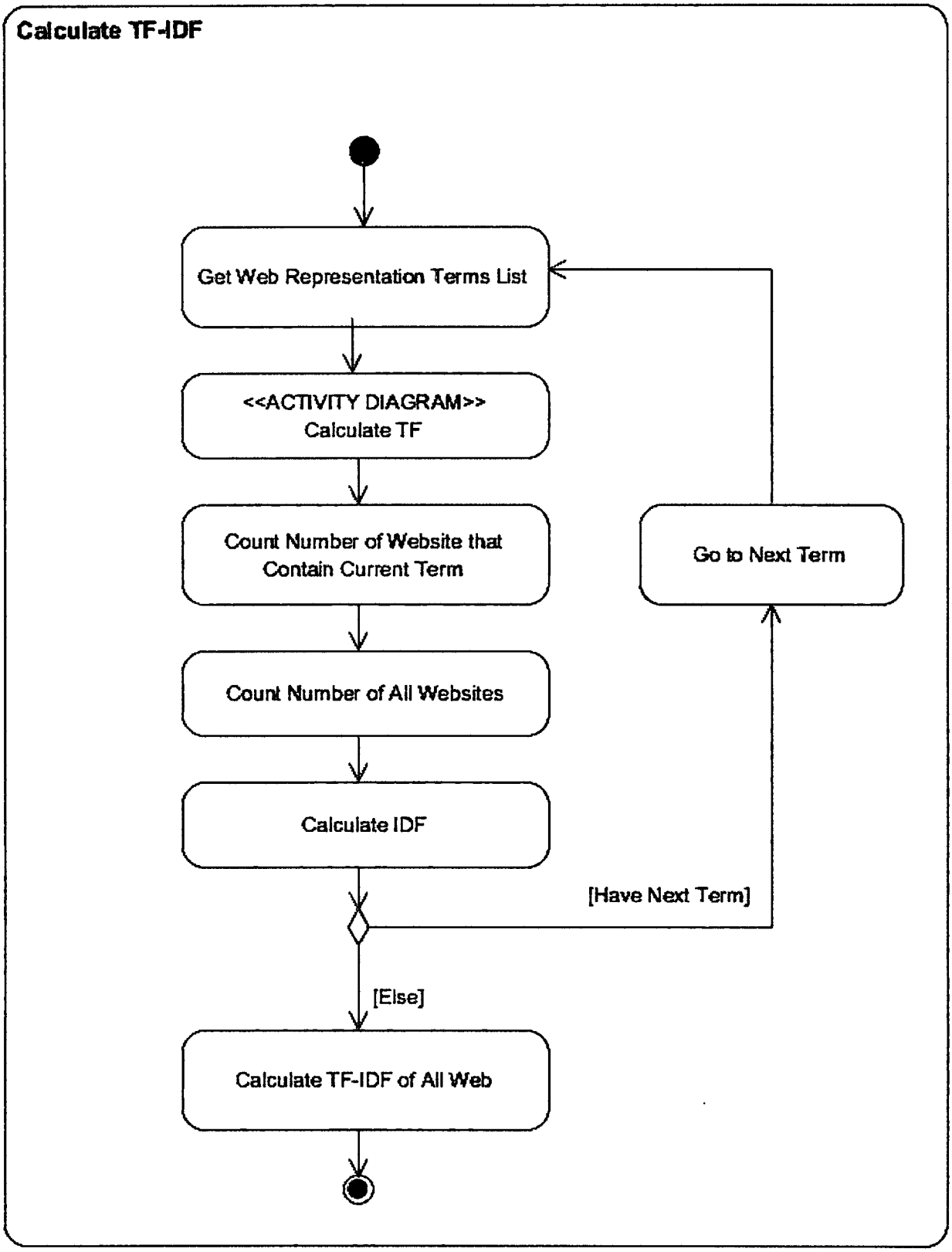


รูปที่ 4.16 ไดอะแกรมลำดับการทำงานของกิจกรรมของยูสเคส Find Web Representation

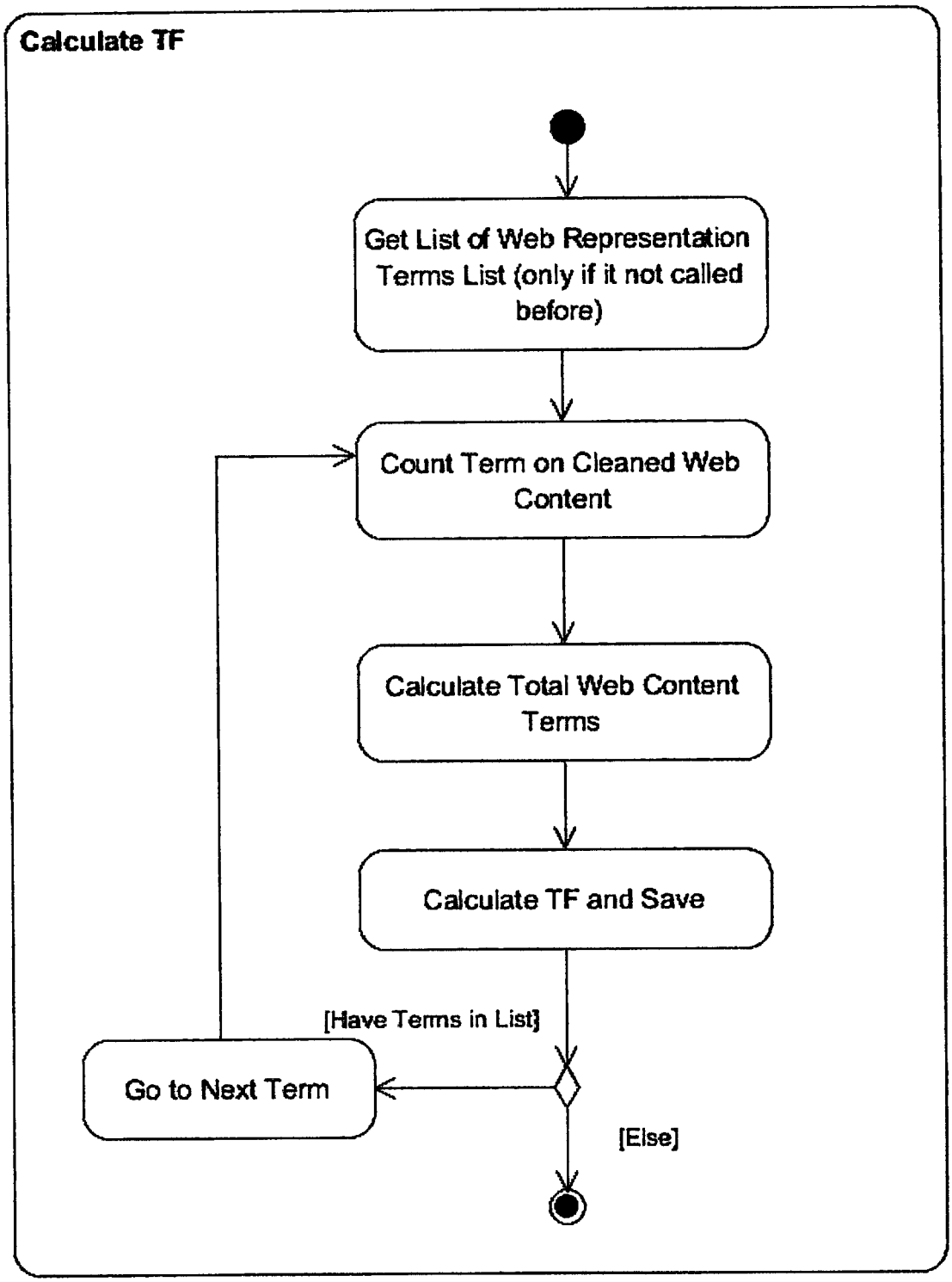
4.1.7 รายละเอียดยูสเคส Calculate TF-IDF

ตารางที่ 4.7 คำอธิบายยูสเคส Calculate TF-IDF

คำอธิบายลำดับกิจกรรมของยูสเคส Calculate TF-IDF
ขั้นตอนหลัก : 1. เรียกรายชื่อคำที่เป็นตัวแทนเอกสารเว็บทั้งหมด 2. คำนวณค่า TF (ไดอะแกรมลำดับกิจกรรมการทำงาน รูปที่ 4.17) 3. คำนวณหาเว็บที่มีคำที่เป็นตัวแทนเอกสารเว็บดังกล่าวอยู่ว่ามีกี่เว็บไซต์ 4. คำนวณหาจำนวนเว็บไซต์ทั้งหมด 5. หาค่า IDF 6. วนกลับไปทำข้อ 2 จนครบทุกคำของคำที่เป็นตัวแทนเอกสารเว็บทั้งหมด 7. คำนวณหาค่า TF-IDF ของทุกเว็บไซต์ที่ระบบหาข้อมูลมาได้ ยกเว้นเว็บไซต์ที่ผู้ใช้เลือก
เงื่อนไขข้อยกเว้น : 1. ถ้าไม่มีคำที่เป็นตัวแทนเอกสารเว็บอยู่เลย a. ระบบจะรายงานว่าเว็บไซต์ที่ผู้ใช้เลือกไม่สามารถหาเว็บไซต์ที่คล้ายคลึงหรือเกี่ยวข้องกันได้ 2. ถ้าเว็บไซต์ที่กำลังคำนวณเป็นเว็บที่ไม่สามารถเก็บข้อมูลได้ a. ระบบจะข้ามการคำนวณเว็บไซต์นั้นไป และเว็บไซต์นั้นจะไม่ถูกนำมาคิดคำนวณในด้านอื่นๆ อีก



รูปที่ 4.17 ไดอะแกรมลำดับการทำงานของกิจกรรมของยูสเคส Calculate TF-IDF

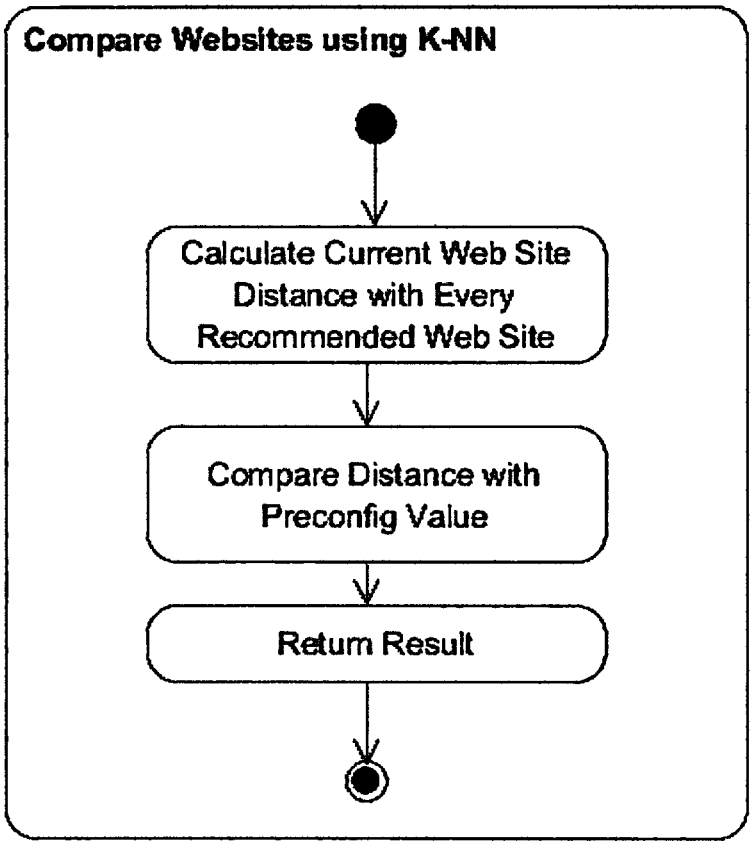


รูปที่ 4.18 ไดอะแกรมลำดับการทำงานของกิจกรรมย่อย Calculate TF ของไดอะแกรมลำดับกิจกรรมการทำงาน Calculate TF-IDF

4.1.8 รายละเอียดยูสเคส Compare Websites using K-NN Model

ตารางที่ 4.8 คำอธิบายยูสเคส Compare Websites using K-NN Model

คำอธิบายลำดับกิจกรรมของยูสเคส Compare Websites using K-NN Model
ขั้นตอนหลัก : 1. คำนวณหาค่าระยะห่างระหว่างเว็บไซต์ที่ต้องการจะเปรียบเทียบกับเว็บไซต์ที่ผู้ใช้เลือก โดยใช้ค่า TF มาคำนวณเป็นตำแหน่งของเว็บไซต์แต่ละเว็บ 2. นำระยะห่างที่ได้มาเปรียบเทียบกับค่าที่ตั้งไว้ถ้ามากกว่าแสดงว่าเว็บไซต์มีความคล้ายคลึงไม่พอ แต่ถ้าน้อยกว่าหรือเท่ากับแสดงว่าเว็บไซต์มีความคล้ายเพียงพอ 3. ส่งผลลัพธ์กลับเป็นค่า จริง ถ้าเหมือนพอ เป็น เท็จ ถ้าเหมือนไม่พอ
เงื่อนไขข้อยกเว้น : -



รูปที่ 4.19 ไดอะแกรมลำดับการทำงานของกิจกรรมของยูสเคส Compare Websites using K-NN Model

4.1.9 รายละเอียดยูสเคส Config Crawler Parameter

ตารางที่ 4.9 คำอธิบายยูสเคส Config Crawler Parameter

คำอธิบายลำดับกิจกรรมของยูสเคส Config Crawler Parameter
ขั้นตอนหลัก : 1. เรียกดูค่าต่างๆ ที่ระบบกำหนดสำหรับการทำงานของ Crawler 2. แก้ไขค่าต่างๆ ที่เกี่ยวกับการทำงานของ Crawler ซึ่งมีค่าที่สามารถแก้ไขได้คือ ความลึก ความกว้างในการค้นหาเว็บ จำนวนเว็บสูงสุดที่เก็บข้อมูลได้ จำนวนเทร็ดในการทำงาน ระยะเวลาในการทำงานมากที่สุดที่ทำงานได้ 3. บันทึกการตั้งค่าสู่ระบบ
เงื่อนไขข้อยกเว้น : 1. ถ้าระบบไม่สามารถบันทึกค่าใหม่ได้ a. ระบบจะใช้ค่าเก่าที่มีอยู่เดิม หรือค่าเริ่มต้นที่ระบบตั้งไว้

