การเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้การเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่

นายณัฐนันท์ ฉันทานุรักษ์

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Concept Coverage Improvement of Video Recommender System in LMS by Using Random Walk and Category Diversity Function

Mr. Natthanun Chantanurak

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2016

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้การเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่
โดย	นายณัฐนันท์ ฉันทานุรักษ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

.....กรรมการภายนอกมหาวิทยาลัย
(ดร.ชัย วุฒิวิวัฒน์ชัย)

ณัฐนันท์ ฉันทานุรักษ์ : การเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้การเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่ (Concept Coverage Improvement of Video Recommender System in LMS by Using Random Walk and Category Diversity Function) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.อติวงศ์ สุชาโต, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร.โปรดปราน บุณยพุกกณะ, 83 หน้า.

องค์ความรู้นอกห้องเรียนจากอินเทอร์เน็ตที่มีจำนวนเพิ่มขึ้นอย่างมาก ส่งผลให้ผู้เรียนมีโอกาสในการเข้าถึงแหล่งความรู้เหล่านั้นได้อย่างกว้างขวางยิ่งขึ้น การพิจารณาเลือกแหล่งเรียนรู้ที่เป็นประโยชน์จึงมีบทบาทสำคัญ ผู้วิจัยจึงได้สร้างระบบแนะนำวีดิทัศน์ขึ้น เพื่อค้นหาและพิจารณาเลือกวีดิทัศน์ที่เป็นประโยชน์และมีความสัมพันธ์กับเอกสารสำหรับวิชาเรียนในระบบจัดการเรียนการสอนเสนอแก่ผู้เรียน งานวิจัยนี้ได้ใช้ความครอบคลุมแนวคิดในเอกสารเพื่อพิจารณาอรรถประโยชน์ของวีดิทัศน์ โดยอาศัยคุณสมบัติความหลากหลายของระบบแนะนำในการพัฒนาเพิ่มความครอบคลุมดังกล่าว มีการประยุกต์ใช้แนวคิดของการเรียงลำดับผลการแนะนำใหม่เพื่อเพิ่มความหลากหลายในหลายงานวิจัย การเดินสุ่มแบบเริ่มต้นใหม่เป็นวิธีการหนึ่งจากแนวคิดนี้ที่เป็นที่นิยม ผู้วิจัยจึงได้เสนออัลกอริทึมใหม่ที่มีชื่อว่า วิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่ เพื่อพัฒนาความสามารถของการเดินสุ่มแบบเริ่มต้นใหม่โดยอาศัยฟังก์ชันวัตถุประสงค์ที่วัดความหลากหลายของหมวดหมู่ของผลการแนะนำ การวัดผลอัลกอริทึมใหม่นี้ ได้ใช้ตัวชี้วัดความหลากหลายที่มีชื่อว่า อัลฟาเอ็นดีซีจี (α-Normalized Discounted Cumulative Gain: α-nDCG) และการวัดเวลาที่ใช้ในการดำเนินการอัลกอริทึมนี้ เปรียบเทียบกับวิธีที่ชื่อว่า เอ็มเอ็มอาร์ (Maximal Marginal Relevance: MMR) ซึ่งเป็นวิธีที่ทันสมัยและถูกใช้เป็นบรรทัดฐานในหลายงานวิจัย จากผลการทดลองแสดงให้เห็นว่าอัลกอริทึมใหม่ให้ค่าอัลฟาเอ็นดีซีจีไม่แตกต่างจากวิธีเอ็มเอ็มอาร์ แต่เวลาที่ใช้น้อยกว่าเป็นอย่างมาก ทางผู้วิจัยจึงได้เปรียบเทียบค่าอัลฟาเอ็นดีซีด้วยการทดสอบความเท่ากัน (Equivalence Test) พบว่าวิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่มีประสิทธิภาพการเพิ่มความหลากหลายไม่แตกต่างจากวิธีเอ็มเอ็มอาร์ ที่ขอบเขตความแตกต่างของค่าเฉลี่ย -0.035 ถึง 0.035

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
ปีการศึกษา	2559	ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5870929721 : MAJOR COMPUTER SCIENCE
KEYWORDS: RECOMMENDER SYSTEM / DIVERSIFICATION / RANDOM WALK WITH RESTART / LEARNING MANAGEMENT SYSTEM / RE-RANKING / CATEGORY DIVERSITY FUNCTION

NATTHANUN CHANTANURAK: Concept Coverage Improvement of Video Recommender System in LMS by Using Random Walk and Category Diversity Function. ADVISOR: ASSOC. PROF. ATIWONG SUCHATO, Ph.D., CO-ADVISOR: ASST. PROF. PROADPRAN PUNYABUKKANA, Ph.D., 83 pp.

With the massive pool of knowledge on the internet, students should be able to learn from these external resources if they can identify the useful ones. In this work, we created Video Recommender System (VRS), that automatically search and select videos using video textual data from YouTube that are relevant to the materials posted on Learning Management System (LMS), a tool common to teachers and students. The usefulness in this study is defined as the coverage of material concepts, and the diversification in Recommender System (RS) is one attribute that can improve its utility. According to previous works, the re-ranking concept has been employed in various algorithms to diverse results; Random Walk with Restart (RWR) is one very successful among them. We propose the new approach, Random Walk with Restart and Category Diversity Function (RWR+CDF), to enhance the capability of RWR by including the objective function called Category Diversity Function (CDF). This new algorithm was evaluated using a diversity metric named α-Normalized Discounted Cumulative Gain (α-nDCG) and the time it takes to finish such a task comparing to the Maximal Marginal Relevance (MMR), the state of the art approach. The results demonstrate that our approach provides as good α-nDCG score as MMR in some conditions, whereas the time spending of the proposed one is less than MMR. Finally, we decided to compare ours to the one from the state of the art using an Equivalence testing, it reveals that our RWR+CDF is equal to MMR at the mean difference between -0.035 and 0.035.

Department:	Computer Engineering	Student's Signature .....
Field of Study:	Computer Science	Advisor's Signature .....
Academic Year:	2016	Co-Advisor's Signature .....

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ รองศาสตราจารย์ ดร. อติวงศ์ สุชาโต อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ผู้ช่วยศาสตราจารย์ ดร. โปรดปราน บุณยพุกกณะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ได้ให้โอกาสในการเข้ามาศึกษา รวมทั้งสละเวลาอันมีค่าเพื่อให้คำปรึกษา คำแนะนำที่เป็นประโยชน์ ชี้แนะแนวทางและตรวจสอบความถูกต้องของเนื้อหาวิทยานิพนธ์นี้ตลอดระยะเวลาการศึกษาของข้าพเจ้า ขอขอบคุณคุณสุภัชชา เหมรัตนธร ที่คอยให้ความช่วยเหลือและอำนวยความสะดวกในการทำวิทยานิพนธ์ของข้าพเจ้า ข้าพเจ้าขอขอบพระคุณ Mr. Levon Ter-Isahakyan หัวหน้าทีมของข้าพเจ้า และเพื่อนร่วมงานทุกคนในทีมคิวปิด ที่บริษัท อโกด้า เซอร์วิสเซส จำกัด ที่เข้าใจและอนุญาตให้ข้าพเจ้าสละเวลาในการทำงานบางส่วน เพื่อใช้ในการทำวิทยานิพนธ์ ขอขอบคุณ คุณพิชญ สุวกูล และคุณอริสรา เอกกุล ผู้มีส่วนในการพิสูจน์อักษรในส่วนของวิทยานิพนธ์และผลงานตีพิมพ์ทางวิชาการของข้าพเจ้า

ข้าพเจ้าขอขอบพระคุณทุกคนในครอบครัวของข้าพเจ้า โดยเฉพาะอย่างยิ่ง บิดามารดา ผู้เป็นแรงบันดาลใจและแรงผลักดันในการศึกษาระดับบัณฑิตศึกษาในครั้งนี้ ขอขอบพระคุณเพื่อนร่วมรุ่นวิทยาศาสตรมหาบัณฑิต หลักสูตรวิทยาศาสตร์คอมพิวเตอร์ ภาคนอกเวลาราชการทุกคน ที่คอยช่วยเหลือและให้กำลังใจแก่กันเสมอมา ซึ่งส่งผลให้การทำวิทยานิพนธ์นี้ของข้าพเจ้าประสบความสำเร็จได้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง 1

สารบัญภาพ 1

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการทำวิจัย 3

1.3 ขอบเขตการทำวิจัย 3

1.4 ขั้นตอนและวิธีการดำเนินการทำวิจัย 4

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย 5

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 6

2.1 ทฤษฎีที่เกี่ยวข้อง 6

2.1.1 ระบบจัดการเรียนการสอน 6

2.1.1.1 มายคอร์สวิลล์ 8

2.1.1.2 ข้อมูลที่นำมาใช้ 9

2.1.2 วีดิทัศน์จากเว็บไซต์ยูทูบ 12

2.1.2.1 ส่วนต่อประสานโปรแกรมประยุกต์ของยูทูบ 14

2.1.3 ระบบแนะนำแบบพิจารณาจากเนื้อหา 16

2.1.4 นิยามความครอบคลุมแนวคิด 18

2.2 งานวิจัยที่เกี่ยวข้อง 19

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

หน้า

2.2.1 งานวิจัยที่เกี่ยวข้องกับการจัดลำดับการแนะนำใหม่ 19
2.2.1.1 การเดินสุ่มแบบเริ่มต้นใหม่ 20
2.2.1.2 เอ็มเอ็มอาร์ 23
2.2.2 งานวิจัยที่เกี่ยวข้องกับฟังก์ชันความหลากหลายของหมวดหมู่ 24
2.2.3 งานวิจัยที่เกี่ยวข้องกับการประเมินผลการแนะนำ 25
2.2.3.1 เซเรนดิพิตี้ 25
2.2.3.2 อัลฟาเอ็นดีซีจี 26
บทที่ 3 แนวคิดในการดำเนินวิจัย 30
3.1 การสร้างระบบแนะนำวีดิทัศน์ 32
3.1.1 ส่วนต่อประสานโปรแกรมประยุกต์ 32
3.1.1.1 ส่วนรับข้อมูลเอกสารจากระบบจัดการเรียนการสอน 32
3.1.1.2 ส่วนสร้างข้อมูลรายการแนะนำวีดิทัศน์ 33
3.1.3 ส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ 34
3.1.3.1 ขั้นตอนการเตรียมข้อมูลและสกัดแนวคิด 34
3.1.3.2 ขั้นตอนการค้นหาวีดิทัศน์ 36
3.1.3.3 ขั้นตอนการสร้างรายการแนะนำวีดิทัศน์ 37
3.2 การเรียงลำดับผลลัพธ์ใหม่ 41
3.2.1 การสร้างกราฟที่มีน้ำหนักสำหรับการเดินสุ่ม 41
3.2.2 วิธีการเดินสุ่มและความหลากหลายของหมวดหมู่ 42
3.3 เครื่องมือที่ใช้ในงานวิจัย 48
3.3.1 ภาษาที่ใช้ในการพัฒนา 48
3.3.2 โครงสร้างพื้นฐานและเทคโนโลยีที่เกี่ยวข้อง 48
3.3.3 เครื่องมืออื่น ๆ 49

หน้า

บทที่ 4 การทดลอง ผลการทดลอง และการวิเคราะห์ผลการทดลอง 50
4.1 ผลการทดลองจากค่าเซเรนดิพิตี้ 50
4.2 กลุ่มเอกสารตัวอย่างที่ใช้ในการทดลองอัลกอริทึมเพิ่มความครอบคลุมแนวคิด 56
4.3 ผลการทดลองจากอัลฟาเอ็นดีซีจี 59
4.3.1 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า c ที่ K แตกต่างกัน 59
4.3.2 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า c เฉลี่ยสำหรับทุก K แตกต่างกัน 63
4.3.3 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า K ณ จุดที่ใกล้ MMR ที่สุด 64
4.3.4 เปรียบเทียบผลการทดลองด้วยการทดสอบทีและทีโอเอสที 66
4.4 ผลการทดลองจากการวัดระยะเวลาดำเนินการ 68
4.4.1 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาในการดำเนินการเมื่อเปลี่ยนค่า c สำหรับทุก K 68
4.4.2 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาในการดำเนินการเมื่อเปลี่ยนค่า K 69
4.4.3 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาของ RWR+CDF_II ที่จำนวนวีดิทัศน์ทั้งหมด N จำนวน ที่แตกต่างกัน 70
บทที่ 5 บทสรุปการวิจัย และข้อเสนอแนะ 74
5.1 สรุปผลการวิจัย 74
5.2 ข้อจำกัด 74
5.2.1 ข้อจำกัดด้านเอกสาร 74
5.2.2 ข้อจำกัดของระบบแนะนำวีดิทัศน์ 75
5.2.3 ข้อจำกัดของอัลกอริทึม 75

หน้า

5.3 ข้อเสนอแนะ 75
5.3.1 ส่วนของการปรับปรุงระบบแนะนำวีดิทัศน์ 75
5.3.2 ส่วนของการปรับปรุงอัลกอริทึม 76
5.3.3 ส่วนของการประเมินผล 76
รายการอ้างอิง 77
ประวัติผู้เขียนวิทยานิพนธ์ 83

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญตาราง

หน้า

ตารางที่ 2.1 ข้อมูลสำหรับคำร้องขอ (Request Data) 14
ตารางที่ 2.2 ข้อมูลสำหรับคำตอบสนอง (Response Data) 14
ตารางที่ 2.3 ข้อมูลสำหรับคำร้องขอ (Request Data) 15
ตารางที่ 2.4 ข้อมูลสำหรับคำตอบสนอง (Response Data) 15
ตารางที่ 2.5 รายละเอียดคำอธิบายของตัวแปรในสมการการเดินสุ่มแบบเริ่มต้นใหม่ 21
ตารางที่ 2.6 รายละเอียดคำอธิบายของตัวแปรในสมการฟังก์ชันความหลากหลายของหมวดหมู่ 24
ตารางที่ 2.7 ตัวอย่างเอกสารและหัวข้อที่เกี่ยวข้องของเอกสาร 27
ตารางที่ 3.1 ข้อมูลของเอกสารที่มายคอร์สวิลล์ต้องส่งให้ระบบแนะนำวีดิทัศน์ 32
ตารางที่ 3.2 ข้อมูลของวีดิทัศน์แต่ละรายการที่ระบบแนะนำวีดิทัศน์ส่งให้มายคอร์สวิลล์ 33
ตารางที่ 3.3 รายละเอียดตัวแปรที่เกี่ยวข้องในการสกัดลักษณะเด่นระหว่างเอกสารและวีดิทัศน์ 38
ตารางที่ 3.4 รายละเอียดตัวแปรที่ใช้ในการคำนวณสมการฟังก์ชันความหลากหลายของหมวดหมู่ 43
ตารางที่ 4.1 ค่าเซเรนดิพิตี้ของเอกสารในภาพรวม จัดประเภทตามผู้เข้าร่วมประเมิน 51
ตารางที่ 4.2 ค่าเฉลี่ยค่าเซเรนดิพิตี้ของเอกสาร จัดประเภทตามผู้เข้าร่วมประเมิน 51
ตารางที่ 4.3 ค่าเฉลี่ยค่าเซเรนดิพิตี้ของรายวิชา 10 วิชา 52
ตารางที่ 4.4 ตัวอย่างค่าเฉลี่ยคำตอบของการประเมินในแต่ละแบบจากรายวิชาที่สนใจ 53
ตารางที่ 4.5 ข้อมูลค่าเซเรนดิพิตี้ที่น่าสนใจ 54
ตารางที่ 4.6 ข้อมูลทางสถิติจากกาทำการทดสอบแซทที่ค่าอัลฟาเท่ากับ 0.05 54
ตารางที่ 4.7 ชื่อเอกสารที่ใช้ในการทดลอง 56
ตารางที่ 4.8 ตารางแสดงค่า P-Value ของการทดสอบที ในแต่ละค่า K 66
ตารางที่ 4.9 ตารางแสดงค่า P-Value ของการทดสอบการเท่ากันด้วยวิธีทีโอเอสทีที่ขอบเขตความแตกต่างของค่าเฉลี่ย -0.035 ถึง 0.035 67

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญภาพ

หน้า

ภาพที่ 2.1 สัดส่วนของการใช้งานระบบจัดการเรียนการสอนในวงการต่าง ๆ [1] 7
ภาพที่ 2.2 ตัวอย่างหน้าจอสำหรับการแนะนำวีดิทัศน์..... 9
ภาพที่ 2.3 ตัวอย่างรายชื่อของเอกสารสำหรับการดาวน์โหลด..... 10
ภาพที่ 2.4 ตัวอย่างรายชื่อวิชาเรียน 10
ภาพที่ 2.5 หน้าจอสำหรับดาวน์โหลดเอกสาร 11
ภาพที่ 2.6 ตัวอย่างเอกสารที่ใช้ในการเรียนการสอน 11
ภาพที่ 2.7 แผนภูมิแท่งแสดงจำนวนชั่วโมงการอัปโหลดวีดิทัศน์ยูทูบใน 1 นาที [4]..... 12
ภาพที่ 2.8 ข้อมูลการดูวีดิทัศน์ในยูทูบของเพศชายแบ่งตามหมวดหมู่ของวีดิทัศน์ [20]..... 13
ภาพที่ 2.9 ข้อมูลการดูวีดิทัศน์ในยูทูบของเพศหญิงแบ่งตามหมวดหมู่ของวีดิทัศน์ [21]..... 13
ภาพที่ 2.10 ตัวอย่างเมทริกซ์ที่เกี่ยวข้องกับการเดินสุ่ม 21
ภาพที่ 2.11 ตัวอย่างเมทริกซ์ผลลัพธ์จากการเดินสุ่ม..... 22
ภาพที่ 3.1 โครงสร้างระบบของส่วนการสร้างระบบแนะนำวีดิทัศน์..... 30
ภาพที่ 3.2 ขั้นตอนการสร้างรายการแนะนำวีดิทัศน์ด้วยการเดินสุ่มร่วมกับ ฟังก์ชันความหลากหลายของหมวดหมู่..... 31
ภาพที่ 3.3 แนวคิดของส่วนต่อประสานโปรแกรมประยุกต์ 32
ภาพที่ 3.4 โครงสร้างโดยรวมของส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ..... 34
ภาพที่ 3.5 การทำงานของขั้นตอนกระบวนการเตรียมข้อมูลและการทำความสะอาดข้อมูล 34
ภาพที่ 3.6 ตัวอย่างการทำความสะอาดข้อมูล 35
ภาพที่ 3.7 ตัวอย่างเอกสารในระบบจัดการเรียนการสอนมายคอร์สวิลล์ 36
ภาพที่ 3.8 การทำงานของกระบวนการสกัดแนวคิดและขั้นตอนการค้นหาวีดิทัศน์..... 36
ภาพที่ 3.9 การส่งคำสำคัญไปยังยูทูบเพื่อรวบรวมข้อมูลวีดิทัศน์ 37
ภาพที่ 3.10 โครงสร้างการสร้างรายการแนะนำวีดิทัศน์ 37

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ภาพที่ 3.11 เมทริกซ์เปลี่ยนสถานะของกราฟมีน้ำหนักที่สร้างจากวีดิทัศน์และเอกสาร 42
ภาพที่ 3.12 เมทริกซ์เปลี่ยนสถานะที่ปรับน้ำหนักบางโหนด..... 42
ภาพที่ 3.13 ตัวอย่างหน้าจอการใช้งานกูเกิ้ลคลาวน์แพลตฟอร์ม..... 48
ภาพที่ 4.1 ตัวอย่างหน้าจอการประเมินเซเรนดิพิตี้ที่แสดงคำแนะนำ..... 50
ภาพที่ 4.2 ตัวอย่างหน้าจอการประเมินเซเรนดิพิตี้ เมื่อทำการประเมิน..... 50
ภาพที่ 4.3 ตัวอย่างเอกสารในวิชา Computer Algorithm 57
ภาพที่ 4.4 ตัวอย่างเอกสารในวิชา Discrete Structure..... 57
ภาพที่ 4.5 ตัวอย่างเอกสารในวิชา Programming Methodology 58
ภาพที่ 4.6 ตัวอย่างเอกสารในวิชา Research Methods in Computer Science 58
ภาพที่ 4.7 ตัวอย่างเอกสารในวิชา User Interface Design..... 58
ภาพที่ 4.8 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 5 61
ภาพที่ 4.9 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 10..... 61
ภาพที่ 4.10 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 15..... 62
ภาพที่ 4.11 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 20..... 62
ภาพที่ 4.12 กราฟแสดงค่าอัลฟาเอ็นดีซีจีเฉลี่ยทุก K ของ c ที่แตกต่างกัน..... 63
ภาพที่ 4.13 กราฟแสดงค่าอัลฟาเอ็นดีซีจี ณ จุดที่ c = 0.2 64
ภาพที่ 4.14 กราฟแสดงค่าอัลฟาเอ็นดีซีจี ณ จุดที่ c = 0.5 65
ภาพที่ 4.15 กราฟเปรียบเทียบระยะเวลาที่ใช้ ของ c ที่แตกต่างกัน..... 69
ภาพที่ 4.16 กราฟเปรียบเทียบระยะเวลาที่ใช้ ของ K ที่ต่างกัน..... 70
ภาพที่ 4.17 กราฟแสดงการเปลี่ยนแปลงระยะเวลาที่ใช้ของ RWR+CDF_II ที่จำนวน N ที่แตกต่างกัน..... 71
ภาพที่ 4.18 กราฟแสดงการเปลี่ยนแปลงของเวลาของ RWR+CDF_II เมื่อเปลี่ยน N 72

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ระบบจัดการเรียนการสอน (Learning Management System: LMS) เป็นซอฟต์แวร์ทางการศึกษาที่ช่วยอำนวยความสะดวกด้านการเรียนการสอนแก่ผู้เรียนและผู้สอนทั้งในระดับโรงเรียนและมหาวิทยาลัย ระบบจัดการเรียนการสอนที่ผ่านมามักถูกใช้สำหรับการจัดการเอกสารในชั้นเรียนและการสอบวัดผลการเรียนเสียเป็นส่วนใหญ่ [1] อย่างไรก็ตาม ครูผู้สอนหลายท่านก็ยังสามารถประยุกต์การใช้งานระบบจัดการเรียนการสอนเพื่อเพิ่มปฏิสัมพันธ์ระหว่างผู้เรียนกับผู้สอนให้มีมากขึ้น โดยอาศัยความสามารถของระบบที่ประกอบไปด้วยคุณสมบัติการใช้งานที่หลากหลาย [2] เช่น กระดานแลกเปลี่ยนความคิดเห็น การประเมินผล รวมไปถึงการประยุกต์ใช้เครือข่ายสังคม (Social network) เช่น เฟซบุ๊ก ในการเรียนการสอน เป็นต้น

ในการเรียนการสอนโดยทั่วไปนั้น ผู้สอนมักเป็นผู้ที่คัดเลือกและจัดเตรียมเนื้อหา เอกสาร หรือสื่อเพื่อประกอบการเรียนการสอนให้แก่ผู้เรียน แต่ด้วยเทคโนโลยีที่พัฒนาอย่างมากในช่วงเวลาที่ผ่านมา ทำให้ผู้เรียนมีช่องทางในการหาความรู้เพิ่มเติมที่มากขึ้น โดยเฉพาะอย่างยิ่ง การเข้ามามีบทบาทสำคัญของอินเทอร์เน็ต ได้เปิดโอกาสให้ผู้เรียนได้เข้าถึงแหล่งความรู้ที่หลากหลายเป็นจำนวนมาก ตัวอย่างเช่น เว็บไซต์การเรียนการสอนออนไลน์ขนาดใหญ่ (Massive Online Open Courses: MOOCs) ที่มีชื่อเสียงและเป็นที่นิยมอย่าง Coursera หรือ Udacity [3] หรือแม้แต่เครือข่ายวีดิทัศน์ออนไลน์อย่างเว็บไซต์ยูทูบ (Youtube) ก็เต็มไปด้วยวีดิทัศน์ที่เกี่ยวข้องกับองค์ความรู้ซึ่งสามารถประยุกต์ใช้เพื่อเสริมเนื้อหาจากในวิชาเรียนได้ สิ่งที่เป็นส่วนสำคัญของการนำแหล่งความรู้ภายนอกมาใช้สำหรับประกอบการเรียนนั้น คือการคัดกรองสิ่งที่เป็นประโยชน์ต่อผู้เรียน มีความเกี่ยวข้องกับเนื้อหาในวิชาเรียน และมีจำนวนเพียงพอที่จะครอบคลุมเนื้อหาของวิชาหรือบทเรียนนั้น ๆ การที่ระบบจัดการเรียนการสอนสามารถดำเนินการคัดเลือกเนื้อหาจากภายนอกที่เหมาะสมได้อย่างอัตโนมัติ จะช่วยเพิ่มช่องทางการเรียนรู้สำหรับผู้เรียนและอำนวยความสะดวกในการเรียนการสอนได้เป็นอย่างมาก

จากเหตุผลดังกล่าว งานวิจัยนี้จึงพัฒนาระบบแนะนำวีดิทัศน์ (Video Recommender System: VRS) ขึ้นมา เนื่องจากวีดิทัศน์ประกอบการสอนเป็นหนึ่งในฟังก์ชันที่ผู้ใช้งานมีความต้องการเป็นอันดับต้น ๆ [1] โดยระบบจะค้นหาวีดิทัศน์ที่มีเนื้อหาเกี่ยวข้องกับเอกสารของบทเรียนในระบบจัดการเรียนการสอน และทำการแนะนำวีดิทัศน์เหล่านั้นให้แก่ผู้เรียน ซึ่งในงานวิจัยนี้จะเลือกใช้วีดิ

ทัศน์จากเว็บไซต์ยูทูบ เป็นแหล่งข้อมูลสำหรับการแนะนำ เนื่องจากวีดิทัศน์ในเว็บไซต์ยูทูบนั้นมีจำนวนมากและมีความหลากหลายเป็นอันดับต้น ๆ ข้อมูลทางสถิติล่าสุดของยูทูบแสดงให้เห็นว่า ในเวลา 1 นาทีจะมีวีดิทัศน์ที่อัปโหลดเข้าสู่ระบบ เป็นจำนวนเฉลี่ยประมาณ 400 ชั่วโมง [4] ด้วยจำนวนวีดิทัศน์ที่มากและมีความหลากหลายดังกล่าว ทำให้การคัดกรองวีดิทัศน์เพื่อให้ครอบคลุมแนวคิด (Concept Coverage) ของเอกสารในบทเรียน กลายเป็นความท้าทายของงานวิจัย ซึ่งความครอบคลุมแนวคิดในงานวิจัยนี้ จะใช้คำสำคัญที่สกัดได้จากเนื้อหาภายในเอกสาร เป็นตัวแทนแนวคิดของเอกสารดังกล่าว และทำการค้นหาวีดิทัศน์ที่เกี่ยวข้องกับเอกสาร ที่มีเนื้อหาของวีดิทัศน์ที่สัมพันธ์กับแนวคิดเหล่านั้น ให้หลากหลายเพียงพอที่จะครอบคลุมทุกแนวคิด และมีจำนวนที่เหมาะสมกับความสำคัญของแต่ละแนวคิดดังกล่าว

ในการพัฒนาระบบดังกล่าว ในเบื้องต้นผู้วิจัยได้ใช้เทคนิคของระบบแนะนำแบบพิจารณาเนื้อหา (Content-based Recommender System) [5] โดยอาศัยลักษณะเด่น (Feature) ที่เป็นข้อความจากเนื้อหาเอกสารในบทเรียนกับชื่อและคำอธิบายของวีดิทัศน์จากยูทูบ และได้ทำการประเมินผลเบื้องต้น โดยใช้ค่าเซเรนดิพิตี้ (Serendipity) [6] ที่เป็นการวัดผลลัพธ์การแนะนำที่ไม่คาดว่าจะพบในรายการแนะนำ แต่ยังคงมีความเกี่ยวข้องกับเนื้อหาของบทเรียน ซึ่งจากการทดลองระหว่างการใช้ระบบแนะนำวีดิทัศน์เปรียบเทียบกับการค้นหาด้วยชื่อของบทเรียนเพียงอย่างเดียว พบว่าวิธีของงานวิจัยนี้ได้ให้ผลลัพธ์ค่าเซเรนดิพิตี้ที่ดีกว่า ถึงกระนั้น ค่าเซเรนดิพิตี้ก็ยังไม่สามารถใช้ชี้วัดความครอบคลุมแนวคิดทั้งหมดได้ เนื่องจากในบางส่วนของรายการแนะนำที่ให้ค่าเซเรนดิพิตี้สูงกลับสัมพันธ์กับแนวคิดของบทเรียนเพียงบางส่วนเท่านั้น

งานวิจัยนี้จึงนำเสนอวิธีการเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ โดยการประยุกต์ใช้การเดินสุ่มร่วมกับฟังก์ชันความหลากหลายของหมวดหมู่ เนื่องจากแนวคิดการพัฒนาระบบแนะนำแบบเดิมนั้น มักเน้นที่การแนะนำโดยอาศัยความคล้ายกันของวัตถุที่จะแนะนำ (Recommended Items) กับข้อมูลประวัติของผู้ใช้งาน (User Profile) ซึ่งวิธีการหาความคล้ายกันนั้นอาจใช้แนวคิดความคล้ายกันของเนื้อหาวัตถุแนะนำ (Content-based Approach) หรือความคล้ายกันของผู้ใช้งาน (Collaborative Filtering Approach) หรือแม้กระทั่งการใช้แนวคิดทั้งสองร่วมกัน [7] อย่างไรก็ตาม ปัญหาหนึ่งที่งานวิจัยด้านระบบแนะนำให้ความสนใจอยู่ในปัจจุบันนี้ คือการเพิ่มความหลากหลายของผลการแนะนำให้มากยิ่งขึ้น [8] ซึ่งการให้ความสนใจเฉพาะการแนะนำสิ่งที่คล้ายกัน (ในกรณีนี้คือวีดิทัศน์ที่มีความคล้ายกันกับเนื้อหาในเอกสารบทเรียนเท่านั้น) อาจให้ผลลัพธ์ที่เกี่ยวข้องกับแนวคิดบางส่วนของบทเรียน แต่ก็ไม่หลากหลายเพียงพอที่จะครอบคลุมแนวคิดทั้งหมดได้ การใช้แนวคิดสร้างกราฟความสัมพันธ์ของวีดิทัศน์กับเอกสารบทเรียน และใช้วิธีการเดินสุ่มแบบเริ่มต้นใหม่ (Random Walk with Restart) ในการจัดลำดับการแนะนำใหม่ (Re-ranking)

จึงเป็นแนวทางหนึ่งที่หลายงานวิจัยใช้ในการเพิ่มความหลากหลายของผลการสืบค้นเอกสาร (Document Retrieval) [9, 10] และในระบบแนะนำ (Recommender System) [11, 12] ถึงกระนั้น การเดินสุ่มก็ยังมีข้อเสียคือ ไม่มีเครื่องมือที่สามารถชี้วัดความหลากของผลลัพธ์การแนะนำที่สภาวะคงที่ (Steady State) ว่าให้ค่าความครอบคลุมแนวคิดที่ดีที่สุดหรือไม่ งานวิจัยนี้จึงเพิ่มส่วนของฟังก์ชันความหลากหลายของหมวดหมู่ (Category Diversity Function) [13] เข้ามา เพื่อใช้วัดผลการเดินสุ่มในแต่ละครั้งควบคู่ไปกับการตรวจสอบการเข้าสู่สภาวะคงที่ เพื่อให้ผลลัพธ์การแนะนำมีความครอบคลุมแนวคิดของเอกสารในบทเรียนมากยิ่งขึ้น

การวัดประสิทธิภาพการทำงานของวิธีที่นำเสนอนี้ จะแบ่งออกเป็น 2 ส่วน ส่วนแรกคือการวัดความครอบคลุมแนวคิดของผลการแนะนำโดยใช้เครื่องมือวัดความหลากหลายที่มีชื่อว่า อัลฟาเอ็นดีซีจี (α-Normalized Discounted Cumulative Gain: α-nDCG) [14] ซึ่งเป็นมาตรฐานที่ใช้วัดความหลากหลายของผลการแนะนำที่ใช้ในการประชุมทางวิชาการที่ชื่อว่า ทีอาร์อีซี (Text REtrieval Conference: TREC) ส่วนที่สองคือการวัดระยะเวลาที่ใช้ในการดำเนินการอัลกอริทึมใหม่นี้ โดยการวัดค่าทั้งสองส่วนนี้จะถูกนำมาเปรียบเทียบผลลัพธ์กับวิธีที่ให้ผลลัพธ์ที่ดีที่สุดในตอนนี้ [15] คือวิธีที่มีชื่อว่า เอ็มเอ็มอาร์ (Maximal Marginal Relevance: MMR) ซึ่งเป็นวิธีที่งานวิจัยหลาย ๆ งานใช้เป็นบรรทัดฐานในการเปรียบเทียบ โดยการดำเนินการทดลองเพื่อวัดผลดังกล่าวทำเพื่อพิสูจน์ว่าวิธีการที่นำเสนอให้ผลลัพธ์ได้ใกล้เคียงกับวิธีทีดีที่สุด

1.2 วัตถุประสงค์ของการทำวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเพิ่มความครอบคลุมแนวคิดที่เกี่ยวข้องในเอกสาร สำหรับการแนะนำวีดิทัศน์ให้แก่ผู้เรียน รวมไปถึงการศึกษาผลลัพธ์และการเปลี่ยนแปลงของค่าตัวชี้วัดความหลากหลายที่ชื่อว่า อัลฟาเอ็นดีซีจี โดยเปรียบเทียบกับวิธีที่ให้ผลดีที่สุด เพื่อพัฒนาคุณภาพของผลลัพธ์การแนะนำวีดิทัศน์สำหรับระบบจัดการเรียนการสอนให้มีประโยชน์และเกิดคุณค่าในด้านการเรียนการสอนต่อผู้ใช้งาน ทั้งในแง่ของความเกี่ยวข้องและความครอบคลุมแนวคิดของเอกสารให้ทั่วถึงในระดับที่เหมาะสม

1.3 ขอบเขตการทำวิจัย

งานวิจัยหัวข้อ การเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้วิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่ มีขอบเขตการ

ดำเนินงานใน 3 ส่วนหลัก คือขอบเขตการดำเนินงานในส่วนของการพัฒนาระบบแนะนำวีดิทัศน์ (ข้อ 1-3) ส่วนของแนวคิดการวิจัย (ข้อ 4) และส่วนของการประเมินผล (ข้อ 5 เป็นต้นไป)

1. ใช้ข้อมูลจากระบบจัดการเรียนการสอนมายคอร์สวิลล์ในการประเมินผล
2. ใช้เอกสารและข้อมูลส่วนที่เป็นตัวอักษรที่อยู่ภายในระบบจัดการเรียนการสอนดังกล่าว สำหรับการสกัดคุณลักษณะ เพื่อใช้ในการค้นหาและการแนะนำวีดิทัศน์
3. นำเสนอวีดิทัศน์จากเว็บไซต์ยูทูบและใช้คุณลักษณะที่เป็นตัวอักษรของวีดิทัศน์ ประกอบไปด้วยชื่อและคำอธิบายเท่านั้น โดยไม่พิจารณาเนื้อหาการนำเสนอภายในตัววีดิทัศน์
4. กราฟถ่วงน้ำหนักที่สร้างเพื่อทำการเดินสุ่มในการแนะนำวีดิทัศน์ จะใช้วีดิทัศน์ที่ค้นหามาจากหัวข้อที่สกัดมาจากเอกสาร
5. <u>ประชากร</u> ที่ใช้ในการประเมินผลการแนะนำในส่วนของเซเรนดิพิตี้ คือผู้ใช้งานระบบจัดการเรียนการสอนมายคอร์สวิลล์ ประกอบด้วยนิสิตและอาจารย์ที่เป็นเจ้าของวิชาเรียนหรือเจ้าของเอกสารที่อัปโหลด (ในกรณีที่มีอาจารย์หลายท่าน) ในส่วนของการประเมินผลความครอบคลุมแนวคิดด้วยค่าอัลฟาเอ็นดีซีจี ประชากรจะเป็นเอกสารรูปแบบพีดีเอฟในระบบจัดการเรียนการสอน มายคอร์สวิลล์
6. <u>กลุ่มตัวอย่างที่ทำการศึกษา</u> ในส่วนของเซเรนดิพิตี้คือนิสิตและอาจารย์จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ใช้ระบบจัดการเรียนการสอนดังกล่าว จำนวน 15 วิชา สำหรับส่วนของนิสิตนั้น จะใช้นิสิตที่มีผลการเรียนอยู่ในระดับดีในวิชานั้นในปีการศึกษาที่ผ่านมา วิชาละ 2 คน และนิสิตที่สุ่มมาจากนิสิตที่กำลังเรียนอยู่ในปีการศึกษาปัจจุบันวิชาละ 2 คน ในส่วนของการประเมินผลความครอบคลุมแนวคิดด้วยค่าอัลฟาเอ็นดีซีจี จะเป็นเอกสารจากวิชาเรียนจำนวน 5 วิชาเรียน จำนวนวิชาละ 5 เอกสาร ซึ่งเป็นวิชาของภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

1.4 ขั้นตอนและวิธีการดำเนินการทำวิจัย

การดำเนินการทำวิจัย สามารถแบ่งเป็นขั้นตอนได้ดังนี้

1. ศึกษางานวิจัยที่เกี่ยวข้องกับการสร้างระบบแนะนำเบื้องต้น และวิธีการที่เป็นไปได้สำหรับลักษณะของข้อมูลในมายคอร์สวิลล์
2. พัฒนาระบบแนะนำวีดิทัศน์จากยูทูบ
3. เก็บรวบรวมข้อมูลการประเมินผลการแนะนำจากกลุ่มตัวอย่าง เพื่อประเมินค่าเซเรนดิพิตี้

