การทำนายคะแนนสินค้าโดยใช้ค่าความถี่สัมพัทธ์สูงสุดในระบบแนะนำ

PRODUCT SCORE PREDICTION USING HIGHEST RELATIVE FREQUENCY IN RECOMMENDATION SYSTEM

พีระศักดิ์ แซ่ตัน

PEERASUK SAETAN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-SC-M-002-022

การทำนายคะแนนสินค้าโดยใช้ค่าความถี่สัมพัทธ์สูงสุดในระบบแนะนำ

PRODUCT SCORE PREDICTION USING HIGHEST RELATIVE FREQUENCY IN RECOMMENDATION SYSTEM

พีระศักดิ์ แซ่ตั้น

PEERASUK SAETAN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-SC-M-002-022

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PRODUCT SCORE PREDICTION USING HIGHEST RELATIVE FREQUENCY IN RECOMMENDATION SYSTEM

PEERASUK SAETAN

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

2013

KMITL-2013-SC-M-002-022

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2013

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การทำนายคะแนนสินค้าโดยใช้ค่าความถี่สัมพัทธ์สูงสุดในระบบแนะนำ
Product Score Prediction Using Highest Relative Frequency in Recommendation System

นักศึกษา นายพีระศักดิ์ แซ่ตั้น
รหัสประจำตัว 51067505
ปริญญา วิทยาศาสตรมหาบัณฑิต
สาขาวิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.จีรพร	วีระพันธุ์	
ผศ.ดร.กรกช	ประชุมรักษ์	
ดร.ชาคริต	วัชโรภาส	
ผศ.ดร.นวลสวาท	หิรัญสกลวงศ์	

วัน / เดือน/ ปี ที่สอบ 15 พฤษภาคม พ.ศ. 2556 เวลา 13.00 – 16.00 น.
สถานที่สอบ ณ ห้อง 304 ชั้น 3 อาคารจุฬาภรณวลัยลักษณ์ 1

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 27 **เดือน** พ.ค. **พ.ศ.** 56

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อวิทยานิพนธ์	การทำนายคะแนนสินค้าโดยใช้ค่าความถี่สัมพัทธ์สูงสุดในระบบแนะนำ
นักศึกษา	นาย พีระศักดิ์ แซ่ตั้น
รหัสประจำตัว	51067505
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2556
อาจารย์ที่ปรึกษา	ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

บทคัดย่อ

ในปัจจุบันมีสินค้าให้ผู้บริโภคเลือกซื้อได้หลากหลายและมีบริการมากมายในระบบพาณิชย์อิเล็กทรอนิกส์จึงจำเป็นต้องมีระบบแนะนำสินค้าให้แก่ผู้ใช้หรือผู้บริโภค โดยการทำนายคะแนนความชอบสินค้าที่จะทำการแนะนำ ซึ่งวิธีที่นิยมใช้ คือ การกรองข้อมูลผู้ใช้ร่วมแบบเดิม (Collaborative Filtering: CF) แต่วิธีนี้ยังประสบปัญหาความถูกต้องและการประมวลผลเบื้องต้นเมื่อมีการใช้ฐานข้อมูลที่มีขนาดใหญ่ ซึ่งต้องเปรียบเทียบหาค่าความคล้ายคลึงข้อมูลของสมาชิกผู้ใช้ร่วม งานวิจัยนี้จึงนำเสนอวิธีการทำนายคะแนนความชอบสินค้าของผู้ใช้โดยใช้ค่าความถี่สัมพัทธ์สูงสุดของสินค้าที่จะแนะนำมาปรับค่าฉลี่ยของผู้ใช้ เพื่อให้ค่าคะแนนความชอบสินค้าใกล้เคียงกับความต้องการของผู้ใช้มากที่สุด ซึ่งวิธีการใหม่นี้ลดความซับซ้อนในการคำนวณและใช้เวลาในการคำนวณได้รวดเร็วกว่าการกรองข้อมูลแบบผู้ใช้ร่วมแบบเดิม โดยในค่า time complexity ของวิธีที่ใช้ค่าความถี่สัมพัทธ์สูงสุด O(m) หรือ O(n) ซึ่ง m คือจำนวนของผู้ใช้ทั้งหมดที่เคยให้คะแนน และ n คือ จำนวนสินค้าทั้งหมดที่ผู้ใช้เคยให้คะแนน ในขณะที่วิธีกรองข้อมูลผู้ใช้ร่วมแบบเดิม(CF) มีค่า time complexity เท่ากับ O(mn) ผลลัพธ์ที่ได้พบว่าวิธีการใช้ค่าความถี่สัมพัทธ์สูงสุดมีประสิทธิภาพดีกว่า โดยดูจากค่าความคลาดเคลื่อนสัมบูรณ์ซึ่งน้อยกว่าวิธีการกรองข้อมูลแบบผู้ใช้ร่วมแบบเดิม

คำสำคัญ: ระบบช่วยแนะนำ, การกรองข้อมูลผู้ใช้ร่วม, ค่าความถี่สัมพัทธ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Thesis Title	PRODUCT SCORE PREDICTION USING HIGHEST RELATIVE FREQUENCY IN RECOMMENDATION SYSTEM
Student	Mr. Peerasuk SaeTan
Student ID	51067505
Degree	Master of Science
Program	Computer Science
Year	2013
Thesis Advisor	Asst. Prof. Dr. Nualsawat Hiransakolwong

Abstract

Currently, there are so many products and services that consumers choose to buy in electronic commerce. Users require a recommender system to filter only items that hit their interests. Most techniques in recommender system are used in combination Collaborative Filtering (called CF), but these techniques are so complicated and spend too much time to compute the preprocessing step with a large database. This research proposes a new method predicted scores of users' interest by using highest relative frequency in products. This new method does not complicate and can calculate faster than the CF. The proposed approach has time complexity O(m) or O(n) where m represents for the number of users and n represents for the number of products. The CF approach has time complexity O(mn). The experimental results show the new approach gaining more performance with less error and faster computation time than the CF.

Keywords: Recommender System, Collaborative Filtering, Relative Frequency

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VII

สารบัญรูป VIII

บทที่ **1** บทนำ

1.1 ความสำคัญและที่มาของปัญหา 1

1.2 วัตถุประสงค์ 1

1.3 ขอบเขตงานวิจัย 2

1.4 ส่วนประกอบของวิทยานิพนธ์ 2

บทที่ **2** ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ระบบแนะนำ 3

2.1.1 ระบบแนะนำด้วยวิธีการกรองผู้ใช้ร่วม (Collaborative Filtering: CF) 5

2.1.2 ระบบช่วยแนะนำด้วยวิธีการกรองเนื้อหา (Content-based Filtering: CB) 6

2.1.3 ระบบช่วยแนะนำแบบผสม (Hybrid Recommender System) 7

2.2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 7

2.2.1 การแจกแจงความถี่สัมพัทธ์ (Relative Frequency) 7

2.2.2 ค่าฐานนิยม (mode) 8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

2.2.3 ขั้นตอนระบบการแนะนำสินค้าโดยใช้เทคนิคการกรองข้อมูลผู้ใช้ร่วม.....8

2.2.3.1 ขั้นตอนการเตรียมข้อมูล 8

2.2.3.2 ขั้นตอนการค้นหาสมาชิกใกล้เคียง.....9

2.2.3.3 ขั้นตอนการทำนายคะแนนสินค้า 11

2.2.2.4 ขั้นตอนการเตรียมข้อมูล 12

2.3 งานวิจัยที่เกี่ยวข้อง.....12

บทที่ **3** การทำนายคะแนนความชอบโดยใช้ค่าความถี่สัมพัทธ์สูงสุดของสินค้า

3.1 ข้อมูลที่ใช้ในการเรียนรู้และทดสอบ 14

3.2 ตัววัดประสิทธิภาพการทำนายคะแนนความชอบสินค้า.....14

3.3 ปัญหาการทำนายคะแนนความชอบสินค้า 15

3.4 การทำนายคะแนนความชอบด้วยค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมสินค้า (Relative Frequency:RF) 17

3.5 การทำนายคะแนนความชอบด้วยค่าความถี่สัมพัทธ์โดยเปรียบเทียบค่าระหว่างค่าฐานนิยมและค่าเฉลี่ยของผู้ใช้ (Relative Frequency Between Point :RFBP) 21

3.6 การทำนายคะแนนความชอบด้วยค่าเฉลี่ยของสินค้า(Mean Item) 23

บทที่ **4** ผลการทดลอง

4.1 ขั้นตอนการเตรียมชุดข้อมูล.....25

4.2 ผลการทดลอง.....26

4.2.1 ผลการทดลองด้วยชุดข้อมูลทดสอบ Movie Lens.....27

4.2.2 ผลการทดลองด้วยชุดข้อมูลทดสอบ Book Crossing.....28

บทที่ **5** สรุปและข้อเสนอแนะ 31

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

เอกสารอ้างอิง 34

ภาคผนวก งานวิจัยที่ตีพิมพ์ 35

ประวัติผู้เขียน 44

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

หน้า

ตารางที่ 2.1: ตัวอย่างข้อมูลหาความถี่สัมพัทธ์ 8

ตารางที่ 2.2 ตัวอย่างข้อมูลหาความค่าความคลึงของสมาชิก 10

ตารางที่ 2.3: ตัวอย่างข้อมูลทำนายคะแนนสินค้า 10

ตารางที่ 3.1: ตัวอย่างข้อมูลชุดทดสอบ 22

ตารางที่ 3.2: ตัวอย่างตารางแสดงค่าคะแนนความชอบสินค้าที่ผู้ใช้ให้กับสินค้าในระบบ 22

ตารางที่ 3.3 ข้อมูลการให้คะแนนสินค้าของผู้ใช้ สำหรับ ตัวอย่างที่ 3.1 22

ตารางที่ 3.4 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.2 19

ตารางที่ 3.5 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.3 20

ตารางที่ 3.6 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.4 21

ตารางที่ 3.7 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.5 23

ตารางที่ 3.8 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.6 24

ตารางที่ 4.1: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบย่อยของ Movie Lens 27

ตารางที่ 4.2: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบของ Book Crossing 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

หน้า

รูปที่ 2.1 ตัวอย่างของส่วนระบบช่วยแนะนำจาก www.amazon.com 3

รูปที่ 2.2 ตัวอย่างของส่วนระบบช่วยแนะนำจาก www.facebook.com 4

รูปที่ 2.3 ตัวอย่างของหน้าเว็บไซต์ www.youtube.com 5

รูปที่ 2.4 ระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วม โดยอ้างอิงจากผู้ใช้ 6

รูปที่ 2.5 ระบบแนะนำด้วยวิธีการกรองเนื้อหา โดยอ้างอิงสินค้า 7

รูปที่ 2.6 ตัวอย่างข้อมูลคะแนนสินค้าของผู้ใช้แต่ละคน 9

รูปที่ 4.1: กราฟเปรียบเทียบประสิทธิภาพการทำนายแต่ละชุดข้อมูลของ Movie Lens 26

รูปที่ 4.2 กราฟเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยของ Movie Lens 29

รูปที่ 4.3: กราฟเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยของ Book Crossing 25

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3 ขอบเขตงานวิจัย

งานวิจัยนี้ปรับปรุงวิธีการทำนายค่าความชอบสินค้าของผู้ใช้ โดยการใช้ค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมในการคำนวณ การวัดผลความผิดพลาดในการทำนายคะแนนสินค้าโดยวัดจากค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error : MAE) ซึ่งได้ทำการทดลองกับชุดข้อมูล 2 ชุด คือ Movie Lens[5] จากเว็บไซต์ http://www.grouplens.org/node/73 เป็นชุดข้อมูลขนาดเล็กประกอบด้วยข้อมูลความชอบจำนวน 100,000 ชุด จำนวนผู้ใช้ 943 คน ที่มีต่อภาพยนตร์จำนวน 1,682 และ Book Crossing เป็นชุดข้อมูลขนาดใหญ่จากhttp://www.grouplens.org/node/74 [5] โดยมีจำนวนผู้ใช้ 278,858 คน ให้คะแนนสินค้า 1,149,780 ชุด กับหนังสือจำนวน 271,379 เล่ม โดยให้คะแนนสินค้าจำนวนเต็มตั้งแต่ 1 จนถึง 10 โดยเรียงลำดับคะแนนจากค่าน้อยไปค่ามากโดยข้อมูลที่นำใช้มีทั้งหมด 361,370 ชุด ในการทดสอบเพราะได้เลือกเฉพาะข้อมูลที่มีผู้ใช้ให้คะแนนหนังสือที่ไม่มีเป็นค่า 0 และที่มีผู้ใช้ในข้อมูลชุดทดสอบ

1.4 ส่วนประกอบวิทยานิพนธ์

ส่วนประกอบที่เหลือของวิทยานิพนธ์ฉบับนี้ประกอบด้วย

บทที่ 2 กล่าวถึงทฤษฎีที่ใช้และงานวิจัยที่เกี่ยวข้องกับระบบช่วยแนะนำด้วยวิธีการกรองข้อมูลผู้ใช้ร่วมพร้อมทั้งยกตัวอย่างวิธีการคำนวณค่าทำนายคะแนนสินค้า และกล่าวถึงงานวิจัยที่เกี่ยวข้อง

บทที่ 3 อธิบายเกี่ยวกับขั้นตอนการทำนายคะแนนความชอบในการกรองข้อมูลผู้ใช้ร่วม โดยใช้วิธีใช้ค่าความถี่สัมพัทธ์สูงสุดในการทำนายคะแนนความชอบของผู้ใช้ที่มีต่อสินค้า ซึ่งจะมีวิธีที่นำเสนอทั้งหมด 3 วิธี การใช้ค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมสินค้า (Relative Frequency: RF) วิธีเปรียบเทียบเปรียบเทียบค่าเฉลี่ยข้อมูลสินค้า (Mean Item) วิธีการเปรียบเทียบค่าความถี่สัมพัทธ์โดยใช้ค่าระหว่างคะแนน (Relative Frequency Between Point :RFBP)

บทที่ 4 นำเสนอรายละเอียดการศึกษาการทดลองและวิธีการวัดประสิทธิภาพของคำตอบที่ได้จากวิธีการที่พัฒนาและเปรียบเทียบกับวิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิมและวิธีที่เสนอในงานวิจัยนี้ทั้ง 3วิธี

บทที่ 5 สรุปผลการทดลองและข้อเสนอแนะเกี่ยวกับเทคนิคทางเลือกที่ใช้ในระบบช่วยแนะนำเพื่อทำนายคะแนนสินค้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่ใช้และงานวิจัยที่เกี่ยวข้องกับระบบช่วยแนะนำด้วยวิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิม(CF) พร้อมทั้งยกตัวอย่างวิธีการคำนวณหาค่าความคล้ายระหว่างผู้ใช้สินค้าและการทำนายคะแนนสินค้า พร้อมทั้งกล่าวถึงงานวิจัยที่เกี่ยวข้อง

2.1 ระบบช่วยแนะนำ

ในยุคแห่งการสื่อสารด้วยระบบเครือข่ายหรือที่เรียกว่าอินเตอร์เน็ต(Internet) ที่ได้เข้ามีบทบาทสำคัญในชีวิตประจำวันไม่ว่าจะเป็นธุรกิจการค้า การติดต่อสื่อสารที่สามารถเข้าถึงได้อย่างง่ายดายทุกที่ทุกเวลา ทำให้การทำธุรกิจผ่านระบบเครือข่ายนี้ เป็นที่นิยมมากขึ้นด้วย เนื่องจากสามารถเข้าถึงได้ทุกกลุ่มเป้าหมายและประหยัดต้นทุน แต่ข้อจำกัดของการทำธุรกิจผ่านระบบเครือข่าย หากเป็นสิ่งของ ลูกค้าจะไม่สามารถจับต้องได้ หรือ สินค้าอื่นๆก็เช่นกัน เพื่อสร้างความพึงพอใจให้กับลูกค้า จึงได้มีการนำระบบแนะนำข้อมูล(Recommender System) เข้ามาใช้ในการแนะนำให้ตรงกับความต้องการของลูกค้าแต่ละคน ตัวอย่างของเว็บไซต์พาณิชย์อิเล็กทรอนิกส์ เช่น เว็บไซท์อเมซอนดอทคอม(www.amazon.com) เป็นเว็บไซต์ที่ขายหนังสือและสินค้าต่างๆบนอินเตอร์เน็ตดังรูปที่ 2.1 ซึ่งเว็บไซต์ดังกล่าวใช้ระบบช่วยแนะนำสินค้าให้กับลูกค้า เมื่อลูกค้ากดเลือกดูสินค้าที่สนใจ รายการสินค้าที่มีความใกล้เคียงหรือคล้ายคลึงกับสินค้ารายการ ที่ลูกค้าเลือกดูจะปรากฏเป็นสินค้าแนะนำให้ลูกค้าได้เลือกชมในส่วนด้านล่าง