4.1.10 รายละเอียดยูสเคส Config Web Representation Parameter

ตารางที่ 4.10 คำอธิบายยูสเคส Config Web Representation Parameter

คำอธิบายลำดับกิจกรรมของยูสเคส Config Web Representation Parameter
ขั้นตอนหลัก : 1. เรียกดูค่าต่างๆ ที่ระบบกำหนดสำหรับการคำนวณหา Web Representation 2. แก้ไขค่าต่างๆ ที่เกี่ยวกับการคำนวณหา Web Representation ซึ่งมีค่าที่สามารถแก้ไขได้คือ จำนวนคำที่จะนำมาเป็นตัวแทนสูงสุดของแต่ละเว็บ และค่าความถี่ขั้นต่ำที่คำๆ นั้นจะสามารถนำมาเป็นคำตัวแทนเอกสารเว็บได้ 3. บันทึกการตั้งค่าสู่ระบบ
เงื่อนไขข้อยกเว้น : 1. ถ้าระบบไม่สามารถบันทึกค่าใหม่ได้ a. ระบบจะใช้ค่าเก่าที่มีอยู่เดิม หรือค่าเริ่มต้นที่ระบบตั้งไว้

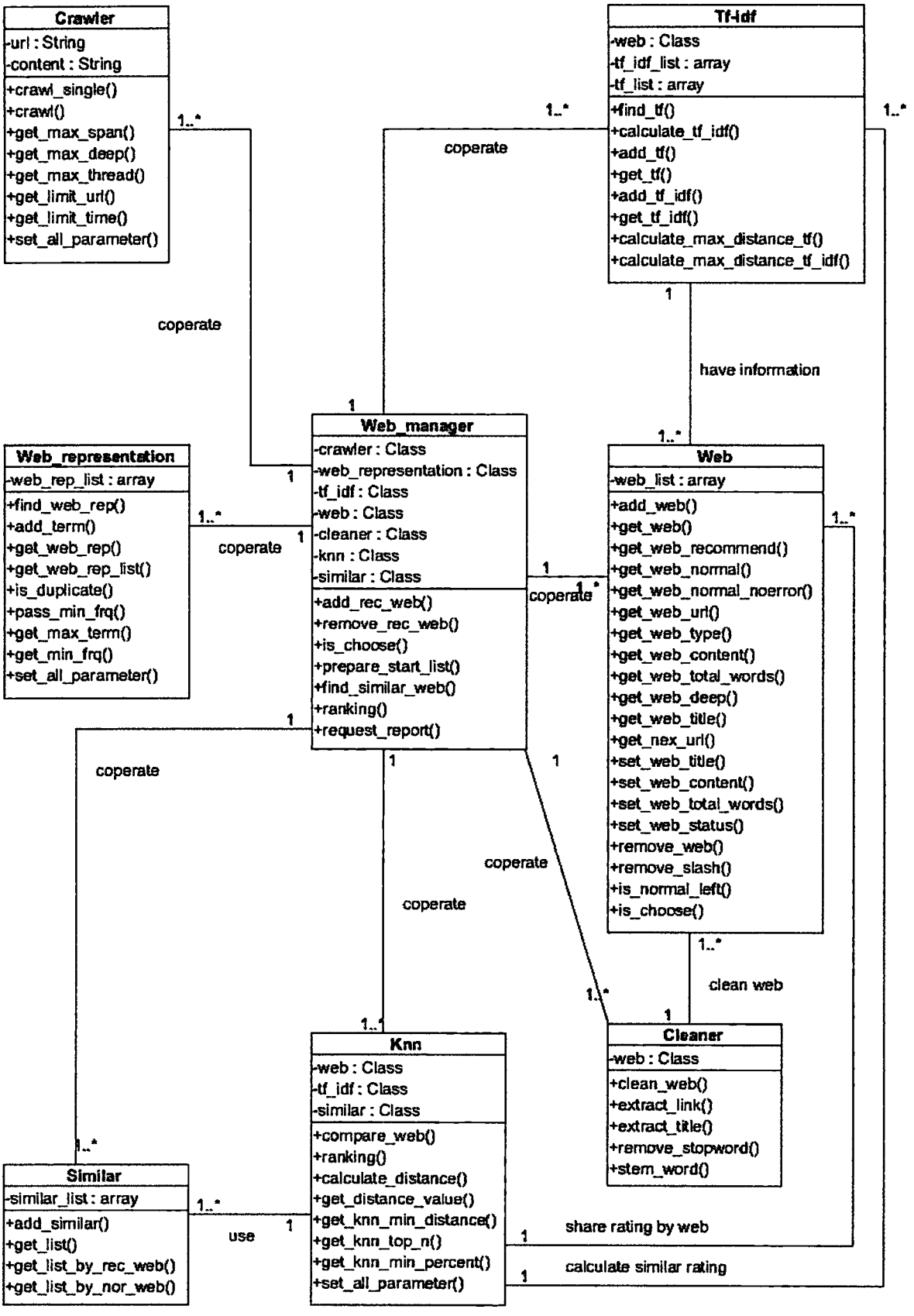
4.1.11 รายละเอียดยูสเคส Config K-NN Parameter

ตารางที่ 4.11 คำอธิบายยูสเคส Config K-NN Parameter

คำอธิบายลำดับกิจกรรมของยูสเคส Config K-NN Parameter
ขั้นตอนหลัก : 1. เรียกดูค่าต่างๆ ที่ระบบกำหนดสำหรับการทำงานของแบบจำลอง K-NN 2. แก้ไขค่าต่างๆ ที่เกี่ยวกับการทำงานของแบบจำลอง K-NN ซึ่งมีค่าที่สามารถแก้ไขได้คือ ระยะห่างมากที่สุดที่สามารถยอมรับได้ ค่าความเหมือนอย่างน้อยที่สุดที่จะนำเว็บไซต์นั้นมาแสดง จำนวนเว็บไซต์ที่จะนำมาแสดงสูงสุด 3. บันทึกการตั้งค่าสู่ระบบ
เงื่อนไขข้อยกเว้น : 1. ถ้าระบบไม่สามารถบันทึกค่าใหม่ได้ a. ระบบจะใช้ค่าเก่าที่มีอยู่เดิม หรือค่าเริ่มต้นที่ระบบตั้งไว้

4.2 คลาสไดอะแกรม

คลาสไดอะแกรมของระบบเราที่ได้ออกแบบมานี้เป็นคลาสไดอะแกรมในระดับ Data Model ซึ่งแสดงความสัมพันธ์ของข้อมูลต่างๆ ที่ระบบเก็บและนำมาใช้ในรูปแบบความสัมพันธ์ในแบบ OOP แสดงได้ดังรูปด้านในหน้าถัดไป



รูปที่ 4.20 คลาสไดอะแกรมระบบค้นหาเว็บแบบเจาะจงในระดับ Data Model

ตารางที่ 4.12 CRC ของ Class Web_manager

Class : Web_manager	
Responsibility :	**Collaborator :**
สร้างรายงานรายชื่อเว็บไซต์ที่คล้ายกับเว็บไซต์ที่ผู้ใช้เลือกเรียงตามลำดับความคล้ายคลึง	
บันทึกเว็บที่ผู้ใช้เลือกว่าตรงความต้องการ	Web
ลบเว็บที่ผู้ใช้เลือกว่าไม่ตรงความต้องการ	Web
ตรวจสอบว่าเว็บไซต์นี้ผู้ใช้เลือกไปหรือยัง	Web
เลือก URL ที่ Crawler ต้องไปเก็บข้อมูล	Web
บันทึก URL ที่อยู่ภายในเนื้อหาเว็บที่คล้ายกับเว็บที่ผู้ใช้เลือก	Web
ดึงเนื้อหาของเว็บไซต์ที่ผู้ใช้เลือก	Crawler
ให้ Crawler ไปเก็บข้อมูลเว็บปกติ	Crawler
เตรียมข้อมูลหน้าเว็บให้พร้อมสำหรับการประมวลผล	Cleaner
หาคำที่เป็นตัวแทนเว็บไซต์ที่ผู้ใช้เลือก	Web_representation
คำนวณค่า TF ของเว็บที่ผู้ใช้เลือก	Tf-idf
คำนวณค่า TF-IDF ของเว็บไซต์ที่มีอยู่ทั้งหมด	Tf-idf
เปรียบเทียบความเหมือนของเว็บที่หามาได้กับเว็บที่ผู้ใช้เลือก	Knn
เรียงลำดับความคล้ายคลึงของเว็บไซต์ที่มีทั้งหมดกับเว็บที่ผู้ใช้เลือก	Knn

ตารางที่ 4.13 CRC ของ Class Web

Class : Web	
Responsibility :	**Collaborator :**
บันทึกเว็บที่ผู้ใช้เลือก	
ลบเว็บที่ผู้ใช้เลือก	
ส่งข้อมูลเว็บตามรหัสเว็บที่ป้อนเข้ามา	
คัดเลือกและส่งรายชื่อ URL ที่ยังไม่ได้ไปเก็บข้อมูล	
ส่งข้อมูลเว็บทั้งหมดที่บันทึกอยู่ไปเป็น array	

ตารางที่ 4.14 CRC ของ Crawler

Class : Crawler	
Responsibility :	**Collaborator :**
ดึงเนื้อหาเว็บ 1 เว็บ	
ดึงเนื้อหาเว็บหลายเว็บ	

ตารางที่ 4.15 CRC ของ Class Cleaner

Class : Cleaner	
Responsibility :	**Collaborator :**
ทำความสะอาดและเตรียมข้อมูลเว็บ	
ถ้าเว็บนั้นเป็นเว็บที่ผู้ใช้เลือกก็จะบันทึกเว็บนั้นไว้ในระบบทันที	Web

ตารางที่ 4.16 CRC ของ Class Web_representation

Class : Web_representation	
Responsibility :	**Collaborator :**
บันทึกคำที่เป็นตัวแทนเว็บ	
ส่งรายชื่อคำที่เป็นตัวแทนเว็บ	
ค้นหาคำที่เป็นตัวแทนเว็บ	Web

ตารางที่ 4.17 CRC ของ Class Knn

Class : Knn	
Responsibility :	**Collaborator :**
คำนวณความคล้ายคลึงของเว็บไซต์โดยใช้ค่า TF-IDF	
ค้นคืนข้อมูลเว็บทั้งหมดที่ระบบหามาได้	Web
ดึงข้อมูลค่า TF-IDF	Tf-idf
บันทึกค่าคล้ายคลึงของแต่ละเว็บกับเว็บที่ผู้ใช้เลือกลงระบบ	Similar

ตารางที่ 4.18 CRC ของ Class Similar

Class : Similar	
Responsibility :	**Collaborator :**
บันทึกค่าความคล้ายคลึงของเว็บที่ระบบหามาได้กับเว็บที่ผู้ใช้เลือก	
ส่งคืนค่าความคล้ายคลึงของเว็บที่ระบบหามาได้กับเว็บที่ผู้ใช้เลือก เพื่อไปออกรายงาน	

บทที่ 5

การออกแบบส่วนติดต่อผู้ใช้และ

ซอฟต์แวร์สำหรับการพัฒนาระบบ

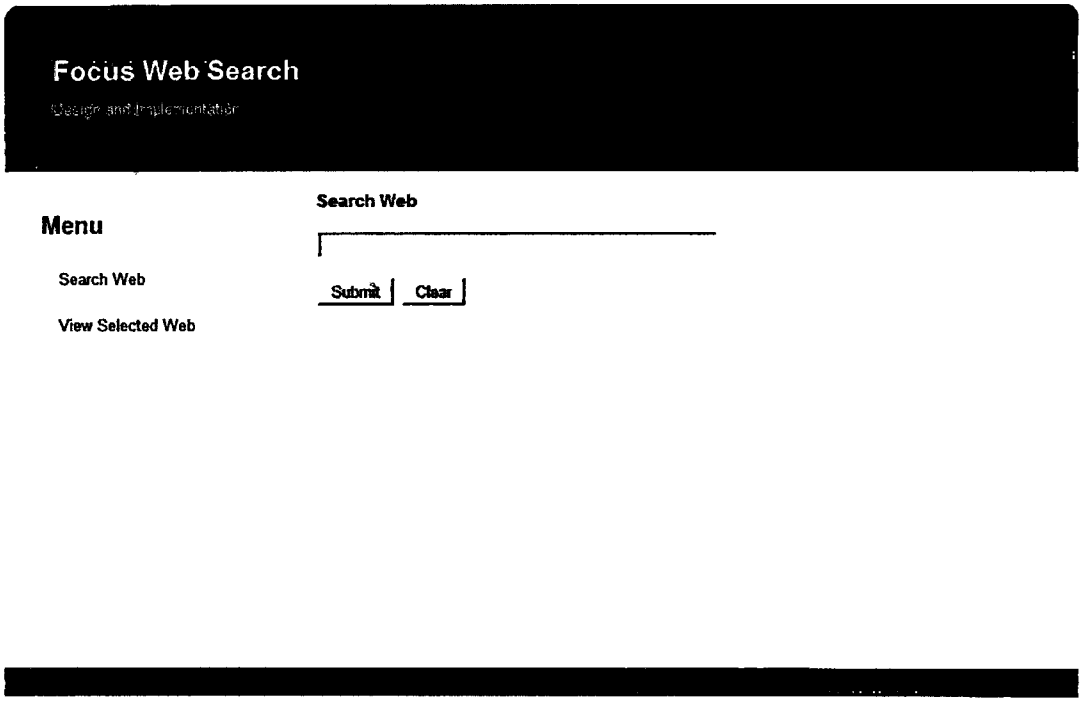
ในบทนี้เราจะแสดงให้เห็นถึงหน้าจอในส่วนที่ทำการรับข้อมูลและแสดงข้อมูลกลับไปให้ผู้ใช้งานในขั้นตอนต่างๆ และเราจะบอกถึงซอฟแวร์ต่างๆ ที่จำเป็นต่อการพัฒนาระบบของ ถ้าขาดซอฟแวร์เหล่านี้ระบบของเราก็ยากที่จะสำเร็จขึ้นมาได้

