การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ

COLLABORATIVE FILTERING WITH NAÏVE BAYES WEIGHING

กรรัฏฐ์ หล่อวิทยาลิศนภา
KORRANAT LORWITTAYALERTNAPA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2554
KMITL-2011-SC-M-002-048

การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ

Collaborative Filtering with Naïve Bayes Weighing

กรนัฏฐ์ หล่อวิทยาเลิศนภา

KORRANAT LORWITTAYALERTNAPA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2554

KMITL-2011-SC-M-002-048

Collaborative Filtering with Naïve Bayes Weighing

KORRANAT LORWITTAYALERTNAPA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2011

KMITL-2011-SC-M-002-048

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ
(Collaborative Filtering with Naive Bayes Weighing)

นักศึกษา นายกรนัฏฐ์ หล่อวิทยาเลิศนภา

รหัสประจำตัว 50067501

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.วีระ บุญจริง

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ผศ.ดร.จีรพร	วีระพันธุ์	
ผศ.ดร.ศรัณย์	อินทโกสุม	
ดร.เฉลิมศักดิ์	เลิศวงศ์เสถียร	
รศ.ดร.วีระ	บุญจริง	

วัน / เดือน/ ปี ที่สอบ 9 มีนาคม พ.ศ. 2554 เวลา 17.00 – 19.00 น.

สถานที่สอบ ณ ห้อง 316 ชั้น 3 อาคารปฏิบัติการใหม่

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)

คณบดีคณะวิทยาศาสตร์

วันที่ 31 **เดือน** มี.ค. **พ.ศ.** 54

หัวข้อวิทยานิพนธ์	การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ
นักศึกษา	นายกรนัฏฐ์ หล่อวิทยาเลิศนภา
รหัสประจำตัว	50067501
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2554
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.วีระ บุญจริง

บทคัดย่อ

การกรองร่วมเป็นเทคนิคอย่างหนึ่งในระบบช่วยแนะนำ ซึ่งเป็นระบบที่ช่วยแนะนำสินค้าหรือบริการให้กับลูกค้าในระบบพาณิชย์อิเล็กทรอนิกส์ซึ่งมักมีฐานข้อมูลขนาดใหญ่ทำให้ลูกค้ายากที่จะเข้าถึงข้อมูลสินค้าได้ทั้งหมดเพื่อทำการตัดสินใจซื้อสินค้าหรือบริการได้ ดังนั้นเพื่อให้ง่ายต่อลูกค้าในการเข้าถึงข้อมูลที่ลูกค้าต้องการจึงได้มีการคิดค้นระบบช่วยแนะนำขึ้น ซึ่งงานวิจัยนี้นำเสนอวิธีการทำนายความนิยมในระบบช่วยแนะนำด้วยการกรองร่วม โดยนำเสนอวิธีทางเลือกโดยใช้ทฤษฎีความน่าจะเป็นเบส์แบบสามัญ ร่วมกับการแปลงลาปลาซแก้ไขปัญหาค่าศูนย์ของความน่าจะเป็นเบส์ที่ได้ มาใช้ในการถ่วงน้ำหนักค่าความนิยมและใช้ผลรวมเชิงเส้นของค่าที่ได้เป็นตัวทำนายความนิยม ทดลองด้วยชุดข้อมูลทดสอบสองชุด ผลลัพธ์ที่ได้พบว่าวิธีที่นำเสนอมีประสิทธิภาพที่ดีในการทำนายความนิยม และได้ผลลัพธ์ที่ดีมากขึ้นเมื่อทดลองกับชุดข้อมูลที่มีขนาดใหญ่กว่า

Thesis	Collaborative Filtering with Naïve Bayes Weighing
Student	Mr. Korranat Lorwittayalertnapa
Student ID	50067501
Degree	Master of Science
Program	Computer Science
Year	2011
Thesis Advisor	Assoc. Prof. Dr. Veera Boonjing

ABSTRACT

Collaborative filtering is a technique that is used for recommender system which is a system for recommending products or service to customers that confront with a large amount of data on e-commerce system. The difficulty is that customers might not find what they are looking for before making the decision. Therefore, recommender system was invented to make such that issue easier. This paper proposes a choice of preferences prediction method in collaborative filtering system using naïve Bayes theorem with Laplace smoothing for weighing preference. Then, linear combination is used to combine all weighed preference as preference predicted. The experiment is based on 2 data sets. The results show that using naïve Bayes weighing with Laplace smoothing is effective to be used as a predictor. Moreover, the results from the bigger data set outperform another one.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้มิอาจจะสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำชี้แนะ และคำแนะนำจาก รศ.ดร.วีระ บุญจริง ผู้เป็นอาจารย์ที่ปรึกษา ที่ปลูกฝังและสั่งสอนแนวทางการทำวิจัยมาตั้งแต่วันแรกที่ได้มาเป็นนักศึกษาของที่นี่ ท่านได้เอาใจใส่กับนักศึกษาทุกคนเป็นอย่างดี ข้าพเจ้าขอขอบพระคุณท่านอาจารย์เป็นอย่างสูง

ขอขอบพระคุณ ผ.ศ.ดร.ศรัณย์ อินทโกสุม ผศ.ดร.จีรพร วีระพันธ์ และดร.เฉลิมศักดิ์ เลิศวงศ์เสถียร คณะกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำตลอดจนข้อชี้แนะจนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ขอขอบพระคุณบิดา มารดา ที่ส่งเสริมข้าพเจ้าในด้านการเรียนเสมอมา อีกทั้งยังดูแลข้าพเจ้าเป็นอย่างดีในทุก ๆ เรื่อง และคอยเป็นกำลังใจให้ข้าพเจ้าเป็นอย่างดียิ่ง

หวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์ให้กับนักวิจัยและผู้ที่สนใจไม่มากก็น้อย สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพยิ่ง ตลอดจน ญาติพี่น้องและเพื่อนๆ ทุกคน ข้าพเจ้ามีความซาบซึ้งในความกรุณาอันดียิ่งจากทุกท่านที่ได้กล่าวนามมา และขอกราบขอบพระคุณมา ณ โอกาสนี้

กรนัฏฐ์ หล่อวิทยาเลิศนภา

พฤษภาคม 2554

สารบัญ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ระบบช่วยแนะนำ (Recommender System) เป็นกระบวนการวิธีการเรียนรู้ของระบบจากพฤติกรรมจากเลือกซื้อสินค้าหรือใช้บริการของผู้ใช้ในระบบ เพื่อช่วยแนะนำสินค้าหรือบริการที่เหมาะสมหรือน่าสนใจให้กับผู้ใช้อื่น ๆ ภายในระบบ ซึ่งปัจจุบันกลายมาเป็นส่วนสำคัญของระบบธุรกิจพานิชย์อิเล็กทรอนิกส์ (E-Commerce) เนื่องจากฐานข้อมูลของสินค้าหรือบริการนั้นมักมีขนาดใหญ่ และผู้ใช้ไม่สามารถจะเข้าถึงข้อมูลสินค้าและบริการทั้งหมดภายในระบบก่อนการตัดสินใจได้ ดังนั้นเพื่อเป็นการลดเวลาที่ใช้ในการตัดสินใจ และเพื่อเป็นการสร้างความประทับใจให้กับลูกค้า จึงมีการนำเอาระบบช่วยแนะนำเข้ามาใช้เพื่อช่วยให้ลูกค้าสามารถเข้าถึงข้อมูลที่ตัวเองสนใจได้ง่ายมากขึ้น

งานวิจัยมีขึ้นเพื่อพัฒนาขั้นตอนวิธีทางเลือก เพื่อใช้ในการทำนายความนิยม (Preference Prediction) ของผู้ใช้ที่มีกับตัวสินค้าในระบบ จากเดิมการทำนายความนิยมของผู้ใช้ที่มีกับตัวสินค้านั้นจะใช้การวัดความคล้าย (Similarity Measurement) ของพฤติกรรมการให้ความนิยมของผู้ใช้ในระบบเปรียบเทียบกับผู้ใช้คนอื่นในระบบ และเลือกผู้ใช้คนที่มีพฤติกรรมการให้ความนิยมใกล้เคียงกันมากที่สุด มาคำนวณค่าทำนายความนิยมให้กับผู้ใช้คนนั้น ๆ ส่วนวิธีที่ใช้ในงานวิจัยนี้ใช้หลักการคิดที่ว่า สินค้าที่ดีและมีคุณภาพก็ย่อมจะมีความน่าจะเป็นที่จะได้รับความนิยมจากผู้ใช้สินค้ามากกว่าสินค้าที่ไม่มีคุณภาพ โดยไม่จำเป็นต้องไปเปรียบเทียบกับผู้ใช้คนอื่น ๆ จึงได้นำมาพัฒนาด้วยหลักความน่าจะเป็น โดยนำเอาความน่าจะเป็นเบส์แบบสามัญเข้ามาประยุกต์ใช้กับการทำนายความนิยม เนื่องจากความน่าจะเป็นเบส์แบบสามัญนั้นมีการคำนวณที่เข้าใจได้ง่ายและเป็นที่ยอมรับอย่างแพร่หลาย

1.2 วัตถุประสงค์

งานวิจัยนี้นำเสนอการกระบวนการวิธีทางเลือกเพื่อใช้ในการคำนวณค่าความนิยมที่ผู้บริโภคมีต่อสินค้าหรือบริการในระบบช่วยแนะนำด้วยการกรองร่วม (Collaborative Filtering) ซึ่งจากเดิมที่ใช้แนวคิดของการถ่วงน้ำหนักค่าความนิยมด้วยค่าความคล้ายของผู้ใช้อื่นในระบบที่มีความคล้ายกันมากที่สุด เปลี่ยนมาใช้ทฤษฎีของความน่าจะเป็นของผู้ใช้คนหนึ่งที่จะให้ความนิยมกับสินค้าตัวหนึ่งด้วยค่าความนิยมต่าง ๆ โดยคำนวณจากทฤษฎีความน่าจะเป็นเบส์แบบสามัญ แล้วนำค่าความน่าจะเป็นเหล่านั้นคำนวณเป็นค่าทำนายอีกครั้ง โดยใช้การจัดกลุ่มแบบเคมีน (K-Mean Clustering) และการแปลง

ลาปลาซ (Laplace Smoothing) เพื่อแก้ไขปัญหาของสินค้าใหม่ในระบบที่ทำให้ไม่สามารถทำนายค่าความนิยมได้ ซึ่งผลที่ได้น่าจะมีประสิทธิภาพที่ดีกว่ากับการคำนวณค่าทำนายแบบดั้งเดิม

1.3 ขอบเขตงานวิจัย

งานวิจัยนี้ทำนายความคิดเห็นของผู้บริโภคในระบบช่วยแนะนำโดยเทคนิคการกรองร่วมโดยคำนวณความน่าจะเป็นจากทฤษฎีความน่าจะเป็นเบส์แบบสามัญ (Naïve Bayes Theorem) ที่ผู้ใช้คนหนึ่งที่จะให้ความนิยมกับสินค้าตัวหนึ่งด้วยค่าความนิยมต่าง ๆ ที่มีในระบบ จากนั้นใช้ค่าที่ได้มาถ่วงน้ำหนักของค่าความนิยมนั้น ๆ โดยมีการใช้การจัดกลุ่มแบบเคมีนและการแปลงลาปลาซ เพื่อแก้ไขปัญหาสำหรับสินค้าที่เพิ่งเข้ามาในระบบครั้งแรกซึ่งทำให้ค่าความน่าจะเป็นที่ระบบคำนวณได้เป็นศูนย์ เปรียบเทียบประสิทธิภาพที่ได้จากขั้นตอนวิธีทางเลือกนี้ กับขั้นตอนวิธีดั้งเดิม โดยการทดลองกับชุดข้อมูลสองชุด คือ Movie Lens ซึ่งเป็นชุดข้อมูลขนาดเล็ก และ Book Crossing ซึ่งเป็นชุดข้อมูลขนาดใหญ่ วัดประสิทธิภาพในการทำนายด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error)

1.4 ส่วนประกอบของวิทยานิพนธ์

ส่วนที่เหลือของวิทยานิพนธ์นี้ประกอบด้วยบทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้องซึ่งจะเสนอทฤษฎีและงานวิจัยต่างๆ ที่เกี่ยวข้องกับระบบช่วยแนะนำ เสนอรายละเอียดของเทคนิคต่าง ๆ ที่นิยมใช้ในระบบช่วยแนะนำ บทที่ 3 นำเสนอรายละเอียดและขั้นตอนวิธีการทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ บทที่ 4 นำเสนอรายละเอียดของชุดข้อมูลในการทดลอง วิธีการทดลอง และผลการทดลอง และบทที่ 5 เป็นการสรุปและเสนอแนะเกี่ยวกับงานวิจัย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะเสนอทฤษฎีและงานวิจัยที่เกี่ยวข้องที่ใช้ในการทำนายความนิยมจากโดยวิธีกรองร่วมด้วยทฤษฎีความน่าเป็นเบส์แบบสามัญ ซึ่งจะแบ่งเนื้อหาออกเป็นสองส่วนคือ ส่วนของทฤษฎีที่เกี่ยวข้อง และนำเสนองานวิจัยต่างๆ ที่เกี่ยวข้อง

2.1 ระบบช่วยแนะนำ

กิจการค้าในโลกยุคปัจจุบันนี้ไม่ได้มีเพียงการเดินจับจ่ายใช้สอยกันตามห้างสรรพสินค้า ร้านค้าหรือตามตลาดนัดทั่วไปเพียงเท่านั้น เพราะการถือกำเนิดของระบบเครือข่ายที่เชื่อมต่อกันระหว่างเครื่องคอมพิวเตอร์ด้วยกันหรือที่เรียกกันว่าอินเตอร์เน็ต (Internet) นั้นได้เข้ามามีบทบาทสำคัญอย่างยิ่งในเรื่องธุรกิจการค้า เนื่องจากทำให้การติดต่อสื่อสารกันระหว่างผู้ผลิตหรือผู้ค้าขายกับผู้บริโภคหรือลูกค้านั้นทำได้อย่างง่ายดายไร้ข้อจำกัดทางระยะทางอีกต่อไป จึงทำให้เกิดธุรกิจการค้าในอีกรูปแบบหนึ่งเกิดขึ้นเรียกว่า ธุรกิจพาณิชย์อิเล็กทรอนิกส์ (E-Commerce) ซึ่งเป็นที่นิยมสูงขึ้นเรื่อย ๆ เนื่องมากจากการที่ธุรกิจพาณิชย์อิเล็กทรอนิกส์สามารถเข้าถึงกลุ่มเป้าหมายได้กว้างขวางและรวดเร็ว ช่วยให้ผู้ค้าประหยัดค่าใช้จ่ายในการเปิดร้านเนื่องจากค่าเช่าพื้นที่ในระบบอินเตอร์เน็ตนั้นถูกกว่าการเช่าพื้นที่ร้านค้าจริงเป็นอย่างมาก รวมทั้งช่วยประหยัดเวลาที่ผู้ค้าจะบริหารร้านค้าเพราะแค่เพียงมีคอมพิวเตอร์ที่สามารถต่อสัญญาณอินเตอร์เน็ตได้ก็สามารถเข้ามาบริหารจัดการร้านค้าได้ทันที ในมุมของผู้บริโภคนั้นก็ช่วยให้ประหยัดเวลาในการเลือกซื้อสินค้าและบริการเพราะสามารถเปรียบเทียบคุณสมบัติของสินค้าหรือเปรียบเทียบราคาระหว่างร้านค้าต่าง ๆ ได้อย่างรวดเร็ว เนื่องด้วยข้อดีหลากหลายประการนี้ ทำให้ธุรกิจพาณิชย์อิเล็กทรอนิกส์มีการขยายตัวขึ้นอย่างรวดเร็ว ทั้งในแง่จำนวนร้านค้าออนไลน์ที่มากขึ้นหรือในแง่จำนวนสินค้าและบริการของร้านค้าแต่ละแห่งก็มากขึ้นเช่นกัน ส่งผลให้ฐานข้อมูลของสินค้าและบริการมีขนาดใหญ่มากขึ้นเป็นเงาตามตัว ซึ่งในยุคแห่งข้อมูลข่าวสารที่มีมากมายมหาศาลนี้ ปัญหาที่สำคัญมากอย่างหนึ่งที่ผู้บริโภคนั้นจะต้องเผชิญก็คือ การที่ผู้บริโภคจะสามารถค้นหาข้อมูลสินค้าและบริการให้ตรงกับความต้องการให้ตรงกับสิ่งที่ผู้บริโภคนั้นสนใจจริง ๆ จากข้อมูลที่มีอยู่มากมายมหาศาลนั้นได้อย่างไร อีกทั้งกลุ่มผู้บริโภคแต่ละกลุ่มก็มีความต้องการที่หลากหลายและแตกต่างกัน ปัญหานี้จึงเป็นโจทย์ที่ท้าทายและถือเป็นการแข่งขันของร้านค้าแต่ละร้านที่จะสร้างความพึงพอใจให้กับผู้บริโภค เพื่อช่วงชิงส่วนแบ่งของตลาดออกมาให้ได้มากที่สุด ดังนั้นระบบช่วยแนะนำ (Recommender System) จึงถูกออกแบบมาเพื่อช่วยแนะนำสินค้าที่คาดว่าผู้บริโภคน่าจะสนใจให้ เพื่อลดเวลาที่ผู้บริโภค

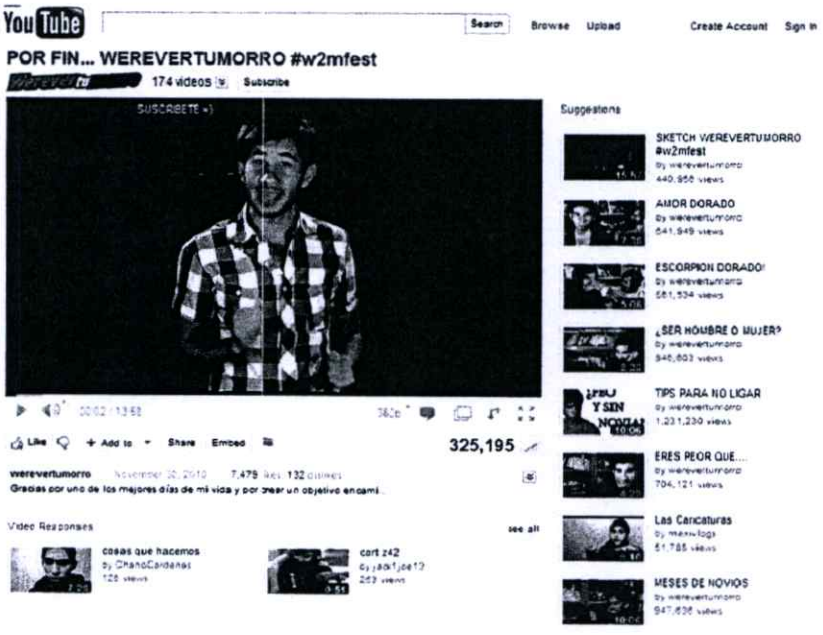
จะเข้าหาข้อมูลสินค้าและบริการทั้งหมดด้วยตนเอง และยังเป็นการสร้างความพึงพอใจให้กับผู้บริโภคอีกด้วย

ตัวอย่างของร้านค้าออนไลน์ที่เป็นที่รู้จักอย่างกว้างขวางและมีการใช้ระบบช่วยแนะนำมาใช้อย่างเป็นรูปธรรมอย่างเห็นได้ชัดเจนก็เช่น เว็บไซท์อเมซอนดอทคอม (www.amazon.com) ซึ่งเป็นเว็บไซท์ที่ขายสินค้าออนไลน์หลากหลายชนิดและเป็นที่รู้จักอับดับต้น ๆ ในโลกอินเตอร์เน็ตตัวอย่างของหน้าเว็บไซต์อเมซอนดอทคอมสามารถดูได้จากรูปที่ 2.1 ซึ่งในเว็บไซท์ดังกล่าวระบบช่วยแนะนำจะทำการแนะนำสินค้าที่ผู้บริโภคมักจะซื้อด้วยเสมอหากผู้บริโภคนั้นซื้อสินค้าชนิดใด ๆ ในเว็บไซท์นั้น ซึ่งเมื่อผู้บริโภคเข้าไปดูสินค้าชนิดหนึ่ง จะปรากฏแถบสินค้าทางด้านล่างของหน้าเว็บนั้น ๆ ดังตัวอย่างจากรูปที่ 2.2 ซึ่งเป็นการดึงข้อมูลจากระบบช่วยแนะนำเพื่อช่วยให้ลูกค้านั้นสามารถได้เห็นสินค้าที่ระบบคาดว่าลูกค้าที่กำลังเข้าชมสินค้าอยู่นั้นจะสนใจมากที่สุด

