การกรองด้วยความร่วมมือโดยใช้ระยะทาง

DISTANCE BASED COLLABORATIVE FILTERING

นายจรุพงษ์ นาคนพคุณ
นายอาณัติ ระฆังสมบูรณ์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชา วิทยาการคอมพิวเตอร์ วิทยาศาสตร์บัณฑิต
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2552

DISTANCE BASED COLLABORATIVE FILTERING

Mr. JARUPONG NARKNOPPAKOON

Mr. ANAT RAKHUNGSOMBOON

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIRMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2009

หัวข้อปัญหาพิเศษ	การกรองด้วยความร่วมมือโดยใช้ระยะทาง Distance Based Collaborative Filtering	
ชื่อนักศึกษา	นายจรุพงษ์ นาคนพคุณ	49050025
	นายอาณัติ ระฆังสมบูรณ์	49050362
ปริญญา	วิทยาศาสตรบัณฑิต	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ประจำปีการศึกษา 2552

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ดร.จีรพร วีระพันธุ์	
อ.วีระชัย ตันยะสิทธิ์	
รศ.ดร.วีระ บุญจริง	

หัวข้อปัญหาพิเศษ	การกรองด้วยความร่วมมือโดยใช้ระยะทาง	
ชื่อนักศึกษา	นายจรุพงษ์ นาคนพคุณ	49050025
	นายอาณัติ ระฆังสมบูรณ์	49050362
ปริญญา	วิทยาศาสตรบัณฑิต	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
ปีการศึกษา	2552	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

บทคัดย่อ

ปัญหาพิเศษนี้เป็นการศึกษาทดลองเรื่องการนำวิธีการวัดระยะทางมาใช้ในการกรองด้วยความร่วมมือ โดยนำการวัดระยะทางแบบมินโคว์สคิอันดับพีเข้ามาใช้ในการคำนวณ ปัญหาพิเศษนี้ได้ทำการทดลองกับชุดข้อมูลจาก Movielens เพื่อวัดความแม่นยำของการทำนายการให้คะแนนด้วยค่าเฉลี่ยความผิดพลาดสัมบูรณ์เมื่อทำการเปลี่ยนแปลงค่าอันดับพีของวิธีการวัดระยะทางแบบมินโคว์สคิอันดับพี จากการศึกษาและทดลองพบว่าค่าอันดับ p ที่มีประสิทธิภาพมากที่สุดคือค่าอันดับ p ที่ 125 ซึ่งให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ 0.80870

คำสำคัญ : การกรองด้วยความร่วมมือ, การวัดระยะทางแบบมินโคว์สคิอันดับพี

Title	Distance Based Collaborative Filtering		
Students	Mr. Jarupong	Narknoppakoon	49050025
	Mr. Anat	Rakhungsomboon	49050362
Degree	Bachelor of Science		
Major Program	Computer Science		
Academic Year	2009		
Advisor	Assoc. Prof. Dr. Veera Boonjing		

ABSTRACT

This special problem is an experimental study on distance based collaborative filtering using Minkowski p-norm distance. The special problem made experiments on Movielens dataset to measure prediction accuracy using Mean Absolute Error when adjust the level of p-norm in Minkowski p-norm distance. The experimental result shows that the best level of p-norm is 125, MAE value is 0.80870.

Keywords : Collaborative Filtering, Minkowski p-norm distance

กิตติกรรมประกาศ

ในการจัดทำปัญหาพิเศษนี้ คณะผู้จัดทำขอขอบพระคุณบุคคลสำคัญหลายท่านที่คอยช่วยเหลือให้ข้อมูลและคำแนะนำต่างๆ อันได้แก่ บิดา มารดา และญาติพี่น้อง ผู้คอยให้กำลังใจ และให้โอกาสทางการศึกษา ขอขอบพะคุณ รศ.ดร.วีระ บุญจริง อาจารย์ที่ปรึกษา ที่คอยให้คำแนะนำชี้แนะถึงปัญหา เสนอแนะแนวทางและวิธีการในการดำเนินงานต่างๆ รวมถึง อ.วีระชัย ตันยะสิทธิ์ และ ผศ.ดร.จีรพร วีระพันธุ์ ที่มีส่วนช่วยในการตรวจสอบ แนะนำ แก้ไขข้อผิดพลาดและให้แนวคิดในการดำเนินงาน สุดท้ายนี้ต้องขอขอบคุณเพื่อนๆ และรุ่นพี่ รวมถึงบุคคลทุกท่านที่ช่วยชี้แนะ คอยให้กำลังใจ และสนับสนุนการทำปัญหาพิเศษนี้ให้สำเร็จลุล่วงไปได้ด้วยดี

คณะผู้จัดทำ

สารบัญ

สารบัญ(ต่อ)

สารบัญตาราง

สารบัญตาราง(ต่อ)

สารบัญรูป

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

ระบบช่วยแนะนำ เป็นระบบที่สร้างขึ้นมาเพื่อช่วยเหลือผู้ใช้โดยแนะนำข้อมูลที่คาดว่าผู้ใช้น่าจะสนใจ ข้อมูลที่นำมาแนะนำนี้ได้มาจากการสอบถามและการเก็บข้อมูลของผู้ใช้แต่ละคนที่เข้ามาใช้บริการ โดยข้อมูลจะถูกเก็บไว้ในฐานข้อมูล ซึ่งข้อมูลต่างๆในฐานข้อมูลจะถูกนำมากรองเพื่อสร้างเป็นข้อมูลแนะนำที่ตรงต่อความต้องการของผู้ใช้ โดยทั่วไปเทคนิคที่ใช้ในระบบช่วยแนะนำแบ่งออกเป็น 2 วิธี ได้แก่ การกรองตามเนื้อหา (Content-based Filtering) และการกรองด้วยความร่วมมือ (Collaborative Filtering) โดยวิธีการกรองตามเนื้อหานั้นจะสร้างการแนะนำโดยคำนวณความเหมือนกันของผลิตภัณฑ์จากองค์ประกอบหลักของผลิตภัณฑ์นั้นๆ เช่น การแนะนำภาพยนตร์อาจใช้ ชื่อภาพยนตร์ ชื่อผู้กำกับ ชื่อนักแสดง ประเภทของภาพยนตร์ ในการคำนวณเพื่อสร้างการแนะนำ แต่วิธีการกรองด้วยความร่วมมือจะมีการเก็บข้อมูลคะแนนความชอบของผู้ชมแต่ละคนตามภาพยนตร์เรื่องต่างๆ และสร้างการแนะนำจากข้อมูลเหล่านี้แทน

การกรองด้วยความร่วมมือนั้นจำเป็นต้องทราบข้อมูลเบื้องต้นของผู้ใช้ก่อนว่าแต่ละคนมีความชอบต่อข้อมูลใดบ้าง ในงานวิจัยนี้ ได้นำฐานข้อมูลของการให้คะแนนภาพยนต์มาใช้ ดังนั้นข้อมูลเบื้องต้นจึงเป็นข้อมูลการให้คะแนนภาพยนต์เรื่องต่างๆ ในลักษณะของคะแนนเรตติ้ง จากนั้นจึงนำข้อมูลของผู้ใช้มาทำการคำนวณหาความสัมพันธ์ โดยมีหลักการที่สำคัญคือ มีความเป็นไปได้ที่ผู้ใช้สองคนได้แก่ ผู้ใช้เป้าหมาย ซึ่งเป็นผู้ใช้คนที่เราสนใจจะแนะนำข้อมูลให้ และผู้ใช้อ้างอิง ซึ่งเป็นผู้ใช้คนอื่นๆที่นำมาเปรียบเทียบกับผู้ใช้เป้าหมายนั้นมีพฤติกรรมความชอบในภาพยนตร์คล้ายกันหรือให้คะแนนความชอบภาพยนตร์เรื่องเดียวกันไว้ใกล้เคียงกันในหลายๆเรื่องนั้น น่าจะมีความชอบในภาพยนตร์เรื่องอื่นๆ เหมือนกันต่อไป ทำให้สามารถทำนายการให้คะแนนภาพยนต์ในเรื่องที่ผู้ใช้เป้าหมายยังไม่เคยให้คะแนนมาก่อนได้ โดยดูจากการให้คะแนนของผู้ใช้อ้างอิงที่มีการให้คะแนนคล้ายกันกับผู้ใช้เป้าหมาย

ในงานวิจัยนี้ได้นำการวัดระยะทางแบบมินโคว์สคิอันดับพี (Minkowski p-norm distance) เข้ามาใช้ในการคำนวณหาผู้ใช้อ้างอิงที่มีการให้คะแนนไว้ใกล้เคียงกัน โดยจะทำการทดลองปรับค่าอันดับ p ของวิธีการวัดระยะทาง เพื่อทำการทดสอบประสิทธิภาพของการทำนายการให้คะแนน โดยจะพิจารณาจากค่าเฉลี่ยความผิดพลาดสัมบูรณ์ (Mean Absolute Error : MAE)

1.2 วัตถุประสงค์ของงานวิจัย

ศึกษาการวัดระยะทางแบบมินโคว์สคิอันดับพี โดยทดลองปรับค่าอันดับ p เพื่อทดสอบประสิทธิภาพในการทำนายว่าที่อันดับ p ใดๆ จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด

1.3 ขอบเขตของงานวิจัย

สร้างแบบจำลองของขั้นตอนวิธีโดยทดลองกับชุดข้อมูลจาก Movielens เพื่อหาสมาชิกใกล้ที่สุด 5 รายโดยใช้การวัดระยะทางแบบมินโคว์สคิอันดับพีในการหาระยะห่างระหว่างข้อมูลคะแนนของผู้ใช้เป้าหมายที่สนใจ กับข้อมูลของผู้ใช้อ้างอิง แล้วนำข้อมูลคะแนนของผู้ใช้อ้างอิงที่มีระยะทางที่ใกล้ที่สุดจำนวน 5 รายมาหาค่าเฉลี่ย จะได้คะแนนที่ใช้ในการทำนาย ในกรณีที่ไม่สามารถหาผู้ใช้อ้างอิงได้เนื่องจากการให้คะแนนของผู้ใช้เป้าหมายไม่มีความสัมพันธ์กับผู้ใช้คนอื่นๆเลยจะทำให้ไม่สามารถใช้วิธีการวัดระยะทางได้ และกรณีที่จำนวนผู้ใช้อ้างอิงที่มีความสัมพันธ์กับผู้ใช้เป้าหมายมีน้อยกว่าจำนวนสมาชิกใกล้ที่สุดที่กำหนด จะทำนายการให้คะแนนโดยใช้คะแนนเฉลี่ยของภาพยนตร์เป้าหมายที่ผู้ใช้ทุกคนเคยให้ไว้และคะแนนเฉลี่ยของผู้ใช้เป้าหมายที่เคยให้ไว้กับภาพยนตร์ทุกเรื่องมาใช้ในการทำนายแทน จากนั้นจึงทำการวัดประสิทธิภาพของการทำนายการให้คะแนนโดยใช้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ เพื่อทดสอบว่าที่ค่าอันดับ p ใด จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในการทำนายน้อยที่สุด