5.1 การออกแบบส่วนติดต่อผู้ใช้

ในส่วนนี้เราจะแสดงให้เห็นถึงหน้าจอที่แสดงผลต่อผู้ใช้ในขั้นตอนต่างๆ ที่ระบบทำงาน ตั้งแต่การค้นหาเว็บไซต์ การแสดงรายเว็บไซต์ที่ผู้ใช้เลือก การทำงานงานของระบบ และการแสดงผลในรูปแบบต่างๆ

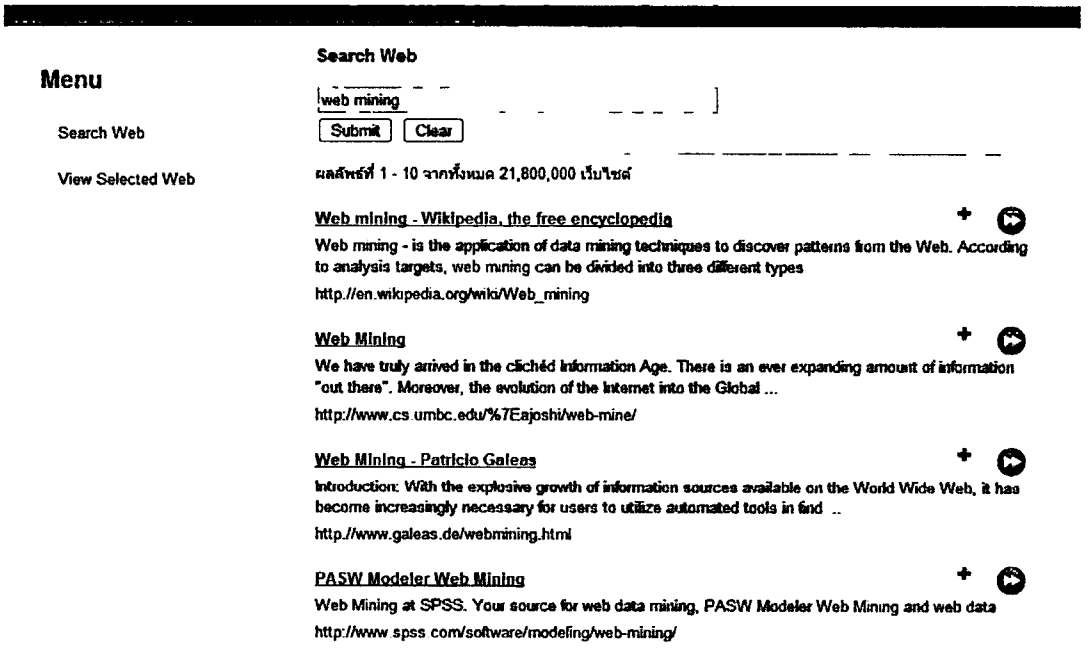
5.1.1 การค้นหา แสดงผล และเลือกเว็บไซต์

หน้าจอนี้แสดงถึงการค้นหาเว็บไซต์ผ่าน Bing API ของเรา การค้นหาสามารถทำได้โดยใส่คีย์เวิร์ดลงไปในช่องรับข้อความ จากนั้นก็กดปุ่ม "[Submit]" ระบบ ก็จะค้นหาเว็บไซต์ที่เกี่ยวข้องกับเว็บไซต์นั้นมาให้ สำหรับปุ่ม "[Cancel]" มีไว้สำหรับทำให้ช่องรับข้อความเป็นค่าว่าง ซึ่งแสดงอยู่ในรูป 5.1 ส่วนหน้าจอแสดงผลลัพธ์การค้นหาอยู่ที่รูป 5.2



รูปที่ 5.1 หน้าจอการค้นหาเว็บไซต์

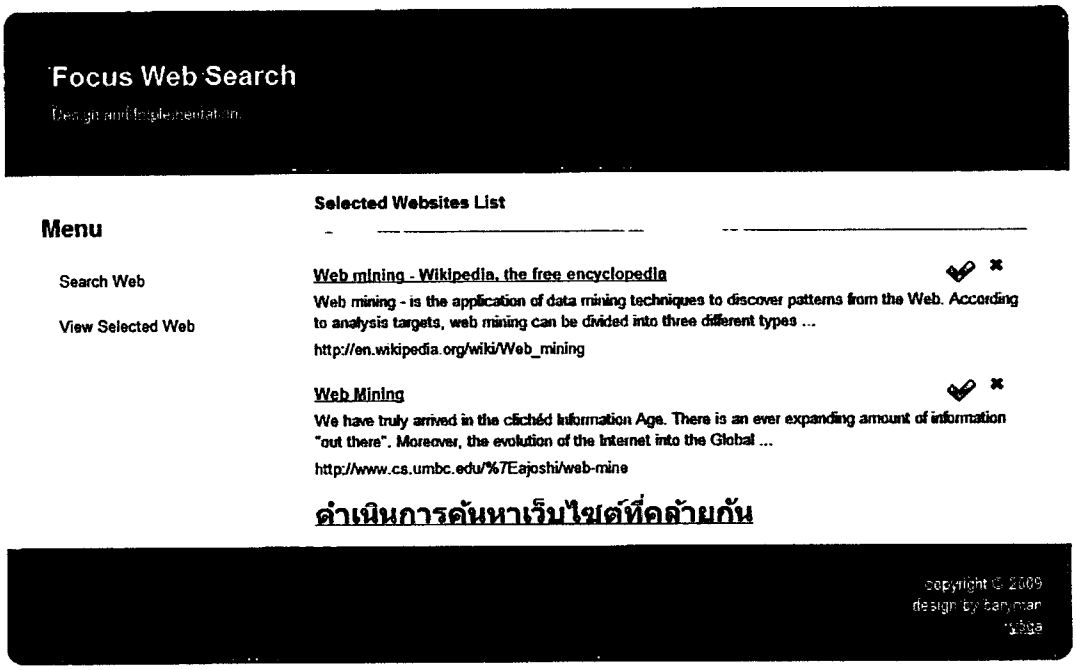
หน้าผลลัพธ์การค้นหาเว็บไซต์ ในแต่ละเว็บไซต์ที่ออกมาเป็นผลลัพธ์ผู้ใช้สามารถเลือกเว็บไซต์ที่ผู้ใช้เห็นว่าเป็นเว็บไซต์ที่ตรงความต้องการได้



รูปที่ 5.2 ผลลัพธ์การค้นหาเว็บไซต์

ในผลการค้นหาตามรูปที่ 5.2 ถ้าผู้ใช้ต้องการเลือกเว็บไซต์ใดให้เป็นเว็บไซต์ที่ตรงความต้องการ ก็ให้คลิกที่เครื่องหมาย "[บวก]" ถ้าต้องการให้เลือกเว็บไซต์ดังกล่าว และค้นหาทันทีให้คลิกที่เครื่องหมาย "[>>]" ซึ่งอยู่ข้างๆ กันในทุกๆ เว็บไซต์ที่ค้นหาได้

5.1.2 รายชื่อเว็บไซต์ที่ผู้ใช้เลือกและการยกเลิกการเลือกเว็บไซต์

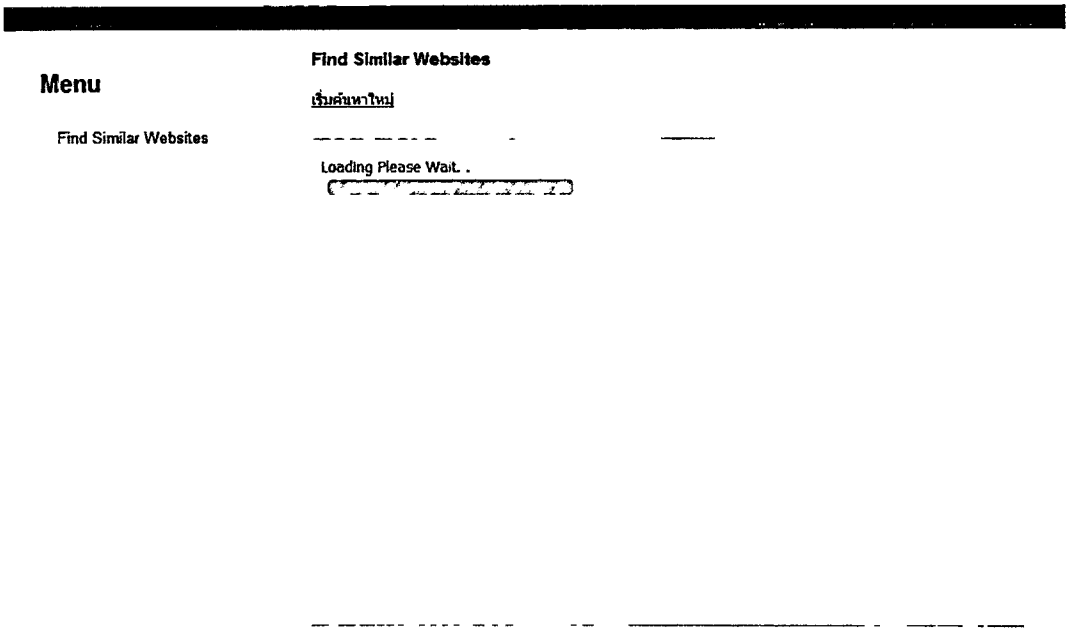


รูปที่ 5.3 เว็บไซต์ที่ผู้ใช้เลือก

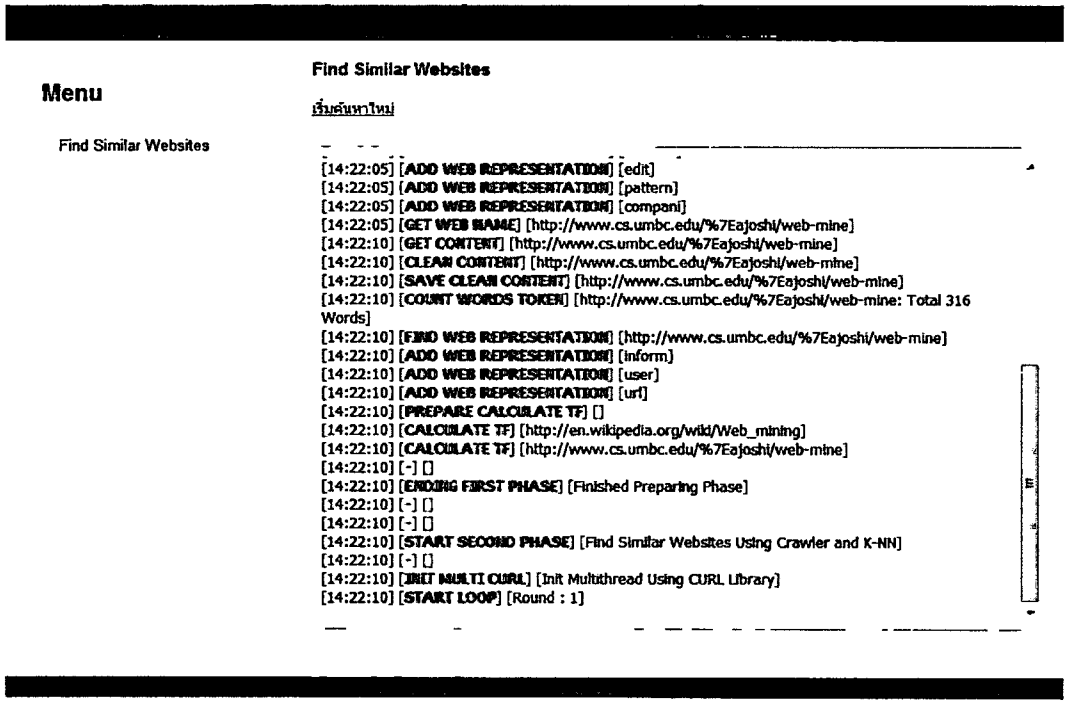
ในรูปที่ 5.3 หน้าจอแสดงเว็บไซต์ที่ผู้ใช้เลือก ในหน้านี้เราสามารถยกเลิกการเลือกเว็บไซต์ได้โดยการกดปุ่ม "[X]" ที่อยู่ทางด้านขวาบนของแต่ละรายชื่อเว็บไซต์ เมื่อเลือกเว็บไซต์เพียงพอแล้วให้คลิกที่ข้อความด้านล่างที่เขียนว่า "[ดำเนินการค้นหาเว็บไซต์ที่คล้ายกัน]" ระบบก็จะเริ่มทำการค้นหาเว็บไซต์ที่คล้ายกันทันที

5.1.3 การคำนวณความคล้ายกันของเว็บไซต์

สำหรับรูปในหัวข้อนี้มี 2 แบบ แบบแรกคือรูปที่ 5.4 เป็นการทำงานของระบบในแบบปกติ และในรูปที่ 5.5 เป็นการทำงานแบบในโหมด Debug เพื่อให้ผู้ใช้เห็นว่าในขณะนั้นระบบกำลังทำงานอะไรอยู่ ซึ่งส่วนใหญ่ในโหมด Debug นั้นจะใช้ในการตรวจสอบการทำงานของผู้ดูแลระบบเท่านั้น