รูปที่ 2.1 ตัวอย่างของหน้าเว็บไซต์ www.amazon.com

รูปที่ 2.2 ตัวอย่างของส่วนระบบช่วยแนะนำจาก www.amazon.com

และอันที่จริงแล้วทุกวันนี้ไม่เฉพาะเพียงระบบธุรกิจพาณิชย์อิเล็กทรอนิกส์เท่านั้นที่นิยมใช้ระบบช่วยแนะนำ แต่เว็บไซท์ในลักษณะอื่น ๆ ก็เริ่มหันมาใช้ระบบนี้ในการดึงดูดผู้เข้าชมเว็บไซท์ให้เข้ามาชมหน้าเว็บไซท์ของตนเองเป็นจำนวนมาก ตัวอย่างเช่น ระบบช่วยแนะนำคลิปวิดีโอที่เกี่ยวข้องในเว็บไซท์ยูทูบดอทคอม (www.youtube.com) ซึ่งเป็นเว็บไซท์อันดับ 1 ในการรับชมหรือรับฝากไฟล์วีดิโอในปัจจุบัน ตัวอย่างของหน้าเว็บไซท์ยูทูบดอทคอมและส่วนของระบบช่วยแนะนำสามารถดูได้จากรูปที่ 2.3 และ 2.4 ตามลำดับ หรือระบบช่วยแนะนำเพื่อนที่คุณอาจรู้จักในเว็บไซท์สังคมออนไลน์ขนาดใหญ่ที่กำลังเป็นที่นิยมในปัจจุบันหรือเฟซบุคดอทคอม (www.facebook.com)



รูปที่ 2.3 ตัวอย่างของหน้าเว็บไซต์ www.youtube.com

ซึ่งเว็บไซท์แต่ละแห่งก็จะใช้ระบบช่วยแนะนำที่มีเทคนิคที่ใช้แตกต่างกันออกไปตามข้อมูลที่ระบบใช้พิจารณาให้เหมาะสมและเข้ากับคุณลักษณะของระบบมากที่สุด ระบบช่วยแนะนำที่คำนึงถึงพฤติกรรม ความคิดเห็น หรือรสนิยมของผู้ใช้ในระบบในการคำนวณเป็นหลักนั้น จะเรียกว่าระบบช่วยแนะนำด้วยวิธีการกรองร่วม (Collaborative Filtering: CF) ในอีกด้านหนึ่งระบบช่วยแนะนำที่คำนึงถึงคุณสมบัติของสินค้า หรือข้อความที่ใช้บรรยายลักษณะของตัวสินค้าในการแนะนำนั้น จะเรียกว่าระบบช่วยแนะนำโดยการกรองเนื้อหา (Content-based Filtering: CB) ซึ่งทั้งสองเทคนิคนี้ก็มีจุดดีและจุดด้อยแตกต่างกันไป การกรองเนื้อหาเน้นไปที่การดึงเอาข้อมูลที่ได้จากตัวเนื้อหาของสินค้าหรือบริการนั้นเข้ามาใช้ในการแนะนำ ส่วนการกรองร่วมใช้ข้อมูลความนิยมที่ผู้บริโภคให้ไว้ในระบบ นำไปคำนวณหาผู้บริโภคที่มีรสนิยมในการบริโภคสินค้าหรือใช้บริการต่าง ๆ ในระบบนี้ที่เหมือนหรือคล้ายกันมากที่สุด

และแนะนำสินค้าที่ผู้บริโภคคนหนึ่งสนใจร่วมกันให้กับผู้บริโภคอีกคนหนึ่ง นอกจากนี้ยังมีระบบช่วยแนะแบบผสม (Hybrid Recommender System) ซึ่งรวมเอาคุณลักษณะที่ดีของทั้งสองวิธีนี้เข้าไว้ด้วยกันอีกด้วย ในส่วนของงานวิจัยในครั้งนี้จะเน้นที่เทคนิคการกรองร่วมเป็นสำคัญ ซึ่งรายละเอียดจะกล่าวในบทถัดไป

	i_1	i_2	...	i_m	...	i_M
u_1				$R_{1,m}$		
u_2						
⋮						
u_k	$R_{k,1}$			$R_{k,m}$?		$R_{k,M}$
⋮						
u_K				$R_{K,m}$		

ภาพที่ 2.4 ตัวอย่างตารางความนิยมของผู้ใช้ U_K ที่มีต่อสินค้า I_M ด้วยค่าความนิยม $R_{K,M}$

ในปีค.ศ.2005 Felicia Poe [5] ได้ทำการรวบรวมข้อดีข้อด้อยของแต่ละเทคนิคของระบบช่วยแนะนำเอาไว้ ซึ่งมีผู้ที่ให้คำนิยามเอาไว้หลากหลายด้วยกัน ดังนี้

2.2 ระบบช่วยแนะนำด้วยการกรองร่วม

ระบบช่วยแนะนำด้วยเทคนิคการกรองร่วมนั้นใช้การพิจารณาคะแนนความนิยมของผู้ใช้ที่มีต่อสินค้าในระบบ และนำไปพิจารณากับสินค้าที่ผู้ใช้ยังไม่ให้ความนิยมไว้ แต่มีความคล้ายกับสินค้าที่ผู้ใช้นั้นเคยใช้แล้ว เทคนิคที่นิยมกันมากในการกรองร่วมนี้คือการใช้ตัววัดความคล้าย (Similarity Measurement) ในการวัดความคล้ายของข้อมูลความนิยมของผู้ใช้แต่ละคนในระบบเพื่อใช้ในการถ่วงน้ำหนักของสินค้าตัวที่ต้องการจะทำนายความนิยม เพื่อใช้ในการทำนายความนิยม (Preference Prediction) ให้กับสินค้าตัวเดียวกันนั้นให้กับผู้ใช้อีกคนหนึ่ง ซึ่งอย่างที่ได้กล่าวไว้ในบทที่แล้วนั้นการวัดความคล้ายนี้จะต้องแปลงข้อมูลความนิยมในระบบให้สามารถวัดออกมาเป็นตัวเลขได้ และเนื่องจากผู้ใช้และสินค้าในระบบที่ถูกให้ความนิยมมีมากมายดังนั้นจึงทำให้อยู่ในรูปของเวกเตอร์ความนิยม และใช้ตัววัดความคล้ายแบบต่าง ๆ เพื่อคำนวณหาค่าความคล้าย จากนั้นใช้การพิจารณาย่านใกล้เคียงที่สุด K ตัว ในการเลือกผู้ใช้หรือสินค้าที่มีความคล้ายมากที่สุด K ตัวมาทำการถ่วงน้ำหนักด้วยค่าความคล้ายที่วัดได้และหาผลรวมเชิงเส้นของผู้ใช้หรือสินค้าเพื่อใช้เป็นค่าทำนาย จะเห็นได้ว่าการทำนายด้วยเทคนิควิธีนี้อาจจะอ้างอิงผู้ใช้ในการทำนาย หรืออ้างอิงสินค้าในการทำนาย ซึ่งเรียกว่าเทคนิคการกรองร่วมตามสินค้า(Item-based Collaborative Filtering) หรือเทคนิคการกรองร่วมตามผู้ใช้(User-based

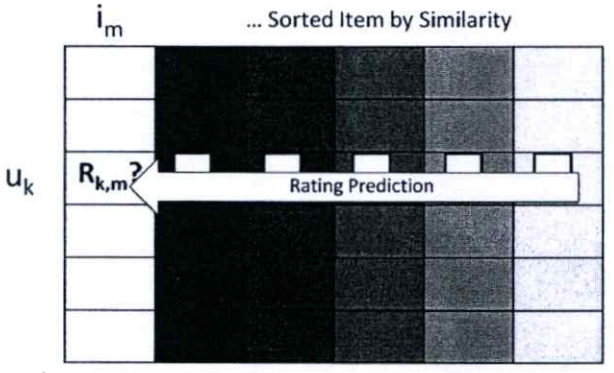
Collaborative Filtering) หรืออาจจะอ้างอิงทั้งสินค้าและผู้ใช้ไปพร้อมกันด้วยก็สามารถทำได้เช่นกัน [3],[9]

> *"เทคนิคการช่วยแนะนำโดยการกรองร่วมทำงานโดยการแนะนำสินค้าให้กับผู้ใช้โดยอ้างอิงจากผู้ใช้คนอื่น ๆ ที่มีความชอบสินค้าคล้ายกันมาก่อนหน้านี้ ระบบการกรองร่วมจะสร้างผู้ใช้ย่านใกล้เคียงสำหรับผู้ใช้แต่ละคนในระบบ และแนะนำสินค้ากับผู้ใช้คนใดคนหนึ่งถ้าผู้ใช้ย่านใกล้เคียงนั้นให้ความนิยมสูง"* (Torres 2004) [18]

> *"ระบบจะเก็บเอาฐานข้อมูลของคะแนนความนิยมของผู้ใช้แต่ละคนและค้นหาผู้ใช้คนอื่น ๆ ที่มีคะแนนความนิยมคล้ายคลึงอย่างมีนัยสำคัญกับผู้ใช้ที่เลือกไว้ จากนั้นจะแนะนำสินค้าอื่นที่ชื่นชอบโดยผู้ใช้อื่นที่คล้ายกับผู้ใช้ที่เลือกให้กับผู้ใช้นั้น"* (Mooney 2000) [14]

2.2.1 การกรองร่วมตามสินค้า

การกรองร่วมตามสินค้า (Item-based Collaborative Filtering) เป็นเทคนิคหนึ่งที่นิยมใช้ในระบบช่วยแนะนำโดยการพิจารณากลุ่มสินค้ากลุ่มหนึ่งที่ผู้ใช้คนหนึ่งให้คะแนนความนิยมเปรียบเทียบกับผู้ใช้อื่นในระบบที่เคยให้ความนิยมกับตัวสินค้าที่ต้องการจะทำนายเอาไว้ จากนั้นคำนวณค่าความคล้ายของชุดสินค้าที่ได้เป็นตัวถ่วงน้ำหนักและใช้ผลรวมเชิงเส้นของค่าถ่วงน้ำหนักทั้งหมดนั้นเป็นค่าความนิยมที่ระบบทำนาย

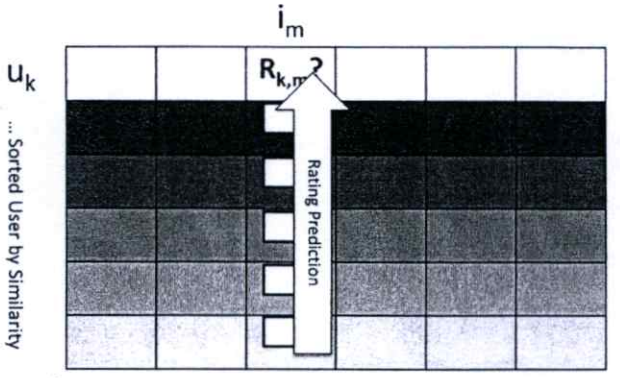


ภาพที่ 2.5 การทำนายความนิยมด้วยการกรองร่วมโดยอ้างอิงสินค้า

2.2.2 การกรองร่วมตามผู้ใช้

การกรองร่วมตามผู้ใช้ (User-based Collaborative Filtering) เป็นอีกเทคนิคหนึ่งที่ใช้ในระบบช่วยแนะนำ มีแนวคิดที่เหมือนกับการกรองร่วมตามสินค้าแต่จะพิจารณากลุ่มของผู้ใช้กลุ่มหนึ่งที่ให้ความ

นิยมต่อสินค้าตัวหนึ่งไว้ เปรียบเทียบกับสินค้าอื่นในระบบแทน ส่วนวิธีของการคำนวณนั้นเหมือนกับการกรองร่วมโดยอ้างอิงสินค้าทุกประการ



ภาพที่ 2.6 การทำนายความนิยมด้วยการกรองร่วมโดยอ้างอิงผู้ใช้

2.2.3 ข้อด้อยของเทคนิคการกรองร่วม

จุดด้วยที่สำคัญของเทคนิคนี้มีอยู่ด้วยกันสองวิธี ได้แก่ ปัญหาการเริ่มต้นยาก (Cold Start Problem) หรือปัญหาผู้ให้ความนิยมแรก (First Rater Problem)เนื่องจากเทคนิคการกรองร่วมนี้ใช้การคำนวณจากข้อมูลความนิยมที่มีอยู่ในระบบ แต่เมื่อมีผู้ใช้ใหม่เข้ามาในระบบ ระบบจะไม่สามารถแนะนำสินค้าใด ๆ ให้ได้เลย หรือในกรณีเดียวกับสินค้าที่เข้ามาในระบบครั้งแรกนั้นจะไม่ได้ถูกแนะนำให้กับผู้ใช้รายใดเลยเช่นกัน เนื่องจากไม่สามารถเปรียบเทียบความคล้ายระหว่างผู้ใช้หรือสินค้าที่เข้ามาใหม่กับผู้ใช้หรือสินค้าที่มีอยู่แล้วในระบบได้ และปัญหาที่สำคัญอีกอย่างหนึ่งก็คือ ปัญหาความเบาบาง (Sparsity Problem) คือการที่ข้อมูลความนิยมในระบบนั้นมีน้อยมาก ผู้ใช้นั้นอาจจะไม่ได้ให้ความนิยมในสินค้าแต่ละตัวมากเพียงพอหรือในทางกลับกันสินค้าก็ไม่ได้ถูกให้ความนิยมจากผู้ใช้ได้มากพอที่จะทำไปเปรียบเทียบความคล้ายซึ่งกันและกัน ทำให้การทำนายคะแนนความนิยมในระบบนั้นเป็นไปได้ยาก

> *"ปัญหาผู้ให้ความนิยมแรก (First-rater problem): สินค้าจำเป็นจะต้องถูกให้ความนิยมโดยผู้ใช้อย่างน้อย 1 คน มิเช่นนั้นแล้วสินค้าจะไม่สามารถถูกแนะนำได้จนกระทั่งจะมีใครมาให้ความนิยมกับสินค้านั้นก่อน"* (Torres 2004) [18]

> *"ปัญหาความเบาบาง (Sparsity problem) ในระบบที่มีข้อมูลจำนวนมาก ผู้ใช้คนหนึ่งดูเหมือนว่าจะให้คะแนนความนิยมสินค้านับเป็นเปอร์เซ็นต์เพียงเล็กน้อยถ้าเทียบสินค้าทั้งหมด สิ่งนี้ทำให้ยากที่จะหาความสอดคล้องในรสนิยมการบริโภคระหว่างของผู้ใช้แต่ละคนได้ เนื่องด้วยพวกเขาได้ให้คะแนนความนิยมกับสินค้าตัวเดียวกันเพียงเล็กน้อย"* (Torres 2004) [18]

"เมื่อระบบการกรองร่วมหนึ่งได้ถูกสร้างครั้งแรก มีสินค้ามากมายอยู่ในระบบ มีผู้ใช้เล็กน้อยในระบบ และไม่มีการให้ความนิยมใดเลย เมื่อปราศจากคะแนนความนิยมเหล่านั้นระบบไม่สามารถสร้างการแนะนำใด ๆ และผู้ใช้ก็จะไม่เห็นข้อดีของระบบ เมื่อปราศจากผู้ใช้ ก็ไม่มีทางที่จะมีความนิยมใดเข้าสู่ระบบ เมื่อประยุกต์ใช้ระบบการกรองร่วมแก่ระบบ ก็มีความจำเป็นที่จะต้องป้อนข้อมูลความนิยมเริ่มต้นเข้าสู่ระบบเสียก่อน"
(McNee 2002) [2]

2.2.4 ข้อดีของเทคนิคการกรองร่วม

สินค้าประเภทต่าง ๆ กันที่ยังไม่เคยถูกเลือกซื้อหรือเปิดดูโดยผู้ใช้มาก่อน ก็สามารถถูกแนะนำได้โดยการเปรียบเทียบกับผู้ใช้คนอื่น ๆ ที่เคยใช้สินค้านั้น ๆ และผู้ใช้นั้นมีรสนิยมการบริโภคสินค้าที่มีความคล้ายคลึงกันโดยวัดจากข้อมูลคะแนนความนิยมที่ผู้ใช้นั้นได้เคยให้ไว้กับสินค้าอื่น ๆ ในระบบ ซึ่งทำให้มีการแนะนำที่หลากหลาย

2.3 ระบบช่วยแนะนำด้วยการกรองเนื้อหา

เทคนิคการกรองเนื้อหาจะใช้ข้อมูลที่ได้จากตัวสินค้ามาพิจารณา โดยจะแนะนำสินค้าที่มีรายละเอียดเนื้อหาของตัวสินค้าที่เหมือนหรือคล้ายกับสินค้าที่ผู้ใช้เคยซื้อหรือใช้บริการให้กับผู้ใช้ ซึ่งจะใช้กับสินค้าที่มีเนื้อหาเป็นคำหรือข้อความ ที่สามารถเรียนรู้ด้วยกระบวนการทางคอมพิวเตอร์ได้เท่านั้น จุดด้อยของเทคนิคการกรองเนื้อหาจะอยู่ที่ข้อจำกัดของเทคนิคนี้ คือรายละเอียดหรือเนื้อหาของผู้ใช้และสินค้าที่จะสามารถใช้กับเทคนิคการกรองเนื้อหานั้น จะต้องเป็นเนื้อหาที่สามารถเรียนรู้ได้ด้วยระบบคอมพิวเตอร์ โดยใช้เทคนิคการดึงข้อมูล (Information Retrieval) เพื่อเปรียบเทียบแยกแยะความเหมือนและความแตกต่างระหว่างผู้ใช้หรือสินค้า แต่หากเป็นสินค้าที่เนื้อหานั้นขึ้นอยู่ความคิดเห็นเห็นของผู้ใช้เป็นหลัก เช่น เรื่องขำขันหรือเพลงต่าง ๆ เป็นต้น เทคนิคนี้จะไม่สามารถแยกแยะได้ว่าเรื่องขำขันเรื่องไหนตลกหรือสนุก หรือแยกแยะได้ว่าเพลงนั้นมีทำนองที่น่าฟังหรือถูกใจผู้ใช้อย่างไร และที่สำคัญอีกอย่างหนึ่งก็คือ สิ่งที่ได้จากระบบช่วยแนะนำแบบนี้จะมีความเฉพาะเจาะจงสูง (Overspecialization) กล่าวคือ เทคนิคนี้ทำให้ระบบไม่สามารถแนะนำสินค้าที่มีรายละเอียดเนื้อที่แตกต่างออกไปแก่ผู้ใช้ได้เลย และจะคำนวณจากสินค้าที่เคยผู้ใช้เคยใช้และให้ความนิยมไว้แล้วเท่านั้น ซึ่งทำให้ขอบเขตของการแนะนำนั้นแคบเกินไป

"หนึ่งในกลยุทธ์พื้นฐานที่ใช้ข้อความของสินค้าต่าง ๆ ที่ผู้ใช้ชื่นชอบนำมาสร้างเป็นข้อมูลคำสำคัญ และจากนั้นก็แนะนำสินค้นที่ตรงกับข้อมูลคำสำคัญนั้น เทคนิคการช่วย

แนะนำโดยการกรองเนื้อหานี้จะทำงานได้ดีเมื่อข้อมูลของสินค้านั้นสามารถเรียนรู้ในกระบวนการทางคอมพิวเตอร์ได้" (McNee 2002) [2]