รูปที่ 2.1 ตัวอย่างของส่วนระบบช่วยแนะนำจาก www.amazon.com

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

นอกจากนี้ยังมีเว็บไซต์ที่เกี่ยวกับสังคมออนไลน์หรือที่เรียกว่า Social Network ใช้ระบบช่วยแนะนำเพื่อนให้กับเพื่อนคนต่อไปโดยอาศัยหลักการที่ว่า เป็นเพื่อนหรือเพื่อนของเพื่อนอาจจะเป็นที่เรารู้จักหรือมีความชอบในสิ่งเดียวกันหรือมีแนะนำกลุ่มหน้าแฟนเพจ ทำให้โลกของคนที่รู้จักนั้นกว้างขวางขึ้นและสามารถแลกเปลี่ยนความคิดเห็น ข่าวสาร ความรู้ ได้หลากหลายขึ้น ตัวอย่างเช่น เว็บไซต์เฟซบุคดอทคอม (www.facebook.com) ที่จะมีหน้าแนะนำเพื่อน หรือแนะนำสิ่งที่คาดว่าจะสนใจให้แต่ละบุคคล เช่น รูปที่ 2.2 แนะนำสื่อโฆษณาที่คล้ายคลึงกับที่สนใจที่เคยได้กดแสดงความชอบไปก่อนหน้านี้ หรือแม้กระทั่งเว็บไซต์ที่ได้รับความนิยมไปทั่วโลกเปรียบเสมือนช่องทางหนึ่งของการแสดงออกในกิจกรรมต่างๆ เช่น เว็บไซต์ยูทูปดอทคอม (www.youtube.com) นั้นยังนำระบบแนะนำไฟล์วิดีโอเข้ามาใช้เพื่อแนะนำให้ตรงกับความต้องการของผู้ดูเป็นต้น ดังรูปที่ 2.3

รูปที่ 2.2 ตัวอย่างของส่วนระบบช่วยแนะนำจาก www.facebook.com

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 2.3 ตัวอย่างของหน้าเว็บไซต์ www.youtube.com

ซึ่งระบบช่วยแนะนำมีเทคนิคต่างกันออกไปตามความเหมาะสมของประเภทธุรกิจโดยแบ่งออกเป็น 3 กลุ่มหลักดังนี้

2.1.1 ระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วม (Collaborative Filtering: CF)

ระบบแนะนำนี้จะคำนึงถึงพฤติกรรม ความคิดเห็นหรือรสนิยมของผู้ใช้ระบบในการคำนวณเป็นหลัก โดยจะใช้หลักการพิจารณาคะแนนความชอบของผู้ใช้ต่อสินค้าในระบบโดยสังเกตว่าสินค้าใดที่ผู้ใช้ร่วมให้คะแนนไป นำไปเปรียบเทียบกับผู้ใช้ที่ใกล้เคียงแล้วนำไปพิจารณากับสินค้าที่ผู้ใช้ยังไม่ให้คะแนนความชอบ แต่มีความคล้ายคลึงกับสินค้าที่ผู้ใช้นั้นเคยใช้ โดยเทคนิคใช้ตัววัดความคล้ายคลึง (Similarity Measurement) ในการวัดความคล้ายคลึงของข้อมูลความนิยมของผู้ใช้แต่ละคนเพื่อใช้ถ่วงน้ำหนักของสินค้าตัวที่ต้องการทำนายความนิยม ดังที่ Torres ในปี 2004 [4,7] กล่าวว่า *"เทคนิคการช่วยแนะนำโดยการกรองผู้ใช้ร่วมทำงานโดยการแนะนำสินค้าให้กับผู้ใช้โดยอ้างอิงจากผู้ใช้คนอื่น ๆ ที่มีความชอบสินค้าคล้ายกันมาก่อนหน้านี้ ระบบการกรองผู้ใช้ร่วมจะสร้างผู้ใช้ย่านใกล้เคียงสำหรับผู้ใช้แต่ละคนในระบบ และแนะนำสินค้ากับผู้ใช้คนใดคนหนึ่งถ้าผู้ใช้ย่านใกล้เคียงนั้นให้ความนิยมสูง"*

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

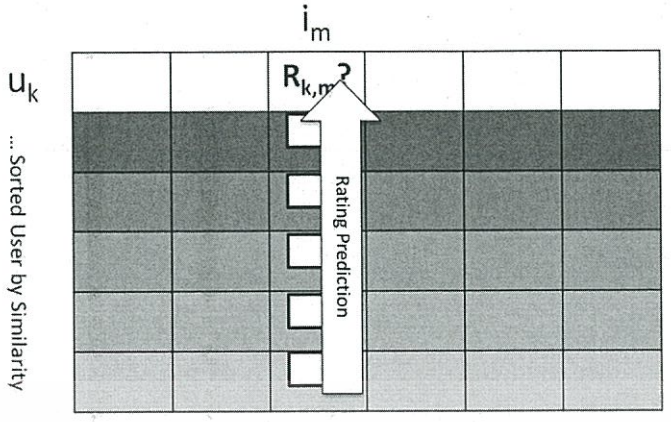
ระบบช่วยแนะนำ (Recommender System) [1] เป็นระบบแนะนำให้แต่ละบุคคล โดยเรียนรู้จากประวัติพฤติกรรม หรือความคล้ายคลึงของแต่ละบุคคลมาช่วยในการตัดสินใจเลือกคำตอบหรือแนะนำเพื่อให้ตรงกับความต้องการของผู้ใช้งานมากที่สุด ซึ่งระบบช่วยแนะนำถูกนำมาประยุกต์ใช้ในสายงานธุรกิจการค้าและงานบริการ ในการแนะนำสินค้าให้ตอบโจทย์และตรงความต้องการลูกค้ามากที่สุด โดยเฉพาะธุรกิจพาณิชย์อิเล็กทรอนิกส์ (E-Commerce) ที่มีฐานข้อมูลลูกค้าขนาดใหญ่ ซึ่งข้อมูลเหล่านี้สามารถนำมาช่วยในการตัดสินใจของลูกค้าได้รวดเร็วขึ้น นอกจากนั้นยังสามารถสร้างความประทับใจให้กับลูกค้าอีกด้วย

วิธีการกรองข้อมูลผู้ใช้ร่วม (Collaborative Filtering : CF) เป็นเทคนิคหนึ่งในการสำรวจพฤติกรรมการใช้สินค้าของผู้ใช้โดยสังเกตว่าสิ่งใดที่ผู้ใช้ทำร่วมกันหรือให้คะแนนสินค้าที่เท่ากันแล้วทำการเปรียบเทียบกับผู้ใช้ที่ใกล้เคียงกันเพื่อทำนายสินค้าและแนะนำให้กับผู้ใช้ ซึ่งวิธีการดังกล่าวผู้วิจัยพบว่ายังประสบปัญหาด้านความถูกต้องและการประมวลผลที่ยังล่าช้าเนื่องจากการทดลองใช้ข้อมูลขนาดใหญ่ในการเปรียบเทียบ งานวิจัยนี้จึงขอเสนอการปรับปรุงวิธีการทำนายคะแนนความชอบสินค้าของผู้ใช้โดยใช้ค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมซึ่งในที่นี้จะเรียกว่า (Relative Frequency: RF) ในการทำนายคะแนนความชอบสำหรับแนะนำสินค้า วิธีที่นำเสนอ (RF) ในงานวิจัยนี้สามารถลดความขั้นตอนการหาความคล้ายคลึงของผู้ใช้ซึ่งส่งผลให้เวลาในการคำนวณเร็วขึ้น ค่า time complexity เท่ากับ O(m) หรือ O(n) ซึ่ง m คือจำนวนของผู้ใช้สินค้าทั้งหมดที่เคยให้คะแนนและ n คือ จำนวนสินค้าทั้งหมดที่ผู้ใช้เคยให้คะแนน ในขณะที่วิธีกรองข้อมูลผู้ใช้ร่วมแบบเดิม(CF) มีค่า time complexity เท่ากับ O(mn) ซึ่งวิธีที่นำเสนอ RF ให้ค่าความคลาดเคลื่อนในการทำนายน้อยกว่าวิธีกรองข้อมูลผู้ใช้ร่วมแบบเดิม

1.2 วัตถุประสงค์

งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพของระบบแนะนำสินค้าที่ใช้วิธีการกรองข้อมูลผู้ใช้ร่วม (CF) เพื่อลดเวลาในการคำนวณและเพิ่มประสิทธิภาพให้ผลการทำนายการแนะนำสินค้าให้มีความถูกต้องมากขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.4 ระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วม โดยอ้างอิงจากผู้ใช้

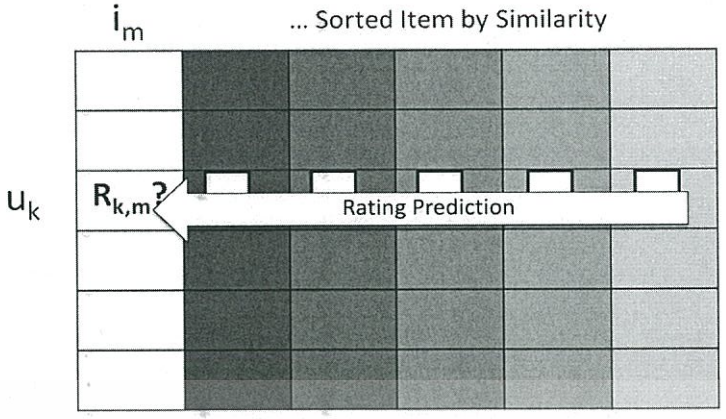
ข้อดีของระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วมคือ สินค้าใดๆที่ยังไม่เคยถูกเลือกซื้อหรือเปิดดูโดยผู้ใช้มาก่อนจะสามารถถูกแนะนำได้โดยการเปรียบเทียบกับผู้ใช้คนอื่นๆที่เคยใช้สินค้านั้นๆ มาแล้วจากผู้ใช้ที่มีรสนิยมการบริโภคสินค้าที่มีความคล้ายคลึงกันโดยวัดจากข้อมูลคะแนนความชอบที่ผู้ใช้นั้นได้เคยให้ไว้กับสินค้าอื่นๆในระบบ

ข้อด้อยของระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วมคือ ปัญหาผู้ให้ความนิยมแรก(First Rater Problem) เมื่อผู้ใช้ใหม่เข้ามาในระบบ ระบบจะไม่สามารถแนะนำสินค้าใดๆ ให้ได้เลยหรือในกรณีเดียวกับสินค้าที่เข้ามาในระบบครั้งแรกนั้นจะไม่ได้ถูกแนะนำให้กับผู้ใช้ เนื่องจากไม่สามารถเปรียบเทียบความคล้ายคลึงระหว่างผู้ใช้หรือสินค้าที่เข้ามาใหม่กับผู้ใช้หรือสินค้าที่มีอยู่แล้วในระบบได้และปัญหาความเบาบางของข้อมูล (Sparsity Problem) ซึ่งจะทำให้การเปรียบเทียบนั้นทำได้ยาก ดังที่ Torres ในปี 2004 กล่าวว่า *"ปัญหาผู้ให้ความนิยมแรก (First-rater problem): สินค้าจำเป็นจะต้องถูกให้ความนิยมโดยผู้ใช้อย่างน้อย 1 คน มิเช่นนั้นแล้วสินค้าจะไม่สามารถถูกแนะนำได้จนกระทั่งจะมีใครมาให้ความนิยมกับสินค้านั้นก่อน"*

2.1.2 ระบบช่วยแนะนำด้วยวิธีการกรองเนื้อหา (Content-based Filtering: CB)

ระบบช่วยแนะนำที่คำนึงถึงคุณสมบัติของสินค้าหรือข้อความบรรยายลักษณะตัวสินค้าโดยจะพิจารณากลุ่มสินค้าที่ผู้ใช้คนหนึ่งได้เคยให้คะแนนความชอบแล้วนำมาเปรียบเทียบกับผู้ใช้อื่นในระบบที่เคยให้ความนิยมกับตัวสินค้าที่ต้องการจะทำนายจากนั้นทำการคำนวณค่าความคล้ายคลึงของชุดสินค้าที่ได้เป็นตัวถ่วงน้ำหนักและใช้ผลรวมเชิงเส้นของค่าถ่วงน้ำหนักทั้งหมดนั้นเป็นค่าความนิยมที่จะทำการทำนาย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.5 ระบบแนะนำด้วยวิธีการกรองเนื้อหาโดยอ้างอิงสินค้า

ข้อดีของระบบแนะนำด้วยวิธีการกรองเนื้อหาคือ ระบบสามารถแนะนำสินค้าอื่นได้โดยการพิจารณาจากสินค้าที่มีเนื้อหาที่ใกล้เคียงกันแม้ว่าสินค้านั้นจะยังไม่เคยถูกเปิด โดยพิจารณาที่รายละเอียดของสินค้าเป็นหลัก

ข้อด้อยของระบบแนะนำด้วยวิธีการกรองเนื้อหาคือ เทคนิคนี้จะใช้แนะนำในลักษณะที่เฉพาะเจาะจงมีการอธิบายรายละเอียดที่ชัดเจน แต่จะไม่สามารถแนะนำรายละเอียดที่คลุมเครือได้ เช่นเพลง เป็นไปได้ยากที่จะแนะนำเพลงตามเนื้อหา หรือ ทำนอง แล้วตัดสินได้ว่าเพลงใดสนุก สุข หรือเศร้า แล้วจะตรงกับความชอบของผู้ใช้แต่ละบุคคล

2.1.3 ระบบช่วยแนะแบบผสม (Hybrid Recommender System)

เป็นระบบช่วยแนะนำแบบผสมที่รวมเอาข้อดีของทั้งสองวิธีนี้เข้าไว้ด้วยกันเพื่อพัฒนาเป็นเป็นระบบใหม่ที่ดีขึ้น เช่น ระบบแนะนำแบบผสมด้วยการถ่วงน้ำหนัก (Weight Hybrid Recommender) จะรวมคะแนนความชอบที่ทำนายได้จากแต่ละเทคนิคที่ใช้ด้วยผลรวมเชิงเส้น ซึ่งคะแนนที่นำมารวมกันนี้ต้องสามารถรวมได้ในระบบเชิงเส้น และควรจะมีความสอดคล้องกัน เช่น มีขอบเขตของค่าทำนายเหมือนกัน หรือ ระบบช่วยแนะนำแบบผสมด้วยเทคนิคการเพิ่มเติมคุณลักษณะ (Feature Augmentation Hybrid Recommender) คือ เทคนิคในระบบช่วยแนะนำแบบผสมอีกอย่างหนึ่งที่คล้ายกับระบบช่วยแนะนำแบบผสมด้วยเทคนิคการรวมคุณลักษณะ เพียงแต่ว่าคุณลักษณะที่ใช้เป็นข้อมูลเข้านั้น จะเป็นคุณลักษณะแบบใหม่ เป็นต้น

2.2 ทฤษฎีที่เกี่ยวข้อง

2.2.1 การแจกแจงความถี่สัมพัทธ์ (Relative Frequency)

ความถี่สัมพัทธ์ (Relative Frequency) คือ อัตราส่วนของความถี่ในแต่ละลำดับข้อมูลเทียบกับจำนวนความถี่ทั้งหมด ผลรวมของความถี่สัมพัทธ์ทั้งหมดมีค่าเท่ากับ 1เสมอ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.1 ตัวอย่างข้อมูลหาความถี่สัมพัทธ์

ลำดับชั้นที่	ความถี่ f_i	ความถี่สัมพัทธ์ (f_i / Sf_i)
1	3	3/13 = 0.23
2	4	4/13 = 0.307
3	2	2/13 = 0.153
4	1	1/13 = 0.077
รวม	**Sf_i = 13**	**1**

2.2.2 ค่าฐานนิยม (mode)

ค่าฐานนิยม คือ ตัวเลขหรือข้อมูลที่มีค่าความถี่มากที่สุด ตัวอย่างเช่น ค่าฐานนิยมตารางที่ 1 คือลำดับชั้นที่ 2 เพราะมีค่าความถี่เท่ากับ 4 เป็นค่าที่เกิดขึ้นมากที่สุดในกลุ่มข้อมูล ถ้าแต่ละกลุ่มข้อมูลไม่มีค่าซ้ำกันเลยหรือค่าความถี่เท่ากับหนึ่ง จะไม่สามารถบอกค่าฐานนิยมได้

