การศึกษาการวัดความคล้ายคลึงในกระบวนการกรองแบบร่วมมือบนพื้นฐานความจำสำหรับระบบแนะนำ

Studying Similarity Measures in Memory-based Collaborative Filtering Technique for Recommendation Systems

กิตติศักดิ์ อ่อนเอื้อน[1], สุนันฑา สดสี[2], พยุง มีสัจ[3]
Kittisak Onuean[1], Sunantha Sodsee[2], Phayung Meesad[3]
Received: 17 March 2017 ; Accepted: 15 September 2017

บทคัดย่อ

ปัจจุบันระบบให้คำแนะนำเป็นระบบที่มีประโยชน์ และสร้างมูลค่าให้กับธุรกิจ ช่วยให้คำแนะนำที่เหมาะสม และตรงตามความต้องการสำหรับผู้ใช้ในยุคของภาวะข้อมูลท่วมท้น โดยมีกิจกรรมหลักคือการเรียงลำดับรายการสินค้าเพื่อแนะนำให้กับผู้ใช้ในแต่ละบุคคลได้อย่างเหมาะสม ซึ่งเทคนิคที่นิยมใช้ในการสร้างระบบให้คำแนะนำคือ เทคนิคตัวกรองแบบร่วมมือ ประกอบด้วยขั้นตอนหลัก 3 ขั้นตอน ได้แก่ 1) การหาค่าความคล้ายคลึง 2) การหาเพื่อนบ้านใกล้เคียง และ 3) การพยากรณ์และให้คำแนะนำ โดยกระบวนการสร้างระบบยังคงพบปัญหาในการสร้างได้แก่ ความเบาบางของข้อมูล ลักษณะของข้อมูลที่เหมือนกัน และขนาดมิติของข้อมูล เป็นต้น ดังนั้นบทความนี้ได้ศึกษาเกี่ยวกับเทคนิคการหาค่าความคล้ายคลึงเพื่อใช้สร้างระบบให้คำแนะนำแบบการกรองแบบร่วมมือบนพื้นฐานความจำซึ่งมีการพัฒนาและปรับปรุงกระบวนการหาค่าความคล้ายคลึง ได้แก่ 1) Pearson's Correlation (COR) 2) Cosine (COS) 3) Adjusted Cosine (ACOS) for similarity between items 4) Distance-based similarity 5) Constrained Pearson's Correlation (CPC) 6) Spearman's Rank Correlation (SRC) 7) Jaccard (Jacc) และ 8) Mean squared difference (MSD) และส่วนของนำกระบวนการมาปรับปรุงเพื่อสร้างกระบวนการใหม่ ได้แก่ 1) Proximity-Impact-Popularity (PIP) 2) Bhattacharyya coefficient 3) Linkelihood Ratio Similarity (LiRa) และ 4) Fuses user and item information (FUIR) เพื่อส่งผลต่อประสิทธิภาพของการสร้างระบบให้คำแนะนำ

คำสำคัญ: ระบบให้คำแนะนำ เทคนิคการกรองแบบร่วมมือ ความคล้ายคลึง

Abstract

As generally known, recommender systems have been applied to increase benefits and add values in many business fields. The systems can recommend appropriate items for users in an information overload era. In application, the systems provide personalized suggestions to fulfill consumer needs. This technique is classified as a collaborative filtering technique and is commonly used in building recommender systems that can be divided into three main steps: 1) finding the similarity, 2) selecting neighbors, and 3) predicting and recommending. However, there are some problems in system performance, such as sparsity, synonymy and scalability. In this article, we focused studies on similarity techniques for further developing recommender systems by using a memory-based collaborative filtering technique comprised of 1) Pearson's Correlation (COR) 2) Cosine (COS) 3) Adjusted Cosine (ACOS) for similarity between items 4) Distance-based similarity 5) Constrained Pearson's Correlation (CPC) 6) Spearman's Rank

[1] *นักศึกษาปริญญาเอก, [2]ผู้ช่วยศาสตราจารย์, [3]รองศาสตราจารย์, คณะเทคโนยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร 10800*

[1] *Doctoral student, [2]Assistant Professor, [3]Associate Professor, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bankok 10800. E-mail: kittisak@buu.ac.th*

Correlation (SRC) 7) Jaccard (Jacc) and 8) Mean squared difference (MSD). Moreover, four algorithms were improved based on traditional process; 1) Proximity-Impact-Popularity (PIP) 2) Bhattacharyya coefficient 3) Linkelihood Ratio Similarity (LiRa) and 4) Fuses user and item information (FUIR) were also evaluated performances.

Keywords: Recommender System Collaborative Filtering Technique Similarity

บทนำ

ระบบให้คำแนะนำ เป็นระบบที่ใช้ในการเก็บรวมสารสนเทศจากความพึงพอใจของผู้ใช้ในตัวสินค้าและบริการที่สนใจ เพื่อทำนายสิ่งที่ผู้ใช้สนใจในตัวสินค้าหรือบริการจากข้อมูลที่มีความสัมพันธ์กัน ซึ่งปัจจุบันได้มีการนำเทคนิคต่างๆ มาใช้ในการสร้างระบบให้คำแนะนำในสินค้าต่างๆ เช่น ภาพยนตร์ เพลง หนังสือ และโปรแกรม เป็นต้น ซึ่งระบบให้คำแนะนำจะนำข้อมูลการให้อันดับของผู้ใช้ต่อสินค้า หรือพฤติกรรมของผู้ใช้ มาใช้ในการคำนวณหาความสัมพันธ์จากสารสนเทศเพื่อแนะนำสินค้าให้กับผู้ใช้ใหม่ที่เข้ามาใช้ระบบ[1] ซึ่งปัจจุบันหน่วยงานต่างๆ มีการนำระบบให้คำแนะนำมาใช้เพื่อสร้างความได้เปรียบทางการแข่งขัน สนับสนุนการเลือกสิ่งที่เหมาะสมให้กับผู้ใช้[2] ซึ่งมีการพัฒนาระบบให้คำแนะนำในรูปแบบต่างๆ เช่น ระบบให้คำแนะนำรัฐบาลอิเล็กทรอนิกส์ ระบบให้คำแนะนำธุรกิจอิเล็กทรอนิกส์ ระบบให้คำแนะนำพาณิชย์อิเล็กทรอนิกส์ เป็นต้น โดยในการสร้างระบบให้คำแนะนำในลักษณะต่างๆ นั้นได้มีการนำเทคนิคมาใช้ในการสร้างระบบให้คำแนะนำโดยประกอบด้วยประเภทต่างๆ ได้แก่ 1) Content-based (CB) recommendation techniques เป็นเทคนิคที่นำค่าความคล้ายของสินค้าที่ผู้ใช้เลือกก่อนหน้ามาแนะนำ 2) Collaborative filtering (CF)-based recommendation techniques เป็นการแบ่งข้อมูลออกเป็น ผู้ใช้ และสินค้า โดยพิจารณาบนพื้นฐานของข้อมูลผู้ที่เข้ามาใช้ 3) Knowledge-based (KB) recommendation techniques เป็นระบบแนะนำสินค้าให้กับผู้ใช้โดยใช้พื้นฐานองค์ความรู้ หรือข้อมูลที่มีความสัมพันธ์เกี่ยวกับผู้ใช้ที่เข้ามาใช้ในการสร้างระบบแนะนำ 4) Hybrid recommendation techniques เป็นการนำเทคนิคมาผสมผสานในการสร้างระบบ โดยอาจมีการผสมผสานเทคนิคต่างๆ มาร่วมกันสร้างระบบให้คำแนะนำ 5) Computational intelligence-based recommendation techniques เป็นเทคนิคการสร้างระบบให้คำแนะนำจากการใช้ปัญญาประดิษฐ์ โดยมีการใช้เทคนิคประกอบด้วยเทคนิคปัญญาประดิษฐ์แบบต่างๆ มาช่วยในการสร้างระบบให้คำแนะนำ 6) Social network-based recommendation technique เป็นเทคนิคที่ใช้ข้อมูลเครือข่ายสังคมมาเป็นเครื่องมือในการสร้างระบบให้คำแนะนำซึ่งจะดูถึงความสัมพันธ์ในลักษณะของเครือข่ายสังคมมาใช้ประกอบการตัดสินใจให้คำแนะนำกับผู้ใช้ 7) Context awareness-Based recommendation techniques เป็นเทคนิคในการสร้างระบบให้คำแนะนำโดยการนำข้อมูล เช่น ข้อมูลเวลา ข้อมูลสถานที่ เป็นต้น มาใช้ในการสร้าง ซึ่งจะมีลักษณะที่มีข้อมูลมากกว่า 2 มิติ ใช้ในการสร้างระบบให้คำแนะนำ 8) Group recommendation techniques เป็นเทคนิคการสร้างระบบให้คำแนะนำสินค้าสำหรับกลุ่มผู้ที่สนใจเฉพาะกลุ่ม