2.3.1 ข้อด้อยของเทคนิคการกรองเนื้อหา

"แนวคิดของระบบช่วยแนะนำโดยการกรองเนื้อหาจะอยู่บนพื้นฐานของข้อมูลเชิงวัตถุเกี่ยวกับตัวสินค้า ข้อมูลเหล่านี้จะถูกดึงมาจากแหล่งข้อมูลที่หลากหลาย (เช่น เว็บไซท์) หรือดึงมาจากแหล่งข้อมูลของสินค้า (เช่น ฐานข้อมูลของสินค้า) ข้อมูลเชิงความคิดเห็นจะไม่นำมาใช้ในกระบวนการนี้" (Montaner 2003)[13]

"เทคนิคของระบบช่วยแนะนำโดยการกรองเนื้อหาจะไม่มีการสืบทอดวิธีการสร้างการค้นพบโดยบังเอิญ ระบบจะแนะนำมากขึ้นในสิ่งที่ผู้ใช้นั้นได้พบเห็นไปแล้วหรือบ่งบอกว่าคล้ายกัน" (Montaner 2003)[13]

2.3.2 ข้อดีของเทคนิคการกรองเนื้อหา

ข้อดีของระบบช่วยแนะนำโดยการกรองเนื้อหาก็คือ แม้ว่าจะเป็นสินค้าที่ยังไม่เคยถูกเลือกซื้อหรือเปิดดูหรือให้ความคิดเห็นโดยผู้ใช้มาก่อนหรือแม้กระทั่งเป็นสินค้าที่เพิ่งจะเข้ามาใหม่ในระบบ ระบบก็สามารถแนะนำสินค้าอื่นได้โดยการพิจารณาจากสินค้าที่มีเนื้อหาที่ใกล้เคียงกัน

2.4 ระบบช่วยแนะนำแบบผสม

ระบบช่วยแนะนำแบบผสมนี้เป็นการผสมผสานหลายเทคนิคของระบบช่วยแนะนำเข้าไว้ด้วยกัน โดยการนำจุดเด่นของแต่ละเทคนิคนั้น มาช่วยเสริมจุดด้อยของอีกเทคนิคหนึ่ง ซึ่งสองเทคนิคที่นิยมนำมาใช้ร่วมกันในระบบช่วยแนะนำแบบผสมนี้ ก็คือการเทคนิคกรองร่วมและการกรองเนื้อหา ซึ่งวิธีการผสมของแต่ละเทคนิคก็มีวิธีแตกต่างกันออกไปในปีค.ศ.2002 Robin Burke [16], [17] ได้รวมรวมเทคนิคที่ใช้ในระบบช่วยแนะนำแบบผสมเอาไว้ ดังนี้

2.4.1 ระบบช่วยแนะนำแบบผสมด้วยการถ่วงน้ำหนัก

ระบบช่วยแนะนำแบบผสมด้วยการถ่วงน้ำหนัก (Weighed Hybrid Recommender) คือระบบช่วยแนะนำที่จะรวมคะแนนความนิยมที่ทำนายได้จากแต่ละเทคนิคที่ใช้ด้วยผลรวมเชิงเส้น ซึ่งคะแนนที่นำมารวมกันนี้ต้องสามารถรวมได้ในระบบเชิงเส้น และควรจะมีความสอดคล้องกัน เช่น มีขอบเขตของค่าทำนายเหมือนกัน เป็นต้น

2.4.2 ระบบช่วยแนะนำแบบผสมด้วยการสับเปลี่ยน

ระบบช่วยแนะนำแบบผสมด้วยการสับเปลี่ยน (Switching Hybrid Recommender) คือจะทำการเลือกวิธีที่จะใช้ในการทำนายความนิยม ปัญหาของเทคนิคนี้จะอยู่ที่เกณฑ์ที่ใช้ในการตัดสินใจเลือกวิธีที่จะใช้ทำนายความนิยมระหว่างวิธีต่าง ๆ ซึ่งประสิทธิภาพของการทำนายก็จะต่างกันออกไปขึ้นอยู่กับสถานการณ์และกฎเกณฑ์ที่เราเลือกใช้

2.4.3 ระบบช่วยแนะนำแบบผสมด้วยการผสม

ระบบช่วยแนะนำแบบผสมด้วยการผสม (Mixed Hybrid Recommender) คือระบบช่วยแนะนำด้วยเทคนิคนี้จะคำนวณหาค่าความนิยมที่ทำนายได้จากแต่ละวิธีแยกกัน และมีกระบวนการที่รวมค่าความนิยมของทุกวิธีเข้าไว้ด้วยกัน ตัวอย่างง่าย ๆ อย่างเช่น การกรองร่วมให้ค่าความนิยมที่ 3 หน่วย ส่วนการกรองเนื้อทำนายความนิยมที่ 4 หน่วย ระบบช่วยแนะนำแบบผสมด้วยเทคนิคนี้ก็จะทำนายด้วยค่าความนิยม 7 เป็นต้น

2.4.4 ระบบช่วยแนะนำแบบผสมด้วยการรวมคุณลักษณะ

ระบบช่วยแนะนำแบบผสมด้วยเทคนิคการรวมคุณลักษณะ (Feature Combination Hybrid Recommender) คือระบบช่วยแนะนำที่นำเอาคุณลักษณะที่ได้จากระบบช่วยแนะนำหนึ่งเข้าไปเป็นข้อมูลเข้าในระบบช่วยแนะนำอีกวิธีหนึ่งเพื่อทำนายคะแนนความนิยมออกมานั่นเอง

2.4.5 ระบบช่วยแนะนำแบบผสมด้วยการเพิ่มเติมคุณลักษณะ

ระบบช่วยแนะนำแบบผสมด้วยเทคนิคการเพิ่มเติมคุณลักษณะ (Feature Augmentation Hybrid Recommender) คือเทคนิคในระบบช่วยแนะนำแบบผสมอีกอย่างหนึ่งที่คล้ายกับระบบช่วยแนะนำแบบผสมด้วยเทคนิคการรวมคุณลักษณะ เพียงแต่ว่าคุณลักษณะที่ใช้เป็นข้อมูลเข้านั้น จะเป็นคุณลักษณะแบบใหม่

2.4.6 ระบบช่วยแนะนำแบบผสมแบบลำดับชั้น

ระบบช่วยแนะนำแบบผสมตามลำดับชั้น (Cascade Hybrid Recommender) คือระบบที่มีการเลือกระบบหลักและระบบรองที่ใช้ในการทำนายค่าความนิยม ในกรณีที่การทำนายความนิยมของระบบช่วย

แนะนำที่เป็นส่วนประกอบในระบบทำนายได้ค่าที่เท่ากัน ระบบช่วยแนะนำที่เป็นระบบรองจะถูกใช้เป็นตัวปรับแต่งค่าทำนายจากระบบช่วยแนะนำหลัก

2.4.7 ระบบช่วยแนะนำแบบผสมระดับเมทา

ระบบช่วยแนะนำแบบผสมระดับเมทา (Meta-Level Hybrid Recommender) คือการที่ทั้งระบบช่วยแนะนำหลักและระบบรองนั้นจะทำการทำนายความนิยมเหมือนกัน แต่ข้อมูลที่ได้จากระบบใดให้ผลที่ดีกว่า จะถูกนำไปแทนที่ทันที

2.5 เทคนิคที่ใช้ในระบบช่วยแนะนำ

ระบบช่วยแนะนำได้นำเอาทฤษฎีและเทคนิคต่าง ๆ เข้ามาประยุกต์ใช้ให้เข้ากับแนวคิดของระบบช่วยแนะนำแต่ละระบบ ดังนี้

2.5.1 เทคนิคที่ใช้ในระบบช่วยแนะนำโดยการกรองเนื้อหา

ในปีค.ศ.2006 Joel Bennett [8] ได้ทำการรวมรวบเทคนิคที่ใช้ในระบบช่วยแนะนำด้วยการกรองเนื้อหาเอาไว้ ดังนี้

2.5.1.1 การกรองด้วยการค้นหา (Search as Filter)

เป็นวิธีการที่ง่ายที่สุดในการแนะนำโดยการให้ผู้ใช้ใส่คำสำคัญ (Key Word) ในการค้นหาและดึงรายชื่อข้อมูลที่มีความเกี่ยวข้องกับคำสำคัญเหล่านั้นในระบบออกมาทั้งหมด ซึ่งจะให้ค่าที่เหมือนเดิมออกมาทุกครั้งหากใส่คำสำคัญเหมือนเดิม และจะไม่ปรับแต่งข้อมูลที่ดึงออกมาตามผู้ใช้แต่ละบุคคลเลย ซึ่งเทคนิคนี้ห่างไกลจากคำนิยามของระบบช่วยแนะนำในปัจจุบันยิ่งนัก

2.5.1.2 การกรองด้วยการจัดกลุ่ม (Clustering as Filter)

เป็นเทคนิคหนึ่งที่นำมช่วยจัดกลุ่มข้อมูล โดยจะจัดกลุ่มข้อมูลตามเนื้อหาของข้อมูลนั้น โดยที่ผู้ใช้จำเป็นต้องให้ข้อมูลความนิยมในเบื้องต้นของตัวเองต่อระบบก่อน และระบบจะทำการแนะนำกลุ่มสินค้าให้ตรงกับข้อมูลความนิยมของผู้ใช้

2.5.1.3 TF-IDF

ในการการสืบค้นสารสนเทศ วิธีที่นิยมใช้กันมากที่สุดวิธีหนึ่งคือ TF-IDF ซึ่งแนวคิดก็คือ คำหรือวลีใด ๆ ที่ปรากฏอยู่ในข้อมูลรายละเอียดของสินค้ามากเท่าไรก็ยิ่งใช้คำหรือวลีนั้นในการอธิบายสินค้าได้ดีเท่านั้น แต่ในบางครั้งคำที่ใช้บ่อยที่สุดนั้นอาจจะปรากฏอยู่ในรายละเอียดของสินค้าได้หลายตัว

2.5.2 เทคนิคที่ใช้ในระบบช่วยแนะนำโดยการกรองร่วม

ในปีค.ศ. 2009 Xiaoyuan Su and Taghi M. Khoshgoftaar [19] ได้รวบรวมเทคนิคต่าง ๆ ที่ใช้ระบบช่วยแนะนำด้วยการกรองร่วมเอาไว้ ซึ่งสามารถแบ่งได้เทคนิคที่ใช้ออกเป็นประเภทใหญ่ ๆ ได้ 2 วิธี คือ การกรองร่วมโดยการอ้างอิงหน่วยความจำ (Memory-based Collaborative Filtering) และการกรองร่วมโดยการอ้างอิงแบบจำลอง (Model-based Collaborative Filtering)

2.5.2.1 การกรองร่วมโดยอาศัยหน่วยความจำ

การกรองร่วมโดยอาศัยหน่วยความจำ (Memory-based Collaborative Filtering) เป็นกระบวนการวิธีของระบบช่วยแนะนำที่ใช้ข้อมูลของระบบทั้งหมด หรือบางส่วนมาเป็นส่วนประกอบของการทำนายความนิยม ซึ่งข้อมูลความนิยมนั้นจะมาจากผู้ใช้หรือสินค้าที่ความเหมือนหรือคล้ายกันในการช่วยทำนาย หรืออาจจะเรียกอีกอย่างได้ว่าเป็นการใช้เทคนิคการอ้างอิงย่านใกล้เคียง (Neighborhood-based) ซึ่งมีขั้นตอนการคำนวณ โดยการคำนวณค่าความคล้ายของผู้ใช้หรือสินค้าที่จะทำนายกับผู้ใช้หรือสินค้าย่านใกล้เคียงเพื่อใช้ในการถ่วงน้ำหนัก จากนั้นคำนวณค่าทำนายโดยถ่วงน้ำหนักจากค่าจากตัววัดความคล้ายที่ได้ก่อนหน้านี้ ซึ่งตัววัดความคล้าย (Similarity Measurement) ก็สามารถคำนวณได้หลากหลายวิธี ตัวอย่างของตัววัดความคล้ายที่นิยมใช้ในงานวิจัยในระบบช่วยแนะนำด้วยการกรองร่วมมี ดังนี้

- ตัววัดความคล้ายแบบสหสัมพันธ์ (Correlation-Based Similarity)
- ตัววัดความคล้ายแบบโคซายน์ (Vector Cosine-Based Similarity)
- ตัววัดความคล้ายแบบโคซายน์ปรับแก้ (Adjusted Cosine Similarity)

2.5.2.2 การกรองร่วมโดยอาศัยแบบจำลอง (Model-based Collaborative Filtering)

การกรองร่วมโดยอาศัยแบบจำลอง (Model-based Collaborative Filtering) เป็นกระบวนการวิธีของระบบช่วยแนะนำที่ใช้ข้อมูลของระบบทั้งหมด หรือบางส่วนมาเป็นเป็นตัวที่ใช้สร้างการทำนายความนิยม ซึ่งข้อมูลความนิยมนั้นจะมาจากผู้ใช้หรือสินค้าที่ความเหมือน โดยจะมีการสร้างแบบจำลองจากข้อมูลความนิยมของผู้ใช้เพื่อให้ในการแนะนำ โดยใช้กลไกการเรียนรู้ในการสร้างแบบจำลองความนิยมของผู้ใช้ในระบบเข้ามา กลไกการเรียนรู้ดังกล่าวเช่น โครงข่ายแบบเบส์ (Bayesian network) การจัดกลุ่ม (Clustering) หรือระบบการสร้างกฎ (Rule-based System)

2.6 การจำแนกประเภทด้วยความน่าจะเป็นเบส์แบบสามัญ

การจำแนกประเภทด้วยความน่าจะเป็นเบส์แบบสามัญ ใช้แนวคิดจากทฤษฎีความน่าจะเป็นเบส์ในการจำแนกข้อมูลที่ความหลากหลายสมมติว่าข้อมูลที่จะทำการจำแนกนั้นประกอบด้วยข้อมูลของสัตว์ชนิดต่าง ๆ โดยมีส่วนของการอธิบายรายละเอียดเรื่องลักษณะภายนอกที่สามารถเห็นได้ เช่น จำนวนขาเท่าไร มีขนหรือไม่มีขน และกินพืชหรือสัตว์เป็นอาหาร เป็นต้น ถ้าเราจะพิจารณาสัตว์ที่มีสี่ขาและมีขน เราจะจำแนกประเภทให้สัตว์ที่เราจะพิจารณานั้นเป็นสัตว์ชนิดใด ความยากของการจำแนกแบบนี้จะอยู่ที่เมื่อข้อมูลย่อยของสิ่งที่เราจะพิจารณานั้นมีจำนวนมากและประเภทที่สามารถใช้จำแนกมีจำนวนมากขึ้นด้วย ซึ่งจำเป็นที่จะต้องใช้ข้อมูลเรียนรู้จำนวนมากไปด้วยเช่นกัน เพราะจำเป็นต้องเรียนรู้ถึงความขึ้นแก่กันของแต่ละเหตุการณ์ ทำให้การคำนวณนั้นซับซ้อนมากขึ้นตามจำนวนเหตุการณ์ที่พิจารณาก่อนนั้น ดังนั้นแนวคิดความน่าจะเป็นเบส์แบบสามัญจึงถูกคิดขึ้นมาเพื่อลดทอนความซับซ้อนที่เกิดขึ้นนี้

2.6.1 ทฤษฎีความน่าจะเป็นของเบส์

ทฤษฎีความน่าจะเป็นของเบส์ (Bayes Theorem) กำหนดให้ X คือข้อมูลที่ยังไม่ได้ถูกจำแนก P(H) คือสมมติฐานที่ข้อมูล X จะถูกจำแนกอยู่ในกลุ่ม C ก่อนที่จะทำการจำแนกประเภทของข้อมูลนั้นเราจะต้องพิจารณาค่า P(H|X) ซึ่งเป็นความน่าจะเป็นของเหตุการณ์ที่เป็นไปได้หลังทราบข้อมูลหรือความน่าจะเป็นโดยประสบการณ์ (Posterior Probability) ซึ่งหมายถึงความน่าจะเป็นที่สมมติฐาน H จะเกิดขึ้นเมื่อเกิดเหตุการณ์ X ตัวอย่างเช่น หากจะหาความน่าจะเป็นที่ชนิดของสัตว์ที่เราจะพิจารณาอยู่เป็นสุนัข โดยให้เงื่อนไขว่าชนิดของสัตว์ที่กำลังพิจารณาอยู่นั้นมีสี่ขาและมีขน ในทางกลับกันเราพิจารณา P(H) ซึ่งเป็นความน่าจะเป็นของเหตุการณ์ที่เป็นไปได้ก่อนทราบข้อมูลหรือความน่าจะเป็นโดยหลักเกณฑ์ (Prior Probability หรือ Apriori Probability) โดยคำนวณหาความน่าจะเป็นที่ชนิดของสัตว์ที่พิจารณานั้นคือสุนัข จะเห็นได้ว่าความน่าจะเป็นโดยประสบการณ์ P(H|X) จะขึ้นอยู่กับเหตุการณ์ X ที่เกิดขึ้นก่อน ส่วนความน่าจะเป็นโดยหลักเกณฑ์ P(H) นั้นจะเป็นอิสระจากเหตุการณ์ X และในทางเดียวกัน P(X|H) ก็เป็นความน่าจะเป็นเหตุการณ์ X ที่เกิดหลังเหตุการณ์ H ทฤษฎีความน่าจะเป็นของเบส์ สามารถคำนวณได้จากสมการ

$$p(H|X) = \frac{p(H)p(X|H)}{p(X)}$$

ในการคำนวณความน่าจะเป็นใช้ตัวแบบของทฤษฎีความน่าจะเป็นเบส์ข้างต้น หากมีเงื่อนไขเหตุการณ์ X หลายเหตุการณ์ที่เกิดขึ้นก่อน หรืออาจจะกล่าวได้ว่าเหตุการณ์ X ประกอบด้วยเหตุการณ์ $x_1, x_2, x_3, \ldots x_n$ เมื่อนำมาคำนวณหาความน่าจะเป็นตามสมการข้างต้นทำให้เกิดการคำนวณที่ซับซ้อน

มากขึ้น เนื่องจากการขึ้นต่อกันของข้อมูล (Dependence) ของการเกิดของเหตุการณ์แต่ละเหตุการณ์นั้นจะส่งผลต่อเหตุการณ์อื่นซึ่งกันและกัน ดังนั้นจึงได้มีการปรับปรุงตัวแบบความน่าจะเป็นนี้ให้ง่ายขึ้นในชื่อความน่าจะเป็นเบส์อย่างง่าย หรือความน่าจะเป็นเบส์แบบสามัญ

2.6.2 ตัวจำแนกประเภทเบส์แบบสามัญ

ตัวจำแนกประเภทเบส์แบบสามัญนั้นใช้ความน่าจะเป็นของเบส์จากสมมติฐานที่ว่าแต่ละเหตุการณ์ย่อยที่เกิดขึ้นนั้นเป็นอิสระต่อกันซึ่งสามารถคำนวณหาความน่าจะ P(H|X) ได้จากสมการ

$$p(H|X) = p(H) * p(H|x_1) * p(H|x_2) * \dots * p(H|x_n)$$

เมื่อ $X = \{x_1, x_2, x_3, \dots x_n\}$ คือเหตุการณ์จะใช้ในการจำแนกประเภท

n คือจำนวนเหตุการณ์ย่อยทั้งหมด

$H = \{c_1, c_2, c_3, \dots c_m\}$ คือสมมติฐานที่เหตุการณ์ที่เกิดขึ้นนั้นจะถูกจำแนกประเภทเป็น $c_1, c_2, c_3, \dots c_m$

m คือจำนวนประเภททั้งหมดที่จะจำแนก

ในขั้นตอนการจำแนกประเภทด้วยตัวจำแนกประเภทเบส์แบบสามัญจะพิจารณาค่า $p(H = c_m|X)$ ที่มีค่ามากที่สุดและจะจัดให้เหตุการณ์ที่เกิดขึ้นนั้นอยู่ในประเภท c_m และเนื่องด้วยการคำนวณความน่าจะเป็นด้วยทฤษฎีของเบส์นั้น จะมีค่า p(X) เหมือนกันทั้งหมด เพราะการเปรียบเทียบนั้นย่อมจะเปรียบเทียบจากเหตุการณ์ที่เหมือนกันนั่นเอง จึงไม่จำเป็นต้องนำค่า p(X) มาคำนวณ ดังนั้นการจำแนกประเภทของเหตุการณ์แสดงได้ดังสมการ

$$\text{Classify}(x_i, \dots x_n) = \max\left(p(H = c) * \prod_{i=1}^{n} p(X = x_i|H = c)\right)$$