1.4 ส่วนประกอบของปัญหาพิเศษ

ปัญหาพิเศษนี้ ประกอบด้วยส่วนสำคัญ 4 ส่วน ได้แก่

- บทที่ 2 ระบบช่วยแนะนำ, การวัดระยะทาง, วิธีหาสมาชิกใกล้ที่สุด, การทำนายการให้คะแนน, การวัดประสิทธิภาพด้วยค่าเฉลี่ยความผิดพลาดสัมบูรณ์
- บทที่ 3 การออกแบบการทดลองการกรองด้วยความร่วมมือโดยใช้ระยะทาง การวัดประสิทธิภาพของการทำนาย พร้อมทั้งตัวอย่างการคำนวณ
- บทที่ 4 วิธีการที่ใช้ในการทดลอง ผลการทดลอง และการเปรียบเทียบประสิทธิภาพของการกรองด้วยความร่วมมือโดยใช้ระยะทาง
- บทที่ 5 สรุปผลการทดลอง และข้อเสนอแนะเพิ่มเติมของปัญหาพิเศษนี้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ระบบช่วยแนะนำ

ระบบช่วยแนะนำ เป็นระบบที่ถูกสร้างขึ้นมาเพื่อช่วยเหลือผู้ใช้ โดยการแนะนำข้อมูลที่คาดว่าผู้ใช้น่าจะสนใจ และตรงกับความต้องการให้กับผู้ใช้ ซึ่งระบบช่วยแนะนำสามารถเรียนรู้ข้อมูลของผู้ใช้ และสามารถแนะนำข้อมูลที่ผู้ใช้ต้องการจากข้อมูลทั้งหมดที่มีอยู่

เทคนิคที่นิยมใช้ในระบบช่วยแนะนำส่วนใหญ่ จะแบ่งออกเป็น 2 ประเภท ได้แก่

1. การกรองตามเนื้อหา
2. การกรองด้วยความร่วมมือ

2.1.1 การกรองตามเนื้อหา

การกรองตามเนื้อหาเป็นวิธีแนะนำข้อมูลให้กับผู้ใช้โดยดูจากความสัมพันธ์ของรายละเอียดข้อมูลของผลิตภัณฑ์มาเปรียบเทียบกับสิ่งที่ผู้ใช้ต้องการ เพื่อนำมาเป็นข้อมูลที่จะแนะนำผลิตภัณฑ์ให้กับผู้ใช้ต่อไป การสร้างคำแนะนำจะเริ่มจากการเปรียบเทียบประวัติผู้ใช้กับรายละเอียดต่างๆของผลิตภัณฑ์โดยรายละเอียดของผลิตภัณฑ์สามารถเก็บให้อยู่ในรูปของคำศัพท์ ซึ่งสามารถสกัดได้จากการนำเอกสารต่างๆ ไปผ่านกระบวนการวิเคราะห์คำ จากนั้นจะนำคำศัพท์ที่ได้มาสร้างดัชนีเพื่อใช้ในการค้นหาข้อมูลตามผู้ใช้ที่ต้องการ จากนั้นทำการให้คะแนนและจัดลำดับผลลัพธ์ที่ได้จากการค้นหาตามความเกี่ยวข้องกันของข้อมูล ว่าตรงกับที่ผู้ใช้ต้องการมากเท่าไร เพื่อแสดงผลลัพธ์ให้ตรงกับความต้องการของผู้ใช้ให้มากที่สุด

สำหรับการกรองตามเนื้อหา โดยทั่วไปจะวิเคราะห์ประเภทของเนื้อหาแบบหยาบๆ จึงทำให้ในบางระบบก็ไม่สามารถนำไปใช้งานได้อย่างมีประสิทธิภาพ อีกทั้งยังไม่สามารถนำไปใช้ในระบบที่เกี่ยวข้องกับมัลติมีเดียซึ่งมีทั้งตัวอักษร รูปภาพ และเสียงรวมกันอยู่ และยังมีปัญหาอีกอย่างคือ ผู้ใช้สามารถเห็นการแนะนำได้มากที่สุดตามจำนวนที่ระบบได้กำหนดไว้เท่านั้น

2.1.2 การกรองด้วยความร่วมมือ

การกรองด้วยความร่วมมือ เป็นวิธีแนะนำข้อมูลให้กับผู้ใช้ โดยดูจากความสัมพันธ์กันของกลุ่มผู้ใช้ สำหรับการสร้างคำแนะนำจะมีหลักการคร่าวๆ คือ เริ่มจากการนำคะแนนที่ผู้ใช้เคยให้ไว้กับผลิตภัณฑ์มาเปรียบเทียบกับคะแนนของผู้ใช้คนอื่นๆที่ให้ไว้กับผลิตภัณฑ์นั้นๆร่วมกัน โดยดูว่าคะแนนของผู้ใช้คนใดคล้ายคลึงกับคะแนนของผู้ใช้เป้าหมายมากที่สุด จากนั้นจึงนำคะแนนของผู้ใช้คนนั้นมาเป็นคะแนนอ้างอิงเพื่อใช้ทำนายการให้คะแนนแก่ผู้ใช้เป้าหมาย ในกรณีที่ยังไม่เคยให้

คะแนนกับผลิตภัณฑ์นั้น หรือยังไม่เคยใช้งานผลิตภัณฑ์นั้นมาก่อน ซึ่งเมื่อพบว่าผลิตภัณฑ์ใดที่มีผลคะแนนการทำนายออกมาสูง แสดงว่าผู้ใช้เป้าหมายน่าจะชอบ หรือสนใจในผลิตภัณฑ์นั้นๆ จึงนำผลิตภัณฑ์นั้นมาแนะนำให้แก่ผู้ใช้

2.1.3 การวัดระยะทาง

การวัดระยะทาง เป็นแนวคิดเกี่ยวกับการวัดระยะห่างระหว่างจุดสองจุดในปริภูมิเวกเตอร์เชิงเส้น สามารถทำได้โดยใช้นอร์ม (Norm) ระยะระหว่างจุดสองจุด โดยx และ y นิยามเป็นฟังก์ชันสเกลาร์ [3]

$$\rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$$

การวัดระยะแบบมินโคว์สคิอันดับพี

การวัดระยะแบบมินโคว์สคิอันดับพี จะมีการเปลี่ยนแปลงค่าอันดับ p เมื่อค่าอันดับ p มีการเปลี่ยนแปลงค่าระยะทางที่คำนวณได้จะเปลี่ยนไปด้วย

p-norm distance

$$\rho(\mathbf{x},\mathbf{y})_p = \sqrt[p]{\sum_{i=1}^{n}|x_i-y_i|^p} = \left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$$

1-norm distance (Manhattan distance)

$$\rho(\mathbf{x},\mathbf{y})_1 = \sum_{i=1}^{n}|x_i-y_i|$$

2-norm distance (Euclidean distance)

$$\rho(\mathbf{x},\mathbf{y})_2 = \sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} = \left(\sum_{i=1}^{n}|x_i-y_i|^2\right)^{1/2}$$

∞-norm distance (Chebyshev distance)

$$\rho(\mathbf{x},\mathbf{y})_\infty = \lim_{p\to\infty}\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} = \max_{i=1}^{n}\{|x_i - y_i|\}$$
$$= \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\}$$

2.1.4 วิธีหาสมาชิกใกล้ที่สุด (k-Nearest Neighbors)

วิธีหาสมาชิกใกล้ที่สุดเป็นเทคนิคที่เหมาะสำหรับงานการจำแนกประเภทข้อมูล (Classification) [5] ในการหาสมาชิกใกล้ที่สุดนั้นต้องระบุค่าตัวเลขจำนวนเต็มบวก ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณีที่จะต้องค้นหาในการทำนายกรณีใหม่

สำหรับวิธีหาสมาชิกใกล้สุดนั้น มีขั้นตอนการดำเนินการดังนี้

1. กำหนดค่า k ซึ่งเป็นจำนวนสมาชิกใกล้สุดที่ต้องการหา
2. คำนวณหาระยะห่างระหว่างข้อมูลที่สนใจกับข้อมูลอื่นๆ ทุกตัว
3. เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ

วิธีหาสมาชิกใกล้ที่สุดเป็นวิธีที่ค่อนข้างใช้เวลาในการคำนวณบนคอมพิวเตอร์สูง เพราะเวลาที่ใช้สำหรับการคำนวณจะเพิ่มขึ้นแบบแฟคทอเรียลตามจำนวนจุดทั้งหมด

2.1.5 การทำนายการให้คะแนน

เป็นการทำนายคะแนนที่คาดว่า ผู้ใช้จะให้ต่อผลิตภัณฑ์ใดผลิตภัณฑ์หนึ่ง ที่ไม่เคยให้คะแนนมาก่อนโดยจะแบ่งออกเป็น 3 กรณีดังต่อไปนี้

2.1.5.1 การทำนายคะแนนในกรณีปรกติ

เป็นการทำนายการให้คะแนนในกรณีที่สามารถหาความสัมพันธ์ของการให้คะแนนระหว่างผู้ใช้เป้าหมายและผู้ใช้อ้างอิงจากการวัดระยะทางได้ และมีจำนวนผู้ใช้อ้างอิงไม่น้อยกว่าจำนวนสมาชิกใกล้สุดที่กำหนด จะทำการทำนายการให้คะแนนโดยหาค่าเฉลี่ยคะแนนของกลุ่มผู้ใช้อ้างอิงที่ได้ให้คะแนนไว้แล้วจากสมาชิกใกล้สุด ดังสมการ

$$P = \frac{\sum_{i=1}^{N} R_i}{k}$$

โดยที่

P คือค่าคะแนนที่คาดว่าผู้ใช้เป้าหมายจะให้แก่ผลิตภัณฑ์หรือชิ้นข้อมูลเป้าหมาย

R_i คือค่าคะแนนที่ผู้ใช้อ้างอิงคนที่ i ให้แก่ผลิตภัณฑ์หรือชิ้นข้อมูลเป้าหมาย

k คือจำนวนผู้ใช้อ้างอิงที่จะนำมาทำนายการให้คะแนน

2.1.5.2 การทำนายคะแนนในกรณีที่มีจำนวนผู้ใช้อ้างอิงน้อยกว่าจำนวนสมาชิกใกล้สุด

เป็นการทำนายการให้คะแนนในกรณีที่พบว่าผู้ใช้คนอื่นๆ ที่เคยให้คะแนนกับภาพยนต์เป้าหมาย มีน้อยกว่าจำนวนสมาชิกใกล้ที่สุด ทำให้ไม่เป็นไปตามเงื่อนไขของจำนวนสมาชิกใกล้ที่สุดที่กำหนด และไม่จำเป็นต้องใช้การวัดระยะทางเพื่อหาความสัมพันธ์ ในกรณีนี้จะใช้วิธีการทำนายการให้คะแนนอีกวิธีหนึ่ง ดังต่อไปนี้