4. ศึกษางานวิจัยที่เกี่ยวข้องกับการจัดลำดับการแนะนำใหม่ การเพิ่มความครอบคลุมแนวคิด ความหลากหลายของผลลัพธ์ และวิธีการเดินสุ่มแบบเริ่มต้นใหม่
5. ปรับปรุงระบบแนะนำวีดิทัศน์เพื่อรองรับอัลกอริทึมการจัดลำดับการแนะนำใหม่โดยใช้การเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่
6. ทดสอบความถูกต้อง และความครอบคลุมแนวคิดของผลการแนะนำวีดิทัศน์ จากการประเมินผลโดยใช้ค่าอัลฟาเอ็นดีซีจี
7. วิเคราะห์ผลการทดลอง
8. สรุปและเรียบเรียงวิทยานิพนธ์
9. เผยแพร่ผลงานตีพิมพ์

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัยเรื่องนี้ คือคุณภาพที่เพิ่มขึ้นของวีดิทัศน์ที่ทำการแนะนำให้แก่ผู้เรียน ทั้งคุณภาพในด้านของความเกี่ยวข้องกันระหว่างเอกสารและวีดิทัศน์ และความหลากหลายที่เพิ่มขึ้นอย่างเพียงพอที่จะครอบคลุมเนื้อหาและแนวคิดโดยรวมของเอกสาร ซึ่งจะเป็นประโยชน์แก่ผู้ใช้งานระบบจัดการเรียนการสอนทั้งผู้เรียนและผู้สอน โดยเฉพาะผู้เรียนที่มีช่องทางการเรียนรู้ที่หลากหลายมากยิ่งขึ้น นอกเหนือไปจากเอกสารที่ทางครูผู้สอนจัดเตรียมไว้ และระบบแนะนำวีดิทัศน์นี้จะเป็นแนวทางสำหรับการพัฒนาระบบแนะนำอื่น ๆ ที่จะเพิ่มเข้ามาเพื่อเพิ่มศักยภาพให้แก่ระบบจัดการเรียนการสอนต่อไป

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

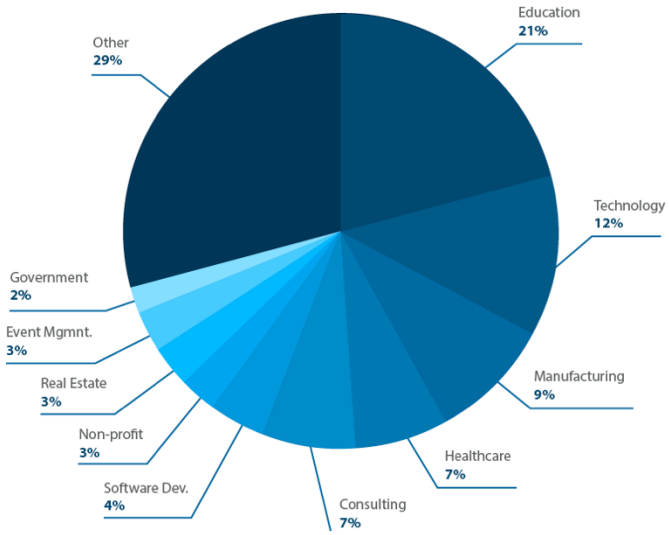
ในบทที่ 2 ของงานวิจัยเรื่อง “การเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้วิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่” ประกอบด้วยเนื้อหาสองส่วน ได้แก่ ส่วนการอธิบายทฤษฎีต่าง ๆ ที่ใช้ในงานวิจัยฉบับนี้ ซึ่งประกอบไปด้วย ข้อมูลเกี่ยวกับระบบจัดการเรียนการสอน, วีดิทัศน์จากเว็บไซต์ยูทูบ, ระบบแนะนำแบบพิจารณาจากเนื้อหา และนิยามความครอบคลุมแนวคิด และอีกส่วนหนึ่งคือส่วนงานวิจัยที่เกี่ยวข้อง ซึ่งแบ่งเป็น งานวิจัยที่เกี่ยวข้องกับการจัดลำดับการแนะนำใหม่, งานวิจัยที่งานวิจัยที่เกี่ยวข้องกับฟังก์ชันความหลากหลายของหมวดหมู่ และงานวิจัยที่เกี่ยวข้องกับการประเมินผลการแนะนำทฤษฎีที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ระบบจัดการเรียนการสอน

ระบบจัดการเรียนการสอน (Learning Management System: LMS) เป็นระบบสารสนเทศ (Information System) ที่ใช้สำหรับจัดการเนื้อหา (Learning Content) และกิจกรรมทางการเรียน (Learning Activity) [16] รวมไปถึงการบริหารจัดการ, ติดตาม, ประเมิน และรายงานผลการเรียนการสอน [17] อย่างไรก็ตาม ระบบจัดการเรียนการสอนของสโตนและเซงใน [16] กลายเป็นคำที่ใช้เรียกระบบทั่วไปที่สนับสนุนกระบวนการการเรียนการสอนทั้งทางตรงและทางอ้อม ซึ่งประกอบด้วย การบริหารจัดการการเรียน (Learning Management), การบริหารจัดการเนื้อหา (Content Management) และการบริหารจัดการรายวิชา (Course Management) นอกจากนี้ เนื่องจากระบบจัดการเรียนการสอนแบบดั้งเดิม มักเป็นระบบปิดที่ใช้ภายในองค์กรเท่านั้น ระบบจัดการเรียนการสอนสมัยใหม่จึงได้พยายามพัฒนาระบบให้เป็นระบบเปิดมากยิ่งขึ้น เพื่อเพิ่มความยืดหยุ่นในการใช้งาน เช่น การพัฒนาระบบในรูปแบบแอปพลิเคชันบนเว็บไซต์ หรือบนโทรศัพท์มือถือ รวมถึงการผนวกอรรถประโยชน์ของเครือข่ายสังคมออนไลน์ (Social Network) เข้ามา เพื่อให้เกิดปฏิสัมพันธ์ระหว่างผู้เรียนกับผู้สอน และระหว่างผู้เรียนด้วยกันเองมากยิ่งขึ้น และในขณะที่การเรียนการสอนแบบดั้งเดิม ทั้งที่ใช้ระบบจัดการเรียนการสอนหรือไม่ใช้ก็ตามนั้น ผู้สอนมักเป็นผู้กำหนดเนื้อหาบทเรียน และเอกสารสำหรับการเรียนเสียเป็นส่วนใหญ่ ระบบจัดการเรียนการสอนสมัยใหม่จึงพยายามเปิดโอกาสให้ผู้เรียนได้มีส่วนกำหนดแนวทางการเรียนของตนเอง และมีความเฉพาะต่อพฤติกรรมของผู้เรียนแต่ละคน (Personalization) มากขึ้น ยิ่งไปกว่านั้น ระบบจัดการเรียนการสอน

ต้องสามารถรองรับปริมาณและความหลากหลายที่เพิ่มขึ้นของข้อมูลในระบบ และนำข้อมูลเหล่านั้นมาใช้วิเคราะห์ เพื่อปรับปรุงและพัฒนารูปแบบของการเรียนการสอนได้ จากข้อมูลใน [1] จะพบว่า มีการใช้งานระบบจัดการเรียนการสอนในหลายวงการด้วยกัน ดังแสดงในภาพที่ 2.1 โดยเฉพาะวงการการศึกษา (ร้อยละ 21) มีมากถึง 1 ใน 5 ซึ่งเป็นตัวชี้วัดหนึ่ง ถึงการให้ความสำคัญของ



ภาพที่ 2.1 สัดส่วนของการใช้งานระบบจัดการเรียนการสอนในวงการต่าง ๆ [1]

สถาบันการศึกษาในการนำระบบจัดการเรียนการสอนมาใช้ ทั้งในระดับโรงเรียนและมหาวิทยาลัย อย่างไรก็ตาม จากงานวิจัยของ Adzharuddin [18] พบว่า มีอุปสรรคหลายประการ สำหรับการใช้งานระบบจัดการเรียนการสอนในมหาวิทยาลัย ประการแรกคือ การที่ผู้สอนใช้ประโยชน์จากระบบจัดการเรียนการสอนเพียงเพื่ออัปโหลดเอกสารสำหรับการเรียนการสอนเท่านั้น แต่ไม่ได้ใช้ความสามารถอื่น ๆ ของระบบ เช่น กระดานถามตอบ (Discussion Board) และการส่งข้อความส่วนตัว (Message) อย่างเต็มที่ อุปสรรคอีกประการหนึ่งคือ งบประมาณที่จำกัดในการจัดซื้อระบบจัดการเรียนการสอนที่มีอยู่ในท้องตลาด ก่อให้เกิดข้อจำกัดในการเลือกผลิตภัณฑ์ที่เหมาะสมกับลักษณะเฉพาะของแต่ละสถาบัน และอาจทำให้เกิดความไม่มั่นใจของทั้งผู้เรียนและผู้สอนในการใช้งานระบบ ในแง่ที่ว่าระบบอาจไม่มีคุณภาพเพียงพอ หรือระบบที่เลือกมามีส่วนต่อประสานกับผู้ใช้ (User interface) ที่ยากในการใช้งาน ซึ่งเป็นสาเหตุที่ทำให้ผู้เรียนและผู้สอนหลีกเลี่ยงที่จะใช้ระบบเหล่านี้ เพื่อแก้ปัญหาดังกล่าว ในหลายมหาวิทยาลัยจึงได้มีการริเริ่มพัฒนาระบบจัดการเรียนการสอนขึ้นเอง หนึ่งในนั้นคือระบบจัดการเรียนสอนของจุฬาลงกรณ์มหาวิทยาลัยที่มีชื่อว่า มายคอร์สวิลล์ (myCourseVille) และเป็นระบบจัดการเรียนการสอนหลักที่จะใช้ในงานวิจัยนี้

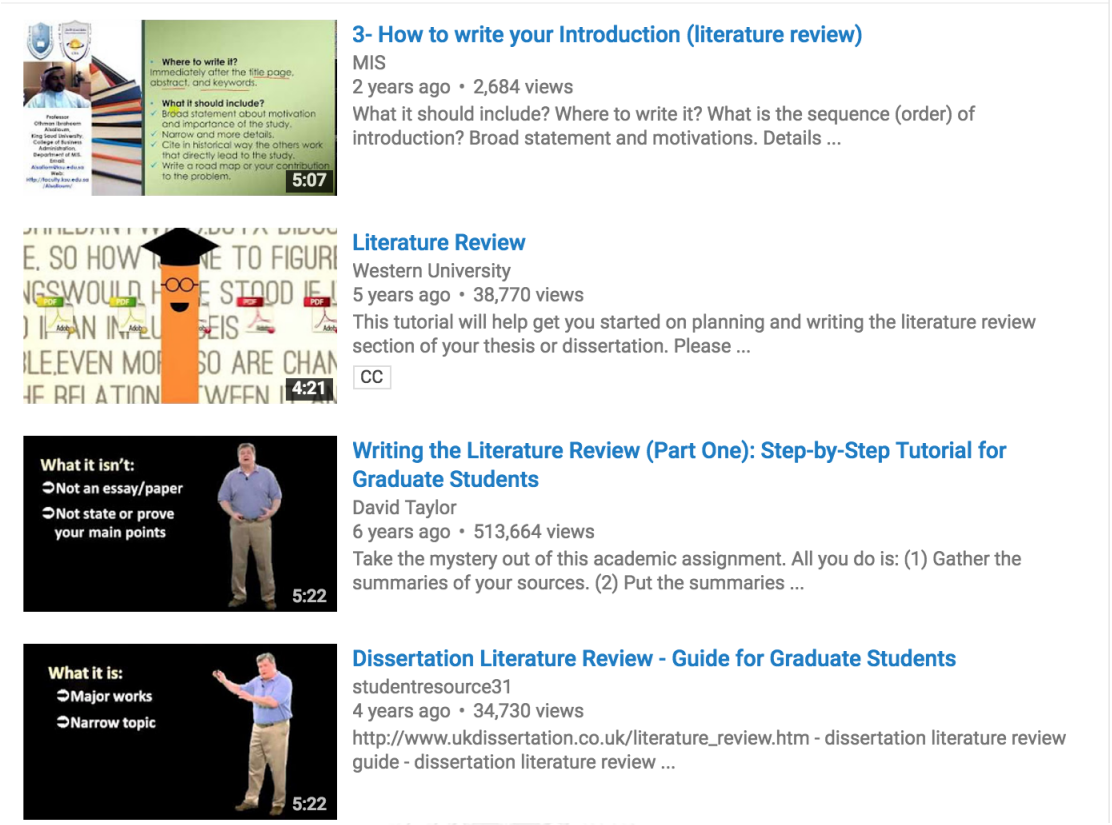
2.1.1.1 มายคอร์สวิลล์

ระบบจัดการเรียนการสอนมายคอร์สวิลล์เป็นระบบจัดการเรียนการสอนที่พัฒนาโดยคณาจารย์และบุคลากรของคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ร่วมกับศูนย์นวัตกรรมการเรียนรู้ แห่งจุฬาลงกรณ์มหาวิทยาลัย ในโครงการพัฒนาระบบสนับสนุนการเรียนการสอนที่เชื่อมต่อกับเครือข่ายสังคมออนไลน์ โดยให้บริการการใช้งานในรูปแบบบริการซอฟต์แวร์ SaaS (Software as a Service) [19] โดยผู้ใช้งาน ไม่จำเป็นต้องติดตั้งหรือ ดูแลระบบด้วยตนเอง การลงทะเบียนใช้งานจะอาศัยบัญชีของเฟซบุ๊ก ซึ่งเป็นเครือข่ายสังคมออนไลน์ที่ใหญ่ที่สุดและมีผู้ใช้งานทั่วโลก ผู้ใช้งานสามารถใช้งานผ่านบราวเซอร์ได้ทันทีจากทุกแพลตฟอร์ม การใช้งานมายคอร์สวิลล์เกิดขึ้นครั้งแรกในปีพุทธศักราช 2554 ภายในคณะวิศวกรรมศาสตร์เท่านั้น 5 ปีต่อมา การใช้งานมายคอร์สวิลล์ได้ขยายไปสู่หลากหลายคณะในมหาวิทยาลัย รวมไปถึงมหาวิทยาลัยอื่น ๆ ในประเทศไทยอีกด้วย

การพัฒนามายคอร์สวิลล์ตั้งอยู่บนแนวคิดการเรียนแบบมีส่วนร่วม (Active Learning) ทำให้ความสามารถของระบบส่วนใหญ่ เน้นไปที่การให้ผู้เรียนมีส่วนร่วมและมีปฏิสัมพันธ์ในชั้นเรียน ได้แก่

1. การจัดการและแจกจ่ายเอกสาร
2. การจัดแผนการสอน / การประกาศข่าวสาร
3. การจัดการการบ้านและงานที่ได้รับมอบหมาย
4. การดำเนินกิจกรรมแบบกลุ่ม
5. การจัดการคะแนนและการเข้าชั้นเรียน

จากข้อมูลเบื้องต้น จะพบว่ามายคอร์สวิลล์ค่อนข้างมีความสามารถที่หลากหลาย และเพียงพอต่อการใช้งานในการจัดการเรียนการสอนแล้ว แต่ในงานวิจัยฉบับนี้จะเพิ่มความสามารถใหม่ให้กับมายคอร์สวิลล์ ได้แก่ความสามารถในการแนะนำวีดิทัศน์ที่เป็นประโยชน์ต่อบทเรียนนั้น ๆ ให้แก่ผู้เรียน ตัวอย่างเช่น ในบทเรียนเรื่อง การเขียนบทนำและงานวิจัยที่เกี่ยวข้อง (Writing Introduction and Literature Reviews) ในวิชาวิธีการทำวิจัย (Research Method) จะได้รายการแนะนำวีดิทัศน์ที่มีเนื้อหาเกี่ยวข้องกับการเขียนบทนำและการเขียนงานวิจัยที่เกี่ยวข้อง ดังภาพที่ 2.



ภาพที่ 2.2 ตัวอย่างหน้าจอสำหรับการแนะนำวีดิทัศน์

2.1.1.2 ข้อมูลที่นำมาใช้

มายคอร์สวิลล์มีข้อมูลที่เกี่ยวข้องกับบทเรียนแต่ละบทเรียนเก็บไว้ในระบบ โดยเฉพาะส่วนของเอกสารบทเรียน ข้อมูล ณ ขณะที่ดำเนินงานวิจัยนี้ มายคอร์สวิลล์มีเอกสารในระบบทั้งหมด 29,697 ชิ้น จากจำนวนวิชาทั้งหมด 4,206 วิชา (นับวิชาเดียวกันในต่างภาคการศึกษา แต่ละวิชาแยกกัน) ในระยะเวลา 5 ปี โดยข้อมูลของเอกสารที่จะนำมาใช้ในงานวิจัยนี้ ได้แก่

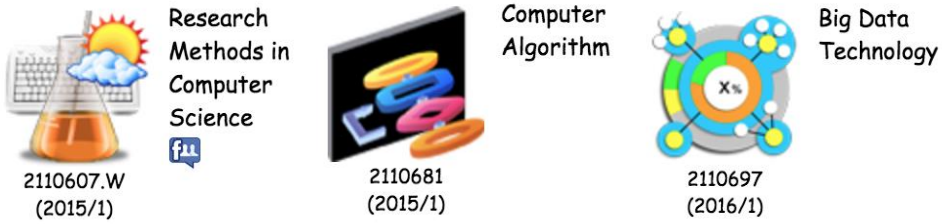
1. ชื่อของเอกสาร (Document Title)

ชื่อของเอกสารเป็นชื่อที่ผู้สอนกำหนดไว้ ณ ตอนที่ทำการอัปโหลดเอกสารเข้าสู่มายคอร์สวิลล์ ซึ่งอาจจะไม่ตรงกับชื่อไฟล์ที่แนบมาก็ได้ โดยจะใช้ชื่อที่ผู้สอนกำหนดเพียงอย่างเดียว ตัวอย่างรายการชื่อของเอกสาร แสดงในภาพที่ 2.3

ภาพที่ 2.3 ตัวอย่างรายชื่อของเอกสารสำหรับการดาวน์โหลด

2. ชื่อวิชาของเอกสาร (Document Course Title)

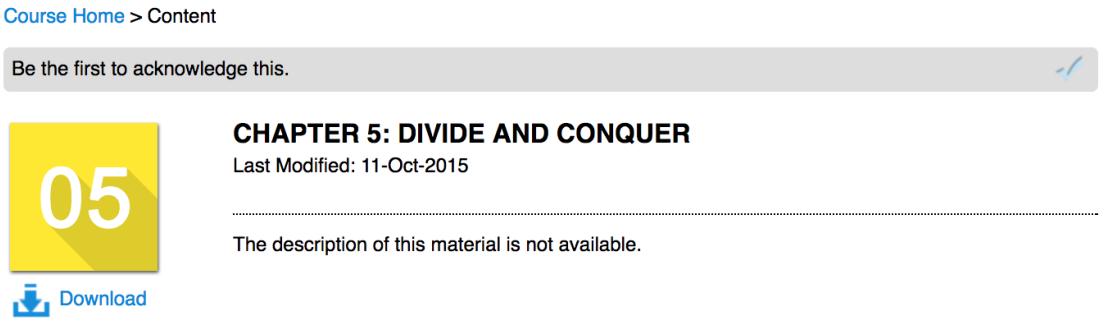
ชื่อวิชาของเอกสารจะใช้ชื่อที่ผู้สอนกำหนดไว้ตอนสร้างรายวิชาขึ้นในมายคอร์สวิลล์ ในกรณีที่มีการใส่รหัสและภาคการศึกษาเข้ามาเป็นส่วนหนึ่งของชื่อวิชา ก็จะนับว่าเป็นส่วนหนึ่งของชื่อวิชาด้วย ตัวอย่างชื่อวิชาแสดงในภาพที่ 2.4



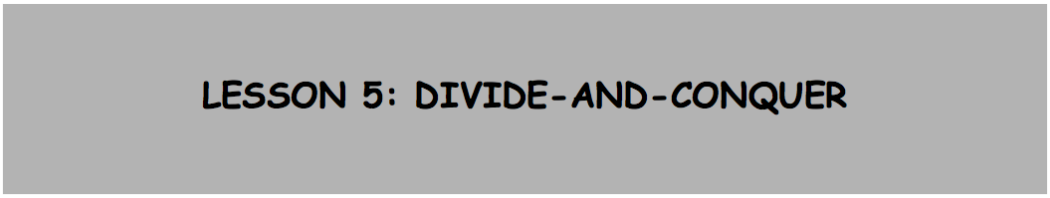
ภาพที่ 2.4 ตัวอย่างรายชื่อวิชาเรียน

3. เนื้อหาของเอกสาร (Document Content)

ข้อมูลส่วนนี้ถือเป็นส่วนที่มีความสำคัญมากที่สุด เพราะเนื้อหาของเอกสารจะถูกนำมาใช้ในการสกัดเอาคำสำคัญซึ่งจะใช้แทนแนวคิดของเอกสารดังกล่าว ซึ่งเนื้อหาในเอกสารนี้จะถูกเก็บอยู่ในรูปแบบของไฟล์นามสกุลต่าง ๆ แต่ที่นำมาใช้ในงานวิจัยนี้ จะเป็นส่วนของเอกสารนามสกุลพีดีเอฟ (Portable Document Format: PDF) นำมาแปลงเป็นเอกสารที่เป็นข้อความ (Text) ก่อนที่จะนำไปใช้ในขั้นตอนการสกัดลักษณะเด่นต่อไป ในภาพที่ 2.5 จะเป็นหน้าสำหรับการดาวน์โหลดเอกสาร ซึ่งในอนาคตจะนำรายการแนะนำวีดิทัศน์จากเว็บไซต์ยูทูบที่ได้จากงานวิจัยนี้มาแสดงในหน้านี้ด้วย ภาพที่ 2.6 จะเป็นตัวอย่างเอกสารที่จะใช้ในรูปแบบพีดีเอฟ ก่อนทำการแปลงเอกสาร



ภาพที่ 2.5 หน้าจอสำหรับดาวน์โหลดเอกสาร



Divide and conquer is a technique for designing of algorithms by divid a problem (large size) into many small problems that is easier to be solved. The whole solution can be obtained by combining all solutions of small problems.

1. INTRODUCTION

The principle behind the Divide-and-Conquer algorithm design paradigm is that it is (often) easier to solve several small instances of a problem than one large one. Algorithm in this sections use the Divide-and-Coquer approach. They divide the problem into smaller instances of the same problem (for example, in the case of sorting, divide into smaller sets to be sorted), then solve (conquer) the smaller instances recursively (i.e., by the same mothod), and finally combine the solutions to obtain the solution for the original input.

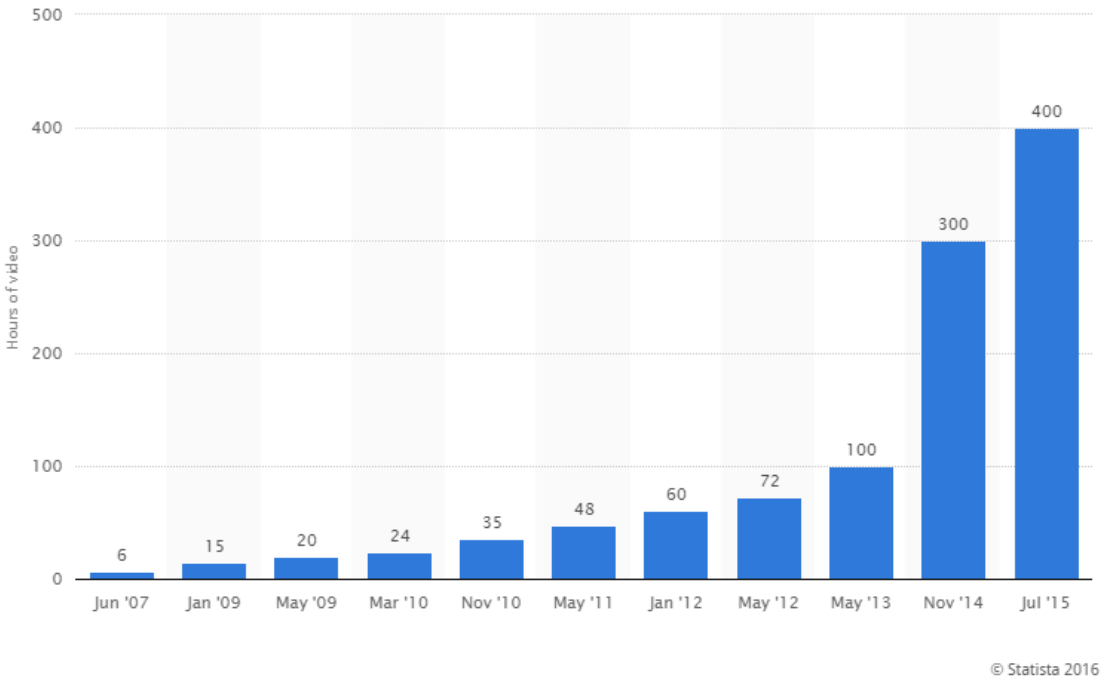
2. DIVIDE AND CONQUER

Definition: An algorithmic technique. To solve a problem on an instance of size n, a solution is found either directly because solving that instance is easy (typically, because the instance is small) or the instance is *divided* into two or more smaller

ภาพที่ 2.6 ตัวอย่างเอกสารที่ใช้ในการเรียนการสอน

2.1.2 วีดิทัศน์จากเว็บไซต์ยูทูบ

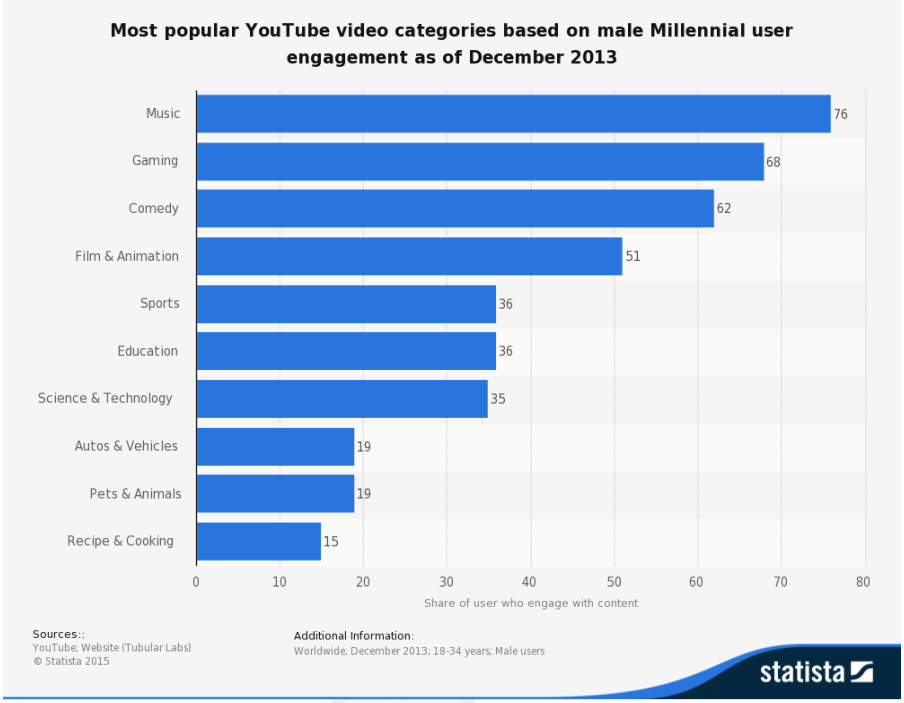
ยูทูบเป็นแพลตฟอร์มสำหรับการแบ่งปันวีดิทัศน์ซึ่งอนุญาตให้ผู้ใช้งานทำการอัปโหลดวีดิทัศน์ของตัวเองขึ้นไปในระบบได้ ยูทูบก่อตั้งในปีพุทธศักราช 2548 และถูกซื้อโดยกูเกิ้ล (Google) ในปีถัดมา จำนวนชั่วโมงของวีดิทัศน์ที่ถูกอัปโหลดใน 1 นาที จากภาพที่ 2.7 [4] เพิ่มจาก 6 ชั่วโมงในปีพุทธศักราช 2550 (คริสต์ศักราช 2007) เป็น 400 ชั่วโมงในปีพุทธศักราช 2558 (คริสต์ศักราช 2015) และมีแนวโน้มที่จะเพิ่มสูงขึ้นเรื่อย ๆ



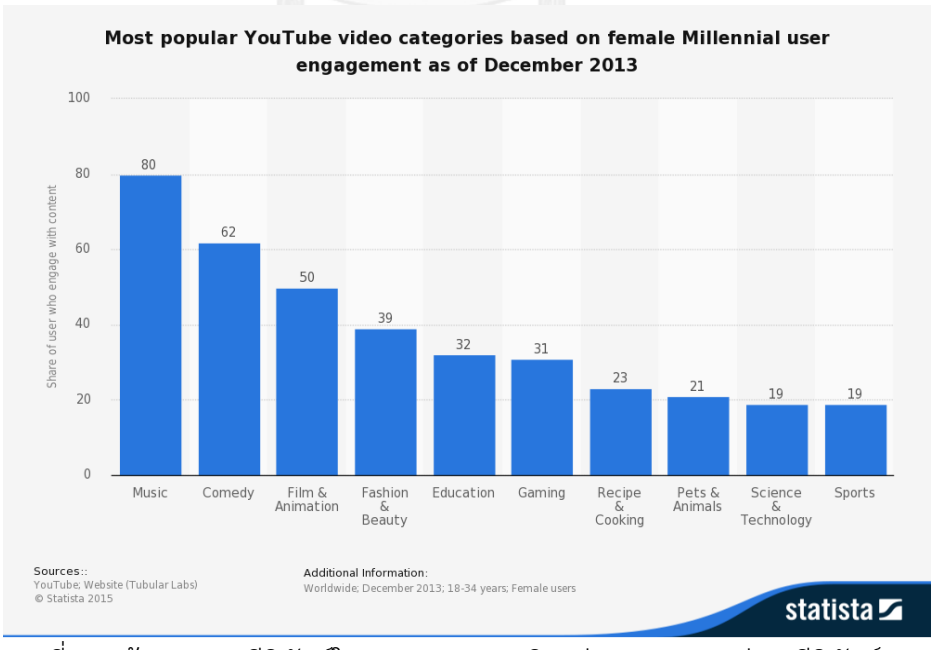
ภาพที่ 2.7 แผนภูมิแท่งแสดงจำนวนชั่วโมงการอัปโหลดวีดิทัศน์ยูทูบใน 1 นาที [4]

สิ่งที่งานวิจัยฉบับนี้ให้ความสนใจคือวีดิทัศน์ที่อยู่ในแวดวงการศึกษา อาจจะเป็นวีดิทัศน์การสอน (Tutorial Videos) วีดิทัศน์บทเรียนออนไลน์ (Online Course Videos) หรือวีดิทัศน์งานประชุมทางวิชาการ (Conference Videos) ซึ่งข้อมูลจาก [20] ในภาพที่ 2.8 และ [21] ภาพที่ 2.9 ที่แสดงถึงความนิยมในการดูวีดิทัศน์ในยูทูบแยกตามหมวดหมู่ในปีพุทธศักราช 2556 (คริสต์ศักราช 2013) โดยแบ่งระหว่างเพศชายและหญิงตามลำดับ พบว่าวีดิทัศน์ในหมวดหมู่การศึกษาได้รับความนิยมอยู่ในลำดับที่ 5 สำหรับเพศหญิงและลำดับที่ 6 สำหรับเพศชาย จากข้อมูลนี้ แสดงให้เห็นว่า การที่ยูทูบมีความหลากหลายในด้านเนื้อหามากกว่าเว็บไซต์การเรียนการสอนออนไลน์ขนาดใหญ่ อย่าง Udacity, Coursera หรือ Udemy ทำให้มีผู้ใช้งานในหมวดหมู่ด้านการศึกษาเป็นอันดับต้น ๆ นอกจากนั้น ยูทูบยังมีลักษณะเด่นคือ เข้าถึงง่าย และไม่ซับซ้อน สามารถเลือกดูวีดิทัศน์ใดก็ได้ โดยไม่ต้องลงทะเบียนวิชาเรียนดังเช่นเว็บไซต์การเรียนการสอนออนไลน์ข้างต้น เพราะฉะนั้นการค้นหาวีดิทัศน์ด้วยคำสำคัญ (Keyword) จึงทำได้ง่ายกว่าเว็บไซต์อื่น ซึ่งในส่วนของการค้นหาวีดิทัศน์นั้น ทางยู

ทูบได้เตรียมส่วนต่อประสานโปรแกรมประยุกต์ ไว้ให้นักพัฒนาได้ใช้ จากเหตุผลที่กล่าวมา ทำให้ยูทูบเป็นตัวเลือกที่มีความเหมาะสมสำหรับการใช้เป็นแหล่งข้อมูลวีดิทัศน์ภายนอกเพื่อใช้ในการแนะนำสำหรับงานวิจัยนี้



ภาพที่ 2.8 ข้อมูลการดูวีดิทัศน์ในยูทูบของเพศชายแบ่งตามหมวดหมู่ของวีดิทัศน์ [20]



ภาพที่ 2.9 ข้อมูลการดูวีดิทัศน์ในยูทูบของเพศหญิงแบ่งตามหมวดหมู่ของวีดิทัศน์ [21]

2.1.2.1 ส่วนต่อประสานโปรแกรมประยุกต์ของยูทูบ

ส่วนต่อประสานโปรแกรมประยุกต์ของยูทูบ (YouTube API) ที่ทางยูทูบเตรียมไว้ให้นักพัฒนานั้น มีอยู่ 3 ส่วนด้วยกัน คือ ส่วนการเล่นวีดิทัศน์ของยูทูบ (YouTube Player), ส่วนข้อมูลและทรัพยากรของยูทูบ (YouTube Data and Resources) และส่วนเครื่องมือและวิดเจตส์ (YouTube Tools and Widgets) ซึ่งส่วนต่อประสานโปรแกรมประยุกต์ที่ใช้ในงานวิจัยนี้ มีชื่อว่า ส่วนต่อประสานโปรแกรมประยุกต์ข้อมูลของยูทูบ (YouTube Data API) ที่อยู่ภายใต้ส่วนข้อมูลและทรัพยากรของยูทูบ ซึ่งปัจจุบันเป็นนับรุ่นที่ 3 ส่วนต่อประสานโปรแกรมประยุกต์ข้อมูลนี้ ประกอบไปด้วยฟังก์ชันให้ใช้งานมากมาย ทั้งฟังก์ชันการค้นหารายการวีดิทัศน์ ค้นหาข้อมูลเฉพาะวีดิทัศน์รายตัว รวมไปถึงการเพิ่ม ลบ หรือเปลี่ยนแปลงข้อมูลวีดิทัศน์ โดยฟังก์ชันที่ใช้ในงานวิจัยนี้ มี 2 ฟังก์ชันด้วยกันคือ ฟังก์ชันค้นหา (Search) และฟังก์ชันวีดิทัศน์ (Videos)

1. ฟังก์ชันค้นหา (Search)

ฟังก์ชันค้นหาจะให้ผลลัพธ์เป็นข้อมูลสารสนเทศที่เกี่ยวข้องกับวีดิทัศน์ (Video) เพลย์ลิสต์ (Playlist) หรือช่อง (Channel) ของยูทูบที่ตรงตามเงื่อนไขการค้นหาที่ใช้ในฟังก์ชันนี้ สำหรับการใช้งานฟังก์ชันนี้มีข้อมูล 2 ส่วนที่สำคัญดังแสดงในตารางที่ 2.1 และ ตารางที่ 2.2

ตารางที่ 2.1 ข้อมูลสำหรับคำร้องขอ (Request Data)

ชื่อตัวแปร	คำอธิบาย
part	part เป็นตัวแปรที่ระบุข้อมูลที่ต้องการในคำตอบสนอง (Response) ประกอบด้วย id – ข้อมูลไอดีของวีดิทัศน์ snippet – ข้อมูลพื้นฐานของวีดิทัศน์ เช่น ชื่อ คำอธิบาย ภาพประกอบ และวันที่เผยแพร่ เป็นต้น
q	q เป็นตัวแปรที่ใช้สำหรับระบุข้อความที่จะใช้ค้นหา ซึ่งสามารถระบุในรูปแบบพิเศษได้ เช่น เครื่องหมาย – แปลว่ายกเว้นคำที่ต่อจากเครื่องหมายนี้ เครื่องหมาย | แปลว่าหรือ

ตารางที่ 2.2 ข้อมูลสำหรับคำตอบสนอง (Response Data)

ชื่อตัวแปร	คำอธิบาย
id.kind	id.kind เป็นส่วนหนึ่งในข้อมูลที่ชื่อว่า id ที่ยูทูบส่งกลับมาให้ ข้อมูลนี้เป็นส่วนที่ระบุว่าข้อมูลไอดีนี้เป็นข้อมูลประเภทวีดิทัศน์ เพลย์ลิสต์หรือช่อง ซึ่งในงานวิจัยนี้จะใช้ข้อมูลประเภทวีดิทัศน์เท่านั้น
id.videoId	id.videoId เป็นส่วนหนึ่งในข้อมูลที่ชื่อว่า id ที่ยูทูบส่งกลับมาให้ ข้อมูลนี้เป็นส่วนที่ระบุว่า ไอดีของวีดิทัศน์นี้คืออะไร ในกรณีที่ข้อมูลนี้เป็นข้อมูลประเภท วีดิทัศน์