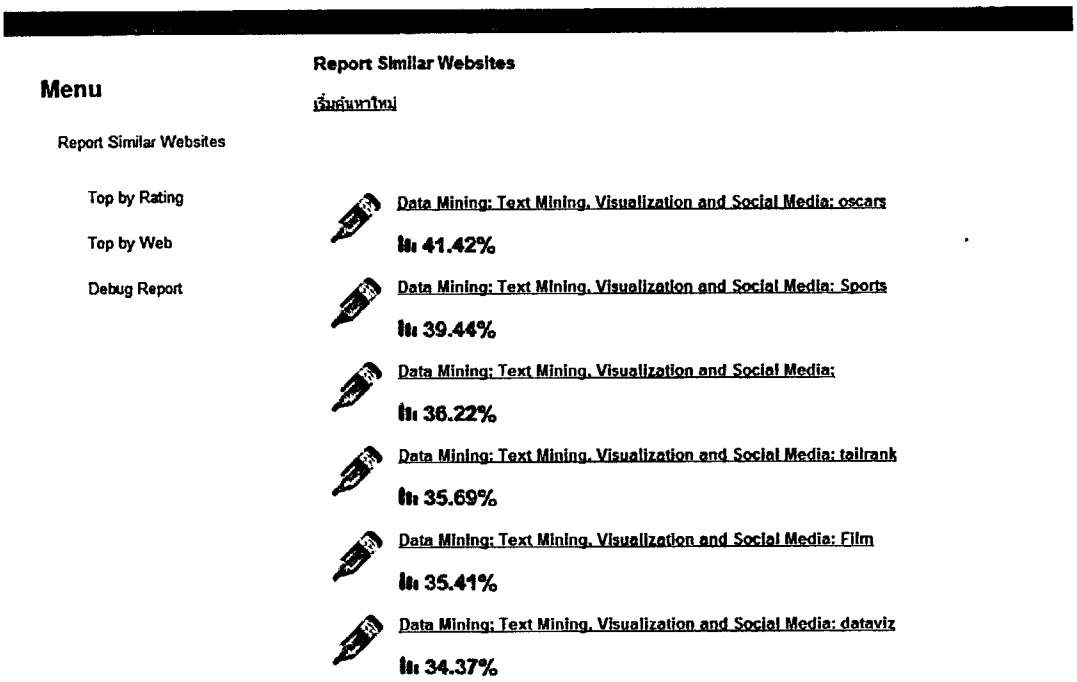
รูปที่ 5.4 การทำงานของระบบในแบบปกติ



รูปที่ 5.5 การทำงานในโหมด Debug

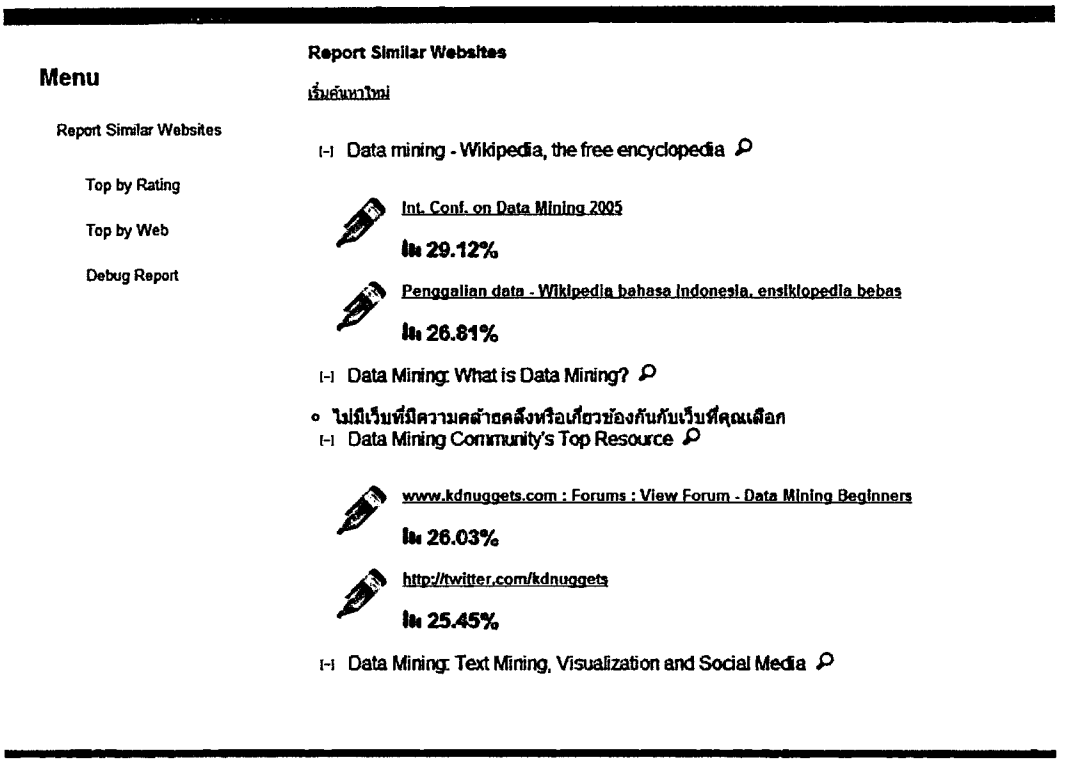
5.1.4 การแสดงผลลัพธ์เว็บไซต์ที่คล้ายกันในรูปแบบต่างๆ

การแสดงผลลัพธ์ของระบบของเรามี 3 แบบ แบบที่หนึ่งเป็นการแสดงผลลัพธ์โดยเรียงลำดับเว็บไซต์ที่ค้นหามาได้ตามความเหมือนของเว็บไซต์ดังกล่าวกับเว็บไซต์ต้นแบบที่ผู้ใช้เลือก



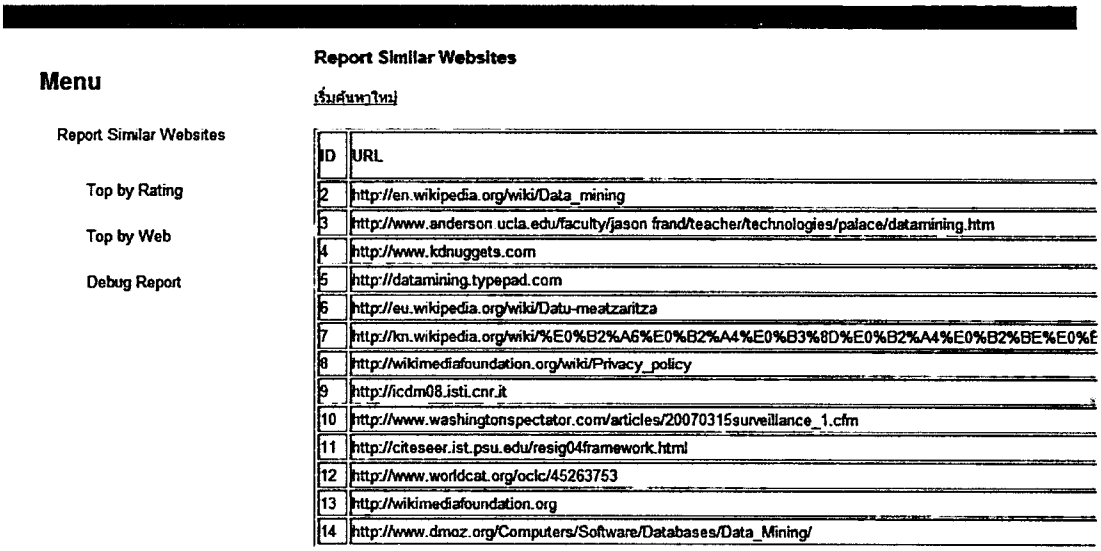
รูปที่ 5.6 แสดงผลลัพธ์ตามลำดับความเหมือนของเว็บไซต์

แบบที่สองนั้นจะเรียงผลลัพธ์ตามความเหมือนของเว็บไซต์ต่อเว็บไซต์ที่ผู้ใช้เลือกเช่นกัน แต่จะแยกตามเว็บไซต์ที่ผู้ใช้เลือกแต่ละเว็บไซต์ด้วย



รูปที่ 5.7 แสดงผลลัพธ์เรียงตามลำดับความเหมือนตามแต่ละเว็บไซต์ที่ผู้ใช้เลือก

แบบสุดท้ายแบบที่สาม แบบนี้จะใช้เฉพาะในโหมด Debug เท่านั้น แบบนี้มีไว้ให้ผู้ดูแลระบบใช้ดูผลลัพธ์การทำงานในขั้นตอนต่างๆ ของระบบเพื่อให้ทราบถึงค่าต่างๆ ที่ระบบเก็บคำนวณและประมวลผล เพื่อหาจุดที่ทำงานผิดพลาดของระบบ



รูปที่ 5.8 รายชื่อเว็บไซต์ในการแสดงผลลัพธ์ในโหมด Debug

TERM ID	TERM VALUE	TERM IDF
0	data	0.17327172127404
1	mine	0.2578291093021
2	pattern	1.4816045409242
3	inform	0.52609309589678
4	discoveri	1.145132304303
5	knowledg	0.8938178760221
6	edit	1.3862943611199
7	analysi	0.73966719619484
8	set	0.73966719619484
9	applic	0.73966719619484
10	process	1.3862943611199
11	analyz	2.1747517214842
12	relationship	1.9924301646902
13	softwar	0.45198512374306
14	system	0.8938178760221
15	kdnugget	1.8382794848629
16	analyt	1.7047480922384
17	latest	1.9924301646902
18	new	0.64869541798911
19	jul	3.7841896339183
20	dec	0

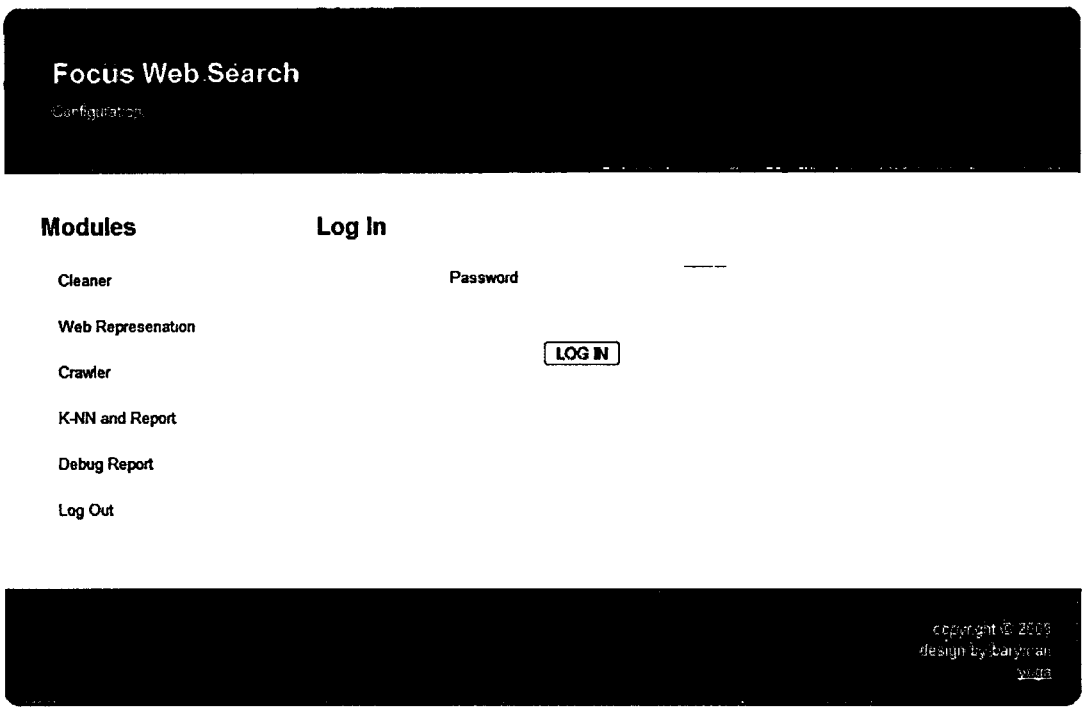
รูปที่ 5.9 รายชื่อคำที่เป็นตัวแทนเอกสารเว็บในการแสดงผลลัพธ์ในโหมด Debug

WEB KEY	2	3	4	5	ENGAGE
6	0.081204	0.089578	0.077345	0.042647	FAILED
7	0.059191	0.070197	0.062707	0.040267	FAILED
8	0.076825	0.082642	0.080012	0.045457	FAILED
9	0.048835	0.059607	0.055269	0.037390	FAILED
10	0.087537	0.095268	0.084732	0.051520	FAILED
11	0.083851	0.091297	0.080208	0.047247	FAILED
12	0.087244	0.095743	0.083470	0.043024	FAILED
13	0.083860	0.091333	0.079201	0.038082	FAILED
14	0.040460	0.045973	0.069190	0.088345	FAILED
15	0.083613	0.086608	0.085247	0.055988	FAILED
16	0.043258	0.053366	0.053298	0.042655	FAILED
17	0.085792	0.093424	0.081908	0.044130	FAILED
18	0.085925	0.093404	0.081246	0.040298	FAILED
19	0.088522	0.094646	0.088107	0.050730	FAILED
20	0.051957	0.060916	0.057091	0.039913	FAILED
21	0.049717	0.061603	0.056121	0.044555	FAILED
22	0.086802	0.094087	0.082777	0.049587	FAILED
23	0.047354	0.055998	0.055898	0.042432	FAILED
24	0.100406	0.106812	0.088743	0.067965	FAILED
25	0.050425	0.064925	0.057966	0.049929	FAILED

