



การแนะนำรายวิชาออนไลน์โดยวิธีการแนะนำแบบผสม
ONLINE COURSE RECOMMENDATION
USING A HYBRID RECOMMENDATION TECHNIQUE



ปรีชา รัตนเดชอมร

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2564

การแนะนำรายวิชาออนไลน์โดยวิธีการแนะนำแบบผสม



สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2564
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

ONLINE COURSE RECOMMENDATION
USING A HYBRID RECOMMENDATION TECHNIQUE



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
2021
Copyright of Srinakharinwirot University

สารนิพนธ์
เรื่อง
การแนะนำรายวิชาออนไลน์โดยวิธีการแนะนำแบบผสม
ของ
ปรีชา รัตนเดชอมร

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....
(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)
คณบดีบัณฑิตวิทยาลัย
.....

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก	 ประธาน
(ผู้ช่วยศาสตราจารย์ ดร.นุรีย์ วิวัฒน์วัฒนา)	(อาจารย์ ดร.ดวงดาว วิชาดากุล)
	 กรรมการ
	(อาจารย์ ดร.ศุภร คนธภักดี)

ชื่อเรื่อง	การแนะนำรายวิชาออนไลน์โดยวิธีการแนะนำแบบผสม
ผู้วิจัย	ปรีชา รัตนเดชมร
ปริญญา	วิทยาศาสตร์มหาบัณฑิต
ปีการศึกษา	2564
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. นุรีย์ วิวัฒน์วัฒนา

ปัจจุบันสื่อการเรียนและแหล่งเรียนรู้ออนไลน์มีเพิ่มมากขึ้น โดยแต่ละแหล่งต่างมีรายวิชาจำนวนมาก ทำให้ยากต่อการค้นหาและตัดสินใจลงทะเบียนเรียน ระบบแนะนำจึงเป็นส่วนหนึ่งในการแนะนำรายวิชาอัตโนมัติให้ตรงกับความต้องการของผู้เรียน โดยอ้างอิงมาจากประวัติของผู้เรียนในระบบที่เคยลงทะเบียนและให้คะแนนในแต่ละรายวิชา ทำให้ผู้เรียนสามารถตัดสินใจลงทะเบียนได้ง่ายยิ่งขึ้น งานวิจัยนี้เป็นการวิจัยเชิงกรณีศึกษารายวิชาบนเว็บไซต์ Coursera ซึ่งมีวัตถุประสงค์เพื่อทำการพัฒนาระบบแนะนำและเปรียบเทียบเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) ซึ่งใช้รายละเอียดข้อมูลรายวิชาเป็นหลัก โดยใช้หลักการหาความคล้ายคลึงกันของโคไซน์ และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) ซึ่งใช้ประวัติการให้คะแนนรายวิชาเป็นหลัก มาทำการสร้างแบบจำลองโดยใช้อัลกอริทึม SVD, SVD++, NMF, SlopeOne, Co-Clustering และ BaselineOnly เมื่อเปรียบเทียบประสิทธิภาพโดยใช้ Root Mean Square Error (RMSE) และ Mean Absolute Error (MAE) พบว่าอัลกอริทึมที่ให้ผลลัพธ์ที่ดีที่สุดคือ SVD ซึ่งมีค่า RMSE และ MAE เท่ากับ 0.6779 และ 0.4375 ตามลำดับ จากนั้นจึงนำเทคนิคการกรองข้อมูลตามเนื้อหาและเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมมาทำงานร่วมกันเป็นเทคนิคการกรองข้อมูลแบบผสม เพื่อทำให้เกิดการแนะนำรายวิชาที่ทำให้ผู้เรียนสามารถทำการค้นหาได้แม้มีข้อจำกัดในข้อมูลที่ต้องการของเทคนิคแต่ละแบบ

คำสำคัญ : ระบบแนะนำ, การกรองข้อมูลตามเนื้อหา, การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม, ระบบแนะนำแบบผสม

Title	ONLINE COURSE RECOMMENDATION USING A HYBRID RECOMMENDATION TECHNIQUE
Author	PARISA RHATTANADEJAMORN
Degree	MASTER OF SCIENCE
Academic Year	2021
Thesis Advisor	Assistant Professor Nuwee Wiwatwattana , Ph.D.

Online learning resources and learning media are growing more popular. Each site has a variety of courses. This makes it difficult to obtain information and decide whether or not to enroll in a course. The recommendation system is a system component that automatically recommends courses based on the learners who enrolled in and rated each course. This work is presented as a Coursera platform case study. The purpose of this research is to compare techniques such as content-based filtering using cosine similarity techniques with a course detail dataset and collaborative filtering with learner ratings using Surprise libraries to build a model using algorithms such as SVD, SVD++, NMF, SlopeOne, Co-Clustering, and BaselineOnly. The accuracy of the collaborative filtering model was determined by comparing the predicted ratings of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) (MAE). The SVD algorithm got the lowest RMSE and MAE values of 0.6779 and 0.4374, respectively, according to the results. Learners can search for and receive recommendations that go beyond the boundaries of a single methodology by merging content-based filtering and collaborative filtering strategies into a hybrid recommendation system.

Keyword : Content-based Filtering, Collaborative Filtering, Hybrid Recommendation System

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยดี เพราะความอนุเคราะห์จาก ผศ. ดร.นุรีย์ วิวัฒน์วัฒนา อาจารย์ที่ปรึกษาที่ให้คำแนะนำและให้ความช่วยเหลือ ทำให้สารนิพนธ์ฉบับนี้เสร็จสมบูรณ์ ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ที่ได้ให้คำแนะนำและข้อเสนอแนะสำหรับการปรับปรุงสารนิพนธ์ รวมถึงคณะอาจารย์ทุกท่าน ที่ได้ให้คำแนะนำและความรู้ในวิชาต่างๆ ตลอดการศึกษาค้นคว้าวิจัยขอกราบขอบพระคุณด้วยความเคารพอย่างสูง

ขอขอบคุณเพื่อนๆ ในสาขาวิชาที่คอยให้ความช่วยเหลือ ตลอดจนคำแนะนำที่เป็นประโยชน์ทั้งในการเรียนและการทำงานวิจัยนี้

สุดท้ายนี้ขอขอบพระคุณครอบครัว ที่ให้การสนับสนุนและเป็นกำลังใจเสมอมาจนสำเร็จการศึกษา

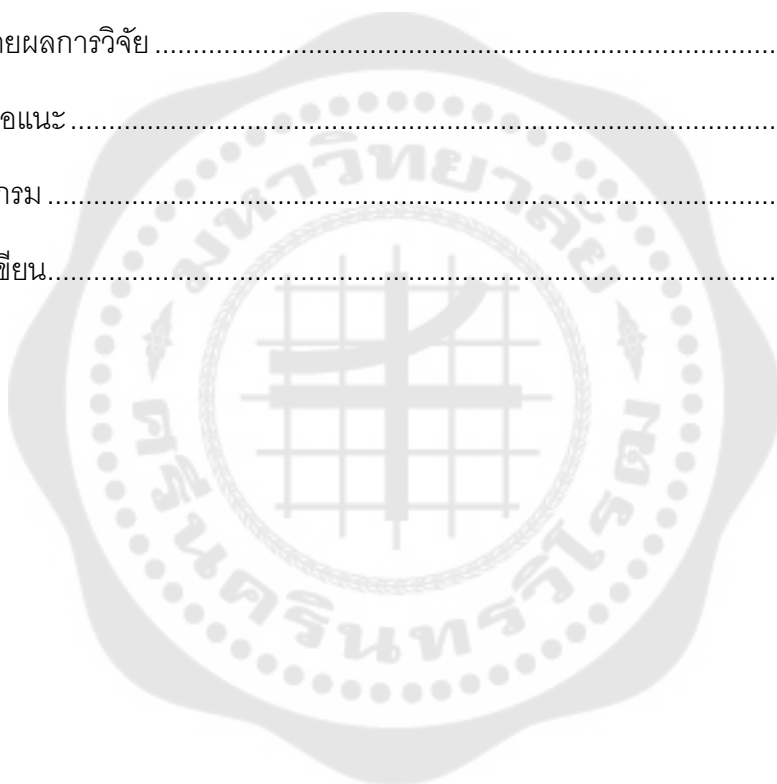
ปรีชา รัตนเดชอมร

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ	ช
สารบัญตาราง.....	ญ
สารบัญรูปภาพ	ฎ
บทที่ 1 บทนำ.....	1
1.1 ความสำคัญและความเป็นมาของงานวิจัย.....	1
1.2 วัตถุประสงค์ในการวิจัย.....	4
1.3 ขอบเขตของการวิจัย	4
1.4 ประโยชน์ที่คาดว่าจะได้รับ.....	5
1.5 ข้อจำกัดในการวิจัย.....	5
บทที่ 2 งานวิจัยที่เกี่ยวข้อง.....	7
2.1 ระบบแนะนำ	7
2.1.1 การกรองข้อมูลตามเนื้อหา (Content-based Filtering)	8
2.1.2 การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)	9
2.1.2.1 Memory-based	12
2.1.2.2 Model-based	14
2.1.2.3 Hybrid.....	14
2.1.3 การกรองข้อมูลแบบผสม (Hybrid Filtering)	14
2.2 ทฤษฎีในการสร้างแบบจำลอง	15

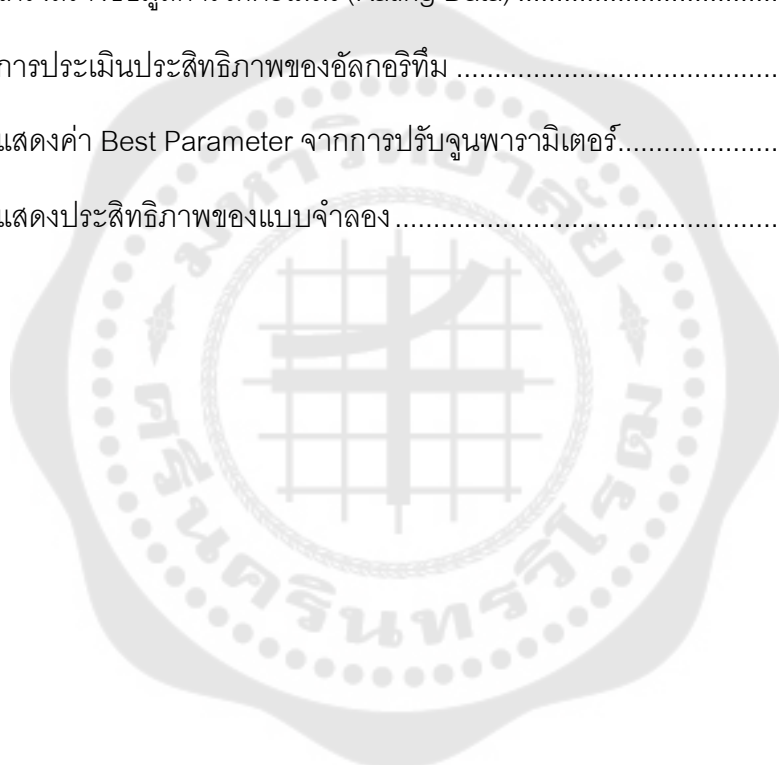
2.2.1 เทคนิคการกรองข้อมูลตามเนื้อหา	15
2.2.1.1 Term Frequency-Inverse Document Frequency (TF-IDF)	15
2.2.1.2 การวัดความคล้ายคลึง (Cosine Similarity).....	16
2.2.2 เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม	17
2.2.2.1 การแยกตัวประกอบเมทริกซ์ (Matrix Factorization)	17
2.2.2.1.1 Singular-value decomposition (SVD)	18
2.2.2.1.2 Non-negative Matrix Factorization (NMF).....	19
2.3 การวัดประสิทธิภาพของแบบจำลอง.....	20
2.3.1 Root Mean Squared Error (RMSE)	20
2.3.2 Mean Absolute Error (MAE).....	20
2.4 Surprise Library	21
2.5 งานวิจัยที่เกี่ยวข้อง	21
บทที่ 3 วิธีดำเนินการวิจัย.....	27
3.1 กระบวนการสร้างแบบจำลอง	27
3.2 การเก็บรวบรวมข้อมูล (Data Collection).....	30
3.3 การทำความสะอาดข้อมูล (Data Cleansing).....	32
3.4 การสำรวจข้อมูล (Exploratory Data Analysis: EDA).....	33
3.5 การเตรียมข้อมูล (Data Preparation)	39
3.6 การสร้างแบบจำลอง	41
3.6.1 เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering).....	41
3.6.2 เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)	42
3.6.3 เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering).....	43
3.7 สรุปผลการดำเนินงาน.....	44

บทที่ 4 ผลการดำเนินงานวิจัย.....	45
4.1 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา.....	45
4.2 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม.....	46
4.3 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบผสม	50
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	53
สรุปผลการวิจัย.....	53
อภิปรายผลการวิจัย	54
ข้อเสนอแนะ	57
บรรณานุกรม	58
ประวัติผู้เขียน.....	61



สารบัญตาราง

	หน้า
ตาราง 1 สรุปงานวิจัยที่เกี่ยวข้อง.....	26
ตาราง 2 แหล่งที่มาและขอบเขตข้อมูล.....	30
ตาราง 3 โครงสร้างข้อมูลรายวิชา (Course Data)	33
ตาราง 4 โครงสร้างข้อมูลการให้คะแนน (Rating Data)	36
ตาราง 5 การประเมินประสิทธิภาพของอัลกอริทึม	47
ตาราง 6 แสดงค่า Best Parameter จากการปรับจูนพารามิเตอร์.....	48
ตาราง 7 แสดงประสิทธิภาพของแบบจำลอง	48



สารบัญรูปภาพ

	หน้า
ภาพประกอบ 1 เว็บไซต์การเรียนรู้ Coursera.....	2
ภาพประกอบ 2 ตัวอย่างปัญหาจาก Coursera Community	3
ภาพประกอบ 3 แสดงการแนะนำภาพยนตร์โดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา	9
ภาพประกอบ 4 การแนะนำภาพยนตร์โดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม	10
ภาพประกอบ 5 การพิจารณาคะแนนแบบเกณฑ์เดียว (Single-criteria)	11
ภาพประกอบ 6 การพิจารณาคะแนนแบบหลายเกณฑ์ (Multi-criteria).....	11
ภาพประกอบ 7 การแนะนำสินค้าโดยใช้วิธี User-based แสดงความสัมพันธ์ระหว่างผู้ใช้.....	12
ภาพประกอบ 8 การแนะนำสินค้าโดยใช้วิธี Item-based แสดงความสัมพันธ์ระหว่างสินค้า	13
ภาพประกอบ 9 แสดงความสัมพันธ์ระหว่างผู้ใช้งาน และความสัมพันธ์ระหว่างผลิตภัณฑ์	13
ภาพประกอบ 10 เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering).....	14
ภาพประกอบ 11 การให้คะแนนของผู้ใช้งาน	17
ภาพประกอบ 12 แสดงความสัมพันธ์ของมุมระหว่าง Ratings Vectors	17
ภาพประกอบ 13 Matrix Factorization	18
ภาพประกอบ 14 แสดง Singular-value decomposition (SVD).....	19
ภาพประกอบ 15 Non-negative Matrix Factorization (NMF)	20
ภาพประกอบ 16 กระบวนการสร้างแบบจำลอง.....	29
ภาพประกอบ 17 การเชื่อมโยงความสัมพันธ์ของข้อมูล	32
ภาพประกอบ 18 ตัวอย่างข้อมูลรายวิชา	34
ภาพประกอบ 19 แสดงข้อมูลของภาษาที่ใช้เรียนในรายวิชา.....	34
ภาพประกอบ 20 แสดงจำนวนวิชาหลักและวิชารองที่เกี่ยวข้องในรายวิชา.....	35
ภาพประกอบ 21 จำนวนความถี่ของคำที่ปรากฏจากชื่อรายวิชา.....	35

ภาพประกอบ 22 จำนวนความถี่ของคำที่ปรากฏจากการเลือกข้อมูลในรายละเอียดที่มากขึ้น	36
ภาพประกอบ 23 ตัวอย่างข้อมูลการให้คะแนนจากผู้เรียน.....	37
ภาพประกอบ 24 ข้อมูลจำนวนการให้คะแนนจากผู้เรียน	37
ภาพประกอบ 25 รายการ 5 อันดับรายวิชาที่มีผู้เรียนให้คะแนนมากที่สุด	38
ภาพประกอบ 26 รายการ 15 อันดับสาขาวิชาที่มีจำนวนการให้คะแนนจากผู้เรียนมากที่สุด.....	38
ภาพประกอบ 27 รายการ 10 อันดับผู้เรียนที่มีการให้คะแนนรายวิชามากที่สุด	38
ภาพประกอบ 28 แสดงตัวอย่างข้อมูลรายละเอียดของรายวิชา.....	39
ภาพประกอบ 29 การทำ Text-Preprocessing ข้อมูล course_name	40
ภาพประกอบ 30 การทำ Text-Preprocessing โดยใช้ข้อมูลรายละเอียดของรายวิชา.....	40
ภาพประกอบ 31 ตัวอย่างข้อมูลการให้คะแนนรายวิชาจากผู้ใช้งาน.....	41
ภาพประกอบ 32 การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลตามเนื้อหา.....	42
ภาพประกอบ 33 กระบวนการแตกเวกเตอร์คำโดยใช้เทคนิค TF-IDF	42
ภาพประกอบ 34 การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม.....	43
ภาพประกอบ 35 แสดงแบบจำลองของเทคนิคการกรองแบบผสม	44
ภาพประกอบ 36 การแนะนำรายวิชาโดยใช้ข้อมูลชื่อรายวิชา.....	46
ภาพประกอบ 37 การแนะนำรายวิชาโดยใช้รายละเอียดรายวิชา	46
ภาพประกอบ 38 แสดงการเปรียบเทียบประสิทธิภาพของอัลกอริทึมโดยใช้ RMSE และ MAE ...	47
ภาพประกอบ 39 การค้นหาประวัติของผู้เรียน 286193	49
ภาพประกอบ 40 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมของผู้เรียน 286193.....	49
ภาพประกอบ 41 การค้นหาประวัติของผู้เรียน 4412.....	49
ภาพประกอบ 42 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมของผู้เรียน 4412	50
ภาพประกอบ 43 การค้นหาประวัติของผู้เรียน 95424	50

ภาพประกอบ 44 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมของผู้เรียน 95424	50
ภาพประกอบ 45 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 286193	51
ภาพประกอบ 46 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 4412	51
ภาพประกอบ 47 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 95424	52



บทที่ 1

บทนำ

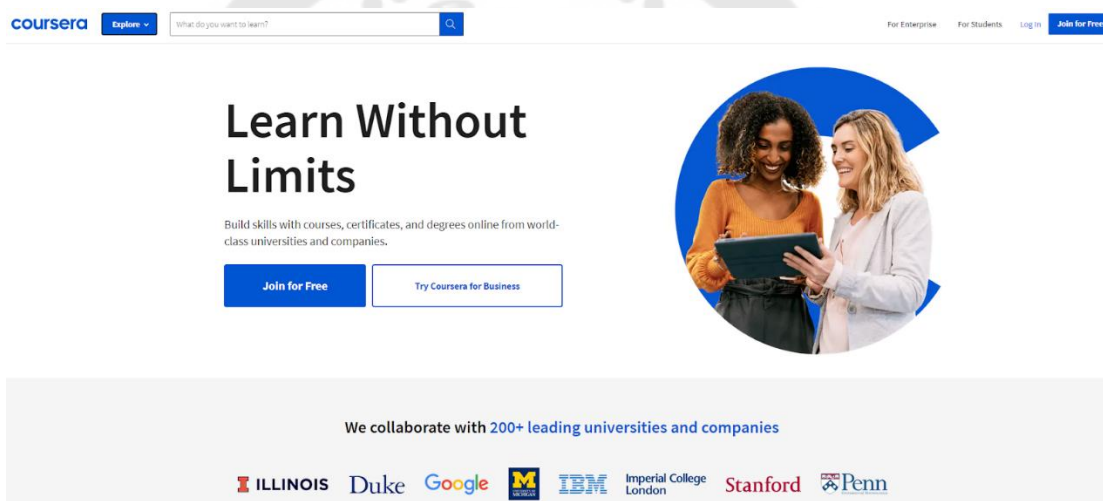
1.1 ความสำคัญและความเป็นมาของงานวิจัย

การพัฒนาทางด้านเทคโนโลยีสารสนเทศและการสื่อสารมีมาอย่างต่อเนื่อง ทำให้มีสินค้าหรือบริการใหม่ๆ เกิดขึ้นอยู่ตลอดเวลา การนำเสนอสินค้าหรือบริการเหล่านี้ก็เพื่อตอบสนองต่อความต้องการของคนในทุกกลุ่ม โดยไม่จำกัดเพศหรือวัย จนกระทั่งวิวัฒนาการของอินเทอร์เน็ตได้เข้ามามีบทบาทต่อชีวิตประจำวันของมนุษย์ ส่งผลให้เกิดการแลกเปลี่ยน การกระจายข้อมูลข่าวสาร การซื้อขายสินค้าผ่านทางอินเทอร์เน็ต การเรียนรู้ผ่านทางช่องทางออนไลน์ นอกจากนี้ยังทำให้เกิดเครือข่ายทางสังคมระหว่างผู้ใช้งานอินเทอร์เน็ตขึ้นอีกด้วย การเติบโตดังกล่าวนี้ทำให้จำนวนข้อมูลมีปริมาณเพิ่มขึ้นอย่างรวดเร็ว ส่งผลให้ผู้ใช้งานเกิดการปรับตัวต่อการเปลี่ยนแปลงเพื่อก้าวให้ทันต่อความต้องการทางด้านสินค้า การบริการ ที่มีความหลากหลายมากยิ่งขึ้น รวมไปถึงรูปแบบการเรียนการสอนในแหล่งเรียนรู้ออนไลน์ก็มีการพัฒนาขึ้นมาบนเว็บไซต์จำนวนมากเช่นกัน ส่งผลทำให้สถาบันต่างๆ ทั้งในประเทศและต่างประเทศได้พัฒนาระบบหลักสูตรออนไลน์ขึ้นมามากมาย โดยใช้อินเทอร์เน็ตเป็นเครื่องมือในกระบวนการเรียนการสอน

ระบบการศึกษาออนไลน์ (Online Learning) เป็นการเรียนรู้ที่สามารถถ่ายทอดเนื้อหา รูปภาพ วิดีโอ หรือการใช้สื่อหลายๆ ประเภท ร่วมกับการสนทนาแลกเปลี่ยนความคิดเห็นผ่านทางเทคโนโลยีสมัยใหม่ ซึ่งปัจจุบันการเรียนรู้เฉพาะหลักสูตรภายในห้องเรียนอาจไม่เพียงพอต่อการทำงานหรือนำมาประยุกต์ใช้งานได้ ดังนั้นการเรียนรู้จากแหล่งเรียนรู้ภายนอก ได้เข้ามามีบทบาทสำคัญทำให้เกิดการเรียนรู้ ยกระดับความรู้ ความคิด การวิเคราะห์ เพื่อพัฒนาทักษะของผู้เรียนให้มากยิ่งขึ้น ซึ่งการศึกษาออนไลน์เป็นส่วนหนึ่งที่ผู้คนนิยมค้นหาและใช้บริการทางอินเทอร์เน็ตเพื่อการเรียนรู้ และมีความจำเป็นมากในปัจจุบัน เพราะการใช้เทคโนโลยีสมัยใหม่ช่วยส่งเสริมให้เกิดการเรียนรู้ตลอดชีวิต (Lifelong Learning) จึงทำให้เกิดสิ่งที่จะช่วยขับเคลื่อนการเรียนรู้ดังกล่าว นั่นคือ แหล่งเรียนรู้ออนไลน์ (Online Learning Platform)

แหล่งเรียนรู้ออนไลน์ หรือที่รู้จักกันอีกชื่อหนึ่งคือ MOOC (Massive Open Online Course) เป็นแพลตฟอร์มที่ผู้ให้บริการทำการรวบรวมคอร์สเรียน โดยมีผู้เชี่ยวชาญในสาขาวิชาต่างๆ ทั้งจากสถาบันการศึกษาระดับโลก และบริษัทเอกชนชั้นนำ มาทำการเปิดหลักสูตรออนไลน์รายวิชาต่างๆ บนแพลตฟอร์มของผู้ให้บริการ เพื่อให้ผู้ใช้งานหรือผู้เรียนทั่วโลกสามารถเข้ามาศึกษาหาความรู้ได้จากแพลตฟอร์มดังกล่าวผ่านช่องทางออนไลน์ นอกจากนี้แหล่งเรียนรู้ออนไลน์ทำให้เกิดสังคมแห่งการเรียนรู้ โดยมุ่งเน้นให้บริการทางการศึกษาที่เหมาะสมกับคนทุกกลุ่มวัย

และมีหลากหลายสาขาวิชาซีพีให้ทำการเรียนรู้ ผู้เรียนสามารถเลือกเรียนรู้โดยไม่มีข้อจำกัดในเรื่องของเวลาและสถานที่มาเกี่ยวข้อง อีกทั้งยังสามารถเรียนรู้ด้วยตนเองได้ตามความต้องการ โดยมีเทคโนโลยีเป็นเครื่องมือในการเข้าถึงการให้บริการทางการศึกษา เพื่อให้ผู้เรียนได้เรียนรู้ด้วยตนเองอย่างต่อเนื่อง ทำให้ผู้เรียนเกิดความรู้ สามารถพัฒนาทักษะใหม่ๆ และสร้างแรงจูงใจในการเรียนรู้ให้กับผู้เรียน ตัวอย่างเว็บไซต์การเรียนรู้ที่เป็นที่รู้จักในประเทศไทยได้แก่ Thai Mooc, Chula Mooc, SkillLane เป็นต้น และตัวอย่างเว็บไซต์การเรียนรู้ที่เป็นที่รู้จักทั่วโลกได้แก่ Coursera, Edx, Udemy, Udacity เป็นต้น จากภาพประกอบ 1 คือตัวอย่างเว็บไซต์การเรียนรู้ Coursera.org ที่มีรายวิชาครอบคลุมทุกสาขาวิชา สามารถทำการเรียนรู้ได้โดยไม่มีข้อจำกัดทั้งเรื่องเวลาและสถานที่ ผู้เรียนสามารถเข้าเรียนในรายวิชาที่ต้องการได้อย่างอิสระ



ภาพประกอบ 1 เว็บไซต์การเรียนรู้ Coursera

ที่มา : ("Coursera.org," 2021)

สื่อการเรียนและแหล่งเรียนรู้ที่มีจำนวนเพิ่มมากขึ้นนั้น ข้อมูลส่วนใหญ่ถูกจัดเก็บในรูปแบบดิจิทัล การใช้ประโยชน์จากข้อมูลที่มีมากมายเหล่านี้ สามารถนำมาวิเคราะห์หาพฤติกรรม การใช้ข้อมูลเพื่อหาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลนั้น และสามารถนำมาต่อยอดเพื่อใช้เป็นแนวทางในการแก้ปัญหาหรือช่วยในการตัดสินใจได้ ในขณะเดียวกันการที่หลักสูตรการเรียนรายวิชาต่างๆ และแหล่งเรียนรู้มีมากมายนั้น การที่ผู้เรียนจะเลือกเรียนวิชาใดวิชาหนึ่งได้ตรงกับความต้องการจะเป็นเรื่องยาก ซึ่งผู้เรียนอาจจะได้รับการเรียนในวิชาที่ไม่เกี่ยวข้องหรือไม่ตรงกับความต้องการ หลังจากที่มีการค้นหาและเลือกลงทะเบียนเรียนไปแล้วในช่วงเวลาหนึ่ง จากปัญหา

ดังกล่าวนี้ทำให้เกิดแนวคิดในการคัดกรองข้อมูล โดยอาศัยเทคโนโลยีการคัดเลือกรูปแบบของข้อมูลที่มีความเหมาะสมตรงกับที่ผู้เรียนต้องการ เพื่อส่งผลให้เกิดความพึงพอใจและตัดสินใจเลือกเรียนในรายวิชาที่มีการแนะนำได้ จากภาพประกอบ 2 คือ ตัวอย่างปัญหาในเว็บไซต์ Coursera Community ที่มีผู้เรียนมาตั้งคำถามเกี่ยวกับการเลือกเรียนในรายวิชาที่สนใจ แต่ไม่สามารถทำการเลือกเรียนในวิชาใดได้ เนื่องจากมีวิชาที่เกี่ยวข้องมากเกินไป

How to choose the right Python course(s)

Hi

I'm new to Coursera and interested in learning Python. I've already had some training in a boot camp for about a week so I'm probably still a beginner.

