ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

Information Extraction สำหรับร้านหนังสือออนไลน์

Information Extraction for Online Book Store

โดย

นาย มนต์ชัย พจนาสมสมาน

รหัส 42067012

อาจารย์ที่ปรึกษา

ดร. โชติพัชร์ ภรณวลัย

วัน เดือน ปี 19 ม.ค. 2550
เลขทะเบียน 01855
เลขเรียกหนังสือ วพ. ม1460 2544
"ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล."

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 1 ปีการศึกษา 2544

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

H001855

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อหัวข้อ	Information Extraction สำหรับร้านหนังสือออนไลน์
นักศึกษา	นายมนต์ชัย พจนาสมสมาน
อาจารย์ที่ปรึกษา	ดร. โชติพัชร์ ภรณวลัย
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2544

บทคัดย่อ

ปัจจุบันร้านหนังสือออนไลน์ มีอยู่มากมายหลายแห่งซึ่งในแต่ละแห่งราคาหนังสือและการบริการจะมีความแตกต่างกัน ทำให้การค้นหาหนังสือที่มีการบริการที่ดีและมีราคาที่ถูกเป็นไปได้ด้วยความยากลำบาก

สำหรับโครงงาน Information Extraction สำหรับร้านหนังสือออนไลน์นี้เป็นโครงงานพัฒนา Web application ที่ทำหน้าที่เป็นตัวกลางในการค้นหาหนังสือโดยเปรียบเทียบราคาและการบริการของร้านหนังสือออนไลน์ต่าง ๆ โดยนำวิธีตัดส่วนข้อมูล (Information Extraction) ซึ่งเป็นวิธีการที่มีความสามารถในการตัดส่วนของข้อมูลอย่างมีประสิทธิภาพ มาช่วยในการตัดส่วนข้อมูลราคา ทำให้การค้นหาเพื่อเลือกซื้อหนังสือทางอินเตอร์เน็ตมีความสะดวกสบายมากยิ่งขึ้น และในการพัฒนาระบบเราได้ปรับปรุงการตัดส่วนข้อมูล โดยจะทำการหากลุ่มของข้อมูลก่อนการตัดส่วนข้อมูลจริงซึ่งผลที่ได้ ทำให้ประสิทธิภาพในการตัดส่วนข้อมูลมีความถูกต้องมากขึ้นอีกทั้งยังมีความเร็วในการตัดส่วนข้อมูลได้ดีขึ้นอีกด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Information Extraction for online bookstore
Student	Mr. Monchai Photchanasomsaman
Advisor	Dr. Chotipat Pornavalai
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2001

ABSTRACT

Today, there are several online bookstores that each different service consequently, good and cheap services are rare for searching.

Information Extraction project for online bookstores has developed Web application for searching by prices and services comparison of various online bookstores using Information Extraction technique which has potentially abilities for extraction prices and services for comfortable online book searching. We have improved information extraction in developing phase by grouping data before extracting. By the way this significantly improves information extraction accuracy and improves information extraction speed.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ในการพัฒนาระบบค้นหา เปรียบเทียบราคาหนังสือ และ การบริการจากร้านค้าหนังสือออนไลน์นี้ จะต้องอาศัยแหล่งความรู้ต่าง ๆ ทั้งคำแนะนำและคำปรึกษา ทั้งในภาคทฤษฎีและภาคปฏิบัติ รวมทั้งอุปกรณ์ฮาร์ดแวร์และซอฟต์แวร์ที่จำเป็น ทั้งนี้ โครงการนี้ สำเร็จลุล่วงลงได้ด้วยกำลังใจ, คำแนะนำและคำปรึกษาจากบุคคลต่าง ๆ เหล่านี้ ที่สมควรได้รับการขอบคุณเป็นพิเศษ ดังนี้

1. บิดา มารดา ผู้ให้กำเนิด เลี้ยงดูเอาใจใส่และดูแล อบรมให้ประพฤติในสิ่งที่ดีและถูกต้องตลอดจนส่งเสริมในด้านการศึกษาได้อย่างดีที่สุด
2. อาจารย์โชติพัชร์ ภรณวลัย ที่ให้คำปรึกษาต่าง ๆ จนกระทั่งโครงการนี้ สำเร็จลุล่วงลงได้อย่างดี
3. อาจารย์ทุก ๆ ท่านที่ได้ประสิทธิประสาทวิชาความรู้ หลักวิชาการต่าง ๆ สำหรับใช้เพื่อเป็นพื้นฐานในการใช้ชีวิต และการทำงาน
4. พี่ น้อง และ เพื่อน ๆ ทุกท่าน ที่ให้กำลังใจในการศึกษาไม่ว่าจะมีปัญหาอะไรเกิดขึ้นก็ตาม

นายมนต์ชัย พจนาสมสมาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูปภาพ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันความนิยมของเครือข่ายอินเตอร์เน็ตทำให้การติดต่อกันระหว่างธุรกิจพัฒนาไปอย่างรวดเร็วและมีการใช้งานอย่างรวดเร็วและมีการใช้งานอย่างกว้างขวาง ทำให้เกิดการค้าอิเล็กทรอนิกส์แบบใหม่หรือวิธีการทำธุรกิจใหม่ ๆ เช่น ร้านหนังสือออนไลน์ต่าง ๆ การค้าอิเล็กทรอนิกส์ก่อให้เกิดให้เกิดประโยชน์ต่อผู้ประกอบการ ธุรกิจขนาดเล็กรวมทั้งบุคคลทั่วไปที่ใช้อินเตอร์เน็ต ทำให้อินเตอร์เน็ตเป็นสื่อใหม่ในการเป็นศูนย์กลางที่นำผู้ขายมาพบผู้ซื้อ โดยผู้ซื้อสามารถหาข้อมูลของสินค่าต่าง ๆ เพื่อเปรียบเทียบสินค้า, ราคาและการบริการ

การขยายตัวของอินเตอร์เน็ตอย่างมากทำให้มีบริการประเภทเดียวกันเพิ่มมากขึ้น ผู้ซื้อมีโอกาสเลือกสินค้า ราคาและการบริการมากยิ่งขึ้น จึงทำให้เกิดบริการบนอินเตอร์เน็ตรูปแบบใหม่เพื่อเป็นตัวกลาง (Agent) ในการเปรียบเทียบราคาและการบริการของสินค้า ผู้ซื้อสามารถเลือกซื้อสินค้าที่มีการบริการที่ดีและมีสินค้าราคาถูกได้สะดวกขึ้น

ในโครงการนี้เป็นการพัฒนาตัวกลางในการเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ขนาดใหญ่ที่มีอยู่ในอินเตอร์เน็ตได้แก่ amazon.com, bn.com, bookpool.com โดยจะทำการปรับปรุงวิธี Information Extraction ที่ใช้ในการตัดส่วนของข้อมูลข่าวสารราคาและการบริการให้เหมาะสมกับปัญหาของเรามากยิ่งขึ้น

1.2 วัตถุประสงค์ของโครงการ

การพัฒนาระบบหรือโครงงานนี้เพื่อวัตถุประสงค์ดังนี้

1. เพื่อเป็นการนำเทคโนโลยีหรืออาศัยเทคโนโลยีที่มีอยู่ในปัจจุบันมาพัฒนาและประยุกต์ใช้งานกับระบบให้เกิดประโยชน์สูงสุด
2. เพื่ออำนวยความสะดวกให้ผู้ใช้สามารถเปรียบเทียบราคาหนังสือได้สะดวกรวดเร็วยิ่งขึ้น
3. เพื่อปรับปรุงและพัฒนาวิธีการ Information Extraction ที่มีอยู่ในปัจจุบันให้เหมาะสมกับงานด้านเอกสาร HTML มากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3 ขั้นตอนการดำเนินการโครงการ

เพื่อให้ระบบสามารถนำไปใช้ในการปฏิบัติงานให้บรรลุตามวัตถุประสงค์ข้างต้น จึงได้สรุปขั้นตอนในการดำเนินงานดังนี้

1. ศึกษาวิธีการตัดส่วนของข้อมูลข่าวสาร(Information Extraction)ที่มีอยู่ในปัจจุบัน รวมถึงศึกษาการนำวิธีการตัดส่วนของข้อมูลข่าวสารไปใช้กับงานด้านเอกสาร HTML เพื่อพัฒนาและปรับปรุงให้เหมาะสมโครงงานนี้
2. วิเคราะห์และออกแบบระบบงาน
3. ศึกษาและคัดเลือกเทคโนโลยีที่เหมาะสมในการพัฒนาระบบ
4. พัฒนาโปรแกรมประยุกต์
5. ทดสอบระบบงาน และปรับปรุงแก้ไข
6. ติดตั้งระบบเพื่อนำไปใช้งาน

1.4 เป้าหมายของโครงการ

เป้าหมายของโครงการ การพัฒนาตัวกลางในการเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ คือ เพื่อพัฒนาตัวกลาง ที่มีประสิทธิภาพในการเปรียบเทียบราคาและมีความถูกต้องและสามารถให้บริการผู้ใช้ได้ตลอดเวลา

1.5 ขอบเขตของโครงการ

โครงงานนี้มีของเขตการศึกษาและพัฒนาดังนี้

1. ระบบตัดส่วนข้อมูลด้วย Information Extraction
2. ระบบการสร้างฐานข้อมูลที่ใช้ในการค้นหาข้อมูลก่อนการเปรียบเทียบ
3. ระบบการร้องขอข้อมูลจากร้านค้าหนังสือออนไลน์ต่าง ๆ ได้แก่
    1. ร้าน amazon.com
    2. ร้าน bn.com
    3. ร้าน bookpool.com
4. ระบบเปรียบเทียบราคาและการบริการ และการแสดงผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากการพัฒนาตัวกลางในการเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ มีดังนี้

1. ได้แนวทางการพัฒนาพัฒนาตัวกลางในการเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์และปัญหาในการพัฒนา
2. ได้แนวทางในการตัดส่วนของข้อมูลข่าวสารสำหรับเอกสาร HTML ที่มีประสิทธิภาพ
3. ช่วยลดเวลาและค่าใช้จ่ายในการค้นหาหนังสือบนร้านหนังสือออนไลน์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ความรู้พื้นฐานเกี่ยวกับการตัดส่วนของข้อมูล

การตัดส่วนตัดส่วนของข้อมูลที่สนใจจากเอกสารหนึ่ง (Information Extraction) เป็นส่วนหนึ่งที่สำคัญในการพัฒนาตัวกลางในการเปรียบเทียบราคาและบริการจากร้านหนังสือออนไลน์ โดยความรู้พื้นฐานที่จำเป็นในการพัฒนาระบบตัดส่วนข้อมูลจะประกอบด้วยหัวข้อดังนี้

2.1 การแบ่งส่วนย่อยของข้อมูล

2.2 ความรู้ทั่วไปเกี่ยวกับเอกสารในแง่มุมของการตัดส่วนของข้อมูลข่าวสาร

2.3 พื้นฐานเกี่ยวกับการตัดส่วนของข้อมูล

2.1 การแบ่งส่วนย่อยของข้อมูล

การตัดส่วนของข้อมูลข่าวสารที่สนใจจากเอกสารหนึ่ง ๆ เป็นส่วนหนึ่งของการแบ่งประเภทข้อมูล (Information Classification) ซึ่งความแตกต่างกันระหว่างงานสองประเภทนี้คืองานของการแบ่งประเภทข้อมูลจะรู้ขอบเขตข้อมูลที่แน่นอนแต่ในขณะที่งานของการตัดส่วนของจะไม่รู้ขอบเขตของข้อมูลแน่นอน ทำให้ต้องมีการแบ่งข้อมูลในเอกสารเป็นส่วนย่อย ๆ เพื่อใช้ในการประมวลผลและรวมส่วนย่อย ๆ เหล่านั้นให้เป็นคำตอบที่เราต้องการ

การตัดส่วนของข้อมูลข่าวสารจึงต้องมีการแยกข้อมูลในเอกสาร ออกเป็นกลุ่มเราจะเรียกกลุ่มของข้อมูลในเอกสารว่า เท็กแฟรกเมนต์ (Text Fragment) และในแต่ละเท็กแฟรกเมนต์จะประกอบด้วยกลุ่มของตัวอักษรที่ติดกันโดยมีเงื่อนไขในการแบ่งได้แก่

- กลุ่มของตัวอักษรที่ถูกแบ่งโดยช่องว่างหรือการขึ้นบันทัดใหม่เช่นเท็กแฟรกเมนต์ He is a man. จะประกอบด้วย 5 กลุ่มได้แก่ “He”, “is”, “a”, “man” และ “.”
- กลุ่มของตัวอักษรที่มีสัญลักษณ์แทรกอยู่โดยเช่น He’s จะประกอบด้วย 3 กลุ่มได้แก่ “He”, “’” และ โทเคน “s”
- กลุ่มของตัวอักษรที่มีตัวเลขแทรกอยู่ เช่น ME101 จะประกอบด้วย 2 กลุ่มได้แก่ “ME” และ “101”

เราจะเรียกกลุ่มของตัวอักษรนี้ว่าโทเคน (Token) ซึ่งเราจะถือว่าโทเคนเป็นส่วนย่อยสุดของข้อมูลที่ไม่สามารถแบ่งได้อีก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 ความรู้ทั่วไปเกี่ยวกับเอกสารในแง่มุมของการตัดส่วนของข้อมูลข่าวสาร

รูปแบบของเอกสารมีความสัมพันธ์กับการตัดส่วนของข้อมูลข่าวสารการที่เราสามารถเข้าใจรูปแบบและลักษณะของเอกสารจะทำให้เราสามารถตัดส่วนของข้อมูลข่าวสารได้มีประสิทธิภาพยิ่งขึ้นโดยเราสามารถแบ่งประเภทของเอกสารได้ 5 ประเภทดังนี้

2.2.1 Term View

2.2.2 Mark-Up View

2.2.3 Layout View

2.2.4 Typographic View

2.2.5 Linguistic View

2.2.1 Terms View

ในเอกสารที่แสดงในรูปแบบ Terms view เอกสารจะประกอบด้วยลำดับของโทเคนต่อเนื่องกันดังแสดงในรูปที่ 2. 1 โดยที่รูปแบบธรรมดาของเอกสารแสดงดังรูปที่ 2. 2 โดยมุมมองของเอกสารประเภทนี้จะสนใจเฉพาะโทเคนของเอกสารไม่ใจลักษณะการจัดเรียงตัว ย่อหน้าหรือการขึ้นบรรทัดใหม่

```
< 0 . 31 . 3 . 95 . 12 . 27 . 07 .cd 0 w + @ andrew . cmu . edu . 0 >
Type :    cmu . andrew . org . epp
Topic :   Distinguished Lecture Series
Dates :   3 - Apr - 95
Time :    4 : 30
```

รูปที่ 2. 1 แสดงเอกสารเกี่ยวกับการสัมมนาซึ่งถูกแสดงในรูปแบบโทเคน view

```
<0.31.3.95.12.27.07.cd0w+@andrew.cmu.edu.0>
Type:    cmu.andrew.org.epp
Topic:   Distinguished Lecture Series
Dates:   3-Apr-95
Time:    4:30
```

รูปที่ 2. 2 แสดงเอกสารเกี่ยวกับการสัมมนา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.2 Mark Up View