ฟังก์ชันค้นหา ส่วนที่ใช้ในงานวิจัยนี้เป็นส่วนการค้นหาไอดีของวีดิทัศน์ที่ตรงกับคำสำคัญที่ใส่ไปในเงื่อนไขการค้นหา แล้วนำไอดีของแต่ละวีดิทัศน์ไปค้นหาข้อมูลเฉพาะของแต่ละวีดิทัศน์ต่อด้วยฟังก์ชันวีดิทัศน์

2. ฟังก์ชันวีดิทัศน์ (Videos)

ฟังก์ชันวีดิทัศน์เป็นส่วนการจัดการเกี่ยวกับวีดิทัศน์ในยูทูบ มีทั้งความสามารถในการค้นหาข้อมูลรายละเอียดของวีดิทัศน์ (List method) การเพิ่มวีดิทัศน์ (Insert method) การเปลี่ยนแปลงข้อมูลวีดิทัศน์ (Update method) การลบวีดิทัศน์ (Delete method) เป็นต้น โดยงานวิจัยนี้จะใช้เพียงการค้นหาข้อมูลรายละเอียดของวีดิทัศน์เท่านั้น สำหรับการใช้งานฟังก์ชันนี้มีข้อมูล 2 ส่วนที่สำคัญดังแสดงในตารางที่ 2.3 และ ตารางที่ 2.4

ตารางที่ 2.3 ข้อมูลสำหรับคำร้องขอ (Request Data)

ชื่อตัวแปร	**คำอธิบาย**
part	part เป็นตัวแปรที่ระบุข้อมูลที่ต้องการในส่วนของคำตอบสนอง (Response) ประกอบไปด้วย snippet – ข้อมูลพื้นฐานของวีดิทัศน์ เช่น ชื่อ คำอธิบาย ภาพประกอบ และวันที่เผยแพร่ เป็นต้น
type	type เป็นตัวแปรที่ระบุประเภทของข้อมูลที่จะค้นหา ในที่นี้คือ video
id	id เป็นตัวแปรที่ระบุไอดีของวีดิทัศน์ที่ต้องการค้นหา โดยสามารถระบุได้มากกว่า 1 ไอดี โดยขั้นแต่ละไอดีด้วยเครื่องหมาย ','

ตารางที่ 2.4 ข้อมูลสำหรับคำตอบสนอง (Response Data)

ชื่อตัวแปร	**คำอธิบาย**
snippet.title	snippet.title เป็นข้อมูลชื่อของวีดิทัศน์
snippet.description	snippet.description เป็นข้อมูลคำอธิบายของวีดิทัศน์
snippet.tags	snippet.tags เป็นข้อมูลแถบป้ายข้อมูลของวีดิทัศน์

2.1.3 ระบบแนะนำแบบพิจารณาจากเนื้อหา

ระบบแนะนำ เป็นระบบที่สร้างขึ้นเพื่อนำเสนอวัตถุ (Item) ที่ต้องการแนะนำแก่ผู้ใช้งาน เช่น หนังสือ ภาพยนตร์ บริการ ข้อมูลข่าวสาร เนื้อหาทางการศึกษา เป็นต้น ซึ่งเป็นการแนะนำที่เฉพาะเจาะจงในแต่ละบุคคล โดยพิจารณาจาก ข้อมูลของผู้ใช้งาน ความคล้ายคลึงกันของเนื้อหา ความคล้ายคลึงกันของผู้ใช้งานที่มีประวัติเหมือนกัน หรือแม้กระทั่งความพึงพอใจของผู้ใช้งานที่ได้รับการแนะนำดังกล่าว

ระบบแนะนำ สามารถอธิบายให้อยู่ในสมการที่ 1 ได้ดังนี้ [22]

$$\forall c \in C , s'_c = \arg\max_{s \in S} u(c,s) \qquad (1)$$

โดยที่

C = เซตของผู้ใช้งานทั้งหมด

S = เซตของวัตถุแนะนำที่เป็นไปได้ทั้งหมด

c = ผู้ใช้งานใด ๆ

s = วัตถุแนะนำใด ๆ

S’c = วัตถุแนะนำ s ที่เหมาะสมกับผู้ใช้งาน c ที่สุด

u = ฟังก์ชันที่วัดความเหมาะสมต่อการแนะนำวัตถุ s ให้กับผู้ใช้งาน c

จากสมการข้างต้น สามารถอธิบายได้ว่า วัตถุแนะนำที่เหมาะสมนั้น หาได้จากฟังก์ชันที่ให้ค่าวัดความเหมาะสมมากที่สุด สำหรับวัตถุแนะนำ s ให้กับผู้ใช้งาน c ซึ่งในส่วนของฟังก์ชันหาความเหมาะสมนี้ ถือเป็นหัวข้อหลักของงานวิจัยทางด้านการสร้างระบบแนะนำ ในด้านของการพัฒนาอัลกอรึทึมในการหาความเหมาะสม ซึ่งระบบแนะนำในปัจจุบันมีวิธีการทำงานหลัก ๆ อยู่ 3 วิธีคือ ระบบแนะนำโดยพิจารณาจากเนื้อหา (Content-Based Recommender System) ระบบแนะนำโดยการคัดกรองร่วม (Collaborative Filtering Recommender System) และระบบแนะนำแบบผสม (Hybrid Recommender System) ส่วนที่งานวิจัยนี้ใช้คือระบบแนะนำโดยพิจารณาจากเนื้อหา ซึ่งเป็นวิธีการนำเสนอโดยอาศัยพื้นฐานจากรายละเอียด เนื้อหาหรือคุณสมบัติของข้อมูลที่มีความคล้ายคลึงกัน ระหว่างข้อมูลเนื้อหาของผู้ใช้งานที่มีอยู่เดิมกับเนื้อหาใหม่ที่จะนำเสนอ [23] โดยให้ u(c,s) เป็นฟังก์ชันหาความเหมาะสมในการแนะนำวัตถุ s ให้กับผู้ใช้งาน c วิธีแนะนำโดยพิจารณาจากเนื้อหานั้น จะหา u(c,s) ด้วยข้อมูลที่มาจาก $u(c,s_i)$ โดยที่ s_i คือวัตถุแนะนำที่เป็นสมาชิกในเซตของวัตถุแนะนำที่เป็นไปได้ทั้งหมด S ที่มีความคล้ายคลึงกับ s ตัวอย่างเช่น การแนะนำภาพยนตร์

ให้กับผู้ใช้งาน ก โดยพิจารณาจากภาพยนตร์ที่ผู้ใช้งาน ก ได้รับชม และมีความพึงพอใจ ระบบแนะนำก็จะเสนอภาพยนตร์อีกเรื่องที่มีคุณสมบัติคล้ายคลึงกัน เช่น ชื่อเรื่อง ประเภทของภาพยนตร์ ผู้กำกับ นักแสดงนำ ฯลฯ ให้แก่ผู้ใช้งาน

ในการสกัดลักษณะเด่น (Feature Extraction) ของระบบแนะนำโดยพิจารณาจากเนื้อหานั้น ส่วนใหญ่จะใช้ลักษณะเด่นที่เป็นตัวอักษร (Textual Feature) ซึ่งวิธีการหนึ่งที่เป็นที่นิยมคือ การหาคำสำคัญในเนื้อหาของวัตถุแนะนำ โดยอาศัยสมการหาความถี่และน้ำหนักของคำที่เรียกว่า ทีเอฟ-ไอดีเอฟ (Term Frequency-Inverse Document Frequency: TF-IDF) [24] โดยที่ tf หมายถึงความถี่ที่คำที่ i ปรากฏในเอกสาร j ในขณะที่ df คือจำนวนของเอกสารทั้งหมดที่มีคำที่ i ปรากฏอยู่ และ N คือจำนวนเอกสารทั้งหมด ซึ่งคะแนนความสำคัญของคำที่ i ในเอกสาร j คำนวณได้จากสมการที่ 2

$$w_{i,\ j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (2)$$

หลังจากได้คะแนนของลักษณะเด่น ซึ่งในที่นี้คือค่าความสำคัญของคำในเอกสารจากการคำนวณด้วยทีเอฟ-ไอดีเอฟแล้ว การหาวัตถุแนะนำที่มีความคล้ายกัน เพื่อใช้แนะนำให้แก่ผู้ใช้งานนั้นจะใช้ฟังก์ชันความคล้ายกัน ซึ่งมีอยู่หลายวิธี โดยวิธีที่เป็นที่นิยมหลัก ๆ คือ ความคล้ายกันของโคไซน์ (Cosine Similarity) ความคล้ายกันของแจคการ์ด (Jaccard Similarity) และสัมประสิทธิ์สหสัมพันธ์ (The coefficient of Correlation) เมื่อได้คะแนนความคล้ายกันระหว่างวัตถุของผู้ใช้งาน กับวัตถุแนะนำแล้ว จะทำการเลือกวัตถุที่จะแนะนำ ซึ่งส่วนใหญ่จะใช้วิธีที่เรียกว่า K-Nearest Neighbor (KNN) [25] โดยเลือกแนะนำวัตถุแนะนำที่มีความคล้ายกัน K อันดับจากมากไปน้อยให้แก่ผู้ใช้งาน

แม้ที่ผ่านมาระบบแนะนำแบบพิจารณาจากเนื้อหาจะไม่เป็นที่นิยมนัก เนื่องจากการหาความคล้ายของวัตถุแนะนำ ไม่สามารถนำมาประยุกต์ใช้สำหรับการแนะนำแบบเฉพาะบุคคล (Personalization) ได้ดีเท่าระบบแนะนำโดยการคัดกรองร่วม และผลลัพธ์ของการแนะนำก็ไม่ได้สดใหม่ หรือหลากหลาย อีกทั้งยังมีประเด็นของลักษณะเด่นที่ใช้เนื้อหาของวัตถุที่ยากต่อการคำนวณและส่งผลกระทบต่อประสิทธิภาพการทำงานของการแนะนำ อย่างไรก็ตาม แนวคิดเรื่องการจัดลำดับการแนะนำใหม่ของงานวิจัยนี้ อาจช่วยแก้ปัญหาส่วนนี้ในระบบแนะนำแบบพิจารณาจากเนื้อหาได้

2.1.4 นิยามความครอบคลุมแนวคิด

โดยทั่วไปแล้วนิยามคำว่าความครอบคลุม (Coverage) ในงานวิจัยด้านระบบแนะนำนั้น มักจะหมายความถึงรายการแนะนำของอัลกอริทึมใด ๆ ก็ตาม ที่รายการแนะนำเหล่านั้น ปรากฏอยู่ในรายการแนะนำที่ควรจะได้รับการแนะนำให้แก่ผู้ใช้งาน [8] ตามวัตถุประสงค์ของการแนะนำ ตัวอย่างเช่น วัตถุประสงค์ของการแนะนำคือผลการแนะนำจะต้องครอบคลุมภาพยนตร์จากทุกประเทศที่สนใจ ซึ่งเป็นเซตของประเทศ = [ไทย, จีน, อินเดีย, มาเลเซีย] แต่รายการแนะนำได้ได้จากอัลกอริทึมที่ทำการทดสอบนั้น ประกอบไปด้วยภาพยนตร์จากไทยและจีนเท่านั้น หมายความว่าอัลกอริทึมที่เสนอนั้นให้ผลการแนะนำมีความครอบคลุมภาพยนตร์จากทุกประเทศที่สนใจอยู่ที่ร้อยละ 50 ถึงแม้ว่าการแนะนำครั้งนั้น จะให้ความแม่นยำ (Accuracy) อยู่ที่ร้อยละ 100 ก็ตาม

จากลักษณะของการพิจารณาความครอบคลุมนี้ หลายงานวิจัยได้ใช้คุณสมบัติหนึ่งของระบบแนะนำ คือความหลากหลาย ในการเพิ่มความครอบคลุมตามวัตถุประสงค์ของการแนะนำ [8, 11] ซึ่งคุณสมบัติด้านความหลากหลายในผลการแนะนำนั้น เป็นหัวข้อที่มีอัลกอริทึมเสนอขึ้นมาเพื่อเพิ่มประสิทธิภาพด้านความหลากหลายดังกล่าว ซึ่งกล่าวต่อไปในส่วนของงานวิจัยที่เกี่ยวข้อง ตัวอย่างความสามารถของความหลากหลายในการเพิ่มความครอบคลุมจากตัวอย่างข้างต้นนั้น อาจจะเป็นอัลกอริทึมที่ทำให้ผลการแนะนำ มีภาพยนตร์จากทุก ๆ ประเทศในแถบเอเชียมากที่สุด แม้ผลลัพธ์อาจจะมีประเทศที่ไม่เกี่ยวข้องมาด้วย ซึ่งอาจจะทำให้ความแม่นยำลดลงตามมา แต่ก็ถือว่าครอบคลุมเซตของประเทศที่สนใจ

ส่วนต่อมาคือ นิยามของคำว่าแนวคิดของเอกสารในงานวิจัยนี้ หมายถึงแนวคิดที่เกี่ยวข้องภายในเอกสารที่ระบบแนะนำวีดิทัศน์นี้พิจารณา ตัวอย่างเช่น เอกสารเรื่องการวิเคราะห์อัลกอริทึม ก็จะประกอบไปด้วยแนวคิดที่เกี่ยวกับ ภาระการทำงานเชิงเวลา การวิเคราะห์ประสิทธิภาพอัลกอริทึม บิ๊กโอ เป็นต้น ซึ่งแนวคิดเหล่านี้ เทียบเคียงได้กับคำสำคัญ (Keyword) ที่อาจปรากฏได้ในเอกสารที่จะใช้ในการสร้างรายการแนะนำ ส่วนแนวคิดของวีดิทัศน์ในงานวิจัยนี้ จะเทียบเคียงกับคำสำคัญที่ใช้ในการค้นหาวีดิทัศน์เหล่านั้น ตัวอย่างเช่น วีดิทัศน์ ก ถูกค้นหาด้วยคำสำคัญ ค1 ค2 ค3 ซึ่งหมายความว่าวีดิทัศน์ ก ประกอบไปด้วยแนวคิด ค1 ค2 ค3

จากคำอธิบายของคำว่าครอบคลุมกับแนวคิดข้างต้น สามารถสรุปนิยามของคำว่าความครอบคลุมแนวคิดในงานวิจัยนี้ได้ว่า ความครอบคลุมแนวคิดของรายการแนะนำคือผลการแนะนำที่คำสำคัญของวีดิทัศน์ที่ถูกเลือกเพื่อแนะนำนั้นมีความหลากหลายเพียงพอที่จะครอบคลุมคำสำคัญของเอกสาร เพราะฉะนั้น การที่รายการวีดิทัศน์ที่แนะนำนั้นมีความเกี่ยวข้องกับคำสำคัญทุกคำซึ่งปรากฏในเอกสาร แสดงว่ารายการแนะนำนี้ครอบคลุมแนวคิดของเอกสารทั้งหมด

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยที่เกี่ยวข้องกับการจัดลำดับการแนะนำใหม่

การเพิ่มความหลากหลายเป็นหนึ่งในวิธีการเพิ่มความพึงพอใจให้แก่ผู้ใช้งาน โดยการแนะนำวัตถุที่มีลักษณะหลายอย่างแตกต่างกัน งานวิจัยด้านการเพิ่มความหลากหลายนั้น ส่วนใหญ่จะปรากฏอยู่ในหัวข้อที่เกี่ยวข้องกับการค้นคืนเอกสารและการสร้างระบบแนะนำ ในส่วนของระบบแนะนำนั้นก่อนที่ความหลากหลายจะเข้ามามีบทบาทสำคัญ นอกเหนือไปจากวิธีพื้นฐานอย่าง Content-based, Collaborative filtering และ Hybrid แล้ว ส่วนใหญ่มักมุ่งค้นหาวัตถุแนะนำที่มีความสัมพันธ์ และมีความคล้ายกันกับวัตถุที่ผู้ใช้งานมีประวัติสนใจ ซึ่งมาจากแนวคิดที่พิจารณาให้ความคล้ายกันดังกล่าวเป็นปัจจัยหลักในการสร้างความพึงพอใจให้แก่ผู้ใช้งาน แต่ในภายหลัง เมื่อความหลากหลายได้รับการเสนอให้เป็นปัจจัยหนึ่งของความพึงพอใจของผู้ใช้งาน [26] หลายงานวิจัยก็ได้เสนออัลกอริทึมในการเพิ่มความหลากหลายซึ่งส่วนใหญ่จะตั้งอยู่บนแนวคิดการเรียงลำดับผลลัพธ์ใหม่ (Re-ranking) [8]

อัลกอริทึมที่ใช้แนวคิดการเรียงลำดับผลลัพธ์ใหม่นั้นมักถูกสร้างใน 2 กระบวนทัศน์ (Paradigm) ด้วยกัน คือ การเดินสุ่ม ซึ่งถูกใช้ในงานด้านการสืบค้นเอกสารรวมทั้งระบบแนะนำในหลายงานวิจัย [27-29] และ อัลกอริทึมที่ครอบคลุมเซตที่มากที่สุด (Maximum set cover algorithm) บนการใช้งานทั้งออนไลน์และออฟไลน์ [10, 30, 31] ซึ่งวิธีอัลกอริทึมที่ครอบคลุมเซตที่มากที่สุดนั้น จัดอยู่ในปัญหา NP-hard อัลกอริทึมที่ถูกเสนอด้วยกระบวนทัศน์นี้จึงมักจะเป็นอัลกอริทึมเชิงละโมบ (Greedy algorithm)

หนึ่งในอัลกอริทึมจากการเดินสุ่มที่ประสบความสำเร็จอย่างมากคือ PageRank [32] ที่กูเกิ้ลใช้ในการแสดงผลการค้นหา โดยอาศัยการเดินสุ่มในการเปลี่ยนลำดับการแสดงผลแล้วจึงแสดงผล k รายการที่มีความหลากหลายกว่าลำดับเดิม จากความสำเร็จดังกล่าว จึงมีการพัฒนาอัลกอริทึมที่ใช้วิธีการเดินสุ่มนี้เพิ่มขึ้นอย่างต่อเนื่อง ตัวอย่างเช่น GRASSHOPPER [28], DRAGON [33] และ GenDeR [34]

ในทางตรงกันข้าม อัลกอริทึมที่ครอบคลุมเซตที่มากที่สุดนั้นก็มีการเสนออัลกอริทึมเชิงละโมบหลายอัลกอริทึมเพื่อแก้ปัญหาความหลากหลายที่ไม่น้อยไปกว่าการเดินสุ่ม ได้แก่ เอ็มเอ็มอาร์ (Maximal Marginal Relevance: MMR) [9], แม็กซ์ซัม (Max-Sum) [35], แม็กซ์มิน (Max-Min) [35] และโมโนอ็อบเจ็กทีฟ (MonoObjective) [35] ซึ่งการวัดผลบนเซตข้อมูลเอกสารทางกฎหมาย [15] ได้เปรียบเทียบทุกอัลกอริทึม ทั้งจากการเดินสุ่มและการครอบคลุมเซตที่มากที่สุด และให้ผลลัพธ์ที่แสดงให้เห็นว่า เอ็มเอ็มอาร์ มีประสิทธิภาพในการสร้างรายการที่มีความหลากหลายได้มากกว่าวิธี

อื่น โดยใช้ตัวชี้วัดที่การประชุมทางวิชาการทีอาร์อีซีกำหนดไว้ให้ ในงานการค้นคืนเอกสารในหัวข้อความหลากหลาย

อย่างไรก็ตาม แม้เอ็มเอ็มอาร์จะเป็นอัลกอริทึมเชิงละโมบที่ให้ผลลัพธ์ที่ดีที่สุด แต่วิธีนี้ยังคงมีจุดอ่อนในด้านระยะเวลาในการดำเนินการ เช่นเดียวกับอัลกอริทึมเชิงละโมบตัวอื่น ที่มีภาระขึ้นกับจำนวนของรายการ k ที่ต้องการแนะนำ กับจำนวนวัตถุที่เป็นไปได้ทั้งหมด N ตัว โดยภาระการทำงานของเอ็มเอ็มอาร์จะเท่ากับ $O(k^2N)$ [36] ในขณะที่ภาระการทำงานของการเดินสุ่มแบบเริ่มต้นใหม่ในอัลกอริทึมอื่น ๆ นั้นน้อยกว่า ตัวอย่างเช่น DRAGON ใช้ $O(m + nk)$ ส่วน GenDeR แบบปกติใช้ $O(n^2 + nk)$ แต่ในกรณีที่เมทริกซ์นี่นำมาใช้มีลักษณะที่เป็นสปาร์ส (Sparse Matrix) อัตราการเติบโตจะลดลงมาเท่ากับอัลกอริทึม DRAGON ส่วนการเดินสุ่มแบบดั้งเดิมนั้นใช้เวลาเพียง O(N) [37] จากข้อเท็จจริงดังกล่าว จึงทำให้เอ็มเอ็มอาร์ต้องใช้เวลาเพิ่มขึ้นทุกครั้งที่มีการเพิ่มค่า k ต่างจากการเดินสุ่มแบบดั้งเดิมที่คำนวณคะแนนเก็บไว้ครั้งเพียงเดียวแล้วสามารถใช้ได้สำหรับค่า k ทุกค่า

ดังนั้น แทนที่จะหาวิธีพัฒนาอัลกอริทึมเพื่อลดระยะเวลาการทำงานของเอ็มเอ็มอาร์ งานวิจัยนี้เลือกที่จะเพิ่มประสิทธิภาพด้านความหลากหลายของการเดินสุ่มแบบเริ่มต้นใหม่แบบดั้งเดิม ให้ใกล้เคียงกับวิธีเอ็มเอ็มอาร์แทน เนื่องจากข้อได้เปรียบของการเดินสุ่มที่ใช้การคำนวณครั้งเดียวแล้วปรับใช้ได้กับทุกจำนวนการแนะนำ k โดยจะเพิ่มในส่วนของฟังก์ชันวัตถุประสงค์เพื่อใช้คัดเลือกผลการแนะนำที่ได้จากการเดินสุ่มในแต่ละครั้งระหว่างการเดินสุ่ม ก่อนที่จะหยุดการเดินตามเงื่อนไขแล้วเก็บรายการแนะนำที่ให้ค่าฟังก์ชันวัตถุประสงค์ที่ดีที่สุด มาใช้เสนอแก่ผู้ใช้งานต่อไป ไม่ว่าจะที่ค่า k เท่ากับเท่าใดก็ตาม

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2.1.1 การเดินสุ่มแบบเริ่มต้นใหม่

การดำเนินการเดินสุ่มนั้น สามารถทำได้หลายวิธีด้วยกัน วิธีหนึ่งที่เป็นที่นิยมในวงกว้างคือ วิธีการวนซ้ำ (Iteration Method) จนกว่าค่าจะลู่เข้า หรือค่า L2 norm ของค่า $\vec{r}_i$ รอบที่ t และรอบที่ t-1 ในสมการที่ 3 [38] มีค่าน้อยกว่าค่า ξ หรือจนกว่าการสุ่มวนไปจนถึงรอบที่ m ซึ่งวิธีการนี้ ถือเป็นวิธีที่คำนวณไว้ก่อน (Pre-storage)

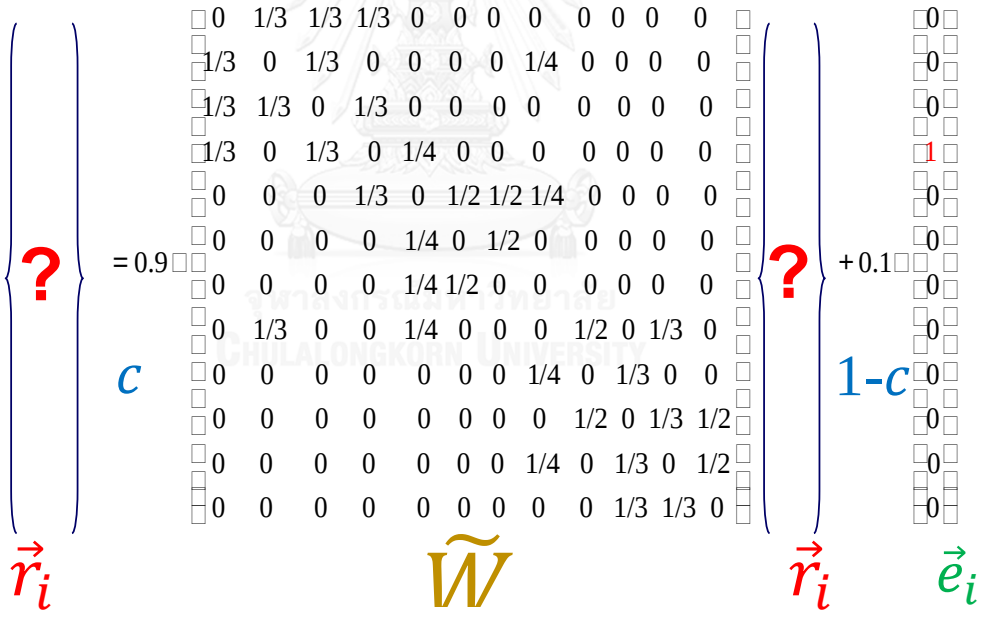
$$\vec{r}_i\,[t] = c\widetilde{W}\,\vec{r}_i\,[t\text{-}1]+(1\text{-}c)\,\vec{e}_i \qquad (3)$$

โดยตัวแปรต่าง ๆ ได้นิยามไว้ในตารางที่ 2.5

ตารางที่ 2.5 รายละเอียดคำอธิบายของตัวแปรในสมการการเดินสุ่มแบบเริ่มต้นใหม่

ตัวแปร	คำอธิบาย
$W= [w_{i,j}]$	เมทริกซ์ของกราฟที่มีน้ำหนัก โดยที่ $1 \le i, j \le n$
$\widetilde{W}$	เมทริกซ์ที่เกิดจากการทำให้เป็นมาตรฐาน (Normalization) ของเมทริกซ์ W
$\vec{e}_i$	เมทริกซ์ตั้งต้นสำหรับการสืบค้น
$\vec{r}_i= [r_{i,j}]$	เมทริกซ์ผลลัพธ์การแนะนำ
c	ค่าคงที่ในการเดินสุ่ม
n	จำนวนวัตถุในการคำนวณทั้งหมด
k	จำนวนรายการที่จะเลือกเพื่อทำการแนะนำ
t	รอบที่ทำการเดินสุ่ม
m	จำนวนรอบสำหรับหยุดการเดินสุ่ม
ξ	ค่าขีดแบ่ง (Threshold) สำหรับการหยุดการเดินสุ่ม

โดยตัวอย่างการคำนวณวิธีการเดินสุ่มแบบเริ่มต้นใหม่ในภาพที่ 2.10 [39]



ภาพที่ 2.10 ตัวอย่างเมทริกซ์ที่เกี่ยวข้องกับการเดินสุ่ม

จากภาพที่ 2.10 สิ่งที่ต้องหาในการเดินสุ่ม คือ $\vec{r}_i$ ที่ i = 3 จาก $\vec{e}_i$ วนซ้ำไปเรื่อย ๆ จนกว่าจะตรงกับเงื่อนไขการหยุดการเดินสุ่ม วิธีหนึ่งที่ทำได้ คือการนำผลลัพธ์ของครั้งที่ t ไปใช้ต่อในครั้งที่ t+1 ดังแสดงในสมการที่ 3 ซึ่งผลลัพธ์ที่คาดหวังจากการดำเนินการสุดท้าย ค่า $\vec{r}_i$ ที่ได้จะมีลักษณะเช่นเดียวกับภาพที่ 2.11

$$\begin{bmatrix}0.13\\0.10\\0.13\\0.22\\0.13\\0.05\\0.05\\0.08\\0.04\\0.03\\0.04\\0.02\end{bmatrix} = 0.9 \begin{bmatrix}0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0\\ 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 1/3 & 0 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1/3 & 0 & 1/2 & 1/2 & 1/4 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1/4 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1/3 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/2 & 0 & 1/3 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/3 & 1/2\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 0 & 1/2\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0\end{bmatrix} \begin{bmatrix}0.13\\0.10\\0.13\\0.22\\0.13\\0.05\\0.05\\0.08\\0.04\\0.03\\0.04\\0.02\end{bmatrix} + 0.1 \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

ภาพที่ 2.11 ตัวอย่างเมทริกซ์ผลลัพธ์จากการเดินสุ่ม

ซึ่งค่า $\vec{r}_i$ นี้จะเป็นคะแนนที่จะใช้ในการคัดเลือกวัตถุทัศน์ k อันดับแรก เรียงตามคะแนนจากมากไปน้อย เพื่อใช้เสนอให้แก่ผู้เรียน

นอกจากวิธีดังกล่าว ยังมีวิธีที่ชื่อว่า B_LIN [40] ที่มีการแบ่งการคำนวณบางส่วนเป็นแบบออฟไลน์ และปรับปรุงผลของการเดินสุ่มตามการค้นหาแบบออนไลน์ และอีกหลายวิธีที่จะกล่าวถึงภายหลังในส่วนงานวิจัยที่เกี่ยวข้อง ซึ่งประเด็นในการพิจารณาวิธีดำเนินการเดินสุ่ม มีเรื่องที่ต้องคำนึงถึงสองเรื่องด้วยกัน ได้แก่

1. ประเด็นการคำนวณแบบออนไลน์ (Online computational cost)

การคำนวณแบบออนไลน์จะให้ผลลัพธ์การจัดลำดับการแนะนำใหม่ที่ทันสมัย และปรับเปลี่ยนได้ตามข้อมูลใหม่ที่เพิ่มเข้ามา หรือเปลี่ยนแปลงได้ทันที อย่างไรก็ดี การคำนวณแบบออนไลน์นั้น จำเป็นต้องพิจารณาถึงเวลาที่ใช้ในการคำนวณ ซึ่งหากผลการจัดลำดับการแนะนำใหม่ใช้เวลานาน จนทำให้ผู้ใช้งานต้องรอผลการคำนวณ จะไม่ส่งผลดีนักในมุมมองของการพัฒนาระบบเพื่อตอบสนองต่อการใช้งานจริง

2. ประเด็นการคำนวณไว้ก่อน (Pre-storage cost)

แม้การคำนวณไว้ก่อน จะลดระยะเวลาการรอของผู้ใช้งาน และสามารถเรียกใช้ผลการเรียงลำดับได้ในทันทีโดยไม่ต้องคำนวณใหม่ทุกครั้ง วิธีการนี้ก็ยังมีจุดอ่อนในเรื่องข้อมูลไม่ทันสมัย ไม่เปลี่ยนแปลงตามข้อมูลใหม่ได้อย่างทันเวลา และเรื่องพื้นที่ในการจัดเก็บผลการเรียงลำดับข้อมูลไว้ใช้ โดยข้อมูลที่เพิ่มขึ้นเรื่อย ๆ ย่อมส่งผลให้พื้นที่ที่ต้องจัดเก็บผลการเรียงลำดับมากขึ้นตามไปด้วย

2.2.1.2 เอ็มเอ็มอาร์

อัลกอริทึมเอ็มเอ็มอาร์ (Maximal Marginal Relevance: MMR) เป็นวิธีบรรทัดฐานที่จะใช้ในการทดลองของงานวิจัยนี้ ซึ่งวิธีเอ็มเอ็มอาร์ เป็นอัลกอริทึมที่ทำการเลือกวัตถุแนะนำในแต่ละลำดับโดยพิจารณาจากความไม่เหมือนกัน ระหว่างวัตถุที่ทำการเลือกไปแล้ว กับวัตถุที่กำลังพิจารณา โดยจะเลือกวัตถุที่จะพิจารณาก็ต่อเมื่อให้ค่า $f_{MMR}(u, q)$ สูงสุด โดยค่าดังกล่าวหาได้จากสมการที่ 4

$$f_{MMR}(u, q) = (1 - c)\, sim(u, q) + c \sum_{v \in S} dis(u, v) \qquad (4)$$

โดยที่ sim(u, q) คือค่า Cosine Similarity ระหว่างวัตถุที่กำลังพิจารณากับข้อมูลที่ใช้ค้นหา ในที่นี้ข้อมูลที่ใช้ค้นหาคือเอกสารในมายคอร์สวิลล์ ส่วนวัตถุที่กำลังพิจารณาคือวีดิทัศน์จากยูทูบ ในขณะที่ dis(u, v) คือความไม่เหมือนกันของวีดิทัศน์ทำการพิจารณาอยู่กับรายการวีดิทัศน์ที่ทำการเลือกไปแล้ว ความไม่เหมือนกันนี้ สามารถหาได้จาก 1 – sim(u,v) โดยที่ค่า c เป็นค่าคงที่ของสมการ ซึ่งวัตถุแรกที่จะได้รับการเลือกก่อนคือวัตถุที่ให้ค่า sim(u, q) สูงสุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2.2 งานวิจัยที่เกี่ยวข้องกับฟังก์ชันความหลากหลายของหมวดหมู่

งานวิจัยเรื่อง Evolving Mobile App Recommender Systems: An Incremental Multi-objective Approach [13] เป็นงานวิจัยที่เกี่ยวกับการแนะนำแอปพลิเคชันสำหรับการใช้งานในโทรศัพท์มือถือ โดยเซียและคณะ (Xia et al.) พยายามที่จะเพิ่มฟังก์ชันวัตถุประสงค์ให้กับการแนะนำ งานวิจัยนี้ได้เสนอวิธีการใหม่ที่ชื่อว่า Rank aggregation based evolving scheme ซึ่งหนึ่งในฟังก์ชันวัตถุประสงค์ของวิธีนี้ คือการขยายขอบเขต (Range) ของการแนะนำ เพื่อให้แอปพลิเคชันใหม่ ๆ ได้มีโอกาสปรากฏอยู่ในรายการแนะนำ ซึ่งในฟังก์ชันวัตถุประสงค์นี้ มีฟังก์ชันย่อยหนึ่งที่ชื่อว่าฟังก์ชันความหลากหลายของหมวดหมู่ (Category diversity) ซึ่งอธิบายด้วยสมการที่ 5

$$D_c(i) = \left(\frac{C_d \times (1 - e_i)}{C_d \times 1}\right) \times (C_n \times 1) \qquad (5)$$

โดยที่ $C_d = r_i^T \times C$ เป็นตัวบ่งชี้ว่าแอปพลิเคชันที่ปรากฏในรายการแนะนำในแต่ละหมวดหมู่มีจำนวนเท่าใด ส่วน C_n เป็นการแปลงรูปมาจาก C_d โดยที่ $C_n(k) = 1$ ถ้า $C_d(k) \geq 1$ และ $C_n(k) = 0$ ถ้า $C_d(k) \leq 1$ ซึ่ง C_n แสดงถึงจำนวนหมวดหมู่ที่การแนะนำนี้ครอบคลุม โดยตัวแปรที่เกี่ยวข้องนิยามไว้ในตารางที่ 2.6

ตารางที่ 2.6 รายละเอียดคำอธิบายของตัวแปรในสมการฟังก์ชันความหลากหลายของหมวดหมู่

ตัวแปร	คำอธิบาย
A	เซตของแอปพลิเคชันทั้งหมด
N_A	จำนวนสมาชิกของ A
R_i	แอปพลิเคชันที่แนะนำสำหรับแอปพลิเคชัน i
N_R	จำนวนของแอปพลิเคชันที่แนะนำ
r_i	เวกเตอร์ขนาด $N_A \times 1$ โดยที่ $r_i(k)=1$ ถ้า $k \in R_i$
C	เมทริกซ์ขนาด $N_A \times N_c$ C(i, j) = 1 ถ้าแอปพลิเคชัน i อยู่ในหมวดหมู่ j หรือ C(i, j)=0 ในทางตรงกันข้าม
e_i	e(i)=1, e(j)=0 สำหรับ j ที่ $j \neq i$ (กรณีที่แอปพลิเคชัน i ไม่อยู่ในหมวกหมู่ j)
1	1(i)= 1 สำหรับทุกค่า i

ฟังก์ชันความหลากหลายของหมวดหมู่นี้ จะถูกใช้ในการคำนวณคะแนนความหลากหลายจากการดำเนินการเดินสุ่มในแต่ละครั้ง

2.2.3 งานวิจัยที่เกี่ยวข้องกับการประเมินผลการแนะนำ

2.2.3.1 เซเรนดิพิตี้

ในการประเมินผลระบบแนะนำวีดิทัศน์เบื้องต้นนั้น ทางผู้วิจัยต้องการวัดผลรายการแนะนำที่ได้จากการสร้างระบบแนะนำวิธีทัศน์ที่เสนอ ว่ามีความแตกต่างกันในส่วนของแนวคิดมากพอหรือไม่ เนื่องจากหากรายการวีดิทัศน์แตกต่างกันไม่มากพอ อาจส่งผลให้แนวคิดของวีดิทัศน์ที่เสนอ ปรากฏเพียงรายการที่เกี่ยวข้องกับเนื้อหาหลัก ทำให้มีแนวคิดไม่หลากหลายเพียงพอที่จะดำเนินการอัลกอริทึม ทางผู้วิจัยจึงได้หาเครื่องมือที่ใช้วัดผลการแนะนำที่เป็นประโยชน์ แต่ไม่ได้เกี่ยวข้องกับเนื้อหาหลักของเอกสารโดยตรง

ในงานวิจัยของ Mouzhi Ge [6] ได้เสนอการวัดผลการแนะนำมาสองวิธีด้วยกัน แต่วิธีที่น่าสนใจคือการวัดผลที่ชื่อว่า เซเรนดิพิตี้ (Serendipity) ในงานวิจัยนี้ได้อธิบายความหมายของเซเรนดิพิตี้เอาไว้ว่า เป็นการค้นหาความไม่คาดฝันจากผลลัพธ์การแนะนำ โดยผลลัพธ์ที่ไม่คาดฝันนี้จะเป็นประโยชน์แก่ผู้ใช้งาน ตัวอย่างเช่น ผลการแนะนำวีดิทัศน์ที่ชื่อ Analysis and Design Algorithm จากเอกสารที่เกี่ยวกับบทเรียนวิชาอัลกอริทึมเรื่อง Divide and Conquer เราสามารถประเมินได้ว่าเป็นผลการแนะนำที่ไม่คาดคิดมาก่อน เนื่องจากวีดิทัศน์นี้ไม่มีส่วนที่กล่าวถึงอัลกอริทึมดังกล่าว แต่เนื่องจากเนื้อหาของวีดิทัศน์ ช่วยเสริมความเข้าใจเกี่ยวกับการวิเคราะห์อัลกอริทึม จึงสามารถกล่าวได้ว่า เป็นผลการแนะนำไม่คาดฝันเป็นประโยชน์