I've just done a search for Python and find 430 courses ! My problem is trying to choose the right beginner course.

"Programming for Everybody (Getting Started with Python)" but that seems to be a bit too basic for me.

Can anyone advise me on how to select a course at the appropriate level without having to enrol, watch videos and read for a few hours before i can decide whether or not its right for me.

I'm looking for a learning path, probably to take me up to advanced level.

Thank you

ภาพประกอบ 2 ตัวอย่างปัญหาจาก Coursera Community

ที่มา : ("Coursera.org," 2021)

ระบบแนะนำ (Recommendation System) คือระบบที่ช่วยนำเสนอสิ่งที่คาดว่าผู้ใช้น่าจะสนใจ เพื่อเพิ่มโอกาสที่ผู้ใช้จะเลือกในรายการนั้นๆ ระบบแนะนำมีบทบาทอย่างมากและได้ถูกนำไปประยุกต์ใช้กับทั้งภาคธุรกิจ ภาคบริการ รวมถึงทางภาคการศึกษา ซึ่งได้ถูกนำมาใช้กันอย่างแพร่หลาย โดยทั่วไปแล้วพื้นฐานของ Recommendation System จะใช้เทคนิคการกรองแบบมีเนื้อหา (Content-based Filtering) วิธีการนี้ได้ทำการแนะนำข้อมูล โดยหาความสัมพันธ์ระหว่างเนื้อหาข้อมูลหรือคุณลักษณะของข้อมูลกับสิ่งที่ผู้ใช้น่าจะสนใจ และเทคนิคการกรองร่วมกัน (Collaborative Filtering) โดยพิจารณาจากประวัติผู้ใช้คนอื่นที่มีพฤติกรรมคล้ายกัน และเคยให้คะแนนความนิยมกับข้อมูลสินค้าหรือการบริการนั้นๆ จึงทำการแนะนำสินค้านั้นให้กับลูกค้าคนปัจจุบัน

นอกจากนี้ปัญหาหรือความท้าทายที่อาจพบได้ในระบบผู้แนะนำ เช่น ปัญหาความเบาบางของข้อมูล (Data Sparsity) ซึ่งอาจเกิดจากข้อมูลที่มีอยู่น้อยเกินไป และปัญหาจากการที่ผู้ใช้หรือมีรายการใหม่เกิดขึ้น (Cold Start) ทำให้ไม่สามารถแนะนำสินค้าหรือบริการนั้นได้ เนื่องจากไม่มีประวัติการใช้งานหรือไม่มีผู้ใช้งานใดๆ ให้คะแนนในรายการนั้นๆ นอกจากปัญหา

ดังกล่าวที่พบ ระบบผู้แนะนำยังมีความท้าทายสำหรับการปรับปรุงกระบวนการแนะนำ และทำการเพิ่มประสิทธิภาพความแม่นยำของการแนะนำ โดยมีการใช้เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering) เป็นวิธีการที่ผสมผสานกับวิธีในข้างต้น หรือวิธีอื่นๆ ที่มีขั้นตอนหรือกระบวนการแนะนำที่ซับซ้อน เพื่อแก้ปัญหาที่อาจเกิดขึ้นจากเทคนิคที่กล่าวไปในข้างต้น รวมทั้งส่งผลให้ผู้มีความพึงพอใจและได้เลือกวิชาในการเรียนที่ตรงกับความต้องการของตนเองได้อีกด้วย

โดยสรุปงานวิจัยครั้งนี้มีแนวคิดในการนำเสนออัลกอริทึมที่ใช้ในการแนะนำรายวิชา เพื่อช่วยให้ระบบนำเสนอวิชาที่ตรงกับความต้องการของผู้ใช้อย่างแท้จริง ซึ่งใช้ทั้งข้อมูลจากรายวิชาและการให้คะแนนของผู้เรียน โดยคาดหวังว่าอัลกอริทึมที่พัฒนาขึ้นมาสามารถนำมาเปรียบเทียบประสิทธิภาพการแนะนำของแต่ละวิธี และเพิ่มความน่าสนใจในการแนะนำรายวิชา ซึ่งก่อให้เกิดประโยชน์ช่วยอำนวยความสะดวก ลดระยะเวลาในการค้นหา และทำให้ได้รายวิชาการเรียนรู้ที่ตรงกับความต้องการของผู้เรียนมากยิ่งขึ้น นอกจากนี้ยังสามารถนำความรู้จากการศึกษาระบบการแนะนำไปประยุกต์ใช้กับชุดข้อมูลอื่นๆ ในรูปแบบสินค้าหรือบริการได้ ตัวอย่างเช่น ข้อมูลสินค้า ข้อมูลภาพยนตร์ ร้านอาหาร การท่องเที่ยว หนังสือ เครื่องสำอาง เพลง เป็นต้น

1.2 วัตถุประสงค์ในการวิจัย

เพื่อศึกษาและเปรียบเทียบเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) มาใช้ในการแนะนำรายวิชาบนเว็บไซต์ Coursera เพื่อสามารถทำให้ผู้เรียนสามารถตัดสินใจลงทะเบียนได้ง่ายขึ้น และตรงกับความต้องการในการค้นหามากที่สุด

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้เป็นการศึกษาและพัฒนาวิธีการแนะนำแบบผสมผสาน โดยใช้ข้อมูลจากคุณลักษณะของเนื้อหาและคะแนนของผู้ใช้งาน เพื่อให้ได้ผลการแนะนำที่หลากหลายและมีแนวโน้มตรงต่อความต้องการของผู้ใช้ โดยมีขั้นตอน ดังต่อไปนี้

1. ข้อมูลที่ใช้ในงานวิจัยนี้ เป็นข้อมูลรายวิชาบนเว็บไซต์ Coursera ซึ่งเป็นชุดข้อมูลสาธารณะจาก kaggle.com โดยผู้วิจัยได้ทำการรวบรวมข้อมูลที่ได้ทั้งหมด 3 แหล่งชุดข้อมูลประกอบด้วย ข้อมูลรายวิชา ข้อมูลสถาบัน ระดับรายวิชา วิชาที่เกี่ยวข้อง การให้คะแนนรายวิชาของผู้เรียน เมื่อนำชุดข้อมูลทั้งหมดมารวมกันได้ชุดข้อมูลที่ประกอบด้วยข้อมูลรายวิชาทั้งหมด

5,599 รายวิชา การให้คะแนนรายวิชาจากผู้ใช้งานที่ทำการให้คะแนนจำนวน 506,116 รายการ จำนวนผู้ให้คะแนน 287,808 คน และจำนวนรายวิชาที่มีการให้คะแนน 604 รายวิชา

2. การเลือกข้อมูลทั้ง 3 แหล่งนี้ เนื่องจากพบว่าในแหล่งข้อมูลที่ 1 มี 2 ชุดข้อมูล ซึ่งมีข้อมูลจำนวนรายวิชาทั้งหมด 623 รายวิชา และข้อมูลที่มีการให้คะแนนจากผู้เรียนมีเพียง 604 รายวิชา แหล่งข้อมูลที่ 2 มีจำนวน 5,164 รายวิชา และแหล่งข้อมูลที่ 3 มีจำนวน 891 รายวิชา และเนื่องจากเป็นข้อมูลที่ได้มาจากหลายแหล่งข้อมูล ซึ่งมี Feature ที่เหมือนและแตกต่างกัน จึงทำการเลือกแหล่งข้อมูลดังกล่าว เพื่อทำให้เกิดข้อมูลที่มีความหลากหลาย และมีตัวเลือกที่มากขึ้น โดยการนำชุดข้อมูลที่ได้มาทำการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม และเทคนิคการกรองข้อมูลแบบผสม

3. ข้อมูลที่นำมาใช้ในการทดลองเป็นชุดข้อมูลที่ได้มาจากหลายแหล่งข้อมูล จึงต้องมีการทำการรวบรวมข้อมูลเพื่อให้ข้อมูลมีความหลากหลาย จากนั้นต้องทำการกระบวนการจัดการข้อมูลเช่น การทำความสะอาดข้อมูล การสำรวจข้อมูล และการเตรียมข้อมูล เป็นต้น เพื่อให้ข้อมูลมีความสอดคล้องกันและไม่มีปัจจัยในเรื่องของข้อมูลซ้ำซ้อน ซึ่งอาจส่งผลทำให้ผลการสร้างระบบคลาดเคลื่อนได้

4. ทำการพัฒนาสร้างแบบจำลอง เพื่อเปรียบเทียบระบบแนะนำข้อมูลด้วยวิธีการแบบผสมผสาน งานวิจัยนี้ปรับปรุงประสิทธิภาพของระบบคำแนะนำที่ใช้วิธีพิจารณาจากข้อมูลรายวิชาหรือเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยการใช้ข้อมูลจากการให้คะแนนของผู้เรียนในรายวิชานั้นๆ และการค้นหาจากรายละเอียดรายวิชาเพื่อทำการแนะนำได้มากยิ่งขึ้น

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำแบบจำลองที่พัฒนาขึ้น มาช่วยเพิ่มประสิทธิภาพในการแนะนำรายวิชาให้ตรงกับความต้องการของผู้เรียนมากยิ่งขึ้น
2. สามารถนำแบบจำลองที่พัฒนาขึ้น ปรับปรุงให้เป็นแนวทางในการประยุกต์ใช้กับระบบแนะนำหรือชุดข้อมูลในโดเมนอื่นๆ ในรูปแบบสินค้าหรือบริการได้ ตัวอย่างเช่น การแนะนำข้อมูลผลิตภัณฑ์ ข้อมูลภาพยนตร์ ร้านอาหาร การท่องเที่ยว หนังสือ เครื่องสำอาง เพลง เป็นต้น

1.5 ข้อจำกัดในการวิจัย

ข้อจำกัดในเรื่องของภาษา นอกเหนือจากอักขระภาษาอังกฤษซึ่งเป็นภาษาสากล ซึ่งในบางรายวิชามีตัวอักษรที่เป็นอักขระภาษาอื่น ตัวอย่างเช่น รัสเซีย สเปน จีน ฝรั่งเศส ญี่ปุ่น เป็นต้น

ทำให้ไม่สามารถจัดการกับรายวิชาที่มีอักขระภาษาอื่นได้ หรือได้ผลไม่ดีนัก จึงต้องนำข้อมูลใน
รายละเอียดอื่นๆ มาทำการสร้างแบบจำลองร่วมด้วย



บทที่ 2

งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ศึกษางานที่เกี่ยวข้องกับการพัฒนาแบบจำลองการผสมผสานในการแนะนำแหล่งความรู้ออนไลน์รายวิชาภายในเว็บไซต์ Coursera โดยใช้เทคนิค การกรองข้อมูลตามเนื้อหา (Content-based Filtering) และการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) มาทำงานร่วมกัน โดยนำเสนอตามหัวข้อดังต่อไปนี้

1. ระบบแนะนำ
2. ทฤษฎีการสร้างแบบจำลอง
3. การวัดประสิทธิภาพของแบบจำลอง
4. Surprise Library
5. งานวิจัยที่เกี่ยวข้อง

2.1 ระบบแนะนำ

ระบบแนะนำ (Recommendation System) คือ ระบบการนำเสนอข้อมูล ผลิตภัณฑ์และบริการที่คาดว่าจะตรงกับความต้องการหรือความสนใจของผู้ใช้ ทำให้มีผลต่อการตัดสินใจในด้านต่างๆ และเพิ่มโอกาสที่ผู้ใช้จะทำการเลือกกรายการนั้นๆ ระบบแนะนำได้ดำเนินการทำงานในเชิงรุกเพื่อนำเสนอข้อมูลให้กับผู้ใช้อย่างอัตโนมัติ ระบบการแนะนำนี้ล้วนถูกคัดสรรมาเพื่อตอบโจทย์ความต้องการของผู้ใช้ ช่วยให้ผู้ใช้สามารถค้นหาผลิตภัณฑ์ที่ต้องการได้อย่างรวดเร็ว ลดภาระและระยะเวลาในการค้นหา โดยใช้การอ้างอิงข้อมูลจากสิ่งที่ผู้ใช้สนใจ และเรียนรู้ผ่านพฤติกรรมจากประวัติของผู้ใช้ นอกจากนี้ยังสามารถทำการคาดเดาได้ว่าในอนาคต ผู้ใช้น่าจะสนใจผลิตภัณฑ์อื่นที่คล้ายคลึงกันกับสิ่งที่เคยเลือกมาก่อนหน้านี้แล้ว นอกจากนี้ประโยชน์ของระบบการแนะนำผลิตภัณฑ์นี้ ยังช่วยขับเคลื่อนข้อมูลในเชิงลึกและสร้างความภักดีในผลิตภัณฑ์ได้

ระบบแนะนำมีบทบาทอย่างมากและได้ถูกนำไปประยุกต์ใช้กับทั้งภาคธุรกิจ ภาคบริการ รวมถึงภาคการศึกษา ซึ่งระบบแนะนำที่เป็นที่รู้จักกันดีได้แก่ ระบบแนะนำของ Amazon การชมภาพยนตร์สตรีมมิงบนแพลตฟอร์ม Netflix เพลย์ลิสต์แนะนำเพลงจาก Spotify หรือการแนะนำบทความจากเว็บไซต์ Medium เป็นต้น

ระบบแนะนำดำเนินการทำงานโดยใช้หลักการซึ่งเป็นที่นิยมใช้ 3 วิธีดังนี้

1. การกรองข้อมูลตามเนื้อหา (Content-based Filtering)
2. การกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering)

3. การกรองข้อมูลแบบผสม (Hybrid Filtering)

นอกจากนี้ในบางงานวิจัยยังมีการใช้เทคนิค Knowledge-based เป็นการแนะนำที่คล้ายคลึงกับเทคนิคการกรองข้อมูลตามเนื้อหา แต่พิจารณาระหว่างข้อมูลคุณลักษณะของผลิตภัณฑ์หรือการบริการในเรื่องใดเรื่องหนึ่งที่ตรงกับลักษณะของผู้ใช้งานในปัจจุบันประกอบกัน เพื่อให้ได้คำแนะนำที่ตอบปัญหาและตรงกับความต้องการของผู้ใช้มากที่สุด มักใช้กับผลิตภัณฑ์หรือการบริการที่ไม่ได้มีการซื้อขายกันบ่อยนัก และมักจะมีเงื่อนไขที่ซับซ้อน ตัวอย่างเช่น การซื้อรถยนต์ บ้าน หรือตั๋วเครื่องบิน เป็นต้น จากตัวอย่างเหล่านี้ การให้คะแนนจากผู้ใช้งานไม่เพียงพอในการตัดสินใจ จึงจำเป็นต้องอาศัยข้อมูลอื่นมาประกอบเพิ่มเติม โดยที่ผู้ใช้งานอาจต้องระบุเงื่อนไขที่ต้องการเช่น วันเดินทาง ประเภทตัว หรืองบประมาณในการซื้อทรัพย์สิน เมื่อได้ข้อมูลที่ครบถ้วนจึงจะนำข้อมูลเหล่านั้นมาใช้ในการคำนวณ เพื่อให้ได้คำแนะนำที่ตรงกับเงื่อนไขของผู้ใช้งานมากที่สุด

นอกจากนี้ยังมีวิธี Demographic-based ที่คล้ายคลึงกับเทคนิค Collaborative Filtering แต่เป็นการเปรียบเทียบความคล้ายคลึงกันของผู้ใช้งาน โดยอาศัยลักษณะเฉพาะข้อมูลตัวบุคคลเข้ามาเกี่ยวข้องด้วย ตัวอย่างเช่น อายุ เพศ อาชีพ เชื้อชาติ พื้นที่ เป็นต้น (เปรมชัยสวัสดิ์, 2020)

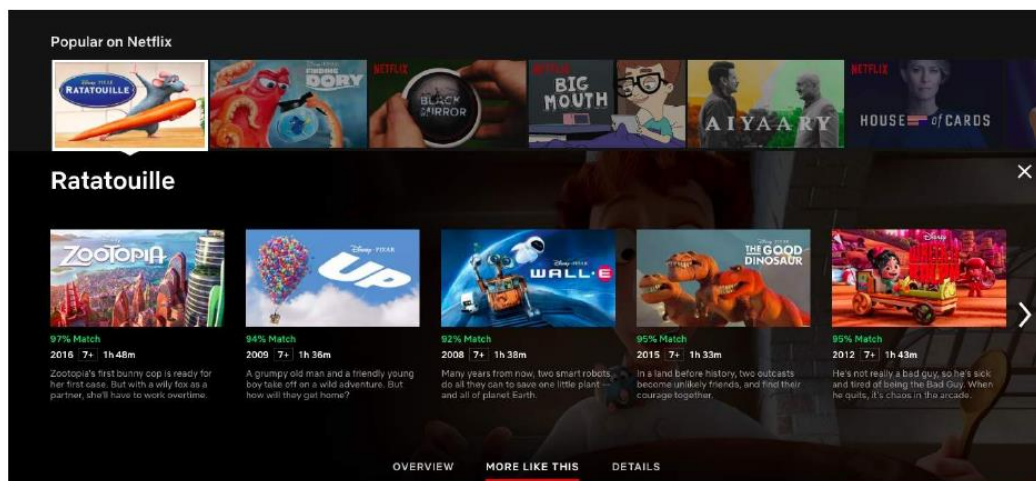
2.1.1 การกรองข้อมูลตามเนื้อหา (Content-based Filtering)

เทคนิคการกรองข้อมูลตามเนื้อหาเป็นวิธีการพื้นฐานโดยนำข้อมูลที่ได้มาจากเนื้อหาของข้อมูลรายการนั้นๆ มาหาความสัมพันธ์ระหว่างเนื้อหาของข้อมูลกับสิ่งที่ผู้เรียนสนใจ ซึ่งวิธีการนี้จะให้ความสนใจกับคุณลักษณะของข้อมูลเป็นส่วนสำคัญ ตัวอย่างเช่นระบบแนะนำภาพยนตร์ ระบบจะทำการแนะนำจากเนื้อหาของภาพยนตร์ที่มีความคล้ายคลึงกันกับสิ่งที่ผู้ใช้เคยดูมาก่อน โดยอ้างอิงจากประเภทของภาพยนตร์ ผู้กำกับ นักแสดง เป็นต้น

เทคนิคการกรองข้อมูลตามเนื้อหา ไม่ต้องการโปรไฟล์ของผู้ใช้งานรายอื่น เนื่องจากไม่มีผลต่อการแนะนำ ภาพยนตร์ที่ได้ทำการแนะนำค่อนข้างตรงกับความต้องการและรสนิยมของผู้ใช้งาน และการแนะนำภาพยนตร์รายการใหม่สามารถให้คำแนะนำโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหาได้ แม้ว่าจะไม่มีผู้ใช้คนใดเคยให้คะแนนหรือทำการรีวิวมาก่อน เพราะจะพิจารณาจากความคล้ายคลึงกันของคุณลักษณะกับภาพยนตร์เรื่องเดิมที่เคยดูมาก่อนหน้านั้น

จากตัวอย่างภาพประกอบ 3 คือการแนะนำภาพยนตร์ในแพลตฟอร์มของ Netflix โดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา จะเห็นได้ว่าเมื่อมีการชมภาพยนตร์เรื่อง Ratatouille ระบบจะแนะนำภาพยนตร์ที่คล้ายกับ Ratatouille ให้กับผู้ใช้โดยเป็นการแนะนำ 5 อันดับแรกซึ่ง

เป็นภาพยนตร์ประเภท Animation และผลิตโดย Disney Pixar ซึ่งคุณลักษณะของข้อมูลเหล่านี้ ถูกเลือกมาจากข้อมูลภาพยนตร์เรื่อง Ratatouille



ภาพประกอบ 3 แสดงการแนะนำภาพยนตร์โดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา

ที่มา : (Banik, 2018, p. 197)

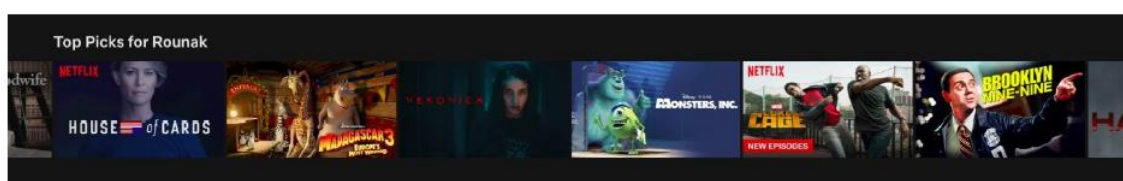
สำหรับข้อเสียของการใช้วิธีนี้ อาจพบได้ว่าไม่มีความหลากหลายเนื่องจากไม่สามารถแนะนำภาพยนตร์ที่แตกต่างจากเรื่องเดิมมากจนเกินไป ทำให้ผู้ใช้งานได้รับการแนะนำภาพยนตร์ประเภทเดิมซึ่งมีความหลากหลายค่อนข้างน้อย และสำหรับวิธีนี้ใช้เวลาในการเตรียมข้อมูลมาก เพื่อทำการสร้างคุณลักษณะที่เหมาะสมกับภาพยนตร์ สำหรับการใช้นี้ใช้เทคนิคของการกรองข้อมูลตามเนื้อหา อาจไม่เหมาะสมกับการแนะนำข้อมูลที่ไม่ได้อยู่ในรูปแบบของข้อความ เช่น วิดีโอ เพลง เป็นต้น