ในเอกสารที่แสดงในรูปแบบ Mark up View เอกสารจะประกอบด้วยลำดับของเทอม(Terms) และ เมต้าเทอม(Meta Terms) ซึ่งเมต้าเทอม จะทำหน้าที่กำหนดรูปแบบของเทอม โดยรูปที่ 2. 3 แสดงเอกสาร HTML และ รูปที่ 2. 4 แสดง Mark up view ของเอกสาร HTML

```
<html>
<head>
<title>Dayne Freitag's Home Page</title>
</head>
<body bgcolor="#FFFFFF">
<center><h2>Dayne Freitag</h2>
<hr>
<h3><font face="Helvetica">Contents</font></h3></center>
</body>
</html>
```

รูปที่ 2. 3 แสดงเอกสาร HTML

```
<html>
<head>
<title>***** ********* **** ****</title>
</head>
<body bgcolor="#FFFFFF">
<center><h2>***** *******</h2>
<hr>
<h3><font face="Helvetica">********</font></h3></center>
</body>
</html>
```

รูปที่ 2. 4 แสดงเอกสารที่แสดงในรูปแบบของ Mark Up View โดย ส่วนที่ไม่ใช่ Mark up จะถูกแสดงด้วยดอกจันทร์ (*)

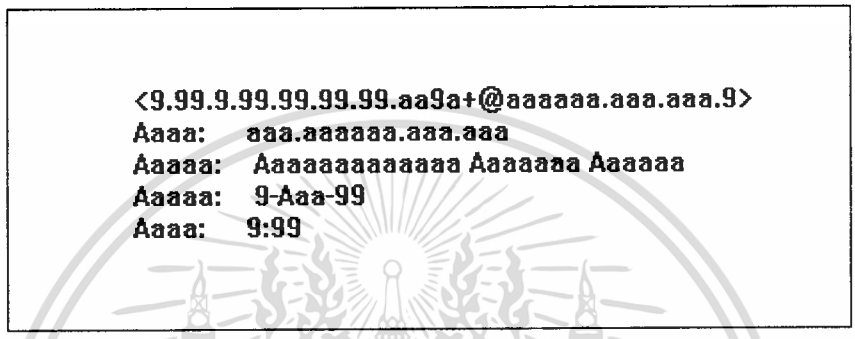
2.2.3 Layout View

ในเอกสารที่แสดงในรูปแบบ Layout View จะสนใจมิติของเอกสารที่ถูกจัดเรียง และ ขนาดของเทอมโดยข้อมูลข่าวสารบางส่วนที่มองในมุมมองนี้จะสามารถพิจารณาได้ง่ายเช่น ย่อหน้า, ตาราง, หัวข้อของซองจดหมาย โดยรูปที่ 2. 2 แสดงเอกสารการสัมมนาจากรูปที่ 2. 1 ที่ถูกแสดงในมุมมอง Layout View นี้

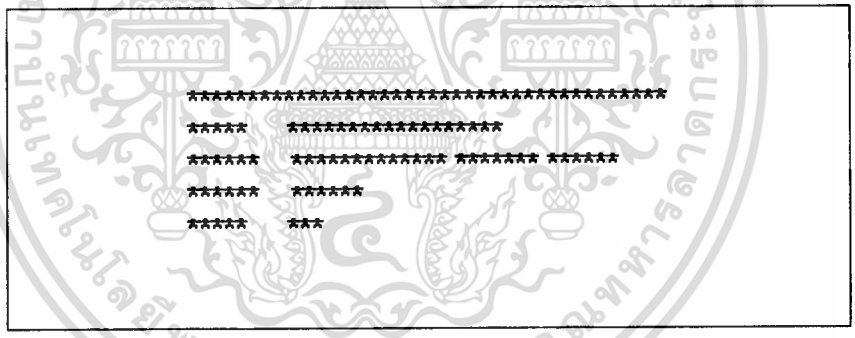
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.4 Typographic View

ในเอกสารที่แสดงในรูปแบบ Typographic View จะใช้ฟังก์ชันเป็นตัวแทนของดเทอม โดยที่ฟังก์ชันจะประกอบด้วยสมาชิกของตัวอักษรประกอบกัน โดยแทนสมาชิกของตัวอักษรด้วยตัวอักษรพิเศษแทนกลุ่มของตัวอักษร เช่น ตัวอักษรใหญ่, ตัวอักษรตัวเล็ก หรือ ตัวเลข ดังแสดงในรูปที่ 2. 6 โดยใช้ตัวอักษร 9 แทนตัวเลขและ A แทนตัวอักษรตัวใหญ่ และ a แทนตัวอักษรตัวเล็ก



รูปที่ 2. 5 แสดงเอกสารสัมมนาที่ถูกแสดงในรูปแบบ Layout view



รูปที่ 2. 6 แสดงเอกสารสัมมนาที่ถูกแสดงในรูปแบบ Typographic View

2.2.5 Linguistic View

ในเอกสารที่แสดงในรูปแบบ Linguistic View จะสนใจความหมาย และ โครงสร้างของประโยคเพียงอย่างเดียว

2.3 พื้นฐานเกี่ยวกับการตัดส่วนของข้อมูล

พื้นฐานในการพัฒนาระบบตัดส่วนข้อมูลเป็นสิ่งที่จำเป็นต้องศึกษาก่อนการพัฒนาระบบ เพื่อทำความเข้าใจถึงปัญหาในการตัดส่วนของข้อมูล ประเภทของปัญหา การประเมินประสิทธิภาพ และวิธีในการตัดส่วนข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3.1 ฟิลด์และฟิลด์อินสแตนด์

ในเอกสารหนึ่ง ๆ ที่เราต้องการตัดส่วนของข้อมูล เราจะเรียกสิ่งที่เราต้องการตัดส่วนของข้อมูลว่าฟิลด์(Field) และ เรียกคำตอบที่ได้จากการตัดส่วนของข้อมูลนั้นว่าฟิลด์อินสแตนด์(Field Instance)

ข้อมูลของเอกสารหนึ่งอาจจะประกอบด้วยหลายฟิลด์และบางฟิลด์อาจจะมีหลายฟิลด์อินสแตนด์ ตัวอย่างเช่น เอกสารเกี่ยวกับการแนะนำหนังสือ "Distributed Systems Concepts and Design" แต่งโดย Coulouris, Dollimore และ Kindberg ในเอกสารนี้เราอาจจะต้องการทราบรายละเอียดเกี่ยวกับ ชื่อหนังสือและผู้แต่ง จะเห็นได้ว่าในเอกสารฉบับนี้ประกอบด้วย 2 ฟิลด์ไดแก่ ฟิลด์ ชื่อหนังสือ และ ฟิลด์ผู้แต่ง โดยที่ในฟิลด์หนังสือมี 1 ฟิลด์อินสแตนด์คือ Distributed Systems Concepts and Design และ ฟิลด์ผู้แต่งมี 3 ฟิลด์อินสแตนด์คือ Coulouris, Dollimore และ Kindberg

2.3.2 การประเมินประสิทธิภาพของระบบตัดส่วนของข้อมูล

ในเอกสาร D หนึ่ง ๆ จะประกอบไปด้วย Set F(D) ซึ่งมีสมาชิกเป็นฟิลด์อินสแตนด์ของเอกสาร D ในเซ็ตของ F(D) อาจจะประกอบด้วยสมาชิกเพียง 1 ตัว (OPD; One pre document) หรือ มากกว่า 1 ตัวก็ได้ (MPD; Many per Document) ตัวอย่างของเอกสารที่ F(D) มีสมาชิก 1 ตัว เช่น เวลาเริ่มต้นของการจัดอบรมสัมมนา ถึงแม้ว่าเอกสารเกี่ยวกับการจัดอบรมอาจจะพบเวลาเริ่มของการจัดอบรมมากกว่า 1 ครั้งก็ตามแต่ในแต่ละครั้งที่พบจะมีค่าเหมือนกันหรือเท่ากัน เช่น เวลาเริ่มต้นอบรมอาจจะมีค่าเป็น 3:00 PM หรือ 3:00 p.m. ก็มีค่าเท่ากัน ตัวอย่างของเอกสารที่ F(D) มีสมาชิกมากกว่า 1 ตัวเช่น สมาชิกในกลุ่มวิจัยงาน จะมีสมาชิกมากกว่า 1 คน

การประเมินประสิทธิภาพของระบบการตัดส่วนของข้อมูลที่แก้ปัญหาประเภท OPD จะให้คะแนนแก่ระบบก็ต่อเมื่อระบบให้คำตอบออกมาถูกต้องเพียงตัวเดียว และถ้าเป็นระบบ MPD จะให้คะแนนแก่ระบบก็ต่อเมื่อคำตอบของระบบออกมาครบทุกสมาชิกในเซ็ต F(D) หรือ ถ้าระบบไม่สามารถให้คำตอบออกมาได้ครบทั้งหมดระบบอาจจะให้คะแนนเพียงส่วนหนึ่ง แล้วแต่การออบแบบระบบตัดส่วนของข้อมูลว่าจะพิจารณาอย่างไร ในโครงงานนี้เราจะใช้ระบบ OPD ในการออกแบบระบบการตัดส่วนข้อมูลสำหรับร้านหนังสือออนไลน์เท่านั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3.3 ผลลัพธ์ของเอกสาร (Document Outcome)

ผลลัพธ์ที่ได้จากเอกสารหนึ่ง ๆ ที่ผ่านการทดสอบโดยระบบแสดงถึงความสามารถของระบบในหัวข้อนี้เราจะแบ่งแยกลักษณะของผลลัพธ์ของเอกสารที่ได้จากการทดสอบได้ 4 ประเภท โดยสมมุติให้ระบบให้คำตอบ P และเอกสารที่ใช้ทดสอบ D ได้แก่

(a) Correct หมายความว่า P เป็นคำตอบของฟิลด์ในเอกสาร D นี้ ($P \in F(D)$)

(b) Wrong หมายความว่า P ไม่เป็นคำตอบของฟิลด์ในเอกสาร D โดยที่ฟิลด์ในเอกสาร D นี้มีคำตอบ ($P \notin F(D)) \wedge (F(D) \neq \varnothing$)

(c) Spurious หมายความว่า P ที่ได้ไม่เป็นคำตอบของฟิลด์ในเอกสาร D โดยที่ในเอกสาร D ไม่มีคำตอบ ($P \neq nil) \wedge (F(D) = \varnothing$)

(d) No prediction หมายความ ($P = nil$)

จากประเภทของผลลัพธ์ที่ได้เราสามารถกำหนดความเชื่อมั่นของคำตอบ (Prediction) ได้จาก

$$precision \equiv \frac{Correct}{Correct + Wrong + Spurious} \quad \ldots(2.1)$$

2.3.4 ผลลัพธ์ของคำตอบ (Fragment Outcome)

ผลลัพธ์ของคำตอบจะพิจารณาผลลัพธ์ของเอกสารประเภท Correct ว่าขอบเขตของการยอมรับว่าคำตอบที่ได้ถูกต้องหรือไม่โดยพิจารณาจากเกณฑ์ที่ใช้ดังนี้

1. Exact คำตอบที่ได้จากระบบ (Field instance) ตรงกับคำตอบจริง ๆ (Actual instance)

2. Contain คำตอบที่ได้จากระบบเป็นส่วนหนึ่งหรืออยู่ภายในคำตอบจริง ๆ เช่น เวลาเริ่มต้นการประชุมสัมมนา เป็น 2:00 pm คำตอบที่ได้จากระบบเป็น 2:00 เป็นต้น

3. Overlap คำตอบที่ได้ซ้อนทับกับคำตอบจริง ๆ เช่น เวลาเริ่มต้นการประชุมสัมมนาเป็น 2:00 pm คำตอบที่ได้จากระบบเป็น Time: 2:00

กฎเกณฑ์ทั้ง 3 สามารถเลือกใช้ได้ขึ้นอยู่กับสถานการณ์ และกฎเกณฑ์ทั้ง 3 ยังสามารถทำให้เราทราบถึงประสิทธิภาพที่เปลี่ยนไป ถ้าเราใช้เกณฑ์ในการพิจารณาแบบ Overlap แสดงถึง ระบบเราต้องการความสามารถในการหาตำแหน่งของคำตอบว่ามีประสิทธิเพียงใด โดยไม่สนใจขอบเขตของคำตอบ สำหรับเกณฑ์ในการพิจารณาแบบ Contain สามารถเลือกใช้ได้ในบางโปรแกรมประยุกต์ โดยให้โปรแกรมประยุกต์เติมส่วนที่ขาดหายไปเองซึ่งเป็นเรื่องไม่ยากนัก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำหรับในโครงงานนี้เราเลือกใช้เกณฑ์ในการพิจารณาแบบ Exact เพียงอย่างเดียว กล่าวคือคำตอบจะถูก ก็ต่อเมื่อคำตอบที่ได้จากระบบตรงกับคำตอบจริง ๆ เท่านั้น

2.3.5 ความเชื่อมั่นและความถูกต้อง (Precision and Recall)

Precision ที่กำหนดไว้ก่อนหน้านี้จะไม่นับจำนวนเอกสารที่ทำการทดสอบจึงไม่สามารถหาความประสิทธิภาพในการหาความถูกต้องของระบบได้ ดังนั้นเราจึงได้นิยามประสิทธิภาพในการหาความถูกต้องของระบบซึ่งหาได้จาก

$$recall \equiv \frac{Correct}{|\{D \mid F(D) \neq \phi\}|} \quad \ldots(2.2)$$

จะเห็นว่า Precision และ Recall มีความสัมพันธ์กันทำให้เราต้องพิจารณาทั้งสองค่าพร้อม ๆ กัน

โดยทั่วไประบบจะให้คำตอบออกมาพร้อมกับคะแนน(Instance value) ซึ่งแสดงถึงความสามารถของคำตอบนั้นว่ามีความน่าจะเป็นเพียงใดในการที่จะเป็นตอบของระบบ ซึ่งเราจะกำหนดขอบเขตของคะแนน(Threshold) ที่เราจะยอมรับให้คำตอบที่มีคะแนนอยู่ในช่วงที่เรากำหนด ให้เป็นคำตอบของระบบ

เมื่อระบบให้คำตอบซึ่งมีคะแนนออกมาต่ำกว่ากว่าขอบเขตที่กำหนดไว้ เราจะถือว่าระบบไม่ได้ให้คำตอบ ซึ่งจะทำให้ค่าความเชื่อมั่นเพิ่มขึ้น โดยทั่วไปขอบเขตของคะแนนจะแปรผันโดยตรงกับค่าความเชื่อมั่น กล่าวคือถ้าขอบเขตที่เราตั้งสูง ค่าความเชื่อมั่นจะมากตามหรือกล่าวอีกในหนึ่งคือ คำตอบจะมีความน่าเชื่อถือสูง แต่ผลลัพธ์ที่ได้โดยทั่วไปจากการปรับขอบเขตของคะแนนเพิ่มขึ้นจะทำให้ ความถูกต้องของระบบโดยรวมลดลงเพราะจำนวนคำตอบที่ได้จากระบบจะมีน้อยลง แต่ไม่จำเป็นเสมอที่ระบบจะให้คำตอบน้อยลงขึ้นอยู่กับความประสิทธิภาพของระบบ