ส่วนที่ 1 ทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์เรื่องนั้นๆจากการให้คะแนนของผู้ใช้ทุกคนที่เคยให้คะแนนไว้ โดยให้น้ำหนัก 70% ของคะแนนที่จะทำนาย

ส่วนที่ 2 ทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมายคนนั้นๆ เพื่อดูแนวโน้มของการให้คะแนน โดยให้น้ำหนัก 30% ของคะแนนที่จะทำนาย

จากนั้นจึงนำคะแนนทั้งสองส่วนมารวมกันเพื่อทำนายการให้คะแนนของผู้ใช้เป้าหมาย

2.1.5.3 การทำนายคะแนนในกรณีที่ไม่มีผู้ใช้คนอื่นเคยให้คะแนนภาพยนต์เป้าหมาย

เป็นการทำนายการให้คะแนนในกรณีที่พบว่า ภาพยนต์เป้าหมายที่ต้องการจะทำนายการให้คะแนนนั้น ไม่มีผู้ใช้อ้างอิงคนใดเลยที่เคยให้คะแนนไว้ ทำให้ไม่มีข้อมูลที่จะนำมาใช้อ้างอิง และทำนายการให้คะแนนแก่ผู้ใช้เป้าหมายได้ จึงไม่จำเป็นต้องใช้การวัดระยะทางเพื่อหาผู้ใช้อ้างอิงใกล้ที่สุด ในกรณีนี้ จะทำนายโดยอ้างอิงจากข้อมูลการให้คะแนนภาพยนต์ทุกเรื่องที่ผู้ใช้เป้าหมายได้ให้คะแนนไว้ เพื่อดูแนวโน้มของการให้คะแนนโดยการหาค่าเฉลี่ย ซึ่งใช้วิธีเดียวกันกับส่วนที่ 2 ของหัวข้อ *2.1.5.2* แต่ในกรณีนี้ จะคิดเป็น 100% แทน

2.1.6 การวัดประสิทธิภาพด้วยค่าเฉลี่ยความผิดพลาดสัมบูรณ์ (Mean Absolute Error)

ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ เป็นการวัดความถูกต้องทางสถิติของความผิดพลาดในการทำนาย ซึ่งวิธีนี้จะทำการวัดผลการเกิดความผิดพลาดในการทำนาย โดยเปรียบเทียบกับค่าสัมบูรณ์ของผลต่างระหว่างคะแนนจริงของผลิตภัณฑ์ที่ผู้ใช้เคยให้ไว้ในระบบกับคะแนนที่ได้จากการทำนายด้วยวิธีการกรองด้วยความร่วมมือ หลังจากนั้นจึงนำค่าที่ได้มาคำนวณหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์ ของการทำนายทั้งหมดตามสมการ [1, 4]

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

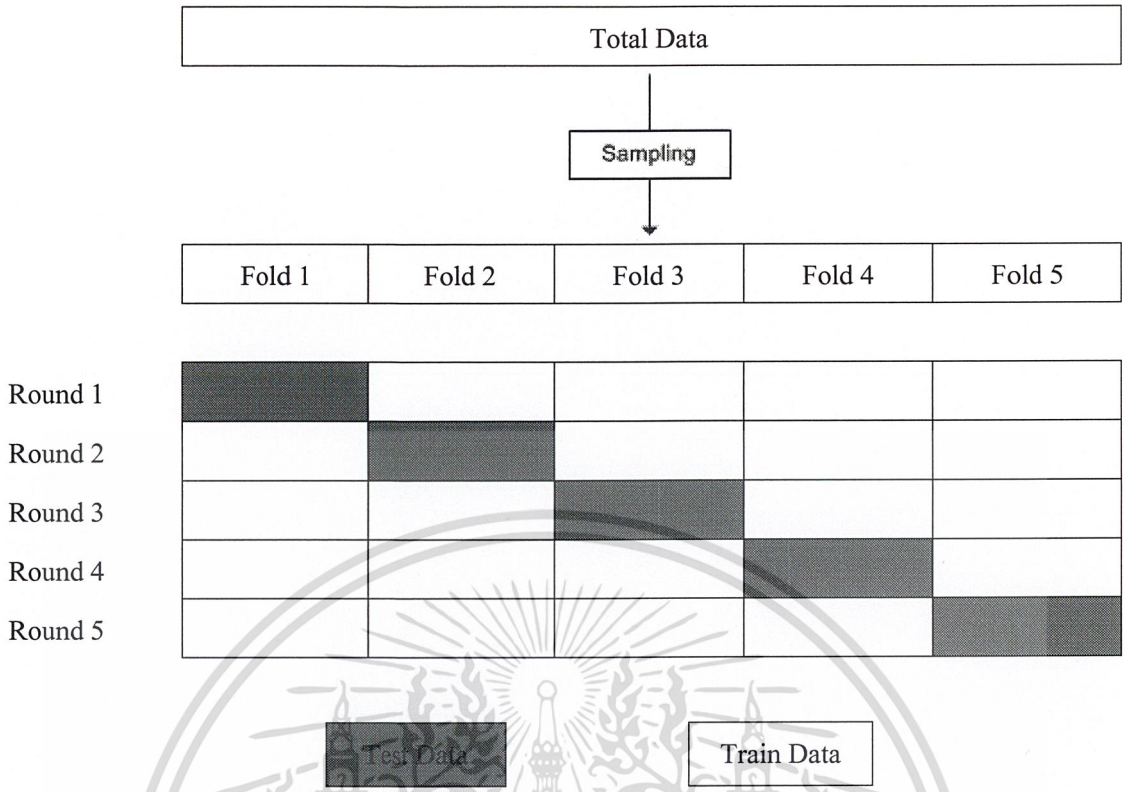
โดยที่ p คือคะแนนที่ได้จากการทำนาย ส่วน r คือคะแนนที่ควรจะเป็นจากชุดข้อมูลทดสอบ และ N คือจำนวนของค่าที่นำมาเปรียบเทียบ

สำหรับค่าเฉลี่ยความผิดพลาดสัมบูรณ์ ยิ่งมีค่าน้อยก็จะยิ่งดี คือค่าคะแนนที่ได้จากการทำนายนั้นจะผิดเพี้ยนไปจากค่าคะแนนจริงเพียงเล็กน้อย ซึ่งเป็นการบ่งบอกถึงประสิทธิภาพที่ดีในการทำนายการให้คะแนน

2.1.7 การวัดประสิทธิภาพแบบไม่เอนเอียง (K-fold cross validation)

Cross validation คือวิธีการตรวจสอบความถูกต้องของแบบจำลองสร้างขึ้น โดยมีพื้นฐานอยู่ที่การสุ่มตัวอย่าง และแบ่งข้อมูลออกเป็นส่วนๆ โดยในการทดลองแต่ละรอบจะมีข้อมูลส่วนหนึ่งถูกเก็บไว้เป็นข้อมูลทดสอบ และข้อมูลส่วนที่เหลือจะเป็นข้อมูลที่ใช้ในการเรียนรู้ของแบบจำลอง [2]

การวัดประสิทธิภาพแบบไม่เอนเอียงจะแบ่งข้อมูลออกเป็น k ส่วน แต่ละส่วนมีจำนวนข้อมูลเท่าๆกัน แล้วนำไปใช้ในการทดลองโดยแต่ละรอบของการทดลอง จะมีข้อมูล 1 ส่วนถูกเก็บไว้เป็นข้อมูลทดสอบ และข้อมูล k - 1 ส่วนเป็นข้อมูลสำหรับการเรียนรู้ของแบบจำลอง และในการทดลองรอบต่อไปจะเปลี่ยนข้อมูลส่วนที่ใช้ในการทดสอบเป็นข้อมูลส่วนถัดไป เมื่อทำการทดลอง k รอบ ข้อมูลทุกส่วนจะถูกนำไปเป็นข้อมูลทดสอบจนครบ



รูปที่ 2.1 การแบ่งข้อมูลออกเป็น 5-fold

2.2 งานวิจัยที่เกี่ยวข้อง

การศึกษาทดลองตัววัดความเหมือนสำหรับการกรองด้วยความร่วมมือ [2]

งานวิจัยนี้มีจุดประสงค์เพื่อจะหาว่า วิธีคำนวณหามาตรวัดความเหมือนทั้ง 4 วิธี ได้แก่ ค่าสัมประสิทธ์สหสัมพันธ์ของเพียร์สัน, ค่าความคล้ายคลึงด้วยโคไซน์, ค่าสัมประสิทธิ์ของไดซ์ และ ค่าสหสัมพันธ์โดยการจัดลำดับของสเปียร์แมน เพื่อให้ทราบว่าวิธีการใดที่มีประสิทธิภาพในการให้ความแม่นยำมากที่สุด โดยการนำเอาชุดข้อมูลทดสอบจาก Movielens มาทำการทดลองตามลำดับขั้นตอนในงานวิจัย ซึ่งวิธีที่ใช้ในการทดสอบความแม่นยำ คือ การเปรียบเทียบค่าความผิดพลาดสัมบูรณ์เฉลี่ย ซึ่งวิธีการใดที่มีความผิดพลาดเกิดขึ้นน้อยที่สุด ก็แสดงให้เห็นว่า วิธีการนั้นเป็นวิธีที่มีประสิทธิภาพในการหาผลลัพธ์ที่แม่นยำมากที่สุดนั่นเอง

การประมวลผลในงานวิจัยนี้ จะได้ค่าของมาตรวัดความเหมือน ที่เป็นเลขจำนวนจริง ซึ่งไม่สามารถนำมาใช้เป็นคำตอบในการอ้างอิงได้ทันที จะต้องมีการปรับค่าตัวเลขก่อน โดยแบ่งออกเป็น 3 วิธี คือ Floor, Round และ Ceil หลังจากทำการทดลองพบว่า การปรับค่าที่ได้ผลดีที่สุด