2.1.2 การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม วิธีการนี้ได้ทำการแนะนำข้อมูลให้กับผู้ใช้งาน โดยพิจารณาจากประวัติของผู้ใช้ที่มีพฤติกรรมที่คล้ายคลึงกัน มาประกอบในคำแนะนำ หากเคยค้นหาหรือมีประสบการณ์ในผลิตภัณฑ์เหล่านั้นและทำการให้คะแนน (Rating) ไว้ก่อนหน้านี้แล้ว ผู้ใช้งานจะได้รับการแนะนำในผลิตภัณฑ์นั้น แม้ว่าไม่เคยให้คะแนนมาก่อน แต่ได้รับการให้คะแนนมาจากผู้ใช้งานรายอื่นที่มีความสนใจหรือมีความชอบที่เหมือนกัน

ตัวอย่างเช่นจากภาพประกอบ 4 พบว่าระบบของ Netflix ได้ทำการแนะนำภาพยนตร์อื่นๆ เพิ่มเติม เนื่องจากว่าผู้ใช้งานไม่ได้ดูภาพยนตร์ประเภทเดียว แต่อาจมีการรับชมประเภทอื่นด้วยเช่น Drama, Comedy เป็นต้น ระบบ Netflix จึงทำการแนะนำภาพยนตร์แยก

ออกมาต่างหาก ซึ่งเป็นภาพยนตร์ที่ผู้ใช้งานอาจไม่เคยดูมาก่อน แต่ใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม โดยการระบุผู้ใช้ที่คล้ายกับผู้ใช้งานปัจจุบัน และแนะนำภาพยนตร์ที่ผู้ใช้คนอื่นดูแล้ว แต่ผู้ใช้งานปัจจุบันยังไม่ได้ดู



ภาพประกอบ 4 การแนะนำภาพยนตร์โดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม

ที่มา : (Banik, 2018, p. 198)

นอกจากนี้หากระบบมีการให้คะแนนกับผลิตภัณฑ์หรือบริการนั้นๆ ก็สามารถนำมาคิดคำนวณ เพื่อหาความชอบหรือความสนใจที่เหมือนกัน โดยแบ่งประเภทของการให้คะแนนออกเป็น 2 แบบคือ

1. การให้คะแนนจากผู้ใช้งานทางตรง (Explicit Rating) เป็นการให้คะแนนสินค้าหรือบริการ เพื่อวัดระดับความพึงพอใจโดยมีค่าคะแนนตั้งแต่ 0-5 คะแนน ตัวอย่างเช่น หากผู้ใช้งาน A มีการให้คะแนนวีวผลิตภัณฑ์ M 5/5 คะแนน หมายความว่าผู้ใช้งาน A มีแนวโน้มที่ชื่นชอบผลิตภัณฑ์ M มาก แต่หากผู้ใช้งาน A มีการให้คะแนนผลิตภัณฑ์ N ด้วยคะแนน 1/5 ทำให้ทำการคาดเดาได้ว่าผู้ใช้ A อาจไม่ชอบในสินค้า N ก็เป็นไปได้

การให้คะแนนจากผู้ใช้งานทางตรงที่ใช้เกณฑ์คะแนนแบบนี้ เป็นการใช้ค่าคะแนนโดยรวมเพียงค่าเดียว สามารถพิจารณาว่าเป็นระบบผู้แนะนำแบบเกณฑ์เดียว (Single-criteria) ดังแสดงในภาพประกอบ 5 เป็นการให้คะแนนภาพยนตร์เรื่องต่างๆ จากผู้ใช้งาน ในขณะที่บางระบบสามารถแยกความพึงพอใจของผู้ใช้ โดยพิจารณาจากความชอบที่ผู้ใช้มีต่อคุณลักษณะต่างๆ ของไอเทมที่มีมากกว่าหนึ่งคุณลักษณะ เรียกระบบการแนะนำแบบหลายเกณฑ์นี้ว่า (Multi-criteria) ดังแสดงในภาพประกอบ 6 ซึ่งแบ่งการให้คะแนนลึกลงไปในแต่ละคุณลักษณะของภาพยนตร์ เช่น Overall, Story, Action, Direction และ Visual Effect เป็นต้น

	Wanted	WALL-E	Star Wars	Seven	Fargo	
Target user → Alice	5	7	5	7	?	← Ratings to be predicted
User most similar to the target user → John	5	7	5	7	9	← Ratings to be used in prediction
Mason	6	6	6	6	5	
:	:	:	:	:	:	

ภาพประกอบ 5 การพิจารณาคะแนนแบบเกณฑ์เดียว (Single-criteria)

ที่มา : (Ricci, Rokach, & Shapira, 2015, p.853)

	Wanted	WALL-E	Star Wars	Seven	Fargo	
Target user → Alice	5,2,2,8,8	7,5,5,9,9	5,2,2,8,8	7,5,5,9,9	7,2,2,2,2	← Ratings to be predicted
User most similar to the target user → John	5,8,8,2,2	7,9,9,5,5	5,8,8,2,2	7,9,9,5,5	9,8,8,10,10	
Mason	6,3,3,9,9	6,4,4,8,8	6,3,3,9,9	6,4,4,8,8	5,2,2,8,8	← Ratings to be used in prediction
:	:	:	:	:	:	

ภาพประกอบ 6 การพิจารณาคะแนนแบบหลายเกณฑ์ (Multi-criteria)

ที่มา : (Ricci et al., 2015, p.854)

ปัญหาจากการให้คะแนนสินค้าเหล่านี้ ผู้ใช้งานอาจไม่ได้ทำการซื้อสินค้าหรือใช้บริการทั้งหมดภายในระบบ ซึ่งเกิดขึ้นในกรณีที่มีสินค้าปริมาณมาก จึงทำให้ข้อมูลเกิดความ Sparse ในสินค้าบางรายการที่ไม่ได้มีการให้คะแนนได้

2. การให้คะแนนจากผู้ใช้งานทางอ้อม (Implicit Rating) เป็นการให้คะแนนโดยวัดจากเหตุการณ์ที่เกิดขึ้นจากการที่ผู้ใช้งานทำรายการในระบบเช่น การซื้อ การใส่สินค้าไว้ในตะกร้าหรือการเข้าดูผลิตภัณฑ์นั้นๆ การให้คะแนนในทางอ้อมนี้มักจะมีปริมาณข้อมูลที่เยอะกว่าการให้คะแนนในทางตรง มีข้อดีคือสามารถใช้ได้กับผู้ใช้ใหม่ในระบบ แต่ข้อเสียคือทำให้ไม่สามารถทราบได้ว่าผู้ใช้ชอบในผลิตภัณฑ์นี้หรือไม่ ทำได้แค่เดาเท่านั้น

ข้อดีของการใช้วิธีการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ทำให้สามารถประยุกต์ใช้ได้กับทุกประเภทของผลิตภัณฑ์หรือการบริการ ไม่ต้องทำรายการของข้อมูลเพื่ออธิบายคุณลักษณะของสินค้า แต่มักมีข้อจำกัดของวิธีการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมคือ เมื่อสินค้านั้นมีการให้คะแนนน้อย หรือสินค้าที่ยังไม่เคยมีผู้ใช้งานให้คะแนนทำให้สามารถทำการแนะนำในรายการนั้น

ได้ยาก ซึ่งเรียกปัญหานี้ว่า Cold-Start หมายความว่า หากมีสินค้าใหม่เข้ามาในระบบ หรือผู้ใช้งานรายใหม่อาจทำให้ไม่สามารถทำการแนะนำรายการได้ เนื่องจากไม่เคยมีผู้ใช้งานใดให้คะแนนมาก่อนและผู้ใช้งานใหม่นั้นไม่เคยมีประวัติการซื้อสินค้าหรือการให้คะแนนในสินค้าใดๆเลย

เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ร่วมได้มีการนำข้อมูลของผู้ใช้งานรายอื่นมาช่วยในการให้คำแนะนำผู้ใช้งานปัจจุบันว่ามีความชื่นชอบหรือสนใจในสินค้าใด รูปแบบของวิธีการกรองข้อมูลแบบพึ่งพาผู้ร่วม แบ่งออกเป็น 3 ประเภท คือ Memory-based, Model-based และ Hybrid

2.1.2.1 Memory-based

การหาความสัมพันธ์จากข้อมูลระหว่างผู้ใช้งานกับสินค้า เป็นการคาดเดาคะแนนจากผู้ใช้ หรือผลิตภัณฑ์ที่มีความใกล้เคียงกัน โดยแบ่งออกเป็น 2 ประเภท ดังนี้

1. User-based คือ การประมาณค่าคะแนนของผู้ใช้ที่เคยให้ค่าคะแนนใกล้เคียงกัน ผู้ใช้จะถูกจัดกลุ่มให้อยู่ในกลุ่มเดียวกัน เนื่องจากมีพฤติกรรมหรือความชอบสินค้าที่มีลักษณะคล้ายๆกัน ตัวอย่างเช่น ผู้ใช้ที่มีการซื้อสินค้าที่เหมือนกัน ระบบได้แนะนำสินค้าที่มีความน่าจะเป็นที่ผู้ใช้งานปัจจุบันน่าจะซื้อเช่นเดียวกัน ทำให้ระบบแนะนำสินค้าประเภทอื่นให้กับผู้ใช้ด้วย ดังแสดงในภาพประกอบ 7

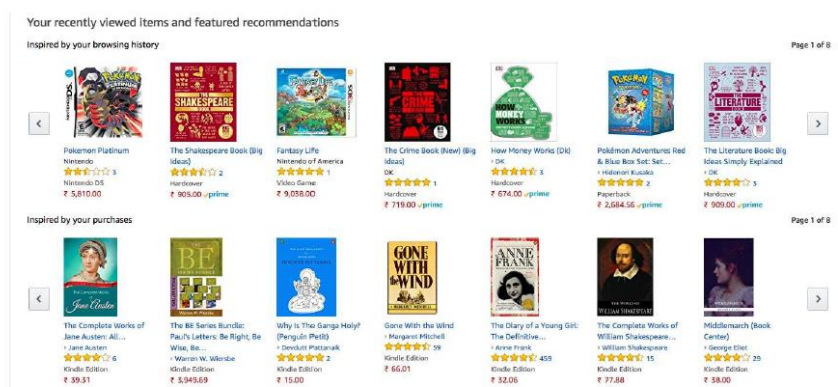


ภาพประกอบ 7 การแนะนำสินค้าโดยใช้วิธี User-based แสดงความสัมพันธ์ระหว่างผู้ใช้

ที่มา : (Banik, 2018, p. 43)

2. Item-based การพิจารณาจากข้อมูลสินค้า อธิบายได้ว่าถ้าผลิตภัณฑ์ A มีความคล้ายกับผลิตภัณฑ์ B แล้วเมื่อใดก็ตามที่ผู้ใช้งานเลือกผลิตภัณฑ์ A ระบบจะทำการแนะนำผลิตภัณฑ์ B ให้อัตโนมัติเพราะมีความเป็นไปได้ว่าผู้ใช้งานอาจชอบผลิตภัณฑ์ B ด้วย จาก

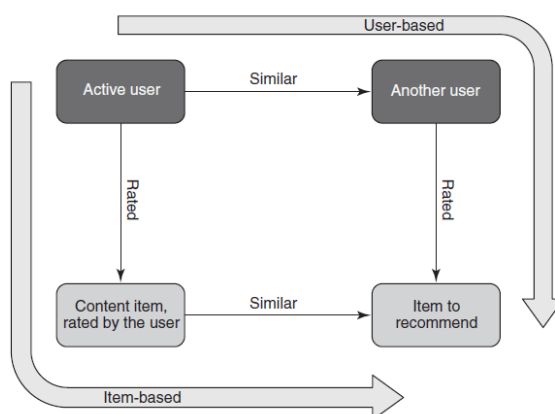
ภาพประกอบ 8 ระบบจะแนะนำสินค้าจากประวัติการค้นหาสินค้า และการสั่งซื้อของผู้ใช้งาน ระบบจึงแนะนำสินค้าที่มีลักษณะคล้ายกัน โดยมาจากลูกค้าในกลุ่มเดียวกัน



ภาพประกอบ 8 การแนะนำสินค้าโดยใช้วิธี Item-based แสดงความสัมพันธ์ระหว่างสินค้า

ที่มา : (Banik, 2018, p. 44)

จากภาพประกอบ 9 เป็นเทคนิคที่แสดงความสัมพันธ์ระหว่างผู้ใช้งาน (User-based) และความสัมพันธระหว่างผลิตภัณฑ์ (Item-based) ที่ทำให้มองเห็นในภาพรวมง่ายยิ่งขึ้น



ภาพประกอบ 9 แสดงความสัมพันธ์ระหว่างผู้ใช้งาน และความสัมพันธ์ระหว่างผลิตภัณฑ์

ที่มา : (Falk, 2019, p. 184)

2.1.2.2 Model-based

วิธีการของ Model-based เป็นการใช้แบบจำลองการเรียนรู้ของเครื่องและการทำเหมืองข้อมูลซึ่งใช้ในบริบทของแบบจำลองการทำนาย โดยจะเป็นการทำนายคะแนนที่ผู้ใช้งานให้คะแนนกับรายการสินค้า และใช้โมเดลในการหาเวกเตอร์ที่มีขนาดเล็กกว่า ในการหา User Weight และ Item Weight เพื่อให้สามารถประเมินความชอบของผู้ใช้ได้ใกล้เคียงกับความเป็นจริงที่สุด

2.1.2.3 Hybrid

การนำเทคนิคของวิธีการจาก Memory-based และ Model-based มาใช้งานร่วมกัน โดยเลือกนำข้อดีของทั้งสองวิธีการมาใช้ และลดปัญหาจากข้อเสียของแต่ละวิธีการที่อาจเกิดขึ้น มาเพิ่มประสิทธิภาพให้กับแบบจำลอง

2.1.3 การกรองข้อมูลแบบผสม (Hybrid Filtering)

เทคนิคการกรองข้อมูลแบบผสมเป็นวิธีการโดยรวมเอาเทคนิคการกรองข้อมูลตามเนื้อหา และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมมาผสมผสานการทำงานเข้าด้วยกัน หรือการนำวิธีอื่นๆ ตั้งแต่สองวิธีขึ้นไปนอกเหนือจากวิธีข้างต้นมาใช้ในการแก้ปัญหา เพื่อสร้างคำแนะนำรายการข้อมูลให้กับผู้ใช้งาน การกรองข้อมูลแบบผสมนี้อาจมีวิธีการที่ซับซ้อน แต่ทำให้ช่วยลดข้อบกพร่องหรือหลีกเลี่ยงในข้อจำกัดของแต่ละเทคนิคเช่น เมื่อมีข้อมูลใหม่เข้ามาในระบบ โดยที่ยังไม่เคยมีการให้คะแนนจากผู้ใช้งานมาก่อน ซึ่งเป็นการใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม จึงทำให้ไม่สามารถแนะนำผลิตภัณฑ์นี้ได้ ดังนั้นจึงต้องนำเทคนิคการกรองข้อมูลตามเนื้อหาเข้ามาช่วยในระบบแนะนำ เพื่อให้ระบบสามารถแนะนำข้อมูลได้อย่างมีประสิทธิภาพ และส่งเสริมการสร้างโอกาสในการแนะนำผลิตภัณฑ์ใหม่ที่เข้ามาในระบบ



ภาพประกอบ 10 เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering)

ที่มา : (Bai et al., 2019)

2.2 ทฤษฎีในการสร้างแบบจำลอง

2.2.1 เทคนิคการกรองข้อมูลตามเนื้อหา

2.2.1.1 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency คือ เทคนิคที่ทำการพิจารณาองค์ประกอบของคำภายในประโยคหรือเอกสาร โดยไม่นำลำดับของคำภายในเอกสารมาใช้ในการวิเคราะห์ ประกอบด้วย 2 ส่วนคือ Term Frequency (TF) และ Inverse Document Frequency (IDF)

Term Frequency (TF)

คือจำนวนของคำแต่ละคำที่ปรากฏขึ้นในเอกสารหนึ่ง โดยคำนวณจากการนำความถี่ของคำแต่ละคำที่ปรากฏในเอกสารหนึ่ง นำมาหารด้วยจำนวนคำทั้งหมดในเอกสาร ดังสมการ (1)

$$TF(t, d) = \frac{f(t_i, d_j)}{\sum_{t \in d} f(t_i, d_j)} \quad (1)$$

โดยที่

t คือ Term หรือ คำ

d คือ Document หรือเอกสาร

$f(t_i, d_j)$ คือ จำนวนของคำที่ปรากฏในเอกสาร

Inverse Document Frequency (IDF)

คือการคำนวณค่าน้ำหนักของแต่ละคำที่ปรากฏในเอกสารทั้งหมด โดยเป็นการพิจารณาจำนวนเอกสารเข้ามาเกี่ยวข้องด้วย วิธีการ TF คำนวณในแต่ละเอกสาร แต่ IDF เป็นการคำนวณค่าของแต่ละคำจากเอกสารทั้งหมด คำที่ปรากฏในเอกสารจำนวนน้อยๆ ทำให้มีน้ำหนักที่มากขึ้นและคำที่ปรากฏในเอกสารจำนวนมากมีน้ำหนักที่น้อยลง โดยสามารถคำนวณได้จากสมการ (2)

$$IDF(t_i) = \log \frac{N(d)}{N(t_i, d)} \quad (2)$$

โดยที่

d คือ Document หรือเอกสารทั้งหมด

$N(t_i, d_j)$ คือ จำนวนเอกสารที่มีคำปรากฏ

จากนั้นจึงนำผลของ TF และ IDF มาทำการคูณกันเพื่อหา TF-IDF

Weight ของคำนั้นๆ ทำให้สามารถแยกคำที่สำคัญออกมาได้จริงจากสมการ (3)

$$TF - IDF = TF(t, d) * IDF(t_i) \quad (3)$$

ถ้าคำมีความถี่มากในเอกสารจำนวนไม่มากนัก จะมี Weight สูง ซึ่งมีแนวโน้มว่าคำนั้นอาจเป็นใจความสำคัญสามารถใช้เป็นตัวแทนของเอกสารได้ แต่จะมี Weight ที่ต่ำลง ถ้าคำนั้นมีความถี่น้อยหรือปรากฏบ่อยครั้งในเอกสารจำนวนมาก แสดงว่าคำนั้นเป็นคำที่ไม่สามารถใช้เป็นตัวแทนของเอกสารได้ และมี Weight ที่ต่ำสุด ถ้าคำนั้นปรากฏในทุกๆ เอกสาร เนื่องจากคำเหล่านี้ไม่สามารถที่จะช่วยแยกแยะใจความสำคัญของเอกสาร อาจเป็นคำที่ไม่สำคัญหรือใช้เป็นตัวแทนเอกสารใดได้เลย

2.2.1.2 การวัดความคล้ายคลึง (Cosine Similarity)

เป็นการหาความคล้ายคลึงกันของข้อมูลจากค่าความต่างของมุมจากวัตถุ 2 อันที่เกิดขึ้นบนพื้นที่เวกเตอร์ เพื่อวัดว่ามีลักษณะไปในทิศทางเดียวกันหรือไม่ โดยสามารถคำนวณได้จากสมการ (4) ค่าของ Cosine มีค่าระหว่าง 0 ถึง 1 โดยที่ 1 หมายถึงข้อมูล A และข้อมูล B มีลักษณะของข้อมูลที่เหมือนกัน ในขณะที่ค่า 0 หมายถึงข้อมูล A และข้อมูล B ไม่มีความเหมือนกันเลย

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

โดยที่

A คือ รายการของชุด A

B คือ รายการของชุด B

n คือ จำนวนคู่ของรายการชุด A และรายการชุด B ทั้งหมด

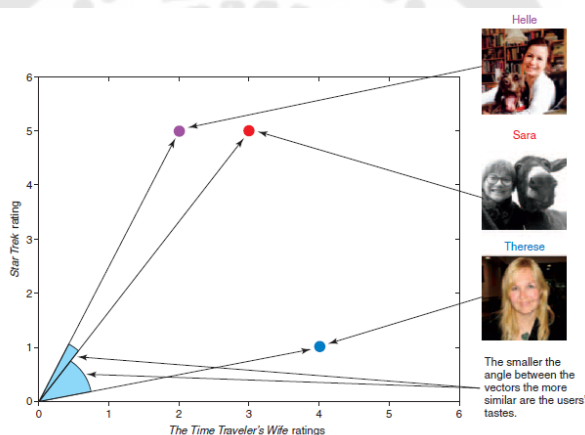
การประยุกต์ใช้ Cosine Similarity ค่อนข้างหลากหลายและสามารถนำมาใช้ในการ Recommendation เพื่อทำการหาความคล้ายคลึงกันของข้อมูลในเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) หรือหาความคล้ายคลึงกันระหว่างผู้ใช้งานกับข้อมูลการให้คะแนนสินค้าในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) ได้

จากตัวอย่างภาพประกอบ 11 คือการดูข้อมูลของแบบเมทริกซ์ซึ่งเป็นการให้คะแนนภาพยนตร์จากผู้ใช้งาน เราสามารถนำมาทำการคำนวณหา Cosine Similarity จากภาพประกอบ 12 พบว่า จากมุมระหว่าง Vector ของ Sara และ Helle มีความใกล้เคียงกันมากสามารถสรุปได้ว่ารสนิยมของทั้งสองคล้ายคลึงกันมาก เมื่อเปรียบเทียบกับ Sara และ Therese

	Comedy	Action
Sara	3	5
Therese	4	1
Helle	2	5

ภาพประกอบ 11 การให้คะแนนของผู้ใช้งาน

ที่มา : (Falk, 2019, p. 159)



ภาพประกอบ 12 แสดงความสัมพันธ์ของมุมระหว่าง Ratings Vectors

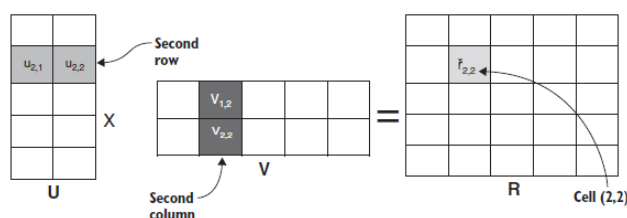
ที่มา : (Falk, 2019, p. 160)

2.2.2 เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

2.2.2.1 การแยกตัวประกอบเมทริกซ์ (Matrix Factorization)

Matrix Factorization คือ การแยกตัวประกอบของเมทริกซ์ให้มีขนาดที่เล็กลงจากเดิม โดยผลคูณของเมทริกซ์ย่อย จะมีค่าเท่ากับเมทริกซ์ดั้งเดิม การใช้ Matrix Factorization เป็นเหมือนกับการค้นหา Utility Matrix หรือ R Matrix โดยที่ R มีสมาชิกที่ i (แถว) และ j (หลัก) ของ R คือ Rating ซึ่งเป็นการให้คะแนนจาก User i ให้กับ Item j หากมองในภาพรวมนั้น Matrix ในแต่ละแถว แสดงถึงการ Represent User และหลักแสดงถึงการ Represent Item ซึ่งการทำ Matrix Factorization นี้ จะเป็นการ Represent Latent Feature ของ User และ Item นั้นๆ

จากภาพประกอบ 13 จากเมทริกซ์ User U และเมทริกซ์ของ Item V ผลคูณของทั้งสองเมทริกซ์จะมีความใกล้เคียงกับเมทริกซ์ Rating ของจริงมากที่สุด ซึ่ง Matrix Factorization อาจมีได้หลากหลายวิธีด้วยกัน ตัวอย่างเช่น NMF, SVD เป็นต้น



ภาพประกอบ 13 Matrix Factorization

ที่มา : (Falk, 2019, p. 293)

2.2.2.1.1 Singular-value decomposition (SVD)

เป็นวิธีการ Factorization ยอดนิยม ใช้เทคนิคทาง Linear Algebra ในการลดขนาดของเวกเตอร์คุณลักษณะ โดยแปลงจาก 1 เมทริกซ์ขนาดใหญ่ (M) และทำการแยกเมทริกซ์คุณลักษณะของ (M) ออกเป็น 3 เมทริกซ์ย่อย จากนั้นจะทำการลดขนาดของเมทริกซ์ M โดยทำการเก็บไว้เฉพาะค่า k อันดับแรกเท่านั้น 3 เมทริกซ์ใหม่ที่ได้ออกมา มีคุณสมบัติพิเศษบางอย่าง ซึ่งเป็นส่วนหนึ่งที่ใช้ชี้หาความสัมพันธ์ที่ซ่อนเร้นอยู่ ทำให้สามารถนำมาใช้งาน และวิเคราะห์ข้อมูลได้ดีขึ้น ตัวอย่างของการใช้ SVD ในงาน Recommender System คือการค้นหาค่าความสัมพันธ์ระหว่าง Item ต่างๆ ทำให้สร้างการแนะนำแบบง่ายได้ SVD เป็นที่นิยมและมีประสิทธิภาพมาก และถูกนำไปใช้ในการแข่งขันชิงรางวัลของ Netflix นอกจากนี้ SVD สามารถเรียกใช้ได้ใน แพ็คเกจ Surprise ของ Sci-kit สามารถทำการคำนวณ SVD ได้ในสมการ (5)

$$M = U\Sigma V^T \quad (5)$$

โดยที่

M คือ เมทริกซ์เดิมขนาด $m \times n$

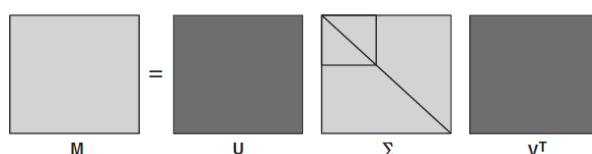
U คือ เมทริกซ์ขนาด $m \times m$

Σ คือ เมทริกซ์ขนาด $m \times n$ ที่มีค่าในแนวเส้นทแยงมุมเท่านั้น และ

ไม่เป็นค่าศูนย์

V คือ เมทริกซ์ขนาด $n \times n$

จากภาพประกอบ 14 เป็นการแสดง Rating เมทริกซ์ M โดยเป็นการสร้างเมทริกซ์ 2 ตัว ที่แสดงถึง Customer's taste และ Item profile โดยการใช้ SVD ซึ่งเป็นการสร้าง เมทริกซ์ U, Σ และ V^T ดังที่แสดงในภาพ



