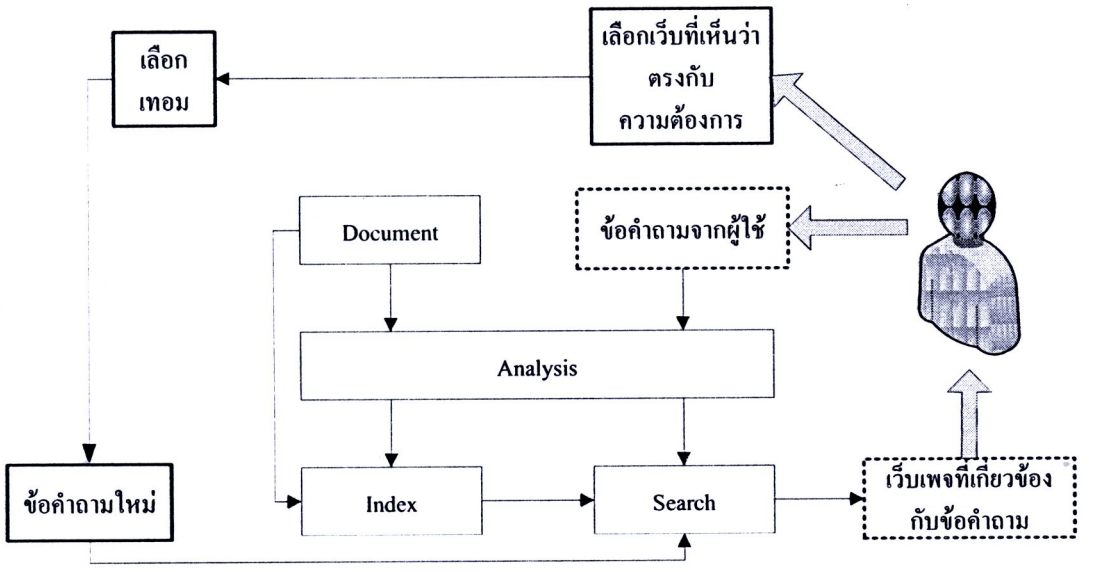
บทที่ 3
ผลการวิจัย

3.1 ส่วนประกอบของระบบ



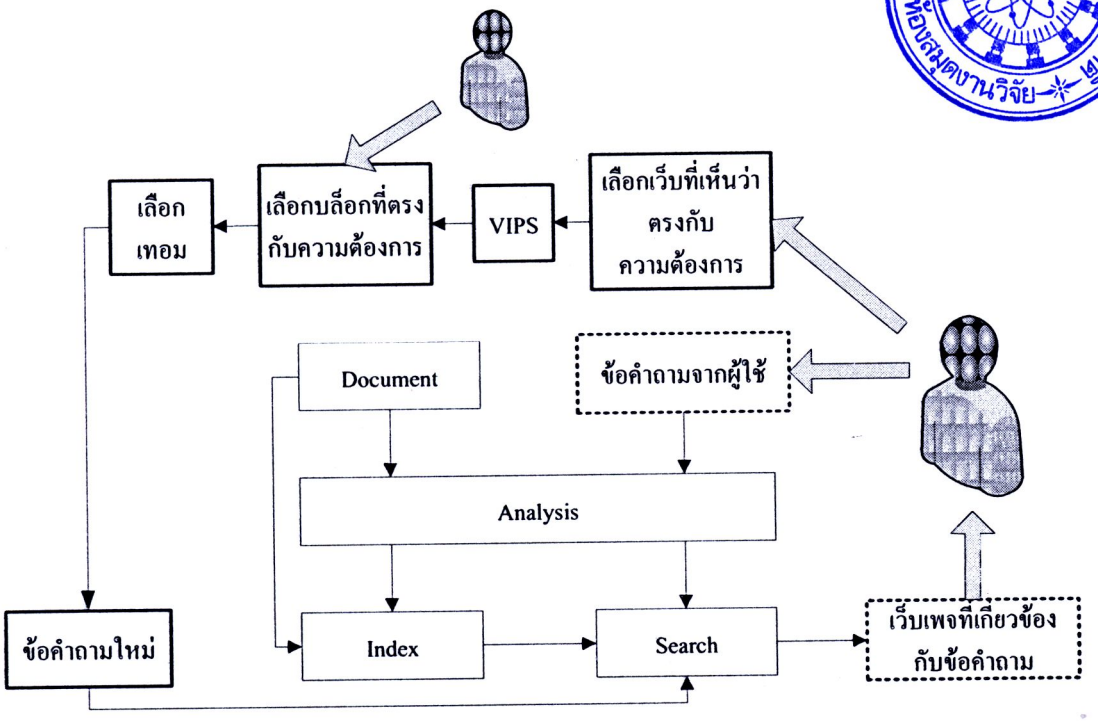
รูปที่ 3.1 แผนภาพทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ

จากรูปที่ 3.1 การค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับจะเริ่มการทำงานจากการรับข้อคำถามจากผู้ใช้ จากนั้นจะใช้ลูซีนเพื่อหาเว็บเพจที่เกี่ยวข้องกับข้อคำถาม เมื่อได้เว็บที่เกี่ยวข้องกับข้อคำถามแล้ว ผู้ใช้ก็จะทำการเลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ และนำเว็บเพจที่เลือกไปหาข้อคำถามใหม่โดยการเพิ่มคำใหม่เข้าไปในข้อคำถามเดิม ซึ่งคำใหม่ที่เพิ่มเข้าไปจะอยู่ภายในเว็บเพจที่ผู้ใช้เลือกเท่านั้น

เพื่อเป็นการเปรียบเทียบประสิทธิภาพการหาข้อคำถามใหม่ จึงมีการออกแบบระบบค้นคืนสารสนเทศบนเว็บโดยใช้อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับดังรูปที่ 3.2 การทำงานจะเริ่มจากการใช้ลูซีนเพื่อหาเว็บเพจที่เกี่ยวข้องกับข้อคำถาม จากนั้นจะใช้ลูซีนเพื่อหาเว็บเพจที่เกี่ยวข้องกับข้อคำถาม เมื่อได้เว็บที่เกี่ยวข้องกับข้อคำถามแล้วผู้ใช้ก็จะทำการเลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ และนำเว็บเพจที่เลือกไปทำการแบ่งออกเป็นบล็อกโดยใช้อัลกอริทึมวีไอพีเอส ผู้ใช้ก็จะทำการเลือกบล็อกที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการอีกครั้งหนึ่ง ระบบก็จะหาข้อคำถามใหม่โดยการเพิ่มคำใหม่เข้าไปในข้อคำถามเดิม โดยคำใหม่ที่เพิ่มเข้าไปจะอยู่ภายในบล็อกที่ผู้ใช้เลือกเท่านั้น

เราสามารถแบ่งระบบค้นคืนออกเป็นระบบย่อยได้ 3 ส่วนด้วยกันคือ

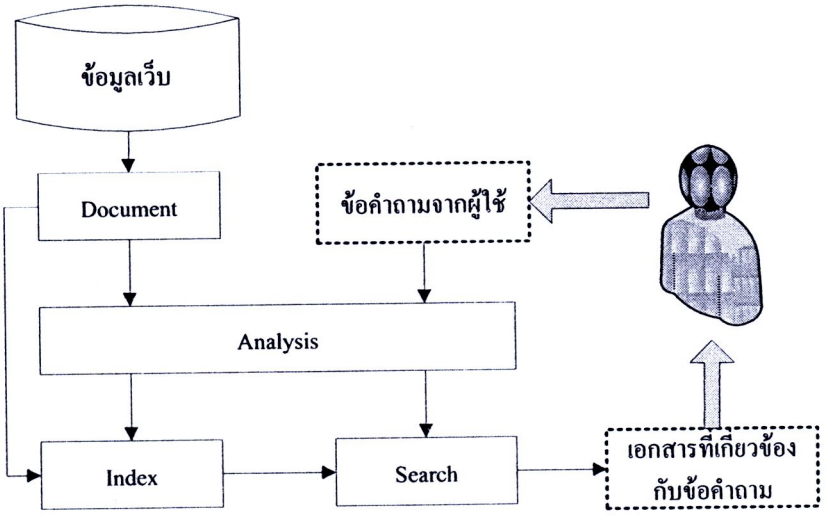
1. การค้นคืนสารสนเทศ
2. การแบ่งเว็บเพจออกเป็นบล็อก
3. การหาข้อคำถามใหม่



รูปที่ 3.2 แผนภาพทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับ

3.1.1 การค้นคืนสารสนเทศ

การค้นคืนสารสนเทศในส่วนแรกนี้เราจะทำการศึกษาโดยใช้ลูซีน ซึ่งมีการกำหนดค่าการทำงานในลูซีนดังนี้



รูปที่ 3.3 การค้นคืนสารสนเทศโดยลูซีน

1) ส่วนการจัดการเอกสาร (Document)

ดังนั้นในการค้นหาข้อมูลจะทำการค้นได้จาก 4 ส่วนนั้นคือ ไทเทิล (Title) เมตทา (Meta) บอดี้ (Body) และยูอาร์แอล (URL) โดย ฟิลด์ยูอาร์แอลและไทเทิลถือว่ามีความสำคัญเราจะไม่ทำการวิเคราะห์คำส่วนข้อมูลในฟิลด์บอดี้มีจำนวนมากจะทำการบีบอัดข้อมูล เพื่อประหยัดพื้นที่ในการเก็บข้อมูล ส่วนการจัดการเอกสารที่ทำการเก็บเอกสารแต่ละหน้าเว็บเพจจะประกอบไปด้วยฟิลด์ดังต่อไปนี้

ตารางที่ 3.1 ฟิลด์ที่กำหนด

Field	ค่าของ field
title	ทำเป็นดัชนีแต่ไม่ต้องวิเคราะห์คำ, ทำการเก็บข้อมูลดั้งเดิม
meta	ทำการวิเคราะห์คำ ก่อนทำเป็นดัชนี , ทำการเก็บข้อมูลดั้งเดิม
body	ทำการวิเคราะห์คำก่อนทำเป็นดัชนี , ทำการเก็บข้อมูลดั้งเดิมและบีบอัดข้อมูล
url	ทำเป็นดัชนีแต่ไม่ต้องวิเคราะห์คำ , ทำการเก็บข้อมูลดั้งเดิม

2) ส่วนการวิเคราะห์คำ (Analysis)

วิเคราะห์คำที่ใช้ในการวิเคราะห์ข้อความและข้อคำถามจะใช้ StandardAnalyzer ของลูซีน คือ

