การใช้จีนีติกอัลกอริทึมและยูสเซอร์โปรไฟล์เพื่อการสืบค้นสารสนเทศจาก WWW

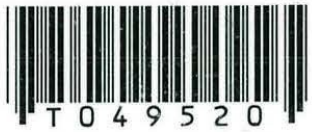
INTELLIGENT WEB SEARCH USING GENETIC ALGORITHMS AND USER'S PROFILE

ไพฑูรย์ ศรีนิล
PHAITOON SRINIL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2546
ISBN 974-324-518-9

การใช้จีนีติกอัลกอริทึมและยูสเซอร์โปรไฟล์เพื่อการสืบค้นสารสนเทศจาก WWW

INTELLIGENT WEB SEARCH USING GENETIC ALGORITHMS AND USER'S PROFILE



ไพฑูรย์ ศรีนิล

PHAITOON SRINIL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2546

ISBN 974-324-518-9

INTELLIGENT WEB SEARCH USING GENETIC ALGORITHMS AND USER'S PROFILE

PHAITOON SRINIL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN COMPUTER ENGINEERING

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2003

ISBN 974-324-518-9

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การใช้จีนีติกอัลกอริทึมและยูเซอร์โปรไฟล์เพื่อการสืบค้นสารสนเทศจาก WWW
INTELLIGENT WEB SEARCH USING GENETIC ALGORITHMS AND USER'S PROFILE

ชื่อนักศึกษา นายไพฑูรย์ ศรีนิล

รหัสประจำตัว 44061632

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมคอมพิวเตอร์

อาจารย์ผู้ควบคุมวิทยานิพนธ์ รศ.ดร.เอื้อน ปิ่นเงิน

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.บุญวัฒน์	อัตชู	
อาจารย์กฤตวัน	ศิริบูรณ์	
รศ.ดร.ศุภมิตร	จิตตะยโศธร	
รศ.ดร.เอื้อน	ปิ่นเงิน	

วัน/เดือน/ปี ที่สอบ 21 พฤษภาคม 2546 เวลา 14.00-16.00 น.

สถานที่สอบ ณ อาคาร 12 ชั้น ชั้น 4 (ห้อง E12-402)

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.บุญวัฒน์ อัตชู)

คณบดีบัณฑิตวิทยาลัย

วันที่ 30 **เดือน** พฤษภาคม **พ.ศ.** 2546

หัวข้อวิทยานิพนธ์	การใช้จีนีติกอัลกอริทึมและยูเซอร์โปรไฟล์เพื่อการสืบค้นสารสนเทศจาก WWW
นักศึกษา	นาย ไพฑูรย์ ศรีนิล
รหัสนักศึกษา	44061632
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขา	วิศวกรรมคอมพิวเตอร์
พ.ศ.	2546
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร. เอื้อน ปิ่นเงิน

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อปรับปรุงประสิทธิภาพของระบบสืบค้นสารสนเทศจาก WWW โดยใช้จีนีติกอัลกอริมและยูเซอร์โปรไฟล์ โดยสารสนเทศที่อยู่ในรูปแบบของ HTML ที่เหมือนกันหรือคล้ายคลึงกันและอยู่ในข่ายความสนใจของผู้ใช้จะถูกระบบสืบค้นดึงออกมารายงานต่อผู้ใช้ ระบบจะเข้ารหัสชุดคำค้นจากผู้ใช้และเทอมต่างๆในยูเซอร์โปรไฟล์ให้อยู่ในรูปโครโมโซมแล้วนำเข้ากระบวนการจีนีติกอัลกอริทึมจนได้ลักษณะที่ดีเหมาะสมแล้วจึงใช้เป็นชุดคิวรีสำหรับสืบค้นข้อมูลจาก WWW ผลรายงานที่ได้จากการสืบค้นจะนำมาจัดเรียงลำดับตามค่าความคล้ายที่มีต่อยูเซอร์โปรไฟล์

Thesis Title	INTELLIGENT WEB SEARCH USING GENETIC ALGORITHMS AND USER'S PROFILE
Student	Mr. Phaitoon Srinil
Student ID.	44061632
Degree	Master of Engineering
Programme	Computer Engineering
Year	2003
Thesis Advisor	Assoc. Prof. Dr. Ouen Pinngern

ABSTRACT

This research proposes the performance improvement of the information retrieval from World Wide Web by developing GA agent using Genetic Algorithms (GAs) and Intelligent user's profile. The documents, in HTML form, retrieved will have same topics, relative fields and interested by user that are interested by the agent. A set of keywords (query) and user's profile are converted to chromosome by GA encoding and using GA operators to improve the query. The GA agent uses this query as a keyword in the searching process. The results derived by this searching are ranked according to their degree of similarity to the user profile.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยพระคุณของพ่อ-แม่และพระคุณจากครู-อาจารย์ที่อบรมสั่งสอนให้ข้าพเจ้าเป็นคนดี มีความรู้ ขอบคุณสำหรับคำแนะนำของเพื่อนร่วมชั้นเรียน กำลังใจที่ดีเสมอจากคนรัก และวิทยานิพนธ์ได้รับแนวคิดและคำแนะนำจากท่าน รศ.ดร เอื้อน ปิ่นเงิน ซึ่งเป็นอาจารย์ควบคุมวิทยานิพนธ์จึงกราบขอบคุณ ณ. โอกาสนี้ด้วย

ไพฑูรย์ ศรีนิล

สารบัญ

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันอินเทอร์เน็ท(Internet) ได้เข้ามามีบทบาทในชีวิตประจำวันมากขึ้น ด้วยความสามารถของอินเทอร์เน็ททำให้การติดต่อสื่อสารข้อมูลต่างๆสามารถกระทำได้อย่างรวดเร็วและประหยัดค่าใช้จ่ายเป็นอย่างมาก การบริการข้อมูลความรู้ผ่านทางอินเทอร์เน็ทในรูป WWW (World Wide Web) ก็เป็นอีกบริการหนึ่งที่กำลังได้รับความนิยมเป็นอย่างมากในปัจจุบัน ข้อมูลใน WWW เป็นฐานข้อมูลแบบกระจายอยู่บนเครื่องเซิร์ฟเวอร์ต่างๆที่ให้บริการเว็บเซิร์ฟเวอร์ซึ่งมีกระจายอยู่ทั่วโลก ด้วยปริมาณข้อมูลที่ขนาดใหญ่และกระจายนี้เองทำให้ผู้ใช้หรือผู้ค้นหาข้อมูลต้องประสบปัญหาในการค้นหาข้อมูลที่ต้องการ เนื่องจากข้อมูลบน WWW มิได้ถูกจัดเก็บอย่างเป็นระบบ ไม่มีการจัดเรียงเป็นหมวดหมู่ ไม่มีเลขอ้างอิงอย่างเป็นระบบอย่างเช่นเอกสารในห้องสมุด ผู้ใช้อาจจะต้องใช้เวลานานมากในการค้นหาข้อมูลที่ต้องการว่าอยู่ที่ใด หรือบางครั้งอาจจะไม่เจอข้อมูลที่ต้องการเลยก็เป็นไปได้ ปัจจุบันการค้นหาข้อมูลบน WWW มีเครื่องมือค้นหา (Search engine) หรืออาจจะท่องไปตามกลุ่มข้อมูลที่ถูกจัดหมวดหมู่ไว้อย่างกว้างๆ เช่น Yahoo แต่บางครั้งก็ไม่สามารถค้นหาข้อมูลที่ต้องการได้เนื่องจากเครื่องมือช่วยค้นหาไม่มีประสิทธิภาพเพียงพอ และ/หรือ ข้อมูลที่เครื่องมือช่วยค้นหารายงานออกมา(เว็บเพจ) มีจำนวนมากมายเกินไปจึงยากต่อการตัดสินใจว่ารายงานหัวข้อใดเป็นหัวข้อที่ตรงกับความต้องการ

งานวิจัยนี้ได้นำเสนอการปรับปรุงประสิทธิภาพของเครื่องมือค้นหาข้อมูลบน WWW ให้มีความฉลาดในการค้นหาสารสนเทศที่มีเนื้อหาคล้ายคลึงกันหรืออยู่ในหัวข้อเดียวกันโดยอาศัยการวิเคราะห์ด้วยจีนีติกอัลกอริทึม ในกระบวนการจีนีติกอัลกอริทึม คำค้น (keyword) จากผู้ใช้จะถูกแทนด้วยโครโมโซมแล้วนำเข้าสู่กระบวนการจีนีติกโอเปอร์เรชัน อันได้แก่ การครอสโอเวอร์ มิว-เตชัน และการคัดเลือก เพื่อปรับปรุงคิวรี ระบบจะใช้ชุดคิวรีที่ปรับปรุงแล้วเป็นคีย์เวิร์ดป้อนให้เครื่องมือค้นหา ผลรายงานที่ได้จากเครื่องมือค้นหาจะถูกนำมาวิเคราะห์ค่าความคล้ายที่มีต่อยูเซอร์โปรไฟล์ ในขั้นตอนการวิเคราะห์ความคล้ายจะพิจารณาโครงสร้างของเอกสาร HTML ร่วมกับวิธีการวิเคราะห์แบบระบบ IR เดิม[6,7]

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

ข้อมูลบน WWW เป็นแหล่งความรู้ที่สำคัญ มีมายมาย และเป็นประโยชน์ต่อการศึกษาหาความรู้มาก แต่ในปัจจุบันยังไม่มีระบบที่มีประสิทธิภาพที่ดีเพียงพอในการดึงเอาข้อมูลต่างๆเหล่านั้นออกมาได้อย่างครบถ้วน ในงานวิจัยนี้มุ่งเน้นปรับปรุงประสิทธิภาพของเครื่องมือค้นหาให้

สามารถค้นหาเอกสารที่มีความคล้ายกับชุดคิวรีจากผู้ใช้และยูสเซอร์โปรไฟล์ หรือกล่าวอีกนัยหนึ่งว่าสามารถค้นหาเอกสารที่มีความคล้ายกับความต้องการของผู้ใช้ได้ โดยประยุกต์ใช้จีนีติกอัลกอริทึมซึ่งเป็นการเลียนแบบธรรมชาติและกระบวนการทางพันธุศาสตร์นำมาตัดสินว่าเอกสารใดบ้างที่มีความคล้ายกับความต้องการของผู้ใช้ องค์ประกอบที่สำคัญอีกส่วนหนึ่งของระบบคือยูสเซอร์โปรไฟล์เพราะเป็นส่วนที่ระบุความสนใจของผู้ใช้ ดังนั้นการออกแบบยูสเซอร์โปรไฟล์และการบริหารยูสเซอร์โปรไฟล์จึงเป็นขั้นตอนที่สำคัญมาก วิทยานิพนธ์นี้ได้นำเสนอการออกแบบและการบริหารยูสเซอร์โปรไฟล์ให้ระบบสามารถเรียนรู้ความสนใจและสามารถปรับปรุงตัวเองไปตามความสนใจของผู้ใช้ได้

งานวิจัย [4] ประยุกต์ใช้จีนีติกอัลกอริทึมเพื่อการค้นคืนเอกสารจากฐานข้อมูลออฟไลน์(off line) ซึ่งกระทำกับเอกสารธรรมดาเช่น เอกสารในห้องสมุด งานวิจัย [4] มีการปรับปรุงคิวรีก่อนจะส่งให้เครื่องมือค้นหา งานวิจัย [5] ประยุกต์ใช้จีนีติกอัลกอริทึมเพื่อการค้นคืนเอกสาร HTML บน WWW งานวิจัยนี้มีการใช้งานยูสเซอร์โปรไฟล์สำหรับใช้ปรับปรุงคิวรี มีการบริหารยูสเซอร์โปรไฟล์โดยใช้ข้อมูลป้อนกลับจากผู้ใช้ งานวิจัยนี้ใช้เวคเตอร์โมเดลแบบเดิม[6,7] โดยไม่มีการพิจารณาโครงสร้างของเอกสาร HTML

งานวิจัยนี้ได้ปรับปรุงจุดอ่อนของระบบดังกล่าวโดยประยุกต์ใช้จีนีติกอัลกอริทึม การใช้ยูสเซอร์โปรไฟล์และการวิเคราะห์ค่าความคล้ายด้วยเวคเตอร์โมเดลแบบพิจารณาโครงสร้าง HTML ด้วยวิธีการที่นำเสนอในวิทยานิพนธ์นี้ทำให้ค่า Precision ของระบบเพิ่มขึ้นเป็น 82.04 เปอร์เซ็นต์

1.3 แนวคิดที่ใช้ในงานวิจัย

จีนีติกอัลกอริทึมเป็นทฤษฎีที่จำลองกระบวนการทางธรรมชาติ คือการคัดเลือกทางธรรมชาติและอาศัยพื้นฐานความคิดทางพันธุกรรมในการถ่ายทอดลักษณะต่างๆ จากพ่อ-แม่ ไปยังลูกหลานที่สามารถนำมาพัฒนาใช้หาคำตอบที่ใกล้เคียง หรือดีที่สุดของปัญหา จีนีติกอัลกอริทึมเป็นวิธีค้นหาคำตอบโดยพิจารณาจากกลุ่มคำตอบของปัญหาที่สร้างขึ้น และใช้วิธีเข้ารหัส(encode) คือ การแปลงค่าตัวแปร หรือพารามิเตอร์ต่างๆของปัญหา ให้อยู่ในรูปโครงสร้างของโครโมโซมตามที่กำหนด เพื่อคัดเลือกโครโมโซมคำตอบที่เหมาะสมสำหรับสร้างวิวัฒนาการคำตอบให้ดีขึ้นตามกระบวนการทางพันธุศาสตร์โดยการแลกเปลี่ยนค่าพารามิเตอร์ต่างๆระหว่างโครโมโซมที่ถูกคัดเลือก จะทำให้คำตอบของปัญหาถูกปรับปรุงให้ดีขึ้น

งานวิจัยนี้ใช้จีนีติกอัลกอริทึมสำหรับปรับปรุงคิวรีเพื่อใช้เป็นอินพุทให้กับเครื่องมือค้นหาและใช้ยูสเซอร์โปรไฟล์สำหรับคัดเลือกเอกสารคำตอบรายงานต่อผู้ใช้ คิวรีจากผู้ใช้และชุด Indexed Term ที่ได้จากฐานข้อมูล(เทอมต่างๆ ถูกสุ่มเพื่อสร้างชุดคิวรีคำตอบ)จะถูกเข้ารหัสอยู่ในรูปโครโมโซม จากนั้นนำโครโมโซมดังกล่าวเข้าสู่กระบวนการจีนีติกอัลกอริทึมได้แก่ การครอส

โอเวอร์ การมิวเตชัน และการคัดเลือกจนได้โครโมโซมคำตอบ นำโครโมโซมคำตอบถอดรหัสก็จะได้ชุดคิวรีที่เหมาะสม

1.4 ขอบเขตของการวิจัย

งานวิจัยนี้นำเสนอการปรับปรุงประสิทธิภาพของเครื่องมือค้นหา ผลรายงานจากเครื่องมือค้นหาจะถูกนำมาวิเคราะห์ความ ความคล้ายกับความสนใจของผู้ใช้หรือยูสเซอร์โปรไฟล์ วิทยานิพนธ์เล่มนี้ได้สร้างระบบจำลอง(prototype) สำหรับทดสอบการทำงานของอัลกอริทึมที่ผู้เขียนได้นำเสนอ ระบบจำลองพัฒนาด้วยภาษาจาวาบนเครื่องไมโครคอมพิวเตอร์ AMD Duron 1.2 GHz ระบบปฏิบัติการ MS WindowsXP และ Linux RedHat 7.1

รายละเอียดของวิทยานิพนธ์ฉบับนี้แบ่งเป็นส่วนๆ คือ กล่าวถึงความรู้พื้นฐานเกี่ยวกับจีนีติกอัลกอริทึม การประยุกต์ใช้จีนีติกอัลริทึมในการสืบค้นสารสนเทศบน WWW ผลการทดสอบประสิทธิภาพของระบบ สรุปผลการทดลอง และแนวทางการปรับปรังพัฒนาให้มีประสิทธิภาพต่อไป

1.5 ขั้นตอนของการวิจัย

การดำเนินการสร้างระบบติดต่อกับเครื่องมือค้นหา(เช่น Yahoo) เริ่มต้นจากศึกษาการทำงานของเครื่องมือค้นหาที่มีใช้อยู่ในปัจจุบัน ศึกษาการทำงานของเอเจนท์หน้าร้านและข้อกำหนดในการร้องขอใช้บริการสำหรับแต่ละผู้ให้บริการ สร้างระบบปรับปรุงชุดคิวรีให้มีความเหมาะสมด้วยจีนีติกอัลกอริทึม สร้างยูสเซอร์โปรไฟล์เพื่อเก็บข้อมูลความสนใจของผู้ใช้แต่ละคน และระบบบริหารยูสเซอร์โปรไฟล์ จากนั้นทดสอบระบบตามขั้นตอนต่อไปนี้

1. รับข้อมูลจากผู้ใช้(ลงทะเบียน)เพื่อนำมาเป็นข้อมูลในการสร้างยูสเซอร์โปรไฟล์ การสร้างยูสเซอร์โปรไฟล์เป็นขั้นตอนสำคัญเพราะยูสเซอร์โปรไฟล์จะเป็นตัวบ่งบอกความสนใจของแต่ละผู้ใช้และในขั้นตอนการออกผลรายงานจะจัดลำดับของเอกสารโดยอ้างอิงจากความคล้ายกับยูสเซอร์โปรไฟล์
2. ป้อนคำค้นให้กับระบบเพื่อทดสอบการทำงาน
3. เมื่อได้ผลรายงานที่ส่งคืนมาจากระบบแล้ว ผู้ใช้ต้องแจ้งผลการรายงานในรูปข้อมูลป้อนกลับ(feedback)เพื่อเป็นข้อมูลแก่ระบบในการปรับปรุงยูสเซอร์โปรไฟล์
4. บันทึกค่า Precision ของผลรายงานสำหรับแต่ละรอบการทำงานของระบบ

1.6 ข้อจำกัดของระบบ

เนื่องจากงานวิจัยนี้ได้สร้างระบบจำลองการทำงานของเครื่องมือค้นหา ระบบจะส่งคำร้องขอไปยังเอเจนท์ของเว็บไซท์ที่ให้บริการใช้เครื่องมือค้นหา เช่น Yahoo ดังนั้นถ้ารูปแบบของคำร้อง