ภาพประกอบ 14 แสดง Singular-value decomposition (SVD)

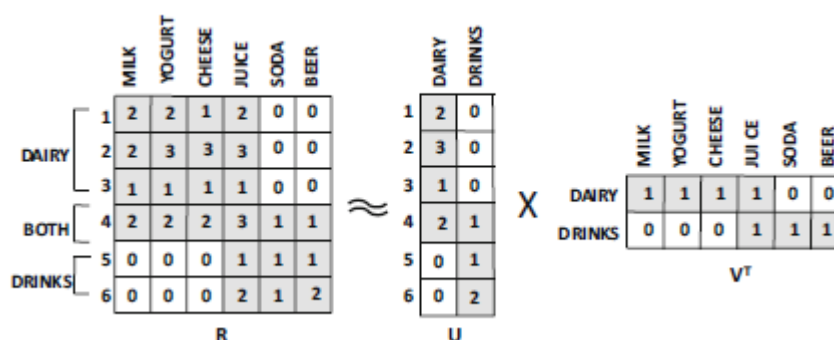
ที่มา : (Falk, 2019, p. 294)

2.2.2.1.2 Non-negative Matrix Factorization (NMF)

คือหนึ่งในเทคนิคของ Matrix Factorization ทำหน้าที่ลดขนาดของข้อมูลขนาดใหญ่จากเมทริกซ์ต้นทางหรือข้อมูลที่มีอยู่ ให้อยู่ในรูปผลคูณของ Matrix ใหม่สองอัน ผลลัพธ์ที่ได้จะมีค่าเป็นบวกเท่านั้น ซึ่งมาแก้ปัญหาค่าติดลบจากเทคนิค SVD ทำให้มีปัญหาในการตีความ ข้อดีของ NMF คือใช้งานง่ายและทำงานได้อย่างรวดเร็ว มีความสามารถในการตีความระดับสูงช่วยแก้ปัญหาและทำความเข้าใจ User-Item Interactions ซึ่งเป็นประโยชน์ในการจับคู่ระบบผู้แนะนำ แต่ข้อจำกัดของเทคนิค NMF คือ ผลลัพธ์ที่ได้เป็นค่าประมาณทำให้ไม่สามารถรวมกลับเป็นเมทริกซ์ต้นฉบับเดิมได้เหมือนกับ SVD

NMF มีความแตกต่างจาก Matrix Factorization วิธีอื่นคือ factor U และ V ไม่เป็นค่าลบ แม้ว่า NMF จะใช้กับการให้คะแนนที่ไม่เป็นลบใด เช่น การให้คะแนนตั้งแต่ 1-5 แต่การตีความเกิดขึ้นในกรณีที่ผู้ใช้ระบุความชอบของรายการ แต่ไม่ได้ระบุว่าไม่ชอบในรายการใด

ตัวอย่างภาพประกอบ 15 เมทริกซ์ขนาด 6×6 ที่มี 6 รายการ และผู้ใช้งาน 6 คน ดังแสดงในภาพประกอบ 15 แสดงพฤติกรรมการซื้อของลูกค้าโดยพิจารณาจากประเภทของสินค้าของ ผลิตภัณฑ์สองประเภทสอดคล้องกับนม (Daily) และเครื่องดื่ม (Drink) ตามลำดับ ลูกค้า 1,2,3,4 ชอบในผลิตภัณฑ์นม ในขณะที่ลูกค้า 4,5,6 ชอบเครื่องดื่ม จากรายการดังกล่าวสะท้อนให้เห็นว่าเป็น user factor เมทริกซ์ U ขนาด 6×2 จากตัวอย่างนี้แสดงค่าการแยกตัวประกอบทั้งหมดใน U และ V ขนาดของรายการของผู้ใช้ในแต่ละคอลัมน์วัดระดับความสนใจในด้านที่เกี่ยวข้อง ในทำนองเดียวกัน Matrix V แสดงให้เห็นว่ารายการนั้นสัมพันธ์กับแง่มุมต่างๆ อย่างไร



ภาพประกอบ 15 Non-negative Matrix Factorization (NMF)

ที่มา : (Aggarwal, 2016, p.122)

2.3 การวัดประสิทธิภาพของแบบจำลอง

2.3.1 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) คือ ความคลาดเคลื่อนเฉลี่ยกำลังสอง เป็นวิธีการวัดความคลาดเคลื่อนจากค่าที่ได้ทำนายจากแบบจำลองกับค่าจริง (Rating ที่ให้โดย User) ผลลัพธ์ของ RMSE มีค่าน้อยแสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับความจริง ดังนั้นหากมีค่าเท่ากับ 0 แสดงว่าไม่เกิดความคลาดเคลื่อนในแบบจำลองนั้นเลย ซึ่งสามารถคำนวณได้จากสมการ (6)

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}} \quad (6)$$

โดยที่

n คือ จำนวนข้อมูลทั้งหมด

y_i คือ ค่าจริง

$\hat{y}_i$ คือ ผลลัพธ์ที่ได้จากการทำนาย

2.3.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) คือ ค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ เป็นวิธีการที่ใช้วัดค่าประเมินผลความแม่นยำของแบบจำลอง โดยวัดจากความแตกต่างระหว่างค่าจริงและค่าที่ประมาณที่ได้จากแบบจำลอง หากค่าผลลัพธ์ของ MAE มีค่าน้อยแสดงถึงแบบจำลองนั้น

ประมาณค่าได้ใกล้เคียงกับค่าจริง และถ้ามีค่าเท่ากับ 0 แสดงว่าแบบจำลองนี้ไม่เกิดความคลาดเคลื่อนเลย สามารถคำนวณได้จากสมการ (7)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

โดยที่

y_i คือ ค่าจริง

$\hat{y}_i$ คือ ผลลัพธ์ที่ได้จากการทำนาย

n คือ จำนวนข้อมูลทั้งหมดที่ใช้ในการประมาณแบบจำลอง

2.4 Surprise Library

Surprise เป็น Library หนึ่งของโมดูล Sci-kit ในภาษา Python Programming มาจากคำว่า Simple Python Recommendation System Engine ใช้สำหรับนำมาวิเคราะห์และสร้างระบบแนะนำโดยเฉพาะ นำมาใช้ในการจัดการกับข้อมูลที่เป็น Rating ซึ่งใช้ในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) มีอัลกอริทึมที่ใช้ในการทำนาย และหาความคล้ายคลึงกันระหว่างผู้ใช้หรือรายการ นอกจากนี้ยังมีเครื่องมือในการวัดผล วิเคราะห์ และทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมต่างๆ โดยปกติแล้ว Surprise แบ่งอัลกอริทึมออกเป็น 5 กลุ่ม ได้แก่

1. Basic Algorithms เป็นอัลกอริทึมพื้นฐานที่ไม่ได้ผลมากนัก มักไม่เป็นที่นิยม ได้แก่ อัลกอริทึม NormalPredictor และ BaselineOnly
2. k-NN Algorithm สำหรับการหาความคล้ายคลึงกันของข้อมูล ได้แก่ อัลกอริทึม KNNBasic, KNNWithMeans, KNNWithZScore และ KNNBaseline เป็นต้น
3. Matrix Factorization-based Algorithms วิธีการแยกตัวประกอบของเมทริกซ์ ได้แก่ อัลกอริทึม SVD, SVD++, NMF เป็นต้น
4. Slope One ได้แก่ อัลกอริทึม SlopeOne
5. Co-clustering ใช้สำหรับจำแนกกลุ่มข้อมูลที่มีคุณสมบัติคล้ายกัน ได้แก่ อัลกอริทึม Coclustering

2.5 งานวิจัยที่เกี่ยวข้อง

ในการทบทวนวรรณกรรมของงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับการแนะนำหลักสูตรบนแหล่งความรู้ออนไลน์ งานวิจัยที่เกี่ยวข้องมีรายละเอียดดังนี้

1. บทความการวิจัยเรื่อง A Hybrid Recommender System for E-learning based on Context Awareness and Sequential Pattern Mining โดย Tarus, J. K., et al. (Tarus, Niu, & Kalui, 2017)

งานวิจัยนี้ได้ทำการศึกษาวิธีการแนะนำแหล่งความรู้ให้กับผู้เรียนโดยใช้แนวทางแบบผสมผสานอัลกอริทึมซึ่งได้แก่ Context Awareness (CA) , Sequential Pattern Mining (SPM) และ Collaborative Filtering (CF) Algorithm

แหล่งการเรียนรู้ที่มีปริมาณมากขึ้น ส่งผลให้ผู้เรียนมีปัญหาในการเลือกแหล่งข้อมูลเพื่อการเรียนรู้ เนื่องจากมีแหล่งข้อมูลที่มากเกินไป การที่ผู้เรียนมีระดับความรู้ในบริบทเบื้องต้นที่แตกต่างกัน อาจมีความพึงพอใจต่อแหล่งการเรียนรู้ที่แตกต่างกัน แต่เมื่อระดับความรู้ของผู้เรียนเปลี่ยนไปเป็นระดับที่สูงขึ้น ได้มีการนำวิธีการมาใช้เพื่อแก้ปัญหา และแนะนำแหล่งข้อมูลการเรียนรู้ให้กับผู้เรียนในสภาพแวดล้อม E-learning โดยคำแนะนำถูกปรับเปลี่ยนให้มีความเป็นส่วนตัวตามบริบทของผู้เรียนและเป็นรูปแบบในการเข้าถึงตามลำดับ

แนวทางการแนะนำแบบผสมในงานวิจัยนี้เป็นการแนะนำแบบผสมผสานของวิธี CF, CA และ SPM ของแหล่งการเรียนรู้ E-learning โดยมีการใช้ข้อมูลของโปรไฟล์ผู้เรียน ซึ่งเก็บข้อมูลและความชอบของผู้เรียน ได้แก่ ข้อมูลส่วนบุคคล (ชื่อ, เพศ, อายุ) ตลอดจนข้อมูลบริบทของผู้เรียนเช่น ระดับความรู้และเป้าหมายของผู้เรียน เป็นต้น นอกจากนี้ยังใช้ข้อมูลจาก Web Log ซึ่งเป็นการเตรียมข้อมูลบริบทของผู้เรียน นำไปวิเคราะห์ข้อมูลตามบริบทที่เพิ่มขึ้นของผู้เรียน และการใช้ข้อมูล Rating เพื่อคำนวณความคล้ายคลึงกันและนำมาทำนาย เพื่อสร้างคำแนะนำสูงสุดของแหล่งการเรียนรู้ ตามบริบทความชอบของผู้เรียน ในส่วนสุดท้ายคือการใช้ SPM Algorithm สำหรับการหาเหมือนข้อมูล เพื่อค้นหารูปแบบของการเข้าถึงตามลำดับ โดยนำไปหาผลลัพธ์เพื่อสร้างคำแนะนำในขั้นตอนสุดท้าย เพื่อเป้าหมายของผู้เรียนให้ได้รับคำแนะนำตามบริบทปัจจุบันของผู้เรียน

ในส่วนของคุณสมบัติข้อมูลที่มีการนำมาใช้เป็นข้อมูลที่เกิดขึ้นรวบรวมในระยะเวลา 6 เดือนของระบบ E-learning ผู้เรียนทั้งหมด 1,200 คน จำนวน 756 แหล่งการเรียนรู้ ข้อมูลการให้คะแนน 57,153 แบ่งเป็น Rating Scale คือ 1-5 และระดับการเรียนรู้ 3 ระดับ (Beginner, Intermediate, Advanced) แบ่งออกเป็น Training Set และ Test Set ในอัตราส่วน 80 และ 20

ผลจากการวิจัยการทดลองค่า kNN มีค่าเท่ากับ 25 เป็นขนาดที่เหมาะสมและอัลกอริทึมที่นำเสนอ (GSP-CA-CF) ให้ความแม่นยำที่ดีกว่าอัลกอริทึมอื่นๆ ซึ่งเปรียบเทียบกับวิธีอื่นได้แก่ CF-CA, CF, GSP สำหรับ kNN ที่จำนวนเท่ากัน นอกจากนี้การวัดประสิทธิภาพ

จากการใช้ Precision, Recall และ F1-Measure แนวทางการแนะนำที่น่าเสนอ ยังแสดงให้เห็นถึงประสิทธิภาพที่ดีเมื่อเปรียบเทียบกับวิธีอื่น

2. บทความการวิจัยเรื่อง Hybrid Recommender System for Learning Material Using Content-Based Filtering and Collaborative Filtering with Good Learners' Rating โดย R. Turnip, D. Nurjanah and D. S. Kusumo (Turnip, Nurjanah, & Kusumo, 2017)

งานวิจัยนี้ได้ทำการศึกษาวิธีการแนะนำแบบผสมผสานโดยใช้อัลกอริทึมซึ่งได้แก่ Content-based Filtering และ Collaborative Filtering มีการใช้คะแนนจากผู้เรียนที่ดี (Good Learner) ซึ่งเรียกอัลกอริทึมนี้ว่า CBF-CF-GL และได้ทำการทดลองและเปรียบเทียบกับอัลกอริทึม CBF-GL โดยใช้ข้อมูลชุดเดียวกัน

การศึกษานี้ใช้การให้คะแนนของผู้เรียนที่ดี คือผู้เรียนที่เชี่ยวชาญด้านสื่อการเรียนรู้ เพื่อเป็นคำแนะนำการให้คะแนนที่จะช่วยให้ผู้เรียนคนอื่นได้รับสื่อการเรียนรู้ที่เหมาะสม

จากการทดลองแสดงให้เห็นว่าวิธีการแนะนำโดยใช้อัลกอริทึม CBF-CF-GL นั้นให้ค่า MAE ที่ดีกว่าวิธี CBF-GL ที่คะแนน 0.447 และ 0.542 ตามลำดับ ซึ่งสรุปได้ว่า ผู้เรียนจะชอบสื่อการเรียนรู้ที่มีเนื้อหา มีความเหมาะสมและสอดคล้องกับความสามารถของตนเอง การให้คะแนนสื่อการเรียนรู้และความชอบจากผู้เรียนที่ดี จะสร้างคำแนะนำที่เหมาะสมและมีอิทธิพลในเชิงบวกตรงกับความสามารถของผู้เรียน โดยอ้างอิงข้อมูลมาจากผู้เรียนที่ดีและมีรสนิยมความชอบที่คล้ายคลึงกับผู้เรียนด้วย

3. บทความการวิจัยเรื่อง A Personalized Recommendation System with Combinational Algorithm for Online Learning โดย Xiao, J., Wang, M., Jiang, B. et al. (Xiao, Wang, Jiang, & Li, 2017)

งานวิจัยนี้ได้ทำการศึกษาวิธีการแนะนำแบบผสมผสานโดยใช้วิธี Association Rule, Content-based Filtering และ Collaborative Filtering โดยนำอัลกอริทึมนี้รวมเข้ากับเครือข่ายการเรียนรู้ตลอดชีวิตของ Shanghai เพื่อดำเนินการแนะนำนักเรียนที่ลงทะเบียนในหลักสูตรออนไลน์ เมื่อผู้เรียนทำการลงทะเบียนข้อมูลส่วนบุคคลและข้อมูลทางวิชาชีพ จะดำเนินการสร้างข้อมูลของผู้เรียนลงในฐานข้อมูลของระบบ รวมถึงบันทึกพฤติกรรมการเรียนรู้และเกิดการแนะนำหลักสูตรตามปัจจัย 3 ด้าน ได้แก่ อาชีพของผู้เรียน, ความสนใจในการเรียน และเนื้อหาหรือแหล่งทรัพยากรที่ผู้เรียนเข้าถึงในระบบ การแนะนำแหล่งเรียนรู้ที่ตรงกับความต้องการของผู้เรียน สามารถอำนวยความสะดวกในการเรียนรู้โดยการปรับปรุงประสบการณ์การดูหลักสูตรของผู้ใช้ ซึ่งทำให้ประหยัดเวลาในการค้นหาหลักสูตรที่เหมาะสม

ผลลัพธ์ที่ได้ทำให้เกิดการดึงดูดผู้ใช้งานจำนวนมากเข้าสู่เครือข่ายการเรียนรู้ตลอดชีวิต ผ่านคำแนะนำแบบปากต่อปากบนโลกแห่งความเป็นจริง ความพึงพอใจของผู้ใช้ จากผลการวิจัยพบว่าเกือบ 90% ของผู้ตอบแบบสอบถาม รู้จักกับเครือข่ายการเรียนรู้ตลอดชีวิตของ Shanghai และวางแผนที่จะกลับมาเรียนบนแพลตฟอร์มนี้เพื่อการศึกษาอย่างต่อเนื่อง

4. บทความการวิจัยเรื่อง A Hybrid Recommendation Algorithm Adapted in E-learning Environments โดย Chen, W., Niu, Z., Zhao, X. et al. (Chen, Niu, Zhao, & Li, 2012)

งานวิจัยนี้ได้ทำการศึกษาวิธีการแนะนำทรัพยากรการเรียนรู้แบบผสม โดยใช้เทคนิค Collaborative Filtering (CF) และ Sequential Pattern Mining (SPM) ประกอบด้วย 2 ขั้นตอน คือ การค้นหาชุดของรายการที่เกี่ยวข้อง โดยใช้ Item-based Collaborative Filtering (Item-based CF) และการใช้ชุดรายการข้อมูลกับ SPM Algorithm เพื่อกรองรายการตามลำดับของการเรียนรู้ทั่วไป การใช้เทคนิคร่วมกันนี้ เพื่อแนะนำรายการการเรียนรู้ที่อาจเป็นประโยชน์และเป็นแนวทางให้ผู้ใช้ในกระบวนการเรียนรู้ในปัจจุบัน

งานวิจัยได้ทำการประเมินการแนะนำโดยใช้ Recommendation Error Metrics (REM) เพื่อวัดประสิทธิภาพโดยคำนวณและสรุป Recommendation-error ของคำแนะนำคู่ของ User-Item แต่ละรายการก่อน แล้วจึงคำนวณค่าเฉลี่ย ยิ่งค่า REM มีค่าต่ำมากเท่าไร Algorithm ก็ยิ่งมีความแม่นยำมากยิ่งขึ้น

การทดลองนี้ได้พัฒนาระบบ E-learning เพื่อใช้ประเมินประสิทธิภาพของ CF, SPM และ Hybrid Algorithm ก่อนที่ผู้เรียนจะเริ่มเรียนรู้ในระบบ ต้องทำการการให้คะแนนรายการที่น่าสนใจ โดย Rating เหล่านี้ใช้เป็นเกณฑ์มาตรฐานในการประเมินคุณภาพของ Hybrid Algorithm

จากผลการทดลอง ได้นำเสนอผลการทดสอบ Hybrid Algorithm ซึ่งถูกเปรียบเทียบกับ CF, SPM Algorithm แยกกัน โดยมุ่งเน้นพารามิเตอร์ในส่วนของเวลาที่ใช้, ขนาดของรายการที่คล้ายกัน และค่าของ Min-support โดยวัดประเมินจากค่า REM ที่ต่ำสุดและมีค่าพารามิเตอร์ที่ดีที่สุด ซึ่งได้ผลลัพธ์คือ Similar Size ที่มีบทบาทสำคัญในคุณภาพของการแนะนำมีค่าเท่ากับ 6 ขนาดของ Min-support มีค่าเท่ากับ 0.15 ส่วนในเรื่องของเวลาที่ใช้โดยแต่ละอัลกอริทึมที่ทำงานตามขนาดต่างๆ และรายการที่คล้ายคลึงกันพบว่า Hybrid ใช้เวลามากกว่าอัลกอริทึมอื่น แต่ไม่มากนัก

5. บทความการวิจัยเรื่อง A Recommender Model in E-learning Environment

โดย Bourkhoukou, O., El Bachari, E. & El Adnani, M. (Bourkhoukou, El Bachari, & El Adnani, 2016)

งานวิจัยนี้ได้ทำการศึกษาระบบผู้แนะนำสำหรับการเรียนรู้ส่วนบุคคลในสภาพแวดล้อม E-learning โดยใช้เทคนิคการผสมผสานร่วมกันระหว่าง Collaborative Filtering (CF) และ Sequential Pattern Mining (SPM) เพื่อปรับปรุงการแนะนำให้ได้ประสิทธิภาพการเรียนรู้ที่มีประสิทธิภาพ

โดยการทดลองนี้ได้ใช้ชุดข้อมูลจริง 2 ชุดคือ Algebra 2005–2006 (Alg) และ Geometry 2006–2007 (Geo) การทดลองดำเนินการเพื่อตอบคำถามที่ว่า พารามิเตอร์ เช่น Similarity Metrics, จำนวนของ Neighborhood และขนาดของชุดข้อมูล ส่งผลต่อผลลัพธ์อย่างไร และประสิทธิภาพของ Collaborative Filtering ของการทดลองนี้ใช้สามารถสำเร็จผลได้อย่างไร เมื่อเปรียบเทียบกับเทคนิค Collaborative Filtering อื่นๆ ในชุดข้อมูลต่างๆ โดยผลลัพธ์ที่ได้สอดคล้องกับการทดลองหลายครั้ง และแสดงให้เห็นว่าการใช้ Tanimoto Similarity Measure มีประสิทธิภาพมากในกรณีของชุดข้อมูลที่มี Sparsity และประสิทธิภาพที่ดีที่สุดเกิดขึ้นเมื่อ k มีค่าประมาณ 190 สำหรับชุดข้อมูล Geo และ 150 สำหรับชุดข้อมูล Alg

นอกจากนี้ยังมีการทดสอบครั้งที่ 2 เพื่อประเมินประสิทธิภาพของแนวทางที่นำเสนอ โดยนำไปเปรียบเทียบกับประสิทธิภาพกับวิธีการต่างๆ พบว่า kNN-Pearson ทำงานได้ดีในชุดข้อมูลของ Alg และ kNN-Cosine ทำงานได้ดีในชุดข้อมูลของ Geo

6. บทความการวิจัยเรื่อง Recommender System for Big Data in Education

โดย Surabhi Dwivedi, Dr Kumari Roshni (Dwivedi & Roshni, 2017)

งานวิจัยนี้ได้ทำการศึกษาระบบการแนะนำการลงทะเบียนเรียน เพื่อให้สถานศึกษาสามารถออกแบบหลักสูตรให้นักเรียนมีทางเลือกในการลงทะเบียนเรียนมากยิ่งขึ้น ผลลัพธ์ที่ได้จากการแนะนำทำให้นักเรียนมีทางเลือกในการลงทะเบียนเรียนมากขึ้น ทำให้เป็นการช่วยตัดสินใจในการลงทะเบียนหลักสูตรที่สนใจ โดยใช้เทคนิค Collaborative Filtering ซึ่งนำข้อมูลจากโรงเรียนต่างๆ ของหลักสูตรโดยคณะกรรมการกลางมัธยมศึกษาทั่วประเทศอินเดีย และใช้ Hadoop Framework เพื่อจัดการกับข้อมูล

คำแนะนำที่สร้างโดยระบบดังกล่าวเป็นประโยชน์ต่อสถาบันการศึกษาในการปรับปรุงประสิทธิภาพของนักเรียน โรงเรียน และครูผู้สอน การใช้ระบบการแนะนำจาก Collaborative Filtering และ Item-based เพื่อแนะนำวิชาเลือกให้กับนักเรียน โดยอ้างอิงจาก

เกรดที่ได้รับในวิชาอื่นๆ สามารถใช้กับข้อมูลการศึกษานานาชาติใหญ่ได้ และสามารถระบุได้ว่าข้อมูลการศึกษาสามารถจับคู่รายการในระบบคำแนะนำได้อย่างไร

ตัวอย่างของข้อมูลที่น่าสนใจคือ รหัสนักเรียน, วิชา, เกรดที่ได้รับ จากการประเมินผลลัพธ์ใช้ RMSE ในการวัดประสิทธิภาพของการทำนายระหว่างข้อมูลจริงและข้อมูลจากระบบแนะนำซึ่งมีค่าเท่ากับ 0.425 แสดงถึงความเหมาะสมของระบบคำแนะนำที่สร้างขึ้น