จากการสร้างระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือนั้น ในการดำเนินการจัดทำระบบมีกระบวนการที่สำคัญคือ การหาค่าความคล้ายคลึงของผู้ใช้ หรือสินค้า เพื่อนำค่าความคล้ายคลึงที่ได้ไปดำเนินการหากลุ่มผู้ใช้ที่ต้องการนำข้อมูลมาจัดทำการพยากรณ์ (Top-N) ซึ่งในการหาค่าความคล้ายคลึงนั้นหากมิติข้อมูลที่นำมาใช้ในการคำนวณมีความเบาบางของข้อมูลสูง จะส่งผลต่อประสิทธิภาพการคำนวณค่าความคล้ายคลึง และประสิทธิภาพของระบบให้คำแนะนำ โดยปัจจุบันมีนักวิจัยทำการพัฒนา และปรับปรุงวิธีการหาค่าความคล้ายคลึงแบบต่างๆ เพื่อรองรับกับข้อมูลที่มีความเบาบาง ในการคำนวณหาค่าความคล้ายคลึงของข้อมูลเพื่อให้ได้ระบบให้คำแนะนำที่มีประสิทธิภาพมากขึ้น

ดังนั้นบทความนี้เป็นบทความที่ศึกษาเกี่ยวกับวิธีการหาค่าความคล้ายคลึงของผู้ใช้ และสินค้าเพื่อใช้ในการสร้างระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ โดยประกอบด้วยกระบวนการหลักได้แก่ การหาค่าความคล้ายคลึง การพยากรณ์และให้คำแนะนำ และการทดสอบประสิทธิภาพของระบบให้คำแนะนำ

ปัญหาและความท้าทายในการสร้างระบบให้คำแนะนำแบบใช้เทคนิคการกรองแบบร่วมมือ

ในการสร้างระบบให้คำแนะนำโดยใช้เทคนิคการกรองแบบร่วมมือปัจจุบันมีปัญหาที่เกิดขึ้น และส่งผลต่อประสิทธิภาพของการสร้างระบบให้คำแนะนำ[3] โดยปัญหาต่างๆ เกิดขึ้นจากข้อมูลที่นำเข้ามาเพื่อทำการสร้างระบบ และทำการพยากรณ์อันดับความสนใจให้กับผู้ใช้ ดังนี้

1. ปัญหาความเบาบางของข้อมูล (Data Sparsity)

ปัญหาด้านความเบาบางของข้อมูลเป็นปัญหาที่ผู้จัดทำระบบให้คำแนะนำพบเป็นประจำในการสร้างระบบ ซึ่งปัญหานี้เกิดจากการสร้างระบบให้คำแนะนำที่มีจำนวนของ

สินค้า หรือสิ่งที่สนใจในปริมาณมาก ทำให้เกิดมิติของข้อมูลที่มีขนาดใหญ่แปรผันตามข้อมูลของสินค้า ซึ่งจะส่งผลต่อประสิทธิภาพการสร้างระบบให้คำแนะนำแบบการกรองแบบร่วมมือ[4]

ในการหาค่าความคล้ายคลึงของผู้ใช้ หรือสินค้าในการพยากรณ์อันดับความสนใจต่อตัวสินค้าที่เข้ามาใหม่ โดยปัญหาความเบาบางของข้อมูลเป็นปัญหาที่ได้รับความนิยมในการจัดทำงานวิจัย เพื่อแก้ปัญหาซึ่งเกิดขึ้นได้จากสถานการ เช่น ปัญหาการมีสินค้าใหม่ หรือผู้ใช้งานใหม่ (Cold start problem) โดยเกิดจากการมีสินค้าใหม่เข้ามาในระบบให้คำแนะนำ ซึ่งสินค้า หรือผู้ใช้ นั้นยังมิได้มีข้อมูลการจัดอันดับความสนใจ ทำให้ขาดข้อมูลที่นำมาสร้างระบบให้คำแนะนำ[5] แสดงได้ดัง Figure 1

User	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	…
U1	1		2	1		2	1	5			
U2	4	3	5	4	2	5	4		1		
U3	5		4	5		4	5	1		2	
U4		2			3			5	4	5	
U5		2			3				4	3	
…	…	…	…	…	…	…	…	…	…	…	

Figure 1 An Example of Data Sparsity Table

Figure 1 แสดงตัวอย่างข้อมูลที่เกิดปัญหาความเบาบางของข้อมูลนั้นจะแสดงให้เห็นได้ว่าหากมีการหาค่าความคล้ายคลึง ของผู้ใช้ U1 กับ U2 จะได้ค่าอันดับความสนใจของแต่ละคู่ดังนี้ U1(1,2,1,2,1) และ U2(4,5,4,5,4) ซึ่งหมายถึง ผู้ใช้ที่ U1 และ U2 มีการให้ค่าอันดับความสนใจของผู้ใช้ทั้งสองที่เหมือนกันในรายการสินค้าที่ 1 3 4 6 และ 7 โดยผู้ใช้คนที่ 1 ให้ค่าอันดับความสนใจกับสินค้าเท่ากับ 1 2 1 2 และ 1 ตามลำดับ และ U1 กับ U3 ได้ U1(1,2,1,2,1,5) และ U3(5,4,5,4,5,1) และ U1 กับ U4 ได้ U1(5) และ U4(5) ส่วน U4 กับ U5 ได้ U4(2,3,4,5) และ U5(2,3,4,3) โดยจากค่าอันดับความสนใจที่ผู้ใช้ให้อันดับร่วมกัน หากคำนวณโดยใช้การหาค่าความคล้ายคลึง จะได้ ผู้ใช้ที่ U1 มีค่าความคล้ายคลึงกับผู้ใช้ U4 มากกว่า ผู้ใช้ U5 โดยสามารถทำการคำนวณค่าความเบาบางของข้อมูล[6] ได้โดยใช้สมการ

$$S_g = 1 - \frac{u_r}{u \times i} \quad (1)$$

คือระดับ sparsity ให้ u_r คือส่วนของจำนวนที่ผู้ใช้ให้อันดับความสนใจต่อสินค้า u คือจำนวนผู้ใช้ และ i คือจำนวนสินค้า ดังนั้นจาก (Figure 1) สามารถหาค่าความเบาบางของข้อมูลคือ S_g= 0.4 จากสมการที่ (1) โดยหากค่า S_g มีค่าน้อยแสดงว่าชุดข้อมูลจะมีค่าความเบาบางน้อยทำให้มีประสิทธิภาพในการพยากรณ์มากขึ้น

2. ปัญหามิติของข้อมูลที่นำมาสร้างระบบให้คำแนะนำมีขนาดใหญ่ (Scalability)

ปัญหามิติของข้อมูลที่นำมาสร้างระบบให้คำแนะนำมีขนาดใหญ่ อันเนื่องมาจากจำนวนของผู้ใช้ และสินค้ามีปริมาณมาก ส่งผลทำให้ต้องใช้ทรัพยากรณ์ในการสร้างระบบให้คำแนะนำที่มีความจำเป็นต้องการใช้ทรัพยากรณ์ในการประมวลผลมาก ซึ่งในการจัดทำงานวิจัยต้องมีการลดมิติของข้อมูลลงเพื่อนำข้อมูลที่มีความสำคัญมาใช้ในการพยากรณ์ ซึ่งมีผู้ทำการแก้ไขปัญหาโดยการลดมิติของข้อมูลลงโดยมีการจัดกลุ่มของมิติข้อมูลที่นำมาทำการคำนวณหาค่าความคล้ายคลึงตามกลุ่มของสินค้า[7] โดยมิติของข้อมูลจะมีขนาดเท่ากับ $u \times i$ ซึ่งหากมีจำนวนผู้ใช้ในปริมาณเพิ่มขึ้น และในทิศทางเดียวกันหากมีสินค้าในปริมาณเพิ่มขึ้นจะทำให้มิติของข้อมูลมีขนาดเพิ่มขึ้น ซึ่งจะทำให้ใช้ทรัพยากรณ์ในการประมวลผลหาค่าความคล้ายคลึงสูงขึ้นตาม แสดงได้ Figure 2

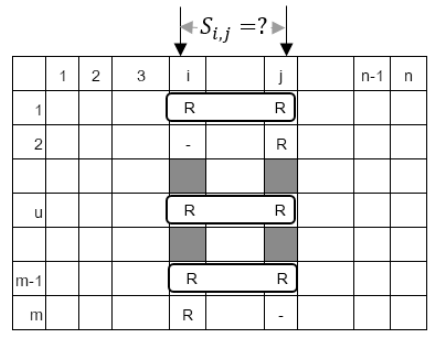

Figure 2 The Similarity Computation of Collaborative Filtering Process

Figure 2 แสดงถึงกระบวนการของระบบให้คำแนะนำใช้เทคนิคการกรอบแบบร่วมมือ[8] ซึ่งในกระบวนการจัดทำต้องมีการหาค่าความคล้ายคลึงของผู้ใช้ หรือสินค้า เพื่อนำมาใช้ในการเลือกชุดข้อมูลมาทำการพยากรณ์สำหรับผู้ใช้ที่เข้ามาใช้ในระบบรายใหม่ ซึ่งในการหาค่า $S_{i,j}$ ได้จากการคำนวณหาค่าความคล้ายคลึงจากการให้อันดับความสนใจระหว่างสินค้าที่ i และสินค้าที่ j โดยมีการคำนวณจากผู้ใช้ตั้งแต่คนที่ 1 ถึงคนที่ m และจากสินค้าที่ 1 ถึงสินค้าที่ n ดังนั้นแสดงให้เห็นได้ว่าหากมีจำนวนผู้ใช้ และสินค้าในปริมาณมากจะส่งผลต่อมิติของข้อมูลที่มีขนาดใหญ่ขึ้น และจะส่งผลต่อประสิทธิภาพของการใช้ทรัพยากรณ์ในการหาค่าความคล้ายคลึง