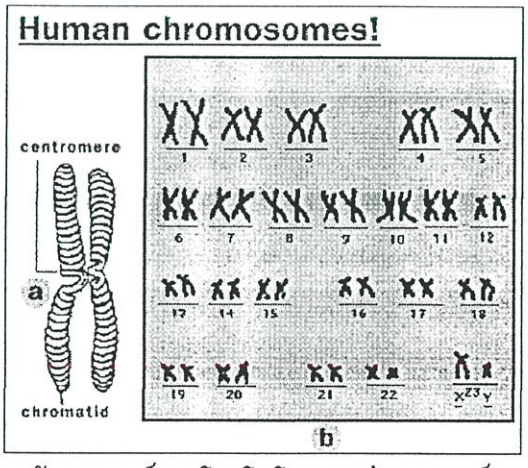
ขอในแต่ละเอเจนท์เปลี่ยนแปลงไปหรือพารามิเตอร์ที่จะต้องส่งแนบไปกับคำร้องขอนั้นไม่ถูกต้องแล้ว ระบบจำลองในวิทยานิพนธ์นี้จะติดต่อกับเอเจนท์ตัวนั้นไม่ได้หรือระบบจะไม่ได้รับผลรายงานจากเอเจนท์ตัวนั้น ในงานวิจัยนี้ได้สร้างฐานข้อมูลขนาดเล็กโดยใช้ MySQL เพื่อสำเนาเอกสารที่ได้จากการติดต่อกับเอเจนท์ของ Yahoo, Google, HotBot, Altavista และ Lycos เพื่อลดปัญหากรณีที่ติดต่อเอนเจนท์ดังกล่าวติดต่อไม่ได้และเพื่อความเร็วในการทำงานเพราะการติดต่อโดยตรงกับเอเจนท์เหล่านั้นต้องใช้เวลามาก ดังนั้นข้อมูลที่ได้จากการสืบค้นในวิทยานินพนธ์นี้อาจจะไม่ครอบคลุมปริมาณข้อมูลทั้งหมดที่มีบน WWW ด้วยสาเหตุนี้อาจจะมีผลกระทบต่อการประเมินประสิทธิภาพของระบบด้วยการวัดจากค่า Precision ได้

บทที่ 2
ความรู้พื้นฐานเกี่ยวกับจีนีติกอัลกอริทึม

ปัญหาที่ต้องการคำตอบที่ดีที่สุด(Optimal Solution) ในทางวิทยาศาสตร์สามารถหาคำตอบได้หลายวิธีซึ่งแตกต่างกันไปตามชนิดของปัญหา โดยมีการนำทฤษฎีที่เกี่ยวข้องกับการเลียนแบบทางธรรมชาติมาช่วยในการวิจัย เช่น การประมวลผลภาษาธรรมชาติ(Natural Language Processing) ระบบผู้เชี่ยวชาญ(Expert System) ฟัซซีลอจิก(Fuzzy Logic) นิวรอลเน็ตเวิร์ค(Neural Network) เป็นต้น จีนีติกอัลกอริทึม(Genetic Algorithm) เป็นอีกวิธีหนึ่งที่จำลองการทำงานทางชีววิทยาในการให้กำเนิดประชากรรุ่นใหม่ ซึ่งอาศัยพื้นฐานของการวิวัฒนาการทางพันธุกรรมในการถ่ายทอดลักษณะต่างๆไปยังรุ่นลูกหลาน โดยปฏิบัติตามหลักการพันธุศาสตร์ นำมาประยุกต์ใช้ในการแก้ปัญหาเพื่อหาคำตอบที่ดีที่สุดหรือใกล้เคียงที่สุด

2.1 พันธุศาสตร์ทางชีววิทยา

ยีนส์(Genes) เป็นหน่วยเก็บลักษณะทางกรรมพันธุ์ ซึ่งค้นพบโดยนักวิทยาศาสตร์ที่ชื่อ เมนเดล(Mendel)และเป็นตัวกำหนดลักษณะรูปร่างภายนอกของสิ่งมีชีวิต ซึ่งยีนส์จะเรียงตัวกันอยู่บนโครโมโซม(Chromosome) อีกที ในเซลของสิ่งมีชีวิตโครโมโซมจะจับกันเป็นคู่ๆแต่จะแตกต่างกันที่ค่าลักษณะต่างๆในแต่ละยีนส์เรียกว่า แอลลี(Allele) ซึ่งแบบต่างๆของยีนส์ที่มีแอลลีต่างกันในแต่ละตำแหน่งเรียกว่า ยีโนไทป์(Genotype) สำหรับลักษณะภายนอกที่ปรากฏออกมาให้เห็นเรียกว่า ฟีโนไทป์(Phenotype) ตัวอย่างคู่โครโมโซมทั้ง 23 คู่ของมนุษย์ ดังแสดงในรูปที่ 2.1 ซึ่งประกอบด้วยยีนส์ที่มีลักษณะสีผมและอื่นๆ

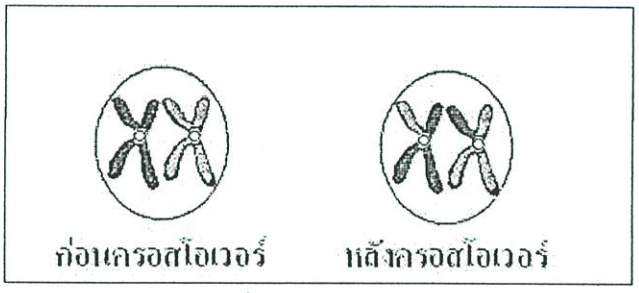


รูปที่ 2.1 แสดงลักษณะทางพันธุศาสตร์ของโครโมโซม 23 คู่ของมนุษย์

การถ่ายทอดลักษณะทางพันธุกรรมเป็นการถ่ายทอดลักษณะต่างๆของสิ่งมีชีวิตที่เกิดขึ้นเมื่อมีการแบ่งตัวของเซล ซึ่งมี 2 แบบคือ

1. การแบ่งตัวแบบไมโตสิส(Mitosis) เป็นการเพิ่มจำนวนเซล โดยโครโมโซมแต่ละตัวจะเพิ่มจำนวนตัวเองเป็นสอง และเยื่อหุ้มนิวเครียสจะสลายลงเพื่อแยกโครโมโซมที่เพิ่มจำนวนขึ้นออกจากโครโมโซมเดิมแล้วเยื่อหุ้มนิวเครียสจะถูกสร้างขึ้นใหม่เป็น 2 เซล
2. การแบ่งตัวแบบไมโอสิส(Meiosis) เป็นการแบ่งตัวของเซลสืบพันธุ์ โดยโครโมโซมจากเซลพ่อและแม่ อย่างละ 1 โครโมโซม จับคู่กันและต่างก็จำลองแบบของตนเพิ่มมาอีก ทำให้ได้โครโมโซมทั้งหมดเพิ่มขึ้นเป็น 2 เท่า โครโมโซมพ่อและแม่พร้อมทั้งแบบจำลองจะแยกคู่ไปรวมกันเป็น 2 นิวเครียสกลายเป็นเซลใหม่ 2 เซล ซึ่งแบ่งตัวต่อไป โดยโครโมโซมพ่อและแม่แยกตัวออกจากแบบจำลอง ทำให้ได้เซลใหม่ 4 เซล

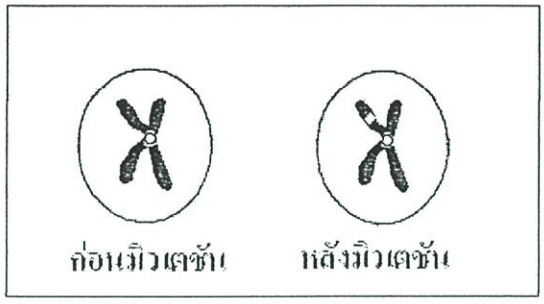
วิธีการแบบไมโอสิสโครโมโซมจะมีโอกาสแลกเปลี่ยนบางส่วนซึ่งกันและกัน เรียกว่าครอสโอเวอร์(Crossover) การครอสโอเวอร์จะเกิดขึ้นระหว่างโครโมโซมพ่อกับโครโมโซมแม่ เนื่องจากยีนส์แต่ละยีนส์ที่เรียงตัวกันบนโครโมโซมไม่ได้อยู่ห่างกันหนาแน่น และมีระยะห่างไม่สม่ำเสมอ ช่องว่างระหว่างยีนส์นี้เองที่สามารถแตกออกมาได้ขณะที่มีการครอสโอเวอร์ และมีการแลกเปลี่ยนยีนส์ของโครโมโซม โดยส่วนที่อยู่หลังรอยแตกทั้งหมดจะถูกย้ายไปอยู่อีกโครโมโซม นอกจากนี้ยังสามารถแตกอีกกี่แห่งก็ได้ ซึ่งขึ้นอยู่กับความสามารถที่จะเชื่อมกันได้มากน้อยเพียงไร ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 แสดงก่อนและหลังการครอสโอเวอร์ของโครโมโซม

ประโยชน์ที่เกิดจากการครอสโอเวอร์คือ ได้ลักษณะต่างๆมาอยู่รวมกันทำให้สิ่งมีชีวิตรุ่นลูกหลานมีความหลากหลายและอาจจะทำให้เกิดโอกาสเกิดสิ่งมีชีวิตที่มีลักษณะที่ดีรวมกันอยู่ได้อย่างพอเหมาะ ถ้าเซลเกิดใหม่โดยไม่มีการครอสโอเวอร์แล้ว โครโมโซมใดที่เคยมียีนส์ลักษณะใดก็จะมีลักษณะนั้นอยู่เรื่อยๆ โอกาสที่สิ่งมีชีวิตนั้นจะปรับตัวให้ดีขึ้นย่อมมีได้ยากกว่า นอกจากนี้ยังมีการผ่าเหล่า(Mutation) คือการเปลี่ยนแปลงของยีนส์ที่มีลักษณะต่างไปจากเดิมที่ควรจะเป็น ซึ่งเป็นต้นเหตุให้เกิดลักษณะที่แปลกออกไป เท่ากับเป็นโอกาสในการเลือกลักษณะแปลกๆมากขึ้น เนื่องจากกระบวนการวิวัฒนาการทางธรรมชาตินั้นช้ามาก เพราะกว่าที่ธรรมชาติจะปรับ

สภาพแวดล้อมให้สิ่งมีชีวิตค่อยๆปรับตัวเองให้เหมาะสมนั้นมีโอกาสน้อยมาก การผ่าเหล่านั้นทุกลักษณะในแต่ละยีนส์ย่อมมีโอกาสที่จะเกิดการเปลี่ยนแปลงไปจากเดิมได้พอๆกัน และถ้าเหมาะสมกับสภาพแวดล้อมในขณะนั้นก็จะคงอยู่ต่อไป แต่ถ้าการเปลี่ยนแปลงใดเกิดผิดจังหวะ คือไม่เหมาะสมกับสภาพแวดล้อมขณะนั้นๆก็จะไม่ถูกคัดเลือกและหายไปในที่สุด ดังรูปที่ 2.3



รูปที่ 2.3 แสดงการเกิดมิวเตชันของโครโมโซม

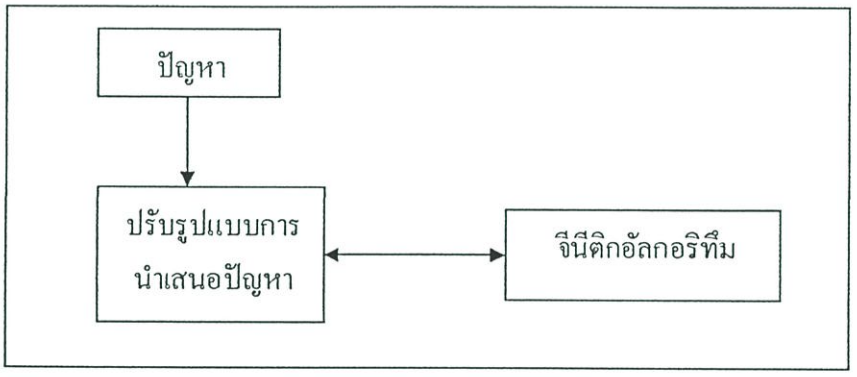
2.2 จีนีติกอัลกอริทึม

ปี ค.ศ. 1975 John Holland ได้ศึกษาเกี่ยวกับวิวัฒนาการทางธรรมชาติ(Natural Evolution) ในการให้กำเนิดประชากรสิ่งมีชีวิตในรุ่นต่อๆไปโดยกระบวนการทางชีววิทยาประกอบด้วยการคัดเลือกทางธรรมชาติ(Natural Selection) คือ สิ่งมีชีวิตที่แข็งแรงกว่าย่อมมีโอกาศอยู่รอดกว่าสิ่งมีชีวิตที่อ่อนแอ นั่นคือโครโมโซมที่ประกอบด้วยยีนส์ต่างๆที่มีลักษณะที่ดีสามารถอยู่รอดได้มากกว่า โครโมโซมที่อยู่รอดได้ก็จะถ่ายทอดยีนส์ที่มีลักษณะที่ดีเหล่านั้นไปยังลูกหลานได้มากกว่าเช่นกัน และกระบวนการทางพันธุกรรศาสตร์(Genetic Operation) คือการกำเนิดโครโมโซมใหม่จากการครอสโอเวอร์(Crossover)หรือมิวเตชัน(Mutation)

จากแนวคิดดังกล่าว Holland จึงได้นำปรับมาใช้กับการแก้ปัญหาด้วยคอมพิวเตอร์เพื่อหาคำตอบที่ดีที่สุดหรือใกล้เคียงที่สุดโดยมีจุดมุ่งหมายที่จะศึกษาระบบปรับปรุงประมวลผลเอง (Self Adaptive Process)และสร้างระบบเชี่ยวชาญ(Artificial System) โดยอาศัยแนวคิดของระบบการคัดเลือกทางธรรมชาติเรียกว่า จีนีติกอัลกอริทึม(Genetic Algorithms : GA) เพื่อปรับปรุงการหาคำตอบที่ดีขึ้น หลักการเบื้องต้นในการใช้จีนีติกอัลกอริทึมแก้ปัญหา คือต้องปรับปรุงรูปแบบปัญหาให้เหมาะสมกับการนำเสนอของจีนีติกอัลกอริทึม ดังรูปที่ 2.4

จีนีติกอัลกอริทึมเป็นวิธีการค้นหาคำตอบโดยการเลียนแบบการคัดเลือกทางธรรมชาติและธรรมชาติทางพันธุกรรมซึ่งอาศัยหลักการสุ่มเพื่อปรับปรุงความสามารถในการค้นหาคำตอบที่ดีขึ้น โดยมีวิธีการคือ

1. จีนีติกอัลกอริทึม ค้นหาคำตอบภายใต้โครงสร้างของปัญหา อันเกิดจากการกำหนดรหัส (Coding) รูปแบบแบบโครงสร้างจากกลุ่มตัวแปรต่างๆของปัญหานั้น ไม่ใช่ค้นหาคำตอบจากค่าของกลุ่มตัวแปรนั้น
2. จีนีติกอัลกอริทึม ค้นหาคำตอบโดยพิจารณาจากประชากรคำตอบ หรือ กลุ่มคำตอบ ไม่ใช่จากค่าของกลุ่มตัวแปร
3. จีนีติกอัลกอริทึม ค้นหาคำตอบจากผลลัพธ์ของกลุ่มค่าตัวแปรที่เป็นฟังก์ชันเป้าหมายของปัญหา
4. จีนีติกอัลกอริทึม ค้นหาคำตอบโดยอาศัยการถ่วงน้ำหนักความเหมาะสมของแต่ละคำตอบจากกลุ่มคำตอบนั้นๆ



รูปที่ 2.4 แสดงหลักการเบื้องต้นของจีนีติกอัลกอริทึม

2.3 ฟังก์ชันเป้าหมายกับฟังก์ชันความเหมาะสม

จีนีติกอัลกอริทึมจะพิจารณาคำตอบที่ผ่านมาว่าใกล้เคียงกับคำตอบที่ต้องการหรือไม่ จากฟังก์ชันเป้าหมาย(Objective Function: f) เนื่องจากปัญหาสามารถกำหนดฟังก์ชันเป้าหมาย ซึ่งแสดงความสัมพันธ์ของแต่ละตัวแปร พารามิเตอร์ เงื่อนไข หรือ ข้อกำหนดต่างๆ สำหรับฟังก์ชันความเหมาะสม (Fitness Function: F) เป็นฟังก์ชันที่ใช้เป็นตัวกำหนดค่าความเหมาะของแต่ละโครโมโซมว่ามีโอกาสถูกคัดเลือกมากน้อยเพียงใด โดยทั่วไปมักใช้ฟังก์ชันเป้าหมายเป็นฟังก์ชันความเหมาะสม หรืออาจใช้ฟังก์ชันเป้าหมายที่ถูกปรับให้เป็นฟังก์ชันความเหมาะสมก็ได้

2.4 รูปแบบการแทนโครโมโซม

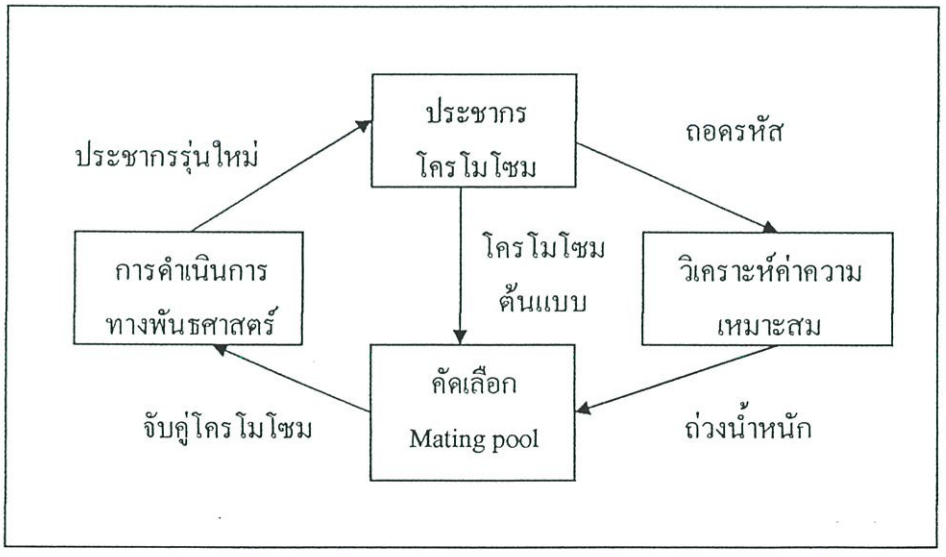
สำหรับจีนีติกอัลกอริทึม ยีนส์ซึ่งเป็นตัวแสดงคำตอบของปัญหาที่เปลี่ยนไปตามการประยุกต์ใช้งาน ซึ่งโดยทั่วไปยีนส์หมายถึงตัวแปร พารามิเตอร์ เงื่อนไขหรือข้อกำหนดต่างๆ ลำดับของยีนส์บนโครโมโซมจะอยู่ในรูปแบบสตริง(String) ประกอบด้วยบิท(bit) ซึ่งลักษณะที่เป็นไปได้

เรียกว่า ค่าของบิท(bit value) รูปแบบค่าบิทที่จัดเรียงบนโครโมโซมคือ ยีโนไทป์(genotype) ที่แสดงถึงค่าตัวแปร พารามิเตอร์ต่างๆที่เป็นไปได้ชุดหนึ่ง หรือฟีโนไทป์(phenotype)นั่นเอง

2.5 การทำงานของจีนีติกอัลกอริทึม

เมื่อกำหนดรูปแบบโครโมโซมและฟังก์ชันความเหมาะสมได้แล้วจีนีติกอัลกอริทึมจะสร้างวิวัฒนาการของกลุ่มคำตอบในรุ่นต่อๆไปโดยมีการทำงาน ดังรูปที่ 2.5 ซึ่งมี 4 ขั้นตอนคือ