ตาราง 1 สรุปงานวิจัยที่เกี่ยวข้อง

ปี	ผู้วิจัย	บทความเรื่อง	Technique	ผลลัพธ์
2017	Tarus, J. K., et al.,	A hybrid recommender system for e-learning based on context awareness and sequential pattern mining	Context Awareness (CA), Sequential Pattern Mining (SPM) และ Collaborative Filtering (CF) Algorithm	kNN = 25
2017	R. Turnip, D. Nurjanah and D. S. Kusumo	Hybrid Recommender System for Learning Material Using Content-Based Filtering and Collaborative Filtering with Good Learners' Rating	Content-based Filtering, Collaborative Filtering, Good Learner's Rating	CBF-CF-GL ดีกว่าที่ MAE = 0.447
2018	(Xiao, J., Wang, M., Jiang, B. et al.	A personalized recommendation system with combinational algorithm for online learning	Association Rule, Collaborative Filtering	ดึงดูดผู้ใช้จำนวนมากให้เข้ามาใช้ระบบ
2014	Chen, W., Niu, Z., Zhao, X. et al.	A hybrid recommendation algorithm adapted in e-learning environments	Item-based Collaborative Filtering, Sequential Pattern Mining, Hybrid	Hybrid ทำงานได้ดีกว่า ค่าที่เหมาะสมคือ Similar Size = 6, Min-support = 0.15
2017	Bourkhouk, O., El Bachari, E. & El Adnani, M.	A Recommender Model in E-learning Environment	Collaborative Filtering, Sequential Pattern Mining (SPM)	Algebra Dataset k = 150, kNN-Pearson Geometry Dataset k = 190, kNN-Cosine
2017	Surabhi Dwivedi, Dr Kumari Roshni,	Recommender System for Big Data in Education	Collaborative Filtering, Item-based	RMSE = 0.425

บทที่ 3

วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยครั้งนี้ ผู้วิจัยได้นำเสนอการให้คำแนะนำในรายวิชาการเรียนออนไลน์บนเว็บไซต์ Coursera โดยได้ใช้ชุดข้อมูลของ Kaggle มาใช้ในงานวิจัยโดยใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งได้ดำเนินการตามขั้นตอนดังต่อไปนี้

1. กระบวนการสร้างแบบจำลอง
2. การเก็บรวบรวมข้อมูล (Data Collection)
3. การทำความสะอาดข้อมูล (Data Cleansing)
4. การสำรวจข้อมูล (Exploratory Data Analysis: EDA)
5. การเตรียมข้อมูล (Data Preparation)
6. การสร้างแบบจำลองเพื่อการทำนาย
7. สรุปผลการดำเนินงาน

3.1 กระบวนการสร้างแบบจำลอง

จากภาพประกอบ 16 ได้อธิบายถึงกระบวนการสร้างแบบจำลอง โดยกระบวนการสร้างแบบจำลองข้อมูล เริ่มต้นมาจากขั้นตอนการเก็บรวบรวมข้อมูลที่เกี่ยวข้อง โดยทำการเชื่อมโยงข้อมูลจากหลายแหล่งข้อมูลเข้าด้วยกัน ซึ่งรวมไปถึงกระบวนการจัดการข้อมูลในเบื้องต้นเช่น การทำความสะอาดข้อมูล การสำรวจข้อมูล และการเตรียมข้อมูล จากนั้นเมื่อข้อมูลพร้อมใช้งาน ได้นำไปสู่กระบวนการสร้างแบบจำลองเพื่อการทำนายรายวิชา โดยข้อมูลถูกแบ่งออกมาเพื่อนำมาใช้ในเทคนิคที่แตกต่างกัน ซึ่งได้แก่

1. เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) ใช้ชุดข้อมูลจากชุดข้อมูลรายละเอียดของรายวิชา (Course Data) มาทำการสร้างแบบจำลอง โดยใช้หลักการของ TF-IDF ทำกระบวนการแตก Vector ของคำจากข้อมูลที่เตรียมไว้ และนำมาทำการหาความคล้ายคลึงของโคไซน์ (Cosine Similarity) ที่มีค่าสูงสุดมาจัดอันดับ และสร้างคำแนะนำรายวิชาให้กับผู้เรียน

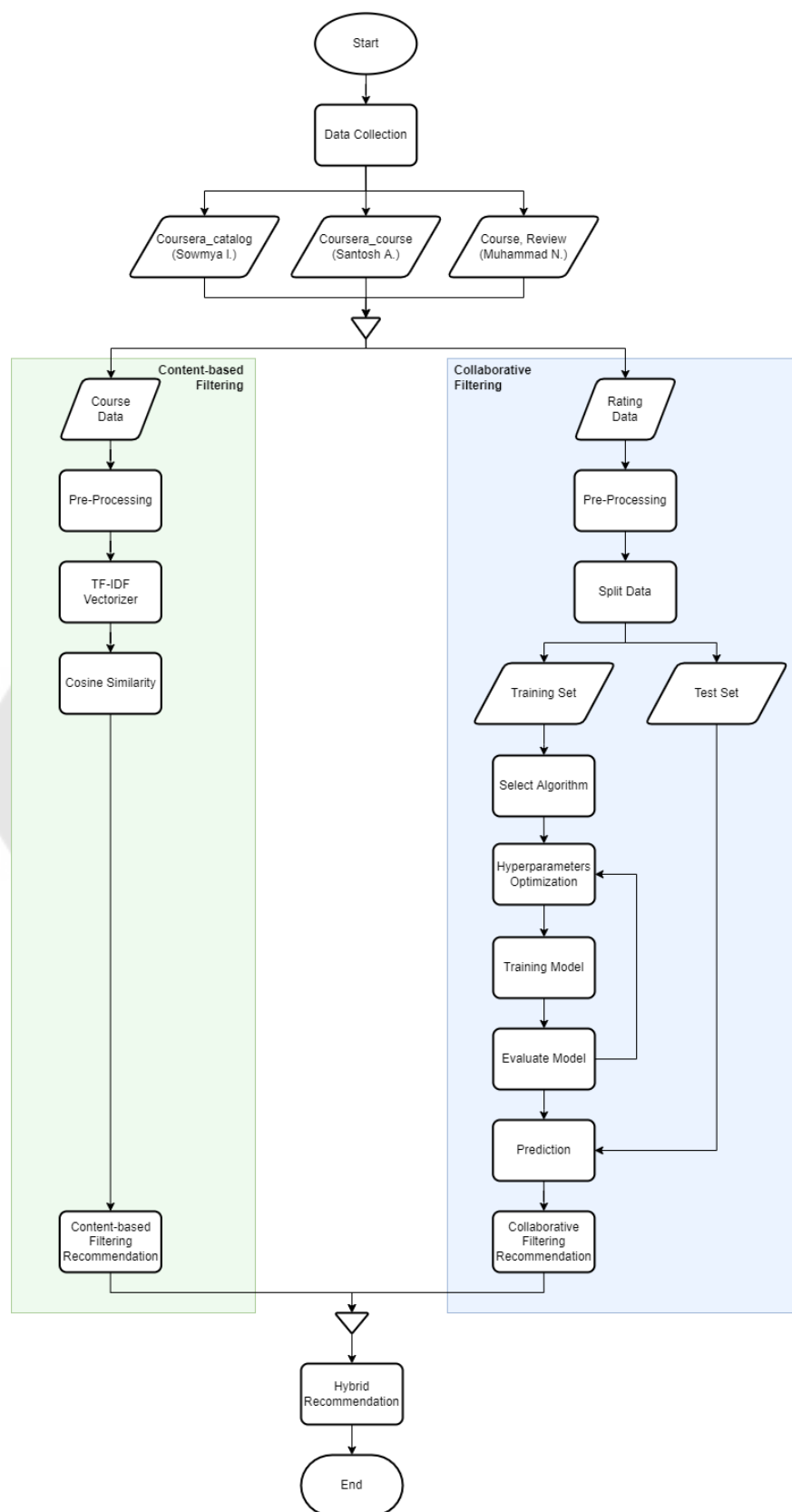
2. เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) โดยใช้ชุดข้อมูลจากการให้คะแนนรายวิชาจากผู้เรียน (Rating Data) มาทำการสร้างเป็นแบบจำลอง โดยใช้วิธีการแบบ Model-based เพื่อทำการหา User Weight และ Item Weight ให้สามารถประเมินความชอบของผู้เรียนได้ใกล้เคียงกับค่าคะแนนของความเป็นจริงที่สุด โดยทำการแบ่งข้อมูล

ออกเป็น 2 ส่วนคือ ข้อมูลการสอน (Training Data) และข้อมูลการทดสอบ (Testing Data) ในอัตราส่วน 80:20 โดยนำข้อมูลจาก Training Data มาเปรียบเทียบประสิทธิภาพของแบบจำลองในไลบรารี Surprise และทำการวัดประสิทธิภาพโดยใช้ค่า RMSE และ MAE โดยทำการเลือกอัลกอริทึมที่มีค่าดีที่สุด มาทำการปรับจูนคัดเลือกพารามิเตอร์ที่ดีที่สุด หลังจากนั้นจึงนำพารามิเตอร์ที่ได้มาทำการฝึกสอนข้อมูลและนำไปใช้กับชุดข้อมูลทดสอบ และนำอัลกอริทึมที่ได้มาสร้างแบบจำลองเพื่อทำการแนะนำรายวิชาให้กับผู้เรียน

3. เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering) คือการนำเทคนิคการกรองข้อมูลตามเนื้อหา และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม มาทำงานร่วมกันและทำการแนะนำค่าสูงสุด 5 อันดับในการแนะนำรายวิชาให้กับผู้เรียน

โดยขั้นตอนของการสร้างแบบจำลองสามารถดูรายละเอียดได้จากภาพประกอบ 16





ภาพประกอบ 16 กระบวนการสร้างแบบจำลอง

3.2 การเก็บรวบรวมข้อมูล (Data Collection)

ในการวิจัยครั้งนี้มีการใช้ชุดข้อมูลรายวิชาออนไลน์บนเว็บไซต์ Coursera ซึ่งเป็นข้อมูลสาธารณะจาก Kaggle.com โดยเลือกชุดข้อมูลทั้งหมด 3 แหล่งข้อมูล ประกอบด้วย ข้อมูลรายวิชาภายในเว็บไซต์ ซึ่งมีรายละเอียดเกี่ยวกับรายวิชาต่างๆ และการให้คะแนนรายวิชาจากผู้เรียน ตัวอย่างของขอบเขตข้อมูลแสดงดังตาราง 2

ตาราง 2 แหล่งที่มาและขอบเขตข้อมูล

แหล่งข้อมูล	ข้อมูล	จำนวนรายวิชา
Muhammad Nakhaee (Nakhaee, 2020)	การให้คะแนน (Rating) และชื่อรายวิชา (Course)	623
Santosh A Patil (Patil, 2020)	รายละเอียดของรายวิชา (Course) เช่น ชื่อรายวิชา สถาบันการสอน ประเภทของรายวิชา ภาษาที่ใช้ในการสอน ทักษะที่ใช้ คะแนนโดยรวมของรายวิชา ระดับของรายวิชา วิชาหลัก วิชารอง เป็นต้น	5,164
Sowmya Iyer (Iyer, 2020)	รายละเอียดของรายวิชา (Course) เช่น ชื่อรายวิชา สถาบันการสอน ประเภทของรายวิชา	891

ผู้วิจัยได้ทำการเลือกข้อมูลทั้ง 3 แหล่งนี้ เนื่องจากพบว่าในแหล่งข้อมูลที่ 1 มี 2 ชุดข้อมูล ซึ่งมีข้อมูลจำนวนรายวิชาทั้งหมด 623 รายวิชา และข้อมูลที่มีการให้คะแนนจากผู้เรียนมีเพียง 604 รายวิชาเท่านั้น ในขณะที่แหล่งข้อมูลที่ 2 มีจำนวน 5,164 รายวิชา และแหล่งข้อมูลที่ 3 มีจำนวน 891 รายวิชา จากรายการข้อมูลพบว่าจำนวนรายวิชามีความแตกต่างกันมาก ถ้านำข้อมูลมาใช้เพียงเทคนิคเดียว อาจพบว่าได้ข้อมูลที่ไม่หลากหลายนัก ผู้วิจัยจึงนำชุดข้อมูลเหล่านี้ซึ่งมี Feature ที่เหมือนและแตกต่างกัน มาทำการแนะนำรายวิชาแบบใช้เทคนิคการกรองแบบผสม โดยเลือกแหล่งข้อมูลชุดรายละเอียดของรายวิชา และชุดข้อมูลการให้คะแนนจากผู้เรียนมาทำการแนะนำร่วมกัน เพื่อทำให้เกิดข้อมูลที่หลากหลายและมีตัวเลือกที่มากขึ้น โดยสามารถนำชุดข้อมูลที่ได้จากรายละเอียดของรายวิชามาทำการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา และนำข้อมูลจากการให้คะแนนของผู้เรียนมาทำการแนะนำโดยใช้เทคนิคการกรองข้อมูลแบบพึงพาผู้ใช้ร่วมกัน จากนั้นจึงนำเทคนิคทั้งสองนี้มาผสมผสานกันเพื่อให้เกิดการแนะนำรายวิชาได้มากยิ่งขึ้น

ด้วยเหตุนี้ผู้วิจัยจึงต้องทำกระบวนการในการจัดการข้อมูลโดยเริ่มต้นจากการเชื่อมโยงข้อมูล (Combining Data) จากข้อมูลหลายแหล่งเข้าด้วยกัน และพบว่าชุดข้อมูลทั้งหมดมีขอบเขตข้อมูลที่เหมือนกันคือ Course Name (ชื่อรายวิชา), Course Institution (สถาบันเจ้าของรายวิชา) และ Course URL โดยได้ดำเนินการรวบรวมข้อมูลตามขั้นตอนดังต่อไปนี้

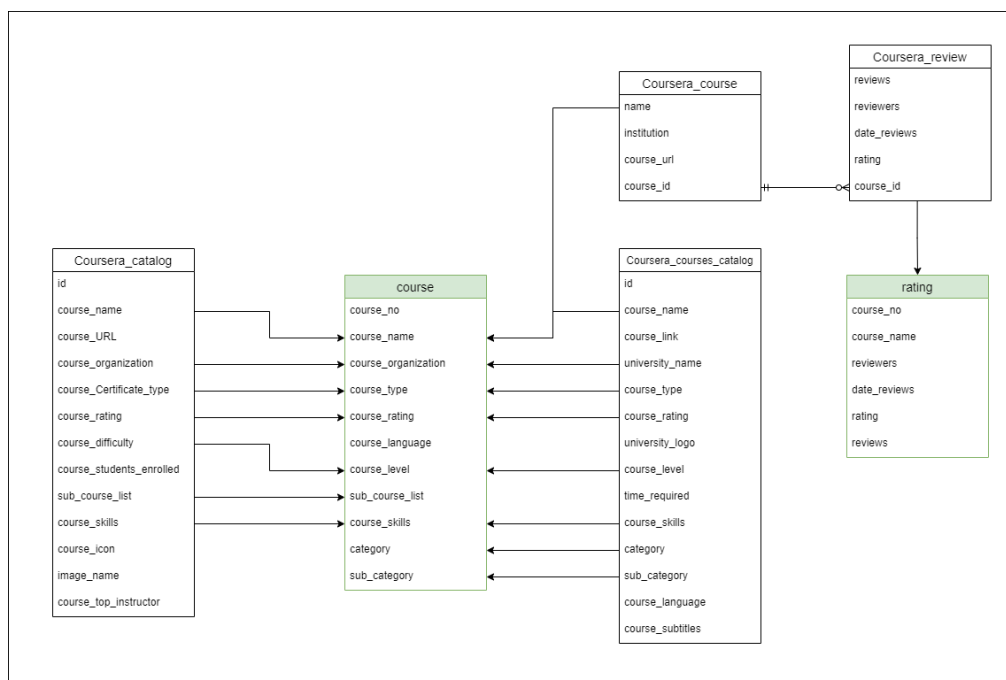
1. รวบรวมแหล่งข้อมูลที่ 2 และ 3 เข้าด้วยกันเป็นอันดับแรก เนื่องจากว่ามีรายละเอียดของข้อมูลครบถ้วนมากกว่าแหล่งข้อมูลที่ 1 ซึ่งได้ใช้ข้อมูลของ Course Name, Course Organization และ Course Type เป็นข้อมูลหลักในการเชื่อมโยงความสัมพันธ์ทั้งหมดของข้อมูล และได้ใช้ชุดข้อมูลของแหล่งข้อมูลที่ 2 เป็นข้อมูลหลัก จากนั้นจึงทำความสะอาดข้อมูลในเบื้องต้นบางส่วน เช่น การจัดการกับค่าว่าง การทำข้อมูลให้เป็นมาตรฐานมีความหมายตรงกัน และการแทนที่ค่าในกรณีที่ชุดข้อมูลชุดที่ 2 ไม่มีข้อมูล ตัวอย่างเช่น course_level, course_rating เป็นต้น โดยนำข้อมูลจากชุดข้อมูลที่ 3 มาแทนที่ เพื่อให้ข้อมูลมีค่าว่างน้อยลง และข้อมูลมีความสมบูรณ์มากยิ่งขึ้น

2. นำข้อมูลที่ได้มาจากการรวบรวมชุดข้อมูล 2 และ 3 มาทำการรวมกับแหล่งข้อมูลที่ 1 ที่เป็นชุดข้อมูลรายวิชา โดยใช้ข้อมูลของ Course Name, Course Organization และ Course URL เป็นเขตข้อมูลในการเชื่อมโยงความสัมพันธ์

3. ทำการ Generate ข้อมูลที่ชื่อว่า course_no เพื่อทำให้ข้อมูลรายวิชาทั้งหมดมี Identifier และมีความ Unique ดังภาพประกอบที่ 17

4. จากแหล่งข้อมูลที่ 1 มีทั้งหมด 2 ชุดข้อมูลคือ ข้อมูลรายวิชา และข้อมูลการให้คะแนนจากผู้เรียน เนื่องจากข้อมูลจากแหล่งที่ 1 เป็นข้อมูลที่เชื่อมโยงกันด้วยคอลัมน์ course_id จึงทำการเชื่อมชุดข้อมูลทั้ง 2 นี้เข้าด้วยกัน โดยใช้ข้อมูลของ course_id เพื่อให้ชุดข้อมูลที่เป็นข้อมูลการให้คะแนนจากผู้เรียน มี course_no ที่ได้จากขั้นตอนที่ 3 และสามารถนำข้อมูลของ course_no ไปใช้ในเทคนิคการกรองข้อมูลแบบฟังก์ชันร่วมในกระบวนการต่อไป

การนำข้อมูลทั้ง 3 แหล่งข้อมูลนี้มาเชื่อมโยงความสัมพันธ์กัน เป็นขั้นตอนการรวบรวมข้อมูล และมีการทำความสะอาดข้อมูลในบางส่วน เพื่อให้สามารถทำการเชื่อมโยงความสัมพันธ์ของข้อมูลได้ เพราะฉะนั้นในกระบวนการเก็บรวบรวมข้อมูลนี้ ทางผู้วิจัยได้ทำการเชื่อมโยงข้อมูลทั้งหมด และได้ชุดข้อมูลทั้งหมด 2 ชุด คือ ชุดข้อมูลรายละเอียดรายวิชา(Course Data) และชุดข้อมูลการให้คะแนนรายวิชาจากผู้เรียน (Rating Data)



ภาพประกอบ 17 การเชื่อมโยงความสัมพันธ์ของข้อมูล

3.3 การทำความสะอาดข้อมูล (Data Cleansing)

เป็นกระบวนการทำงานที่เกี่ยวข้องกับการตรวจสอบและแก้ไขข้อมูล เพื่อให้ข้อมูลอยู่ในรูปแบบที่ถูกต้องและสมบูรณ์ ซึ่งขั้นตอนของการทำความสะอาดข้อมูลถือเป็นขั้นตอนที่สำคัญของการเตรียมพร้อมข้อมูล ผู้วิจัยได้ดำเนินการทำความสะอาดข้อมูลโดยมีขั้นตอนดังนี้

1. จัดการกับข้อมูลที่สูญหาย (Missing Values) โดยเปลี่ยนแปลงค่าข้อมูลจาก NaN หรือค่าว่างให้เป็น Undefined เพื่อให้ข้อมูลไม่เป็นค่าว่าง ซึ่งจะทำการกระบวนการนี้ในข้อมูลประเภท Text เช่นเดียวกับข้อมูลประเภทตัวเลขเช่น ข้อมูลคะแนนของรายวิชา (Rating) ในชุดข้อมูลรายละเอียดของรายวิชา ซึ่งบางรายวิชาไม่มีข้อมูลการให้คะแนน แต่ไม่ได้หมายความว่ารายวิชานี้ไม่ดี จึงทำการให้คะแนนโดยใช้ค่าเฉลี่ยเลขคณิต (Mean) จากคะแนนรายวิชาทั้งหมด

2. จัดการกับข้อมูลที่มีความซ้ำซ้อน (Duplicated Data) โดยทำการลบข้อมูลส่วนนี้ออก เพื่อลดความซ้ำซ้อนของข้อมูล ตัวอย่างเช่น ข้อมูลการให้คะแนนรายวิชาจากผู้เรียน (Rating Data) เนื่องจากมีผู้เรียนบางคนทำการเรียนซ้ำหรือเกิดข้อผิดพลาดของระบบทำให้มีการบันทึกข้อมูลการให้คะแนนรายการซ้ำ และเลือกรายการข้อมูลล่าสุด โดยอ้างอิงจากวันที่การให้คะแนน

3. การทำให้ข้อมูลเป็นรูปแบบเดียวกัน (Standardizing) เนื่องจากมีการนำข้อมูลมาจากหลากหลายแหล่งข้อมูลค่าของข้อมูลที่ได้อาจมีรูปแบบแตกต่างกัน จึงต้องทำการแทนที่

(Replace) ข้อมูลที่มีความหมายหรือความสัมพันธ์ให้เหมือนกัน เพื่อให้ข้อมูลมีความเป็นมาตรฐานมากขึ้น เช่น ประเภทรายวิชา (course_type), ระดับรายวิชา (course_level) เป็นต้น

4. การเปลี่ยนแปลงข้อมูล โดยทำการเปลี่ยนแปลงชื่อผู้ให้คะแนน (Reviewer) จากชื่อที่เป็นข้อมูลประเภท Text และทำการ Generate ข้อมูลของ user_id ให้อยู่ในรูปแบบตัวเลข ทำให้ข้อมูลผู้เรียนมี Identifier และมีความ Unique มากยิ่งขึ้น เพื่อให้สามารถนำไปใช้ในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) ได้

3.4 การสำรวจข้อมูล (Exploratory Data Analysis: EDA)

การดำเนินการสำรวจข้อมูลโดยข้อมูลที่ได้ทำความสะอาดแล้วนั้น ประกอบไปด้วยชุดข้อมูลทั้งหมด 2 ชุดข้อมูลดังนี้

1. ชุดข้อมูลรายละเอียดของรายวิชา Course Data ประกอบด้วยข้อมูลจำนวน 5,599 แถวและ 12 คอลัมน์ ตามตาราง 3 ทางผู้วิจัยได้ใช้ชุดข้อมูลนี้มาทดสอบโดยการใช้เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) ซึ่งใช้ข้อมูลทั้งหมด 6 คอลัมน์ ได้แก่ รหัสรายวิชา (course_no), ชื่อรายวิชา (course_name), วิชาหลัก (category), วิชารอง (sub_category), ทักษะ (course_skills) และรายวิชาที่เกี่ยวข้อง (sub_course_list) เป็นต้น โดยใช้ข้อมูลนี้มาทำการวิเคราะห์ในการแนะนำรายวิชาให้กับผู้เรียน ตัวอย่างข้อมูลรายละเอียดของรายวิชา ดังภาพประกอบ 18

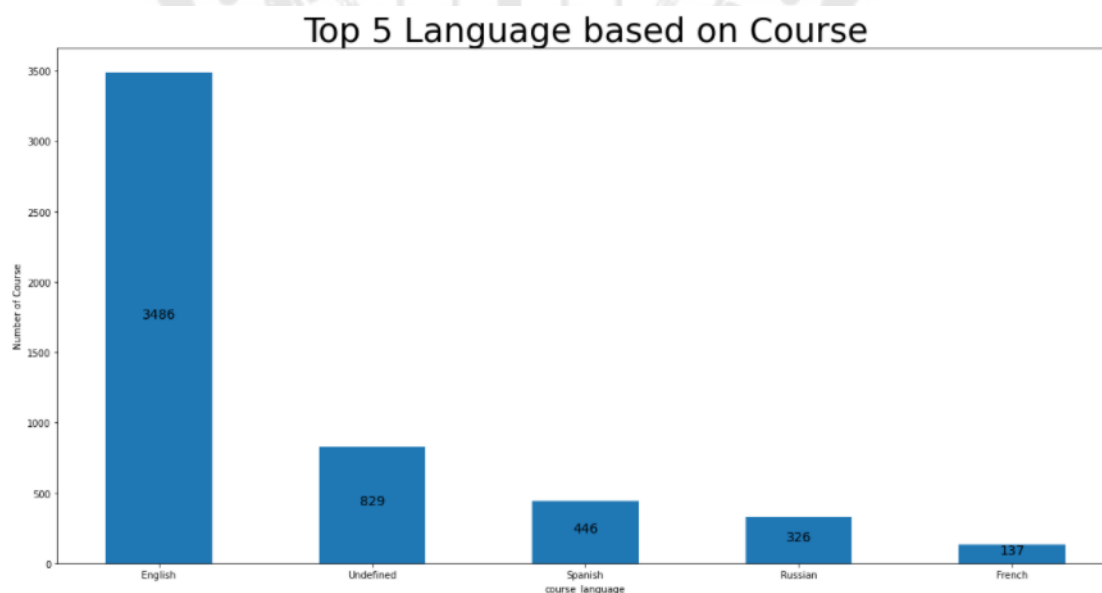
ตาราง 3 โครงสร้างข้อมูลรายวิชา (Course Data)