3. ปัญหาลักษณะของข้อมูลเหมือนกัน (Synonymy)

ปัญหาลักษณะของข้อมูลเหมือนกัน[3] เป็นปัญหาที่เกิดขึ้นจากการสร้างระบบให้คำแนะนำโดยข้อมูลอันดับความสนใจของผู้ใช้ต่อสินค้าแต่ละผู้ใช้เหมือนกัน หรือมีความคล้ายคลึงกันระหว่างสินค้าที่มีความแตกต่างกัน ทำให้ไม่สามารถค้นหาสารสนเทศที่แฝงอยู่ในชุดข้อมูลที่นำมาสร้าง

ระบบให้คำแนะนำ เช่น ในลักษณะของสินค้าในกลุ่มต่างๆ โดยสินค้านั้นอยู่ในกลุ่มหนังสำหรับเด็ก และอยู่ในกลุ่มภาพยนต์สำหรับเด็ก ซึ่งเป็นสินค้าชนิดเดียวกัน แต่อยู่ในกลุ่มที่แตกต่างกัน หรือชื่อมีความแตกต่างกัน ทำให้ระบบนำค่าอันดับความสนใจมาทำการหาค่าความคล้ายคลึงกัน ซึ่งเป็นอันดับความสนใจที่มีค่าเหมือนกันทำให้ส่งผลต่อประสิทธิภาพของการสร้างระบบให้คำแนะนำ

4. ปัญหาการความไม่แน่นอนของผู้ใช้ (Gray Sheep)

ปัญหาความไม่นอนของผู้ใช้[9] เป็นปัญหาที่เกิดจากการตัดสินใจของผู้ใช้ที่ไม่คงเส้นคงวา ซึ่งยากจะจัดผู้ใช้นั้นอยู่ในกลุ่มประชากรไหน และไม่เป็นประโยชน์ต่อการคัดกรองแบบการกรองแบบร่วมมือ ซึ่ง Black sheep เป็นกลุ่มคนที่มีความให้อันดับความพึงพอใจตามลักษณะรสนิยมของผู้ใช้นั้นๆ ซึ่งอาจส่งผลให้การสร้างระบบให้คำแนะนำล้มเหลวได้ ซึ่งมีนักวิจัยได้นำเทคนิคแบบ Content-based มาใช้ร่วมกับ การจัดทำระบบให้คำแนะนำแบบการกรองแบบร่วมมือเพื่อแก้ปัญหาดังกล่าว

5. ปัญหาการเข้าข้างตนเองและโจมตีฝั่งตรงข้าม (Shilling Attacks)

ปัญหาที่เกิดจากการให้อันดับความสนใจในรายการสินค้าของที่ตนเองเป็นเจ้าของสูง แต่ให้อันดับความสนใจในรายการสินค้าของคู่แข่งต่ำมาก ซึ่งจะมีผลต่อค่าอันดับความสนใจที่จะนำมาใช้ในการคำนวณค่าความคล้ายคลึง โดยกระบวนการเข้าข้างตนเองและโจมตีฝั่งตรงข้าม[10] มีลักษณะอยู่ 2 ลักษณะ ได้แก่ การเพิ่มค่าอันดับความสนใจสูงสุด เพื่อให้เกิดการแนะนำสินค้าที่ต้องการ (a push attack) และการโจมตีแบบให้อันดับความสนใจเพื่อไม่ได้ระบบเลือกขึ้นมาให้คำแนะนำ (a nuke attack) แสดงตัวอย่างการโจมตีดัง Figure 3

	Item1	Item2	Item3	Item4	Item5	Item6	Correlation with Alice
Alice	5	2	3	3		?	
User1	2		4		4	1	-1.00
User2	3	1	3		1	2	0.76
User3	4	2	3	1		1	0.72
User4	3	3	2	1	3	1	0.21
User5		3		1	2		-1.00
User6	4	3		3	3	2	0.94
User7		5		1	5	1	-1.00
Attack1	**5**		**3**		**2**	**5**	1.00
Attack2	**5**	**1**	**4**		**2**	**5**	0.89
Attack3	**5**	**2**	**2**	**2**		**5**	0.93
Correlation with Item6	0.85	-0.55	0.00	0.48	-0.59		

Figure 3 An Example of a Push Attack Favoring the Target Item[10]

Figure 3 แสดงให้เห็นถึงความสำคัญของค่าที่ได้จากการหาค่าความสัมพันธ์ของผู้ใช้ที่เข้ามาใช้ในระบบใหม่ (Alice) กับผู้ใช้งานเดิมที่อยู่ในระบบโดยแสดงในคอลัมน์ Correlation with Alice ซึ่งผู้ใช้ที่มีความสัมพันธ์มากที่สุดจะมีค่าเท่ากับ 1 ซึ่งกระบวนการโจมตีเพื่อให้เกิดการแนะนำสินค้าที่ต้องการนี้มีการโจมตีโดยการเพิ่มข้อมูลการโจมตีใน Attack1 ถึง Attack3 เพื่อทำให้ค่าความคล้ายคลึงของผู้ใช้ (Attack1) มีความคล้ายคลึงมากที่สุดกับผู้ใช้ที่เข้ามาใหม่ในการพยากรณ์อันดับความสนใจของสินค้าที่ 6

เทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ (Memory-Based Collaborative Filtering Technique)

การสร้างระบบให้คำแนะนำในปัจจุบันนิยมใช้ เทคนิคการกรองแบบร่วมมือ (Collaborative filtering techniques : CF)[3] ซึ่งเป็นการสร้างระบบให้คำแนะนำโดยการนำข้อมูลการให้อันดับความสนใจของผู้ใช้ต่อสินค้ามาใช้ในการสร้างระบบให้คำแนะนำ สำหรับพยากรณ์สินค้า หรือสิ่งที่ผู้ใช้สนใจให้กับผู้ใช้รายใหม่ โดยมีขั้นตอนในการจัดทำโดยการใช้เทคนิคนี้ดังนี้ 1) กำหนดจำนวนผู้ใช้ที่ใกล้เคียงกับผู้ใช้ที่กำลังใช้งาน 2) รวมการจัดอันดับความสนใจ (Rating) สำหรับสิ่งของที่ใกล้เคียง 3) พยากรณ์จากขั้นตอนที่ 2 และเลือกสินค้ามาให้คำแนะนำ โดยมีเทคนิคหลักดังนี้

เทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ (Memory-based Collaborative Filtering Technique) เป็นเทคนิคที่ใช้ข้อมูลจากประวัติการให้อันดับความสนใจของผู้ใช้ต่อสินค้า และทำให้ได้ตารางมิติอันดับความสนใจของผู้ใช้ต่อสินค้า จากนั้นทำการหาค่าความคล้ายคลึง ระหว่างผู้ใช้ใหม่กับผู้ใช้ที่อยู่ในระบบ หรือหาค่าความคล้ายคลึงสำหรับสินค้าใหม่กับสินค้าในระบบ และทำการเลือกผู้ใช้หรือสินค้าที่มีค่าความคล้ายคลึงสูงสุดมาทำการพยากรณ์โดยจะได้มาจำนวน N ชุดข้อมูล และให้คำแนะนำ ซึ่งปัญหาในเรื่องความเบาบางของข้อมูลจะเป็นข้อจำกัดในการจัดทำโดยการใช้วิธีการนี้

เทคนิคการกรองแบบร่วมมือบนพื้นฐานแบบจำลอง (Model-Based Collaborative Filtering Technique) เป็นเทคนิคที่ใช้วิธีการเรียนรู้ของเครื่องมาใช้ในการสร้างระบบให้คำแนะนำ โดยนำมาใช้เมื่อมีผู้ใช้ใหม่เข้ามาซึ่งไม่เคยมีผู้ใช้ใดในระบบมาก่อน ซึ่งมีการใช้อัลกอริธึมมาใช้ในการสร้างระบบให้คำแนะนำประเภทนี้ได้แก่ การจำแนกข้อมูลแบบเบย์ (Bayesian classifiers) โครงข่ายประสาทเทียม (Neural network) ระบบฟัซซี่หรือระบบแบบคลุมเครือ (Fuzzy system) ขั้นตอนวิธีเชิงพันธุกรรม (Genetic algorithms) การหา

คุณลักษณะแฝง (Latent features) และ การแยกตัวประกอบเมตริกซ์ (Matrix factorization) เป็นต้น

เทคนิคการกรองแบบร่วมมือแบบผสม (Hybrid Collaborative Filtering Technique) เป็นการสร้างระบบให้คำแนะนำโดยมีการทำงานร่วมกันระหว่างวิธีการกรองแบบร่วมมือ (CF) กับตัวกรองทางประชากรศาสตร์ (Demographic filtering) หรือวิธีการกรองแบบร่วมมือ (CF) กับตัวกรองเนื้อหา (Content-based filtering) มาใช้ในการสร้างระบบให้คำแนะนำเพื่อประสิทธิภาพที่ดีขึ้น