การคำนวณเซเรนดิพิตี้นั้น สามารถคำนวณได้จากสมการที่ 6

$$\text{Serendipity} = \frac{\sum_{i=1}^{N} u(RS_i)}{N} \qquad (6)$$

โดยที่ $u(RS_i)$ เป็นฟังก์ชันที่ใช้วัดความเป็นประโยชน์ของผลการแนะนำที่ไม่คาดหวัง (นิยามของฟังก์ชันจะกล่าวถึงในส่วนของการประเมินผล) ในขณะที่ N คือจำนวนผลการแนะนำที่ไม่คาดหวังทั้งหมด

การผู้วิจัยจึงได้เลือกใช้ค่าเซเรนดิพิตี้นี้ เพื่อวัดว่าวิธีการสร้างระบบแนะนำวีดิทัศน์ที่ได้เสนอนั้น มีวีดิทัศน์ที่นอกเหนือไปจากเนื้อหาหลักของเอกสารหรือไม่ และวีดิทัศน์เหล่านั้นเป็นประโยชน์กับผู้เรียนหรือไม่ การที่ผลการแนะนำมีรายการแนะนำที่เป็นประโยชน์นอกเหนือไปจากเนื้อหาหลักของวีดิทัศน์นั้น ช่วยรับประกันว่าวิธีการสร้างระบบแนะนำวีดิทัศน์ที่เสนอนี้ มีความหลากหลายเพียงพอที่จะทำการเรียงลำดับใหม่ด้วยอัลกอริทึมที่เสนอ

อย่างไรก็ตาม การวัดด้วยเซเรนดิพิตี้เพียงอย่างเดียวนั้น อาจส่งผลให้ความแม่นยำลดลงได้ ในกรณีที่ค่าเซเรนดิพิตี้มีค่าสูง ต้องพิจารณาต่อมาว่าค่าดังกล่าวมีจำนวนผลการแนะนำที่ไม่คาดหวังเท่าใด หากจำนวนผลการแนะนำที่ไม่คาดหวังสูงมาก หรือเทียบเท่ากับจำนวนวัตถุที่แนะนำทั้งหมด ก็จะกลายเป็นว่า การแนะนำนั้นมีผลลัพธ์ที่เกี่ยวข้องกับเอกสารน้อยมาก และทำให้ความแม่นยำลดลง ยกตัวอย่างเช่น หากมีจำนวนวัตถุที่แนะนำทั้งหมด 10 ชิ้น โดยแบ่งเป็นผลการแนะนำที่ไม่คาดหวัง 8 ชิ้น และผลการแนะนำตามที่คาดหวัง 2 ชิ้น จะถือว่าการแนะนำมีความแม่นยำน้อย แม้ว่าในจำนวนผลการแนะนำที่ไม่คาดหวัง จะมีวัตถุที่เป็นประโยชน์ต่อผู้ใช้งานจำนวนมาก หรือเทียบเท่ากับ 8 ชิ้นก็ตาม

2.2.3.2 อัลฟาเอ็นดีซีจี

โดยทั่วไปแล้ว ตัวชี้วัดหนึ่งที่ได้รับความนิยมในการประเมินผลประสิทธิภาพอัลกอริทึมที่ใช้ในการค้นคืนเอกสารและระบบแนะนำ คือ เอ็นดีซีจี (Normalized Discounted Cumulative Gain: nDCG) ซึ่งจะพิจารณาจากลำดับรายการแนะนำที่ได้จากอัลกอริทึม คำนวณหาค่าเกน (Gain) ของวัตถุแนะนำในแต่ละลำดับจากสมการ f(i) ที่กำหนดไว้ แล้วลดสัดส่วนของค่าเกนตามตำแหน่งของวัตถุนั้น ๆ ในรายการแนะนำ แล้วเทียบกับค่าที่คำนวณได้จากรายการแนะนำในอุดมคติ (Ideal list) การคำนวณค่าเอ็นดีซีจีของการเรียงลำดับ K ลำดับ อธิบายได้จากสมการที่ 7

$$nDCG_K = \frac{DCG_K}{IDCG_K} \qquad (7)$$

โดยที่ค่า DCG_K คือค่า Discounted Cumulative Gain ณ ตำแหน่งที่ K ซึ่งคำนวณได้จากสมการที่ 8 ในขณะที่ $IDCG_K$ คำนวณจากรายการแนะนำในอุดมคติที่ให้ DCG_K ที่มากที่สุดที่เป็นไปได้

$$DCG_K = \sum_{i=1}^{K} \frac{G(i)}{\log_2(i+1)} \qquad (8)$$

ในการคำนวณค่าเอ็นดีซีจี ทั่วไปแล้วสมการ G(i) นิยมใช้เป็น $2^{rel_i}-1$ โดยค่า rel_i จะเป็นค่าลำดับความเกี่ยวข้องของวัตถุแนะนำ (graded relevance) ณ ตำแหน่ง i ที่กำหนดเป็นช่วงได้ หรืออาจค่าเป็น 0 กับ 1 แทนการเลือกหรือไม่เลือกวัตถุนั้นในรายการแนะนำ

อย่างไรก็ตาม G(i) ที่ใช้ค่าลำดับความเกี่ยวข้องของวัตถุแนะนำนั้น ไม่สามารถใช้วัดความหลากหลายของผลการแนะนำได้ ในงาน [14] ได้ปรับปรุงวิธีการวัดผลแบบเอ็นดีซีจี โดยการปรับให้ค่า $G(i)$ มีการพิจารณาในส่วนของความครอบคลุมหมวดหมู่ของผลการแนะนำด้วย ดังสมการที่ 9

$$G(i) = \sum_{j=1}^{m} J(d_i, j)\,(1 - \alpha)^{r_{j,\ i-1}} \qquad (9)$$

โดยที่ *m* คือจำนวนหมวดหมู่ทั้งหมด $J(d_i, j)$ หมายถึงวัตถุ d ลำดับที่ i ที่เป็นส่วนหนึ่งของหมวดหมู่ j และ $r_{j,\ i-1}$ คำนวณได้จากสมการที่ 10

$$r_{j,\ i-1} = \sum_{l=1}^{i-1} J(d_l, j) \qquad (10)$$

นอกจากสมการดังกล่าว สิ่งที่จำเป็นต้องกำหนดให้ชัดเจนสำหรับการคำนวณค่าเอ็นดีซีจี คือ การหารายการแนะนำที่มีการเรียงลำดับที่ให้ค่า DCG_K สูงสุด ซึ่งในงานวิจัยที่เสนอวิธีการอัลฟาเอ็นดีซีจี ได้เสนอวิธีการกำหนดรายการในอุดมคติดังกล่าวมาสองวิธีด้วยกัน ได้แก่

1. การเรียงลำดับตามผลรวมจำนวนหมวดหมู่ที่วัตถุนี้ครอบคลุม กรณีที่มีความครอบคลุมเท่ากัน ให้เรียงตามลำดับก่อนหลังปกติ
2. การเรียงลำดับตามผลรวมจำนวนหมวดหมู่ที่วัตถุนี้ครอบคลุม กรณีที่มีความครอบคลุมเท่ากัน ให้เลือกวัตถุที่อยู่ในหมวดหมู่ที่ไม่เคยปรากฏมาก่อน หรือวัตถุที่ทำให้ค่าเฉลี่ยความครอบคลุมหมวดหมู่พอ ๆ กัน

ซึ่งในงานวิจัยดังกล่าวได้เลือกใช้แบบที่สอง เนื่องจากทำให้ผลลัพธ์ในแต่ละลำดับ มีการพิจารณาถึงความแปลกใหม่ร่วมด้วย

ตัวอย่างการคำนวณค่าเอ็นดีซีจีจากงานวิจัยข้างต้น [14]

ตารางที่ 2.7 ตัวอย่างเอกสารและหัวข้อที่เกี่ยวข้องของเอกสาร

ชื่อเอกสาร	หมายเลขหัวข้อของเอกสาร						รวม
	85.1	85.2	85.3	85.4	85.5	85.6	
a. Carnival Re-Enters Norway Bidding		X		X			2
b. NORWEGIAN CRUISE LINE SAYS...		X					1
c. Carnival, Star Increase NCL Stake		X					1
d. Carnival, Star Solidify Control							0
e. HOUSTON CRUISE INDUSTRY GETS...	X					X	2
f. TRAVELERS WIN IN CRUISE TUG-OF-WAR	X						1
g. ARMCHAIR QUARTERBACKS NEED...			X				1
h. EUROPE, CHRISTMAS ON SALE	X						1
i. TRAVEL DEALS AND DISCOUNTS							0
j. HAVE IT YOUR WAY ON THIS SHIP							0

จากตารางที่ 2.7 ซึ่งเป็นตัวอย่างเอกสารในรายการค้นหาในงานวิจัย ซึ่งเรียงลำดับผลการค้นหาดังตารางคือ a-b-c-d-e-f-g-h ในงานวิจัยได้เลือกวิธีการสร้างรายการค้นหา (หรือแนะนำ) ใน

อุดมคติ ด้วยวิธีที่ 2 จาก 2 วิธีข้างต้น คือคำนึงถึงความแปลกใหม่ร่วมด้วย ผลการค้นหาในอุดมคติจึงเป็น a-e-g-b-f-c-h-i-j จากสมการที่ 9 สามารถคำนวณค่า G จะได้ค่า ณ ที่ k ใด ๆ เป็นดังนี้

$$G = < 2, \frac{1}{2}, \frac{1}{4}, 0, 2, \frac{1}{2}, 1, \frac{1}{4}, ... >$$

โดยที่ G คือเซตของค่า G(k) โดยที่สมาชิกแต่ละตัว คือค่า G ที่ตำแหน่ง k ใด ๆ เช่น G(1) = 2, G(3) = 0.25 เป็นต้น ซึ่งมีวิธีการคำนวณดังนี้

ตัวอย่างที่ k = 3 จากข้อมูล จำนวนหัวข้อทั้งหมดคือ 6 เพราะฉะนั้น m = 6

$$G(3) = \sum_{j=1}^{6} J(d_3, j)(1 - \alpha)^{r_{j,\ 3-1}}$$

ซึ่ง d_3 ในกรณีนี้คือ c. Carnival, Star Increase NCL Stake ลำดับต่อไปต้องทำการหาค่า $J(d_3, j)$ และ $r_{j,\ 3-1}$ซึ่งจากสมการจะพบว่า ค่าหลังคือผลรวมจำนวนเอกสารที่มีหัวข้อ j นี้มาแล้วก่อนหน้าข้อมูลจะได้ว่าในตัวอย่างนี้คิดแค่ j เท่ากับ 2 คือหัวข้อที่ 85.2 ก็พอ เนื่องจากเอกสารนี้มีปรากฏแค่หัวข้อ 85.2 เพราะฉะนั้น จะได้

$$J(d_3, 2) = 1$$

(เนื่องจาก เอกสารที่ 3 นี้เกี่ยวข้องเพียงหัวข้อ 85.2)

$$r_{2,\ 3-1} = 2$$

(เนื่องจาก หัวข้อ 85.2 ปรากฏในเอกสารก่อนหน้ามาแล้ว 2 เอกสารด้วยกัน)

จากค่าที่คำนวณได้ทั้งสอง และค่า α = 0.5 ทำให้ G(3) คำนวณได้เท่ากับ

$$G(3) = \frac{1}{4}$$

เมื่อนำไปหาค่า DCG ด้วยสมการที่ 8 ของทุกค่า K จะได้เซตของค่า DCG ที่ K ต่าง ๆ เป็น

$$DCG = <2, 2.315, 2.440, ...>$$

เช่นเดียวกันกับ ค่า IDCG ใช้วิธีคำนวณแบบเดียวกันกับ DCG เพียงเปลี่ยนรายการค้นหาจาก a-b-c-d-e-f-g-h เป็น a-e-g-b-f-c-h-i-j ทำให้เซตของค่า IG (Ideal Gain) ที่ค่า K ต่าง ๆ กัน เป็น

$$IG = < 2, 2, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, ... >$$

เมื่อนำไปหาค่า IDCG ด้วยสมการที่ () ของทุกค่า K จะได้เซตของค่า IDCG ที่ K ต่าง ๆ เป็น

IDCG = <2, 3.262, 3.762, ...>

สุดท้ายจึงทำค่า DCG และ IDCG มาหา nDCG ด้วยสมการ () ได้เซตของค่า nDCG ที่ K ต่าง ๆ เป็น

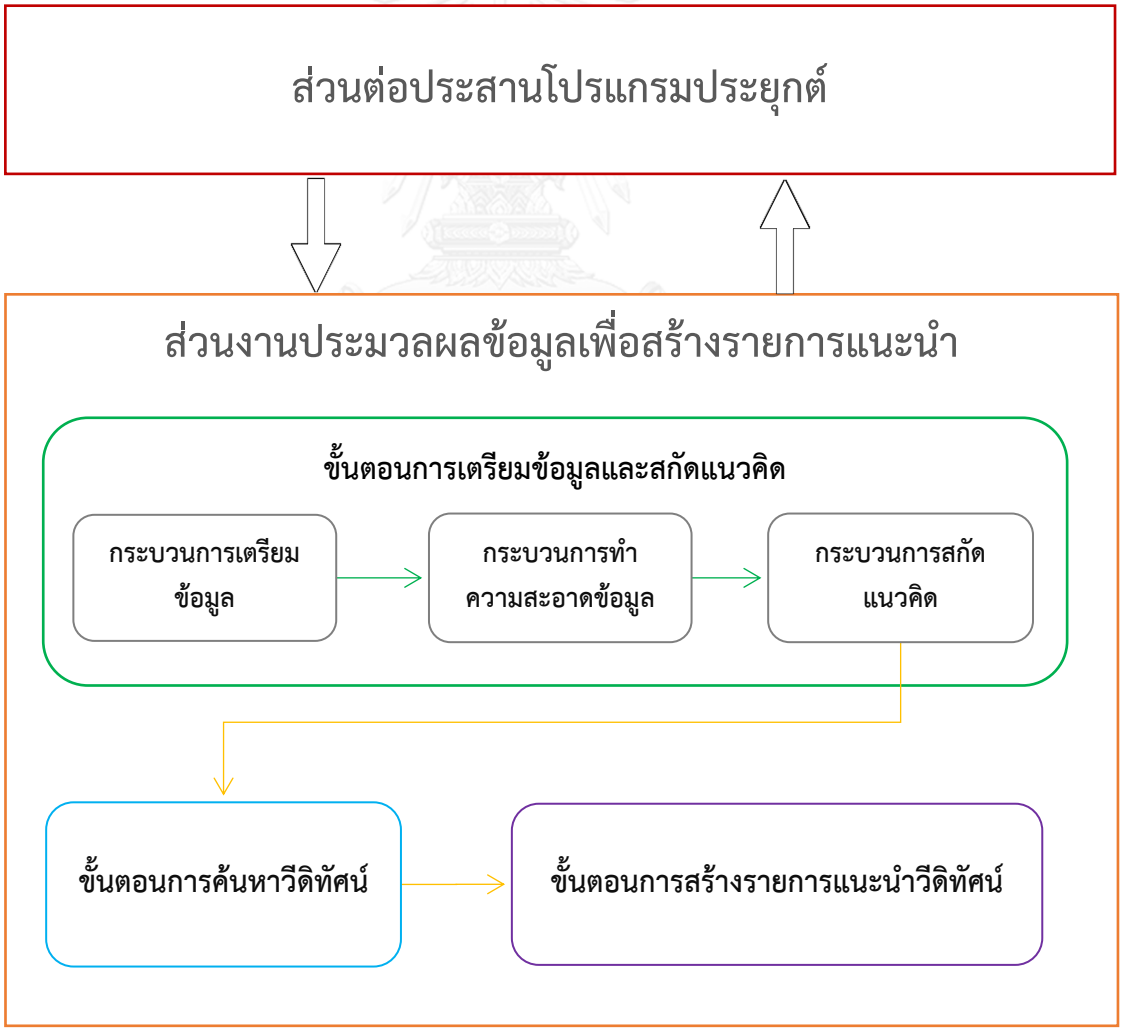
nDCG = <1, 0.710, 0.649, ...>

โดยที่ nDCG(i) คือค่า nDCG เมื่อทำการคำนวณจนถึงรายค้นหาลำดับที่ i เช่น nDCG(1) = 1 และ nDCG(3) = 0.649

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

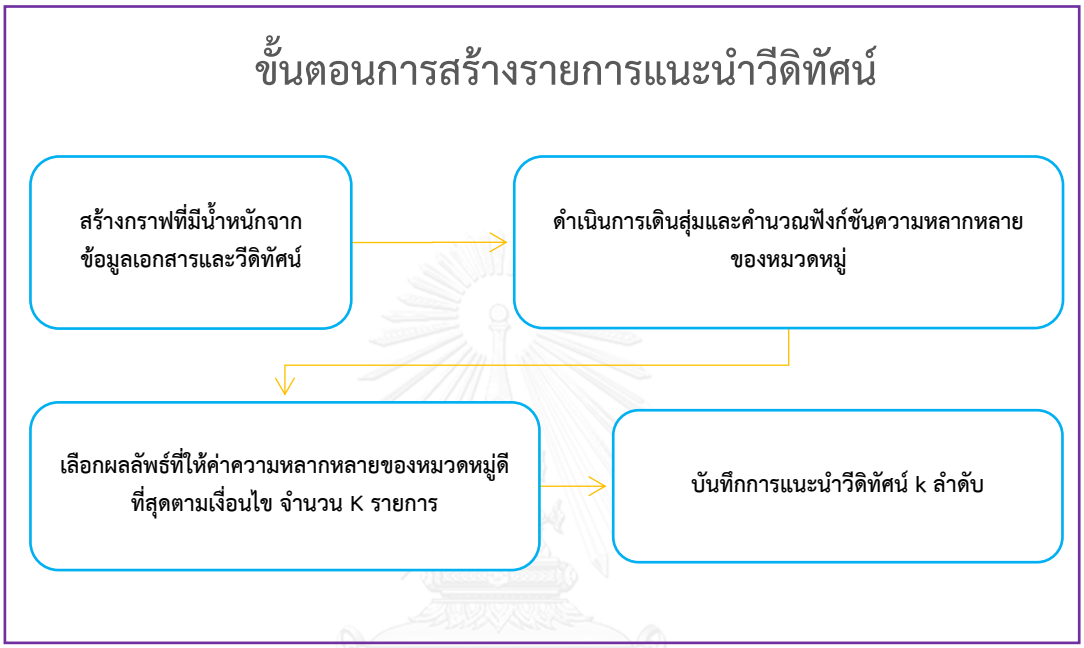
บทที่ 3
แนวคิดในการดำเนินวิจัย

แนวคิดในการดำเนินวิจัยนั้น ประกอบไปด้วย 2 ส่วน คือ ส่วนของการสร้างระบบแนะนำวีดิทัศน์ และส่วนของการเรียงลำดับผลลัพธ์ใหม่ ซึ่งในส่วนแรกจะแบ่งออกเป็น 2 ส่วนย่อย คือส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) สำหรับสื่อสารและส่งข้อมูลระหว่างระบบแนะนำวีดิทัศน์และระบบจัดการเรียนการสอน ส่วนที่สอง จะเป็นส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ ซึ่งจะแบ่งกระบวนการทำงานออกเป็น 3 ขั้นตอนได้แก่ ขั้นตอนการเตรียมข้อมูลและสกัดแนวคิด (Data Preparation and Concept Extraction), ขั้นตอนการค้นหาวีดิทัศน์ (Video Searching), และขั้นตอนการสร้างรายการแนะนำวีดิทัศน์ (Video Recommendation Creating) ดังแสดงในภาพที่ 3.1



ภาพที่ 3.1 โครงสร้างระบบของส่วนการสร้างระบบแนะนำวีดิทัศน์

ในส่วนของการเรียงลำดับผลลัพธ์ใหม่ จะเป็นการปรับปรุงวิธีการในขั้นตอนการสร้างรายการแนะนำวีดิทัศน์ โดยจะนำผลลัพธ์เดิมจากขั้นตอนนี้ มาทำการจัดลำดับผลลัพธ์ใหม่โดยใช้วิธีเดินสุ่มแบบเริ่มต้นใหม่ร่วมกับฟังก์ชันความหลากหลายของหมวดหมู่ ดำเนินการเช่นนี้วนไปจนกระทั่งผลการสุ่มเข้าสู่สภาวะคงที่ แล้วจึงเลือกผลการจัดลำดับการแนะนำใหม่ที่ให้ค่าฟังก์ชันความหลากหลายของหมวดหมู่มากที่สุด k ลำดับ เพื่อใช้แนะนำให้แก่ผู้เรียน ดังภาพที่ 3.2



ภาพที่ 3.2 ขั้นตอนการสร้างรายการแนะนำวีดิทัศน์ด้วยการเดินสุ่มร่วมกับฟังก์ชันความหลากหลายของหมวดหมู่

Chulalongkorn University

3.1 การสร้างระบบแนะนำวิดิทัศน์

3.1.1 ส่วนต่อประสานโปรแกรมประยุกต์

ภาพที่ 3.3 แนวคิดของส่วนต่อประสานโปรแกรมประยุกต์

3.1.1.1 ส่วนรับข้อมูลเอกสารจากระบบจัดการเรียนการสอน

ส่วนรับข้อมูลเอกสารจากระบบจัดการเรียนการสอน ดังแนวคิดในภาพที่ 3.3 เป็นส่วนที่ระบบจัดการเรียนการสอนใช้ส่งข้อมูลเอกสารที่ผู้สอนทำการอัปโหลดมายังระบบแนะนำวิดิทัศน์ ซึ่งในขั้นตอนนี้ ทางฝั่งระบบแนะนำจะบันทึกข้อมูลดังกล่าวไว้ เพื่อใช้ในส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ โดยข้อมูลเหล่านี้ จะแสดงในตารางที่ 3.1

ตารางที่ 3.1 ข้อมูลของเอกสารที่มายคอร์สวิลล์ต้องส่งให้ระบบแนะนำวิดิทัศน์

สดมภ์	ชนิด	คำอธิบาย	ตัวอย่าง
id	int(10)	ไอดีของเอกสาร	11,12,13
cv_file_path	longtext	ยูอาร์แอลที่เก็บเอกสาร	http://www...../lesson5.pdf
course_id	int(10)	ไอดีของวิชาเรียนของเอกสารที่พิจารณา	1
document_title	text	ชื่อของเอกสาร	Lesson 5: Divide and Conquer
course_title	text	ชื่อของวิชาเรียนของเอกสารที่พิจารณา	Computer Algorithms
document_desc	longtext	คำอธิบายของเอกสาร	In this lesson, we introduce...

3.1.1.2 ส่วนสร้างข้อมูลรายการแนะนำวีดิทัศน์

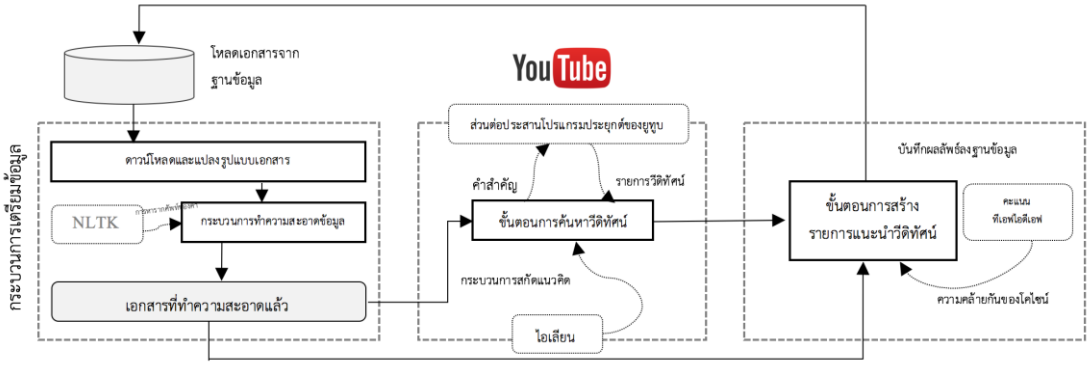
ส่วนสร้างข้อมูลรายการแนะนำวีดิทัศน์จะรับข้อมูลไอดีของเอกสาร จากระบบจัดการเรียนการสอนเมื่อมีผู้สอนหรือผู้เรียนเปิดหน้าเอกสารเพื่ออ่านหรือทำการดาวน์โหลดเอกสาร หลังจากนั้นระบบแนะนำจะคืนค่ารายการแนะนำวีดิทัศน์ที่ประกอบไปด้วยข้อมูลดังตารางที่ 3.2

ตารางที่ 3.2 ข้อมูลของวีดิทัศน์แต่ละรายการที่ระบบแนะนำวีดิทัศน์ส่งให้มายคอร์สวิลล์

สดมภ์	ชนิด	คำอธิบาย	ตัวอย่าง
title	text	ไอดีของเอกสาร	Divide And Conquer....
description	text	ยูอาร์แอลที่เก็บเอกสาร	This video is part of....
URL	text	ไอดีของวิชาเรียนของเอกสารที่พิจารณา	https://www.youtube...
channel title	text	ชื่อของเอกสาร	Udacity
thumbnails	text	ชื่อของวิชาเรียนของเอกสารที่พิจารณา	https://i.ytimg.com/vi/Ftm2uv7..
tags	array[text]	คำอธิบายของเอกสาร	[algorithms, divide, conquer]

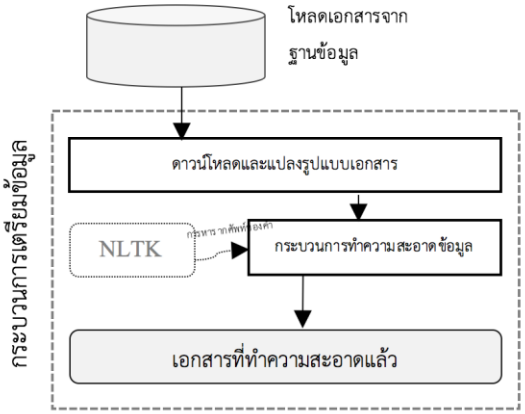
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.1.3 ส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ



ภาพที่ 3.4 โครงสร้างโดยรวมของส่วนงานประมวลผลข้อมูลเพื่อสร้างรายการแนะนำ

3.1.3.1 ขั้นตอนการเตรียมข้อมูลและสกัดแนวคิด



ภาพที่ 3.5 การทำงานของขั้นตอนกระบวนการเตรียมข้อมูลและการทำความสะอาดข้อมูล

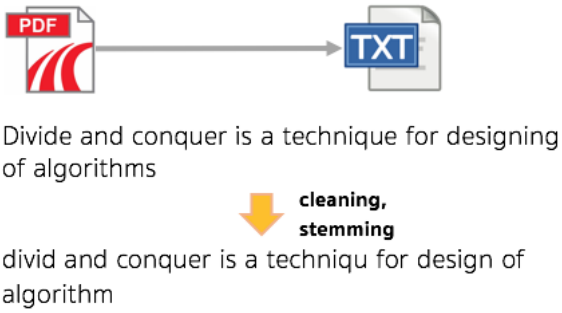
1. กระบวนการเตรียมข้อมูล

กระบวนการเตรียมข้อมูล (Data Preparation) จะทำการดาวน์โหลดและแปลงรูปแบบเอกสาร โดยรวบรวมเอกสารทั้งหมดที่ยังไม่เคยทำการสร้างรายการแนะนำ และใช้ข้อมูลยูอาร์แอลที่ได้จากการรับข้อมูลในส่วนต่อประสานโปรแกรมประยุกต์เพื่อดาวน์โหลดเอกสารเหล่านั้นมา หลังจากนั้น กระบวนการนี้จะทำการแปลงรูปแบบเอกสารข้อมูลจากรูปแบบพีดีเอฟ ให้เป็นรูปแบบข้อความแล้วจึงบันทึกเอกสารรูปแบบข้อความไว้สำหรับใช้ในขั้นตอนต่อไป

2. กระบวนการทำความสะอาดข้อมูล

กระบวนการทำความสะอาดข้อมูล (Data Cleansing) จะใช้เนื้อหาในเอกสารจากกระบวนการก่อนหน้า เพื่อกรองข้อความให้เหลือเพียงรูปแบบที่ประกอบไปด้วยตัวอักษร

ภาษาอังกฤษเท่านั้น (a-z) อย่างไรก็ตาม พบว่าคำที่ปรากฏในเอกสารนั้น มีบางคำที่มีความหมายในทำนองเดียวกันแต่มีรูปแบบการเขียนที่ไม่ต่างกันมาก เช่น คำว่า “student” กับ “students” หรือคำว่า “fly” กับ “flies” หรือแม้กระทั่งคำที่มีประเภทและหน้าที่ของคำที่ต่างกัน แต่มีความหมายในลักษณะเดียวกัน เช่นคำว่า “satisfy” กับ “satisfaction” จากเหตุผลดังกล่าวจึงจำเป็นที่จะต้องมีการหารากศัพท์ของคำ (Stemming Word) ซึ่งในงานวิจัยนี้ ได้ใช้เครื่องมือที่มีชื่อว่า Natural Language Processing Toolkit (NLTK) [41] โดยเครื่องมือนี้จะทำการค้นหารากศัพท์ของแต่ละคำ ซึ่งผู้วิจัยได้นำคำศัพท์เหล่านั้นไปแทนที่คำต้นฉบับดังตัวอย่างในภาพที่ 3.6



ภาพที่ 3.6 ตัวอย่างการทำความสะอาดข้อมูล

3. กระบวนการสกัดแนวคิด

กระบวนการสกัดแนวคิด (Concept Extraction) จะใช้เพื่อหาแนวคิดสำคัญของเอกสาร ในที่นี้คือคำสำคัญ (Keywords) ในเอกสาร โดยงานวิจัยนี้ได้ใช้เครื่องสำหรับสกัดแนวคิดของเอกสารที่มีชื่อว่าไอเลียน (Aylien) ซึ่งมีความสามารถในการดึงคำสำคัญของเอกสารออกมาเพื่อใช้เป็นคำสำคัญด้วยการใช้กระบวนการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ร่วมกับการเรียนรู้ด้วยเครื่อง (Machine Learning) หลายวิธีร่วมกันโดยใช้ชุดข้อมูลการเรียนรู้จาก DBPedia ซึ่งในส่วนของวิธีการที่ชัดเจนในการดำเนินการไม่ได้ระบุไว้ แต่มีงานวิจัยที่วัดประสิทธิภาพว่าเครื่องมือไอเลียนให้ผลค่าความแม่นยำ (Precision) ค่ารีคอล (Recall) และเอฟเมเชอร์ (F-measure) ดีที่สุดในกลุ่มเครื่องมือสำหรับการสกัดแนวคิด [42] ตัวอย่างการสกัดแนวคิดเอกสารของ Hemostasis and Abnormality (Competency skills) ในวิชา Clinical Hematology and Systemic Infection ดังภาพที่ 3.7

คู่มือปฏิบัติการ

เรื่อง การทดสอบ Whole blood clotting time, clot retraction time และ clot lysis time

1. หลักการ

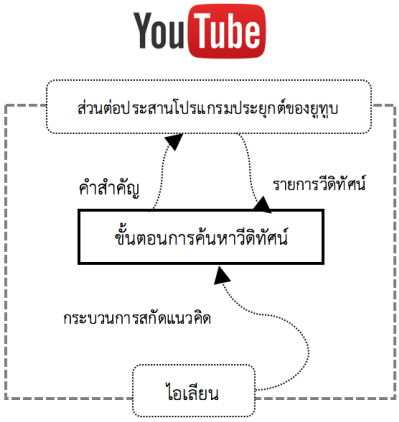
Whole blood clotting time (WBCT) หรือ Venous clotting time (VCT) เป็นการวัดระยะเวลาตั้งแต่เลือดที่เจาะจากหลอดเลือดดำ กลายเป็นลิ่มเลือด (fibrin clot) เลือดดำถูกกระตุ้นให้เกิดเป็นลิ่มเลือดเมื่อสัมผัสกับผิวขรุขระของหลอดทดลองที่เป็นแก้ว (Factor XII) การเกิดลิ่มเลือดนี้เกิดโดยการทำงานของปัจจัยแข็งตัวของเลือดในส่วนของ intrinsic pathway และ common pathway ได้แก่ ปัจจัยการแข็งตัวของเลือดที่ (Factor) XII, XI, IX, VIII, X, V, II และ I

เนื่องจากการทดสอบนี้มีความไวต่ำ (low sensitivity) ดังนั้นในผู้ป่วยที่มีเกล็ดเลือดผิดปกติ (platelet dysfunction) หรือมีภาวะพร่องปัจจัยการแข็งตัวของเลือดชนิดไม่รุนแรง (mild coagulation defect) จะมีค่า

ภาพที่ 3.7 ตัวอย่างเอกสารในระบบจัดการเรียนการสอนมายคอร์สวิลล์

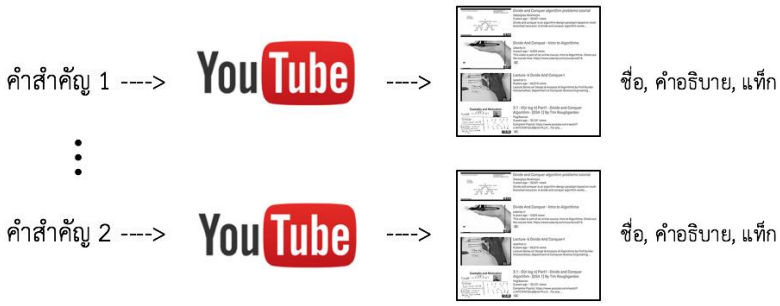
สามารถสกัดคำสำคัญด้วยเครื่องมือไอเลียนได้คำว่า world health organization, wbct, platelet, bloodclot, adp, placenta, thyroidectomy, fibrin, tissue factor, fibrinogen

3.1.3.2 ขั้นตอนการค้นหาวีดิทัศน์



ภาพที่ 3.8 การทำงานของกระบวนการสกัดแนวคิดและขั้นตอนการค้นหาวีดิทัศน์

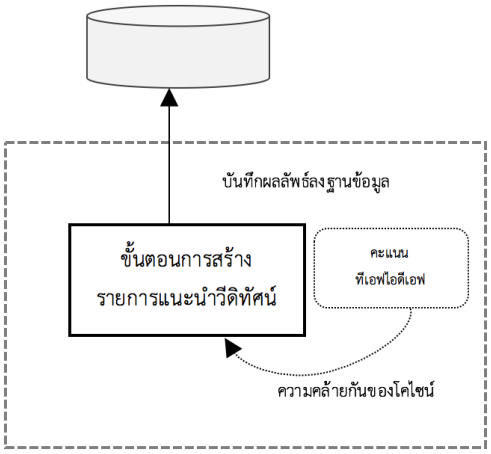
ขั้นตอนการค้นหาวีดิทัศน์นั้น จะใช้คำสำคัญที่สกัดได้จากเอกสารจากกระบวนการสกัดแนวคิดดังแสดงในภาพที่ 3.8 แล้วทำการส่งคำสำคัญเหล่านี้ไปยังส่วนต่อประสานโปรแกรมประยุกต์ของยูทูบเพื่อทำการค้นหาวีดิทัศน์ตามคำสำคัญเหล่านั้น ซึ่งคำสำคัญที่ส่งไปจะรับข้อมูลวีดิทัศน์มา 50 วีดิทัศน์ต่อคำสำคัญ แล้วเก็บเมตาดาตาของวีดิทัศน์เหล่านั้นไว้เพื่อนำมาใช้ในการสร้างรายการแนะนำ ดังตัวอย่างในภาพที่ 3.9



ภาพที่ 3.9 การส่งคำสำคัญไปยังยูทูบเพื่อรวบรวมข้อมูลวีดิทัศน์

3.1.3.3 ขั้นตอนการสร้างรายการแนะนำวีดิทัศน์

ในขั้นตอนการสร้างรายการแนะนำวีดิทัศน์นั้น จะใช้แนวคิดของการสร้างระบบแนะนำโดยพิจารณาจากเนื้อหา เพื่อทำการกรองรายการวีดิทัศน์สำหรับแนะนำให้เหลือเพียงรายการที่มีความเกี่ยวข้องกับเอกสารในจำนวนที่เหมาะสม และเพียงพอสำหรับการสร้างกราฟที่มีน้ำหนัก เพื่อใช้ในส่วนของการเรียงลำดับผลใหม่ (ภาพที่ 3.10)



ภาพที่ 3.10 โครงสร้างการสร้างรายการแนะนำวีดิทัศน์

การกรองรายการวีดิทัศน์ในงานวิจัยนี้จะใช้ลักษณะเด่นที่เป็นข้อความที่แสดงถึงเนื้อหาของเอกสารและวีดิทัศน์ มาคำนวณหาคะแนนความเหมือนกัน (Matching Score) ระหว่างวัตถุทั้งสอง ซึ่งการสกัดคะแนนความเหมือนกันของลักษณะเด่นที่ปรากฏของคู่เอกสารและวีดิทัศน์นั้น มีตัวแปรที่เกี่ยวข้องนิยามไว้ในตารางที่ 3.3