ลำดับ	ข้อมูลตัวแปร (Variable)	คำอธิบายข้อมูล (Description)
1	course_no	รหัสรายวิชา
2	course_name	ชื่อรายวิชา
3	course_organization	สถาบันหรือองค์กร
4	course_type	ประเภทรายวิชา
5	course_language	ภาษาที่ใช้ในการเรียน
6	course_skills	ทักษะที่ใช้ในวิชา
7	category	วิชาหลัก
8	sub_category	วิชารอง
9	sub_course_list	รายวิชาที่เกี่ยวข้อง
10	course_url	ลิงก์รายวิชา
11	c_level	ระดับรายวิชา
12	rating	คะแนนของรายวิชาที่ทำการ Normalize แล้ว

course_no	course_name	course_organization	course_type	course_language	course_skills	category	sub_category	sub_course_list	course_link	c_level	rating
1	Atención prehospitalaria del ictus agudo y sel...	Universitat de Barcelona	COURSE	Spanish		Health	Patient Care		https://www.coursera.org/learn/ictus-agudo-esc...	Intermediate	3.139989
2	Brand & Content Marketing	IE Business School	COURSE	English		Arts and Humanities	History		https://www.coursera.org/learn/brand-and-conte...	Intermediate	4.200000
3	Cluster Analysis using RCmdr	Coursera Project Network	GUIDED PROJECT	English	['Cluster Analysis', 'Dendrogram', 'K-Means Cl...	Data Science	Data Analysis		https://www.coursera.org/learn/cluster-analysi...	Undefined	3.139989
4	Comercio, Inmigración y Tipos de Cambio en un ...	IE Business School	COURSE	Spanish		Social Sciences	Economics		https://www.coursera.org/learn/comercio-inmigr...	Beginner	3.139989
5	Create a Budget with Google Sheets	Coursera Project Network	GUIDED PROJECT	English	['Spreadsheet', 'Budget', 'Google Sheets', 'Te...	Business	Finance		https://www.coursera.org/learn/create-a-budget...	Undefined	4.800000

ภาพประกอบ 18 ตัวอย่างข้อมูลรายวิชา

ในการสำรวจข้อมูลเพื่อทำการพิจารณาว่าข้อมูลใดมีความจำเป็นที่สามารถนำข้อมูลเหล่านั้นไปใช้ในแบบจำลองได้ จากข้อมูลรายวิชาทั้งหมด มีข้อมูลรายวิชาจำนวน 5,599 แถวและ 12 คอลัมน์ ค้นพบว่าจำนวนภาษาที่ใช้ในการเรียนสูงสุดคือภาษาอังกฤษ ซึ่งเป็นภาษาสากล และภาษาอื่นๆ ที่อาจเป็นข้อจำกัดในการนำไปเข้าแบบจำลองได้ ดังภาพประกอบ 19 จำนวนวิชาหลัก (category) ทั้งหมด 12 วิชา และวิชารอง (sub_category) ทั้งหมด 46 วิชา ดังตัวอย่างภาพประกอบ 20 แสดงจำนวนวิชาหลักและวิชารอง นอกจากนี้ยังมีข้อมูลที่เป็นตัวอักษร ยกตัวอย่างเช่น ข้อมูลของทักษะที่ใช้ในรายวิชา (course_skills) และรายวิชาที่เกี่ยวข้อง (sub_course_list) ตามภาพประกอบที่ 18 ซึ่งสามารถนำมาใช้ในการวิเคราะห์ข้อความและนำไปใช้ในแบบจำลองเพื่อสร้างการแนะนำรายวิชาได้

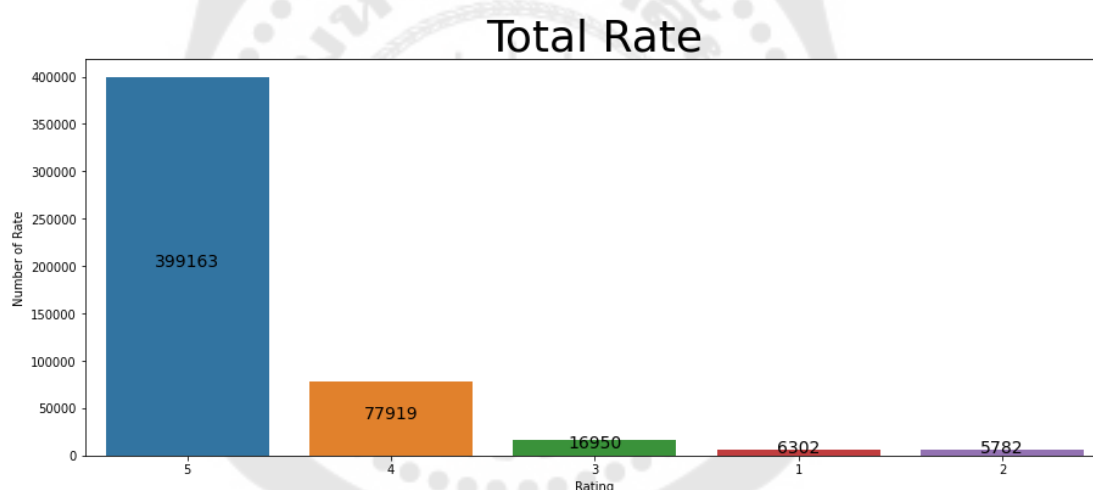


ภาพประกอบ 19 แสดงข้อมูลของภาษาที่ใช้เรียนในรายวิชา

course_no	course_name	date_reviews	rating	reviews	user_id
66	A Crash Course in Causality: Inferring Causal...	Sep 29, 2019	1	Can not download slides which make the source ...	72312
66	A Crash Course in Causality: Inferring Causal...	Nov 28, 2019	3	Clear deliver of engaging content. Very disapp...	21226
1633	English for Career Development	Aug 08, 2020	5	A simple yet great course for people who are s...	21226
1879	Forensic Accounting and Fraud Examination	Apr 02, 2018	5	A good course which outline basics and fundame...	21226
2104	Getting Started with Google Kubernetes Engine	Dec 09, 2019	5	Learned Kubernetes and CD	21226

ภาพประกอบ 23 ตัวอย่างข้อมูลการให้คะแนนจากผู้เรียน

จากข้อมูลการให้คะแนนนี้พบว่ามียาการที่เคยทำการให้คะแนนรายวิชาทั้งหมด 506,116 รายการ ซึ่งมาจากผู้เรียนทั้งหมด 287,808 คน และเป็นรายวิชาที่มีผู้ใช้งานให้คะแนนทั้งหมด 604 รายวิชา จากทั้งหมด 5,599 รายวิชา จำนวนการให้คะแนนของผู้เรียนถูกแบ่งออกเป็น 5 ระดับด้วยระดับ 1-5 คะแนน ดังภาพประกอบที่ 24

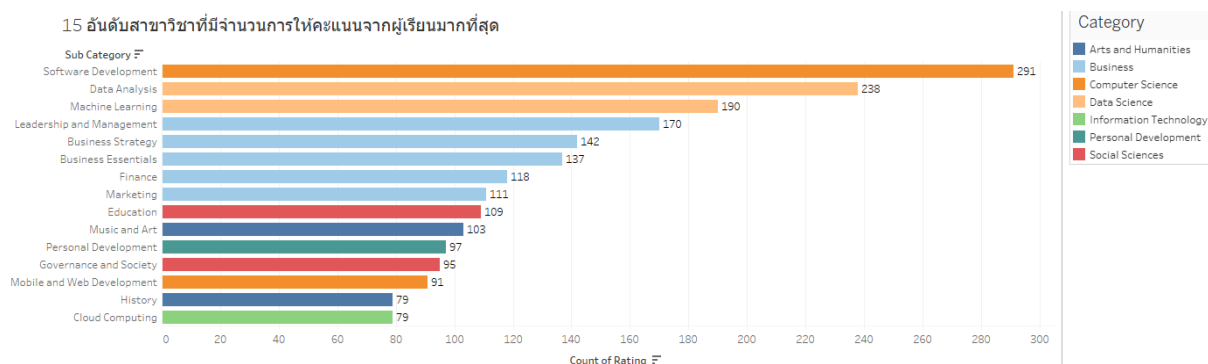


ภาพประกอบ 24 ข้อมูลจำนวนการให้คะแนนจากผู้เรียน

โดยที่ข้อมูลการให้คะแนนจากผู้เรียนมีคะแนนต่ำสุดคือ 1 และคะแนนสูงสุดคือ 5 โดยมีค่าเฉลี่ยอยู่ที่ 4.69 นอกจากนี้จากภาพประกอบ 25 ค้นพบว่า 5 อันดับรายวิชาที่มีจำนวนการให้คะแนนจากผู้เรียนมากที่สุด อันดับที่ 1 มีจำนวน 14,063 ครั้ง แต่ไม่สามารถระบุสาขาวิชาได้ ส่วนอันดับอื่นๆ เป็นสาขาวิชาที่เกี่ยวข้องกับ Data Science หรือสาขาวิชาในทาง Information Technology และจากภาพประกอบที่ 26 ได้แสดง 15 อันดับสาขาวิชาที่มีการให้คะแนนจากผู้เรียนมากที่สุด โดยแสดงรายการที่ระบุสาขาวิชา พบว่าสาขาวิชาที่ได้รับความสนใจในลำดับแรกคือ Computer Science, Data Science และสาขาวิชา Business ตามลำดับ

course_no	course_name	Count	category	sub_category
5214	Programming for Everybody (Getting Started wit...	14063.0	Undefined	Undefined
2997	Machine Learning	11487.0	Data Science	Machine Learning
3303	Neural Networks and Deep Learning	11330.0	Data Science	Machine Learning
5268	Learning How to Learn: Powerful mental tools t...	11048.0	Undefined	Undefined
4226	Technical Support Fundamentals	10903.0	Information Technology	Support and Operations

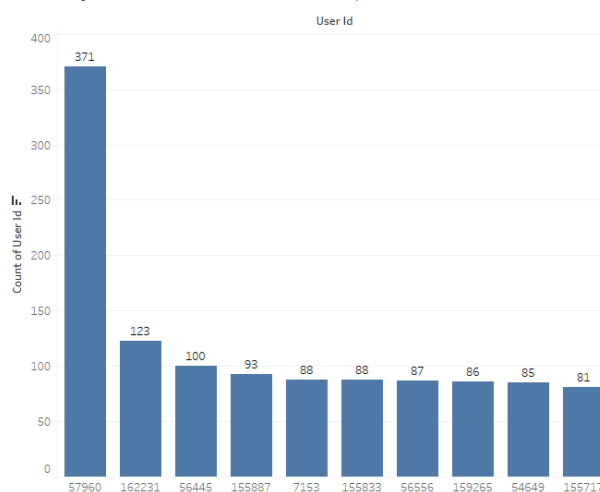
ภาพประกอบ 25 รายการ 5 อันดับรายวิชาที่มีผู้เรียนให้คะแนนมากที่สุด



ภาพประกอบ 26 รายการ 15 อันดับสาขาวิชาที่มีจำนวนการให้คะแนนจากผู้เรียนมากที่สุด

นอกจากนี้ได้ทำการค้นหา 10 อันดับผู้เรียน ที่มีจำนวนการให้คะแนนรายวิชามากที่สุด พบว่ามีผู้เรียนอันดับที่ 1 ให้คะแนนรายวิชาทั้งสิ้น 371 รายวิชา ดังภาพประกอบ 27 อาจมีความหมายได้ว่า ผู้เรียนคนนี้เป็นผู้ที่เรียนในรายวิชาของเว็บไซต์ Coursera มาก ซึ่งอาจมีอิทธิพลในการแนะนำรายวิชาให้กับผู้อื่นด้วย

10 อันดับผู้เรียน ที่มีจำนวนการให้คะแนนมากที่สุด



ภาพประกอบ 27 รายการ 10 อันดับผู้เรียนที่มีการให้คะแนนรายวิชามากที่สุด

ชุดข้อมูลทั้ง 2 ชุดนี้ถูกนำมาสำรวจและทำการวิเคราะห์ข้อมูลเบื้องต้น เพื่อหาความสัมพันธ์ของข้อมูล โดยทำข้อมูลให้มีความพร้อมในการใช้งานมากยิ่งขึ้น และทำการลบข้อมูลคอลัมน์ที่ไม่ใช้ได้แก่ course_organization, course_type, course_language, c_level เป็นต้น ซึ่งไม่มีผลในการนำไปค้นหาข้อมูล หรือไม่จำเป็นในการนำมาสร้างแบบจำลองในการให้คำแนะนำ จากนั้นจึงนำข้อมูลที่ได้ทำการคัดเลือกหรือพร้อมใช้งาน นำมาสร้างแบบจำลอง โดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) และเทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering) เพื่อทำการแนะนำรายวิชาให้กับผู้เรียน เมื่อผู้เรียนทำการค้นหารายวิชาที่ต้องการ

3.5 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเพื่อนำมาสร้างแบบจำลองการแนะนำ โดยการใช้เทคนิคการกรองข้อมูลตามเนื้อหา ใช้ชุดข้อมูลรายละเอียดของรายวิชา ไฟล์ Course Data โดยเลือกคอลัมน์ที่เกี่ยวข้องกับรายวิชามาใช้ในการวิเคราะห์ ได้แก่ course_name, category, sub_category, course_skills, sub_course_list ซึ่งข้อมูลที่เลือกมานี้เป็นส่วนหนึ่งของรายละเอียดของรายวิชาที่สามารถนำมาใช้ในการคำนวณหาความใกล้เคียงของรายวิชา หรือรายวิชาที่เกี่ยวข้องกันได้ ตัวอย่างข้อมูลของรายวิชา แสดงดังภาพประกอบ 28 โดยที่เมื่อนำข้อมูลมารวมกันแล้ว การนำข้อมูลจะไม่นำข้อความที่มีค่า Undefined มาคิดในการคำนวณครั้งนี้ด้วย

course_no	course_name	course_skills	category	sub_category	sub_course_list
5159	竞争策略 (中文版)	['Game Theory', 'Strategic Thinking', 'Competi...	Business	Business Strategy	[]
5160	算法基础	[]	Computer Science	Algorithms	[]
5161	算法设计与分析 Design and Analysis of Algorithms	[]	Computer Science	Algorithms	[]
5162	系统平台与C#程序基础	[]	Computer Science	Software Development	[]
5163	系统平台与计算环境	[]	Computer Science	Software Development	[]
5164	红楼梦(The Red Chamber Dream)	[]	Arts and Humanities	Philosophy	[]
5165	红楼梦-母体美译 (The Red Chamber Dream (2))	[]	Arts and Humanities	Philosophy	[]
5166	经济学概论: 贫困与市场 (Introduction to Economics: Incent...	[]	Social Sciences	Economics	[]
5167	结构生物化学 (Structural Biochemistry)	[]	Health	Basic Science	[]
5168	Google IT Support	['Binary Code', 'Customer Support', 'Linux', 'I...	Undefined	Undefined	['Technical Support Fundamentals', 'The Bits...
5169	IBM Data Science	['Data Science', 'Statistical Analysis', 'Mach...	Undefined	Undefined	['What is Data Science?', 'Tools for Data S...
5170	Introduction to Data Science	['Data Science', 'Relational Database Manageme...	Undefined	Undefined	['What is Data Science?', 'Tools for Data S...

ภาพประกอบ 28 แสดงตัวอย่างข้อมูลรายละเอียดของรายวิชา

ในการใช้เทคนิคการกรองข้อมูลตามเนื้อหานั้นแบ่งออกเป็น 2 ขั้นตอน โดยขั้นตอนที่ 1 ได้เลือกข้อมูลชื่อรายวิชา (course_name) และขั้นตอนที่ 2 ทำการเลือกข้อมูลรายละเอียดวิชา ซึ่งประกอบด้วย ชื่อรายวิชา(course_name), วิชาหลัก (category), วิชารอง (sub_category), ทักษะที่ใช้ในรายวิชา (course_skills) และรายวิชาที่เกี่ยวข้อง (sub_course_list) นำมารวมกัน

ข้อมูลทั้ง 2 ขั้นตอนนี้ทั้งข้อมูลรายวิชาและข้อมูลรายละเอียดของรายวิชาที่เลือกมา ได้ทำการรวบรวมข้อมูลคอลัมน์ที่เลือกมาทำการรวมกัน และทำ Text-Preprocessing เช่น การตัดคำที่ไม่สำคัญ (Stopwords) การทำ Regular Expression โดยลบอักขระ, สัญลักษณ์, ตัวเลขออก และเลือกใช้เพียงคำสำคัญ เพื่อนำไปสร้าง Vector แบบ TF-IDF ใช้ในการคำนวณในเทคนิคการกรองข้อมูลตามเนื้อหา และสร้างคำแนะนำหลักสูตรที่เกี่ยวข้องกับรายวิชาที่ผู้เรียนทำการค้นหา โดยทำการสร้างคอลัมน์เป็นข้อมูลใหม่เพื่อพร้อมนำไปสร้างแบบจำลอง

จากตัวอย่างภาพประกอบ 29 คอลัมน์ที่ชื่อว่า corpus_course_name ในคอลัมน์นี้ได้เตรียมไว้เพื่อนำไปสร้าง Vector TF-IDF ในขั้นตอนแรก โดยใช้ course_name เพียงข้อมูลเดียว และตัวอย่างภาพประกอบที่ 30 เป็นขั้นตอนที่ 2 ของการปรับปรุงการแนะนำรายวิชาของเทคนิคการกรองข้อมูลตามเนื้อหา คอลัมน์ที่ชื่อว่า corpus เป็นการเตรียมข้อมูลเพื่อนำไปสร้าง Vector TF-IDF เช่นเดียวกัน โดยใช้ข้อมูลคือ ชื่อรายวิชา (course_name), วิชาหลัก (category), วิชารอง (sub_category), ทักษะที่ใช้ในรายวิชา (course_skills), รายวิชาที่เกี่ยวข้อง (sub_course_list) ซึ่งเป็นการนำข้อมูลทั้งหมดมารวมกัน

course_no	course_name	corpus_course_name
51	3D Interaction Design in Virtual Reality	design interaction virtual reality
52	3D Interactions and Navigation	interactions navigation
53	3D Model Creation with Autodesk Fusion 360	autodesk creation fusion model
54	3D Models for Virtual Reality	reality models virtual
55	3D Printing Applications	printing applications
56	3D Printing Capstone	capstone printing

ภาพประกอบ 29 การทำ Text-Preprocessing ข้อมูล course_name

course_no	course_name	course_skills	category	sub_category	sub_course_list	corpus
5162	红楼梦 (The Red Chamber Dream) (2)		Arts and Humanities	Philosophy		red chamber dream artsandhumanities philosophy
5163	經濟學概論: 誘因與市場 (Introduction to Economics: Incenti...		Social Sciences	Economics		socialsciences markets introduction economics ...
5164	结构生物化学 (Structural Biochemistry)		Health	Basic Science		biochemistry health structural basicscience
5165	Google IT Support	['Binary Code', 'Customer Support', 'Linux', '...	Undefined	Undefined	['Technical Support Fundamentals', 'The Bits...	user support domain customer ldap protocol dir...
5166	IBM Data Science	['Data Science', 'Statistical Analysis', 'Mach...	Undefined	Undefined	['What is Data Science?', 'Tools for Data S...	intelligence cloud relational analysis sql pro...
5167	Introduction to Data Science	['Data Science', 'Relational Database Manage...	Undefined	Undefined	['What is Data Science?', 'Tools for Data S...	sql programming methodology python rdms lpyth...
5168	Python for Everybody	['Json', 'Xml', 'Python Programming', 'Databas...	Undefined	Undefined	['Programming for Everybody (Getting Started ...	scrapping dbms json getting started sql xml pro...

ภาพประกอบ 30 การทำ Text-Preprocessing โดยใช้ข้อมูลรายละเอียดของรายวิชา

สำหรับข้อมูลการให้คะแนนรายวิชาจากผู้เรียน ซึ่งใช้ข้อมูล Rating Data ได้ทำการเลือกคอลัมน์ที่เกี่ยวข้องได้แก่ user_id, course_no และ rating โดยที่คอลัมน์ rating มีค่าคะแนนระหว่าง 1-5 คะแนน เป็นตัวกำหนดหรือวัดความพึงพอใจของผู้เรียน และใช้ข้อมูลจากคอลัมน์

ดังกล่าวนี้มาทำการวิเคราะห์ข้อมูลโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ตัวอย่างข้อมูลการให้คะแนนแสดงดังภาพประกอบ 31 ซึ่งในชุดข้อมูลนี้ไม่มีค่าว่างและได้มีการจัดการกับข้อมูลที่มีความซ้ำซ้อนไปแล้ว

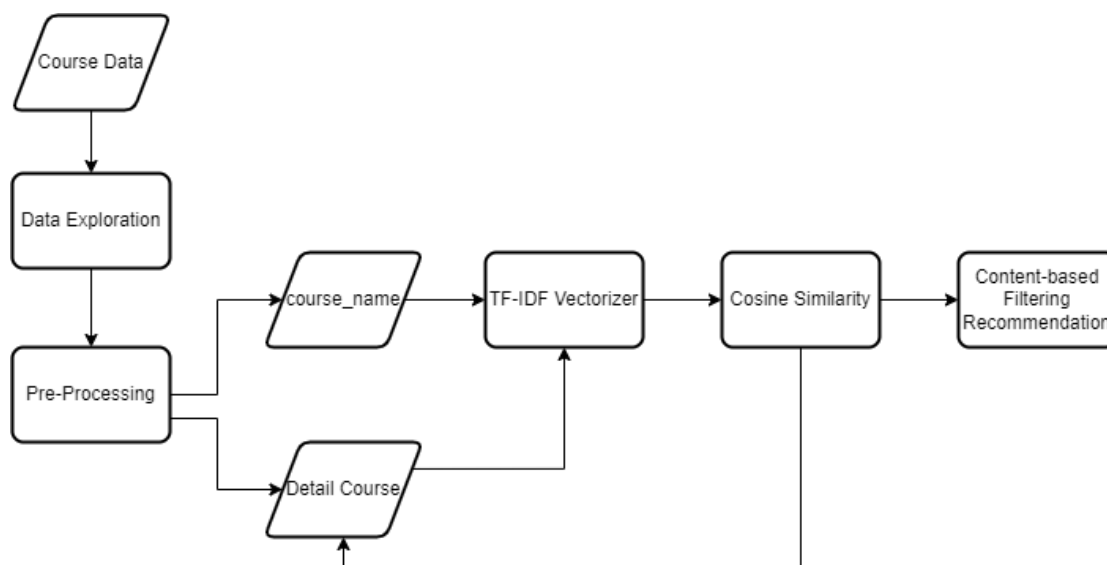
user_id	course_no	rating
151503	4405	5
151503	5241	3
236548	66	4
236548	2163	4
15217	66	4

ภาพประกอบ 31 ตัวอย่างข้อมูลการให้คะแนนรายวิชาจากผู้ใช้งาน

3.6 การสร้างแบบจำลอง

3.6.1 เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering)

การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) จากตัวอย่างภาพประกอบ 32 เริ่มต้นจากการสำรวจชุดข้อมูล Course Data และทำการวิเคราะห์ข้อมูลที่สามารถนำมาใช้ในการแนะนำรายวิชา จากนั้นทำการเตรียมข้อมูลเพื่อพร้อมที่จะนำมาใช้งาน โดยแบ่งข้อมูลออกเป็น 2 ชุดและนำไปทำกระบวนการแตกเวกเตอร์ของคำจากข้อมูลที่เตรียมไว้โดยใช้เทคนิคของ TF-IDF ซึ่งเป็นการคำนวณค่าน้ำหนักคำแต่ละคำ ดังตัวอย่างภาพประกอบ 33 คำในแต่ละคอลัมน์หมายถึงคำที่ได้มาจากคำในชื่อของแต่ละรายวิชา (course_name) โดยจะทำการทดสอบข้อมูล 2 ชุด ชุดแรกคือ course_name และชุดที่ 2 คือรายละเอียดของรายวิชา โดยทำกระบวนการเดียวกัน แต่ละคำที่ได้จะนำมาหาค่าความคล้ายคลึงกัน (Cosine Similarity) จากนั้นได้ทำการสร้างเป็นคำแนะนำของรายวิชาจากค่าความคล้ายคลึงของ Cosine ที่มีค่าสูงสุด 10 อันดับ



