การวิเคราะห์เปรียบเทียบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม เพื่อเพิ่มประสิทธิภาพในระบบแนะนำร้านอาหาร

A COMPARATIVE ANALYSIS OF COLLABORATIVE FILTERING TECHNIQUES FOR PERFORMANCE OPTIMIZATION IN RESTAURANT RECOMMENDATION SYSTEMS

ภัทรพล เดโช

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

การวิเคราะห์เปรียบเทียบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม เพื่อเพิ่มประสิทธิภาพในระบบแนะนำร้านอาหาร

ภัทรพล เดโช

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

A COMPARATIVE ANALYSIS OF COLLABORATIVE FILTERING TECHNIQUES FOR PERFORMANCE OPTIMIZATION IN RESTAURANT RECOMMENDATION SYSTEMS

PATTARAPHON DACHO

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

สารนิพนธ์

เรื่อง

การวิเคราะห์เปรียบเทียบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม เพื่อเพิ่มประสิทธิภาพในระบบแนะนำร้านอาหาร

ของ

ภัทรพล เดโช

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญากุล)
คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	การวิเคราะห์เปรียบเทียบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมเพื่อเพิ่มประสิทธิภาพในระบบแนะนำร้านอาหาร
ผู้วิจัย	ภัทรพล เดโช
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา

ปัจจุบันระบบแนะนำถูกนำมาใช้ในธุรกิจและอุตสาหกรรมต่างๆ อย่างแพร่หลาย เพื่อคาดการณ์หรือสามารถแนะนำสิ่งที่ผู้ใช้งานหรือลูกค้าชื่นชอบได้อย่างถูกต้องและใกล้เคียงกับความชื่นชอบของผู้ใช้งานได้มากที่สุด โดยงานวิจัยฉบับนี้เป็นการวิจัยเชิงวิเคราะห์เปรียบเทียบวิธีการแนะนำในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม และดำเนินการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยวิธี Grid Search Optimization เพื่อให้แบบจำลองสามารถทำงานได้อย่างมีประสิทธิภาพมากที่สุดเมื่อนำมาใช้กับชุดข้อมูลในระบบแนะนำร้านอาหาร โดยมีวัตถุประสงค์เพื่อแสดงให้เห็นข้อดี ข้อเสีย และโอกาสในการพัฒนาแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมให้มีประสิทธิภาพที่มากยิ่งขึ้น งานวิจัยนี้ใช้ข้อมูลความคิดเห็นและการให้คะแนนของผู้บริโภคจาก Yelp โดยนำเสนอการเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆ ทั้งก่อนและหลังการปรับค่าพารามิเตอร์ที่เหมาะสม รวมถึงการเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง ผ่านการใช้ค่า Mean Absolute Error (MAE) และ Root Mean Square Error (RMSE) เพื่อประเมินประสิทธิภาพในด้านของความแม่นยำของแบบจำลอง และใช้ค่า Fit Time และ Test Time ในการเปรียบเทียบความเร็วในการประมวลผลของแบบจำลอง รวมถึงมีการเปรียบเทียบด้วยค่า Precision@k และ Recall@k เพื่อประเมินความถูกต้องแม่นยำของแบบจำลองในการคาดการณ์การแนะนำร้านอาหาร ซึ่งผลลัพธ์จากการเปรียบเทียบแบบจำลองแสดงให้เห็นความสำคัญในการพิจารณาการสูญเสียประสิทธิภาพบางอย่างของแบบจำลองเพื่อให้สอดคล้องจุดประสงค์ของงานในบริบทที่แตกต่างกัน

คำสำคัญ: ระบบแนะนำ, การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม, การปรับค่าพารามิเตอร์ที่เหมาะสม

Title	A COMPARATIVE ANALYSIS OF COLLABORATIVE FILTERING TECHNIQUES FOR PERFORMANCE OPTIMIZATION IN RESTAURANT RECOMMENDATION SYSTEMS
Author	PATTARAPHON DACHO
Degree	MASTER OF SCIENCE
Academic Year	2024
Advisor	Assistant Professor Dr. Nuwee Wiwatwattana, Ph.D.

Recommendation systems have been widely adopted across diverse sectors to predict and suggest items that match users' preferences. The objective of the study is to investigate and compare the performance of collaborative filtering techniques, emphasizing model refinement through hyperparameter tuning via Grid Search Optimization to enhance the accuracy and reliability of restaurant recommendations. The experiments utilize consumer reviews and ratings data from Yelp. This study presents a comprehensive comparative analysis of the performance of various recommendation models, both before and after hyperparameter optimization, as well as a comparison between models using key evaluation metrics, including Mean Absolute Error and Root Mean Square Error for model accuracy, while computational efficiency is measured through Fit Time and Test Time. Furthermore, Precision@k and Recall@k metrics are employed to evaluate the effectiveness of each model in generating restaurant recommendations. The findings from the comparison emphasize the importance of acknowledging and managing inevitable trade-offs in performance to ensure that the chosen model remains consistent with the intended objectives and the contextual demands of different application scenarios.

Keywords: Recommendation System, Collaborative Filtering, Hyperparameter Optimization

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้ได้รับความอนุเคราะห์จาก ผศ.ดร.นุวีย์ วิวัฒนวัฒนา อาจารย์ที่ปรึกษาที่ได้ให้คำปรึกษา คำแนะนำ และข้อคิดเห็นต่างๆ อันเป็นประโยชน์อย่างยิ่งในการทำสารนิพนธ์ อีกทั้งขอขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ผศ.ดร.อัครินทร์ ไพบูลย์พานิช และ ผศ.ดร.ศิริสรรพ เหล่าหะเกียรติ ที่ได้เสียสละเวลาอันมีค่าในการให้คำแนะนำ แนวทาง และข้อเสนอแนะสำหรับการปรับปรุงสารนิพนธ์ รวมถึงคณะอาจารย์ทุกท่านที่ได้ให้คำแนะนำและความรู้ในวิชาต่างๆ ตลอดการศึกษา ผู้วิจัยขอกราบขอบพระคุณด้วยความเคารพอย่างสูง

ขอขอบคุณคุณพ่อ คุณแม่และครอบครัวที่คอยให้การสนับสนุนมาโดยตลอด ขอบคุณเพื่อนๆ ที่เคยช่วยเหลือไม่ว่าทางตรงหรือทางอ้อม ตลอดจนคำแนะนำที่เป็นประโยชน์ทั้งในการเรียนและการทำสารนิพนธ์ฉบับนี้

ผู้วิจัยหวังเป็นอย่างยิ่งว่าสารนิพนธ์ฉบับนี้จะมีประโยชน์ไม่มากก็น้อย สำหรับข้อบกพร่องต่างๆ ที่อาจเกิดในในสารนิพนธ์ผู้วิจัยขอน้อมรับผิดเพียงผู้เดียว และยินดีรับฟังความคิดเห็นและข้อเสนอนแนะจากผู้ที่สนใจที่เข้ามาศึกษาในสารนิพนธ์ฉบับนี้เพื่อเป็นประโยชน์ในการพัฒนางานวิจัยต่อไป

ภัทรพล เดโช

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและความเป็นมาของงานวิจัย

ปัจจุบันธุรกิจหลายประเภทไม่ว่าจะเป็นธุรกิจออฟไลน์ หรือออนไลน์อาศัยคำแนะนำในการสร้างความน่าเชื่อถือและการรับรู้ของธุรกิจให้เป็นที่รู้จักและมีชื่อเสียง โดยเฉพาะอย่างยิ่งธุรกิจที่ได้รับอิทธิพลอย่างมากจากคำแนะนำ ยกตัวอย่างเช่น ธุรกิจอีคอมเมิร์ซ หรือธุรกิจร้านอาหาร ไม่ว่าจะมาจากการให้คะแนน และการแสดงความคิดเห็นจากผู้บริโภคหรือผู้ใช้บริการซึ่งเป็นปัจจัยสำคัญที่จะช่วยให้ธุรกิจนำไปสู่การพัฒนาและปรับปรุงสินค้าและบริการให้ดียิ่งขึ้น ในส่วนของผู้บริโภคหรือผู้ใช้บริการสามารถใช้ประโยชน์จากการแสดงความคิดเห็นหรือคำแนะนำจากผู้บริโภคหรือผู้ใช้บริการที่เคยบริโภคและใช้บริการเหล่านี้เข้ามาเป็นตัวเลือกในการตัดสินใจในการเลือกบริโภคและใช้บริการธุรกิจต่างๆได้เช่นเดียวกัน

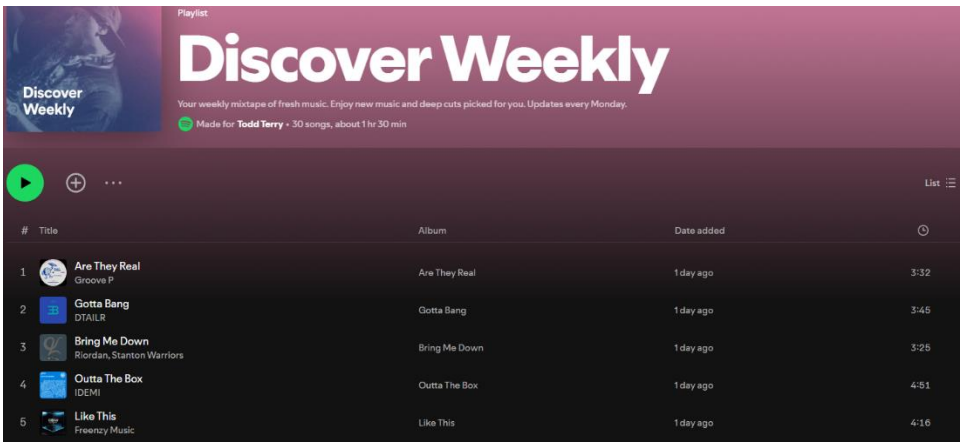
สำหรับธุรกิจร้านอาหารในประเทศไทยมีจำนวนธุรกิจร้านอาหารและเครื่องดื่มของไทยที่มีการจดทะเบียนทั้งหมด 382,038 ร้านมีแนวโน้วเพิ่มขึ้นอย่างต่อเนื่องในปี 2567-2569 (ศุภกร, 2024) ในขณะที่ร้านอาหารหลากหลายสัญชาติและหลายประเภทเกิดขึ้นจำนวนมาก ตัวเลือกของผู้บริโภคมีทิศทางที่เพิ่มขึ้นเช่นเดียวกัน แต่อย่างไรก็ตามปัจจัยหนึ่งที่เป็นส่วนสำคัญในการเลือกรับประทานอาหาร คือการแสดงความคิดเห็นและการให้คะแนนทั้งจากผู้ที่เคยบริโภคหรือใช้บริการทั้งบนอินเตอร์เน็ต เว็ปไซต์ และเครือข่ายสังคมต่างๆ โดยในส่วนนี้จะเข้ามาช่วยในการตัดสินใจของผู้บริโภคมากยิ่งขึ้น ยกตัวอย่างเช่น Yelp หนึ่งในแพลตฟอร์มที่รวบรวมความคิดเห็นและการให้คะแนนของผู้บริโภคที่มีชื่อเสียงมากที่สุดในประเทศสหรัฐอเมริกา โดยผู้บริโภคหรือผู้ที่สนใจสามารถค้นหาร้านอาหาร ข้อมูล การให้คะแนนและการแสดงความคิดเห็นของผู้บริโภคผู้อื่นที่เคยใช้บริการหรือบริโภคร้านอาหารนั้นได้ เพื่อนำข้อมูลมาประกอบการตัดสินใจในการรับบริการหรือรับประทานร้านอาหารนั้นๆ สำหรับในประเทศไทยมีแอปพลิเคชันในลักษณะดังกล่าวข้างต้นเช่นเดียวกันมีชื่อว่า Wongnai ซึ่งเป็นแหล่งรวบรวมร้านอาหาร ข้อมูล การให้คะแนนและการแสดงความคิดเห็นของผู้ใช้งานที่มีโอกาสใช้บริการหรือรับประทานร้านอาหารต่างๆในประเทศไทย

ข้อมูลการแสดงความคิดเห็นและการให้คะแนนของผู้บริโภคเป็นสิ่งที่สำคัญปัจจัยหนึ่งต่อธุรกิจร้านอาหาร ไม่ว่าจะเป็นความคิดเห็นในเชิงชื่นชม หรือคำวิจารณ์ในมุมต่างๆของธุรกิจร้านอาหาร ไม่ว่าจะเป็น รสชาติ การบริการ หรือแม้กระทั่งบรรยากาศภายในร้านอาหาร โดยแสดงผ่านทั้งการให้คะแนนที่ชัดเจน (Explicit Rating) หมายถึงการแสดงความคิดเห็นในรูปแบบของตัวเลขหรือรูปแบบต่างๆที่อยู่ในลักษณะจำนวนที่สามารถระบุค่าได้ ซึ่งตัวเลขที่มีค่ามากแสดงถึง

แนวโน้มระดับความชอบหรือพึงพอใจอยู่ในระดับที่สูงสูง หากตัวเลขมีค่าน้อยสามารถแสดงถึงแนวโน้มในการแสดงความคิดเห็นต่อสิ่งนั้นอยู่ในระดับที่ไม่พึงพอใจหรือไม่ชอบ ตัวอย่างเช่น การให้คะแนนร้านอาหาร 4 จาก 5 ดาว แสดงให้เห็นว่าผู้ใช้งานมีความชื่นชอบหรือพึงพอใจในร้านอาหารในระดับที่ค่อนข้างสูง รวมถึงการให้คะแนนแบบโดยนัย (Implicit Rating) ซึ่งหมายถึงสิ่งที่แสดงถึงความชื่นชอบหรือความพึงพอใจของผู้ใช้งานที่ไม่ได้แสดงออกมาอย่างตรงไปตรงมา โดยจะคาดการณ์ผ่านปัจจัยอื่น อย่างพฤติกรรมของผู้ใช้งาน ตัวอย่างเช่น การเข้าชมผลิตภัณฑ์หรือการเพิ่มผลิตภัณฑ์ที่ผู้ใช้งานสนใจลงในรายการชื่นชอบ (Favorite) เป็นต้น

ระบบแนะนำ (Recommendation System) เป็นระบบในการแนะนำสิ่งที่ตรงกับความต้องการหรือมีความเกี่ยวข้องกับความสนใจของผู้ใช้งาน เพื่อเข้ามาช่วยในการตัดสินใจให้กับผู้ใช้งาน ไม่ว่าจะเป็นการแนะนำสิ่งของ หรือสถานที่ โดยในช่วงหลายปีที่ผ่านมาระบบแนะนำได้เข้ามามีบทบาทอย่างมากต่อธุรกิจต่างๆ ถูกนำมาใช้ด้วยเทคนิคที่หลากหลาย ไม่ว่าจะเป็น เทคนิคการกรองข้อมูลอิงจากเนื้อหา (Content-based Filtering) ที่เป็นเทคนิคในการแนะนำผู้ใช้งานโดยอ้างอิงจากเนื้อหา ข้อมูลหรือคุณลักษณะของสิ่งที่ผู้ใช้งานเคยสนใจ ในส่วนของอีกเทคนิคที่มีความนิยมสูงอย่าง เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เป็นเทคนิคในการแนะนำที่คาดการณ์จากข้อมูลผู้ใช้งานผู้อื่นที่มีรสนิยม ความชื่นชอบ หรือความสนใจที่คล้ายกัน รวมถึงเคยมีการให้คะแนน (Rating) กับสิ่งที่ผู้ใช้งานเคยใช้งาน บริโภค หรือใช้บริการและทำการแนะนำให้กับผู้ใช้งานปัจจุบัน (Active User)

ธุรกิจสตรีมมิงวิดีโอ (Video Streaming) อย่างบริษัท Netflix หรือบริษัทที่ให้บริการสตรีมมิงเพลง (Music Streaming) อย่าง Spotify ล้วนใช้ระบบแนะนำเพื่อให้สามารถแนะนำสิ่งที่สอดคล้องกับความสนใจหรือรสนิยมของผู้ใช้งานมากที่สุดเช่นเดียวกัน โดย Spotify จะมีการใช้ระบบแนะนำ จากภาพประกอบที่ 1 Discover Weekly ซึ่งเป็นฟีเจอร์การแนะนำเพลงที่ผู้ใช้งานไม่เคยรับฟังมาก่อน แต่เป็นเพลงที่มีความใกล้เคียงกับรสนิยม หรือความชื่นชอบของผู้ใช้งานนั้น ซึ่งหลักงานทำงานของฟีเจอร์ดังกล่าวจะเป็นการใช้ข้อมูลจากการฟังเพลงของผู้ใช้งานปัจจุบันในสัปดาห์ที่ผ่านมา แล้วนำเอารูปแบบ (Pattern) การฟังเพลงนั้นไปเปรียบเทียบกับผู้ใช้งานผู้อื่นที่ฟังเพลงในลักษณะคล้ายกัน หรือมีรสนิยมในการฟังเพลงในลักษณะใกล้เคียงกัน หลังจากนั้นระบบจะแนะนำเพลงผู้ใช้งานนั้นๆที่ไม่เคยฟังมาก่อนแต่คาดการณ์ว่าผู้ใช้งานนั้นจะสนใจหรือชื่นชอบ



ภาพประกอบที่ 1 ตัวอย่างฟีเจอร์ Discover Weekly ของแอปพลิเคชันสตรีมมิงเพลง Spotify

ที่มา: https://open.spotify.com/playlist/37i9dQZEVXcQ9COmYvdajy

ดังนั้นงานวิจัยนี้จึงมีแนวคิดที่จะศึกษาและเปรียบเทียบระบบแนะนำร้านอาหารโดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ซึ่งเป็นเทคนิคในการแนะนำโดยใช้ข้อมูลจากผู้ใช้งานรายอื่น โดยใช้ชุดข้อมูลจาก Yelp ที่มีการแสดงความคิดเห็นและข้อมูลร้านอาหารที่หลากหลายประเภทในประเทศสหรัฐอเมริกา รวมถึงพัฒนาระบบแนะนำร้านอาหารให้ได้ผลลัพธ์ที่มีประสิทธิภาพมากยิ่งขึ้น และสามารถแนะนำร้านอาหารได้อย่างถูกต้องแม่นยำ

1.2 วัตถุประสงค์ของงานวิจัย

ศึกษาและเปรียบเทียบประสิทธิภาพการทำงานของระบบแนะนำโดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ในแบบจำลองหลากหลายรูปแบบ ได้แก่ Baseline, Item-based Collaborative Filtering, User-based Collaborative Filtering, Co-Clustering, Singular Value Decomposition (SVD), SVD++, และ Non-negative Matrix Factorization (NMF) และมีการปรับปรุงประสิทธิภาพของแบบจำลองด้วยวิธีหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization) เพื่อเพิ่มประสิทธิภาพการแนะนำร้านอาหารให้มีความถูกต้องและแม่นยำมากยิ่งขึ้นในการศึกษาและวิเคราะห์เชิงเปรียบเทียบ

1.3 ขอบเขตงานวิจัย

งานวิจัยนี้นำข้อมูลจาก Yelp Dataset มาทำการศึกษา และพัฒนาแบบจำลองในการแนะนำร้านอาหาร โดยใช้เทคนิคการรองข้อมูลแบบพึ่งพาร่วม (Collaborative Filtering) ซึ่งมีขอบเขตงานวิจัย ดังนี้

1. ชุดข้อมูลที่ใช้ในงานวิจัย ได้แก่ ชุดข้อมูลการแสดงความคิดเห็นและการให้คะแนนของผู้ใช้งานที่มีต่อร้านอาหารและธุรกิจบริการจากเว็บไซต์และแอปพลิเคชัน Yelp ทั้งหมดกว่า 6,990,280 การแสดงความคิดเห็น 150,346 ร้านอาหารและธุรกิจ และผู้ใช้งานจำนวน 1,987,897 ราย

2. ผู้วิจัยได้ทำการเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับร้านอาหาร ที่อยู่เฉพาะในรัฐฟลอริดา ประกอบด้วย ข้อมูลร้านอาหารจำนวน 4,065 ร้าน และข้อมูลการแสดงความคิดเห็นและการให้คะแนน (Review) จำนวน 162,124 ครั้ง จากผู้ใช้งานจำนวน 88,986 คน

3. พัฒนาแบบจำลอง และเปรียบเทียบระบบแนะนำร้านอาหารโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ด้วยขั้นตอนวิธีหาค่าพารามิเตอร์ที่เหมาะสม (Optimization) และดำเนินการเปรียบเทียบประสิทธิภาพของแบบจำลองในมิติต่างๆ ได้แก่ การวประเมินประสิทธิภาพในด้านของความแม่นยำ (Accuracy Metrics) การประเมินประสิทธิภาพในด้านของการจัดอันดับการแนะนำ (Top-N Metrics) และการประเมินประสิทธิภาพในด้านของประสิทธิภาพ (Performance Metrics)

1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถนำแบบจำลองที่มีประสิทธิภาพ และความแม่นยำสูงเมื่อเทียบกับแบบจำลองอื่น มาประยุกต์ใช้กับชุดข้อมูล หรือระบบแนะนำเทคนิคอื่นๆได้อย่างมีประสิทธิภาพ

2. สามารถนำแบบจำลองไปปรับใช้ เพื่อช่วยเพิ่มประสิทธิภาพการแนะนำร้านอาหารที่หลากหลาย เหมาะสม และสอดคล้องกับความชื่นชอบหรือรสนิยมของผู้ใช้งานที่มีความเปลี่ยนแปลงอย่างต่อเนื่อง

3. เป็นแนวทางในการประเมินข้อดี และข้อเสียในแง่ของประสิทธิภาพในการตัดสินใจนำแบบจำลองไปประยุกต์ใช้ทั้งในด้านการเวลาการประมวลผล ความคลาดเคลื่อนของแบบจำลองและความถูกต้องแม่นยำ

4.เป็นแนวทางในการพัฒนาและต่อยอดแบบจำลอง เพื่อให้ระบบแนะนำในรูปแบบของเทคนิคต่างๆมีประสิทธิภาพที่สูงขึ้น หรือสามารถนำแบบจำลองไปประยุกต์ใช้กับวิธีการอื่นๆเพื่อปรับปรุงผลลัพธ์ระบบแนะนำให้มีความแม่นยำยิ่งขึ้น

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

งานวิจัยครั้งนี้ทางผู้วิจัยได้ทำการศึกษาและนำเสนอรายละเอียดทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการระบบการแนะนำร้านอาหาร (Restaurant Recommendation System) โดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เพื่อนำมาต่อยอดและเป็นแนวทางในการพัฒนางานวิจัยต่อไป โดยประกอบด้วยหัวข้อดังนี้

1. ระบบแนะนำ (Recommendation System)
2. วิธีการประเมินค่าความคล้ายคลึง (Similarity Measures)
3. ทฤษฎีในการสร้างแบบจำลอง
4. การปรับหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization)
5. การประเมินประสิทธิภาพของแบบจำลอง
6. งานวิจัยที่เกี่ยวข้อง

2.1 ระบบแนะนำ (Recommendation System)

ระบบแนะนำ คือ ระบบที่นำเสนอหรือแนะนำผลิตภัณฑ์ หรือการบริการให้กับผู้ใช้งาน ณ ขณะนั้น (Active User) โดยอ้างอิงจากข้อมูลหรือประวัติการใช้งานของผู้ใช้งานในอดีต หรือคาดการณ์จากข้อมูลการใช้งานของผู้ใช้งานรายอื่น ๆ ที่มีคุณลักษณะที่คล้ายคลึงกับผู้ใช้งาน ซึ่งช่วยให้ผู้ใช้งานสามารถตัดสินใจ หรือเพิ่มโอกาสในการเลือกผลิตภัณฑ์ หรือการบริการมากขึ้น รวมทั้งตอบโจทย์ผู้ใช้งานที่สนใจ หรือมีความชื่นชอบได้อย่างถูกต้องและรวดเร็วโดยลดเวลาในการสืบค้นข้อมูลของผู้ใช้งานลง

ระบบแนะนำมีบทบาทที่สำคัญต่อภาคธุรกิจในอุตสาหกรรมต่างๆ ไม่ว่าจะเป็น ธุรกิจแพลตฟอร์มซื้อขายสินค้าออนไลน์ อย่างบริษัท Amazon และ Netflix แพลตฟอร์มสตรีมมิงภาพยนตร์ชื่อดัง หรือแพลตฟอร์มวิดีโอที่ใหญ่ที่สุดในโลกอย่าง YouTube ที่นำระบบแนะนำมาปรับใช้ในการแนะนำวิดีโอให้มีความเหมาะสมเฉพาะกับแต่ละบุคคลและมีความเกี่ยวข้องหรือเชื่อมโยงกับพฤติกรรมการรับชมวิดีโอที่ผ่านมาของผู้ใช้งาน โดยสามารถตอบสนองความต้องการและความสนใจของผู้ใช้งาน รวมถึงสามารถนำเสนอผลิตภัณฑ์ต่างๆที่ตรงกับความชื่นชอบหรือรสนิยมของผู้ใช้งานได้ดียิ่งขึ้น