2.2.3 ขั้นตอนระบบการแนะนำสินค้าโดยใช้เทคนิคการกรองข้อมูลผู้ใช้ร่วม

ในระบบแนะนำสินค้าจะใช้ข้อมูลของสินค้าและข้อมูลคะแนนของผู้ที่เคยใช้สินค้า มาทำการคำนวณสร้างรายการแนะนำให้กับผู้ใช้อื่นที่ยังไม่เคยใช้สินค้า ด้วยเทคนิควิธีการกรองข้อมูลผู้ใช้ร่วม เป็นวิธีการที่นิยมนำมาใช้ในระบบแนะนำสินค้า เนื่องจากวิธีการนี้จะใช้วิธีการค้นหาข้อมูลของผู้ใช้ที่เป็นสมาชิกข้างเคียง โดยมีความชอบสินค้าที่คล้ายกันกับผู้ใช้ ซึ่งข้อมูลของผู้ใช้ที่เป็นสมาชิกใกล้เคียงจะถูกนำมาวิเคราะห์และแนะนำสินค้าให้ตรงกับความต้องการของผู้ใช้ แบ่งเป็น 4 ขั้นตอน ดังนี้

2.2.3.1 ขั้นตอนการเตรียมข้อมูล

ข้อมูลตัวอย่างที่นำมาทดลองนี้จะมีข้อมูลสินค้าและผู้ใช้สินค้า ซึ่งผู้ใช้แต่ละคนจะมีค่าหรือไม่มีค่าของคะแนนความชอบที่มีต่อสินค้า ดังตัวอย่างใน รูปที่ 2.6 ซึ่งเป็นตัวอย่างคะแนนที่ผู้ใช้แต่ละคนมีต่อสินค้าในแต่ละชิ้น ในที่นี้สมมุติให้มีผู้ใช้ 4 คน และมีสินค้าอยู่ 5 ชิ้น โดยค่าคะแนนความชอบมีช่วงตั้งแต่ 1 ถึง 5 ซึ่งค่าคะแนนที่ 1 แสดงว่าไม่ชอบ ส่วนค่าคะแนนที่ 5 คือชอบมากที่สุด โดยที่เครื่องหมาย “?” หมายความว่ายังไม่มีการให้คะแนนกับสินค้าชิ้นนั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผู้ใช้ \ สินค้า	1	2	3	4	5
1	4	5	1	3	3
2	4	3	2	3	3
3	2	?	2	3	2
4	1	3	1	4	1

No Co-rate Item | Co-rated Item

รูปที่ 2.6 ตัวอย่างข้อมูลคะแนนสินค้าของผู้ใช้แต่ละคน

2.2.3.2 ขั้นตอนการหาค่าความคล้ายคลึงของสมาชิก

ขั้นตอนนี้เป็นกระบวนการค้นหาสมาชิกที่อยู่ในระบบที่มีความชอบคล้ายคลึงกันกับผู้ใช้ที่ต้องการค่าแนะนำอยู่ในขณะนี้ จากรูปที่2.6 ถ้าต้องการหาว่า ผู้ใช้ที่ 4 มีความชอบคล้ายคลึงกับ ผู้ใช้ที่ 3 หรือไม่ จะต้องพิจารณาจากค่าคะแนนความชอบของ ผู้ใช้ที่ 4 กับ ผู้ใช้ที่ 3 ที่มีให้กับสินค้าชิ้นเดียวกัน ซึ่งจะเรียกว่า co-rated items ในการหาความคล้ายคลึงของผู้ใช้ทั้งสองผู้ใช้ จะนิยมใช้สหสัมพันธ์ของเพียร์สัน (Pearson correlation coefficient) [1] ซึ่งค่าที่คำนวณได้มีค่ามากแสดงว่าผู้ใช้สองคนนี้มีความคล้ายคลึงกัน แต่ถ้าค่าที่คำนวณได้มีค่าน้อยแสดงว่าผู้ใช้สองคนนี้มีความคล้ายคลึงกันน้อย คำนวณตามสมการที่ 2.1

$$sim_{(u_i,u_j)} = \frac{\sum_{k=1}^{n}\left(r_{u_{i,k}} - \overline{r_{u_i}}\right)\times\left(r_{u_{j,k}} - \overline{r_{u_j}}\right)}{\sigma_{u_i} \times \sigma_{u_j}} \qquad (2.1)$$

โดยที่

$sim_{(u_i,u_j)}$ คือ ค่าความคล้ายระหว่างผู้ใช้ u_i และ u_j

$r_{u_{i,k}}$ คือ ค่าคะแนนความชอบที่ผู้ใช้ u_i มีต่อสินค้าชิ้นที่ k

$r_{u_{j,k}}$ คือ ค่าคะแนนความชอบที่ผู้ใช u_j มีต่อสินค้าชิ้นที่ k

$\overline{r_{u_i}}$ คือ ค่าเฉลี่ยของค่าคะแนนความชอบผู้ใช้ u_i มีต่อสินค้าทุกชิ้น

$\overline{r_{u_j}}$ คือ ค่าเฉลี่ยของค่าคะแนนความชอบที่ผู้ใช้ u_j มีต่อสินค้าทุกชิ้น

σ_{u_i} คือ ค่าความเบี่ยงเบนมาตรฐานของค่าคะแนนที่ผู้ใช้ u_i มีต่อสินค้าทุกชิ้น

σ_{u_j} คือ ค่าความเบี่ยงเบนมาตรฐานของค่าคะแนนที่ผู้ใช้ u_j มีต่อสินค้าทุกชิ้น

n คือ จำนวนสินค้าทั้งหมดที่ผู้ใช้เคยให้คะแนนชิ้นเหมือนกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากสมการการหาค่าความคล้ายคลึงของสมาชิกใช้ time complexity มีค่าเท่ากับ O(n) เนื่องจากต้องคำนวณหาค่าความคล้ายของทุกสินค้าที่ผู้ใช้เคยให้คะแนน

ตัวอย่าง 2.1 การคำนวณหาค่าความคล้ายคลึงของสมาชิก

หลังจากที่ค้นหาสมาชิกที่อยู่ในระบบที่มีความชอบคล้ายคลึงกับผู้ใช้เป้าหมายได้แล้ว วิธีการต่อไปคือการทำนายค่าความชอบของผู้ใช้เป้าหมายต่อสินค้า โดยพิจารณาจากการให้คะแนน

ตารางที่ 2.2 ตัวอย่างข้อมูลหาความค่าความคลึงของสมาชิก

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4	I_5	I_6	I_7
u_i	2	3	1	2	5	4	2
u_j	5	3	2	4	3	2	?

ในการคำนวณจะนำค่าคะแนน ของสินค้า ที่ $I_1 - I_6$ เพราะเป็นจำนวนสินค้าที่เคยให้ไว้ชิ้นเหมือนกัน ส่วนใน I_7 ไม่ได้นำไม่ใช้ เพราะ u_2 ไม่เคยให้คะแนน

การคำนวณหาค่าต่างๆ ดังนี้

$\sum r_{u_i}$ = 2+3+1+2+5+4 = 17

$\sum r_{u_j}$ = 5+3+2+4+3+2 = 19

$\overline{r_{u_i}}$ = 17/6 = 2.83

$\overline{r_{u_i}}$ = 19/6 = 3.17

$\sigma_{u_i} = \sqrt{((2-2.83)^2 + (3-2.83)^2 + (1-2.83)^2 + (2-2.83)^2 + (5-2.83)^2 + (4-2.83)^2)/6}$

$= \sqrt{1.80} = 1.34$

$\sigma_{u_j} = \sqrt{((5-3.17)^2 + (3-3.17)^2 + (2-3.17)^2 + (4-3.17)^2 + (3-3.17)^2 + (2-3.17)^2)/6}$

$= \sqrt{1.138} = 1.067$

$\sum_{k=1}^{6} \left(r_{u_i} - \overline{r_{u_i}}\right)\left(r_{u_j} - \overline{r_{u_j}}\right)$ = ((2 - 2.83) * (5 - 3.17)) + ((3 - 2.83) * (3 - 3.17)) +

((1 - 2.83) * (2 - 3.17)) + ((2 - 2.83) * (4 - 3.17)) +

((5 - 2.83) * (3 - 3.17)) + ((4 - 2.83) * (2 - 3.17))

= - 1.8334

$$sim_{(u_i,u_j)} = \frac{-1.8334}{1.37 \times 1.067}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$sim_{(u_i,u_j)} = -1.254$$

2.2.3.3 ขั้นตอนการทำนายคะแนนสินค้า

หลังจากที่ค้นหาสมาชิกที่อยู่ในระบบที่มีความชอบคล้ายคลึงกับผู้ใช้เป้าหมายได้แล้ว วิธีการต่อไปคือการทำนายค่าความชอบของผู้ใช้เป้าหมายต่อสินค้า โดยพิจารณาจากการให้คะแนนความคล้ายคลึงระหว่างสินค้าชิ้นนั้นกับสินค้าอื่นๆ ที่เคยให้คะแนนแล้วนำมาคำนวณตามสมการที่ 2.2

$$P(u_i,k) = \overline{r_{u_i}} + \frac{\sum_{j=1}^{m} sim_{(u_i,u_j)} \left(r_{u_j,k} - \overline{r_{u_j}}\right)}{\sum_{j=1}^{m} sim_{(u_i,u_j)}} \quad (2.2)$$

โดยที่

$P(u_i,k)$ คือ ค่าคะแนนทำนายของผู้ใช้ u_i ต่อสินค้าชิ้นที่ k ของผู้ใช้ u_i

$\overline{r_{u_i}}$ คือ ค่าเฉลี่ยของคะแนนสินค้าของผู้ใช้ u_i

m คือ จำนวนของผู้ใช้ทั้งหมดที่ให้คะแนนสินค้าชิ้นที่ k

$r_{u_{j,k}}$ คือ ค่าคะแนนสินค้าของผู้ใช้ u_j ที่ให้กับสินค้าชิ้นที่ k

$sim_{(u_i,u_j)}$ คือ ค่าความคล้ายคลึงระหว่างผู้ใช้ u_i และผู้ใช้ u_j

ตารางที่ 2.3 ตัวอย่างข้อมูลทำนายคะแนนสินค้า

ผู้ใช้ \ สินค้า	I_1	I_2	I_3	I_4	I_5	I_6	I_7
u_i	2	3	1	2	5	4	2
u_j	5	3	2	4	3	2	?

จากตารางตัวอย่างจะทำนายคะแนนสินค้าของผู้ใช้ u_j ในสินค้าที่ I_7 (m = 1)

$$P(u_j, I_7) = \overline{r_{u_j}} + \frac{\sum_{i=1}^{m} sim_{(u_j,u_i)} \left(r_{u_{j,I_7}} - \overline{r_{u_i}}\right)}{\sum_{i=1}^{m} sim_{(u_i,u_j)}}$$

$\overline{r_{u_i}}$ = 17/6 = 2.83

$\overline{r_{u_j}}$ = 19/6 = 3.17

$$P(u_i, i_7) = 3.17 + \frac{-1.254*(2-2.83)}{-1.254}$$

$P(u_j, i_7) = 4$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการคำนวณทำนายคะแนนสินค้า ผู้ใช้ u_j สินค้าที่ $I_i = 4$

จากสมการทำนายคะแนนสินค้าใช้ m คือ จำนวนผู้ใช้ทั้งหมดที่หาค่าความคล้าย และ n คือ จำนวนสินค้าที่ผู้ใช้เคยให้คะแนนสินค้า ค่า time complexity มีค่า O(mn) เนื่องจากมีการคำนวณหาค่าความคล้ายคลึงของสมาชิกและมีการคำนวณความคล้ายของทุกผู้ใช้ที่ให้คะแนนสินค้าที่ต้องการทำนาย

2.2.3.4 ขั้นตอนการสร้างรายการแนะนำ

เมื่อทำการทำนายคะแนนสินค้าในชุดข้อมูลเรียบร้อย ระบบแนะนำจะทำการเรียงลำดับสินค้าที่จะแนะนำให้กับผู้ใช้จากลำดับตั้งแต่ คะแนนมากไปหาน้อย

2.3 งานวิจัยที่เกี่ยวข้อง

ระบบช่วยแนะนำมีแนวคิดและงานวิจัยที่หลากหลายซึ่งงานวิจัยที่เกี่ยวข้องเกี่ยวกับระบบช่วยแนะนำด้วยวิธีการกรองผู้ใช้ร่วมที่ได้ศึกษามีดังต่อไปนี้

1. Melville, P., Mooney, R. J. และ Nagarajan, R. [6] นำเสนอ Content-Boosted Collaborative Filtering for Improved Recommendations ระบบช่วยแนะนำแบบผสมที่ใช้ตัวจำแนกประเภทเบส์แบบสามัญในการคัดกรองเนื้อหาเพื่อทำนายความนิยมของผู้ใช้เพิ่มเติมบางส่วน ทำให้มีข้อมูลความนิยมในระบบมากขึ้นโดยการใช้ระบบช่วยแนะนำแบบผสมเพื่อแก้ปัญหาข้อมูลเบาบาง ด้วยวิธีการกรองเนื้อหาสำหรับทำนายความนิยมของสินค้าในระบบบางส่วนก่อน หลังจากนั้นการกรองผู้ใช้ร่วมโดยนำผลลัพธ์ที่ได้ไปพิจารณาย่านใกล้เคียงที่สุด k ตัว โดยการถ่วงน้ำหนักและวัดความคล้ายของเวกเตอร์ความนิยมในการกรองผู้ใช้ร่วมซึ่งวิธีนี้คือสามารถแก้ไขปัญหาความเบาบางของข้อมูลในระบบได้ดีแต่ข้อด้อยของวิธีนี้ก็คือหากวิธีการทำนายความนิยมที่ใช้ครั้งแรกมีความคลาดเคลื่อนสูงเมื่อนำมาเป็นข้อมูลเข้าในการทำนายอีกวิธีหนึ่งก็ทำให้ผลการทำนายนั้นคลาดเคลื่อนสูงตามไปด้วย

2. วงกต ศรีอุไร [8] นำเสนอการเปรียบเทียบเทคนิคการแทนค่าข้อมูลที่ขาดหายสำหรับการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมโดย 1)การเลือกรายการสินค้าที่ต้องการแทนค่า(Item Selection) และ 2) คำนวณค่าที่ใช้ในการแทนค่าข้อมูล (Imputed Value Calculation) จากนั้นเปรียบเทียบสองเทคนิคที่ใช้ในการเลือกรายการสินค้าระหว่างเลือกโดยการสุ่มและเลือกโดยการวิเคราะห์ค่าความแปรปรวนแล้วคำนวณค่าที่ต้องการจะแทนที่ด้วยสองวิธีการ ได้แก่ 1) คำนวณด้วยการหาค่าเฉลี่ย และ2) คำนวณด้วยวิธีสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor: KNN) พบว่า วิธีการเลือกรายการโดยวิเคราะห์ค่าความแปรปรวนและการแทนค่าข้อมูลด้วยวิธีสมาชิกที่ใกล้ที่สุดให้ผลลัพธ์ที่ดีที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. นิรภัฏ อุทัยฉาย [9] นำเสนอระบบช่วยแนะนำแบบผสมโดยใช้การสืบค้นกฎการเกิดร่วมกัน และ การจัดกลุ่มโดยการนำ k-Means Algorithm มาใช้ในการจัดการกลุ่มข้อมูลเพื่อใช้แก้ปัญหาการเริ่มต้นการทำงานล่าช้า ร่วมกับการนำทฤษฎีกฎความสัมพันธ์มาประยุกต์ใช้ร่วมกัน พบว่าสามารถลดความล่าช้าลงได้

4. Torres, Roberto [7] "เทคนิคการช่วยแนะนำโดยการกรองร่วมทำงานโดยการแนะนำสินค้าให้กับผู้ใช้โดยอ้างอิงจากผู้ใช้คนอื่น ๆ ที่มีความชอบสินค้าคล้ายกันมาก่อนหน้านี้ ระบบการกรองร่วมจะสร้างผู้ใช้ย่านใกล้เคียงสำหรับผู้ใช้แต่ละคนในระบบและสินค้ากับผู้ใช้คนใดคนหนึ่งถ้าผู้ใช้ย่านใกล้เคียงนั้นให้ความนิยมสูง"

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
การทำนายคะแนนความชอบโดยใช้ค่าความถี่สัมพัทธ์สูงสุดของสินค้า

ในบทนี้จะนำเสนอขั้นตอนการทำนายคะแนนความชอบในการกรองข้อมูลผู้ใช้ร่วม โดยใช้วิธีใช้ค่าความถี่สัมพัทธ์สูงสุด ในการทำนายคะแนนความชอบของผู้ใช้ที่มีต่อสินค้า โดยการนำค่าเฉลี่ยของผู้ใช้ที่ทำนายมาปรับค่าคะแนนด้วยเงื่อนไขที่กำหนดซึ่งจะดูจากค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยม ค่าคะแนนความชอบที่น้อย หมายถึงผู้ใช้สินค้ามีความชอบในตัวสินค้านั้นน้อย ส่วนค่าคะแนนความชอบที่มาก หมายถึงผู้ใช้สินค้ามีความชอบต่อตัวสินค้านั้นมากตามลำดับ