1. สร้างประชากรรุ่นเก่าตามรูปแบบที่กำหนดไว้ โดยประชากรต้นกำเนิด(Initial Popular) เกิดจากการสร้างชุดโครโมโซมโดยการสุ่มค่าแต่ละบิท
2. วิเคราะห์ค่าความเหมาะสมของแต่ละโครโมโซมโดยถอดรหัสค่าตัวแปร พารามิเตอร์ต่างๆของแต่ละบิทและคำนวณค่าความเหมาะสม
3. สร้าง mating pool คือชุดโครโมโซมต้นแบบหรือโครโมโซม พ่อ-แม่ ที่สามารถอยู่รอดเป็นต้นแบบ โดยพิจารณาถ่วงน้ำหนักจากค่าความเหมาะสมของแต่ละโครโมโซม
4. ดำเนินการทางพันธุศาสตร์ โดยสุ่มจับคู่โครโมโซมต้นแบบใน mating pool เพื่อสร้างประชากรรุ่นใหม่ ซึ่งตัวดำเนินการทางพันธุศาสตร์ประกอบด้วยครอสโอเวอร์หรือมิวเตชัน



รูปที่ 2.5 การทำงานของจีนีติกอัลกอริทึม

การค้นหาคำตอบของจีนีติกอัลกอริทึมจะประมวลผลซ้ำๆจนกว่าได้รับคำตอบที่พอใจตามเงื่อนไขที่วางไว้ หรือในระยะเวลาตามจำนวนรุ่นที่ต้องการ ซึ่งแสดงอัลกอริทึมการทำงนของจีนีติกอัลกอริทึมได้ดังนี้

```
Simple Genetic Algorithm
Begin
     Initial population;
     Evaluate population;
     While termination criterion not reached
     Begin
          Select solution for next population;
          Crossover;
          Mutation;
          Evaluate population;
     End;
End.
```

สำหรับจีนีติกอัลกอริทึม ตัวแปรหรือพารามิเตอร์ของปัญหาจะถูกแปลงให้อยู่ในรูปของสตริง ซึ่งมักเรียกว่าโครโมโซม ประกอบด้วยอักขระหรือบิท แต่ละตำแหน่งของโครโมโซมประกอบด้วยค่าของอักขระหรือค่าของบิทที่แสดงถึงโครงสร้างของแต่ละโครโมโซมที่มีตัวแปรหรือพารามิเตอร์ของปัญหาที่แตกต่างกัน และเป็นตัวกำหนดค่าความเหมาะสมตามฟังก์ชันความเหมาะสมของแต่ละปัญหา ซึ่งในตารางที่ 2.1 อธิบายคำศัพท์เปรียบเทียบที่ใช้ทางพันธุศาสตร์กับจีนีติกอัลกอริทึม

Natural Genetic	Genetic Algorithm
Chromosome	String
Gene	Character, Bit
Allele	Character value, Bit value
Locus	String position
Genotype	Structure
Phenotype	Decode structure

ตารางที่ 2.1 แสดงคำศัพท์ที่ใช้ในทางพันธุศาสตร์ กับ จีนีติกอัลกอริทึม

จีนีติกอัลกอริทึมในยุคเริ่มแรกของ Holland นั้นคือ จีนีติกอัลกอริทึมแบบง่าย (Simple Genetic Algorithm: SGA) ซึ่งมีขั้นตอนพื้นฐานไม่มากนัก แบ่งออกเป็น 2 ส่วนคือการเตรียมการและการทำงาน

สำหรับในส่วนการเตรียมการนี้เป็นขั้นตอนการปรับรูปแบบของปัญหาให้เหมาะสมสำหรับการนำเสนอเพื่อใช้สำหรับแก้ปัญหานั้นๆประกอบด้วย

1. กำหนดฟังก์ชันความเหมาะสม เพื่อความสะดวกและง่ายต่อการเข้าใจ จะกำหนดตัวอย่างการหาคำตอบของปัญหาค่าสูงสุดของฟังก์ชัน $y = 2x^2$ ที่ x มีค่าระหว่างจำนวนเต็ม I[0,63]

<u>ตัวอย่าง</u> ฟังก์ชันเป้าหมาย คือ $f = 2x^2$

กำหนดฟังก์ชันความเหมาะสม คือ $F = 2x^2$

ซึ่งคำตอบที่ดีที่สุดคือค่า x ที่มีค่าความเหมาะสมสูงสุด Max(F)

2. กำหนดรูปแบบโครโมโซม โดยค่าตัวแปรหรือพารามิเตอร์ของปัญหาจะถูกแปลงให้อยู่ในรูปไบนารีโครโมโซมและมีความยาว(Chromosome Length : lchrom) ตามที่กำหนดเช่น

$$B_1 B_2 B_3 \ldots B_{lchrom} \text{ ซึ่ง } B_i \in \{0, 1\}$$

<u>ตัวอย่าง</u> วิธีแปลงค่าพารามิเตอร์ x ให้อยู่ในรูปไบนารี 6 บิท(lchrom = 6) ดังนั้น โครโมโซมของปัญหาจะมีค่าอยู่ระหว่าง 000000 ถึง 111111 ซึ่งเมื่อถอดรหัสแล้วจะมีค่าอยู่ช่วง 0 ถึง 63 สำหรับในส่วนการทำงานจะประกอบด้วย

1. ประชาการรุ่นเก่า(Old Population) เป็นชุดโครโมโซมที่ถูกคัดเลือกไปเป็นโครโมโซมต้นแบบสำหรับสร้างประชากรรุ่นใหม่(New Population) ในวิวัฒนาการรุ่น (Generation : gen) ต่อไป โดยประชากรเริ่มต้นที่ gen = 0 และจะถูกสร้างขึ้นโดยการสุ่มตามจำนวนโครโมโซมในแต่ละรุ่น(Population Size : popsize) ที่กำหนด

<u>ตัวอย่าง</u>

ลำดับ	โครโมโซม
1	101110
2	111001
3	101000
4	110011

ชุดโครโมโซมเริ่มต้นนี้เป็นชุดโครโมโซมที่กำหนดให้ในแต่ละรุ่น จะประกอบด้วย 4 โครโมโซม ซึ่งแต่ละโครโมโซมเกิดจากการสุ่มค่าไบนารีจำนวน 6 ครั้ง

2. การวิเคราะห์ค่าความเหมาะสม เป็นขั้นตอนการถอดรหัสจากโครงสร้างโครโมโซมที่กำหนดไว้ เพื่อคำนวณค่าความเหมาะสมตามฟังก์ชันความเหมาะสมของปัญหา ในที่นี้ฟังก์ชันความเหมาะสมคือ $F = 2x^2$

ตัวอย่าง	ลำดับ	โครโมโซม	x	ค่าความเหมาะสม(F)
	1	101110	46	4232
	2	111001	57	6498
	3	101000	40	3200
	4	110011	51	5202

2.5.1 การคัดเลือก

เป็นขั้นตอนการคัดเลือกโครโมโซมต้นแบบเพื่อสร้าง mating pool โดยโครโมโซมรุ่นเก่าเป็นโครโมโซมต้นแบบหรือโครโมโซมพ่อ-แม่ เพื่อใช้ในการสร้างโครโมโซมรุ่นลูก-หลานต่อไป โครโมโซม ที่มีค่าความเหมาะสมที่ดีจะถูกกำหนดน้ำหนักค่าความน่าจะเป็นที่จะถูกคัดเลือกสูง การกำหนดค่าความน่าจะเป็นที่จะถูกคัดเลือกแต่ละครั้ง(Probability of Selected Value : pselect) ของแต่ละโครโมโซมโดยกำหนดจากค่าความเหมาะสมเทียบกับผลรวมของค่าความเหมาะสมทั้งหมด ดังสมการที่ 2.1

$$pselect = F_i / \sum F \tag{2.1}$$

ซึ่งสามารถคำนวณค่าที่จะสุ่มได้(Expected Value: E) ของแต่ละโครโมโซมในแต่ละรุ่นดังสมการที่ 2.2

$$E_i = pselect_i * popsize = F_i / \overline{F} \tag{2.2}$$

สำหรับวิธีการสุ่มโครโมโซมต้นแบบ เป็นการจำลองการหมุนวงล้อถ่วงน้ำหนัก(Rolette Wheel: RW) ซึ่งกำหนดขนาดแต่ละช่องของวงล้อตามความน่าจะเป็นที่สุ่มได้ในแต่ละครั้งของโครโมโซม ซึ่งมีวิธีการดังนี้

1. หาค่าความเหมาะสมของแต่ละโครโมโซม
2. หาค่าความน่าจะเป็นที่สุ่มได้ในแต่ละครั้งของแต่ละโครโมโซม
3. หาความถี่สะสม(q) ของค่าความน่าจะเป็นของแต่ละโครโมโซมดังสมการที่ 2.3

$$q_i = \sum_{j=1}^{i} pselect_j \quad (2.3)$$

4. สร้างเลขสุ่มจำนวนจริง(r) ที่มีค่าอยู่ในช่วง[0.0 , 1.0]
5. เลือกโครโมโซมลำดับที่ r ซึ่งมีค่าอยู่ระหว่าง q_{i-1} และ q_i

ตัวอย่าง

ลำดับ	โครโมโซม	x	ค่าความเหมาะสม F	ค่าความน่าจะเป็น $pselect_j$	จำนวนที่คาดหวัง E_i	จำนวนที่สุ่มได้จาก RW
1	101110	46	4232	0.221	0.885	1
2	101101	57	6498	0.340	1.359	2
3	101000	40	3200	0.167	0.669	0
4	110011	51	5202	0.272	1.088	1
	รวม		19132	1.000	4.000	
	ค่าเฉลี่ย		4783	0.25	1.000	
	ค่าสูงสุด		6498	0.340	1.359	

ตารางที่ 2.2 การคัดเลือกด้วยวิธีหมุนวงล้อ

ตัวอย่าง การกำหนดค่าความน่าจะเป็น โดยกำหนดค่าความเหมาะสมเทียบกับผลรวมของค่าความเหมาะสมทั้งหมด จะเห็นได้ว่าในการคัดเลือกโครโมโซมต้นแบบจาก 4 โครโมโซมนี้โอกาสที่จะสุ่มได้โครโมโซมลำดับที่ 1, 2, 3, 4 ต่อการสุ่มแต่ละครั้งเท่ากับ 0.221, 0.340, 0.167, และ 0.272 ตามลำดับ และจำนวนโครโมโซมต้นแบบที่สุ่มได้จากการหมุนวงล้อมีดังนี้

ลำดับโครโมโซม	1	2	3	4
ค่าความเหมาะสม(F)	4232	6498	3200	5202
ค่าความน่าจะเป็นที่สุ่มได้แต่ละครั้ง($pselect_i$)	0.221	0.340	0.167	0.272
ความถี่สะสมค่าความน่าจะเป็น(q_i)	0.221	0.561	0.728	1.000
เลขสุ่มจากการหมุนวงล้อแต่ละครั้ง(r)	0.333	0.844	0.456	0.128
ลำดับโครโมโซมที่ถูกเลือก($q_{i-1} \le r \le q_i$)	2	4	2	1

ตารางที่ 2.3 โครโมโซมต้นแบบที่สุ่มได้จากการหมุนวงล้อ

ซึ่งจำนวนที่สุ่มได้เป็นโครโมโซมต้นแบบใน mating pool ของแต่ละโครโมโซมเป็น 1, 2, 0, และ 1 ตามลำดับ จะเห็นว่าโครโมโซมลำดับที่ 2 มีค่าความเหมาะสมสูงสุดจะมีโอกาสถูกเลือกมากที่สุด

2.5.2 การดำเนินการทางพันธุศาสตร์

การดำเนินการทางพันธุศาสตร์เป็นขั้นตอนการจำลองแบบทางพันธุกรรม ซึ่งมีตัวดำเนินการทางพันธุศาสตร์คือ *การครอสโอเวอร์* และ *การมิวเตชัน* โดยมีรายละเอียดดังนี้

การครอสโอเวอร์: เป็นตัวดำเนินการในการแลกเปลี่ยนส่วนของโครโมโซมพ่อ-แม่ ตามอัตราความน่าจะเป็นของการครอสโอเวอร์(Probability of Crossover : P_c) เพื่อสร้างชุดโครโมโซมรุ่นใหม่ มีขั้นตอนดังนี้คือ

ขั้นตอนที่ 1: จับคู่โครโมโซมพ่อ-แม่ ใน mating pool ที่สร้างไว้จากการคัดเลือก

ขั้นตอนที่ 2: สร้างเลขสุ่มจำนวนจริง(r) ที่มีค่าอยู่ในช่วง[0.0 , 1.0] โดยถ้า $r \leq P_c$ แล้วโครโมโซมพ่อ-แม่ นั้นจึงมีการครอสโอเวอร์

ขั้นตอนที่ 3: ครอสโอเวอร์โดยแลกเปลี่ยนส่วนคู่ของโครโมโซมพ่อ-แม่นั้น ซึ่งการครอสโอเวอร์ของจีนีติกอัลกอริทึมแบบง่ายเป็นแบบ 1 จุด(One-point Crossover) ดังนี้

คู่โครโมโซมพ่อ-แม่	$d_1\ d_2\ d_3 \ldots d_{pos}\ d_{pos+1} \ldots d_{lchrom}$
	$m_1\ m_2\ m_3 \ldots m_{pos}\ m_{pos+1} \ldots m_{lchrom}$

ทำการครอสโอเวอร์โดยสุ่มเลือกตำแหน่ง pos เป็นตำแหน่งที่จะครอสโอเวอร์ ซึ่งมีค่าอยู่ในช่วง[1 , lchrom] และแลกเปลี่ยนค่าในแต่ละบิทของคู่โครโมโซมพ่อ-แม่ ตั้งแต่ตำแหน่งที่ pos+1 ถึง lchrom ซึ่งจะทำให้เกิดโครโมโซมลูก 2 โครโมโซมคือ

คู่โครโมโซมลูก	$d_1\ d_2\ d_3 \ldots d_{pos}\ m_{pos+1} \ldots m_{lchrom}$
	$m_1\ m_2\ m_3 \ldots m_{pos}\ d_{pos+1} \ldots d_{lchrom}$

จำนวนการครอสโอเวอร์ในแต่ละรุ่นขึ้นอยู่กับค่า P_c ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรในแต่ละรุ่น popsize เท่ากับ 40 โครโมโซมและกำหนดให้ $P_c = 0.6$ แล้วจำนวนการครอสโอเวอร์ในแต่ละรุ่นเท่ากับ P_c*(popsize/2) = 0.6*(40/2) = 12 ครั้ง(การครอสโอเวอร์ 1 ครั้งเกิดจาก 2 โครโมโซม)

ตัวอย่าง กำหนด $P_c = 0.5$ โครโมโซมพ่อ-แม่ ใน mating pool จากการครอสโอเวอร์ดังนี้

ลำดับ	mating pool	สุ่มจับคู่พ่อ-แม่	เลขสุ่ม (r)	ก่อนครอสโอเวอร์	ตำแหน่ง pos	หลังครอสโอเวอร์	x	F	ลำดับโครโมโซมลูก
2	111001	1,2	0.321	**101|110**	3	**101**|001	41	3362	1
4	110011		≤0.5	111|001		111|**110**	62	7688	2
2	111001	2,4	0.654	ไม่ครอสโอเวอร์		111001	57	6498	3
1	101110		≥0.5			110011	51	5202	4
	รวม							22750	
	ค่าเฉลี่ย							5688	
	ค่าสูงสุด							7688	

ตารางที่ 2.4 การครอสโอเวอร์โดยใช้ค่าความน่าจะเป็นในการครอสโอเวอร์ $P_c = 0.5$

จากการจับคู่ใน mating pool ได้โครโมโซมลำดับที่ 1 คู่กับโครโมโซมลำดับที่ 2 และลำดับที่ 1 คู่กับลำดับที่ 4 แต่เฉพาะโครโมโซม คู่แรกที่มีการครอสโอเวอร์เพราะค่า $r \leq 0.5$ โดยตำแหน่งที่มีการครอสโอเวอร์คือ pos = 3 หลังจากการครอสโอเวอร์มีค่าความเหมาะสมดีกว่าโครโมโซมพ่อ-แม่ทั้งหมด คือ 7688 ซึ่งแสดงให้เห็นว่าการจำลองกระบวนการครอสโอเวอร์ตามธรรมชาติช่วยสร้างคำตอบที่ดีขึ้นได้

การมิวเตชัน: เป็นการดำเนินการผ่าเหล่าตัวหนึ่งที่ให้โครโมโซมมีค่าความเหมาะสมดีขึ้นหลังการครอสโอเวอร์โดยกลับค่าบิทที่สุ่มได้ตามอัตราความน่าจะเป็นของการมิวเตชัน(Probability of Mutation : P_m) ที่กำหนดเช่น

ก่อนการมิวเตชัน	$c_1\ c_2\ c_3\ \ldots\ c_{pos}\ \ldots\ c_{lchrom}$

สำหรับการมิวเตชันของ SGA เป็นแบบไบนารีมิวเตชัน คือการกลับค่าคอมพลีเมนต์จาก 0 เป็น 1 ดังนั้นจะได้โครโมโซมใหม่คือ

หลังการมิวเตชัน	$c_1\ c_2\ c_3\ \ldots\ c'_{pos}\ \ldots\ c_{lchrom}$

จำนวนการมิวเตชันในแต่ละรุ่นขึ้นอยู่กับค่า P_m ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรในแต่ละรุ่น popsize เท่ากับ 40 โครโมโซม แต่ละโครโมโซมประกอบด้วย 6 บิทและกำหนดให้ P_m= 0.02 แล้วจำนวนการมิวเตชันเท่ากับ P_m*popsize*lchrom = 0.02*40*6 = 5 บิท

ตัวอย่าง กำหนด P_m=0.1 การดำเนินการมิวเตชันโครโมโซมลูกที่ได้จากการครอสโอเวอร์ดังนี้

ลำดับ	ก่อนมิวเตชัน	เลขสุ่ม	หลังมิวเตชัน	x	F
1	101001	0.896, 0.254, 0.753, **0.062**, 0.351, 0.684	101101	45	4050
2	111110	0.984, 0.421, 0.564, 0.241, 0.958,0.547	111110	62	7688
3	111001	0.552, 0.637, 0.258, 0.491, 0.746, 0.029	111000	56	6272
4	110011	0.951, 0.874, 0.464, 0.829, 0.648, 0.214	110011	51	5202
	รวม				23212
	ค่าเฉลี่ย				5803
	ค่าสูงสุด				7688

ตารางที่ 2.5 การมิวเตชันโดยใช้ค่าความน่าจะเป็นในการมิวเตชัน P_m=0.1

จากการสุ่มตำแหน่งที่จะมีการมิวเตชัน โดยสร้างเลขสุ่ม r ของแต่ละบิทโครโมโซมแล้ว ตำแหน่งที่ 3 ของโครโมโซมที่ 1 และตำแหน่งที่ 4 ของโครโมโซมที่ 3 เป็นตำแหน่งที่ค่า $r \leq 0.1$ ตามอัตราการมิวเตชัน ทำให้ค่าความเหมาะสมจาก 3362 และ 6498 เป็น 4050 และ 6272 ตามลำดับ จะเห็นได้ว่าการมิวเตชันเป็นตัวดำเนินการที่ทำให้เกิดค่าความเหมาะสมสูงขึ้นหรือต่ำลงได้ แต่อย่างไรก็ตามตามค่าเฉลี่ยของความเหมาะสมดีขึ้นจาก 5688 เป็น 5803 แสดงถึงการหาคำตอบของจีนีติกอัลกอริทึมแบบง่ายดีขึ้น