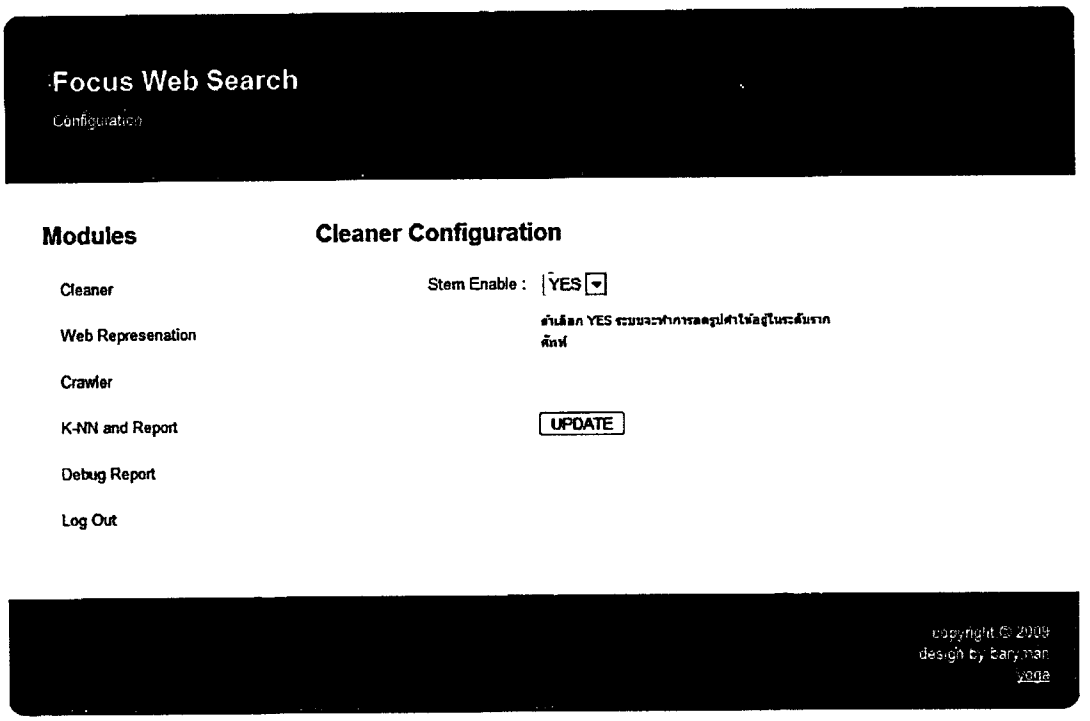
รูปที่ 5.10 แสดงระยะห่างระหว่างเว็บไซต์ที่ค้นหาได้กับเว็บไซต์ที่ผู้ใช้เลือกในการแสดงผลลัพธ์ในโหมด Debug

5.1.5 การตั้งค่าระบบสำหรับผู้ดูแลระบบ

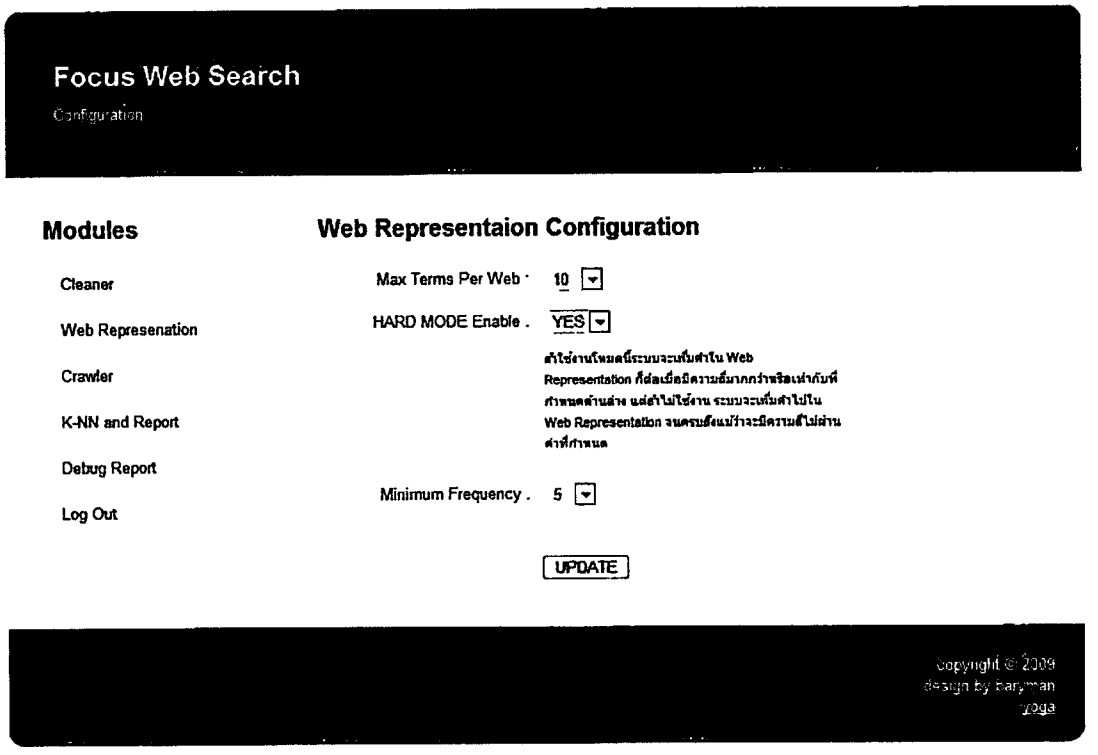
ส่วนนี้เป็นหน้าจอสำหรับให้ผู้ดูแลระบบตั้งค่าการทำงานให้กับระบบ เพื่อให้เหมาะกับสภาพแวดล้อมที่ระบบทำงานอยู่ รวมถึงความเหมาะสมในการแสดงผลลัพธ์ต่อผู้ใช้งาน มีอยู่ 5 หน้าจอ คือรูปที่ 5.11 เป็นหน้าจอการล็อกอินของผู้ใช้ระบบ รูปที่ 5.12 เป็นการตั้งค่าการทำงานของการเตรียมข้อมูล รูปที่ 5.13 เป็นการตั้งค่าการทำงานการค้นหาคำที่เป็นตัวแทนเว็บไซต์ รูปที่ 5.14 เป็นการตั้งค่าการทำงานของ Crawler รูปที่ 5.15 เป็นการตั้งค่าการทำงานของแบบจำลอง K-NN และการแสดงผลลัพธ์ต่อผู้ใช้ รูปที่ 5.16 เป็นการตั้งค่าการแสดงผลในแบบโหมด Debug



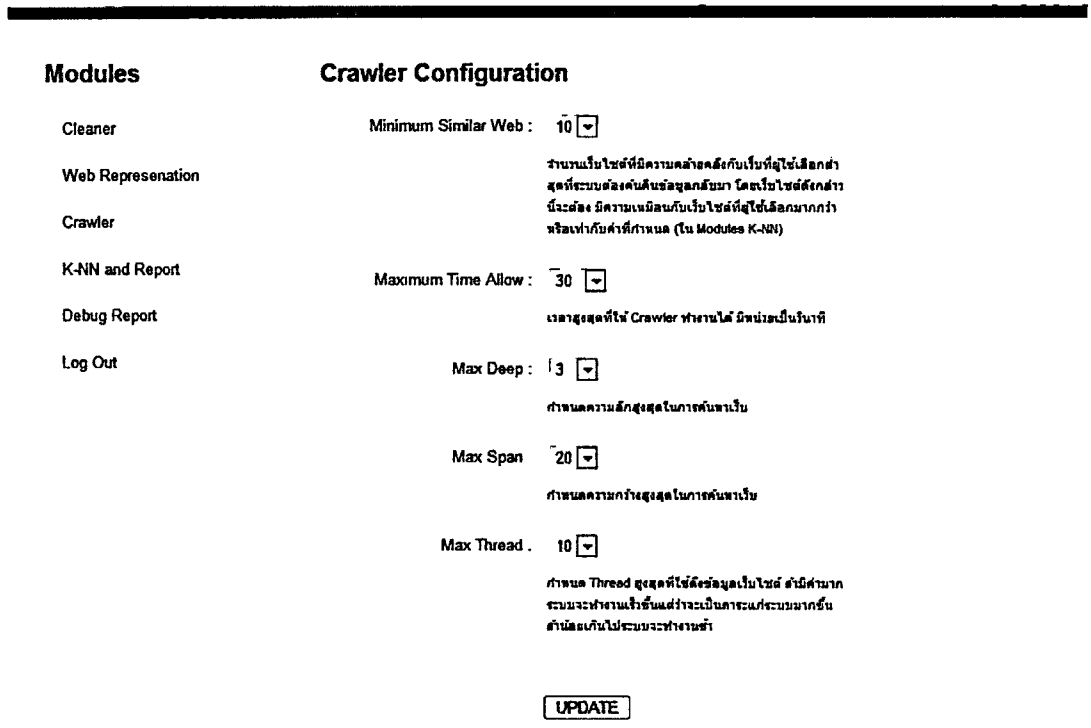
รูปที่ 5.11 การเข้าสู่การตั้งค่าระบบของผู้ดูแลระบบ



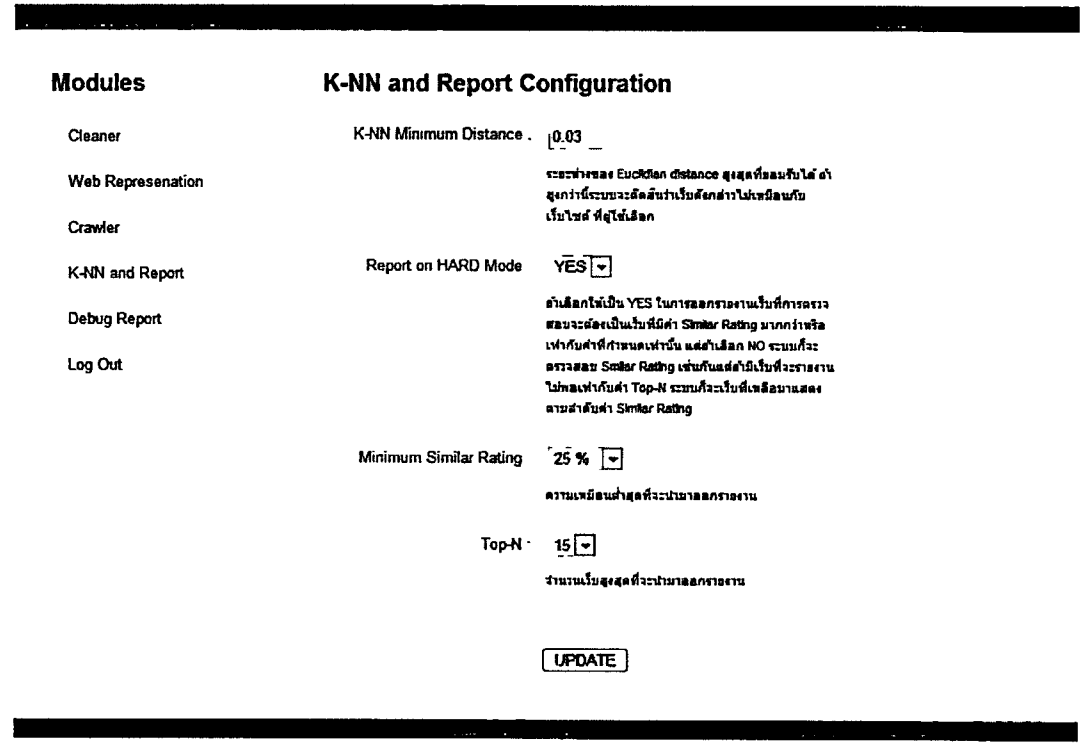
รูปที่ 5.12 หน้าจอตั้งค่าการเตรียมข้อมูลเว็บไซต์



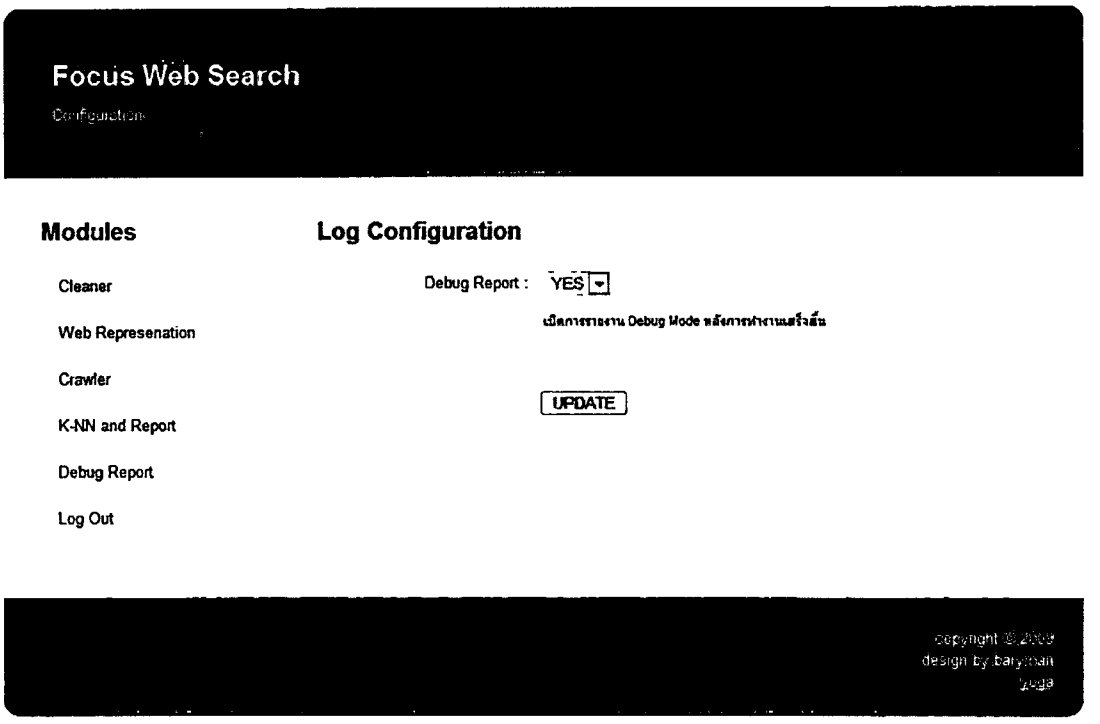
รูปที่ 5.13 หน้าจอการตั้งค่าการหาตัวแทนเว็บไซต์



รูปที่ 5.14 หน้าจอการตั้งค่าการทำงานของ Crawler



รูปที่ 5.15 หน้าจอการตั้งค่าการทำงานของแบบจำลอง K-NN และการแสดงผลลัพธ์ต่อผู้ใช้



รูปที่ 5.16 หน้าจอการตั้งค่าการแสดงผลในโหมด Debug

5.2 ซอฟแวร์สำหรับการพัฒนาระบบ

ในการพัฒนาระบบค้นหาเว็บไซต์แบบเจาะจงของเรานั้นจำเป็นจะต้องใช้ซอฟแวร์หลายตัวในหลายๆ ขั้นตอนของการพัฒนาระบบ ตั้งแต่การออกแบบ การพัฒนา และการใช้งานจริง ซึ่งแต่ละขั้นตอนก็ต้องการซอฟแวร์เหมาะสมสำหรับแต่ละขั้นตอน