ปัจจุบันระบบแนะนำ ประกอบไปด้วยหลากหลายเทคนิค โดยในแต่ละเทคนิคจะมีแบบจำลองที่ทำงานเบื้องหลังแตกต่างกันออกไป ขึ้นอยู่กับจุดประสงค์และข้อมูลของผู้ใช้งาน โดย

เทคนิคที่มีความนิยมและถูกนำมาปรับใช้ในปัจจุบันมีด้วยกัน 3 เทคนิคหลัก ได้แก่ การกรองข้อมูลตามเนื้อหา (Content-based Filtering) การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) และการกรองข้อมูลแบบผสม (Hybrid Filtering)

2.1.1 เทคนิคการกรองแบบอิงเนื้อหา (Content-Based Filtering)

เทคนิคในการแนะนำตามคุณสมบัติของเนื้อหาข้อมูลนั้นๆ โดยอ้างอิงจากคุณลักษณะบางประการของสิ่งที่ผู้ใช้งานเคยใช้งานหรือใช้บริการในอดีต กับสิ่งที่ระบบจะทำการแนะนำ ยกตัวอย่างเช่น ผู้ใช้งานมีประวัติความชอบทานอาหารอย่าง ก๋วยเตี๋ยวต้มยำ โดยเป็นอาหารประเภทเส้น ประกอบด้วย เส้น และเนื้อสัตว์ ระบบจะแนะนำเป็นอาหารประเภทเส้นเช่นเดียวกัน อาจจะเป็นก๋วยเตี๋ยวน้ำตก เนื่องจากเป็นอาหารประเภทเดียวกัน มีเส้น และเนื้อสัตว์เป็นส่วนประกอบ

2.1.2 เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

เทคนิคการแนะนำโดยอ้างอิงจากข้อมูลพฤติกรรมของผู้ใช้ หรือผู้ใช้งานรายอื่นๆ ที่มีความคล้ายคลึงกับผู้ใช้งาน ณ ขณะนั้น ซึ่งเทคนิคนี้มีปัจจัยสำคัญในการแนะนำคือ การให้คะแนน (Rating) ที่ได้มาจากผู้ใช้งานรายอื่นที่ทำการให้คะแนนผลิตภัณฑ์ หรือบริการมาก่อนหน้า เพื่อทำการคาดการณ์และแนะนำรายการที่ผู้ใช้อาจสนใจ ซึ่งการให้คะแนนแสดงถึงความพึงพอใจในสิ่งต่างๆของผู้ใช้งาน โดยสามารถแบ่งการให้คะแนนเป็น 2 ประเภท ดังนี้

1. การให้คะแนนแบบชัดเจน (Explicit Rating) หมายถึง การให้คะแนนผลิตภัณฑ์หรือบริการ เพื่อแสดงถึงความชื่นชอบหรือพึงพอใจ โดยสามารถระบุระดับค่าคะแนนผ่านตัวเลข 0-5 คะแนนได้ ซึ่งแสดงให้เห็นความพึงพอใจระดับต่ำที่สุดเท่ากับ 0 และความพึงพอใจสูงที่สุดเท่ากับ 5 ยกตัวอย่างเช่น ผู้ใช้งานมีโอกาสรับประทานอาหารร้าน A โดยมีการให้คะแนนความพึงพอใจต่อร้านอาหาร 5 คะแนน หมายความ ผู้ใช้งานมีความพึงพอใจต่อร้านอาหาร A อยู่ในระดับสูงที่สุด อย่างไรก็ตาม หากผู้ใช้งานให้คะแนนร้าน A เท่ากับ 1 คะแนน สามารถอธิบายได้ว่าผู้ใช้งานอาจไม่พึงพอใจอย่างมากต่อการรับประทานอาหารในร้าน A

2. การให้คะแนนแบบโดยนัย (Implicit Rating) หมายถึง การคาดการณ์การให้คะแนนผลิตภัณฑ์หรือการบริการของผู้ใช้งาน ผ่านปัจจัยต่างๆที่ไม่ใช่ค่าคะแนนในลักษณะตัวเลขรูปแบบเดียวกับการให้คะแนนแบบชัดเจน (Explicit Rating) ซึ่งสามารถอธิบายเป็นนัยได้ว่าผู้ใช้งานพึงพอใจต่อผลิตภัณฑ์หรือการบริการมากน้อยเพียงใด ยกตัวอย่างเช่น พฤติกรรมการเสิร์ชหาผลิตภัณฑ์ใดผลิตภัณฑ์หนึ่งซ้ำๆ หรือมีการเพิ่มผลิตภัณฑ์นั้นลงในรายการที่ชื่นชอบ (Favorite) หรือเพิ่มเข้าไปในตะกร้าสินค้าแต่ยังไม่มีการชำระเงิน เป็นต้น

2.1.3 **เทคนิคการกรองแบบผสม** (Hybrid Filtering)

เทคนิคการแนะนำที่เป็นการนำเทคนิคสองประเภทขึ้นไปมาปรับใช้ให้ทำงานร่วมกัน เพื่อเพิ่มประสิทธิภาพให้ระบบแนะนำดียิ่งขึ้น โดยเป็นการนำข้อดีของแต่ละเทคนิคมาประกอบกัน และลดข้อเสียในแต่ละเทคนิค โดยทั่วไปนิยมนำเทคนิคการกรองแบบอิงเนื้อหา (Content-Based Filtering) ประกอบกับเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เพื่อประยุกต์แบบจำลองให้สามารถทำงานร่วมกัน ยกตัวอย่างเช่น เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม มีปัญหาในกรณีเมื่อมีผู้ใช้งานใหม่เข้ามาในระบบ ซึ่งทำให้มีข้อมูลเกี่ยวกับผู้ใช้งานใหม่ไม่มากพอที่จะสามารถเปรียบเทียบความคล้ายคลึงผู้ใช้งานอื่นกับผู้ใช้งานใหม่ได้ (Cold-Start Problem) การใช้เทคนิคการกรองแบบอิงเนื้อหาจะเข้ามาช่วยแก้ปัญหานี้ ตัวอย่างเช่น การแนะนำโดยการจัดอันดับสิ่งที่มีผู้ใช้งานสนใจหรือชื่นชอบมากที่สุดในช่วงเวลานั้นๆ 10 อันดับแรก

2.2 **วิธีการประเมินค่าความคล้ายคลึง** (Similarity Measures)

2.2.1 **ค่าความคล้ายคลึงแบบโคไซน์** (Cosine Similarity)

การประเมินค่าความคล้ายคลึงแบบโคไซน์ คือการหาค่าความคล้ายคลึงของข้อมูลจากค่าความต่างของมุมระหว่าง 2 เวคเตอร์ รวมถึงเปรียบเทียบความใกล้กันขององศา ซึ่งโคไซน์จะมีค่าอยู่ระหว่าง 0 ถึง 1 กล่าวคือ กรณีค่าโคไซน์มีค่าใกล้หรือเท่ากับ 0 หมายถึง เวคเตอร์ทั้ง 2 มีความคล้ายคลึงกันต่ำหรือไม่มีความคล้ายคลึงกัน ในกรณีที่ค่าโคไซน์มีค่าใกล้หรือเท่ากับ 1 หมายถึง ทั้ง 2 เวคเตอร์มีความคล้ายคลึงกัน โดยมีสมการ ดังนี้

$$cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

โดยที่

A และ B คือ เวคเตอร์ (Vector)

$A \cdot B$ คือ ผลคูณจุด (Dot Product)

$\| A \|$และ $\| B \|$ คือ ขนาดของเวคเตอร์ (Magnitude)

2.2.2 Mean Squared Difference (MSD)

การประเมินค่าความคล้ายคลึงโดยใช้ค่าเฉลี่ยของผลต่างกำลังสอง ใช้ในการเปรียบระหว่างผู้ใช้งาน (User) หรือผลิตภัณฑ์ (Item) โดยอ้างอิงจากความแตกต่าง (Difference)

ระหว่างการให้คะแนน (Rating) ซึ่งวิธีนี้จะประเมิณความไม่คล้ายคลึงกัน (Dissimilarity) โดยคำนวณจากค่าเฉลี่ยของผลต่างกำลังสองระหว่างการให้คะแนน (Rating) สำหรับค่า MSD หากมีค่ายิ่งสูง สามารถอธิบายได้ว่าผู้ใช้งาน (User) หรือผลิตภัณฑ์ (Item) มีความไม่คล้ายคลึงหรือมีความคล้ายคลึงกันต่ำ แต่หากค่า MSD มีค่าที่ต่ำ หมายถึงผู้ใช้งานหรือผลิตภัณฑ์มีความคล้ายคลึงกันสูง มีสมการ ดังนี้

$$MSD(A,B) = \frac{1}{N}\sum_{i=1}^{N}(A_i - B_i)^2$$

โดยที่

A_i และ B_i คือ การให้คะแนน (Rating) ของผู้ใช้งาน (User) A และ B สำหรับผลิตภัณฑ์ i

N คือ จำนวนของผลิตภัณฑ์ที่ถูกให้คะแนนจากผู้ใช้งาน

สำหรับการประมเนิค่าความคล้ายคลึงทั้งสองวิธีดังกล่าวข้างต้น สามารถนำมาประยุกต์ใช้ในระบบแนะนำแบบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ในอัลกอริทึมที่มีการคำนวณหาความคล้ายคลึงของการให้คะแนน (Rating) ระหว่างผู้ใช้งาน หรือผลิตภัณฑ์ หากปรับใช้วิธีต่างๆข้างต้นให้เหมาะสมกับชุดข้อมูลจะช่วยให้อัลกอริทึมสามารถประเมินค่าความคล้ายคลึงได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

2.3 ทฤษฎีในการสร้างแบบจำลอง

2.3.1 Baseline Model

แบบจำลองพื้นฐานคำนวณค่าเฉลี่ยของการให้คะแนนทั้งหมด (Global Average Rating) และค่าอคติ (Bias) ที่คำนวณจากค่าความเบี่ยงเบนของการให้คะแนนของผู้ใช้งาน และผลิตภัณฑ์จากค่าเฉลี่ยของการให้คะแนนทั้งหมด โดย Baseline จะมุ่งเน้นกับการอ้างอิงกับความเป็นจริงที่ผู้ใช้งานบางรายมีแนวโน้มที่จะให้คะแนนในระดับที่สูงหรือต่ำกว่าผู้ใช้งานรายอื่นมากกว่าที่ควรจะเป็น หรือผลิตภัณฑ์บางอย่างที่เป็นกระแส หรือเป็นที่นิยมย่อมมีแนวโน้มที่จะได้รับการให้คะแนนมากกว่าผลิตภัณฑ์อื่นๆทั่วไป การคำนวณค่าอคติ จึงเข้ามาช่วยสร้างความสมมุตให้กับแบบจำบนบนพื้นฐานความเป็นจริง ซึ่งจุดประสงค์ของแบบจำลองนี้เพื่อเป็นพื้นฐานในการเปรียบเทียบประสิทธิภาพกับแบบจำลองอื่นๆ โดยมีสมการ ดังนี้

$$\hat{r}_{ui} = \mu + b_u + b_i$$

โดยที่

$\hat{r}_{ui}$ คือ การคาดการณ์การให้คะแนน (Rating) โดยผู้ใช้งาน u สำหรับผลิตภัณฑ์ i

μ คือ ค่าเฉลี่ยของการให้คะแนนทั้งหมดของผู้ใช้งานและผลิตภัณฑ์

b_u คือ ค่าเบี่ยงเบนการให้คะแนนของผู้ใช้งาน u จากค่าเฉลี่ยทั้งหมด (Global Average)

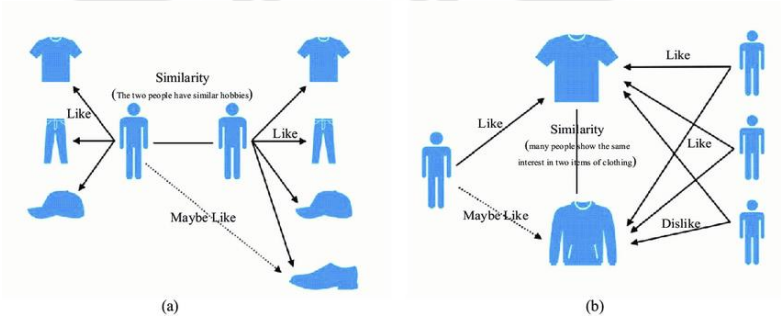
b_i คือ ค่าเบี่ยงเบนการให้คะแนนของผู้ใช้งาน i จากค่าเฉลี่ยทั้งหมด (Global Average)

2.3.2 เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

งานวิจัยนี้มุ่งเน้นการพัฒนาระบบแนะนำด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยอัลกริทึมในเทคนิคนี้สามารถแบ่งออกเป็น 2 วิธีหลัก คือ วิธีแบบอ้างอิงจากความจำ (Memory-based Method) และวิธีแบบอ้างอิงจากแบบจำลอง (Model-based Method)

2.3.2.1 วิธีแบบอ้างอิงจากความจำ (Memory-based Method)

วิธีแบบอ้างอิงจากความจำ เป็นวิธีการพิจารณาหาความคล้ายคลึง (Similarity) ระหว่างผู้ใช้งานและผลิตภัณฑ์ในการคาดการณ์การให้คะแนน (Rating) และแนะนำแก่ผู้ใช้งาน ตามภาพประกอบที่ 2 โดยสามารถแบ่งได้เป็น 2 ประเภท ดังนี้



ภาพประกอบที่ 2 วิธีในการประเมินค่าความคล้ายคลึงกันระหว่างผู้ใช้งาน และระหว่างผลิตภัณฑ์

ที่มา: (Ni, Jianjun & Cai, Yu & Tang, Guangyi & Xie, Yingjuan, 2021, p.3)

2.3.2.1.1 User-based Collaborative Filtering

เป็นวิธีในการประเมินค่าความคล้ายคลึงกันระหว่างผู้ใช้งาน ณ ขณะนั้น (Active User) กับผู้ใช้งานที่มีรสนิยมหรือความชื่นชอบ (User Preference) ที่คล้ายคลึงกัน และทำการแนะนำผลิตภัณฑ์ที่ผู้ใช้งานทั้งสองรายมีความชื่นชอบที่คล้ายกัน ยกตัวอย่างเช่น ผู้ใช้งาน A และผู้ใช้งาน B มีการให้คะแนนผลิตภัณฑ์หลายๆอย่างในลักษณะคล้ายกัน วิธีดังกล่าวจะทำการแนะนำผลิตภัณฑ์ที่ผู้ใช้งาน B ให้คะแนนสูงหรือมีความชื่นชอบอย่างมากให้กับผู้ใช้งาน A

2.3.2.1.2. Item-based Collaborative Filtering

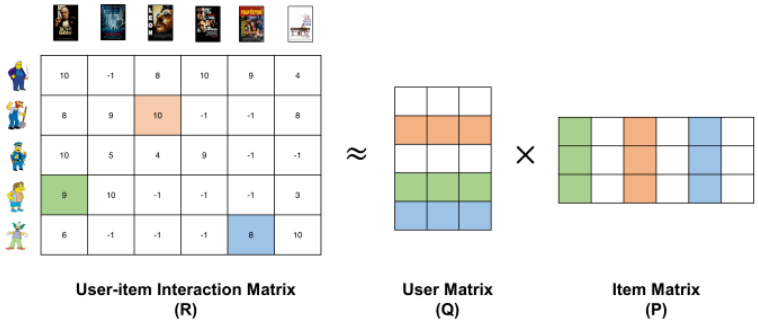
คือวิธีการพิจารณาความคล้ายคลึงระหว่างผลิตภัณฑ์ ยกตัวอย่างเช่น กรณีผลิตภัณฑ์ A และผลิตภัณฑ์ B ถูกให้คะแนนโดยผู้ใช้งานหลายรายในระดับคะแนนที่สูง ดังนั้นทั้งสองผลิตภัณฑ์นี้มีความคล้ายคลึงกัน หากผู้ใช้งานชื่นชอบผลิตภัณฑ์ B วิธีนี้จะทำการแนะนำผลิตภัณฑ์ A แก่ผู้ใช้งาน

2.3.2.2 **วิธีแบบอ้างอิงจากแบบจำลอง** (Model-based Method)

วิธีแบบอ้างอิงจากแบบจำลอง เป็นวิธีในการสร้างแบบจำลองด้วยวิธีการเรียนรู้ด้วยเครื่อง (Machine Learning) เพื่อให้สามารถคาดการณ์รสนิยมหรือความชื่นชอบของผู้ใช้งานโดยอ้างอิงจากรูปแบบของข้อมูล (Pattern) ที่เกิดขึ้นจากการฝึกฝนข้อมูลในชุดข้อมูลผู้ใช้งานและผลิตภัณฑ์ อย่างการให้คะแนน (Rating) ของผู้ใช้งานในผลิตภัณฑ์ต่างๆ โดยงานวิจัยนี้นำอัลกอริทึมที่มีความนิยม 2 รูปแบบมาประยุกต์ใช้ในงาน ดังนี้

2.3.2.2.1 **การแยกตัวประกอบเมทริกซ์** (Matrix Factorization)

การแยกตัวประกอบของเมทริกซ์ เป็นวิธีในการแยกให้เมทริกซ์ที่มีขนาดใหญ่มีขนาดที่เล็กลงจากเมทริกซ์เดิม โดยที่ผลคูณของเมทริกซ์ที่ถูกแยกออกมาจะเท่ากับเมทริกซ์ตั้งต้น (โดยประมาณ) สำหรับระบบแนะนำโดยปกติจะนำเมทริกซ์ของการให้คะแนนของผู้ใช้งานและผลิตภัณฑ์ (User-Item Interaction Matrix) แยกตัวประกอบออกมาเป็น 2 ส่วน คือ เมทริกซ์ผู้ใช้งาน (User Matrix) และเมทริกซ์ผลิตภัณฑ์ (Item Matrix) ตามภาพประกอบที่ 3 โดยวิธีนี้จะช่วยให้สามารถจำแนกปัจจัยซ่อนเร้น (Latent Factor) ที่มีความสัมพันธ์ระหว่างผู้ใช้งานและผลิตภัณฑ์ที่ไม่สามารถพบได้ในเมทริกซ์ตั้งต้น รวมถึงยังสามารถช่วยลดปัญหาข้อมูลเบาบาง หรือมีค่าเป็นศูนย์จำนวนมาก (Data Sparsity) ซึ่งวิธีดังกล่าว สามารถทำได้หลายวิธี ดังนี้

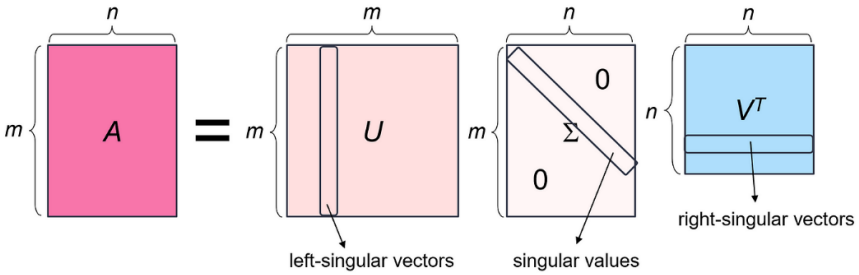


ภาพประกอบที่ 3 วิธีการแยกตัวประกอบของเมทริกซ์

ที่มา: https://www.linkedin.com/pulse/fundamental-matrix-factorization-recommender-system-saurav-kumar

2.3.2.2.1.1 Singular Value Decomposition (SVD)

Singular Value Decomposition หรือ SVD เป็นวิธีในการแยกตัวประกอบที่มีความนิยมและมีประสิทธิภาพสูงวิธีหนึ่ง โดยการทำงานของวิธีนี้จะทำการแยกตัวประกอบของเมทริกซ์ของการให้คะแนนของผู้ใช้งานและผลิตภัณฑ์ (User-Item Interaction Matrix) ออกมาเป็น 3 เมทริกซ์ย่อย ตามภาพประกอบที่ 4



ภาพประกอบที่ 4 การแยกตัวประกอบด้วยวิธี Singular Value Decomposition

ที่มา: https://towardsdatascience.com/singular-value-decomposition-svd-demystified-57fc44b802a0

เพื่อทำการลดของของมิติของข้อมูล (Dimensionality Reduction) หาปัจจัยซ่อนเร้น (Latent Factor) และทำการเก็บเฉพาะคุณสมบัติที่มีความสำคัญและเชื่อมโยงกับรสนิยมหรือความชื่นชอบของผู้ใช้งาน (User Preference) และคุณลักษณะของผลิตภัณฑ์ (Item Attribute) ซึ่งมีสมการ ดังนี้

$$A = U\Sigma V^t$$

โดยที่

A คือ เมทริกซ์ตั้งต้นขนาด m x n

U คือ เมทริกซ์ผู้ใช้งาน (User) ขนาด m x k

Σ คือ เมทริกซ์ทแยงมุมขนาด k x k

V คือ เมทริกซ์ผลิตภัณฑ์ (Item) ขนาด k x n

k คือ จำนวนปัจจัยซ่อนเร้น (Latent factor)

2.3.2.2.1.2 Singular Value Decomposition (SVD++)

เป็นวิธีที่ถูกพัฒนาจาก Singular Value Decomposition (SVD) โดยมีการเพิ่มข้อมูลในส่วนของการให้คะแนนแบบโดยนัย (Implicit Rating) ตัวอย่างเช่น จำนวนผู้ชมหรือการจำนวนการคลิกเพื่อชมผลิตภัณฑ์ :ซึ่งมีจุดประสงค์ในการเพิ่มประสิทธิภาพให้กับการแนะนำ ในกรณีที่การให้คะแนนแบบชัดเจนมีจำนวนน้อย แต่การมีปฏิสัมพันธ์ระหว่างผู้ใช้งานกับผลิตภัณฑ์กลับมีจำนวนมาก โดยมีสมการ ดังนี้

$$\hat{r}_{ui} = \mu + b_u + b_i + Q_i \cdot \left(P_u + \frac{1}{\sqrt{N(u)}} \sum_{j \in N(u)} y_i\right)$$

โดยที่

μ คือ ค่าเฉลี่ยของการให้คะแนนทั้งหมดของผู้ใช้งานและผลิตภัณฑ์

b_u และ b_i คือ ค่าเบี่ยงเบนการให้คะแนนของผู้ใช้งาน u และผลิตภัณฑ์ i จากค่าเฉลี่ยทั้งหมด (Global Average)

Q_i คือ เวกเตอร์คุณลักษณะซ่อนเร้น (Latent feature vector) ของผลิตภัณฑ์ i

P_u คือ เวกเตอร์คุณลักษณะซ่อนเร้น (Latent feature vector) ของผู้ใช้งาน u ซึ่งเรียนรู้มาจากการให้คะแนนแบบชัดเจน (Explicit Rating)

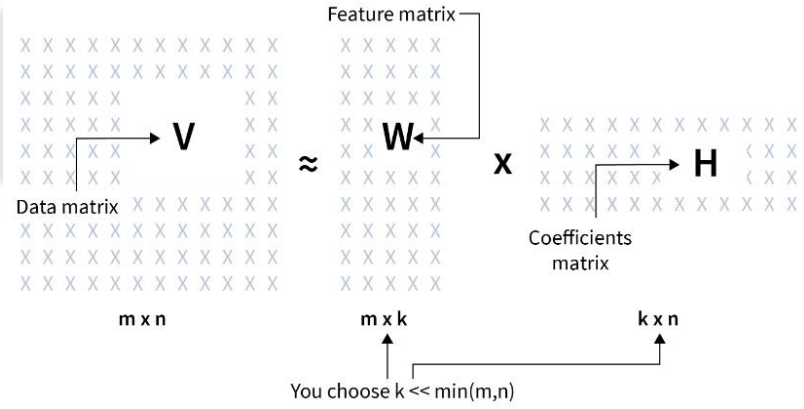
$N(u)$ คือ เซ็ตของผลิตภัณฑ์ที่คาดการณ์การให้คะแนนแบบโดยนัย (Implicit Rating) จากผู้ใช้งาน

y_i คือ เวกเตอร์คุณลักษณะซ่อนเร้น (Latent feature vector) ของผลิตภัณฑ์ j คาดการณ์การให้คะแนนแบบโดยนัยจากผู้ใช้งาน

$\frac{1}{\sqrt{N(u)}}$ คือ คะแนนแบบโดยนัยที่ผ่านการลดความซับซ้อนของข้อมูล (Normalization)

2.3.2.2.1.3 Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization เป็นวิธีการแยกตัวประกอบเมทริกซ์วิธีหนึ่งที่เข้ามาแก้ปัญหาของวิธีที่มีความนิยมสูงอย่าง Singular Value Decomposition (SVD) ในกรณีที่ผลลัพธ์ที่ออกมามีค่าติดลบ ซึ่งวิธีนี้มีความสามารถในการลดขนาดของข้อมูล และแยกเมทริกซ์ให้อยู่ในรูปของเมทริกซ์ย่อย 2 เมทริกซ์ ตามภาพประกอบ 5