2.5.3 ประชากรรุ่นใหม่

เป็นชุดโครโมโซมลูกที่เกิดจากการวิวัฒนาการต่างๆทั้งหมด ซึ่งประชากรรุ่นใหม่ทั้งหมดที่เกิดขึ้นจะถูกถ่ายทอดกลายเป็นประชากรรุ่นเก่าสำหรับวิวัฒนาการในรุ่นต่อไป ซึ่งเรียกวิวัฒนาการ

แบบนี้ว่า การรีโพรดักชัน(Reproduction) กระบวนการต่างๆจะถูกปฏิบัติซ้ำๆจนกระทั่งถึงรุ่นที่มากที่สุด(Max generation : maxgen) ที่ต้องการ

จะเห็นได้ว่าจีนีติกอัลกอริทึมเป็นวิวัฒนาการทางธรรมชาติที่นำมาประยุกต์ใช้กับปัญหาต่างๆ ซึ่งมีการทำงานเบื้องต้นเป็นจีนีติกอัลกอริทึมแบบง่าย โดยมีโครโมโซมเป็นแบบไบนารีและมีตัวดำเนินการทางพันธุศาสตร์ คือ การครอสโอเวอร์ และ การมิวเตชัน จะเห็นว่าค่าความเหมาะสมของจีนีติกอัลกอริทึมแบบง่าย มีค่าสูงขึ้นและลดลงได้ด้วยวิธีการสุ่ม เพื่อช่วยให้จีนีติกอัลกอริทึมสามารถหาคำตอบที่ดีขึ้นได้โดยปรับปรุงการทำงานของจีนีติกอัลกอริทึมดังนี้

1. รีโพรดักชันแบบรักษาค่าความเหมาะสมที่ดี เนื่องจากในการค้นหาคำตอบของจีนีติกอัลกอริทึมแบบง่ายนั้นมีโอกาสที่จะสูญเสียโครโมโซมรุ่นเก่าที่มีค่าความเหมาะสมที่ดีไปได้ ซึ่งจะทำให้คำตอบในรุ่นถัดไปมีค่าความเหมาะสมดีมากขึ้นหรือน้อยลงได้ ดังนั้นหากปรับปรุงจีนีติกอัลกอริทึมแบบง่ายให้ควบคุมการหาคำตอบโดยรักษาโครโมโซมที่ดีไว้แล้วจะช่วยให้วิวัฒนาการในการหาคำตอบในรุ่นถัดไปดีขึ้นเรื่อยๆโดยวิธีการดังนี้

   - กำหนดจำนวนโครโมโซมที่ดีที่สุด(#best) ของรุ่นเก่าที่ต้องการรักษาเป็น 1, 2, 4, ...
   - ถ้าจำนวนโครโมโซมที่กำหนดเป็น 1 ให้สร้างชุดโครโมโซมรุ่นใหม่ทั้งหมด แล้วจึงคัดลอก(copy) โครโมโซมที่ดีที่สุดของรุ่นเก่ามากแทนที่โครโมโซมรุ่นใหม่ที่มีค่าความเหมาะสมน้อยที่สุด
   - ถ้าจำนวนโครโมโซมที่กำหนดเป็น 2, 4, ... ให้คัดลอกโครโมโซมที่ดีที่สุดในรุ่นเก่าตามจำนวนที่กำหนดมาเป็นโครโมโซมรุ่นใหม่ แล้วจึงสร้างโครโมโซมรุ่นใหม่ส่วนที่เหลือต่อไป

2. การครอสโอเวอร์แบบ 2 จุด การแลกเปลี่ยนส่วนของโครโมโซมพ่อ-แม่ บางครั้งหากแลกเปลี่ยนเพียงบางช่วงของโครโมโซมแล้วจะสร้างโครโมโซมที่ดีกว่าเช่น การหาค่าสูงสุดของฟังก์ชัน $y = 2x^2$ ของคู่โครโมโซมพ่อ-แม่ 101110 และ 111001 ซึ่งมีค่าความเหมาะสมเป็น 4232 และ 6498 แลกเปลี่ยนค่าบิทตำแหน่งที่ 4 และ 5 เท่านั้น จำทำให้เกิดโครโมโซมลูกคือ 101000 และ 111111 มีค่าความเหมาะสมเป็น 3200 และ 7938 ซึ่งโครโมโซม 111111 เป็นโครโมโซมที่ให้คำตอบที่ดีที่สุดที่ต้องการ ดังนั้นการพัฒนาตัวดำเนินการครอสโอเวอร์เป็นแบบ 2 จุด(Two-point Crossover) จะทำให้จีนีติกอัลกอริทึมแบบง่ายค้นหาคำตอบที่ดีขึ้นได้ ดังรูปที่ 2.6 มีวิธีการดังนี้

- สุ่มเลือกตำแหน่ง pos_1, pos_2 คือตำแหน่งเริ่มต้นและตำแหน่งสุดท้ายที่จะครอสโอเวอร์ตามลำดับ ซึ่ง pos_1 มีค่าอยู่ระหว่าง [1 , lchrom-1] และ pos_2 มีค่าอยู่ระหว่าง [1 , lchrom] โดยที่ pos_1 มีค่าน้อยกว่า pos_2
- แลกเปลี่ยนค่าในแต่ละบิทของคู่โครโมโซมพ่อ-แม่ ตั้งแต่ตำแหน่งที่ pos_1+1 ถึง pos_2

คู่โครโมโซม $d_1\ d_2\ d_3 \dots d_{pos1}\ d_{pos1+1} \dots d_{pos2}\ d_{pos2+1} \dots d_{lchrom}$

$m_1\ m_2\ m_3 \dots m_{pos1}\ m_{pos1+1} \dots m_{pos2}\ m_{pos2+1} \dots m_{lchrom}$

คู่โครโมโซมลูก $d_1\ d_2\ d_3 \dots d_{pos1}\ m_{pos1+1} \dots m_{pos2}\ d_{pos2+1} \dots d_{lchrom}$

$m_1\ m_2\ m_3 \dots m_{pos1}\ d_{pos1+1} \dots d_{pos2}\ m_{pos2+1} \dots m_{lchrom}$

(pos_1 , pos_2 คือตำแหน่งที่สุ่มได้ซึ่ง $pos_1 \in [1, lchrom-1]$, $pos_2 \in [1, lchrom]$ และ $pos_1 \neq pos_2$)

รูปที่ 2.6 การครอสโอเวอร์แบบ 2 จุด

3. ไบนารีมิวเตชันแบบกำหนดค่าบิท เนื่องจากการหาคำตอบของจีนีติกอัลกอริทึมแบบง่ายนั้น กระบวนการไบนารีมิวเตชันอาจทำให้โครโมโซมที่เปลี่ยนแปลงไปให้คำตอบที่ดีลดลงและทำให้สูญเสียโครโมโซมที่ดีไป เช่น โครโมโซม 111110 มีค่าความเหมาะสมเป็น 7688 หากสุ่มได้ตำแหน่งบิทที่ 1 เกิดมิวเตชันแล้ว โครโมโซมที่เกิดขึ้นจากการมิวเตชันคือ 011110 ทำให้มีค่าความเหมาะสมลดลงเป็น 1800 แต่บางครั้งข้อดีหรือจุดเด่นของปัญหาจะสามารถนำมาปรับให้เข้ากับการค้นหาคำตอบที่ดีขึ้นได้ สำหรับการหาค่าสูงสุดของฟังก์ชัน $y = 2x^2$ นี้ค่าบิทของโครโมโซมที่เป็น 1 จะทำให้ค่าความเหมาะสมสูงขึ้นเสมอ ดังนั้นหากปรับปรุงไบนารีมิวเตชันให้เป็นแบบกำหนดค่าแน่นอนให้กับบิทที่เกิดมิวเตชันโดยกำหนดให้บิทที่เกิดมิวเตชันมีค่าบิทเป็น 1 เสมอจะช่วยปรับแนวทางการค้นหาคำตอบของจีนีติกอัลกอริทึมแบบง่ายได้ดีขึ้น เช่น หากประยุกต์ไบนารีมิวเตชันที่กำหนดค่าบิทให้เป็น 1 เสมอกับโครโมโซม 111110 ในตำแหน่งที่ 1 แล้วโครโมโซมที่เกิดขึ้นจะเหมือนเดิมและยังเป็นการรักษาโครโมโซมที่มีค่าความเหมาะสมที่ดีไว้ด้วย

4. อินเวอร์ชัน(Inversion) เป็นตัวดำเนินการที่ประยุกต์เพิ่มเติมในจีนีติกอัลกอริทึมแบบง่าย โดยจำลองแบบลักษณะของการอินเวอร์ชันในทางพันธุศาสตร์ที่เป็นลักษณะของการกลับหัวกลับหางส่วนของยีนส์ภายในโครโมโซมที่อาจช่วยให้เกิดโครโมโซมที่ดีขึ้นได้ โดยการกลับส่วนค่า

บิทภายในช่วงตำแหน่งของโครโมโซมที่สุ่มได้ตามอัตรค่าความน่าจะเป็นของการอินเวอร์ชันแต่ละโครโมโซม(Probability of Inversion : P_i) ที่กำหนดดังรูปที่ 2.7 มีขั้นตอนดังนี้

- สุ่มเลือกตำแหน่ง pos_1, pos_2 คือตำแหน่งเริ่มต้นและตำแหน่งสุดท้ายที่จะอินเวอร์ชันตามลำดับซึ่ง pos_1 และ pos_2 ค่าอยู่ในช่วง [1, lchrom] โดยที่ pos_1 มีค่าน้อยกว่า pos_2
- กลับค่าบิทในช่วงของตำแหน่งที่ pos_1 ถึง pos_2 ของโครโมโซม โดยสลับค่าบิท pos_1 กับ pos_2 , pos_2+1 กับ pos_2-1, pos_1+2 กับ pos_2-2, ...

ก่อนอินเวอร์ชัน	$c_1\ c_2\ c_3 \cdots c_{pos1}\ c_{pos1+1} \cdots c_{pos2-1}\ c_{pos2} \cdots c_{lchrom}$
หลังอินเวอร์ชัน	$c_1\ c_2\ c_3 \cdots c_{pos2}\ c_{pos2-1} \cdots c_{pos1+1}\ c_{pos1} \cdots c_{lchrom}$
(pos_1 , pos_2 คือตำแหน่งที่สุ่มได้ซึ่ง pos_1 , $pos_2 \in [1 , lchrom]$ และ $pos_1 \neq pos_2$)	

รูปที่ 2.7 การอินเวอร์ชัน

ตัวอย่างเช่น สุ่มโครโมโซมที่จะอินเวอร์ชันคือ 101010 มีค่าความเหมาะสมเป็น 3528 โดยสุ่มตำแหน่ง pos_1 เท่ากับ 2 และ pos_2 เท่ากับ 5 แล้ว จะเห็นว่าการอินเวอร์ชันทำให้เกิดโครโมโซม **110100** ซึ่งมีค่าความเหมาะสมดีขึ้นเป็น 5408 เป็นต้น สำหรับจำนวนการอินเวอร์ชันในแต่ละรุ่นขึ้นอยู่กับการกำหนดค่า P_i ซึ่งแตกต่างกันในแต่ละปัญหา เช่น ถ้าจำนวนประชากรในแต่ละรุ่น popsize เท่ากับ 40 โครโมโซมและกำหนดให้ P_i=0.1 แล้ว จำนวนการอินเวอร์ในแต่ละรุ่นเท่ากับ P_i*popsize = 0.1*40 = 4 ครั้ง

จะเห็นว่าจีนีติกอัลกอริทึม เป็นการเลียนแบบการวิวัฒนาการทางธรรมชาติที่สามารถนำมาประยุกต์ใช้กับคอมพิวเตอร์เพื่อช่วยแก้ปัญหาในการหาคำตอบต่างๆ ซึ่งพื้นฐานการทำงานเบื้องต้นเป็นจีนีติกอัลกอริทึมแบบง่าย โดยมีรูปแบบโครโมโซมเป็นไบนารีและตัวดำเนินการทางพันธุศาสตร์ที่สำคัญคือ การครอสโอเวอร์และการมิวเตชันที่ไม่ซับซ้อนแต่สามารถปรับปรุงให้เข้ากับปัญหาเพื่อช่วยให้จีนีติกอัลกอริทึมมีการค้นหาคำตอบที่ดีขึ้น

บทที่ 3
โมเดลที่ใช้ในระบบค้นคืนสารสนเทศ

ในงานวิจัยเกี่ยวกับการค้นคืนสารสนเทศ(Information Retrieval) นั้นเอกสารต่างๆที่จะนำมาประมวลผลจะต้องจัดเก็บให้อยู่ในรูปต่างๆ เรียกว่า IR model ปัจจุบันมีการคิดค้นโมเดลต่างๆมาสนับสนุนมากมาย ซึ่งทุกๆโมเดลจะอ้างอิงจากโมเดล คือ บูลลีนโมเดล(Boolean model) เวคเตอร์โมเดล(Vector model) และแบบจำลองเชิงความน่าจะเป็น(Probabilistic model) ซึ่งจะกล่าวถึงในหัวข้อต่อไป

3.1 บูลลีนโมเดล

บูลลีนโมเดล(Boolean model) เป็นโมเดลในระบบ IR ที่ทำงานอยู่บนพื้นฐานของเซตและพีชคณิต(Boolean algebra) ซึ่งเป็นโมเดลที่ง่ายต่อการทำความเข้าใจและง่ายต่อการนำมาใช้งาน ด้วยคุณสมบัติดังกล่าวทำให้บูลลีนโมเดลได้รับความนิยมในเชิงพาณิชย์เป็นอย่างมาก แต่ประสิทธิภาพของโมเดลนี้ไม่ค่อยดีมากนักดังสรุปได้ดังนี้

1. การได้มาซึ่งข้อมูลจะอยู่บนพื้นฐานของการตัดสินใจแบบไบนารี(binary decision) กล่าวคือเอกสารจะถูกจัดอยู่ในกลุ่มเอกสารที่สัมพันธ์กันหรือไม่สัมพันธ์กันอย่างใดอย่างหนึ่ง โดยไม่มีการแบ่งลำดับชั้นย่อยเลย ทำให้ไม่สามารถใช้เป็นข้อมูลในการจัดอันดับเอกสารที่สืบค้นได้
2. ในบางครั้งการที่จะแปลเอาความต้องการมาของผู้ใช้ให้อยู่ในรูปแบบของบูลีนนั้นสามารถทำได้ยาก

บูลลีนโมเดลนั้นจะพิจารณาถึงตัวอินเด็กซ์เทอม(index term)ว่าปรากฏอยู่ในเอกสารหรือไม่ นั้นคือน้ำหนักของเทอมจะมีค่าเป็น 0,1 เท่านั้น $w_{ij} \in \{0,1\}$ คิวรีนั้นจะประกอบไปด้วยอินเด็กซ์เทอมที่เชื่อมด้วย and , or , not ดังนั้นจึงสามารถแทนคิวรีได้ในรูป disjunction of conjunction vector (disjunctive normal from : DNF) เช่น $[q = k_a \wedge (k_b \vee \neg k_c)]$ สามารถเขียนให้อยู่ในรูป DNF ได้เป็น $[\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$ เมื่อแต่ละองค์ประกอบของ binary weight vector แทนด้วยทัปเปิล (k_a , k_b , k_c)

นิยาม 3.1 : ระบบ IR model แทนด้วยจตุรกูณคือ $[D,Q,F,R(q_i,d_j)]$ *เมื่อ*

(1) D *คือ เซตของกลุ่มเอกสารทั้งหมดในระบบ*

(2) Q *คือ เซตของคิวรี(query) ที่ป้อนโดยผู้ใช้*

(3) F *คือ รูปแบบของโมเดลที่ใช้ในระบบ*

(4) $R(q_i,d_j)$ *คือ ฟังก์ชันที่ชันจัดอันดับผลการค้นคืนเอกสาร* $d_j \in D$ *ที่สัมพันธ์กับคิวรี* $q_i \in Q$ *ที่เรียงตามลำดับความสัมพันธ์ที่มีต่อกัน*

เมื่อเอกสาร d_j ถูกแทนด้วยอินเด็กซ์เทอม(Index terms) ดังนั้น D ก็คือเซตของทุกอินเด็กซ์เทอมที่มีอยู่ใน d_j กำหนดให้ k_i เป็นอินเด็กซ์เทอมและ d_j เป็นเอกสาร $d_j \in D$ ดังนั้น $w_{i,j} \geq 0$ คือค่าน้ำหนักของความสัมพันธ์ระหว่าง k_i ที่มีต่อ d_j

นิยาม 3.2: กำหนดให้ t *คือจำนวนอินเด็กซ์เทอมทั้งหมดในระบบ,* $t \in D$*, และ* $K=\{k_1, \ldots, k_t\}$ *ถ้าหากค่าน้ำหนัก* $w_{i,j} > 0$ *แสดงว่าอินเด็กซ์เทอม* k_i *มีความสัมพันธ์ต่อเอกสาร* d_j

นิยาม 3.3: ให้น้ำหนักของอินเด็กซ์เทอมเป็นแบบไบนารี $w_{i,j} \in \{0,1\}$ *และคิวรี* q *ถูกแทนด้วยเวคเตอร์* $\vec{q}_{dnf}$ *ในรูปของ Disjunctive Normal Form : DNF และให้เวคเตอร์* $\vec{q}_{cc}$ *คือองค์ประกอบที่อยู่ใน* $\vec{q}_{dnf}$ *ดังนั้นสามารถคำนวณค่า similarity ระหว่างเอกสาร dj กับคิวรี q ได้คือ*

$$sim(d_j,q) = \begin{cases} 1 & if\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0 & otherwise \end{cases}$$

ถ้าค่า $sim(d_j,q)=1$ *แล้วบูลลีนโมเดลจะทำนายว่าเอกสาร* d_j *จะมีความเกี่ยวข้อง(Relevant)กับคิวรี* q *ถ้านอกเหนือไปจากนี้จะสรุปว่าเอกสาร* d_j *ไม่มีความเกี่ยวข้องกับคิวรี* q

บูลลีนโมเดลมีความสามารถทำนายได้แค่ว่าแต่ละเอกสาร relevant หรือ non-relevant เท่านั้น ไม่สามารถระบุความเกี่ยวข้องแบบ partial matching ได้ ตัวอย่างเช่น เอกสาร d_j มี k_b เป็นอินเด็กซ์เทอมแทนด้วย $d_j = (0,1,0)$ ดังนั้นบูลลีนโมเดลจะสรุปว่า d_j non-relevant กับคิวรี $[q=k_a \wedge (k_b \vee \neg k_c)]$ เป็นต้น