หรือได้ค่าความผิดพลาดสัมบูรณ์เฉลี่ยน้อยที่สุด คือ การปรับค่าที่ round และพบว่า เมื่อมีการเพิ่มขนาดของผู้ใช้ใกล้เคียงหรือจำนวนผู้ใช้ที่มีพฤติกรรมการให้คะแนนใกล้เคียงกันมากยิ่งขึ้น ก็จะทำให้การคำนวณค่ามีความแม่นยำมากยิ่งขึ้นตามไปด้วย โดยสังเกตุจากค่า ความผิดพลาดสัมบูรณ์เฉลี่ยของแต่ละวิธี จะลดลงตามลำดับ

ผลการเปรียบเทียบผลลัพธ์วิธีการคำนวณทั้ง 4 วิธีที่ได้ทำการทดลองด้วยชุดข้อมูลที่ได้กล่าวมาแล้วข้างต้นนั้น จะเห็นได้ว่าผลลัพธ์ที่แม่นยำที่สุดคือ วิธีหาค่าสัมประสิทธิ์ของไดซ์ซึ่งจากการวิจัยพบว่าเนื่องจากมีการนำเอาเฉพาะโคเรตที่มีคะแนนที่เท่ากันมาคำนวณในสูตร ทำให้ได้ค่ามาตรวัดความเหมือนที่เน้นเฉพาะผู้ใช้ที่มีความคล้ายคลึงกันมากที่สุดของพฤติกรรมการให้เรตติ้งมาใช้ในการคำนวณเหตุนี้จึงทำให้ได้ผลลัพธ์ที่แม่นยำตามไปด้วย ซึ่งสิ่งจำเป็นอีกอย่างหนึ่ง คือ ในชุดข้อมูลนี้มีช่วงคะแนนอยู่ระหว่าง 1 ถึง 5 คะแนน ซึ่งทำให้การจับคู่โคเรตไม่มีความห่างกันมากนัก คู่โคเรตที่เหมือนกันจึงสามารถพบได้ในจำนวนที่เหมาะสมกับวิธีหาค่าสัมประสิทธิ์ของไดซ์ และไม่ว่าจะทดลองด้วยขนาดผู้ใช้ที่ใกล้เคียงขนาด 10 ถึง 70 ก็ตามผลที่ได้ วิธีหาค่าสัมประสิทธิ์ของไดซ์ก็ยังคงมีค่าความผิดพลาดสัมบูรณ์เฉลี่ยที่ต่ำที่สุด ทั้งนี้ขึ้นอยู่กับชุดข้อมูลที่นำมาทดลองด้วย

บทที่ 3
การออกแบบการทดลอง

3.1 ขั้นตอนการทดลอง

3.1.1 ขั้นตอนการเตรียมข้อมูล

ชุดข้อมูลที่ใช้ทดสอบในการทดลองนี้ มาจาก Movielens ซึ่งสามารถดาวน์โหลดได้จาก www.grouplens.org ในชุดข้อมูลที่เลือกใช้ จะประกอบด้วยผู้ใช้จำนวน 943 ราย, ภาพยนต์ 1682 เรื่อง และคะแนนเรตติ้งจำนวน 100,000 ข้อมูล โดยมีระดับการให้คะแนน 1 – 5 คะแนน

การเตรียมข้อมูล จะเริ่มจากการนำชุดข้อมูลที่ดาวน์โหลดมาแล้ว ซึ่งจะอยู่ในรูปของไฟล์ตัวอักษรมาจัดเก็บลงในฐานข้อมูล โดยในการทดลอง จะใช้การวัดประสิทธิภาพแบบไม่เอนเอียงในการแบ่งชุดข้อมูลในการทดลองออกเป็น 5 กลุ่ม หรือ 5 - fold ซึ่งจะมีชุดข้อมูลสำหรับทดสอบ 20 เปอร์เซ็นต์ และชุดข้อมูลสำหรับเรียนรู้ 80 เปอร์เซ็นต์ โดยจะนำข้อมูลหมายเลขของผู้ใช้, รหัสของภาพยนต์ และคะแนนที่ผู้ใช้ได้ให้แก่ภาพยนต์ มาเก็บลงในฐานข้อมูล

3.1.2 ขั้นตอนการทำนายเรตติ้งด้วยการวัดระยะทาง

1. ดึงข้อมูลในฐานข้อมูลที่เป็นชุดทดสอบออกมาครั้งละ 1 แถว แล้วกำหนดให้เป็นผู้ใช้เป้าหมายและภาพยนต์เป้าหมายที่จะทำการทำนายเรตติ้ง

2. ทำการวัดระยะทางของผู้ใช้เป้าหมายกับผู้ใช้คนอื่นๆ ด้วยวิธีการวัดระยะแบบมินโคว์สคิอันดับพี

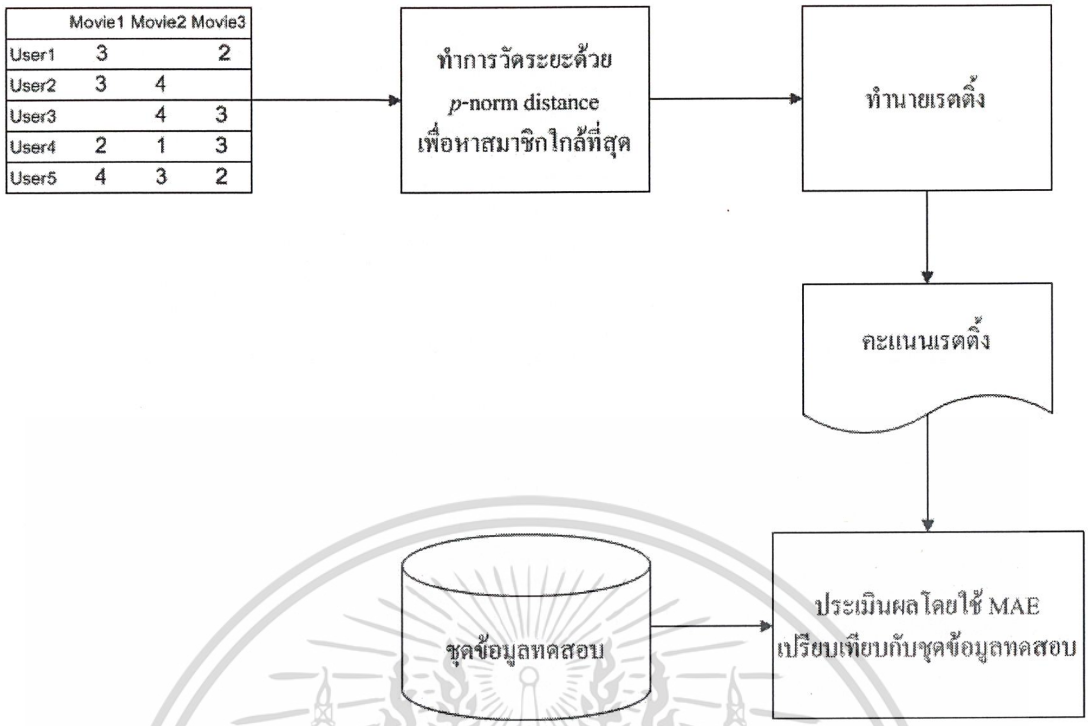
3. เรียงลำดับค่าระยะห่างที่น้อยที่สุดไปหามากด้วย Selection sort

4. เลือกผู้ใช้ที่มีระยะทางห่างจากผู้ใช้เป้าหมายน้อยที่สุดจำนวน k ราย มาทำนายการให้คะแนนของผู้ใช้เป้าหมาย ซึ่งในการทดลองนี้ กำหนดจำนวนสมาชิกใกล้ที่สุด เป็นจำนวน 5 ราย

5. เก็บค่าคะแนนที่ทำนายออกมาลงในฐานข้อมูล

6. ทำขั้นตอนที่ 1 ถึง 5 โดยทำการเปลี่ยน fold ของข้อมูลจนครบทั้งหมด 5 fold

7. ทำขั้นตอนที่ 1 ถึง 6 โดยทำการเปลี่ยนแปลงค่าอันดับ p



รูปที่ 3.1 ภาพรวมการทำงานของการทดลอง

3.1.3 ตัวอย่างการกรองด้วยความร่วมมือโดยการวัดระยะทาง

กระบวนการทำงานของการกรองด้วยความร่วมมือโดยการวัดระยะทาง ในตัวอย่างนี้จะกำหนดจำนวนใกล้ที่สุดเป็น 3 ราย ในขั้นตอนแรกของกระบวนการ จะทำการดึงข้อมูลจากชุดทดสอบมาทีละแถว แล้วกำหนดให้เป็นผู้ใช้เป้าหมาย และภาพยนต์เป้าหมายที่ต้องการจะทำนายคะแนน ดังตัวอย่างต่อไปนี้

ตารางที่ 3.1 ตัวอย่างข้อมูลที่ต้องการทำนายจากชุดข้อมูลทดสอบ

user_id	movie_id	rating
1	6	5

จากตัวอย่าง กำหนดผู้ใช้คนที่ 1 เป็นผู้ใช้เป้าหมาย จากนั้นจึงนำ movie_id ซึ่งเป็นภาพยนต์เป้าหมายที่ได้ ไปตรวจสอบกับข้อมูลการให้คะแนนของของผู้ใช้คนอื่นๆ ในชุดข้อมูลสำหรับการเรียนรู้ ว่ามีผู้ใช้คนใดบ้างที่เคยให้คะแนนภาพยนต์เรื่องที่ 6 เอาไว้แล้วจากนั้นจึงทำการดึงข้อมูลการให้คะแนนภาพยนต์เรื่องต่างๆของผู้ใช้อ้างอิงคนนั้นๆออกมา แล้วทำการตรวจสอบว่ามีเรื่องใดบ้างที่ผู้ใช้เป้าหมายและผู้ใช้อ้างอิงได้ให้คะแนนไว้แล้ว โดยจะแสดงให้เห็นในรูปตารางดังต่อไปนี้

ตารางที่ 3.2 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 18

movie_id / user_id	4	6	8	9	15	22
1	3		1	5	5	4
18	3	5	5	5	4	5

ตารางที่ 3.3 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 71

movie_id / user_id	6	52	153	168	181	197
1		4	3	5	5	5
71	3	4	4	5	3	5

ตารางที่ 3.4 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 76

movie_id / user_id	6	7	42	77	93	137
1		4	5	4	5	5
76	5	4	3	2	4	5

ตารางที่ 3.5 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 181

movie_id / user_id	6	15	18	19	93	105
1		5	4	5	5	2
181	1	3	1	1	1	1

ตารางที่ 3.6 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 198

movie_id / user_id	1	4	6	25	50	55
1	5	3		4	5	5
198	4	3	2	2	5	3

เมื่อได้ข้อมูลการให้คะแนนของผู้ใช้เป้าหมายและผู้ใช้อ้างอิงแล้วจึงนำข้อมูลมาทำการคำนวณด้วยการวัดระยะทางเพื่อหาผู้ใช้อ้างอิงที่มีระยะทางห่างจากผู้ใช้เป้าหมายน้อยที่สุด ดังตัวอย่างการคำนวณต่อไปนี้