ภาพประกอบที่ 5 การแยกตัวประกอบด้วยวิธี Non-negative Matrix Factorization

ที่มา: https://www.scaler.com/topics/nlp/non-negative-matrix-factorization

ซึ่งผลลัพธ์ที่ออกมาจะมีค่าเป็นบวกเท่านั้น แต่อย่างไรก็ตามค่าที่ออกมานั้นจะเป็นค่าประมาณการ ส่งผลให้การแยกตัวประกอบเมทริกซ์วิธีนี้ไม่สามารถทำให้กลับไปอยู่ในรูปของผลคูณของเมทริกซตั้งต้นได้ รวมถึงการทำงานของวิธีนี้ในแต่ละครั้ง อาจจะแสดงผลลัพธ์ที่ไม่เท่ากันได้ (Non-deterministic) โดยวิธีดังกล่าวมีสมการ ดังนี้

$$V \approx W \cdot H$$

โดยที่

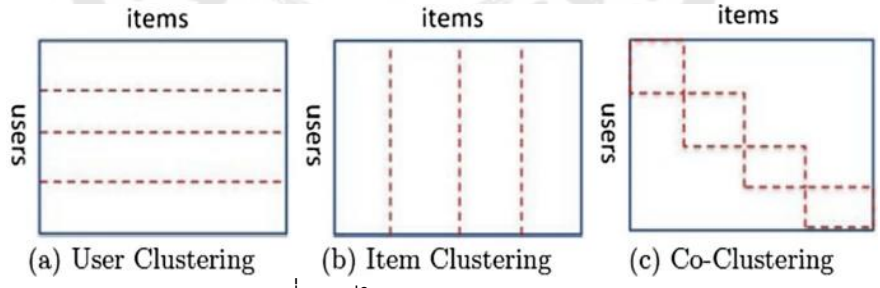
V คือ เมทริกซ์ของการให้คะแนนของผู้ใช้งานและผลิตภัณฑ์ ขนาด m x n

W คือ เมทริกซ์ซ่อนเร้นของผู้ใช้งานขนาด m x k

H คือ เมทริกซ์ซ่อนเร้นของผลิตภัณฑ์ขนาด k x n

2.3.2.2.2 Co-Clustering

เป็นวิธีในการจัดกลุ่มความสัมพันธ์ระหว่างผู้ใช้งาน และผลิตภัณฑ์ โดยผ่านเมทริกซ์ขนาด 2 มิติเพื่อพิจารณารูปแบบ (Pattern) ร่วมกันระหว่างผู้ใช้งานและผลิตภัณฑ์ กล่าวคือ ผู้ใช้งานและผลิตภัณฑ์จะถูกจับกลุ่ม (Clustering) ให้อยู่ในกลุ่มเดียวกันหากมีการปฏิสัมพันธ์ระหว่างกัน ยกตัวอย่างเช่น ผู้ใช้งาน A, B และผลิตภัณฑ์ X, Y อยู่ในกลุ่มร่วม (Co-Cluster) เดียวกัน เนื่องจากผู้ใช้งาน A และ B ให้คะแนน (Rating) ผลิตภัณฑ์ X และ Y ในระดับที่สูงเช่นเดียวกัน โดยอัลกอริทึมจะทำการหาค่าเฉลี่ยการให้คะแนนในแต่ละกลุ่มร่วม (Co-Cluster) เพื่อใช้ในการคาดการณ์การให้คะแนนของผู้ใช้งาน ตามภาพประกอบที่ 6



ภาพประกอบที่ 6 วิธีในการจัดกลุ่มแบบ Co-Clustering

ที่มา: (AbbasiRad, E., Keyvanpour, M.R. & Tohidi, 2024, p.6)

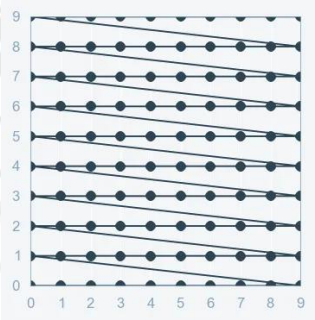
2.4 การปรับหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization)

ในการพัฒนาแบบจำลอง พารามิเตอร์เป็นหนึ่งสิ่งที่สำคัญในการช่วยให้แบบจำลองมีประสิทธิภาพ ซึ่งมีหน้าที่ในการกำกับอัลกอรทึมในแบบจำลองให้เป็นการดำงานที่แตกต่างกันตามการตั้งค่าพารามิเตอร์ โดยการดำเนินการที่จะทำให้แบบจำลองมีประสิทธิภาพสูงที่สุดเพื่อให้ตอบโจทย์ในการประยุกต์ใช้แบบจำลองคือ การหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter

Optimization) สามารถดำเนินการโดยการพิจารณาหาค่าพารามิเตอร์ให้เหมาะสมกับแบบจำลอง ซึ่งจะส่งผลให้แบบจำลองมีความแม่นยำที่มากขึ้น และค่าความคลาดเคลื่อนที่ต่ำที่สุด วิธีการประกอบด้วย 3 วิธีหลัก ดังนี้

1. Manual Search คือ วิธีในการหาค่าพารามิเตอร์ด้วยตัวเองโดยผ่านการทดลองการผสมผสานพารามิเตอร์ต่างๆที่กำหนดขึ้นเอง และทำการฝึกฝนแบบจำลองเพื่อค้นหาค่าพารามิเตอร์ที่ส่งผลให้แบบจำลองมีความแม่นยำสูงและความคลาดเคลื่อนน้อยที่สุด และพิจารณาเลือกจากประสิทธิภาพของแบบจำลองที่ดีที่สุดจากการปรับใช้ค่าพารามิเตอร์นั้นเพื่อนำมาประยุกต์ใช้

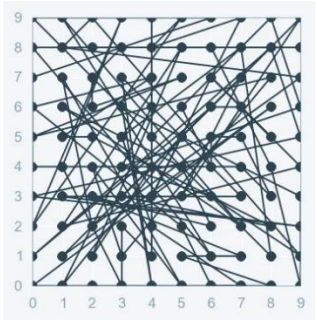
2. Grid Search คือ วิธีการค้นหาพารามิเตอร์แบบกริด (Grid) เป็นหนึ่งในวิธีที่มีความนิยมสูง เนื่องจากเป็นวิธีที่ไม่ซับซ้อน โดยเป็นการกำหนดช่วงของค่าพารามิเตอร์ทุกค่าที่เป็นไปได้ทั้งหมด และดำเนินการสร้างแบบจำลองจากชุดของค่าพารามิเตอร์ที่ถูกกำหนดข้างต้นทุกชุด และนำมาประเมินประสิทธิภาพแบบจำลองที่ดีที่สุดในลักษณะของกริด (Grid) ตามภาพประกอบที่ 7



ภาพประกอบที่ 7 ตัวอย่างการทำงานของ Grid Search

ที่มา: https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318

3. Random Search คือ วิธีการหาค่าพารามิเตอร์ที่มีลักษณะคล้ายกับวิธีแบบ Grid Search แต่มีความแตกต่างกันตรงวิธีนี้จะเป็นการสุ่มเลือกค่าพารามิเตอร์ แทนการกำหนดช่วงของค่าพารามิเตอร์ ตามภาพประกอบที่ 8 ซึ่งจะส่งผลให้วิธีลักษณะนี้มีข้อดีในเรื่องของเวลาในการประมวลผล แต่อย่างไรก็ตามในด้านของประสิทธิภาพวิธีเช่นนี้จะไม่สามารถการันตีได้ว่าจะสามารถค้นหาแบบจำลองที่มีประสิทธิภาพดีที่สุดเหมือนกับวิธี Grid Search ได้



ภาพประกอบที่ 8 ตัวอย่างการทำงานของ Random Search

ที่มา: https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318

2.5 การประเมินประสิทธิภาพของแบบจำลอง

2.5.1 การประเมินประสิทธิภาพในด้านของความแม่นยำ (Accuracy Metrics)

2.5.1.1 Mean Absolute Error (MAE)

ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error) เป็นวิธีที่คำนวณค่าสัมบูรณ์ (Absolute) ของค่าความแตกต่างระหว่างค่าจริง (Actual) กับค่าคาดการณ์ที่ได้จากแบบจำลอง (Predicted) โดยเฉลี่ย เพื่อแสดงให้เห็นค่าความคลาดเคลื่อน (Error) ของแบบจำลอง สามารถคำนวณได้จากสมการ

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

โดยที่

y_i คือ ค่าจริง (Actual)

$\hat{y}_i$ คือ ค่าคาดการณ์ที่ได้จากแบบจำลอง (Predicted)

n คือ จำนวนข้อมูลทั้งหมด

2.5.1.2 Root Mean Squared Error (RMSE)

ค่ารากที่สองของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (Root Mean Squared Error) คือวิธีในการประเมินความคลาดเคลื่อนเพื่อพิจารณาหาความแตกต่างระหว่างค่า

คาดการณ์ที่ได้จากแบบจำลอง (Predicted) และค่าจริง (Actual) โดยเฉลี่ยกำลังสอง และนำมาหาค่ารากที่สอง ตามสมการ ดังนี้

$$RMSE = \sqrt{\sum_{i=2}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

โดยที่

y_i คือ ค่าจริง (Actual)

$\hat{y}_i$ คือ ค่าคาดการณ์ที่ได้จากแบบจำลอง (Predicted)

n คือ จำนวนข้อมูลทั้งหมด

2.5.2 การประเมินประสิทธิภาพในด้านของการจัดอันดับการแนะนำ (Top-N Metrics)

2.5.2.1 Precision@k

Precision@k คือค่าความแม่นยำ หรือการประเมินความแม่นยำของการแนะนำ โดยเป็นการการประเมินสัดส่วนของรายการที่แนะนำซึ่งมีความเกี่ยวข้องจากรายการแนะนำ k อันดับแรก ซึ่งมีวิธีในการประเมินอัตราส่วนของจำนวนผลลัพธ์ หรือผลิตภัณฑ์ที่ถูกแนะนำจากอันดับการแนะนำทั้งหมดที่มีความเกี่ยวข้องกับผู้ใช้งาน (Relevant Items) เทียบกับจำนวนอันดับการแนะนำทั้งหมด โดยมุ่งเน้นไปที่การแนะนำผลลัพธ์ หรือผลิตภัณฑ์ที่มีความเกี่ยวข้องกับผู้ใช้งานมากที่สุด โดย k อันดับการแนะนำทั้งหมดคือ การกำหนดค่าคงที่ค่าหนึ่ง (Threshold) เพื่อกำหนดจำนวนอันดับที่เป็นผลลัพธ์จากการแนะนำ ยกตัวอย่างเช่น k = 5 หมายถึงการกำหนดผลลัพธ์ให้แบบจำลองคาดการณ์การแนะนำผลิตภัณฑ์ที่มีความเกี่ยวข้องกับผู้ใช้งานออกมาเท่ากับ 5 ผลิตภัณฑ์ ซึ่งเมื่อ k มีจำนวนเพิ่มมากขึ้น ค่า Precision@k มักจะลดลง เนื่องจากผลลัพธ์ หรือผลิตภัณฑ์ที่ถูกแนะนำที่ไม่เกี่ยวข้องมีโอกาสปรากฏมากขึ้น โดย Precision@k สามารถคำนวณได้ ดังนี้

$$Precision@k = \frac{Number\ of\ Relevant\ Items\ in\ k}{Total\ Number\ of\ Items\ in\ k}$$

2.5.2.2 Recall@k

Recall@k ค่าความถูกต้อง หรือการประเมินความครอบคลุมของรายการแนะนำ โดยเป็นการประเมินสัดส่วนของรายการที่เกี่ยวข้องทั้งหมดที่ปรากฏในรายการแนะนำ k อันดับแรก ซึ่งมีวิธีในการประเมินอัตราส่วนของจำนวนผลลัพธ์ หรือผลิตภัณฑ์ที่มีความเกี่ยวข้องกับผู้ใช้งาน (Relevant Items) ทั้งหมดที่ปรากฏในอันดับ k ทั้งหมดที่ถูกแนะนำ เทียบกับจำนวนผลลัพธ์ หรือผลิตภัณฑ์ที่มีความเกี่ยวข้องกับผู้ใช้งานทั้งหมด สำหรับ Recall@k จะมุ่งเน้นไปในส่วนของผลลัพธ์ หรือผลิตภัณฑ์ที่มีความเกี่ยวข้องกับผู้ใช้งานให้ครอบคลุมมากที่สุด โดย Recall@k มักจะเพิ่มขึ้นเมื่อ k เพิ่มขึ้น เพราะมีโอกาสครอบคลุมรายการที่เกี่ยวข้องมากขึ้น ซึ่งสามารถคำนวณได้ ดังนี้

$$Recall@k = \frac{Number\ of\ Relevant\ Items\ in\ k}{Total\ Number\ of\ Relevant\ Items}$$

2.5.3 **การประเมินประสิทธิภาพในด้านของเวลาในการประมวลผล** (Processing Time Performance Metric)

2.5.3.1 Fit Time

Fit Time คือ การประเมินประสิทธิภาพด้านเวลาในการประมวลผลในช่วงเวลาฝึกฝนของแบบจำลอง ในช่วงเวลาดังกล่าวจะเป็นช่วงที่อัลกอริทึมเรียนรู้รูปแบบของข้อมูล (Pattern) หรือสร้างแบบจำลอง ยกตัวอย่างเช่น การคำนวณหาค่าความคล้ายคลึง (Similarity), การฝึกฝนปัจจัยซ่อนเร้น (Latent factor) ในการกรองแบบพึ่งพาผู้ใช้ร่วม และการแยกตัวประกอบของเมทริกซ์ต่างๆ ซึ่งหาก Fit Time ใช้เวลานานเกินไปจะส่งผลให้ไม่เหมาะกับการนำแบบจำลองไปปรับใช้กับงานบางประเภท ตัวอย่างเช่น งานที่มีความถี่ในประมวลผลสูง หรืองานที่ใช้ชุดข้อมูลขนาดใหญ่ เป็นต้น

2.5.3.2 Test Time

Test Time คือ การประเมินประสิทธิภาพด้านเวลาที่แบบจำลองประมวลผลเพื่อคาดการณ์ หรือสร้างรายการแนะนำขึ้นมาสำหรับข้อมูลชุดใหม่ซึ่งใช้แบบจำลองเดียวกับเวลาในการฝึกฝน ซึ่งเวลาในช่วงนี้จะใช้เวลาน้อยกว่าเวลาในการฝึกฝน (Fit Time) เนื่องจากเป็นช่วงเวลาที่แบบจำลองทำงานกับข้อมูลขนาดเล็กกว่าชุดข้อมูลฝึกฝน โดยส่วนใหญ่จะเป็นการประมวลผลในส่วนของการคาดการณ์การให้คะแนนของผู้ใช้งานจะที่มีโอกาสจะให้กับผลิตภัณฑ์ใดๆ หรือการ

สร้างรายการอันดับของผลิตภัณฑ์ในการแนะนำ ซึ่งเมื่อยิ่ง Test Time ใช้เวลาน้อยจะยิ่งช่วยให้แบบจำลองสามารถทำงานกับงานที่ต้องการผลลัพธ์ถี่ หรืองานที่ต้องการผลลัพธ์ในทันที

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับระบบแนะนำโดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยงานวิจัยที่เกี่ยวข้องมีรายละเอียด ดังนี้

1. **บทความวิจัยเรื่อง** A Book Recommender System Using Collaborative Filtering Method (Khalifeh & Almousa, 2021)

งานวิจัยนี้นำเสนอระบบแนะนำด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยมุ่งเน้นที่ชุดข้อมูลหนังสือภาษาอาหรับ (Arabic books) ซึ่งมีจุดประสงค์เพื่อยกระดับประสบการณ์การอ่านสำหรับนักอ่านชาวอาหรับด้วยการนำเสนอรายการแนะนำหนังสือที่เหมาะสมกับแต่ละบุคคล ซึ่งงานวิจัยนี้ประเมินและเปรียบเทียบแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมที่แตกต่างกัน

ในส่วนของชุดข้อมูลที่มีการนำมาใช้เป็นชุดข้อมูลการรีวิวหนังสือ (Books Reviews in Arabic Dataset) จากเว็บไซต์ GoodReads.com (2016) ประกอบด้วย ข้อมูลการรีวิวหนังสือ 510,600 รีวิว, 76,530 ผู้ใช้งาน, และหนังสือ 4,993 รายการ

งานวิจัยนี้ศึกษาและเปรียบเทียบแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ผู้วิจัยเลือกใช้แบบจำลองหลักทั้งหมด 3 แบบ ได้แก่ User-Based Collaborative Filtering, Item-Based Collaborative Filtering, และ Matrix Factorization ด้วยแบบจำลองแบบ Singular Value Decomposition (SVD)

ผลลัพธ์ของงานวิจัยในแต่ละแบบจำลอง ผู้วิจัยใช้ค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) ในการประเมินความแม่นยำของแบบจำลอง สำหรับการประเมินประสิทธิภาพของแบบจำลองในมิติของเวลาประมวลผลใช้ค่า Fit Time และ Test Time รวมถึงการประเมินประสิทธิภาพรายการแนะนำผู้วิจัยใช้ค่า Precision@k และ Recall@k โดยสามารถสรุปได้ว่า KNN User-Based และ KNN Item-Based มีประสิทธิภาพในเรื่องของเวลาในการประมวลผลทำได้รวดเร็วกว่าแบบจำลองอื่น สำหรับประเด็นของความแม่นยำของแบบจำลองแสดงให้เห็นว่า Singular Value Decomposition (SVD) มีความแม่นยำที่สูงกว่าแบบจำลองอื่น ประเมินจากค่า MAE และ RMSE ที่ต่ำที่สุด แม้ว่าจะใช้เวลาในการประมวลผลนานกว่าแบบจำลองอื่น โดยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมในแต่ละแบบจำลองมีข้อดีที่

แตกต่างกันในการะนำไปประยุกต์ใช้ ดังเช่นงานวิจัยนี้ที่แสดงให้เห็นถึงการแลกเปลี่ยนระหว่างข้อดีของแบบจำลองที่ไม่สามารถได้มาพร้อมกันได้อย่าง ความรวดเร็วในการประมวลผลและความแม่นยำ

2. **บทความวิจัยเรื่อง** A Movie and Book Recommender System using Surprise Recommendation Kit (Ananth & Raghuveer, 2020)

งานวิจัยนี้ได้ทำการนำเสนอการพัฒนาระบบแนะนำภาพยนตร์และหนังสือ โดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) และเทคนิคการกรองตามเนื้อหา (Content-based Filtering) โดยผู้วิจัยจะมุ่งเน้นไปที่การประยุกต์ใช้แบบจำลองต่างๆในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมบน Surprise Library เพื่อนำไปใช้ในการแนะนำภาพยนตร์และหนังสือรวมถึงมีการประเมินผลแบบจำลองกับชุดข้อมูล 2 ชุด ได้แก่ ชุดข้อมูล MovieLens และ ชุดข้อมูล Books-Crossing ซึ่งผู้วิจัยมีความประสงค์ให้ระบบแนะนำนี้คาดการณ์รายการแนะนำที่เหมาะสมกับแต่ละบุคคล และสุดท้ายผู้วิจัยประเมินผลโดยใช้ตัวชี้ประเมินต่าง ๆ เช่น RMSE , Test Time, และ อัตราความผิดพลาดในการคาดการณ์ที่แย่ที่สุด (Worst Predictions error rate)

สำหรับชุดข้อมูลแรกที่ใช้ในงานวิจัยนี้ ได้แก่ MovieLens ml-20m Dataset ประกอบด้วยการให้คะแนน (ratings) 20 ล้านครั้ง, ภาพยนต์ 27,000 เรื่อง, และ ผู้ใช้งานที่ให้คะแนนภาพยนต์จำนวน 138,000 คน และชุดข้อมูลที่สอง คือ Books-Crossing Dataset ประกอบด้วยรายการหนังสือประมาณ 200,000 รายการ, 200,000 ผู้ใช้งาน, และการให้คะแนน (ratings) ประมาณ 1 ล้านครั้ง

ผู้วิจัยมีการนำแบบจำลองไปใช้ตามชุดข้อมูล (Dataset) ที่แตกต่างกัน ดังนี้ MovieLens ml-20m Dataset ใช้แบบจำลองทั้งหมด 6 แบบจำลอง ได้แก่ BaselineOnly, CoClustering, SlopeOne, KNNWithMeans, SVD++, และ NMF สำหรับชุดข้อมูล Books-Crossing Dataset ใช้แบบจำลองทั้งหมด 5 แบบจำลอง ได้แก่ KNNBaseline, KNNWithZScore, SVD, KNNBasic, และ SVD++ อีกทั้งผู้วิจัยยังมีการใช้การสรุปข้อมูลให้ออกมาอยู่ในรูปแบบของภาพ (Visualization) ด้วย WordCloud, Seaborn, และ Matplotlib

ผลลัพธ์ของงานวิจัย ผู้วิจัยใช้การประเมินผลด้วยค่า Root Mean Squared Error (RMSE) และ Test Time หรือการประเมินประสิทธิภาพด้านเวลาที่แบบจำลองประมวลผลเพื่อคาดการณ์ รวมถึงการประเมินประสิทธิภาพจากการคาดการณ์รายการแนะนำด้วยค่าอัตราความผิดพลาด (Error Rate) ในชุดข้อมูล MovieLens ml-20m Dataset ปรากฏว่าแบบจำลอง BaselineOnly มีค่า RMSE ต่ำที่สุดแสดงให้เห็นความแม่นยำที่มีประสิทธิภาพของแบบจำลอง ในขณะค่า Test Time แบบจำลอง SlopeOne ใช้เวลาน้อยที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ ใน

ส่วนของชุดข้อมูล Books-Crossing Dataset แสดงให้เห็นว่า แบบจำลอง KNNBaseline มีค่า RMSE ต่ำที่สุด และในส่วนของ Test Time แบบจำลอง SVD++ ใช้เวลาน้อยที่สุด

งานวิจัยนี้ยังแสดงให้เห็นถึงความท้าทายในระบบแนะนำ ยกตัวอย่างเช่น ชุดข้อมูลที่มีขนาดใหญ่ และข้อจำกัดด้านข้อมูลทางประชากร (Demographic data) อย่างเพศ อายุ เชื้อชาติ เป็นต้น โดยงานวิจัยนำเสนอข้อแนะนำที่ควรจัดการกับข้อมูลทางประชากรและการทดสอบด้วยชุดข้อมูลขนาดใหญ่ขึ้นเพื่อเพิ่มประสิทธิภาพในการแนะนำของระบบ

3. **บทความวิจัยเรื่อง Collaborative filtering and kNN based recommendation to overcome cold start and sparsity issues: A comparative analysis (Anwar, Uma, Hussain, & Pantula, 2022)**

งานวิจัยนี้นำเสนอการวิเคราะห์เชิงเปรียบเทียบระหว่างระบบแนะนำแบบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) รวมถึงใช้วิธี K-Nearest Neighbors (KNN) เพื่อแก้ไขปัญหาข้อมูลเบาบาง (Data Sparsity) และปัญหาผู้ใช้งานระบบรายใหม่ (Cold-start Problem) ผู้วิจัยได้ประเมินวิธีแบบอ้างอิงจากความจำ (Memory-based Method) หลายวิธี เช่น KNNBasic, KNNBaseline, KNNWithMeans, SVD, SVD++ และวิธีแบบอ้างอิงจากแบบจำลอง (Model-based Method) อย่าง Co-Clustering ผลการวิจัยแสดงให้เห็นว่าวิธี KNNBaseline สามารถลดอัตราความผิดพลาดและเพิ่มความแม่นยำในการแนะนำได้อย่างมีประสิทธิภาพบนชุดข้อมูลที่มีปัญหาข้อมูลเบาบาง (Data Sparsity)

งานวิจัยนี้ใช้ชุดข้อมูล FilmTrust Dataset ประกอบด้วยข้อมูลภาพยนต์ 2,071 เรื่องจากทั้งหมด 1,508 ผู้ใช้งาน รวมถึงการให้คะแนน (Ratings) 35,497 ครั้ง โดยมีช่วงการให้คะแนนอยู่ระหว่าง 0.5 – 4 คะแนน และมีข้อมูลเบาบาง (Data Sparsity) ประมาณ 98.86%

ผู้วิจัยเลือกใช้แบบจำลองในวิธีแบบอ้างอิงจากความจำ (Memory-based Method) ประกอบด้วย KNNBasic, KNNBaseline, KNNWithMeans, SVD, SVD++ และวิธีแบบอ้างอิงจากแบบจำลอง (Model-based Method) ได้แก่ SVD, SVD++, และ Co-Clustering