3.2 เวคเตอร์โมเดล

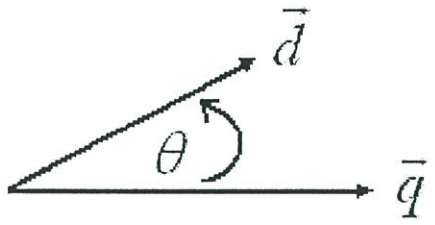
เอกสารที่พิจารณาจะถูกจัดให้อยู่ในรูปเวคเตอร์เรียกว่า Vector Model[6,7] โดยคุณสมบัติของโมเดลนี้จะมีประสิทธิภาพดีกว่าบูลลีนโมเดลเนื่องจากมีความสามารถทำ partial matching ได้ เพราะใช้ตัวเลขจำนวนจริงบวกแทนค่าน้ำหนักของเทอม ซึ่งต่างจากบูลลีนโมเดล ที่ใช้ค่าน้ำหนักของเทอมด้วยเลขไบนารี่, $w_{i,j} \in \{0,1\}$, เอกสารทั้งหมดที่อยู่ในฐานข้อมูลของระบบ(collection)จะถูกแทนด้วยเวคเตอร์น้ำหนักของเทอม ดังนั้นทำให้สามารถคำนวณหาค่าความคล้าย(similarity) ระหว่างเอกสาร d_j กับชุดคำค้นจากผู้ใช้ q (user query ต่อไปจะเรียกว่า คิวรี) ได้โดยแทน q ให้อยู่ในรูปเวคเตอร์เดียวกับ d_j ซึ่งค่าความคล้ายนี้จะเป็นตัวบ่งบอกว่าเอกสาร d_j คล้ายกับคิวรี q เพียงไรและมีประโยชน์ในการเรียงลำดับ(ranking)เอกสารในการออกผลรายงานต่อไป

นิยาม 3.4: กำหนดให้ค่าน้ำหนัก w_{ij} คือค่าความสัมพันธ์ของคู่ลำดับ (k_i,d_j) แสดงความมีอิทธิพลของเทอม k_i ที่มีต่อเอกสาร d_j มีค่าเป็นจำนวนจริงบวกแบบ non-binary และกำหนดให้ค่าน้ำหนัก $w_{i,q}$ค่าความสัมพันธ์ของคู่ลำดับ (k_i,q) แสดงความมีอิทธิพลของเทอม k_i ที่มีคิวรี q มีค่าเป็นจำนวนจริงบวก จะได้เวคเตอร์แทนเอกสารเป็น $\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ และเวคเตอร์แทนคิวรีเป็น $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ ตามลำดับ

ดังนั้นทั้งเอกสาร d_j และคิวรี q ต่างถูกแทนด้วยเวคเตอร์ในระบบ t-มิติ ดังแสดงในรูป 3.1 จากทฤษฎีของเวคเตอร์ทำให้สามารถคำนวณหา similarity ได้จากโอเปอร์เรเตอร์ที่ทำบนเวคเตอร์ เช่น ใช้ฟังก์ชัน cosine of angle:

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \qquad (3.1)$$

เนื่องจาก $w_{i,j} \geq 0$ และ $w_{i,q} \geq 0$ ดังนั้นค่า $sim(d_j,q)$ จึงมีค่าอยู่ในช่วง 0 ถึง +1 ซึ่งสามารถทำนายความเกี่ยวข้องกันระหว่างเอกสาร d_j กับ q ได้จากค่า $sim(d_j,q)$ หรือเรียกค่านี้ว่า degree of similarity



รูปที่ 3.1 การแทนคิวรีและเอกสารให้อยู่ในรูปเวคเตอร์ *t*-มิติ

นิยาม 3.5: *กำหนดให้ N คือจำนวนเอกสารทั้งหมดใน collection และ n_i คือจำนวนของเอกสารที่มี k_i ปรากฏอยู่ ดังนั้น $freq_{i,j}$ คือความถี่(ทางสถิติ)ของเทอม k_i ที่ปรากฏในเอกสาร d_j (จำนวนครั้งที่เทอม k_i ถูกกล่าวถึงในเอกสาร d_j) ดังนั้นค่า normalized frequency, $f_{i,j}$,ของเทอม k_i ในเอกสาร d_j แสดงได้เป็น $f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$ เมื่อ max_l คือค่าความถี่ของเทอมที่สูงที่สุดในเอกสาร d_j*

นิยาม 3.6: *ถ้าเทอม k_i ปรากฏอยู่ในทุกเอกสารใน collection แล้ว เทอม k_i จะไม่มีอำนาจในการจำแนกเอกสาร d_j ออกจาก collection ได้ หรือถ้าเทอม k_i ปรากฏในหลายๆเอกสารใน collection แล้ว เทอม k_i จะมีความสำคัญต่อเอกสาร d_j น้อยลง เรียกค่านี้ว่า inverse document frequency ของเทอม k_i $idf_i = \log(N / n_i)$ ดังนั้นค่า w_{ij} ในนิยาม 3.4 คำนวณได้จาก $w_{i,j} = f_{i,j} \times idf_i$ เมื่อ f_{ij} นิยามตามนิยามที่ 3.5 เรียกวิธีการคำนวณแบบนี้ว่า if-idf*

<u>ตัวอย่าง</u> : กำหนดให้ใน collection ประกอบด้วยเอกสารจำนวน 10,000 ชุด และเอกสาร d_j มีเทอม A ปรากฏ 3 ครั้ง เทอม B ปรากฏ 2 ครั้ง และเทอม C ปรากฏ 1 ครั้ง จะได้

$$tf_{Aj}= 3 , tf_{Bj}= 2 \text{ และ } tf_{Cj}=1$$

และจำนวนเอกสารใน collection ที่มีเทอม A ปรากฏอยู่ 50 เอกสาร, เทอม B ปรากฏอยู่ 1,300 เอกสาร และเทอม C ปรากฏอยู่ 250 เอกสาร จะได้

$$df_A= 50 , df_B= 1300 \text{ และ } df_C= 250$$

ดังนั้นเทอม A, B และ C จะมีน้ำหนักความสำคัญต่อเอกสาร d_j เป็น:

w_{Aj}: tf = 3/3; idf = log(10000/50) = 5.3; tf-idf = 5.3

w_{Bj}: tf = 2/3; idf = log(10000/1300) = 2.0; tf-idf = 1.3

w_{Cj}: tf = 1/3; idf = log(10000/250) = 3.7; tf-idf = 1.2

จากตัวอย่างจะพบว่าเทอม B จะมีค่า *idf* เท่ากับ 2.0 ซึ่งน้อยกว่าเทอมอื่นๆ เพราะเทอม B ปรากฏในหลายๆเอกสารจึงทำให้อำนาจจำแนกเอกสาร d_j ออกจาก collection นั้นมีค่าน้อยลงตามลำดับ

สรุปการทำงานของเวคเตอร์โมเดลคือ การแทนเอกสารใน Collection และ คิวรี ให้อยู่ในรูปเวคเตอร์ เมื่อทั้งเอกสารและคิวรีอยู่ในรูปเวคเตอร์แล้ว ทำให้สามารถใช้โอเปอร์ทางคณิตศาสตร์มาใช้ประมวลผลในกระบวนการวัดความคล้ายได้ ประโยชน์ของเวคเตอร์คือ 1). ด้วยวิธีการ *tf-idf* ทำให้สามารถคำนวณ degree of similarity ระหว่างคิวรี q กับเอกสาร d_j ด้วยโอเปอร์เรเตอร์ทางคณิตศาสตร์ได้ 2). สามารถค้นคืนเอกสารแบบ partial matching ได้ และ 3). สามารถนำค่า degree of similarity มาใช้ในกระบวน ranking เพื่อเรียงลำดับเอกสารก่อนรายงานต่อผู้ใช้ อย่างไรก็ตาม มีโมเดลอื่นๆอีกหลายแบบได้ถูกนำเสนอขึ้นมาใช้งานใน IR system เช่น Probabilistic model เป็นต้น แต่ในวิทยานิพนธ์เล่มนี้ได้แสดงเพียงโมเดลพื้นฐานไว้เท่านั้นครับ

3.3 เวคเตอร์โมเดลสำหรับเอกสาร HTML

จากหัวข้อ 2.2 เป็นเวคเตอร์โมเดลที่ใช้สำหรับเอกสารที่เป็น Text ธรรมดา ไม่มีโครงสร้างที่มีความหมาย ซึ่งต่างจากเอกสาร HTML ที่เป็นเอกสารที่มีโครงสร้างที่มีความหมาย เช่น โครงสร้างบ่งบอกว่าเอกสารนั้นมีเนื้อหากล่าวถึงเรื่องอะไร โครงสร้างบ่งบอกมีเอกสารอื่นๆที่อยู่ในข่ายเดียวกัน หรือโครงสร้างบ่งบอกหัวเรื่อง หัวข้อ ของเนื้อหาภายในเป็นต้น

ข้อแตกต่าง 2 ประการของเอกสารจาก HTMLจากเอกสารธรรมดาในระบบ Traditional IR คือ

1. เอกสาร HTML มีโครงสร้างตาม HMTL Tags ซึ่งโครงสร้างนี้จะเป็นตัวบ่งบอกเนื้อหาภายในเอกสารได้ ขณะที่เอกสารธรรมดาในระบบ Traditional IR นั้นไม่มีโครงสร้างของเนื้อหา
2. ใน HTML Collection จะมีกลุ่มลิงค์ ซึ่งสามารถนำมาวิเคราะห์เนื้อหาของแต่เอกสารได้ โดยทั่วไปแล้วผู้เขียนเว็บเพจจะเพิ่มเติมข้อมูลสำคัญบางประการลงใน Anchor Tag เพื่ออธิบายลิงค์ดังกล่าวแทนที่จะมีเพียงแค่ URLs ในขณะเอกสารธรรมดาในระบบ IR ดั้งเดิมนั้นมีเพียงแค่เทอมต่างๆสำหรับอธิบายเนื้อหาในเอกสารเท่านั้น

ด้วยเหตุผลดังที่กล่าวมานี้จึงจัดกลุ่มของเทอมในเอกสารออกเป็น 6 กลุ่มจำแนกตามโครงสร้าง HTML tags: Title, Header, Anchor, Strong, List และ Plain Text โดยแต่ละคลาสจะประกอบด้วยเทอมต่างๆที่ปรากฏในแต่ละ Tag ดังแสดงในตารางที่ 4.1 แนวคิดพื้นฐานคือเทอมใดๆก็ตามหากปรากฏในโครงสร้างของเอกสารที่ต่างกันแล้วอาจจะมีค่าความสำคัญ(class importance value: CIV) ต่อเอกสารนั้นต่างกัน คือ: เทอมที่ปรากฏในคลาส Title จะให้ข้อมูลเกี่ยวกับเอกสารว่าได้กล่าวอะไร, เทอมที่ปรากฏในคลาส Header จะให้ข้อมูลเกี่ยวกับโครงสร้างหลักและหัวข้อหลักของเอกสาร, เทอมที่ปรากฏในคลาส Anchor จะให้ข้อมูลเกี่ยวกับหัวเรื่องของเอกสารอื่นๆใน collection ที่มีสัมพันธ์กันดังนั้นจึงสามารถให้เป็นข้อมูลเพิ่มเติมเกี่ยวกับหัวเรื่องของเอกสารนั้นได้, เทอมที่ปรากฏในคลาส Strong จะให้ข้อมูลที่จำเพาะเจาะจง(เน้น)เกี่ยวกับเนื้อหาของเอกสาร, เทอมที่

ปรากฏในคลาส List จะให้ข้อมูลเกี่ยวกับภาพรวมและเนื้อความที่ได้จากการสรุป และเทอมที่ปรากฏในคลาส Plain Text คือเนื้อหาของเอกสารซึ่งประกอบด้วยเทอมจำนวนมาก

ชื่อคลาส	ชื่อ Tag
Title	TITILE
Header	H1, H2, H3, H4, H5, H6
Anchor	A
Strong	STRONG, B, EMM, I, U
List	DL, OL, UL
Plain Text	Text

ตารางที่ 3.1 คลาสของเทอมในเอกสาร HTML

เนื่องจากวิธีนี้จะการพิจาณาโครงสร้างของเอกสารต่างๆซึ่งต่างจากเวคเตอร์โมเดลในหัวข้อที่ 3.2 ดังนั้นค่าพารามิเตอร์ต่างๆ จึงต้องนิยามใหม่ด้วย

นิยาม 3.7: *กำหนดให้ CIV คือค่าความสำคัญของแต่ละเทอมในแต่ละคลาส* $CIV=(c_1,c_2,c_3,c_4,c_5,c_6)$ *เมื่อ* $c_n \in \{1,2,3,...,10\}$ *และ term-frequency ของเทอมในแต่ละคลาสแสดงด้วย Term Frequency Vector* $TFV=(tf_{c1}, tf_{c2}, tf_{c3}, tf_{c4}, tf_{c5}, tf_{c6},)$ *เมื่อ* tf_{c1} *แทนค่าความถี่ของ term i ที่ปรากฏในคลาส c1 หรือคลาส Title และ* tf_{c2} *แทนค่าความถี่ของ term i ที่ปรากฏในคลาส c2 หรือคลาส Header ตามลำดับในตารางที่ 3.1*

เมื่อนิยามของ term-frequency ถูกนิยามใหม่ดังนั้นวิธีคำนวณค่า weight-term จึงต้องนิยามใหม่ คือ

$$\begin{aligned} w_{i,j} &= (TFV_{i,j} \bullet CIV) * idf_i \\ &= TF_{i,j} * idf_i \end{aligned} \tag{3.2}$$

เมื่อ

$$TF_{i,j} = \sum_{n=1..6} \frac{tf_{i,j,n}}{\max l_n} * civ_n$$

เมื่อ *TFV* และ *CIV* นิยามด้วย *นิยาม 3.7* $tf_{i,j,n}$ คือความถี่ของ term *i* ที่ปรากฏใน class *n* ของ document d_j และ $\max l_n$ คือ term ที่มีความถี่สูงสุดใน class *n* และ idf_i นิยามตาม *นิยาม 3.6* การนำค่าความ CIV มาพิจารณาในกระบวนการคำนวณ weight-term ทำให้ระบบที่ทำงานกับเอกสาร HTML มีความเที่ยงตรงเพิ่มขึ้น

บทที่ 4
การประเมินประสิทธิภาพของระบบค้นคืนสารสนเทศ

การวัดประสิทธิภาพเป็นขั้นตอนในการทดสอบระบบว่ามีประสิทธิภาพในการทำงานเป็นเช่นไร โดยทั่วไปจะพิจารณาจาก 2 ปัจจัยคือ เวลาที่ใช้ (time) และหน่วยความจำที่ต้องการในการประมวลผล(space) ระบบที่ใช้เวลาในการประมวลผลน้อย และระบบที่ใช้หน่วยความจำในการประมวลผลน้อยกว่าจะถือว่าระบบนั้นมีประสิทธิภาพดีกว่าอีกระบบหนึ่ง แต่ในระบบ IR นั้นนิยมพิจารณาประสิทธิภาพของระบบในอีกรูปแบบหนึ่งนอกเหนือจากพิจารณาจากเวลาที่ใช้และหน่วยความจำที่ต้องการในการทำงานของระบบ ในระบบ IR นิยมใช้ค่า Recall และ Precision เป็นตัววัดประสิทธิภาพของระบบ

4.1 การวัดประสิทธิภาพด้วยค่า Recall

การวัดประสิทธิภาพของระบบ IR จากค่า Recall กระทำโดยป้อนคิวรี q จากผู้ใช้ให้กับระบบค้นคืนสารสนเทศ $\boldsymbol{S}$ แล้วพิจารณาว่าระบบ $\boldsymbol{S}$ สามารถดึงข้อมูลที่เกี่ยวข้องกับคิวรีออกมาจากฐานข้อมูลได้มากน้อยเพียงไร

การทดสอบระบบด้วยค่า Recall กระทำโดยป้อนคิวรี q ให้กับระบบ $\boldsymbol{S}$ แล้วตรวจสอบความสามารถของระบบ $\boldsymbol{S}$ ในการดึงเอกสารจากเซต R มาใส่เซต A ได้มากน้อยเพียงไรดังแสดงในรูปที่ 4.1 และสมการ (4.1)

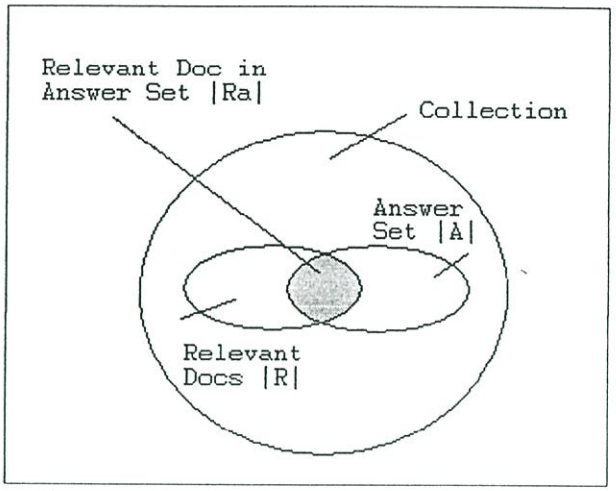
$$\text{Re}call = \frac{|Ra|}{|R|} \tag{4.1}$$

เมื่อ R คือ เซตของเอกสารใน collection D ที่เกี่ยวข้องกับคิวรี q

A คือ เซตของเอกสารคำตอบที่ได้จากระบบ $\boldsymbol{S}$

Ra คือ เซตที่ได้จาก $R \cap A$

ถ้าหากคา Recall มีค่ามาก (เข้าใกล้ 1) แสดงว่าระบบ S มีประสิทธิภาพในเชิง Recall ดีตามลำดับ



รูปที่ 4.1 ค่า Recall และค่า Precision

4.2 การวัดประสิทธิภาพด้วยค่า Precision

การวัดประสิทธิภาพของระบบ IR จากค่า Precision กระทำโดยป้อนคิวรี q จากผู้ใช้ให้กับระบบ S แล้วตรวจสอบว่าเซตคำตอบ A ที่ระบบค้นคืนมาได้นั้นมีเอกสารที่เกี่ยวข้องกับคิวรี q อยู่ในเป็นอัตราส่วนเป็นเท่าไร ดังแสดงตามสมการ (4.2)

$$\Pr ecision = \frac{|Ra|}{|A|} \tag{4.2}$$

เมื่อ R คือ เซตของเอกสารใน collection D ที่เกี่ยวข้องกับคิวรี q

A คือ เซตของเอกสารคำตอบที่ได้จากระบบ S

Ra คือ เซตที่ได้จาก $R \cap A$

4.3 ความสัมพันธ์ระหว่าง Recall กับ Precision

พิจารณาความสัมพันธ์ของค่า Recall และ Precision แล้วจะพบว่าความสัมพันธ์ของค่าทั้งสองจะเป็นในลักษณะแปรผกผันดังแสดงในรูปที่ 4.2 นั่นคือถ้าต้องการให้ระบบมีค่า Recall สูงแล้วระบบจะให้ค่า Precision ต่ำ ในทางกลับกันถ้าต้องการให้ระบบมีค่า Precision สูงแล้วระบบจะให้ค่า Recall ต่ำ ลองพิจารณาความสัมพันธ์ระหว่างค่า Recall และค่า Precision จากตัวอย่างต่อไปนี้