- ทำการแบ่งเป็นคำตามช่องว่าง และอักขระพิเศษที่ไม่ใช่ตัวอักษร
- เปลี่ยนตัวอักษรภาษาอังกฤษตัวใหญ่เป็นตัวเล็กทั้งหมด
- ตัดคำที่มีอยู่ในเอกสารมากแต่ไม่ แสดงความหมายที่สำคัญ
- วิเคราะห์หลักไวยากรณ์ คือ สามารถรู้ ลักษณะอีเมล์ ไอพีแอดเดรส ตัวย่อและตัวอักษรที่ประกอบกับตัวเลข

ตัวอย่างการวิเคราะห์คำโดย StandardAnalyzer

The XY&Z Corporation – XYZ@example.com

[xy&z] [corporation] [xyz@example.com]

จากตัวอย่าง จะเห็นได้ว่าการใช้ StandardAnalyzer สามารถวิเคราะห์ข้อความโดยแบ่งเป็นคำตามช่องว่าง และอักขระพิเศษที่ไม่ใช่ตัวอักษรเปลี่ยนตัวอักษรภาษาอังกฤษตัวใหญ่เป็นตัวเล็กทั้งหมด และจะไม่ทำการแบ่งข้อความที่เป็นอีเมล์

3) ส่วนดัชนี (Index)

เก็บข้อมูลดัชนีแบบแฟ้มข้อมูลผกผันตามรูปแบบของลูซีน

4) ส่วนค้นคืน (Search)

การค้นคืนข้อมูลโดยหาเอกสารที่เกี่ยวข้องกับข้อคำถามของผู้ใช้จากการเปรียบเทียบกับ ดัชนีที่มี

- คิวรี่เฟสเซอร์ ที่ใช้ทำการวิเคราะห์ข้อคำถามใช้รูปแบบเดียวกับการวิเคราะห์เอกสาร (StandardAnalyzer)
- ใช้ เทอมคิวรี่ (TermQuery) ในการค้นหาเอกสารที่เกี่ยวข้องกับข้อคำถาม
- การคำนวณค่าสกอร์เพื่อนำไปจัดลำดับการแสดงผลจะกำหนดให้ ทุกๆคำในข้อคำถาม ฟิลด์ทุกฟิลด์และเอกสารทุกเอกสารมีค่าน้ำหนักเท่ากัน

3.1.2 การแบ่งเว็บเพจออกเป็นบล็อก

การแบ่งเว็บเพจออกเป็นบล็อกมีขั้นตอนดังนี้ คือ

1. หลังจากที่ได้ทำการค้นคืนเอกสารในครั้งแรกแล้ว ก็จะได้รายชื่อของเว็บเพจที่ระบบค้นคืนกลับมาให้ผู้ใช้ จากนั้นผู้ใช้จะทำการเลือกเว็บเพจที่คิดว่ามีความเกี่ยวข้องกับข้อคำถาม โดยเว็บเพจที่ผู้ใช้เลือกนั้น จะนำมาทำการแบ่งเป็นบล็อกด้วย วีไอพีเอสอัลกอริทึม ซึ่งมีขั้นตอนการทำงานเริ่มจาก การตัดออกเป็นบล็อก โดยจะเป็นการนำดอมทรีของเว็บเพจมาพิจารณาโดยใช้กฎในการแบ่งออกเป็นบล็อก จากกฎจะได้ค่า ดีโอซีของแต่ละบล็อก และเราจะกำหนดค่าพีดีโอซี ให้มีค่าเท่ากับ 4 เสมอ เพื่อกำหนดโครงสร้างเนื้อหา
2. การหาตัวแบ่ง จะเป็นการหาตัวแบ่งระหว่างแต่ละบล็อกที่ได้จากข้อหนึ่งเพื่อแยกแต่ละบล็อกให้ออกจากกัน และทำการให้ค่าน้ำหนักแต่ละตัวแบ่ง
3. การสร้างโครงสร้างเนื้อหา ในขั้นตอนนี้จะทำการรวมบล็อกโดยดูจากค่าน้ำหนักของตัวแบ่งแต่ละตัว โดยตัวแบ่งที่มีค่าน้ำหนักน้อยจะสามารถทำการรวมกันเป็นบล็อก
4. ผู้ใช้ทำการเลือกบล็อกที่เห็นว่าตรงตามความต้องการ โดยบล็อกที่ถูกเลือกจะนำไปทำการเพิ่มข้อคำถามใหม่

3.1.3 การหาข้อคำถามใหม่

การหาข้อคำถามใหม่มีขั้นตอนดังนี้ คือ

1. นำข้อมูลจากเว็บที่ถูกเลือก หรือข้อมูลจากบล็อกที่ถูกเลือกที่มาจากการใช้วีไอพีเอสอัลกอริทึมมาทำการวิเคราะห์หาคำใหม่ โดยเทอมที่จะเพิ่มเข้าไปนั้นใช้หลักการของเว็กเตอร์โมเดล สามารถคำนวณได้จากสูตรซอสออเดอร์ คือ

$$s(t_k) = \frac{1}{n_k} df_k \left| R^{t_k} \right| \tag{3.1}$$

โดยที่ n_k คือค่านอยซ์เมทชัวร์ของเทอมที่ k

df_k คือความถี่ของเทอมที่ k ที่ปรากฏในหน้าเว็บที่ถูกเลือก หรือความถี่ของเทอมที่ k ที่ปรากฏในบล็อกที่ถูกเลือก

R^{t_k} คือจำนวนเอกสารที่ค้นคืนได้ที่มีเทอมที่ k ปรากฏอยู่

คำนวณหาค่านอยซ์เมทชัวร์ได้จากสูตร

$$n_k = \sum_{i=1} N \times \frac{tf_{ik}}{f_k} \times \log(tf_{ik} f_k) \tag{3.2}$$

โดยที่ N คือจำนวนเทอมทั้งหมดภายในหน้าเว็บที่ถูกเลือก หรือจำนวนเทอมทั้งหมดภายในบล็อกที่ถูกเลือก

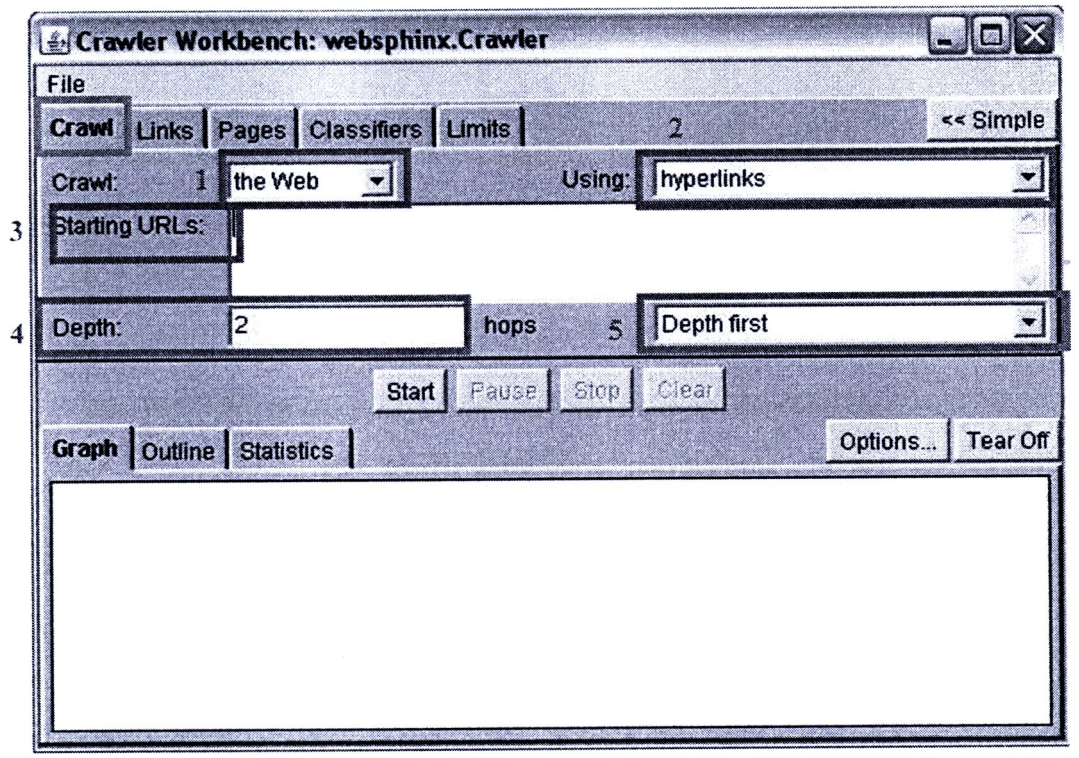
tf_{ik} คือความถี่ของเทอมที่ k ที่ปรากฏในเอกสารที่ i

f_k คือความถี่ของเทอมที่ k ที่ปรากฏในหน้าเว็บที่ถูกเลือก หรือ ความถี่ของเทอมที่ k ที่ปรากฏในบล็อกที่ถูกเลือก

2. เมื่อได้ค่าซอสออเดอร์ของแต่ละคำออกมาแล้ว นำมาเรียงลำดับ แล้วเลือกเพียงลำดับสูงสุดเพียง 3 ค่า จากนั้นนำมาเพิ่มในข้อคำถาม ได้เป็นข้อคำถามใหม่ขึ้นมา

3.2 การเตรียมข้อมูล

การเตรียมข้อมูลจะใช้ ครอเลอร์เว็บสฟิงซ์ (Website-Specific Processors for HTML Information Extraction: WebSPHINX) ซึ่งเป็นโอเพนซอร์สโปรแกรมครอเลอร์มาทำการเก็บข้อมูลเว็บเพจ