5.2.1 ซอฟแวร์สำหรับการทำให้ระบบทำงานได้

- Apache Web Server [Version 2.2.12]
    - สำหรับใช้ในการพัฒนาและใช้งานจริงของระบบ เพื่อให้ระบบทำงานในแบบ Web Application ได้
- PHP [Version 5.3.0]
    - จำเป็นสำหรับการพัฒนาในแบบ Web Application เป็นภาษาที่เหมาะที่สุดภาษาหนึ่งที่ใช้ทำ Web Application ในปัจจุบัน
- MySQL [Version 5.1.37]
    - สำหรับใช้เก็บการตั้งค่าการทำงานของระบบต่างๆ

- ทางเลือกอื่นๆ สำหรับซอฟแวร์ตัวนี้
  - File System
  - SQLite

5.2.2 ซอฟแวร์สำหรับการออกแบบระบบ

- Microsoft Visio 2007
  - สำหรับการวาดไดอะแกรมต่างๆ ซึ่งสามารถโหลดมาใช้ได้ฟรีสำหรับนักศึกษาสถาบันเทคโนโลยีเจ้าคุณทหารลาดกระบัง
- Visual Paradigm for UML 7.1 Community Edition
  - สำหรับวาดไดอะแกรม UML ทั้งหมด เป็นเครื่องที่วาด UML ไดอะแกรมได้ง่ายมาก สำหรับในอิดิชั่นคอมมูนิตี้นั้นทางผู้ขายให้ใช้งานฟรีด้วย แต่ถ้าจะปริ้นตัวรูปจะติด Watermark ออกมาด้วย
  - ทางเลือกอื่นๆ สำหรับซอฟแวร์ตัวนี้
    - Violet UML (Freeware)
    - ArgoUML (Freeware)
    - Star UML (Freeware)
- Microsoft Office
  - สำหรับตัวนี้ใช้ในการจัดทำเอกสารต่างๆ ออกมาเป็นรูปเล่ม
  - ทางเลือกอื่นๆ สำหรับซอฟแวร์ตัวนี้
    - Open Office (Freeware)
    - Google DOC (Freeware, ต้องการการเชื่อมต่ออินเตอร์เน็ต)

5.2.3 ซอฟแวร์สำหรับการพัฒนาระบบ

- Netbean IDE [Version 6.7.1]
  - สำหรับใช้ในการเขียนโค้ดภาษา PHP ซึ่ง Netbean นั้นสามารถรองรับการพัฒนาได้หลายภาษามาก เช่น Java, Python, Ruby โดยต้องดาวโหลดปลั๊กอินเข้ามาเพิ่ม แต่ข้อเสียของซอฟแวร์ตัวนี้คือ ในช่วงเริ่มทำงานจะช้ามาก แต่ข้อดีคือฟรี
  - ทางเลือกอื่นๆ สำหรับซอฟแวร์นี้

- Aptana Studio (Freeware)
- Notepad++ (Freeware)
- Microsoft Expression Web 3 (Free for IT KMITL)
- Dreamweaver CS 4 (Commercial)

- phpMyAdmin [Version 3.2.0.1]
  - สำหรับสร้างและจัดการฐานข้อมูล MySql ของการพัฒนาระบบของเรา ข้อดีคือฟรี และสามารถทำงานบนเว็บเบราเซอร์ได้
  - ทางเลือกอื่นๆ สำหรับซอฟแวร์นี้
    - Mysql Front (Commercial)
    - Navicate (Commercial)
    - Toad for Mysql (Commercial)

บทที่ 6

การทดสอบและการเปรียบเทียบการทำงาน

เราจะทำการเปรียบเทียบผลการค้นหาเว็บของระบบค้นหาเว็บแบบเจาะจงซึ่งเป็นผลงานในการศึกษานี้ โดยเราจะใช้การวิธีการคำนวณค่า IDF และวิธีการคำนวณความคล้ายของเว็บไซต์ที่ต่างกัน ส่วนพารามิเตอร์ที่เหลือให้ตั้งไว้เหมือนกันทุกการทดสอบ

6.1 การตั้งค่าการทำงาน

ค่าพารามิเตอร์ที่เราต้องตั้งให้คงที่ตลอดทุกการทดลองมีดังนี้

- จำนวนคำที่เป็นตัวแทนเอกสารเว็บ 15 คำ
- จำนวน URL ที่นำมาใช้งานมากที่สุดต่อ 1 เว็บ : 30 URL
- จำนวนชั้นความลึกในการค้นหาเว็บของ Crawler : 3 ชั้น

ค่าพารามิเตอร์ที่แปรเปลี่ยนไปตามแต่ละการทดลองมีดังนี้

- การคำนวณค่า IDF เราจะใช้ 2 ค่าคือ Ln กับ Log_{10} กับทั้งสองระบบ
- การเปรียบเทียบความเหมือนของเว็บเราจะใช้ K-NN กับ Cosine Similarity (CS)

6.2 คำและเว็บที่จะนำมาทดสอบเพื่อเปรียบเทียบระบบ

คำที่เรานำมาใช้ค้นหานั้นจะเป็นคำนามที่กำลังได้รับความนิยมอยู่ในปัจจุบัน กับคำนามที่เกี่ยวกับข้องกับเทคโนโลยีสารสนเทศ ที่เราเลือกคำนามเพราะว่า คำนามมักเป็นคำที่มีโอกาสที่จะมีความนิยมสูงมากกว่าคำกิริยาหรือคำประเภทอื่นๆ และเพื่อลดปัจจัยที่จะทำให้ผลลัพธ์แปรปรวนเราจึงกำหนดอีกว่าให้เลือกเว็บที่ได้จากการค้นหาด้วยคำๆ นั้นด้วย ซึ่งมีคำและเว็บต่างๆ ดังนี้

- ipad
    - http://www.apple.com/ipad/design/
- 4g
    - http://www.networkworld.com/news/2007/052107-special-focus-4g.html

- intel
    - http://en.wikipedia.org/wiki/Intel
- content management system
    - http://webdesign.about.com/od/contentmanagement/a/aa031300a.htm
- firefox
    - http://www.mozilla.com/en-US/firefox/firefox.html
- software engineering
    - http://www.answers.com/topic/software-engineering
- data mining
    - http://www.kdnuggets.com/
- information security
    - http://searchsecurity.techtarget.com/
- avatar
    - http://www.comingsoon.net/news/videonews.php?id=60735
- raid
    - http://en.wikipedia.org/wiki/RAID

6.3 ผลการทำงาน

เราจะแสดงผลการทดสอบระบบตามคีย์เวิร์ดที่แสดงไว้ข้างต้น โดยแต่ละคีย์เวิร์ดจะมี 4 การทดลองด้วยกัน ซึ่งจะต่างกันตรงวิธีการคำนวณค่า IDF และการคำนวณความคล้ายคลึงของเว็บไซต์ และตามด้วยความคิดเห็นของผู้ที่ทดลองใช้โปรแกรม โดยแต่ละคอลัมน์ของผลการทดสอบมีความหมายดังนี้

- คอลัมน์ที่ 1 รายชื่อเว็บ – เป็นรายชื่อเว็บไซต์ที่ปรากฏผลตามการทดลองแต่ละครั้ง โดยรายชื่อเว็บนี้ได้ถูกเรียงตามความคล้ายกับเว็บที่ผู้ใช้เลือกจากมากสุดไปน้อยสุดในทิศทางบนลงล่าง โดยจะแสดงเป็น URL ของเว็บไซต์นั้นๆ
- คอลัมน์ที่ 2 ปรากฏผลในการทดลองอื่น – ในการทดสอบแต่ละคีย์เวิร์ดเราจะทดสอบทั้งหมด 4 ครั้งตามการแปรผันของวิธีการคำนวณค่า IDF และวิธีการเปรียบเทียบความ

คล้ายคลึงกับเว็บไซต์ที่ผู้ใช้เลือก ซึ่งถ้าเว็บไซต์อาจจะมีการปรากฏผลในการทดลองแต่ละครั้งไม่เหมือนกัน ซึ่งมีค่าต่างๆ ดังนี้

- 1 – ปรากฏแค่การทดสอบเดียวในการทดสอบคีย์เวิร์ดนั้น
- 2 – ปรากฏสองการทดสอบในการทดสอบคีย์เวิร์ดนั้น
- 3 – ปรากฏสามการทดสอบในการทดสอบคีย์เวิร์ดนั้น
- 4 – ปรากฏทั้งหมดสี่การทดสอบในการทดสอบคีย์เวิร์ดนั้น

- คอลัมน์ที่ 3 ผู้ใช้พอใจ – ถ้าผู้ใช้ดูเนื้อเว็บแล้วปรากฏว่าเว็บนี้มีเนื้อหาตรงความต้องการของผู้ใช้ คอลัมน์นี้จะมีค่า FAV แต่ถ้าผู้ใช้รู้สึกว่าเนื้อหาไม่ตรงความต้องการก็จะปล่อยว่างไว้

6.3.1 ผลการทดสอบกับคีย์เวิร์ด ipad

ตารางที่ 6.1 ผลการทดสอบด้วยคีย์เวิร์ด ipad ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://store.apple.com/us/browse/home/shop_ipad/family/ipad	3	
http://store.apple.com/us_smb_78313?cid=AOSA10000022131	4	
http://store.apple.com/us	4	
http://www.apple.com/ipad/features/	2	FAV
http://www.apple.com/retail/geniusbar/	1	FAV

ตารางที่ 6.2 ผลการทดสอบด้วยคีย์เวิร์ด ipad ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://store.apple.com/us/browse/home/shop_ipad/family/ipad	3	
http://www.apple.com/ipad/features/	2	FAV
http://store.apple.com/us_smb_78313?cid=AOSA10000022131	4	
http://store.apple.com/us	4	
http://www.apple.com/ipad/	2	FAV

ตารางที่ 6.3 ผลการทดสอบด้วยคีย์เวิร์ด ipad ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://store.apple.com/us/browse/home/shop_ipad/family/ipad	3	
http://store.apple.com/us_smb_78313?cid=AOSA10000022131	4	
http://store.apple.com/us	4	
http://www.apple.com/ipad/app-store/	1	FAV
http://www.apple.com/ipad/3g/	1	FAV

ตารางที่ 6.4 ผลการทดสอบด้วยคีย์เวิร์ด ipad ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.apple.com/ipad/	2	FAV
http://store.apple.com/us_smb_78313?cid=AOSA10000022131	4	
http://store.apple.com/us	4	
http://www.apple.com/ipad/features/ibooks.html	1	FAV
http://developer.apple.com/	1	FAV

การทดสอบกับคีย์เวิร์ด ipad มีชื่อเว็บไซต์ที่ปรากฏเหมือนกันทั้งสี่การทดสอบอยู่ 2 ซึ่งเป็นเว็บเกี่ยวกับการขายสินค้าซึ่งแสดงว่าหน้าเว็บนั้นมีเนื้อหาที่เกี่ยวข้องอย่างชัดเจนอยู่แต่ว่าผู้ใช้ไม่ได้เลือกอาจจะเป็นเพราะว่าผู้ใช้อาจจะไม่ได้มีจุดประสงค์ที่เข้าเว็บมาเพื่อซื้อสินค้าแต่มาเพื่อหาข้อมูล ซึ่งดูได้จากลักษณะของเนื้อหาของเว็บที่ผู้เลือกว่าเป็นที่พอใจ

6.3.2 ผลการทดสอบกับคีย์เวิร์ด 4g

ตารางที่ 6.5 ผลการทดสอบด้วยคีย์เวิร์ด 4g ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.networkworld.com/news/2006/081406-sprint-nextel.html	3	FAV
http://www.networkworld.com/news/2009/010609-sling-media-plans-iphone-client.html	1	
http://www.networkworld.com/news/2010/031510-virtual-server-security.html?t51hb	2	
http://www.networkworld.com/newsletters/converg/2009/031609converge1.html	1	FAV
http://www.networkworld.com/news/2010/031510-apple-eu-stores-closed-ipad.html?t51hb	2	

ตารางที่ 6.6 ผลการทดสอบด้วยคีย์เวิร์ด 4g ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.networkworld.com/news/2010/012610-cisco-netapp-vmware-security.html	1	
http://www.networkworld.com/news/2006/081406-sprint-nextel.html	3	FAV
http://www.networkworld.com/news/2010/031210-cisco-chambers-compete.html?t51hb	2	
http://www.networkworld.com/news/2010/031510-apple-eu-stores-closed-ipad.html?t51hb	2	
http://www.networkworld.com/news/2010/031510-google-launches-street-view-in.html?t51hb	2	

ตารางที่ 6.7 ผลการทดสอบด้วยคีย์เวิร์ด 4g ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.networkworld.com/news/2010/031210-cisco-chambers-compete.html?t51hb	2	
http://www.networkworld.com/news/2006/081406-sprint-nextel.html	3	FAV
http://www.networkworld.com/news/2010/031510-google-launches-street-view-in.html?t51hb	2	
http://www.networkworld.com/news/2010/031510-windows-phone-apps.html?t51hb	1	
http://www.networkworld.com/news/2010/031510-fccs-broadband-plan-155-billion.html?t51hb	1	FAV

ตารางที่ 6.8 ผลการทดสอบด้วยคีย์เวิร์ด 4g ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.cio.com/	1	
http://www.networkworld.com/news/2010/031210-layer8-fbi-internet-scams.html?t51hb	1	FAV
http://www.networkworld.com/news/2010/031510-virtual-server-security.html?t51hb	2	
http://www.networkworld.com/slideshows/2010/031510-facebook.html?t51hb	1	
http://www.networkworld.com/subnets/opensource/	1	