<u>ตัวอย่าง:</u> ให้ S ระบบซึ่งเมื่อป้อนคิวรี q ให้กับ S แล้ว S จะส่งคืนเซตคำตอบ A ออกมาโดยมีจัดลำดับดังนี้

Ranking for query q:

1. d_{123} •	4. d_6	7. d_{511}	10. d_{25} •	13. d_{250}
2. d_{84}	5. d_8	8. d_{129}	11. d_{38}	14. d_{113}
3. d_{56} •	6. d_9 •	9. d_{187}	12. d_{48}	15. d_3 •

ในขณะที่เซตของเอกสารใน R ที่เกี่ยวข้องกับคิวรี q: R_q={d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{89}, d_{123}} สังเกตเอกสารในเซตคำตอบ A ที่มีจุดอยู่ข้างหลัง คือเอกสารที่เกี่ยวข้องกับคิวรี q หรือเอกสารนั้นอยู่ใน R_q นั่นเอง

เริ่มพิจารณาเอกสารในเซตคำตอบ A ทีละเอกสาร จะได้

1. เอกสาร d_{123} เป็นเอกสารที่อยู่ใน R_q ถูกจัดลำดับให้อยู่ในลำดับที่ 1 ดังนั้น ณ. จุดนี้ระบบจะให้

   Recall = 1/10 = 10% และ Precision = 1/1 = 100%

2. เอกสาร d_{56} เป็นเอกสารที่อยู่ใน R_q ถูกจัดลำดับให้อยู่ในลำดับที่ 3 ดังนั้น ณ. จุดนี้ระบบจะให้

   Recall = 2/10 = 20% และ Precision = 2/3 = 66%

3. เอกสาร d_9 เป็นเอกสารที่อยู่ใน R_q ถูกจัดลำดับให้อยู่ในลำดับที่ 6 ดังนั้น ณ. จุดนี้ระบบจะให้

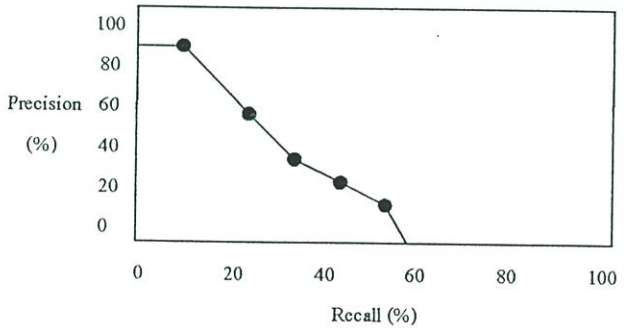
   Recall = 3/10 = 30% และ Precision = 3/6 = 50%

4. เอกสาร d_{25} เป็นเอกสารที่อยู่ใน R_q ถูกจัดลำดับให้อยู่ในลำดับที่ 10 ดังนั้น ณ. จุดนี้ระบบจะให้

   Recall = 4/10 = 40% และ Precision = 4/10 = 40%

5. เอกสาร d_3 เป็นเอกสารที่อยู่ใน R_q ถูกจัดลำดับให้อยู่ในลำดับที่ 15 ดังนั้น ณ. จุดนี้ระบบจะให้

   Recall = 5/10 = 20% และ Precision = 5/15 = 33.33%



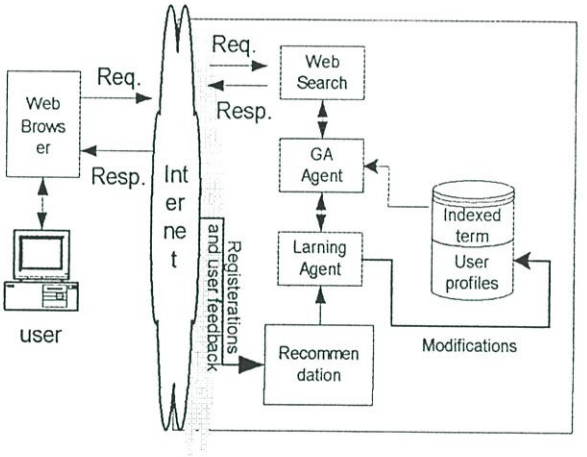
รูปที่ 4.2 ความสัมพันธ์ระหว่างค่า Recall และ Precision

เมื่อนำความสัมพันธ์ดังกล่าวมาเขียนเป็นกราฟ จะได้ความสัมพันธ์ดังแสดงในรูปที่ 4.2 ค่า Precision ในระดับที่ค่า Recall มากกว่า 50% นั้นจะมีค่าเข้าสู่ค่า 0 เพราะไม่ว่าเอกสารใดๆที่เกี่ยวข้องกับคิวรี q ถูก retrieve ออกมาอีกแล้ว

บทที่ 5
การออกแบบและสร้างระบบ

5.1 โครงสร้างของระบบ

โครงสร้างระบบแสดงดังในรูปที่ 5.1 จุดประสงค์หลักของระบบคือ ร้องขอใช้บริการเครื่องมือค้นหาจากผู้ให้บริการจำนวน 5 ตัวคือ Yahoo, Altavista, Lycos, Hotbot และ Google และนำผลรายงานที่ได้เครื่องมือค้นหาดังกล่าวมาวิเคราะห์เพื่อจำแนกเอกสารหรือเลือกเฉพาะเอกสารที่ตรงกับความต้องการของผู้ใช้ ระบบมีการทำงานแบ่งออกเป็น 5 ขั้นตอนดังนี้



รูปที่ 5.1 โครงสร้างของระบบ

1. รับข้อมูลจากการลงทะเบียนของผู้ใช้ จัดกลุ่มผู้ใช้ตามสาขาวิชาที่ผู้ใช้สนใจ
2. นำข้อมูลลงทะเบียนจากผู้ใช้มาสร้างเป็นยูสเซอร์โปรไฟล์และใช้เป็นข้อมูลสำหรับสอนระบบ
3. สร้างชุดคิวรีจากชุดคำค้นของผู้ใช้เพื่อป้อนให้กับ Metasearch เพื่อค้นหาข้อมูลจาก WWW
4. กรองผลรายงานโดยเลือกเอาเฉพาะเอกสารที่ตรงตามความสนใจของผู้ใช้และรายงานผลการสืบค้นต่อผู้ใช้
5. รับข้อมูลป้อนกลับจากผู้ใช้เพื่อนำมาปรับปรุงยูสเซอร์โปรไฟล์ให้มีความทันสมัยตลอดเวลา วนกลับไปทำงานในขั้นตอนที่ 3.

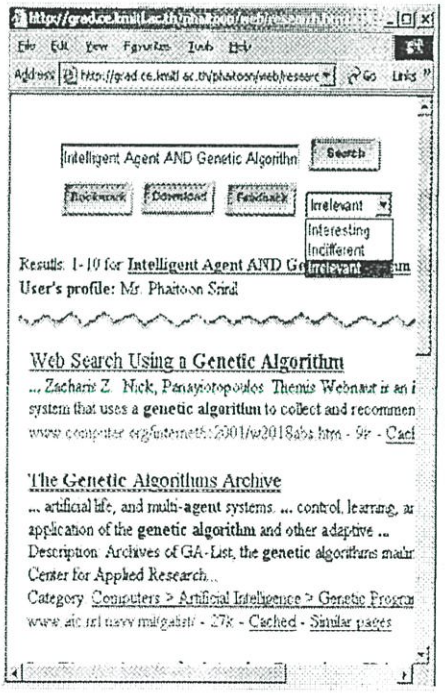
การทำงานของระบบจะวนรอบอยู่ระหว่างขั้นตอนที่ 3 ถึงขั้นตอนที่ 5 การทำงานในขั้นตอนที่ 5 คือการปรับปรุงยูสเซอร์โปรไฟล์ ซึ่งกระบวนการนี้ดำเนินการโดย Learning Agent จากการทดลองจะพบว่าในยูสเซอร์โปรไฟล์หนึ่งๆจะมีความสนใจในสาขาวิชาที่เฉพาะเจาะจง(เช่น

Database System, Fuzzy Control เป็นต้น) ภายใต้โดเมนที่เฉพาะเจาะจงเช่นนี้ เมื่อผู้ใช้ได้รับข้อมูลที่ต้องการไปจำนวนหนึ่งหรือมากเพียงพอแล้วก็จะเปลี่ยนความสนในไปยังอีกสาขาวิชาอื่นที่ใกล้เคียงภายใต้โดเมนเดียวกัน ดังนั้นกระบวนการปรับปรุงยูสเซอร์โปรไฟล์(update)จึงเป็นขั้นตอนที่สำคัญมากเพราะความสนใจของผู้ใช้จะเปลี่ยนแปลงตลอดเวลา และจากการทดลองพบว่าการปรับปรุงยูสเซอร์โปรไฟล์ให้ทันสมัยอยู่เสมอจะทำให้ประสิทธิภาพของระบบเพิ่มขึ้นงานวิจัยนี้ใช้ 3 เหตุการณ์ในปรับปรุงยูสเซอร์โปรไฟล์คือ User Feedback, User Bookmark และ User Downloadable.

ระบบที่นำเสนอมีองค์ประกอบที่สำคัญคือ

5.1.1 Web Search

การทำงานของ Web Search คือตอบโต้กับผู้ใช้ รับข้อมูลลงทะเบียนจากผู้ใช้ รับคำค้นจากผู้ใช้ รับข้อมูลป้อนกลับจากผู้ใช้และแสดงผลรายงานแก่ผู้ใช้ รูปที่ 5.2 แสดงหน้าตาของ Web Search หน้าจอแบ่งออกเป็นสองส่วนคือ เฟรม(Frame)ส่วนบนสำหรับรับข้อมูลจากผู้ใช้(รับชุดคำค้นและส่งข้อมูลป้อนกลับ)และเฟรมส่วนล่างสำหรับแสดงผลรายงานต่อผู้ใช้



รูปที่ 5.2 ส่วนตอบโต้ผู้ใช้ของ Web Search

5.1.2 GA Agent

GA Agent มีหน้าที่หลัก 2 ประการคือ (1)ใช้กระบวนการจีนีติกอัลกอริทึมสร้างชุดคิวรีที่เหมาะสมจากคำค้นที่ป้อนโดยผู้ใช้ป้อนให้กับ Metasearch โดยใช้เครื่องมือค้นหาจำนวน 5 ตัวคือ Yahoo, Altavista, Lycos, Hotbot และ Google และ (2) ค้นหาค่า CIV ที่เหมาะสมที่สุดสำหรับแต่ละยูสเซอร์โปรไฟล์

5.1.3 Learning Agent

Learning Agent คือหน่วยการเรียนรู้ความต้องการของผู้ใช้ ข้อมูลของผู้ใช้แต่ละคนจะถูกเก็บไว้ในยูสเซอร์โปรไฟล์ โดยข้อมูลในยูสเซอร์โปรไฟล์จะถูกปรับปรุงให้ทันสมัยอยู่

5.1.4 Recommendation

ผู้ใช้สามารถส่งข้อมูลป้อนกลับเพื่อปรับปรุงยูสเซอร์โปรไฟล์โดยป้อนข้อมูลดังกล่าวผ่านทางเฟรมส่วนบนของ Web Search รูปแบบข้อมูลป้อนกลับแบ่งออกเป็น 3 ประเภทคือ

1. *UF*(User Feedback) ผู้ใช้สามารถกดปุ่ม **Feedback** และเลือกให้คะแนนกับโฮมเพจในผลรายงานเป็น Interesting, Indifferent หรือ Irrelevant โดยระบบจะส่งคืนค่า "1", "0" และ "–1" ตามลำดับ
2. *UD*(User Downloadable) ถ้าหากผู้ใช้กดปุ่ม **Download** เพื่อบันทึกโฮมเพจ(HTML document) ที่สนใจเก็บลงที่เครื่องคอมพิวเตอร์ของผู้ใช้แล้ว ระบบจะส่งคืนค่า "1"
3. *UB*(User Bookmark) ถ้าหากผู้ใช้กดปุ่ม **Bookmark** เพื่อบันทึก URL ของโฮมเพจที่สนใจลงใน bookmark ของเว็บบราวเซอร์ได้ ระบบจะส่งคืนค่า "1"

ตารางที่ 5.1 แสดงข้อมูลป้อนกลับในแต่ละรูปแบบ ระบบ Learning Agent จะนำข้อมูลป้อนกลับที่ได้นำมาใช้ในการปรับปรุงยูสเซอร์โปรไฟล์ซึ่งจะกล่าวโดยละเอียดในหัวข้อที่ 5.4 เรื่อง Learning Agent Implementations.

User Feedback	c		
	UF	UD	UB
Interesting	1	-	-
Indifferent	0	-	-
Irrelevant	-1	-	-
True	-	1	1

ตารางที่ 5.1 ค่า Recommendation URL แสดงความสนใจที่ป้อนกลับจากผู้ใช้

5.2 การสร้างและการทำงานของ GA Agent

การทำงานของ GA Agent มีการทำงานแบ่งออกเป็นขั้นตอนคือ (1) สร้างชุดคิวรีที่เหมาะสมจากคำค้นที่ป้อนจากผู้ใช้และ (2) ค้นหาค่า CIV ที่เหมาะสมที่สุดสำหรับแต่ละยูสเซอร์โปรไฟล์ (ดู CIV ให้หัวข้อ 3.3)

5.2.1 การปรับปรุงคิวรีก่อนนำไปใช้งาน

การทำงานในขั้นตอนนี้คือ สร้างชุดคิวรีเหมาะสมจากคำค้นที่ป้อนโดยผู้ใช้ GA Agent จะรับคำค้นจากผู้ใช้ผ่านทาง Web Search นำมาเข้ารหัสเป็นโครโมโซมของประชากรต้นกำเนิดเข้าสู่กระบวนการจีนีติกอัลกอริทึม เมื่อสิ้นสุดกระบวนการจีนีติกอัลกอริทึมระบบจะได้โครโมโซมที่มีลักษณะดี จากนั้นนำโครโมโซมที่ได้มาถอดรหัสเป็นชุดคิวรีที่เหมาะสมเพื่อป้อนให้กับเครื่องมือค้นหาต่อไป

5.2.2 การเข้ารหัสโครโมโซมสำหรับปรับปรุงคิวรี

นำคำค้นที่ป้อนโดยจากผู้ใช้มาเทียบเคียงกับชุดเอกสารใน Indexed term แล้วจะได้ชุดเอกสารจำนวนหนึ่ง (ดูรูปที่ 5.1 ประกอบ) ที่มีความเกี่ยวข้องกับชุดคำค้นของผู้ใช้

นิยาม 5.1: *กำหนดให้ Q เป็นเซตของชุดคำค้นที่ป้อนโดยผู้ใช้ $Q = \{q_1, q_2, \ldots, q_n\}$ และ Doc เป็นเซตของ term ที่ได้จากการเทียบเคียงสมาชิกใน Q กับชุดเอกสารใน Indexed term, $Doc_j = \{t_{1j}, t_{2j}, \ldots, t_{mj} \mid m,j \in$ positive number$\}$เมื่อ t_{ij} คือ termที่ i ของเอกสารหมายเลข j*

ตัวอย่างเช่นผู้ใช้ส่งชุดคำค้นเป็น

$$Q=\{Data\ mining\ ,\ Genetic\ algorithms\ ,\ Expert\ system\}$$

เมื่อนำชุดคำค้น Q ไปเทียบเคียงกับชุดเอกสารใน Indexed term จะได้ชุดของเอกสารที่เกี่ยวข้องกับ Q ดังนี้

Doc_1 = (*Internet, genetic algorithm, search engine, intelligent agent*)

Doc_2 = (*automated test generation, dialog model specification, genetic algorithm, software engineering test process*)

Doc_3 = (*expert system, decision support system, natural language processing, data mining*)

Doc_4 = (*artificial intelligence, information retrieval, expert system, data mining*)

นิยาม 5.2: *ให้* $K=\{t_1, t_2, ..., t_k\}$ *เป็นเซตของ term ต่างๆที่ปรากฏใน ทุก Doc เมื่อ* $k \in$ *positive number และกำหนดให้* $Chr_j=\{c_{1j}, c_{2j}, ..., c_{kj}\}$ *คือ โครโมโซมที่ได้จากการเข้ารหัส* Doc_j *แบบไบนารี่ เมื่อ* $c_{ij} \in \{0,1\}$ *คือยีนส์โลคัสที่ i ของโครโมโซม* Chr_j

เมื่อนำทุก term ในทุกๆ Doc มาจัดเรียงตามลำดับตัวอักษรจะได้

K={ artificial intelligent , automated test generation , data mining , decision support system , dialog model specification , expert system , genetic algorithm , information retrieval , intelligent agent , Internet , natural language processing , search engine , software test engineering process}

เมื่อ

$$|K| = |Chr_j| \tag{5.1}$$

จะได้

$K=\{t_1, t_2, \ldots, t_{13}\}$

$Chr_j=\{c_{1j}, c_{2j}, \ldots, c_{13j}\}$

สามารถคำนวณค่า c_{ij} ได้จาก

$$c_{ij} = \begin{cases} 1 & if \quad t_i \in Doc_j \\ 0 & otherwise \end{cases} \tag{5.2}$$

จากสมการ (5.1)และ(5.2)จะได้

Chr_Q= {0010011000000}

Chr_1 ={0000001011010}

Chr_2 ={0100101000001}

Chr_3 ={0011010000100}

Chr_4 ={1010010100000}

จะได้ชุดโครโมโซมของประชากรต้นกำเนิดสำหรับเข้าสู่กระบวนการจีนีติกอัลกอริทึม

5.2.3 การถอดรหัสโครมโซมสำหรับปรับปรุงคิวรี

หลังจากประชากรต้นกำเนิดถูกนำเข้าสู่กระบวนการจีนีติกอัลกอริทึม สุดท้ายจะได้โครโมโซมที่มีลักษณะดีเหมาะสมกับปัญหา จากนั้นนำโครโมโซมดังกล่าวมาถอดรหัสกลับเป็นชุดคิวรีเหมาะสมเพื่อใช้ในเครื่องมือค้นหาต่อไป

นิยาม 5.3: *ให้* $Chr_{op}=\{c_1, c_2, \ldots, c_k \mid c_i \in \{0,1\}\}$ *คือโครโมโซมที่ถูกเลือกเป็นคำตอบจากกระบวนการจีนีติกอัลกอริทึมและ* $Q_{op}=\{q_1, q_2, \ldots, q_n\}$*คือเซตของ key term* t_i *ที่ได้จากการถอดรหัสไบนารี่* Chr_{op} *เมื่อ n จำนวนสมาชิกใน* Chr_{op} *ที่มีค่าเป็น "1"*

จากตัวอย่างและสมการ (5.1) สามารถเขียนได้เป็น

$$K=\{t_1, t_2, \ldots, t_{13}\}$$
$$Chr_{op}=\{c_1, c_2, \ldots, c_{13}\}$$

ถอดรหัสไบนารี่ด้วยแทน Q_{op} ด้วย term t_i เมื่อโลคัส c_i เป็น "1"

$$Q_{op}=\forall t_i \mid c_i=1, i=\{1, 2, \ldots, k\} \tag{5.3}$$

คำตอบของกระบวนการจีนีติกอัลกอริทึมได้เป็น

$$Chr_{op}=\{1000001001010\}$$
$$Q_{op}=\{\textit{artificial intelligent}, \textit{genetic algorithm}, \textit{intelligent agent}, \textit{search engine}\}$$

