MIKE Lucene Search Engine Prototype

From MikeWiki

Contents

- 1 สาระสำคัญ
- 2 หลักการและเหตุผล
- 3 วัตถุประสงค์
- 4 ปัญหาหรือประโยชน์ที่เป็นเหตุผลให้ควรพัฒนาโปรแกรม
- 5 เป้าหมายและขอบเขตของโครงการ
- 6 รายละเอียดของการพัฒนา
  - 6.1 ส่วนของโปรแกรมที่ทำการพัฒนา
  - 6.2 การวัดประสิทธิภาพ
  - 6.3 ข้อมูลเข้า
  - 6.4 ข้อมูลออก
- 7 Progress
  - 7.1 @Jan 2009

สาระสำคัญ

> การที่มีคลังข้อมูลปริมาณมากๆ จะไม่เป็นประโยชน์เลยถ้าไม่สามารถค้นหาเพื่อนำไปใช้งานได้อย่างสะดวก ในขณะที่ข้อมูลบนโลกมีปริมาณมากขึ้นในอัตราเร็วที่สูงขึ้นตลอดเวลา

รวมถึงความต้องการในการบริโภคข้อมูลข่าวสารที่มีความเป็นปัจจุบันที่สุด ดังนั้นการพัฒนาเครื่องมือเพื่อช่วยด้านการสืบค้นข้อมูลที่มีประสิทธิภาพสูง ใช้เวลาประมวณผลต่ำ และตอบสนองอย่างรวดเร็ว จึงเป็นสิ่งที่มีความจำเป็นอย่างหลีกเลี่ยงไม่ได้ทั้งในยุคปัจจุบันและในอนาคต โครงการนี้ได้เลือกพัฒนาระบบสืบค้นข้อมูลระบบหนึ่ง ที่เรียกว่า Search Engine ซึ่งปัจจุบันเป็นสิ่งที่เข้ามาอยู่ในชีวิตประจำวันเสียแล้ว Search Engine นี้เองเป็น ระบบสืบค้นข้อมูลที่แทนที่ระบบสืบค้นแบบเก่าๆ เนื่องด้วยมีความสามารถในการจัดการข้อมูลปริมาณมากและหลากหลาย แต่ความสามารถย่อมมีขีดจำกัด ในโครงการนี้จะเน้นศึกษากรณีข้อมูลมากเกินกว่าจะใช้งานบน Search Engine ธรรมดาทั่วๆไปได้ โดยพัฒนาขีดความสามารถของ Search Engine

หลักการและเหตุผล

> พิจารณาระบบ Search Engine ใดๆ เมื่อปริมาณข้อมูลที่ต้องจัดการมีปริมาณสูงเกินกว่าหน่วยประมวลผลที่มีสมรรถณะสูงหน่วยหนึ่งของระบบนั้น จะทำงานได้สำเร็จในระยะเวลาอันสั้น

แนวทางการแก้ปัญหาหนึ่งคือการเพิ่มขนาดและขีดความสามารถในหน่วยการประมวลนั้นๆ แต่นั่นไม่ใช่วิธีที่ควรใช้ในการแก้ไขปัญหานี้เท่าไรนัก เนื่องจากอัตราการเพิ่มขึ้นของข้อมูลมีสูงมาก และการตอบสนองต่อความต้องการของผู้ใช้งานอยู่ในระดับสูง การแก้ปัญหาด้วยวิธีดังกล่าวจึงอาจส่งผลให้ถึงทางตันในเวลาอันใกล้ จึงต้องค้นคว้าหาวิธีการเพิ่มขีดความสามารถในรูปแบบอื่น ในโครงการนี้จะศึกษาถึงกรณีการเพิ่มจำนวนหน่วยประมวลผล ซึ่งสามารถเพิ่มได้เรื่อยๆ ทำให้ขีดจำกัดของระบบอยู่สูงกว่าการแก้ไขปัญหาในรูปแบบแรก

วัตถุประสงค์

> ศึกษาและพัฒนาระบบ Search Engine ที่มีความสามารถดำเนินการจัดการกับข้อมูลที่มีปริมาณมากเกินกว่าที่คอมพิวเตอร์เครื่องหนึ่งๆจะประมวลผลและตอบสนองได้

ในเวลาอันจำกัด อย่างมีประสิทธิภาพ

ปัญหาหรือประโยชน์ที่เป็นเหตุผลให้ควรพัฒนาโปรแกรม

> เมื่อการสืบค้นข้อมูลเป็นสิ่งสำคัญในชีวิตประจำวัน แต่อัตราการเพิ่มของประมาณข้อมูลบนเครือข่ายอินเตอร์เนตเพิ่มขึ้นตลอดเวลา ความต้องการในการใช้งานมีมาก