เมื่อ p คือ ความน่าจะเป็นของเหตุการณ์

c คือ ประเภทที่จำแนก

x_i คือ เหตุการณ์ที่ i

n คือ จำนวนเหตุการณ์ย่อยทั้งหมด

2.6.3 ความน่าจะเป็นเบส์แบบสามัญด้วยการแปลงลาปลาซ

หลาย ๆ ครั้งที่การคำนวณความน่าจะเป็นจากทฤษฎีความน่าจะเป็นเบส์นั้นมีผลลัพธ์เป็น 0 ซึ่งอาจจะเป็นผลมาจากข้อมูลที่นำมาคำนวณนั้นมีจำนวนข้อมูลที่น้อยมาก ทำให้ความน่าจะเป็นของเหตุการณ์บางเหตุการณ์นั้นมีค่าเป็น 0 ซึ่งเมื่อนำมาคำนวณหาความน่าจะเป็นเบส์แบบสามัญร่วมกับความน่าจะเป็นของเหตุการณ์อื่น ๆ จะได้ค่าเป็น 0 ด้วยเช่นกัน ซึ่งทำให้ข้อมูลส่วนที่น่าจะมีผลสำคัญต่อการคำนวณนั้นสูญเสียไปอย่างน่าเสียดาย ดังนั้นจึงมีการนำเอาทฤษฎีแปลงลาปลาซเข้ามาใช้เพื่อปรับค่าข้อมูลที่เป็น 0 นั้น [15]

การแปลงลาปลาซในความน่าจะเป็นเบส์แบบสามัญนั้นทำได้โดยการเพิ่มเหตุการณ์ที่เป็นไปได้เพิ่มเข้าไปอย่างละ 1 เหตุการณ์

- ความน่าจะเป็นโดยไม่มีการปรับแต่ง $\mathrm{p}(\mathrm{H}=\mathrm{c})=\frac{\mathrm{n_c}}{\mathrm{N}}$
- ความน่าจะเป็นโดยการแปลงลาปลาซ $\mathrm{p}(\mathrm{H}=\mathrm{c})=\frac{\mathrm{n_c}+1}{\mathrm{N}+\mathrm{K}}$

เมื่อ p คือ ความน่าจะเป็นของเหตุการณ์

c คือ ประเภทที่จำแนก

n_c คือ จำนวนเหตุการณ์ c

N คือ จำนวนเหตุการณ์ทั้งหมด

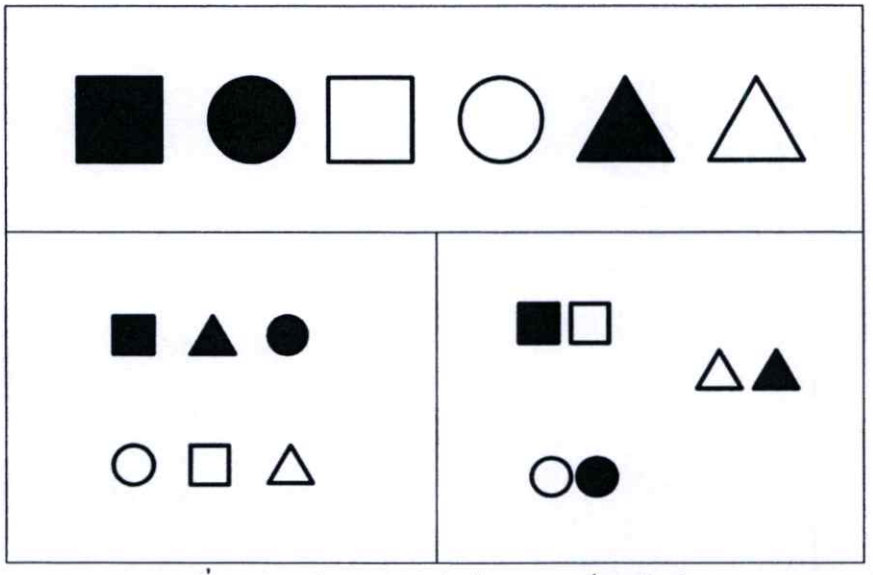
K คือ จำนวนของสมมติฐาน H ที่เป็นไปได้ทั้งหมด

2.7 การจัดกลุ่ม

เทคนิคการจัดกลุ่มนั้นเป้าหมายหลักนั้นอยู่ที่การแยกกลุ่มข้อมูลที่เรากำลังสนใจออกเป็นกลุ่มตามที่เราต้องการซึ่งกลุ่มที่แบ่งนั้นต้องตรงตามเกณฑ์ดังต่อไปนี้

- ข้อมูลในแต่ละกลุ่มต้องมีความเหมือนกัน
- ข้อมูลระหว่างกลุ่มต่าง ๆ จะต้องมีความแตกต่างกัน

กระบวนการวิธีพื้นฐานที่สุดของการจัดกลุ่มนี้ คือ การจัดกลุ่มแบบเคมีน ซึ่งใช้กันอย่างกว้างขวางและเป็นวิธีที่ใช้ในการอธิบายแนวคิดของการจัดกลุ่มได้เป็นอย่างดี ตัวอย่างการจัดกลุ่มแสดงให้เห็นดังภาพที่ 2.5



ภาพที่ 2.5 ตัวอย่างการจัดกลุ่มข้อมูลออกเป็นกลุ่มต่างๆ

2.7.1 การจัดกลุ่มแบบเคมีน

การจัดกลุ่มแบบเคมีนเป็นวิธีหนึ่งของการจัดกลุ่มข้อมูล (Data Clustering) ซึ่งเป็นวิธีที่ง่ายและสามารถใช้อธิบายแนวคิดของทฤษฎีการจัดกลุ่มได้อย่างง่ายที่สุด ซึ่งคำว่า *"เค (K)"* ในชื่อนั้นเป็นตัวแปรตัวค่าของจำนวนกลุ่มที่ต้องการจะแบ่งนั่นเอง ส่วนคำว่า *"มีน (Mean)"* นั้นหากแปลตรงตัวตามพจนานุกรมจะแปลว่าค่าเฉลี่ย ซึ่งในที่นี้จะหมายถึง ค่าเฉลี่ยของระยะทางวัดเทียบกับจุดศูนย์กลางกลุ่มของสมาชิกในกลุ่มทั้งหมด การวัดระยะทางระหว่างข้อมูลนั้นมีด้วยกันอยู่หลายมาตรวัดด้วยกันซึ่งจะกล่าวถึงต่อไป มาตรวัดเหล่านั้นจำคำนวณค่าระยะทางออกมาเป็นตัวเลข ดังนั้นข้อมูลที่จะใช้ในการจัด

กลุ่มนั้นจะต้องเป็นข้อมูลที่เป็นตัวเลขและนำไปวัดค่าได้เช่นกัน โดยมากมักจะทำให้อยู่ในรูปเวกเตอร์ หากข้อมูลที่จะใช้วัดเป็นข้อมูลรูปแบบอื่นจะต้องทำการปรับค่าให้เป็นตัวเลขเสียก่อน

กระบวนการวิธีของการจัดกลุ่มแบบเคมีนนั้นสามารถอธิบายได้อย่างเป็นขั้นตอน ดังนี้

1) สุ่มเลือกข้อมูลขึ้นมาตามจำนวนกลุ่มที่เราต้องการจะแบ่ง เพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นของกลุ่ม
2) ข้อมูลส่วนที่เหลือทั้งหมดจะถูกจัดเข้าไปอยู่ในกลุ่มที่อยู่ใกล้ที่สุด โดยวัดระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางที่เลือกเอาไว้ตอนต้น
3) คำนวณจุดศูนย์กลางของทุกกลุ่มใหม่อีกครั้ง และคำนวณหาค่าเฉลี่ยของระยะห่างของข้อมูลในกลุ่มทั้งหมดอีกครั้ง
4) เริ่มทำขั้นตอนที่ 2 จนกว่าจุดศูนย์กลางจะไม่เปลี่ยนแปลง

การวัดระยะห่างระหว่างข้อมูลที่ใช้ในกระบวนการวิธีของการจัดกลุ่มแบบเคมีนนั้นมีด้วยกันอยู่หลายวิธี เช่น การวัดระยะแบบยูคลิด (Euclidean Distance) การวัดระยะแบบแมนฮัตตัน (Manhattan Distance) และการวัดระยะแบบเชบิเชฟ (Chebychev Distance) เป็นต้น โดยในงานวิจัยในครั้งนี้ใช้การวัดระยะแบบยูคลิดในการจัดกลุ่ม

การวัดระยะแบบยูคลิด (Euclidean Distance)

คือระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง ซึ่งอาจสามารถวัดได้ด้วยไม้บรรทัด ระยะทางแบบยุคลิดระหว่างจุดสองจุด p และ q คือความยาวของส่วนของเส้นตรง pq หรือก็คือระยะห่างระหว่างจุด p และ q ซึ่งสามารถคำนวณได้จากสมการนี้

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i + q_i)^2}$$

เมื่อ d คือ ระยะห่างระหว่างเวกเตอร์

n คือ จำนวนมิติของเวกเตอร์ (จำนวนความนิยมของสินค้าที่นำมาใช้วัด)

p_i และ q_i คือ ค่าของเวกเตอร์ในมิติต่าง ๆ (ค่าความนิยมของสินค้าชิ้นที่ i)

2.8 ชุดข้อมูลที่ใช้ในการเรียนรู้และทดสอบ

2.8.1 ชุดข้อมูลทดสอบ Movie Lens

ชุดข้อมูลทดสอบที่ทำมาใช้ในการทำนายความนิยมนั้น เป็นชุดข้อมูลที่ชื่อว่า Movie Lens ซึ่งประกอบด้วยข้อมูลความนิยมจำนวน 100,000 ชุด จากผู้ใช้ 943 คน ที่มีต่อภาพยนตร์จำนวน 1,682 เรื่อง ซึ่งคะแนนความนิยมนั้นใช้เป็นเลขจำนวนเต็มตั้งแต่ 1 จนถึง 5 โดยค่าคะแนนความนิยม 1 หมายถึงผู้ใช้นั้นมีความชื่นชอบให้กับสินค้าน้อยที่สุด ส่วนค่าคะแนนความนิยม 5 นั้นหมายถึงผู้ใช้มีความชื่นชอบในสินค้านั้นมากที่สุด ซึ่งชุดข้อมูลนี้จะถูกนำมาแบ่งออกเป็นข้อมูลที่ใช้ทดสอบ 5 ชุดด้วยกัน โดยทำการทดลองโดยวิธีการตรวจสอบข้ามชุด (5-fold Cross Validation) แต่ละรอบการทดลองนั้น 1 ใน 5 ของข้อมูลนั้นจะถูกนำมาใช้ทดสอบและข้อมูลที่เหลือ 4 ใน 5 จะถูกใช้เป็นข้อมูลสอนทั้งหมด และทำการวนชุดข้อมูลที่ใช้ทดสอบไปจนครบ 5 ชุด โดนแต่ละชุดจะไม่ใช้ข้อมูลที่ซ้ำกันเลยมาทดสอบ ชุดข้อมูล Movie Lens นี้ สามารถดาวน์โหลดได้จาก http://www.grouplens.org/node/73

นอกจากข้อมูลความนิยมที่ผู้ใช้แต่ละคนมีให้กับสินค้าในระบบแล้วชุดข้อมูลของ Movie Lens ยังได้มีข้อมูลส่วนตัวของผู้ใช้ คือ อายุ เพศ และอาชีพ และข้อมูลรายละเอียดของสินค้า ในที่นี้คือชนิดของภาพยนตร์ เช่น ภาพยนตร์แนวชีวิต (Drama) ภาพยนต์แนวแฟนตาซี (Fantasy) หรือภาพยนตร์แนวตลก (Comedy) เป็นต้น ซึ่งงานวิจัยนี้ได้นำเอาข้อมูลของสินค้านี้มาทำการจัดกลุ่มสินค้า เพื่อช่วยในการทำนายความนิยมอีกทางหนึ่ง ชนิดของภาพยนตร์ทั้งหมดที่มีอยู่ในชุดข้อมูล Movie Lens มีด้วยกันทั้งหมด 19 ชนิด ได้แก่ Unknown, Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War และ Western

2.8.2 ชุดข้อมูลทดสอบ Book Crossing

ชุดข้อมูลอีกชุดหนึ่งที่ได้นำมาใช้ในการทำนายความนิยมได้แก่ชุดข้อมูล Book Crossing (BX) ซึ่งประกอบด้วยข้อมูลความนิยมที่ผู้ใช้ให้กับหนังสือ รวบรวมโดย Cai-Nicolas Ziegler ในช่วงระยะเวลา 4 สัปดาห์ระหว่างเดือนสิงหาคมและกันยายน ค.ศ. 2004 ซึ่งประกอบไปด้วยผู้ใช้จำนวน 278,858 คน ซึ่งให้ความนิยมหนังสือ 271,379 เล่มด้วยจำนวนถึง 1,149,780 ความนิยม ซึ่งคะแนนความนิยมนั้นใช้เป็นเลขจำนวนเต็มตั้งแต่ 1 จนถึง 10 โดยค่าคะแนนความนิยม 1 หมายถึงผู้ใช้นั้นมีความชื่นชอบให้กับสินค้าน้อยที่สุด ส่วนค่าคะแนนความนิยม 10 นั้นหมายถึงผู้ใช้มีความชื่นชอบในสินค้านั้นมากที่สุด ซึ่งชุดทดสอบนี้ถือว่าเป็นฐานข้อมูลที่มีขนาดใหญ่มาก การทดลองนั้นใช้วิธีการตรวจสอบข้ามชุด (5-fold

Cross Validation) แบ่งข้อมูลออกเป็น 5 กลุ่มโดยการสุ่มด้วยตนเอง ผู้วิจัยสามารถดาวน์โหลดข้อมูลได้จาก http://www.grouplens.org/node/74

ภาพที่ 2.6: หน้าเพจแรกของเว็บไซท์ www.grouplens.org

2.9 ตัววัดประสิทธิภาพการทำนายความนิยม

ตัววัดประสิทธิภาพของขั้นตอนวิธี ใช้ค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) และทำการทดลองโดยวิธีการตรวจสอบข้ามชุด K ชุด (K-fold Cross Validation) งานวิจัยนี้แบ่งข้อมูลทดสอบเป็น 5 ชุด แต่ละรอบการทำงานข้อมูลที่ไม่ได้นำมาใช้ทดสอบจะใช้เป็นข้อมูลสอนทั้งหมด

$$MAE = \frac{\sum_{i=1}^{n} |P_i - Q_i|}{n}$$

เมื่อ P_i คือค่าความนิยมที่ทำนาย

Q_i คือค่าความนิยมที่ถูกต้อง

n คือจำนวนความนิยมที่ทำนายทั้งหมด

2.10 งานวิจัยในระบบช่วยแนะนำ

งานวิจัยในระบบช่วยแนะนำนั้นที่นิยมทำกันมีอยู่ 2 ปัญหา ปัญหาแรกคือการทำนายความนิยม (Preference / Rating Prediction) เป็นการใช้กระบวนการวิธีต่าง ๆ ในการทำนายความนิยมที่ขาดหายไปในระบบ วัดประสิทธิภาพโดยใช้ค่าความคลาดเคลื่อนเทียบจากความนิยมจริงในระบบ ส่วนปัญหาที่สอง คือการแนะนำสินค้า N อันดับ (Top-N Recommendation) เป็นการใช้กระบวนการวิธีต่าง ๆ ในการทำนายหาสินค้า N อันดับแรกที่จะแนะนำให้กับผู้ใช้ วัดประสิทธิภาพโดยความแม่นยำในการทำนายเทียบกับสินค้าที่ผู้ใช้ในระบบสนใจจริง [4] งานวิจัยที่นำทฤษฎีความน่าจะเป็นเบส์แบบสามัญมาใช้มีไม่มากนัก และส่วนใหญ่จะนำไปใช้ในระบบช่วยแนะนำแบบผสม โดยตัวจำแนกประเภทเบส์แบบสามัญ (Naïve Bayes Classifier) ถูกใช้ในส่วนของการกรองเนื้อหา

2.10.1 Content-Boosted Collaborative Filtering for Improved Recommendations

ในปี ค.ศ. 2002 Melville, P., Mooney, R. J. และ Nagarajan, R. [12] ได้ทำงานวิจัยเรื่อง "Content-Boosted Collaborative Filtering for Improved Recommendations" ซึ่งเป็นระบบช่วยแนะนำแบบผสมที่ใช้ตัวจำแนกประเภทเบส์แบบสามัญในการคัดกรองเนื้อหาเพื่อทำนายความนิยมของผู้ใช้เพิ่มเติมบางส่วนทำให้มีข้อมูลความนิยมในระบบมากขึ้น แนวคิดของงานวิจัยนี้ก็คือการใช้ระบบช่วยแนะนำแบบผสมเพื่อแก้ปัญหาข้อมูลเบาบาง โดยการใช้วิธีการกรองเนื้อหาเพื่อทำนายความนิยมของสินค้าในระบบบางส่วนก่อน จากนั้นใช้วิธีการกรองร่วมโดยนำผลลัพธ์ที่ได้ไปคำนวณด้วยวิธีพิจารณาย่านใกล้เคียงที่สุด K ตัว โดยการถ่วงน้ำหนักและวัดความคล้ายของเวกเตอร์ความนิยมในการกรองร่วม เพื่อทำนายความนิยมของผู้ใช้ที่เหลือในระบบทั้งหมด จุดเด่นของวิธีนี้คือสามารถแก้ไขปัญหาความเบาบางของข้อมูลในระบบได้ดี แต่จุดด้อยของวิธีนี้ก็คือ หากวิธีการทำนายความนิยมที่ใช้ครั้งแรกมีความคลาดเคลื่อนสูง เมื่อนำมาเป็นข้อมูลเข้าในการทำนายอีกวิธีหนึ่ง แม้ว่าอีกวิธีหนึ่งจะเป็นวิธีที่มีประสิทธิภาพดี ก็อาจจะทำให้ผลการทำนายนั้นคลาดเคลื่อนสูงตามไปด้วยได้เช่นกัน

2.10.2 A Simple Hybrid Movie Recommender System

ในปีค.ศ. 2004 Jaldert Rombouts และ Tessa Verhoef [7] ได้มีงานวิจัยที่ชื่อว่า "A Simple Hybrid Movie Recommender System" ซึ่งใช้ระบบช่วยแนะนำแบบผสมโดยใช้ผลรวมเชิงเส้นของทั้งการกรองร่วมและการกรองเนื้อหาเข้าด้วยกันด้วยอัตราส่วน 50:50 โดยในส่วนของการกรองเนื้อหานั้น ระบบพิจารณาจากคำสำคัญ ชนิดของภาพยนตร์ และรายชื่อนักแสดงนำ และทำนายความนิยมด้วยตัวจำแนกเบส์แบบสามัญ ในส่วนของการกรองร่วมนั้นใช้วิธีพิจารณาย่านใกล้เคียงที่สุด K ตัวและถ่วงน้ำหนักความคล้ายเวกเตอร์ความนิยมในการทำนาย แนวคิดของงานวิจัยนี้ก็คือ การทำนายความนิยมโดยใช้ทั้ง

สองเทคนิคร่วมกันอย่างง่าย คือ การกรองเนื้อหาและการกรองร่วมในอัตราส่วนที่เท่ากัน ข้อดีของวิธีนี้ คือ เป็นขั้นตอนที่ง่ายและไม่ยุ่งยาก ส่วนของด้อยของวิธีนี้ คือ การทำนายความนิยมจะไม่เด่นไปทางด้านใดด้านหนึ่ง คือ การแนะนำจะไม่มุ่งเน้นไปที่พฤติกรรมผู้ใช้โดยชัดเจนเกินไป หรือไม่เน้นไปที่เนื้อหาของสินค้าอย่างเดียว