จากเทคนิคดังที่กล่าวมาข้างต้นในบทความนี้จะอธิบายถึงเทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำสำหรับสร้างระบบให้คำแนะนำ โดยอธิบายถึงกระบวนการและงานวิจัยที่เกี่ยวข้องในการจัดทำระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำได้ดังนี้

1. กระบวนการจัดทำระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ

ในกระบวนการจัดทำระบบให้คำแนะนำแบบการกรองแบบร่วมมือบนพื้นฐานความจำ[11] มีกระบวนการหลักในการจัดทำระบบให้คำแนะนำแบ่งออกเป็น 3 ขั้นตอนได้แก่ 1) การหาค่าความคล้ายคลึงระหว่างผู้ใช้ 2) เลือกผู้ใช้ที่มีความใกล้เคียงมาใช้ในการพยากรณ์ และ 3) การพยากรณ์และให้คำแนะนำ แสดงกระบวนการทำงานดัง Figure 4

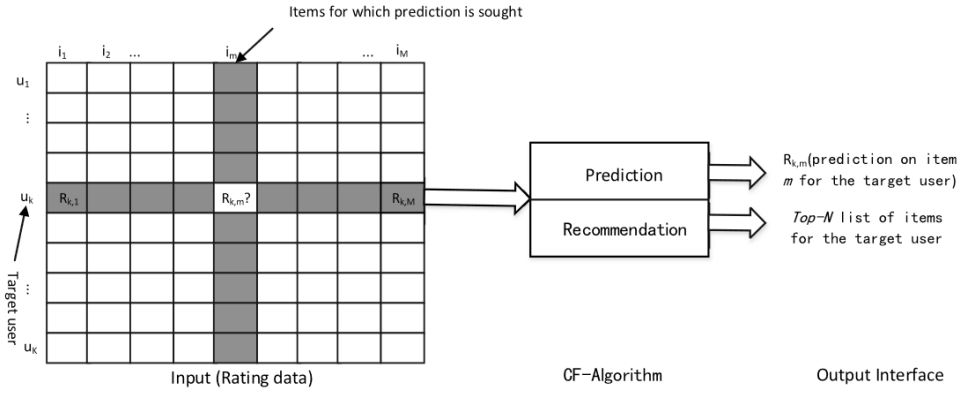

Figure 4 The Collaborative Filtering Process

Figure 4 แสดงถึงกระบวนการจัดทำระบบให้คำแนะนำโดยการใช้เทคนิคการกรองแบบร่วมมือ[12] ซึ่งจะเป็นการพยากรณ์อันดับความสนใจในสินค้าที่ I_m ให้กับผู้ใช้ U_k ที่เข้ามาใช้ในระบบใหม่คือค่า $R_{k,m}$ ซึ่งมีการนำเอาข้อมูลเข้าในรูปแบบของมิติข้อมูลของผู้ใช้-สินค้า (User-Item) โดยนำค่าอันดับความสนใจของแต่ละผู้ใช้ที่มีต่อตัวสินค้ามาทำการหาค่าความคล้ายถึงเพื่อทำการเลือกผู้ใช้หรือสินค้าที่มีการคล้ายคลึงกันมากที่สุดมาทำการพยากรณ์และให้คำแนะนำสำหรับผู้ใช้รายใหม่

การหาค่าความคล้ายคลึง (Similarity) การจัดทำระบบให้คำแนะนำแบบการกรองแบบร่วมมือบนพื้นฐานความจำนั้น เป็นการนำเอาข้อมูลประวัติของผู้ใช้ที่เคยให้อันดับความสนใจของสินค้ามาใช้ในการคำนวณหาค่าความคล้ายคลึงกับผู้ใช้ใหม่ที่เข้ามาในระบบ โดยมีผู้วิจัยได้จัดทำระบบให้คำแนะนำโดยใช้อัลกอริธึมต่างๆ ได้แก่ Pearson's Correlation (COR) Cosine (COS) Adjusted Cosine (ACOS) for similarity between เป็นต้น ซึ่งในแต่ละอัลกอริธึมมีผู้วิจัยได้นิยมใช้สร้างระบบให้คำแนะนำในการหาค่าความคล้ายของผู้ใช้เดิมกับผู้ใช้ใหม่ที่เข้ามาในระบบ และจากนั้นจะหาเพื่อนบ้านใกล้เคียงเพื่อนำข้อมูลอันดับมาใช้ในการพยากรณ์อันดับความสนใจของผู้ใช้ใหม่ต่อไป

การหาเพื่อนบ้านใกล้เคียง (Neighbors) การเลือกข้อมูลมาเพื่อใช้ในการพยากรณ์นั้น ได้มีผู้วิจัยได้จำแนกการเลือกข้อมูลเพื่อมาใช้ในการพยากรณ์ผู้ใช้ใหม่ไว้ 2 ลักษณะ[13] ได้แก่ 1) Nearest Neighbor algorithms เป็นการหาค่าความคล้ายของผู้ใช้กับผู้ใช้ปัจจุบัน โดยจะนำข้อมูลเดิมที่มีอยู่มาใช้ในการหาค่าความคล้ายกับผู้ใช้ใหม่ที่เข้ามา และหาสินค้าที่มีความใกล้เคียงแนะนำให้กับผู้ใช้ใหม่ เนื่องจากผู้ใช้ใหม่น่าจะมีความชอบในสินค้าที่ใกล้เคียงกับประวัติของผู้ใช้เดิมในระบบ โดยจะแบ่งการหาเพื่อนบ้านใกล้เคียงออกเป็น หาเพื่อนบ้านใกล้เคียงในมุมมองของผู้ใช้ และหาเพื่อบ้านใกล้เคียงในมุมมองของสินค้ากับผู้ใช้ใหม่ที่เข้ามาในระบบ 2) Top-N recommendation เป็นวิธีการนำข้อมูลมาใช้ในการพยากรณ์โดยมีการกำหนดค่า N คือจำนวนชุดข้อมูลของผู้ใช้ หรือสินค้าที่มีค่าความคล้ายคลึงสูง ซึ่งจะทำการวิเคราะห์จากตารางมิติของผู้ใช้-สินค้า โดยหาค่าความสัมพันธ์ระหว่างผู้ใช้กับผู้ใช้ หรือสินค้ากับสินค้า และนำมาใช้ในคำนวณหาค่าความคล้ายกับผู้ใช้ที่เข้ามาใหม่

การพยากรณ์และให้คำแนะนำ (Prediction and Recommendation Computation) เป็นส่วนที่สำคัญเพื่อที่จะแนะนำและพยากรณ์อันดับความสนใจให้กับผู้ใช้ใหม่ที่เข้ามาในระบบ ซึ่งมีวิธีที่นิยมด้วยกันอยู่ 2 วิธี[3] ได้แก่ การใช้ผลรวมน้ำหนัก (Weighted Sum of Others' Ratings) เป็นการพยากรณ์อันดับความสนใจโดยการใช้ผลรวมของค่าถ่วงน้ำหนักของอันดับความสนใจ และวิธีค่าเฉลี่ยน้ำหนักแบบง่าย (Simple Weighted Average) เป็นการพยากรณ์อันดับความสนใจโดยการใช้วิธีการแบบคิดค่าเฉลี่ย

จากกระบวนการที่กล่าวมาส่วนของการหาค่าความคล้ายคลึงของผู้ใช้รายใหม่กับผู้ใช้งานเดิม ซึ่งเป็นขั้นตอนแรกในการจัดทำระบบให้คำแนะนำแบบการกรองแบบร่วมมือ ได้มีนักวิจัยได้ใช้อัลกอริธึมในการหาค่าความคล้ายคลึงแบบต่างๆ มาใช้ในการหาค่าความคลึงซึ่งจะอธิบายในหัวข้อถัดไปในส่วนของการเปรียบเทียบของการหาค่าความคล้ายคลึง