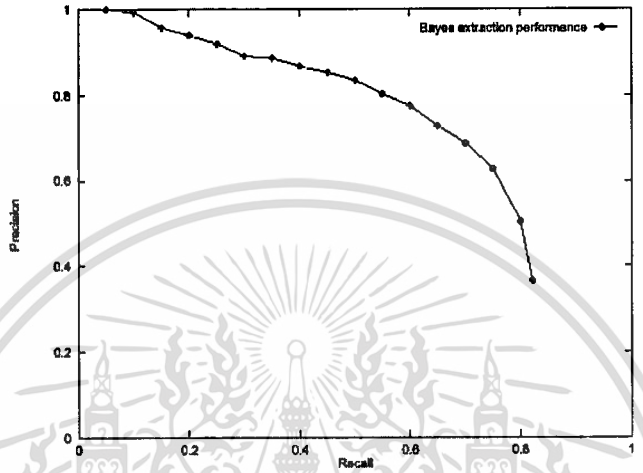
เราสามารถเขียนกราฟแสดงความสัมพันธ์ระหว่างความเชื่อมั่นกับความถูกต้องของระบบได้โดยปรับเปลี่ยนค่า Threshold ต่าง ๆ ดังแสดงในรูปที่ 2. 7

จากความสัมพันธ์ความเชื่อมั่นกับความถูกต้องของระบบดังกล่าวข้างต้น ทำให้การเปรียบเทียบหรือการวัดประสิทธิภาพของระบบที่แตกต่างเป็นไปด้วยความยากลำบาก จึงได้มีการนำค่าความเชื่อมั่นกับค่าความถูกต้องของระบบมาคิดร่วมกันเป็นเพียงค่าเดียว โดยเรียกค่านี้ว่า F-measure ซึ่งสามารถหาได้จาก

$$F = \frac{(\beta^2 + 1.0)PR}{(\beta^2 P) + R} \quad \ldots(2.3)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยที่ P หมายถึงค่าความเชื่อมั่นของคำตอบ R หมายถึงค่าความถูกต้องในการให้คำตอบของระบบและ β หมายถึง อัตราส่วนความสำคัญของ Recall ส่วน Precision โดยทั่วไปเราจะให้ค่า $\beta = 1$ ซึ่งให้ความสำคัญของความเชื่อมั่นของคำตอบเท่ากับค่าความถูกต้องในการให้คำตอบของระบบ



รูปที่ 2. 7 กราฟแสดงความสัมพันธ์ระหว่างความเชื่อมั่นกับความถูกต้องของระบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีการตัดส่วนข้อมูล

วิธีในการตัดส่วนข้อมูลในปัจจุบันมีอยู่มากมายหลายวิธี ซึ่งแต่ละวิธีมีข้อดีข้อเสียแตกต่างกันทำให้การเลือกวิธีในการตัดส่วนข้อมูลที่เหมาะสมกับลักษณะงานเป็นสิ่งที่สำคัญอย่างยิ่ง ในโครงงานฉบับนี้จะใช้วิธีในการตัดส่วนข้อมูลที่สนใจเฉพาะโทเคนของเอกสาร (Term-Space Learning) เป็นวิธีพื้นฐานในการพัฒนาโครงงาน

ในวิธี Term-Space Learning นี้เอกสารที่ต้องการพิจารณาจะมองเอกสารในลักษณะโทเคน View โดยจะสนใจเฉพาะข้อมูลในลักษณะของโทเคน และจะไม่สนใจลักษณะของรูปแบบของเอกสาร ๆ ย่อหน้า ช่องว่าง

วิธี Term-Space Learners จะใช้โทเคนของข้อมูลข่าวสารที่เรียงกันอยู่ในเอกสาร เป็นตัวช่วยในการค้นหาข้อมูลซึ่งมีข้อดีได้แก่

1. เอกสารที่ต้องการพิจารณาจะไม่ยึดติดกับรูปแบบของเอกสารเพราะวิธี Term-space Learning มีข้อบังคับในการตัดส่วนของข้อมูลค่อนข้างน้อยทำให้สามารถนำไปใช้ในการตัดส่วนของข้อมูลได้อย่างกว้างขวาง

2. เป็นวิธีการที่มีประสิทธิภาพในการประมวลผลเพราะข้อบังคับในการตัดส่วนของข้อมูลค่อนข้างน้อย

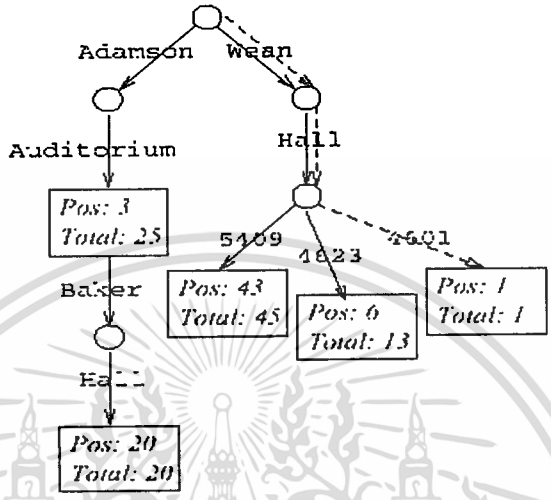
3.1 Rote Learning

เป็นวิธีแรกเริ่มที่มีการนำมาแก้ปัญหาการตัดส่วนข้อมูล โดยหลักการทำงานของ วิธีการนี้คือ ให้ระบบเรียนรู้เก็บโทเคนทุกโทเคนที่เคยพบ แล้วทำการนับจำนวนครั้งที่โทเคนนั้นเป็นคำตอบ เมื่อต้องการทดสอบระบบจะทำการตัดเอกสารออกเป็นโทเคน และ ทำการเปรียบเทียบกับโทเคนที่เก็บไว้ในฐานข้อมูลว่าพบหรือไม่ ถ้าพบว่ามี จะทำการเปรียบเทียบที่ละโทเคนและนับจำนวนครั้งที่โทเคนนั้นเป็นคำตอบ ถ้าอัตราส่วนระหว่างจำนวนครั้งทีเป็นคำตอบ กับ จำนวนครั้งที่พบมากกว่าค่าขอบเขตต่ำสุดที่ยอมรับได้ก็จะถือว่ากลุ่มของโทเคนนั้นเป็นคำตอบโดยสามารอธิบายการทำงานได้ดังรูปที่ 3. 1

จากการขั้นตอนการทำงานดังกล่าว จะเห็นได้ว่าวิธีการนี้สามารถใช้ได้กับเอกสารทั่วไปโดยไม่จำกัดหัวข้อของเอกสาร โดยประสิทธิภาพที่ได้จะขึ้นอยู่กับรูปแบบของ คำตอบถ้ารูปแบบของคำตอบมีรูปแบบที่จำกัดก็วิธีการนี้จะเป็นวิธีการที่มีประสิทธิภาพสูง แต่ในทางตรงกันข้ามถ้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปแบบของเอกสารมีลักษณที่หลากหลายก็จะเป็นวิธีที่ไม่เหมาะสมในการตัดส่วน ข้อเสียของวิธีการนี้ก็คือจะไม่สามารถให้คำตอบ ที่ไม่เคยพบมาก่อนได้ จึงทำให้วิธีการนี้ต้องการข้อมูลที่ใช้ในการสอนมากและมีลักษณะที่เหมาะสม



รูปที่ 3. 1 แสดงการทำงานของ Rote learning โดยเท็กแฟรกเมนต์ เป็น "Wean Hall 4601"

3.2 Naive Bayese Learning

Naive Bayes Learning เป็นวิธีการในการตัดส่วนข้อมูลที่แตกต่าง จากวิธี Rote Learning ที่ใช้การเปรียบเทียบโทเคนของเอกสารที่ละโทเคน ในการคาดเดาคำตอบ ในวิธีการนี้จะใช้การเปรียบเทียบโทเคนของคำตอบ และโทเคนรอบ ๆ คำตอบจาก Training data และใช้ค่าทางสถิติช่วยในการคาดเดาคำตอบ ทำให้การคาดเดาคำตอบที่ไม่เคยพบมาก่อนให้มีประสิทธิภาพที่ดีขึ้น

3.2.1. สมมุติฐานในการคาดเดาคำตอบ

จากกฎของ Bayes เราสามารถหาความน่าจะเป็นของสมมุติฐานที่ เราคาดว่าจะเป็นคำตอบในเอกสาร D ได้จาก

$$\Pr(H \mid D) = \frac{\Pr(D \mid H)\Pr(H)}{\Pr(D)} \qquad \ldots(3.1)$$

Pr(H) หมายถึง ความน่าจะเป็นที่สมมุติฐานที่ตั้งไว้เป็นคำตอบ

Pr(D|H) หมายถึง ความน่าจะเป็นที่เอกสาร D เป็นไปตามเงื่อนไข H ที่

P(D) หมายถึง ความน่าจะเป็นของเอกสารที่ใช้ในการสอนระบบ(Training data) ซึ่งจะมีค่าเท่ากันทุก H_i เนื่องจากเราใช้เอกสารชุดเดียวกันในการคาดเดาคำตอบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สมมุติว่าเราต้องการหาชื่อผู้อภิปรายในเอกสารเกี่ยวกับการอบรมสัมมนาฉบับหนึ่ง เราจะทำการสร้างแบบจำลองซึ่งประกอบไปด้วยกลุ่มของสมมุติฐาน H ซึ่งแต่ละสมมุติฐาน H จะแทนกลุ่มของโทเคน(เท็กแฟรกเมนต์) ที่เราคาดว่าจะเป็น ชื่อผู้อภิปราย ซึ่งกลุ่มของโทเคนจะถูกกำหนดโดยใช้ตำแหน่งของโทเคนตัวแรกกับจำนวนโทเคนเช่น $H_{309,2}$ หมายถึง สมมุติฐาน H ที่พิจารณา เท็กแฟรกเมนต์ที่ประกอบด้วยโทเคนที่ 309 ของเอกสารและมีความยาว 2 โทเคน(โทเคนที่ 309 กับโทเคนที่ 310) โดยสรุปแล้วเราสามารถอธิบายสมมุติฐาน $H_{i,k}$ ที่เราจะพิจารณาในรูปแบบคำพูดได้ว่า "สมมุติฐาน ที่เราคาดว่าคำตอบประกอบด้วยโทเคนที่ i กับโทเคนถัดไปอีก k-1 โทเคน"

ในการพิจารณาหาคำตอบเราจะเลือกเอาสมมุติฐานที่ให้ค่าความน่าจะเป็นมากที่สุด (สมมุติฐานที่ให้ค่า $Pr(D|H_{p,k})Pr(H_{p,k})$ มากที่สุด) และเปรียบเทียบกับค่าขอบเขตของคะแนนที่ตั้งไว้ก่อนเลือกคำตอบ

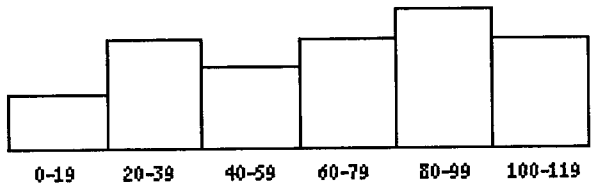
3.2.2. การหาค่าความน่าจะเป็นจากสมการ Bayes

การหาค่าความน่าจะเป็นของสมมุติฐานในเอกสาร D สิ่งที่เราต้องการระบุในการตั้งสมมุติฐานของเราก็คือตำแหน่ง และ ความยาว ซึ่งอิสระจากกัน ดังนั้นจากกฎของ Bayes ความน่าจะเป็น $Pr(H_{p,k})$ ก็จะสามารถหาได้จาก

$$Pr(H_{p,k})=Pr(\text{position} = i)Pr(\text{length} = k) \quad \ldots(3.2)$$

โดยทั่วไปคำตอบที่เราจะพิจารณาจะมีความยาวไม่มากนักดังนั้น Pr(length=k) ก็จะสามารถหาได้จากจำนวนคำตอบที่มีความยาว k หารด้วยจำนวนคำตอบที่เราใช้ในการสอนระบบ

ส่วนถัดมาที่เราต้องหาก็คือ Pr(position=k) โดยวิธี Bayes จะเรียงลำดับตำแหน่งของคำตอบ แล้วแบ่งออกเป็น n ช่วงโดย n มีขนาดเล็กกว่าขนาดของเอกสาร วิธี Bayes จะใช้ความถี่ของคำตอบที่อยู่ในช่วงที่แบ่งไว้ก่อนหน้านี้ ในการหา Pr(position=k) โดย Pr (position=k) จะเท่ากับความถี่รวมในช่วงที่ k อยู่หารด้วยจำนวนคำตอบทั้งหมด ซึ่งแสดงดังรูปที่ 3. 2 โดยแกนนอนแสดงช่วงของคำตอบ แกนตั้งแสดงความถี่ของคำตอบที่อยู่ในช่วงตำแหน่ง ใด ๆ



รูปที่ 3. 2 แสดง histogram ที่ถูกใช้ในการประมาณเพื่อหาค่า Pr(Postion = k)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Bayes ใช้วิธีการประมาณค่า $Pr(D|H_{p,k})$ ของโทเคนที่อยู่ใกล้เท็กแฟรกเมนต์ของสมมุติฐานที่เรากำลังพิจารณา แต่ก่อนที่จะประมาณค่า $Pr(D|H_{p,k+})$เราต้องกำหนดค่า w ขึ้นมาก่อนซึ่ง w เป็นค่าที่บอกถึงจำนวนโทเคนก่อนและหลังเท็กแฟรกเมนต์ที่เรากำลังพิจารณาในแต่ละโทเคนที่อยู่ในเท็กแฟรกเมนต์ ก่อนหน้าเท็กแฟรกเมนต์และหลังเท็กแฟรกเมนต์จำนวน w โทเคน เราจะสมมุติให้ทุกโทเคนอิสระจากกัน สำหรับโทเคน t เราสามารถหา Pr(D|H) ที่ตำแหน่ง i ใด ๆ ได้จาก

$$Pr(D \mid H_{p,k}) = \prod_{p-w \le i \le p+k+w-1} Pr(t_i \mid H_{p,k}) \qquad \ldots(3.3)$$

เราจะทำการแบ่งผลคูณของ Pr(t|Hp, k) ออกจากกันโดยแบ่งเป็น 3 ส่วนได้แก่ ผลคูณของความน่าจะเป็นของโทเคนก่อนหน้าเท็กแฟรกเมนต์ ผลคูณของความน่าจะเป็นของโทเคนหลังจากเท็กแฟรกเมนต์ และค่าความน่าจะเป็นของโทเคนภายในเท็กแฟรกเมนต์ซึ่งสามารถเขียนเป็นสมการได้เป็น

$$Pr(D|H_{p,k}) = [\prod_{j=1}^{w} Pr(before_j = t_{p-j})] \bullet [\prod_{j=1}^{k} Pr(in = t_{p+j-1})] \bullet [\prod_{j=1}^{w} Pr(after_j = t_{p+k+j-1})] \qquad \ldots(3.4)$$

ข้อแตกต่างระหว่างตัวแปร before, after กับตัวแปร in ก็คือ ตัวแปร before, after เป็นกลุ่มของตัวแปรซึ่งจะเก็บค่าเป็นความถี่ของแต่ละโทเคนก่อนหน้าและตามหลัง เท็กแฟรกเมนต์แต่ in จะเป็นตัวแปรเพียงตัวเดียวที่เก็บความถี่ของเท็กแฟรกเมนต์ทั้งก้อนเราสามารถเขียน Algorithm ในการสร้างฐานข้อมูลเก็บความถี่ของคำก่อนหน้าฟิลด์อินสแตนด์ ตามหลังฟิลด์อินสแตนด์และความถี่ของฟิลด์อินสแตนด์ได้ดังรูปที่ 3. 3