บทที่ 3
การทำนายความนิยมโดยใช้เทคนิคการถ่วงน้ำหนักด้วยความน่าจะเป็นเบส์แบบสามัญ

ในบทนี้จะนำเสนอขั้นตอนการทำนายความนิยมในการกรองร่วม โดยใช้ตัวถ่วงน้ำหนักที่ได้จากทฤษฎีความน่าจะเป็นเบส์แบบสามัญเพื่อใช้ในซึ่งผลลัพธ์ที่ได้ คือความนิยมของผู้ใช้คนหนึ่งที่มีต่อสินค้าตัวหนึ่ง โดยค่านิยมที่จะใช้ในระบบนั้นขึ้นอยู่กับผู้ที่สร้างระบบขึ้นมาจะกำหนด โดยทั่วไปค่าความนิยมที่น้อย หมายถึงผู้ใช้มีความชอบในตัวสินค้านั้นน้อย ส่วนค่าความนิยมที่มาก หมายถึงผู้ใช้มีความชอบต่อตัวสินค้านั้นมากขึ้นตามลำดับ

3.1 ปัญหาการทำนายความนิยม

ปัญหาการทำนายความนิยมในงานวิจัยนี้ นำเสนอในรูปแบบของเมตริกซ์หรือตาราง 2 มิติ โดยที่ในแถวนั้นแสดงรายการผู้ใช้ (User) และคอลัมน์แสดงรายการสินค้า (Item) โดยข้อมูลภายในตารางคือความนิยมที่ผู้ใช้คนหนึ่งให้กับสินค้าตัวหนึ่ง ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างตารางแสดงค่าความนิยมที่ผู้ใช้ให้กับสินค้าในระบบ

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4	I_5
U_1	2	3	4	?	2
U_2	4	3	2	4	?
U_3	2	2	4	5	1
U_4	1	2	2	3	5
U_5	3	2	4	5	2

จากข้อมูลในตารางที่ 3.1 นั้น คือระบบที่ผู้ใช้ 5 คนให้ความนิยมแก่ตัวสินค้าต่าง ๆ 5 ชนิด โดยมีที่มีค่าความนิยมตั้งแต่ 1 จนถึง 5 ซึ่งในโลกแห่งความเป็นจริงแล้วจำนวนผู้ใช้และจำนวนสินค้านั้นมีมากกว่านี้หลายเท่าตัวนัก และค่าความนิยมที่มีในระบบอาจจะไม่หนาแน่นมากนักทำให้การทำนายความนิยมเป็นไปได้ลำบากดังที่กล่าวไว้แล้วในบทที่ 2

3.2 การทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ

ตัวอย่างที่ใช้ในการทำนายความนิยมจะอ้างอิงจากตัวอย่างตารางที่ 3.1 โดยวิธีที่นำเสนอนั้นคำนวณมาจากทฤษฎีความน่าจะเป็นเบส์แบบสามัญที่ใช้ในตัวจำแนกประเภทเบส์แบบสามัญ โดยสมมติให้เหตุการณ์ที่ผู้ใช้ให้ความนิยมกับตัวสินค้ากับเหตุการณ์ที่สินค้าถูกให้ความนิยมโดยผู้ใช้นั้นเป็นอิสระต่อกัน คำนวณความน่าจะเป็นที่ความนิยม 1 จนถึง 5 จะถูกให้กับสินค้าที่ระบุโดยผู้ใช้ที่ระบุ จากนั้นนำค่าที่ได้ไปถ่วงน้ำหนักกับความนิยมแต่ละค่านั้น และนำผลรวมที่ได้มาเป็นค่าทำนายความนิยมที่ต้องการ ดังนี้

- **ตัวอย่างที่ 1** การคำนวณความน่าจะเป็นของแต่ละค่าความนิยมที่ผู้ใช้ U_1 จะให้กับสินค้า I_4

$$p(c=1|User=1, Item=4)$$
$$= p(c=1) \times p(User=1|c=1) \times p(Item=4|c=1) = \frac{2}{23} \times \frac{0}{4} \times \frac{0}{4} = 0$$
$$p(c=2|User=1, Item=4)$$
$$= p(c=2) \times p(User=1|c=2) \times p(Item=4|c=2) = \frac{9}{23} \times \frac{2}{4} \times \frac{0}{4} = 0$$
$$p(c=3|User=1, Item=4)$$
$$= p(c=3) \times p(User=1|c=3) \times p(Item=4|c=3) = \frac{4}{23} \times \frac{1}{4} \times \frac{1}{4} = 0.0109$$
$$p(c=4|User=1, Item=4)$$
$$= p(c=4) \times p(User=1|c=4) \times p(Item=4|c=4) = \frac{5}{23} \times \frac{1}{4} \times \frac{1}{4} = 0.0136$$
$$p(c=5|User=1, Item=4)$$
$$= p(c=5) \times p(User=1|c=5) \times p(Item=4|c=5) = \frac{3}{23} \times \frac{0}{4} \times \frac{2}{4} = 0$$

ตารางที่ 3.2 ค่าการทำนายค่าความนิยมที่ผู้ใช้ U_1 จะให้กับสินค้า I_4 จากตัวอย่างที่ 1

ความนิยม	1	2	3	4	5
ค่าถ่วงน้ำหนัก	0	0	0.0109	0.0136	0
ความนิยม x ค่าถ่วงน้ำหนัก	0	0	0.0282	0.05	0
ค่าทำนายความนิยม	$\frac{\sum(\text{ความนิยม}*\text{ค่าถ่วงน้ำหนัก})}{\sum \text{ค่าถ่วงน้ำหนัก}} = 3.5551$				

○ **ตัวอย่างที่ 2** การคำนวณความน่าจะเป็นของแต่ละค่าความนิยมที่ผู้ใช้ U_2 จะให้กับสินค้า I_5

$$p(c=1|User=2, Item=5)$$

$$= p(c=1) \times p(User=2|c=1) \times p(Item=5|c=1) = \frac{2}{23} \times \frac{0}{4} \times \frac{1}{4} = 0$$

$$p(c=2|User=2, Item=5)$$

$$= p(c=2) \times p(User=2|c=2) \times p(Item=5|c=2) = \frac{9}{23} \times \frac{1}{4} \times \frac{2}{4} = 0.0489$$

$$p(c=3|User=2, Item=5)$$

$$= p(c=3) \times p(User=1|c=3) \times p(Item=4|c=3) = \frac{4}{23} \times \frac{1}{4} \times \frac{0}{4} = 0$$

$$p(c=4|User=2, Item=5)$$

$$= p(c=4) \times p(User=1|c=4) \times p(Item=4|c=4) = \frac{5}{23} \times \frac{2}{4} \times \frac{0}{4} = 0$$

$$p(c=5|User=2, Item=5)$$

$$= p(c=5) \times p(User=1|c=5) \times p(Item=4|c=5) = \frac{3}{23} \times \frac{0}{4} \times \frac{1}{4} = 0$$

ตารางที่ 3.3 ค่าการทำนายค่าความนิยมที่ผู้ใช้ U_2 จะให้กับสินค้า I_5 จากตัวอย่างที่ 2

ความนิยม	1	2	3	4	5
ค่าถ่วงน้ำหนัก	0	0.0489	0	0	0
ความนิยม x ค่าถ่วงน้ำหนัก	0	0.0978	0	0	0
ค่าทำนายความนิยม	$\frac{\sum(\text{ความนิยม}*\text{ค่าถ่วงน้ำหนัก})}{\sum \text{ค่าถ่วงน้ำหนัก}} = 2$				

3.3 การทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการจัดกลุ่มแบบเคมีน

เนื่องจากเมื่อสินค้าใหม่ถูกเพิ่มเข้ามาในระบบจะยังไม่ถูกให้คะแนนความนิยม ดังนั้นเมื่อนำไปคำนวณจากวิธีดังกล่าวข้างต้นนั้นจะทำให้ไม่สามารถทำการคำนวณได้เลย ดังนั้นเพื่อเป็นการแก้ไขปัญหาในจุดนี้ จึงได้นำเอาการจัดกลุ่มของสินค้าเข้ามาช่วย การจัดกลุ่มนั้นทำได้โดยการนำเอาเนื้อหาของสินค้าที่เราจะจัดกลุ่มเข้ามาแปลงในรูปแบบของเวกเตอร์ และจัดกลุ่มโดยใช้วิธีการจัดกลุ่มแบบเคมีนจัดกลุ่มสินค้าใหม่และสินค้าที่มีอยู่เดิมในระบบ วัดความคล้ายด้วยการวัดระยะแบบยูคลิเดียน (Euclidean Distance) ตัวอย่างตารางการจัดกลุ่มสินค้าแสดงดังตารางที่ 3.4 และทำการคำนวณโดยการถ่วงน้ำหนักด้วยความน่าจะเป็นเบส์แบบสามัญเหมือนเดิม แต่ใช้เหตุการณ์ที่เป็นกลุ่มของสินค้านั้นมาคำนวณแทนตัวสินค้านั้นโดยตรง

ตารางที่ 3.4 ตัวอย่างตารางแสดงการจัดกลุ่มของสินค้าใหม่เข้ากับสินค้าเดิมในระบบ

กลุ่มสินค้า	กลุ่มสินค้าที่ 1 ($Cluster_1$)			กลุ่มสินค้าที่ 2 ($Cluster_2$)		
สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4	I_5	I_6
U_1	2	3	4	-	2	?
U_2	4	3	2	4	-	?
U_3	2	2	4	5	1	4
U_4	1	2	2	3	5	4
U_5	3	2	4	5	2	5

- **ตัวอย่างที่ 3** การคำนวณความน่าจะเป็นของแต่ละค่าความนิยมที่ผู้ใช้ U_1 จะให้กับสินค้า I_6

$$p(c=1|User=1, Cluster=2)$$
$$= p(c=1) \times p(User=1|c=1) \times p(Cluster=2|c=1) = \frac{2}{23} \times \frac{0}{4} \times \frac{1}{11} = 0$$
$$p(c=2|User=1, Cluster=2)$$
$$= p(c=2) \times p(User=1|c=2) \times p(Cluster=2|c=2) = \frac{9}{23} \times \frac{2}{4} \times \frac{2}{11} = 0.0356$$
$$p(c=3|User=1, Cluster=2)$$
$$= p(c=3) \times p(User=1|c=3) \times p(Cluster=2|c=3) = \frac{4}{23} \times \frac{1}{4} \times \frac{1}{11} = 0.004$$
$$p(c=4|User=1, Cluster=2)$$
$$= p(c=4) \times p(User=1|c=4) \times p(Cluster=2|c=4) = \frac{5}{23} \times \frac{1}{4} \times \frac{3}{11} = 0.0148$$

$$p(c=5|User=1, Cluster=2)$$

$$= p(c=5) \times p(User=1|c=5) \times p(Cluster=2|c=5) = \frac{3}{23} \times \frac{0}{4} \times \frac{4}{11} = 0$$

ตารางที่ 3.5 ค่าการทำนายค่าความนิยมที่ผู้ใช้ U_2 จะให้กับกลุ่มสินค้า I_6 จากตัวอย่างที่ 3

ความนิยม	1	2	3	4	5
ค่าถ่วงน้ำหนัก	0	0.0356	0.004	0.0148	0
ความนิยม x ค่าถ่วงน้ำหนัก	0	0.712	0.012	0.0592	0
ค่าทำนายความนิยม	$\frac{\sum(\text{ความนิยม}*\text{ค่าถ่วงน้ำหนัก})}{\sum \text{ค่าถ่วงน้ำหนัก}} = 2.6176$				

- **ตัวอย่างที่ 4** การคำนวณความน่าจะเป็นของแต่ละค่าความนิยมที่ผู้ใช้ U_2 จะให้กับสินค้า I_6

$$p(c=1|User=2, Cluster=2)$$

$$= p(c=1) \times p(User=2|c=1) \times p(Cluster=2|c=1) = \frac{2}{23} \times \frac{0}{4} \times \frac{1}{11} = 0$$

$$p(c=2|User=2, Cluster=2)$$

$$= p(c=2) \times p(User=2|c=2) \times p(Cluster=2|c=2) = \frac{9}{23} \times \frac{1}{4} \times \frac{2}{11} = 0.0178$$

$$p(c=3|User=2, Cluster=2)$$

$$= p(c=3) \times p(User=2|c=3) \times p(Cluster=2|c=3) = \frac{4}{23} \times \frac{1}{4} \times \frac{1}{11} = 0.004$$

$$p(c=4|User=2, Cluster=2)$$

$$= p(c=4) \times p(User=2|c=4) \times p(Cluster=2|c=4) = \frac{5}{23} \times \frac{2}{4} \times \frac{3}{11} = 0.0296$$

$$p(c=5|User=2, Cluster=2)$$

$$= p(c=5) \times p(User=2|c=5) \times p(Cluster=2|c=5) = \frac{3}{23} \times \frac{0}{4} \times \frac{4}{11} = 0$$

ตารางที่ 3.6 ค่าการทำนายค่าความนิยมที่ผู้ใช้ U_2 จะให้กับกลุ่มสินค้า I_6 จากตัวอย่างที่ 4

ความนิยม	1	2	3	4	5
ค่าถ่วงน้ำหนัก	0	0.0178	0.004	0.0296	0
ความนิยม x ค่าถ่วงน้ำหนัก	0	0.0356	0.012	0.1184	0
ค่าทำนายความนิยม	$\frac{\sum(\text{ความนิยม}*\text{ค่าถ่วงน้ำหนัก})}{\sum \text{ค่าถ่วงน้ำหนัก}} = 3.2296$				

3.4 การทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซ

เนื่องจากข้อมูลในระบบนั้นอาจจะมีมีความเบาบางมาก ทำให้ความน่าจะเป็นที่นำมาคำนวณได้นั้นมีค่าเป็น 0 ดังที่ได้กล่าวไปแล้วในบทที่ 2 ซึ่งทำให้ข้อมูลความน่าจะเป็นส่วนที่ไม่ได้เป็น 0 เมื่อถูกนำมาคำนวณด้วยความน่าจะเป็นเบส์แบบสามัญแล้วทำให้ผลลัพธ์กลายเป็น 0 ไปด้วย ซึ่งในบางครั้งอาจจะทำให้เราสูญเสียข้อมูลส่วนที่สำคัญไป ดังนั้นจึงได้นำเอาการปรับแต่งแบบลาปลาซมาช่วย โดยตัวอย่างที่เราจะนำมาคำนวณนั้น จะใช้ข้อมูลจากตารางที่ 3.1 และนำตัวอย่างที่ 1 ที่เสนอไปข้างต้นมาเพิ่มส่วนของการแปลงลาปลาซเข้าไป ดังนี้

- **ตัวอย่างที่ 5** การคำนวณความน่าจะเป็นด้วยการแปลงลาปลาซของแต่ละค่าความนิยมที่ผู้ใช้ U_1 จะให้กับสินค้า I_4

$$p(c=1|User=1, Item=4)$$

$$= p(c=1) \times p(User=1|c=1) \times p(Item=4|c=1) = \frac{2+1}{23+5} \times \frac{0+1}{4+5} \times \frac{0+1}{4+5}$$

$$= \frac{3}{28} \times \frac{1}{9} \times \frac{1}{9} = 0.00132275$$

$$p(c=2|User=1, Item=4)$$

$$= p(c=2) \times p(User=1|c=2) \times p(Item=4|c=2) = \frac{9+1}{23+5} \times \frac{2+1}{4+5} \times \frac{0+1}{4+5}$$

$$= \frac{10}{28} \times \frac{3}{9} \times \frac{1}{9} = 0.01322751$$

$$p(c=3|User=1, Item=4)$$

$$= p(c=3) \times p(User=1|c=3) \times p(Item=4|c=3) = \frac{4+1}{23+5} \times \frac{1+1}{4+5} \times \frac{1+1}{4+5}$$

$$= \frac{5}{28} \times \frac{2}{9} \times \frac{2}{9} = 0.00881834$$

$$p(c=4|User=1, Item=4)$$

$$= p(c=4) \times p(User=1|c=4) \times p(Item=4|c=4) = \frac{5+1}{23+5} \times \frac{1+1}{4+5} \times \frac{1+1}{4+5}$$

$$= \frac{6}{28} \times \frac{2}{9} \times \frac{2}{9} = 0.01058201$$

$$p(c=5|User=1, Item=4)$$

$$= p(c=5) \times p(User=1|c=5) \times p(Item=4|c=5) = \frac{3+1}{23+5} \times \frac{0+1}{4+5} \times \frac{2+1}{4+5}$$

$$= \frac{4}{28} \times \frac{1}{9} \times \frac{3}{9} = 0.00529101$$

ตารางที่ 3.7 ค่าการทำนายค่าความนิยมที่ผู้ใช้ U_1 จะให้กับสินค้า I_4 จากตัวอย่างที่ 5

ความนิยม	1	2	3	4	5
ค่าถ่วงน้ำหนัก	0.00132	0.01322	0.00881	0.01058	0.00529
ความนิยม x ค่าถ่วงน้ำหนัก	0.00132	0.02645	0.02645	0.04232	0.02645
ค่าทำนายความนิยม	$\frac{\sum(\text{ความนิยม} * \text{ค่าถ่วงน้ำหนัก})}{\sum \text{ค่าถ่วงน้ำหนัก}} = 3.1348$				

จะเห็นได้ว่าจากตัวอย่างเดิมนั้นค่าถ่วงน้ำหนักของความนิยม 1, 2 และ 5 นั้นมีค่าเป็น 0 และผลลัพธ์ของค่าทำนายความนิยมอยู่ที่ 3.5551 แต่หลังจากใช้การแปลงลาปลาซเพิ่มเข้าไปทำให้น้ำหนักของค่าความนิยม 1, 2 และ 5 นั้นได้ถูกนำมาใช้ปรับปรุงผลลัพธ์ด้วย ดังนั้นจึงถือว่าเป็นวิธีการที่ทำให้สามารถดึงข้อมูลที่มีอยู่ในระบบเข้ามาร่วมใช้ในการคำนวณได้อย่างครบถ้วนที่สุด

บทที่ 4

ผลการทดลอง

บทที่ 3 ได้อธิบายแนวคิดในการใช้ความน่าจะเป็นเบส์แบบสามัญในการทำนายความนิยมประยุกต์ใช้ร่วมกับจัดกลุ่มแบบเคมีน รวมถึงแสดงตัวอย่างในการคำนวณเพื่อให้ง่ายต่อความเข้าใจไปแล้วนั้น ในบทที่ 4 ได้นำเสนอผลการทดลองที่ได้จากชุดทดสอบที่ใช้ในงานวิจัยนี้

4.1 ขั้นตอนการทดลอง

4.1.1 การทดลองด้วยชุดข้อมูลทดสอบ Movie Lens

1. นำชุดข้อมูลที่ถูกแบ่งไว้อย่างละเท่าๆกันจำนวน 5 ชุดข้อมูลมาทำการทดลอง โดยที่จะเลือก 4 ใน 5 ชุดข้อมูลทดสอบมาทำการเรียนรู้ ส่วนอีกชุดหนึ่งที่เหลือจะทำเป็นชุดข้อมูลทดสอบ เวียนไปจนครบ 5 ชุดข้อมูล
2. แต่ละชุดข้อมูลทดสอบจะถูกนำมาคำนวณความน่าจะเป็นที่ผู้ใช้ u_j จะให้คะแนนความนิยมแก่สินค้า i_k ด้วยค่าคะแนนความนิยม r ดังสมการความน่าจะเป็นเบส์แบบสามัญ

$$p(r|f_i) = p(c)\prod_{i=1}^{n} p(f_i|r)$$

เมื่อ p คือความน่าจะเป็นของเหตุการณ์
r คือคะแนนความนิยม (Rating/Preference)
f_i คือเหตุการณ์ที่ i
n คือจำนวนเหตุการณ์ทั้งหมด

เหตุการณ์ f_i ของปัญหาการทำนายความนิยมมีด้วยกันสองเหตุการณ์คือ เหตุการณ์ที่ผู้ใช้ u_j จะให้คะแนนความนิยม r และเหตุการณ์ที่สินค้า i_k จะถูกให้ความนิยมด้วยด้วยคะแนน r