3.1 ข้อมูลที่ใช้ในการเรียนรู้และทดสอบ

3.1.1 ชุดข้อมูลทดสอบ Movie Lens

ชุดข้อมูลทดสอบที่ทำมาใช้ในการทำนายคะแนนความชอบคือชุดข้อมูลMovie Lens [5] จาก http://www.grouplens.org/node/73 ซึ่งประกอบด้วยข้อมูลคะแนนความชอบสินค้าจำนวน 100,000 ชุด จากจำนวนผู้ใช้ 943 คน ที่มีต่อภาพยนตร์จำนวน 1,682 เรื่องโดยที่คะแนนความชอบนั้นใช้เป็นเลขจำนวนเต็มตั้งแต่ 1 จนถึง 5 เรียงจากคะแนนที่มีค่าน้อยไปคะแนนที่มีค่ามากประกอบไปด้วยข้อมูลส่วนตัวของผู้ใช้ คือ อายุ เพศ อาชีพ และข้อมูลรายละเอียดของสินค้า ประเภทของภาพยนต์ ชนิดของภาพยนตร์ทั้งหมดที่มีอยู่ในชุดข้อมูล Movie Lens มีทั้งหมด 19 ชนิด ได้แก่ Unknown, Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War และ Westernซึ่งชุดข้อมูลนี้จะถูกนำมาแบ่งออกเป็นข้อมูลที่ใช้ทดสอบ 5 ชุดด้วยกัน โดยทำการทดลองเป็นข้อมูลสอน 80% จะมีทั้งหมด 80,000 แถวข้อมูล และชุด-ทดสอบ 20 % จะมี 20,000 แถวข้อมูล ซึ่งจะทำการทดสอบที่ละชุดจนครบ 5 ชุด โดยแต่ละชุดจะไม่ใช้ข้อมูลที่ซ้ำกันเลยในการทดสอบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.1 ตัวอย่างข้อมูลชุดทดสอบ

user id	Item id	rating	timestamp
1	1	5	874965758
1	2	3	876893171
1	3	4	878542960
1	4	3	876893119
1	5	3	889751712
1	6	5	887431973
1	7	4	875071561
1	8	1	875072484
1	9	5	878543541
1	10	3	875693118
1	11	2	875072262
1	12	5	878542960
1	13	5	875071805
1	14	5	874965706
1	15	5	875071608
1	16	5	878543541
1	17	3	875073198
1	18	4	887432020
1	19	5	875071515
1	20	4	887431883

ในตัวอย่างข้อมูลคอลัมน์ที่ 1 คือรหัสของผู้ใช้สินค้า คอลัมน์ที่ 2 คือรหัสสินค้า(ภาพยนตร์) คอลัมน์ที่ 3 คือ คะแนนความชอบสินค้า คอลัมน์ที่ 4 คือวันเวลาที่ให้คะแนนสินค้า

3.1.2 ชุดข้อมูลทดสอบ Book Crossing

ชุดข้อมูล Book Crossing (BX) จาก http://www.grouplens.org/node/74โดยCai-Nicolas Ziegler [5] เป็นผู้รวบรวมระยะเวลา 4 สัปดาห์ระหว่างเดือนสิงหาคมและกันยายน ค.ศ. 2004 มีจำนวนผู้ใช้ 278,858 คน ให้คะแนนสินค้า 1,149,780 ชุด กับหนังสือจำนวน 271,379 เล่ม โดยให้คะแนนสินค้าจำนวนเต็มตั้งแต่ 1 จนถึง 10 โดยเรียงลำดับคะแนนจากค่าน้อยไปค่ามากโดย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลที่นำใช้มีทั้งหมด 361,370 ชุด ในการทดสอบเพราะได้เลือกเฉพาะข้อมูลที่มีผู้ใช้ให้คะแนนหนังสือที่ไม่มีเป็นค่า 0 และที่มีผู้ใช้ในข้อมูลชุดทดสอบ

3.2 ตัววัดประสิทธิภาพการทำนายคะแนนความชอบสินค้า

สำหรับการวัดประสิทธิภาพของขั้นตอนวิธี จะใช้ค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) และทำการทดลองโดยข้อมูลนี้จะถูกแบ่งข้อมูลออกเป็น 5 ชุด แต่ละชุดข้อมูลจะแบ่งออกเป็นชุดฝึกสอน (training set) 80% และชุดทดสอบ(test set) 20% สำหรับข้อมูล Movie Lens

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

เมื่อ P_i คือ ค่าคะแนนความชอบสินค้าที่ทำนาย

Q_i คือ ค่าคะแนนความชอบสินค้าที่ถูกต้อง

n คือ จำนวนคะแนนสินค้าที่ทำนายทั้งหมด

3.3 ปัญหาการทำนายคะแนนความชอบสินค้า

ปัญหาการทำนายคะแนนความชอบสินค้าในงานวิจัยนี้นำเสนอในรูปแบบของเมตริกซ์หรือตาราง 2 มิติโดยที่ในแถวนั้นแสดงรายการผู้ใช้ (User) และคอลัมน์แสดงรายการสินค้า (Item) :ซึ่งข้อมูลภายในตารางคือคะแนนความชอบของผู้ใช้ที่ผู้ใช้คนหนึ่งให้กับสินค้าตัวหนึ่งดังแสดงตารางที่ 3.1

ตารางที่ 3.2 ตัวอย่างตารางแสดงค่าคะแนนความชอบสินค้าที่ผู้ใช้ให้กับสินค้าในระบบ

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 3.2 นั้นคือข้อมูลที่มีผู้ใช้ 5 คนให้คะแนนความชอบแก่ตัวสินค้าต่าง ๆ 5 ชิ้น โดยมีที่มีค่าคะแนนความชอบสินค้าตั้งแต่ 1 จนถึง 5 ซึ่งในข้อมูลสินค้าส่วนใหญ่เจริงแล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จำนวนผู้ใช้และจำนวนสินค้านั้นมีมากกว่านี้หลายเท่าตัวนัก และค่าคะแนนความชอบสินค้าที่มีในระบบอาจจะไม่หนาแน่นมากนักทำให้การทำนายคะแนนความชอบสินค้าเป็นไปด้วยความไม่สะดวกดังที่กล่าวไว้แล้วในบทที่ 2

3.4 การทำนายคะแนนความชอบด้วยค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมสินค้า (Relative Frequency:RF)

งานวิจัยนี้เสนอการทำนายคะแนนสินค้าชิ้นที่ k ของผู้ใช้ u_i โดยการนำค่าฐานนิยม(mode) ของคะแนนสินค้าที่ผู้ใช้อื่นเคยให้คะแนนในสินค้าที่จะทำนาย ซึ่งในตัวอย่างนี้จะมีการให้คะแนนตั้งแต่ 1-5 ค่าคะแนนที่ 1 แสดงว่าชอบน้อยที่สุดส่วนค่าคะแนนที่ 5 คือชอบมากที่สุดวิธีนี้จะใช้การเปรียบเทียบค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมของคะแนนสินค้าในการปรับค่าคะแนนโดยจะมีเงื่อนไขและขั้นตอนดังต่อไปนี้

1. นำค่าเฉลี่ยของผู้ใช้สินค้ามาปัดทศนิยม (หากเกิน 0.5 ปัดขึ้น มิฉะนั้นปัดลง)ให้เป็นค่าจำนวนเต็มของคะแนนสินค้า $\overline{r_{u_i}}$

2. แจกแจงความถี่สัมพัทธ์ของคะแนนสินค้าที่ทำนาย ซึ่งจะได้ค่าฐานนิยมของคะแนนในสินค้าที่ต้องการทำนาย mode

ถ้าความถี่ของคะแนนสินค้ามีค่าความถี่ที่เท่ากันให้เลือกค่า mode ที่ใกล้เคียงกับค่าเฉลี่ยของผู้ใช้สินค้า

3. การคำนวณคะแนนทำนายสินค้ามีเงื่อนไขดังนี้

หาก $\overline{r_{u_i}}$ > mode คำนวณ คะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่คงค่าทศนิยม) ลบด้วยค่าความถี่สัมพัทธ์สูงสุด

หาก $\overline{r_{u_i}}$ < mode คำนวณ คะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่คงค่าทศนิยม) บวกด้วยค่าความถี่สัมพัทธ์สูงสุด

หาก $\overline{r_{u_i}}$ = mode = คะแนนที่ทำนาย

จากสมการของวิธีการใช้ค่าความถี่สัมพัทธ์สูงสุดในการปรับค่าคำทำนายคะแนนสินค้าจะมีค่า time complexity ดังนี้

กำหนดให้

m = จำนวนสมาชิกผู้ใช้สินค้าที่เคยให้คะแนน

n = จำนวนสินค้าทั้งหมดที่เคยให้คะแนน

ค่า time complexity $= \begin{cases} O(m); & m \geq n \\ O(n); & m < n \end{cases}$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างที่ 3.1 การคำนวณคะแนนความชอบสินค้าโดยการทำนายความชอบสินค้าชิ้นที่ 3 ของผู้ใช้ที่ 3

ตารางที่ 3.3 ข้อมูลการให้คะแนนสินค้าของผู้ใช้ สำหรับตัวอย่างที่ 3.1

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

ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 8/3 = 2.67 ปัดทศนิยม $\approx$ 3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน

คะแนน 2 จำนวน 1 คน และ คะแนน 4 จำนวน 3 คน

ค่าฐานนิยมของคะแนน คือ 4 มากกว่าค่าเฉลี่ยคะแนนของผู้ใช้ที่ 3 ซึ่งมีค่าเท่ากับ 3

กำหนดให้

m = จำนวนของผู้ใช้ที่ให้คะแนนทั้งหมด

FP = จำนวนความถี่ของคะแนน

Pn = ค่าทำนายคะแนนสินค้า

Avp = ค่าความถี่สัมพัทธ์ = FP/m = 3 / 4 = 0.75

ค่าฐานนิยม มากกว่าค่าเฉลี่ยของผู้ใช้ที่ 3 จึงนำค่าความถี่สัมพัทธ์ 0.75 ไปบวกกับ 2.67

Pn = 2.67 + 0.75 = 3.42 ทศนิยมต่ำกว่า 0.5 จึงปัดลง = 3

ดังนั้นค่าทำนายคะแนนความชอบสินค้า 3 ของผู้ใช้ที่ 3 นี้ คือ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.4 ข้อมูลการให้คะแนนความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.2

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4
U_1	2	5	1	2
U_2	4	3	1	5
U_3	2	4	?	4
U_4	4	5	1	2
U_5	5	2	4	5
U_6	3	1	4	4
U_7	1	2	4	4

ตัวอย่างที่ 3.2 ทำนายคะแนนความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3 ในกรณีที่มีค่าคะแนนความถี่เท่ากัน จะนำค่าฐานนิยมคะแนนนั้นๆ มาหาค่าใกล้เคียงกับค่าเฉลี่ยของผู้ใช้ที่ 3 จากตารางที่ 3.4

ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 3.34 ปัดทศนิยม $\approx$ 3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่3มีผู้ให้คะแนน

คะแนน 1 จำนวน 3 คน

คะแนน 4 จำนวน 3 คน

ดังนั้นค่าฐานนิยมของคะแนนมีค่าเป็น 1 และ 4 จึงทำการหาค่าความใกล้เคียง กับ ค่าคะแนนเฉลี่ยของผู้ใช้ ซึ่งมีค่าคะแนนเฉลี่ยเท่ากับ 3 โดยการนำค่าคะแนนเฉลี่ยมาลบกับค่าฐานนิยมของคะแนนทั้งคู่ แล้วทำเป็นค่าสัมบูรณ์ นำค่าที่น้อยที่สุดเป็นค่าใกล้เคียงกับค่าคะแนนเฉลี่ยของผู้ใช้ เช่น $|3 - 4| = 1$, $|3 - 1| = 2$ ผลคือคะแนน 4 มีค่าใกล้กับคะแนนเฉลี่ย มากกว่า (เพราะ 1 มีค่าน้อยกว่า 2)

ดังนั้นค่าฐานนิยมคะแนนจึงเป็น 4

เมื่อคำนวณ Avp = FP/m = 3 / 6 = 0.50 ค่าฐานนิยมคะแนน มากกว่าค่าคะแนนเฉลี่ยของผู้ใช้ที่ 3 จึงนำค่าความถี่สัมพัทธ์ (Avp) 0.5 ไปบวกกับ 3.34 (ถ้าเกิดค่าฐานนิยมคะแนน น้อยกว่าค่าคะแนนเฉลี่ยของผู้ใช้ จะนำค่าความถี่สัมพัทธ์ไปลบ)

Pn = 3.34 + 0.50 = 3.84 ปัดทศนิยม $\approx$ 4

ค่าคะแนนทำนายความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3 นี้ คือ 4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.5 ข้อมูลการให้คะแนนความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.3

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4
U_1	2	5	2	2
U_2	4	3	2	5
U_3	2	4	?	4
U_4	4	5	2	2
U_5	5	2	4	5
U_6	3	1	4	4
U_7	1	2	4	4

ตัวอย่างที่ 3.3 ทำนายคะแนนความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3ในกรณีที่มีค่าคะแนนความถี่เท่ากัน จะนำค่าฐานนิยมคะแนนนั้นๆ มาหาค่าใกล้เคียงกับค่าเฉลี่ยของผู้ใช้ที่ 3

จากตารางที่ 3.5 ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 3.34 ปัดทศนิยม ≈ 3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มีผู้ให้คะแนน 2 เป็นจำนวน 3 คน และ คะแนน 4 เป็นจำนวน 3 คน ดังนั้นค่าฐานนิยมของคะแนนมีค่าเป็น 2 และ 4 จึงทำการหาค่าความใกล้เคียง กับค่าคะแนนเฉลี่ยของผู้ใช้ ซึ่งมีค่าคะแนนเฉลี่ยเท่ากับ 3 โดยการนำค่าคะแนนเฉลี่ยมาลบกับค่าฐานนิยมของคะแนนทั้งคู่ แล้วทำเป็นค่าสัมบูรณ์นำค่าที่น้อยที่สุดเป็นค่าใกล้เคียงกับค่าคะแนนเฉลี่ยของผู้ใช้ เช่น $|3 - 4| = 1$, $|3 - 2| = 1$ ผล คือ คะแนนมีค่าใกล้เคียงที่เท่ากันดังนั้นจึงจะนำค่าคะแนนเฉลี่ยของผู้ใช้ที่ปัดทศนิยมแล้ว เป็น ค่าคะแนนทำนายความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3 นี้ คือ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างที่ 3.4 ในกรณีที่ค่าฐานนิยมคะแนนมีค่าเท่ากับคะแนนเฉลี่ยของผู้ใช้ให้นำค่าฐานนิยมคะแนนไปทำนายคะแนนสินค้า

ตารางที่ 3.6 ข้อมูลการให้คะแนนสินค้าของผู้ใช้ สำหรับตัวอย่างที่ 3.4

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

จากตารางที่ 3.6 ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 2.67 ปัดทศนิยม $\approx$ 3วิธีการคำนวณโดยจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน 2 เป็นจำนวน 1 คน คะแนน 3 เป็นจำนวน 2 คน และ คะแนน 4 เป็นจำนวน 1 คน ค่าฐานนิยมคะแนนคือ 3 ซึ่งมีค่าเท่ากับค่าเฉลี่ยของผู้ใช้ที่ 3 ที่ปัดทศนิยมแล้ว จึงใช้ค่าฐานนิยมคะแนนเป็นค่าทำนายคะแนนสินค้าชิ้นที่ 3

3.5 การทำนายคะแนนความชอบด้วยค่าความถี่สัมพัทธ์สูงสุดโดยเปรียบเทียบค่าระหว่างค่าฐานนิยมและค่าเฉลี่ยของผู้ใช้ (Relative Frequency Between Point :RFBP)

เนื่องจากสินค้ามีระดับการให้คะแนนที่ไม่เท่ากันในวิธีข้างต้นนั้นไม่สนับสนุนการให้คะแนนที่มีระยะห่างที่มาก เช่น 1-10 , 1-25 ทำให้การแนะนำหรือทำนายคะแนนสินค้ามีความผิดพลาด จึงต้องพัฒนาวิธีทำการทำนายโดยวัดค่าระยะห่างของคะแนนด้วยการนำไปหาค่าระยะห่างของคะแนนสินค้าแล้วนำไปคูณกับความถี่สัมพัทธ์

ขั้นตอนวิธีที่นำเสนอมีดังต่อไปนี้

1. นำค่าเฉลี่ยของผู้ใช้สินค้ามาปัดทศนิยม (หากเกิน 0.5 ปัดขึ้น มิฉะนั้นปัดลง) ให้เป็นค่าจำนวนเต็มของคะแนนสินค้า $\overline{r_{u_i}}$