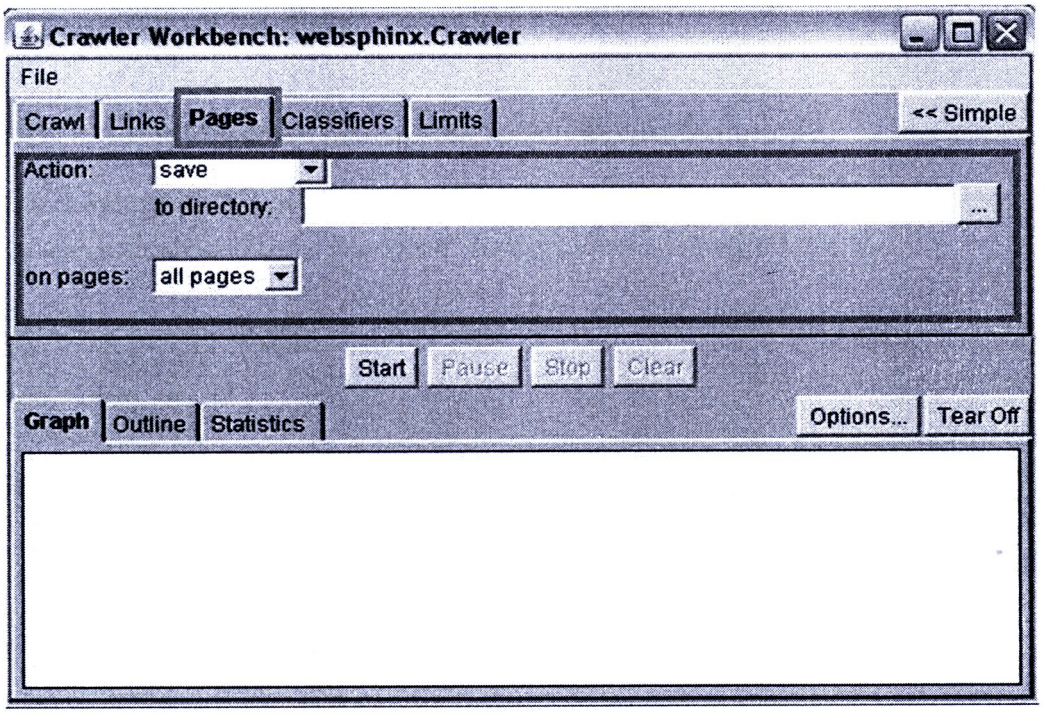


รูปที่ 3.4 แสดงโปรแกรมเว็บสฟิงซ์

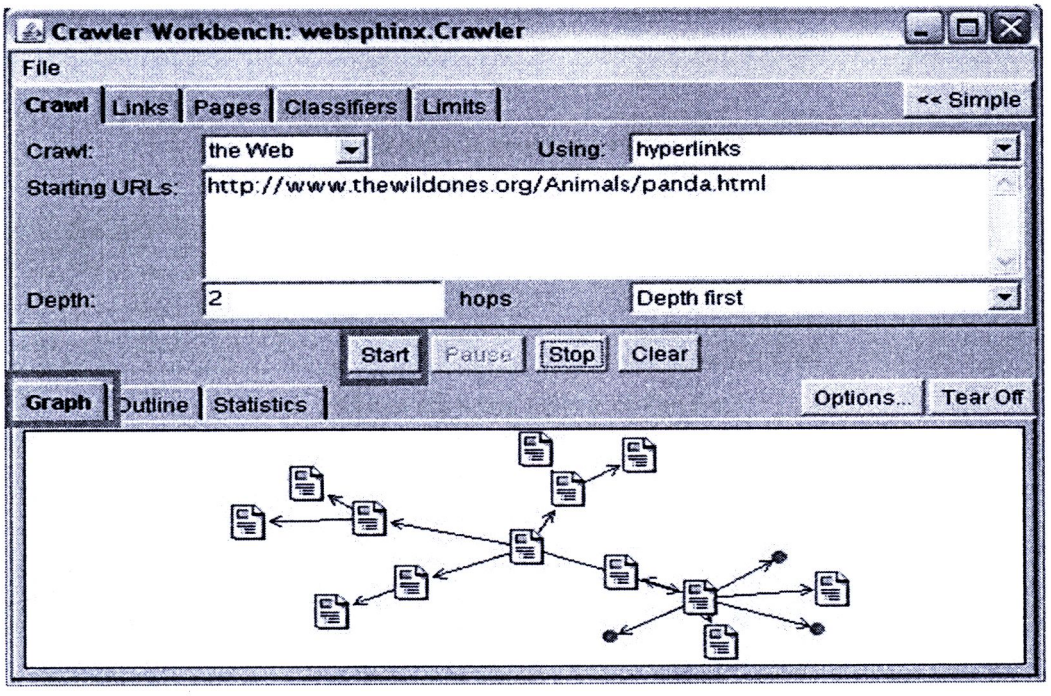
เมื่อเปิดโปรแกรมเว็บสฟิงซ์จะได้หน้าจอดังรูปที่ 3.4 โดยในส่วนของแทป crawl นี้มีการกำหนดค่าดังต่อไปนี้

1. เลือกรูปแบบการครอเลอร์ ซึ่งจะเลือกแบบ the Web เพื่อทำการเก็บข้อมูลของหน้าเพจทั้งหมด
2. เก็บข้อมูลของลิ้งค์ทุกลิ้งค์ที่ออกจากเพจนั้น โดยใช้ hyperlink
3. สำหรับใส่ยูอาร์แอลเริ่มต้นในการเก็บข้อมูล
4. กำหนดค่าความลึกในการค้นหาให้มีค่าเป็น 2 ซึ่งเว็บสฟิงซ์จะเก็บข้อมูลเว็บเพจจากหน้าเริ่มต้นและอีกสองหน้าถัดไปจากลิ้งค์นั้นๆ
5. กำหนดให้ทำการเก็บข้อมูลตามแนวดิ่ง คือจะเก็บข้อมูลเพจเริ่มต้นให้เรียบร้อยก่อน และไปยังลิ้งค์อื่นๆ ตามที่กำหนดค่าความลึก

บันทึกข้อมูลของเว็บเพจที่ทำการครอเลอร์มาได้ โดยคลิ๊กที่แทป Pages เลือก Action เป็น Save และใส่ไดเรกทอรี่ที่ต้องการเก็บข้อมูล ไปที่ช่อง to directory ซึ่งเก็บทุกเพจที่ครอเลอร์ได้ เลือก all pages ใน on pagesดังรูปที่ 3.5



รูปที่ 3.5 กำหนดการใช้งานของแทป Pages

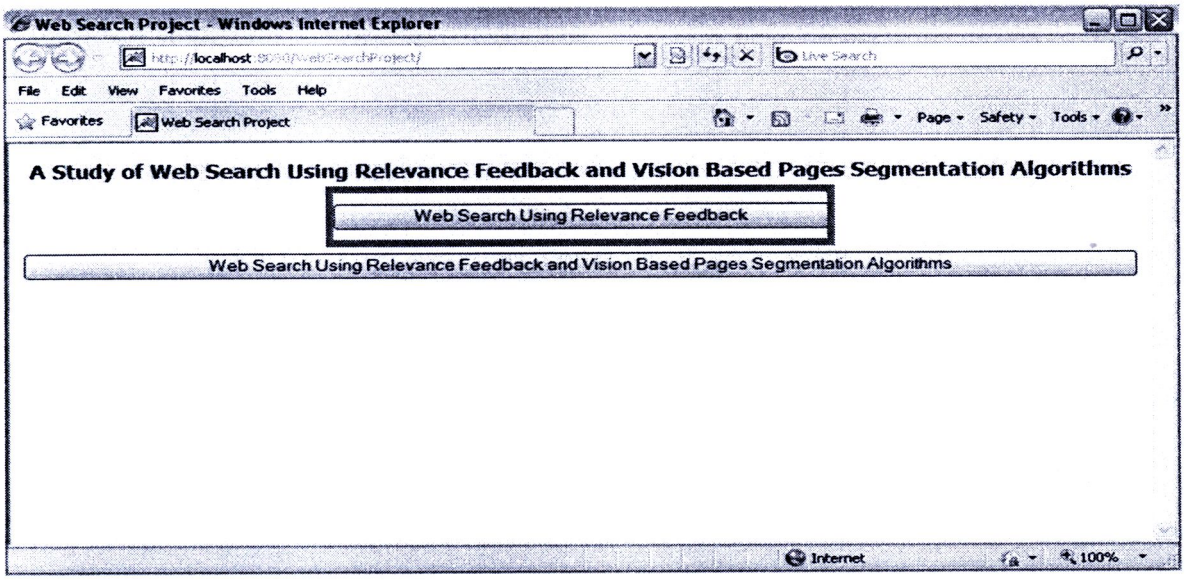


รูปที่ 3.6 แสดงการทำงานของเว็บสฟิงซ์

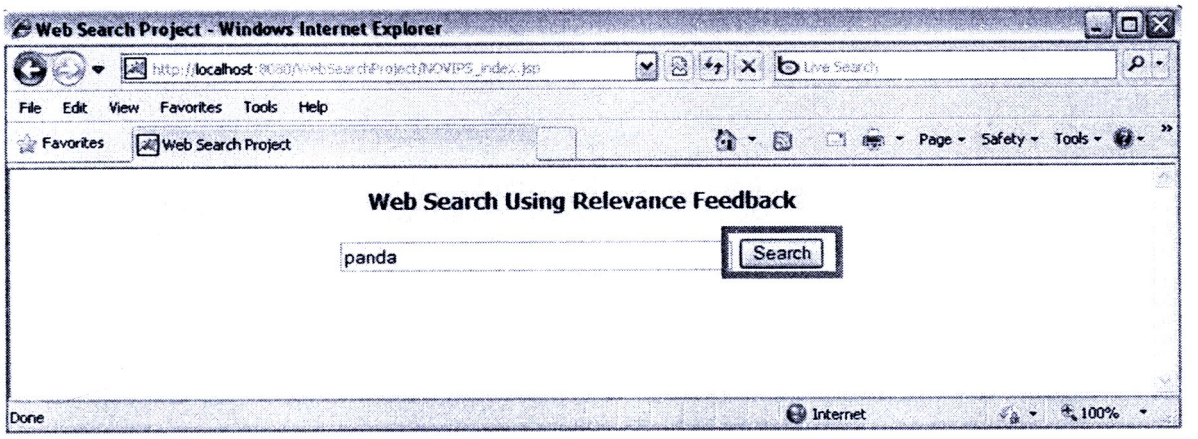
เมื่อกำหนดค่าข้างต้นแล้ว ก็สามารถ กดปุ่ม Start เพื่อเริ่มทำการเก็บข้อมูลเว็บเพจได้เลย โดยแทป Graph จะแสดงเส้นทางที่เว็บสฟิงซ์ได้ทำการเก็บข้อมูล ดังรูปที่ 3.6 ยูอาร์แอลของเว็บที่นำมาเก็บข้อมูลนี้ จะเป็นยูอาร์แอลของเว็บที่เกี่ยวข้องกับเรื่อง panda aids java sushi และ titanic ซึ่งเมื่อทำการครอเลอร์เรียบร้อยแล้วจะต้องนำเพจทุกเพจ มาวิเคราะห์ว่าเกี่ยวข้องกับเรื่องที่กำหนดหรือไม่ โดยให้แต่ละเรื่องมีเพจที่เกี่ยวข้อง 30 เพจจากอย่างน้อย 5 เว็บไซต์

3.3 การทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ

การเริ่มต้นการทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ จะต้องการเลือกปุ่ม “Web Search Using Relevance Feedback” ในหน้าเริ่มต้นของเว็บดังรูปที่ 3.7