3. คำนวณค่าความน่าจะเป็นของทุกคะแนนความน่าจะเป็นที่มีในระบบตั้งแต่ 1 ถึง 5 และนำค่าความน่าจะเป็นที่ได้มาถ่วงน้ำหนักด้วยคะแนนความนิยมนั้น และใช้ผลรวมเชิงเส้นเป็นค่าทำนายความนิยม (ตัวอย่างโดยละเอียดแสดงในบทที่ 3)
4. ข้อมูลส่วนที่ไม่สามารถทำนายได้ จะถูกนำมาจัดกลุ่มด้วยเนื้อหาของสินค้าก่อน โดยการจัดกลุ่มข้อมูลของชุดทดสอบ Movie Lens นั้นจัดกลุ่มด้วยชนิดของภาพยนต์ ซึ่ง

ภาพยนตร์แต่ละเรื่องจะถูกระบุชนิดของภาพยนตร์ตั้งแต่ 1 ถึง 3 ชนิดด้วยกัน และนำมาแปลงเป็นเวกเตอร์ชนิดของภาพยนตร์ ดังนี้

(Unknown, Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western)

หากภาพยนตร์เรื่องใดถูกระบุให้อยู่ในชนิดภาพยนตร์ใดจะแทนค่าด้วย 1 และที่เหลือจะแทนด้วยค่า 0 แล้วนำเอาเวกเตอร์ที่ได้เหล่านี้ไปนำการจัดกลุ่มด้วยวิธีการจัดกลุ่มแบบเคมีน

5. จำนวนกลุ่มที่ใช้แบ่งกลุ่มในการทดลองครั้งนี้ คือ 5 กลุ่ม
6. นำเอากลุ่มของภาพยนตร์ c_i ที่จัดเอาไว้ ไปคำนวณแทน ตั้งแต่ข้อ 2 ถึง 4 อีกครั้ง
7. นำค่าการทำนายความนิยมที่ได้ไปคำนวณหาค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์
8. เปรียบเทียบผลการทดลองระหว่างกับวิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญที่นำเสนอทั้งแบบใช้การแปลงลาปลาซ และแบบไม่ใช้ เปรียบเทียบกับเทคนิคการกรองร่วมแบบดั้งเดิม (Traditional Collaborative Filtering) ที่ใช้การถ่วงน้ำหนักด้วยการวัดความคล้ายและพิจารณาย่านใกล้เคียง 20 ตัว และนำไปเปรียบเทียบกับวิธีการทำนายโดยใช้ตัวจำแนกประเภทแบบเบส์
9. ทดลองโดยปรับปรุงตัวแบบการทำนายความนิยมด้วยตัวถ่วงน้ำหนักเบส์แบบสามัญอีกครั้ง โดยใช้การแปลงลาปลาซเพื่อแก้ไขค่าข้อมูลที่เป็น 0 แทนการใช้การจัดกลุ่ม และทดลองโดยใช้การแปลงลาปลาซในทุกข้อมูลทดสอบ

4.1.2 การทดลองด้วยชุดข้อมูลทดสอบ Book Crossing

1. แบ่งชุดข้อมูลออกเป็น 5 ชุดข้อมูลด้วยวิธีการสุ่ม โดยที่จะเลือก 4 ใน 5 ชุดข้อมูลทดสอบมาทำการเรียนรู้ ส่วนอีกชุดหนึ่งที่เหลือจะทำเป็นชุดข้อมูลทดสอบ เวียนไปจนครบ 5 ชุดข้อมูล
2. แปลงข้อมูลความนิยมในระบบจาก 1-10 ให้เป็น 1-5 โดยที่
   - คะแนนความนิยม 1 และ 2 จะแทนด้วยคะแนนความนิยม 1
   - คะแนนความนิยม 3 และ 4 จะแทนด้วยคะแนนความนิยม 2
   - คะแนนความนิยม 5 และ 6 จะแทนด้วยคะแนนความนิยม 3
   - คะแนนความนิยม 7 และ 8 จะแทนด้วยคะแนนความนิยม 4
   - คะแนนความนิยม 9 และ 10 จะแทนด้วยคะแนนความนิยม 5

3. แต่ละชุดข้อมูลทดสอบจะถูกนำมาคำนวณความน่าจะเป็นที่ผู้ใช้ u_j จะให้คะแนนความนิยมแก่สินค้า i_k ด้วยค่าคะแนนความนิยม r ดังสมการความน่าจะเป็นเบส์แบบสามัญ

$$p(r|f_i) = p(c)\prod_{i=1}^{n} p(f_i|r)$$

เมื่อ p คือความน่าจะเป็นของเหตุการณ์
r คือคะแนนความนิยม (Rating/Preference)
f_i คือเหตุการณ์ที่ i
n คือจำนวนเหตุการณ์ทั้งหมด

เหตุการณ์ f_i ของปัญหาการทำนายความนิยมมีด้วยกันสองเหตุการณ์คือ เหตุการณ์ที่ผู้ใช้ u_j จะให้คะแนนความนิยม r และเหตุการณ์ที่สินค้า i_k จะถูกให้ความนิยมด้วยด้วยคะแนน r

4. คำนวณค่าความน่าจะเป็นของทุกคะแนนความน่าจะเป็นที่มีในระบบตั้งแต่ 1 ถึง 5 และนำค่าความน่าจะเป็นที่ได้มาถ่วงน้ำหนักด้วยคะแนนความนิยมนั้น และใช้ผลรวมเชิงเส้นเป็นค่าทำนายความนิยม
5. นำค่าการทำนายความนิยมที่ได้ไปคำนวณหาค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ โดยค่าความนิยมที่ทำนายไม่ได้เนื่องจากปัญหาความเบาบางของข้อมูลจนทำให้ความน่าจะเป็นรวมมีค่าเป็นศูนย์นั้นจะถูกนำเข้าไปคำนวณความน่าจะเป็นเบส์แบบสามัญด้วยการแปลงลาปลาซ เพื่อแก้ไขปัญหาค่าข้อมูลที่เป็นศูนย์

4.2 ผลการทดลอง

4.2.1 ผลการทดลองด้วยชุดข้อมูลทดสอบ Movie Lens

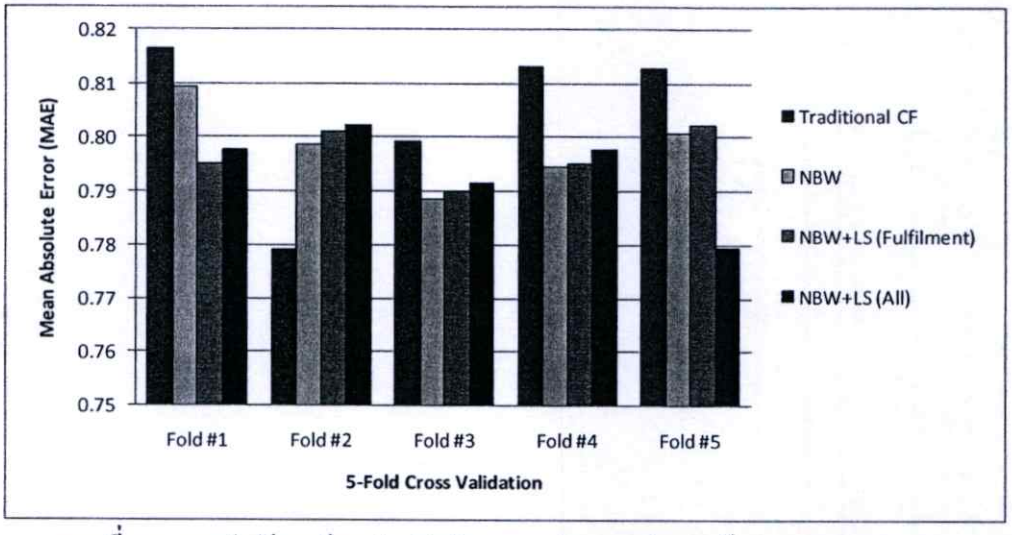
การทดสอบประสิทธิภาพในการทำนายเปรียบเทียบกันในแต่ละชุดข้อมูลด้วยชุดทดสอบ Movie Lens ระหว่างเทคนิคการกรองร่วมแบบดั้งเดิม กับวิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญที่นำเสนอทั้งแบบใช้การแปลงลาปลาซและแบบไม่ใช้การแปลงลาปลาซ ผลการทดลองค่าความคลาดเคลื่อนสัมบูรณ์ที่ได้แสดงดังตารางที่ 4.1

ตารางที่ 4.1: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบย่อยของ Movie Lens

ชุดข้อมูลทดสอบย่อย Movie Lens	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญ	0.80933	0.79875	0.78863	0.79466	0.81310
เทคนิคการกรองร่วมแบบดั้งเดิม	0.81650	0.77910	0.79920	0.81330	0.81310
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญด้วยการแปลงลาปลาซแก้ไขค่าที่เป็นศูนย์	0.79516	0.80105	0.79019	0.79516	0.80234
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญด้วยการแปลงลาปลาซ	0.79788	0.80224	0.79170	0.79788	0.77944

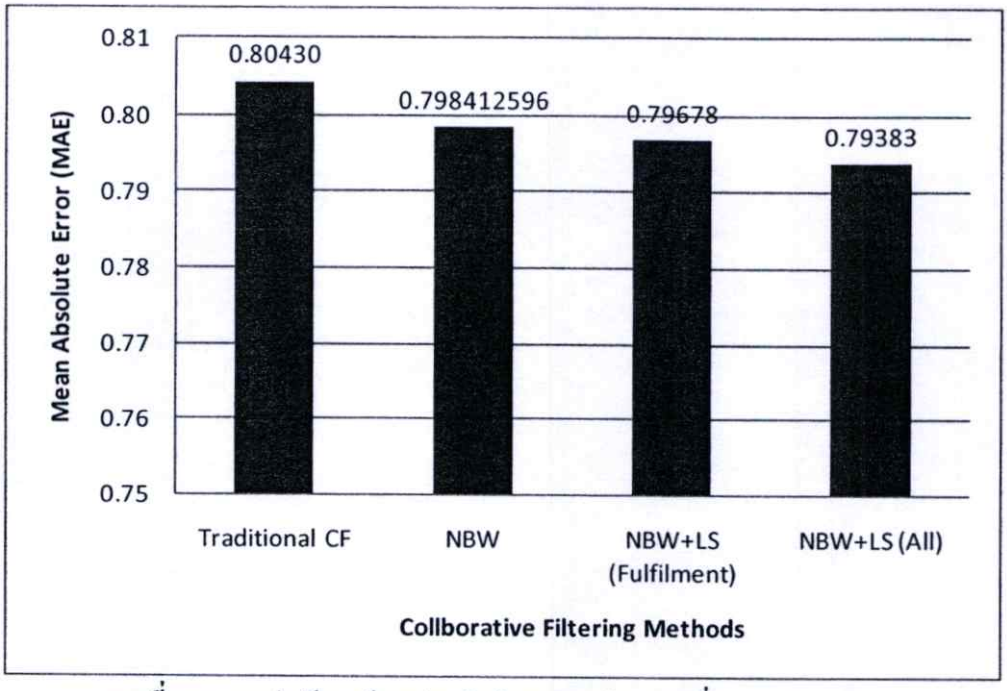
จากการทดลองวัดประสิทธิภาพการทำนายด้วย 4 วิธี คือ วิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ (Naïve Bayes Weighing: NBW) วิธีการกรองร่วมแบบดั้งเดิม (Traditional Collaborative Filtering: TCF) นั้น วิธีการทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซ (Naïve Bayes Weighing with Laplace Smoothing: NBW+LS) โดยการทดลองใช้แก้ไขเฉพาะค่าข้อมูลที่เป็นศูนย์ (NBW+LS – Fulfillment) และสุดท้ายคือวิธีการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซแต่ทดลองด้วยการใช้กับทั้งชุดข้อมูล (NBW+LS - All) ผลการทดลองที่ได้ในชุดข้อมูลย่อยที่ 1 จะเห็นว่าการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซแก้ค่าศูนย์นั้นมีประสิทธิภาพในการทำนายมากที่สุด ส่วนวิธีการกรองร่วมแบบดั้งเดิมนั้นมีประสิทธิภาพในการทำนายน้อยที่สุด แต่ในชุดข้อมูลย่อยที่ 2 นั้นวิธีการกรองร่วมแบบดั้งเดิมมีประสิทธิภาพมากที่สุด ส่วนการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซแบบทั้งชุดข้อมูลนั้นมีประสิทธิภาพที่น้อยที่สุด ในชุด

ข้อมูลย่อยที่ 3 และ 4 การถ่วงน้ำหนักเบส์แบบสามัญแบบปกติให้ประสิทธิภาพการทำนายที่ดีที่สุด ส่วนวิธีการกรองร่วมแบบดั้งเดิมนั้นมีประสิทธิภาพในการทำนายน้อยที่สุด และสุดท้ายกับชุดข้อมูลย่อยที่ 5 การถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซแบบทั้งชุดข้อมูลนั้นมีประสิทธิภาพที่ดีที่สุด แต่วิธีการกรองร่วมแบบดั้งเดิมนั้นมีประสิทธิภาพในการทำนายน้อยที่สุดภาพที่ 4.1 แสดงการเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์แต่ละชุดข้อมูลย่อยของชุดข้อมูล Movie Lens โดยแยกตามแต่ละเทคนิคที่ใช้ในการเปรียบเทียบ



ภาพที่ 4.1: กราฟเปรียบเทียบประสิทธิภาพการทำนายแต่ละชุดข้อมูลของ Movie Lens

จากผลการทดสอบประสิทธิภาพในการทำนายที่ได้ในแต่ละชุดข้อมูลย่อยของ Movie Lens นั้น เมื่อนำผลมารวมเฉลี่ยในแต่ละชุดข้อมูลแล้วจะเห็นได้จากภาพที่ 4.2 ซึ่งแสดงค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์รวมทุกชุดข้อมูลของแต่ละวิธีการกรองร่วม ผลที่ได้จะเห็นว่าสำหรับชุดข้อมูล Movie Lens นั้น วิธีการกรองร่วมแบบดั้งเดิมนั้นมีประสิทธิภาพการทำนายน้อยที่สุดโดยมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.80430 รองลงมาคือวิธีการถ่วงน้ำหนักเบส์แบบสามัญ และวิธีการถ่วงน้ำหนักเบส์แบบสามัญโดยการแปลงลาปลาซเพื่อแก้ไขค่าศูนย์โดยมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.79841 และ 0.79678 ตามลำดับ และสุดท้ายคือวิธีการถ่วงน้ำหนักเบส์แบบสามัญโดยการแปลงลาปลาซทั้งชุดข้อมูลมีประสิทธิภาพการทำนายมากที่สุดโดยมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.79383



ภาพที่ 4.2: กราฟเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยของ Movie Lens

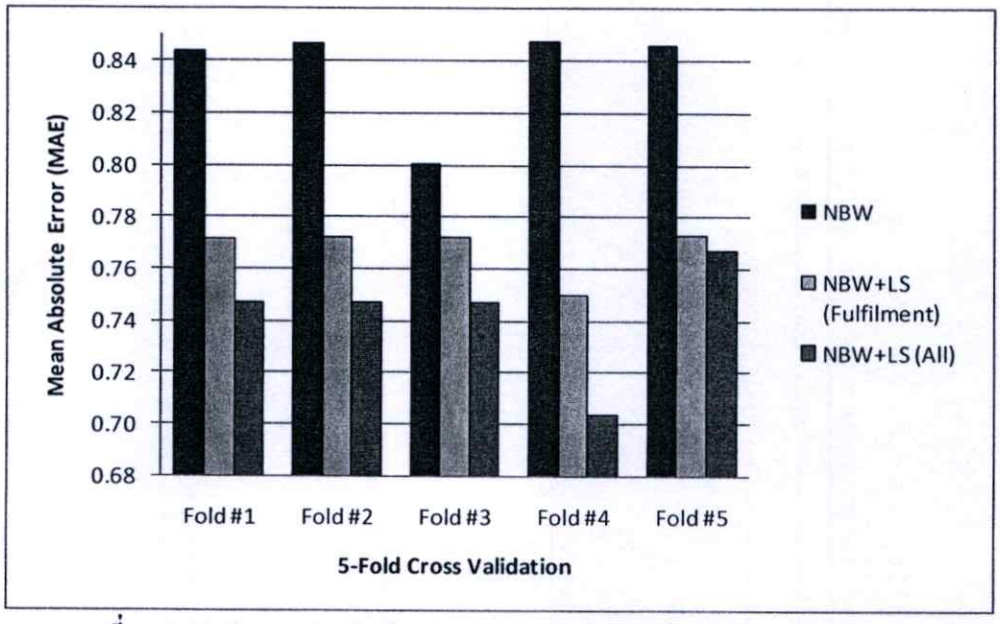
4.2.2 ผลการทดลองด้วยชุดข้อมูลทดสอบ Book Crossing

การทดสอบประสิทธิภาพในการทำนายเปรียบเทียบกันในแต่ละชุดข้อมูลด้วยชุดทดสอบ Book Crossing ระหว่างเทคนิควิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญที่นำเสนอทั้งแบบใช้การแปลงลาปลาซและแบบไม่ใช้การแปลงลาปลาซ ผลการทดลองค่าความคลาดเคลื่อนสัมบูรณ์ที่ได้แสดงดังตารางที่ 4.2

ตารางที่ 4.2: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบย่อยของ Book Crossing

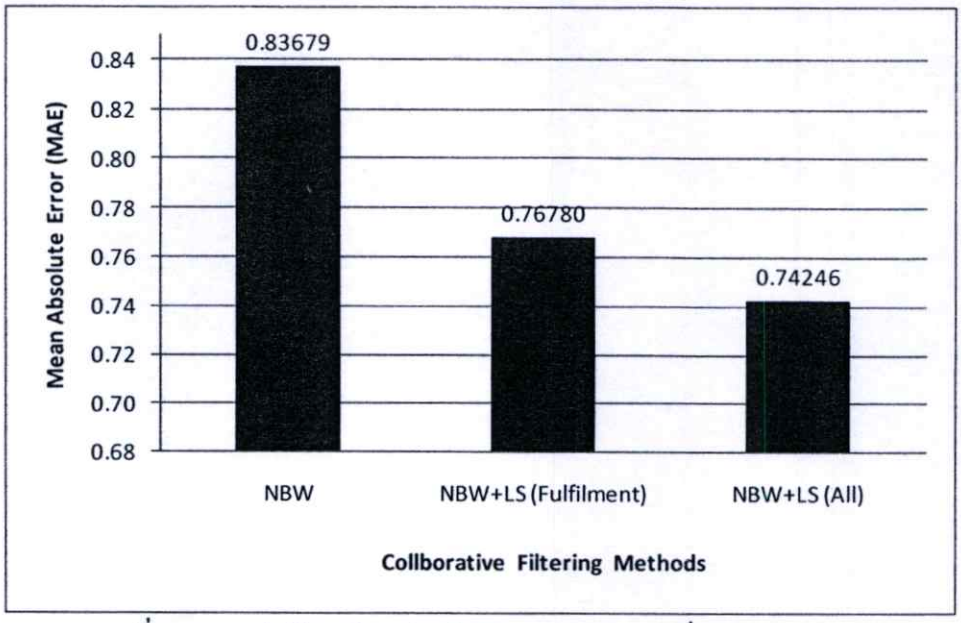
ชุดข้อมูลทดสอบย่อย Book Crossing	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญ	0.84387	0.84660	0.80064	0.84739	0.84598
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญด้วยการแปลงลาปลาซแก้ไขค่าที่เป็นศูนย์	0.77125	0.77255	0.77255	0.74966	0.77298
เทคนิคการถ่วงน้ำหนักเบส์แบบสามัญด้วยการแปลงลาปลาซ	0.74680	0.74718	0.74718	0.70381	0.76735