สำหรับ Algorithm ในการประมาณค่าคำตอบในรูปที่ 3. 4 Algorithm จะมีตัวแปร in, before, after และ all ทั้ง ตัวแปร in และ ตัวแปร all จะเป็นตัวแปรประเภท hash table ที่ดึงความถี่ของโทเคนที่ดึงมาจากเอกสาร ส่วนตัวแปร before, after เป็น array ของ hash table ที่เก็บความถี่ของโทเคนที่ตำแหน่งก่อนหน้า ตามเท็กแฟรกเมนต์จำนวน i term สำหรับตัวแปร totalFIcount หมายถึง จำนวนหน้าฟิลด์อินสแตนด์ทั้งหมดที่สอนให้กับระบบ ตัวแปร totalFieldTokens หมายถึงจำนวนโทเคนทั้งหมดก่อนหน้าฟิลด์อินสแตนด์ และ ฟังก์ชัน positionPrior กับ ฟังก์ชัน LengthPrior เป็นฟังก์ชันที่ให้ค่าความน่าจะเป็นของตำแหน่งและความน่าจะเป็นของความยาวของเท็กแฟรกเมนต์ ที่กำลังพิจารณา ฟังก์ชัน MEst เป็นฟังก์ชันชดเชยค่าความถี่เมื่อความถี่มีค่าน้อย ๆ ผลลัพธ์ของ ฟังก์ชัน BayesEstimate จะมีค่าน้อยกว่า 0 เพราะความน่าจะเป็นมีค่าอยู่ระหว่าง 0 – 1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
1 Procedure BayesAccount(doc, fieldname)
2 fbounds = FieldInstanceBounds(doc, fieldname)
3 For (firsti, lasti) in fbounds/*for each index*/
4
5 PositionAccount(firsti) /* For position prior */
6 LengthAccount(lasti - firsti + 1)/* For length*/
7 /* Update in */
8 For i = firsti to lasti
9 token = TokenAt(doc, i)
10 in{token} = in{token} + 1
11 End For
12 Fori =1 to$w$
13 /* Update before */
14 tab = before[i]
15 index = firsti – I
16 token = TokenAt(doc, index)
17 tab{token} = tab{token} + 1
18 /* Update after */
19 tab = after[i]
20 index = lasti + I
21 token = TokenAt(doc, index)
22 tab{token} = tab{token} + 1
23 End For
24 End For
25 /* Update all */
26 For i = 0 to LastTokenIndex(doc)
27 token = TokenAt(doc, i)
28 all{token} = all{token} + 1
29 End For
30 End Procedure
```

รูปที่ 3. 3 แสดง Alogrithm ในการสอนระบบตัดส่วนข้อมูลด้วยวิธี Bayes

```
1 Function BayesEstimate(doc, firsti, lasti)
2 logprob = log(PositionPrior(firsti))
3 + log(LengthPrior(lasti - firsti + 1))
4 For i = 1 to $w$
5 tab = before[i]
6 token = TokenAt(doc, firsti - i)
7 count = tab{token}
8 logprob = logprob + log(MEst(count, totalFIcount))
9 End For
10 For i = firsti to lasti
11 token = TokenAt(doc, i)
12 count = in{token}
13 logprob = logprob + log(MEst(count, totalFieldTokens))
14 End For
15 For i = 1 to $w$
16 tab = after[i]
17 token = TokenAt(doc, lasti + i)
18 count = tab{token}
19 logprob = logprob + log(MEst(count, totalFIcount))
20 End For
21 Return logprob
22 End Function
```

รูปที่ 3. 4 แสดง Algorithm ในประมาณค่าเท็กแฟรกเมนต์โดยวิธี Bayes

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

วิธี Bayes จะไม่สนใจคำตอบที่ได้ถ้าค่า log ของค่าความจะเป็นน้อยกว่าขอบเขตที่กำหนด(Threshold) T โดยค่า T สามารถหาได้จากค่า log ของความน่าจะเป็นน้อยที่สุดที่ให้คำตอบได้ถูกต้องหรือสามารถเขียนเป็นสมการได้เป็น

$$T = \alpha \min P(f) \quad \ldots(3.5)$$

โดยที่ α เป็นค่าที่ผู้ใช้ตั้ง ถ้าเพิ่มค่า α จะทำให้มีคำตอบออกมาจากระบบมากขึ้น ในทางกลับกันถ้าลด α คำตอบที่ได้จากระบบจะลดน้อยลง

3.2.3. การปรับปรุงประสิทธิภาพของ Bayes

จากหลักการทำงานของ Bayes Learning ที่แสดงในรูปที่ 3. 4 จะเห็นว่าความน่าจะเป็นของเท็กแฟรกเมนต์ที่ยาวกว่าจะมีความได้เปรียบกว่าเท็กแฟรกเมนต์ที่สั้นกว่า ทั้ง ๆ ที่ความเป็นจริงคำตอบของฟิลด์อินสแตนด์ อาจจะเป็นคำตอบที่สั้นกว่าก็ได้ดังนั้นเราจะถ่วงน้ำหนักของเท็กแฟรกเมนต์ที่สั้นยาวไม่เท่ากันด้วยการหาค่าความน่าจะเป็นเฉลี่ยต่อหน่วยความยาว โดย Algorithm ที่ได้แสดงดังรูปที่ 3. 5

```
1 Function BayesLNEstimate(doc, firsti, lasti)
2 logprob = log(PositionPrior(firsti))
3 + log(LengthPrior(lasti – firsti + 1))
4 For i = 1 to $w$
5 tab = before[i]
6 token = TokenAt(doc, firsti - i)
7 count = tab{token}
8 logprob = logprob + log(Mest(count, totalFIcount))
9 End For
10 probsum = 0
11 probcount = 0
12 For i = firsti to lasti
13 token = TokenAt(doc, i)
14 count = in{token}
15 probsum = probsum + log(MEst(count, totalFieldTokens))
16 probcount = probcount + 1
17 End For
18 logprob = logprob + avgFIlength * probsum / probcount
19 For i = 1 to $w$
20 tab = after[i]
21 token = TokenAt(doc, lasti + i)
22 count = tab{token}
23 logprob = logprob + log(MEst(count, totalFIcount))
24 End For
25 Return logprob
26 End Function
```

รูปที่ 3. 5 แสดง Algorithm ในประมาณค่าเท็กแฟรกเมนต์โดยวิธี BayesLn

โดยส่วนที่ปรับปรุงมี 2 ตำแหน่งก็คือบรรทัดที่ 10 และ 18 จะกำหนดค่าตัวแปร probsum ขึ้นมาเพื่อเก็บความน่าจะเป็นทั้งหมดของเท็กแฟรกเมนต์ที่กำลังพิจารณาและหลัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ออกจาก Loop ก็จะทำการหาค่าความน่าจะเป็นรวมด้วยความยาวของเท็กแฟรกเมนต์ที่กำลังพิจารณาทำให้ความน่าจะเป็นของเท็กแฟรกเมนต์ที่มีความยาวไม่เท่ากัน ไม่ได้เปรียบเสียเปรียบกัน เราจะเรียกวิธีการนี้ว่า BayesLn

จากการที่เราปรับปรุง Algorithm เป็น Bayesln แล้วยังพบปัญหาที่ว่าโทเคนที่มีการพบบ่อยเช่น ในเอกสารการอบรมสัมมนา ในฟิลด์ชื่อผู้อภิปราย จะพบโทเคน "Dr." ได้บ่อยมากทำให้เท็กแฟรกเมนต์ที่อยู่ใกล้หรือประกอบด้วยโทเคนที่พบบ่อย เกิดการโอนเอียง (Bias) จากความเป็นจริงมีผลให้คำตอบที่ได้ผิดจากความเป็นจริง ดังนั้นเราจึงทำการปรับปรุง Algorithm Bayesln ใหม่ โดยหาความน่าจะเป็นของโทเคนนั้นจากจำนวนครั้งที่พบโทเคนนั้นเป็นโทเคนส่วนประกอบของฟิลด์อินสแตนด์หารด้วยจำนวนครั้งที่พบโทเคนนั้นในเอกสารทั้งหมดที่สอนให้กับระบบ(Corpus) ดังแสดงในรูปที่ 3. 6 ซึ่งจุดที่เปลี่ยนแปลงไปจาก Bayesln ก็คือบรรทัดที่ 8, 15, 23 จากเดิมจะใช้จำนวนคำตอบทั้งหมดของเอกสารที่สอนให้กับระบบ และ จำนวนความยาวของฟิลด์อินสแตนด์ทั้งหมด เปลี่ยนเป็นจำนวนครั้งที่พบโทเคนนั้นในเอกสารทั้งหมดที่สอนให้กับระบบ ซึ่ง Algorithm นี้เราจะเรียกว่า BayesIDF โดยผลจากการที่เราเปลี่ยนวิธีการหาความน่าเป็นวิธีใหม่ จะทำให้โทเคนที่ถูกพบบ่อยถูกเฉลี่ยให้มีค่าลดลง