ในการทดสอบชุดนี้เว็บที่ปรากฏจะเป็นข่าวสารเป็นหลักเว็บที่ผู้ใช้พอใจส่วนใหญ่จะเป็นข่าวสารที่เกี่ยวข้องกับระบบ network ซึ่งมีเนื้อหาที่คล้ายคลึงกับคีย์เวิร์ด 3g เว็บที่ปรากฏซ้ำกันในแต่ละการทดสอบมีน้อย และไม่มีเว็บที่ปรากฏทั้งสี่การทดสอบเหมือนกัน

6.3.3 ผลการทดสอบกับคีย์เวิร์ด intel

ตารางที่ 6.9 ผลการทดสอบด้วยคีย์เวิร์ด intel ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://fa.wikipedia.org/wiki/%D8%A7%DB%8C%D9%86%D8%AA%D9%84	2	
http://www.intel.com/pressroom/intel_inside.htm	1	FAV
http://mk.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D0%BB	1	
http://www.intel.com/intel/	1	FAV
http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/745&format=HTML&aged=0&language=EN&guiLanguage=en	1	

ตารางที่ 6.10 ผลการทดสอบด้วยคีย์เวิร์ด intel ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://en.wikipedia.org/wiki/Intel_Corporation	2	FAV
http://www.reuters.com/article/idUSL2783620520070727?sp=true	1	FAV
http://simple.wikipedia.org/wiki/Intel	2	FAV
http://www.time.com/time/business/article/0,8599,1897913,00.html	1	FAV
http://www.michaelrobertson.com/archive.php?minute_id=56	1	

ตารางที่ 6.11 ผลการทดสอบด้วยคีย์เวิร์ด intel ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://en.wikipedia.org/wiki/Intel_Corporation	2	FAV
http://www.reuters.com/article/governmentFilingsNews/idUSL2783620520070727?sp=true	1	FAV
http://www.eetimes.com/news/latest/showArticle.jhtml?articleID=201303681	1	FAV
http://simple.wikipedia.org/wiki/Intel	2	FAV
http://fa.wikipedia.org/wiki/%D8%A7%DB%8C%D9%86%D8%AA%D9%84	2	

ตารางที่ 6.12 ผลการทดสอบด้วยคีย์เวิร์ด intel ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://ms.wikipedia.org/wiki/Intel_Corporation	1	
http://www.pcmag.com/article2/0,2817,2348923,00.asp	1	FAV
http://www.intel.com/community/selectacommunity.htm?iid=intel_comm+comm_select	1	FAV
http://eo.wikipedia.org/wiki/Intel	1	
http://www.old-computers.com/MUSEUM/computer.asp?c=754&st=1	1	FAV

ในการทดสอบกับคีย์เวิร์ด intel ในแต่ละการทดสอบย่อยนั้นเว็บที่ปรากฏเหมือนกันในหลายๆ การทดสอบมีน้อยซึ่งเป็นเพราะหน้าเว็บที่ผู้ใช้เลือกมี URL อยู่มากทำให้ระบบต้องทำการสุ่ม URL ขึ้นมาบางส่วนเพื่อนำมาใช้งาน

6.3.4 ผลการทดสอบกับคีย์เวิร์ด content management system

ตารางที่ 6.13 ผลการทดสอบด้วยคีย์เวิร์ด content management system ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://webdesign.about.com/od/contentmanagement/a/aa031300b.htm	4	FAV
http://webdesign.about.com/od/contentmanagement/a/aa021802a.htm	3	FAV
http://webdesign.about.com/od/contentmanagement/a/aa102504.htm	2	FAV
http://webdesign.about.com/od/contentmanagement/a/cms_resistance.htm	1	FAV
http://webdesign.about.com/cs/contentmgmt/bb/aab-cms.htm	2	FAV

ตารางที่ 6.14 ผลการทดสอบด้วยคีย์เวิร์ด content management system ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://webdesign.about.com/od/contentmanagement/a/aa031300a.htm	3	FAV
http://webdesign.about.com/od/contentmanagement/a/aa031300b.htm	4	FAV
http://webdesign.about.com/od/contentmanagement/a/aa021802a.htm	3	FAV
http://webdesign.about.com/cs/contentmgmt/bb/aab-cms.htm	3	FAV
http://webdesign.about.com/od/contentmanagement/a/aa031300c.htm	2	FAV

ตารางที่ 6.15 ผลการทดสอบด้วยคีย์เวิร์ด content management system ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://webdesign.about.com/od/contentmanagement/a/aa031300b.htm	4	FAV
http://webdesign.about.com/od/contentmanagement/a/aa102504.htm	2	FAV
http://webdesign.about.com/cs/contentmgmt/bb/aab-cms.htm	2	FAV
http://webdesign.about.com/od/contentmanagement/a/aa031300c.htm	2	FAV
http://webdesign.about.com/od/contentmanagement/a/content_probs.htm	1	FAV

ตารางที่ 6.16 ผลการทดสอบด้วยคีย์เวิร์ด content management system ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://webdesign.about.com/od/contentmanagement/a/aa031300a.htm	3	FAV
http://pcworld.about.com/news/Jul192004id116933.htm	1	
http://webdesign.about.com/od/contentmanagement/a/aa031300b.htm	4	FAV
http://webdesign.about.com/od/contentmanagement/a/content_probs_2.htm	1	FAV
http://webdesign.about.com/od/contentmanagement/a/aa021802a.htm	3	FAV

การทดสอบกับคีย์เวิร์ดนี้แสดงให้เห็นว่าเว็บที่ผู้ใช้เลือกมี URL ที่มีเนื้อหาคล้ายคลึงกับเว็บที่ผู้ใช้เลือกอยู่มาก และมีปริมาณ URL ที่น้อยทำให้แต่ละการทดลองนั้นผลลัพธ์ที่ออกมาจะเห็นว่ามีเว็บเดียวกันปรากฏในหลายๆ การทดสอบ

6.3.5 ผลการทดสอบกับคีย์เวิร์ด firefox

ตารางที่ 6.17 ผลการทดสอบด้วยคีย์เวิร์ด firefox ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://blog.mozilla.com/	3	FAV
http://www.mozilla.com/en-US/firefox/all.html	4	
http://www.getpersonas.com/en-US/	3	
http://www.mozilla.org/contribute/	1	FAV
http://support.mozilla.com/th/kb/	3	FAV

ตารางที่ 6.18 ผลการทดสอบด้วยคีย์เวิร์ด firefox ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.getpersonas.com/en-US/	3	FAV
http://blog.mozilla.com/	3	FAV
http://support.mozilla.com/th/kb/	3	FAV
http://www.mozilla.com/en-US/firefox/all.html	4	
http://www.spreadfirefox.com/	2	

ตารางที่ 6.19 ผลการทดสอบด้วยคีย์เวิร์ด firefox ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://planet.mozilla.org/	1	FAV
http://www.mozilla.com/en-US/firefox/all.html	4	
http://support.mozilla.com/th/kb/	3	FAV
http://support.mozilla.com/en-US/kb/	1	FAV
http://www.getpersonas.com/en-US/	3	FAV

ตารางที่ 6.20 ผลการทดสอบด้วยคีย์เวิร์ด firefox ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.mozilla.com/th/firefox/fastest/	1	FAV
http://www.mozilla.com/en-US/firefox/all.html	4	
http://blog.mozilla.com/	3	FAV
http://www.mozilla.com/en-US/press/	1	FAV
http://www.spreadfirefox.com/	2	FAV

การทดสอบกับคีย์เวิร์ด firefox แสดงให้เห็นว่าเว็บที่ผู้ใช้เลือกมี URL ที่มีเนื้อหาที่ผู้ใช้พอใจอยู่มาก และปริมาณ URL ทั้งหมดมีไม่มากนักทำให้เว็บปรากฏอยู่ในหลายๆ การทดลอง

6.3.6 ผลการทดสอบกับคีย์เวิร์ด software engineering

ตารางที่ 6.21 ผลการทดสอบด้วยคีย์เวิร์ด software engineering ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.answers.com/main/what_content.jsp	2	
http://www.answers.com/topic/answertips	1	
http://wiki.answers.com/Q/What_is_the_difference_between_a_Software_engineer_and_Computer_Software_engineer&src=ansTT	1	FAV
http://wiki.answers.com/Q/What_is_the_aim_of_software_engineering..what_does_the_discipline_of_software_engineering_discuss&src=ansTT	1	
http://wiki.answers.com/Q/What_is_software_engineering&src=ansTT	1	FAV

ตารางที่ 6.22 ผลการทดสอบด้วยคีย์เวิร์ด software engineering ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.answers.com/main/sitemap.jsp	3	
http://reference.answers.com/	2	
http://wiki.answers.com/Q/How_do_you_engineer_a_software&src=ansTT	1	FAV
http://wiki.answers.com/Q/How_can_be_a_software_engineer&src=ansTT	1	FAV
http://wiki.answers.com/Q/What_is_software_and_software_engineering&src=ansTT	1	FAV

ตารางที่ 6.23 ผลการทดสอบด้วยคีย์เวิร์ด software engineering ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.answers.com/main/answers_rss.jsp	1	
http://www.answers.com/main/sitemap.jsp	3	
http://reference.answers.com/	2	
http://www.computer.org/portal/web/swebok	1	
http://wiki.answers.com/Q/Software_Engineering_is_related_to_Engineering_or_not&src=ansTT	2	FAV

ตารางที่ 6.24 ผลการทดสอบด้วยคีย์เวิร์ด software engineering ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.answers.com/main/what_content.jsp	2	
http://www.answers.com/main/sitemap.jsp	3	
http://www.barronseduc.com/	1	
http://www.springer.com/?SGWID=0-102-0-0-0	1	
http://wiki.answers.com/Q/Software_Engineering_is_related_to_Engineering_or_not&src=ansTT	2	FAV

การทดสอบกับคีย์เวิร์ด software engineering ในเว็บที่กำหนดเว็บที่ปรากฏในหลายๆ การทดสอบในชุดนี้เว็บที่ปรากฏเพียงครั้งเดียวในแต่ละการทดลองย่อยมักจะเป็นเว็บที่ผู้ใช้พอใจ ซึ่งมักเป็นเว็บที่เป็นการถามและตอบซึ่งเป็นลักษณะเด่นของเว็บไซต์นี้

6.3.7 ผลการทดสอบกับคีย์เวิร์ด data mining

ตารางที่ 6.25 ผลการทดสอบด้วยคีย์เวิร์ด data mining ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://feeds.feedburner.com/kdnuggets-data-mining-analytics	4	FAV
http://twitter.com/kdnuggets	4	FAV
http://www.kdnuggets.com/phpBB/viewforum.php?f=2	4	
http://www.kdnuggets.com/phpBB/viewforum.php?f=11	4	
http://www.kdnuggets.com/phpBB/index.php	4	

ตารางที่ 6.26 ผลการทดสอบด้วยคีย์เวิร์ด data mining ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://feeds.feedburner.com/kdnuggets-data-mining-analytics	4	FAV
http://twitter.com/kdnuggets	4	FAV
http://www.kdnuggets.com/phpBB/viewforum.php?f=2	4	
http://www.kdnuggets.com/phpBB/viewforum.php?f=11	4	
http://www.kdnuggets.com/phpBB/index.php	4	

ตารางที่ 6.27 ผลการทดสอบด้วยคีย์เวิร์ด data mining ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://twitter.com/kdnuggets	4	FAV
http://feeds.feedburner.com/kdnuggets-data-mining-analytics	4	FAV
http://www.kdnuggets.com/phpBB/viewforum.php?f=2	4	
http://www.kdnuggets.com/phpBB/viewforum.php?f=11	4	
http://www.kdnuggets.com/phpBB/index.php	4	

ตารางที่ 6.28 ผลการทดสอบด้วยคีย์เวิร์ด data mining ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://twitter.com/kdnuggets	4	FAV
http://feeds.feedburner.com/kdnuggets-data-mining-analytics	4	FAV
http://www.kdnuggets.com/phpBB/viewforum.php?f=2	4	
http://www.kdnuggets.com/phpBB/viewforum.php?f=11	4	
http://www.kdnuggets.com/phpBB/index.php	4	

การทดสอบกับคีย์เวิร์ด data mining แสดงให้เห็นว่าในเว็บที่ผู้ใช้เลือกนั้นมี URL อยู่น้อย ทำให้เว็บในสี่การทดสอบเหมือนกันทั้งหมดต่างกันที่ลำดับความคล้ายคลึงเท่านั้น การทดสอบนี้แสดงให้เห็นว่าค่า Ln กับ log10 นั้นไม่ส่งผลกระทบต่ออันดับความคล้ายคลึงของเว็บแต่วิธีการคำนวณความคล้ายคลึงส่งผลต่อความคล้ายคลึงเท่านั้น

6.3.8 ผลการทดสอบกับคีย์เวิร์ด information security

ตารางที่ 6.29 ผลการทดสอบด้วยคีย์เวิร์ด information security ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://searchsecurity.techtarget.com/home/0,289692,sid14,00.htm	4	FAV
http://searchsecurity.techtarget.com/topics/0,295493,sid14_tax314032,00.html	2	
http://searchsecurity.techtarget.com/news/article/0,289142,sid14_gci1420681,00.html	2	FAV
http://securitymanagement.searchsecurity.com/kw;Network+Security/Network+Security/security.htm	2	
http://twitter.com/SearchSecurity	1	

ตารางที่ 6.30 ผลการทดสอบด้วยคีย์เวิร์ด information security ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://searchsecurity.techtarget.com/home/0,289692,sid14,00.html	4	FAV
http://searchsecurity.techtarget.com/topics/0,295493,sid14_tax314032,00.html	2	
http://searchsecurity.techtarget.com/news/article/0,289142,sid14_gci1405679,00.html	2	FAV
http://securitymanagement.searchsecurity.com/kw;Security+Solutions/Security+Solutions/security.htm	1	
http://itknowledgeexchange.techtarget.com/security-wire-weekly/google-attacks-and-infrastructure-insecurities/	1	

ตารางที่ 6.31 ผลการทดสอบด้วยคีย์เวิร์ด information security ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://searchsecurity.techtarget.com/home/0,289692,sid14,00.html	4	FAV
http://searchmidmarketsecurity.techtarget.com/	1	
http://searchsecurity.techtarget.com/magazineFeature/0,296894,sid14_gci1357847,00.html	1	FAV
http://securitymanagement.searchsecurity.com/kw;Internet+Security/Internet+Security/security.htm	2	
http://searchsecurity.techtarget.com/video/0,297151,sid14_gci1355568,00.html	1	FAV