สำหรับผลลัพธ์ของงานวิจัย ผู้วิจัยมีการประเมินผลด้วยค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) ร่วมกับการใช้ Cross Validation ในการแบ่งข้อมูลออกเป็นหลายส่วนเพื่อทดสอบประสิทธิภาพของแบบจำลอง โดยกำหนดให้มีค่าเท่ากับ 5, 10, และ 15 ตามลำดับ ซึ่งแสดงให้เห็นว่า KNNBaseline มีประสิทธิภาพที่สูงกว่าแบบจำลองประเภทอื่นๆทั้งในแง่ของ RMSE และ MAE ในทุกๆจำนวนของการแบ่งข้อมูล (Cross Validation) สำหรับส่วนของ Co-Clustering มีค่าอัตราความผิดพลาดสูงที่สุดเมื่อเทียบกับแบบจำลองอื่น

แสดงให้เห็นถึงประสิทธิภาพของแบบจำลองที่ไม่สามารถรองรับต่อปัญหาข้อมูลเบาบาง (Data Sparsity) และปัญหาผู้ใช้งานระบบรายใหม่ (Cold-start Problem) ได้

4. **บทความวิจัยเรื่อง** Recommending Restaurants: A Collaborative Filtering Approach (Tripathi & Sharma, 2020)

งานวิจัยนี้นำเสนอระบบแนะนำร้านอาหารโดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) รวมถึงศึกษาแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมหลากหลายรูปแบบ ซึ่งประกอบไปด้วยแบบจำลอง K-Nearest Neighbors (KNN) และแบบจำลอง Multiclass Support Vector Machine (SVM) ซึ่งนำมาปรับใช้กับเทคนิคการกรองแบบพึ่งพาผู้ใช้ 2 แบบจำลอง ได้แก่ User-Based Collaborative Filtering และ Item-Based Collaborative Filtering เพื่อคาดการณ์คะแนนของผู้ใช้สำหรับร้านอาหาร อีกทั้งงานวิจัยนี้ยังได้มีการการเปรียบเทียบประสิทธิภาพของแบบจำลองข้างต้น รวมถึงแสดงให้เห็นถึงประสิทธิผลของ Multiclass Support Vector Machine (SVM) ในการนำมาปรับใช้กับระบบการแนะนำร้านอาหารด้วยวิธีการประเมินประสิทธิภาพแบบจำลอง ร่วมกับสัดส่วนชุดข้อมูลฝึกฝนที่แตกต่างกัน (Training data)

ผู้วิจัยมีการใช้แบบจำลอง KNN โดยอาศัยความคล้ายคลึงของร้านอาหารซึ่งใช้ Cosine Similarity เข้ามาช่วยในการคำนวณเพื่อที่จะแบ่งประเภทร้านอาหาร และใช้แบบจำลอง SVM แบบการจำแนกประเภทหลายคลาส (Multi-Class Classification) เข้ามาใช้ร่วมกับ User-Based และ Item-Based Collaborative Filtering ในการคาดการณ์การให้คะแนนโดยใช้วิธีแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งงานวิจัยนี้มีการใช้การแปลงคลาสแบบ 2 ประเภทด้วยกัน ได้แก่ One-versus-All คือ การแบ่งปัญหาออกเป็น 1 ต่อทั้งหมด และ One-versus-One คือ การแบ่งปัญหาออกเป็นคู่

สำหรับผลลัพธ์ของงานวิจัย ผู้วิจัยใช้ค่า Mean Squared Error (MSE) ในการประเมินประสิทธิภาพของแบบจำลองโดยแบ่งเป็น 2 ส่วน ได้แก่ จำนวนเพื่อนบ้าน (Neighbors for KNN) โดยมีการกำหนดค่า k ให้มีค่าระหว่าง 5 – 50 รวมถึงมีการประเมินแบบจำลองด้วยสัดส่วนชุดข้อมูลที่ต่างกัน โดยผู้วิจัยแบ่งจำนวนของชุดข้อมูลฝึกฝน (Training data) ออกเป็นตั้งแต่ 60% - 90% ซึ่งMulticlass Support Vector Machine (SVM) ที่ทำการประยุกต์ให้ทำงานร่วมกับ User-Based และ Item-Based Collaborative Filtering แสดงให้เห็นประสิทธิภาพของแบบจำลองว่า User-based SVM มีประสิทธิภาพดีกว่า Item-based SVM เมื่อทดลองบนชุดข้อมูลฝึกฝนที่แตกต่างกันค่า MSE ในขณะที่ User-based SVM มีค่า MSE ที่น้อยที่สุด ซึ่งหมายถึงความแม่นยำ

และประสิทธิภาพของแบบจำลองดีที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ รวมถึงแบบจำลอง K-Nearest Neighbors

5. **บทความวิจัยเรื่อง** GSO-CRS: grid search optimization for collaborative recommendation system (Behera & Nain, 2022)

งานวิจัยนี้ได้ทำการศึกษาและนำเสนอวิธีหาค่าพารามิเตอร์ที่เหมาะสม (Optimization) ระบบแนะนำแบบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ด้วยวิธี Grid Search Optimization (GSO) เพื่อเพิ่มประสิทธิภาพของแบบจำลองที่ใช้วิธีการแยกตัวประกอบเมทริกซ์ (Matrix Factorization) ผู้วิจัยมุ่งเน้นในส่วนของเพิ่มความแม่นยำของระแบบแนะนำ สามารถจัดการกับปัญหาข้อมูลเบาบาง (Data Sparsity) และปรับค่า Hyperparameter เพื่อเพิ่มประสิทธิภาพแบบจำลองให้ได้มากที่สุด งานวิจัยนี้ยังนำเสนอแนวทางใหม่ในการปรับค่า Hyperparameter ตัวอย่างเช่น อัตราการเรียนรู้ (Learning Rate) การกำกับ (Regularization) ปัจจัยแฝง (Latent Factors) และรอบการเรียนรู้ (Epochs) โดยแสดงให้เห็นว่า Grid Search Optimization (GSO) มีประสิทธิภาพที่ดีกว่าเมื่อเทียบกับเทคนิคอื่นๆล่าสุด (State-Of-The-Art)

ชุดข้อมูลในงานวิจัยนี้มีทั้งหมด 2 ชุดข้อมูล ได้แก่ MovieLens 100K Dataset ประกอบด้วยข้อมูล 100,000 การให้คะแนน (Ratings) จาก 943 ผู้ใช้งาน (Users) จากภาพยนต์ทั้งหมด 1,682 เรื่อง และ MovieLens 1M Dataset ประกอบด้วยข้อมูล 1 ล้านการให้คะแนน (Ratings) จาก 6,040 ผู้ใช้งาน (Users) จากภาพยนต์ทั้งหมด 3,952 เรื่อง

ผู้วิจัยนำแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมใช้วิธีการแยกตัวประกอบเมทริกซ์ (Matrix Factorization) อย่าง Singular Value Decomposition (SVD) และ Probabilistic Matrix Factorization (PMF) โดยทำการหาค่าพารามิเตอร์ที่เหมาะสม (Optimization) ด้วยวิธี Stochastic Gradient Descent (SGD), Alternating Least Squares (ALS), Grid Search Optimization, และ Random Search Optimization เพื่อลดค่าความคลาดเคลื่อน (Error) และแยกเมทริกซ์การมีปฏิสัมพันธ์ระหว่างผู้ใช้กับผลิตภัณฑ์ให้เป็นปัจจัยแฝง (Latent Factors)

ผู้วิจัยได้ออกแบบพื้นที่การค้นหาแบบตารางขนาด 4x4 เพื่อปรับค่า Hyperparameter ทั้งหมด 4 ได้แก่ จำนวนปัจจัยแฝง (nfactors) จำนวนรอบการเรียนรู้ (nepochs) อัตราการเรียนรู้ (lr) และค่าการกำกับ (kr)

สำหรับการประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยมีการประเมินและเปรียบเทียบกับทั้ง 2 ชุดข้อมูล (Dataset) ได้แก่ MovieLens 100K และ 1M Dataset แบ่งเป็น 2

วิธีการประเมินหลัก ได้แก่ การประเมินประสิทธิภาพในส่วนของความแม่นยำ (Accuracy) ประกอบด้วย ค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) และการประเมินในส่วนของการคาดการณ์รายการแนะนำ (Top-N) ประกอบด้วย การประเมินความแม่นยำเฉลี่ยของ k รายการ (MAP@k) และการประเมินคุณภาพของผลลัพธ์จากรายการแนะนำโดยอ้างอิงความเกี่ยวข้องกับผู้ใช้งาน (NDCG@k) รวมถึงค่า Precision@k และ Recall@k ในการประเมินความแม่นยำและความครอบคลุมของการคาดการณ์รายการแนะนำ

ผลลัพธ์ของงานวิจัย แบบจำลอง SVD และ PMF ที่ผ่านวิธีการปรับหาค่าพารามิเตอร์ที่เหมาะสมที่สุดด้วยวิธี Grid Search Optimization มีประสิทธิภาพสูงที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ โดยเฉพาะอย่างยิ่งในประเด็นของความแม่นยำ และมีประสิทธิภาพที่เพิ่มขึ้นเมื่อเทียบกับแบบจำลอง SVD และ PMF แบบดั้งเดิมในส่วนของค่า Precision@k และ NDCG@k

งานวิจัยดังกล่าวแสดงให้เห็นว่าการปรับหาค่าพารามิเตอร์ที่เหมาะสมด้วย Grid Search Optimization ในการหาค่า Hyperparameter ในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ช่วยลดค่าความคลาดเคลื่อน และช่วยเพิ่มความแม่นยำในการคาดการณ์รายการแนะนำของแบบจำลองได้มีประสิทธิภาพมากขึ้น

6. **บทความวิจัยเรื่อง** Restaurant Recommender System Using UserBased Collaborative Filtering Approach: A Case Study at Bandung Raya Region (Fakhri, Baizal, & Setiawan, 2019)

งานวิจัยนี้นำเสนอระบบแนะนำร้านอาหารโดยใช้เทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมที่อิงผู้ใช้เป็นหลัก (User-based Collaborative Filtering) ในเขตบันดุงรายา ประเทศอินเดีย ซึ่งระบบนี้แนะนำร้านอาหารโดยอาศัยคะแนนจากผู้ใช้งาน ซึ่งมีการนำค่าความคล้ายคลึง (Similarity) เข้ามาปรับใช้ในระบบแนะนำด้วยการถ่วงน้ำหนักความคล้ายคลึง (Weighted Similarity) ด้วยค่าสัมประสิทธิ์ (Coefficient) ระหว่าง 0-1 ได้แก่ ค่าความคล้ายคลึงของการให้คะแนนของผู้ใช้งาน (User Rating Similarity) และความคล้ายคลึงของคุณลักษณะของผู้ใช้งาน (User Attribute Similarity) โดยมีจุดมุ่งหมายเพื่อช่วยให้ผู้ใช้งานค้นพบร้านอาหารที่เหมาะกับรสนิยมและความชื่นชอบของแต่ละบุคคลมากยิ่งขึ้น อีกทั้งงานวิจัยนี้ยังมีการเปรียบเทียบประสิทธิภาพของเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วมโดยไม่อิงกับคุณลักษณะของผู้ใช้งาน และเทคนิคการกรองแบบพึ่งพาผู้ใช้แบบอิงคุณลักษณะของผู้ใช้งาน ได้แก่ อายุ และเพศ

สำหรับชุดข้อมูลงานวิจัยนี้ใช้ Zomato Restaurant Dataset: ประกอบด้วยข้อมูลร้านอาหารจำนวน 86 ร้านในเขตบันดุงรายา โดยมีการให้คะแนน (Rating) ทั้งหมด 50,998 ครั้ง มี

และผู้ใช้งาน 593 คน ที่มีคุณลักษณะของผู้ใช้งาน โดยอายุของผู้งานแบ่งเป็น 3 กลุ่ม ได้แก่ กลุ่มอายุต่ำกว่า 20 ปี, อายุระหว่าง 20–40 ปี, และอายุมากกว่า 40 ปี รวมถึงประกอบไปด้วยทั้งเพศหญิงและเพศชาย

ผู้วิจัยได้คำนวณหาค่าความคล้ายคลึงระหว่างผู้ใช้งานในแบบจำลองแบบ User-Based Collaborative Filtering โดยใช้การคำนวณความคล้ายคลึงด้วยวิธี Pearson Correlation ที่อ้างอิงจากการให้คะแนนร้านอาหารของผู้ใช้งานแต่ละบุคคน นอกจากนี้ ยังมีการเพิ่มปัจจัยจากคุณลักษณะของผู้ใช้งาน (User Attributes) เช่น เพศและอายุ เข้ามาในการคำนวณเพื่อเพิ่มความแม่นยำ โดยถ่วงน้ำหนัก (Weight) ระหว่างค่าความคล้ายคลึงของการให้คะแนน (Rating Similarity) และค่าความคล้ายคลึงด้านคุณลักษณะผู้ใช้งาน (User Attribute Similarity) เพื่อให้ได้ค่าความคล้ายคลึงสุดท้ายที่จะนำมาปรับใช้กับแบบจำลองเพื่อการคาดการณ์รายการการแนะนำ ตัวอย่างเช่น หากผู้ใช้งานมีเพศเดียวกัน การถ่วงน้ำหนักค่าความคล้ายคลึงจะเท่ากับ 1 แต่หากเพศต่างกันจะมีค่าถ่วงน้ำหนักเป็น 0 ในกระบวนการถ่วงน้ำหนักนี้ ผู้วิจัยได้ใช้ค่าสัมประสิทธิ์ที่มีค่าอยู่ในช่วง 0 ถึง 1 เพื่อควบคุมสัดส่วนความสำคัญของแต่ละปัจจัยในการคำนวณค่าความคล้ายคลึง

ผลลัพธ์ของงานวิจัย การคำนวณความคล้ายคลึงระหว่างผู้ใช้งานโดยโดยไม่ใช้คุณลักษณะของผู้ใช้งาน (Without User Attributes) สามารถให้ผลการแนะนำที่มีความแม่นยำสูงกว่าการคำนวณความคล้ายคลึงโดยใช้คุณลักษณะของผู้ใช้งาน (With User Attributes) ได้แก่ เพศและอายุ ดังนั้นการเพิ่มคุณลักษณะของผู้ใช้งานเข้ามาไม่ได้ช่วยเพิ่มความแม่นยำได้ของแบบจำลองมากนัก เนื่องจากปัจจัยด้านอายุและเพศอาจไม่ได้มีอิทธิพลต่อความชื่นชอบของร้านอาหารมากเท่ากับการให้คะแนนจริงสำหรับการแนะนำร้านอาหารในบริบทของการวิจัยนี้

ตารางที่ 1 สรุปงานวิจัยที่เกี่ยวข้องในงานวิจัย

ปี	ผู้วิจัย	งานวิจัย	เทคนิคของแบบจำลอง	ผลลัพธ์
2021	Khalifeh, S., Almousa, A.	A Book Recommender System Using Collaborative Filtering Method	User-Based และ Item-Based Collaborative Filtering, และ Matrix Factorization (SVD)	SVD มีประสิทธิภาพและแม่นยำที่สุด มีค่า MAE = 0.7907, RMSE = 0.9928
2020	Ananth, G. S., Raghuveer	A Movie and Book Recommender System using Surprise Recommendation Kit	k-NN inspired algorithms on Surprise library, SVD, SVDpp, BaselineOnly, CoClustering, และ SlopeOne	BaselineOnly มีค่า RMSE ต่ำที่สุดอยู่ที่ 3.38 และ SlopeOne ใช้เวลาทดสอบ น้อยที่สุด 0.29 วินาที
2022	Anwar, T., Uma, V., Hussain, M. I., Pantula, M.	Collaborative filtering and kNN based recommendation to overcome cold start and sparsity issues: A comparative analysis	Memory-based และ Model-based Collaborative Filtering	KNNBaseline ที่ CV = 5: มีค่า RMSE = 0.8222, และค่า MAE = 0.6246.ต่ำที่สุด
2020	Tripathi, A., Sharma, A. K.	Recommending Restaurants: A Collaborative Filtering Approach	Multiclass SVM ใช้ร่วมกับ User-Based และ Item-Based Collaborative Filtering	User-based SVM มีประสิทธิภาพดีที่สุด โดยมีค่า MSE = 0.76
2022	Behera, G., Nain, N.	GSO-CRS: grid search optimization for collaborative recommendation system	Matrix Factorization อย่าง SVD และ PMF และใช้ Grid Search Optimization ร่วมในแบบจำลองในลักษณะ GSVD และ GPMF	GSVD ให้ผลลัพธ์ดีที่สุด โดยมีค่า RMSE = 0.8525 และ MAE = 0.6711
2019	Fakhri, A. A., Baizal, Z. K. A., Setiawan, E. B.	Restaurant Recommender System Using UserBased Collaborative Filtering Approach: A Case Study at Bandung Raya Region	User-Based Collaborative Filtering ด้วยเทคนิค Pearson Correlation ร่วมกับ User Attributes	User Attributes เข้ามาถ่วงน้ำหนักค่าความคล้ายไม่สามารถช่วยเพิ่มประสิทธิภาพของแบบจำลองได้

บทที่ 3
วิธีดำเนินการวิจัย

ผู้วิจัยได้ศึกษา และนำเสนอระบบการแนะนำร้านอาหารด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยใช้ชุดข้อมูล Yelp Open Dataset ซึ่งดำเนินการวิจัยตามขั้นตอนดังต่อไปนี้

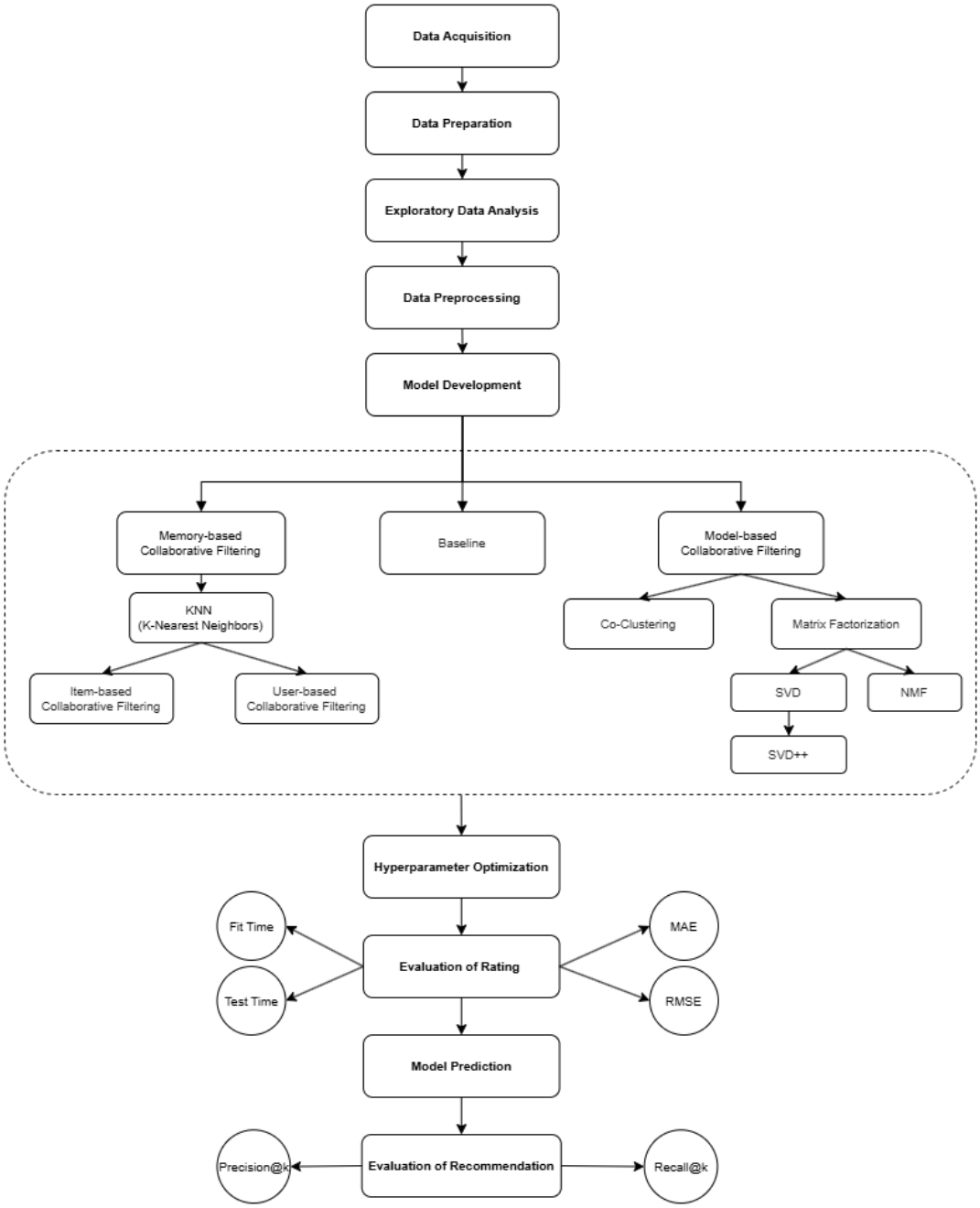
1. กระบวนการสร้างแบบจำลอง
2. การได้มาซึ่งข้อมูล (Data Acquisition)
3. การจัดเตรียมและรวมรวมข้อมูล (Data Preparation)
4. การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)
5. การจัดการข้อมูล (Data Preprocessing)
6. การพัฒนาแบบจำลอง (Model Development)
7. การหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization)
8. การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)
9. สรุปวิธีดำเนินงานวิจัย

3.1 กระบวนการพัฒนาแบบจำลอง

กระบวนการสร้างแบบจำลองสำหรับระบบแนะนำร้านอาหาร โดยการนำเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยใช้ชุดข้อมูล Yelp Open Dataset ในส่วนของข้อมูลธุรกิจ (Business) และข้อมูลความคิดเห็นและการให้คะแนน (Rating) เพื่อพัฒนาและฝึกฝนแบบจำลอง โดยดำเนินการผ่านแบบจำลองที่ถูกสร้างขึ้นมาใน Surprise Library ซึ่งประกอบด้วยแบบจำลอง Baseline ที่ใช้ค่าเฉลี่ยโดยรวม (Global Average) ของการให้คะแนนและปรับค่าอคติ (Bias) ที่เกี่ยวข้องกับผู้ใช้งานและผลิตภัณฑ์ (ร้านอาหาร) และแบบจำลองในวิธีแบบอ้างอิงจากความจำ (Memory-based Method) ที่ใช้การคำนวณความคล้ายคลึงระหว่างผู้ใช้งานหรือผลิตภัณฑ์ เพื่อสร้างการคาดการณ์คะแนนหรือรายการแนะนำ รวมถึงแบบจำลองในวิธีแบบอ้างอิงจากแบบจำลอง (Model-based Method) ใช้การเรียนรู้ของเครื่อง (Machine Learning) เพื่อค้นหารูปแบบหรือปัจจัยแฝง (Latent Factor) ที่ซ่อนอยู่ในข้อมูลการให้คะแนน

อีกทั้งมีการปรับพารามิเตอร์เพื่อหาค่าที่เหมาะสม(Hyperparameter Optimization) เพื่อให้แบบจำลองสามารถดำเนินการได้อย่างมีประสิทธิภาพและมีความเหมาะสมกับชุดข้อมูล หลังจากการปรับเพื่อหาค่าพารามิเตอร์ที่ดีที่สุด ส่วนต่อมาจะเป็นการประเมินประสิทธิภาพของ

แบบจำลอง โดยใช้การประเมินประสิทธิภาพในมิติต่างๆ ได้แก่ การประเมินประสิทธิภาพในด้านของความแม่นยำ (Accuracy Metrics), การประเมินประสิทธิภาพในด้านของการจัดอันดับการแนะนำ (Top-N Metrics), และการประเมินประสิทธิภาพในด้านของเวลาในการประมวลผล (Processing Time Performance Metric) โดยมีขั้นตอนตามภาพประกอบที่ 9



ภาพประกอบที่ 9 กระบวนการพัฒนาแบบจำลอง

3.2 **การได้มาซึ่งข้อมูล** (Data Acquisition)