```
1 Function BayesIDFEstimate(doc, firsti, lasti)
2 logprob = log(PositionPrior(firsti))
3 + log(LengthPrior(lasti - firsti + 1))
4 For i = 1 to $w$
5 tab = before[i]
6 token = TokenAt(doc, firsti - i)
7 count = tab{token}
8 logprob = logprob + log(MEst(count, all{token}))
9 End For
10 probsum = 0
11 probcount = 0
12 For i = firsti to lasti
13 token = TokenAt(doc, i)
14 count = in{token}
15 probsum = probsum + log(MEst(count, all{token}))
16 probcount = probcount + 1
17 End For
18 logprob = logprob + avgFIlength * probsum / probcount
19 For i = 1 to $w$
20 tab = after[i]
21 token = TokenAt(doc, lasti + i)
22 count = tab{token}
23 logprob = logprob + log(MEst(count, all{token}))
24 End For
25 Return logprob
26 End Function
```

รูปที่ 3. 6 แสดง Algorithm ในประมาณค่าเท็กแฟรกเมนต์โดยวิธี BayesIDF

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3 ตัวอย่างผลการพัฒนาระบบโดยวิธีการต่าง ๆ

เอกสารที่ใช้ในการพัฒนาระบบนี้ เป็นเอกสารเกี่ยวกับการจัดอบรมสัมมนาโดยเราจะเปรียบเทียบประสิทธิภาพของระบบการตัดส่วนข้อมูลที่ใช้ Algorithm Rote Learning และระบบการตัดส่วนข้อมูลที่ใช้ Algorithm Bayes Learning ทั้งสามวิธี เพื่อแสดงประสิทธิภาพของระบบการตัดส่วนข้อมูลที่ใช้วิธีการต่าง ๆ ในฟิลด์ (Field) ต่าง ๆ ได้แก่ เวลาเริ่มอบรมการสัมมนา เวลาเลิกการอบรมสัมมนา ผู้อภิปราย และ สถานที่อบรม โดยผลการพัฒนาระบบแสดงดังตารางที่ 3. 1 ตารางที่ 2 แสดงความเชื่อมั่นในคำตอบเมื่อกำหนดค่าความถูกต้องที่ 25% และตารางที่ 3 แสดงค่า F-measure

	ผู้อภิปราย		สถานที่การสัมมนา		เวลาเริ่มการสัมมนา		เวลาเลิกการสัมมนา	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Rote	55.1	6.8	89.5	58.1	73.7	73.4	37.4	95.7
Bayes	10.0	11.8	32.8	34.3	96.2	96.2	42.6	91.7
BayesLN	11.5	13.6	44.8	46.9	98.1	98.1	44.4	95.6
BayesIDF	28.8	27.4	57.3	58.8	98.2	98.2	46.8	95.7

ตารางที่ 3. 1 แสดงค่าความเชื่อมั่นในคำตอบ (Prec) และค่าความถูกต้อง (Rec) ใน Algorithm แบบต่าง ๆ

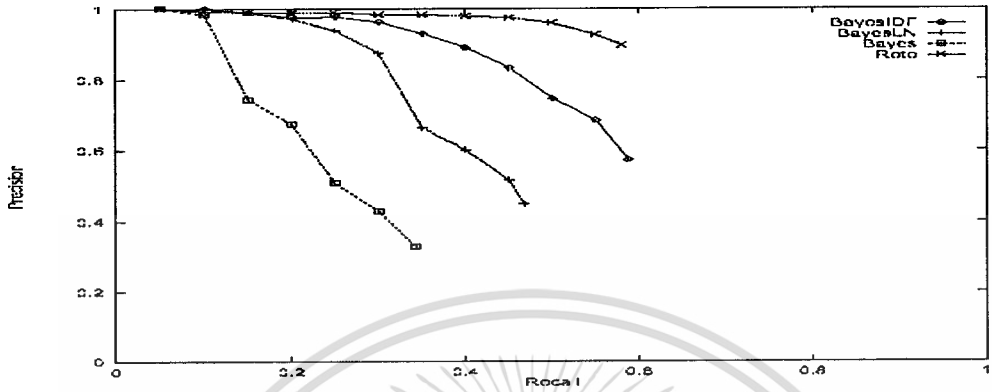
	ผู้อภิปราย		สถานที่การสัมมนา		เวลาเริ่มการสัมมนา		เวลาเลิกการสัมมนา	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Rote	-	-	99.2±1.2	24.8	78.2±4.0	26.3	79.4±5.7	27.3
Bayes	-	-	50.7±4.1	25.0	100.0±0.0	25.1	100.0±0.0	25.7
BayesLN	-	-	93.9±2.7	25.0	100.0±0.0	25.1	100.0±0.0	25.0
BayesIDF	35.6±3.5	25.0	97.7±1.7	25.2	100.0±0.0	25.3	100.0±0.0	25.0

ตารางที่ 3. 2 แสดงค่าความเชื่อมั่นในคำตอบ (Prec) ที่ค่าความถูกต้อง (Rec) 25 % ใน Algorithm แบบต่าง ๆ

จากผลการทดลองทำให้เราทราบว่าวิธี BayesIDF ให้คำตอบได้ดีที่สุดทุกฟิลด์ที่ทำการทดสอบ และ Algorithm แบบ Bayes ทั้งสามแบบให้คำตอบได้เกือบถูกต้อง 100% ในฟิลด์เวลาเริ่มอบรมสัมมนา เมื่อเปรียบเทียบประสิทธิภาพระหว่าง BayesIDF กับ Rote จะเห็นว่า BayesIDF ให้

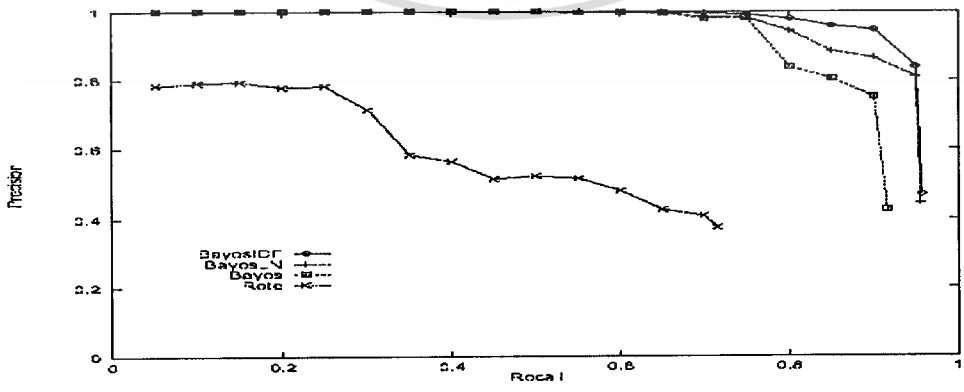
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่า F1 ได้มากที่ฟิลด์ ผู้อภิปราย ฟิลด์เวลาเริ่มอบรมสัมมนา และ ฟิลด์เวลาเลิกอบรมสัมมนา มีเพียงฟิลด์สถานที่จัดอบรมสัมมนาที่วิธี Rote ให้คำตอบได้ดีกว่า



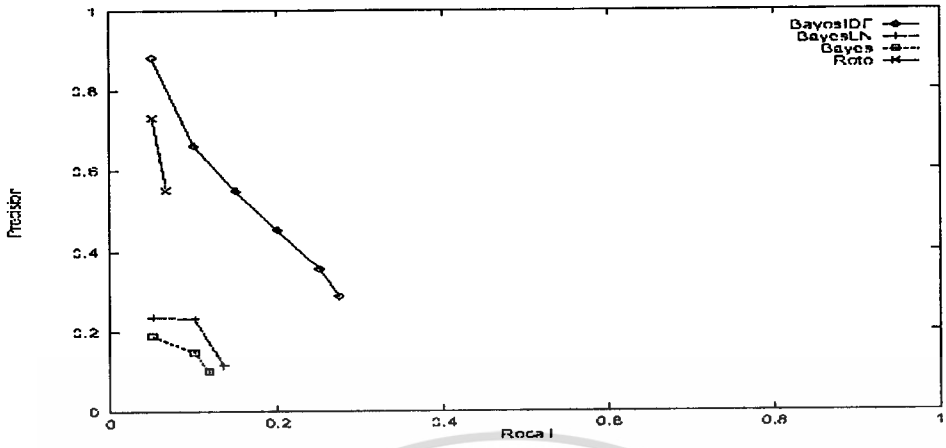
รูปที่ 3. 7 แสดงความสัมพันธ์ระหว่างค่าความเชื่อมั่นในคำตอบกับค่าความถูกต้องในหัวข้อสถานที่จัดอบรมสัมมนา

รูปที่ 3. 8 แสดงความสัมพันธ์ระหว่างค่าความเชื่อมั่นในคำตอบกับค่าความถูกต้องในหัวข้อสถานที่เวลาเริ่มจัดอบรมสัมมนา



รูปที่ 3. 9 แสดงความสัมพันธ์ระหว่างค่าความเชื่อมั่นในคำตอบกับค่าความถูกต้องในหัวข้อสถานที่เวลาเลิกจัดอบรมสัมมนา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3. 10 แสดงความสัมพันธ์ระหว่างค่าความเชื่อมั่นในคำตอบกับค่าความถูกต้องในหัวข้อผู้อภิปรายอบรมสัมมนา

จากรูปที่ 3. 7 ถึงรูปที่ 3. 10 แสดงกราฟความสัมพันธ์ระหว่างค่าความเชื่อมั่นในคำตอบกับค่าความถูกต้องที่ค่า ๆ ต่างโดยกราฟที่ได้หาได้โดยการเปลี่ยนค่าขอบเขตของคะแนนที่ยอมรับได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
การพัฒนาระบบ

ในโครงงานพัฒนาระบบการตัดส่วนข้อมูลสำหรับร้านหนังสือออนไลน์ ได้พัฒนาระบบค้นหาและเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ผ่านระบบเว็บโดยจะเน้นการพัฒนาเพื่อให้สามารถนำไปใช้งานได้จริง

4.1 ส่วนประกอบของระบบค้นหาและเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ผ่านระบบเว็บ

การทำงานของระบบค้นหาและเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์ผ่านระบบเว็บสามารถแสดง Context Diagram ได้ดังรูปที่ 4. 1

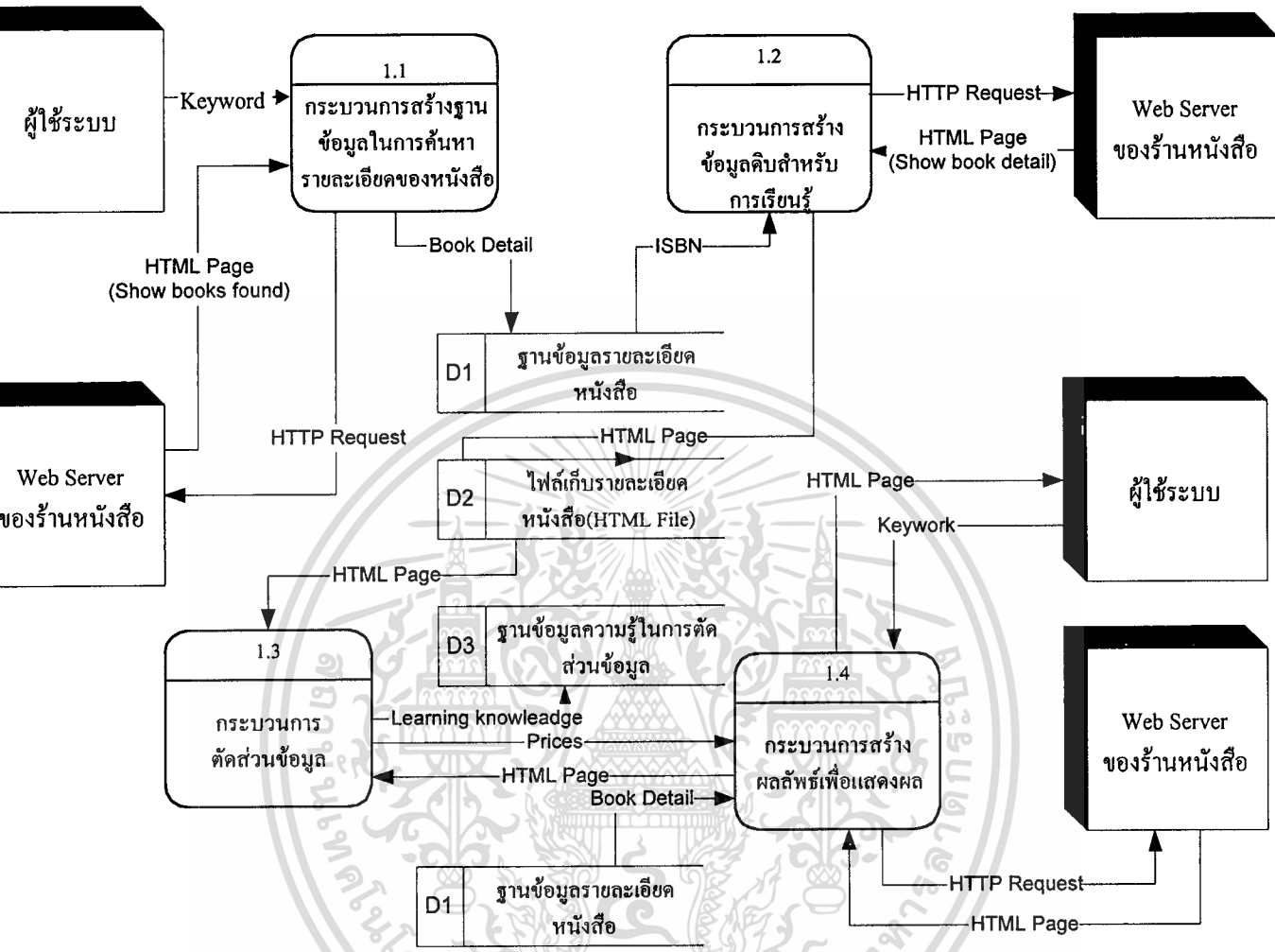
รูปที่ 4. 1 แสดง Context Diagram ของระบบ

จากการทำงานรูปที่ 4.1 สามารถแบ่งฟังก์ชันการทำงานย่อยได้อีก 4 ฟังก์ชันดังนี้

1. ฟังก์ชันการสร้างฐานข้อมูลในการค้นหารายละเอียดของหนังสือ
2. ฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้
3. ฟังก์ชันการตัดส่วนข้อมูล
4. ฟังก์ชันการแสดงผล

ซึ่งสามารถแสดง Data Flow Diagram Level 1 ได้ดังรูปที่ 4. 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4. 2 แสดง Data Flow Diagram Level 1 ของระบบ

4.2 ฟังก์ชันการสร้างฐานข้อมูลในการค้นหารายละเอียดของหนังสือ

ในการให้บริการค้นหาข้อมูลเพื่อใช้ในการเปลี่ยนเทียบราคา จำเป็นต้องมีระบบฐานข้อมูลเพื่อใช้เก็บข้อมูลเกี่ยวกับหนังสือต่าง ๆ อาทิเช่น ชื่อหนังสือ, ผู้แต่ง, ISBN และราคาหน้าปกหนังสือ

4.2.1 ส่วนประกอบของฟังก์ชันการสร้างฐานข้อมูลในการค้นหารายละเอียดของหนังสือ

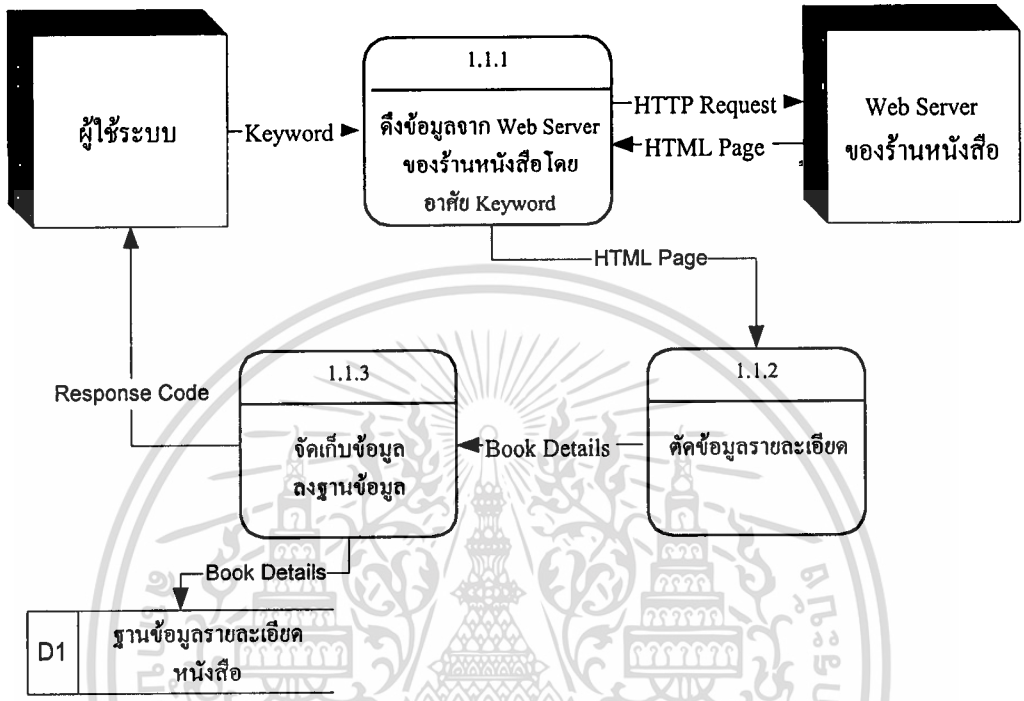
จากการทำงานที่แสดงดังรูปที่ 4. 2 แสดงให้เห็นการทำงานอย่างคร่าว ๆ ของระบบโดยรวม ซึ่งในฟังก์ชันการสร้างฐานข้อมูลในการค้นหารายละเอียดของหนังสือสามารถแบ่งเป็นฟังก์ชันการทำงานได้ 3 ส่วนดังนี้

4.2.1.1 ฟังก์ชันการทำงานในการดึงข้อมูลโดยอาศัย Keyword

4.2.1.2 ฟังก์ชันการทำงานในการตัดส่วนข้อมูลรายละเอียด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.3 ฟังก์ชันการทำงานในการจัดเก็บข้อมูลลงฐานข้อมูลโดยสามารถเขียน Data Flow Diagram Level 2 ซึ่งสามารถอธิบายเฉพาะการทำงานของฟังก์ชันได้ดังรูปที่ 4. 3



รูปที่ 4. 3 แสดง Data Flow Diagram Level 2 ของระบบเฉพาะฟังก์ชันการสร้างฐานข้อมูลในการค้นหารายละเอียดของหนังสือ

4.2.1.1 ฟังก์ชันการทำงานในการดึงข้อมูลโดยอาศัย Keyword

ฟังก์ชันการทำงานนี้ทำหน้าที่ในการดึงข้อมูลจาก Web Server ของร้านหนังสือออนไลน์ขนาดใหญ่สำหรับในโครงงานนี้ใช้ Web Server ของ Amazon ซึ่งมีจำนวนหนังสือมาก ความเร็วในการตอบสนองสูง มีหนังสือหลากหลายประเภทและมีโครงสร้างของ HTML Page ที่แสดงผลให้กับผู้ใช้ง่าย