แต่ละผู้ใช้งานต้องการการตอบสนองต่อการใช้งานอย่างรวดเร็ว เป็นการยากที่จะสร้างระบบแบบทั่วๆไปในการบริหารจัดการให้เกิดประสิทธิภาพสูงสุดทั้งสามสิ่ง ดังนั้นการศึกษาค้นคว้าพัฒนาระบบสืบค้นข้อมูลที่มีสมรรถภาพสูงขึ้นจึงเป็นสิ่งพึงกระทำ

เป้าหมายและขอบเขตของโครงการ

> การพัฒนาแบ่งเป็น 4 ช่วง โดยช่วงแรก เป็นการศึกษาการทำงานของ Lucene Library บนระบบปฏิบัติการ UNIX โดยใช้ชุดข้อมูลสืบค้นที่มีปริมาณไม่มากนัก

และเป็น Single Language จากนั้นทำการเพิ่มปริมาณชุดข้อมูลขึ้นเรื่อยๆ แล้วทำการวัดประสิทธิภาพ ในช่วงที่สองของการพัฒนา จะเน้นการศึกษาในเรื่องการประมวลผลการทำงานแบบ Multi Language ช่วงที่สามเป็นการศึกษาค้นคว้า การแก้ปัญหากรณีที่ปริมาณข้อมูลมากเกินกว่าจะใช้คอมพิวเตอร์เพียงเครื่องหนึ่งจะสามารถประมวลผลได้ดีในเวลาอันจำกัด ในการพัฒนาช่วงที่สามนี้จะศึกษาเรื่องการแก้ปัญหาโดยใช้ระบบเครือข่าย ร่วมกัน ประมวลผล ศึกษาถึงวิธีการ ข้อดีและข้อเสียของการติดต่อระหว่างเครื่องแบบต่างๆ ช่วงสุดท้ายของโครงการจะทำการศึกษาในเรื่องการนำระบบไปทำงานบนระบบเครือข่ายการประมวลผลขนาดใหญ่ เพื่อไปสู่การทำงานบนประสิทธิภาพสูงสุด โดยช่วงที่สามและสี่จะศึกษาบนสมมติฐานที่ว่าจำนวนเครื่องมากขึ้นจะทำให้ความเร็วในการประมวลผลสูงขึ้น

รายละเอียดของการพัฒนา

ส่วนของโปรแกรมที่ทำการพัฒนา

1. ส่วนของการแยกข้อความออกจาก html code ที่ได้รับจาก Spider
2. ส่วนของการทำ Indexing และ Analyzing ข้อความจากข้อแรก โดยใช้ Lucene
3. Web สำหรับการ Searching เขียนโดยใช้ JSP

โดยการพัฒนาโปรแกรมในแต่ละช่วงของโครงการ จะดำเนินการพัฒนาทั้งสามส่วน ใช้ Lucene Library ซึ่งเป็น Library บนภาษา Java และทำงานบนระบบปฏิบัติการ Linux

การวัดประสิทธิภาพ

- ช่วงแรกและช่วงที่สองของโครงการ
  - โปรแกรมสามารถทำงานได้ถูกต้อง และสามารถใช้งานได้
- ช่วงที่สามและสี่
  - การใช้เวลาในการทำ Indexing ต่อจำนวนเครื่องที่ใช้ ลดลงตามสมมติฐาน
  - เวลาที่ใช้ในการ Search ต่อจำนวนเครื่องที่ใช้ ลดลงตามสมมติฐาน

ทั้งนี้รูปแบบการเชื่อมต่อที่ต้องทำการออกแบบ เป็นตัวแปรที่มีผลต่อประสิทธิภาพเช่นกัน