ตัวอย่างการวัดระยะทางด้วยค่าอันดับ *p* เป็น 1 (Manhattan distance)

ตารางที่ 3.7 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 18

movie_id / user_id	4	8	9	15	22
1	3	1	5	5	4
18	3	5	5	4	5

จากสูตร

$$\rho(\mathbf{x},\mathbf{y})_1 = \sum_{i=1}^{n} |x_i - y_i|$$

นำข้อมูลไปแทนค่าในสมการ จะได้

$$p(18,1)_1 = (|3-3| + |5-1| + |5-5| + |4-5| + |5-4|)$$

$$= 0 + 4 + 0 + 1 + 1$$

$$= 6$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 18 จะมีค่าเป็น 6

ทำการคำนวณระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนอื่นๆ ดังนี้

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมาย กับผู้ใช้คนที่ 71

ตารางที่ 3.8 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 71

movie_id / user_id	52	153	168	181	197
1	4	3	5	5	5
71	4	4	5	3	5

$$p(71,1)_1 = (|4-4| + |4-3| + |5-5| + |3-5| + |5-5|)$$

$$= 3$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 71 จะมีค่าเป็น 3

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 76

ตารางที่ 3.9 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 76

movie_id / user_id	7	42	77	93	137
1	4	5	4	5	5
76	4	3	2	4	5

$$p(76,1)_1 = (|4-4| + |3-5| + |2-4| + |4-5| + |5-5|)$$

$$= 5$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 76 จะมีค่าเป็น 5

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 181

ตารางที่ 3.10 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 181

movie_id / user_id	15	18	19	93	105
1	5	4	5	5	2
181	3	1	1	1	1

$$p(181,1)_1 = (|3-5| + |1-4| + |1-5| + |1-5| + |1-2|$$
$$= 14$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 181 จะมีค่าเป็น 14

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 198

ตารางที่ 3.11 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 198

movie_id / user_id	1	4	25	50	55
1	5	3	4	5	5
198	4	3	2	5	3

$$p(198,1)_1 = (|4-5| + |3-3| + |2-4| + |5-5| + |3-5|$$
$$= 6$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 198 จะมีค่าเป็น 6

ตัวอย่างการวัดระยะทางด้วยค่าอันดับ p เป็น 2 (Euclidian distance)

ตารางที่ 3.12 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 18

movie_id / user_id	4	8	9	15	22
1	3	1	5	5	4
18	3	5	5	4	5

จากสูตร

$$\rho(\mathbf{x},\mathbf{y})_2 = \sqrt{\sum_{i=1}^{n}|x_i - y_i|^2} = \left(\sum_{i=1}^{n}|x_i - y_i|^2\right)^{1/2}$$

นำข้อมูลไปแทนค่าในสมการ จะได้

$$p(18,1)_2 = \sqrt{(|3-3|^2 + |5-1|^2 + |5-5|^2 + |4-5|^2 + |5-4|^2)}$$

$$= \sqrt{0 + 4^2 + 0 + 1^2 + 1^2}$$

$$= 4.2426$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 18 จะมีค่าเป็น 4.2426

ทำการคำนวณระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนอื่นๆ ดังนี้

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 71

ตารางที่ 3.13 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 71

movie_id / user_id	52	153	168	181	197
1	4	3	5	5	5
71	4	4	5	3	5

$$p(71,1)_2 = \sqrt{(|4-4|^2 + |4-3|^2 + |5-5|^2 + |3-5|^2 + |5-5|^2)}$$

$$= 2.2361$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 71 จะมีค่าเป็น 2.2361

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 76

ตารางที่ 3.14 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 76

movie_id / user_id	7	42	77	93	137
1	4	5	4	5	5
76	4	3	2	4	5

$$p(76,1)_2 = \sqrt{(|4-4|^2 + |3-5|^2 + |2-4|^2 + |4-5|^2 + |5-5|^2)}$$

$$= 3$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 76 จะมีค่าเป็น 3

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 181

ตารางที่ 3.15 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 181

movie_id / user_id	15	18	19	93	105
1	5	4	5	5	2
181	3	1	1	1	1

$$p(181,1)_2 = \sqrt{(|3-5|^2 + |1-4|^2 + |1-5|^2 + |1-5|^2 + |1-2|^2)}$$

$$= 5.6569$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 181 จะมีค่าเป็น 5.6569

- คำนวณหาระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 198

ตารางที่ 3.16 ตัวอย่างข้อมูลการให้คะแนนของผู้ใช้คนที่ 1 และ 198

movie_id / user_id	1	4	25	50	55
1	5	3	4	5	5
198	4	3	2	5	3

$$p(198,1)_2 = \sqrt{(|4-5|^2 + |3-3|^2 + |2-4|^2 + |5-5|^2 + |3-5|^2)}$$

$$= 3$$

ดังนั้น ระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้คนที่ 198 จะมีค่าเป็น 3

3.1.4 ตัวอย่างการทำนายคะแนนของผู้ใช้เป้าหมาย

3.1.4.1 การทำนายคะแนนในกรณีปรกติ

หลังจากทำการวัดระยะทางระหว่างผู้ใช้เป้าหมายกับผู้ใช้อ้างอิงแล้ว จะทำการเรียงลำดับค่าระยะห่างจากน้อยไปมากด้วย Selection Sort เพื่อหาผู้ใช้อ้างอิงที่มีระยะห่างน้อยที่สุด k ราย

ตารางที่ 3.17 ตัวอย่างค่าระยะห่างจากการคำนวณด้วยค่าอันดับ p ที่ 1

	user_id 71	user_id 76	user_id 198	user_id 18	user_id 181
Distance	2.2361	3	3	4.2426	5.6569

จากตัวอย่าง กำหนดค่า k หรือจำนวนผู้ใช้อ้างอิงที่มีระยะทางห่างจากผู้ใช้เป้าหมายน้อยที่สุดเป็น 3 ราย จะพบว่า ผู้ใช้คนที่ 71, 76 และ 198 มีระยะห่างน้อยที่สุด

ตารางที่ 3.18 การให้คะแนนภาพยนต์เรื่องที่ 6 โดยผู้ใช้อ้างอิง 3 ราย ที่มีระยะห่างน้อยที่สุด

	user_id 71	user_id 76	user_id 198
rating	4	4	3

จากนั้นจึงนำค่าคะแนนภาพยนต์เรื่องที่ 6 ที่ให้ไว้โดยผู้ใช้อ้างอิงทั้งสามคน มาทำการหาค่าเฉลี่ยเพื่อทำนายการให้คะแนน

Predict Rating – Movie_id 6 = (4+4+3)/3

= 3.67

คะแนนที่ทำนายได้ จะนำไปเก็บไว้ในฐานข้อมูล เพื่อเตรียมการหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์ ซึ่งคะแนน 3.67 ถือว่าเป็นคะแนนที่ค่อนข้างสูง เมื่อเทียบกับช่วงคะแนน 1 – 5

3.1.4.2 การทำนายคะแนนในกรณีที่มีจำนวนผู้ใช้อ้างอิงน้อยกว่าจำนวนสมาชิกใกล้สุด

ในตัวอย่างนี้ จะสมมุติว่ามีเพียงผู้ใช้คนที่ 18 และ 198 เท่านั้นที่เคยให้คะแนนภาพยนต์เรื่องที่ 6 ไว้ จากนั้น การทำนายจะคิดคะแนนที่จะทำนายจากคะแนนสองส่วน ดังต่อไปนี้

ส่วนที่ 1

ตารางที่ 3.19 ตัวอย่างการให้คะแนนภาพยนต์เรื่องที่ 6 ของผู้ใช้แต่ละคน

	user_id 18	user_id 198
rating	5	2

ทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์เรื่องที่ 6 จากการให้คะแนนของผู้ใช้ทุกคนที่เคยให้คะแนนไว้

Predict Rating (1) = (5+2)/2

= 3.5

ส่วนที่ 2

ตารางที่ 3.20 ตัวอย่างการให้คะแนนของผู้ใช้เป้าหมาย

	movie_id 4	movie_id 7	movie_id 15	movie_id 25	movie_id 52
rating	3	4	5	4	4

ทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมายเพื่อดูแนวโน้มของการให้คะแนน

Predict Rating (2) = (3+4+5+4+4)/5

= 4

เมื่อได้คะแนนจากทั้งสองส่วนแล้ว จึงนำคะแนนมารวมกันเป็นคะแนนที่จะทำนาย โดยให้น้ำหนักของคะแนนในส่วนที่ 1 เป็น 70% และน้ำหนักของคะแนนในส่วนที่ 2 เป็น 30%

Predict Rating – Movie_id 6 = (3.5 x 0.7) + (4 x 0.3)

= 3.65

3.1.4.1 การทำนายคะแนนในกรณีที่ไม่มีผู้ใช้คนอื่นเคยให้คะแนนภาพยนต์เป้าหมาย

ในกรณีนี้ จะใช้ข้อมูลการให้คะแนนของผู้ใช้เป้าหมายเพียงอย่างเดียวในการทำนายโดยการหาค่าเฉลี่ยจากการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมาย ซึ่งใช้วิธีเดียวกันกับส่วนที่ 2 ในหัวข้อ *3.1.4.2* แต่คิดเป็น 100% แทน

Predict Rating – Movie_id 6 = (3+4+5+4+4)/5

= 4

3.2 ขั้นตอนการประเมินผล

3.2.1 ตัวอย่างการประเมินผลด้วยค่าเฉลี่ยความผิดพลาดสัมบูรณ์

ทำการนำคะแนนที่ทำนายได้มาวัดประสิทธิภาพของการทำนายโดยคำนวณหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์ จากการเปรียบเทียบคะแนนที่ทำนาย กับข้อมูลจริงที่ผู้ใช้เป้าหมายได้ให้ไว้แก่ผลิตภัณฑ์นั้นๆ ในชุดข้อมูลทดสอบ

ตารางที่ 3.21 ตัวอย่างข้อมูลในชุดข้อมูลทดสอบ

user_id	movie_id	rating
1	6	5
1	10	3
1	12	5
1	14	5
1	17	3

ตารางที่ 3.22 ตัวอย่างข้อมูลที่ทำนายออกมา

user_id	movie_id	rating
1	6	3.6
1	10	3.4
1	12	4.6
1	14	4
1	17	3.4

ข้อมูลที่ทำนายออกมาได้ จะถูกนำไปปัดเศษด้วยค่ากลางก่อน เพื่อให้เป็นเลขจำนวนเต็มซึ่งเป็นคะแนนที่ใช้ทำนายจริง ก่อนนำมาคำนวณหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์