ชุดข้อมูลที่นำมาใช้ในการวิจัยในครั้งนี้ เป็นข้อมูลจากแพลตฟอร์มการแสดงความคิดเห็นและให้คะแนนร้านอาหารและธุรกิจบริการชื่อดังในประเทศสหรัฐอเมริกา ซึ่งเปิดเผยข้อมูลเป็นสาธารณะ ประกอบด้วยไฟล์ทั้งหมด 6 ไฟล์ ได้แก่ ชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ (Business Data) ชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน (Review Data) ชุดข้อมูลผู้ใช้งาน (User Data) ชุดข้อมูลการเช็กอินในธุรกิจต่างๆ (Check-in Data) ชุดข้อมูลการให้ทิป (Tip Data) และชุดข้อมูลเกี่ยวกับภาพ (Photo Data) โดยหลักผู้วิจัยทำการเลือกชุดข้อมูลออกมา 2 ชุดข้อมูลที่มีความเกี่ยวข้องและสามารถนำไปปรับใช้ในงานวิจัย ได้แก่ ชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ (Business Data) มีจำนวนธุรกิจทั้งหมด 150,346 ธุรกิจ ประกอบด้วยข้อมูลพื้นฐานเกี่ยวกับธุรกิจ ยกตัวอย่างเช่น ชื่อธุรกิจ ที่อยู่ ประเภทธุรกิจ และชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน (Review Data) ซึ่งมีจำนวนการแสดงความคิดเห็นและให้คะแนนทั้งหมด 6,990,280 การแสดงความคิดเห็น ประกอบด้วยข้อมูลการแสดงความคิดเห็นเกี่ยวกับประสบการณ์การรับประทานในร้านอาหาร หรือการใช้บริการธุรกิจ รวมถึงมีข้อมูลเกี่ยวกับการให้คะแนนของผู้ใช้งานแต่ละบุคคลต่อธุรกิจที่ผู้ใช้งานได้รับประสบการณ์หรือเคยใช้บริการเช่นเดียวกัน โดยมีระดับการให้คะแนนอยู่ที่ 1-5 คะแนน

3.3 **การจัดเตรียมและรวมรวมข้อมูล** (Data Preparation)

โดยทางผู้วิจัยจะมีการจัดเตรียมข้อมูลเบื้องต้น (Data Preparation) กับทั้งชุดข้อมูลเกี่ยวกับธุรกิจและชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน รวมถึงทำความสะอาดข้อมูล เพื่อเป็นการตรวจสอบและแก้ไขข้อมูลให้อยู่ในลักษณะที่มีความพร้อมในการปรับใช้ในขั้นตอนอื่นต่อไปในรูปแบบที่ถูกต้องและครบถ้วนสมบูรณ์ โดยมีวิธีการทำความสะอาดข้อมูล (Data Cleaning) ดังนี้

1. การลดขนาดข้อมูล (Downsizing) เนื่องจากชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน (Review Data) มีขนาดใหญ่ ผู้วิจัยจำเป็นต้องลดขนาดข้อมูลลง โดยจำกัดจำนวนการนำเข้าข้อมูลอยู่ที่จำนวน 200,000 การแสดงความคิดเห็น เพื่อให้สะดวกต่อการพัฒนาแบบจำลองและลดทรัพยากรให้การประมวลผลลง

2. การกรองข้อมูล (Filtering) ในส่วนของชุดข้อมูลเกี่ยวกับธุรกิจ (Business Data) ผู้วิจัยเลือกเฉพาะข้อมูลที่มีความเกี่ยวข้องและสำคัญต่อจุดประสงค์ของงานวิจัย โดยการกรองข้อมูลประเภทธุรกิจ (categories) ที่เป็นธุรกิจร้านอาหารเท่านั้น และเมื่อตรวจสอบข้อมูลเบื้องต้นพบว่าชุดข้อมูลประกอบด้วยร้านอาหารที่ดำเนินการให้บริการอยู่ และร้านอาหารที่ปิดกิจการไป

เรียบร้อยแล้ว เพื่อให้ชุดข้อมูลมีความเหมาะสมมากยิ่งขึ้นผู้วิจัยจึงเลือกเฉพาะร้านอาหารที่ดำเนินการอยู่เท่านั้น อย่างไรก็ตามข้อมูลดังกล่าวยังคงมีลักษณะขนาดใหญ่ และอาจเป็นอุปสรรคต่อการพัฒนาและเปรียบเทียบแบบจำลองหลายในรูปแบบ ผู้วิจัยจึงดำเนินการกรองข้อมูลเพิ่มเติมโดยอิงจาก รัฐ (State) ในที่นี้ผู้วิจัยเลือก รัฐฟลอริดา (Florida) เนื่องจากรัฐดังกล่าวเป็นรัฐท่องเที่ยวที่มีนักท่องเที่ยวจำนวนมาก และมีความหลายหลายของร้านอาหารที่มีความใกล้เคียงกับประเทศไทย ในแง่ของการท่องเที่ยว โดยเฉพาะอย่างยิ่งร้านอาหารในประเทศไทยมีความหลากหลายประเภทและสัญชาติ ยกตัวอย่างเช่น อาหารสัญชาติจีน อาหารสัญชาติญี่ปุ่น และอาหารสัญชาติเกาหลี เป็นต้น

3. การจัดการกับข้อมูลที่สูญหาย (Missing Values) ชุดข้อมูลเกี่ยวกับธุรกิจ (Business Data) มีข้อมูลบางส่วนอยู่ในลักษณะของค่าว่าง ตัวอย่างเช่น ฟีเจอร์ประเภทธุรกิจ (categories) แต่เนื่องด้วยข้อมูลที่อยู่ในลักษณะของค่าว่างมีจำนวนไม่มาก ผู้วิจัยจึงดำเนินการจัดการด้วยวิธีการกำจัดข้อมูลธุรกิจที่มีค่าว่างในฟีเจอร์ประเภทธุรกิจออกจากชุดข้อมูล เพื่อความครบถ้วนของข้อมูล และไม่เป็นอุปสรรคต่อการพัฒนาแบบจำลอง

4. การจัดการกับข้อมูลซ้ำซ้อน (Duplicated Value) ข้อมูลร้านอาหารประกอบด้วยธุรกิจร้านอาหารหลายรูปแบบ โดยเฉพาะร้านอาหารที่เป็นลักษณะแฟรนไชส์ มีจำนวนมาก เพื่อให้ข้อมูลมีความหลากหลาย (Variety) และให้ชุดข้อมูลมีความเหมาะสมสำหรับแบบจำลองในการคาดการณ์รายการแนะนำไม่ให้แสดงรายการร้านอาหารที่เป็นร้านอาหารเดียวกันหรือแฟรนไชส์มาแนะนำผู้ใช้งาน จึงจำเป็นต้องจัดการกับข้อมูลในส่วนนี้เช่นเดียวกัน

โดยผลลัพธ์เมื่อผ่านจัดเตรียมข้อมูลเบื้องต้น และการทำความสะอาดข้อมูลเพื่อนำมาใช้ในงานวิจัยประกอบด้วย จำนวนการแสดงความคิดเห็นและให้คะแนน 162,124 ครั้ง จำนวนผู้ใช้งาน 88,986 ราย และจำนวนร้านอาหารทั้งหมด 4,065 ร้านอาหาร

3.4 **การวิเคราะห์ข้อมูลเชิงสำรวจ** (Exploratory Data Analysis)

สำหรับการวิเคราะห์ข้อมูลเชิงสำรวจ เพื่อสำรวจข้อมูลที่มีความสำคัญก่อนนำไปใช้กับแบบจำลอง จากข้อมูลทั้งหมดเมื่อดำเนินการวิเคราะห์ข้อมูลเชิงสำรวจในส่วนของข้อมูลเกี่ยวกับธุรกิจร้านอาหาร และข้อมูลการแสดงความคิดเห็น สามารถแสดงโครงสร้างของข้อมูลได้ ดังนี้

1. ชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ (Business Data) ประกอบด้วยข้อมูลทั้งหมดจำนวน 4,065 แถว 8 คอลัมน์ ได้แก่ รหัสธุรกิจ (business_id), ชื่อธุรกิจ (name), ที่อยู่ธุรกิจ (address), เมือง (city), รัฐ (state), การให้คะแนนธุรกิจ (stars), ลักษณะของธุรกิจ (attributes), และ ประเภทของธุรกิจ (categories) โดยมีตัวอย่างดังภาพประกอบที่ 10

	business_id	name	address	city	state	stars	attributes	categories
0	pJfh3Ct8iL58NZa8ta-a5w	Top Shelf Sports Lounge	3173 Cypress Ridge Blvd	Wesley Chapel	FL	5	{'BestNights': '{'monday': False, 'tuesday': F...	Burgers, Sports Bars, Bars, Lounges, Restauran...
1	vje0KIiE7vtpx7JzmBx5LQ	The Pearl	163 107th Ave	Treasure Island	FL	4	{'WiFi': "free", 'NoiseLevel': 'u'average",...	Restaurants, French, Moroccan, Seafood, Medite...
2	aNtKyc2rr-uK5cqzY9TVQQ	Chipotle Mexican Grill	10160 Ulmerton Rd	Largo	FL	3	{'RestaurantsPriceRange2': '1', 'Caters': 'Tru...	Mexican, Fast Food, Restaurants
3	QjV4v7q_pt7tt3K1US7lHg	PDQ Temple Terrace	5112 E Fowler Ave	Tampa	FL	3	{'RestaurantsReservations': 'False', 'Business...	Fast Food, Sandwiches, Chicken Shop, Restaurants
4	CtMEJxpVMlNzFpB4PtFjfA	Aussie Grill	25340 Sierra Center Blvd	Lutz	FL	4	{'BYOB': 'False', 'OutdoorSeating': 'True', 'B...	Restaurants, American (New), Burgers, Fast Foo...

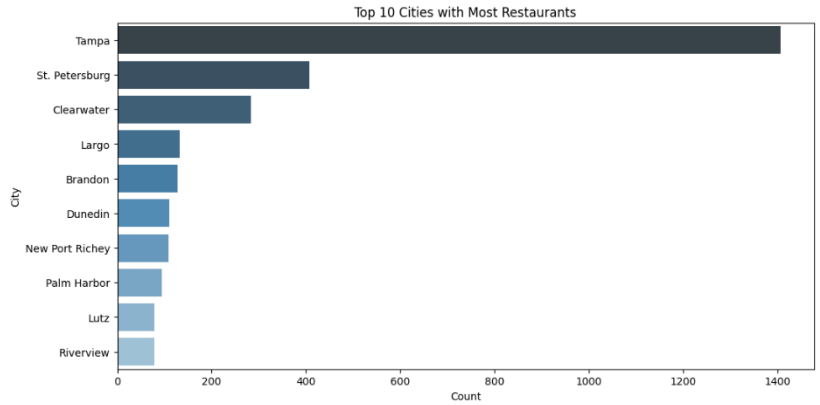
ภาพประกอบที่ 10 ตัวอย่างชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ

2. ชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน (Review Data) ประกอบด้วยข้อมูลจำนวน 162,124 แถว 9 คอลัมน์ ได้แก่ รหัสการแสดงความคิดเห็น (review_id), รหัสผู้ใช้งาน (user_id), รหัสธุรกิจ (business_id), การให้คะแนนธุรกิจ (stars), การแสดงความคิดเห็นมีประโยชน์ (useful), การแสดงความคิดเห็นน่าขบขัน (funny), การแสดงความคิดเห็นดีเยี่ยม (cool), ข้อความการแสดงความคิดเห็น (text), และวันที่แสดงความคิดเห็น (date) โดยมีตัวอย่างข้อมูล ตามประภาพกอบที่ 11

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	OAhBYw8IQ6wIfw1owXWRWw	1C2lxzUo1Hyye4RFIXly3g	BVndHaLihEYbr76Z0CMEGw	5	0	0	0	Great place for breakfast! I had the waffle, w...	2014-10-11 16:22:06
1	meGaFP7yxQdjyABrYDVeoQ	_jaJDV-qTBafatbObmtzpA	cg4JFJcCxRTTMmcg9O9KtA	1	0	1	0	Skip this train wreck if you are looking for d...	2018-02-11 03:11:20
2	FMFZoES5LfumuwhsZbIDlQ	4ubLHlnMFFw4JikcXr-F4w	f4PA-f1tcN1blpZJLdFsQQ	5	3	0	0	I must admit, I wasn't expecting much. This pl...	2017-05-26 02:32:01
3	PPgbLBvi34A6m7bKJfTwhw	3TL6HZ1JrKcNTvGDWKlrow	GyC36Pn0Q1-qHnqXys6yFg	1	0	0	0	Service and management terrible... After messi...	2013-12-07 13:17:13
4	OJO7m2zn3LAr011J1I7ppQ	JriXL8qqw_tJ1mpwtlBabg	1QVB0_-piu0GXes87BXeGw	5	1	0	0	Love this place...best hot dogs and chili dogs...	2017-10-11 01:27:15

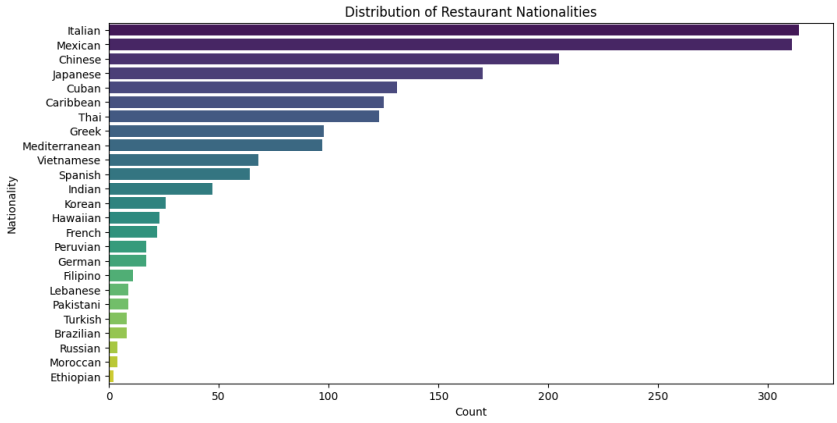
ภาพประกอบที่ 11 ตัวอย่างชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการให้คะแนน

จากชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ เมื่อดำเนินการสำรวจจำนวนของร้านอาหารโดยอิงจากเมืองที่ตั้งของธุรกิจร้านอาหาร ตามภาพประกอบ 12 พบว่าเมือง Tempa มีสัดส่วนของร้านอาหารมากสุด รองลงมาเป็นเมือง Clearwater และ Saint Petersburg แสดงให้เห็นถึงการกระจุกตัวของร้านอาหารที่ตั้งอยู่ในเมือง Tempa เป็นส่วนใหญ่



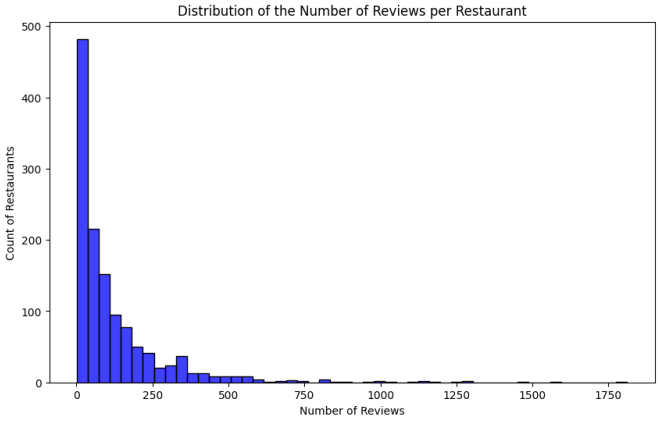
ภาพประกอบที่ 12 กราฟแสดงจำนวนร้านอาหารในแต่ละเมือง 10 อันดับแรก

ในส่วนของการสำรวจการกระจายตัวของจำนวนร้านอาหารโดยแบ่งตามสัญชาติของร้านอาหาร พบว่าร้านอาหารสัญชาติอิตาลี (Italian) มีจำนวนมากที่สุด รองลงมาจะเป็นร้านอาหารสัญชาติเม็กซิโก (Mexican) และร้านอาหารสัญชาติจีน (Chinese) ซึ่งจะเห็นได้ว่าร้านอาหารในรัฐฟลอริดามีความหลากหลายของสัญชาติจำนวนมาก รวมถึงแสดงถึงความหลากหลายและเป็นรัฐที่มีความเป็นรัฐที่เป็นจุดหมายปลายทางของนักท่องเที่ยวเช่นเดียวกัน ตามภาพประกอบที่ 13



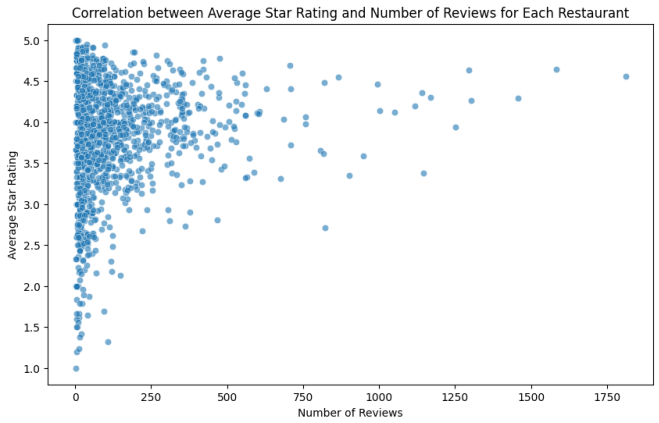
ภาพประกอบที่ 13 กราฟแสดงการกระจายตัวของร้านอาหารแบ่งตามสัญชาติของร้านอาหาร

นอกจากนี้ผู้วิจัยได้ทำการสำรวจการกระจายตัวของจำนวนการแสดงความคิดเห็นต่อร้านอาหาร ตามภาพประกอบที่ 14 พบว่าการกระจายตัวอยู่ในลักษณะเบ้ขวา ตามภาพประกอบแสดงให้เห็นว่าร้านอาหารส่วนใหญ่มีการแสดงความคิดเห็นและการให้คะแนน (Rating) ไม่มาก อย่างไรก็ตามมีร้านอาหารจำนวนหนึ่งที่มีการแสดงความคิดเห็นค่อนข้างสูงในอีกด้านหนึ่งของของกราฟ การกระจายตัวของชุดข้อมูลยังมีไม่มาก



ภาพประกอบที่ 14 กราฟแสดงการกระจายตัวของจำนวนการแสดงความคิดเห็นต่อร้านอาหาร

รวมถึงผู้วิจัยมีการสำรวจการกระจายตัวของค่าเฉลี่ยการให้คะแนน (Average Rating) พบว่าการให้คะแนนกระจุกตัวอยู่บริเวณ 3-5 คะแนน โดยร้านอาหารที่มีค่าเฉลี่ยการให้คะแนนสูงมีแนวโน้มที่มีจำนวนการแสดงความคิดเห็นที่ต่ำ ร้านอาหารที่มีจำนวนการแสดงความคิดเห็นที่ต่ำจะมีช่วงของการให้คะแนนตั้งแต่ 3-5 คะแนนเป็นหลัก แสดงให้เห็นถึงแนวโน้มการให้คะแนนที่ค่อนข้างสูง และร้านอาหารที่มีจำนวนการแสดงความคิดเห็นที่สูงส่วนใหญ่มีไม่มากนัก ตามภาพประกอบที่ 15



ภาพประกอบที่ 15 กราฟแสดงการกระจายตัวของค่าเฉลี่ยการให้คะแนน

3.5 **การจัดการข้อมูล** (Data Preprocessing)

การจัดการข้อมูลในการพัฒนาแบบจำลองด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เพื่อให้ชุดข้อมูลที่นำไปปรับใช้ในแบบจำลองมีความครบถ้วนและเหมาะสม โดยทางผู้วิจัยดำเนินการจัดการข้อมูลด้วย 5 วิธี ดังนี้

1. การรวมข้อมูล (Combining) เป็นการรวมชุดข้อมูล 2 ชุดเข้าด้วยกัน ได้แก่ ชุดข้อมูลเกี่ยวกับข้อมูลธุรกิจ (Business Data) และชุดข้อมูลเกี่ยวกับการแสดงความคิดเห็นและการ

ให้คะแนน (Review Data) เพื่อให้มั่นใจว่าข้อมูลจากทั้งสองชุดข้อมูลมีความถูกต้องสมบูรณ์ รวมถึงมีการเปลี่ยนแปลงชื่อคอลัมน์ที่มีความสำคัญในแบบจำลองเพื่อความเหมาะสมของข้อมูล

2. การเลือกข้อมูลที่เกี่ยวข้อง ในการพัฒนาแบบจำลองด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม สิ่งที่สำคัญที่สุด คือ การให้คะแนนต่อธุรกิจหรือร้านอาหารต่างๆ ดังนั้นผู้วิจัยได้ทำการเลือกเฉพาะคอลัมน์ที่เกี่ยวข้องและเหมาะสมในด้านการประมวลผลของแบบจำลองมากที่สุด ประกอบด้วย รหัสผู้ใช้งาน (user_id), รหัสธุรกิจ (business_id), และการให้คะแนนธุรกิจ (rating)

3. การทำข้อมูลให้เป็นมาตรฐานเดียวกัน (Standardization) ผู้วิจัยพบว่าข้อมูลในส่วนของคอลัมน์การให้คะแนน (rating) มีค่าที่มีความแตกต่างกันทั้งเพิ่มขึ้นและลดลงทีละ 0.5 คะแนน ซึ่งทำให้ช่วงของการให้คะแนนมีลักษณะค่อนข้างกว้าง เพื่อทำให้ข้อมูลมีประสิทธิภาพและง่ายต่อการพัฒนาและการประมวลผลของแบบจำลอง ผู้วิจัยจึงดำเนินการเปลี่ยนแปลงข้อมูลให้เป็นมาตรฐานด้วยการเปลี่ยนแปลงค่าให้เป็นค่าคงที่อยู่ระหว่าง 1-5 คะแนน และมีช่วงความแตกต่างของคะแนนอยู่ที่ 1 คะแนน

4. การกรองข้อมูล (Filtering) เพื่อให้การทำงานของแบบจำลองที่ผู้วิจัยมีความประสงค์ในการพัฒนาและเปรียบเทียบในแบบจำลองหลากหลายประเภท ผู้วิจัยจึงทำการกรองข้อมูลในส่วนของการให้คะแนนธุรกิจสำหรับผู้ใช้งานที่มีการให้คะแนนต่อธุรกิจมากกว่า 2 ครั้งขึ้นไป และธุรกิจที่ถูกให้คะแนนจากผู้ใช้งานมากกว่า 3 ครั้งขึ้นไป เพื่อเป็นการลดข้อมูลหรือความแปรปรวนที่ไม่เกี่ยวข้อง (Noise) ที่อาจเกิดขึ้นระหว่างการพัฒนาหรือประมวลผลของแบบจำลอง

5. การแบ่งสัดส่วนของข้อมูล (Data Splitting) ผู้วิจัยได้ทำการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลฝึกฝน (Training Data) และชุดข้อมูลทดสอบ (Test Data) โดยมีสัดส่วนของข้อมูลอยู่ที่ 80:20 ตามลำดับ

3.6 การพัฒนาแบบจำลอง (Model Development)

การพัฒนาแบบจำลองด้วยเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เริ่มจากการนำชุดข้อมูลที่ผ่านการรวมข้อมูล (Combined Data) และถูกแบ่งข้อมูลออกเป็นชุดข้อมูลสำหรับฝึกฝน 80% และชุดข้อมูลสำหรับทดสอบ 20% จากนั้นนำข้อมูลในส่วนชุดข้อมูลฝึกฝน (Training Data) มาปรับใช้กับแบบจำลองที่ถูกพัฒนาเรียบร้อยแล้วบน Surprise Library ซึ่งสามารถแบ่งแบบจำลองเป็น 3 ประเภท ได้แก่

1. แบบจำลองประเภท Baseline ผู้วิจัยใช้แบบจำลอง BaselineOnly ซึ่งเป็นแบบจำลองพื้นฐานคำนวณค่าเฉลี่ยของการให้คะแนนทั้งหมด (Global Average Rating) และค่าอคติ (Bias) ที่คำนวณจากค่าความเบี่ยงเบนของการให้คะแนนของผู้ใช้งาน และธุรกิจร้านอาหาร

จากค่าเฉลี่ยของการให้คะแนนทั้งหมด เพื่อนำมาใช้เป็นพื้นฐานในการเปรียบเทียบกับแบบจำลองประเภทอื่น

2. แบบจำลองในวิธีแบบอ้างอิงจากความจำ (Memory-based Method) ประกอบด้วยแบบจำลองที่มีการใช้วิธีการหาเพื่อนบ้านที่ใกล้เคียงที่สุด (K-nearest neighbors : KNN) ซึ่งประกอบด้วย 2 แบบจำลองหลัก คือ User-based Collaborative Filtering โดยใช้อัลกอริทึม KNNBasic โดยตั้งค่า user_based ให้มีค่าเป็น True และใช้ค่าความคล้ายคลึงเป็นค่าเริ่มต้น ในส่วนของ Item-based Collaborative Filtering ผู้วิจัยใช้อัลกอริทึม KNNBasic เช่นดียวกับ User-based Collaborative Filtering แต่มีการปรับเปลี่ยนการตั้งค่าแบบจำลองในส่วนของ user_based .ให้มีค่าเป็น False อัลกอริทึมจะทำการเปลี่ยนแปลงเป็น Item-based

3. แบบจำลองในวิธีแบบอ้างอิงจากแบบจำลอง (Model-based Method) ที่ใช้การเรียนรู้ของเครื่องเข้ามาช่วยในการคาดการณ์รายการการแนะนำร้านอาหาร โดยในงานวิจัยนี้มีแบบจำลองลักษณะดังกล่าวทั้งหมด 4 แบบจำลอง คือ Singular Value Decomposition, Singular Value Decomposition++, Non-Negative Matrix Factorization, และ Co-Clustering