1.1 เทคนิคการหาค่าความคล้ายคลึง

เทคนิคการหาค่าความคล้ายคลึง[14] เป็นการคำนวณหาค่าความคล้ายคลึงจากตารางมิติของ User-Item ซึ่งเป็นขั้นตอนที่สำคัญของการสร้างระบบให้คำแนะ สำหรับการคำนวณแบบมิติพื้นฐานของสินค้า Item-based เป็นการหาค่าความคล้ายคลึงระหว่างสินค้า i_a และ สินค้า i_b โดยจะทำการคำนวณจากอันดับความสนใจที่ผู้ใช้ให้ในสินค้าที่มีความสัมพันธ์กัน และแบบมิติพื้นฐานของผู้ใช้ User-based มีลักษณะการคำนวณค่าความคล้ายคลึงเหมือนกันแต่จะเป็นการหาค่าความคล้ายคลึงของผู้ใช้ ซึ่งมีผู้วิจัยใช้ และทำการเปรียบเทียบการหาค่าความคล้ายคลึงแบบต่างๆ มาใช้การสร้างระบบให้คำแนะนำ โดยงานวิจัยส่วนมากจะมีการหาค่าความคล้ายคลึง[1,3] แบบ ได้แก่ 1) Pearson's Correlation (COR) 2) Cosine (COS) 3) Adjusted Cosine (ACOS) for similarity between items 4) Distance-based similarity 5) Constrained Pearson's Correlation (CPC) 6) Spearman's Rank Correlation (SRC) 7) Proximity-Impact-Popularity (PIP) 8) Bhattacharyya coefficient และ 9) Linkelihood Ratio Similarity (LiRa) เป็นต้น ซึ่งในการหาค่าความคล้ายคลึงแบบต่างๆ ได้มีการนำข้อมูลอันดับความสนใจของผู้ใช้ต่อสินค้ามาใช้ในการคำนวณซึ่งจะมีการนำค่าผลต่างหรือผลคูณของคู่อันดับความสนใจสินค้า ค่าเฉลี่ย ค่ามัธยฐาน หรือจำนวนการให้อันดับความสนใจของสินค้าที่เหมือนกันของแต่ละผู้ใช้ (Co-rate) มาใช้เป็นข้อมูลเข้าเพื่อใช้ในการคำนวณหาค่าความคล้ายคลึงของผู้ใช้ หรือของสินค้า โดยแสดงได้ดัง Table 1 แสดงถึงสมการในการหาค่าความคล้ายคลึงสำหรับเทคนิคการกรองแบบร่วมมือโดยแต่ละสมการจะมีการนำข้อมูลอันดับความสนใจในรูปแบบ ผู้ใช้-สินค้า มาทำการหาค่าความคล้ายคลึงโดยแต่ละสมการจะมีการนำเอาค่าของอันดับความสนใจมาทำการหาค่าต่างกับค่าที่ใช้ได้แก่ ค่าปกติ ค่าเฉลี่ย ค่ามัธยฐาน การนับจำนวน

1.2 การพยากรณ์และให้คำแนะนำ (Prediction and Recommendation Computation)

ในการพยากรณ์และให้คำแนะนำของระบบให้คำแนะนำ นิยมใช้สองวิธีการในการพยากรณ์ค่าอันดับความสนใจให้กับผู้ใช้รายใหม่ที่เข้ามาใช้ในระบบ แสดงดัง Table 2 แสดงถึงวิธีการในการพยากรณ์อันดับความสนใจให้กับผู้ใช้รายใหม่ที่เข้ามาใช้ระบบให้คำแนะนำ ซึ่งมีวิธีการที่นิยมใช้ 2 วิธีการ โดยจะเห็นได้ว่าวิธีการแบบวิธีผลรวมน้ำหนักของอันดับความสนใจ นิยมใช้กับการพยากรณ์จากผู้ใช้ ส่วนวิธีค่าเฉลี่ยน้ำหนักแบบง่ายจะนิยมใช้กับการพยากรณ์สำหรับสินค้า

1.3 การทดสอบและประเมินผลระบบให้คำแนะนำ

ในการทดสอบและประเมินผลระบบให้คำแนะนำนั้นได้มีผู้นิยมใช้กระบวนการทดสอบการจำแนกหรือการพยากรณ์ที่ได้จากการสร้างระบบให้คำแนะนำ[1,3,15] ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความเที่ยงตรงแม่นยำ (Precision) ค่าความระลึก (Recall) และ ค่า (F-measure) ส่วนการประเมินประสิทธิภาพของระบบให้คำแนะนำ นิยมทดสอบโดยการใช้ ค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ (Mean Absolute Error: MAE) และค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE)

Table 1 The Equations of Similarity of Collaborative Filtering Techniques

การหาค่าความคล้ายคลึง	รายละเอียด	ค่าปกติ (R)	ค่าเฉลี่ย (AVG)	มัธยฐาน (Med)	การนับจำนวน (C)	ประเภท
Pearson's Correlation (COR)	$S(u_a,u_b)=\frac{\sum_{h=1}^{n}(r_{u_a i_h}-\bar{r}_{u_a})(r_{u_b i_h}-\bar{r}_{u_b})}{\sqrt{\sum_{h=1}^{n}(r_{u_a i_h}-\bar{r}_{u_a})^2}\sqrt{\sum_{h=1}^{n}(r_{u_b i_h}-\bar{r}_{u_b})^2}}$ เป็นการหาค่าความคลายคลึงระหว่างผู้ใช้ u_a และ u_b โดยกำหนดให้ $r_{u,i}$ เป็นอันดับความสนใจของสินค้า i โดยผู้ใช้ u และ $\bar{r}$ คือค่าเฉลี่ยของอันดับความสนใจทั้งหมดของผู้ใช้ u ที่มีการให้อันดับสินค้าเหมือนกันของผู้ใช้ (co-rate) และ n คือจำนวนของสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate)		x			S
Cosine (COS)	$S(u_a,u_b)=\frac{\sum_{h=1}^{n}(r_{u_a i_h})(r_{u_b i_h})}{\sqrt{\sum_{h=1}^{n}r_{u_a i_h}^2}\sqrt{\sum_{h=1}^{n}r_{u_a i_h}^2}}$ เป็นการหาค่าความคล้ายคลึงเชิงมุมที่กำหนดให้ $r_{u,i}$ เป็นอันดับความสนใจของสินค้า i โดยผู้ใช้ u และ n คือจำนวนของสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate)	x				S
Adjusted Cosine (ACOS) for similarity between items	$S(i_a,i_b)=\frac{\sum_{h=1}^{n}(r_{u_h i_a}-\bar{r}_{u_h})(r_{u_h i_b}-\bar{r}_{u_h})}{\sqrt{\sum_{h=1}^{n}(r_{u_h i_a}-\bar{r}_{u_h})^2}\sqrt{\sum_{h=1}^{n}(r_{u_h i_b}-\bar{r}_{u_h})^2}}$ เป็นการหาค่าความคล้ายคลึงสำหรับสินค้า i_a และ i_b โดยกำหนดให้ $r_{u,i}$ เป็นอันดับความสนใจของสินค้า i โดยผู้ใช้ u และ $\bar{r}_u$ คือค่าเฉลี่ยของอันดับความสนใจทั้งหมดของผู้ใช้ u ที่มีการให้อันดับสินค้าโดยผู้ใช้นั้น และ n คือจำนวนของผู้ใช้ที่ให้อันดับความสำคัญต่อสินค้านั้น		x			S
Constrained Pearson's Correlation (CPC)	$S(u_a,u_b)=\frac{\sum_{h=1}^{n}(r_{u_a i_h}-r_{med})(r_{u_b i_h}-r_{med})}{\sqrt{\sum_{h=1}^{n}(r_{u_a i_h}-r_{med})^2}\sqrt{\sum_{h=1}^{n}(r_{u_a i_h}-r_{med})^2}}$ เป็นการหาค่าความคลายคลึงระหว่างผู้ใช้ u_a และ u_b โดยกำหนดให้ $r_{u,i}$ เป็นอันดับความสนใจของสินค้า i โดยผู้ใช้ u และ r_{med} คือค่ามัธยฐานของอันดับความสนใจ (เช่น ค่าเป็น 3 ในอันดับความสนใจที่ 5) และ n คือจำนวนของสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate)			x		S
Spearman's Rank Correlation (SRC)	$S(u_a,u_b)=1-\frac{6\sum_{h=0}^{n}d_h^2}{n(n^2-1)}$ เป็นการหาค่าความคล้ายคลึงระหว่างผู้ใช้ u_a และ u_b โดยกำหนดให้ d_n เป็นค่าความแตกต่างในช่วงสำหรับสินค้า h สำหรับผู้ใช้สองคน และ n คือจำนวนของสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate)	x				S

หมายเหตุ การหาค่าความคล้ายคลึงโดยพิจารณาจาก ค่าความสัมพันธ์ (S) ค่าความแตกต่าง (D) ค่าอัตราส่วน (R) แบบผสมผสาน (M) และการนำชุดข้อมูลภายนอกมาใช้ในการสร้างระบบให้คำแนะนำ (S+E)

Table 1 The Equations of Similarity of Collaborative Filtering Techniques (Cont.)