2. แจกแจงความถี่สัมพัทธ์ของคะแนนสินค้าที่ทำนาย ซึ่งจะได้ค่าฐานนิยมของคะแนนในสินค้าที่ต้องการทำนาย mode ถ้าความถี่ของคะแนนสินค้ามีค่าความถี่ที่เท่ากันให้เลือกค่า mode ที่ใกล้เคียงกับค่าเฉลี่ยของผู้ใช้สินค้า

3. การคำนวณคะแนนทำนายสินค้ามีเงื่อนไขดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1 หาก $\overline{r_{u_i}}$ > mode คำนวณคะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่คงค่าทศนิยม) ลบด้วยค่าความถี่สัมพัทธ์สูงสุดที่คูณกับค่าระยะห่างคะแนน $\overline{r_{u_i}}$ กับ mode

สมการ :คะแนนทำนาย = $\overline{r_{u_i}}$ - ($|\overline{r_{u_i}}$ - mode| * ค่าความถี่สัมพัทธ์สูงสุด)

3.2 หาก $\overline{r_{u_i}}$ < mode คำนวณคะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่คงค่าทศนิยม) บวกด้วยค่าความถี่สัมพัทธ์สูงสุดคูณกับค่าระยะห่างคะแนน $\overline{r_{u_i}}$ กับ mode

สมการ :คะแนนทำนาย = $\overline{r_{u_i}}$ + ($|\overline{r_{u_i}}$ - mode| * ค่าความถี่สัมพัทธ์สูงสุด)

3.3 หาก $\overline{r_{u_i}}$ = mode = คะแนนที่ทำนาย

ตารางที่ 3.7 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.5 ซึ่งจะมีคะแนนตั้งแต่ 1-10

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4	I_5
U_1	1	5	8	2	6
U_2	4	3	5	5	2
U_3	2	4	8	4	6
U_4	4	5	8	2	1
U_5	5	6	4	5	5
U_6	7	4	2	8	8
U_7	1	2	4	4	4

ตัวอย่างที่ 3.5 ทำนายคะแนนความชอบสินค้า สินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 6 จะใช้วิธีค่าความถี่สัมพัทธ์ในการคำนวณและใช้ค่าระยะห่างของค่าฐานนิยมกับค่าเฉลี่ยของผู้ใช้นำมาคูณด้วยเพื่อลดปัญหาความผิดพลาดของการทำนายคะแนนสินค้า

ค่าเฉลี่ยของ ผู้ใช้คนที่ 6 = 27/5 = 5.4 ปัดทศนิยม $\approx$ 5

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน

คะแนน 4 เป็นจำนวน 2 คน

คะแนน 5 เป็นจำนวน 1 คน

คะแนน 8 เป็นจำนวน 3 คน

ค่าฐานนิยมของสินค้าชิ้นที่ 3 คือ 8 ซึ่งมากว่าค่าเฉลี่ยของผู้ใช้ที่มีค่าเป็น 5 ดังนั้นการปรับค่าจะนำค่าเฉลี่ยของผู้ใช้นำไปบวก หากน้อยกว่าให้นำไปลบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คำนวณแบบค่าระยะห่างของคะแนนฐานนิยมและค่าเฉลี่ยของผู้ใช้

กำหนดให้

m = จำนวนของผู้ให้คะแนนทั้งหมด

FP = จำนวนความถี่ของคะแนน

Pn = ค่าทำนายคะแนนสินค้า

Pnb = ค่าเฉลี่ยของผู้ใช้ +(ค่าระยะห่างคะแนนฐานนิยมและค่าเฉลี่ยของผู้ใช้ * ค่าความถี่สัมพัทธ์)

Avp = ค่าความถี่สัมพัทธ์ = FP/m = 3 / 6 = 0.5

ค่าฐานนิยม= 8

ค่าเฉลี่ยของผู้ใช้ = 5.4

ค่าระยะห่างของคะแนน = |8-5.4| = 2.6

Pnb = 5.4 + (0.5*2.6) = 6.7 ทศนิยมมากกว่า 0.5จึงปัดขึ้น = 7

ดังนั้นค่าทำนายคะแนนความชอบสินค้าชิ้นที่ 3 ของผู้ใช้ที่ 6 นี้ คือ 7

3.6 การทำนายคะแนนความชอบด้วยค่าเฉลี่ยของสินค้า(Mean Item)

ในวิธีขั้นตอนนี้จะใช้ค่าเฉลี่ยของคะแนนสินค้าแทนค่าความถี่สัมพัทธ์ของค่าฐานนิยมสินค้าโดยจะนำค่าเฉลี่ยคะแนนสินค้าที่จะทำนายมาเปรียบเทียบกับค่าคะแนนเฉลี่ยของผู้ใช้สินค้า ถ้าหากมีค่ามากกว่าให้นำค่าเฉลี่ยผู้ใช้ไปบวก หรือหากมีค่าน้อยกว่าให้นำไปลบ และถ้ามีค่าเท่ากัน(ค่าที่ปัดทศนิยมแล้ว) ให้นำค่านั้นเป็นค่าทำนายคะแนนสินค้า

ตารางที่ 3.8 ข้อมูลการให้ความชอบสินค้าของผู้ใช้ ตัวอย่างที่ 3.6 ซึ่งจะมีคะแนนตั้งแต่ 1-10

สินค้า / ผู้ใช้	I_1	I_2	I_3	I_4
U_1	9	5	4	8
U_2	8	3	4	5
U_3	2	7	2	5
U_4	4	5	3	6
U_5	7	6	4	5
U_6	4	4	?	2
U_7	8	5	4	4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างที่ 3.6 ใช้วิธีค่าเฉลี่ยสินค้าในการคำนวณและใช้ค่าคะแนนฐานนิยมมาคูณในการเพิ่มลดคะแนนเฉลี่ยของผู้ใช้ในการทำนายคะแนนสินค้าเพื่อลดปัญหาระยะห่างของคะแนน

ค่าเฉลี่ยของ ผู้ใช้คนที่ 6 = 10/3 = 3.4 ปัดทศนิยม ≈ 3

ซึ่งจะดูจากค่าเฉลี่ยของคะแนนสินค้าที่ 3 (4+4+2+3+4+4)/6 = 3.5

ค่าเฉลี่ยคะแนนสินค้ามากว่าค่าเฉลี่ยของผู้ใช้ที่มีค่าเป็น 3 ดังนั้นการปรับค่าจะนำค่าเฉลี่ยของผู้ใช้นำไปบวก หากน้อยกว่าให้นำไปลบ

กำหนดให้

m = จำนวนของผู้ให้คะแนนทั้งหมด

Pnb = ค่าเฉลี่ยของผู้ใช้ +(ค่าระยะห่างของค่าเฉลี่ยสินค้าและค่าเฉลี่ยของผู้ใช้ * ค่าเฉลี่ยสินค้า)

ค่าเฉลี่ยของผู้ใช้ = 3

ค่าเฉลี่ยสินค้า = 3.5

ค่าระยะห่างของคะแนน = |3.5 -3| = 0.5

Pnb = 3.5 + (0.5*3.5) = 5.25

ดังนั้นค่าทำนายคะแนนความชอบสินค้าชิ้นที่ 3 ของผู้ใช้ที่ 6 นี้ คือ 5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการทดลอง

จากบทที่ 3 ได้กล่าวถึงการทำนายคะแนนสินค้าทั้งหมด 3 วิธี การใช้ค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยมสินค้า (Relative Frequency: RF) วิธีเปรียบเทียบเปรียบเทียบค่าเฉลี่ยข้อมูลสินค้า (Mean Item) และวิธีการเปรียบเทียบค่าความถี่สัมพัทธ์โดยใช้ค่าระหว่างคะแนน (Relative Frequency Between Point :RFBP) ในส่วนบทที่2 วิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิม (Collaborative Filtering: CF) ในบทนี้จะแสดงผลการทดลองทั้ง 4 วิธี ซึ่งจะแสดงผลค่าความคลาดเคลื่อนสัมบูรณ์ที่วัดประสิทธิภาพการแนะนำสินค้า

4.1 ขั้นตอนการเตรียมชุดข้อมูล

4.1.1 การทดลองด้วยชุดข้อมูลทดสอบ Movie Lens และ Book Crossing

1.ชุดข้อมูลทดลอง ของ Movie Lens 100,000 ชุด ซึ่งการให้คะแนนภาพยนต์มีเกณฑ์ตั้งแต่ 1-5 ค่าระดับคะแนนความชอบ 1 คือชอบน้อยที่สุดค่าระดับความชอบ 5 คือชอบมากที่สุด โดยข้อมูลนี้จะถูกแบ่งข้อมูลออกเป็น 5 ชุด แต่ละชุดข้อมูลจะแบ่งออกเป็นชัอมูลฝึกสอน (training set) 80% และข้อมูลทดสอบ(test set) 20%

2.ชุดข้อมูลทดลอง Book Crossing ข้อมูล 1,149,780 ซึ่งการให้คะแนนหนังสือที่มีเกณฑ์ตั้งแต่ 1 – 10 โดย 1 คะแนน คือความชอบน้อยสุดและ 10 คะแนน คือชอบมากที่สุด โดยจะนำข้อมูลที่ผู้ใช้ให้คะแนนสินค้าที่มีค่าไม่ใช่ 0 มาใช้ในการทดลอง ซึ่งจะมีอยู่ 361,370 ชุด โดยจะทำการ เอาชุดข้อมูลแถวที่ทดสอบออกมาที่ละ 1 แถว และชุดแถวข้อมูลที่เหลือเป็นชุดข้อมูลฝึกสอน (training set)

3.โดยในแต่ละชุดข้อมูลทดสอบจะถูกนำมาคำนวณความน่าจะเป็นที่ผู้ใช้ u_j จะให้คะแนนความชอบแก่สินค้า i_k ด้วยค่าคะแนนความชอบ r ในกาทำนายคะแนนสินค้าด้วยวิธีการที่เสนอไปทั้งหมด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2 ผลการทดลอง

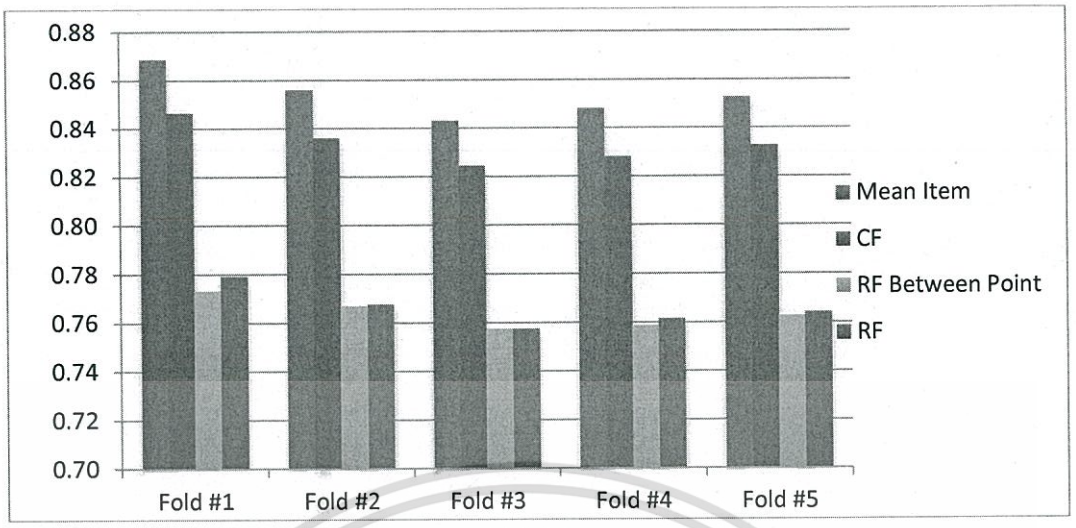
4.2.1 ผลการทดลองด้วยชุดข้อมูลทดสอบ Movie Lens

การทดสอบประสิทธิภาพในการทำนายเปรียบเทียบกันในแต่ละชุดข้อมูลด้วยชุดทดสอบ Movie Lens ระหว่างเทคนิคการกรองข้อมูลผู้ใช้ร่วมแบบเดิม กับ วิธีการใช้ค่าความถี่สัมพัทธ์สูงสุด คะแนนสินค้า การใช้ค่าเฉลี่ยคะแนนสินค้า และวิธีสุดท้าย วิธีการใช้ค่าความถี่สัมพัทธ์โดยวัดค่าระหว่างคะแนน ผลการทดลองค่าความคลาดเคลื่อนสัมบูรณ์ที่ได้แสดงดังตารางที่ 4.1

ตารางที่ 4.1: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบย่อยของ Movie Lens

ชุดข้อมูลทดสอบย่อย Movie Lens	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	ผลรวม
วิธีกรองผู้ใช้ร่วมแบบเดิม (Traditional Collaborative Filtering: TCF)	0.85	0.84	0.82	0.83	0.83	0.83
วิธีการใช้ค่าความถี่สัมพัทธ์สูงสุด (Relative Frequency: RF)	0.78	0.77	0.78	0.76	0.76	0.77
วิธีการใช้ค่าเฉลี่ยข้อมูลสินค้า (Mean Item)	0.86	0.85	0.84	0.85	0.85	0.85
วิธีการใช้ค่าความถี่สัมพัทธ์โดยใช้ค่าระหว่างคะแนน (Relative Frequency Between Point :RFBP)	0.77	0.76	0.76	0.76	0.76	0.76

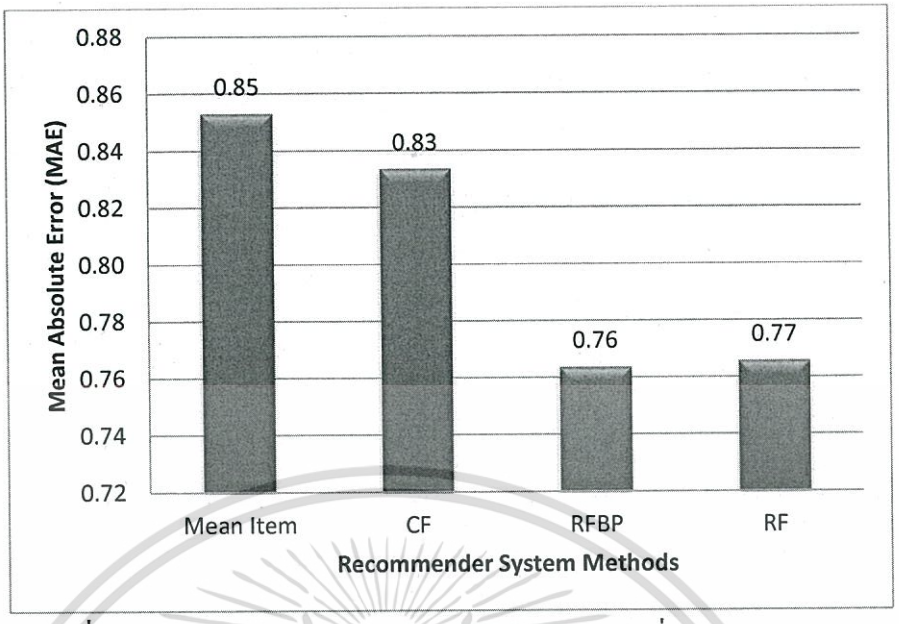
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.1: กราฟเปรียบเทียบประสิทธิภาพการทำนายแต่ละชุดข้อมูลของ Movie Lens

จากการทดลองวัดประสิทธิภาพการทำนายด้วย 4 วิธี คือ วิธีการกรองข้อมูลผู้ใช้ร่วมด้วยการเปรียบเทียบค่าคะแนนของค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยม (RF) วิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิม (CF) วิธีเปรียบเทียบเปรียบเทียบค่าเฉลี่ยข้อมูลสินค้า (Mean Item) โดยการนำค่าของคะแนนสินค้ามาเฉลี่ยเปรียบเทียบค่าเฉลี่ยของผู้ใช้สินค้า และสุดท้ายคือวิธีการค่าความถี่สัมพัทธ์สูงสุดโดยเปรียบเทียบค่าระหว่างค่าฐานนิยมและค่าเฉลี่ยของผู้ใช้ (RFBP) ซึ่งวิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิมนั้นมีประสิทธิภาพในการทำนายน้อยที่สุด กับชุดข้อมูลย่อยที่ 5 การเปรียบเทียบค่าคะแนนของค่าความถี่สัมพัทธ์ร่วมกับการหาค่าระหว่างคะแนนของแบบทั้งชุดข้อมูลนั้นมีประสิทธิภาพที่ดีที่สุด แต่วิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิมนั้นมีประสิทธิภาพในการทำนายน้อยกว่าวิธีที่นำเสนอ (RF) และ(RFBP) ตามรูปที่ 4.1 แสดงการเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์แต่ละชุดข้อมูลย่อยของชุดข้อมูล Movie Lens โดยแยกตามแต่ละเทคนิคที่ใช้ในการเปรียบเทียบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.2 กราฟเปรียบเทียบประสิทธิภาพการทำนายค่าเฉลี่ยของ Movie Lens