ภาพประกอบ 32 การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลตามเนื้อหา

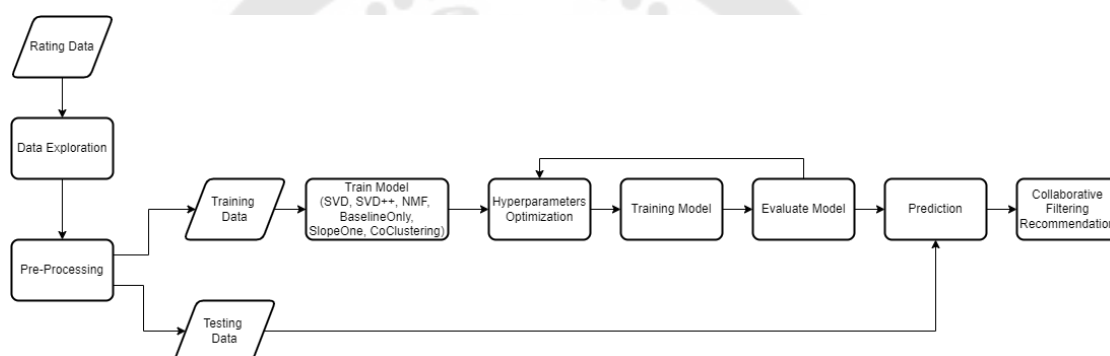
	ab	abc	abdomen	ableton	abogado	aborigin	abstract	ac	acad	academ
113	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.594398
114	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.561659
115	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.518411
116	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.711078
117	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.529595
1630	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.437405
2037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.641174
4658	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.553785
5225	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.675173
5396	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.470245
5529	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.566432

ภาพประกอบ 33 กระบวนการแตกเวกเตอร์คำโดยใช้เทคนิค TF-IDF

3.6.2 เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) จากภาพประกอบ 34 เริ่มต้นจากการนำชุดข้อมูล Rating Data มาทำการสำรวจข้อมูล และวิเคราะห์ข้อมูล จากนั้นเตรียมข้อมูลให้พร้อมก่อนนำไปใช้งานโดยทำการลบข้อมูลซ้ำ ซึ่งเป็นรายการที่มีผู้เรียนได้ทำการให้คะแนนซ้ำ โดยเลือกคอลัมน์ที่จำเป็นในการสร้างแบบจำลอง สำหรับการให้ Library ของ Scikit Learn ที่มีชื่อว่า Surprise เป็นไลบรารีที่เป็นการสร้าง Recommender System โดยเฉพาะ โดยใช้ข้อมูลที่เป็นคะแนน (rating), รหัสผู้ใช้งาน (user_id), และรายการ

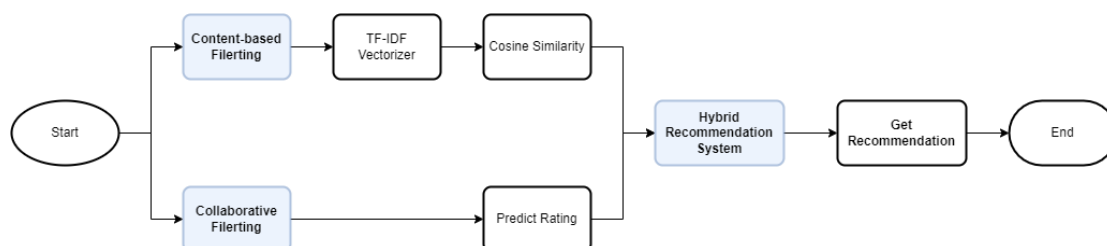
(item_id) โดยที่ item_id ของข้อมูลที่ทำกรวิจัยนี้คือรายวิชาหรือ course_no ดังนั้นข้อมูลที่จะนำมาใช้ในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมในงานวิจัยนี้ใช้ข้อมูล user_id, course_no และ rating เท่านั้น จากนั้นทำการแบ่งสัดส่วนของข้อมูล (Split Data) ออกเป็น ข้อมูลการสอน (Training Data) และข้อมูลการทดสอบ (Testing Data) ในอัตราส่วน 80:20 ตามลำดับ โดยนำ Training Data มาทดสอบกับอัลกอริทึมของ Surprise ได้แก่ SVD, SVD++, NMF, Slope One, Co-Clustering และ BaselineOnly เป็นต้น จากนั้นทำการเปรียบเทียบประสิทธิภาพการทำงานของแต่ละอัลกอริทึม และประเมินประสิทธิภาพของอัลกอริทึม โดยใช้ค่า Root Mean Square Error (RMSE) และ Mean Absolute Error (MAE) ที่มีค่าน้อยที่สุดซึ่งเป็นค่าที่ดีแสดงว่าแบบจำลองที่ได้ประมาณค่าได้ใกล้เคียงกับความจริง



ภาพประกอบ 34 การสร้างแบบจำลองตามเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม

3.6.3 เทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering)

การสร้างแบบจำลองของเทคนิคการกรองข้อมูลแบบผสม ได้นำวิธีการจากเทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) และเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) มาทำงานร่วมกัน โดยเริ่มต้นจากข้อมูลผู้ใช้คนหนึ่ง โดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมที่ให้ผลลัพธ์การประเมินประสิทธิภาพของแบบจำลองที่มีค่า RMSE และ MAE ที่ดีที่สุดหรือมีค่าต่ำที่สุด จากนั้นนำผลลัพธ์ที่ได้มาประมวลผลรวมกับการใช้เทคนิคการกรองข้อมูลตามเนื้อหา เมื่อแบบจำลองทั้งสองทำงานร่วมกัน จึงได้ทำการแนะนำค่าสูงสุด 5 อันดับในการแนะนำรายวิชาให้กับผู้เรียน ดังภาพประกอบที่ 36



ภาพประกอบ 35 แสดงแบบจำลองของเทคนิคการกรองแบบผสม

3.7 สรุปผลการดำเนินงาน

เทคนิคการกรองข้อมูลตามเนื้อหา ได้ทำการสร้างคำแนะนำรายวิชาจากค่าความคล้ายคลึงกันของโคไซน์ (Cosine Similarity) ที่มีค่าสูงสุด 10 อันดับ โดยใช้ค่าความคล้ายคลึงกันเป็นตัววัดประสิทธิภาพ ส่วนเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม ทำการเลือกอัลกอริทึมจากการเปรียบเทียบประสิทธิภาพการทำงานของแต่ละอัลกอริทึม และประเมินประสิทธิภาพของอัลกอริทึมที่มีค่า RMSE และ MAE ดีที่สุดซึ่งคือค่าที่ต่ำที่สุด มาทำการปรับพารามิเตอร์เพื่อนำไปสร้างแบบจำลอง ส่วนเทคนิคการแนะนำแบบผสมได้ใช้วิธีการของการทำงานของทั้งเทคนิคการกรองข้อมูลตามเนื้อหา และการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมที่สร้างมาจากขั้นตอนก่อนหน้านี้ มาทำงานร่วมกันเพื่อสร้างคำแนะนำรายวิชาสูงสุด 5 อันดับ

บทที่ 4

ผลการดำเนินงานวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงกรณีศึกษาข้อมูลรายวิชาบนเว็บไซต์ Coursera เพื่อมุ่งเน้นการศึกษาระบบการแนะนำรายวิชา ผู้วิจัยได้ดำเนินการศึกษาตามกระบวนการของแต่ละวิธีเพื่อให้บรรลุวัตถุประสงค์ในการวิจัยที่ได้กำหนดไว้ ดังนี้

1. ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา
2. ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม
3. ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสม

4.1 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา

เทคนิคการกรองข้อมูลตามเนื้อหา คือ เทคนิคที่ใช้คุณลักษณะข้อมูลที่เกี่ยวข้องกับรายวิชามาใช้ในการวิเคราะห์ค่า และทำการแนะนำรายวิชาที่มีความเกี่ยวข้องกัน ผู้วิจัยได้ใช้เทคนิคของ TF-IDF ซึ่งเป็นการคำนวณค่าน้ำหนักคำแต่ละคำ โดยมาจากกระบวนการแตก Vector ของคำจากข้อมูลที่เตรียมไว้ก่อนหน้านี้ มาทำการแบ่งข้อมูลออกเป็น 2 ชุดซึ่งได้แก่ ข้อมูลชื่อรายวิชา (course_name) และข้อมูลรายละเอียดรายวิชา ซึ่งประกอบด้วยคอลัมน์ดังนี้ ชื่อรายวิชา (course_name), ทักษะที่ใช้ในรายวิชา (course_skills), วิชาหลัก (category), วิชารอง (sub_category) และรายวิชาที่เกี่ยวข้อง (sub_course_list) เป็นต้น จากนั้นจึงนำชุดข้อมูลแต่ละชุดมาทำการคำนวณค่าความคล้ายคลึงของ Cosine ที่มีค่าสูงสุด 10 อันดับ เพื่อทำการแนะนำรายวิชาให้กับผู้เรียน

จากการทดลองนำข้อมูลที่ได้มาทำการคำนวณค่าความคล้ายคลึงแบบโคไซน์ โดยได้ค้นหารายวิชาที่ชื่อว่า Creating Toolpaths for a CNC Lathe ซึ่งเป็นวิชาหลักคือ Physical Science and Engineering และวิชารองคือ Mechanical Engineering จากภาพประกอบ 36 พบว่าการใช้ข้อมูลสำหรับการค้นหาจากชื่อรายวิชาเพียงข้อมูลเดียว ทำให้เกิดการแนะนำรายวิชาโดยมีรายวิชาอื่นเข้ามาปะปนด้วย และเมื่อทำการคำนวณในขั้นตอนที่ 2 โดยใช้รายละเอียดข้อมูลของรายวิชาพบว่าชุดข้อมูลที่ได้ทำการเลือกมาแนะนำ มีค่าความคล้ายคลึงกันของโคไซน์ที่สูงกว่าการใช้เพียงแค่ชื่อรายวิชา และทำการแนะนำรายการของรายวิชาได้ตรงกับหมวดหมู่มารายวิชาที่ทำการค้นหา ดังตัวอย่างในภาพประกอบ 37

course_no	course_name	category	sub_category	similar_score
3238	Multi-Axis CNC Toolpaths	Physical Science and Engineering	Mechanical Engineering	0.542246
2588	Introduction to CAD, CAM, and Practical CNC Ma...	Physical Science and Engineering	Mechanical Engineering	0.276406
1177	Creating Innovation	Business	Leadership and Management	0.195944
1116	Create Your First Python Program	Computer Science	Software Development	0.172020
1188	Creating a Portfolio	Business	Finance	0.169083
3650	Project: Creating Your First C++ Application	Computer Science	Software Development	0.168240
1118	Create a 5s System in Google Sheets	Business	Business Strategy	0.165317
1113	Create Your First Game with Python	Computer Science	Software Development	0.165105
1110	Create Your First Application with Java Using ...	Computer Science	Software Development	0.126441
1168	Create basic behavior with C# in Unity	Computer Science	Software Development	0.123424

ภาพประกอบ 36 การแนะนำรายวิชาโดยใช้ข้อมูลชี้อรายวิชา

course_no	course_name	category	sub_category	course_skills	sub_course_list	similar_score
3238	Multi-Axis CNC Toolpaths	Physical Science and Engineering	Mechanical Engineering	['Computer-aided manufacturing', 'Engineering', ...	[]	0.815374
2679	Introduction to Mechanical Engineering Design ...	Physical Science and Engineering	Mechanical Engineering	['Autodesk', 'Manufacturing Process Management', ...	[]	0.725241
44	3-Axis Machining with Autodesk Fusion 360	Physical Science and Engineering	Mechanical Engineering	['Manufacturing Process Management', 'Computer...	[]	0.714286
2043	Generative Design for Additive Manufacturing	Physical Science and Engineering	Mechanical Engineering	['Manufacturing Process Management', 'Computer...	[]	0.714286
2588	Introduction to CAD, CAM, and Practical CNC Ma...	Physical Science and Engineering	Mechanical Engineering	['Manufacturing Process Management', 'Computer...	[]	0.713024
5539	Autodesk CAD/CAM for Manufacturing	Undefined	Undefined	['Autodesk', 'Computer-Aided Design (CAD)', 'C... ['Introduction to CAD, CAM, and Practical CNC...	[]	0.709208
3083	Manufacturing Process with Autodesk Fusion 360	Physical Science and Engineering	Mechanical Engineering	[]	[]	0.654654
754	CAM and Design Manufacturing for Mechanical En...	Physical Science and Engineering	Mechanical Engineering	['Computer-Aided Design (CAD)', 'Computer-Aide...	[]	0.642857
2044	Generative Design for Industrial Applications	Physical Science and Engineering	Mechanical Engineering	['Computer-Aided Design (CAD)', 'Autodesk Fusi...	[]	0.642857
2046	Generative Design for Performance and Weight R...	Physical Science and Engineering	Mechanical Engineering	['Manufacturing Process Management', 'product ...	[]	0.621059

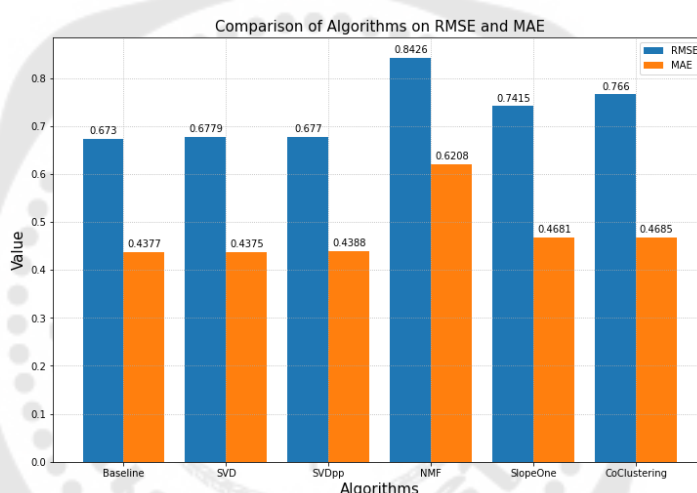
ภาพประกอบ 37 การแนะนำรายวิชาโดยใช้รายละเอียดรายวิชา

4.2 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบพึงพาผู้เข้าร่วม

เทคนิคการกรองข้อมูลแบบพึงพาผู้เข้าร่วม คือ เทคนิคที่ใช้การให้คะแนนรายการจากผู้ใช้งานคนอื่นมาทำการแนะนำ ข้อมูลที่ใช้ในแบบจำลองนี้ประกอบด้วย ข้อมูลที่เป็นคะแนน (rating), รหัสผู้ใช้งาน (user_id), และรายวิชา (course_no) จากนั้นทำการแบ่งสัดส่วนของข้อมูล (Split Data) ออกเป็นข้อมูลการสอน (Training Data) และข้อมูลการทดสอบ (Testing Data) ในอัตราส่วน 80:20 ตามลำดับ และนำชุดข้อมูลการสอน มาทำการทดสอบกับอัลกอริทึมของ Surprise ได้แก่ SVD, SVD++, NMF, SlopeOne, Co-Clustering และ BaselineOnly โดยประเมินประสิทธิภาพด้วยการทำ Cross-Validation ทั้งหมด 5 Folds จากนั้นทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึม โดยใช้ Root Mean Square Error (RMSE) และ Mean Absolute Error (MAE) ที่มีค่าน้อยที่สุดซึ่งเป็นค่าที่ดีแสดงว่าแบบจำลองที่ได้ประมาณค่าได้ใกล้เคียงกับความจริง ดังที่แสดงในตาราง 5 และภาพประกอบ 38

ตาราง 5 การประเมินประสิทธิภาพของอัลกอริทึม

Algorithm	RMSE	MAE
Baseline	0.673	0.4377
SVD	0.6779	0.4375
SVD++	0.677	0.4388
NMF	0.8462	0.6208
Slope One	0.7415	0.4681
Co-Clustering	0.766	0.4685



ภาพประกอบ 38 แสดงการเปรียบเทียบประสิทธิภาพของอัลกอริทึมโดยใช้ RMSE และ MAE

จากผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพบว่าอัลกอริทึม BaselineOnly มีค่าดีที่สุด ซึ่งเป็นอัลกอริทึมพื้นฐานที่ไม่ได้ผลมากนัก เนื่องจากไม่เป็นที่นิยมและอาจถูกเบี่ยงเบนค่าจากผู้ที่ใช้ที่มีการให้คะแนนหลากหลายรายวิชาว่าผู้เรียนคนอื่น หรือรายวิชาที่มีความนิยมมากกว่ารายวิชาอื่น แต่มีประโยชน์สำหรับการเปรียบเทียบความแม่นยำ ซึ่งได้ค่า RMSE ที่ต่ำที่สุดในบรรดาอัลกอริทึมทั้งหมด ผู้วิจัยจึงได้ทำการเลือกอัลกอริทึมตัวอื่นที่มีประสิทธิภาพดีที่สุด รองลงมาโดยใช้ค่า RMSE และ MAE โดยได้เลือกอัลกอริทึม SVD และ SVD++ เนื่องจากมีค่าประสิทธิภาพดีที่สุดจากอัลกอริทึมทั้งหมด มาทำการปรับพารามิเตอร์เพื่อนำไปใช้แนะนำรายวิชา และได้ทำการใช้เทคนิค GridSearch ทั้งหมด 5 Folds Cross-Validation สำหรับขั้นตอน Training Data เพื่อทำการปรับพารามิเตอร์ที่เหมาะสมของอัลกอริทึม SVD, SVD++ จากค่าพารามิเตอร์ดังนี้

n_epochs: [15, 25]

lr_all: [0.009, 0.01]

reg_all: [0.5, 0.9]

n_factors: [15, 30]

สำหรับผลลัพธ์ของการปรับพารามิเตอร์โดยการทำ 5 Folds Cross-Validation จากการทำ GridSearch ของอัลกอริทึม SVD, SVD++ พบว่าค่าของพารามิเตอร์ไม่ต่างกันมากนักดังตาราง 6

ตาราง 6 แสดงค่า Best Parameter จากการปรับพารามิเตอร์

Algorithm	Evaluation	n_epochs	lr_all	reg_all	n_factors
SVD	RMSE	15	0.009	0.5	15
	MAE	25	0.01	0.5	15
SVD++	RMSE	15	0.009	0.5	30
	MAE	25	0.01	0.5	15

จากนั้นจึงนำพารามิเตอร์ที่ได้จากการปรับพารามิเตอร์มาใช้กับชุดข้อมูล Testing Data และวัดประสิทธิภาพของอัลกอริทึมของ SVD และ SVD++ พบว่าไม่แตกต่างกัน ซึ่งทำการเปรียบเทียบประสิทธิภาพโดยใช้ RMSE ค่าที่ได้ทำการเปรียบเทียบแสดงดังตาราง 7

ตาราง 7 แสดงประสิทธิภาพของแบบจำลอง

Algorithm	Train Set	Test Set
SVD	0.6755	0.6768
SVD++	0.6759	0.6768

จากนั้นผู้วิจัยได้ทำการเลือกอัลกอริทึม SVD มาสร้างแบบจำลองของเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้งาน พบว่าการนำเสนอรายวิชาให้กับผู้เรียน มีค่าการทำนายคะแนนรายวิชาค่อนข้างสูง รวมไปถึงการแนะนำรายวิชาในหลักสูตรอื่นๆ ทำให้มีการแนะนำรายวิชาที่หลากหลายมากยิ่งขึ้น นอกเหนือจากที่ผู้เรียนเคยลงทะเบียนเรียนมาก่อน ตัวอย่างจากภาพประกอบ 39 ได้ทำการค้นหาผู้เรียนรหัส 286193 พบว่าได้มีการลงทะเบียนเรียน 1 วิชาและมีการให้คะแนนเท่ากับ 5

สำหรับเทคนิคการกรองข้อมูลแบบพึ่งพาแบบผู้ใช้นี้ ได้นำข้อมูลดังกล่าวไปทำการแนะนำรายวิชาที่ผู้เรียนมีความชอบคล้ายกัน และแนะนำรายวิชาดังภาพประกอบ 40 ซึ่งข้อมูล predictions คือค่าคะแนนที่ทำนายจากแบบจำลองของ SVD

```
1 hb_user , hb_collaborative, hb_recommend = hybrid_recommend(286193,interactions,course_df,ratings_df,5)
2 hb_user.head()
```

User 286193 has already rated 1 Course.
Succeeded 1 Course.

	user_id	course_no	rating_x	course_name	category	sub_category
0	286193	2703	5	Introduction to Programming with MATLAB	Computer Science	Software Development

ภาพประกอบ 39 การค้นหาประวัติของผู้เรียน 286193

```
1 hb_collaborative
```

	course_no	predictions	course_name	category	sub_category
0	1232	4.947103	Cultural Competence - Aboriginal Sydney	Arts and Humanities	History
1	579	4.919763	Bugs 101: Insect-Human Interactions	Health	Basic Science
2	1577	4.917183	El Abogado del Futuro: Legaltech y la Transfor...	Business	Leadership and Management
3	612	4.906832	Build a Modern Computer from First Principles:...	Computer Science	Software Development
4	4658	4.900740	Writing in the Sciences	Health	Basic Science

ภาพประกอบ 40 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เรียน 286193

ตัวอย่างการค้นหาจากผู้เรียนคนที่ 2 ผู้เรียนรหัส 4412 พบว่าได้มีการลงทะเบียนเรียน 6 วิชา และระบบได้ทำการเลือกรายวิชาที่มีการให้คะแนนสูงสุด 5 อันดับแรก โดยมีประวัติการเรียนรู้จากรายวิชาตามภาพประกอบ 41 จากนั้นระบบได้ทำการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เรียนจากแบบจำลองของ SVD ตามภาพประกอบ 42

```
1 hb_user2 , hb_collaborative2, hb_recommend2 = hybrid_recommend(4412,interactions, course_df, ratings_df,5)
2 hb_user2.head()
```

User 4412 has already rated 6 Course.
Succeeded 6 Course.

	user_id	course_no	rating_x	course_name	category	sub_category
0	4412	2162	5	Google Cloud Platform Fundamentals: Core Infra...	Information Technology	Cloud Computing
1	4412	3344	5	Operating Systems and You: Becoming a Power User	Information Technology	Support and Operations
2	4412	4223	5	Technical Support Fundamentals	Information Technology	Support and Operations
3	4412	4275	5	The Bits and Bytes of Computer Networking	Information Technology	Networking
4	4412	5325	5	System Administration and IT Infrastructure Se...	Undefined	Undefined

ภาพประกอบ 41 การค้นหาประวัติของผู้เรียน 4412

```
1 hb_collaborative2
```

	course_no	predictions	course_name	category	sub_category
0	3798	5.0	Removing Barriers to Change	Business	Leadership and Management
1	4658	5.0	Writing in the Sciences	Health	Basic Science
2	1577	5.0	El Abogado del Futuro: Legaltech y la Transfor...	Business	Leadership and Management
3	1735	5.0	Everyday Excel, Part 2	Computer Science	Design and Product
4	5544	5.0	Aprendiendo a aprender: Poderosas herramientas...	Undefined	Undefined

ภาพประกอบ 42 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพื้พหุผู้เข้าร่วมของผู้เรียน

4412

ตัวอย่างการค้นหาจากผู้เรียนคนที่ 3 ผู้เรียนรหัส 95424 พบว่าได้มีการลงทะเบียนเรียน 1 วิชา และมีการให้คะแนนเท่ากับ 5 จากภาพประกอบ 43 จากนั้นได้ทำการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบพื้พหุแบบผู้เข้าร่วมจากแบบจำลองของ SVD ตามภาพประกอบ 44

```
1 hb_user3 , hb_collaborative3, hb_recommend3 = hybrid_recommend(95424,interactions, course_df, ratings_df,5)
2 hb_user3.head()
```

User 95424 has already rated 1 Course.
Succeeded 1 Course.

	user_id	course_no	rating_x	course_name	category	sub_category
0	95424	5418	5	Oil & Gas Industry Operations and Markets	Undefined	Undefined

ภาพประกอบ 43 การค้นหาประวัติของผู้เรียน 95424

```
1 hb_collaborative3
```

	course_no	predictions	course_name	category	sub_category
0	579	4.915119	Bugs 101: Insect-Human Interactions	Health	Basic Science
1	1577	4.910701	El Abogado del Futuro: Legaltech y la Transfor...	Business	Leadership and Management
2	1232	4.909364	Cultural Competence - Aboriginal Sydney	Arts and Humanities	History
3	612	4.904944	Build a Modern Computer from First Principles:...	Computer Science	Software Development
4	2296	4.904400	How to Manage a Remote Team	Business	Leadership and Management

ภาพประกอบ 44 การแนะนำรายวิชาโดยเทคนิคการกรองข้อมูลแบบพื้พหุผู้เข้าร่วมของผู้เรียน