การหาค่าความคล้ายคลึง	รายละเอียด	ค่าปกติ (R)	ค่าเฉลี่ย (AVG)	มัธยฐาน (Med)	การนับจำนวน (C)	ประเภท
Euclidean[16,17]	$Distanct(u_a,u_b)=\sqrt{\sum_{i=1}^{n}(r_{u_a,i}-r_{u_b,i})^2}$ และจะได้สมการในการหาค่าความคล้ายคลึงดังนี้ (18) $S(u_a,u_b)=\frac{1}{1+Distanct(u_a,u_b)}$ เป็นการหาค่าความแตกต่างระหว่างผู้ใช้ u_a กับผู้ใช้ u_b ในการให้อันดับความสนใจของสินค้า i โดยค่าความแตกต่างน้อยจะมีความคล้ายคลึงสูง	x				D
Jaccard (Jacc)[19]	$S(u_a,u_b)=\frac{\lvert I_{u_a}\cap I_{u_b}\rvert}{\lvert I_{u_a}\cup I_{u_b}\rvert}$ เป็นการหาค่าความคล้ายโดยการกำหนดให้ I_{u_a} คือชุดข้อมูลของสินค้าที่ถูกจัดอันดับความสำคัญโดยผู้ใช้ u_a				x	R
Mean squared difference (MSD)	$S(u_a,u_b)=1-\frac{\sum_{h\in n}(r_{u_a i_h}-r_{u_b i_h})^2}{\lvert n\rvert}$ เป็นการหาค่าความคล้ายคลึงโดยกำหนดให้ n เป็นชุดของสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate) และ $r_{u,i}$ คืออันดับความสนใจของผู้ใช้ u บนสินค้า i	x				S
Proximity-Impact-Popularity (PIP)[14]	$S(u_a,u_b)=\sum_{h\in n}PIP(r_{u_a i_h},r_{u_b,i_h})$ เป็นการหาค่าความคล้ายคลึงของผู้ใช้ u_a และ u_b ซึ่งกำหนดให้ โดยมีการกำหนด ให้ r_{ah} และ r_{bh} คืออันดับความสนใจของสินค้าที่ h โดยผู้ใช้ u_a และ u_b และ n คือสินค้าที่มีการให้อันดับความสนใจเหมือนกันของผู้ใช้ (co-rate) ของผู้ใช้ u_a และ u_b โดยจะมีการดำเนินการหาค่าความใกล้ ค่าผลกระทบ และค่าความนิยมมาใช้ในการหาค่าความคล้ายคลึง ซึ่ง $PIP(r_{u_a h},r_{u_b h})$ คือค่าคะแนนสำหรับสองระดับที่ได้จาก $r_{u_a h}$ และ $r_{u_b h}$ และนำค่าสำหรับสองระดับที่ได้ r_1 และ r_2 มาดำเนินการ ดังสมการด้านล่าง $PIP(r_1,r_2)=Proximity(r_1,r_2)\times Impact(r_1,r_2)\times Popularity(r_1,r_2)$ โดยจากสมการด้านบนในการทำงานจะมีการแยกออกเป็นสองกรณี ซึ่งได้แก่กรณีที่เป็นจริง และกรณีที่เป็นเท็จ โดยมีการนำค่ากลางมาใช้ในการพิจารณากับค่า r_1,r_2 ว่ามากหรือน้อยกว่าค่ากลางหรือไม่	x	x	x		M
Bhattacharyya coefficient (BC)[19]	$S(u_a,u_b)=Jacc(u_a,u_b)+\sum_{i\in i_{u_a}}\sum_{j\in i_{u_b}}BC(i,j)\,loc(r_{u_a},r_{u_b})$ เป็นการหาค่าความคล้ายคลึงโดยผู้ใช้ u_a และผู้ใช้ u_b ซึ่ง $BC(i,j)$ เป็นค่าอันดับความสนใจระหว่างสินค้า i และ j ซึ่งเป็นการตรวจสอบการหาค่า Co-rate ของสินค้าที่มีการให้อันดับความสนใจ โดยจะมีค่าเป็น 0 และ 1 หากมีค่าเป็น 0 จะใช้ค่าของ Jaccard เป็นค่าความคล้ายคลึง และ $loc(r_{u_a},r_{u_b})$ เป็นค่าความคล้ายคลึงระหว่างอันดับความสนใจของทั้งสองวิธี ได้แก่การใช้การหาค่าความคล้ายคลึงโดยใช้ค่าเฉลี่ย หรือการหาค่าความคล้ายคลึงโดยใช้ค่ามัธยฐาน และ $Jacc(u_a,u_b)$ เป็นการคำนวณค่าความคล้ายคลึงของผู้ใช้แบบ Jaccard	x	x	x	x	M

หมายเหตุการหาค่าความคล้ายคลึงโดยพิจารณาจาก ค่าความสัมพันธ์ (S) ค่าความแตกต่าง (D) ค่าอัตราส่วน (R) แบบผสมผสาน (M) และการนำชุดข้อมูลภายนอกมาใช้ในการสร้างระบบให้คำแนะนำ (S+E)

Table 1 The Equations of Similarity of Collaborative Filtering Techniques (Cont.)

การหาค่าความคล้ายคลึง	รายละเอียด	ค่าปกติ (R)	ค่าเฉลี่ย (AVG)	มัธยฐาน (Med)	การนับจำนวน (C)	ประเภท
Linkelihood Ratio Similarity (LiRa)[20]	$S(u_a,u_b) = \log_{10} \frac{p(differences\ in\ u_a and\ u_b \mid same\ cluster)}{p(differences\ in\ u_a and\ u_b \mid pure\ chance)}$ เป็นการหาค่าความคล้ายคลึงจากการคำนวณอัตราส่วนของค่าความน่าจะเป็นของ u_a และ u_b				x	R
Fuses user and item information (FUIR)[12]	$S(u_a,u_b) = \frac{\sum_{k=1}^{\lvert F\rvert} CTRI(u_a,k) \cdot CTRI(u_b,k)}{\sqrt{\sum_{k=1}^{\lvert F\rvert} CTRI(u_a,k)^2}\sqrt{\sum_{k=1}^{\lvert F\rvert} CTRI(u_b,k)^2}}$ เป็นการหาค่าความคล้ายคลึงของผู้ใช้ซึ่งจะมีการดำเนินการลำดับแรก โดยการหาค่า Customer Relative Interest (CTRI) สำหรับเมตริกของ ผู้ใช้-สินค้า โดยคำนวณได้จากสมการ $CTRI(u_a,k) = \frac{\sum_{i \in u_{a,b}} RP_{a,i} \cdot v_{i,k}}{\sum_{i \in u_{a,b}} RP_{a,i}}$ โดยจากสมการ CTRI เป็นสมการในการหาค่าความคล้ายคลึงระหว่างผู้ใช้ u_a และ u_b โดยเบื้องต้นจะทำการคำนวณบนพื้นฐานของสินค้าของผู้ใช้แต่ละคนโดยจะพิจารณาจาก k ที่เป็นคุณลักษณะของสินค้าที่ใช้ในการคำนวณ		x	x		S+E

หมายเหตุการหาค่าความคล้ายคลึงโดยพิจารณาจาก ค่าความสัมพันธ์ (S) ค่าความแตกต่าง (D) ค่าอัตราส่วน (R) แบบผสมผสาน (M) และการนำชุดข้อมูลภายนอกมาใช้ในการสร้างระบบให้คำแนะนำ (S+E)

Table 2 Prediction and Recommendation Computation

วิธีการพยากรณ์	รายละเอียด
วิธีผลรวมน้ำหนักของอันดับความสนใจ (Weighted Sum of Others' Ratings : WS)	$P_{u_a,i_a} = \bar{r}_{u_a} + \frac{\sum_{h=1}^{n}(r_{u_h,i_a} - \bar{r}_{u_h}) \cdot S(u_a,u_h)}{\sum_{h=1}^{n} \lvert S(u_a,u_h)\rvert}$ จากสมการด้านบน กำหนดให้ $\bar{r}_{u_a}$ คือค่าเฉลี่ยอันดับความสนใจของผู้ใช้ a ในทุกการให้อันดับความในใจทุกสินค้า และ n คือจำนวนของผู้ใช้ที่ให้อันดับความสนใจกับสินค้า i_a และค่า $S(u_a,u_h)$ คือ ค่าความคล้ายคลึงระหว่างผู้ใช้ a และ ผู้ใช้ h โดยคำนวณทุกสมาชิกที่มีการให้อันดับสินค้า i_a โดยเป็นการพยากรณ์บนพื้นฐานของเพื่อนบ้านใกล้เคียงผู้ใช้ที่เข้ามาใหม่
วิธีค่าเฉลี่ยน้ำหนักแบบง่าย (Simple Weighted Average: SWA)	$P_{u_a,i_a} = \frac{\sum_{h=1}^{n} r_{u_a,i_h} \cdot S(i_a,i_h)}{\sum_{h=1}^{n} \lvert S(i_a,i_h)\rvert}$ เป็นการพยากรณ์สำหรับ Item-based ซึ่งเป็นการหาโดยค่าเฉลี่ยน้ำหนักแบบง่าย สำหรับ ผู้ใช้ u_a ใน สินค้า i_a โดยจะทำการหาอันดับความสนใจทั้งหมดในแต่ละสินค้าสำหรับผู้ใช้ $S(i_a,i_h)$ คือค่าความคล้ายคลึงระหว่างสินค้า i_a และ i_h ส่วน r_{u_a,i_h} คืออันดับความสนใจของ ผู้ใช้ u_a กับสินค้า i_h