เมื่อดำเนินการพัฒนาแบบจำลองด้วยค่าพารามิเตอร์ที่ถูกตั้งค่าไว้เป็นค่าตั้งต้น (Default) ผู้วิจัยดำเนินการทดสอบแบบจำลองด้วยชุดข้อมูลทดสอบ (Test Data) ในเบื้องต้น โดยนำค่า RMSE มาใช้ในการประเมินผล และเพื่อนำข้อมูลประสิทธิภาพจากการทดสอบแบบจำลองดังกล่าวใช้ในการเปรียบเทียบกับแบบจำลองที่ถูกพัฒนาด้วยค่าพารามิเตอร์ที่เหมาะสมที่สุดต่อไป

3.7 การหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization)

สำหรับงานวิจัยนี้มีการดำเนินการปรับค่าพารามิเตอร์เพื่อหาพารามิเตอร์ที่เหมาะสม (Hyperparameter) ในแต่ละแบบจำลอง เพื่อให้แบบจำลองสามารถประมวลผลและให้ผลลัพธ์ได้อย่างมีประสิทธิภาพมากที่สุด ซึ่งจากงานวิจัย GSO-CRS: grid search optimization for collaborative recommendation system (Behera & Nain, 2022) นำเสนอวิธีการหาค่าพารามิเตอร์ที่เหมาะสมวิธีต่างๆกับในแบบจำลองในกลุ่มของอัลกอริทึมแบบ Matrix Factorization อย่างแบบจำลอง Singular Value Decomposition (SVD) และ Probabilistic Matrix Factorization (PMF) พบว่า Grid Search Optimization ช่วยเพิ่มประสิทธิภาพให้กับแบบจำลอง Singular Value Decomposition (SVD) ในการลดค่าความคลาดเคลื่อน และช่วยเพิ่มความแม่นยำในการคาดการณ์รายการแนะนำยิ่งขึ้น ผู้วิจัยจึงประยุกต์ใช้วิธีการปรับค่าพารามิเตอร์แบบ Grid Search Optimization โดยอ้างอิงจากค่า Root Mean Squared Error (RMSE) ที่ดีที่สุดหรือมีค่าต่ำ รวมถึงตั้งค่า Cross-Validation เท่ากับ 5 เพื่อหาค่าเฉลี่ยของผลของ

การพัฒนาแบบจำลองทั้งหมด 5 ครั้ง เพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพที่สุดและไม่เกิดการที่แบบจำลองทำงานได้มีประสิทธิภาพมากกับชุดข้อมูลฝึกฝน แต่ในทางกลับกันทำงานได้ไม่ดีกับชุดข้อมูลทดสอบ (Overfit) โดยในแต่ละแบบจำลองมีการปรับค่าพารามิเตอร์แตกต่างกันออกไป ยกตัวอย่างเช่น แบบจำลอง SVD มีการปรับเปลี่ยนพารามิเตอร์ในส่วนของ n_factors คือ การกำหนดจำนวนฟีเจอร์แฝง (Latent Features) ที่ใช้ในวิธีการแยกตัวประกอบของเมทริกซ์ n_epochs คือ จำนวนรอบ (Iterations) ที่แบบจำลองจะทำการฝึกฝน โดยวนซ้ำไปบนข้อมูล user-item matrix เพื่อปรับค่าพารามิเตอร์ให้เหมาะสม lr_all คือ ควบคุมขนาดการเปลี่ยนแปลงของการอัปเดตพารามิเตอร์ระหว่างการฝึกแบบจำลอง และ reg_all คือ การกำหนดค่า Regularization หรือเป็นการลดความซับซ้อนของแบบจำลองเพื่อให้แบบจำลองลดค่าความแปรปวนลง (Variance) ไม่ให้เกิดการ Overfitting โดยแบบจำลองและพารามิเตอร์ในแต่ละแบบจำลองผู้วิจัยมีการปรับเปลี่ยน ตามตารางที่ 2 ดังนี้

ตารางที่ 2 รายชื่อแบบจำลองและพารามิเตอร์ที่ดำเนินการปรับเปลี่ยนด้วยวิธี Grid Search Optimization เพื่อหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter)

แบบจำลอง	พารามิเตอร์
Baseline	• bsl_options - method ['sgd'] - reg_u [10] - reg_i [5] - learning_rate [0.01]
User-based Collaborative Filtering (KNN User-based)	• k [30] • min_k [10] • sim_options - name ['msd'] - user_based [True]
Item-based Collaborative Filtering (KNN Item-based)	• k [35] • min_k [1] • sim_options - name ['msd'] - user_based [False]

ตารางที่ 2 (ต่อ)

แบบจำลอง	พารามิเตอร์
Singular Value Decomposition (SVD)	• n_factors [150] • n_epochs [20] • lr_all [0.01] • reg_all [0.2]
Singular Value Decomposition++ (SVD++)	• n_factors [50] • n_epochs [20] • lr_all [0.01] • reg_all [0.1]
Non-negative Matrix Factorization (NMF)	• n_factors [20] • n_epochs [20] • reg_pu [0.1] • reg_qi [0.1]
Co-Clustering	• n_cltr_u': [2] • n_cltr_i': [2] • n_epochs': [20]

3.8 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)

การประเมินประสิทธิภาพของแบบจำลองในงานวิจัย ประกอบด้วยตัวชี้วัด 3 ประเภทหลัก ดังนี้ การประเมินประสิทธิภาพในด้านของความแม่นยำ (Accuracy Metrics) ในส่วนนี้จะใช้ค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) ในการประเมินประสิทธิภาพแบบจำลองเพื่อหาค่าความคลาดเคลื่อนในการคากการณ์ผลลัพธ์ของแบบจำลองในส่วนของการคาดการณ์คะแนนเป็นหลัก

อีกทั้งมีการใช้การประเมินประสิทธิภาพในด้านของการจัดอันดับการแนะนำ (Top-N Metrics) เพื่อทำการประเมินประสิทธิภาพของแบบจำลองในการคาดการณ์รายการแนะนำในเรื่องของความแม่นยำด้วยค่า Precision@k และในส่วนของความถูกต้องและครอบคลุมรายการแนะนำที่เกี่ยวข้องกับผู้ใช้งาน (Relevant Item) ด้วยค่า Recall@k โดยมีการกำหนดค่าคะแนน (Threshold) ในการบ่งชี้ว่ารายการแนะนำมีความเกี่ยวข้องกับผู้ใช้งานมากน้อยเพียงใด ซึ่ง

กำหนดให้คะแนนที่แสดงถึงความเกี่ยวข้องกับผู้ใช้งาน หากแบบจำลองคาดการณ์มากกว่าหรือเท่ากับ 3 คะแนน ถือว่ามีความเกี่ยวข้องกับผู้ใช้งาน โดยผู้วิจัยมองเห็นว่าคะแนนที่กำหนดอยู่ที่ 3 คะแนน เป็นคะแนนที่อยู่ในระดับกลาง อีกทั้งจากการสำรวจข้อมูลเชิงวิเคราะห์ การกระจายตัวของคะแนน คะแนนเริ่มกระจุกตัวอยู่ในระดับตั้งแต่ 3 คะแนนขึ้นไปจึงเป็นอีกปัจจัยหนึ่งในการกำหนดค่าคะแนนดังกล่าว ซึ่งช่วยให้สามารถประเมินการจัดอันดับรายการแนะนำได้หลากหลายและครอบคลุม รวมถึงให้ค่า Precision@k และ Recall@k มีความสมดุลกัน ในงานวิจัยนี้ผู้วิจัยเลือกในมีการจัดอันดับรายการแนะนำ 10 รายการ หรือกำหนดให้ค่า k=10 โดยอ้างอิงจากการจัดอันดับรายการแนะนำในเว็บไซต์ yelp.com ที่มีการแสดงผลรายการแนะนำร้านอาหารตั้งต้นที่ 10 ร้านบนหน้าเว็บเพจสำหรับการแนะนำ

รวมถึงมีการประเมินในด้านของเวลาในการประมวลผล (Processing Time Performance Metric) ด้วยค่า Fit Time และ Test Time ที่ใช้ในการประเมินระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองและการคาดการณ์ผลลัพธ์จากแบบจำลอง

3.9 **สรุปวิธีดำเนินงานวิจัย**

การเปรียบเทียบประสิทธิภาพและการประมวลผลแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม 3 ประเภทหลัก ได้แก่ แบบจำลองประเภท Baseline แบบจำลองในวิธีแบบอ้างอิงจากความจำ และแบบจำลองในวิธีแบบอ้างอิงจากแบบจำลอง รวมทั้งหมด 7 แบบจำลอง โดยดำเนินการผ่านแบบจำลองที่ผ่านการพัฒนาใน Surprise Library และนำมาปรับใช้กับชุดข้อมูลการแสดงความคิดเห็นและการให้คะแนนในธุรกิจร้านอาหาร รวมถึงมีการปรับค่าพารามิเตอร์เพื่อหาค่าพารามิเตอร์ที่เหมาะสมในแต่ละแบบจำลองด้วยวิธี Grid Search Optimization โดยอิงจากค่า RMSE ที่ดีที่สุด เพื่อให้แบบจำลองประมวลผลออกมาได้ผลลัพธ์ที่ดีและเหมาะสมที่สุดกับชุดข้อมูลในงานวิจัยนี้ และมีการประเมินผลในแต่ละอัลกอริทึมด้วยค่า MAE และ RMSE อีกทั้งการคาดการณ์รายการแนะนำของแบบจำลองด้วยค่า Precision@k และ Recall@k ในอับดับการแนะนำ 10 อันดับแรก และเวลาในการประมวลผลของแบบจำลอง

บทที่ 4
ผลการดำเนินงานวิจัย

ในบทที่ 4 ผู้วิจัยนำเสนอผลการดำเนินงานวิจัย เรื่องการวิเคราะห์เปรียบเทียบเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม เพื่อเพิ่มประสิทธิภาพในระบบแนะนำร้านอาหาร โดยแบ่งผลการดำเนินงานวิจัยเป็น 3 ส่วน ดังนี้

1. ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลอง
2. ผลลัพธ์การเปรียบเทียบความถูกต้องแม่นยำในการแนะนำของแบบจำลอง
3. ผลลัพธ์การแนะนำร้านอาหารโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

4.1 ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม ประกอบด้วยแบบจำลอง 7 แบบ ได้แก่ KNN User-based, KNN Item-based, Singular Value Decomposition, Singular Value Decomposition++, Non-negative Matrix Factorization, และ Co-Clustering รวมทั้งแบบจำลองแบบ Baseline ที่คำนวณมาจากค่าเฉลี่ยของการให้คะแนนทั้งหมด (Global Average Rating) และค่าอคติ (Bias) ซึ่งแบบจำลองดังกล่าวทั้งหมดจะถูกประเมินประสิทธิภาพด้วยการทำ Cross-Validation จำนวนทั้งหมด 5 Folds ด้วยการใช้ค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) ในการประเมินประสิทธิภาพในเรื่องของความแม่นยำของแบบบจำลอง และค่า Fit Time และ Test Time ในการประเมินประสิทธิภาพในด้านของเวลาในการประมวลผลของแบบจำลอง

4.1.1 การเปรียบเทียบประสิทธิภาพของแบบจำลองระหว่างก่อนและหลังการปรับปรุงค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization)

ผู้วิจัยได้ดำเนินการเปรียบเทียบประสิทธิภาพของแบบจำลองระหว่างก่อนและหลังการหาค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization) โดยการเปรียบเทียบอัตราการเปลี่ยนแปลงของประสิทธิภาพของแบบจำลอง ซึ่งแบ่งเป็น 4 ค่าการประเมินประสิทธิภาพที่แตกต่างกัน ดังนี้

4.1.1.1 Mean Absolute Error (MAE)

สำหรับการเปรียบเทียบประสิทธิภาพในด้านของความแม่นยำของแบบจำลอง (Accuracy Matric) โดยประเมินด้วยค่า Mean Absolute Error (MAE) ระหว่างก่อนและหลังการหาค่าพารามิเตอร์ที่เหมาะสม

เพื่อแสดงให้เห็นการเปลี่ยนแปลงก่อนค่าความคลาดเคลื่อน (Error) ของแบบจำลอง จากตารางที่ 3 แบบจำลอง NMF เป็นแบบจำลองที่มีการเปลี่ยนแปลงในเรื่องของค่าความคลาดเคลื่อนที่ลดลงมากที่สุด เมื่อดำเนินการปรับค่าพารามิเตอร์ให้เหมาะสม จากเดิมค่า MAE อยู่ที่ 1.0526 ลดลงกว่า 9.36% ส่งผลให้แบบจำลองมีความแม่นยำมากขึ้นอยู่ที่ 0.9541 สำหรับแบบจำลอง Baseline ถือเป็นอีกหนึ่งแบบจำลองที่มีประสิทธิภาพสูงขึ้นในด้านของค่า MAE ลดลงอยู่ที่ 2.32% รวมถึงแบบจำลอง SVD++ และ KNN User-based ที่มีค่า MAE ที่ลดลงประมาณ 1.38%

ตารางที่ 3 การเปรียบเทียบการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยค่า MAE

แบบจำลอง	MAE Default Parameters	MAE Best Parameters	MAE Improvement (%)
Baseline	0.9109	0.8897	2.32%
SVD	0.9006	0.8922	0.93%
SVD++	0.9012	0.8888	1.38%
Co-Clustering	0.9230	0.9167	0.68%
NMF	1.0526	0.9541	9.36%
KNN User-based	0.9442	0.9312	1.38%
KNN Item-based	0.9576	0.9557	0.19%

4.1.1.2 Root Mean Squared Error (RMSE)

ในส่วนของ Root Mean Squared Error (RMSE) เป็นอีกหนึ่งค่าที่นิยมในการประเมินประสิทธิภาพในด้านของความแม่นยำของแบบจำลอง โดยเป็นค่าที่แสดงถึงความแตกต่างระหว่างค่าคาดการณ์ที่ได้จากแบบจำลอง (Predicted) และค่าจริง (Actual) จากตารางที่ 4 พบว่าแบบจำลอง NMF เป็นแบบจำลองที่มีประสิทธิภาพเพิ่มมากขึ้นสูงที่สุดหลังจากดำเนินการปรับค่าพารามิเตอร์ให้เหมาะสม ซึ่งมีค่า RMSE ที่ต่ำลงกว่าพารามิเตอร์เริ่มต้นอยู่ที่ 4.29% หรือมีค่า RMSE อยู่ที่ 1.3206 สำหรับแบบจำลอง KNN User-based มีค่า RMSE ลดลงเช่นเดียวกัน

เมื่อผ่านการปรับค่าพารามิเตอร์ที่เหมาะสมประมาณ 2.87% และแบบจำลอง SVD++ มีลักษณะเช่นเดียวกับแบบจำลองข้างต้นโดยมีค่า RMSE ลดลงจากเดิมประมาณ 1.33% โดยภาพรวมในแบบจำลองต่างๆ มีค่า Root Mean Squared Error (RMSE) ที่ลดลงเมื่อผ่านการปรับค่าพารามิเตอร์ที่เหมาะสมในแต่ละแบบจำลอง

ตารางที่ 4 การเปรียบเทียบการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยค่า RMSE

แบบจำลอง	RMSE Default Parameters	RMSE Best Parameters	RMSE Improvement (%)
Baseline	1.1549	1.1443	0.92%
SVD	1.1509	1.1395	0.99%
SVD++	1.1573	1.1419	1.33%
Co-Clustering	1.3059	1.2981	0.60%
NMF	1.3798	1.3206	4.29%
KNN User-based	1.2537	1.2178	2.87%
KNN Item-based	1.3610	1.3603	0.06%

4.1.1.3 Fit Time

การประเมินประสิทธิภาพในด้านของเวลาในการประมวลผลของแบบจำลอง Fit Time เป็นหนึ่งในค่าที่ใช้ในการประเมินเวลาของแบบจำลองระหว่างการฝึกฝน โดยบางแบบจำลองเมื่อดำเนินการปรับค่าพารามิเตอร์ให้เหมาะสมแล้ว แม้ว่าผลลัพธ์จากการฝึกฝนแบบจำลองจะมีประสิทธิภาพที่เพิ่มขึ้น อย่างไรก็ตามกลับทำให้เวลาในการฝึกฝนเพิ่มมากขึ้นเช่นกัน จากตารางที่ 5 แบบจำลอง SVD ใช้เวลาในการฝึกฝนแบบจำลองนานขึ้นเมื่อผ่านการปรับค่าพารามิเตอร์ที่เหมาะสมกว่า 97.98% เมื่อเทียบกับแบบจำลองที่ประมวลผลด้วยพารามิเตอร์ตั้งต้น, ส่วนของแบบจำลอง Baseline พบว่าเวลาในการประมวลผลในช่วงฝึกฝนนานขึ้นเช่นเดียวกันถึง 50.63% และอีกแบบจำลองที่ใช้เวลาใน

การประมวลผลช่วงฝึกฝนแบบจำลองเพิ่มขึ้นได้แก่แบบจำลอง SVD ที่ใช้เวลานานขึ้นประมาณ 26.84% ในทางกลับกันแบบจำลอง NMF ใช้เวลาในการประมวลผลในช่วงฝึกฝนแบบจำลองรวดเร็วขึ้นต่างสูงที่สุดเมื่อเทียบกับแบบจำลองอื่นสูงถึง 53.40%

ตารางที่ 5 การเปรียบเทียบการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยค่า Fit Time

แบบจำลอง	Fit Time Default Parameters	Fit Time Best Parameters	Fit Time Improvement (%)
Baseline	0.4495	0.6951	-50.63%
SVD	1.1212	1.4222	-26.84%
SVD++	2.1068	4.1712	-97.98%
Co-Clustering	2.9629	2.9372	0.86%
NMF	2.9834	1.3904	53.40%
KNN User-based	8.0403	7.9870	0.66%
KNN Item-based	0.0785	0.0806	-2.70%

4.1.1.4 Test time

ค่าที่แสดงถึงเวลาในการทดสอบ (Test Time) เป็นค่าในการประเมินประสิทธิภาพด้านเวลาที่แบบจำลองประมวลผลเพื่อคาดการณ์หรือทำนาย ซึ่งเวลาในช่วงนี้จะใช้เวลาน้อยกว่าเวลาในการฝึกฝน เนื่องจากเป็นช่วงเวลาที่แบบจำลองทำงานกับข้อมูลขนาดเล็กกว่าชุดข้อมูลฝึกฝน โดยเมื่อปรับค่าพารามิเตอร์ให้เหมาะสม ในบางแบบจำลองอาจมีเวลาในการทดสอบเพิ่มมากขึ้น จากตารางที่ 6 แบบจำลอง Co-Clustering ใช้เวลาในการทดสอบนานสูงถึง 65.95% , SVD++ ใช้เวลาในการทดสอบนานขึ้น 15.68%, และ Baseline ใช้เวลาในการประมวลผลเพื่อคาดการณ์นานขึ้น 11.55% โดยแบบจำลองดังกล่าวถือว่าเป็นแบบจำลองได้ได้รับผลกระทบมากที่สุดในด้านของเวลาในการทำนายหรือคาดการณ์เมื่อเทียบกับแบบจำลองอื่น อย่างไรก็ตามในส่วนของ

แบบจำลอง SVD เมื่อมีการปรับค่าพารามิเตอร์ ส่งผลให้แบบจำลองสามารถประมวลผลในช่วงทดสอบที่เร็วมากยิ่งขึ้นกว่าเดิมถึง 18.05%

ตารางที่ 6 การเปรียบเทียบการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยค่า Test Time

แบบจำลอง	Test time Default Parameters	Test time Best Parameters	Test time Improvement (%)
Baseline	0.0856	0.0955	-11.55%
SVD	0.1475	0.1209	18.05%
SVD++	0.3663	0.4238	-15.68%
Co-Clustering	0.0704	0.1170	-65.95%
NMF	0.1250	0.1222	2.25%
KNN User-based	2.7311	2.6572	2.71%
KNN Item-based	0.2761	0.2965	-7.36%

4.1.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการการหาค่าพารามิเตอร์ที่เหมาะสม

ผู้วิจัยดำเนินการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการปรับค่าพารามิเตอร์ที่เหมาะสม โดยแบ่งเป็นการเปรียบเทียบในส่วนของความแม่นยำของแบบจำลองด้วยค่าความคลาดเคลื่อน (Error) ได้แก่ ความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) รวมถึงการเปรียบเทียบประสิทธิภาพด้านของเวลาในการประมวลผล ได้แก่ การเปรียบเทียบประสิทธิภาพด้านเวลาในการประมวลผลในช่วงเวลาฝึกฝนของแบบจำลอง (Fit Time) และการเปรียบเทียบประสิทธิภาพด้านเวลาในการประมวลผลในช่วงเวลาทดสอบของแบบจำลอง (Test Time)

4.1.2.1 การเปรียบเทียบความแม่นยำของแบบจำลองด้วยค่าความคลาดเคลื่อน (Error)

ในส่วนของการเปรียบเทียบความแม่นยำของแบบจำลองด้วยค่าความคลาดเคลื่อน ผู้วิจัยดำเนินการเปรียบเทียบด้วยค่า ความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) จากตารางที่

6 พบว่า แบบจำลองที่มีค่า MAE น้อยที่สุด คือ SVD++ ซึ่งมีค่าเท่ากับ 0.8889 รองลงมา คือ แบบจำลอง Baseline มีค่า MAE เท่ากับ 0.8898 และแบบจำลอง SVD มีค่า MAE เท่ากับ 0.8923 ในขณะที่แบบจำลองที่เหลืออีก 4 แบบจำลอง มีค่า MAE ที่สูงกว่า 0.9 ขึ้นไป แต่หากเปรียบเทียบแบบจำลองในอีกส่วนหนึ่งที่เป็นค่า RMSE แบบจำลองที่มีค่าต่ำที่สุด ได้แก่ SVD โดยมีค่าอยู่ที่ 1.1395 รองลงมา คือ SVD++ มีค่า RMSE เท่ากับ 1.1420 และแบบจำลองที่มีค่า RMSE ใกล้เคียงกับอันดับก่อนหน้าได้แก่ Baseline ซึ่งมีค่าเท่ากับ 1.1444

ตารางที่ 7 การเปรียบเทียบประสิทธิภาพแบบจำลองของค่า MAE และ RMSE

แบบจำลอง	Mean Average Error (MAE)	Root Mean Squared Error (RMSE)
Baseline	0.8898	1.1444
SVD	0.8923	1.1395
SVD++	0.8889	1.1420
Co-Clustering	0.9168	1.2981
NMF	0.9541	1.3206
KNN User-based	0.9312	1.2178
KNN Item-based	0.9558	1.3603

4.1.2.2 การเปรียบเทียบประสิทธิภาพด้านของเวลาในการประมวลผล

สำหรับการเปรียบเทียบประสิทธิภาพด้านของเวลาในการประมวลผลของแบบจำลองต่างๆ ผู้วิจัยใช้ค่าที่ใช้ในการประเมินเวลาของแบบจำลองระหว่างการฝึกฝน (Fit Time) และค่าที่ใช้ในการประเมินเวลาของแบบจำลองระหว่างการทดสอบ (Test Time) โดยแบบจำลองที่ใช้เวลาในช่วงการฝึกฝนน้อยที่สุด คือ KNN Item-based มีค่าเท่ากับ 0.0806 ในทางกลับกันแบบจำลองที่ใช้เวลานานที่สุดในช่วงเวลาการฝึกฝน คือ KNN User-based โดยมีค่า Fit Time สูงถึง 7.8970 สำหรับเวลาในการทดสอบ แบบจำลอง Baseline ใช้เวลาน้อยที่สุดขณะเดียวกัน KNN User-based ยังคงใช้เวลาทดสอบนานที่สุดเช่นเดียวกัน

ตารางที่ 8 การเปรียบเทียบประสิทธิภาพแบบจำลองของค่า Fit Time และ Test Time

แบบจำลอง	Fit Time	Test Time
Baseline	0.6951	0.0955
SVD	1.4222	0.1209
SVD++	4.1712	0.4238
Co-Clustering	2.9372	0.1170
NMF	1.3904	0.1222
KNN User-based	7.9870	2.6572
KNN Item-based	0.0806	0.2965

4.2 ผลลัพธ์การเปรียบเทียบประสิทธิภาพการจัดอันดับการแนะนำของแบบจำลอง

การประเมินประสิทธิภาพการจัดอันดับการแนะนำของแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม โดยผู้วิจัยดำเนินเปรียบเทียบประสิทธิภาพโดยการใช้ค่าที่มีความสำคัญในการประเมินผลลัพธ์รายการแนะนำ 2 แบบ ได้แก่ Precision@k กล่าวคือ ค่าความแม่นยำของการแนะนำจากรายการแนะนำ k อันดับแรก และ Recall@k กล่าวคือ ค่าความถูกต้องของการแนะนำจากรายการแนะนำ k อันดับแรก สำหรับการเปรียบเทียบผู้วิจัยดำเนินการแจกแจงการเปรียบเทียบในแต่ละแบบจำลองแบ่งตามจำนวนของรายการแนะนำ (k) 10 อันดับแรก เพื่อแสดงให้เห็นการเปรียบเทียบเชิงผลลัพธ์ของแบบจำลองต่อรายการแนะนำที่เปลี่ยนแปลงไป