การทำงานของฟังก์ชันการทำงานนี้เริ่มจากรับข้อมูล Keyword จากผู้ใช้เพื่อทำการสร้าง URL Request ที่จำเป็นต้องใช้ในการส่งให้กับ Web Server เพื่อขอผลลัพธ์ที่ได้จากการค้นหาหนังสือด้วย Keyword ดังกล่าวในรูปของ HTML Page แล้วทำการส่ง HTML Page ให้กับฟังก์ชันการทำงานในการตัดส่วนข้อมูลรายละเอียดเพื่อทำงานต่อไป แต่โดยทั่วไปข้อมูลที่ค้นพบจาก Keyword ดังกล่าวจะถูกจำกัดจำนวนการแสดงผลทำให้ข้อมูลที่ได้รับในรูปแบบของ HTML Page มีปริมาณน้อยไม่เพียงพอต่อการใช้งานจึงต้องมีการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Rerequest กลับไปเพื่อขอรายละเอียดของหนังสือในส่วนที่เหลือจนกว่าจะครบจำนวนหนังสือทั้งหมดที่ค้นพบด้วย Keyword ดังกล่าว

จากลักษณะการทำงานดังกล่าวทำให้ Keyword ที่ใช้ในการค้นหาจำเป็นต้อง Keyword ที่ทำให้ค้นหาหนังสือได้มากโดยในโครงงานนี้ใช้ Keyword ในการค้นหา 4 คำ ได้แก่ Life, Food, Computer และ Java

4.2.1.2 ฟังก์ชันการทำงานในการตัดส่วนข้อมูลรายละเอียด

หลังจากที่ฟังก์ชันการทำงานในการดึงข้อมูลโดยอาศัย Keyword ส่งผลลัพธ์ที่เป็น HTML Page กลับมาแล้ว ข้อมูลที่เราได้มายังไม่สามารถเก็บเข้าฐานข้อมูลได้ทำให้เราทำการตัดส่วนข้อมูลก่อนส่งให้ฟังก์ชันการทำงานในการเก็บข้อมูลลงในฐานข้อมูล

การทำงานของฟังก์ชันการทำงานในการตัดส่วนข้อมูลรายละเอียดจะใช้วิธี Pattern Matching เพราะรูปแบบของ Web Page มีลักษณะค่อนข้างคงที่แต่จะมีข้อมูลบางส่วนที่ไม่สามารถตัดส่วนย่อยได้เราจะไม่สนใจข้อมูลดังกล่าว

4.2.1.3 ฟังก์ชันการทำงานในการจัดเก็บข้อมูลลงฐานข้อมูล

หลังจากที่ได้ข้อมูลจากฟังก์ชันการทำงานในการตัดส่วนข้อมูลรายละเอียดแล้วเราจะนำข้อมูลดังกล่าวจัดเก็บลงฐานข้อมูลเพื่อใช้ในการค้นหาข้อมูลอีกที่หนึ่ง

4.3 ฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้

ฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้เป็นส่วนประกอบหนึ่งที่จำเป็นในระบบการตัดส่วนข้อมูลสำหรับร้านหนังสือออนไลน์ เนื่องจากว่าการตัดส่วนข้อมูลจำเป็นต้องอาศัยข้อมูลดิบเพื่อใช้ในการสร้าง Prior Knowledge สำหรับข้อมูลดิบที่ใช้จะอาศัยข้อมูลรายละเอียดหนังสือจาก Web Site ที่ให้บริการสั่งซื้อหนังสือผ่านระบบออนไลน์ 4 แห่งได้แก่

1. www.amazon.com
2. www.barnandnobel.com
3. www.1bookstreet.com
4. www.powells.com

โดยประเภทหนังสือที่จะใช้ในการสร้างจะได้จากการสุ่ม เพื่อให้ข้อมูลดิบมีลักษณะแตกต่างกันและปริมาณข้อมูลที่เตรียมสำหรับการเรียนรู้จะใช้จำนวน 50 เล่ม เหมือน ๆ กันในทุก Site เพื่อให้สามารถเปรียบเทียบประสิทธิภาพในการทำงานโดยรวมของระบบได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การทำงานของระบบนี้ต้องอาศัยการแท็กข้อมูลคำตอบ โดยอาศัยคนเพื่อให้ข้อมูลที่ได้มีความถูกต้องสูงและมีความน่าเชื่อถือ

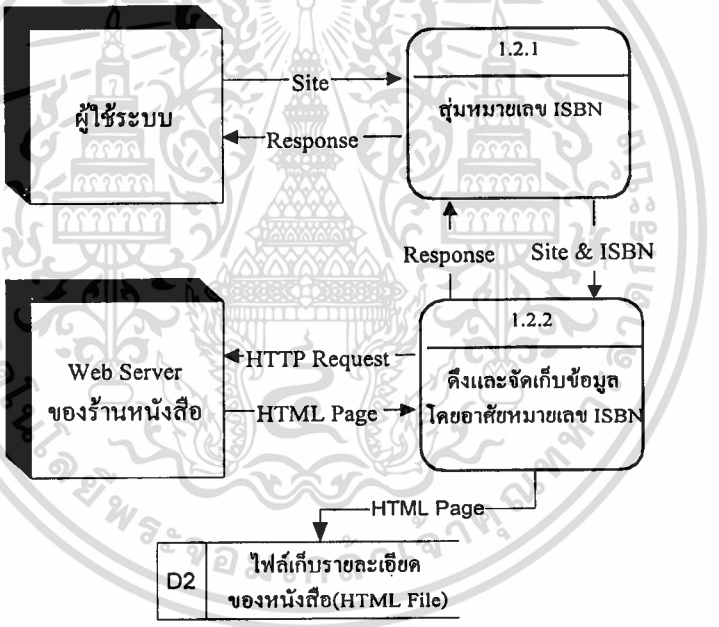
4.3.1 ส่วนประกอบของฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้

จากการทำงานที่แสดงดังรูปที่ 4. 2 แสดงให้เห็นการทำงานอย่างคร่าว ๆ ของระบบ ซึ่งในฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้สามารถแบ่งเป็นฟังก์ชันการทำงานได้ 2 ส่วนดังนี้

4.3.1.1 ฟังก์ชันการทำงานในการสุ่มหมายเลข ISBN

4.3.1.2 ฟังก์ชันการทำงานในการดึงและจัดเก็บข้อมูลโดยอาศัยหมายเลข ISBN

โดยการทำงานฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้สามารถแสดง Data Flow Diagram Level 2 แสดงเฉพาะส่วนการทำงานของฟังก์ชันได้ดังรูปที่ 4. 4



รูปที่ 4. 4 แสดง Data Flow Diagram Level 2 ของระบบเฉพาะฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้

4.3.1.1 ฟังก์ชันการทำงานในการสุ่มหมายเลข ISBN

ฟังก์ชันการทำงานนี้ทำหน้าที่ในการสุ่มหมายเลข ISBN จากฐานข้อมูลรายละเอียดของหนังสือที่ได้จากฟังก์ชันการสร้างฐานข้อมูลรายละเอียดหนังสือในหัวข้อที่ 4.2 เพื่อทำการส่งหมายเลข ISBN ให้กับฟังก์ชันการทำงานถัดไป โดยจะทำการสุ่มหมายเลข ISBN จำนวนทั้งสิ้น 50 ครั้งเหมือน ๆ กันในทุก Site

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.1.2 ฟังก์ชันการทำงานในการดึงและจัดเก็บข้อมูลโดยอาศัยหมายเลข ISBN

ฟังก์ชันการทำงานของฟังก์ชันนี้จะคล้ายกับฟังก์ชันการทำงานในหัวข้อที่ 4.2.1.1 แตกต่างกันเพียง URL Request ที่ส่งให้กับ Web Server จะใช้หมายเลข ISBN แทน

หลังจากที่ได้ HTML Page แล้วก็จะทำการเก็บลงแยกไว้เป็นไฟล์ ๆ และทำการตัด Tag ต่าง ๆ ในเอกสาร HTML เพื่อเข้าสู่ขั้นตอนการแท็กข้อมูลคำตอบโดยข้อมูลที่ได้จากการแท็กจะแยกเก็บคนละส่วนกับเอกสาร HTML

4.4 ฟังก์ชันการตัดส่วนข้อมูล

ฟังก์ชันการตัดส่วนข้อมูลเป็นหัวใจสำคัญของระบบการตัดส่วนข้อมูลของร้านหนังสือออนไลน์ โดยฟังก์ชันนี้จะใช้วิธีการจากบทที่ 3 เป็นหลัก

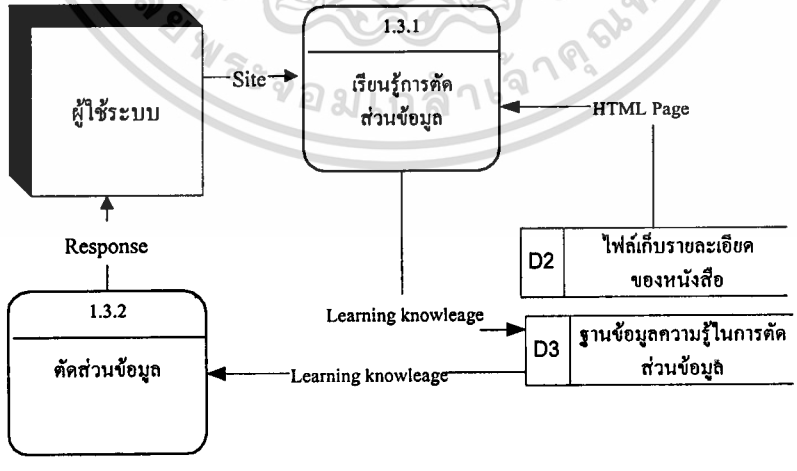
4.4.1 ส่วนประกอบของฟังก์ชันการตัดส่วนข้อมูล

จากการทำงานที่แสดงดังรูปที่ 4. 2 แสดงให้เห็นการทำงานอย่างคร่าว ๆ ของระบบโดยรวม ซึ่งในฟังก์ชันการตัดส่วนข้อมูลสามารถแบ่งเป็นฟังก์ชันการทำงานได้ 2 ส่วนดังนี้

4.3.1.1.1 ฟังก์ชันการเรียนรู้การตัดส่วนข้อมูล

4.3.1.1.2 ฟังก์ชันการตัดส่วนข้อมูล

โดยการทำงานฟังก์ชันการตัดส่วนข้อมูลสามารถแสดง Data Flow Diagram Level 2 แสดงเฉพาะส่วนการทำงานของฟังก์ชันได้ดังรูปที่ 4. 5



รูปที่ 4. 5 แสดง Data Flow Diagram Level 2 ของระบบเฉพาะฟังก์ชันการตัดส่วนข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.1.1 ฟังก์ชันการเรียนรู้การตัดส่วนข้อมูล

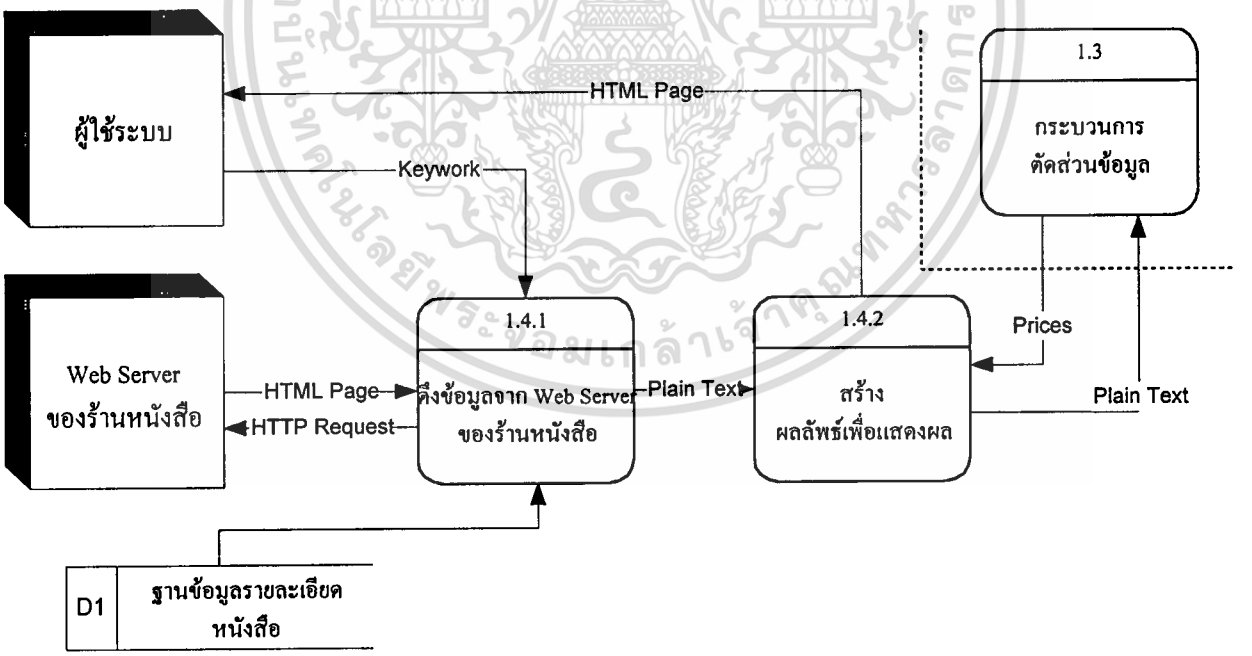
เป็นฟังก์ชันการทำงานซึ่งทำหน้าที่นำข้อมูลที่ได้จากการฟังก์ชันการสร้างข้อมูลดิบสำหรับการเรียนรู้ในหัวข้อที่ 4.3 มาทำการเรียนรู้โดยใช้ Algorithm จากตารางที่ 3. 1 ในบทที่ 3 ซึ่งผลลัพธ์ที่ได้จากการเรียนรู้จะถูกเก็บลงในฐานข้อมูลการเรียนรู้เพื่อสะดวกในการใช้งาน

4.4.1.2 ฟังก์ชันการตัดส่วนข้อมูล

ฟังก์ชันการทำงานนี้ทำหน้าที่นำข้อมูลที่ได้จากการฟังก์ชันในการเรียนรู้การตัดส่วนข้อมูลซึ่งเก็บอยู่ในฐานข้อมูลมาทำการตัดส่วนข้อมูลโดยอาศัย Algorithm จากบทที่ 3 ซึ่งได้แก่ ตารางที่ 3.3 ตารางที่ 3.4 และ ตารางที่ 3.5

4.5 ฟังก์ชันการแสดงผล

ฟังก์ชันการแสดงผลเป็นระบบที่ทำหน้าที่ในการติดต่อสื่อสารกับผู้ใช้งาน โดยการทำงานจะเป็นอาศัยเทคโนโลยี CGI



รูปที่ 4. 6 แสดง Data Flow Diagram Level 2 ของระบบเฉพาะฟังก์ชันการแสดงผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.1 ส่วนประกอบของฟังก์ชันการแสดงผล

จากการทำงานที่แสดงดังรูปที่ 4. 2 แสดงให้เห็นการทำงานอย่างคร่าว ๆ ของระบบโดยรวม ซึ่งในฟังก์ชันการแสดงผลสามารถแบ่งเป็นฟังก์ชันการทำงานได้ 2 ส่วนดังนี้

