

ระบบรวบรวมข้อมูลจาก Deep Web (Deep Web Crawler)

นางสาวกัลยารัตน์ ศรีชัย
นายประพัฒน์ วิริยะรุ่งเรืองชัย
อาจารย์ที่ปรึกษาโครงการ :
ผศ.ดร.กุลวดี สมบูรณ์วิวัฒน์
ภาควิชาวิศวกรรมศาสตร์
ศรีราชา มหาวิทยาลัย
เกษตรศาสตร์ วิทยาเขตศรีราชา



ที่มาและความสำคัญ

Deep Web เป็นข้อมูลหรือเว็บไซต์ที่ไม่ถูกจัดทำ Index โดย Search Engine มาตรฐานทั่วไป นั่นหมายความว่า ผู้ใช้ทั่วไปจะไม่สามารถค้นหาเว็บไซต์เหล่านั้นได้เจอผ่าน Search Engine แต่ยังสามารถเข้าผ่าน URL ได้ตามปกติ เว็บไซต์ที่จำเป็นต้อง Login ก่อน, เว็บเมล, เพจที่อยู่ด้านหลัง Firewall เป็นต้น ซึ่งส่วนใหญ่จะเก็บข้อมูลเกี่ยวกับเอกสารการศึกษา บันทึกการแพทย์ เอกสารกฎหมาย

ผู้จัดทำเห็นความมีประโยชน์ของ Deep Web จึงต้องการจัดทำโครงการนี้ขึ้นมา

เป้าหมายของโครงการ

- 1.ออกแบบและพัฒนาระบบ Deep Web โดยใช้ภาษา Python
- 2.พัฒนา Web Site ที่สามารถเรียกดูข้อมูลจาก Database

ประโยชน์ของข้อมูลที่ได้จาก Deep Web

ข้อมูลที่ได้จาก Deep Web สามารถนำข้อมูลมารวบรวมสร้างขึ้นเป็น Web Portal ได้ไม่ว่าจะเป็นข้อมูลเกี่ยวกับร้านอาหาร เพลง หนังสือ และซีรี่ย์ภาพยนตร์

- 1.ข้อมูลที่ได้เกี่ยวกับร้านอาหาร เราสามารถนำข้อมูลเหล่านั้น มาแยกหมวดหมู่ร้านอาหาร
- 2.ข้อมูลที่ได้เกี่ยวกับเพลง เราสามารถนำข้อมูลเหล่านั้น มาจัดสรรแยกหมวดหมู่
- 3.ข้อมูลที่ได้เกี่ยวกับหนังสือ เราสามารถนำข้อมูลเหล่านี้มาแยกหมวดประเภทต่าง ๆ
- 4.ข้อมูลที่ได้เกี่ยวกับซีรี่ย์ภาพยนตร์ สามารถนำข้อมูลมาแยกประเภทของภาพยนตร์

ขั้นตอนการพัฒนาระบบ

- 1.หา code basic ของ crawler มาเป็นขั้นเริ่มต้น
- 2.ทำความเข้าใจการทำงานของโค้ด และเขียนออกมาเป็นflowchart
3. ทำการ Crawler ข้อมูลในเว็บไซต์ และดึงข้อมูลเก็บลงในดาต้าเบส
- 4.เขียนโปรแกรมให้สามารถค้นหาคำค้นหาจากไฟล์ได้แบบอัตโนมัติ และดึงข้อมูลเก็บลงในดาต้าเบส
- 5.สร้างโครงสร้างดาต้าเบส
- 6.สร้าง Class Diagram
- 7.พัฒนาเว็บไซต์เพื่อแสดงข้อมูลในดาต้าเบส

วิธีการทำงานของระบบ

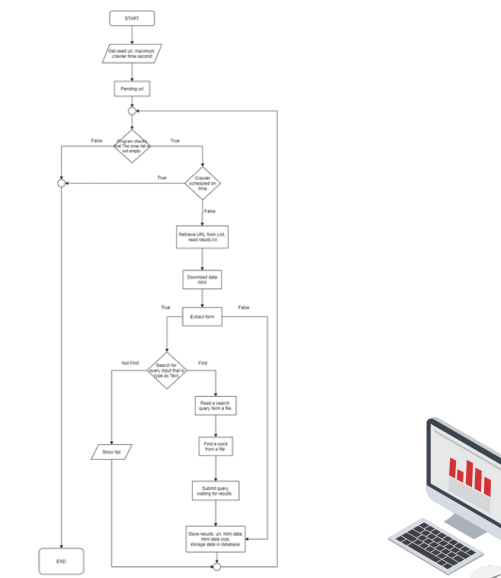
ขั้นตอนการทำงานของ Deep Web Crawler

- 1.ใส่ URLและเวลา(ชั่วโมง) ที่ต้องการ Crawl
- 2.โปรแกรมจะทำการ Crawl url, data ของ html, size ของ html, form ของ html, วันเวลาที่ Crawl เก็บลงในดาต้าเบส
- 3.โปรแกรมจะหยุดทำงานตามเวลาที่ใส่ไว้ตามขั้นตอนที่ 1 หรือ ไม่มีข้อมูลให้ Crawl

ขั้นตอนการทำงานของเว็บไซต์สำหรับแสดงข้อมูลที่ Crawl

1. ใส่ URL, สิ่งที่เราต้องการค้นหา เช่น เพลง , <https://www.sanook.com>
2. เว็บจะนำข้อมูลจากดาต้าเบสมาแสดงบนหน้าเว็บไซต์

Flowchart อธิบายการทำงานของ Deep Web Crawler



สรุป

Deep Web เป็นเว็บไซต์ที่ผู้ใช้ทั่วไปจะไม่สามารถค้นหาเว็บไซต์เหล่านั้นได้เจอผ่าน Search Engine แต่ยังสามารถเข้าผ่าน URL ได้ตามปกติ ประโยชน์ของ Deep Web นอกจากจะสามารถเข้าถึงได้เฉพาะบุคคล บุคคลขององค์กรนั้นๆ เพื่อความปลอดภัยของฐานข้อมูลของผู้ใช้, หรือจะเก็บข้อมูลเกี่ยวกับเอกสารการศึกษา บันทึกการแพทย์ เอกสารกฎหมาย รายงานทางวิทยาศาสตร์ ข้อมูลรัฐบาลแล้ว ประโยชน์ของมันยังสามารถที่จะรวบรวมข้อมูลต่าง ๆ เพื่อสร้างขึ้นเป็น Web Portal ได้เพื่อความสะดวกสบายแก่ผู้ใช้งาน

การจัดทำโครงการนี้ขึ้น เพื่อสามารถนำประโยชน์ของ Deep Web Crawler มาใช้ให้เกิดประโยชน์สูงสุด และเพื่ออำนวยความสะดวกสบายของมนุษย์

สรุป