ตารางที่ 3.3 รายละเอียดตัวแปรที่เกี่ยวข้องในการสกัดลักษณะเด่นระหว่างเอกสารและวีดิทัศน์

ตัวแปร	คำอธิบาย	ตัวอย่างข้อมูล
T_d	เซตของคำที่ปรากฏในชื่อของเอกสาร	writing, introduction, and, Literature, Reviews
T_v	เซตของคำที่ปรากฏในชื่อของวีดิทัศน์	writing, introduction
K_d	เซตของคำสำคัญที่ปรากฏในชื่อและเนื้อหาของเอกสาร	introduction, literature, research method, related theory
K_v	เซตของคำสำคัญที่ปรากฏในชื่อและเนื้อหาของวีดิทัศน์	introduction, standford, method, theory
W_d	เซตของคำที่ไม่ซ้ำกันทั้งหมดที่ปรากฏในชื่อและเนื้อหาของเอกสาร	how, to, write, an, good, in, research, paper, introduction
W_v	เซตของคำที่ไม่ซ้ำกันทั้งหมดที่ปรากฏในชื่อและเนื้อหาของวีดิทัศน์	introduction, in, research, paper, standford
d	เอกสารใด ๆ ที่พิจารณา	เอกสารเรื่อง Writing Introduction and Literature Reviews
v	วีดิทัศน์ใด ๆ ที่พิจารณา	-
V	เซตทีเอฟ-ไอดีเอฟของลักษณะเด่นของวีดิทัศน์	-
D	เซตทีเอฟ-ไอดีเอฟของลักษณะเด่นของเอกสาร	-

ซึ่งการคำนวณคะแนนความเหมือนกันของลักษณะเด่นสามารถหาได้จากข้อมูลดังต่อไปนี้

1. ความตรงกันของคำสำคัญในเอกสารกับชื่อวีดิทัศน์ (Document Keyword and Video Title matching)

ความตรงกันของคำสำคัญกับชื่อวีดิทัศน์ เป็นการแสดงถึงความสัมพันธ์ของเอกสารและวีดิทัศน์ที่บ่งบอกว่า ชื่อของวีดิทัศน์นั้นเป็นส่วนหนึ่งของคำสำคัญในเอกสาร โดยจะใช้จำนวนสมาชิกของเซตที่นำส่วนของคำสำคัญที่ปรากฏในชื่อและเนื้อหาของเอกสาร K_d และคำที่ปรากฏในชื่อของวีดิทัศน์ T_v มาอินเทอร์เซกชันกัน แล้วหาอัตราส่วนเทียบกับจำนวนสมาชิกจากเซตของคำสำคัญในเอกสาร ดังสมการที่ 11

$$f_1 = \frac{|T_v \cap K_d|}{|K_d|} \qquad (11)$$

2. ความตรงกันของคำสำคัญของเอกสารกับวีดิทัศน์ (Document-Video Keyword matching)

ความตรงกันของคำสำคัญที่ปรากฏในเนื้อหาของเอกสารกับวีดิทัศน์ แสดงถึงความสัมพันธ์ของเอกสารและวีดิทัศน์ผ่านการเกิดร่วมกันของคำสำคัญของทั้งคู่ โดยจะใช้จำนวนคำในเซตที่สร้างจากการอินเทอร์เซกชันกันของคำสำคัญที่ปรากฏในชื่อและเนื้อหาของเอกสาร K_d และคำสำคัญที่ปรากฏในชื่อและเนื้อหาของวีดิทัศน์ K_v แล้วหาอัตราส่วนเทียบกับจำนวนสมาชิกจากเซตของคำสำคัญในเอกสาร ดังสมการที่ 12

$$f_2 = \frac{|K_v \cap K_d|}{|K_d|} \qquad (12)$$

3. ความตรงกันของชื่อหัวข้อ (Document Title Matching)

ความตรงกันของชื่อหัวข้อนั้น บ่งบอกถึงความเหมือนกันของชื่อเอกสารและชื่อของวีดิทัศน์ ซึ่งเป็นส่วนหนึ่งในการชี้วัดความเกี่ยวข้องกัน ในส่วนนี้จะใช้จำนวนที่เกิดจากการนำเซตคำที่ปรากฏในชื่อหัวข้อของทั้งเอกสารและวีดิทัศน์มาทำการอินเทอร์เซกชันอีกเช่นกัน แล้วหาอัตราส่วนเทียบกับจำนวนเซตของคำที่ปรากฏในชื่อของเอกสาร ดังสมการที่ 13

$$f_3 = \frac{|T_v \cap T_d|}{|T_d|} \qquad (13)$$

4. ความตรงกันของคำที่แตกต่างกันของเอกสารกับวีดิทัศน์ (Document-Video Distinct Word matching)

ความตรงกันของคำที่แตกต่างกันในเอกสารกับวีดิทัศน์ ส่วนนี้จะเป็นการค้นหาการเกิดร่วมกันของคำทั้งในส่วนของชื่อและคำอธิบายของวีดิทัศน์กับชื่อและเนื้อหาของเอกสาร ซึ่งจะแตกต่างจากการหาการเกิดร่วมกันของคำสำคัญ เนื่องจากในส่วนนี้จะใช้ทุกคำที่เกิดขึ้นไม่พิจารณาแค่คำสำคัญเท่านั้น โดยการคำนวณในส่วนนี้จะใช้เซตที่เกิดจากการอินเทอร์เซ็กชันกันระหว่างเซตของคำที่ไม่ซ้ำกันทั้งหมดที่ปรากฏในชื่อและเนื้อหาของเอกสาร และเซตของคำที่ไม่ซ้ำกันทั้งหมดที่ปรากฏในชื่อและเนื้อหาของวีดิทัศน์ แล้วนำมาเทียบกับจำนวนสมาชิกในส่วนของวีดิทัศน์แทน เพื่อหาอัตราส่วน ดังแสดงในสมการที่ 14

$$f_4 = \frac{|W_v \cap W_d|}{|W_v|} \qquad (14)$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สาเหตุที่ส่วนนี้หาเทียบกับข้อมูลจากวีดิทัศน์ เนื่องจากข้อมูลในส่วนนี้มีโอกาสที่จะได้ค่าน้อยมาก จากการที่จำนวนของคำที่เกิดในเอกสาร โดยทั่วไปมักมีมากกว่าคำในส่วนของวีดิทัศน์ ความคาดหวังในกรณีนี้คือ เซตของคำที่ปรากฏในเนื้อหาของวีดิทัศน์ทั้งหมด เป็นเซตย่อยของคำที่เกิดในเอกสาร ซึ่งถ้าหาเทียบกับข้อมูลของเอกสาร ก็อาจจะได้ค่าที่น้อยมาก แต่ถ้าเทียบข้อมูลของวีดิทัศน์ ก็จะได้ค่าเท่ากับ 1 ซึ่งแสดงว่าข้อมูลทั้งหมดในส่วนของวีดิทัศน์มีปรากฏอยู่ในเอกสารด้วย

5. ความคล้ายคลึงกันของเนื้อหาจากการคำนวณด้วยวิธีทีเอฟ-ไอดีเอฟ (Term Frequency-Inverse Document Frequency)

การคำนวณในส่วนนี้จะใช้การคำนวณค่าความคล้ายของทีเอฟ-ไอดีเอฟ จากแนวคิดของระบบแนะนำแบบพิจารณาเนื้อหา โดยอาศัยความคล้ายโคไซน์ หาความความคล้ายระหว่างลักษณะเด่นของเอกสารและลักษณะเด่นของวีดิทัศน์ ดังสมการที่ 15

$$f_5 = \text{Similarity}(D,V) = \cos(\theta) = \frac{D \cdot V}{\|D\|\|V\|} \qquad (15)$$

หากค่านี้ยิ่งมาก แสดงว่าเอกสารและวีดิทัศน์มีความคล้ายกันค่อนข้างมาก ซึ่งคะแนนส่วนนี้จะเป็นส่วนแสดงความสัมพันธ์ระหว่างวัตถุทั้งสอง

ซึ่งในส่วนของการกรองผลลัพธ์เบื้องต้นสำหรับการเรียงลำดับผลลัพธ์ใหม่ในขั้นตอนถัดไปนั้น จะใช้คะแนนที่คำนวณได้จากทั้ง 5 ฟังก์ชันนี้รวมกันแล้วเรียยงคะแนนดังกล่าวเพื่อหาลำดับความสัมพันธ์ที่เกี่ยวข้องกับเนื้อหาเบื้องต้นมาทั้งหมด 25 รายการต่อ 1 เอกสาร

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.2 การเรียงลำดับผลลัพธ์ใหม่

3.2.1 การสร้างกราฟที่มีน้ำหนักสำหรับการเดินสุ่ม

หลังจากที่ทำการคำนวณคะแนนดังกล่าวแล้ว จากจำนวนวีดิทัศน์ทั้งหมด N รายการ จะถูกคัดเลือกเหลือเพียง n รายการ ต่อ 1 เอกสาร เรียงตามคะแนนความสัมพันธ์ สำหรับการสร้างกราฟที่มีน้ำหนัก เพื่อจำกัดของเขตของวัตถุแนะนำให้มีเพียงวัตถุที่เกี่ยวข้องกับเอกสารสำหรับการเปลี่ยนลำดับเพื่อเพิ่มความหลากหลาย เพื่อป้องกันผลการแนะนำที่หลากหลายมาก แต่มีความเกี่ยวข้องกับเอกสารตั้งต้นน้อย

กำหนดให้โหนดแต่ละโหนดในกราฟคือวีดิทัศน์โดยที่เส้นเชื่อมของกราฟ ในที่นี้จะใช้คะแนนความคล้ายที่ชื่อความคล้ายกันของเนื้อหา คือคะแนนทีเอฟไอดีเอฟ ในการสร้างน้ำหนักของกราฟ เหตุผลที่เลือกคะแนนส่วนนี้ แทนที่จะใช้คะแนนความสัมพันธ์ที่คำนวณได้ก่อนหน้านี้มี 2 ประการ ประการแรกนั้น คะแนนทีเอฟไอดีเอฟ เป็นค่าบ่งชี้ความคล้ายคลึงของเนื้อหาจริง ๆ ในขณะที่ข้อมูลส่วนอื่นนั้น เป็นองค์ประกอบเสริมสำหรับการจำกัดของเขตของรายการวีดิทัศน์ในการแนะนำ ประการที่สอง เนื่องจากวิธีการที่จะนำมาเปรียบเทียบ ใช้เพียงคะแนนดีเอฟไอดีเอฟเท่านั้น ทางผู้วิจัยจึงตั้งใจที่จะวัดประสิทธิภาพด้านความหลากหลายด้วยข้อมูลลักษณะเดียวกัน

อย่างไรก็ตาม นอกจากต้องสร้างกราฟที่มีเส้นเชื่อมระหว่างวีดิทัศน์ด้วยกันแล้ว กราฟที่จะนำไปใช้ จำเป็นที่จะต้องเพิ่มส่วนของโหนดที่เป็นเอกสารตั้งต้นเข้าไปด้วย โดยใช้ค่าทีเอฟไอดีเอฟเช่นเดียวกัน ในการเชื่อมความสัมพันธ์ระหว่างเอกสารต้นทาง กราฟที่ได้จากเงื่อนไขดังกล่าว สามารถใช้เมทริกซ์เปลี่ยนสถานะ (Transition Matrix) เป็นตัวแทนเพื่ออธิบายได้ ซึ่งเมทริกซ์ดังกล่าวจะมีลักษณะเดียวกันกับภาพที่ 3.11

$$\widetilde{W} = \begin{bmatrix} 1.00, 0.46, 0.45, 0.50, 0.28, 0.15, 0.33, 0.27, 0.10, 0.19, 0.07 \\ 0.46, 1.00, 0.87, 0.72, 0.28, 0.15, 0.25, 0.37, 0.09, 0.13, 0.07 \\ 0.45, 0.87, 1.00, 0.67, 0.30, 0.16, 0.25, 0.36, 0.07, 0.08, 0.08 \\ 0.50, 0.72, 0.67, 1.00, 0.33, 0.15, 0.31, 0.44, 0.12, 0.15, 0.06 \\ 0.28, 0.28, 0.30, 0.33, 1.00, 0.10, 0.68, 0.18, 0.07, 0.05, 0.06 \\ 0.15, 0.15, 0.16, 0.15, 0.10, 1.00, 0.07, 0.09, 0.10, 0.14, 0.05 \\ 0.33, 0.25, 0.25, 0.31, 0.68, 0.07, 1.00, 0.17, 0.08, 0.09, 0.05 \\ 0.27, 0.37, 0.36, 0.44, 0.18, 0.09, 0.17, 1.00, 0.04, 0.05, 0.02 \\ 0.10, 0.09, 0.07, 0.12, 0.07, 0.10, 0.08, 0.04, 1.00, 0.16, 0.06 \\ 0.19, 0.13, 0.08, 0.15, 0.05, 0.14, 0.09, 0.05, 0.16, 1.00, 0.17 \\ 0.07, 0.07, 0.08, 0.06, 0.06, 0.05, 0.05, 0.02, 0.06, 0.17, 1.00 \end{bmatrix}$$

ภาพที่ 3.11 เมทริกซ์เปลี่ยนสถานะของกราฟมีน้ำหนักที่สร้างจากวีดิทัศน์และเอกสาร

จากรูปจะพบว่า ค่าของเมทริกซ์ตามเส้นทแยงมุมนั้น มีค่าเท่ากับ 1 หมายถึงค่าโคไซน์ซิมิลาริตี้ระหว่างโหนดตัวเองกับโหนดตัวเอง ซึ่งเป็นค่าที่ถูกต้อง แต่จากการทดลองดำเนินการเดินสุ่มแล้วพบว่าเมทริกซ์ $\vec{r}_i$ ในครั้งถัดมา มีค่าพุ่งขึ้นสูงอย่างรวดเร็ว จนไม่สามารถลู่เข้าได้ แม้ว่าจะเปลี่ยนค่า c ให้น้อยแล้วก็ตาม จากการวิเคราะห์สาเหตุเกิดจากการเดินสุ่มของผู้เดิน (Walker) ไม่ย่อมเปลี่ยนโหนด เนื่องจากความน่าจะเป็นในการเปลี่ยนสถานะ หรือน้ำหนักเส้นเชื่อมของกราฟนั้นมีค่าสูง ทำให้ผู้เดินวนอยู่ที่โหนดเดิม ทางผู้วิจัยจึงทำการปรับเมทริกซ์เปลี่ยนสถานะดังกล่าว โดยให้ความน่าจะเป็นของการเปลี่ยนสถานะจากโหนดใดวนกลับมาโหนดตัวเองเป็น 0 หรือกล่าวได้ว่า $w_{i,j} = 0$ ที่ $i = j$ ดังแสดงในภาพที่ 3.12

$$\widetilde{W} = \begin{bmatrix} 0.00, 0.46, 0.45, 0.50, 0.28, 0.15, 0.33, 0.27, 0.10, 0.19, 0.07 \\ 0.46, 0.00, 0.87, 0.72, 0.28, 0.15, 0.25, 0.37, 0.09, 0.13, 0.07 \\ 0.45, 0.87, 0.00, 0.67, 0.30, 0.16, 0.25, 0.36, 0.07, 0.08, 0.08 \\ 0.50, 0.72, 0.67, 0.00, 0.33, 0.15, 0.31, 0.44, 0.12, 0.15, 0.06 \\ 0.28, 0.28, 0.30, 0.33, 0.00, 0.10, 0.68, 0.18, 0.07, 0.05, 0.06 \\ 0.15, 0.15, 0.16, 0.15, 0.10, 0.00, 0.07, 0.09, 0.10, 0.14, 0.05 \\ 0.33, 0.25, 0.25, 0.31, 0.68, 0.07, 0.00, 0.17, 0.08, 0.09, 0.05 \\ 0.27, 0.37, 0.36, 0.44, 0.18, 0.09, 0.17, 0.00, 0.04, 0.05, 0.02 \\ 0.10, 0.09, 0.07, 0.12, 0.07, 0.10, 0.08, 0.04, 0.00, 0.16, 0.06 \\ 0.19, 0.13, 0.08, 0.15, 0.05, 0.14, 0.09, 0.05, 0.16, 0.00, 0.17 \\ 0.07, 0.07, 0.08, 0.06, 0.06, 0.05, 0.05, 0.02, 0.06, 0.17, 0.00 \end{bmatrix}$$

ภาพที่ 3.12 เมทริกซ์เปลี่ยนสถานะที่ปรับน้ำหนักบางโหนด

3.2.2 วิธีการเดินสุ่มและความหลากหลายของหมวดหมู่

ในส่วนของวิธีการเดินสุ่มและความหลากหลายของหมวดหมู่นั้น ขั้นตอนการดำเนินการมีลักษณะเช่นเดียวกับวิธีการเดินสุ่มโดยปกติ โดยเพิ่มส่วนของการคำนวณคะแนนความหลากหลายของหมวดหมู่จากรายการแนะนำที่ได้จากการสุ่มในแต่ละรอบ แล้วทำการเก็บรายการที่ให้ค่าของคะแนนความหลากหลายของหมวดหมู่ดีที่สุดไว้ รวมทั้งคำนวณค่า L2Norm เพื่อใช้ตรวจสอบการลู่เข้าทุกครั้ง การดำเนินการเหล่านี้จะกระทำไปจนกระทั่งการเดินสุ่มเข้าเงื่อนไขในการหยุด (ค่า L2Norm < ξ หรือทำไปจนถึงรอบที่ m) หลังจากนั้นอัลกอริทึมจะคืนค่ารายการที่เก็บไว้ในผู้ใช้งาน

ในส่วนของฟังก์ชันความหลากหลายของหมวดหมู่นั้นมีการปรับปรุงเพื่อให้ตรงกับการใช้งานในระบบแนะนำนี้ดังสมการที่ 16

$$D_c = \left(\frac{C_d \times (1 - m)}{C_d \times 1}\right) \times (C_n \times 1) \qquad (16)$$

ซึ่งจะสังเกตว่าเป็นสมการเดียวกันกับที่งานวิจัยที่เกี่ยวข้องได้ใช้ แต่สิ่งที่ปรับคือนิยามของตัวแปรแต่ละตัว เพื่อให้เข้ากับข้อมูลในงานวิจัยนี้ โดยความหมายใหม่ของตัวแปรแต่ละตัว แสดงในตารางที่ 3.4

ตารางที่ 3.4 รายละเอียดตัวแปรที่ใช้ในการคำนวณสมการฟังก์ชันความหลากหลายของหมวดหมู่

ตัวแปร	คำอธิบาย
V	เซตของวีดิทัศน์ทั้งหมด
N_V	จำนวนสมาชิกของ V
R	วีดิทัศน์ที่ได้รับการเลือกจากการเดินสุ่มสำหรับเอกสารใด ๆ
r_b	เวกเตอร์ขนาด $1 \times N_V$ โดยที่ $r_b(k) = 1$ ถ้า $k \in R$
C	เมทริกซ์ขนาด $N_V \times N_C$ C(i, j) = 1 ถ้าวีดิทัศน์ i เกิดจากการค้นหาด้วยคำสำคัญ j หรือ C(i, j)=0 ในทางตรงกันข้าม
m	e_j=1 ยกเว้นกรณีที่คำสำคัญ j ไม่ได้ถูกพิจารณาให้เป็นแนวคิดหลักของเอกสารจะให้เป็น 0
1	1(i)= 1 สำหรับทุกค่า i

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

<u>ขั้นตอนวิธีการดำเนินการเดินสุ่มและตัวอย่างข้อมูล</u>

ตัวอย่างเมทริกซ์ของวัตถุแนะนำ 10 รายการ จากขั้นตอนการสร้างกราฟมีน้ำหนัก

$$\widetilde{W} = \begin{bmatrix} 0.00, 0.46, 0.45, 0.50, 0.28, 0.15, 0.33, 0.27, 0.10, 0.19, 0.07 \\ 0.46, 0.00, 0.87, 0.72, 0.28, 0.15, 0.25, 0.37, 0.09, 0.13, 0.07 \\ 0.45, 0.87, 0.00, 0.67, 0.30, 0.16, 0.25, 0.36, 0.07, 0.08, 0.08 \\ 0.50, 0.72, 0.67, 0.00, 0.33, 0.15, 0.31, 0.44, 0.12, 0.15, 0.06 \\ 0.28, 0.28, 0.30, 0.33, 0.00, 0.10, 0.68, 0.18, 0.07, 0.05, 0.06 \\ 0.15, 0.15, 0.16, 0.15, 0.10, 0.00, 0.07, 0.09, 0.10, 0.14, 0.05 \\ 0.33, 0.25, 0.25, 0.31, 0.68, 0.07, 0.00, 0.17, 0.08, 0.09, 0.05 \\ 0.27, 0.37, 0.36, 0.44, 0.18, 0.09, 0.17, 0.00, 0.04, 0.05, 0.02 \\ 0.10, 0.09, 0.07, 0.12, 0.07, 0.10, 0.08, 0.04, 0.00, 0.16, 0.06 \\ 0.19, 0.13, 0.08, 0.15, 0.05, 0.14, 0.09, 0.05, 0.16, 0.00, 0.17 \\ 0.07, 0.07, 0.08, 0.06, 0.06, 0.05, 0.05, 0.02, 0.06, 0.17, 0.00 \end{bmatrix}$$

จากสมการ $\vec{r}_i[t] = c\tilde{W}\vec{r}_i[t-1]+(1-c)\vec{e}_i$ ให้เมทริกซ์ $\vec{e}_i$ มีค่าเป็น $\vec{e}_i$ = [1.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00] เนื่องจากต้องการหาความสัมพันธ์กับเอกสาร ซึ่งในเมทริกซ์เปลี่ยนสถานะเอกสารถูกใส่ไว้ในตำแหน่งแรก เมทริกซ์ $\vec{e}_i$ จึงถูกใส่ไว้ในตำแหน่งแรกด้วย

ทดสอบการเดินสุ่มรอบที่ 1 ที่ c = 0.5 ในรอบแรกนี้ เนื่องจากยังไม่มี $\vec{r}_i[t-1]$ จึงให้ $\vec{r}_i[t-1]$ เท่ากับ $\vec{e}_i$ ตามแนวทางการเดินสุ่มแบบเริ่มต้นใหม่ในงานวิจัยต่าง ๆ หลังจากการเดินสุ่มในรอบแรกจบลงจะได้ $\vec{r}_i[t]$ มีลักษณะเป็นดังนี้

$\vec{r}_i[t]$ = [0.500, 0.230, 0.221, 0.250, 0.139, 0.071, 0.162, 0.135, 0.048, 0.092, 0.035]

โดยที่แต่ละตำแหน่งใน $\vec{r}_i[t]$ จะเป็นคะแนนลำดับ (Ranking Score) ของการเดินสุ่ม ซึ่งแทนด้วย

$\vec{r}_i[t]$ = [0.500, video1, video2, video3, video4, video5, video6, video7, video8, video9, video10]

โดยที่ตำแหน่งแรก จะเป็นคะแนนลำดับของตัวเอกสารเอง ซึ่งในคำนวณค่าของฟังก์ชันความหลากหลายของหมวดหมู่นั้น จะทำการตัดส่วนที่เป็นเอกสารออก (0.500)

ในขั้นตอนของการคำนวณคะแนนความหลากหลายของหมวดหมู่นั้น สิ่งที่ต้องหาเพิ่มคือเมทริกซ์ R ที่เป็นรายการวีดิทัศน์ที่ถูกเลือก โดยในตัวอย่างนี้จะใช้ค่า K = 5 เพื่อพิจารณาวีดิทัศน์ที่มีคะแนนลำดับสูงสุด 5 อันดับแรก เพื่อเลือกเป็นรายการที่จะแนะนำ กำหนดให้

V = [video1, video2, video3, video4, video5, video6, video7, video8, video9, video10]

จากข้อกำหนดการเลือกรายการแนะนำจะได้

$\vec{r}_i[t]$ = [~~0.500~~, 0.230, 0.221, 0.250, 0.139, 0.071, 0.162, 0.135, 0.048, 0.092, 0.035]

ซึ่งรายการวีดิทัศน์ที่ถูกเลือกจะเป็น

R = [video2, video3, video4, video6]

ซึ่งส่งผลให้ ค่า r_b ที่พิจารณาจาก R มีค่าเป็น

r_b = [1, 1, 1, 1, 0, 1, 0, 0, 0, 0]

ให้เมทริกซ์ C เป็นเมทริกซ์ตัวแปรที่บ่งบอกว่า วีดิทัศน์นี้อยู่ในแนวคิดไหนบ้าง โดยใช้ที่มาของวีดิทัศน์ ว่าเกิดจากการค้นหาด้วยคำสำคัญใดบ้าง กำหนดให้เมทริกซ์ C เป็นดังนี้

$$C = \begin{bmatrix} 1, 1, 0, 0, 0 \\ 0, 1, 0, 1, 0 \\ 1, 0, 1, 0, 0 \\ 1, 0, 0, 0, 0 \\ 1, 0, 0, 0, 0 \\ 0, 1, 0, 1, 0 \\ 0, 0, 0, 1, 1 \\ 0, 0, 0, 0, 1 \\ 1, 1, 0, 0, 0 \\ 0, 1, 0, 1, 0 \end{bmatrix}$$

โดยหมวดหมู่ หรือกรณีนี้คือแนวคิดของเอกสารมีทั้ง หมด 5 หัวข้อ C_1, C_2, C_3, C_4, C_5 การกำหนดแนวคิดหลัก จะใช้วิธีการหาแนวคิดที่มีวิดิทัศน์ปรากฏอยู่มากที่สุด ในกรณีที่มีแนวคิดมีจำนวนวิดิทัศน์ปรากฏอยู่เท่ากันทุกแนวคิด จะถือว่าไม่มีแนวคิดหลักแทน จากตัวอย่างแนวคิดหลักคือ C_1, C_2 เนื่องจากแนวคิดทั้งสองนั้น มีวีดิทัศน์ปรากฏอยู่ 5 วีดิทัศน์เท่ากัน

จากข้อกำหนดและข้อมูลตัวอย่าง สามารถกำหนดค่าของตัวแปรอื่น ๆ ได้เป็นดังนี้

$$m = [1, 1, 0, 0, 0]$$

$$1 = [1, 1, 1, 1, 1]$$

$$1 - m = [0, 0, 1, 1, 1]$$

จากการที่ $C_d = r^T \times C$ แต่เนื่องจากมีการปรับมิติของเมทริกซ์ C จึงสามารถใช้ $C_d = r \times C$ ได้เลย ส่งผลให้สามารถคำนวณค่าของตัวแปร C_d และ C_nได้ดังนี้

$$C_d = [3, 3, 1, 2, 0]$$

$$C_n = [1, 1, 1, 1, 0]$$

จากการคำนวณด้วยตัวแปรที่หาได้จากการเดินสุ่มและข้อมูลเกี่ยวกับหมวดหมู่ของรายการวีดิทัศน์ จะสามารถคำนวณค่าความหลากหลายของหมวดหมู่ได้ดังต่อไปนี้

$$D_c = \left(\frac{C_d \times (1 - m)}{C_d \times 1} \right) * (C_n \times 1)$$

$$D_c = \frac{3}{9} * 4 = 1.33$$

แม้จะสามารถคำนวณคะแนนจากฟังก์ชันความหลากหลายของหมวดหมู่ได้ 1.33 ดังตัวอย่าง แต่ส่วนที่ยังต้องพิจารณาเพิ่มเติมคือ วิธีการใดที่จะใช้สำหรับเลือกค่าความหลากหลายของหมวดหมู่ที่

ดีที่สุด ซึ่งคะแนนความหลากหลายของหมวดหมู่ดีที่สุดนั้น ทางผู้วิจัยได้แบ่งออกเป็น 2 วิธีพิจารณาด้วยกัน

1. วิธีพิจารณาความหลากหลายของหมวดหมู่ดีที่สุดโดยไม่คำนึงถึงเงื่อนไขการลู่เข้า

วิธีนี้จะใช้ฟังก์ชันความหลากหลายของหมวดหมู่แบบดั้งเดิมในการคำนวณคะแนนโดยไม่สนใจการลู่เข้า โดยจะเลือกเก็บรายการแนะนำที่ให้คะแนนจากคำนวณด้วยฟังก์ชันความหลากหลายของหมวดหมู่ที่สูงที่สุด ($D_{c_{max}}$) ไม่ว่าผลลัพธ์ของการแนะนำที่การลู่เข้าเป็นเท่าใด ก็จะไม่มีผลต่อรายการแนะนำที่เก็บไว้แต่อย่างใด เมื่อจบการเดินสุ่ม อัลกอริทึมนี้จะคืนค่ารายการแนะนำที่เก็บไว้ ($\vec{r}_{D_{c_{max}}}$) ซึ่งข้อดีของวิธีนี้ก็คือ รายการที่เก็บไว้นั้นจะให้คะแนนความหลากหลายของหมวดหมู่สูงสุด แต่ก็มีข้อเสียที่ตามมาก็คือ รายการการดังกล่าวจะไม่สะท้อนถึงความสัมพันธ์กับการเดินสุ่มแต่อย่างใด

รหัสเทียมอัลกอริทึม RWR+CDF_I

อัลกอริทึมที่ 1 RWR+CDF_I

สิ่งที่จำเป็น: $\widetilde{W}$ และ C

for each t < iteration **do**
  $\vec{r}_i[t] = c\widetilde{W}\vec{r}_i[t-1] + (1-c)\vec{e}_i$
  find *distance* between $\vec{r}_i[t]$ and $\vec{r}_i[t-1]$
  if *distance* > ξ **then**
    calculate C_d, C_n, D_c
    if $D_{c_{max}} < D_c$ **then**
      $D_{c_{max}} = D_c$
      $\vec{r}_{D_{c_{max}}} = \vec{r}_i[t]$
  else stop iteration
end for
output: $\vec{r}_{D_{c_{max}}}$

2. วิธีพิจารณาความหลากหลายของหมวดหมู่ดีที่สุดโดยนำเงื่อนไขการลู่เข้ามาพิจารณาร่วมด้วย

แต่การเลือกพิจารณาคะแนนความหลากหลายของหมวดหมู่เพียงอย่างเดียวนั้น อาจส่งผลให้ผลลัพธ์การเดินสุ่มถูกละเลย เนื่องจากคะแนนความหลากหลายที่สูง อาจเกิดจากการเดินสุ่มที่ไม่ได้ใกล้เคียงการลู่เข้าหรือไม่อยู่ในสภาวะมั่นคงนั่นเอง ทำให้ค่าของ r ตามแนวคิดของการเดินสุ่มไม่ได้สะท้อนมาจากกราฟที่มีน้ำหนัก ที่คำนวณจากคะแนนความสัมพันธ์ระหว่างเอกสารและวีดิทัศน์ หรืออีกนัยหนึ่ง ผลลัพธ์ที่ให้ค่าคะแนนความหลากหลายของหมวดหมู่มาก อาจจะเป็นผลการแนะนำวีดิ

ทัศน์ที่ไม่เกี่ยวข้องกับเอกสารเท่าที่ควร ดังนั้น ทางผู้วิจัยจึงได้เพิ่มตัวแปรเข้าไปเพื่อให้ผลการเดินสุ่มถูกพิจารณาด้วย ดังสมการที่ 17

$$D_c' = \frac{1}{D_c} \left\| r_{t+1} - r_t \right\| \quad (17)$$

จะเห็นความเปลี่ยนแปลงของสมการ โดยการนำ D_c ไปเป็นตัวหารแทน และมีพจน์ของ $\left\| r_{t+1} - r_t \right\|$ คือการหาความแตกต่างของผลการเดินสุ่ม ครั้งปัจจุบันกับครั้งที่ผ่านมา ส่วนนี้เป็นการนำการลู่เข้ามาพิจารณาเพิ่มเติม เนื่องจาก ยิ่งค่าของ r_{t+1} และ r_t ใกล้เคียงกันมากเท่าใด ก็แสดงว่าการเดินสุ่มใกล้จะลู่เข้าสภาวะมั่นคงแล้วเท่านั้น จากการเปลี่ยนแปลงดังกล่าว แต่เดิมจะพิจารณาค่าความหลากหลายของหมวดหมู่ที่มาก แต่ด้วยสมการ D_c' จะพิจารณาผลการเดินสุ่มที่ให้ค่า D_c' น้อยแทน

รหัสเทียมอัลกอริทึม RWR+CDF_II

อัลกอริทึมที่ 2 RWR+CDF_II

สิ่งที่จำเป็น: $\widetilde{W}$ และ C

```
1:  for each t < iteration do
2:          r_i[t] = c W̃ r_i[t - 1] + (1 - c) e_i
3:      find distance between r_i[t] and r_i[t - 1]
4:      if distance > ξ then
5:          calculate C_d, C_n, D'_c
6:          if D'_cmin > D_c then
7:              D'_cmin = D_c
8:              r_D'cmin = r_i[t]
9:      else stop iteration
10: end for
11: output: r_D'cmin
```

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

3.3 เครื่องมือที่ใช้ในงานวิจัย

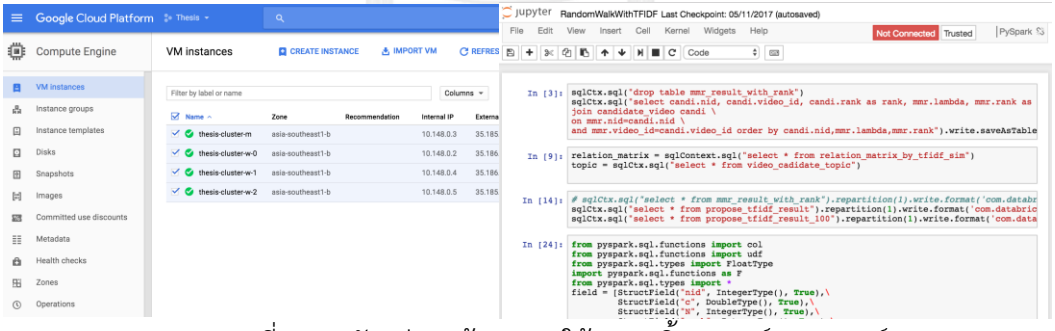
3.3.1 ภาษาที่ใช้ในการพัฒนา

ในส่วนของการสร้างระบบแนะนำวีดิทัศน์ในงานวิจัยนี้ ได้เลือกใช้ภาษาไพธอน ในการพัฒนา เนื่องจากภาษาไพธอนเป็นหนึ่งในภาษาที่ได้รับความนิยมในด้านของการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analysis) รวมทั้งไพธอนยังมีเฟรมเวิร์คสำหรับการสร้างเว็บไซต์สำหรับส่วนต่อประสานโปรแกรมประยุกต์ที่น่าสนใจ เช่น ดีจังโก (Django) ฟลาส์ค (Flask) เป็นต้น ซึ่งทางผู้วิจัยได้เลือกใช้ฟลาส์ค เนื่องจากฟลาส์คมีความกระทัดรัด ใช้สร้างระบบได้ง่าย [43] ในขณะเดียวกัน ก็มีความสามารถที่เพียงพอสำหรับงานวิจัยนี้

เช่นเดียวกันกับส่วนของการเรียงลำดับผลลัพธ์ใหม่นั้น ไพธอนยังคงเป็นภาษาหลักที่ใช้ในการพัฒนา โดยเพิ่มส่วนที่เกี่ยวข้องกับการคำนวณแบบขนาน และการทำแมปรีดิวที่ชื่อว่าสปาร์ค (SPARK) ซึ่งรุ่นที่ใช้เป็นรุ่นที่ 2.0.2

3.3.2 โครงสร้างพื้นฐานและเทคโนโลยีที่เกี่ยวข้อง

การเรียงลำดับผลลัพธ์ใหม่นั้น จะดำเนินการอยู่บนเทคโนโลยีที่ชื่อว่า Apache SPARK ซึ่งเป็นเครื่องมือสำหรับการดำเนินการบนข้อมูลขนาดใหญ่ โดยในงานนี้ ทางผู้วิจัยได้ดำเนินการทดลองในส่วนของการเรียงลำดับผลลัพธ์ใหม่บนกูเกิ้ลคลาวน์แพลตฟอร์ม (Google Cloud Platform) ที่ได้ติดตั้ง IPython Notebook สำหรับการเขียน SRARK program โดยใช้ภาษาไพธอน



ภาพที่ 3.13 ตัวอย่างหน้าจอการใช้งานกูเกิ้ลคลาวน์แพลตฟอร์ม

3.3.3 เครื่องมืออื่น ๆ

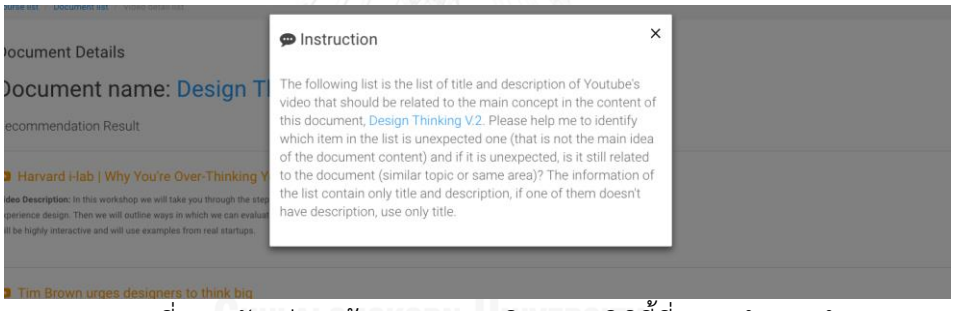
- เครื่องมือพัฒนาโปรแกรมภาษาไพธอนชื่อว่า PyCharm
- โปรแกรม Microsoft Excel 2016
- เว็บโฮสติ้งชื่อว่า Digital Ocean
- GitHub