ตารางที่ 6.32 ผลการทดสอบด้วยคีย์เวิร์ด information security ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://searchsecurity.techtarget.com/home/0,289692,sid14,00.html	4	FAV
http://searchsecurity.techtarget.com/aboutUs/0,289153,sid14,00.html	1	
http://securitymanagement.searchsecurity.com/kw;Internet+Security/Internet+Security/security.htm	2	
http://searchsecurity.techtarget.com/topicsMain/0,295490,sid14,00.html	1	
http://securitymanagement.searchsecurity.com/kw;Network+Security/Network+Security/security.htm	1	

การทดสอบกับคีย์เวิร์ด information security เว็บที่ผู้ใช้เลือกเป็นเว็บที่เป็นแหล่งรวมข่าวสาร ทำให้เว็บที่เป็นผลลัพธ์เป็นลักษณะของบทความต่างๆ ซึ่งก็มีบางบทความผู้ใช้พอใจ

6.3.9 ผลการทดสอบกับคีย์เวิร์ด avatar

ตารางที่ 6.33 ผลการทดสอบด้วยคีย์เวิร์ด avatar ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.comingsoon.net/films.php?id=45067	1	
http://www.avatarmovie.com/	1	FAV
http://www.comingsoon.net/news/movienews.php?id=60853	1	FAV
http://www.comingsoon.net/news/movienews.php?id=60651	1	FAV
http://www.comingsoon.net/news/videonews.php?id=60735	1	FAV

ตารางที่ 6.34 ผลการทดสอบด้วยคีย์เวิร์ด avatar ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.comingsoon.net/films.php?id=59398	1	
http://www.comingsoon.net/films.php?id=39442	2	
http://www.youtube.com/watch?v=TsZILiJBukc	1	FAV
http://www.youtube.com/watch?v=StI04ssV-y8	1	FAV
http://www.youtube.com/watch?v=BfUnoWFf7A8	2	FAV

ตารางที่ 6.35 ผลการทดสอบด้วยคีย์เวิร์ด avatar ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.comingsoon.net/films.php?id=39442	2	
http://www.dvdfile.com/reviews/blurayreviews/25019-i-know-what-you-did-last-summer-bd	1	
http://www.youtube.com/watch?v=ZD8EEJepfiY	2	FAV
http://www.youtube.com/watch?v=t0OCCuqcyGg	1	FAV
http://www.comingsoon.net/dvd/news.php	1	FAV

ตารางที่ 6.36 ผลการทดสอบด้วยคีย์เวิร์ด avatar ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.dvdfile.com/reviews/blurayreviews/23269-four-christmases-bd?start=2	1	
http://www.youtube.com/watch?v=49Hg_4jRHFI	1	FAV
http://www.youtube.com/watch?v=ZD8EEJepfiY	2	FAV
http://www.youtube.com/watch?v=BfUnoWFf7A8	2	FAV
http://www.youtube.com	1	

การทดสอบกับคีย์เวิร์ด avatar เว็บที่ออกมาแต่ละการทดลองมักจะซ้ำกันน้อยแสดงให้เห็นว่าเว็บที่ผู้ใช้เลือกมี URL อยู่มากทำให้ต้องมีการสุ่ม URL ขึ้นมาทำงาน ซึ่งบาง URL ที่สุ่มขึ้นมาก็มักจะเชื่อมต่อไปยังหนังเรื่องอื่น หรือเชื่อมต่อไปยังภาพวีดีโอหนังตัวอย่างของหนังเรื่องดังกล่าว

6.3.10 ผลการทดสอบกับคีย์เวิร์ด raid

ตารางที่ 6.37 ผลการทดสอบด้วยคีย์เวิร์ด raid ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://simple.wikipedia.org/wiki/RAID	1	FAV
http://www.ibizdir.com/business-research/web-hosting/61-introducing-raid.html	2	FAV
http://www.tomshardware.com/reviews/RAID-MIGRATION-ADVENTURE,1640.html	3	FAV
http://www.shub-internet.org/brad/FreeBSD/vinum.html	1	
http://www.dmoz.org/Computers/Hardware/Storage/Subsystems/RAID/	1	FAV

ตารางที่ 6.38 ผลการทดสอบด้วยคีย์เวิร์ด raid ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ K-NN

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.acnc.com/raid.html	2	FAV
http://www.enterprisestorageforum.com/technology/features/article.php/3839636	2	FAV
http://tldp.org/HOWTO/Software-RAID-0.4x-HOWTO-8.html	3	FAV
http://tr.wikipedia.org/wiki/RAID	1	
http://www.tomshardware.com/reviews/RAID-MIGRATION-ADVENTURE,1640.html	3	FAV

ตารางที่ 6.39 ผลการทดสอบด้วยคีย์เวิร์ด raid ค่า IDF ใช้ Ln และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.ibizdir.com/business-research/web-hosting/61-introducing-raid.html	2	FAV
http://www.acnc.com/raid.html	2	FAV
http://www.enterprisestorageforum.com/technology/features/article.php/3839636	2	FAV
http://tldp.org/HOWTO/Software-RAID-0.4x-HOWTO-8.html	3	FAV
http://www.drobo.com/products/drobo.php	1	

ตารางที่ 6.40 ผลการทดสอบด้วยคีย์เวิร์ด raid ค่า IDF ใช้ Log_{10} และการเปรียบเทียบใช้ CS

รายชื่อเว็บ	ปรากฏผลในการทดลองอื่น	ผู้ใช้พอใจ
http://www.i-justblog.com/2009/06/linux-raid-and-lvm-management.html	1	FAV
http://www.tomshardware.com/reviews/RAID-MIGRATION-ADVENTURE,1640.html	3	FAV
http://tldp.org/HOWTO/Software-RAID-HOWTO.html	3	FAV
http://fi.wikipedia.org/wiki/RAID_%28tietotekniikka%29	1	
http://sk.wikipedia.org/wiki/RAID	1	

การทดสอบกับคีย์เวิร์ด raid แสดงให้เห็นว่า crawler ของเราไปนำเว็บที่ไม่ใช่ภาษาอังกฤษแต่เป็นตัวอักษรภาษาอังกฤษมาคำนวณ ซึ่งเนื้อหาของเว็บคล้ายกับเว็บที่ผู้ใช้เลือกจริงแต่ว่าเป็นคนละภาษาทำให้ผู้ใช้ไม่พอใจในเว็บดังกล่าว

6.4 สรุปผลการทำงาน

จากการทดสอบทั้งหมดทำให้เห็นว่าการใช้ Ln กับ Log_{10} ในการคำนวณนั้นไม่ทำให้ลำดับความคล้ายคลึงเปลี่ยนไปเพราะว่าการคำนวณค่า IDF ไม่ว่าใช้วิธีการใดสุดท้ายแล้วผลมันจะทำให้ค่าของ TF-IDF ที่ออกมาเป็นไปในทางทิศทางเดียวกัน

ส่วนการคำนวณความคล้ายคลึงของเว็บไซต์โดยใช้แบบจำลอง K-NN กับ Cosine Similarity นั้น ผลต่างกันเล็กน้อย ซึ่งเว็บที่จะได้เป็นผลลัพธ์กลับคืนมานั้นขึ้นอยู่กับปริมาณ URL ที่เว็บที่ผู้ใช้เลือกมีอยู่ว่ามีเนื้อหาภายใน URL นั้นใกล้เคียงกับเว็บที่ผู้ใช้เลือกหรือไม่เป็นสำคัญ แต่ถ้าเว็บที่ผู้ใช้เลือกมี URL อยู่มากเกินกว่าที่ระบบสามารถรับได้ระบบจะต้องสุ่ม URL เหล่านั้นขึ้นมาทำงานทำให้มีโอกาสได้ URL ที่ไม่มีเนื้อหาที่คล้ายหรือคล้ายน้อยกว่า URL ที่ไม่ถูกเลือก

และอีกปัญหาที่เห็นชัดคือการทำงานของโปรแกรมยังไม่สามารถตรวจสอบได้ว่าเนื้อหาเว็บไซต์นั้นเป็นภาษาอะไร ซึ่งถ้าเรานำเว็บที่มีเนื้อหาคล้ายกับเว็บที่ผู้ใช้เลือกแต่เป็นคนละภาษาไปนำเสนอแก่ผู้ใช้ มีโอกาสสูงที่ผู้ใช้จะไม่พอใจในผลลัพธ์นั้นถึงแม้จะมีเนื้อหาคล้ายกัน เช่น เว็บ Wikipedia ซึ่งจะมีเนื้อหาเรื่องเดียวกัน แต่มีหลายๆ ภาษา ไว้ให้เลือกอ่าน

บทที่ 7

บทสรุปและข้อเสนอแนะ

7.1 สรุปผลการพัฒนาระบบงาน

โครงการพัฒนาระบบงานนี้ มีจุดประสงค์ในการพัฒนาซอฟแวร์ในแบบ Web Aplication เพื่อทำงานต่อยอดจากการค้นหาของ Search Engine ซึ่งหลังจากออกแบบและพัฒนาระบบงานแล้ว ได้ผลออกมาดังนี้

- ระบบค้นหาเว็บไซต์แบบเจาะจงของเราสามารถนำผลลัพธ์จากการค้นหาของ Search Engine มาใช้งานต่อเพื่อค้นหาเว็บไซต์ที่มีความคล้ายคลึงกันได้
- ระบบค้นหาเว็บไซต์แบบเจาะจงของเราสามารถค้นหาเว็บไซต์ที่คล้ายคลึงกับเว็บไซต์ที่ผู้ใช้เลือกจาก Search Engine ได้โดยให้ทำงานร่วมกับ Crawler ที่เราพัฒนาขึ้นเอง
- Crawler ที่เราพัฒนาขึ้นเป็น Crawler ประเภท Topic-Specific Web Crawler ซึ่งจะเลือกเว็บไซต์ที่มีค่าเพียงพอที่จะนำไปทำงานต่อ ซึ่งขณะทำงาน Crawler ได้มีการพิจารณาเลือกเว็บไซต์และไม่เลือกเว็บไซต์ แต่ละเว็บไปทำงานต่อจริงๆ โดยพิจารณาจากค่าความห่างกับเว็บไซต์ที่ผู้ใช้เลือกที่แบบจำลอง K-NN ส่งมาให้
- แบบจำลอง K-NN ทำงานได้อย่างรวดเร็วมาก เนื่องจากเว็บที่เรานำมาเปรียบเทียบนั้นมีจำนวนเล็กน้อยเท่านั้น (ไม่ถึง 100 เว็บในการทำงานแต่ละครั้ง)
- แต่ส่วนที่ทำงานช้าอย่างเห็นได้ชัดในระบบของเราคือการดึงข้อมูลเว็บไซต์แต่ละเว็บของ Crawler ซึ่งทำงานได้ช้ามากกว่าขั้นตอนอื่นๆ อย่างเห็นได้ชัด
- ในการค้นหาเว็บไซต์ที่คล้ายคลึงกับเว็บไซต์ที่ผู้ใช้เลือกครั้งหนึ่งระบบจะทำงานเร็วหรือช้านั้นขึ้นอยู่กับจำนวนเว็บไซต์ที่ผู้ใช้เลือกในตอนแรก กับจำนวนเว็บไซต์ที่ระบบสามารถดึงข้อมูลได้สูงสุดที่เรากำหนดไว้เป็นสำคัญ เพราะสองตัวแปรนี้จะเป็นตัวกำหนดว่า Crawler ของระบบนั้นจะต้องทำงานมากหรือน้อยไม่เท่ากัน

7.2 ข้อเสนอแนะ

ถึงระบบของเราจะพัฒนาออกมาสำเร็จและทำงานได้ก็ตาม แต่ก็ยังบางจุดซึ่งยังถือว่าเป็นจุดด้อยของระบบนี้อยู่ และมีบางจุดที่อยากให้ปรับปรุงเพิ่มเติมในการพัฒนาครั้งต่อๆ ไปดังนี้

- ควรทำงานร่วมกับ Search Engine ตัวอื่นได้ด้วยนอกจาก Bing แล้ว ควรทำงานร่วมกับ Google หรือ Yahoo ได้เช่นกัน
- พิจารณาเลือกภาษาอื่นในการพัฒนาในส่วนของ Crawler เพราะ PHP ไม่ใช่ภาษา Multi-Thread โดยตรง อาจจะพัฒนาโดยให้ภาษา PHP เป็นส่วนติดต่อประสานผู้ใช้ และให้ภาษาอื่นทำงานเป็นตัว Crawler ไป
- ควรปรับปรุงการทำงานในส่วนของการดึงข้อมูลเว็บไซต์แต่ละเว็บไซต์ของ Crawler ให้สามารถดึงข้อมูลแต่ละเว็บไซต์ได้เร็วยิ่งขึ้น
- ควรปรับปรุงให้ Crawler สามารถกรองเว็บที่ไม่ใช่ภาษาที่เราสนใจออกไปได้ สิ่งนี้จะทำให้ผลลัพธ์ของเราดูเกี่ยวกับเว็บที่ผู้ใช้สนใจมากขึ้น เนื่องจากเป็นภาษาเดียว กันกับเว็บไซต์ที่ผู้ใช้เลือก

ประวัติผู้เขียน

ชื่อผู้เขียน	นายวงศกร ตั้งทรงจิตร
สถานที่เกิด	จังหวัดปราจีนบุรี
ระดับประถมศึกษา	โรงเรียนอนุบาลปราจีนบุรี
ระดับมัธยมศึกษาตอนต้น	โรงเรียนกบินทร์วิทยา
ระดับมัธยมศึกษาตอนปลาย	โรงเรียนกบินทร์วิทยา
วุฒิการศึกษาระดับปริญญาตรี	สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ประสบการณ์การทำงาน	ห้างหุ้นส่วนจำกัด เฮลโล โซลูชั่น