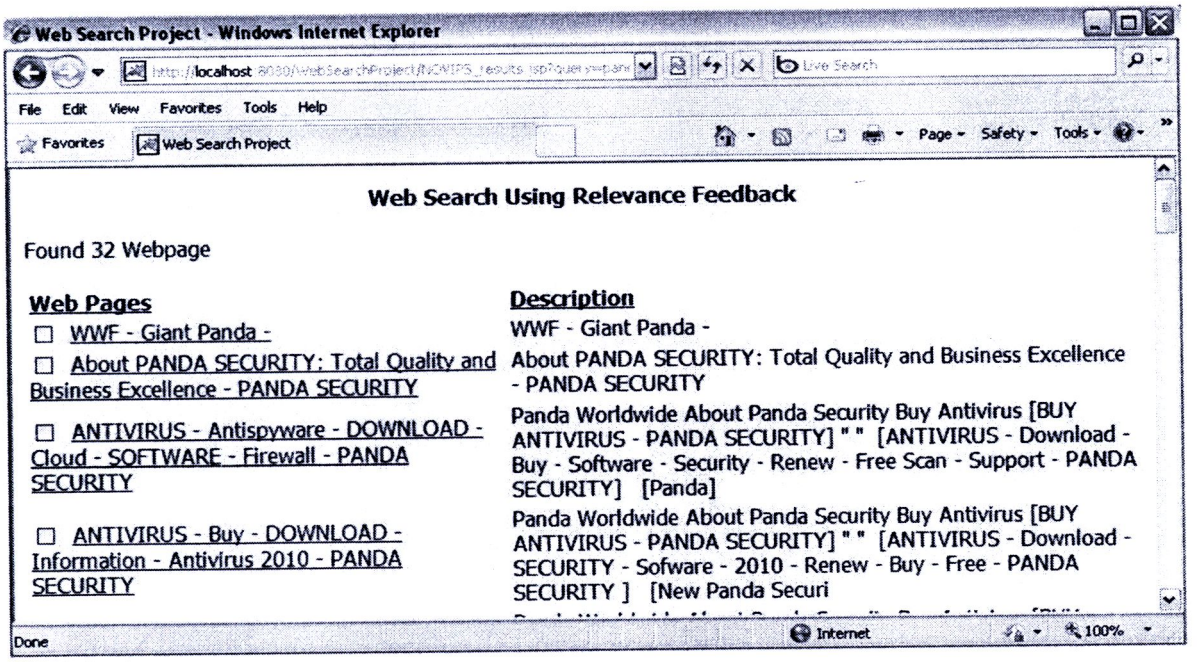
รูปที่ 3.7 หน้าเว็บเพจแรกของระบบจำลอง



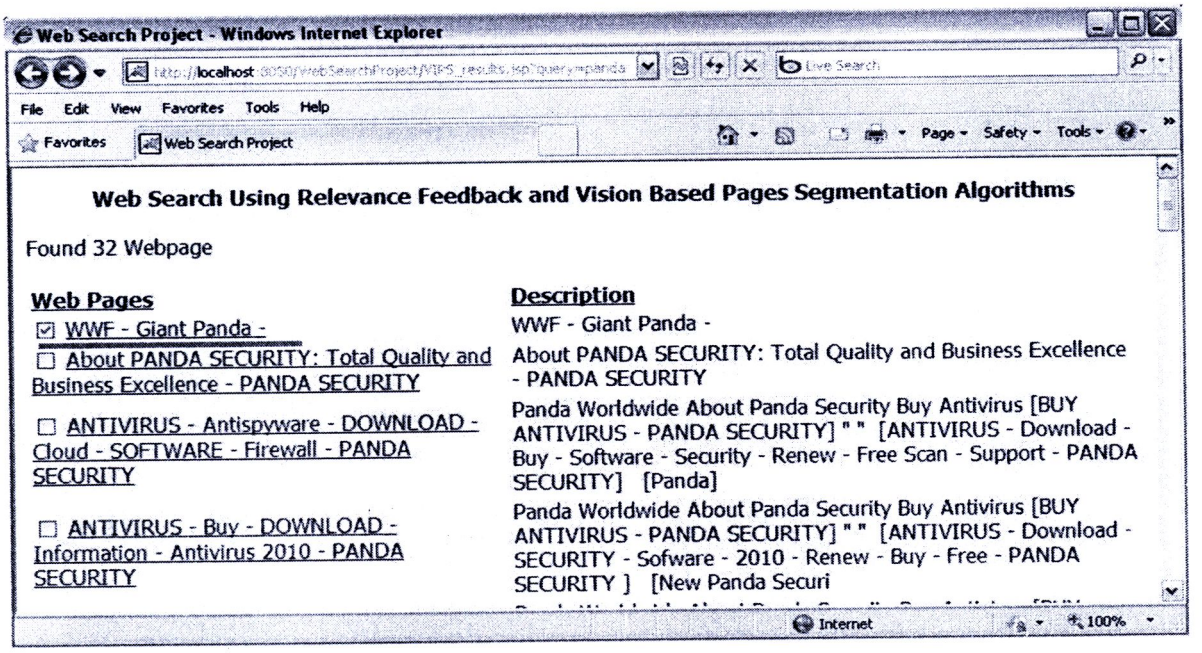
รูปที่ 3.8 หน้าเว็บเพจรับข้อคำถามจากผู้ใช้

3.3.1 ขั้นตอนการทำงานของการค้นคืนสารสนเทศบนเว็บ

ผู้ใช้ใส่ข้อคำถามที่ต้องการทำการค้นคืน ซึ่งระบบจะรองรับเฉพาะข้อคำถามที่เป็นภาษาอังกฤษเท่านั้น แล้วกดปุ่ม Search ดังรูปที่ 3.8 ผู้ใช้ได้ใส่ข้อคำถามคำว่า "panda" เมื่อกดปุ่ม search แล้วก็จะแสดงเว็บเพจที่เกี่ยวข้องกับข้อคำถามออกมาโดย Web Pages คือชื่อเพจ สามารถกดลิ้งค์ไปยังเพจนั้นๆได้ และ Description คือคำอธิบายของเพจนั้นๆ ดังรูปที่ 3.9



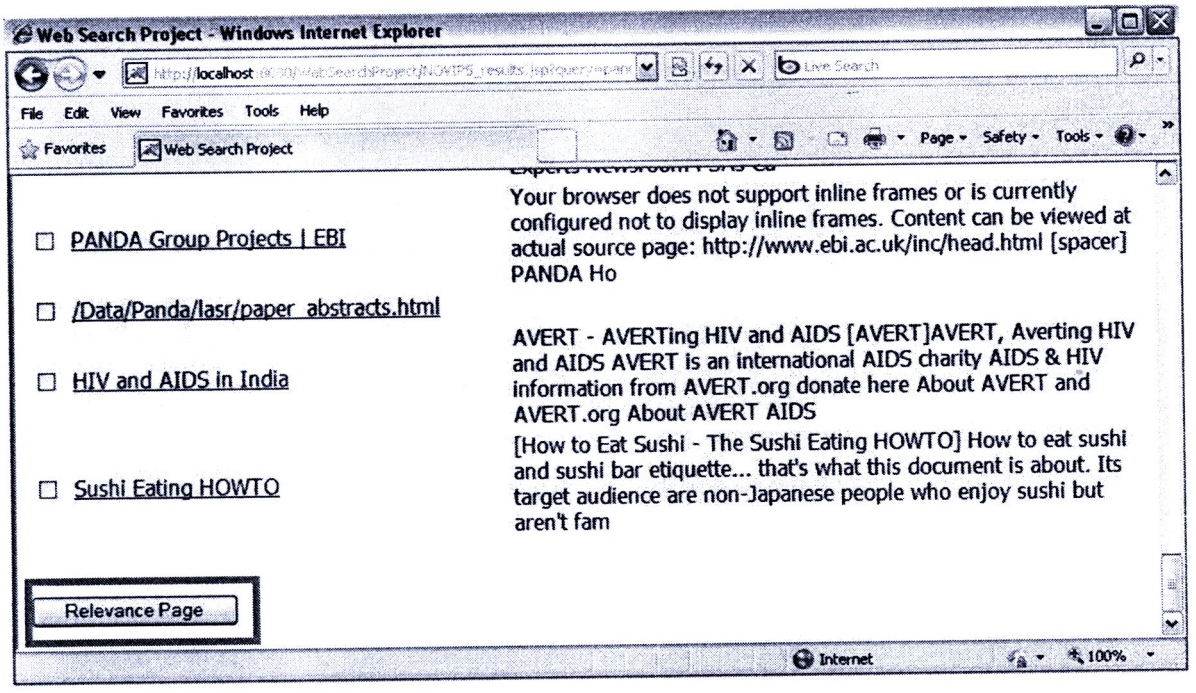
รูปที่ 3.9 ผลการค้นคืนของข้อคำถาม



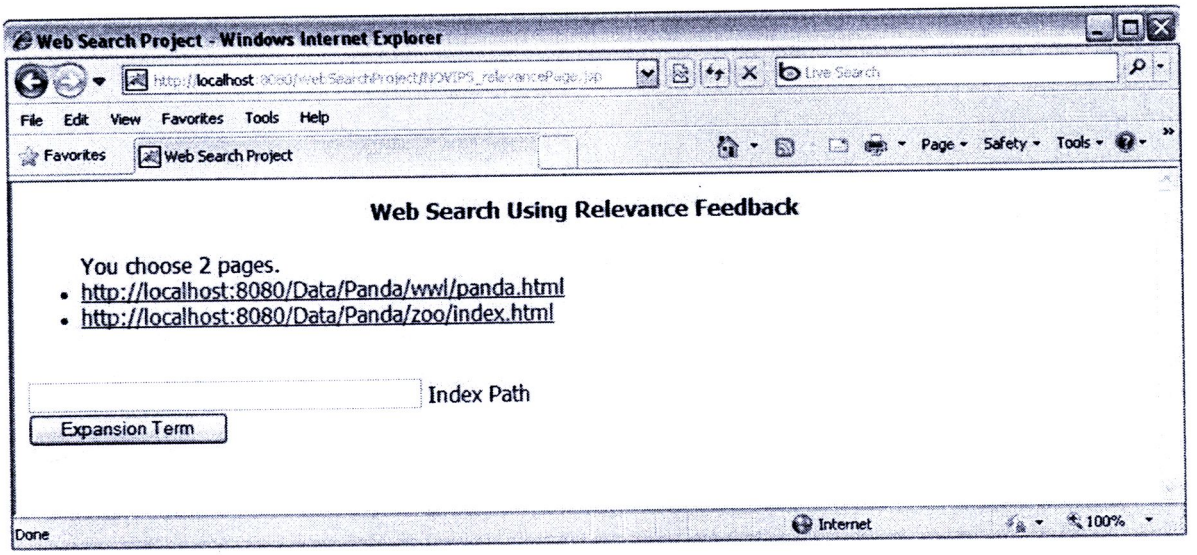
รูปที่ 3.10 หน้าเว็บเพจให้ผู้ใช้เลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ

3.3.2 ขั้นตอนการทำงานของการค้นคืนย้อนกลับ

หากผู้ใช้ไม่พอใจผลการค้นคืนครั้งแรกต้องการทำการค้นคืนอีกครั้ง แล้วระบบจะทำการหาข้อคำถามใหม่ กลับมาให้ผู้ใช้ ซึ่งในขั้นตอนการค้นคืนย้อนกลับนี้ผู้ใช้จะต้องทำการเลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการจากผลการค้นคืนครั้งแรกโดยคลิ๊กปุ่มที่ด้านหน้าของเพจดังรูปที่ 3.10 จากนั้นให้เลื่อนลงมายังด้านล่างของเพจแล้ว กดปุ่ม Relevance Pages ดังรูปที่ 3.11 หลังจากที่ผู้ใช้กดปุ่ม Relevance Page ระบบก็จะแสดงยูอาร์แอลของเพจที่ผู้ใช้เลือกว่าเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ และช่อง "Index Path" สำหรับใส่ไดเรกทอรี่ที่เก็บไฟล์ดัชนีของเว็บที่ผู้ใช้เลือก ดังรูปที่ 3.12



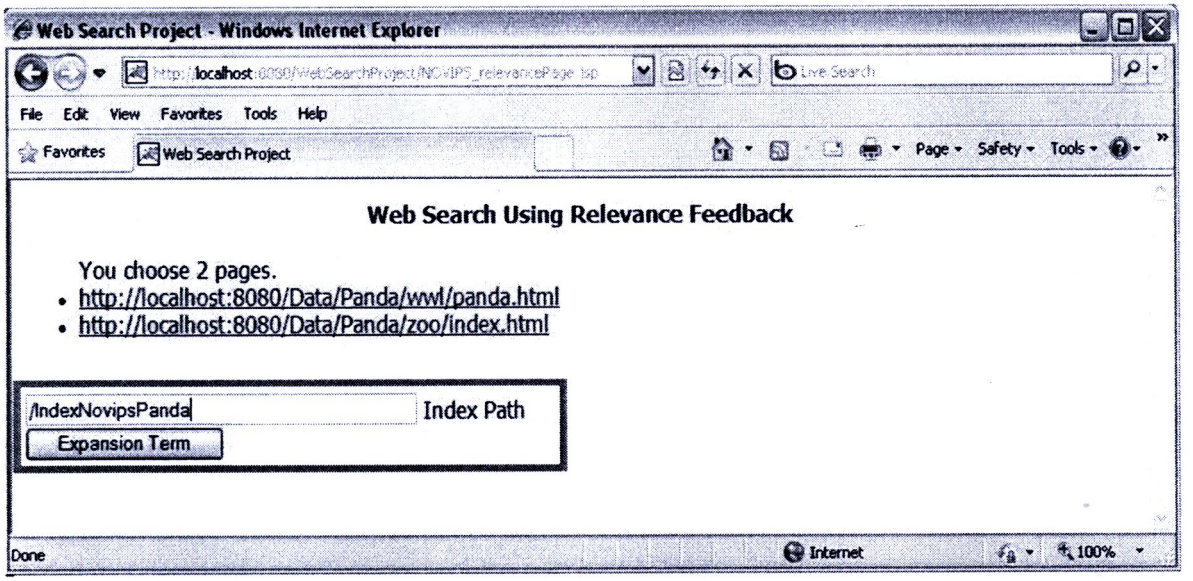
รูปที่ 3.11 หน้าเว็บด้านล่างสุดของเพจที่ทำการค้นคืนได้



รูปที่ 3.12 ผลจากการกดปุ่ม Relevance Page

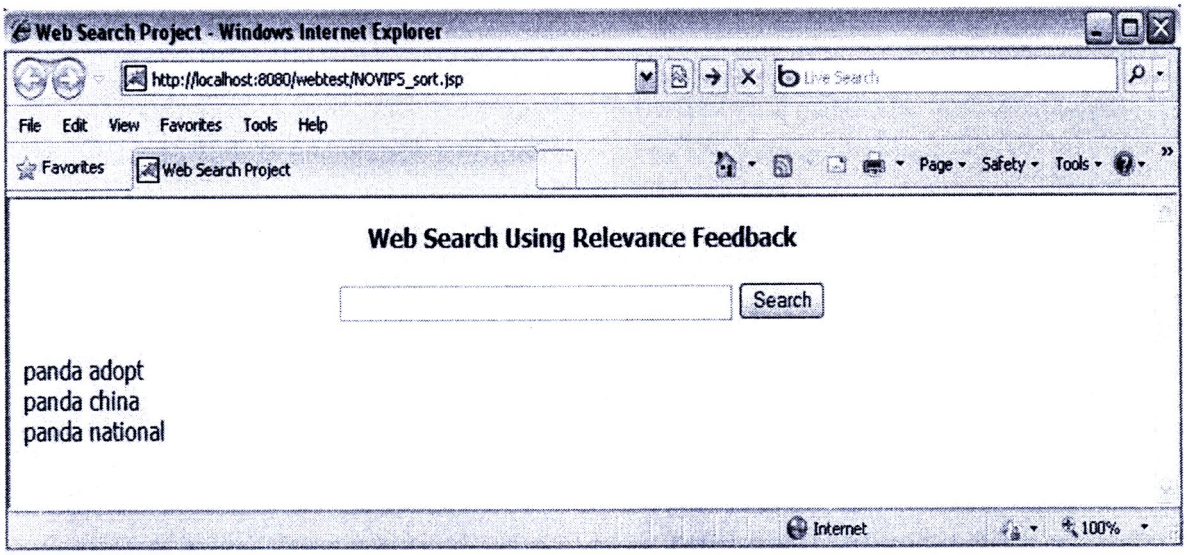
3.3.3 ขั้นตอนการทำงานของการหาข้อคำถามใหม่

ผู้ใช้จะต้องนำเพจที่ผู้ใช้เลือกว่าเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการไปทำเป็นข้อมูลดัชนี และนำไดเรกทอรี่ที่เก็บไฟล์ดัชนีนั้นมาใส่ในช่อง "Index Path" ดังรูปที่ 3.13 ทำการเก็บไฟล์ดัชนีไว้ที่ "/IndexNovipsPanda" เมื่อใส่ไดเรกทอรี่ของไฟล์ดัชนีเรียบร้อยแล้วก็สามารถกดปุ่ม Expansion เพื่อหาข้อคำถามใหม่



รูปที่ 3.13 แสดงยูอาร์แอลที่ผู้ใช้เลือกและการใส่ค่าไดเรกทอรี่เพื่อหาข้อคำถามใหม่

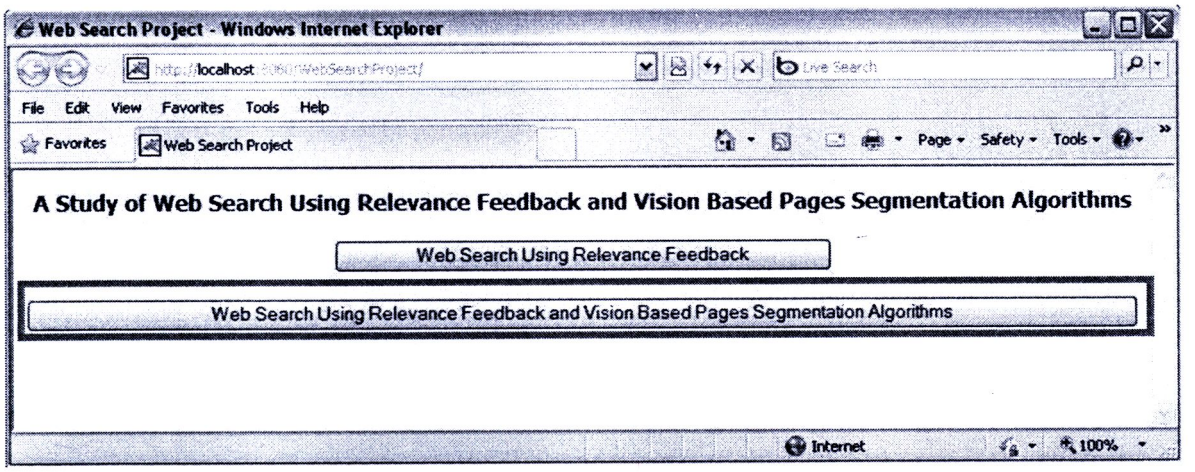
ข้อคำถามใหม่พร้อมให้ผู้ใช้นำกลับไปค้นคืนได้อีกครั้งหนึ่ง ดังรูปที่ 3.14 คำใหม่ที่นำมาเพิ่มในข้อคำถามเดิมเพื่อให้ได้ข้อคำถามใหม่นั้น จะมาจากการเลือกเทอมที่มีความสำคัญในเว็บเพจที่ผู้ใช้ได้ทำการเลือกมา โดยคำนวณตามอัลกอรึทึมการเลือกเทอมของ ดอนน่า ฮาแมน



รูปที่ 3.14 ข้อคำถามใหม่จากเว็บเพจที่ผู้ใช้เลือกว่าเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ

3.4 การทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้ อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับ

การการทำงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับ จะต้องการเลือกปุ่ม "Web Search Using Relevance Feedback" ในหน้าเริ่มต้นของเว็บดังรูปที่ 3.15