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

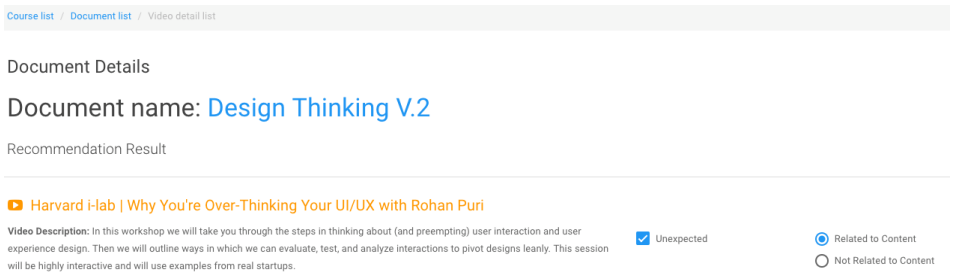
บทที่ 4
การทดลอง ผลการทดลอง และการวิเคราะห์ผลการทดลอง

4.1 ผลการทดลองจากค่าเซเรนดิพิตี้

ในการประเมินผลระบบแนะนำวีดิทัศน์เบื้องต้นด้วยค่าเซเรนดิพิตี้ เพื่อวัดว่าวิธีการสร้างระบบแนะนำวีดิทัศน์ที่ได้เสนอนั้น มีวีดิทัศน์ที่เป็นประโยชน์นอกเหนือไปจากเนื้อหาหลักของเอกสารหรือไม่ โดยการทดลองนี้ จะทำการทดสอบโดยการให้ผู้เรียนและผู้สอนเข้ามาประเมินผลการแนะนำโดยมีเงื่อนไขว่า ให้พิจารณาเลือกวีดิทัศน์ที่ไม่คาดหวังว่าจะพบในรายการแนะนำนี้ก่อน แล้วจึงพิจารณาต่อว่า รายการที่ไม่คาดหวังเหล่านั้นเป็นประโยชน์หรือไม่ ซึ่งความเป็นประโยชน์นี้จะเป็นส่วนของคะแนน $u(RS_i)$ ในสมการ (6) ซึ่งถ้าเป็นประโยชน์ ค่า $u(RS_i)$ จะเป็น 1 ถ้าไม่จะเป็น 0 โดยที่ N คือจำนวนของวีดิทัศน์ที่ไม่คาดหวังว่าจะพบในรายการแนะนำนี้ การพิจารณาดังกล่าว ผู้ประเมินจะใช้ข้อมูลเพียงชื่อและคำอธิบายของวีดิทัศน์เปรียบเทียบกับเอกสารในวิชาเรียนนั้น ๆ ผู้วิจัยได้สร้างหน้าเว็บไซต์ขึ้นเพื่อใช้ในการประเมินครั้งนี้ ดังภาพที่ 4.1 และภาพที่ 4.2



ภาพที่ 4.1 ตัวอย่างหน้าจอการประเมินเซเรนดิพิตี้ที่แสดงคำแนะนำ



ภาพที่ 4.2 ตัวอย่างหน้าจอการประเมินเซเรนดิพิตี้ เมื่อทำการประเมิน

ในการประเมินผลเซเรนดิพิตี้ครั้งนี้ ผู้วิจัยได้ทำการเปรียบเทียบกับรายการแนะนำที่เกิดจากการค้นหาด้วยชื่อเอกสารเพียงอย่างเดียว เพื่อพิสูจน์ว่าวิธีการสร้างระบบแนะนำวีดิทัศน์ที่ได้เสนอนั้นให้วีดิทัศน์ที่เป็นประโยชน์นอกเหนือไปจากการค้นหาด้วยชื่อเอกสาร ซึ่งจะให้ผลลัพธ์ที่เกี่ยวข้องกับชื่อเอกสารเพียงอย่างเดียว โดยในหน้าเว็บไซต์ที่ทำการประเมินนั้น จะทำการสุ่มคละผลลัพธ์โดยไม่

บอกผู้ประเมิน เป็นทดสอบแบบเอบี (A/B Testing) โดยที่ตัวแปรเอ หมายถึงวิธีการค้นหาด้วยชื่อเอกสารเพียงอย่างเดียว ในขณะที่ตัวแปรบี เป็นวิธีการสร้างระบบแนะนำวีดิทัศน์ที่เสนอ (วีอาร์เอส)

ผลการประเมินจากกลุ่มตัวอย่างจำนวน 80 เอกสารจาก 15 รายวิชา ได้ถูกนำมาคำนวณหาค่าเซเรนดิพิตี้ในมุมมองที่แตกต่างกัน จากตัวแปรในการทดลอง (A, B) ที่ได้อธิบายไว้ส่วนก่อนหน้านี้นั้น ทำให้มีความจำเป็นต้องระบุคะแนนจากผู้ใช้งานจริง หากผู้ใช้งานได้ระบุว่า วีดิทัศน์ที่เสนอเป็นวีดิทัศน์ที่ไม่ได้คาดคิดมาก่อน จะนับวีดิทัศน์นั้นว่าเป็นตัวอย่างที่ไม่ตรงกับเนื้อหาหลักของเอกสาร แล้วจึงพิจารณาตัวอย่างที่ไม่คาดคิดเหล่านั้น หากผู้ใช้งานระบุว่ามีความเกี่ยวข้องกับเอกสาร และจะถูกนับเป็นตัวอย่างบวกที่เป็นประโยชน์กับเนื้อหาในเอกสารนั้น ซึ่งหมายถึงเป็นผลลัพธ์ที่ไม่คาดคิดและเป็นประโยชน์

ตารางที่ 4.1 แสดงการเปรียบเทียบค่าเซเรนดิพิตี้โดยรวม จากผลการประเมินเอกสารทั้งหมด โดยแบ่งข้อมูลตามลักษณะของผู้เข้าร่วมประเมินและวิธีการที่ใช้ในการสร้างรายการแนะนำวีดิทัศน์

ตารางที่ 4.1 ค่าเซเรนดิพิตี้ของเอกสารในภาพรวม จัดประเภทตามผู้เข้าร่วมประเมิน

ผู้ประเมิน	**ค่าเซเรนดิพิตี้**	
	ค้นหาด้วยชื่อเอกสาร	**วีอาร์เอส**
ผู้สอน	0.136	**0.524**
ผู้เรียน	0.174	**0.332**
รวมผู้เรียนและผู้สอน	0.164	**0.388**

เพื่อจะแสดงให้เห็นค่าเซเรนดิพิตี้จากมุมมองของแต่ละเอกสาร ตารางที่ 4.2 จึงแสดงค่าสถิติของค่าเซเรนดิพิตี้ที่สำคัญ คือค่าเซเรนดิพิตี้เฉลี่ยกลาง (Mean Average Serendipity) และส่วนเบี่ยงเบนมาตรฐานของเซเรนดิพิตี้เฉลี่ย (Standard Deviation of Average Serendipity) ซึ่งเป็นการหาค่ากลาง (Mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ของคะแนนเซเรนดิพิตี้เฉลี่ยของรายการวีดิทัศน์ที่แนะนำให้ผู้ใช้งานในแต่ละเอกสาร

ตารางที่ 4.2 ค่าเฉลี่ยค่าเซเรนดิพิตี้ของเอกสาร จัดประเภทตามผู้เข้าร่วมประเมิน

ผู้ประเมิน	**ค้นหาด้วยชื่อเอกสาร**		**วีอาร์เอส**	
	ค่าเซเรนดิพิตี้เฉลี่ยกลาง	**ส่วนเบี่ยงเบนมาตรฐาน**	**ค่าเซเรนดิพิตี้เฉลี่ยกลาง**	**ส่วนเบี่ยงเบนมาตรฐาน**
ผู้สอน	0.199	0.352	**0.532**	0.423
ผู้เรียน	0.240	0.382	**0.386**	0.404
รวมผู้เรียนและผู้สอน	0.230	0.372	**0.426**	0.411

จากตารางทั้ง 2 (ตารางที่ 4.1 และตารางที่ 4.2) แสดงถึงค่าที่ได้จากตัวชี้วัดที่น่าสนใจบางประการที่ควรนำมาพิจารณา ประการแรกคือ วิธีวีอาร์เอสให้ค่าเซเรนดิพิตี้ที่ดีกว่าการค้นหาเพียงชื่อเอกสาร ทำให้เกิดสมมติฐานที่ว่า วิธีวีอาร์เอสช่วยปรับปรุงผลลัพธ์ของวีดิทัศน์ที่แนะนำให้ผู้ใช้งานในแง่ของการเพิ่มผลลัพธ์ที่เป็นประโยชน์ที่ผู้ประเมินไม่คาดคิดมาก่อน (ซึ่งจะทำการทดสอบสมมติฐานในขั้นตอนหลังจากนี้) อีกเรื่องหนึ่งที่น่าสนใจ คือ ผู้สอนซึ่งเป็นเจ้าของรายวิชานั้น มีแนวโน้มว่าจะระบุว่าวีดิทัศน์ที่ไม่ได้คาดคิดมาก่อนนั้นเป็นผลลัพธ์ทางบวกสูงกว่าผู้เรียน จากข้อมูลในตารางข้างต้น ช่วยสนับสนุนข้อสรุปที่ว่า การตัดสินใจของผู้สอนช่วยสร้างค่าเซเรนดิพิตี้ได้มากกว่าการตัดสินใจจากผู้เรียน

จากตารางที่ 4.3 ที่แสดงค่าเฉลี่ยค่าเซเรนดิพิตี้จาก 10 รายวิชา พบว่า 7 ใน 10 รายวิชานั้นมีค่าเฉลี่ยค่าเซเรนดิพิตี้จากวิธีวีอาร์เอสสูงกว่าการค้นหาด้วยชื่อเอกสาร ซึ่ง 4 รายวิชา ใน 7 รายวิชาดังกล่าวนั้น (4 รายวิชา คือ Computer Algorithm, Computer Electronics and Interfacing, Performance Analysis and Evaluation, Wireless Computer Networks) วีอาร์เอสให้ค่าเซเรนดิพิตี้ที่มากกว่าถึงสองเท่า และอีก 3 รายวิชาที่เหลือ (Computer System Architectures, Computer Vision, User Interface Design) มีค่าเซเรนดิพิตี้เป็นศูนย์เทียบกับวิธีการค้นหาด้วยชื่อเอกสาร ดังนั้น จึงไม่สามารถสรุปได้อย่างชัดเจนว่า วีอาร์เอสให้ค่าเซเรนดิพิตี้ที่ดีกว่าอีกวิธีหนึ่ง เนื่องจากมี 3 รายวิชาที่ให้ผลลัพธ์ตรงกันข้าม

ตารางที่ 4.3 ค่าเฉลี่ยค่าเซเรนดิพิตี้ของรายวิชา 10 วิชา

รายวิชา	**ค้นหาด้วยชื่อเอกสาร**	**วีอาร์เอส**
Computer Algorithm	0.333	**0.875**
Computer Electronics and Interfacing	0.136	**0.673**
Computer System Architectures	0.000	**1.000**
Computer Vision	0.000	**0.666**
Introduction to Robotics	**1.000**	0.600
Performance Analysis and Evaluation	0.025	**0.454**
Programming Methodology 1	**0.704**	0.286
Research Methods in Computer Science	**0.261**	0.163
User Interface Design	0.000	**0.055**
Wireless Computer Networks	0.367	**0.750**

ตารางที่ 4.4 ตัวอย่างค่าเฉลี่ยคำตอบของการประเมินในแต่ละแบบจากรายวิชาที่สนใจ

รายวิชา	ค้นหาด้วยชื่อเอกสาร		วีอาร์เอส	
	ไม่คาดคิด	สัมพันธ์กับเอกสาร	ไม่คาดคิด	สัมพันธ์กับเอกสาร
Introduction to Robotics	2.000	2.000	5.000	3.000
Programming Methodology 1	4.143	2.143	6.429	0.857
Research Methods in Computer Science	7.500	2.125	7.375	1.375

เมื่อทำการตรวจสอบและวิเคราะห์ 3 รายวิชาที่ทำให้วิธีการค้นหาด้วยชื่อมีค่าเซเรนดิพิตี้ที่ดีกว่าวิธีวีอาร์เอส ตารางที่ 4.4 จึงได้นำเสนอค่าเฉลี่ยของผลตอบรับของผู้ประเมินจากข้อมูลดิบที่มีต่อรายการแนะนำด้วยวิธีวีอาร์เอส เทียบกับวิธีค้นหาด้วยชื่อเอกสาร ในรายวิชา Programming Methodology และวิชา Research Method in Computer Science ค่าเฉลี่ยของจำนวนวีดิทัศน์ที่ผู้ประเมินให้คำตอบว่าไม่คาดคิดและที่ให้คำตอบว่าสัมพันธ์กับเอกสาร วิธีค้นหาด้วยชื่อเอกสารมีค่าสูงกว่าจากวิธีวีอาร์เอส ซึ่งเป็นไปในทิศทางเดียวกันกับค่าเซเรนดิพิตี้ตามปกติ แปลว่าสองวิชานี้ให้ค่าที่แตกต่างจากรายวิชาอื่นจริง แต่วิชา Introduction to Robotics ให้ผลลัพธ์ในทางตรงกันข้าม โดยค่าเฉลี่ยของผลตอบรับที่ไม่คาดคิดจากวิธีนี้ มีค่าเพียง 2.000 ซึ่งมีค่าเท่ากับคะแนนบวก ดังนั้น วิธีการค้นหาด้วยชื่อของรายวิชานี้ มีค่าเซเรนดิพิตี้เท่ากับ 1.000 ซึ่งเป็นค่าสูงสุดที่เป็นไปได้ ในทางกลับกันวิธีวีอาร์เอสให้ค่าเซเรนดิพิตี้ที่น้อยกว่า คือเท่ากับ 0.600 ก็จริง แต่ก็เนื่องมาจาก จำนวนผลตอบรับที่ไม่คาดคิดมากกว่าจำนวนที่สัมพันธ์กับเอกสาร คือ 5 และ 3 ตามลำดับ ซึ่งการที่วีอาร์เอสให้ค่าที่สัมพันธ์กับเอกสารถึง 3 ในขณะที่วิธีที่หาด้วยชื่อเอกสารนั้นน้อยกว่า ทำให้ไม่อาจสรุปได้อย่างชัดเจนว่าวิธีวีอาร์เอสดีกว่าอีกวิธีหนึ่งในวิชานี้

ตารางที่ 4.5 แสดงถึงรายละเอียดของค่าเซเรนดิพิตี้ที่น่าสนใจ ซึ่งได้แก่ค่า 1 และ 0 ซึ่งสถานการณ์ที่ค่าเซเรนดิพิตี้สามารถมีค่าเป็น 1 ได้นั้น คือเมื่อค่าคะแนนที่ประเมินว่าสัมพันธ์กับเอกสาร มีค่าเท่ากับผลลัพธ์ที่ไม่คาดคิดดังแสดงในตาราง ส่วนค่าเซเรนดิพิตี้ที่เป็น 0 จะเกิดได้จากสองสาเหตุด้วยกัน สาเหตุแรก คือเมื่อค่าผลลัพธ์ที่ไม่คาดคิดเป็น 0 ซึ่งทำให้ไม่สามารถคำนวณหาค่าเซเรนดิพิตี้ได้ (ส่วนเป็นศูนย์) อีกสาเหตุหนึ่งเกิดจากการที่ไม่มีค่าคะแนนที่ประเมินว่าสัมพันธ์กับเอกสาร

ตารางที่ 4.5 ข้อมูลค่าเซเรนดิพิตี้ที่น่าสนใจ

วิธีการ	ค่าเซเรนดิพิตี้ = 1	ค่าเซเรนดิพิตี้ = 0	
		ไม่คาดคิด = 0	สัมพันธ์กับเอกสาร = 0
ค้นหาด้วยชื่อเอกสาร	11.11%	16.92%	62.96%
วีอาร์เอส	23.64%	15.38%	32.73%

จากผลลัพธ์ที่ไม่คาดคิดเลย จากตาราง จะเห็นว่าร้อยละของค่าผลลัพธ์ที่ไม่คาดคิดที่เป็น 0 จากวิธีวีอาร์เอสนั้น เหมือนกันกับจากวิธีค้นหาจากชื่อเอกสาร แต่วิธีค้นหาจากชื่อเอกสารมีค่าร้อยละของคะแนนที่ประเมินว่าสัมพันธ์กับเอกสารที่เป็น 0 สูงกว่า หมายความว่า วิธีวีอาร์เอสให้ผลตอบรับที่สัมพันธ์กับเอกสารได้มากกว่าวิธีค้นหาด้วยชื่อเอกสาร

ในการค้นหาวิธีการที่ดีที่สุด จะใช้การทดสอบแซท (Z-test) ที่ความช่วงความเชื่อมั่น (Confidence Interval) เท่ากับร้อยละ 95 ในการหาความแตกต่างของค่าเซเรนดิพิตี้จากวิธีวีอาร์เอสและวิธีค้นหาด้วยชื่อเอกสาร ซึ่งทดสอบแซทเป็นวิธีทดสอบทางสถิติใช้สำหรับตัดสินใจว่าค่าเฉลี่ยประชากรของสองกลุ่มนั้นแตกต่างกันหรือไม่ เมื่อรู้ความแปรปรวนและมีกลุ่มตัวอย่างขนาดใหญ่ สมมติฐานในงานวิจัยนี้ คือ วิธีวีอาร์เอสสามารถให้ค่าเซเรนดิพิตี้ที่สูงกว่าวิธีค้นหาด้วยชื่อเอกสาร ซึ่งตรงกันข้ามกับ Null Hypothesis ที่ทั้งสองวิธีให้ค่าเท่ากัน จากผลลัพธ์ที่ได้จากผู้เข้าร่วมประเมินทั้งหมด และจากเงื่อนไขที่บางผลการประเมินไม่สามารถหาค่าเซเรนดิพิตี้ได้ (ไม่มีผลลัพธ์ที่ไม่คาดคิดเลย) ทำให้จำนวนกลุ่มตัวอย่างที่ใช้ในการประเมินมีค่าแซทไม่เท่ากัน

จากตารางที่ 4.6 จะพบว่า ค่า P(Z<=z) ที่ความช่วงความเชื่อมั่นร้อยละ 95 มีค่าน้อยกว่าอัลฟาที่ α = 0.05 ดังนั้น การทดสอบแซทนี้ปฏิเสธสมมติฐานที่ว่าสองวิธีนี้ให้ค่าเฉลี่ยเท่ากัน หมายความว่า ค่าเซเรนดิพิตี้เฉลี่ยของวิธีวีอาร์เอสสูงกว่าวิธีค้นหาด้วยชื่อเอกสารอย่างมีนัยสำคัญ เท่ากับว่า วิธีวีอาร์เอสนี้ให้ผลลัพธ์ที่ไม่คาดคิดแต่สัมพันธ์กับเอกสารและเป็นประโยชน์มากกว่า จึงเหมาะที่จะใช้เป็นวิธีสร้างรายการแนะนำวีดิทัศน์สำหรับการเรียงลำดับผลลัพธ์ใหม่ในขั้นตอนต่อไป

ตารางที่ 4.6 ข้อมูลทางสถิติจากกาทำการทดสอบแซทที่ค่าอัลฟาเท่ากับ 0.05

	ค้นหาด้วยชื่อเอกสาร	วีอาร์เอส
Mean	0.168	0.386180556
Known Variance	0.09736955	0.16346514
z	-2.44217184	
P(Z<=z) two-tail	0.014599196	

z Critical two-tail	1.959963985

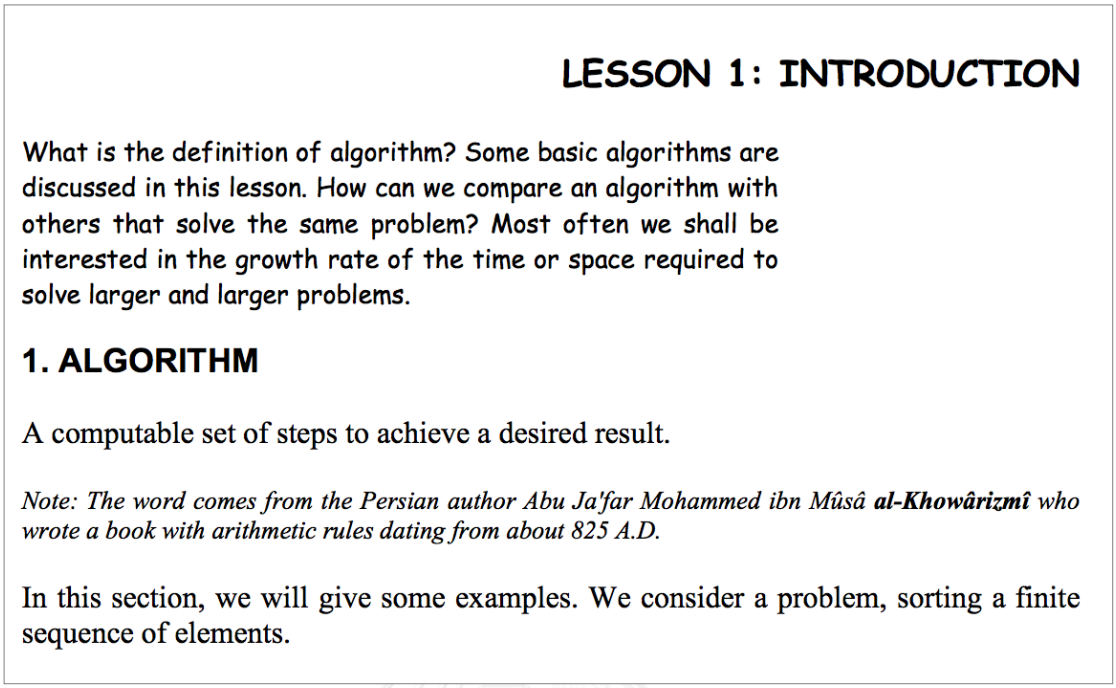
4.2 กลุ่มเอกสารตัวอย่างที่ใช้ในการทดลองอัลกอริทึมเพิ่มความครอบคลุมแนวคิด

เอกสารตัวอย่างที่ใช้ในการทดลองครั้งนี้ คัดเลือกมาจากเอกสารในระบบจัดการเรียนการสอน มายคอร์สวิลล์ โดยเลือกจากวิชาของภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ภาคการศึกษาต้น ปีการศึกษา 2559 เอกสารเหล่านั้นต้องอยู่ในรูปแบบพีดีเอฟ ที่เนื้อหาเป็นภาษาอังกฤษ ทั้งหมด 5 วิชา วิชาละ 5 เอกสาร ดังแสดงในตารางที่ 4.7 เอกสารเหล่านี้ นอกจากจะต้องมีความสมบูรณ์เพียงพอแล้ว จะต้องสามารถแปลงให้เป็นเอกสารที่เป็นข้อความได้ เนื่องจากเอกสารบางส่วนที่เป็นรูปแบบพีดีเอฟ ไม่สามารถแปลงได้ เมื่อทำการแปลงรวมทั้งทำความสะอาดข้อมูลแล้ว จะต้องเหลือข้อมูลเพียงพอสำหรับการสกัดคำสำคัญ

ตารางที่ 4.7 ชื่อเอกสารที่ใช้ในการทดลอง

ชื่อวิชา	ชื่อเอกสาร
Computer Algorithm	CHAPTER 1: INTRODUCTION
	CHAPTER 2: ANALYSIS OF ALGORITHMS
	CHAPTER 3: AMORTIZED ANALYSIS
	CHAPTER 6: DYNAMIC PROGRAMMING
	CHAPTER 5: DIVIDE AND CONQUER
Discrete Structure	Logic
	Rules of Inference
	Sets and Functions
	Other Proof Methods
	SEC 1 LOGIC [Additional note]
Programming Methodology	Lecture 2 Slide
	Self Study Slide
	Lecture 1-2 Coding Style
	Lecture 1-4 Eclipse (extra)
	GitHub
Research Methods in Computer Science	Is Computer Science Science? by Peter J. Denning
	Scientific Methods in Computer Science by Gordana DODIG-CRNKOVIC
	How to read, present, review a research paper
	Lecture 02
	How to read (Dr. Ekapol)
User Interface Design	Introduction
	Users (new version)
	Design Thinking
	Prototyping
	Project Themes

ตัวอย่างเอกสารที่ได้รับการคัดเลือก



LESSON 1: INTRODUCTION

What is the definition of algorithm? Some basic algorithms are discussed in this lesson. How can we compare an algorithm with others that solve the same problem? Most often we shall be interested in the growth rate of the time or space required to solve larger and larger problems.

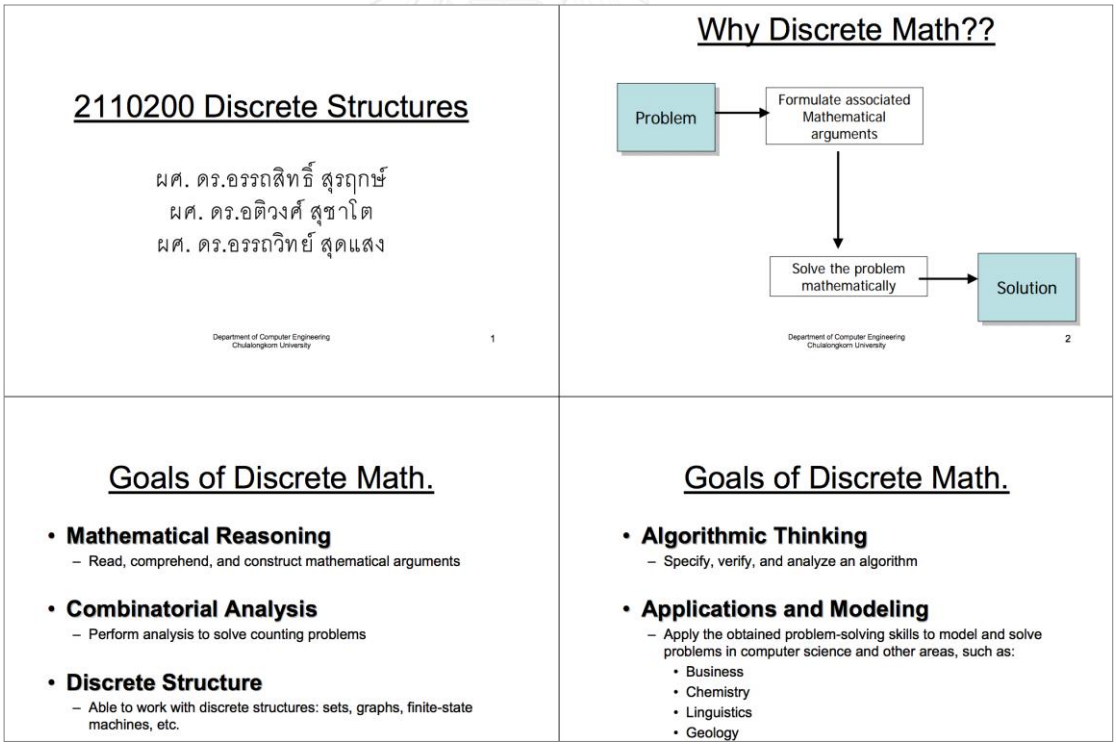
1. ALGORITHM

A computable set of steps to achieve a desired result.

Note: The word comes from the Persian author Abu Ja'far Mohammed ibn Mûsâ ***al-Khowârizmî*** *who wrote a book with arithmetic rules dating from about 825 A.D.*

In this section, we will give some examples. We consider a problem, sorting a finite sequence of elements.

ภาพที่ 4.3 ตัวอย่างเอกสารในวิชา Computer Algorithm



2110200 Discrete Structures

ผศ. ดร.อรรถสิทธิ์ สุรฤกษ์
ผศ. ดร.อติวงศ์ สุชาโต
ผศ. ดร.อรรถวิทย์ สุดแสง

Department of Computer Engineering
Chulalongkorn University 1

Why Discrete Math??

Problem → Formulate associated Mathematical arguments → Solve the problem mathematically → Solution

Department of Computer Engineering
Chulalongkorn University 2

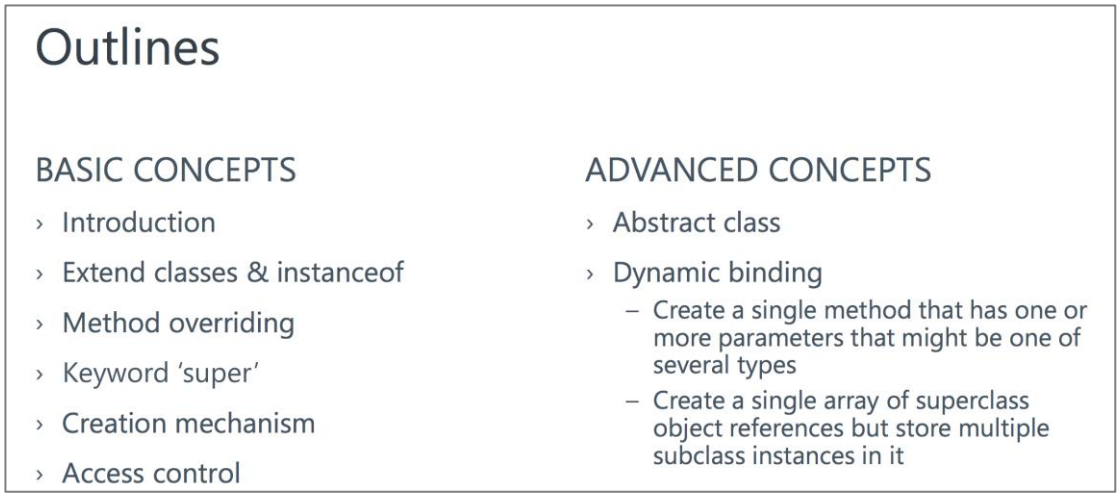
Goals of Discrete Math.

- **Mathematical Reasoning**
  - Read, comprehend, and construct mathematical arguments
- **Combinatorial Analysis**
  - Perform analysis to solve counting problems
- **Discrete Structure**
  - Able to work with discrete structures: sets, graphs, finite-state machines, etc.

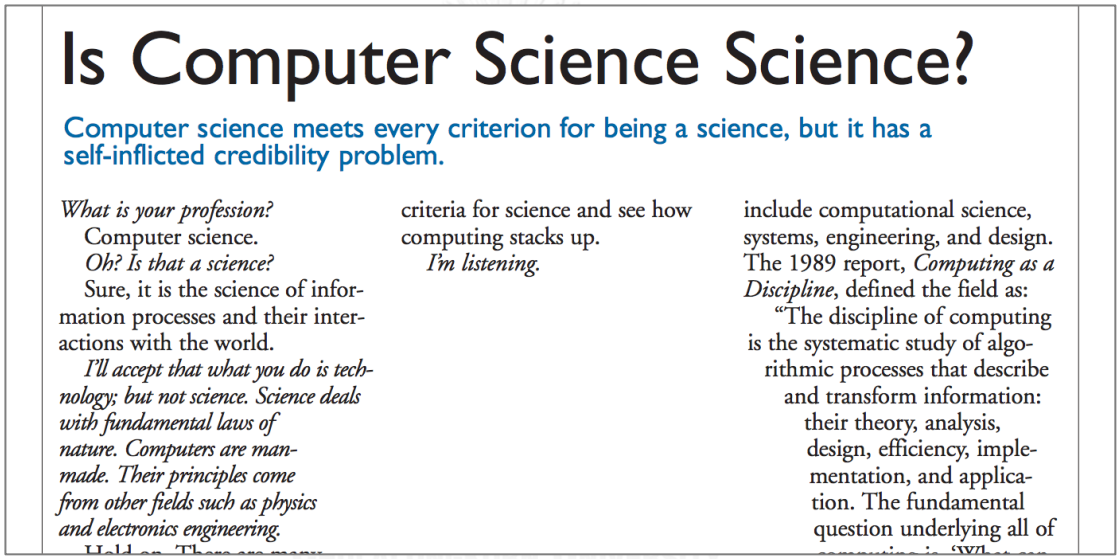
Goals of Discrete Math.

- **Algorithmic Thinking**
  - Specify, verify, and analyze an algorithm
- **Applications and Modeling**
  - Apply the obtained problem-solving skills to model and solve problems in computer science and other areas, such as:
    - Business
    - Chemistry
    - Linguistics
    - Geology

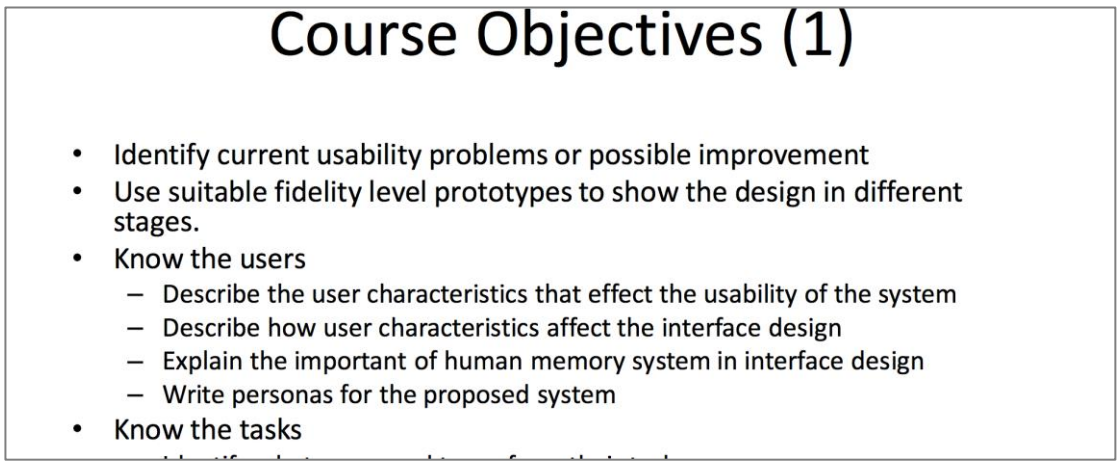
ภาพที่ 4.4 ตัวอย่างเอกสารในวิชา Discrete Structure



ภาพที่ 4.5 ตัวอย่างเอกสารในวิชา Programming Methodology



ภาพที่ 4.6 ตัวอย่างเอกสารในวิชา Research Methods in Computer Science



ภาพที่ 4.7 ตัวอย่างเอกสารในวิชา User Interface Design

4.3 ผลการทดลองจากอัลฟาเอ็นดีซีจี

จุดประสงค์หลักของการวัดผลการทดลองด้วยอัลฟาเอ็นดีซีจีนั้น เพื่อจะทดสอบประสิทธิภาพด้านความหลากหลายของผลการแนะนำของอัลกอริทึมที่เสนอใหม่ทั้งสองวิธีการ เทียบกับวิธีที่ดีที่สุดอย่างเอ็มเอ็มอาร์ และวิธีการเดินสุ่มแบบเริ่มต้นใหม่ดั้งเดิม เป้าหมายเพื่อพิสูจน์ว่า วิธีการใหม่ที่เสนอนั้น ให้ค่าความหลากหลายที่วัดด้วยค่าอัลฟาเอ็นดีซีจีที่ดีกว่า เทียบเท่า หรือต่ำกว่า วิธีอื่น ๆ ที่นำมาเปรียบเทียบหรือไม่ ในกรณีที่วิธีที่เสนอแตกต่างจากวิธีที่เปรียบเทียบ โดยเฉพาะที่ประสิทธิภาพต่ำกว่า จะพิจารณาเพิ่มเติมในส่วนที่ให้ผลลัพธ์ใกล้เคียงกับวิธีเอ็มเอ็มอาร์มากที่สุด ว่าค่าพารามิเตอร์มีลักษณะเป็นอย่างไร ในกรณีที่วิธีที่เสนอดีกว่า เกิดได้ด้วยเงื่อนไขใดบ้างและด้วยสาเหตุอะไร ซึ่งวิธีที่ใช้ในการเปรียบเทียบ ประกอบไปด้วย

- วิธีเอ็มเอ็มอาร์ แสดงด้วยตัวย่อ MMR
- วิธีการเดินสุ่มแบบเริ่มต้นใหม่ดั้งเดิม แสดงด้วยตัวย่อ RWR
- วิธีการเดินสุ่มแบบเริ่มต้นใหม่และฟังก์ชันความหลากหลายของหมวดหมู่ แบบที่ 1 แสดงด้วยตัวย่อ RWR+CDF_I
- วิธีการเดินสุ่มแบบเริ่มต้นใหม่และฟังก์ชันความหลากหลายของหมวดหมู่ แบบที่ 2 แสดงด้วยตัวย่อ RWR+CDF_II

การทดลองจะทำการวัดผลโดยพิจารณาแนวโน้มของค่าอัลฟาเอ็นดีซีจีเปรียบเทียบกันในแต่ละวิธี ที่พารามิเตอร์ต่างกันดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.3.1 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า c ที่ K แตกต่างกัน

การทดลองนี้จะทดสอบผลกระทบของการเปลี่ยนแปลงของค่าพารามิเตอร์ c ในสมการของทุกอัลกอริทึม ว่าทำให้ค่าอัลฟาเอ็นดีซีจีเพิ่ม หรือลด ในลักษณะใดได้บ้าง เนื่องจากการเปลี่ยงแปลงค่า c นั้นมีส่วนที่ทำให้ลำดับการแนะนำเปลี่ยน ตัวอย่างเช่น กรณีของอัลกอริทึมเอ็มเอ็มอาร์นั้น ค่า c จะส่งผลต่อคะแนนเอ็มเอ็มอาร์ที่คำนวณได้ในแต่ละวัตถุแนะนำ ว่าจะถูกครอบงำโดยค่าซิมิลาริตี้ หรือค่าระยะห่างของวัตถุนำกับรายการวัตถุแนะนำที่ได้เลือกไปแล้ว ถ้า c มีค่าที่สูง จะส่งผลให้ระยะห่างมีผลต่อการเลือกวัตถุ ในขณะที่ค่า c น้อย วัตถุที่จะถูกเลือกมักจะมีความเหมือนกับเอกสารมากกว่า ส่วนในกรณีของการเดินสุ่มแบบเริ่มต้นใหม่นั้น ค่า c ที่เปลี่ยนไปนั้น ส่งผลโดยตรงกับผลการคำนวณเมทริกซ์ r ในแต่ละรอบ และส่งผลต่อลำดับของการแนะนำตามมา รวมทั้งยังส่งผลถึงการลู่เข้าของ