ข้อมูลเข้า

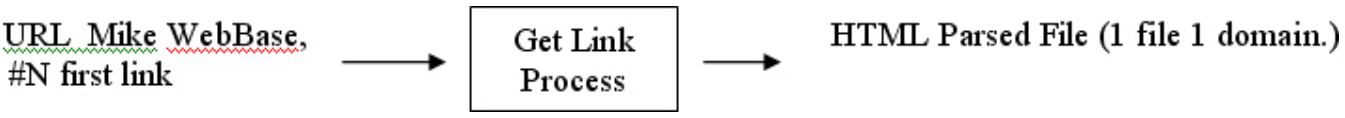
- File จาก Web Spider

ข้อมูลออก

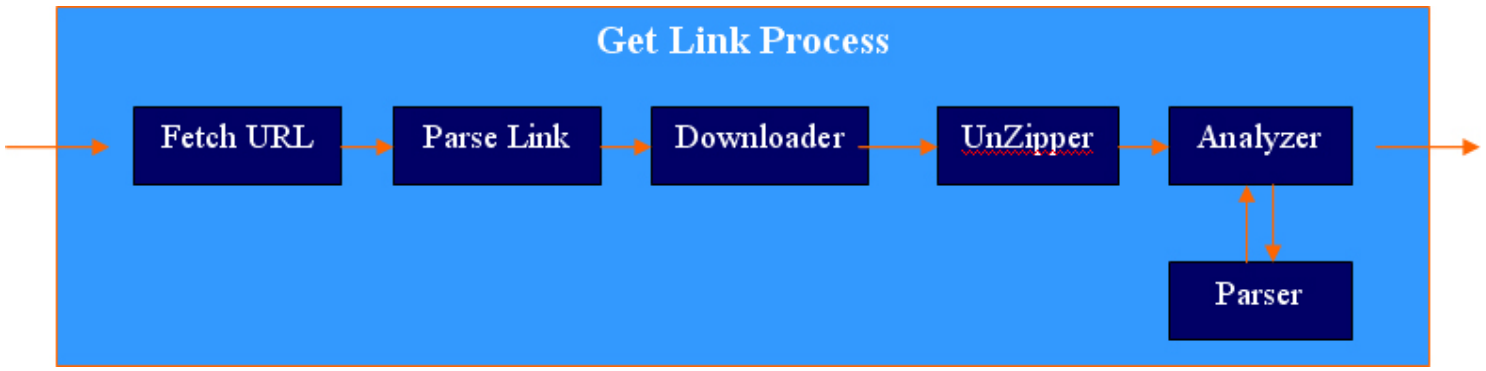
- หน้า Web สำหรับทำการสืบค้นข้อมูล โดยผลการค้นหาประกอบด้วย url และ snippet เรียงลำดับตามความสำคัญ

Progress

@Jan 2009



นำ URL ในรูปแบบ Mike WebBase เข้าเป็น Input



การทำงาน ของ Process คือ Fetch url input มาเป็น raw html file จากนั้น ใช้ Jericho Html Parser ทำการParse เก็บเฉพาะส่วนที่เป็น Link เท่านั้นออกมาเป็น Text File

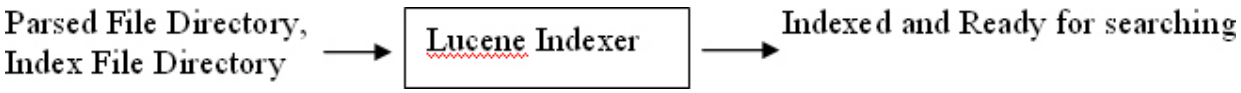
ที่มี List of Link จากนั้นส่งให้ Downloader (โดยส่ง Directory Path และ Path ของ List of Link) ทำการ Download ลงมาที่เครื่องจะได้file .gz เมื่อครบแล้ว จะทำการ Unzip File ทั้งหมด โดย Unzipper (ส่งเฉพาะ Directory Path) ซึ่ง File ที่ unzip มาได้จะเป็น file ที่ได้จาก spider ซึ่งมี รูปแบบ (format) ตายตัว ให้ Analyzer ทำการแยกเป็น pageๆ ซึ่งแต่ละ page เป็น html code ทำการเขียนลง temp file แล้วนำไป Parse เก็บเฉพาะ Url Title และ Body เขียนต่อท้าย File Output เป็น text file

- รูปแบบของ File Output คือ

  url
  title
  body

  url
  title
  body

  ...



สร้าง List File ที่จะนำไปทำ Index ก่อน โดย Traverse Directory จนครบ

การทำงาน นำชื่อ File ออกจาก List เพื่อทำ index ที่ละ file โดยรูปแบบ File นั้น เป็น url title body ต่อๆกัน จึงต้อง อ่าน file ทีละบรรทัด 1 Lucene doc แยกเป็น field 3 field คือ url, body, title (แยก 3 field สำกรับการ weight ในอนาคต) ทุกครั้งที่ add ทั้ง 3 field ลง Lucene doc แล้ว ทำการ write doc ลง จากนั้นอ่าน file นั้นต่อ เมื่อทำงานจนจบ 1 file ทำการ Optimize Index 1 ครั้ง แล้วทำซ้ำจนกระทั่งรายชื่อFile หมด List