จากผลการทดสอบประสิทธิภาพในการทำนายที่ได้ในแต่ละชุดข้อมูลย่อยของ Movie Lens นั้น เมื่อนำผลมารวมเฉลี่ยในทั้งหมด 5 ชุด แล้วจะเห็นผลได้จากรูปที่ 4.2 ซึ่งแสดงค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์รวมทุกชุดข้อมูลของแต่ละวิธีการกรองข้อมูลผู้ใช้ร่วม ผลที่ได้จะเห็นว่าสำหรับชุดข้อมูล Movie Lens นั้น วิธีการปรับค่าเฉลี่ยของผู้ใช้โดยวัดจากของเฉลี่ยของคะแนน (Mean Item) สินค้ามีประสิทธิภาพการทำนายน้อยที่สุด ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.85 รองลงมาคือวิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิม (CF) อยู่ที่0.83 และวิธีการเปรียบเทียบค่าคะแนนของค่าความถี่สัมพัทธ์สูงสุดและค่าฐานนิยม (RF) อยู่ที่ 0.77 และสุดท้ายคือวิธีการค่าความถี่สัมพัทธ์สูงสุดโดยเปรียบเทียบค่าระหว่างค่าฐานนิยมและค่าเฉลี่ยของผู้ใช้ (RFBP) ทั้งชุดข้อมูลมีประสิทธิภาพการทำนายมากที่สุดโดยมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.76

4.2.2 ผลการทดลองด้วยชุดข้อมูลทดสอบ Book Crossing

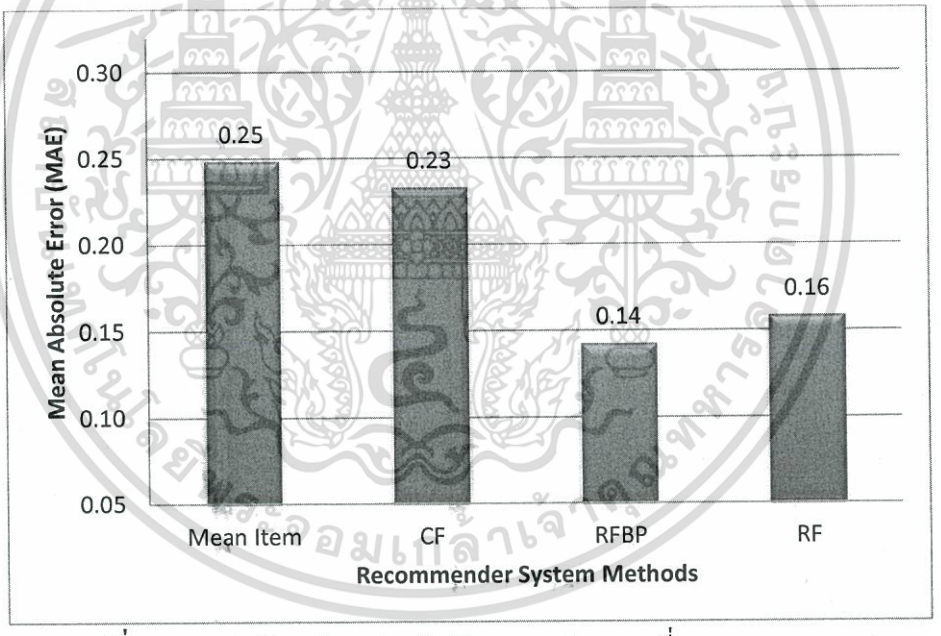
การทดสอบประสิทธิภาพในการทำนายเปรียบเทียบกันในแต่ละชุดข้อมูลด้วยชุดทดสอบ Book crossing ระหว่างเทคนิควิธีการกรองผู้ใช้ร่วมด้วยการเปรียบเทียบค่าคะแนนของค่าความถี่สัมพัทธ์ที่นำเสนอทั้งแบบการเปรียบเทียบความถี่สัมพัทธ์ระยะห่างของคะแนน ผลการทดลองค่าความคลาดเคลื่อนสัมบูรณ์ที่ได้แสดงดังตารางที่ 4.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.2: แสดงความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดทดสอบของ Book Crossing

ชุดข้อมูลทดสอบ Book Crossing	ค่าความคลาดเคลื่อนสัมบูรณ์
วิธีกรองผู้ใช้ร่วมแบบเดิม(Collaborative Filtering: CF)	0.23
วิธีการใช้ค่าความถี่สัมพัทธ์สูงสุด (Relative Frequency:RF)	0.16
วิธีการใช้ค่าเฉลี่ยข้อมูลสินค้า(Mean Item)	0.25
วิธีการใช้ค่าความถี่สัมพัทธ์โดยใช้ค่าระหว่างคะแนน (Relative Frequency Between Point :RFBP)	0.14

ค่าที่ได้จากการทดลองจะเห็นได้ว่าประสิทธิภาพการทำนายวัดโดยค่าความคลาดเคลื่อนสัมบูรณ์



รูปที่ 4.3 กราฟเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยของ Book Crossing

จากผลการทดสอบประสิทธิภาพในการทำนายที่ได้ในชุดข้อมูล ของ Book Crossing นั้น เมื่อนำผลมารวมเฉลี่ยในแต่ละชุดข้อมูลแล้วจะเห็นได้จากภาพที่ 4.3 ซึ่งแสดงค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ของชุดข้อมูลในแต่ละวิธีการกรองข้อมูลผู้ใช้ร่วม ผลที่ได้จะเห็นว่าชุดข้อมูล Book Crossing นั้น วิธีการปรับค่าเฉลี่ยของผู้ใช้โดยวัดจากของเฉลี่ยของคะแนนสินค้า (Mean Item) มีประสิทธิภาพการทำนายน้อยที่สุด ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.24 รองลงมาคือวิธีการกรองข้อมูลผู้ใช้ร่วมแบบเดิม (CF) นั้นอยู่ที่ 0.23 วิธีการเปรียบเทียบค่าคะแนน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ของค่าความถี่สัมพัทธ์สูงสุด (RF) อยู่ที่0.16 และสุดท้ายคือวิธีการค่าความถี่สัมพัทธ์สูงสุดโดยเปรียบเทียบค่าระหว่างค่าฐานนิยมและค่าเฉลี่ยของผู้ใช้ (RFBP) มีประสิทธิภาพการทำนายมากที่สุดโดยมีค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์อยู่ที่ 0.14

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเทคนิคทางเลือกที่ใช้ในระบบช่วยแนะนำเพื่อทำนายคะแนนสินค้าโดยใช้ค่าความถี่สัมพัทธ์สูงสุดของคะแนนสินค้า แทนการใช้การกรองข้อมูลผู้ใช้ร่วมแบบเดิม (CF) ที่ใช้ค่าความคล้ายของพฤติกรรมการให้ความนิยมของผู้ใช้ในระบบ โดยมีแนวคิดที่ว่า สินค้าหรือบริการที่ดีและมีคุณภาพก็ย่อมจะมีความน่าจะเป็นที่จะได้รับความนิยมจากผู้บริโภคมากกว่าสินค้าหรือบริการที่ไม่มีคุณภาพ

โดยชุดข้อมูลที่นำมาวิจัยนั้นมีอยู่ด้วยกัน 2 ชุด ได้แก่ Movie Lens ซึ่งเป็นชุดข้อมูลที่นิยมใช้ในการทดลองระบบช่วยแนะนำสินค้า ชุดข้อมูลที่มีขนาดเล็กเพียง 100,000 ชุดข้อมูล และ Book Crossing ซึ่งไม่ค่อยพบเห็นในการทดลองมากนักเนื่องจากเป็นชุดข้อมูลที่มีขนาดใหญ่มากถึง 1,149,780 ชุดข้อมูล แต่ในการงานวิจัยนี้จะเอาแต่ข้อมูลที่ผู้ใช้สินค้าที่ให้คะแนนไม่ใช่ 0 และมีผู้ให้คะแนนจริงในสินค้า ซึ่งการทดลองนี้มีทั้งหมด 361,370 ชุด โดยการวัดประสิทธิภาพการทำนายคะแนนสินค้าจะคำนวณด้วยค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) คือจะนำผลรวมของค่าการทำนายในแต่ละชุดข้อมูลที่ได้มาคำนวณ โดยหากค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ยิ่งน้อยลงเท่าไรแสดงว่าการทำนายข้อมูลความนิยมนั้นให้ความแม่นยำมากขึ้นเท่านั้น

จากผลการทดลองในบทที่ 4 ที่ได้ทดลองกับชุดข้อมูลดังกล่าวข้างต้นการทำนายคะแนนความชอบสินค้าในระบบช่วยแนะนำด้วยการกรองข้อมูลผู้ใช้ร่วมในงานวิจัยนี้ได้เสนอวิธีใช้ค่าความถี่สัมพัทธ์สูงสุด ในการทำนายคะแนนความชอบของผู้ใช้ที่มีต่อสินค้า โดยการนำค่าเฉลี่ยของผู้ใช้มาปรับค่าคะแนนด้วยการเปรียบเทียบค่าความถี่สัมพัทธ์สูงสุดตามเงื่อนไขที่กำหนด ผลลัพธ์ที่ได้พบว่ามีค่าผิดพลาดในการทำนายคะแนนน้อยกว่า วิธีการกรองข้อมูลแบบผู้ใช้ร่วมแบบเดิมและแก้ไขปัญหาการเริ่มต้นทำงานช้าของระบบได้ โดยในค่า time complexity ของวิธีที่ใช้ค่าความถี่สัมพัทธ์สูงสุด O(m) หรือ O(n) ซึ่ง m คือจำนวนของผู้ใช้ทั้งหมดที่เคยให้คะแนน และ n คือ จำนวนสินค้าทั้งหมดที่ผู้ใช้เคยให้คะแนน ในขณะที่วิธีกรองข้อมูลผู้ใช้ร่วมแบบเดิม(CF) มีค่า time complexity เท่ากับ O(mn) แต่ทั้ง 2 วิธีนี้ยังมีจุดอ่อนอยู่ที่ความเบาบางข้อมูล ถ้าเกิดข้อมูลสินค้าไม่เคยมีผู้ใช้ให้คะแนนมาก่อนจะทำให้ประสิทธิภาพในการทำนายลดลง และในวิธีที่เสนอนี้จะง่ายต่อการเข้าใจและลดเวลาในการคำนวณเพราะจะไม่คำนวณความคล้ายคลึงของผู้ใช้ร่วม จากผลการทดลองจะเห็นว่าค่าผิดพลาดในการทำนายคะแนนสินค้าลดลง

งานวิจัยในอนาคตจะทำการประยุกต์ใช้เทคนิควิธีการกรองแบบผสมผสานโดยการคิดค่าเฉลี่ยของผู้ใช้ให้ดูจากเนื้อหาของสินค้า เพื่อให้การเริ่มต้นค่าทำนายคะแนนมีความถูกต้องยิ่งขึ้น หรือใช้เทคนิค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แนวคิดนี้ร่วมกับการเติมค่าข้อมูลที่ขาดหายเพื่อแก้ไขปัญหาความเบาบางของข้อมูล ซึ่งคิดว่าน่าจะทำให้ระบบมีประสิทธิภาพในการแนะนำมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[1] Paul Resnick,and Hal R. "Recommender System", **COMMUNICATIONS OF THE ACM, vol. 40, Issue 3, March 1997,** 1997.

[2] Robin van Meteren and Meteren van Someren ,"Using Content-Based Filtering for Recommendation" ,**ECML 2000 Workshop,** 2000

[3] J.Ben Schafer,Dan Frankowski,Jon Herlocker,and Dhilad Sen, "Collaborative Filtering Recommender System", P.Brusilovsky, A.Kobsa,and W.Nejdl(Eds.):**The Adaptive Web, LNCS 4321, Pages:291-324, 2007**

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

[5] GroupLens Research, "MovieLens Project Web site", **http://movielens.umn.edu.,1997**

[6] Melville, P., Mooney, R. J. และ Nagarajan, R. "Content-Boosted Collaborative Filtering for Improved Recommendations" , 2002

[7] Torres, Roberto, "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** ,2004

[8] วงกต ศรีอุไร, ชูชาติ หฤไชยะศักดิ์, จิรารัตน์ สิทธิวรชาติ, "การแทนค่าข้อมูลที่ขาดหายเพื่อแก้ไขปัญหาความเบาบางของข้อมูล ในการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม" The Use of Imputation Technique for Solving Sparsity Problem in Collaborative **Filtering , วารสารพระจอมเกล้าลาดกระบัง: ปีที่ 16 ฉบับที่ 1 เดือนเมษายน 2551**

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

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก

งานวิจัยที่ตีพิมพ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

THE 8TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY

PROCEEDINGS OF NCCIT 2012

THE 8TH NATIONAL CONFERENCE ON COMPUTING AND INFORMATION TECHNOLOGY

9-10 MAY 2012
DUSIT THANI HOTEL, PATTAYA CITY, THAILAND

WWW.NCCIT.NET

FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK

บทความวิจัย

การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
ครั้งที่ 8

9-10 พฤษภาคม 2555
โรงแรมดุสิตธานี พัทยา

IEEE
THAILAND SECTION

คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

WWW.IT.KMUTNB.AC.TH

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กฎความสัมพันธ์และค่าฐานนิยมของการกรองข้อมูลผู้ใช้ร่วมในระบบแนะนำ
An Association Rules with Mode BasedCollaborative Filtering in Recommender System

พีระศักดิ์ แซ่ตั้น(Peerasuk SaeTan)[1] และนวลสวาท หิรัญสกลวงศ์ (Nualsawat Hiransakolwong)[2]
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
diewd2a@gmail.com[1], khnualsa@kmitl.ac.th[2]

บทคัดย่อ

ในปัจจุบันมีสินค้าให้ผู้บริโภคเลือกซื้อได้หลากหลายและบริการมากมายในระบบพาณิชย์อิเล็กทรอนิกส์จึงจำเป็นต้องมีระบบแนะนำสินค้าให้แก่ผู้ใช้หรือผู้บริโภค โดยการทำนายคะแนนความชอบสินค้าที่จะทำการแนะนำ ซึ่งวิธีที่นิยมใช้ คือ การกรองข้อมูลผู้ใช้ร่วม (Collaborative Filtering:CF) แต่วิธีนี้ยังประสบปัญหาความถูกต้องและการประมวลผลเบื้องต้นมีความล่าช้าเนื่องจากต้องใช้กับฐานข้อมูลที่มีขนาดใหญ่ซึ่งต้องเปรียบเทียบข้อมูลของผู้ใช้ร่วมงานวิจัยนี้จึงนำเสนอวิธีการทำนายคะแนนความชอบสินค้าของผู้ใช้โดยใช้กฎความสัมพันธ์และค่าฐานนิยมในการแนะนำสินค้าซึ่งวิธีการใหม่นี้ลดความซับซ้อนในการคำนวณและใช้เวลาในการคำนวณได้รวดเร็วกว่าการกรองข้อมูลแบบผู้ใช้ร่วมแบบดั้งเดิม (เมื่อ m แทน จำนวนผู้ใช้ทั้งหมดจะได้ time complexity ของวิธีที่นำเสนอคือO(m) ขณะที่วิธีดั้งเดิม CF มี time complexity เท่ากับ $O(m^2)$)ขณะที่ผลลัพธ์ที่ได้พบว่าวิธีที่นำเสนอนี้มีประสิทธิภาพสูงกว่าวิธีCFโดยพิจารณาจากค่าความคลาดเคลื่อนสมบูรณ์(Mean Absolute Error: MAE)ซึ่งผลการทดลองกับชุดข้อมูล MovieLens [5]วิธีที่นำเสนอมีค่าMAE น้อยกว่าค่าMAE ของวิธี CF

คำสำคัญ:ระบบช่วยแนะนำ การกรองร่วม กฎความสัมพันธ์

Abstract

Currently, there are so many products and services that consumers choose to buy in electronic commerce. Users require a recommender system to filter only items that hit their interests. Most techniques in recommender system are used in combination Collaborative Filtering (called CF), but these techniques are so complicated and spend too much time to compute the preprocessing step with a large database. This research proposes a new method predicting scores of users' interest by using new association rules and popularity frequency in products. This new method does not complicate and can calculate faster than the CF. The proposed approach has time complexity O(m) while The CF approach has time complexity $O(m^2)$, where m is the number of users. The experimental results with MovieLensdataset[5] show that the new approach gains more performance with less error (Mean Absolute Error:MAE) and faster computation time than the CF method.