GA Agent จะใช้สมาชิกใน Q_{op} เป็นชุดคิวรีเหมาะสมสำหรับป้อนให้กับเครื่องมือค้นหาเพื่อใช้ในการสืบค้นข้อมูลต่อไป

5.2.4 ฟังก์ชันความเหมาะสมที่ใช้ในการปรับปรุงคิวรี

ฟังก์ชัน Jaccard Coefficient (JC) เป็นตัววัดลักษณะความเหมาะสมกับปัญหาของแต่ละโครโมโซมในกระบวนการจีนีติกอัลกอริทึม สำหรับปัญหานี้ฟังก์ชัน Jaccard Coefficient จะเป็น

ตัววัดความคล้ายคลึงกันระหว่างโครโมโซม Chr_j กับโครโมโซม Chr_Q โดยถ้ามีความคล้ายคลึงกันมากฟังก์ชันจะส่งคืนค่าออกมามาก(เข้าใกล้ 1) ตามลำดับ

$$JC(Chr_Q, Chr_j) = \frac{\left|Chr_Q \cap Chr_j\right|}{\left|Chr_Q\right| + \left|Chr_j\right| - \left|Chr_Q \cap Chr_j\right|} \tag{5.4}$$

โดยที่ $JC \in [0,1]$ มีค่าอยู่ระหว่าง 0 ถึง 1

5.3 การสร้างและการทำงานของ Learning Agent

หน้าที่ของ Learning Agent คือปรับปรุงยูเซอร์โปรไฟล์ให้มีความทันสมัยตลอดเวลา เมื่อผู้ใช้ให้ข้อมูลป้อนกลับแก่ระบบ (ดูตารางที่ 5.1) Learning Agent จะดำเนินการปรับปรุงค่า iff_i และค่า w_i (ดูตารางที่ 5.2) ในยูสเซอรโปรไฟล์ให้มีความสอดคล้องกับ recommended URL ด้วยสมการ(5.5) และ (5.6) ตามลำดับ

5.3.1 โครงสร้างและการจัดการยูเซอร์โปรไฟล์

ยูเซอร์โปรไฟล์จะเก็บข้อมูลเกี่ยวกับตัวผู้ใช้ซึ่งข้อมูลต่างๆที่อยู่ในยูเซอร์โปรไฟล์จะถูกใช้ในกระบวนการคำนวณ weight-term (ดูหัวข้อที่ 3.3) รูปแบบของยูเซอร์โปรไฟล์ดังแสดงในตารางที่ 5.2 ยูเซอร์โปรไฟล์จะถูกแทนให้อยู่ในรูปเซต $u = (w_1, w_2, ..., w_N)$ เมื่อ N คือจำนวนเทอมทั้งหมดในยูเซอร์โปรไฟล์ สมการ (5.5) ใช้สำหรับคำนวณน้ำหนักของแต่ละเทอมในยูเซอร์โปรไฟล์

$$w_i = \frac{iff_i}{iff_{\max}} \tag{5.5}$$

และสมการ (5.6) สำหรับใช้ในกระบวนการปรับปรุงยูเซอร์โปรไฟล์

$$iff_i = iff_i + (\frac{c}{10})tf_D \tag{5.6}$$

เมื่อ iff_i (influence factor) คือ ค่าความมีอิทธิพลของเทอม i ที่มีต่อยูเซอร์โปรไฟล์ tf_D คือความถี่ของเทอม i ที่ปรากฏใน recommended URL $iff_{\max}$ คือเทอมในยูเซอร์โปรไฟล์ที่มีค่า iff_i สูงที่สุด และค่า $\frac{c}{10}$ คือค่าอัตราการเรียนรู้ ซึ่งถ้า c เป็นค่าลบจะหมายถึงอัตราการลืมเลือนดังแสดงในตารางที่ 5.1 จากสมการ (5.5) และสมการ (5.6) สังเกตได้ว่าการคำนวณ w_i ในยูเซอร์โปรไฟล์จะต่างจากการคำนวณ $w_{i,j}$ ใน d_j (ดูนิยาม 3.6) คือ การคำนวณ w_i ในยูเซอร์โปรไฟล์นั้นไม่ได้นำ *idf*

มาร่วมพิจารณาด้วยเป็นเพราะแต่ละเทอมที่ปรากฏในยูสเซอร์โปรไฟล์นั้นคือคีย์เวิร์ดที่ป้อนโดยผู้ใช้เพื่อแสดงความสนใจ และ/หรือ เป็นเทอมที่คัดเลือกมาจาก recommended URL ซึ่งต่างจากเทอมปกติที่ปรากฏใน d_j ดังนั้นการพิจารณาความสำคัญระหว่างเทอมทั้งสองชนิดนี้จึงต้องแตกต่างกันด้วย

Terms	*iff*	*w*
Intelligent	50	0.625
Learning	20	0.250
Neuron	30	0.375
Algorithm	60	0.750
...	...	...

ตารางที่ 5.2 รูปแบบของการเก็บข้อมูลในยูสเซอร์โปรไฟล์

งานวิจัยนี้ได้กำหนด iff_i จากตอนเริ่มต้นสร้างยูสเซอร์โปรไฟล์ขึ้นมาใหม่ครั้งแรก $iff_i = 1$ เมื่อผู้ใช้ป้อนกลับ recommended URL ยูสเซอร์โปรไฟล์จะปรับปรุงตามสมการ (5.5)(5.6) และระบบจะคัดเลือกเทอมจาก recommended URL (กรณีที่มีค่าป้อนกลับเป็น "1") ที่มีค่า w_{ij} ดีที่สุดจำนวน 5 เทอมใส่เพิ่มเข้าไปในยูสเซอร์โปรไฟล์ ถ้าหากผลจากสมการ (5.5) ทำให้ค่า w_i มีค่าเป็น 0 หรือมีค่าลบแล้ว Learning Agent จะตัดเทอม i ออกจากยูสเซอร์โปรไฟล์ ด้วยวิธีการนี้จะสนับสนุนให้ระบบสามารถเรียนรู้ความสนใจของผู้ใช้หรือสามารถปรับเปลี่ยนข้อมูลในยูสเซอร์โปรไฟล์ไปตามพฤติกรรมของผู้ใช้ได้

5.3.2 การคำนวณค่าความคล้ายระหว่างเอกสาร HTML กับยูสเซอร์โปรไฟล์

การคำนวณค่าความคล้ายระหว่างเอกสาร HTML หรือผลรายงานที่ได้รับจากเครื่องมือค้นหากับยูสเซอร์โปรไฟล์มีประโยชน์ในการจัดเรียงลำดับผลรายงานก่อนแสดงต่อผู้ใช้ โดยเอกสารจะถูกจัดเรียงตามค่าความคล้ายกับยูสเซอร์โปรไฟล์จากค่ามากไปยังค่าน้อย ตามลำดับ

เอกสารที่ได้รับจากเครื่องมือค้นหา(เอกสาร HTML) และยูสเซอร์โปรไฟล์จะถูกแทนให้อยู่ในรูปเวคเตอร์ [6,7] จากนั้นนำมากระทำโอเปอร์เรชันทางคณิตศาสตร์เวคเตอร์โดยนำค่า ***CIV*** (คูนิยาม 3.7)มาร่วมพิจารณา จะได้สมการที่ใช้ในการคำนวณตามสมการ (5.7)

$$
\begin{aligned}
w_{i,j} &= (TFV_{i,j} \bullet CIV) * idf_t \\
&= TF_{i,j} * idf_t \\
TF_{i,j} &= \sum_{n=1..6} \frac{tf_{i,j,n}}{\max l_n} * civ_n
\end{aligned}
\qquad 5.7
$$

เมื่อ $tf_{i,j,n}$ คือความถี่ของ term i ที่ปรากฏใน class n ของ document d_j และ $\max l_n$ คือ term ที่มีความถี่สูงสุดใน class n โดยที่ค่า CIV ที่เหมาะสมสำหรับแต่ละยูสเซอร์โปรไฟล์จะค้นหาด้วยกระบวนการจีนีติกอัลกอริทึม(ดูหัวข้อ 5.3.3)

5.3.3 การค้นหาค่า CIV ที่เหมาะสม

ค่า CIV ที่เหมาะสมคือเซตของตัวเลขจำนวนเต็มช่วง [1, 10] ใช้สำหรับแสดงลำดับความสำคัญของเทอมที่อยู่ต่างคลาสกัน(ดูตารางที่ 3.1) ค่านี้จะถูกนำมาพิจารณาในขั้นตอนคำนวณ weight-term ดังแสดงในสมการ (5.5)(5.6) การกำหนดค่าสำหรับแต่ละ civ_n ได้มาจากกระบวนการจีนีติกอัลกอริทึม ระบบจะสุ่มตัวเลขจำนวนเต็มช่วง [1, 10] เพื่อสร้างโครโมโซมจำนวน 30 โครโมโซมและสร้าง 5 โครโมโซมจากค่า civ_n ที่ได้จากการทดลอง (ผู้วิจัยทดลองกำหนดค่าเองด้วยมือ) กำหนดเจนเนอร์เรชันสูงสุดเป็น 50 รอบ และใช้ฟังก์ชันความเหมาะสมคือ ค่าเฉลี่ย 30 ลำดับแรกของผลรายงานบนฟังก์ชัน Cosine of angle เลือกค่า CIV ที่ทำให้ค่าเฉลี่ยออกเป็นดีที่สุดเป็นค่า CIV ที่เหมาะสม

$$
\begin{aligned}
sim(d_j, u) &= \frac{\vec{d}_j \bullet \vec{u}}{\left|\vec{d}_j\right| \times \left|\vec{u}\right|} \\
&= \frac{\sum_{i=1}^{t} w_{i,j} \times w_i}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_i^2}}
\end{aligned}
\qquad (5.8)
$$

เมื่อ u คือยูสเซอร์โปรไฟล์

เนื่องจากเอกสาร HTML เป็นเอกสารที่มีโครงสร้างที่มีความหมาย จากการทดลองพบว่าการให้ลำดับความสำคัญของเทอมในแต่ละคลาสต่างกันจะทำให้ค่า Precision ของระบบดีกว่าการใช้ weight-term ใน Traditional IR System ตามแบบเดิม

งานวิจัยนี้ได้สร้างระบบจำลองด้วยภาษา JAVA โดยสร้าง Search Engine ด้วยเวคเตอร์โมเดลที่ทำงานบนเอกสาร HTML สร้าง Indexing Engine โดยสำเนาเอกสารบางส่วนจาก Yahoo, Altavista, Lycos, Hotbot และ Google นำมาเก็บในฐานข้อมูลบน DBMS MySQL สร้างระบบปรับปรุงคิวรีจากผู้ใช้ด้วยจีนีติกอัลกอริทึม และสร้างระบบบริหารยูสเซอร์โปรไฟล์เพื่อเรียนรู้ความสนใจของผู้ใช้ ด้วยความสามารถของภาษา JAVA ทำให้ระบบจำลองที่สร้างขึ้นมานี้สามารถทำงานได้บนระบบปฏิบัติ MS Windows, Linux, UNIX หรือระบบปฏิบัติการอื่นๆทุกระบบ

บทที่ 6
ผลการทดลองและการวิเคราะห์

การทดสอบประสิทธิภาพของระบบแบ่งออกเป็น 3 ขั้นตอนคือ 1). ทดสอบการปรับปรุงยูสเซอร์คิวรี 2). ทดสอบการเรียนรู้ของยูสเซอร์โปรไฟล์ และ 3). ทดสอบค่า Precision ของระบบ (ดู Precision ในหัวข้อ 4.2)

6.1 ทดสอบการปรับปรุงยูสเซอร์คิวรี

การปรับปรุงยูสเซอร์คิวรีมีจุดประสงค์เพื่อค้นหาเทอมจาก Indexed term ที่มีความหมายคล้ายกับคิวรีที่ป้อนจากผู้ใช้เพื่อมาขยายจากคิวรีต้นฉบับ ขั้นตอนการทดลองแบ่งออกเป็นขั้นตอนดังนี้

1. สร้างฐานข้อมูล Indexed Term แล้วบรรจุชุด Doc ลงไปในฐานข้อมูล Indexed term
2. ทดสอบค่า Precision ของระบบเมื่อเลือกเทอมโดยใช้จีนีติกอัลกอริทึมเทียบกับค่า Precision เมื่อเลือกเทอมด้วยการสุ่มเทอมจากเซต K โดยตรง (ดูนิยามของเซต K ในนิยาม 5.2) เพื่อนำมาขยายชุดคิวรีต้นฉบับ
3. บันทึกผลการทดสอบลงตารางที่ 6.1

จากตารางที่ 6.1(ก) แสดงการเปรียบเทียบระหว่างการปรับปรุงคิวรีด้วยจีนีติกอัลกอริทึมกับการปรับปรุงคิวรีโดยการสุ่มเลือกเทอมจากเซต K โดยตรง คอลัมน์ที่ 2 แสดงคิวรีจากผู้ใช้ที่ป้อนให้กับระบบ คอลัมน์ที่ 3 แสดงเทอมที่เลือกจากจีนีติกอัลกอริทึม (ดูหัวข้อ 5.2) และคอลัมน์ที่ 4 แสดงเทอมที่ได้จากการสุ่มจากเซต K โดยตรง จากตารางจะพบว่าเทอมต่างๆในสองคอลัมน์ที่ 3 และคอลัมน์ที่ 4 จะต่างกันแม้จะได้มาจากการวัดค่าความเหมาะสมจากฟังก์ชัน Jaccard Coefficient เดียวกัน

<u>ทดสอบปรับปรุงคิวรีด้วยจีนีติกอัลกอริทึม</u> :

เมื่อผู้ใช้ป้อนคิวรีเป็น

Q={*Data mining , Genetic algorithms , Expert system*}

Chr_Q= {0010011000000}

คำตอบจากจีนีติกอัลกอริทึมได้เป็น

Chr_{op} ={1000001001010}

Q_{op}={ *artificial intelligent , genetic algorithm , intelligent agent , search engine* }

ดังนั้น

$$JC = \frac{\#\{0010011000000\} \cap \{1000001001010\}}{4+3-1} = \frac{1}{6} = 0.167$$

และได้คิวรีคำตอบ $Q_{new} = Q \cup Q_{op}$ คือ

Q_{new}= {***Data mining, Genetic algorithms, Expert system***, *artificial intelligent, intelligent agent, search engine*}

<u>ทดสอบปรับปรุงคิวรีด้วยการสุ่มเทอมจากเซต K โดยตรง</u>:

สุ่มเทอมในเซต K ที่ทำให้ค่าของ JC เท่ากับค่าที่ได้จากการใช้จีนีติกอัลกอริทึม

$$JC = \frac{\#\{0010011000000\} \cap \{1110001000000\}}{4+3-1} = \frac{1}{6} = 0.167$$

จะได้

Q_{op}= *{artificial intelligent , automated test generation , data mining, genetic algorithm}*

และได้คิวรีคำตอบ $Q_{new} = Q \cup Q_{op}$ คือ

Q_{new}= {***Data mining, Genetic algorithms, Expert system*** , *automated test generation}*

จากทั้งสองวิธีการจะพบว่าเทอมที่ได้จากการสุ่มมีโอกาสจะได้เทอมที่ไม่เหมาะสมเป็นไปได้มาก หรือกล่าวอีกนัยได้คือวิธีการสุ่มเทอมจากเซต K โดยตรงเป็นการเลียนแบบกระบวนการจีนีติกอัลกอริทึมโดยกำหนด Pc = 0 และกำหนด Pm = 1 ซึ่งการเลือกใช้วิธีนี้อาจจะไม่สามารถทำงานได้หรือเป็นอาจจะไปไม่ได้ในทางปฏิบัติ การปรับปรุงคิวรีเป็นปัญหาที่ค่อนข้างน่าสนใจเพราะถ้าคิวรีคำตอบที่ได้มีความกำกวมแล้วโอกาสที่ระบบจะให้ผลรายงานที่ไม่ตรงกับความสนใจของผู้ใช้นั้นเป็นไปได้สูง

ตารางที่ 6.1(ก) แสดงชุดคิวรีเหมาะสมที่ได้จากจีนีติกอัลกอริทึมและที่ได้จากการสุ่มเทอมโดยตรงจากเซต K และตารางที่ 6.1(ข) แสดงจำนวนผลรายงานและค่า Precision ของผลรายงานที่ได้จากคิวรีที่ป้อนโดยผู้ใช้และที่ได้จากคิวรีเหมาะสมจากทั้งสองวิธีเปรียบเทียบกัน

ลำดับ	คิวรีต้นฉบับ (q)	เทอมที่เลือกโดยจีนีติกอัลกอริทึม (q1)	เทอมที่เลือกจากการสุ่ม (q2)
1.	Data mining, Genetic algorithms, Expert system	artificial intelligent , genetic algorithm , intelligent agent , search engine	artificial intelligent , automated test generation , data mining, genetic algorithm
2.	Data mining, Genetic algorithms, Expert system	decision support system, expert system, genetic algorithm , intelligent agent	Data mining, expert system, natural language processing, information retrieval
3.	Data mining, Genetic algorithms, Expert system	artificial intelligent , decision support system, expert system, genetic algorithm , intelligent agent	automated test generation, Data mining, genetic algorithm, information retrieval, search engine, software test engineering process

ตารางที่ 6.1(ก) แสดงการทดสอบการปรุงยูสเซอร์คิวรี

ลำดับ	ออกผลรายงานโดยอ้างอิงจากคิวรี						ออกผลรายงานโดยอ้างอิงจากยูสเซอร์โปรไฟล์					
	จำนวนผลรายงาน			Precision (%)			จำนวนผลรายงาน			Precision (%)		
	q	q+q1	q+q2	q	q+q1	q+q2	q	q+q1	q+q2	q	q+q1	q+q2
1.	290	380	425	77.58	80.53	78.59	260	451	470	77.31	82.04	78.72
2.	290	392	471	77.58	79.85	78.13	260	480	533	77.31	81.04	77.49
3.	290	375	478	77.58	79.52	77.62	260	462	542	77.31	80.30	76.75

ตารางที่ 6.1(ข) แสดงการทดสอบการปรุงยูสเซอร์คิวรี

6.2 ทดสอบการเรียนรู้ของยูเซอร์โปรไฟล์

การเรียนรู้ของยูเซอร์โปรไฟล์แบ่งออกเป็น 2 กระบวนการคือ การค้นหาค่า ***CIV*** ที่เหมาะสม (ดูหัวข้อที่ 5.3.3) และการปรับปรุง *iff* ของเทอมต่างๆในยูเซอร์โปรไฟล์(ดูหัวข้อที่ 5.3.1) การทดสอบการเรียนรู้ของยูเซอร์โปรไฟล์ด้วยการทดลองสร้าง 3 ยูเซอร์โปรไฟล์ขึ้นมาทดสอบระบบ ดังแสดงในตารางที่ 6.2 ยูเซอร์โปรไฟล์ IA1 กับ IA2 มีความสนใจเหมือนกันคือ Intelligent Agent และยูเซอร์โปรไฟล์ JP มีความสนใจเรื่อง Java Programming ตอนเริ่มต้นผู้ใช้จะต้องลงทะเบียนกับระบบเพื่อสร้างยูเซอร์โปรไฟล์โดยการกรอกเทอมต่างๆเพื่อระบุความสนใจลงในแบบฟอร์มลงทะเบียน ผู้ใช้สามารถป้อนโฮมเพจตัวอย่างเพื่อสอนระบบ (Learning Agent) หรือกล่าวอีกนัยหนึ่งก็คือ ผู้ใช้สามารถป้อนกลับ Recommend URLs ได้ในขั้นตอนสร้างยูเซอร์โปรไฟล์ได้ เช่น ป้อน URL http://www.sun.com ให้กับยูเซอร์โปรไฟล์ JP เป็นต้น