ตารางที่ 3.23 ตัวอย่างข้อมูลที่ทำนายออกมาหลังทำการปัดเศษ

user_id	movie_id	rating
1	6	4
1	10	3
1	12	5
1	14	4
1	17	3

จากนั้นจึงนำข้อมูลมาคำนวณหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์ ด้วยสมการดังต่อไปนี้

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

แทนค่าในสมการ

$$MAE = \frac{(|4-5| + |3-3| + |5-5| + |4-5| + |4-3|)}{5}$$
$$= 0.6$$

ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ ถ้ามีค่าน้อย แสดงว่ามีความผิดพลาดในการทำนายต่ำ ถ้ามีค่ามากแสดงว่ามีความผิดพลาดในการทำนายสูง ซึ่งในการทดลองจะทำการหาค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด โดยจะกล่าวถึงในบทต่อไป

บทที่ 4

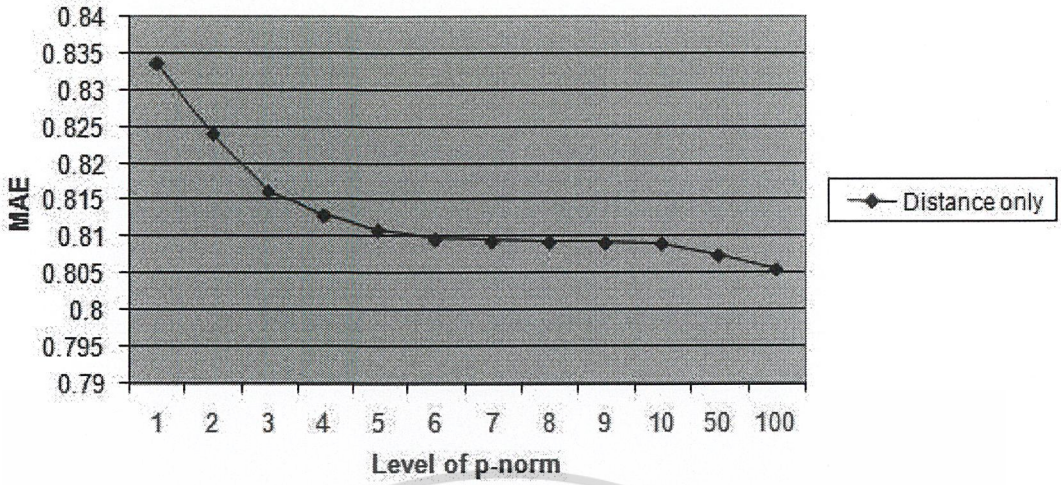
ผลการทดลอง

งานวิจัยนี้ต้องการทดสอบประสิทธิภาพของการทำนายโดยใช้การวัดระยะทางแบบมินโคว์สคิอันดับพี โดยทดลองปรับค่าอันดับ p เพื่อศึกษาประสิทธิภาพในการทำนายว่าที่อันดับ p ใดๆ จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด

ในการทดลองหาค่าเฉลี่ยความผิดพลาดสัมบูรณ์นี้จะกำหนดค่าจำนวนสมาชิกใกล้ที่สุดหรือ k = 5 และมีการปรับค่าอันดับ p ซึ่งหมายถึง จำนวนมิติของข้อมูลที่ใช้ในการหาระยะห่าง ซึ่งในการทดลองนี้จะทดลองกับ p ที่อยู่ในช่วงระหว่าง 1 – 1000 โดยรอบแรกจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50 และ 100 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.1 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 1

k	p	MAE	Predict Amount
5	1	0.833659733235	98676
5	2	0.824134721359	98676
5	3	0.816280082515	98676
5	4	0.813059066338	98676
5	5	0.811021638417	98676
5	6	0.809785336296	98676
5	7	0.809552050220	98676
5	8	0.809420408512	98676
5	9	0.809339356270	98676
5	10	0.809207481637	98676
5	50	0.807605473858	98676
5	100	0.805690171743	98676

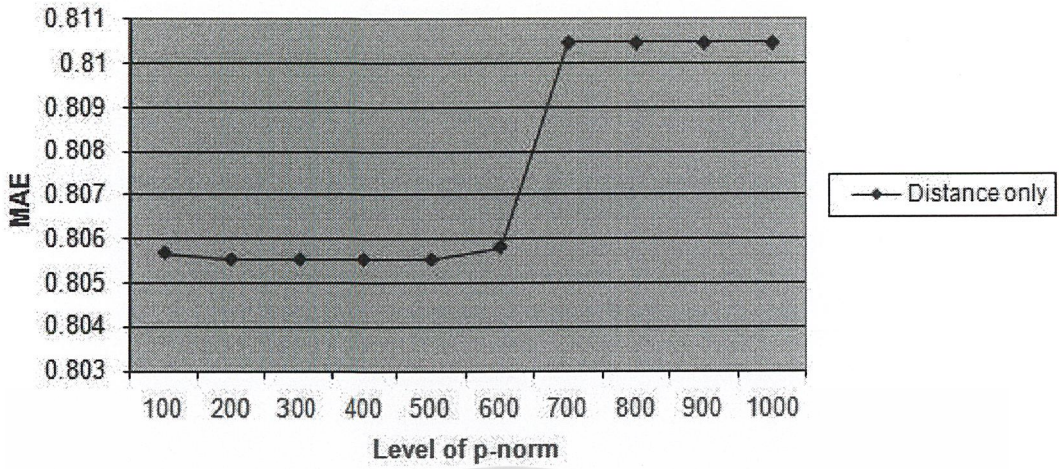


รูปที่ 4.1 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 1

จากรูปที่ 4.1 จะเห็นได้ว่ายิ่งเพิ่มค่าอันดับ p จะยิ่งทำให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยลง ซึ่งค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดคือ p = 100 ซึ่งจากผลการทดลองนี้เป็นไปได้ว่าค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่ p = 100 นี้อาจไม่ใช่ค่าที่น้อยที่สุด จึงได้มีการทดลองในรอบที่สอง โดยจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 100, 200, 300, 400, 500, 600, 700, 800, 900 และ 1000 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.2 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 2

k	p	MAE	Predict Amount
5	100	0.805690171743	98676
5	200	0.805538243544	98676
5	300	0.805538243544	98676
5	400	0.805538243544	98676
5	500	0.805538243544	98676
5	600	0.805821894670	98676
5	700	0.810483810249	98676
5	800	0.810473698003	98676
5	900	0.810473698003	98676
5	1000	0.810473698003	98676

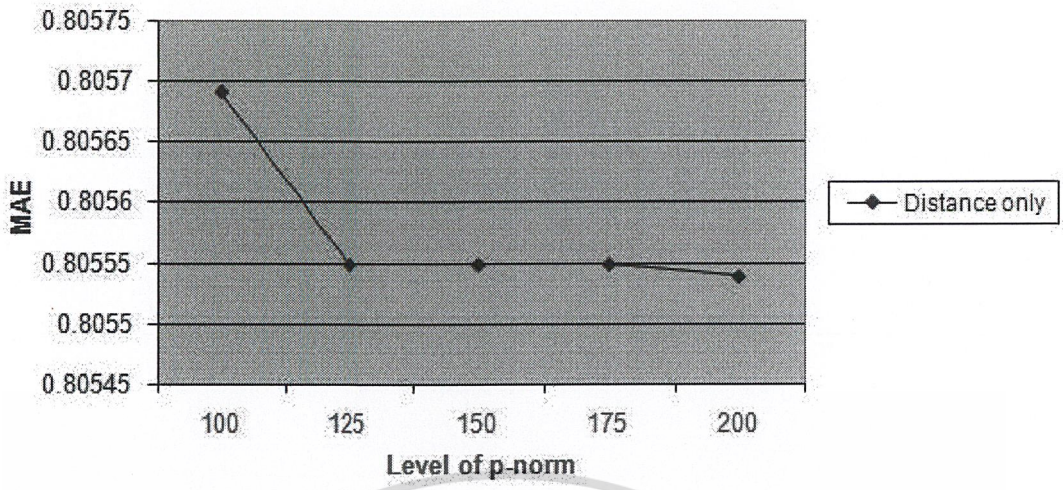


รูปที่ 4.2 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ *p* จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 2

จากรูปที่ 4.2 จะเห็นได้ว่าค่าอันดับ *p* ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง *p* 100 - 200 ดังนั้นจึงได้มีการทดลองในรอบที่สาม โดยจะมีการปรับเปลี่ยนค่าอันดับ *p* ดังนี้คือ 100, 125, 150, 175 และ 200 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.3 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 3

k	p	MAE	Predict Amount
5	100	0.805690171743	98676
5	125	0.805548355790	98676
5	150	0.805548355790	98676
5	175	0.805548355790	98676
5	200	0.805538243544	98676

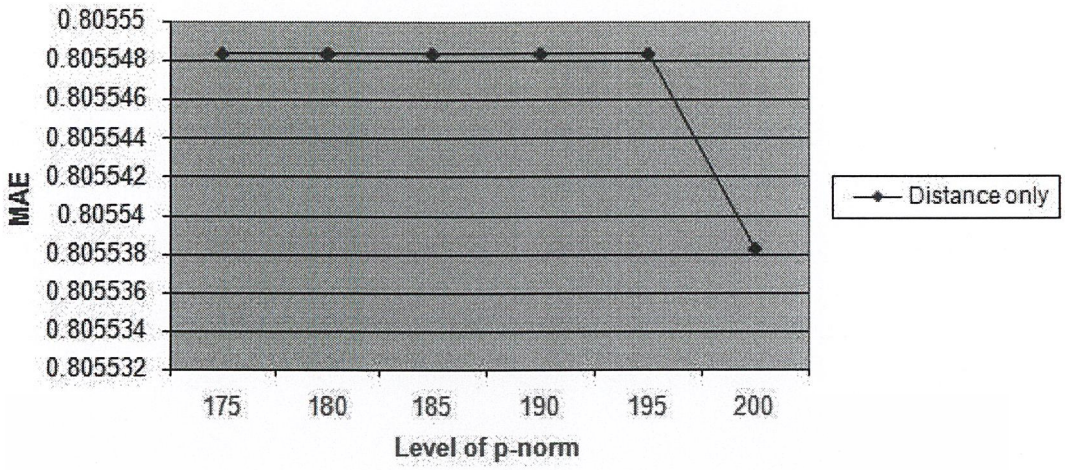


รูปที่ 4.3 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 3

จากรูปที่ 4.3 จะเห็นได้ว่าค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง p 175 - 200 ดังนั้นจึงได้มีการทดลองในรอบที่สี่ โดยจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 175, 180, 185, 190, 195 และ 200 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.4 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 4

k	p	MAE	Predict Amount
5	175	0.805548355790	98676
5	180	0.805548355790	98676
5	185	0.805548355790	98676
5	190	0.805548355790	98676
5	195	0.805548355790	98676
5	200	0.805538243544	98676