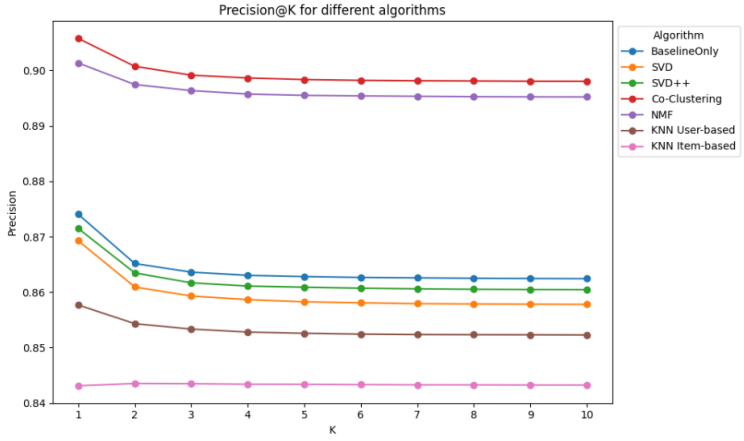
4.2.1 การเปรียบเทียบความแม่นยำของแบบจำลองในการแนะนำจากรายการแนะนำ 10 อันดับแรก (Precision@k)

การเปรียบเทียบความแม่นยำของผลลัพธ์รายการแนะนำจากแบบจำลองแบบต่างๆ ในงานวิจัย รวมถึงเปรียบเทียบรายการแนะนำตั้งแต่จำนวนรายการแนะนำเริ่มต้นรายการที่ 1 จนถึงรายการที่ 10 จากตารางที่ 9 แสดงให้เห็นว่าแบบจำลองที่มีความแม่นยำสูงที่สุด คือ แบบจำลอง Co-Clustering มีค่า Precision เท่ากับ 0.9058 ที่ k=1 รองลงมาจะเป็นแบบจำลอง NMF โดยมีค่า Precision เท่ากับ 0.9013 ที่ k=1 เช่นเดียวกัน ในทางกลับกันกลุ่มของแบบจำลองที่มีความแม่นยำน้อยที่สุดเมื่อเทียบกับแบบจำลองประเภทอื่น คือ KNN Item-based มีค่า Precision ต่ำสุดที่ 0.8431 ที่ k=1 และแบบจำลอง KNN User-based มีค่า Precision ค่อนข้างต่ำเท่ากับ 0.8523 ที่ k=7

ตารางที่ 9 การเปรียบเทียบค่าความแม่นยำของแบบจำลอง 10 อันดับแรก (Precision@k)

Top-N	Baseline	SVD	SVD++	Co-Clustering	NMF	KNN User-based	KNN Item-based
1	0.8740	0.8693	0.8715	0.9058	0.9013	0.8577	0.8431
2	0.8652	0.8609	0.8635	0.9007	0.8975	0.8543	0.8435
3	0.8636	0.8593	0.8617	0.8992	0.8964	0.8533	0.8435
4	0.8630	0.8586	0.8611	0.8986	0.8957	0.8528	0.8434
5	0.8628	0.8582	0.8609	0.8984	0.8955	0.8526	0.8434
6	0.8626	0.8581	0.8607	0.8982	0.8954	0.8524	0.8433
7	0.8626	0.8579	0.8606	0.8981	0.8953	0.8523	0.8433
8	0.8625	0.8579	0.8605	0.8981	0.8953	0.8523	0.8433
9	0.8625	0.8578	0.8605	0.8980	0.8952	0.8523	0.8432
10	0.8624	0.8578	0.8604	0.8980	0.8952	0.8523	0.8432

จากภาพประกอบที่ 16 แสดงให้เห็น กลุ่มของแบบจำลองที่มีค่าความแม่นยำในระดับสูง ได้แก่ แบบจำลอง Co-Clustering และ NMF ที่มีค่า Precision เริ่มต้น (k=1) สูงที่สุดประมาณ 0.90 และลดน้อยลงจนค่อนข้างคงที่เมื่อรายการแนะนำ หรือค่า k เพิ่มขึ้น กลุ่มต่อมาเป็นกลุ่มที่มีค่าความแม่นยำระดับปานกลาง ได้แก่ แบบจำลอง Baseline, SVD++, และ SVD โดยมีค่าเริ่มต้นประมาณ 0.87 และมีแนวโน้มลดลง เมื่อค่า k เพิ่มขึ้น และกลุ่มที่มีค่าความแม่นยำในระดับต่ำ ประกอบด้วย KNN User-based และ KNN Item-based มีค่า Precision เริ่มต้นต่ำกว่าแบบจำลองประเภทอื่นประมาณ 0.84 และเริ่มคงที่เมื่อค่า k=3 เป็นต้นไป



ภาพประกอบที่ 16 กราฟแสดงค่า Precision@k ในแต่ละแบบจำลอง

4.2.2 การเปรียบเทียบความถูกต้องของแบบจำลองในการแนะนำจากรายการแนะนำ 10 อันดับแรก (Recall@k)

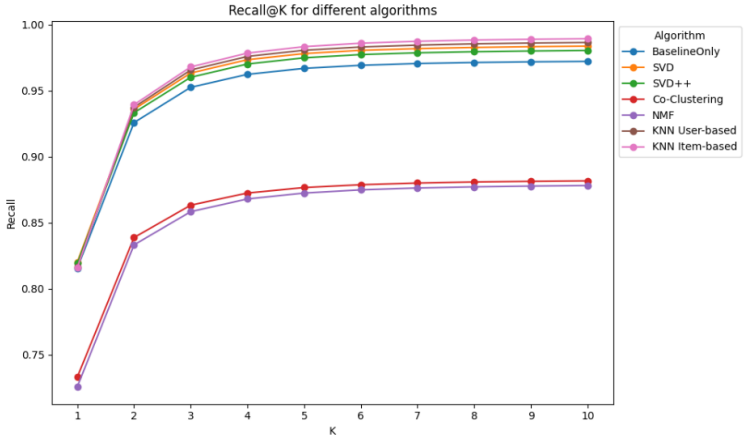
สำหรับการเปรียบเทียบความถูกต้องในการแนะนำของแบบจำลองที่สามารถคาดการณ์รายการแนะนำที่เกี่ยวข้อง และเปรียบเทียบผลลัพธ์ของการแนะนำผ่านรายการแนะนำ 10 อันดับแรก จากตารางที่ 10 พบว่า แบบจำลอง KNN Item-based มีค่า Recall สูงที่สุด เมื่อ k=10 อยู่ที่ 0.9893 รองลงมา คือ แบบจำลอง KNN User-based มีค่า Recall ที่ k=10 เท่ากับ 0.9864 อย่างไรก็ตามเมื่อ k=1 หรือเมื่อมีรายการแนะนำเพียงแค่รายการเดียว แบบจำลอง SVD จะมีค่า Recall สูงที่สุด เท่ากับ 0.8203 และรองลงมาจะเป็นแบบจำลอง SVD++ โดยมีค่า Recall อยู่ที่ 0.8192

ตารางที่ 10 การเปรียบเทียบค่าความถูกต้องของแบบจำลอง 10 อันดับแรก (Recall@k)

Top-N	Baseline	SVD	SVD++	Co-Clustering	NMF	KNN User-based	KNN Item-based
1	0.8157	0.8203	0.8192	0.7332	0.7260	0.8163	0.8161
2	0.9258	0.9359	0.9333	0.8388	0.8332	0.9374	0.9393
3	0.9525	0.9633	0.9601	0.8633	0.8584	0.9656	0.9679
4	0.9623	0.9734	0.9701	0.8725	0.8680	0.9758	0.9784
5	0.9668	0.9780	0.9748	0.8766	0.8725	0.9805	0.9833
6	0.9691	0.9805	0.9772	0.8788	0.8749	0.9830	0.9858
7	0.9705	0.9818	0.9786	0.8800	0.8763	0.9844	0.9873
8	0.9713	0.9827	0.9794	0.8808	0.8772	0.9854	0.9883
9	0.9718	0.9833	0.9800	0.8813	0.8777	0.9860	0.9888
10	0.9721	0.9837	0.9804	0.8817	0.8781	0.9864	0.9893

จากภาพประกอบที่ 17 แสดงให้เห็นการเปรียบเทียบแบบจำลองในด้านของค่าความถูกต้องในการแนะนำของแบบจำลองที่สามารถคาดการณ์รายการแนะนำที่เกี่ยวข้อง โดยแบบจำลอง Co-clustering และ NMF มีค่า Recall เริ่มต้นน้อยที่สุดเมื่อเทียบกับแบบจำลองอื่น โดยที่ k=1 ประมาณ 0.73 และมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง เมื่อค่า k หรือรายการแนะนำเพิ่มมากขึ้น และเริ่มคงที่และถึงจุดอิ่มตัวเร็วกว่าแบบจำลองอื่น อย่างไรก็ตามแบบจำลองอื่นจะมีค่า

Recall เริ่มต้นประมาณ 0.82 สูงกว่าแบบจำลองข้างต้น และเพิ่มขึ้นอย่างต่อเนื่องจนค่อนข้างคงที่หรืออยู่ในจุดอิ่มตัว (Saturation Point)



ภาพประกอบที่ 17 กราฟแสดงค่า Recall@k ในแต่ละแบบจำลอง

4.3 ผลลัพธ์การแนะนำร้านอาหารโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

ผู้วิจัยดำเนินการนำแบบจำลองในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมทั้ง 7 แบบจำลองในงานวิจัยฉบับนี้ ที่ผ่านการปรับค่าพารามิเตอร์ที่เหมาะสมมาดำเนินการคาดการณ์ร้านอาหารโดยอ้างอิงจากข้อมูลพฤติกรรมของผู้ใช้ หรือผู้ใช้งานรายอื่นๆ ที่มีความคล้ายคลึงกับผู้ใช้งาน ณ ขณะนั้น เพื่อแสดงให้เห็นถึงผลลัพธ์ของแบบจำลองในเชิงเปรียบเทียบ จากตัวอย่างในภาพประกอบที่ 18 แสดงถึงตัวอย่างผู้ใช้งานรหัส jCaranTjMIbCU2DbMJ7y5Q พบว่าผู้ใช้งานมีประวัติการให้คะแนน (Rating) ร้านอาหารจำนวน 3 ร้าน ได้แก่ ร้าน TacoSon ให้คะแนน 5 คะแนน, ร้าน Taste Of New York Pizzeria ให้คะแนน 2 คะแนน, และร้าน Miguelitos Taqueria Y Tequilas ให้คะแนน 5 คะแนน ซึ่งแบบจำลองในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมจะดำเนินการคาดการณ์รายการแนะนำตามอัลกอริทึมที่แตกต่างกันของแบบจำลองโดยอ้างอิงจากพฤกรรมของผู้ใช้ที่มีรสนิยม หรือความชื่นชอบใกล้เคียงกัน

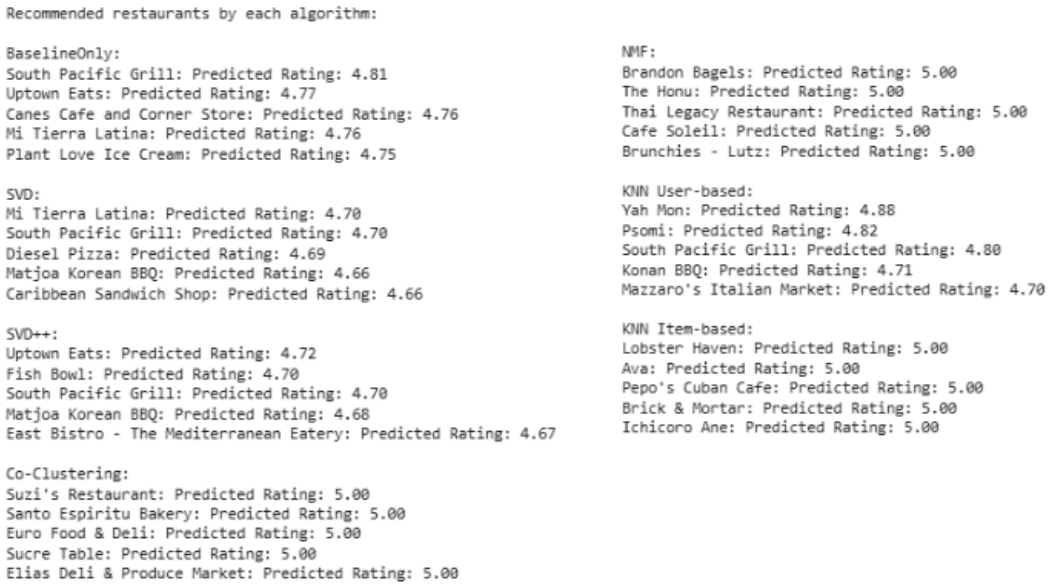
```
User ID: jCaranTjMIbCU2DbMJ7y5Q

User's previous ratings:
                                name  rating
62203                       TacoSon       5
102774     Taste Of New York Pizzeria       2
133644  Miguelitos Taqueria Y Tequilas       5
```

ภาพประกอบที่ 18 ตัวอย่างประวัติการให้คะแนนของผู้ใช้งาน

แบบจำลองดำเนินการคาดการณ์รายการแนะนำให้กับผู้ใช้ดังกล่าว ตามภาพประกอบที่ 19 แสดงให้เห็นถึงร้านอาหารที่แต่ละแบบจำลองแสดงผลลัพธ์รายการแนะนำที่แตกต่างกันออกไป

รวมถึงคะแนนในแต่ละร้านอาหารที่แบบจำลองคาดการณ์ ยกตัวอย่างเช่น ร้าน South Pacific Grill ที่แบบจำลองส่วนใหญ่ให้น้ำหนักและดำเนินการแนะนำให้กับผู้ใช้ในหลายแบบจำลอง ได้แก่ Baseline โดยมีการให้คะแนนเท่ากับ 4.81 คะแนน, SVD เท่ากับ 4.70 คะแนน, SVD++ เท่ากับ 4.70 คะแนน, และ KNN User-based เท่ากับ 4.80 คะแนน ซึ่งแต่ละแบบจำลองมีการให้คะแนนใกล้เคียงกัน มีแนวโน้มว่าเป็นร้านที่ได้รับความนิยมสูงหรือมีคะแนน (Rating) ที่ดีจากผู้ใช้ผู้อื่นที่มีรสนิยม หรือความชื่นชอบที่คล้ายกัน



ภาพประกอบที่ 19 ตัวอย่างผลลัพธ์รายการแนะนำและคะแนน (Rating) จากแบบจำลอง

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ระบบแนะนำร้านอาหาร (Restaurant Recommendation System) ด้วยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม เป็นระบบที่ช่วยในการแนะนำร้านอาหารให้กับผู้ใช้งานโดยอ้างอิงจากพฤติกรรมหรือความชื่นชอบของผู้ใช้รายอื่น โดยอ้างอิงจากข้อมูลการให้คะแนน (Rating) ผู้ใช้งานในอดีต

ผู้วิจัยนำเสนอการวิเคราะห์เชิงเปรียบเทียบแบบจำลองในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมทั้งในด้านของการเปรียบเทียบประสิทธิภาพแบบจำลองทั้งก่อนและหลังการปรับค่าพารามิเตอร์ที่เหมาะสม และการเปรียบเทียบในด้านของประสิทธิภาพของแบบจำลองที่ปรับใช้ค่าพารามิเตอร์ที่เหมาสมในแต่ละประเภท (Hyperparameter Optimization) เพื่อแสดงให้เห็นถึงการทำงานและประสิทธิภาพของแบบจำลองประเภทที่สามารถนำไปประยุกต์ใช้กับชุดข้อมูลจริงได้อย่างเหมาะสม โดยแบ่งการสรุปผล 3 ส่วน ดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

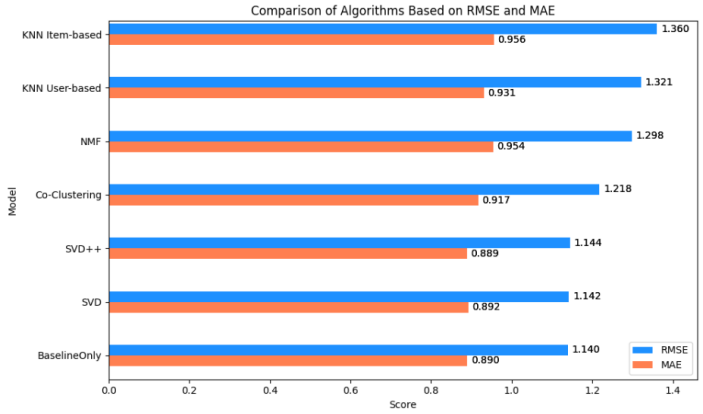
ผู้วิจัยดำเนินการพัฒนาระบบแนะนำร้านอาหารด้วยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ในแบบจำลองที่แตกต่างกัน 7 แบบ ได้แก่ Baseline, KNN User-based, KNN Item-based, Singular Value Decomposition, Singular Value Decomposition++, Non-negative Matrix Factorization, และ Co-Clustering โดยผู้วิจัยได้ดำเนินการปรับค่าพารามิเตอร์ที่เหมาะสมให้แก่แบบจำลองดังกล่าวทั้งหมดโดยอ้างอิงจากค่า RMSE เพื่อให้แบบจำลองสามารถทำงานได้อย่างมีประสิทธิภาพที่สุดบนชุดข้อมูลในงานวิจัย

จากผลการเปรียบเทียบผลลัพธ์จากการดำเนินการปรับค่าพารามิเตอร์ที่เหมาะสมแบบจำลองที่ตอบสนองต่อการปรับปรุงค่าพารามิเตอร์มากที่สุดในด้านของประสิทธิภาพของค่าความคลาดเคลื่อน (Error) ที่ลดลง โดยประเมินจากอัตราการเปลี่ยนแปลงของค่า MAE และ RMSE ได้แก่ แบบจำลอง NMF ซึ่งมีอัตราของค่าความคลาดเคลื่อนดีขึ้นอ้างอิงจากค่า RMSE เท่ากับ 4.29% และหากอ้างอิงจากค่า MAE แบบจำลองที่มีประสิทธิภาพที่สูงขึ้น 9.36%.เนื่องจากการปรับพารามิเตอร์ในส่วนของ n_factors ทำให้แบบจำลองสามารถจับรูปแบบของข้อมูลได้ดี

ยิ่งขึ้น รวมถึงจัดการกับข้อมูลเบาบาง หรือมีค่าเป็นศูนย์จำนวนมาก (Data Sparsity) ได้ดี จึงส่งผลให้แบบจำลองสามารถคาดการณ์รายการแนะนำโดยมีความแม่นยำมากขึ้น ในส่วนของประสิทธิภาพในด้านเวลาการประมวลผลในช่วงการฝึกฝน (Fit Time) แบบจำลอง NMF สามารถประมวลผลได้รวดเร็วขึ้นหลังจากการปรับค่าพารามิเตอร์ที่เหมาะสมที่ 53.40% ซึ่งอาจเป็นเพราะการลดความซับซ้อนของการทำงานของพารามิเตอร์ในส่วนของ n_epochs จึงส่งผลให้แบบจำลองทำงานได้รวดเร็วยิ่งขึ้น แต่ประสิทธิภาพในการประมวลผลไม่ลดน้อยลง สำหรับประสิทธิภาพในด้านเวลาการประมวลผลในช่วงการทดสอบ (Test Time) แบบจำลอง SVD สามารถประมวลผลได้รวดเร็วขึ้นในช่วงทดสอบ 18.05% เนื่องมาจากแบบจำลองมีลักษณะการประมวลผลที่รวดเร็วอยู่แล้วด้วยการทำ Dot Product กับ User-Item Interaction Matrix การปรับค่าพารามิเตอร์ให้อยู่ในจุดที่เหมาะสมในการหาปัจจัยแฝง (Latent Factors) ไม่มากเกินไปจึงส่งผลให้แบบจำลองทำงานในช่วงทดสอบได้เร็วมากขึ้นอีกเช่นเดียวกัน

รวมถึงผู้วิจัยได้ดำเนินการเปรียบเทียบแบบจำลองที่ผ่านการปรับค่าพารามิเตอร์ที่เหมาะสม ด้วยวิธี Grid Search Optimization โดยเป็นการกำหนดช่วงของค่าพารามิเตอร์ทุกค่าที่เป็นไปได้ทั้งหมด และดำเนินการพัฒนาแบบจำลองจากชุดของค่าพารามิเตอร์ที่ถูกกำหนดข้างต้นทุกชุด เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในแต่ละอัลกอริทึมกับชุดข้อมูลในงานวิจัย จากภาพประกอบที่ 20 แสดงให้เห็นว่าแบบจำลอง Baseline, SVD, และ SVD++ สามารถทำงานได้อย่างมีประสิทธิภาพใกล้เคียงกันเมื่อเทียบกับแบบจำลองอื่นไม่ว่าจะเป็นในค่าของ MAE หรือ RMSE ที่แสดงถึงค่าความคลาดเคลื่อนที่ส่งผลความแม่นยำในการคาดการณ์รายการแนะนำ ซึ่งเมื่อดำเนินการคาดการณ์เพื่อสร้างผลลัพธ์ออกมาเป็นรายการแนะนำให้กับผู้ใช้แบบจำลองทั้ง 3 แบบนี้สามารถคาดการณ์ผลลัพธ์ได้ใกล้เคียงกันเช่นเดียวกัน ประกอบด้วยร้านอาหารลักษณะเดียวกันในรายการแนะนำหลายร้านอาหาร ดังตัวอย่างในภาพประกอบที่ 19 ร้านอาหาร South Pacific Grill และ ร้านอาหาร Uptown Eats ล้วนถูกแนะนำด้วยแบบจำลองดังกล่าวทั้งหมด ในทางกลับกันแบบจำลองในกลุ่มของ KNN ประกอบด้วย KNN Item-based และ KNN User-based ที่มีค่า MAE และ RMSE ต่ำที่สุดเมื่อเทียบกับแบบจำลองอื่นและมีผลลัพธ์ในการคาดการณ์รายการแนะนำที่แตกต่างจากแบบจำลองข้างต้น ซึ่งอาจเป็นผลมาจากอัลกอริทึมของแบบจำลองที่อ้างอิงจากการคำนวณค่าความคล้ายคลึง (Similarity) ที่ค่อนข้างอ่อนไหวต่อข้อมูลที่เบาบาง (Sparsity) ยกตัวอย่างเช่น ร้านอาหารบางแห่งมีการให้คะแนนน้อย หรือไม่มีร้านอาหารที่ใกล้เคียงมากพอ รวมถึงอ่อนไหวต่อข้อมูลที่ผิดปกติ (Outliers) ยกตัวอย่างเช่น ร้านอาหารบางร้านได้รับการให้คะแนนสูงจากผู้ใช้งานบางกลุ่ม ดังนั้นแบบจำลองลักษณะนี้จึงมีประสิทธิภาพไม่ดีเท่า

แบบจำลองอย่าง SVD หรือ SVD++ ที่สามารถจัดการกับปัญหาบางอย่างที่แบบจำลองกลุ่ม KNN ไม่สามารถจัดการได้ ยกตัวอย่างเช่น การแยกตัวประกอบเมทริกซ์ เพื่อให้เห็นถึงปัจจัยแฝงอยู่ในชุดข้อมูล ซึ่งส่งผลให้แบบจำลองในกลุ่ม SVD มีประสิทธิภาพมากกว่า