4.2.1.1 ฟังก์ชันการดึงข้อมูลจาก Web Server ของร้านหนังสือ

4.2.1.2 ฟังก์ชันการสร้างผลลัพธ์

โดยการทำงานฟังก์ชันการแสดงผลสามารถแสดง Data Flow Diagram Level 2 แสดงเฉพาะส่วนการทำงานของฟังก์ชันได้ดังรูปที่ 4. 6

4.6 รายละเอียดฐานข้อมูล

ฐานข้อมูลที่ใช้ในโครงงานนี้ใช้ฐานข้อมูล MySQL Version 3.23.38 ซึ่งเป็นฐานข้อมูลที่มีประสิทธิภาพมีความเร็วในการดึงข้อมูลสูง เหมาะกับงานที่ไม่มีความซับซ้อนมาก ซึ่งในโครงงานนี้ได้ออกแบบฐานข้อมูลที่ใช้เก็บรายละเอียดหนังสือและความรู้ในการตัดส่วนข้อมูลให้ไม่มีความซับซ้อน เพื่อความเร็วในดึงข้อมูล

ฐานข้อมูลที่ใช้ในโครงงานนี้ประกอบด้วย

1. ฐานข้อมูลเก็บรายละเอียดหนังสือ
2. ฐานข้อมูลเก็บความรู้ในการตัดส่วนข้อมูล

โดยรายละเอียดของแต่ละฐานข้อมูลสามารถแสดงได้ดังนี้

1. ฐานข้อมูลเก็บรายละเอียดหนังสือประกอบด้วยตารางดังนี้

   ตาราง Book_Detail ซึ่งประกอบด้วย Field ดังนี้

   - Title VARCHAR(50) เก็บรายชื่อหนังสือแต่ละเล่ม
   - Author VARCHAR(50) เก็บรายชื่อผู้แต่งของหนังสือ
   - ISBN VARCHAR(10) เก็บหมายเลขหนังสือ ISBN
   - Price FLOAT เก็บราคาตามปกของหนังสือ

2. ฐานข้อมูลความรู้ในการตัดส่วนข้อมูลประกอบด้วยตารางดังนี้

   ตาราง Before_Data1 ซึ่งประกอบด้วย Field ดังนี้

   - Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
   - Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตาราง Before_Data2 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง Before_Data3 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง Before_Data4 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง After_Data1 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง After_Data2 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง After_Data3 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง After_Data4 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตาราง IN_Data1 ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง Position ซึ่งประกอบด้วย Field ดังนี้

- Position INTEGER เก็บตำแหน่งของคำตอบที่พบในการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

ตาราง Length ซึ่งประกอบด้วย Field ดังนี้

- Length INTEGER เก็บความยาวของคำตอบที่พบในการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

4.7 รายละเอียดของโปรแกรม

โปรแกรมที่พัฒนาในโครงงานนี้พัฒนาด้วยภาษา Perl ซึ่ง Run บน Linux OS โดยได้ติดตั้ง Module เพิ่มเติมดังนี้

1. DBI Module : เป็น Module สำหรับสร้าง Interface ติดต่อฐานข้อมูลสำหรับ Perl
2. Msql-Mysql-modules เป็น Module สำหรับ ติดต่อฐาน MySQL ผ่านทาง Interface ที่ได้จากการติดตั้ง DBI Module สำหรับการเขียนโปรแกรมด้วยภาษา Perl

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
การปรับปรุงระบบ

จากการศึกษาการทำงานของระบบที่ได้พัฒนาในบทที่ 4 พบว่า เวลาที่ใช้ในการประมวลผลและส่งผลลัพธ์ออกมาให้ผู้ใช้เป็นไปด้วยความล่าช้า อีกทั้งเทคนิคในการแสดงผลในรูปแบบของเอกสาร HTML ธรรมดามีความยืดหยุ่นจำกัดทำให้ผู้ใช้ไม่สะดวกสบายในการใช้งานในบทนี้จะนำเสนอวิธีการปรับปรุงระบบให้มีความสามารถในการทำงานที่ดีขึ้น

5.1 การปรับปรุงประสิทธิภาพในการตัดส่วนข้อมูล

จากการศึกษาการทำงานของระบบตัดส่วนข้อมูลสำหรับร้านหนังสือออนไลน์พบว่า เวลาที่ใช้ในการทำงานส่วนใหญ่ของระบบจะอยู่ในฟังก์ชั่นการตัดส่วนข้อมูล เนื่องจากขนาดของข้อมูลที่ใช้ในการทดสอบมีขนาดใหญ่ อีกทั้งระยะเวลาในการรอข้อมูลที่อยู่ในรูปของ HTML Page จากเว็บเซิร์ฟเวอร์แต่ละแห่งใช้เวลาค่อนข้างมากเมื่อการทำงานของระบบเป็นลักษณะ Sequence ทำให้ระยะเวลาในการรอข้อมูลจึงมีมาก

จากเหตุผลดังกล่าวการปรับปรุงประสิทธิภาพในการตัดส่วนข้อมูลจึงสามารถทำได้โดย

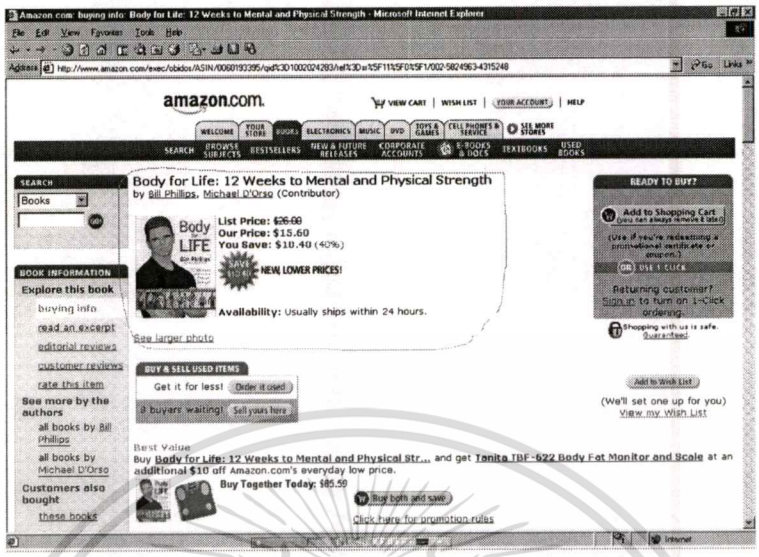
1. ลดขนาดของข้อมูลในการตัดส่วนข้อมูล
2. แตก Process ย่อยให้ทำงานพร้อม ๆ
3. กระจายงานให้ทำงานกันคนละเครื่อง

5.1.1 การลดขนาดของข้อมูลในการตัดส่วนข้อมูล

จากผลการศึกษาการทำงานของระบบพบว่าข้อมูลที่ใช้ในการทดสอบการตัดส่วนของข้อมูลมีขนาดใหญ่ และคำตอบของเอกสารจะอยู่ในส่วนของข้อมูลในช่วง ๆ หนึ่งเท่านั่นดังแสดงได้ดัง

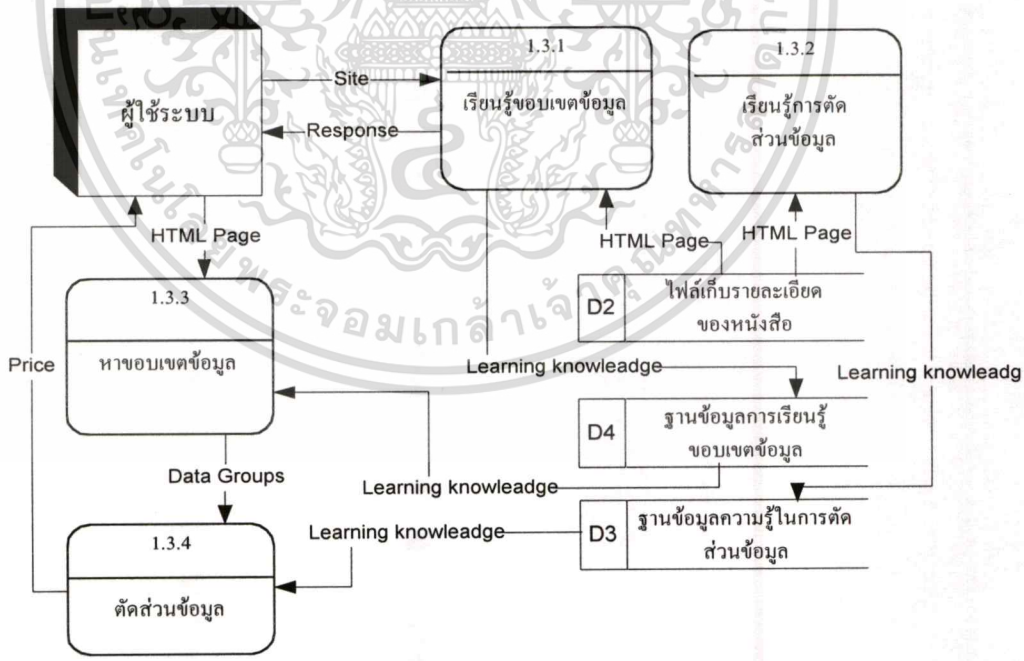
การปรับปรุงวิธีการตัดส่วนข้อมูลเพื่อแก้ปัญหาดังกล่าว สามารถทำได้โดยเพิ่มระบบย่อยในการเรียนรู้เพื่อหาขอบเขตของข้อมูล โดยระบบย่อยการตัดส่วนข้อมูลสามารถหลังการเพิ่มการหาขอบเขตของข้อมูลสามารถแสดงได้ดังรูปที่ 5. 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5. 1 แสดงกลุ่มข้อมูลที่สามารถลดขนาดได้

และจากการทำงานดังรูปที่ 5. 2 แสดงให้เห็นว่าเราต้องมีข้อมูล Split Data เพื่อใช้ในการเรียนรู้ส่วนข้อมูล ซึ่ง Split Data สามารถสร้างได้โดยการสังเกตขอบเขตของข้อมูลดิบโดยฐานข้อมูล D3 ยังมีโครงสร้างเหมือนระบบเดิม



รูปที่ 5. 2 แสดง DFD Level 2 ของระบบย่อยการตัดส่วนข้อมูลหลังจากเพิ่มระบบการหาขอบเขตข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

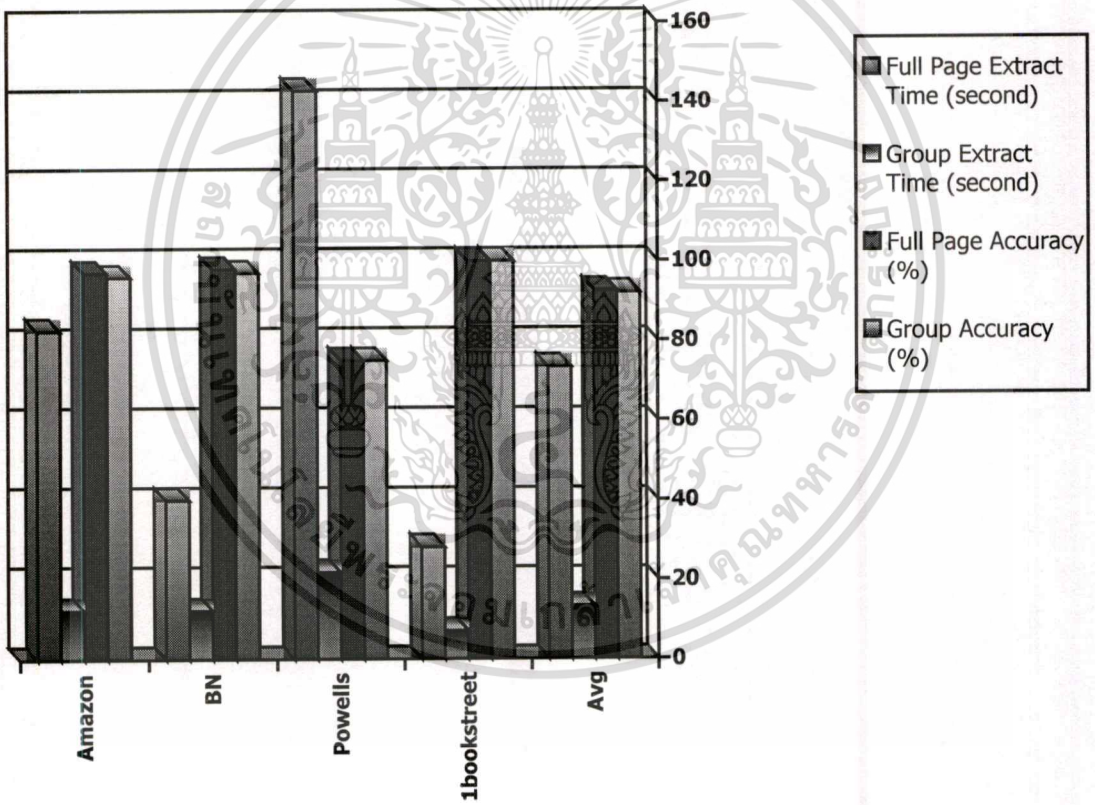
รายละเอียดฐานข้อมูลการเรียนรู้ของเขตข้อมูลประกอบด้วยตารางดังนี้

ตาราง Data_Group ซึ่งประกอบด้วย Field ดังนี้

- Token VARCHAR(50) เก็บ Token ของคำที่ได้จากการเรียนรู้
- Freq INTEGER เก็บความถี่ของคำที่พบในการเรียนรู้

5.1.1.1 ผลการลดขนาดของข้อมูลในการตัดส่วนข้อมูล

ผลการลดขนาดข้อมูลก่อนการตัดส่วนข้อมูลทำให้เวลาที่ใช้ในการทำงานลดลงประมาณ 70% โดยที่ความถูกต้องของการตัดส่วนข้อมูลลดลงไม่เกิน 2 % โดยสามารถแสดงได้ดังรูปที่ 5. 3 และตารางที่ 5. 1



รูปที่ 5. 3 แสดงผลการทำงานหลังจากผ่านการลดขนาดข้อมูลโดยการหากลุ่มข้อมูลที่สนใจ

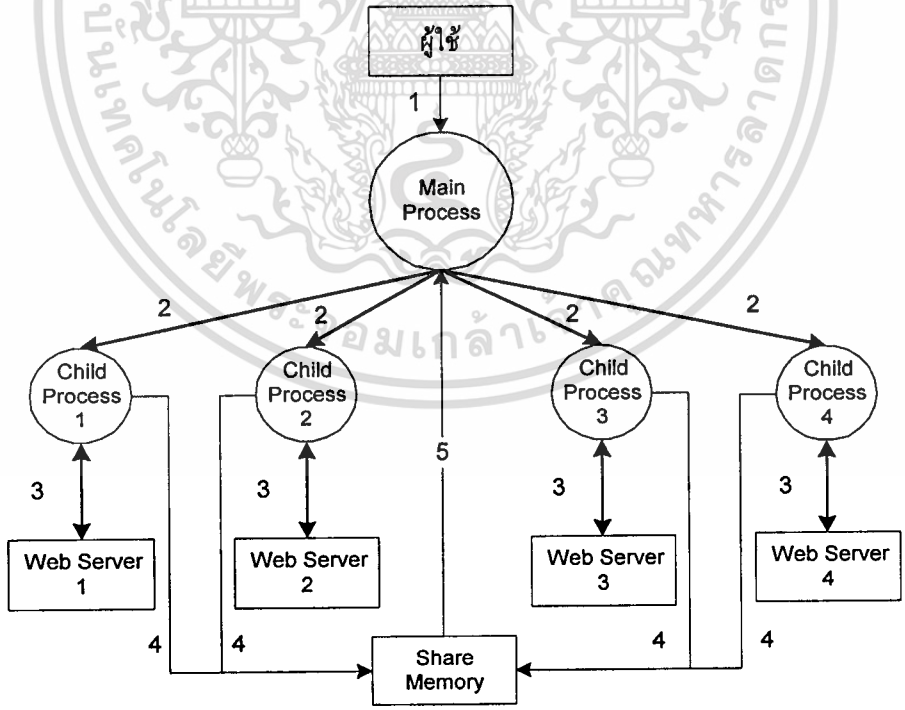
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	Amazon	BN	Powells	1bookstreet	Avg
Full Page Extract Time (second)	82.5	39.99	142.93	28.08	73.375
Group Extract Time (second)	12.34	12.22	21.85	7.15	13.39
Full Page Accuracy (%)	97	98	75	100	92.5
Group Accuracy (%)	96	97	75	100	92

ตารางที่ 5. 1 แสดงผลการทำงานหลังจากผ่านการลดขนาดข้อมูลโดยการหากลุ่มข้อมูลที่สนใจ

5.1.2 การแตกโปรเซสย่อยให้ทำงานพร้อม ๆ กัน

การแตกโปรเซสย่อยให้ทำงานพร้อม ๆ กันทำเพื่อลดระยะเวลาการ Block ที่เกิดจากรอข้อมูลของจากเว็บเซิร์ฟเวอร์ โดยสามารถแสดงการทำงานได้ดังรูปที่ 5. 4



รูปที่ 5. 4 แสดงการทำงานของโปรเซสในการร้องข้อมูลจากจากเว็บเซิร์ฟเวอร์หลังจากปรับปรุง

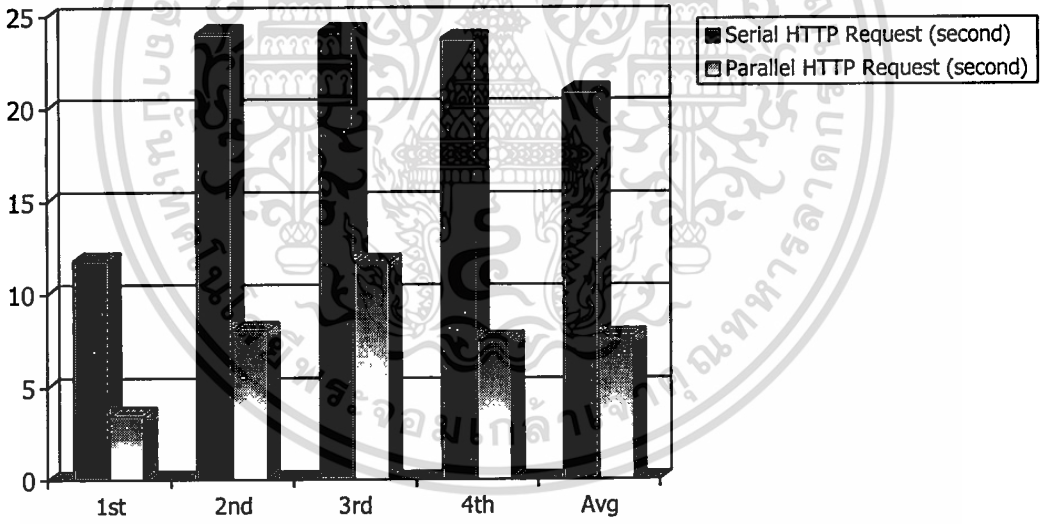
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 5. 4 เราสามารถอธิบายขั้นตอนการทำงานได้ดังนี้