95424

4.3 ผลลัพธ์ของการแนะนำรายวิชาโดยใช้เทคนิคการกรองข้อมูลแบบผสม

เทคนิคการกรองข้อมูลแบบผสม เป็นการทำงานร่วมกันระหว่างเทคนิคการกรองข้อมูลตามเนื้อหาและเทคนิคการกรองข้อมูลแบบพื้พหุผู้เข้าร่วม เพื่อทำให้เกิดการแนะนำที่หลากหลายและลดข้อบกพร่องของแต่ละวิธี โดยเริ่มต้นจากการนำเทคนิคการกรองข้อมูลแบบพื้พหุผู้เข้าร่วม

มาทำการแนะนำรายวิชาที่ผู้เรียนยังไม่เคยเรียนมาก่อน โดยใช้ค่าการทำนายค่าคะแนนสูงสุด 5 รายวิชา จากนั้นจึงนำข้อมูลของรายละเอียดรายวิชาที่มาจากการแนะนำของกระบวนการในเทคนิคการกรองข้อมูลแบบฟิงพาผู้เข้าร่วมข้างต้น มาทำการแนะนำโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา ซึ่งได้มีการกำหนดค่าความคล้ายคลึงกันของโคไซน์ที่มีค่ามากกว่า 0.55 เนื่องจากเป็นค่าที่สามารถทำการแนะนำรายวิชาในจำนวนที่เหมาะสม หากกำหนดค่าโคไซน์ที่มากเกินไประบบทำการแนะนำรายวิชาได้จำนวนค่อนข้างน้อย เนื่องจากการแนะนำบางรายวิชาให้ค่าโคไซน์นี้ไม่มากนัก จากนั้นนำรายวิชาที่ได้รับการแนะนำจากเทคนิคการกรองข้อมูลตามเนื้อหา มาเรียงอันดับตามค่าความเหมือนของ Cosine ที่มีค่าสูงสุด 5 อันดับ

จากการทดลองพบว่า วิธีการผสมผสานระหว่างสองเทคนิคทำให้มีการแนะนำรายวิชาได้หลากหลายมากกว่าการใช้เทคนิคเดียว เป็นการแนะนำในสาขาวิชาอื่นโดยมาจากการอ้างอิงข้อมูลจากผู้เรียนคนอื่นในระบบ และนำเสนอรายวิชาที่ผู้เรียนยังไม่เคยให้คะแนนมาก่อน จากภาพประกอบ 45 คือ การแนะนำรายวิชาแบบผสมโดยใช้รหัสผู้เรียน 286193 ภาพประกอบ 46 คือ การแนะนำรายวิชาแบบผสมโดยใช้รหัสผู้เรียน 4412 และภาพประกอบ 47 คือ การแนะนำรายวิชาแบบผสมโดยใช้รหัสผู้เรียน 95424 ตามลำดับ

1 hb_recommend

course_no	course_name	category	sub_category
604	605	Build a Google Cloud Firestore Web App (Part I)	Computer Science Mobile and Web Development
605	606	Build a Google Firebase Web Application	Computer Science Mobile and Web Development
3645	3646	Project Planning	Business Leadership and Management
2703	2704	Introduction to Project Management	Business Leadership and Management
5132	5133	現代文學導讀：詩、散文、小說	Arts and Humanities History

ภาพประกอบ 45 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 286193

1 hb_recommend2

course_no	course_name	category	sub_category
196	197	Advanced Writing	Language Learning Learning English
116	117	Academic Literacy	Language Learning Learning English
4223	4224	Technical Writing	Language Learning Learning English
2850	2851	L'impact investing : comprendre les fondamentaux	Business Entrepreneurship
2852	2853	L'impact investing : être acteur	Business Entrepreneurship

ภาพประกอบ 46 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 4412

1 hb_recommend3

	course_no	course_name	category	sub_category
604	605	Build a Google Cloud Firestore Web App (Part I)	Computer Science	Mobile and Web Development
605	606	Build a Google Firebase Web Application	Computer Science	Mobile and Web Development
3645	3646	Project Planning	Business	Leadership and Management
1922	1923	From Idea to Startup	Business	Entrepreneurship
2703	2704	Introduction to Project Management	Business	Leadership and Management

ภาพประกอบ 47 การแนะนำรายวิชาโดยใช้เทคนิคการแนะนำแบบผสมของผู้เรียน 95424



บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยเพื่อศึกษาและเปรียบเทียบการแนะนำรายวิชาบนเว็บไซต์ Coursera โดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) และเทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering) ผู้วิจัยได้ทำการประเมินประสิทธิภาพของแบบจำลองแต่ละเทคนิค โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

สรุปผลการวิจัย

ในการวิจัยนี้เป็นการศึกษาระบบการแนะนำโดยใช้เทคนิคของการผสมผสานกันระหว่างเทคนิคการกรองข้อมูลตามเนื้อหา ร่วมกับเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม ซึ่งเป็นการนำจุดเด่นของแต่ละเทคนิคมาผสมผสานกัน ส่งผลให้ระบบแนะนำสามารถนำเสนอรายวิชาได้หลากหลายมากยิ่งขึ้น โดยทำการศึกษา 3 วิธี ดังต่อไปนี้

วิธีที่ 1 การกรองข้อมูลตามเนื้อหาโดยใช้เทคนิคของ TF-IDF ซึ่งเป็นการคำนวณน้ำหนักคำซึ่งมาจากกระบวนการแตกเวกเตอร์ นำมาหาค่าความคล้ายคลึงของ Cosine เพื่อทำการแนะนำรายวิชาให้กับผู้เรียน โดยมีการใช้ข้อมูลทั้งหมด 6 คอลัมน์และแบ่งข้อมูลออกเป็น 2 ขั้นตอน โดยขั้นตอนที่ 1 ได้เลือกข้อมูลชื่อรายวิชา (course_name) และขั้นตอนที่ 2 ทำการเลือกข้อมูลรายละเอียดวิชา ซึ่งประกอบด้วย ชื่อรายวิชา(course_name), วิชาหลัก (category), วิชารอง (sub_category), ทักษะที่ใช้ในรายวิชา (course_skills) และรายวิชาที่เกี่ยวข้อง (sub_course_list) นำข้อมูลรายละเอียดวิชามารวมกันและนำข้อมูลเหล่านี้มาทำการวิเคราะห์ในการแนะนำรายวิชาให้กับผู้เรียน ผลลัพธ์ที่ได้พบว่าการเลือกใช้ชุดข้อมูลรายละเอียดวิชา ได้ทำการแนะนำรายวิชาได้ตรงกับหมวดหมู่ของรายวิชาที่ทำการค้นหา และมีค่าความคล้ายคลึงกันของโคไซน์ที่สูงกว่าการใช้เพียงแค่ชื่อรายวิชา

วิธีที่ 2 การกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม โดยนำข้อมูลการให้คะแนน, รหัสผู้ใช้งาน และรหัสรายวิชา มาทำการทดสอบกับอัลกอริทึมซึ่งได้แก่ SVD, SVD++, NMF, SlopeOne,

Co-Clustering และ BaselineOnly จากนั้นวัดประสิทธิภาพของแบบจำลองด้วยค่า RMSE และ MAE โดยทำการเลือกอัลกอริทึมที่ดีที่สุด ซึ่งพบว่า SVD มีค่า RMSE และ MAE ดีที่สุดในบรรดาอัลกอริทึมทั้งหมดเท่ากับ 0.6779 และ 0.4375 ตามลำดับ จึงนำอัลกอริทึม SVD มาสร้างการแนะนำรายวิชาในเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมและในเทคนิคการกรองข้อมูลแบบผสมในกระบวนการต่อไป

วิธีที่ 3 การกรองข้อมูลแบบผสม เป็นการทำงานที่ผสมผสานกันระหว่างการใช้เทคนิคการกรองข้อมูลตามเนื้อหาและการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วมที่ได้มาจากวิธีก่อนหน้านี้ โดยเริ่มต้นจากการใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม เป็นการนำอัลกอริทึม SVD มาทำการสร้างแบบจำลอง เพื่อทำการแนะนำรายวิชาโดยอ้างอิงข้อมูลมาจากผู้เรียนคนอื่น จากนั้นรายวิชาที่ถูกทำการแนะนำ ได้ถูกนำไปแนะนำรายวิชาในขั้นตอนต่อไปโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา ซึ่งถือว่าเป็นการแนะนำรายวิชาโดยใช้การทำงานร่วมกันระหว่าง 2 เทคนิค กลายเป็นเทคนิคการกรองข้อมูลแบบผสม เพื่อทำการแนะนำรายวิชาให้กับผู้เรียน ให้มีความหลากหลายและอัตโนมัติโดยอ้างอิงจากประวัติของผู้เรียน

อภิปรายผลการวิจัย

งานวิจัยนี้ได้ทำการศึกษาระบบแนะนำโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา (Content-based Filtering) ร่วมกับเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม (Collaborative Filtering) และนำเทคนิคดังกล่าวมาทำงานร่วมกันกลายเป็นเทคนิคการกรองข้อมูลแบบผสม (Hybrid Filtering) ซึ่งเป็นการนำจุดเด่นของแต่ละเทคนิคมาผสมผสานกัน อีกทั้งยังช่วยลดข้อด้อยในแต่ละเทคนิค

โดยเริ่มต้นจากเทคนิคการกรองข้อมูลตามเนื้อหา เป็นการแนะนำในรายวิชาที่เกี่ยวข้องและมีความสัมพันธ์กัน โดยใช้รายละเอียดคุณลักษณะของรายวิชาเป็นส่วนสำคัญ จึงทำให้ทุกรายวิชามีโอกาสถูกนำมาเสนอให้กับผู้เรียนได้ หากคุณลักษณะของข้อมูลรายวิชามีความคล้ายหรือเหมือนกับรายวิชาที่ทำการค้นหา สำหรับอุปสรรคของการใช้เทคนิคการกรองข้อมูลตามเนื้อหา คือ ชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้ บางส่วนของข้อมูลมีข้อมูลไม่ครบถ้วนและคุณลักษณะของคำที่ไม่ชัดเจนหรือซ้ำซ้อน ตัวอย่าง ข้อมูลทักษะที่ใช้ในรายวิชา เช่น Html, HTML Tables ควรระบุเป็น HTML เพราะมีความหมายที่คล้ายคลึงกัน ซึ่งข้อมูลในลักษณะนี้มีการจัดการที่ค่อนข้างยากเนื่องจากข้อมูลมีจำนวนมาก ทำให้ข้อมูลมีคุณภาพไม่สมบูรณ์ในกระบวนการทำการสกัดคุณลักษณะที่สำคัญ กระบวนการใช้เทคนิคการกรองข้อมูลตามเนื้อหาจึงอาจทำได้ไม่ดีนัก และข้อเสียจากเทคนิคการกรองตามเนื้อหานี้ใช้ระยะเวลาในการเตรียมข้อมูลมาก เพื่อทำการ

สำรวจข้อมูลให้ข้อมูลที่ได้ทำการเลือกมาใช้ในเทคนิคมีความครบถ้วนมากที่สุด เพื่อนำมาสร้างคุณลักษณะที่เหมาะสมของการแนะนำรายวิชา

สำหรับเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมเป็นการนำข้อมูลของผู้เรียนคนที่ต้องการค้นหา มาทำการแนะนำรายวิชาอื่นๆ โดยใช้ข้อมูลอ้างอิงจากผู้เรียนคนอื่นที่มีพฤติกรรมหรือความชอบคล้ายคลึงกันมาประกอบในคำแนะนำ ทำให้มีการนำเสนอในรายวิชาที่ผู้เรียนอาจมีความสนใจ โดยที่เทคนิคนี้มีขั้นตอนในการเตรียมข้อมูลไม่มากเนื่องจากต้องการเพียงแค่ 3 ข้อมูลคือ รหัสผู้ใช้งาน, รหัสรายวิชา และการให้คะแนนในรายวิชา แต่ข้อเสียของวิธีนี้เป็นปัญหาในเรื่องของขนาดข้อมูล ทำให้เกิดปัญหาความเบาบางของข้อมูล (Data Sparsity) เนื่องจากผู้เรียนไม่ได้เรียนทุกวิชา ข้อมูลประวัติการเรียนที่มีอยู่อาจน้อยเกินไป และปัญหาจากการที่ผู้เรียนหรือมีรายวิชาใหม่เกิดขึ้น (Cold Start) ทำให้ไม่สามารถแนะนำรายการได้ เนื่องจากไม่มีประวัติการใช้งานหรือไม่ผู้เรียนให้คะแนนในรายวิชานั้นมาก่อน ซึ่งกรณีของผู้เรียนใหม่ที่ยังไม่มีประวัติการใช้งานสามารถทำการแนะนำรายวิชาให้กับผู้เรียนโดยใช้เทคนิคการกรองข้อมูลตามเนื้อหา ซึ่งทำการค้นหารายวิชาที่ผู้เรียนมีความสนใจ สำหรับกรณีรายวิชาใหม่ เป็นความจริงที่ว่ารายวิชานั้นอาจไม่ถูกนำมาทำการแนะนำเนื่องจากไม่มีผู้เรียนให้คะแนนมาก่อน แต่สามารถถูกค้นพบได้จากขั้นตอนของเทคนิคการกรองข้อมูลตามเนื้อหาและเทคนิคการกรองข้อมูลแบบผสม หากรายวิชานั้นมีรายละเอียดข้อมูลหรือคุณลักษณะข้อมูลที่เหมาะสมและครบถ้วน เพื่อนำไปใช้ในเทคนิคดังกล่าวได้ ตัวอย่างเช่นรายวิชาที่ชื่อว่า Introduction to Project Management อยู่ในรายวิชาหลัก Business เป็นวิชาที่ไม่ผู้ทำการให้คะแนน แต่ถูกแนะนำรายวิชาในเทคนิคการกรองข้อมูลแบบผสม ตามตัวอย่างภาพประกอบ 47

จากผลลัพธ์ของทั้งสองเทคนิค ผู้วิจัยจึงนำเทคนิคดังกล่าวมาทำงานร่วมกัน เรียกว่าเทคนิคการกรองข้อมูลแบบผสม ส่งผลให้ระบบแนะนำแบบผสมนี้สามารถนำเสนอรายวิชา ทั้งวิชาหลักหรือวิชารอง และสามารถทำการแนะนำรายวิชาในหมวดหมู่อื่นได้ ซึ่งมีประสิทธิภาพดีกว่าการใช้เทคนิคเดียว รวมถึงยังช่วยลดปัญหาของรายวิชาที่ยังไม่มีการให้คะแนนจากผู้เรียนได้ หากไม่มีการใช้ระบบการแนะนำจากกระบวนการดังกล่าวในข้างต้น การแนะนำแบบวิธีดั้งเดิมคงเป็นการจัดอันดับรายวิชาโดยใช้ค่าเฉลี่ยจากการให้คะแนนของผู้เรียนเท่านั้น รายวิชาใดที่ได้รับการความสนใจจากผู้เรียนมากและมีการให้คะแนนของผู้เรียนสูง มักมีค่าเฉลี่ยมาก ทำให้อยู่ในลำดับต้นๆ ของการแนะนำรายวิชา ตัวอย่างเช่น รายวิชาที่ชื่อว่า Python Data Structures หรือ Neural Networks and Deep Learning ทั้งสองวิชานี้ถูกค้นพบในกระบวนการสำรวจข้อมูล เป็นรายวิชาที่

มีจำนวนการให้คะแนนจากผู้เรียนมากที่สุด โดยรายวิชาที่พบในกระบวนการนี้ส่วนใหญ่เป็นรายวิชาในหมวดหมู่ Data Science และ Computer Science และอาจเป็นรายวิชาในหมวดหมู่เดิมซึ่งไม่มีความแตกต่างมากนัก สำหรับรายวิชาใดที่ไม่มีผู้เข้าเรียนหรือได้รับความสนใจจากผู้เรียนน้อย อาจถูกมองข้ามและไม่ได้รับการแนะนำในรายวิชานั้นเลย

นอกจากนี้ข้อมูลรายวิชาที่ถูกนำมาแนะนำโดยกระบวนการข้างต้น มีรูปแบบข้อมูลที่เป็นที่น่าสังเกตคือ หากผู้เรียนมีประวัติการเรียนรู้ในรายวิชา ที่มาจากวิชาหลักหรือวิชารองซึ่งได้รับความนิยมน้อย มักมีรูปแบบเริ่มต้นของการแนะนำรายวิชาในเทคนิคการกรองข้อมูลผู้เข้าร่วมที่แตกต่างจากผู้เรียนคนอื่น และทำให้การแนะนำรายวิชาในเทคนิคการกรองข้อมูลแบบผสมมีความแตกต่างจากผู้เรียนคนอื่น ซึ่งถือเป็นข้อดีที่ทำให้ค้นพบรายวิชาใหม่ในระบบ แต่เป็นความยากที่ต้องทำให้รายวิชาที่ได้รับความสนใจน้อยได้รับการให้คะแนนจากผู้เรียนเช่นเดียวกัน นอกจากนี้หากผู้เรียนส่วนใหญ่มีการให้คะแนนในรายวิชาต่างๆ เท่ากับ 5 คะแนน ทำให้เกิดปัญหาและส่งผลต่อแบบจำลองสำหรับการแนะนำในรายวิชาต่างๆ เนื่องจากทางผู้วิจัยมีการกำหนดจำนวนการแนะนำรายวิชาเท่ากับ 5 รายวิชา จึงทำให้ส่งผลในการแนะนำรายวิชาจากระบบและมีการแนะนำที่เหมือนกับผู้เรียนคนอื่นค่อนข้างเห็นได้ชัด ดังนั้นการแก้ปัญหาดังกล่าวควรทำให้การให้คะแนนจากผู้เรียนมีความแตกต่างกันพอสมควร โดยเป็นการปรับช่วงคะแนนของข้อมูลให้เป็นระดับ 1-10 คะแนนและปรับคะแนนให้มีความเหมาะสมโดยใช้เทคนิค Normalization เพื่อทำให้ผลลัพธ์ของการแนะนำมีความแตกต่างมากยิ่งขึ้น จากนั้นอาจเพิ่มจำนวนของการแนะนำรายวิชาจาก 5 ให้เป็น 10 รายวิชา ก็สามารถทำให้เห็นความต่างของการแนะนำรายวิชาได้

โดยสรุปแล้วงานวิจัยนี้มีความคล้ายกับบทความวิจัย Hybrid Recommender System for Learning Material Using Content-Based Filtering and Collaborative Filtering with Good Learners' Rating (Turnip, Nurjanah, & Kusumo, 2017) สำหรับการแนะนำรายวิชา โดยใช้เทคนิคกรองข้อมูลตามเนื้อหาและเทคนิคการกรองข้อมูลแบบพึ่งพาผู้เข้าร่วม มาทำงานร่วมกัน เพื่อทำการแนะนำรายวิชาให้กับผู้เรียน โดยอ้างอิงมาจากผู้เรียนคนอื่นที่มีรสนิยมความชอบที่คล้ายกัน นอกจากนี้หากมีการปรับเปลี่ยนขอบเขตข้อมูลโดยเป็นการใช้ข้อมูลในส่วน of ระดับรายวิชา (level) หรือประเภทรายวิชา (course_type) อาจมีความคล้ายกับบทความ A Hybrid Recommender System for E-learning based on Context Awareness and Sequential Pattern Mining (Tarus, Niu, & Kalui, 2017) ในการแนะนำรายวิชาที่อ้างอิงมาจากระบบของผู้เรียน และเป็นการแนะนำในรายวิชาที่อาจถูกปรับเปลี่ยนให้มีความชอบส่วนตัวตาม

บริบทของผู้เรียน เพื่อทำให้เกิดการแนะนำรายวิชาที่หลากหลายและนำคุณลักษณะของข้อมูลมาใช้ให้เกิดประโยชน์ในการแนะนำมากยิ่งขึ้น

ข้อเสนอแนะ

1. ข้อมูลในบางส่วนที่นำมาทดสอบมีข้อมูลไม่ครบถ้วนและคุณลักษณะของค่ายังไม่ชัดเจน เช่น ทักษะที่ใช้ในรายวิชา จึงทำให้ข้อมูลมีคุณภาพไม่สมบูรณ์ในกระบวนการทำการสกัดคุณลักษณะที่สำคัญ กระบวนการใช้เทคนิคการกรองข้อมูลตามเนื้อหาจึงอาจทำได้ไม่ดีนัก

2. ระดับคะแนนที่ผู้เรียนทำการให้ในแต่ละรายวิชามีค่าระดับคะแนนเพียง 1-5 คะแนน โดยมีข้อมูลรายวิชาที่ได้ทำการให้คะแนนรายวิชาเพียง 604 รายวิชา จากจำนวนรายวิชาทั้งหมด 5,599 รายวิชา การนำไปคำนวณในเทคนิคการกรองผู้ใช้ร่วม อาจทำให้พบรายวิชาที่มีความเหมือนกับผู้เรียนรายอื่น เนื่องจากมีค่าการทำนายที่แตกต่างกันเพียงเล็กน้อย

3. สามารถประยุกต์ใช้งานวิจัยนี้ไปพัฒนาแบบจำลอง โดยทำการเพิ่มข้อมูล ลงรายละเอียดของรายวิชาให้มีความถูกต้อง และเปลี่ยนขอบเขตคุณลักษณะของข้อมูลเพื่อทำการสกัดคุณลักษณะของค่าที่สำคัญให้มีความชัดเจน สามารถทำการเพิ่มประสิทธิภาพของแบบจำลองให้สามารถทำการแนะนำได้ดีและหลากหลายมากยิ่งขึ้น ตัวอย่างเช่น ข้อมูลประเภทรายวิชา (course_type) และระดับรายวิชา (level) ให้ระบบสามารถทำการแนะนำรายวิชาตามบริบทปัจจุบันของผู้เรียน และนำเสนอรายวิชาในระดับที่มากกว่าผู้เรียนเคยเรียนมาก่อน เพื่อเป็นแนวทางให้กับผู้เรียนไปสู่จุดหมายของการเรียนในรายวิชานั้นๆ ได้

4. การมีข้อมูลผู้เรียนในขอบเขตข้อมูลอื่น เช่น อายุ ความสนใจในสาขาวิชา อาจนำข้อมูลเหล่านี้มาจัดกลุ่มผู้เรียนและสามารถทำการแนะนำรายวิชาที่ตรงกับความต้องการและความสนใจเฉพาะที่ตรงกับผู้เรียนได้

บรรณานุกรม

- Aggarwal, C. C. (2016). *Recommender Systems* (1): Springer.
- Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific Paper Recommendation: A Survey. *IEEE Access*, 7, 9324-9339.
- Banik, R. (2018). *Hands-On Recommendation Systems with Python: Start building powerful and personalized, recommendation engines with Python*: Packt Publishing.
- Bourkhouk, O., El Bachari, E., & El Adnani, M. (2016). A Recommender Model in E-learning Environment. *Arabian Journal for Science and Engineering*, 42(2), 607-617.
- Chen, W., Niu, Z., Zhao, X., & Li, Y. (2012). A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, 17(2), 271-284.
- Coursera.org. (2021). <https://www.coursera.org/>
- Dwivedi, S., & Roshni, V. S. K. (2017, 3-4 Aug. 2017). *Recommender system for big data in education*. Paper presented at the 2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH).
- Falk, K. (2019). *Practical Recommender Systems* (1st Edition): Manning.
- Iyer, S. (2020). Coursera Courses. Retrieved from <https://www.kaggle.com/sowmyaiyer/coursera-courses-and-details>
- Nakhaee, M. (2020). Course Reviews on Coursera. Retrieved from <https://www.kaggle.com/imuhammad/course-reviews-on-coursera/metadata>
- Patil, S. A. (2020). Coursera All Courses(5164 Courses) with language. Retrieved from <https://www.kaggle.com/santoshapatil31/coursera-courses-all5164-coursesin-all-languages/metadata>
- Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*: Springer Publishing Company, Incorporated.
- Tarus, J. K., Niu, Z., & Kalui, D. (2017). A hybrid recommender system for e-learning based on context awareness and sequential pattern mining. *Soft Computing*,

22(8), 2449-2461.

Turnip, R., Nurjanah, D., & Kusumo, D. S. (2017, 16-17 Nov. 2017). *Hybrid recommender system for learning material using content-based filtering and collaborative filtering with good learners' rating*. Paper presented at the 2017 IEEE Conference on e-Learning, e-Management and e-Services (IC3e).

Xiao, J., Wang, M., Jiang, B., & Li, J. (2017). A personalized recommendation system with combinational algorithm for online learning. *Journal of Ambient Intelligence and Humanized Computing*, 9(3), 667-677.

เปรมชัยสวัสดิ์, น. (2020). A MULTI-CRITERIA RECOMMENDATION SYSTEM BASED ON HYBRID PROFILE. *Suthiparithat Journal*, 26(80), 111-128.



ประวัติผู้เขียน

ชื่อ-สกุล	ปรีชา รัตนเดชอมร
วัน เดือน ปี เกิด	16 มีนาคม 2531
สถานที่เกิด	นครสวรรค์
วุฒิการศึกษา	พ.ศ. 2552 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยนเรศวร