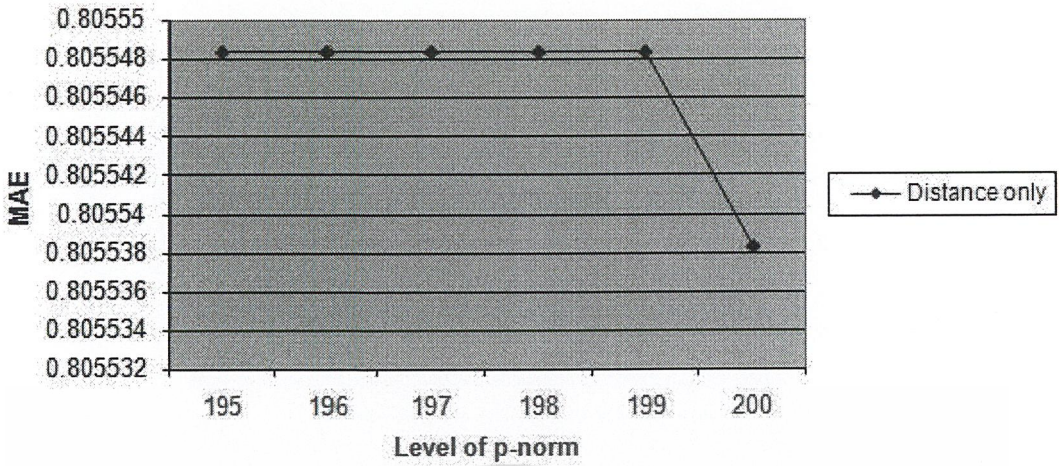


รูปที่ 4.4 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ *p* จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 4

จากรูปที่ 4.4 จะเห็นได้ว่าค่าอันดับ *p* ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง *p* 195 - 200 ดังนั้นจึงได้มีการทดลองในรอบที่ห้า โดยจะมีการปรับเปลี่ยนค่าอันดับ *p* ดังนี้คือ 195, 196, 197, 198, 199 และ 200 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.5 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 5

k	p	MAE	Predict Amount
5	195	0.805548355790	98676
5	196	0.805548355790	98676
5	197	0.805548355790	98676
5	198	0.805548355790	98676
5	199	0.805548355790	98676
5	200	0.805538243544	98676



รูปที่ 4.5 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p
จากการทำนายโดยใช้การวัดระยะทาง รอบที่ 5

จากรูปที่ 4.5 จะเห็นได้ว่าค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดที่อยู่ในช่วงระหว่าง p 1 – 1000 คือ $p = 200$ ซึ่งให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ 0.805538243544

จากการทดลองข้างต้นแม้ว่าจะสามารถหาค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดได้แล้ว แต่เมื่อสังเกตจากผลการทดลองจะพบว่า การทำนายที่ใช้การวัดระยะทางจะเกิดปัญหาทำให้ไม่สามารถทำนายข้อมูลได้ครบทุกค่าตามข้อมูลในชุดข้อมูลทดสอบ ซึ่งเกิดจาก 2 กรณีดังนี้

กรณีแรกเป็นกรณีที่มีจำนวนผู้ใช้อ้างอิงมีน้อยกว่าจำนวนสมาชิกใกล้สุด ซึ่งจะพบว่าผู้ใช้คนอื่นๆ ที่เคยให้คะแนนกับภาพยนต์เป้าหมาย มีน้อยกว่าจำนวนสมาชิกใกล้ที่สุด ทำให้ไม่เป็นไปตามเงื่อนไขที่กำหนด

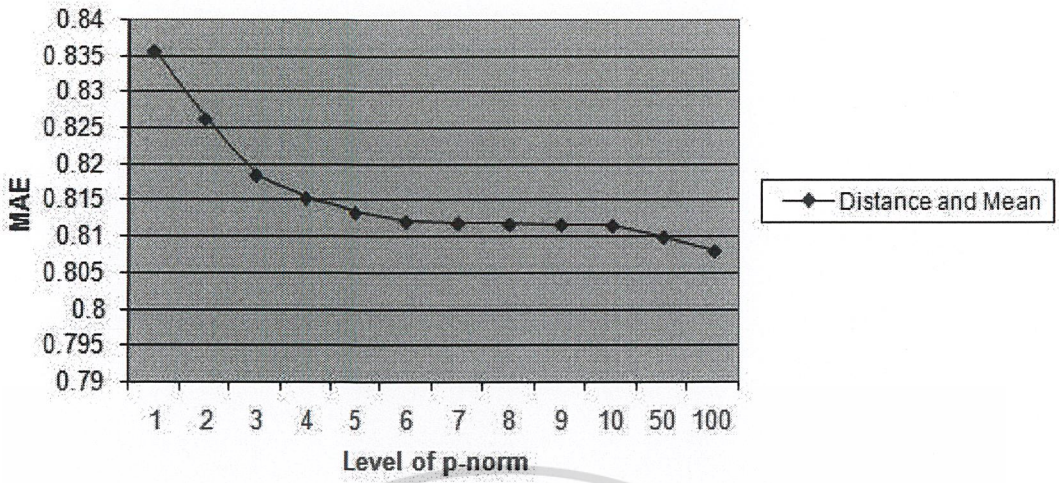
ส่วนกรณีที่สองเป็นกรณีที่ไม่มีผู้ใช้คนอื่นเคยให้คะแนนภาพยนต์เป้าหมาย ทำให้ไม่มีข้อมูลที่จะนำมาใช้อ้างอิง และทำนายการให้คะแนนแก่ผู้ใช้เป้าหมายได้ จึงไม่จำเป็นต้องใช้การวัดระยะทางเพื่อหาผู้ใช้อ้างอิงที่มีระยะห่างน้อยที่สุด

จากปัญหาดังกล่าวสามารถแก้ไขได้ โดยในกรณีแรกจะแบ่งการทำนายออกเป็น 2 ส่วน โดยส่วนที่ 1 จะทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์เรื่องนั้นๆจากการให้คะแนนของผู้ใช้ทุกคนที่เคยให้คะแนนไว้ คิดเป็น 70% ของคะแนนที่จะทำนาย และส่วนที่สองจะทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมายคนนั้นๆ เพื่อดูแนวโน้มของการให้คะแนน คิดเป็น 30% ของคะแนนที่จะทำนาย จากนั้นจึงนำคะแนนทั้งสองส่วนมารวมกันเพื่อทำนายการให้คะแนน ส่วนกรณีที่สองจะทำนายการให้คะแนน โดยทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมายคนนั้นๆ เพื่อดูแนวโน้มของการให้คะแนนเพียงอย่างเดียว

เพื่อเป็นการทดสอบว่าวิธีแก้ปัญหาของทั้งสองกรณีนี้จะมีประสิทธิภาพหรือไม่ จะทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี เพื่อหาค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด โดยรอบแรกจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50 และ 100 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.6 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 1

k	p	MAE	Predict Amount
5	1	0.83555	100000
5	2	0.82615	100000
5	3	0.81840	100000
5	4	0.81522	100000
5	5	0.81321	100000
5	6	0.81199	100000
5	7	0.81176	100000
5	8	0.81163	100000
5	9	0.81155	100000
5	10	0.81142	100000
5	50	0.80984	100000
5	100	0.80795	100000

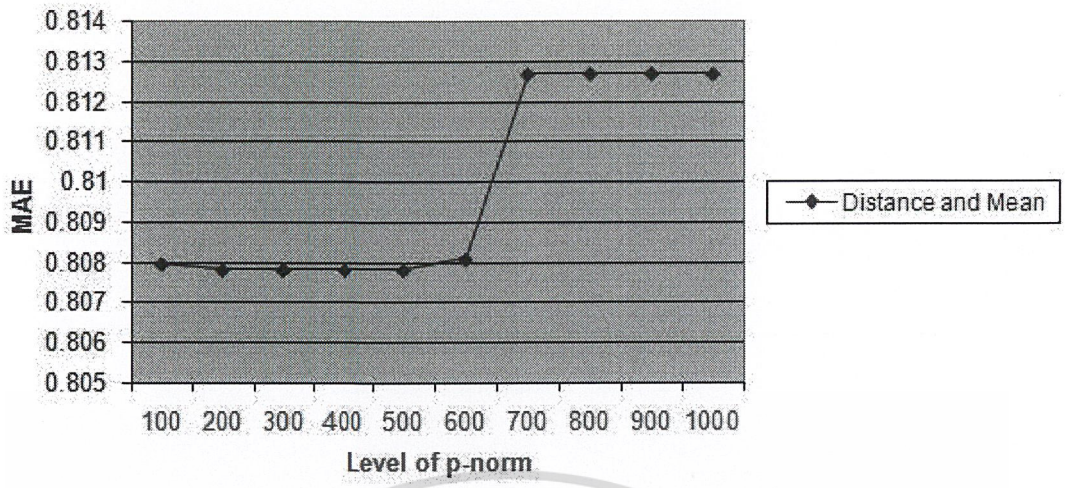


รูปที่ 4.6 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 1

จากรูปที่ 4.6 จะเห็นได้ว่ายิ่งเพิ่มค่าอันดับ p จะยิ่งทำให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยลง ซึ่งค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดคือ p = 100 ซึ่งจากผลการทดลองนี้เป็นไปได้ว่าค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่ p = 100 นี้อาจไม่ใช่ค่าที่น้อยที่สุด จึงได้มีการทดลองในรอบที่สอง โดยจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 100, 200, 300, 400, 500, 600, 700, 800, 900 และ 1000 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.7 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 2

k	p	MAE	Predict Amount
5	100	0.80795	100000
5	200	0.80780	100000
5	300	0.80780	100000
5	400	0.80780	100000
5	500	0.80780	100000
5	600	0.80807	100000
5	700	0.81267	100000
5	800	0.81268	100000
5	900	0.81268	100000
5	1000	0.81267	100000