1. รับ Request ในการร้องขอจากผู้ใช้
2. โปรเซสหลักแตกโปรเซสย่อยเพื่อร้องขอข้อมูลจากเว็บเซิร์ฟเวอร์แต่ละแห่ง
3. แต่ละโปรเซสย่อยร้องขอข้อมูลจากเว็บเซิร์ฟเวอร์
4. แต่ละโปรเซสย่อยจะนำข้อมูลที่ได้จากเว็บเซิร์ฟเวอร์มาเก็บไว้ใน Share Memory
5. โปรเซสหลักหลังจากทราบว่าการทำงานของโปรเซสย่อยทั้งหมดเสร็จสิ้นจะทำการอ่านข้อมูลจาก Share Memory

5.1.2.1 ผลการทำงานจากการการแตกโปรเซสย่อยให้ทำงานพร้อม ๆ กัน

ผลการลดทำงานการแตกโปรเซสย่อยให้ทำงานพร้อม ๆ กันทำให้เวลาที่ใช้ในการรอผลลัพธ์จาก Web Server ของร้านหนังสือลดลงประมาณ 65 % โดยเฉลี่ยโดยสามารถแสดงผลการทำงานได้ดังรูปที่ 5. 5 และตารางที่ 5. 2



รูปที่ 5. 5 แสดงผลการผลการลดทำงานการแตกโปรเซสย่อยให้ทำงานพร้อม ๆ

	1st	2nd	3rd	4th	Avg
Serial HTTP Request (second)	11.81	24.03	24.11	23.72	20.91
Parallel HTTP Request (second)	3.43	8.00	11.71	7.49	7.66

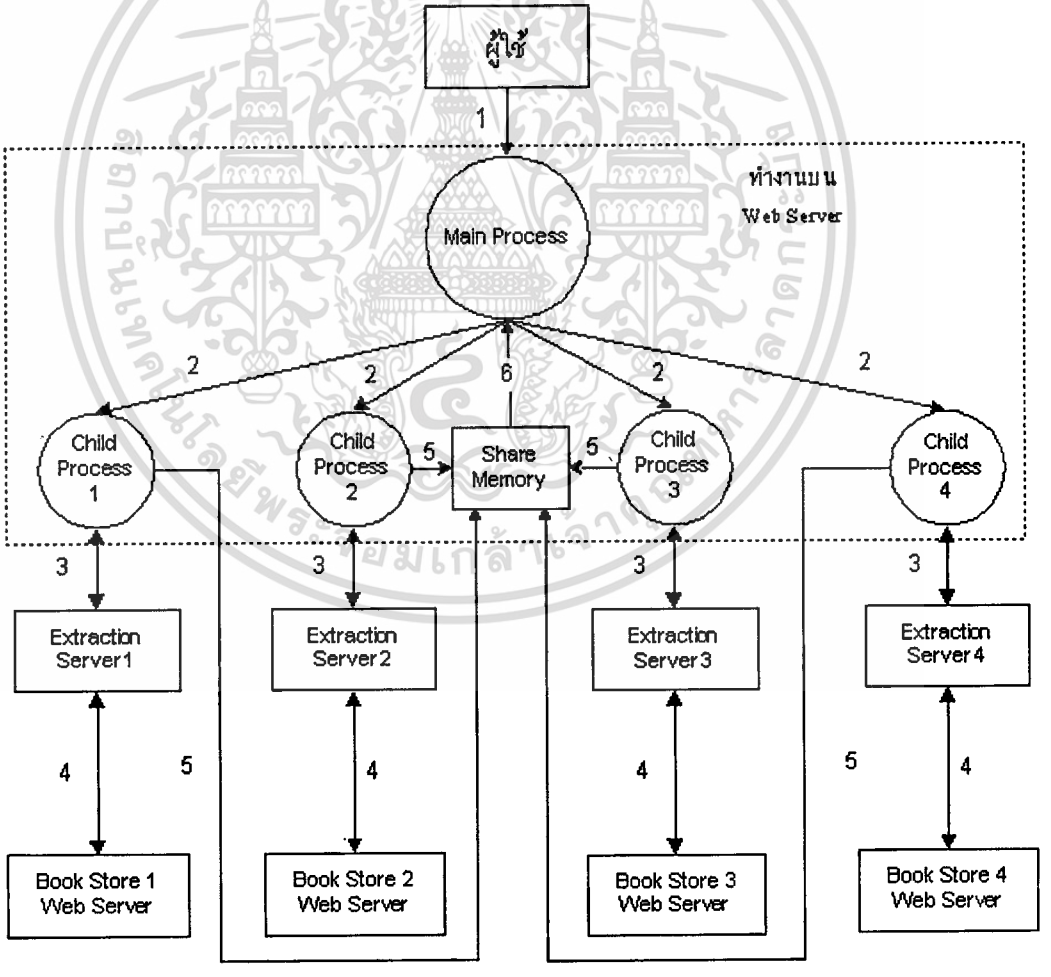
ตารางที่ 5. 2 แสดงผลการผลการลดทำงานการแตกโปรเซสย่อยให้ทำงานพร้อม ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.1.3 กระจายงานให้ทำงานกันคนละเครื่อง

จากการศึกษาผลการทำงานของระบบหลังจากปรับปรุงประสิทธิภาพโดยการลดขนาดข้อมูลและการแตกโปรเซส พบว่าการทำงานของระบบยังไม่นำสามารถนำไปใช้งานได้จริงเนื่องจากต้องการใช้ CPU-Time สูงมาก

การแก้ปัญหาดังกล่าวสามารถทำได้โดยการกระจายงานให้ทำงานกันคนละเครื่อง ในโครงงานนี้จะใช้เครื่องคอมพิวเตอร์ที่ทำหน้าที่ในการบริการจำนวน 5 เครื่อง โดยแบ่งเป็นเครื่องที่ทำหน้าที่เป็นเว็บเซิร์ฟเวอร์จำนวน 1 เครื่องและ เครื่องที่ทำหน้าที่ตัดส่วนข้อมูลจำนวน 4 เครื่องโดยแต่ละเครื่องจะมี Daemon Process รันอยู่โดยการติดต่อสื่อสารจะทำผ่าน Port 2222 และ มีฐานข้อมูลในการเรียนรู้เป็นของตัวเองโดยสามารถแสดงการทำงานได้ดังรูปที่ 5. 6

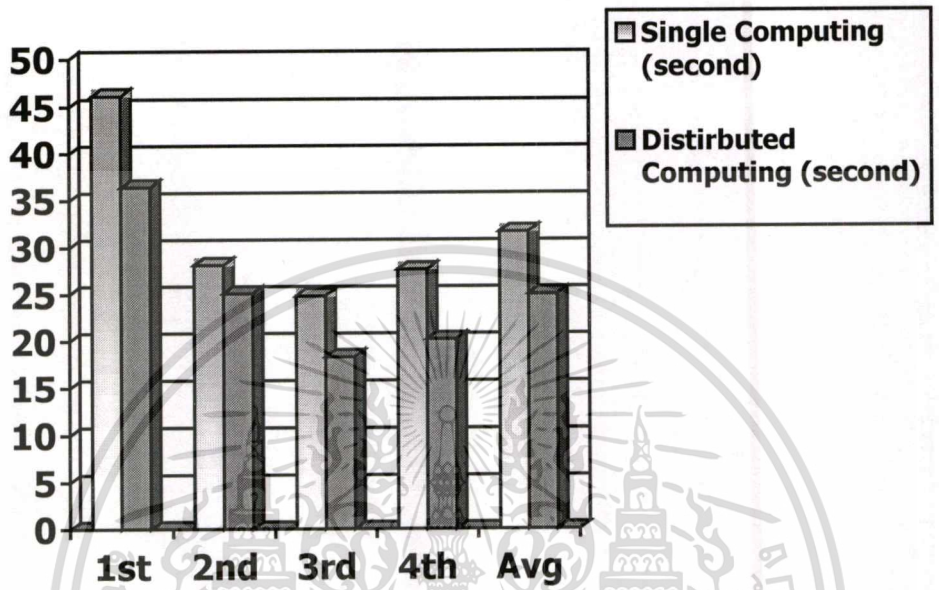


รูปที่ 5. 6 แสดงระบบโดยรวมหลังจากการกระจายงาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.1.3.1 ผลการทำงานกระจายงานให้ทำงานกันคนละเครื่อง

ผลการทำงานจากการกระจายงานให้ทำงานคนละเครื่องทำให้ความเร็วในการตัดส่วนข้อมูลเร็วขึ้นประมาณ 20 % โดยเฉลี่ยซึ่งสามารถแสดงได้รูปที่ 5. 7 และตารางที่ 5. 3



รูปที่ 5. 7 แสดงผลการทำงานกระจายงานให้ทำงานกันคนละเครื่อง

	1st	2nd	3rd	4th	Avg
Single Computing (second)	46.14	28.08	24.8	27.61	31.65
Distirbuted Computing (second)	36.4	25.06	18.32	20.15	24.98

ตารางที่ 5. 3 แสดงผลการทำงานกระจายงานให้ทำงานกันคนละเครื่อง

5.2 การปรับปรุงเทคนิคในการแสดงผล

จากการทดลองการใช้งานระบบโดยรวมพบว่า การแสดงผลยังโดยใช้เอกสาร HTML อย่างเดียวทำให้ใช้งานได้ไม่สะดวกเท่าที่ควรอีกทั้งในปัจจุบันมีเทคโนโลยี XML ที่เป็นเทคโนโลยีในการระบุประเภทของข้อมูล (ซึ่งการที่ข้อมูลไม่ได้ถูกระบุประเภทเป็นปัญหาที่ทำให้เกิดการสกัดข้อมูลขึ้น) และยังสามารถทำให้การแสดงผลมีประสิทธิภาพดีขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.3 รายละเอียดการ Implement ระบบโดยรวม

รายละเอียดการ Implement ระบบโดยรวมประกอบด้วยโปรแกรม 2 ส่วนได้แก่

1. Daemon โปรแกรมในการตัดส่วนข้อมูล
2. CGI โปรแกรมในการให้บริการกับผู้ใช้งาน

โดยระบบนี้ใช้เครื่องในการตัดส่วนข้อมูล 4 เครื่องและใช้เครื่องทำหน้าที่ Web Server ซึ่งเป็นตัวกลางในการกระจายงาน 1 เครื่อง แต่สามารถเพิ่มและลดจำนวนเครื่องโดยการ Config ระบบได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 6
บทสรุปและข้อเสนอแนะ

6.1 บทสรุป

ในการศึกษาโครงการศึกษากรณีพิเศษนี้ เป็นการศึกษาการพัฒนาระบบการตัดส่วนข้อมูลสำหรับร้านหนังสือออนไลน์ สำหรับการเปรียบเทียบราคาหนังสือจากร้านหนังสือออนไลน์จำนวน 4 แห่งได้แก่

1. www.amazon.com
2. www.barnandnobel.com
3. www.1bookstreet.com
4. www.powells.com

โดยใช้ Information Extraction เป็นแกนในการตัดส่วนข้อมูล ทำให้ระบบมีความทนต่อความเปลี่ยนแปลงจากแหล่งข้อมูล

จากการศึกษาการทำงานของระบบที่ได้พัฒนาพบว่า เวลาที่ใช้ในการประมวลผลและส่งผลลัพธ์ออกมาให้ผู้ใช้เป็นไปด้วยความล่าช้า อีกทั้งเทคนิคในการแสดงผลในรูปแบบของเอกสาร HTML ธรรมดามีความยืดหยุ่นจำกัดทำให้ผู้ใช้ไม่สะดวกสบายในการใช้งาน จึงได้ทำการปรับปรุงระบบให้สามารถทำงานได้เร็วขึ้นโดยการใช้เทคนิคการประมวลผลแบบ Parallel และ การกระจายงานโดยแบ่งการทำงานออกเป็น 4 เครื่องทำให้ผลการตอบสนองดีขึ้นอย่างมาก และท้ายสุดได้ทำการเปลี่ยนรูปแบบการแสดงผลจากเอกสาร HTML มาเป็นการแสดงผลโดยใช้ XML ควบคู่ไปกับ XSL ทำให้การใช้งานสะดวกยิ่งขึ้น

6.2 ประโยชน์ที่ได้รับ

1. ทำให้ผู้ใช้งานสามารถเปรียบเทียบราคาหนังสือได้อย่างสะดวกรวดเร็ว
2. ระบบมีความยืดหยุ่นในการแสดงผลสูงทำให้ผู้ใช้สามารถเรียงข้อมูลตามต้องการได้อย่างง่ายดายโดยไม่ต้องขอข้อมูลจากเว็บเซิร์ฟเวอร์ใหม่
3. ระบบมีความทนทานต่อการเปลี่ยนแปลงทำให้เจ้าของระบบไม่ต้องเปลี่ยนแปลงแก้ไขโปรแกรม
4. เนื่องจากระบบแยกการแสดงผลกับข้อมูลออกจากกันโดยสมบูรณ์ทำให้การเปลี่ยนแปลงแก้ไขหน้าตาการแสดงผลไม่มีผลต่อโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

6.3 ข้อเสนอแนะ

- เนื่องจากการทำงานของระบบตัดส่วนข้อมูลต้องการใช้ CPU-Time สูงทำให้จำเป็นต้องทำการกระจายการทำงานซึ่งต้องใช้จำนวนเครื่องคอมพิวเตอร์ค่อนข้างสูงดังนั้นการนำไปใช้งานจริงอาจจะไม่สามารถนำไปใช้ได้จริง
- การนำเทคโนโลยี XML มาใช้ในการแสดงผลอาจจะทำให้ Browser ที่มี Version เก่าไม่สามารถทำงานได้
- การตัดส่วนข้อมูลอาจจะมีข้อผิดพลาดได้จึงควรมีระบบตรวจสอบคำตอบก่อนการแสดงผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

T. M. Mitchell.1997. Machine Learning: McGraw-Hill

Ellen Riloff. 1998. "*Automatically Generationg Extraction Patterns from Untagged Text,*" Department of Computer Science University of Utah Sate Lake City, UT 84112

Freitag, D. 1999. "*Multistrategy Learning for Information Extraction,*" School of Computer Science Carnegic Mollon University Pittsburgh PA 15213

Dayne Freitag "*Machine Learning for Information Extraction in Informal Domains*" Computer Science Department Carnegie Mellon University Pittsburgh, PA

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน	นายมนต์ชัย พจนาสมสมาน
วันเดือนปีเกิด	17 กุมภาพันธ์ พ.ศ. 2520
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	วิศวกรรมศาสตร์บัณฑิต
จากสถานศึกษา	มหาวิทยาลัยเกษตรศาสตร์
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2541

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้