Keyword: Recommender System, Collaborative Filtering, Association Rules

1.บทนำ

ระบบแนะนำข้อมูล (Recommender System) [1] คือ ระบบที่ช่วยในการแนะนำสินค้าที่คาดว่าผู้ใช้จะให้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ความสนใจในการเลือกซื้อหรือพิจารณา ระบบนี้กลายเป็นระบบพื้นฐานที่ช่วยในการเข้าถึงข้อมูล ทำให้สามารถได้รับข้อมูลที่มีคุณภาพจากข้อมูลจำนวนมาก ซึ่งข้อมูลจะตรงกับความต้องการของผู้ใช้ โดยเทคนิคทั่วไปที่ใช้ในระบบช่วยแนะนำแบ่งออกเป็น 2 วิธี ได้แก่ การกรองข้อมูลพื้นฐานสินค้า (Content-Based Filtering) [2] เช่นชื่อสินค้า ประเภทสินค้า เป็นต้น มาเป็นข้อมูลที่ใช้ในการเลือกสินค้าแนะนำ และการกรองข้อมูลผู้ใช้ร่วม(Collaborative Filtering) [3],[4] เช่นการเลือกซื้อสินค้าร่วมกันของผู้ใช้ มาเป็นข้อมูลที่ใช้ในการเลือกสินค้าแนะนำ

การกรองข้อมูลผู้ใช้ร่วมเป็นวิธีการที่นิยมใช้กันมากกว่าการใช้ข้อมูลพื้นฐานของสินค้า เนื่องจากมีความแม่นยำในการแนะนำสินค้ามากกว่าการกรองข้อมูลพื้นฐานสินค้าแต่วิธีนี้ยังมีปัญหาการประมวลผลเบื้องต้นมีความล่าช้า และปัญหาความถูกต้องของข้อมูล ซึ่งในการกรองข้อมูลผู้ใช้ร่วมจะมีขั้นตอนการเปรียบเทียบข้อมูลของผู้ใช้สินค้าจากแต่ละผู้ใช้จึงทำให้ใช้เวลาคำนวณนานเมื่อประมวลผลจากฐานข้อมูลที่มีผู้ใช้จำนวนมาก โดยหลักการทำงานทั่วไปของระบบแนะนำจะต้องเพิ่มข้อมูลผู้ใช้งานอยู่ตลอดเวลาเพื่อเพิ่มความถูกต้องของระบบในการแนะนำ

ดังนั้นในงานวิจัยนี้ได้เสนอวิธีการใหม่ซึ่งช่วยแก้ไขปัญหาการประมวลผลเบื้องต้นมีความล่าช้า และลดความคลาดเคลื่อนในการแนะนำสินค้า โดยใช้วิธีกฎความสัมพันธ์และใช้ค่าฐานนิยมสินค้าเพื่อลดขั้นตอนการเปรียบเทียบข้อมูลของผู้ใช้ร่วมซึ่งทำให้ลดขั้นตอนการทำงานของระบบและเพิ่มประสิทธิภาพในการแนะนำสินค้า

2. ทฤษฎี

2.1 ทฤษฎีพื้นฐาน

ระบบแนะนำเป็นรูปแบบหนึ่งของระบบที่ใช้เทคนิคในการกรองข้อมูล โดยจะทำการกรองข้อมูลพื้นฐานของสินค้าต่างๆ อาทิ เช่น ภาพยนตร์ เพลง หนังสือ ข่าว รูปภาพ หรือเว็บเพจ ที่มีการจัดเก็บไว้เป็นจำนวนมากออก เพื่อให้เหลือสิ่งที่ตรงกับความต้องการหรือความสนใจของผู้ใช้ ซึ่งการทำงานของระบบจะเริ่มจากการนำข้อมูลที่ได้ผ่านการวิเคราะห์และคำนวณค่าคะแนน เพื่อแยกว่าข้อมูลที่มีค่าคะแนนมากจึงจะถือว่าเป็นข้อมูลที่มีความเกี่ยวข้องหรือตรงความต้องการของผู้ใช้ นอกจากนี้ระบบแนะนำทั่วไปจะทำการเปรียบเทียบประวัติผู้ใช้กับลักษณะเฉพาะบางอย่างที่สามารถใช้อ้างอิงถึงสินค้าต่างๆ ที่ผู้ใช้ต้องการโดยลักษณะเฉพาะเหล่านี้อาจใช้สาระสำคัญของข้อมูลหรือใช้ความสัมพันธ์กันของกลุ่มผู้ใช้

ระบบแนะนำส่วนใหญ่ แบ่งตามเทคนิค วิธีการใช้ออกเป็น 3ประเภท ได้แก่

1.การกรองข้อมูลพื้นฐานสินค้า (Content-Based Filtering) ใช้ข้อมูลพื้นฐานของสินค้า เป็นเทคนิควิธีการที่จะเลือกแนะนำสินค้าโดยดูความสัมพันธ์ระหว่างชื่อสินค้า หรือรายละเอียดของสินค้ามาเปรียบเทียบกับสิ่งที่ผู้ใช้ต้องการ เพื่อนำมาเป็นข้อมูลพื้นฐานในการแนะนำสินค้าให้กับผู้ใช้

2.การกรองข้อมูลผู้ใช้ร่วม (Collaborative Filtering) การกรองข้อมูลของกลุ่มผู้ใช้ร่วมเป็นเทคนิควิธีการสำรวจพฤติกรรมการใช้สินค้าของผู้ใช้ โดยจะสังเกตว่าสิ่งใดที่ผู้ใช้ทำร่วมกันหรือการให้คะแนนสินค้าที่เท่ากัน แล้วทำการเปรียบเทียบกับผู้ใช้ที่ใกล้เคียงกัน แล้วนำข้อมูลนั้นมาคำนวณเพื่อทำนายสินค้าที่ควรแนะนำให้กับผู้ใช้

3.การกรองแบบผสมผสาน (Hybrid Recommender Systems)การกรองแบบผสมผสาน [4] เป็นวิธีการที่รวมวิธีการใช้ข้อมูลพื้นฐานของสินค้าและวิธีการกรองข้อมูลผู้ใช้ร่วมเพื่อนำข้อดีของทั้งสองวิธีมาพัฒนาเป็นระบบใหม่ที่ดีขึ้น

2.1.1 การแจกแจงความถี่สัมพัทธ์

ความถี่สัมพัทธ์ (Relative Frequency) คือ อัตราส่วนของความถี่ในแต่ละลำดับข้อมูลเทียบกับจำนวน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ความถี่ทั้งหมดผลรวมของความถี่สัมพัทธ์ทั้งหมดมีค่าเท่ากับ 1 เสมอ

ตารางที่ 1:ตัวอย่างข้อมูลหาความถี่สัมพัทธ์

ลำดับชั้นที่	ความถี่ f_i	ความถี่สัมพัทธ์(f_i / S f_i)
1	3	3/13 = 0.23
2	4	4/13 = 0.307
3	2	2/13 = 0.153
4	1	1/13 = 0.077
รวม	13	1

2.1.2 ค่าฐานนิยม (mode)

ค่าฐานนิยม คือ ตัวเลขหรือข้อมูลที่มีค่าความถี่มากที่สุด ตัวอย่าง เช่น ค่าฐานนิยมในตารางที่ 1 คือลำดับชั้นที่ 2 เพราะมีค่าความถี่เท่ากับ 4 เป็นค่าที่เกิดขึ้นมากที่สุดในกลุ่มข้อมูล ถ้าแต่ละกลุ่มข้อมูลไม่มีค่าซ้ำกันเลยหรือค่าความถี่เท่ากับหนึ่ง จะไม่สามารถบอกค่าฐานนิยมได้

2.2 ขั้นตอนระบบการแนะนำสินค้าโดยใช้เทคนิคการกรองข้อมูลผู้ใช้ร่วม (Collaborative Filtering)

ในระบบแนะนำสินค้าจะใช้ข้อมูลของสินค้าและข้อมูลคะแนนของผู้ที่เคยใช้สินค้ามาทำการคำนวณสร้างรายการแนะนำให้กับผู้ใช้อื่นที่ยังไม่เคยใช้สินค้าด้วยเทคนิควิธีการกรองข้อมูลผู้ใช้ร่วมเป็นวิธีการที่นิยมนำมาใช้ในระบบแนะนำสินค้าเนื่องจากวิธีการนี้จะใช้วิธีการค้นหาข้อมูลของผู้ใช้ที่เป็นสมาชิกข้างเคียงที่มีความชอบสินค้าที่คล้ายกันกับผู้ใช้ซึ่งข้อมูลของผู้ใช้ที่เป็นสมาชิกใกล้เคียงจะถูกนำมาวิเคราะห์และแนะนำสินค้าให้ตรงกับความต้องการของผู้ใช้โดยมี4 ขั้นตอนดังนี้

2.2.1 ขั้นตอนการเตรียมข้อมูล

ตัวอย่างข้อมูลที่ทดลองนี้จะมีข้อมูลสินค้าและผู้ใช้สินค้าซึ่งผู้ใช้แต่ละคนจะมีค่าคะแนนความชอบที่มีต่อสินค้าดังตัวอย่างในภาพที่ 1ซึ่งเป็นตัวอย่างคะแนนที่ผู้ใช้แต่ละคนมีต่อสินค้าในแต่ละชิ้นในที่นี้สมมุติให้มีผู้ใช้ 4 คนและมีสินค้าอยู่ 5 ชิ้น โดยค่าคะแนนความชอบมีช่วงตั้งแต่ 1 ถึง 5 ซึ่งค่าคะแนนที่ 1 แสดงว่าไม่ชอบส่วนค่าคะแนนที่ 5 คือชอบมากที่สุดโดยที่เครื่องหมาย "?" หมายความว่ายังไม่มีการให้คะแนนกับสินค้าชิ้นนั้น

สินค้า / ผู้ใช้	1	2	3	4	5
1	4	5	1	3	3
2	4	3	2	3	3
3	2	?	2	3	2
4	1	3	1	4	1

No Co-rate Item Co-rated Item

ภาพที่ 1 ตัวอย่างข้อมูลคะแนนสินค้าของผู้ใช้แต่ละคน

2.2.2 ขั้นตอนการค้นหาสมาชิกใกล้เคียง

ขั้นตอนนี้เป็นกระบวนการค้นหาสมาชิกที่อยู่ในระบบที่มีความชอบคล้ายคลึงกันกับผู้ใช้ที่ต้องการคำแนะนำอยู่ในขณะนี้จากภาพที่ 1 ถ้าต้องการหาว่าผู้ใช้ที่ 4มีความชอบคล้ายคลึงกับผู้ใช้ที่ 3 หรือไม่จะต้องพิจารณาจากค่าคะแนนความชอบของผู้ใช้ที่ 4 กับผู้ใช้ที่ 3 ที่มีให้กับสินค้าชิ้นเดียวกันซึ่งจะเรียกว่า co-rated items ในการหาความคล้ายคลึงของผู้ใช้ทั้งสองคนจะนิยมใช้สหสัมพันธ์ของเพียร์สัน (Pearson correlation coefficient) [4] ซึ่งค่าที่คำนวณได้มีค่ามากแสดงว่าผู้ใช้สองคนนี้มีความคล้ายคลึงกันแต่ถ้าค่าที่คำนวณได้มีค่าน้อยแสดงว่าผู้ใช้สองคนนี้มีความคล้ายคลึงกันน้อย คำนวณตามสมการที่ 1

$$SIM(U_i, U_j) = \frac{\sum_{i=1}^{n}\left(r_{i,k} - \overline{r_{u_i}}\right) \times \left(r_{u_j,k} - \overline{r_{u_j}}\right)}{\sigma_{u_i} \times \sigma_{u_j}} \quad (1)$$

โดยที่

$sim(u_i,u_j)$ คือค่าความคล้ายระหว่างผู้ใช้u_i และ u_j

$r_{u_i,k}$ คือ ค่าความชอบที่ผู้ใช้ i (u_i) มีต่อสินค้าชิ้นที่ k

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$r_{u_{i,k}}$ คือ ค่าความชอบที่ผู้ใช้ j(u_j) มีต่อสินค้าชิ้นที่ k

$\overline{r_{u_i}}$ คือ ค่าเฉลี่ยของค่าความชอบที่ผู้ใช้u_iมีต่อสินค้าทุกชิ้น

$\overline{r_{u_j}}$ คือค่าเฉลี่ยของค่าความชอบที่ผู้ใช้u_j มีต่อสินค้าทุกชิ้น

σ_{u_i} คือค่าความแปรปรวนของค่าความชอบที่ผู้ใช้u_i

σ_{u_j} คือค่าความแปรปรวนของค่าความชอบที่ผู้ใช้u_j

nคือ จำนวนของ co-rated items

2.2.3 ขั้นตอนการทำนายคะแนนสินค้า

หลังจากที่ค้นหาสมาชิกที่อยู่ในระบบที่มีความชอบคล้ายคลึงกับผู้ใช้เป้าหมายได้แล้ววิธีการต่อไปคือการทำนายค่าความชอบของผู้ใช้เป้าหมายต่อสินค้าโดยพิจารณาจากการให้คะแนนความคล้ายคลึงระหว่างสินค้าชิ้นนั้นกับสินค้าอื่นๆที่เคยให้คะแนนแล้วนำมาคำนวณตามสมการที่ 2

$$P(u_i,k) = \overline{r_{u_i}} + \frac{\sum_{j=1}^{m} W_{u_i,u_j}\left(r_{u_{j,k}} - \overline{r_{u_j}}\right)}{\sum_{j=1}^{m} W_{u_i,u_j}} \quad (2)$$

โดยที่

$P(u_i,k)$คือค่าคะแนนทำนายของผู้ใช้u_iต่อสินค้าชิ้นที่ k

$\overline{r_{u_i}}$คือ ค่าเฉลี่ยของคะแนนสินค้าของผู้ใช้u_i

mคือ จำนวนสมาชิกทั้งหมดที่ให้คะแนนสินค้าชิ้นที่ k

$r_{u_{j,k}}$คือค่าคะแนนสินค้าของผู้ใช้u_jที่ให้กับสินค้าชิ้นที่ k

W_{u_i,u_j}คือค่าความคล้ายคลึงระหว่างผู้ใช้u_iและผู้ใช้u_j

2.2.4 ขั้นตอนการสร้างรายการแนะนำ

เมื่อทำการทำนายคะแนนสินค้าระบบแนะนำจะทำการเรียงลำดับสินค้าที่จะแนะนำจากลำดับคะแนนมากไปน้อย

3.การดำเนินงานวิจัย

จากการวิเคราะห์ชุดข้อมูลที่ใช้ทดลองในงานวิจัยนี้ คือ MovieLens [5] เป็นชุดทดลองที่มีข้อมูลสินค้าเป็นภาพยนตร์ ซึ่งพบว่าข้อมูลคะแนนของสินค้าส่วนใหญ่จะเป็นไปตามค่าฐานนิยมค่าคะแนนของสินค้า และถ้าใช้ค่าฐานนิยมค่าคะแนนของสินค้าช่วยจะทำให้การทำนายคะแนนความชอบสินค้ามีความถูกต้องมากยิ่งขึ้น ดังนั้นการจะทำนายคะแนนสินค้าควรจะให้ความสำคัญถึงความนิยมของสินค้ามากกว่าการใช้ข้อมูลผู้ใช้ร่วมที่ใช้ตัววัดอื่นๆ ซึ่งเป็นข้อมูลที่มากเกินไป ขณะที่ผู้ใช้สินค้าในแต่ละกลุ่มหรือแต่ละผู้ใช้มีระดับการให้คะแนนที่มีความแตกต่างกัน

งานวิจัยนี้เสนอการทำนายคะแนนสินค้าชิ้นที่ k ของผู้ใช้u_iโดยการนำค่าฐานนิยม($mode_k$)ของคะแนนสินค้าที่ผู้ใช้อื่นเคยให้เป็นคะแนน 5 ระดับ ซึ่งค่าคะแนนที่ 1 แสดงว่าไม่ชอบส่วนค่าคะแนนที่ 5 คือชอบมากที่สุดโดยใช้กฎความสัมพันธ์ของค่าฐานนิยมของคะแนนสินค้าในการปรับค่า ซึ่งขั้นตอนวิธีที่นำเสนอมีดังต่อไปนี้

1)ทำค่าเฉลี่ยของผู้ใช้สินค้ามาปัดทศนิยม (หากเกิน 0.5 ปัดขึ้น มิฉะนั้นปัดลง)ให้เป็นค่าจำนวนเต็มของคะแนนสินค้า $\overline{r_{u_i}}$