การเดินสุ่ม โดยค่า c ที่น้อย ทำให้พจน์ (1-c) e มีค่ามาก ทำให้เกิดผลที่ตามมาได้ 2 กรณีคือ การเดินสุ่มลู่เข้าอย่างรวดเร็ว หรือการเดินสุ่มไม่ลู่เข้าเลย

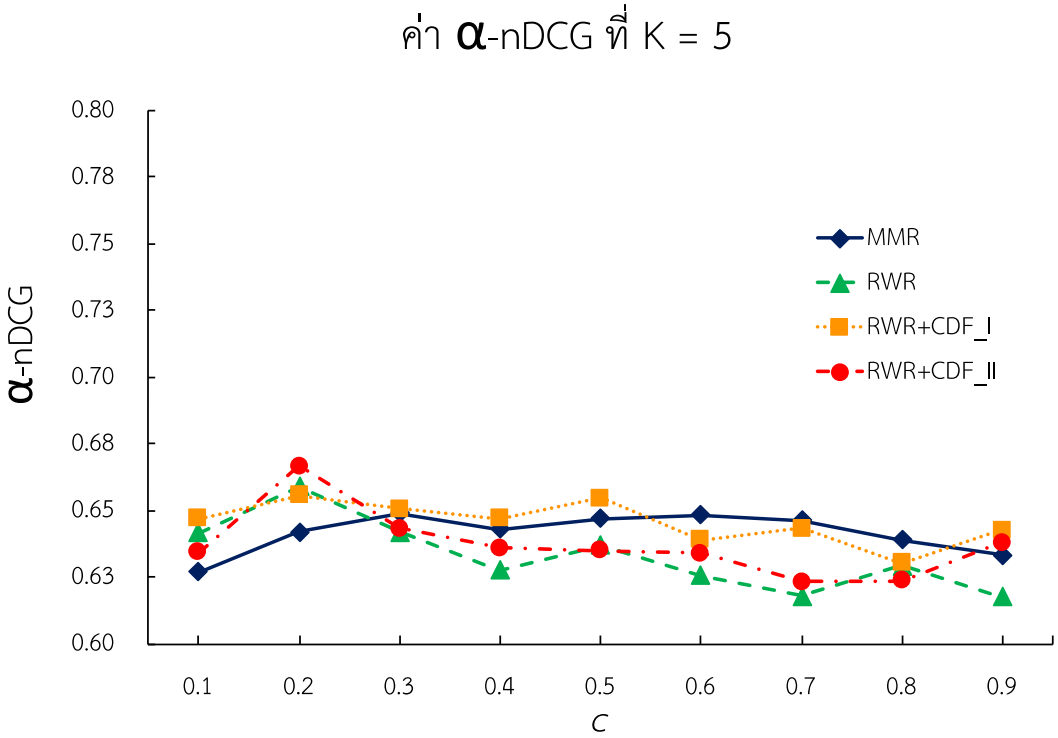
การทดสอบในส่วนนี้จะทำการทดสอบที่ ค่า K ซึ่งหมายถึงจำนวนวีดิทัศน์ที่จะทำการแนะนำที่ต่างกันไปคือที่ K= 5, 10, 15, 20 ตามวิธีการวัดค่าอัลฟาเอ็นดีซีจีที่ K ที่แตกต่างกันจากทีอาร์อีซี ผลการทดลองเป็นดังนี้ ภาพที่ 4.8 ถึง ภาพที่ 4.11 เป็นกราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 5, 10, 15, 20 ตามลำดับ จะเห็นว่า ณ ค่า K ที่ 10 ใน

ภาพที่ 4.9 และ K ที่ 20 ในภาพที่ 4.11ตามลำดับ แม้วิธี MMR จะเป็นวิธีที่ให้ผลลัพธ์ที่ดีที่สุดที่ค่า c ทุกค่า แต่วิธี RWR+CDF_II ซึ่งทำได้ดีกว่าวิธี RWR วิธีอื่นๆ ที่ค่า c ตั้งแต่ 0.3 ขึ้นไปนั้น ก็ให้ผลลัพธ์ที่ใกล้เคียงกับวิธี MMR ได้ ที่ค่า c=0.2 และ 0.5

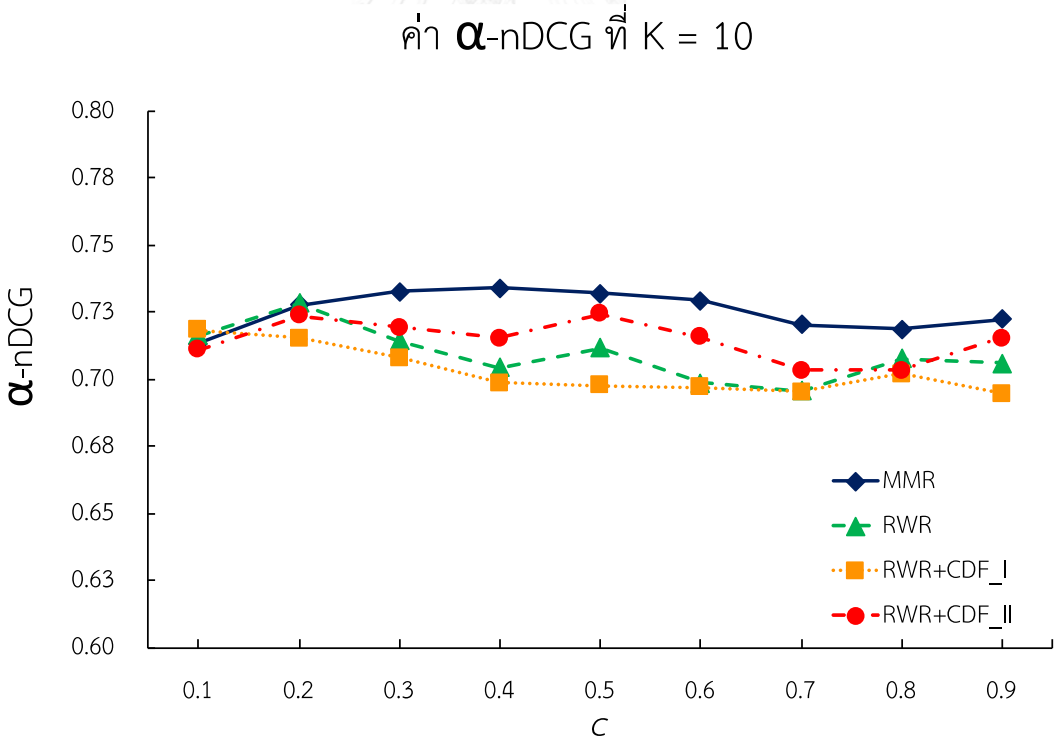
สำหรับภาพที่ 4.8 ซึ่งแสดงผลการทดลองที่ค่า K=5 วิธี RWR+CDF_II ให้ผลลัพธ์ที่เหนือกว่าวิธี MMR ที่ค่า c บางค่า (0.1, 0.2, 0.3, 0.4, 0.5 และ 0.9) ในขณะที่วิธี RWR และวิธี RWR+CDF_I สามารถทำได้ดีกว่าวิธี MMR ที่ค่า C น้อยกว่าหรือเท่ากับ 0.2 เท่านั้น

จากกราฟทั้ง 4 จึงแสดงให้เห็นว่า ผลลัพธ์จากวิธี MMR จะมีค่าต่ำ ในกรณีที่ค่า K ต่ำ เนื่องมาจากจำนวนของวัตถุในรายการที่ถูกจัดลำดับแล้ว ที่ถูกเลือกก่อนหน้าที่ใช้ในการคำนวณระยะทางการเลือกวัตถุต่อไปนั้น มีจำนวนที่น้อย ส่วนของคะแนน $\sum_{v \in S} dis(u, v)$ ที่มีค่าน้อยจากสมการ MMR กับค่า c มีค่าตั้งแต่ 0.1 ถึง 0.5 ส่งผลให้ค่า $f_{MMR}(u, q)$ ถูกครอบงำโดย (1 - c) sim(u, q) ซึ่งส่งผลให้รายการแนะนำเลือกวีดิทัศน์ที่ใกล้เคียงกับเอกสารต้นทางมากกว่าก่อน โดยส่วนของความแตกต่างระหว่างวีดิทัศน์ที่เลือกไปแล้วกับที่พิจารณาอยู่ถูกพิจารณาในสัดส่วนที่น้อยลง

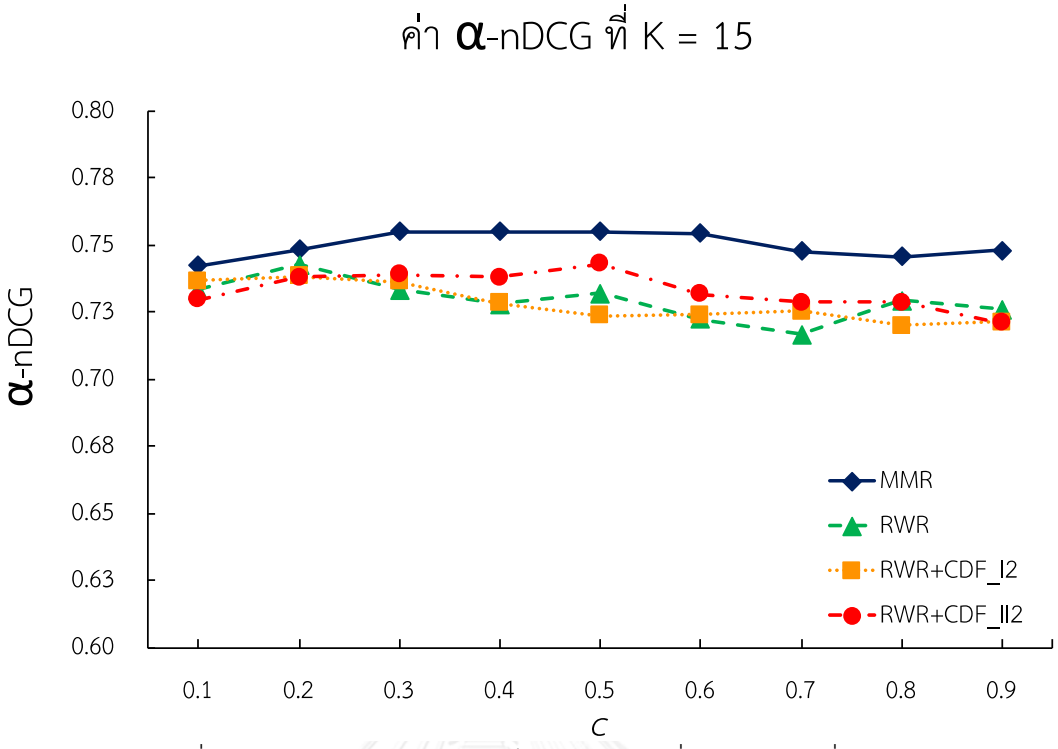
Chulalongkorn University



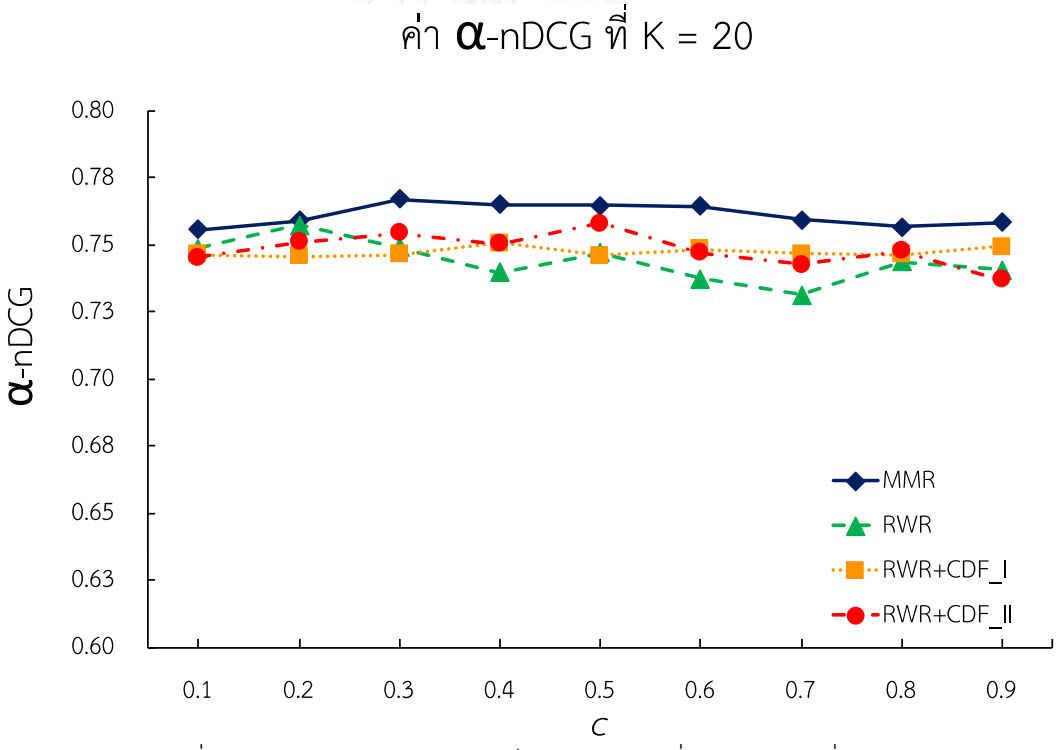
ภาพที่ 4.8 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 5



ภาพที่ 4.9 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 10



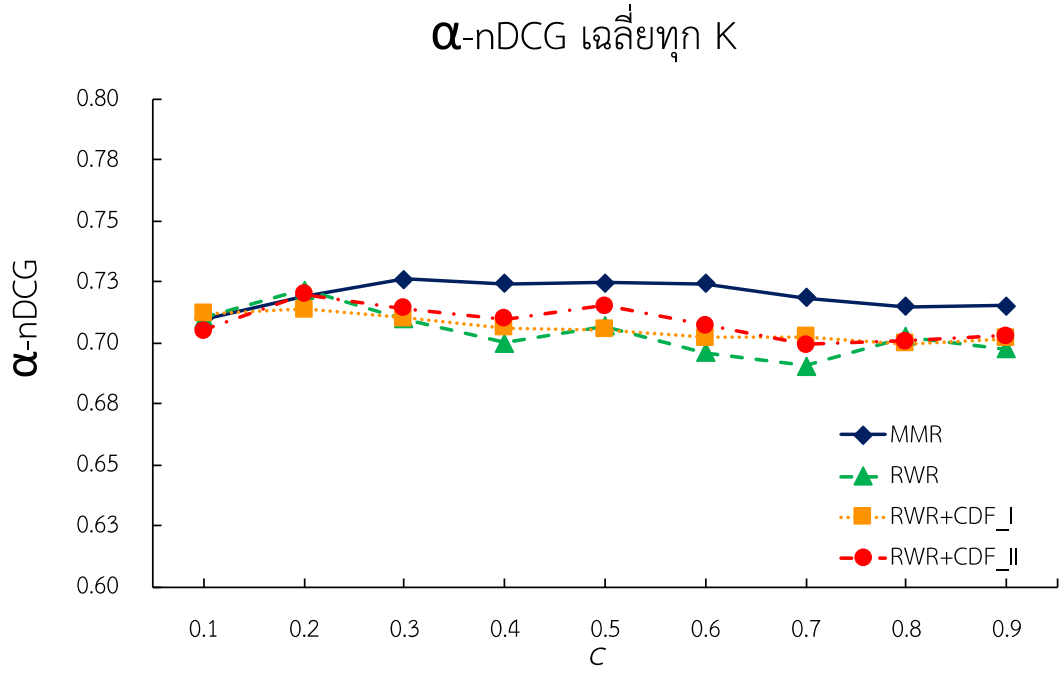
ภาพที่ 4.10 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 15



ภาพที่ 4.11 กราฟแสดงค่าอัลฟาเอ็นดีซีจีของ c ที่แตกต่างกัน ที่ K = 20

4.3.2 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า c เฉลี่ยสำหรับทุก K แตกต่างกัน

หลังจากทำการเปรียบเทียบผลของค่า c แยกตามจำนวนวีดิทัศน์ที่ต้องการแนะนำ K ตัวแล้ว ทางผู้วิจัยได้พิจารณาผลของการเปลี่ยนแปลงของค่า c ต่อคะแนนความหลากหลายอัลฟาเอ็นดีซีจี ที่ค่าเฉลี่ยสำหรับทุก K เพื่อวัดผลโดยรวมว่า ประสิทธิภาพของอัลกอริทึมที่เสนอโดยเฉลี่ยที่ทุกจำนวนของการแนะนำ ยังมีแนวโน้มเช่นเดียวกับการพิจารณาแยกค่า K หรือไม่ ผลการทดลองแสดงในภาพที่ 4.12



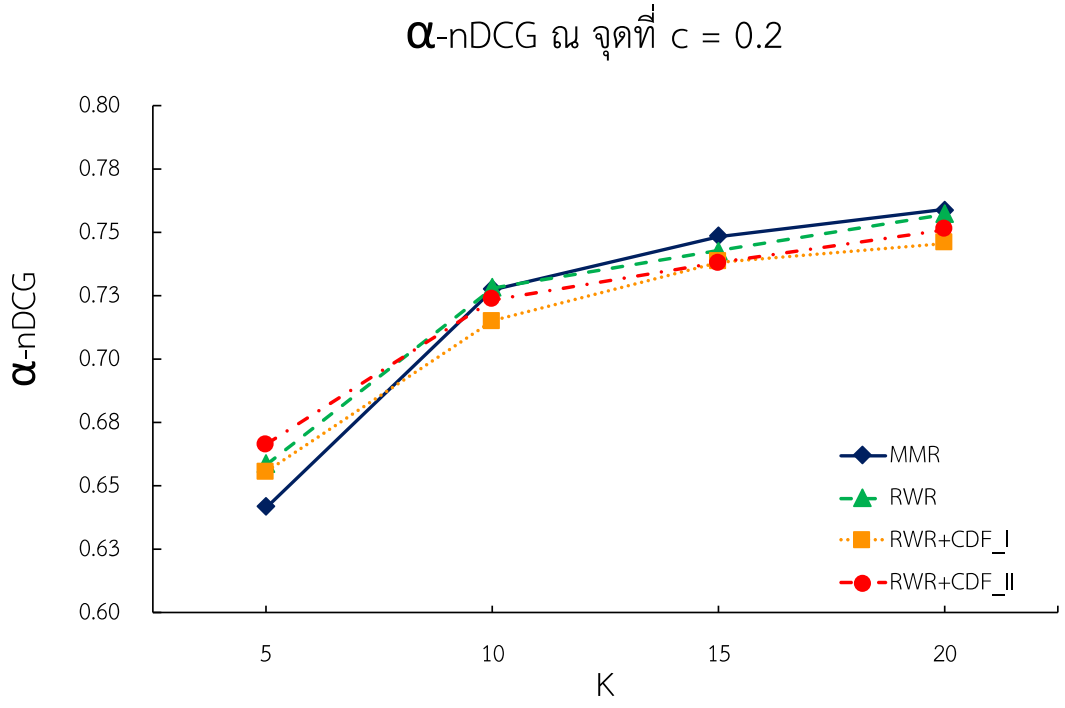
ภาพที่ 4.12 กราฟแสดงค่าอัลฟาเอ็นดีซีจีเฉลี่ยทุก K ของ c ที่แตกต่างกัน

จากภาพที่ 4.12 พบว่า เส้นกราฟมีแนวโน้มไปลักษณะเดียวกันกับกราฟที่ค่า K อื่น ๆ กล่าวคืออัลกอริทึมที่ใช้ RWR ทุกตัวยังคงต่ำกว่า MMR โดยรวม อาจมีค่าที่สูงกว่าที่ c ต่ำ ๆ เพียงเล็กน้อย ในกลุ่มอัลกอริทึม RWR ด้วยกันนั้นส่วนใหญ่ RWR+CDF_II จะให้ค่าที่สูงกว่าในเกือบทุกค่า c โดยเฉพาะ ณ จุดที่ค่า c=0.2 และ 0.5 ผลลัพธ์ที่ได้จากวิธี RWR+CDF_II สูงจนเข้าใกล้กับ MMR มากที่สุด จากลักษณะกราฟในภาพที่ 4.8 และภาพที่ 4.12 อาจสรุปได้ว่า การเปลี่ยนแปลงค่า c ไม่ได้ส่งผลต่ออัลกอริทึมอย่างมีนัยสำคัญ ซึ่งข้อมูลจากการประเมิน แสดงให้เห็นว่าการเพิ่มค่า c ไม่ได้ทำให้อัลฟาเอ็นดีซีจีสูงขึ้นเสมอไป ในทำนองเดียวกัน การลดค่า c ก็ไม่มีผลให้เส้นกราฟตกลงเสมอไปด้วย จากภาพที่ 4.12 และภาพก่อนหน้านี้ จะพบว่า RWR+CDF_II ให้ค่าอัลฟาเอ็นดีซีจีที่ดีที่สุดในกลุ่ม RWR โดยรวม และให้ค่าที่ใกล้เคียงกับ MMR ที่สุดอีกด้วย

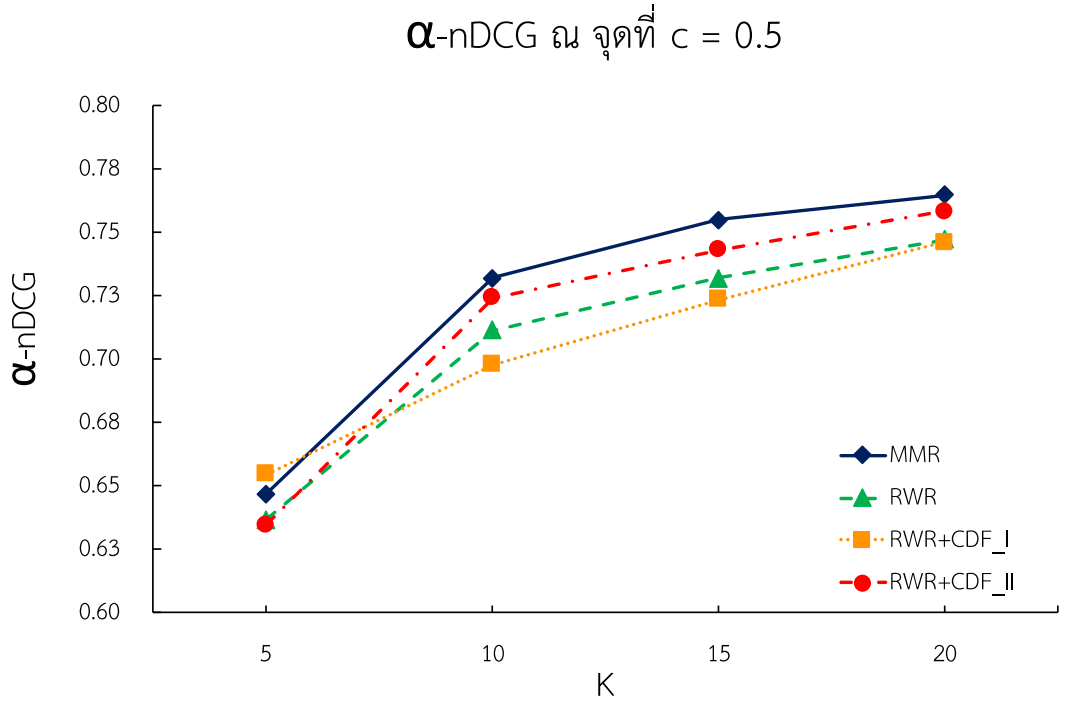
จากการทดลองที่ค่า c = 0.5 ในวิธี RWR+CDF_II ถือได้ว่าเป็นค่า c ที่ให้ค่าอัลฟาเอ็นดีซีจีที่ใกล้เคียงกับวิธี MMR มากที่สุดนอกจากนั้น ที่ค่า K ส่วนใหญ่ วิธี RWR+CDF_II ยังให้ค่าผลลัพธ์ที่สูงกว่าทั้งวิธี RWR และวิธี RWR+CDF_I

4.3.3 การทดลองเปรียบเทียบการเปลี่ยนแปลงของค่าอัลฟาเอ็นดีซีจีเมื่อเปลี่ยนค่า K ณ จุดที่ใกล้ MMR ที่สุด

จากข้อสรุปของการทดลองที่ผ่านมา จะพบว่า มีค่า c อยู่สองจุดที่ทำให้ อัลกอริทึมในกลุ่มของการเดินสุ่ม โดยเฉพาะ RWR+CDF_II มีค่าใกล้เคียงกับวิธีเอ็มเอ็มอาร์ หรือมากกว่า คือจุดที่ c = 0.2 และ 0.5 ทางผู้วิจัยจึงได้ทำการวิเคราะห์เพิ่มเติมที่สองจุดดังกล่าว เพื่อสังเกตแนวโน้มของค่าอัลฟาเอ็นดีซีจีเมื่อจำนวนวิดิทัศน์ที่จะแนะนำ K ตัวเปลี่ยนแปลงไป



ภาพที่ 4.13 กราฟแสดงค่าอัลฟาเอ็นดีซีจี ณ จุดที่ c = 0.2



ภาพที่ 4.14 กราฟแสดงค่าอัลฟาเอ็นดีซีจี ณ จุดที่ c = 0.5

ภาพที่ 4.13 และภาพที่ 4.14 ใช้อธิบายผลกระทบของค่า K ต่ออัลกอริทึมทั้งหมด ซึ่งสังเกตได้ว่า เมื่อเพิ่มค่า K จะทำให้ค่าอัลฟาเอ็นดีซีจีสูงขึ้นอย่างที่คาดหวังไว้ตามไปด้วย และที่ค่า K ต่ำ ๆ พบว่า วิธี MMR ไม่ได้ให้ผลลัพธ์ที่ดีที่สุดเสมอไปเมื่อเทียบกับการใช้วิธีอื่น ๆ ภาพที่ 4.13 ที่ c = 0.2 นั้น นอกจากจะพบว่า วิธี MMR ไม่ให้ค่าอัลฟาเอ็นดีซีจีที่เหนือกว่าวิธีอื่น ๆ ในทุกค่า K แล้ว ยังพบว่าวิธีการในกลุ่มของ RWR เอง ก็ให้ผลลัพธ์ที่สลับไปมา ไม่มีวิธีไหนที่ชนะโดยสมบูรณ์ชัดเจน ที่บางค่า K RWR ก็มากที่สุด บาง K ก็ปรากฏว่าเป็น RWR+CDF_I ในขณะที่บางครั้ง RWR+CDF_II ก็ให้ค่าอัลฟาเอ็นดีซีจีสูงที่สุดได้ ซึ่งแตกต่างจากภาพที่ 4.14 ที่ c = 0.5 นั้น MMR มีแนวโน้มที่ชัดเจนว่าให้ค่าอัลฟาเอ็นดีซีจีที่เหนือว่าวิธีการกลุ่ม RWR แม้ที่ K = 5 RWR+CDF_I จะให้ค่าที่มากกว่าก็ตาม ซึ่งเหตุผลได้อธิบายไปในการทดลองก่อนหน้า ส่วนที่น่าสนใจของกราฟนี้ก็คืออัลกอริทึมกลุ่ม RWR ตั้งแต่ K ที่มากกว่า 5 ขึ้นไปนั้น RWR+CDF_II ให้ค่าอัลฟาเอ็นดีซีจีที่สูงกว่าอัลกอริทึมในกลุ่มเดียวกัน และมีแนวโน้มจะห่างขึ้นเรื่อย ๆ ทำให้จุดที่ c = 0.5 เป็นจุดที่น่าสนใจจุดหนึ่ง

4.3.4 เปรียบเทียบผลการทดลองด้วยการทดสอบทีและทีโอเอสที

เพื่อหาข้อสรุปเกี่ยวกับประสิทธิภาพของวิธีการเดินสุ่มแบบเริ่มต้นใหม่และฟังก์ชั่นความหลากหลายของหมวดหมู่ แบบที่ 2 หรือ RWR+CDF_II ที่ใกล้เคียงกับวิธีพื้นฐานอย่าง MMR มากที่สุด ทางผู้วิจัยจึงได้ทำการเปรียบเทียบค่าเฉลี่ยอัลฟาเอ็นดีซีจีของทั้งสองวิธีการด้วยวิธีการทดสอบที (T-testing) เพื่อพิสูจน์ว่าผลลัพธ์ความหลากหลายของการแนะนำซึ่งวัดด้วยค่าอัลฟาเอ็นดีซีจีของทั้งสองวิธีการต่างกันอย่างมีนัยสำคัญหรือไม่

ตารางที่ 4.8 ตารางแสดงค่า P-Value ของการทดสอบที ในแต่ละค่า K

K	P-value	H0 result
5	0.708	ไม่สามารถปฏิเสธสมมติฐาน
10	0.235	ไม่สามารถปฏิเสธสมมติฐาน
15	0.042	ปฏิเสธสมมติฐาน
20	0.126	ไม่สามารถปฏิเสธสมมติฐาน
รวม	0.030	ปฏิเสธสมมติฐาน

การทดสอบแรก จะทำการทดสอบโดยไม่รวมผลกระทบของค่า c และแยกตามจำนวนของค่า K ที่ช่วงความเชื่อมั่นร้อยละ 90 โดยใช้ α = 0.5 และ H0 (Null Hypothesis) ผลลัพธ์การทดสอบที ได้อธิบายอยู่ในตารางที่ 4.8

เมื่อใช้ H0 การทดสอบด้วยการทดสอบที ให้ผลว่า RWR+CDF_II เป็นไปตามสมมติฐาน คือไม่สามารถบอกได้ว่ามีความแตกต่างจาก MMR อย่างมีนัยสำคัญ (P-value > α) ในบางกรณี คือที่ค่า K = 5, 10 และ 20 ในขณะเดียวกัน การทดสอบก็แสดงให้เห็นอีกว่า ณ ค่า K = 15 และผลการทดสอบรวมทุกค่า K ผลการทดสอบไม่เป็นไปในแนวทางเดียวกัน การศึกษาผลการทดสอบพบว่าสาเหตุที่การทดสอบไม่เป็นไปตามสมมติฐานนั้น เนื่องมาจากความแตกต่างกันอย่างมากของคะแนนอัลฟาเอ็นดีซีจี ที่ K = 15 จากภาพที่ 4.10 จะพบว่าค่าที่ได้จากวิธี RWR+CDF_II ค่อนข้างห่างจาก MMR ในทุกค่า c รวมทั้งที่จุด c = 0.2 และ 0.5 ซึ่งควรจะเป็นจุดที่สองอัลกอริทึมนี้ใกล้กันที่สุด กลับมีระยะห่างระหว่างกันมากกว่าที่ ค่า K อื่น ๆ ไม่เพียงแต่เฉพาะที่ค่า K = 15 เท่านั้น ที่ K อื่น ๆ ยังพบว่า ค่า c บางค่า ทำให้เกิดความแตกต่างอย่างมากระหว่างวิธี RWR+CDF_II และ MMR ซึ่งความแตกต่างนี้ อาจไม่ส่งผลโดยตรงต่อการเปรียบเทียบด้วยการทดสอบทีแบบแยกค่า K แต่ส่งผลโดยตรงต่อผลทดสอบทีที่รวมทุกค่า K ซึ่งระยะห่างที่ค่า c บางค่านี้ ทำให้เกิดความไม่สอดคล้องกันของผลลัพธ์จากการทดสอบทีของอัลกอริทึม 2 ชนิดนี้

จากการใช้การทดสอบทีซึ่งมีวัตถุประสงค์เพื่อวัดความแตกต่างของสองวิธีดำเนินการ ไม่อาจให้คำตอบได้ว่า ทั้งสองวิธีให้ผลลัพธ์ที่เหมือนกันหรือไม่ เนื่องจากในส่วนผลการทดลองที่ให้คำตอบว่าไม่สามารถปฏิเสธสมมติฐานได้นั้น จะสรุปว่ายอมรับสมมติฐานและให้คำตอบสุดท้ายว่าทั้งสองวิธีเหมือนกันไม่ได้ ผู้วิจัยจึงทำการทดสอบเพิ่มเติมด้วยการทดสอบความเท่ากัน (Equivalence Test) ด้วยวิธีทีโอเอสที (Two-One-Sided T-tests: TOST) [44] ซึ่งมีคุณสมบัติในการทดสอบหาความเท่ากันของวิธีการสองวิธีการ ในลักษณะเดียวกันกับการทดสอบที โดยให้สมมติฐาน H0 (Null Hypothesis) เป็น วิธีการทั้งสองแตกต่างกันที่ขอบเขตความแตกต่างของค่าเฉลี่ยที่กำหนดหรือไม่ ซึ่งตรงข้ามกับการทดสอบที ในการทดสอบทีโอเอสทีครั้งนี้ กำหนดขอบเขตความแตกต่างของค่าอัลฟาเอ็นดีซีจีเฉลี่ยอยู่ที่ -0.035 ถึง 0.035 ที่ช่วงความเชื่อมั่นร้อยละ 95 ตารางที่ 4.9 แสดงให้เห็นว่า จากผลการทดสอบด้วยวิธีทีโอเอสที ทั้งวิธี MMR และ RWR+CDF_II ให้ค่าอัลฟาเอ็นดีซีจีไม่ต่างกันอย่างมีนัยสำคัญ ตามเงื่อนไขของการทดสอบ ในทุกค่า K

ตารางที่ 4.9 ตารางแสดงค่า P-Value ของการทดสอบการเท่ากันด้วยวิธีทีโอเอสทีที่ขอบเขตความแตกต่างของค่าเฉลี่ย -0.035 ถึง 0.035

K	Lower bound (95 %)	Upper bound (95 %)	Lower P-value	Upper P-value	Test Interpretation
5	-0.019333	0.02843	0.00629	0.00061	Equivalent
10	-0.00719	0.02914	0.00483	< 0.0001	Equivalent
15	0.00058	0.03382	0.01716	< 0.0001	Equivalent
20	-0.00367	0.02966	0.00488	< 0.0001	Equivalent
รวม	-0.00689	0.02971	0.00582	< 0.0001	Equivalent

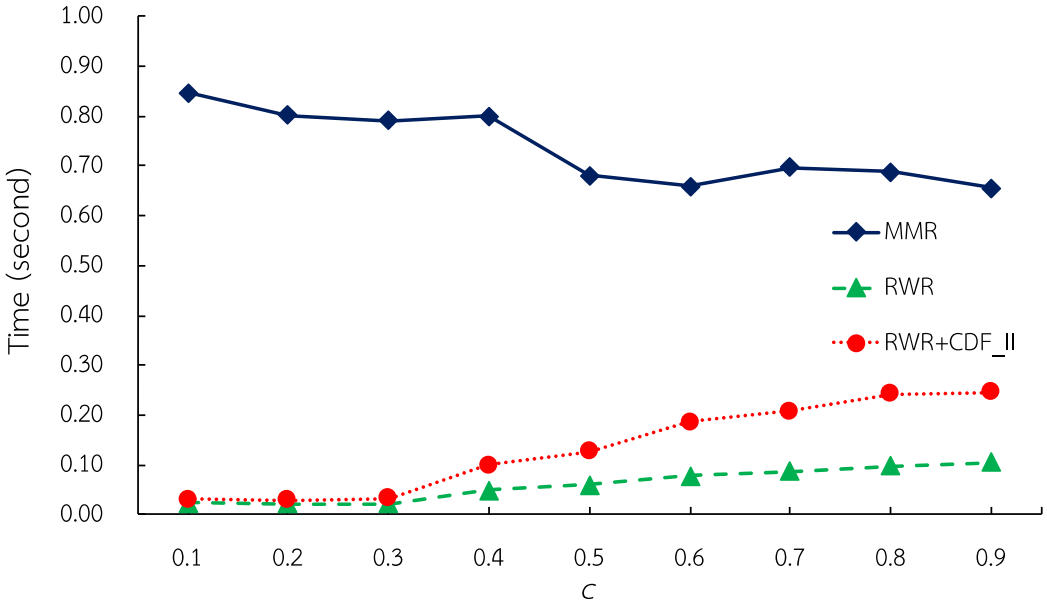
4.4 ผลการทดลองจากการวัดระยะเวลาดำเนินการ

การทดลองในส่วนของระยะเวลาในการดำเนินงานนั้น จุดประสงค์หลักเพื่อวัดภาระงานที่เกิดขึ้นเมื่อเพิ่มส่วนของการคำนวณความหลากหลายของหมวดหมู่เข้ามาในการเดินสุ่มแบบเริ่มต้นใหม่ เมื่อเทียบกับอัลกอริทึมอื่น ๆ ซึ่งในการทดลองหลังจากนี้ จะใช้วิธี RWR+CDF_II เป็นตัวแทนของทั้ง RWR+CDF_I และ RWR+CDF_II เนื่องจากทั้งสองอัลกอริทึมมีลักษณะการทำงานที่เหมือนกัน ต่างกันเพียงวิธีการเลือกผลลัพธ์สุดท้าย