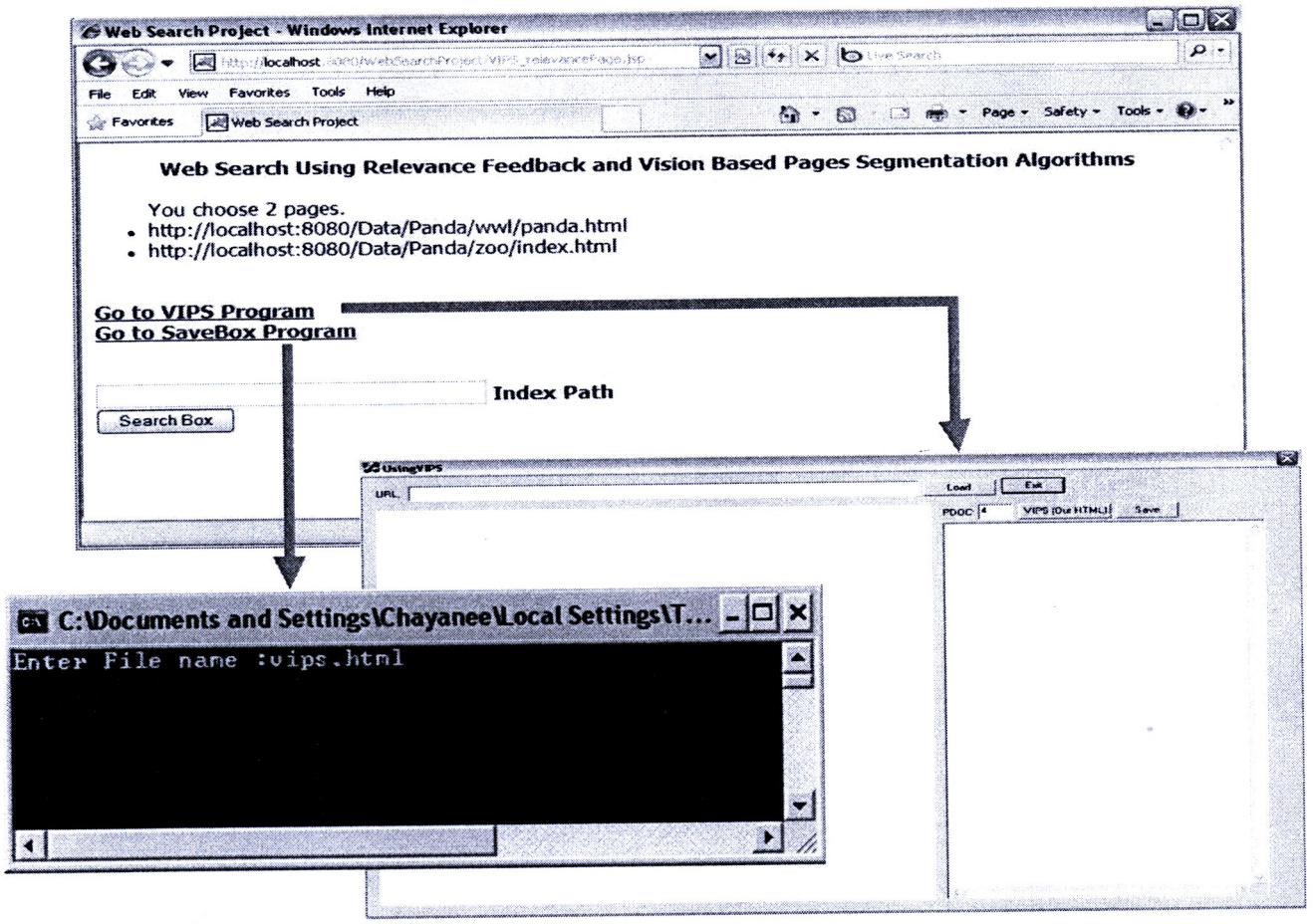
รูปที่ 3.15 หน้าหลักของระบบจำลอง

หลังจากที่ผู้ใช้ได้เลือกแล้ว การทำงานของระบบการงานของระบบค้นคืนสารสนเทศบนเว็บโดยใช้อัลกอริทึมวีไอพีเอสและการค้นคืนย้อนกลับนี้ จะมีการทำงานเช่นเดียวกันกับระบบค้นคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับจนกระทั้งผู้ใช้เลือกเว็บเพจที่เกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ และกดปุ่ม Relevance Page ระบบจะแสดงผลลัพธ์ดังรูปที่ 3.16 โดยระบบจะแสดงยูอาร์แอลของเพจที่ผู้ใช้เลือกว่าเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ จะต้องนำเพจที่ผู้ใช้เลือกเหล่านี้ไปทำการแบ่งเป็นบล็อกโดยใช้โปรแกรมวีไอพีเอส และแยกไฟล์ของแต่ละบล็อกโดยใช้โปรแกรมการบันทึกไฟล์ของแต่ละบล็อก

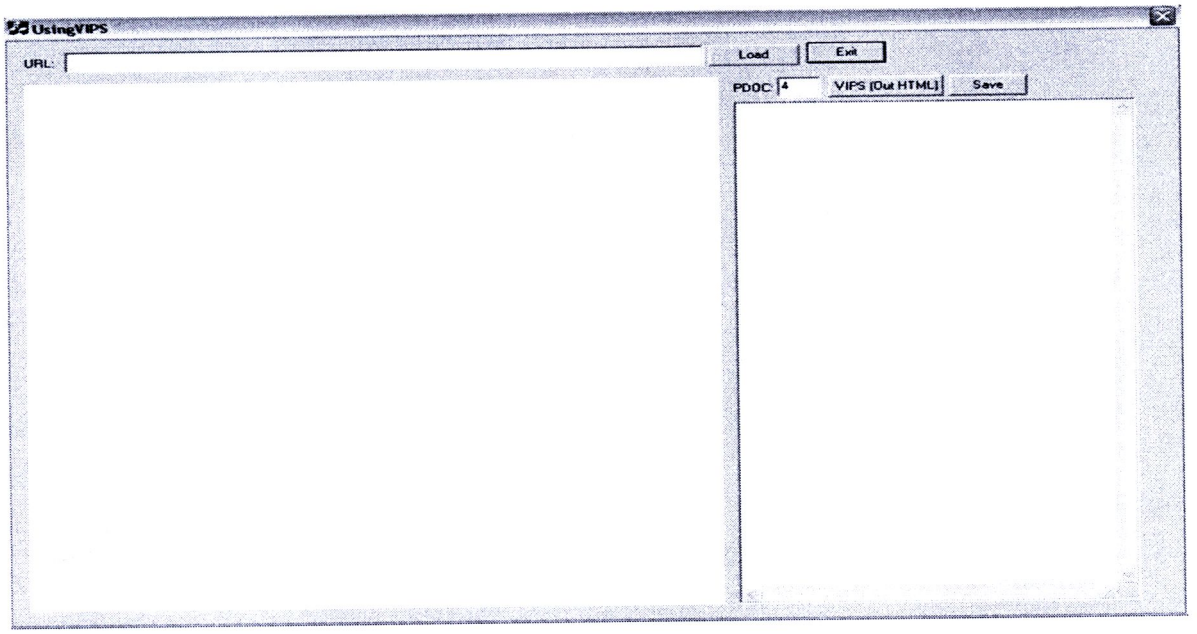
3.4.1 ขั้นตอนการทำงานของโปรแกรมวีไอพีเอส

เมื่อผู้ใช้กดลิ้งค์ "Go to VIPS Program" จะเป็นการเปิดโปรแกรมขึ้นมาและพบกับส่วนติดต่อกับผู้ใช้ของโปรแกรมวีไอพีเอส แสดงดังรูปที่ 3.17 โดยมีการทำงานตามขั้นตอนดังต่อไปนี้คือ

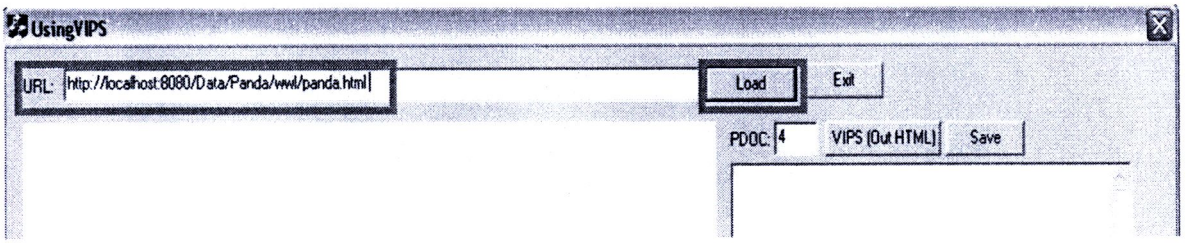
1. ใส่ยูอาร์แอลที่ต้องการแบ่งเป็นบล็อกในช่อง URL และกดปุ่ม Load เพื่ออ่านค่าดังรูปที่ 3.18
2. กำหนดค่าความละเอียดของโครงสร้างเนื้อหาในช่อง PDOC
3. ทำการแบ่งเว็บเพจเป็นบล็อกโดยกดปุ่ม VIPS(Out HTML) จะได้โค้ดเอชทีเอ็มแอลทีมีแท็กวีไอพีเอส แสดงการแบ่งเป็นบล็อก ดังรูปที่ 3.19 จากนั้นกดปุ่ม SAVE เพื่อนำไปทำการแยกไฟล์ของแต่ละบล็อกต่อไป



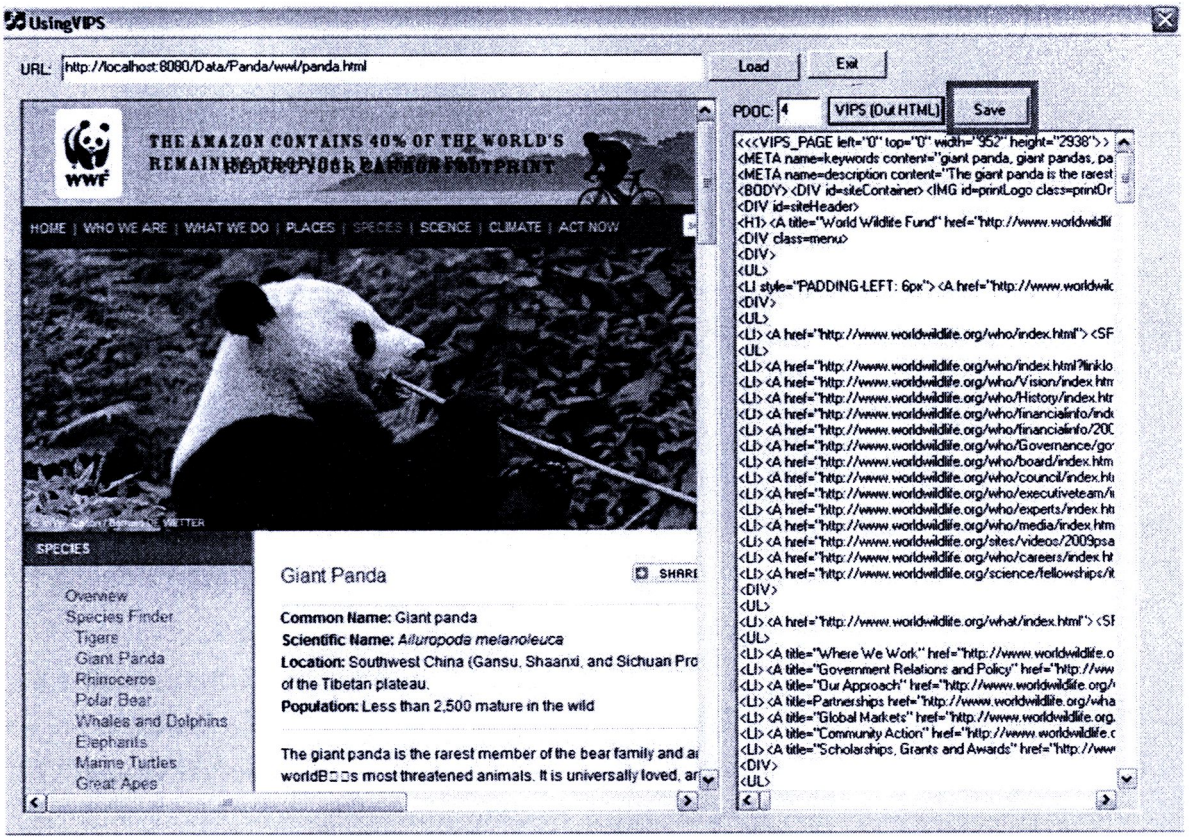
รูปที่ 3.16 แสดงการลิ้งก์ไปยังโปรแกรมวีไอพีเอสและโปรแกรมการบันทึกไฟล์ของแต่ละบล็อก