ค่าที่ได้จากการทดลองจะเห็นได้ว่าประสิทธิภาพการทำนายวัดโดยค่าความคลาดเคลื่อนสัมบูรณ์ของความนิยมที่ทำนายกับความนิยมจริงในระบบโดยใช้วิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ (Naïve Bayes Weighing: NBW) เปรียบเทียบการทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซ (Naïve Bayes Weighing with Laplace Smoothing: NBW+LS) โดยการทดลองใช้แก้ไขเฉพาะค่าข้อมูลที่เป็นศูนย์ (NBW+LS – Fulfillment) และทดลองด้วยการใช้กับทั้งชุดข้อมูล (NBW+LS - All) นั้น สำหรับชุดข้อมูลย่อยทุกชุดข้อมูลนั้น มีผลการทดลองไปในทิศทางเดียวกัน ก็คือ การทำนายความนิยมด้วยเทคนิคการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซสำหรับทั้งชุดข้อมูลให้ประสิทธิภาพในการทำนายความนิยมได้ดีที่สุด และรองลงมาคือ การถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซเพื่อแก้ค่าศูนย์ และการถ่วงน้ำหนักเบส์แบบสามัญที่แบบไม่ได้ใช้การแปลงลาปลาซตามลำดับ ภาพที่ 4.3 แสดงการเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์แต่ละชุดข้อมูลย่อยของชุดข้อมูล Book Crossing แยกตามเทคนิคที่ใช้



ภาพที่ 4.3: กราฟแสดงประสิทธิภาพการทำนายแต่ละชุดข้อมูลของ Book Crossing

เมื่อนำค่าที่ได้จากการทดลองแต่ละชุดข้อมูลมาหาค่าเฉลี่ยรวม แน่นอนว่าค่าที่ได้ก็จะเป็นไปในทิศทางเดียวกัน ซึ่งการทำนายที่ได้เมื่อใช้การแปลงลาปลาซร่วมกับความน่าจะเป็นเบส์แบบสามัญ จะเห็นได้ว่าทำให้เพิ่มประสิทธิภาพในการทำนายได้ค่อนข้างมากอย่างเห็นได้ชัด ภาพที่ 4.4 แสดงค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ที่ได้เฉลี่ยรวมทั้งชุดข้อมูล Book Crossing



ภาพที่ 4.4: กราฟเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยของ Book Crossing

บทที่ 5
สรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเทคนิคทางเลือกที่ใช้ในการทำนายความนิยมในระบบช่วยแนะนำโดยการกรองร่วมโดยใช้ทฤษฎีความน่าจะเป็นของเบส์แทนการใช้การการกรองร่วมแบบดั้งเดิม (Traditional Collaborative Filtering) ที่ใช้ค่าความคล้ายของพฤติกรรมการให้ความนิยมของผู้ใช้ในระบบ โดยมีแนวคิดที่ว่า สินค้าหรือบริการที่ดีและมีคุณภาพก็ย่อมจะมีความน่าจะเป็นที่จะได้รับความนิยมจากผู้บริโภคมากกว่าสินค้าหรือบริการที่ไม่มีคุณภาพ ดั้งนั้นหากเรานำเอาทฤษฎีความน่าจะเป็นเข้ามาทำเป็นแบบจำลองในการทำนายความนิยมก็น่าจะให้ประสิทธิภาพได้ดีในระดับหนึ่ง

เทคนิคที่ใช้ทฤษฎีความน่าจะเป็นของเบส์ในการทำนายค่าความนิยมในงานวิจัยนี้เรียกว่า การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ (Collaborative Filtering using Naïve Bayes Weighing) โดยสร้างแบบจำลองการทำนายโดยการหาค่าความน่าจะเป็นที่ผู้ใช้จะให้ความนิยมกับสินค้าในระบบ โดยหาความน่าจะเป็นในทุกค่าความนิยมที่ผู้ใช้สามารถจะให้ได้ แล้วนำค่าความน่าจะเป็นที่ได้เหล่านั้นคูณกับค่าความนิยมนั้น ๆ และนำผลรวมของผลลัพธ์ที่ได้มาใช้เป็นค่าทำนายของงานวิจัยนี้ โดยมีการใช้การจัดกลุ่มแบบเคมีน (K-Mean Clustering) และการแปลงลาปลาซ (Laplace Smoothing) เพื่อแก้ไขค่าศูนย์ที่อาจเกิดขึ้นได้จากการคำนวณความน่าจะเป็น โดยชุดข้อมูลที่นำมาวิจัยนั้นมีอยู่ด้วยกัน 2 ชุด ได้แก่ Movie Lens ซึ่งเป็นชุดข้อมูลที่นิยมไปใช้เป็นชุดข้อมูลทดลองในสายงานนี้จำนวนมากเป็นชุดข้อมูลที่มีขนาดเล็กเพียง 100,000 ข้อมูลความนิยม และ Book Crossing ซึ่งไม่ค่อยพบเห็นในการทดลองมากนักเนื่องจากเป็นชุดข้อมูลที่มีขนาดใหญ่มากถึง 1,149,780 ข้อมูลความนิยม วัดประสิทธิภาพการทำนายด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ซึ่งก็คือผลรวมของค่าการทำนายที่คลาดเคลื่อนไปของแต่ละข้อมูลความนิยมแล้วหารด้วยจำนวนข้อมูลทั้งหมดที่เราทดลอง โดยหากค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ยิ่งน้อยลงเท่าไรแสดงว่าการทำนายข้อมูลความนิยมนั้นมีความแม่นยำมากขึ้นเท่านั้น

ผลการทดลองค่าทำนายที่ได้จากการใช้การถ่วงน้ำหนักเบส์แบบสามัญพบว่าเกิดปัญหาค่าศูนย์เกิดขึ้น เนื่องจากสินค้าบางตัวไม่ได้ถูกให้ความนิยมมาก่อน และเกิดขึ้นมากตามไปด้วยโดยเฉพาะกับฐานข้อมูลที่มีขนาดใหญ่ งานวิจัยนี้ทดลองใช้การจัดกลุ่มแบบเคมีนมาช่วยแก้ไขปัญหาค่าศูนย์ในชุดข้อมูล Movie Lens แต่ผลที่ได้ไม่ชัดเจน เพราะไม่ให้ประสิทธิภาพที่ดีขึ้นอย่างเห็นได้ชัด น่าจะเป็นเพราะข้อมูลที่นำมาจัดกลุ่มสินค้าที่มีอยู่ในชุดข้อมูล Movie Lens นั้น เป็นเพียงชนิดของภาพยนตร์ซึ่งไม่มีผลในแง่ของคุณภาพของภาพยนตร์ที่มีอิทธิพลต่อความชอบหรือไม่ชอบของผู้บริโภคมากนัก จึงได้ใช้

วิธีอีกวิธีหนึ่งเข้ามาทดสอบการแก้ไขปัญหาค่าศูนย์นี้แทน โดยใช้การแปลงลาปลาซซึ่งผลลัพธ์ที่ได้จากการทดลองกับชุดข้อมูล Movie Lens พบว่าประสิทธิภาพในการทำนายมีแนวโน้มที่ดีขึ้น และดีกว่าการใช้ร่วมกับการจัดกลุ่มแบบเคมีน จากนั้นได้นำการแปลงลาปลาซมาทดสอบกับชุดข้อมูล Book Crossing จะเห็นได้ว่าประสิทธิภาพในการทำนายนั้นดีขึ้นมากอย่างชัดเจน สรุปได้ว่าการทำนายความนิยมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซนั้น สามารถใช้เป็นแบบจำลองหนึ่งในการทำนายความนิยมได้ และจะมีประสิทธิภาพที่ดีขึ้นเมื่อข้อมูลมีขนาดใหญ่ขึ้น

การวิจัยนี้แสดงให้เห็นว่าแบบจำลองการทำนายความนิยมในระบบช่วยแนะนำด้วยการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญร่วมกับการแปลงลาปลาซที่นำเสนอนั้นสามารถเป็นทางเลือกหนึ่งเพื่อใช้ในการทำนายความนิยมได้ โดยมีการคำนวณที่ไม่ยุ่งยากซับซ้อนมากนัก ในด้านของประสิทธิภาพมีแนวโน้มที่จะมีประสิทธิภาพมากขึ้นเมื่อใช้กับฐานข้อมูลที่มีขนาดใหญ่ขึ้น แสดงให้เห็นว่าแบบจำลองการทำนายความนิยมที่นำเสนอนี้จะทำงานได้ดีขึ้น เมื่อมีข้อมูลความนิยมเข้ามาในระบบมากขึ้นตามไปด้วย

บรรณานุกรม

[1] Alpaydin, Ethem. 2004. "Introduction To Machine Learning". Cambridge. Massachusetts: MIT Press.

[2] A.M. Rashid, I. Albert, D. Cosley, S.K. Lam, S.M. McNee, J.A. Konstan, and J. Riedl. 2002. "Getting to Know You: Learning New User Preferences in Recommender Systems". In Proceedings of The 2002 International Conference on Intelligent User Interfaces (IUI 2002). San Francisco CA.

[3] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. "Item-based collaborative filtering recommendation algorithms", In WWW '01: Proceedings of the 10th international conference on World Wide Web (ACM), Pages: 285–295. New York, USA.

[4] Emmanouil Vozalis, Konstantinos G. Margaritis. 2003. "Analysis of Recommender Systems' Algorithms".

[5] Felicia Poe. 2005. "Do You Have Any Recommendation?: An Introduction To Recommender System".

[6] J.MacQUEEN. 1967. "Some Methods for Classification and Analysis of Multivariate Observations", Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press).

[7] Jaldert Rombouts and Tessa Verhoef. "A Simple Hybrid Movie Recommender System".

[8] Joel Bennett. 2006. "A survey of SOM and Recommender techniques", Golisano College of Computing and Information Science.

[9] Jun Wang1, Arjen P. de Vries1,2, Marcel J.T. Reinders. 2006. "Unifying Userbased and Itembased Collaborative Filtering Approaches by Similarity Fusion" SIGIR'06, Seattle, Washington, USA, ACM.

[10] Kevin P. Murphy. 2006. "Naive Bayes classifiers".

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. 2009. "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.

[12] Melville, P., Mooney, R. J. and Nagarajan, R. 2002. "Content-Boosted Collaborative Filtering for Improved Recommendations", In: Proc. of the 18th Natl. Conf. on AI (AAAI-2002), Pages: 187-192.

[13] Montane and Miquel. 2003. "A Taxonomy of Recommender Agents on the Internet", *Artificial Intelligence Review*, Volume 19, Issue 4 (June 2003), Pages: 285 – 330.

[14] Mooney and Raymond J. 2000. "Content-based Book Recommending using Learning for Text Categorization", *International Conference on Digital Libraries Proceedings of the fifth ACM conference on Digital libraries, San Antonio*, Texas, USA, Pages: 195 – 204.

[15] Ng, A.Y. & Jordan, M. I. 2002. On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems, Ng, A.Y., and Jordan, M.

[16] Robin Burke. 2002. "Hybrid Recommender Systems: Survey and Experiments", User Modeling and User-Adapted Interaction.

[17] Robin Burke. 2007. "Hybrid Web Recommender Systems", the Adaptive Web, LNCS 4321, Pages: 377 – 408.

[18] Torres Roberto. 2004. "Enhancing digital libraries with TechLens+", *International Conference on Digital Libraries Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries Tuscon*, AZ, USA. Pages: 228 – 236.

[19] Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. "A Survey of Collaborative Filtering Techniques", *Advances in Artificial Intelligence*, Volume 2009.

การกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญ
Collaborative Filtering with Naïve Bayes Weighing

กรนัฏฐ์ หล่อวิทยาเลิศนภา[1] และ วีระ บุญจริง[2]
ห้องปฏิบัติการวิศวกรรมระบบซอฟต์แวร์ สาขาวิทยาการคอมพิวเตอร์,
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพ 10520
E-mail: korranatl@hotmail.com[1], kbveera@kmitl.ac.th[2]

บทคัดย่อ

การกรองร่วม (Collaborative Filtering: CF) เป็นเทคนิคอย่างหนึ่งที่นำมาใช้ในระบบช่วยแนะนำ (Recommender System) ซึ่งเป็นระบบที่ช่วยแนะนำสินค้าหรือบริการให้กับลูกค้าในระบบพาณิชย์อิเล็กทรอนิกส์ เพื่อให้ง่ายต่อการเข้าถึงข้อมูลที่ลูกค้าต้องการ เมื่อเผชิญกับฐานข้อมูลที่มีขนาดใหญ่ งานวิจัยนี้นำเสนอวิธีการทำนายความนิยมวิธีใหม่โดยใช้ทฤษฎีความน่าจะเป็นเบส์แบบสามัญมาใช้ในการถ่วงน้ำหนักค่าความนิยมและใช้ผลรวมเชิงเส้นของค่าที่ได้เป็นตัวทำนายความนิยมนั้น ผลลัพธ์ที่ได้พบว่าวิธีที่นำเสนอมีประสิทธิภาพดีขึ้นในการทำนายความนิยม

คำสำคัญ: ระบบช่วยแนะนำ, การกรองร่วม, ความน่าจะเป็นเบส์แบบสามัญ, การจัดกลุ่มแบบเคมีน, การทำนายความนิยม

Abstract

Collaborative filtering is a technique that is used for recommender system which is a system for recommending products or service to customers that confront with a large amount of data on e-commerce system. This paper proposes a new preferences prediction method using naïve Bayes theorem for weighing preference in collaborative filtering system. Linear combination is used to combine all weighed preference for prediction. The results show that using naïve Bayes weighing outperforms the others.

Keywords: Recommender System, K-Mean Clustering, Collaborative Filtering, Naïve Bayes Theorem, Preference Prediction

1. คำนำ

ระบบช่วยแนะนำ (Recommender System) เป็นกระบวนการวิธีการเรียนรู้ของระบบเพื่อช่วยแนะนำสินค้าหรือบริการที่เหมาะสมหรือน่าสนใจให้กับลูกค้า [11] ระบบนี้กลายมาเป็นส่วนสำคัญของระบบธุรกิจพานิชย์อิเล็กทรอนิกส์ เพื่อช่วยให้ลูกค้าสามารถเข้าถึงข้อมูลที่ตัวเองสนใจได้ง่ายมากขึ้น เนื่องจากฐานข้อมูลของสินค้าหรือบริการนั้นมักมีขนาดใหญ่ เทคนิคที่นิยมใช้ในระบบช่วยแนะนำจะอยู่ในรูปแบบของการกรองร่วม (Collaborative Filtering) หรือการกรองเนื้อหา (Content-based Filtering) หรือใช้ทั้งสองรูปแบบร่วมกันที่เรียกว่า

ระบบช่วยแนะนำแบบผสม (Hybrid Recommender System) [3] เทคนิคการกรองร่วมนั้นใช้การพิจารณาคะแนนความนิยมของผู้ใช้ที่มีต่อสินค้าในระบบ และนำไปพิจารณากับสินค้าที่ผู้ใช้ยังไม่ให้ความนิยมไว้ แต่มีความคล้ายกับสินค้าที่ผู้ใช้นั้นเคยใช้แล้ว โดยใช้ตัววัดความคล้าย (Similarity Measurement) ในการถ่วงน้ำหนักของสินค้าตัวที่ถูกระบุความนิยม เพื่อใช้ในการทำนายความนิยม (Preference Predition) ให้กับสินค้าอีกตัวหนึ่ง ซึ่งการวัดความคล้ายนี้จะแปลงความนิยมในระบบให้อยู่ในรูปของเวกเตอร์ และใช้ใช้ตัววัดความคล้ายแบบโคซายน์ (Cosine) ตัววัดความคล้ายแบบสหสัมพันธ์ (Correlation) และตัววัดความคล้ายแบบโคซายน์ปรับแก้ (Adjusted Cosine) จากนั้นใช้การพิจารณาย่านใกล้เคียงที่สุด K ตัว ในการเลือกสินค้าที่มีความคล้ายมากที่สุด K ตัวมาทำการถ่วงน้ำหนักและหาผลรวมเชิงเส้น ซึ่งการวัดความคล้ายอาจจะวัดโดยอ้างอิงจากผู้ใช้ (User-based Collaborative Filtering) หรืออ้างอิงจากสินค้า (Item-based Collaborative Filtering) หรืออาจจะใช้สองเทคนิคนี้ร่วมกันก็ได้ [1], [7] และเนื่องจากเทคนิคการกรองร่วมนี้ใช้ข้อมูลความนิยมที่มีอยู่ในระบบ ดังนั้นจุดด้อยที่สำคัญของเทคนิคนี้คือ ปัญหาการเริ่มต้นได้ยาก (Cold Start Problem) หรืออาจจะเรียกว่าปัญหาการให้ความนิยมครั้งแรก (First Rater Problem) ซึ่งทำให้เมื่อผู้ใช้เข้ามาในระบบแล้วจะไม่ได้รับการแนะนำสินค้าใด ๆ เลย หรือสินค้าที่เข้ามาในระบบครั้งแรกนั้นจะไม่ได้ถูกแนะนำให้กับผู้ใช้รายใดเลยเช่นกัน จุดด้อยสำคัญอีกประการหนึ่งก็คือ ปัญหาข้อมูลเบาบาง (Sparsity Problem) ซึ่งทำให้การวัดความคล้ายของคะแนนความนิยมในระบบเป็นไปได้ยาก เนื่องจากการกระจายตัวของข้อมูลมีสูงและหากผู้ใช้ไม่ได้ให้ความนิยมของสินค้าที่ซ้ำกันเลยก็ไม่สามารถวัดความคล้ายของข้อมูลความนิยมนั้นได้ ทำให้การทำนายความนิยมด้วยวิธีนี้เป็นไปได้ยาก

เทคนิคการกรองเนื้อหา จะใช้ข้อมูลที่ได้จากตัวสินค้ามาพิจารณา โดยจะแนะนำสินค้าที่มีรายละเอียดเนื้อหาของตัวสินค้าที่เหมือนหรือคล้ายกับสินค้าที่ผู้ใช้เคยซื้อหรือใช้บริการให้กับผู้ใช้ จุดด้อยของวิธีนี้ คือ หากเป็นสินค้าที่ไม่มีการระบุรายละเอียดเนื้อหาจะทำให้ไม่สามารถแนะนำได้ จุดด้อยอีกอย่างหนึ่งคือ ระบบไม่สามารถแนะนำสินค้าที่มีรายละเอียดเนื้อที่แตกต่างออกไปแก่ผู้ใช้ได้เลย [5] ซึ่งทำให้จำกัดขอบเขตของการแนะนำที่แคบเกินไป

ระบบช่วยแนะนำแบบผสมใช้เทคนิคการกรองร่วมและเทคนิคการกรองเนื้อหามาใช้ร่วมกัน เพื่อชดเชยจุดด้อยของแต่ละวิธี [3] กล่าวคือ เทคนิคการกรองร่วมจะแนะนำสินค้าที่เป็นที่นิยมให้กับผู้ใช้ และเทคนิคการกรองเนื้อหาจะแนะนำสินค้าที่ตรงกับรสนิยมให้กับผู้ใช้งานวิจัยในระบบช่วยแนะนำนั้นเป็นไปได้ทั้งการทำนายความนิยมของผู้ใช้ (Preference Prediction) หรือการแนะนำสินค้า N อันดับ (Top-N Recommendation) [2]

1.1 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่นำทฤษฎีความน่าจะเป็นเบส์แบบสามัญมาใช้ส่วนใหญ่ใช้ในระบบช่วยแนะนำแบบผสมโดยใช้ตัวจำแนกประเภทเบส์แบบสามัญ (Naïve Bayes Classifier) ในส่วนของการกรองเนื้อหา เช่น ระบบช่วยแนะนำแบบผสมด้วยการใช้ตัวจำแนกประเภทเบส์แบบสามัญในการคัดกรองเนื้อหาเพื่อทำนายความนิยมของผู้ใช้เพิ่มเติมบางส่วน ทำให้มีข้อมูลความนิยมในระบบมากขึ้น แก้ปัญหาข้อมูลเบาบาง จากนั้นนำผลลัพธ์ที่ได้ไปคำนวณด้วยวิธีพิจารณาย่านใกล้เคียงที่สุด K ตัวและถ่วงน้ำหนักด้วยความคล้ายของเวกเตอร์ความนิยมในการกรองร่วมเพื่อทำนายความนิยมของผู้ใช้ที่เหลือในระบบทั้งหมด [10] นอกจากนี้ยังมีงานวิจัยที่ใช้ระบบช่วยแนะนำแบบผสมที่ใช้ผลรวมเชิงเส้นของทั้งการกรองร่วมและการกรองเนื้อหาด้วยอัตราส่วน 50:50 โดยการกรองเนื้อหาพิจารณาจากคำสำคัญ ชนิดของภาพยนตร์ และรายชื่อนักแสดงนำ ทำนายความนิยมด้วยตัวจำแนกเบส์แบบสามัญ และส่วนของการกรองร่วมใช้วิธีพิจารณาย่านใกล้เคียงที่สุด K ตัวและถ่วงน้ำหนักความคล้ายเวกเตอร์ความนิยมในการทำนาย [5]