Table 3 Evaluation of Recommendation System

วิธีการทดสอบ	รายละเอียด
ค่าความถูกต้อง (Accuracy) เป็นค่าสำหรับใช้ในการวัดค่าความถูกต้องจากผลที่ได้จากการพยากรณ์อันดับความสนใจให้กับผู้ใช้รายใหม่	$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}$ โดยกำหนดให้ TP ค่าใช่ที่ถูกต้อง TN ค่าใช่ที่ผิด FP ค่าไม่ที่ทายถูก FN ค่าไม่ ที่ทายผิด
ค่าความเที่ยงตรงแม่นยำ (Precision) เป็นค่าที่ใช้สำหรับวัดความแม่นยำของการทำนาย กับค่าที่เป็นจริงทั้งหมดของชุดข้อมูลจริง	$Precision = \frac{TP}{TP + FP}$ โดยกำหนดให้ TP ค่าใช่ที่ถูกต้อง TN ค่าใช่ที่ผิด FP ค่าไม่ที่ทายถูก FN ค่าไม่ ที่ทายผิด
ค่าความระลึก (Recall) เป็นค่าจากการค้นหาข้อมูลที่เป็นค่าอัตราส่วนของการค้นพบค่าความถูกต้องเชิงบวก (True Positives: TP) เทียบกับค่าที่ทำนายถูกต้องทั้งหมด	$Recall = \frac{TP}{TP + FN}$ โดยกำหนดให้ TP ค่าใช่ที่ถูกต้อง TN ค่าใช่ที่ผิด FP ค่าไม่ที่ทายถูก FN ค่าไม่ ที่ทายผิด
ค่า (F-measure) คือการนำค่า Precision และค่า Recall มารวมกันเป็นค่าเฉลี่ยที่ดีที่สุด	$F1 = \frac{2 \times precision \times recall}{precision + recoll}$
ค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ (Mean Absolute Error: MAE) เป็นการวัดค่าความแตกต่างระหว่างค่าจริงและค่าที่ประมาณการได้จากแบบจำลอง โดย หากค่า MAE มีค่าน้อย แสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับผลที่ได้จากการทดลองมีสมการ	$MAE = \frac{1}{S}\sum_{i,j}\lvert R_{i,j} - \hat{R}_{i,j}\rvert$ โดยกำหนดให้ $R_{i,j}$ คือค่าจากการพยากรณ์ที่ได้จากแบบจำลอง $\hat{R}_{i,j}$ คือค่าที่เกิดขึ้นจริง S คือจำนวนข้อมูลที่ใช้ในแบบจำลอง
ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) เป็นการวัดค่าความคลาดเคลื่อนซึ่งมีลักษณะเหมือนรากที่สองของค่าเฉลี่ยส่วนเบี่ยงเบนมาตรฐาน ซึ่งหากมีค่าน้อยแสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริงมีสมการ	$RMSE = \sqrt{\frac{1}{S}\sum_{i,j}\left(R_{i,j} - \hat{R}_{i,j}\right)^2}$ โดยกำหนดให้ $R_{i,j}$ คือค่าจากการพยากรณ์ที่ได้จากแบบจำลอง $\hat{R}_{i,j}$ คือค่าที่เกิดขึ้นจริง S คือจำนวนข้อมูลที่ใช้ในแบบจำลอง

Table 4 Memory-based Collaborative Filtering Technique

ปี	เรื่อง	เทคนิคใช้เปรียบเทียบ	เทคนิคที่งานวิจัยนำเสนอ	การพยากรณ์	การทดสอบ	ข้อมูลที่ใช้	ลักษณะเด่นและเทคนิค
2016	FUIR: Fusing user and item information to deal with data sparsity by using side information in recommendation systems[12]	COS, COR, ACOS	FUIR	SWA	MAE	R	มีการนำ CTRI มาประยุกต์กับการหาค่าความคล้ายคลึงกับ COS ซึ่งเป็นการนำมิติข้อมูลของคุณลักษณะของสินค้ามาทำการพิจารณาในการหาค่า CTRI ระหว่างสินค้าแต่ละรายการ
2016	LiRa: A New Likelihood-Based Similarity Score For Collaborative Filtering[20]	COS, COR, BC	LiRa	BCF	RMSE	C	นำหลักการ Likelihood โดยมีลักษณะเป็นอัตราส่วน มาใช้ในการสร้างสมการหาค่าความคล้ายคลึง ใช้ $\log_{10}$ ในการหาค่าความคล้ายคลึงซึ่งเข้ามาแก้ปัญหาของการใช้การหาค่าความคล้ายคลึงของผู้ใช้ที่มีการให้อันดับความสนใจเหมือนกันทำให้ค่าความคล้ายคลึงเท่ากับ 1 ดังนี้จึงใช้วิธีการนี้ในการหาค่าความคล้ายคลึงซึ่งจะนำจำนวน co-rate ของสินค้ามาใช้ในการพิจารณา
2016	An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering[16]	COR, Euclidean, SRC, TAN, AG	GA_M	SWA	RMSE	R, AVG, C	นำเอาวิธีการขั้นตอนวิธีเชิงพันธุกรรมมาประยุกต์ใช้ในการเลือกค่าผลลัพธ์ที่ได้จากการหาค่าความคล้ายคลึงของเทคนิคในการจัดทำระบบให้คำแนะนำแบบต่างๆ มาผสมกัน บนพื้นฐานเทคนิคการกรองแบบร่วมมือในการสร้างระบบให้คำแนะนำ
2016	Extended Collaborative Filtering Technique for Mitigating the Sparsity Problem[21]	COS	USIM,ISIM	WS	Precision, Recall, F1	R	การหาค่าความคล้ายคลึงบนพื้นฐานของการใช้วิธีการแบบ COS ในการดำเนินการซึ่งจะดำเนินการหา USIM จากการนำเอา ISIM มาใช้ในการคำนวณหา USIM บนสมการ COS

Table 4 Memory-based Collaborative Filtering Technique (Cont.)

ปี	เรื่อง	เทคนิคใช้เปรียบเทียบ	เทคนิคที่งานวิจัยนำเสนอ	การพยากรณ์	การทดสอบ	ข้อมูลที่ใช้	ลักษณะเด่นและเทคนิค
2015	A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data[19]	MJD, PIP, PC, JMSD, NHSM, CPC	BCF {BCF(cor), BCF(med),}	BCF	MAE, RMES, F1	R, AVG, Med, C	การหาค่าความคล้ายคลึง โดยมีลักษณะผสมผสานด้วยเทคนิคต่างๆ โดยลำดับแรกจะเป็นการตรวจสอบการหาค่า Co-rate ของสินค้าที่มีการให้อันดับความสนใจ โดยจะมีค่าเป็น 0 และ 1 หากมีค่าเป็น 0 จะใช้ค่าของ Jaccard เป็นค่าความคล้ายคลึง และหากมีค่าเป็น 1 จะพิจารณาโดยใช้มีลักษณะของการหาค่าความคล้ายคลึงแบบค่าเฉลี่ยและค่าความคล้ายคลึงจากค่ามัธยฐานเข้ามาใช้ในการดำเนินการ
2013	A new similarity function for selecting neighbors for each target item in collaborative filtering[22]	COR(CF_P), COS(CF_C), Distance(CF_D)	CF_P_P, CF_P_C, CF_P_D, CF_C_P, CF_C_C,CF_C_D, CF_D_P, CF_D_C, CF_D_D	SWA	Precision, Recall, F1, Coverage	R, AVG	การหาเพื่อนบ้านใกล้เคียงของผู้ใช้งานที่ต้องการ ในแต่ละสินค้าที่ต้องการที่มีความแตกต่างกัน ซึ่งก่อให้เกิดความถูกต้องในการแนะนำและได้สินค้าที่หลากหลายในกลุ่มผู้ใช้งานเป้าหมาย

หมายเหตุ ข้อมูลหรือตัวแปรที่สำคัญมาใช้ในการหาค่าความคล้ายคลึงได้แก่ อันดับความสนใจ (R) ค่าเฉลี่ยของอันดับความสนใจ (AVG) ค่ามัธยฐานของอันดับความสนใจ (Med) และจำนวน Co-rate ของอันดับความสนใจ (C)

2. งานวิจัยที่เกี่ยวข้องการสร้างระบบให้คำแนะนำโดยใช้เทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ

จากกระบวนการจัดทำระบบให้คำแนะนำโดยใช้เทคนิคการกรองแบบร่วมมือที่กล่าวมาข้างต้นนั้น ได้มีนักวิจัยได้จัดทำระบบให้คำแนะนำโดยใช้เทคนิคต่างๆ มาสร้างระบบให้คำแนะนำซึ่งมีการดำเนินการประกอบด้วยกระบวนการต่างๆ โดยนำเทคนิคมาทำการเปรียบเทียบกับวิธีการเดิม ดัง (Table 4) โดยจากลักษณะของระบบให้คำแนะนำที่ใช้เทคนิคการหาค่าความคล้ายคลึงในลักษณะต่างๆ เป็นตารางที่แสดงถึงงานวิจัยที่นำเสนอเทคนิคการหาค่าความคล้ายคลึงมาใช้ในการจัดทำการหาค่าความคล้ายคลึงในการสร้างระบบให้คำแนะนำ โดยตารางได้แสดงถึงเทคนิคที่ใช้ การเปรียบเทียบกับเทคนิคแบบดังเดิม และข้อมูลนำเข้าที่ใช้ในการคำนวณ ซึ่งในแต่ละเทคนิคมีการนำข้อดีของแต่ละเทคนิคมาผสมผสานทำงานร่วมกัน ซึ่งจากงานวิจัยในเทคนิคใหม่ๆ ส่วนใหญ่จะมีการนำค่ามัธยฐาน หรือค่าการนับจำนวน (จำนวนการให้อันดับความสนใจเหมือนกัน) มาใช้ในการเลือกเทคนิคที่เหมาะสมสำหรับมิติข้อมูลชุดนั้นๆ เพื่อให้ได้ค่าความถูกต้องในการพยากรณ์ และประสิทธิภาพในการให้คำแนะนำของระบบมีค่ามากขึ้น และลดทรัพยากรณ์ที่ใช้ในการประมวลผลลง