รูปที่ 3.17 ส่วนติดต่อกับผู้ใช้ของโปรแกรมวีไอพีเอส



รูปที่ 3.18 แสดงการใส่ค่ายูอาร์แอลของเว็บเพจที่ต้องการแบ่งเป็นบล็อก

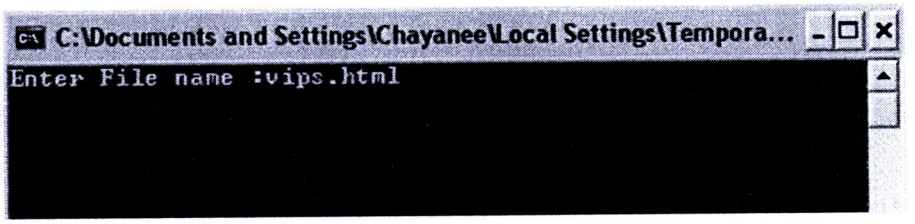


รูปที่ 3.19 ผลของการแบ่งเว็บเพจเป็นบล็อกจากโปรแกรมวีไอพีเอส

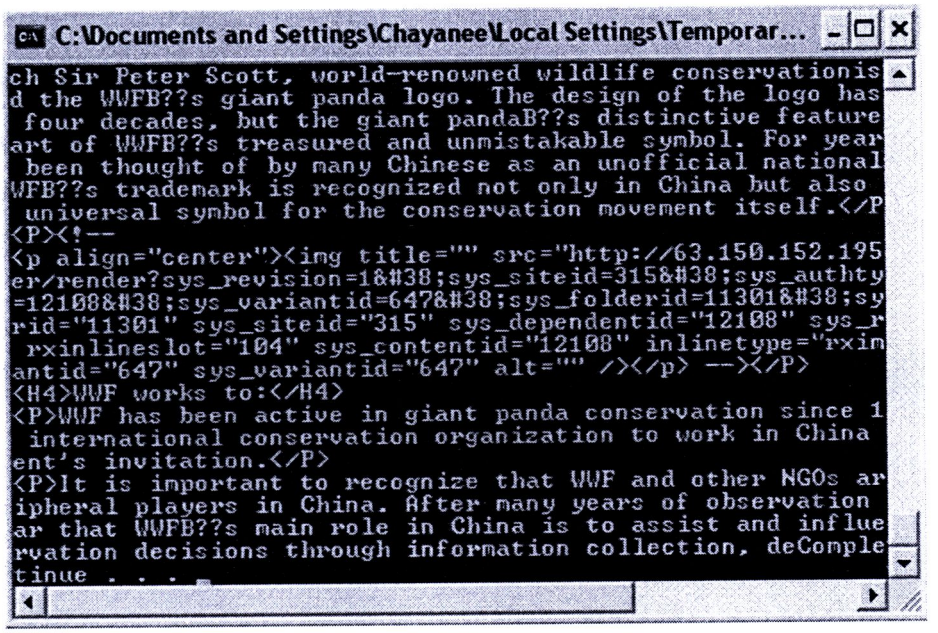
3.4.2 ขั้นตอนการทำงานของโปรแกรมบันทึกไฟล์ของแต่ละบล็อก

จากไฟล์เอชทีเอ็มแอลที่ได้จากโปรแกรมวีไอพีเอสยังไม่ได้ทำการแยกแต่ละบล็อกออกมาเป็นไฟล์จึงใช้โปรแกรมนี้ในการบันทึกไฟล์ของแต่ละบล็อก เมื่อผู้ใช้กดลิ้งค์ “Go to VIPS Program” ก็จะเป็นการเปิดโปรแกรมขึ้นมา โดยมีขั้นตอนการใช้งาน คือ

1. ใส่ชื่อไฟล์ที่ได้จากโปรแกรมวีไอพีเอส ดังรูปที่ 3.20
2. โปรแกรมจะทำการแยกไฟล์เป็นบล็อกตามแท็กที่โปรแกมวีไออพีเอสกำหนด จากนั้นกดปุ่มเอนเทอร์ (Enter) เพื่อจบโปรแกรม ดังรูปที่ 3.21



รูปที่ 3.20 โปรแกรมการบันทึกไฟล์ของแต่ละบล็อกโดยใส่ชื่อไฟล์ที่ต้องการแยกไฟล์ของแต่ละบล็อก



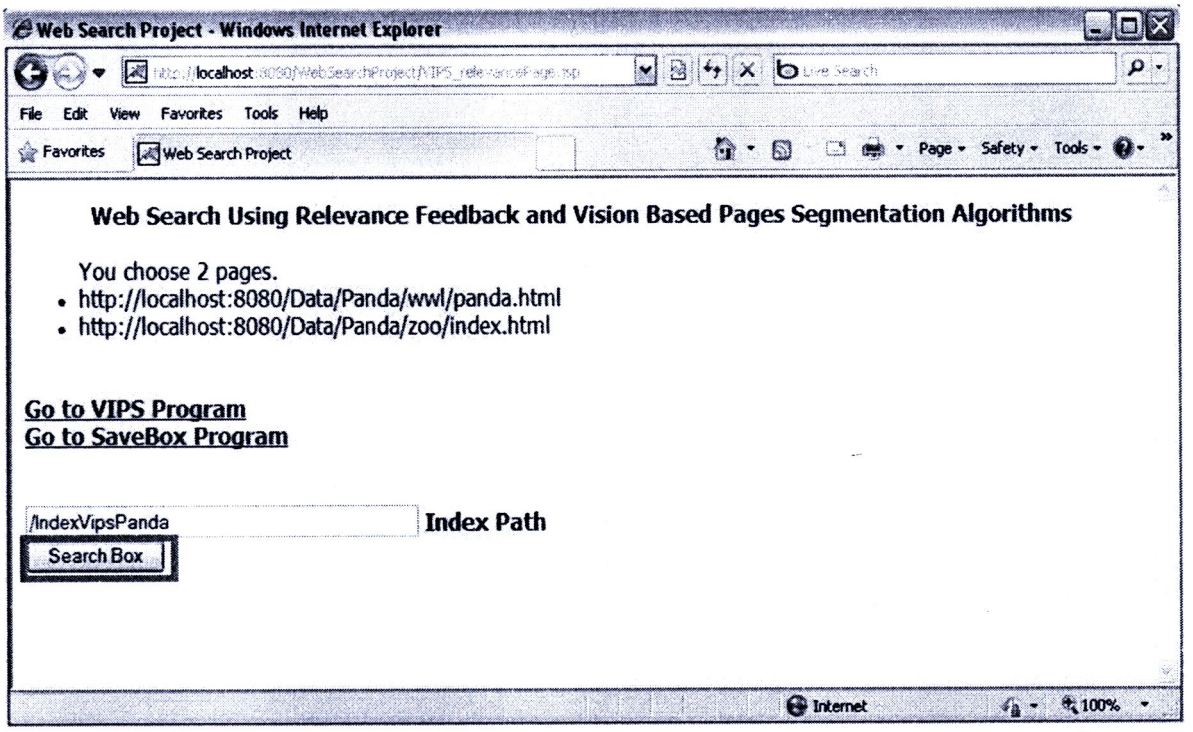
รูปที่ 3.21 การทำงานของโปรแกรมการบันทึกไฟล์ของแต่ละบล็อก

3.4.3 ขั้นตอนการทำงานของการหาข้อคำถามใหม่

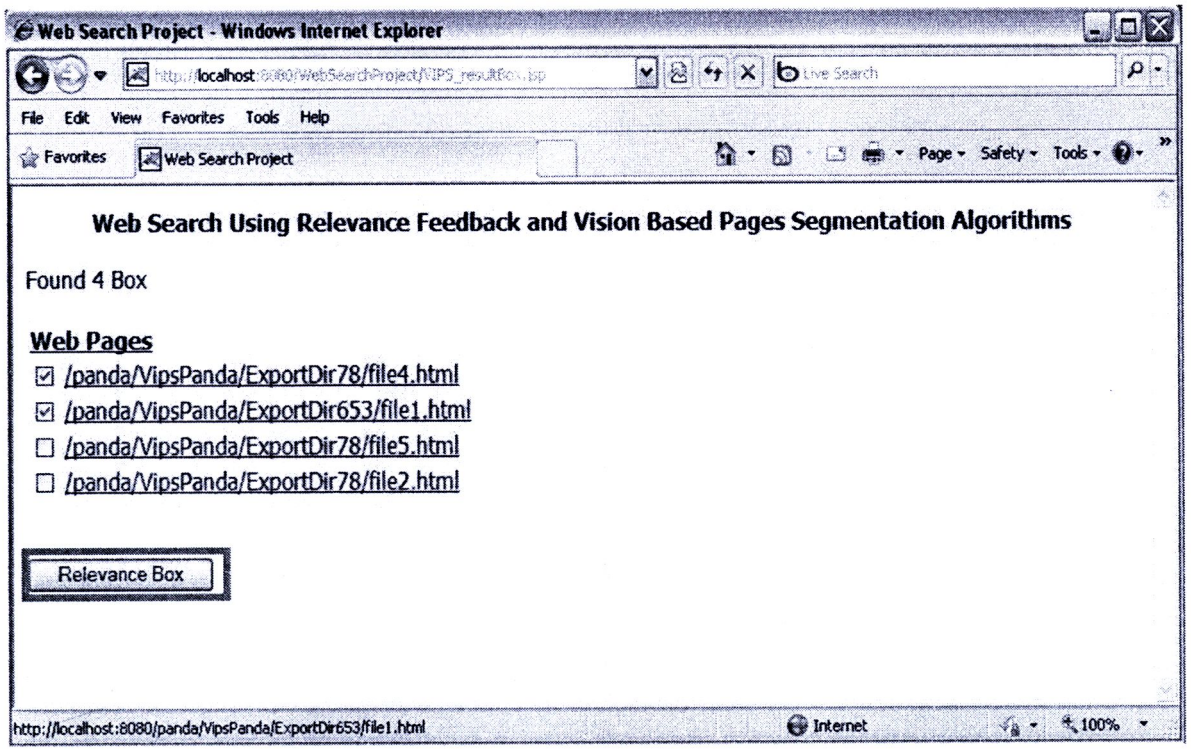
นำบล็อกที่แบ่งได้มาทำเป็นไฟล์ดัชนีและใส่ไดเรกทอรี่ที่เก็บไฟล์นั้นในช่อง "Index Path" ดังรูปที่ 3.22 ไดเรกทอรี่ที่เก็บไฟล์ดัชนีชื่อ "/IndexVipsPanda" จากนั้นกดปุ่ม "Search Box" ระบบจะแสดงบล็อกที่เกี่ยวข้องกับข้อคำถามที่ทำการค้นหาตั้งแต่ต้น ดังรูปที่ 3.23 จากนั้นผู้ใช้จะต้องเลือก บล็อกที่มีข้อมูลเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการจากนั้นกดปุ่ม "Relevance Box" ในที่นี้ได้ทำการเลือกบล็อกที่ 1 และ 2 ว่ามีความเกี่ยวข้องกับสิ่งที่ต้องการ