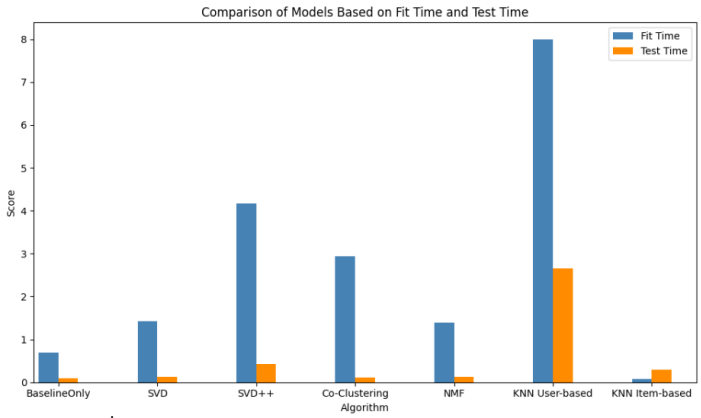


ภาพประกอบที่ 20 การเปรียบเทียบค่าความคลาดเคลื่อนด้วย MAE และ RMSE

เมื่อเปรียบเทียบแบบจำลองที่ผ่านการปรับค่าพารามิเตอร์ที่เหมาะสม ในด้านของเวลาในการประมวลผล จากภาพประกอบที่ 21 แสดงให้เห็นถึงแบบจำลองที่ใช้เวลาประมวลผลทั้งในช่วงเวลาฝึกและและทดสอบนานที่สุดเมื่อเทียบกับแบบจำลองอื่น ได้แก่ KNN User-based เนื่องจากข้อมูลมีผู้ใช้จำนวนมากถึง 88,986 ราย ส่งผลให้แบบจำลองที่มีอัลกอริทึมที่จำเป็นต้องคำนวณค่าความคล้ายคลึงของผู้ใช้งานทุกรายในระหว่างการฝึกฝน (Fit Time) จึงใช้เวลานานกว่าอย่างเห็นได้ชัด รวมถึงแบบจำลองนี้จำเป็นต้องค้นหาเพื่อนบ้าน (Neighbors) ทุกทั้งที่ทำการประมวลผลช่วงทดสอบ จึงทำให้ Test Time สูงขึ้นเช่นเดียวกัน แต่อย่างไรก็ตามประสิทธิภาพของแบบจำลองนี้ไม่ได้ดีเท่าที่ควรแม้ว่าจะใช้เวลาในการประมวลผลค่อนข้างนาน

สำหรับแบบจำลอง SVD++ ที่ใช้เวลาในการประมวลผลช่วงฝึกฝนและทดสอบนานรองลงมาจากแบบจำลอง KNN User-based กลับมีค่า MAE และ RMSE ในระดับที่ดี ซึ่งแสดงให้เห็นเวลาในการประยุกต์ใช้แบบจำลองในบางสถานการณ์จำเป็นต้องแลกเปลี่ยนระหว่างประสิทธิภาพในด้านเวลาและความแม่นยำ

ในทางกลับกันแบบจำลอง SVD เมื่อมีการปรับค่าพารามิเตอร์ ยกตัวอย่างเช่น lr_all ให้มีค่าสูงขึ้นทำให้แบบจำลองมีอัตราการเรียนรู้ (Learning Rate) ที่คงทีมากขึ้น แบบจำลองจึงไม่ต้องใช้เวลาในการประมวลผลมาก ส่งผลให้แบบจำลองสามารถประมวลผลช่วงทดสอบ (Test Time) ที่รวดเร็วมากยิ่งขึ้น



ภาพประกอบที่ 21 การเปรียบเทียบเวลาในการประมวลผล Fit Time และ Test Time

สำหรับผลการเปรียบเทียบด้านความถูกต้องแม่นยำของการแนะนำของแบบจำลอง โดยผู้วิจัยดำเนินการเปรียบเทียบ 10 อันดับรายการแนะนำแรก (k) พบว่า แบบจำลอง KNN User-based และ KNN Item-based เป็นแบบจำลองที่มีค่า Recall สูงที่สุดเมื่อ k=10 กล่าวคือ แบบจำลองดังกล่าวสามารถให้ผลลัพธ์การแนะนำ 10 รายการแรกที่ค่อนข้างครอบคลุมและเกี่ยวข้องกับการรสนิยม หรือความชื่นชอบของผู้ใช้งานในอดีต KNN User-based อ้างอิงจากผู้ใช้งานที่มีพฤติกรรมคล้ายกัน ซึ่งให้ผลลัพธ์รายการแนะนำที่หลากหลาย เนื่องจากแบบจำลองพิจารณาจากพฤติกรรมของผู้ใช้กลุ่มกว้าง และ KNN Item-based แสดงรายการแนะนำที่คล้ายกับสิ่งที่ผู้ใช้งานเคยให้คะแนนสูงมาก่อน โดยแบบจำลองเหล่านี้อาจจะขาดความแม่นยำ (Precision) ในการคาดการณ์คะแนนและการแนะนำร้านอาหารที่ใกล้เคียงกับรสนิยม หรือความชื่นชอบของผู้งานได้ทั้งหมด และอาจมีรายการแนะนำที่ไม่เกี่ยวข้องหลายรายการ

ในส่วนของแบบจำลอง Co-Clustering มีค่า Precision สูงที่สุดที่ k=1 และมีค่าที่ลดลงและเริ่มคงที่เมื่อค่า k เพิ่มขึ้น ซึ่งแบบจำลอง Co-Clustering มีอัลกอริทึมดำเนินการแบ่งข้อมูลออกเป็นกลุ่มย่อย โดยเมื่อดำเนินการกับชุดข้อมูลในงานวิจัย แบบจำลองจะแบ่งออกเป็น กลุ่มผู้ใช้งานและกลุ่มร้านอาหาร ซึ่งการแบ่งกลุ่มจะช่วยคัดกรองเฉพาะข้อมูลที่เกี่ยวข้อง และลดผลกระทบจากการให้คะแนนที่ไม่สอดคล้องกันของผู้ใช้แต่ละคน และผู้ใช้งานที่อยู่ในกลุ่มเดียวกันมักให้คะแนนกับร้านอาหารบางแห่งที่คล้ายกัน ส่งผลให้การแนะนำตรงกับความต้องการของกลุ่มมากขึ้น ด้วยเหตุนี้ส่งผลให้แบบจำลองดังกล่าวมีค่า Precision สูงที่สุด

การแลกเปลี่ยนความถูกต้องในการคาดการณ์รายการแนะนำที่เกี่ยวข้อง (Recall) กับความแม่นยำในสร้างรายการแนะนำให้กับผู้ใช้งานที่ใกล้เคียงกับความชื่นชอบ หรือรสนิยมในอดีต (Precision) เป็นปัจจัยหนึ่งที่ต้องพิจารณาในการแลกเปลี่ยน (Trade-Off) โดยขึ้นอยู่กับจุดประสงค์ของงานที่แตกต่างกัน

โดยสรุปผลลัพธ์จากงานวิจัยในการเปรียบเทียบประสิทธิภาพของแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม สามารถแบ่งได้เป็น 3 กลุ่มหลัก ดังนี้

1. กลุ่มของแบบจำลองที่โดดเด่นด้านความแม่นยำ โดยเฉพาะอย่างยิ่งในการคาดการณ์การให้คะแนนของผู้ใช้งาน (Accuracy-Based Models) ซึ่งอ้างอิงจากค่า MAE และ RMSE ที่ต่ำที่สุด ประกอบด้วย แบบจำลอง BaselineOnly, SVD, และ SVD++ โดยจุดเด่นของแบบจำลองกลุ่มนี้สามารถคาดการณ์ได้แม่นยำ ส่งผลต่อดีการจัดอันดับรายการแนะนำ และเหมาะสำหรับงานที่มีจุดประสงค์ที่เน้นในเรื่องของความแม่นยำของการคาดการณ์การให้คะแนน
2. กลุ่มของแบบจำลองที่โดดเด่นในด้านของความเร็วในการประมวลผล (Processing Time Models) อ้างอิงจากค่า Fit time และ Test time ได้แก่ แบบจำลอง KNN Item-based, BaselineOnly, และ SVD โดยจุดเด่นของกลุ่มนี้เหมาะสำหรับงานที่มีจุดประสงค์ต้องการผลลัพธ์ที่รวดเร็ว หรือทันที (Real-time) และเหมาะสำหรับการทดลองหรือทดสอบแบบจำลอง
3. กลุ่มของแบบจำลองที่โดดเด่นในด้านของการจัดอันดับรายการแนะนำที่ถูกต้องแม่นยำ (Recommendation Performance Models) ซึ่งแบบจำลองที่เด่นในด้านความแม่นยำ (Precision) ในการจัดอันดับรายการแนะนำ ได้แก่ แบบจำลอง Co-Clustering, NMF, และ BaselineOnly ในส่วนของแบบจำลองที่เด่นในด้านความถูกต้อง และครอบคลุม (Recall) ในการจัดอันดับรายการแนะนำ ได้แก่ KNN Item-based, KNN User-based, และ SVD โดยจุดเด่นของแบบจำลองในกลุ่มนี้จะเน้นในด้านของประสิทธิภาพและผลลัพธ์ในการแนะนำ Top-N รวมถึงเหมาะสำหรับงานที่มีจุดประสงค์ในการแนะนำที่หลากหลาย

5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ ผู้วิจัยได้ดำเนินศึกษาเปรียบเทียบระบบแนะนำโดยใช้เทคนิคการรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยมีการปรับค่าพารามิเตอร์ที่เหมาะสม ด้วยวิธี Grid Search Optimization เพื่อความถูกต้องแม่นยำและประสิทธิภาพในการเปรียบเทียบแบบจำลอง ซึ่งจากผลการทดลองแสดงให้เห็นถึงค่าความคลาดเคลื่อน (Error) ของแบบจำลองที่ลดลง ส่งผลต่อความแม่นยำของแบบจำลองที่เพิ่มขึ้น ซึ่งผลลัพธ์ของการปรับค่าพารามเตอร์ที่เหมาะสมมีความคล้ายคลึงกับบทความวิจัยเรื่อง GSO-CRS: grid search optimization for collaborative recommendation system (Behera & Nain, 2022) ที่กล่าวถึงการดำเนินการปรับค่าพารามิเตอร์

ให้กับแบบจำลอง SVD และดำเนินการเปรียบเทียบกับแบบจำลองที่ดำเนินการด้วยวิธีอื่น ซึ่งส่งผลให้มีค่าความแม่นยำที่สูงขึ้น แต่อย่างไรก้ตามผู้วิจัยพบอีกปัจจัยหนึ่งซึ่งอาจต้องแลกเปลี่ยนกับความแม่นยำที่เพิ่มขึ้นมา คือ เวลาในการประมวลผลของแบบจำลองที่เพิ่มสูงขึ้นโดยเฉพาะในช่วงเวลาฝึกฝน (Fit Time)

จากผลการทดลองการปรับค่าพารามิเตอร์ของแบบจำลองมีส่วนช่วยให้แบบจำลองสามารถประมวลผลได้มีประสิทธิภาพมากยิ่งขึ้น แม้ว่าในบางสถานการณ์หรือบางแบบจำลองอาจจะต้องยอมเสียประสิทธิภาพในด้านเวลาในการประมวลผล เพื่อแลกกับประสิทธิภาพในด้านของความแม่นยำของแบบจำลอง

การเปรียบเทียบแบบจำลองในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม โดยผลการเปรียบเทียบ แสดงให้เห็นว่าแบบจำลองอย่าง Singular Value Decomposition (SVD) และ Singular Value Decomposition++ (SVD++) มีความแม่นยำที่ค่อนข้างสูงเมื่อเทียบกับแบบจำลองอื่นอ้างอิงจากค่า MAE และ RMSE ที่มีค่าที่อยู่ในระดับที่ต่ำ แต่อย่างไรก็ตามแบบจำลองดังกล่าวใช้เวลาในการประมวลผลช่วงฝึกฝนและทดสอบนานกว่าหลายแบบจำลอง โดยเฉพาะอย่างยิ่งแบบจำลอง SVD++ เนื่องจากแบบจำลองมีการทำงานแบบแยกเมทริกซ์ ที่สามารถสกัดปัจจัยแฝง (Latent Factors) จากข้อมูลผู้ใช้งานและร้านอาหารได้ ส่งผลให้แบบจำลองสามารถคาดการณ์ค่าคะแนน (Rating) ได้แม่นยำสูงกว่าแบบจำลองอื่นค่อนข้างมาก แต่ใช้เวลาประมวลผลนานเช่นเดียวกัน อย่างไรก็ตามแบบจำลองดังกล่าวมีความซับซ้อนในการทำงาน ยกตัวอย่างเช่น การอัปเดตค่าพารามิเตอร์ซ้ำๆ หลายรอบ (Epochs) เป็นต้น ซึ่งสอดคล้องกับบทความวิจัยเรื่อง A Book Recommender System Using Collaborative Filtering Method (Khalifeh & Almousa, 2021) ที่ชี้ให้เห็นถึงการแลกเปลี่ยนระหว่างข้อดีของแบบจำลองที่ไม่สามารถดำเนินการพร้อมกันได้อย่าง ความรวดเร็วในการประมวลผลและความแม่นยำ

นอกจากนี้การเปรียบเทียบแบบจำลองในเทคนิคการรองแบบพึ่งพาผู้ใช้ร่วมในงานวิจัยแสดงให้เห็นข้อดี และข้อเสียของแต่ละแบบจำลองที่สามารถนำไปประยุกต์ใช้ได้ในระบบแนะนำที่หลากหลายโดยขึ้นอยู่กับจุดประสงค์ในแต่ละงาน รวมถึงแสดงให้เห็นความแตกต่างของอัลกอริทึมในแต่ละแบบจำลองเพื่อให้เห็นถึงการจัดการกับข้อมูล และปัญหาของแบบจำลองได้ เพื่อนำไปสู่การพัฒนาระบบแนะนำร้านอาหารที่มีประสิทธิภาพ

สำหรับผลลัพธ์จากการแนะนำของแบบบจำลองที่ดำเนินงานวิจัย แสดงให้เห็นข้อสังเกตของการแนะนำคือ ในบางแบบจำลองอย่าง Baseline, SVD, และ SVD++ แสดงผลลัพธ์ในการแนะนำที่มีความใกล้เคียงกันทั้งการคาดการณ์คะแนน และรายการแนะนำร้านอาหาร ซึ่งค่อนข้าง

มีประสิทธิภาพหากเปรียบเทียบในการแนะนำระหว่างแบบจำลอง แต่ในทางกลับกันแบบจำลองอื่น ยกตัวอย่างเช่น Co-Clustering, NMF, และ KNN Item-based ผลลัพธ์จากการแนะนำของแบบจำลองเหล่านี้กระจัดกระจาย และมีความหลากหลายค่อนข้างสูง ซึ่งถือเป็นข้อดีในรูปแบบหนึ่งที่แบบจำลองสามารถคาดการณ์ร้านอาหารใหม่ๆ ให้กับผู้ใช้งานได้หลากหลาย และเป็นทางเลือกให้กับผู้ใช้งานได้ แต่มีความเป็นไปได้ที่ร้านอาหารเหล่านี้จะเป็นร้านอาหารที่ได้รับการให้คะแนนหรือความสนใจจากผู้ใช้งานน้อยได้เช่นเดียวกัน รวมถึงกรณีผู้ใช้งานหากมีประวัติการให้คะแนนร้านอาหารร้านต่างๆ ในระดับที่สูง หรือเท่ากับ 5 คะแนน ส่งผลให้เกิดปัญหาในการคาดกาณ์หรือแนะนำร้านอาหารของแบบจำลองที่จะแนะนำร้านอาหารที่เคยได้รับการให้คะแนน 5 คะแนนเพียงอย่างเดียว ทำให้ขาดความหลากหลาย และอาจแนะนำร้านที่ไม่เกี่ยวข้องกับผู้ใช้งาน

5.3 ข้อเสนอแนะ

1. ข้อมูลที่นำมาใช้ในการวิจัยครั้งนี้อาจมีขนาดไม่มากนักและมีความหลากหลายค่อนข้างจำกัด ส่งผลต่อความแม่นยำในการคาดการณ์รายการแนะนำ นอกจากนี้ข้อจำกัดด้านทรัพยากรในการประมวลผลยังทำให้ผู้วิจัยจำเป็นต้องลดขนาดชุดข้อมูลและทำการคัดกรองข้อมูลให้มีความเฉพาะเจาะจงมากยิ่งขึ้น เพื่อให้สามารถพัฒนาและเปรียบเทียบแบบจำลองได้อย่างมีประสิทธิภาพ ในอนาคตหากมีโอกาสต่อยอดงานวิจัยในลักษณะเช่นนี้ การเพิ่มขนาดและความหลากหลายของข้อมูล รวมถึงการใช้ทรัพยากรประมวลผลที่มีประสิทธิภาพมากขึ้น จะสามารถช่วยเพิ่มความหลากหลายในรายการแนะนำ (Diversity) และประสิทธิภาพของแบบจำลองได้อย่างมีนัยสำคัญ

2. แบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) สามารถคาดการณ์รายการแนะนำให้กับผู้ใช้งานได้โดยมีผลลัพธ์ที่แสดงความหลากหลายค่อนข้างสูง อย่างไรก็ตามหากมีการพัฒนาแบบจำลองในเทคนิคดังกล่าวในอนาคต อาจจำเป็นต้องพิจารณาการแลกเปลี่ยนข้อดีข้อเสีย (Trade-Off) ระหว่างความหลากหลายและความแม่นยำของรายการแนะนำ เพื่อให้ได้ผลลัพธ์ที่สอดคล้องกับเป้าหมายของระบบแนะนำในแต่ละบริบทการใช้งาน รวมไปถึงการแลกเปลี่ยนในด้านของเวลาในการประมวลผลและความถูกต้องแม่นยำที่อาจเป็นอีกปัจจัยหนึ่งในการพิจารณาในการพัฒนาแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม เนื่องจากแต่ละแบบจำลองมีความสามารถในการจัดการกับข้อมูล หรือมีอัลกอริทึมที่แตกต่างกัน ยกตัวอย่างเช่น KNN-user-based เป็นหนึ่งในแบบจำลองที่จำเป็นต้องใช้เวลานานในการประมวลผล แต่แบบบจำลองอย่าง SVD ประมวลผลได้รวดเร็วกว่าโดยเฉพาะอย่างยิ่งในช่วงเวลาทดสอบ เนื่องจากอัลกอริทึมของแบบจำลองที่มีการแยกตัวประกอบเมทริกซ์โดยเฉพาะ

3. ในการแนะนำร้านอาหารสำหรับผู้ใช้งานรายใหม่ พบว่าแบบจำลองยังประสบปัญหา Cold Start เนื่องจากข้อมูลผู้ใช้งานมีประวัติการให้คะแนนที่น้อย หรือร้านอาหารที่ไม่ได้รับการให้คะแนน ซึ่งส่งผลต่อความแม่นยำของการคาดการณ์ทำให้บางร้านอาหารไม่ถูกนำถึงมาแนะนำโดยแบบจำลองได้ หากสามารถพัฒนาแบบจำลองเพิ่มเติมในอนาคตอาจจำเป็นต้องมีการนำข้อมูลที่เกี่ยวข้องกับผู้ใช้งานอื่นๆเข้ามาใช้ร่วมด้วย ยกตัวอย่างเช่น อายุ ความชื่นชอบและความสนใจ หรือข้อมูลด้านตำแหน่งที่พักอาศัย

4. ผลลัพธ์ในการแนะนำของแบบจำลองในเทคนิคการกรองแบบพึ่งพาผู้ใช้ร่วม ในบางกรณีอาจมีการแนะนำที่ไม่เกี่ยวข้องเชื่อมโยงกับความชื่นชอบของผู้ใช้งาน ณ ขณะนั้น (Active User) ซึ่งหากมีการพัฒนาแบบจำลองเพิ่มเติม การนำคุณลักษณะ (Features) เพิ่มเติมบางประการเข้ามาร่วมปรับใช้ในแบบจำลอง อาจช่วยเพิ่มประสิทธิภาพความถูกต้องแม่นยำของแบบจำลองได้มากยิ่งขึ้น ยกตัวอย่างเช่น สัญชาติร้านอาหาร ที่ตั้งร้านอาหาร หรือประเภทร้านอาหาร เป็นต้น

5. ในการปรับค่าพารามิเตอร์ที่เหมาะสม (Hyperparameter Optimization) นอกจากวิธีการแบบ Grid Search Optimization ในการพัฒนาแบบจำลองที่มีการปรับค่าพารามิเตอร์ที่เหมาะสม หากมีการต่อยอดงานวิจัยอาจมีการนำวิธีอื่นๆ มาประยุกต์ใช้หรือดำเนินการเปรียบเทียบเพื่อแสดงให้เห็นความแตกต่างการการปรับค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีต่างๆ อย่างชัดเจน และสามารถช่วยให้เกิดการต่อยอดงานวิจัยในอนาคตได้ยิ่งขึ้น

6. การประเมินประสิทธิภาพของแบบจำลองในด้านของประสิทธิภาพในการจัดอันดับรายการแนะนำ ผู้วิจัยดำเนินการใช้ค่า Precision@k และ Recall@k ในการวัดประสิทธิภาพ โดยกำหนดให้มีรายการแนะนำ 10 รายการ หรือค่า k=10 หากมีการพัฒนาต่อยอดงานวิจัยอาจมีการประเมินประสิทธิภาพของค่า k ในค่าที่แตกต่างกัน หรือมีจำนวนที่มากขึ้น เพื่อแสดงให้เห็นผลลัพธ์และประสิทธิภาพของแบบจำลองในการจัดอันดับรายการแนะนำในหลายมิติมากยิ่งขึ้น

7. การวิเคราะห์เปรียบเทียบแบบจำลอง แสดงให้เห็นถึงความสามารถและประสิทธิภาพในการจัดการกับข้อมูลและกาคาดการณ์ผลลัพธ์ที่แตกต่างกันอย่างเห็นได้ชัดในแต่ละแบบจำลอง ซึ่งช่วยให้สามารถประยุกต์ใช้แบบจำลองได้อย่างเหมาะสมในอนาคตกับบริบทของงานหรือชุดข้อมูลที่แตกต่างกันออกไป รวมถึงสามารถนำข้อดีและข้อเสียในแต่ละแบบจำลองไปพัฒนาต่อยอดร่วมกับแบบจำลองหรืออัลกอริทึมอื่นๆ ที่มีศักยภาพและถูกพัฒนาได้ดีขึ้นในปัจจุบัน ยกตัวอย่างเช่น การเรียนรู้แบบเสริมแรง (Reinforcement Learning)

ในส่วนของการพัฒนาแบบจำลอง จากผลลัพธ์ที่แสดงให้เห็นภาพรวมของแบบจำลองในกลุ่มของ Singular Value Decomposition มีประสิทธิภาพภาพรวมที่ค่อนข้างสูง ดังนั้นในอนาคตหากมีการพัฒนาต่อยอดเพื่อเพิ่มประสิทธิภาพในการคาดการณ์ให้มีความถูกต้องแม่นยำมากยิ่งขึ้น อาจมีการนำไปประยุกต์ใช้ร่วมกับ Deep Learning เพื่อเพิ่มศักยภาพของระบบแนะนำในด้านการเรียนรู้คุณลักษณะที่ซับซ้อนและสามารถจับความสัมพันธ์เชิงลึกระหว่างผู้ใช้งานและร้านอาหารได้ดียิ่งขึ้น

บรรณานุกรม

Ananth, S., & Raghuveer, K. (2020). A movie and book Recommender System using Surprise Recommendation. Social Science Research Network. https://doi.org/10.2139/ssrn.3551032

Anwar, T., Uma, V., Hussain, M. I., & Pantula, M. (2022). Collaborative filtering and kNN based recommendation to overcome cold start and sparsity issues: A comparative analysis. Multimedia Tools and Applications, 81(25), 35693–35711. https://doi.org/10.1007/s11042-021-11883-z

Behera, G., & Nain, N. (2022). GSO-CRS: grid search optimization for collaborative recommendation system. Sadhana, 47(3). https://doi.org/10.1007/s12046-022-01924-0

Behera, G., & Nain, N. (2023). Collaborative filtering with temporal features for movie recommendation system. Procedia Computer Science, 218, 1366–1373. https://doi.org/10.1016/j.procs.2023.01.115

Fakhri, A. A., Baizal, Z. K. A., & Setiawan, E. B. (2019). Restaurant recommender system using user-based collaborative filtering approach: A case study at Bandung Raya region. Journal of Physics. Conference Series, 1192, 012023. https://doi.org/10.1088/1742-6596/1192/1/012023

Gupta, M., Thakkar, A., Aashish, Gupta, V., & Rathore, D. P. S. (2020). Movie recommender system using collaborative filtering. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).

Khalifeh, S., & A. Al-Mousa, A. (2021). A book recommender system using collaborative filtering method. International Conference on Data Science, E-Learning and Information Systems 2021.

Munaji, A. A., & Emanuel, A. W. R. (2021). Restaurant recommendation system based on user ratings with collaborative filtering. IOP Conference Series. Materials Science and Engineering, 1077(1), 012026. https://doi.org/10.1088/1757-899x/1077/1/012026

Patoulia, A. A., Kiourtis, A., Mavrogiorgou, A., & Kyriazis, D. (2022). A comparative study of collaborative filtering in product recommendation. Emerging Science Journal, 7(1), 1–15. https://doi.org/10.28991/esj-2023-07-01-01

Tripathi, A., & Sharma, A. K. (2020). Recommending restaurants: A collaborative filtering approach. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 42, 1165–1169.

Yelp. (2024.). Yelp Open Dataset. Yelp. https://www.yelp.com/dataset

ศุภกร กรบุญไตรทศ. (2024). แนวโน้มธุรกิจ/อุตสาหกรรม ปี 2567-2569: ธุรกิจร้านอาหารและเครื่องดื่ม. สืบค้นเมื่อ 19 กุมภาพันธ์ 2568, จาก https://www.krungsri.com/th/research/industry/industry-outlook/services/food-beverages/io/io-food-beverage-restaurant-2024-2026