บทที่ 1

บทนำ

1.1 ความเป็นมา

ปัจจุบันกล่าวได้ว่าเป็นยุคแห่งสารสนเทศ สังคมถูกเรียกได้ว่าเป็น "สังคมแห่งสารสนเทศ" สารสนเทศต่าง ๆ ที่มนุษย์สัมผัสอยู่มีหลายรูปแบบด้วยกัน เช่น ในรูปของตัวอักษรเขียน ในรูปของเสียง และในรูปของรูปภาพ

สารสนเทศในรูปของตัวอักษรหรือข้อความเป็นรูปแบบที่มีความสำคัญ เพราะว่ามนุษย์ใช้ตัวอักษรในการสื่อสารกันอย่างแพร่หลาย ในปัจจุบันสารสนเทศที่เป็นตัวอักษรได้ถูกผลิตออกมามากมาย เช่น หนังสือ วารสาร รายงานผลการวิจัย รายงานการประชุม เอกสารทางวิชาการ บทความ และเอกสารสิ่งพิมพ์ต่าง ๆ เป็นต้น เมื่อมีปริมาณสารสนเทศที่รวบรวม และเก็บรักษาไว้ในปริมาณที่มาก ๆ จึงเป็นการยากลำบากในการค้นหาข้อมูลที่ต้องการ ในระยะเวลาที่กำหนดด้วยเหตุนี้ ระบบการค้นหาโดยใช้เครื่องคอมพิวเตอร์ จึงได้ถูกนำมาใช้อย่างแพร่หลายมากยิ่งขึ้น การใช้เครื่องคอมพิวเตอร์มาดำเนินการกับข้อมูลสารสนเทศเหล่านี้มีข้อเด่นอยู่ 2 ประการ คือ

ก. มีความสะดวกในการรวบรวม การจัดทำสำเนา และการเก็บรักษา ในปริมาณของข้อมูลที่มาก ๆ ได้

ข. มีความสะดวกในการเตรียมวิธีการค้นหา และเรียกคืนสารสนเทศออกมาได้หลายรูปแบบ

ในกรณีการประมวลผลฐานข้อมูลที่เป็นข้อความ (text base) เมื่อพิจารณาเอกสารหนึ่ง ๆ สามารถแบ่งตัวบ่งชี้เอกสาร (Document identifiers) ออกได้เป็น 2 ส่วน [1] คือ

ก. Objective document identifiers เป็นตัวบ่งชี้ที่มองเห็นได้โดยง่าย เช่น ชื่อเรื่อง ชื่อผู้แต่ง ปีพิมพ์ จำนวนหน้า สำนักพิมพ์ เป็นต้น ซึ่งในส่วนนี้ระบบการจัดการฐานข้อมูลสามารถนำมาใช้เพื่อประมวลผล เก็บรักษา และค้นคืนข้อมูลที่ต้องการได้

ข. Nonobjective document identifiers เป็นตัวบ่งชี้ที่บอกถึงเนื้อความของเอกสารที่เกี่ยวข้องอยู่ ตัวบ่งชี้เนื้อความ (Content identifiers) มีชื่อเรียกหลายชื่อ เช่น คำหลัก (keyword) ตัวพรรณนา (descriptor) คำดรรชนี (index term) เป็นต้น ในเอกสารหนึ่ง ๆ อาจมีตัวบ่งชี้เนื้อความได้หลายตัว ด้วยเหตุนี้จึงเป็นการยาก ที่จะจับคู่อย่างสมบูรณ์ (complete match) ในระหว่างการสอบถาม (query) และเซ็ทของตัวบ่งชี้เอกสารเพื่อค้นคืนให้ได้เอกสารที่ตรงกับกรณีที่สุด

สำหรับวิธีการที่ใช้ในการค้นคืนสารสนเทศที่มีใช้กันอยู่ในปัจจุบันมี 2 ระบบใหญ่ ๆ [1] คือ

ก. ระบบการค้นคืนแบบสัญนิยม (Conventional Retrieval System)

เป็นระบบที่ใช้เทคโนโลยีพื้นฐานของแฟ้มข้อมูลผกผัน (Inverted files) ช่วยในการค้นคืนสารสนเทศ

ข. ระบบการค้นคืนชั้นสูง (Advanced Retrieval System)

เป็นระบบที่ใช้เทคโนโลยีพื้นฐานของแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) ในการค้นคืนสารสนเทศ

การวิจัยนี้จะศึกษาการค้นคืนสารสนเทศ โดยอาศัยแนวคิดแบบจำลองปริภูมิเวกเตอร์เป็นพื้นฐาน

1.2 วัตถุประสงค์

สร้างโปรแกรมเพื่อใช้ในการค้นคืนสารสนเทศ โดยอาศัยแนวคิดแบบจำลองปริภูมิเวกเตอร์เป็นพื้นฐาน

1.3 ขอบเขตของการวิจัย

1.3.1 ข้อมูลที่ใช้ในการทดสอบ จะใช้เอกสารรายงานผลการวิจัยจากฐานข้อมูลของสำนักงานคณะกรรมการวิจัยแห่งชาติ

1.3.2 ระบบที่พัฒนาขึ้นสามารถใช้ได้กับข้อความที่เป็นภาษาไทย

1.3.3 ในส่วนของเอกสารที่ใช้ภาษาไทย จะต้องมีการแทรกด้วยอักขระพิเศษ เพื่อแบ่งแยกคำมาให้แล้ว

1.3.4 ภาษาคอมพิวเตอร์ที่จะใช้ในการพัฒนาโปรแกรมจะใช้ภาษาซี และเครื่องคอมพิวเตอร์ที่จะใช้ เป็นเครื่องไมโครคอมพิวเตอร์

1.4 ขั้นตอนของการวิจัย

1.4.1 ศึกษาแนวคิดแบบจำลองปริภูมิเวกเตอร์ ที่ใช้ในการค้นคืนสารสนเทศ

1.4.2 ศึกษาและออกแบบโครงสร้างของข้อมูล โครงสร้างของแฟ้มข้อมูลที่จะใช้ในการประยุกต์ตามแนวคิดนี้

1.4.3 ออกแบบโครงสร้างของโปรแกรม

1.4.4 เขียนโปรแกรม และทดสอบโปรแกรม

1.4.5 สรุป ทดสอบ ประเมินผล และข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ทำให้ทราบปัญหาในการประยุกต์แนวคิดแบบจำลองปริภูมิเวกเตอร์ เพื่อสร้างโปรแกรมในการค้นคืนสารสนเทศ โดยเฉพาะกับข้อมูลภาษาไทย