หลังจากที่ผู้ใช้กดปุ่ม Relevance Box ระบบก็จะแสดงยูอาร์แอลของบล็อกที่ผู้ใช้เลือกว่าเกี่ยวข้องกับสิ่งที่ผู้ใช้ต้องการ จะต้องนำเพจที่ผู้ใช้เลือกเหล่านี้ไปทำเป็นข้อมูลดัชนี และนำไดเรกทอรี่ที่เก็บไฟล์ดัชนีนั้นมาใส่ในช่อง "Index Path" ในที่นี้ทำการเก็บไฟล์ดัชนีไว้ที่ "/IndexBoxPanda" ดังรูปที่ 3.24 เมื่อใส่ไดเรกทอรี่ของไฟล์ดัชนีเรียบร้อยแล้วก็สามารถกดปุ่ม Expansion เพื่อหาข้อคำถามใหม่ ก็จะได้ข้อคำถามใหม่พร้อมให้ผู้ใช้นำกลับไปค้นคืนได้อีกครั้งหนึ่ง ดังรูปที่ 3.25

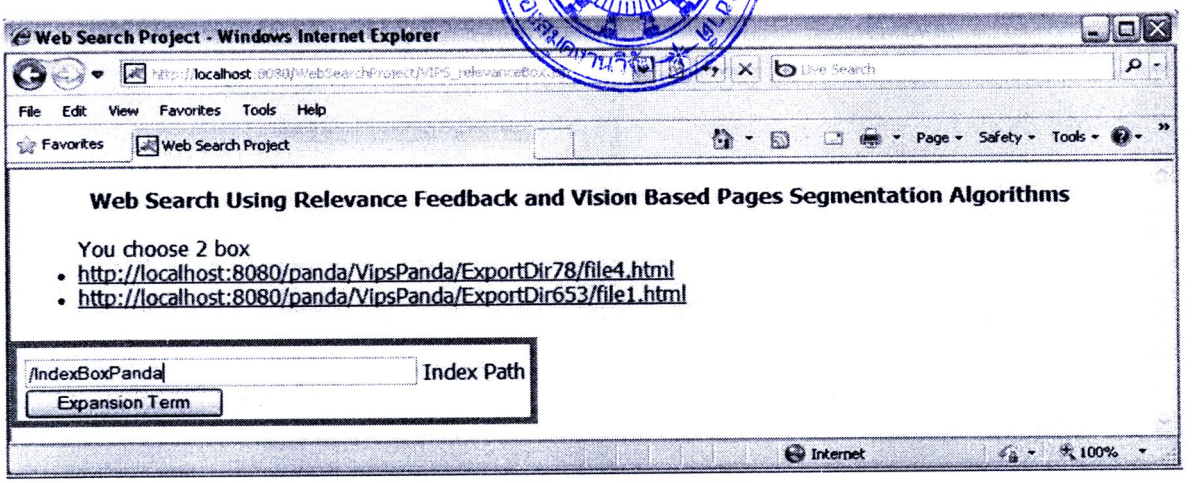
คำใหม่ที่นำมาเพิ่มในข้อคำถามเดิมเพื่อให้ได้ข้อคำถามใหม่นั้นจะมาจากการเลือกเทอมที่มีความสำคัญในเว็บเพจที่ผู้ใช้ได้ทำการเลือกมา โดยคำนวณตามอัลกอรึทึมการเลือกเทอมของ ดอนน่า ฮาแมน เช่นเดียวกับระบบคืนสารสนเทศบนเว็บโดยใช้การค้นคืนย้อนกลับ



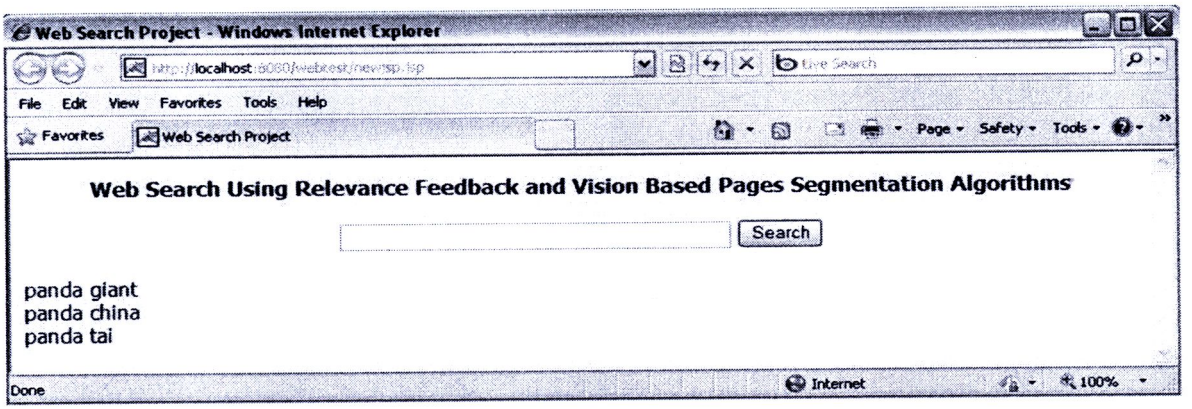
รูปที่ 3.22 การใส่ไดเรกทอรี่ของไฟล์ดัชนีเว็บเพจที่แบ่งเป็นบล็อกแล้ว



รูปที่ 3.23 บล็อกที่เกี่ยวข้องกับข้อคำถามตั้งต้น



รูปที่ 3.24 หน้าเพจหลังจากกดปุ่ม Relevance Box จะแสดงยูอาร์แอลที่ผู้ใช้เลือกและการหาข้อคำถามใหม่

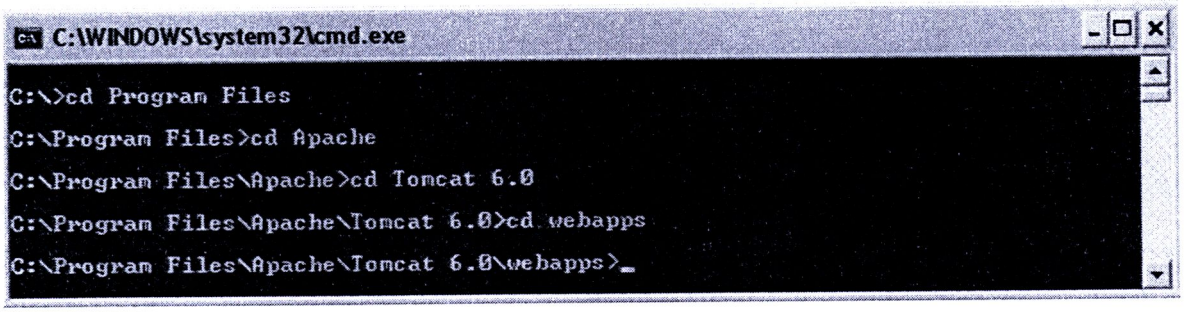


รูปที่ 3.25 ผลการหาข้อคำถามใหม่จากบล็อกที่ผู้ใช้เลือกว่าเกี่ยวข้อง

3.5 การสร้างไฟล์ดัชนี

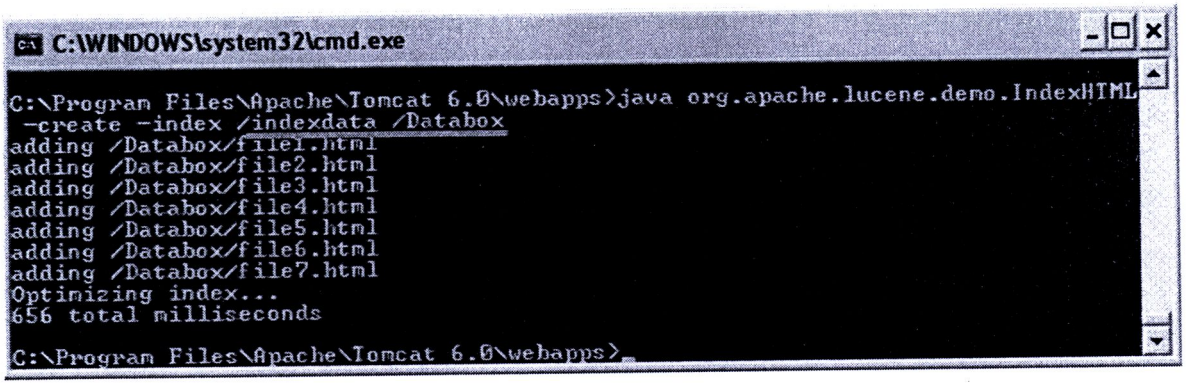
การค้นคืนของลูซีนนั้นจะทำการค้นคืนผ่านทางดัชนีซึ่งเป็นค่าดัชนีที่เป็นตัวแทนเอกสารโดยมีขั้นตอนการสร้างดัชนีคือ

1. เปิด command Prompt
2. เข้าไปยังไดเรกทอรี่ webapps ของ Server ในที่นี้ใช้ Apache Tomcat 6.0 ดังรูปที่ 3.26



รูปที่ 3.26 การเข้าไปในไดเรกทอรี่ของ webapps ของ Apache Tomcat 6.0

3. พิมพ์คำสั่งเพื่อสั่งการสร้างดัชนี คือ "java org.apache.lucene.demo.IndexHTML –create –index ไดเรกทอรี่ที่ต้องการให้เก็บค่าดัชนี ไดเรกทอรี่ที่ต้องการทำไฟล์มาทำเป็นดัชนี ยกตัวอย่างการทำดัชนีของข้อมูลภายในโฟรเดอร์ DataBox ให้ไปเก็บโฟรเดอร์ indexdata ดังรูปที่ 3.27



รูปที่ 3.27 การทำไฟล์ดัชนีของข้อมูลภายในโฟรเดอร์ Databox และนำไปเก็บใน โฟรเดอร์ indexdata