ในการเปรียบเทียบระยะเวลาการทำงานของ RWR+CDF_II นั้น จะไม่สนใจระยะเวลาในขั้นตอนการเตรียมข้อมูลความแตกต่างกันระหว่างวีดิทัศน์และซิมิลาริตี้ของเอกสารกับวีดิทัศน์เก็บไว้ก่อนของอัลกอริทึม MMR และไม่สนใจระยะเวลาในการสร้างเมทริกซ์เปลี่ยนสถานะในอัลกอริทึม RWR+CDF_II โดยการทดลองจะแบ่งเป็น 3 หัวข้อหลัก คือการเปรียบเทียบระยะเวลาเฉลี่ยสำหรับทุกจำนวนวีดิทัศน์ที่แนะนำ K จำนวนกับวิธี MMR และวิธีการเดินสุ่มแบบเริ่มต้นใหม่ดั้งเดิม การเปรียบเทียบผลกระทบของจำนวนวีดิทัศน์ที่แนะนำกับวิธี MMR และการทดลองระยะเวลาในการดำเนินการเมื่อเปลี่ยนแปลงขนาดของวิดีทัศน์ N จำนวน

4.4.1 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาในการดำเนินการเมื่อเปลี่ยนค่า c สำหรับทุก K

ในการทดลองนี้ จะทำการเปรียบเทียบระยะเวลาที่ใช้ของการดำเนินการอัลกอริทึม RWR+CDF_II ซึ่งจะทำการวัดระยะเวลาในการเรียงลำดับรายการแนะนำใหม่จนสำเร็จในแต่ละเอกสาร การวัดระยะเวลาจะเริ่มวัด ณ จุดที่เริ่มการดำเนินการอัลกอริทึมเพื่อสร้างลำดับผลการแนะนำใหม่เท่านั้น ผลการวัดระยะเวลาในการดำเนินการเป็นดังภาพที่ 4.15

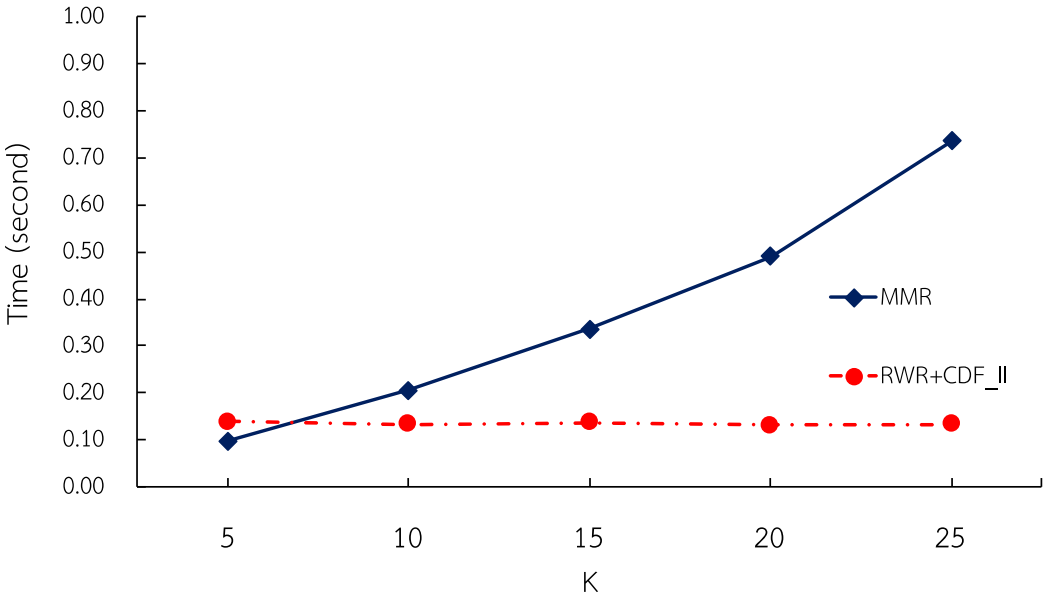


ภาพที่ 4.15 กราฟเปรียบเทียบระยะเวลาที่ใช้ ของ c ที่แตกต่างกัน

ภาพที่ 4.15 ได้แสดงให้เห็นถึงเวลาที่ต้องใช้ในการดำเนินการอัลกอริทึมของวิธี MMR ซึ่งได้ผลลัพธ์ออกมาว่า วิธี MMR ใช้เวลาในการดำเนินการ นานกว่าวิธี RWR และวิธี RWR+CDF_II อย่างเห็นได้ชัด โดยในส่วนของอัลกอริทึมกลุ่ม RWR ถึงแม้ว่าวิธี RWR+CDF_II ใช้เวลาในการดำเนินการนานกว่าวิธี RWR แบบดั้งเดิม แต่ทิศทางการเคลื่อนที่ของเส้นกราฟของทั้งสองวิธีนี้ เป็นไปในทิศทางเดียวกัน ดังนั้น ผลการทดลองนี้จึงชี้ให้เห็นว่า ค่า c เป็นปัจจัยหนึ่งที่ส่งผลต่อเวลาที่ใช้ของอัลกอริทึม RWR ทั้งสอง ซึ่งสมการที่ (2) ได้อธิบายเรื่องนี้ไว้ คือเมื่อค่า c เพิ่มขึ้น ระยะเวลาที่ใช้ในการเข้าสู่จุดที่การเดินสุ่มลู่เข้านั้นก็นานขึ้น เนื่องจากส่วนของ $(1 - c)\vec{e}_i$ มีบทบาทน้อยลงที่ c มากขึ้น อย่างไรก็ตาม แม้ว่าค่า c จะอยู่ในระดับที่สูงที่สุด วิธี RWR+CDF_II ก็ยังใช้เวลาในการดำเนินการน้อยกว่าวิธี MMR เช่นเดิม

4.4.2 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาในการดำเนินการเมื่อเปลี่ยนค่า K

สำหรับการเปรียบเทียบผลของการเปลี่ยนแปลงค่า K ในส่วนนี้ มีจุดประสงค์เพื่อพิสูจน์ความสามารถของอัลกอริทึม RWR+CDF_II เปรียบเทียบกับวิธี MMR ในกรณีที่ค่า K มีการเปลี่ยนแปลง แม้ว่าจะมีงานวิจัยอื่นได้ทำการพิสูจน์และวิเคราะห์ภาระการทำงานในเชิงเวลาของวิธีเอ็มเอ็มอาร์ ว่าเท่ากับ $O(K^2N)$ และของการเดินสุ่มแบบเริ่มต้นใหม่ว่าเป็น O(N) แล้วก็ตาม แต่ทางผู้วิจัยอยากทราบว่า ในขั้นตอนของการเพิ่มฟังก์ชันความหลากหลายของหมวดหมู่เข้าไปนั้น ค่า K มีผลกระทบต่อภาระการทำงานเชิงเวลาโดยรวมหรือไม่



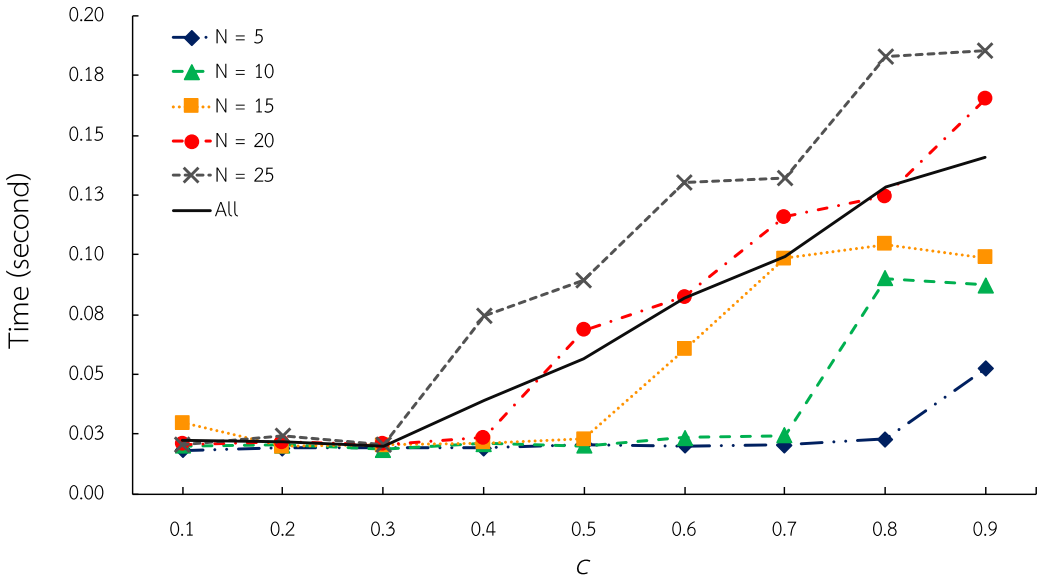
ภาพที่ 4.16 กราฟเปรียบเทียบระยะเวลาที่ใช้ ของ K ที่ต่างกัน

ภาพที่ 4.16 ได้อธิบายถึงผลกระทบที่ค่า K หรือจำนวนของการเรียงลำดับที่ถูกเลือกนั้น มีต่อประสิทธิภาพของเวลาที่ใช้ในการดำเนินงาน ซึ่งกราฟนี้ ได้ช่วยพิสูจน์ให้เห็นว่า ค่า K ไม่มีความสัมพันธ์กับวิธี RWR+CDF_II ซึ่งตรงกันข้ามกับวิธี MMR ที่มีค่า K มาเกี่ยวข้องในอัลกอริทึมด้วยเหตุผลเพราะ วิธี RWR+CDF_II ใช้การคำนวณเพียงครั้งเดียวสำหรับการนำมาใช้กับค่า K ได้ทุกค่า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.4.3 การทดลองเปรียบเทียบการเปลี่ยนแปลงของระยะเวลาของ RWR+CDF_II ที่จำนวนวีดิทัศน์ทั้งหมด N จำนวน ที่แตกต่างกัน

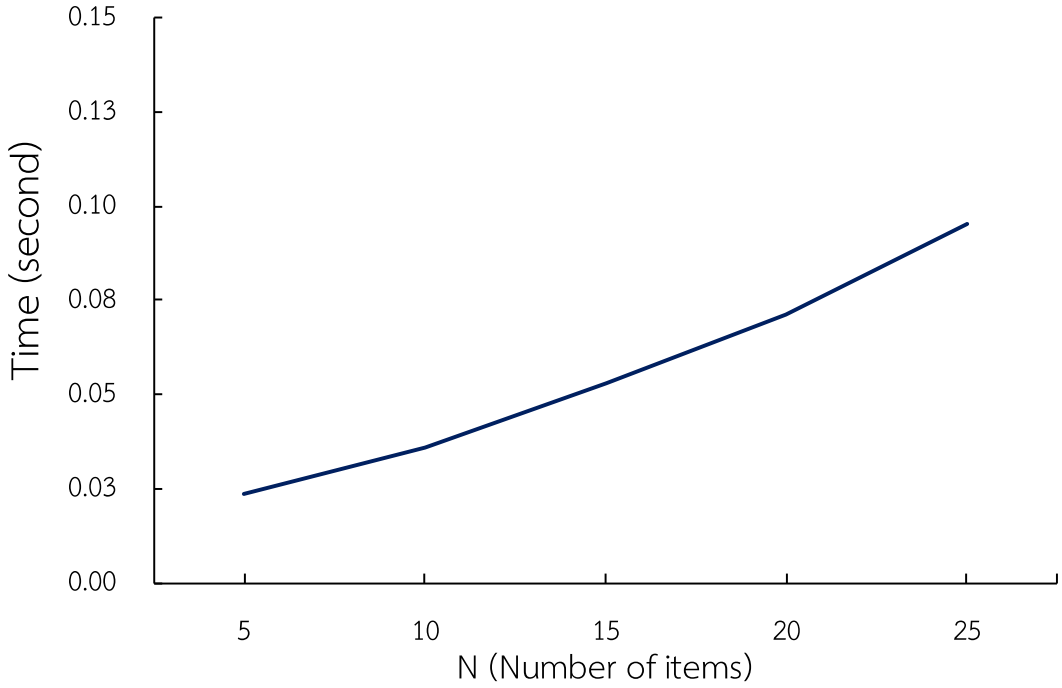
นอกเหนือไปจากค่า K แล้ว ตัวแปรอีกตัวแปรหนึ่งที่มีโอกาสส่งผลต่อภาระการทำงานเชิงเวลาของ RWR+CDF_II คือจำนวนของวีดิทัศน์ทั้งหมด ที่ใช้ในการสร้างเมทริกซ์เปลี่ยนสถานะสำหรับการเดินสุ่ม ซึ่งจะส่งผลกระทบโดยตรงกับขนาดของเมทริกซ์และระยะเวลาที่ใช้ในการคูณเมทริกซ์ในการเดินสุ่ม เช่นเดียวกับการวัดภาระการทำงานเชิงเวลาเมื่อเปลี่ยนค่า K ในหัวที่ผ่านมา ภาระการทำงานของการเดินสุ่มแบบเริ่มต้นใหม่ดั้งเดิมนั้นจะเป็น O(N) แต่การเพิ่มเข้ามาของฟังก์ชันความหลากหลายของหมวดหมู่นั้น ก็ต้องทำการตรวจสอบว่า การเปลี่ยนค่า N ไม่ทำให้ภาระงานเปลี่ยนไปจากเดิมมากนัก รวมทั้งภาระในการทำงานเชิงเวลายังคงเป็นเส้นตรงอยู่ (Linear)



ภาพที่ 4.17 กราฟแสดงการเปลี่ยนแปลงระยะเวลาที่ใช้ของ RWR+CDF_II ที่จำนวน N ที่แตกต่างกัน

ภาพที่ 4.17 แสดงถึงเวลาที่ใช้ในการดำเนินการ ณ ค่า c ที่แตกต่างกัน โดยศึกษากับจำนวนวิดีทัศน์ทั้งหมดที่แตกต่างกัน 5 จำนวน ได้แก่ 5, 10, 15 และ 25 และขนาดค่าเฉลี่ยจากทุกจำนวนดังกล่าว เส้นกราฟที่เป็นตัวแทนของแต่ละจำนวนทั้ง 5 เส้นนี้ และเส้นกราฟของค่าเฉลี่ยรวมทุกจำนวน มีทิศทางไปในทางเดียวกัน คือเพิ่มขึ้นตามค่า c ที่สูงขึ้น เช่นเดียวกับกราฟในภาพที่ 4.15

ความแตกต่างระหว่างเส้นกราฟแต่ละเส้น คือ เวลาที่ใช้ในการดำเนินการ ที่ค่า c ที่เท่ากัน ซึ่งสามารถอธิบายได้จากการเพิ่มขึ้นของจำนวนวีดิทัศน์นั้นส่งผลให้ขนาดของ Transition matrix $\widetilde{W}$, $\vec{r}_i$ และ $\vec{e}_i$ ในสมการ RWR ใหญ่ขึ้น ระยะเวลาในการคูณเมทริกซ์ในสมการดังกล่าวจึงทำให้ในภาพรวมเพิ่มขึ้น เพื่อดำเนินการอัลกอริทึมให้แล้วเสร็จ



ภาพที่ 4.18 กราฟแสดงการเปลี่ยนแปลงของเวลาของ RWR+CDF_II เมื่อเปลี่ยน N

ภาพที่ 4.18 แสดงถึงเวลาที่ใช้ในการดำเนินการที่มีแนวโน้มเพิ่มสูงขึ้น เมื่อมีจำนวนของวีดิทัศน์ที่ใช้คำนวณด้วยอัลกอริทึม RWR+CDF_II ที่ใหญ่ขึ้น โดยกราฟนี้ถูกสร้างขึ้นโดยใช้ค่าเฉลี่ยของเวลาที่ใช้ ณ ค่า c ทุกค่า ในอัลกอริทึม RWR+CDF_II การค้นพบข้อมูลของระยะเวลาที่ใช้ในการดำเนินการนี้ ถือเป็นหลักฐานว่าภาระการทำงานเชิงเวลา หรือ Big-O Notation ของอัลกอริทึม RWR+CDF_II ไม่เปลี่ยนแปลงไป การรวม CDF เข้ามาเป็นส่วนหนึ่งกับ RWR ไม่มีผลกระทบกับเวลาที่ใช้ในการทำ RWR เดิม ภาระการทำงานเชิงเวลายังคงสัมพันธ์กับ N แบบเส้นตรงคือ O(N) เช่นเดิม ซึ่งแตกต่างจากวิธี MMR ที่ไม่เพียงแต่เปลี่ยนแปลงตามจำนวนของวีดิทัศน์ทั้งหมดเท่านั้น แต่ยังขึ้นกับจำนวนของรายการแนะนำที่ได้รับเลือก K จำนวนด้วย

แต่ข้อมูลจากการดำเนินการทั้งหมดพบว่า ระยะเวลาที่ใช้รวมทั้งหมดของการสร้างรายการแนะนำวีดิทัศน์ ตั้งแต่ขั้นตอนการเตรียมข้อมูล การค้นหาวีดิทัศน์ การสร้างรายการแนะนำ ยังคงสูงกว่าเวลาที่ใช้ในส่วนของการเรียงลำดับใหม่ โดยระยะเวลาในการสร้างรายการแนะนำวีดิทัศน์จำนวน 25 ราย เฉลี่ยใช้เวลา 2 นาที โดยส่วนใหญ่นั้น จะใช้เวลามากที่การคำนวณคะแนนความคล้ายกันของโคไซน์ ซึ่งใช้เวลาเฉลี่ยถึง 1.57 นาที ซึ่งคิดเป็นร้อย 75 ของระยะเวลาทั้งหมด ในขณะที่การเรียงลำดับใหม่โดยใช้วิธีเอ็มเอ็มอาร์นั้น ใช้มากสุดคือ 1 วินาทีสำหรับการเรียงลำดับ 25 รายการ ซึ่งการสร้างรายวีดิทัศน์ในส่วนแรก ใช้เวลาเป็น 120 เท่าของการเรียงลำดับใหม่ อาจกล่าวได้ว่าระยะเวลาส่วนใหญ่ใช้ไปกับการเตรียมรายการวีดิทัศน์เพื่อทำการเรียงลำดับใหม่ อย่างไรก็ตาม ในการ

สร้างรายการแนะนำ จำเป็นที่จะต้องเปลี่ยนแปลงข้อมูลตลอดเวลา เพื่อเพิ่มโอกาสให้วีดิทัศน์ใหม่ ๆ ที่อัปโหลดในยูทูบถูกพิจารณาเพิ่มเติมในรายการที่จะแนะนำด้วย ซึ่งในการสร้างรายการแนะนำในครั้งต่อมานั้น ระบบไม่จำเป็นต้องคำนวณคะแนนความคล้ายกันของโคไซน์ของวีดิทัศน์ที่เคยคำนวณไว้แล้ว ส่งผลให้ความสัมพันธ์ของระยะเวลาที่ใช้ในการสร้างรายการแนะนำกับจำนวนวีดิทัศน์ทั้งหมดลดลง เปลี่ยนเป็นระยะเวลาที่ใช้จะเปลี่ยนแปลงตามจำนวนวีดิทัศน์ใหม่ที่เพิ่มเข้ามาในการคำนวณ ในขณะที่เวลาในการเรียงลำดับใหม่ยังคงเปลี่ยนแปลงตามจำนวนรายการวีดิทัศน์ทั้งหมด จากการคำนวณอัตราส่วนระยะเวลาของวิธี MMR และ RWR+CDF_II พบว่า เมื่อเพิ่มจำนวนรายการวีดิทัศน์เป็น 200 รายการ ซึ่ง 200 รายการนี้ ใช้เวลาในการสร้างรายการแนะนำวีดิทัศน์ก่อนการเรียงลำดับใหม่ เพียง 1 นาที เนื่องจากมีวีดิทัศน์ใหม่เพียง 10 รายการ โดยวิธี MMR ใช้เวลาในการเรียงลำดับผลลัพธ์ใหม่สำหรับ 200 รายการนานถึง 9.12 นาที (ทดสอบด้วยเอกสารเพียง 1 ชิ้น ไม่ใช่ค่าเฉลี่ย) เนื่องจากอัตราการเติบโตที่เป็น $O(N^3)$ สำหรับการสร้างลำดับการแนะนำทั้ง 200 รายการ (K = 200) ซึ่งใกล้เคียงกับระยะเวลาที่คำนวณได้ทางทฤษฎี อ้างอิงข้อมูลจากค่าเฉลี่ยในการดำเนินการที่ 25 รายการคือ 1 วินาที โดยค่าทางทฤษฎีที่คำนวณได้เท่ากับ 8.56 นาที ในขณะที่วิธี RWR+CDF_II ใช้เวลาเพียง 1.12 วินาทีเท่านั้นสำหรับวีดิทัศน์ 200 รายการเช่นเดียวกัน เทียบกับค่าที่คำนวณทางทฤษฎีเท่ากับ 0.8 วินาที ซึ่งเร็วกว่าถึง 540 เท่า ทางผู้วิจัยได้ทำการทดลองเพิ่มเติม โดยการเรียงลำดับใหม่สำหรับรายการแนะนำ 25 รายการ (K = 25) จากรายการวีดิทัศน์ทั้งหมด 200 รายการ พบว่า MMR ใช้เวลาทั้งหมด 1.28 นาที ส่วนวิธี RWR+CDF_II ใช้เวลา 1.11 วินาที ไม่ต่างจากการเรียงลำดับ 200 รายการ จากข้อมูลดังกล่าว จึงได้ทดลองกับการสร้างรายการแนะนำ 1 รายการเท่านั้น (K = 1) จากการเรียงลำดับด้วยวิธี MMR พบว่าใช้เวลาถึง 3 วินาที ซึ่งยังคงใช้เวลามากกว่าวิธี RWR+CDF_II

จากข้อมูลในการทดลองจะเห็นว่า ที่จำนวนวีดิทัศน์ที่เพิ่มขึ้น หรือมีการเปลี่ยนแปลงระยะเวลาที่ใช้ในส่วนของการสร้างรายการวีดิทัศน์ก่อนการเรียงลำดับใหม่นั้น มีบทบาทต่อระยะเวลารวมน้อยลง เนื่องจากไม่จำเป็นต้องคำนวณคะแนนความคล้ายกันของโคไซน์ใหม่สำหรับรายการวีดิทัศน์ที่เคยคำนวณไว้แล้ว ส่งผลให้ระยะเวลาส่วนใหญ่มีการเติบโตจากการเรียงลำดับใหม่ ซึ่งสัมพันธ์กับจำนวนวีดิทัศน์ทั้งหมดแทน การที่วิธี RWR+CDF_II ใช้เวลาที่น้อยกว่าวิธี MMR ที่จำนวนรายการวีดิทัศน์สูง ๆ ทำให้วิธี RWR+CDF_II เป็นอีกหนึ่งวิธีที่สามารถนำมาพิจารณาเพื่อใช้ในการเพิ่มความหลากหลายของการแนะนำ โดยที่ให้ค่าความหลากหลายจากการวัดด้วยอัลฟ่าเอ็นดีซีจีไม่ต่างจากวิธี MMR ซึ่งวิธีที่ดีที่สุด

Chulalongkorn University

บทที่ 5
บทสรุปการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้เสนอวิธีการในการเพิ่มความครอบคลุมแนวคิดของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน โดยใช้การเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่ รวมทั้งได้เสนอวิธีการและขั้นตอนในการสร้างระบบแนะนำวีดิทัศน์จากยูทูบ จากการใช้เอกสารของวิชาเรียนในระบบจัดการเรียนการสอนมายคอร์สวิลล์ โดยอาศัยข้อมูลที่เป็นข้อความจากเนื้อหาในเอกสาร และจากชื่อ คำอธิบายและแถบป้ายบอกข้อมูลของวีดิทัศน์ จากการที่งานวิจัยนี้ได้พิจารณาความครอบคลุมแนวคิดให้เทียบเท่ากับความหลากหลาย (ของแนวคิด) ของผลการแนะนำ ผู้วิจัยจึงเลือกการเดินสุ่มแบบเริ่มต้นใหม่ซึ่งเป็นวิธีหนึ่งที่ถูกใช้ในการเพิ่มความหลากหลายของผลการแนะนำ มาพัฒนาต่อโดยการเพิ่มส่วนของฟังก์ชันวัตถุประสงค์ที่ชื่อว่าฟังก์ชันความหลากหลายของหมวดหมู่ จากการทำการทดลองเทียบประสิทธิภาพด้านความหลากหลายโดยใช้ตัวชี้วัดที่ชื่อว่าอัลฟาเอ็นดีซีจี เปรียบเทียบกับวิธีการเรียงลำดับผลการแนะนำใหม่ที่ให้ผลลัพธ์การแนะนำที่ดีที่สุดที่ชื่อว่า เอ็มเอ็มอาร์ รวมทั้งเปรียบเทียบประสิทธิภาพด้านภาระการทำงานเชิงเวลา พบว่าวิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่นั้น ใช้เวลาที่ดีกว่ามาก รวมถึงภาระการทำงานจะไม่เพิ่มขึ้นแม้ว่าจะเปลี่ยนแปลงจำนวนผลลัพธ์การแนะนำเท่าใด ในขณะที่ผลจากการเปรียบเทียบค่าอัลฟาเอ็นดีซีจีด้วยการทดสอบที วิธีการที่เสนอนี้ให้ประสิทธิภาพด้านความหลากหลายไม่แตกต่างจากวิธีเอ็มเอ็มอาร์อย่างมีนัยยะสำคัญ ที่ค่า c = 0.2, 0.5 และ 0.8 จากผลลัพธ์ดังกล่าว การพิจารณาใช้วิธีการเดินสุ่มและฟังก์ชันความหลากหลายของหมวดหมู่ เพื่อเพิ่มความครอบคลุมแนวคิดของของระบบแนะนำวีดิทัศน์ในระบบจัดการเรียนการสอน จึงถือเป็นทางเลือกหนึ่งที่มีภาระการทำงานเชิงเวลาที่น้อยกว่าและให้ผลลัพธ์ด้านความครอบคลุมแนวคิดได้ทัดเทียมกับวิธีที่ดีที่สุดตอนนี้

5.2 ข้อจำกัด

5.2.1 ข้อจำกัดด้านเอกสาร

- เนื่องจากเครื่องมือในการแปลงเอกสารจากรูปแบบพีดีเอฟเป็นข้อความนั้น ส่วนมากยังไม่สนับสนุนภาษาไทย ทำให้การใช้งานยังถูกจำกัดไว้แต่เพียงเอกสารภาษาอังกฤษเท่านั้น

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

- เอกสารที่ใช้ในการสร้างระบบแนะนำนั้น บางส่วนเป็นการสร้างมาจากรูปภาพ ทำให้ไม่สามารถแปลงเป็นข้อความได้
- ชื่อของเอกสาร หลายเอกสารไม่ได้ตั้งชื่อเกี่ยวข้องกับเนื้อหาภายใน ทำให้การค้นหาวีดิทัศน์จากยูทูบที่มีส่วนของชื่อเอกสารด้วยนั้น เกิดความยากลำบาก และได้ผลลัพธ์ที่ไม่แม่นยำ

5.2.2 ข้อจำกัดของระบบแนะนำวีดิทัศน์

- การสกัดคำสำคัญนั้น บางครั้งก็ได้คำที่น้อยเกินไป ทำให้ค้นหาวีดิทัศน์ได้หัวข้อที่ไม่หลากหลาย หรือบางครั้งก็ได้คำสำคัญที่ไม่เกี่ยวข้องกับเอกสารเท่าที่ควร
- วีดิทัศน์จากยูทูบ บางรายการมีเนื้อหาที่ดีและมีคุณภาพมาก แต่ไม่มีคำอธิบาย หรือป้ายติดข้อมูล รวมถึงตั้งชื่อของวีดิทัศน์ได้อย่างไม่สมเหตุสมผล ทำให้ต้องตัดรายการเหล่านั้นออกจากการคัดเลือก

5.2.3 ข้อจำกัดของอัลกอริทึม

- การคำนวณค่าโคไซน์ซิมิลาริตี้ระหว่างวีดิทัศน์นั้นใช้เวลานานกว่าการหาระหว่างเอกสารกับวิดิทัศน์ ทำให้ต้องจำกัดขนาดของวีดิทัศน์ที่จะมาใช้ในการทดลองเพียง 25 วีดิทัศน์ต่อ 1 เอกสาร
- จากข้อจำกัดเรื่องจำนวนของวีดิทัศน์นั้น ทำให้ขนาดของเมทริกซ์ที่ใช้คำนวณมีขนาดเล็กส่งผลให้ในหลาย ๆ ครั้ง การลู่เข้าเกิดเร็วเกินไปทำให้ผลลัพธ์ระหว่างการเดินสุ่มแบบเริ่มต้นใหม่ดั้งเดิม ไม่ต่างจากวิธีที่เพิ่มฟังก์ชันความหลากหลายของหมวดหมู่เข้ามา

5.3 ข้อเสนอแนะ

5.3.1 ส่วนของการปรับปรุงระบบแนะนำวีดิทัศน์

ระบบแนะนำวีดิทัศน์สามารถเพิ่มความแม่นยำได้มากกว่าเดิม ถ้าผู้ใช้งานทำการใส่คำสำคัญที่เกี่ยวข้องมาด้วยเมื่อมีการอัปโหลดเอกสาร รวมถึงการกำหนดว่าคำสำคัญไหนที่เป็นแนวคิดหลักของเอกสาร นอกจากนี้ การตั้งชื่อเอกสารให้สัมพันธ์กับเนื้อหาภายใน จะช่วยเพิ่มความถูกต้องให้กับการค้นหาวีดิทัศน์มากขึ้น

5.3.2 ส่วนของการปรับปรุงอัลกอริทึม

ในส่วนของการปรับปรุงอัลกอริทึมนั้น สามารถปรับเปลี่ยนเปลี่ยนฟังก์ชันวัตถุประสงค์ที่ใช้แทนที่จะใช้ฟังก์ชันความหลากหลายของหมวดหมู่ หรือปรับปรุงฟังก์ชันความหลากหลายของหมวดหมู่เดิมให้เพิ่มส่วนของการคิดลำดับในการแนะนำไปด้วย เนื่องจากฟังก์ชันที่ใช้ในงานวิจัยนี้คำนวณจากรายการแนะนำที่ถูกเลือกเท่านั้น ไม่ได้พิจารณาลำดับก่อนหลัง นอกจากฟังก์ชันความหลากหลายของหมวดหมู่แล้ว ส่วนที่สามารถเพิ่มเติมได้คือการใช้อัลกอริทึม B_LIN ซึ่งได้กล่าวไว้ในส่วนของงานวิจัยที่เกี่ยวข้อง อัลกอริทึมนี้เป็นอีกรูปแบบหนึ่งของการเดินสุ่มแบบเริ่มต้นใหม่ ซึ่งสามารถปรับปรุงให้ใช้ในการคำนวณแบบออนไลน์ได้

5.3.3 ส่วนของการประเมินผล

สำหรับการประเมินผลนั้น นอกจากใช้ค่าอัลฟาเอ็นดีซีจีแล้ว ยังสามารถใช้วิธีการประเมินผลสำหรับการพิจารณาความหลากหลายได้อีกหลายวิธีคือ ERR-IA [45] และ S-Recall [46] รวมทั้งการปรับปรุงค่าอัลฟาเอ็นดีซีจี โดยการเปลี่ยนวิธีการกำหนดรายการแนะนำในอุดมคติ จากการคิดตามงานวิจัยที่อ้างอิง อาจจะเปลี่ยนใช้การประเมินจากผู้ใช้งานจริง เพื่อสอบถามความเห็นจากผู้ใช้งานถึงความหลากหลายของผลการแนะนำในอุดมคติ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

รายการอ้างอิง

[1] J. P. Medved. (2015, 19 Nov). *LMS Industry User Research Report*. Available: http://www.capterra.com/learning-management-system-software/user-research

[2] N. Temiyasathit, P. Punyabukkana, and A. Suchato, "Course Periodic Behavior Modelling and Its Application in LMS Activity Prediction," ed, 2016, pp. 1164-1174.

[3] A. M. Kaplan and M. Haenlein, "Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster," *Business Horizons,* vol. 59, pp. 441-450, 7// 2016.

[4] A. Woodie. (2016). *Analyzing Video, the Biggest Data of Them All*. Available: https://www.datanami.com/2016/05/26/analyzing-video-biggest-data/

[5] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*, ed: Springer, 2011, pp. 73-105.

[6] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 257-260.

[7] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal,* vol. 16, pp. 261-273, 2015.

[8] S. A. P. Parambath, N. Usunier, and Y. Grandvalet, "A Coverage-Based Approach to Recommendation Diversity On Similarity Graph," presented at the Proceedings of the 10th ACM Conference on Recommender Systems, Boston, Massachusetts, USA, 2016.

[9] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," presented at the Proceedings of the 21st annual international ACM SIGIR conference on

Research and development in information retrieval, Melbourne, Australia, 1998.

[10] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," presented at the Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 2009.

[11] Sa, #250, l. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," presented at the Proceedings of the 8th ACM Conference on Recommender systems, Foster City, Silicon Valley, California, USA, 2014.

[12] A. Ashkan, B. Kveton, S. Berkovsky, and Z. Wen, "Optimal greedy diversity for recommendation," presented at the Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015.

[13] X. Xia, X. Wang, and X. Zhou, "Evolving Recommender System for Mobile Apps: A Diversity Measurement Approach," *Smart CR,* vol. 3, pp. 139-154, / 2013.

[14] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher*, et al.*, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 659-666.

[15] M. Koniaris, I. Anagnostopoulos, and Y. Vassiliou, "Evaluation of Diversification Techniques for Legal Information Retrieval," *Algorithms,* vol. 10, p. 22, 2017.

[16] J. G. Z. David E Stone, "Learning Management Systems in a Changing Environment," *Handbook of Research on Education and Technology in a Changing Society,* pp. 756-767, 2014.

[17] R. K. Ellis, "Field Guide to Learning Management Systems," 2009.

[18] N. A. Adzharuddin and L. H. Ling, "Learning Management System (LMS) among University Students: Does It Work? f," *International Journal of e-Education, e-Business, e-Management and e-Learning,* vol. 3, p. 248, 2013.

[19] อ. สุชาโต, "โครงการพัฒนาระบบ สนับสนุนการเรียนการสอน ที่เชื่อมต่อกับเครือข่ายสังคมออนไลน์ ระยะที่ 3 เมษายน 2556 – มีนาคม 2557," ed, pp. 1–32.

[20] YouTube. (2014). *Most popular YouTube video categories based on male Millennial user engagement as of December 2013*. Available: https://www.statista.com/statistics/290404/millennials-popular-youtube-video-categories-male/

[21] YouTube. (2014). *Most popular YouTube video categories based on female Millennial user engagement as of December 2013*. Available: https://www.statista.com/statistics/290394/millennials-popular-youtube-video-categories-female/

[22] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, pp. 734-749, 2005.

[23] S. Jain, A. Grover, P. S. Thakur, and S. K. Choudhary, "Trends, problems and solutions of recommender system," in *International Conference on Computing, Communication & Automation*, 2015, pp. 955-958.

[24] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*: Cambridge University Press, 2014.

[25] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, pp. 175-185, 1992.

[26] K. Bradley and B. Smyth, "Improving recommendation diversity," in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, 2001, pp. 85-94.

[27] M.-X. Gan, L. Sun, and R. Jiang, "Trinity: Walking on a User-Object-Tag Heterogeneous Network for Personalised Recommendations," *Journal of Computer Science and Technology,* vol. 31, pp. 577-594, 2016.

[28] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *NAACL HLT 2007 - Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 2007, pp. 97-104.

[29] T. Bogers, "Movie recommendation using random walks over the contextual graph," in *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems*, 2010.

[30] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335-336.

[31] R. L. T. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," *WWW,* 2010.

[32] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.

[33] H. Tong, J. He, Z. Wen, R. Konuru, and C.-Y. Lin, "Diversified ranking on large graphs: an optimization viewpoint," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1028-1036.

[34] J. He, H. Tong, Q. Mei, and B. K. Szymanski, "GenDeR: A generic diversified ranking algorithm," in *Advances in Neural Information Processing Systems*, 2012, pp. 1142-1150.

[35] S. Gollapudi and A. Sharma, "An axiomatic approach for result diversification," in *WWW'09 - Proceedings of the 18th International World Wide Web Conference*, 2009, pp. 381-390.

[36] P. Fraternali, D. Martinenghi, and M. Tagliasacchi, "Top-k bounded diversification," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 421-432.

[37] A. N. Langville and C. D. Meyer, "Deeper inside pagerank," *Internet Mathematics,* vol. 1, pp. 335-380, 2004.

[38] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 653-658.

[39] C. Faloutsos, G. Miller, and C. B. Tsourakakis, "Large graph-mining: Power tools and a practitioner's guide," in *The 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Paris*, 2009.

[40] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," 2006.

[41] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*: O'Reilly Media, Inc., 2009.

[42] G. Piao and J. G. Breslin, "Interest Representation, Enrichment, Dynamics, and Propagation: A Study of the Synergetic Effect of Different User Modeling Dimensions for Personalized Recommendations on Twitter," in *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, Eds., ed Cham: Springer International Publishing, 2016, pp. 496-510.

[43] R. Brown. (2015). *Django vs Flask vs Pyramid: Choosing a Python Web Framework*. Available: https://www.airpair.com/python/posts/django-flask-pyramid

[44] G. B. Limentani, M. C. Ringo, F. Ye, M. L. Bergquist, and E. O. MCSorley, "Beyond the t-test: statistical equivalence testing," ed: ACS Publications, 2005.

[45] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 621-630.

[46] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 10-17.

ภาคผนวก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียนวิทยานิพนธ์

นายณัฐนันท์ ฉันทานุรักษ์ เกิดวันอาทิตย์ที่ 12 พฤศจิกายน พ.ศ. 2532 สำเร็จการศึกษาระดับมัธยมศึกษาที่ โรงเรียนเบญจมราชูทิศ ราชบุรี จังหวัดราชบุรี สำเร็จการศึกษาระดับปริญญาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร ในปีการศึกษา 2554

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University