1.2 แนวคิดของงานวิจัย

งานวิจัยนี้นำเสนอตัวถ่วงน้ำหนักใหม่เพื่อใช้ในการทำนายความนิยมในเทคนิคการกรองร่วม โดยใช้ความน่าจะเป็นที่ผู้ใช้คนหนึ่งจะให้ความนิยมกับสินค้าชิ้นหนึ่งด้วยคะแนนต่าง ๆ ตามทฤษฎีความน่าจะเป็นเบส์แบบสามัญเป็นตัวถ่วงน้ำหนักของค่าความนิยม และใช้ผลรวมเชิงเส้นของค่าที่ได้เป็นค่าทำนายความนิยมของผู้ใช้ที่มีต่อสินค้านั้น ๆ การทำนายและแก้ปัญหาการเริ่มต้นได้ยากด้วยการจัดกลุ่มสินค้าในระบบเป็นกลุ่มต่าง ๆ สินค้าที่เข้ามาใหม่และยังไม่ถูกให้คะแนนความนิยมจะถูกจัดกลุ่มร่วมกับสินค้าที่มีอยู่เดิมในระบบและถ่วงน้ำหนักจากความน่าจะเป็นของระบบที่มีต่อสินค้าในกลุ่มนั้น ๆ แทนซึ่งต่างกับงานวิจัยอื่นที่ใช้การวัดความคล้ายในการถ่วงน้ำหนัก ซึ่งวิธีที่นำเสนอนี้ใช้วิธีการคำนวณที่ง่ายและพิจารณาจากข้อมูลความนิยมที่มีทั้งระบบ ทำให้ไม่ต้องประสบกับปัญหาข้อมูลเบาบางอีกด้วย

เนื้อหาของบทความวิจัยนี้ประกอบด้วย ทฤษฎีที่เกี่ยวข้องในส่วนที่ 2 วิธีที่นำเสนอในส่วนที่ 3 ผลลัพธ์ที่ได้จากการทดลองในส่วนที่ 4 สรุปผลและงานวิจัยในอนาคตในส่วนที่ 5

2. ทฤษฎีที่ใช้ในงานวิจัย

2.1 ทฤษฎีความน่าจะเป็นเบส์แบบสามัญ

ความน่าจะเป็นเบส์แบบสามัญ (Naïve Bayes: NB) สามารถใช้เป็นแบบจำลองอย่างหนึ่งในการจำแนกประเภทโดยใช้ทฤษฎีความน่าจะเป็นของเบส์ ที่ตั้งอยู่บนสมมติฐานว่าแต่ละเหตุการณ์ f_i เกิดขึ้นอย่างอิสระ [8] โดยทฤษฎีความน่าจะเป็นของเบส์นั้น นิยามได้ดังนี้

$$p(c|f_i) = \frac{p(c)p(f_i|c)}{p(f_i)} \qquad (1)$$

และจากสมมติฐานของเบส์สามัญที่ว่าแต่ละเหตุการณ์ f_i สามารถเกิดขึ้นอย่างอิสระ นิยามดังนี้

$$p(c|f_i) = p(c)\prod_{i=1}^{n} p(f_i|c) \qquad (2)$$

ตัวจำแนกประเภทเบส์แบบสามัญคำนวณความน่าจะเป็นที่จะเกิด c เมื่อกำหนด f_i และใช้ค่า c ที่มีความน่าจะเป็นสูงสุดเป็นประเภทที่จำแนกในแบบจำลองนี้

$$classify(f_i \dots f_n) = max\, p(C = c)\prod_{i=1}^{n} p(F_i = f_i|C = c) \qquad (3)$$

เมื่อ p คือความน่าจะเป็นของเหตุการณ์
c คือประเภทที่จำแนก
f_i คือเหตุการณ์ที่ i
n คือจำนวนเหตุการณ์ทั้งหมด

2.2 การจัดกลุ่มแบบเคมีน (K-mean Clustering)

การจัดกลุ่มเป็นขั้นตอนวิธีที่ใช้ในการจัดกลุ่มข้อมูลที่ความคล้ายคลึงกันไว้ในกลุ่มเดียวกัน [4] วิธีหนึ่งที่ง่ายและเป็นที่นิยมในการจัดกลุ่มของข้อมูล คือ การจัดกลุ่มแบบเคมีน ซึ่งเป็นการแบ่งข้อมูลเป็นกลุ่มตามจำนวนที่เรากำหนด โดยการสุ่มเลือกข้อมูลตามจำนวนที่เราต้องการเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้น และวัดระยะห่างระหว่างข้อมูลที่เหลือกับจุดศูนย์กลาง เพื่อที่จะตัดสินว่าข้อมูลนั้นควรอยู่ในกลุ่มใด ทำซ้ำขั้นตอนนี้ไปจนกว่าจะครบทุกข้อมูลที่ต้องการจัดกลุ่ม จากนั้นคำนวณจุดศูนย์กลางของแต่ละกลุ่มข้อมูลใหม่ และทำการวัดระยะห่างและจัดข้อมูลกลุ่มข้อมูลใหม่อีกครั้ง จนกว่าจุดศูนย์กลางจะไม่มีการเปลี่ยนแปลง ส่วนข้อมูลที่นำมาจัดกลุ่มจะถูกแปลงให้อยู่ในรูปของเวกเตอร์ และใช้ระยะทางแบบยูคลิเดียนในการวัดระยะห่างระหว่างข้อมูล ดังสมการที่ 4

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i + q_i)^2} \qquad (4)$$

เมื่อ d คือ ระยะห่างระหว่างเวกเตอร์
n คือ จำนวนคุณสมบัติของสินค้า
p_i และ q_i คือ สินค้าชิ้นที่ i

3. วิธีที่นำเสนอ

งานวิจัยนี้เสนอตัวถ่วงน้ำหนักที่ได้จากทฤษฎีความน่าจะเป็นเบส์แบบสามัญเพื่อใช้ในการทำนายความนิยมในการคัดกรองร่วม ซึ่งผลลัพธ์ที่ได้ คือความนิยมของผู้ใช้คนหนึ่งที่มีต่อสินค้าตัวหนึ่ง โดยมีค่าตั้งแต่ 1 ถึง 5 โดยที่ 1 หมายถึง ผู้ใช้มีความชอบในตัวสินค้านั้นน้อยที่สุด และ 5 หมายถึง ผู้ใช้มีความชอบในตัวสินค้านั้นมากที่สุด ตามลำดับ รายละเอียดของการคำนวณความนิยมด้วยวิธีที่นำเสนอมีดังต่อไปนี้

3.1 การคำนวณความนิยม

การคำนวณคะแนนความนิยมทำได้โดยการหาความน่าจะเป็นที่ผู้ใช้ในระบบจะให้คะแนนความนิยม 1 – 5 กับสินค้าตัวหนึ่งด้วยทฤษฎีความน่าจะเป็นเบส์แบบสามัญ ดังสมการที่ 2 ตัวอย่างข้อมูลที่ใช้ในการคำนวณแสดงในตารางที่ 1

ตารางที่ 1: ตัวอย่างข้อมูลการให้ความนิยมสินค้าของผู้ใช้ในระบบ

ผู้ใช้ \ สินค้า	1	2	3	4
1	2	3	4	2
2	4	3	2	4
3	4	2	4	?

ตัวอย่างการคำนวณความน่าจะเป็นตามทฤษฎีความน่าจะเป็นเบส์แบบสามัญดังสมการที่ 5 – 9 และตารางที่ 2 แสดงค่าทำนายความนิยมที่ผู้ใช้คนที่ 3 ให้กับสินค้าชิ้นที่ 4

$$p(c=1|User=3, Item=4)$$
$$= p(c=1)p(User=3|c=1)\, p(Item=4\,|c=1)$$
$$= \left(\frac{0}{11}\right)\left(\frac{0}{3}\right)\left(\frac{0}{2}\right) = 0 \qquad (5)$$

$$p(c=2|User=3, Item=4)$$
$$= p(c=2)p(User=3|c=2)\, p(Item=4\,|c=2)$$
$$= \left(\frac{4}{11}\right)\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = 0.0606 \qquad (6)$$

$$p(c=3|User=3, Item=4)$$
$$= p(c=3)p(User=3|c=3)\, p(Item=4\,|c=3)$$
$$= \left(\frac{2}{11}\right)\left(\frac{0}{3}\right)\left(\frac{1}{2}\right) = 0 \qquad (7)$$

$$p(c=4|User=3, Item=4)$$
$$= p(c=4)p(User=3|c=4)\, p(Item=4\,|c=4)$$
$$= \left(\frac{5}{11}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right) = 0.1515 \qquad (8)$$

$$p(c=5|User=3, Item=4)$$
$$= p(c=5)p(User=3|c=5)\, p(Item=4\,|c=5)$$
$$= \left(\frac{0}{11}\right)\left(\frac{0}{3}\right)\left(\frac{0}{2}\right) = 0 \qquad (9)$$

ตารางที่ 2: ตัวอย่างการทำนายความนิยมด้วยวิธีที่นำเสนอ

Rating	1	2	3	4	5
Weighing	0	0.0606	0	0.1515	0
Rating* Weighing	0	0.1818	0	0.6060	0
Rating Prediction	$\frac{\sum(\text{Rating} * \text{Weighing})}{\sum\text{Weighing}} = 3.4286$				

3.2 การคำนวณความนิยมสำหรับสินค้าใหม่

สินค้าใดที่ยังไม่เคยถูกให้ความนิยมจากผู้ใช้คนใดเลย หากนำไปคำนวณตามสมการข้างต้นจะทำให้หาค่าไม่ได้ เนื่องจากถูกหารด้วยค่าศูนย์ ดังนั้นก่อนที่ทำนายความนิยมให้กับสินค้าจะนำสินค้านั้นไปจัดกลุ่มร่วมกันในระบบ และใช้กลุ่มที่ถูกจัดนั้นมาคำนวณตามวิธีในหัวข้อ 3.1 ต่อไป โดยการจัดกลุ่มทำโดยการพิจารณาจากเนื้อหาข้อมูลสินค้า

4. การทดลอง

4.1 รายละเอียดของการทดลอง

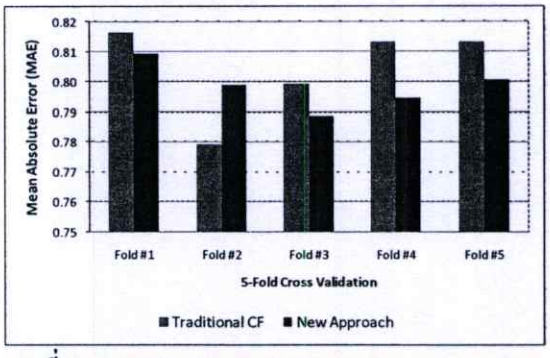
ชุดข้อมูลที่ใช้ทดลองในงานวิจัยนี้ คือ MovieLens ประกอบด้วยข้อมูลความนิยมจำนวน 100,000 ชุด จากผู้ใช้ 943 คน ที่มีต่อภาพยนตร์จำนวน 1,682 เรื่อง ทำนายความนิยมด้วยวิธีการวัดน้ำหนักด้วยความน่าจะเป็นเบส์แบบสามัญ สำหรับตัววัดประสิทธิภาพของขั้นตอนวิธี ใช้ค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ดังสมการที่ 9 และทำการทดลองโดยวิธีการตรวจสอบข้ามชุด K ชุด (K-fold Cross Validation) งานวิจัยนี้แบ่งข้อมูลทดสอบเป็น 5 ชุด แต่ละรอบการทำงานข้อมูลที่ไม่ได้นำมาใช้ทดสอบจะใช้เป็นข้อมูลสอนทั้งหมด

$$MAE = \frac{\sum_{i=1}^{n} |P_i - Q_i|}{n} \qquad (9)$$

เมื่อ P_i คือค่าความนิยมที่ทำนาย
Q_i คือค่าความนิยมที่ถูกต้อง
n คือจำนวนความนิยมที่ทำนายทั้งหมด

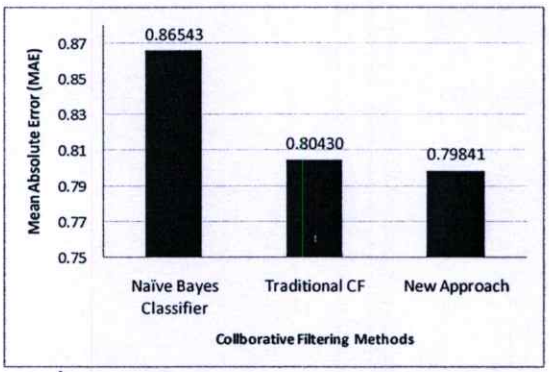
4.2 ผลการทดลองประสิทธิภาพในการทำนาย

ผลการทดลองเปรียบเทียบประสิทธิภาพการทำนายในแต่ละชุดข้อมูล ระหว่างวิธีการกรองร่วมด้วยการถ่วงน้ำหนักเบส์แบบสามัญที่นำเสนอ กับเทคนิคการกรองร่วมแบบดั้งเดิม (Traditional Collaborative Filtering) ที่ใช้การถ่วงน้ำหนักด้วยการวัดความคล้ายและพิจารณา 20 ย่านใกล้เคียง 20 ตัว [13] แสดงดังภาพที่ 1 จะเห็นได้ว่าประสิทธิภาพการทำนายของวิธีที่นำเสนอดีกว่าวิธีการกรองร่วมแบบดั้งเดิมถึง 4 ชุดข้อมูลจาก 5 ชุดข้อมูลที่นำมาทดสอบ



ภาพที่ 1: กราฟเปรียบเทียบประสิทธิภาพการทำนายแต่ละชุดข้อมูล

ผลการทดลองเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยรวมทุกชุดข้อมูล ระหว่างวิธีที่นำเสนอกับเทคนิคการกรองร่วมแบบดั้งเดิม เทียบกับการทำนายความนิยมโดยใช้ตัวจำแนกประเภทเบส์แบบสามัญที่คำนวณจากโปรแกรม WEKA [9] แสดงดังภาพที่ 2 จะเห็นได้ว่าการใช้ตัวจำแนกประเภทเบส์แบบสามัญในการทำนายความนิยมของผู้ใช้จะให้ค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ที่สูงกว่าวิธีอื่น เนื่องจากตัวจำแนกประเภทเบส์แบบสามัญนั้นเป็นการทำนายที่ชี้ตรง คือมีค่าการทำนายเป็นจำนวนเต็ม ดังนั้นหากทำนายคลาดเคลื่อนจะทำให้ความคลาดเคลื่อนในการทำนายมีสูงตามไปด้วย และเมื่อใช้ตัววัดน้ำหนักเบส์แบบสามัญในการทำนายความนิยมของระบบช่วยแนะนำพบว่าให้ประสิทธิภาพในการทำนายที่ดีขึ้น



ภาพที่ 2: กราฟเปรียบเทียบประสิทธิภาพการทำนายรวมชุดข้อมูล

5. บทสรุปและงานวิจัยในอนาคต

การทำนายความนิยมในระบบช่วยแนะนำด้วยการกรองร่วมในงานวิจัยนี้ได้เสนอตัวถ่วงน้ำหนักใหม่ในการทำนายความนิยมของผู้ใช้ที่มีต่อสินค้าในระบบช่วยแนะนำ โดยการนำความน่าจะเป็นเบส์แบบสามัญมาใช้เป็นตัวถ่วงน้ำหนัก ผลลัพธ์ที่ได้พบว่ามีประสิทธิภาพในการทำนายที่ดีกว่าวิธีการกรองร่วมด้วยตัวจำแนกประเภทเบส์แบบสามัญที่มีจุดอ่อนอยู่ที่การทำนายออกเป็นค่าที่ชี้ตรงเกินไป และมีประสิทธิภาพที่ดีกว่าการกรองร่วมแบบดั้งเดิม ข้อดีของวิธีที่นำเสนอนี้ คือ ไม่ประสบกับปัญหาข้อมูลเบาบางในระบบ และแก้ไขปัญหาการเริ่มต้นได้ยากโดยใช้การจัดกลุ่มสินค้าใหม่กับสินค้าในระบบเดิม และทำนายโดยถ่วงน้ำหนักความนิยมในระบบที่มีต่อกลุ่มของสินค้านั้นแทน

การวิจัยในอนาคตอาจจะเป็นการประยุกต์ใช้แนวคิดนี้ร่วมกับระบบช่วยแนะนำแบบผสม หรืออาจนำไปใช้ในการเติมเต็มข้อมูลความนิยมในระบบก่อนที่ใช้เทคนิคต่าง ๆ ในการทำนายความนิยมต่อไป

6. เอกสารอ้างอิง

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms", In WWW '01: Proceedings of the 10th international conference on World Wide Web (ACM), pages 285–295, New York, NY, USA, 2001.

[2] Emmanouil Vozalis, Konstantinos G. Margaritis, "Analysis of Recommender Systems' Algorithms", 2003.

[3] Felicia Poe, "Do You Have Any Recommendation?: An Introduction To Recommender System", 2005.

[4] J.MacQUEEN, "Some Methods for Classification and Analysis of Multivariate Observations", Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press), 1967

[5] Jaldert Rombouts, Tessa Verhoef, "A Simple Hybrid Movie Recommender System".

[6] Joel Bennett, "A survey of SOM and Recommender techniques", Golisano College of Computing and Information Science, 2006.

[7] Jun Wang1, Arjen P. de Vries1,2, Marcel J.T. Reinders, "Unifying Userbased and Itembased Collaborative Filtering Approaches by Similarity Fusion" SIGIR'06, Seattle, Washington, USA, 2006. ACM.

[8] Kevin P. Murphy, "Naive Bayes classifiers", 2006.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1, 2009.

[10] Melville, P., Mooney, R. J. and Nagarajan, R., "Content-Boosted Collaborative Filtering for Improved Recommendations", In: Proc. of the 18th Natl. Conf. on AI (AAAI-2002), pp. 187-192. (2002).

[11] Robin Burke, "Hybrid Recommender Systems: Survey and Experiments", User Modeling and User-Adapted Interaction, 2002.

[12] Robin Burke, "Hybrid Web Recommender Systems", the Adaptive Web, LNCS 4321, pp. 377 – 408, 2007.

[13] นิรภัฏ อุทัยฉาย, วีระพงษ์ ปัญญาเหมือง, สุนาท วนไพศาล และวีระ บุญจริง, "ระบบช่วยแนะนำแบบผสมโดยใช้การสืบค้นกฎการเกิดร่วมกันและการจัดกลุ่ม", The 13th National Computer Science and Engineering Conference (NCSEC 2009), pp. 244-250, 2009.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายกรนัฏฐ์ หล่อวิทยาเลิศนภา
วัน เดือน ปีเกิด	4 กันยายน 2526 ที่กรุงเทพมหานคร
ที่อยู่	15 ซอยเพชรเกษม 110 ถ.เพชรเกษม
	แขวงหนองค้างพลู เขตหนองแขม กรุงเทพฯ 10160 โทร.08-4198-5198
ประวัติการศึกษา	2548 วิทยาศาสตรบัณฑิต ภาควิชาคณิตศาสตร์
	มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
	2554 วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณมหารลาดกระบัง
ประสบการณ์การทำงาน	
พ.ศ.2548-2549	ตำแหน่งเว็บมาสเตอร์ บริษัท บัณฑิตไซเบอร์ จำกัด
พ.ศ.2549-2550	ตำแหน่งที่ปรึกษาด้านการจัดหางานไอที บริษัท แมนพาวเวอร์ จำกัด
ปัจจุบัน	ตำแหน่งนักวิเคราะห์ข้อมูลและรายงาน บริษัท ดิจิทอล อัลเคมี่ จำกัด