จากกระบวนการสร้างระบบให้คำแนะนำที่กล่าวมาข้างต้นแสดงให้เห็นได้ว่ามีการนำสมการการหาค่าความคล้ายคลึงมาใช้ในการสร้างเทคนิคการหาค่าความคล้ายคลึงแบบต่างๆ ซึ่งในแต่ละสมการมีความต้องลักษณะข้อมูลของมิติข้อมูลที่แตกต่างกัน เช่นการจัดทำโดยใช้ FUIR[12] เป็นการนำมิติข้อมูลคุณลักษณะของสินค้าเข้ามาช่วยในการหาค่าความคล้ายคลึง LiRa[20] การนำ $\log_{10}$ ในการหาค่าความคล้ายคลึงซึ่งเข้ามาแก้ปัญหาของการใช้การหาค่าความคล้ายคลึงของผู้ใช้ที่มีการให้อันดับความสนใจเหมือนกันทำให้ค่าความคล้ายคลึงเท่ากับ 1 Bhattacharyya[19] เป็นการนำสมการการหาค่าความคล้ายคลึงที่เหมาะสมสำหรับมิติข้อมูลมาใช้ในการเลือกใช้ PIP[14] เป็นเทคนิคในการแบ่งลักษณะของข้อมูลออกเป็นกรณี และดำเนินจัดการกับข้อมูล 3 กิจกรรมหลัก หรือ GA_M [16] เป็นการนำเอาวิธีการขั้นตอนวิธีเชิงพันธุกรรมมาใช้ในการเลือกค่าผลลัพธ์ที่ได้จากการหาค่าความคล้ายคลึงของเทคนิคในการจัดทำระบบให้คำแนะนำแบบต่างๆ มาผสมกัน ซึ่งในแต่ละเทคนิคนั้นเพื่อใช้ในการเข้ามาแก้ไขปัญหาที่เกิดขึ้นจากความเบาบางของข้อมูล โดยปัญหาดังกล่าวนี้จะส่งผลโดยตรงต่อการสร้างระบบให้คำแนะนำแบบการกรองแบบร่วมมือบนพื้นฐานความจำ เนื่องจากต้องมีการนำมิติข้อมูลที่มีการให้อันดับความสนใจในสินค้าเดียวกันมาใช้ในการหาค่าความคล้ายคลึง

สรุป

บทความนี้เป็นการศึกษาเกี่ยวกับการสร้างระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือพื้นฐานความจำ ซึ่งมีกระบวนการหลัก 3 ขั้นตอน ได้แก่ การหาค่าความคล้ายคลึง การหาเพื่อนบ้านใกล้เคียง และการพยากรณ์และให้คำแนะนำในกระบวนการสร้างดังที่กล่าวมาข้างต้นนั้น ในขั้นตอนของกระบวนการหาค่าความคล้ายคลึง ของผู้ใช้ หรือของสินค้ามีผู้นิยมจัดทำงานวิจัยเพื่อพัฒนาและปรับปรุงวิธีการหาค่าความคล้ายคลึง โดยการจัดทำระบบให้คำแนะนำดังกล่าวจะมีข้อจำกัดของมิติข้อมูล และค่าความเบาบางของข้อมูล ที่จะส่งผลต่อประสิทธิภาพในการหาค่าความคล้ายคลึง การพยากรณ์และประสิทธิภาพในการให้คำแนะนำของระบบ ซึ่งแต่ละเทคนิคมีข้อดี และข้อเสียในการหาค่าความคล้ายคลึงจากชุดข้อมูลที่มีความเบาบาง จึงมีผู้วิจัยได้นำเทคนิคต่างๆ มาผสมผสานในการใช้หาค่าความคล้ายคลึง และนำเสนอเทคนิคใหม่ โดยจากที่ศึกษาในเทคนิคแบบ LiRa จะมีความเหมาะสมสำหรับการหาค่าความคล้ายคลึงของผู้ใช้เป็นหลักซึ่งจะมีการคำนวณในผู้ให้อันดับความชอบของผู้ใช้ที่เหมือนกัน ส่วน BC จะมีการให้ผลการหาค่าความคล้ายคลึงของสินค้าที่สูงเนื่องจากจะมีการนำค่าน้ำหนักของความชอบผู้ใช้ที่เหมือนกันมาร่วมในการคำนวณค่าความคล้ายคลึง และการประเมินผลการทดลองในการทำนายนิยมใช้ทั้งแบบ MAE และ RMSE โดยการวัดประสิทธิภาพการทำนายแบบ RMSE จะให้ค่าที่สูงกว่าแบบ MAE เนื่องจากนำค่าเฉลี่ยความผิดพลาดมาทำการยกกำลังสอง และ MAE จะเหมาะกับข้อมูลที่มีค่าความคลาดเคลื่อนที่มีความสม่ำเสมอ ดังนั้นวิธีการหาค่าความคล้ายคลึงของผู้ใช้ และสินค้าเพื่อใช้ในการสร้างระบบให้คำแนะนำด้วยเทคนิคการกรองแบบร่วมมือบนพื้นฐานความจำ ยังมีความท้าทายในการจัดทำการปรับปรุงการหาค่าความคล้ายคลึงและการทำนายอันดับความชอบให้กับผู้ใช้ในการสร้างระบบให้คำแนะนำเพื่อให้มีประสิทธิภาพมากขึ้นได้ในอนาคต

เอกสารอ้างอิง

1. Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowl-Based Syst. 2013 Jul;46:109–32.
2. Lu J, Wu D, Mao M, Wang W, Zhang G. Recommender system application developments: A survey. Decis Support Syst. 2015 Jun;74:12–32.
3. Su X, Khoshgoftaar TM. A Survey of Collaborative Filtering Techniques. Adv Artif Intell. 2009 Jan;4:2–4.

4. Ma H, King I, Lyu MR. Effective missing data prediction for collaborative filtering. InProceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval 2007 Jul 23 (pp. 39-46). ACM.
5. Guo J, Guo J. Research on Information Entropy Measure based on Collaborative Filtering Algorithm. International Journal of Hybrid Information Technology. 2016 Mar 31;9(3):1-0.
6. Zhang H, Ni W, Li X, Yang Y. Modeling Idle Customers to Tackle the Sparsity Problem in Time-dependent Recommendation. ICIS 2016.
7. Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing. 2003 Jan;7(1):76-80.
8. Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. InProceedings of the 10th international conference on World Wide Web 2001 Apr 1 (pp. 285-295). ACM.
9. Claypool M, Gokhale A, Miranda T, Murnikov P, Netes D, Sartin M. Combining content-based and collaborative filters in an online newspaper. InProceedings of ACM SIGIR workshop on recommender systems 1999 Aug 19 (Vol. 60).
10. Mobasher B, Burke R, Bhaumik R, Williams C. Effective attack models for shilling item-based collaborative filtering systems. InProceedings of the 2005 WebKDD Workshop, held in conjuction with ACM SIGKDD 2005 Aug 21 (Vol. 2005).
11. Zeng C, Xing CX, Zhou LZ. Similarity measure and instance selection for collaborative filtering. InProceedings of the 12th international conference on World Wide Web 2003 May 20 (pp. 652-658). ACM.
12. Niu J, Wang L, Liu X, Yu S. FUIR: Fusing user and item information to deal with data sparsity by using side information in recommendation systems. Journal of Network and Computer Applications. 2016 Jul 31;70:41-50.
13. Tatiya RV, Vaidya AS. A survey of recommendation algorithms. IOSR Journal of Computer Engineeringf. 2014;16(6):16-9.
14. Ahn HJ. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences. 2008 Jan 2;178(1):37-51.
15. Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS). 2004 Jan 1;22(1):5-3.
16. da Silva EQ, Camilo-Junior CG, Pascoal LM, Rosa TC. An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering. Expert Systems with Applications. 2016 Jul 1;53:204-18.
17. Shimodaira H. Similarity and recommender systems. School of Informatics, The University of Eidenburgh. 2014;21.
18. ธนพล พุกเส็ง และ สุนันฑา สดสี. การประยุกต์แนวคิดผู้เชี่ยวชาญเพื่อการแนะนำสินค้า. วารสารวิทยาศาสตร์และเทคโนโลยี. 2017;25(2):361–75.
19. Patra BK, Launonen R, Ollikainen V, Nandi S. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. Knowledge-Based Systems. 2015 Jul 31;82:163-77.
20. Strnadova-Neeley V, Buluc A, Gilbert JR, Oliker L, Ouyang W. LiRa: A New Likelihood-Based Similarity Score for Collaborative Filtering. arXiv preprint arXiv:1608.08646. 2016 Aug 30.
21. Choi K, Suh Y, Yoo D. Extended Collaborative Filtering Technique for Mitigating the Sparsity Problem. International Journal of Computers, Communications & Control. 2016 Oct 1;11(5).
22. Choi K, Suh Y. A new similarity function for selecting neighbors for each target item in collaborative filtering. Knowledge-Based Systems. 2013 Jan 31;37:146-53.