ลำดับ	สาขาวิชาที่สนใจ
1.	Intelligent Agent(IA1)
2.	Intelligent Agent(IA2)
3.	Java Programming(JP)

ตารางที่ 6.2 ยูเซอร์โปรไฟล์ที่ใช้ทดสอบระบบ

6.2.1 ทดสอบค้นหา CIV ที่เหมาะสม

ระบบจะค้นหาค่า CIV ที่เหมาะสมทุกครั้งที่ผู้ใช้ออกจากระบบ (สิ้นสุดการทำงาน)โดยก่อนออกจากระบบมีการทำป้อนกลับ กระบวนการจีนีติกอัลกอริทึมจะค้นหาค่า CIV ที่เหมาะสมตามกระบวนการจากหัวข้อที่ 5.3.3 ตารางที่ 6.3 แสดงค่า ***CIV*** ที่เหมาะสมที่ได้จากจีนีติกอัลกอริทึมสำหรับการค้นหาแต่ละรอบ

ลำดับ	ค่า CIV ที่เหมาะสม
1	(8,7,8,5,4,3)
2	(8,7,6,5,4,2)
3	(8,8,6,5,3,2)
4	(8,7,7,5,3,1)
5	(8,7,7,5,3,1)

ตารางที่ 6.3 ค่า CIV ที่เหมาะสมสำหรับแต่ละรอบการทำงาน

6.2.2 ทดสอบปรับปรุงยูเซอร์โปรไฟล์

ระบบจะปรับปรุง iff_i และ w_i ของแต่ละเทอมในยูเซอร์โปรไฟล์ทุกๆครั้งที่มีการป้อนกลับ ค่า iff_i และ w_i จะถูกปรับปรุงไปตามข้อมูลใน recommended URL โดยใช้สมการ (5.6) และสมการ (5.7) ตารางที่ 6.4 แสดงค่า *iff* สำหรับการปรับปรุงค่าในแต่ละครั้งต่อการป้อนกลับ

ลำดับ	Recommend URLs	*iff*		
		เทอม	ก่อน	หลัง
1.	http://www.sun.com	Java	1	1+2=3
		J2EE	1	1+0.3=1.3
		Technology	1	1+0.5=1.5
		VM	1	1+0.1=1.1
		:	:	:
2.	http://klomp.org/gnujsp/	Java	3	3+7=10
		Jserv	1	1+0.2=1.2
		JSP	1	1+0.4=1.4
		Servlet	1	1+0.8=1.8
		:	:	:
3.	http://java.apache.org/	Java	10	10+0.4=10.4
		Html	1	1+0.4=1.4
		Http	1	1+0.1=1.1
		Web	1	1+0.2=1.2
		:	:	

ตารางที่ 6.4 การปรับปรุงยูเซอร์โปรไฟล์

6.3 ทดสอบค่า Precision ของระบบ

6.3.1 Normal CIV

ถ้าหากกำหนดค่า CIV = (1, 1, 1, 1, 1, 1) หรือกำหนดให้ทุกคลาสจากตารางที่ 3.1 มีค่าความสำคัญเท่ากันหรือกล่าวอีกนัยได้คือ พิจารณาเอกสาร HMTL เหมือนกับเอกสาร text ปกติโดยทำการถอดเอา HTML Tags ออกจากเอกสาร HTML แล้วพิจารณาตามรูปแบบ Traditional IR System แบบเดิม สมการ (6.1) แสดงการพิสูจน์ว่าถ้าหากกำหนดค่า CIV = (1, 1, 1, 1, 1, 1) แล้วจะทำให้การคำนวณ weight-term ตามนิยาม 3.6 เท่ากันกับสมการ (3.2) หรือวิธีการคำนวณ weight-

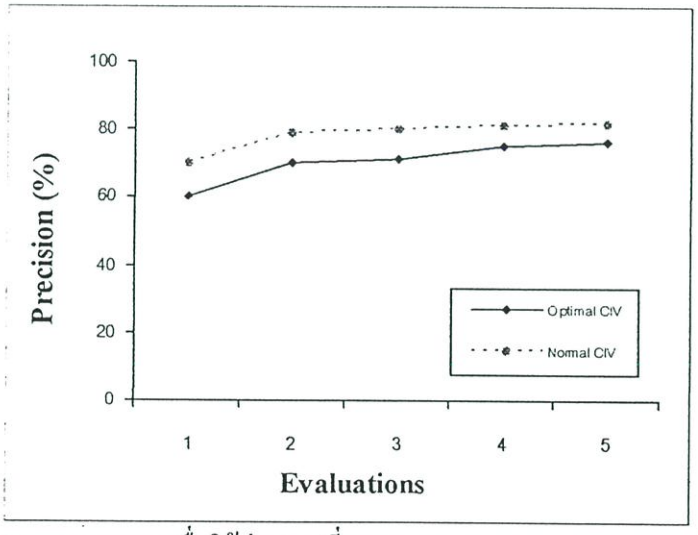
term ที่ผู้เขียนนำเสนอใหม่นี้จะมีความหมายเช่นเดียวกับ Traditional IR System แบบเดิมทุกประการ

$$
\begin{aligned}
w_{i,j} &= (TFV \bullet CIV)\, idf_i \\
w_{i,j} &= ((TFV) \bullet (1,1,1,1,1,1))\, idf_i \\
w_{i,j} &= (tfv_{i,j,c1} + tfv_{i,j,c2} + tfv_{i,j,c3} + tfv_{i,j,c4} + tfv_{i,j,c5} + tfv_{i,j,c6})\, idf_i \\
w_{i,j} &= \sum_{i=1..6} tfv_{i,j,ci} \times idf_j = tf_{i,j} \times idf_j
\end{aligned}
\qquad (6.1)
$$

เมื่อ • คือ Dot product

6.3.2 ทดสอบค่า Precision ของระบบเมื่อใช้ *CIV* ที่เหมาะสมเทียบกับ Normal *CIV*

ขั้นตอนการทดสอบจะเปรียบเทียบค่า Precision ของระบบระหว่างการนำค่า CIV ที่เหมาะสมมาใช้ในกระบวนการคำนวณ weight-term และการนำค่า Normal CIV มาใช้ในกระบวนการคำนวณ weight-term นำค่า Precision ของผลรายงานที่ได้จากทั้งสองกระบวนการมาเปรียบเทียบกัน รูปที่ 6.1 แสดงค่า Precision ของระบบเมื่อใช้ค่า CIV ที่เหมาะสมและเมื่อใช้ค่า Normal CIV



รูปที่ 6.1 ค่า Precision ของระบบเมื่อใช้ค่า CIV ที่เหมาะสมและ Normal CIV

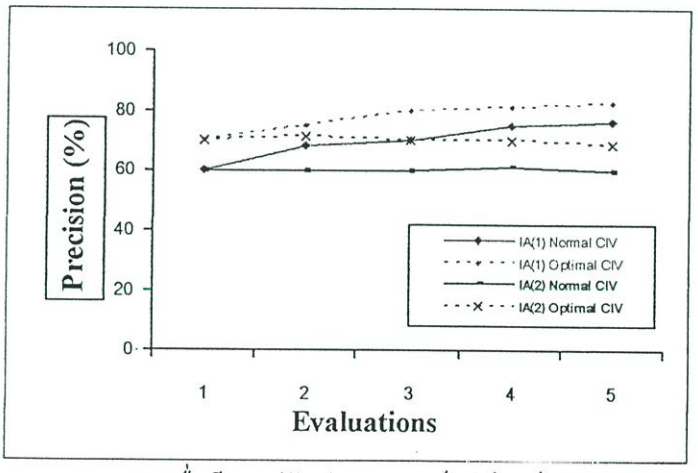
รูปที่ 6.1 ทดสอบใช้ค่า CIV ที่เหมาะสม (Optimal CIV) และ Normal CIV โดยสร้างยูสเซอร์โปรไฟล์ JP (ดูตารางที่ 6.2) กำหนดให้ขนาดสูงสุดของยูสเซอร์โปรไฟล์เป็น 50 เทอมโดยให้เจ้าของยูสเซอร์โปรไฟล์ใส่เทอมลงในยูสเซอร์โปรไฟล์จำนวน 25 เทอมและอีก 25 เทอมที่เหลือจะคัดเลือกเทอมที่ดีที่สุดจากโฮมเพจตัวอย่างที่ระบุในตอนเริ่มต้นสร้างยูสเซอร์โปรไฟล์ และ/หรือคัดเลือกมาจาก recommended URL ที่ป้อนกลับมายังระบบ รูปที่ 6.1 แกนตั้งแสดงค่า Precision คิด

เป็นเปอร์เซ็นต์และแกนนอนคือจำนวนครั้งที่ทดสอบโดยที่การทดสอบแต่ละครั้งจะต้องทำการป้อนกลับทุกครั้ง จากผลการทดลองพบว่าเมื่อใช้ค่า CIV ที่เหมาะสมหรือเมื่อพิจารณาโครงสร้างของเอกสาร HMTL จะทำให้ค่า Precision ของระบบดีกว่าการใช้ Normal CIV หรือพิจารณาตาม Traditional IR System แบบเดิม

เนื่องจากเอกสาร HTML เป็นเอกสารที่มีโครงสร้างที่สามารถสื่อความหมายได้เช่น จากตารางที่ 6.3 เทอมที่อยู่ในคลาส Title คลาส Header และคลาส Anchor (ลิงค์) จะได้รับความสนใจจากระบบเป็นพิเศษหรือให้ค่าความสำคัญมากกว่าเทอมที่อยู่ในคลาสอื่นๆ ในขณะที่ Tradition IR System จะให้ความสำคัญทุกๆเทอมในเอกสารเท่ากันหมด ดังนั้นโอกาสที่เทอมที่มีอิทธิพลต่อเอกสารน้อยแต่กลับได้รับความสนใจจากระบบมากกว่าเทอมที่มีอิทธิพลมากกว่านั้นเป็นไปได้สูง ด้วยสาเหตุนี้ทำให้ระบบอาจจะตัดสินใจผิดพลาดในการออกผลรายงานได้ ซึ่งเป็นเหตุให้ค่า Precision ของระบบจากการใช้ Normal CIV จึงมีค่าน้อยกว่าการใช้ค่า CIV ที่เหมาะสม

6.3.3 ทดสอบค่า Precision ของระบบเมื่อมีการปรับปรุงยูสเซอร์โปรไฟล์

ขั้นตอนการทดสอบจะเปรียบเทียบค่า Precision ของระบบเมื่อมีการปรับปรุงยูสเซอร์โปรไฟล์จากข้อมูลป้อนกลับด้วยสมการ (5.5) และสมการ (5.6) เปรียบเทียบกับเมื่อไม่มีการปรับปรุงยูสเซอร์โปรไฟล์



รูปที่ 6.2 ค่า Precision ของระบบเมื่อมีการปรับปรุงยูสเซอร์โปรไฟล์

รูปที่ 6.2 แสดงค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ IA(1) และ IA(2) โดยยูสเซอร์โปรไฟล์ทั้งสองเหมือนกันทุกประการ (ตอนเริ่มต้นสร้าง) เส้นกราฟไข่ปลา แสดงค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ทั้งสองเมื่อใช้ค่า CIV ที่เหมาะสม โดยที่ IA(1) มีการปรับปรุงยูสเซอร์โปรไฟล์จากข้อมูลป้อนกลับด้วยสมการ (5.5) และสมการ (5.6) ขณะที่ IA(2) ไม่มีการ

ปรับปรุงยูสเซอร์โปรไฟล์หรือไม่มีการป้อนกลับ จากการทดลองพบว่าการปรับปรุงยูสเซอร์โปรไฟล์จะที่ให้ค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ IA(1) เพิ่มขึ้นเป็น 82.04% ในขณะที่ให้ค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ IA(2) อยู่ที่ 70% *กราฟเส้นทึบ* แสดงค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ IA(1) และ IA(2) เมื่อใช้ค่า Normal CIV โดยที่ IA(1) มีการปรับปรุงยูสเซอร์โปรไฟล์จากข้อมูลป้อนกลับด้วยสมการ (5.5) และสมการ (5.6) ขณะที่ IA(2) ไม่มีการปรับปรุงยูสเซอร์โปรไฟล์หรือไม่มีการป้อนกลับ จากการทดลองพบว่าการปรับปรุงยูสเซอร์โปรไฟล์จะที่ให้ค่า Precision ของระบบบนยูสเซอร์โปรไฟล์ IA(1) เพิ่มขึ้นเป็น 78% ในขณะที่ค่า Precision ของยูสเซอร์โปรไฟล์ IA(2) อยู่ที่ 60%

สรุปผลการทดลองคือ เมื่อระบบใช้ค่า CIV ที่เหมาะสมและมีการปรับปรุงยูสเซอร์โปรไฟล์อย่างสม่ำเสมอแล้วจะทำให้ประสิทธิภาพของระบบดีขึ้นเพราะเมื่อระบบสามารถเรียนรู้ความสนใจของผู้ใช้แต่ละคนได้ แล้วระบบจะสามารถออกผลรายงานได้ตรงกับความต้องการของผู้ใช้มากขึ้น ทำให้ค่า Precision ของระบบจะเพิ่มขึ้น ตามลำดับ

บทที่ 7
สรุปผลการวิจัยและข้อเสนอแนะ

7.1 สรุปผลการวิจัย

การประยุกต์ระบบ IR มาใช้บน WWW มีขั้นตอนหลักๆอยู่ 2 ขั้นตอน

1. การสร้างอินเด็กซ์ คือ ใช้ spider วิ่งท่องไปบน WWW เพื่อดาวน์โหลดเอกสาร HTML สร้างอินเด็กซ์ และเก็บเอกสารที่ได้ลงในฐานข้อมูล ในงานวิจัยนี้ได้สำเนาข้อมูลจาก Web Search engine ซึ่งเป็นเอกสารที่ถูกสร้างอินเด็กซ์ไว้เรียบร้อยแล้ว
2. การสร้างเครื่องมือค้นหา คือ ส่วนที่รับคำค้นจากผู้ใช้แล้วดึงข้อมูลที่เกี่ยวข้องกับชุดคิวรีออกมาจากฐานข้อมูลและออกผลรายงาน งานวิจัยนี้ได้สร้างระบบส่วนนี้โดยใช้จีนีติกอัลกอริทึมเข้ามาช่วยปรับปรุงคิวรี และใช้ยูสเซอร์โปรไฟล์กับค่า CIV ที่เหมาะสมสำหรับปรับปรุงการออกผลรายงานของเครื่องมือค้นหา

จากผลการทดลองการใช้จีนีติกอัลกอริทึมมาช่วยในการปรับปรุงคิวรีให้เหมาะสม และการใช้ยูสเซอร์โปรไฟล์กับค่า CIV ที่เหมาะสมมาช่วยในการปรับปรุงการออกผลรายงานจะทำให้ระบบมีค่า Precision เพิ่มขึ้นเป็น 82.04%

7.2 ข้อเสนอแนะ

ปัญหาการทดสอบค่า Precision ของระบบเป็นปัญหาที่ค่อนข้างยาก เนื่องจากข้อมูลบน WWW เพราะความไม่มีมาตรฐานของข้อมูล คือไม่ทราบจำนวนเอกสารทั้งหมด ไม่ทราบจำนวนของเอกสารที่อยู่ในเซต R (ดูรูปที่ 4.1) ผู้ใช้เป็นคนตัดสินใจเองว่าเอกสารใดบ้างที่อยู่ในเซต Ra อย่างไรก็ตามการทดสอบประสิทธิภาพของระบบ IR นั้นสามารถกระทำบนฐานข้อมูลจำลอง TREC[10] ซึ่งเป็นฐานข้อมูลมาตรฐานที่ใช้ทดสอบประสิทธิภาพของระบบ IR ได้ ซึ่งกำลังอยู่ในช่วงดำเนินการจัดซื้อ โอกาสต่อไปผู้เขียนจะทดสอบระบบที่ได้นำเสนอในวิทยานิพนธ์นี้ด้วยฐานข้อมูล TREC ต่อไป

บรรณานุกรม

[1] Goldberg, D., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley, Reading, Massachusetts, USA.

[2] Chen, H., Chung, Y., Ramsey, M., Yang, C., Ma, P., Yen, J., 1997. Intelligent Spider for Internet Searching, Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences, Maui, Hawaii, USA.

[3] M. Cutler, H. Deng, S.S. Maniccam, and W. Meng. A new study on using html structures to improve retrieval. In Proceedings of the Eleventh IEEE Conference on Tools with Artificial Intelligence, pages 406-409, 1999.

[4] Bangorn, K., 2000. Online information retrieval using genetic algorithms, M.Sc. Thesis, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand.

[5] Zacharis Z., Panayiotopoulos, T., *"Web Search Using a Genetic Algorithm,"* IEEE Internet Computing, March-April 2001, pp. 18-26.

[6] G. Salton. Associative document retrieval technique using bibliographic information., Journal of the ACM, 10(4):440-457, October 1963.

[7] G. Salton and M. E. Lesh. Computer Evaluation of indexing and text processing., Journal of the ACM, 15(1):8-36 January 1968.

[8] W. B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structure & Algorithms., Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[9] K. Bharat and A. Z.. Broder. A technique for measuring the relative size and overlap of public Web search engines. In 7th WWW Conference, pages 379-388, Brisbane, Australia, 1998.

ประวัติผู้เขียน

ไพฑูรย์ ศรีนิล เกิดเมื่อวันที่ 25 มีนาคม 2519 ภูมิลำเนา 28 หมู่ 1 ต. ต้นยวน อ. พนม จ. สุราษฎร์ธานี มีประวัติการศึกษาดังนี้

1. วุฒิประถมศึกษา โรงเรียนวัดปากตรัง ต. ต้นยวน อ. พนม จ. สุราษฎร์ธานี
2. วุฒิมัธยมต้น โรงเรียนบ้านตาขุนวิทยา อ. บ้านตาขุน จ. สุราษฎร์ธานี
3. วุฒิ ปวช. อิเล็กทรอนิกส์ วิทยาลัยเทคนิคสุราษฎร์ธานี อ. เมือง จ. สุราษฎร์ธานี
4. วุฒิ ปวส. คอมพิวเตอร์ วิทยาลัยเทคนิคท่าหลวงซิเมนต์ไทยอนุสรณ์ ต. บ้านหมอ อ. บ้านหมอ จ. สระบุรี
5. วุฒิ วศ.บ. วิศวกรรมคอมพิวเตอร์ สถาบันเทคโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เขต. ลาดกระบัง กรุงเทพฯ
6. ปัจจุบันศึกษา ระดับปริญญาโท ภาควิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เขต. ลาดกระบัง กรุงเทพฯ