2)แจกแจงความถี่สัมพัทธ์ของคะแนนสินค้าที่ทำนาย ซึ่งจะได้ค่าฐานนิยมของคะแนนในสินค้าที่ต้องการทำนาย$mode_k$

ถ้าความถี่ของคะแนนสินค้ามีค่าความถี่ที่เท่ากันให้เลือกค่า$mode_k$ที่ใกล้เคียงกับค่าเฉลี่ยของผู้ใช้สินค้า

3) การคำนวณคะแนนทำนายสินค้ามีเงื่อนไขดังนี้

3.1)หาก $\overline{r_{u_i}}$ >$mode_k$คำนวณ คะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่ลงค่าทศนิยม) ลบด้วยความถี่สัมพัทธ์ของคะแนนค่าฐานนิยม

หาก $\overline{r_{u_i}}$ <$mode_k$คำนวณ คะแนนทำนาย มีค่าเท่ากับ $\overline{r_{u_i}}$ (ที่ลงค่าทศนิยม) บวกด้วยความถี่สัมพัทธ์ของคะแนนค่าฐานนิยม

3.2) หาก $\overline{r_{u_i}}$ = $mode_k$ = คะแนนที่ทำนาย

3.1 การทำนายคะแนนความความชอบสินค้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างที่ 1การคำนวณคะแนนความชอบสินค้าโดยการทำนายความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3

ตารางที่ 2:ข้อมูลการให้คะแนนสินค้าของผู้ใช้ สำหรับตัวอย่างที่ 1

ผู้ใช้ \ สินค้า	1	2	3	4
1	2	3	4	2
2	4	3	2	4
3	2	2	?	4
4	4	5	4	2
5	3	2	4	4

ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 8/3 = 2.67 ปัดทศนิยม ≈ 3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน

คะแนน 2 จำนวน 1 คน และ คะแนน 4 จำนวน 3 คน

ค่าฐานนิยมของคะแนน คือ 4 มากกว่าค่าเฉลี่ยคะแนนของผู้ใช้ที่ 3 ซึ่งมีค่าเท่ากับ 3

กำหนดให้

N = จำนวนของผู้ให้คะแนนทั้งหมด

FP = จำนวนความถี่ของคะแนน

Pn =ค่าทำนายคะแนนสินค้า

Avp = ค่าความถี่สัมพัทธ์ = FP/N = 3 / 4 = 0.75

ค่าฐานนิยม มากกว่าค่าเฉลี่ยของผู้ใช้ที่ 3 จึงนำค่าความถี่สัมพัทธ์ 0.75 ไปบวกกับ 2.67

Pn = 2.67 + 0.75 = 3.42 ทศนิยมต่ำกว่า 0.5 จึงปัดลง = 3

ดังนั้นค่าทำนายคะแนนความชอบสินค้า 3 ของผู้ใช้ที่ 3 นี้ คือ 3

ตารางที่ 3:ข้อมูลการให้ความนิยมสินค้าของผู้ใช้ ตัวอย่างที่ 2

ผู้ใช้ \ สินค้า	1	2	3	4
1	2	5	1	2
2	4	3	1	5
3	2	4	?	4
4	4	5	1	2
5	5	2	4	5
6	3	1	4	4
7	1	2	4	4

ตัวอย่างที่ 2 ในกรณีที่มีค่าคะแนนความถี่เท่ากัน จะนำค่าฐานนิยมคะแนนนั้นๆ มาหาค่าใกล้เคียงกับค่าเฉลี่ยของผู้ใช้ที่ 3

จากตารางที่ 3 ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 3.34 ปัดทศนิยม ≈ 3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน คะแนน 1 จำนวน 3 คน และ คะแนน 4 จำนวน 3 คน ดังนั้นค่าฐานนิยมของคะแนนมีค่าเป็น 1 และ 4 จึงทำการหาค่าความใกล้เคียง กับค่าคะแนนเฉลี่ยของผู้ใช้ ซึ่งมีค่าคะแนนเฉลี่ยเท่ากับ 3 โดยการนำค่าคะแนนเฉลี่ยมาลบกับค่าฐานนิยมของคะแนนทั้งคู่ แล้วทำเป็นค่าสัมบูรณ์นำค่าที่น้อยที่สุดเป็นค่าใกล้เคียงกับค่าคะแนนเฉลี่ยของผู้ใช้ เช่น |3 – 4| = 1, |3 – 1|= 2 ผลคือคะแนน 4 มีค่าใกล้กับคะแนนเฉลี่ย มากกว่า (เพราะ 1 มีค่าน้อยกว่า 2) ดังนั้น ค่าฐานนิยมคะแนนจึงเป็น 4เมื่อคำนวณ **Avp** = FP/N = 3 / 6 = 0.50

ค่าฐานนิยมคะแนน มากกว่าค่าคะแนนเฉลี่ยของผู้ใช้ที่ 3 จึงนำค่าความถี่สัมพัทธ์ (**Avp**)0.5 ไปบวกกับ 3.34 (ถ้าเกิดค่าฐานนิยมคะแนน น้อยกว่าค่าคะแนนเฉลี่ยของผู้ใช้ จะนำค่าความถี่สัมพัทธ์ไปลบ)**Pn** = 3.34 + 0.50 = 3.84 ปัดทศนิยม≈4

ดังนั้นค่าคะแนนทำนายความชอบสินค้าชิ้นที่ 3 ของผู้ใช้คนที่ 3 นี้ คือ 4

ตัวอย่างที่ 3 ในกรณีที่ค่าฐานนิยมคะแนน มีค่าเท่ากับคะแนนเฉลี่ยของผู้ใช้ ให้นำค่าฐานนิยมคะแนน ไปทำนายคะแนนสินค้า

ตารางที่ 4:ข้อมูลการให้คะแนนสินค้าของผู้ใช้ สำหรับตัวอย่างที่ 3

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

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3	2	2	?	4
4	4	5	3	2
5	3	2	4	4

จากตารางที่ 4ค่าเฉลี่ยของ ผู้ใช้คนที่ 3 = 2.67 ปัดทศนิยม ≈3

ซึ่งจะดูจากค่าความถี่ของคะแนนสินค้าที่ 3 มี ผู้ให้คะแนน2 จำนวน 1 คน คะแนน 3 จำนวน 2 คน และคะแนน 4 จำนวน 1 คน ค่าฐานนิยมคะแนนคือ 3 ซึ่งมีค่าเท่ากับค่าเฉลี่ยของผู้ใช้ที่ 3 ที่ปัดทศนิยมแล้ว จึงใช้ค่าฐานนิยมคะแนนเป็นค่าทำนายคะแนนสินค้าชิ้นที่ 3

4. ผลการดำเนินงานวิจัย

4.1 รายละเอียดของผลการทดลอง

ชุดข้อมูลที่ใช้ทดลองในงานวิจัยนี้ คือ MovieLens ซึ่งเป็นชุดข้อมูลที่เก็บข้อมูลสินค้าที่เป็นภาพยนตร์จากเว็บไซต์ http://www.movielens.org เป็นเว็บไซต์แนะนำภาพยนต์ ชุดข้อมูลจะประกอบด้วยจำนวนชุดของการดำเนินการ(Transaction) 100,000 ชุด ซึ่งการให้คะแนนภาพยนต์มีเกณฑ์ตั้งแต่ 1 – 5มีผู้ใช้จำนวน 943 คน มีข้อมูลภาพยนตร์ 1,682 เรื่อง โดยข้อมูลนี้จะถูกแบ่งข้อมูลออกเป็น 5 ชุดแต่ละชุดข้อมูลจะแบ่งออกเป็นชุดฝึกสอน (training set) 80% และชุดทดสอบ (test set) 20% โดยการวัดประสิทธิภาพจะใช้ค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ตามสมการที่ 3 ดังนี้

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

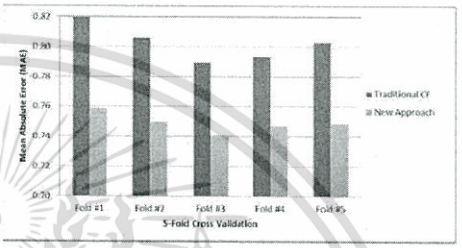
โดยที่ P_iคือค่าคะแนนความชอบที่ทำนาย

Q_iคือค่าคะแนนความชอบที่ถูกต้อง

nคือจำนวนคะแนนสินค้าที่ทำนายทั้งหมด

4.2 ผลการทดลองประสิทธิภาพในการทำนาย

ผลการทดลองเปรียบเทียบประสิทธิภาพการทำนายในแต่ละชุดข้อมูล ระหว่างวิธีการกรองร่วมด้วยการใช้ค่าฐานนิยมตามด้วยกฎความสัมพันธ์ที่นำเสนอ กับเทคนิคการกรองร่วมแบบดั้งเดิม (Traditional Collaborative Filtering : CF) แสดงดังภาพที่ 2 จากกราฟ หากแท่งกราฟที่มีค่า MAE น้อย จะ เป็นวิธีที่มีประสิทธิภาพที่ดีกว่า ซึ่งจะเห็นได้ว่าประสิทธิภาพการทำนายของวิธีที่นำเสนอดีกว่าวิธีการกรองร่วมแบบดั้งเดิมทั้ง 5 ชุดข้อมูลที่นำมาทดสอบ



ภาพที่ 2:กราฟเปรียบเทียบค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) แต่ละชุดข้อมูลทดลองระหว่างวิธีการใหม่กับวิธีการกรองข้อมูลผู้ใช้ร่วม หรือ วิธี CF

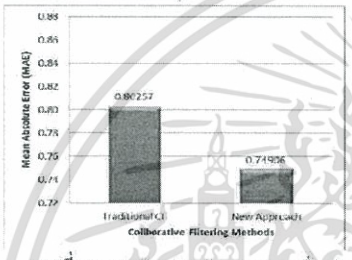
ตารางที่ 5:ผลการทดลองข้อมูลตัวเลขเปรียบเทียบค่าความคลาดเคลื่อนสัมบูรณ์ ทั้ง 5 ชุดข้อมูลระหว่างวิธีการใหม่กับวิธีการกรองข้อมูลผู้ใช้ร่วม

วิธีการ	ชุดที่1	ชุดที่ 2	ชุดที่3	ชุดที่4	ชุดที่5
CF	0.8205	0.806	0.78965	0.7936	0.8031
(3)New Approach	0.75895	0.750	0.74095	0.747	0.7486

ผลการทดลองเปรียบเทียบประสิทธิภาพการทำนายเฉลี่ยรวมทุกชุดข้อมูล ระหว่างวิธีที่นำเสนอกับเทคนิคการกรองร่วมแบบดั้งเดิม เทียบกับการทำนายความนิยมโดยใช้กฎความสัมพันธ์และค่าฐานนิยมโดยผลที่ได้ทำการคำนวณแสดงดังภาพที่ 2 จะเห็นได้ว่าการใช้กฎความสัมพันธ์และค่าฐานนิยมคะแนนในการทำนายความนิยมของผู้ใช้จะให้ค่าเฉลี่ยความคลาดเคลื่อน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สัมบูรณ์ที่น้อยกว่าวิธีอื่น เนื่องจากกฎความสัมพันธ์และค่าฐานนิยมนั้นเป็นการทำนายที่มีการปรับค่าเฉลี่ยของผู้ใช้ตามค่าฐานนิยมของสินค้าซึ่งจะอยู่ในค่าระดับการให้คะแนนของผู้ใช้นั้นๆด้วย คือช่วงระหว่างการให้คะแนนของผู้ใช้แต่ละคนนั้นจะไม่เท่ากัน ดังนั้นการทำนายคะแนนความชอบของสินค้าจะเริ่มต้นค่าจากค่าคะแนนเฉลี่ยของผู้ใช้แล้วปรับค่าตามค่าคะแนนที่เป็นฐานนิยมสินค้าซึ่งจะทำให้พบว่าให้ประสิทธิภาพในการทำนายที่ดีขึ้นและลดเวลาขั้นตอนการคำนวณหาความคล้ายคลึงในแต่ละผู้ใช้



ภาพที่ 3: กราฟเปรียบเทียบค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ของทุกชุดข้อมูลทดสอบ

4.3 วิเคราะห์ time complexity วิธีที่นำเสนอ กับ วิธีของ CF

กำหนดให้

n แทน จำนวนข้อมูลสินค้าทั้งหมด

m แทน จำนวนผู้ใช้ทั้งหมด

วิธีที่นำเสนอจะพิจารณาเฉพาะข้อมูลสินค้าที่สนใจทำนายโดยดูเฉพาะข้อมูลความถี่ของคะแนนที่ผู้ใช้เคยให้ไว้เท่านั้น ดังนั้น time complexityของวิธีที่นำเสนอมีค่าเท่ากับ O(m)

ส่วนวิธี CF เมื่อดูจากสมการที่ 1(ในรายละเอียดที่หัวข้อ 2.2.2) และสมการที่ 2 (ในรายละเอียดที่หัวข้อ 2.2.3)time complexity ของวิธี CF มีค่าเท่ากับ $O(m^2+m)$ หรือ $O(m^2)$ ซึ่งจะเห็นว่าวิธีของ CF ใช้เวลาการคำนวณมากกว่าวิธีที่นำเสนอ

5. สรุปผลและข้อเสนอแนะ

การทำนายคะแนนความชอบสินค้าในระบบแนะนำด้วยการกรองข้อมูลผู้ใช้ร่วมซึ่งงานวิจัยนี้ได้เสนอวิธีใช้กฎความสัมพันธ์และค่าความถี่สัมพัทธ์ของฐานนิยมคะแนน ในการทำนายคะแนนความชอบของผู้ใช้ที่มีต่อสินค้า โดยการนำค่าเฉลี่ยของผู้ใช้มาปรับค่าคะแนนด้วยกฎความสัมพันธ์และข้อมูลฐานนิยมของผู้ใช้ร่วม ผลลัพธ์ที่ได้พบว่ามีประสิทธิภาพในการทำนายคะแนนดีกว่าวิธีการกรองข้อมูลแบบผู้ใช้ร่วมแบบดั้งเดิมและแก้ไขปัญหาการเริ่มต้นทำงานช้าของระบบได้ ซึ่ง time complexity ของวิธีที่นำเสนอคือ O(m) ขณะที่วิธีดั้งเดิม CF มี time complexity เท่ากับ $O(m^2)$ แต่วิธีนี้ยังมีจุดอ่อนอยู่ที่ความเบาบางข้อมูล ถ้าเกิดข้อมูลสินค้าไม่เคยมีผู้ใช้ให้คะแนนมาก่อนจะทำให้ประสิทธิภาพในการทำนายลดลง แต่วิธีการนี้ง่ายต่อการเข้าใจและลดเวลาในการคำนวณความคล้ายคลึงของผู้ใช้ร่วม จากผลการทดลองจะเห็นค่าผิดพลาดในการทำนายคะแนนสินค้าน้อยลง

ข้อเสนอแนะการทำงานวิจัยในอนาคตโดยการประยุกต์ใช้เทคนิควิธีการกรองแบบผสมผสานโดยการคิดค่าเฉลี่ยของผู้ใช้ให้ดูจากเนื้อหาของสินค้า เพื่อให้การเริ่มต้นค่าทำนายคะแนนมีความถูกต้องยิ่งขึ้น หรือใช้เทคนิคแนวคิดนี้ร่วมกับการเติมค่าข้อมูลที่ขาดหายเพื่อแก้ไขปัญหาความเบาบางของข้อมูล ซึ่งคิดว่าน่าจะทำให้ระบบมีประสิทธิภาพในการแนะนำมากยิ่งขึ้น

เอกสารอ้างอิง

[1] Paul Resnick and Hal R. "Recommender System", *COMMUNICATIONS OF THE ACM*, vol. 40, Issue 3, March 1997, 1997

[2] Robin van Meteren and Meterenvan Someren ,"Using Content-Based Filtering for Recommendation" ,*ECML 2000 Workshop*, 2000

[3]BenSchafer,DanFrankowski,JonHerlocker,andDhilaSen, "Collaborative Filtering Recommender System",P.Brusilovsky,A.Kobsa,andW.Nejdl(Eds.):The Adaptive *Web,LNCS 4321*,Pages:291-324,2007

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

[5] GroupLens Research, "MovieLens Project Web site", http://movielens.umn.edu.,1997

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ-สกุล	นาย พีระศักดิ์ แซ่ตั้น
วัน เดือน ปี	21 ตุลาคม 2528
ที่อยู่	308 ภนน ประชาอุทิศ ตำบล ปากแพรก อ.ทุ่งสง จ.นครศรีธรรมราช
E-mail :	diewd2a@gmail.com
ประวัติ	2550 จบการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต
การศึกษา	สาขาวิชาวิศวกรรมมัลติมีเดียและระบบอินเทอร์เน็ต วิทยาเขตรังสิต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้