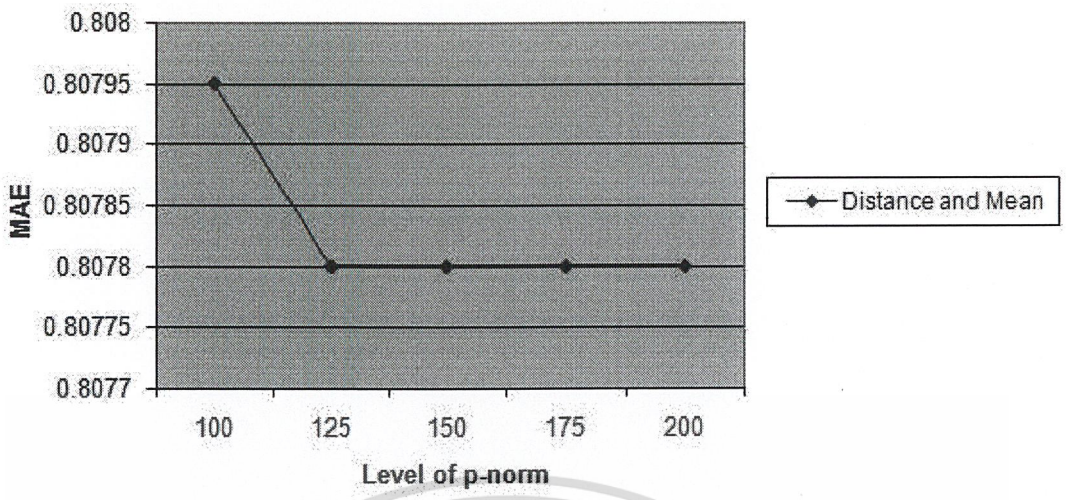


รูปที่ 4.7 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 1

จากรูปที่ 4.7 จะเห็นได้ว่าค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง p 100 - 200 ดังนั้นจึงได้มีการทดลองในรอบที่สาม โดยจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 100, 125, 150, 175 และ 200 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.8 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 3

k	p	MAE	Predict Amount
5	100	0.80795	100000
5	125	0.80780	100000
5	150	0.80780	100000
5	175	0.80780	100000
5	200	0.80780	100000

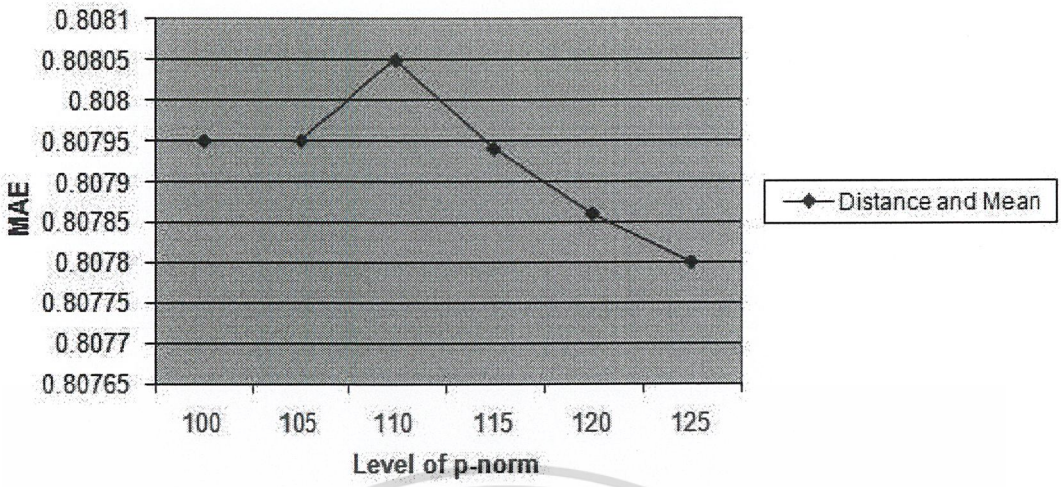


รูปที่ 4.8 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 3

จากรูปที่ 4.8 จะเห็นได้ว่าค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง p 100 - 125 ดังนั้นจึงได้มีการทดลองในรอบที่สี่ โดยจะมีการปรับเปลี่ยนค่าอันดับ p ดังนี้คือ 100, 105, 110, 115, 120 และ 125 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.9 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 4

k	p	MAE	Predict Amount
5	100	0.80795	100000
5	105	0.80795	100000
5	110	0.80805	100000
5	115	0.80794	100000
5	120	0.80786	100000
5	125	0.80780	100000

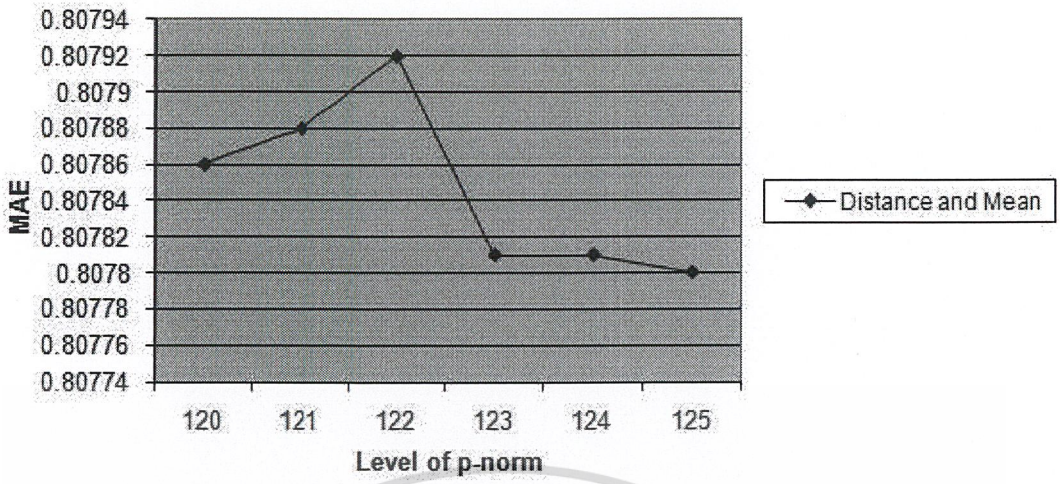


รูปที่ 4.9 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ *p*
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 4

จากรูปที่ 4.9 จะเห็นได้ว่าค่าอันดับ *p* ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดจะอยู่ในช่วงระหว่าง *p* 120 - 125 ดังนั้นจึงได้มีการทดลองในรอบที่ 5 โดยจะมีการปรับเปลี่ยนค่าอันดับ *p* ดังนี้คือ 120, 121, 122, 123, 124 และ 125 ตามลำดับ ซึ่งได้ผลการทดลองออกมาดังนี้

ตารางที่ 4.10 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 5

k	p	MAE	Predict Amount
5	120	0.80786	100000
5	121	0.80788	100000
5	122	0.80792	100000
5	123	0.80781	100000
5	124	0.80781	100000
5	125	0.80780	100000



รูปที่ 4.10 แสดงค่าเฉลี่ยความผิดพลาดสัมบูรณ์ในแต่ละอันดับ p
จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี รอบที่ 5

จากรูปที่ 4.10 จะเห็นได้ว่าค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดที่อยู่ในช่วงระหว่าง p 1 – 1000 คือ $p = 125$ ซึ่งให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ 0.80870

จากการทดลองนี้พบว่า การทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณีนั้น สามารถทำนายข้อมูลได้ครบทุกค่า ถึงแม้การทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณี จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่มากกว่าการทำนายโดยใช้การวัดระยะทางเพียงอย่างเดียว แต่ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่ได้จากการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณีนั้น เป็นค่าที่คำนวณได้จากข้อมูลทั้งหมด จึงถือว่าค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่ได้นี้เป็นค่าที่มีประสิทธิภาพมากกว่าการทำนายโดยใช้การวัดระยะทางเพียงอย่างเดียว ดังนั้นค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดในการทดลองนี้ คือ $p = 125$ ซึ่งให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ 0.80870

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุป

การกรองด้วยความร่วมมือโดยใช้ระยะทางได้นำการวัดระยะทางแบบมินโคว์สคิอันดับพี เข้ามาใช้ในการทำนาย จะกำหนดค่า k = 5 และทำการทดลองปรับค่าอันดับ p ในช่วงระหว่าง 1 – 1000 เพื่อศึกษาประสิทธิภาพในการทำนายว่าที่อันดับ p ใดๆ จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด ซึ่งในการทดลองจะแบ่งการทดลองออกเป็นรอบๆ โดยเริ่มจากทดลองปรับค่าอันดับ p แบบกว้างๆก่อน แล้วพิจารณาผลการทดลองที่ได้ว่าค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดอยู่ในช่วงอันดับ p ใด จากนั้นจึงทำการทดลองต่อจากช่วงอันดับ p นั้น ไปจนกว่าจะพบค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุด

แต่เนื่องจากการทำนายที่ใช้การวัดระยะทางเกิดปัญหาไม่สามารถทำนายข้อมูลได้ครบทุกค่าตามข้อมูลในชุดข้อมูลทดสอบ ซึ่งเกิดจาก 2 กรณี คือ กรณีแรกเป็นกรณีที่มีจำนวนผู้ใช้อ้างอิงน้อยกว่าจำนวนสมาชิกใกล้สุด หรือค่า k ที่กำหนดไว้ ส่วนกรณีที่สองเป็นกรณีที่ไม่มีผู้ใช้คนอื่นเคยให้คะแนนภาพยนตร์เป้าหมาย จากปัญหาดังกล่าวสามารถแก้ไขได้ โดยในกรณีแรกจะแบ่งการทำนายออกเป็น 2 ส่วน ส่วนแรกจะทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์เรื่องนั้นๆจากการให้คะแนนของผู้ใช้ทุกคนที่เคยให้คะแนนไว้ คิดเป็น 70% ของคะแนนที่จะทำนาย และส่วนที่สองจะทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนต์ทั้งหมดของผู้ใช้เป้าหมายคนนั้นๆ เพื่อดูแนวโน้มของการให้คะแนน คิดเป็น 30% ของคะแนนที่จะทำนาย จากนั้นจึงนำคะแนนทั้งสองส่วนมารวมกันเพื่อทำนายการให้คะแนน ส่วนกรณีที่สองจะทำนายการให้คะแนน โดยทำการหาค่าเฉลี่ยของการให้คะแนนภาพยนตร์ทั้งหมดของผู้ใช้เป้าหมายคนนั้นๆเพียงอย่างเดียว

ซึ่งจากการศึกษาประสิทธิภาพในการทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณีนั้น พบว่าสามารถทำนายข้อมูลได้ครบทุกตัว ถึงแม้จะให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่มากกว่าการทำนายโดยใช้การวัดระยะทางเพียงอย่างเดียว แต่ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่ได้จากการทำนายนั้น เป็นค่าที่คำนวณได้จากข้อมูลทั้งหมด ดังนั้นจึงสรุปได้ว่า การทำนายโดยใช้การวัดระยะทางร่วมกับวิธีแก้ปัญหาของทั้งสองกรณีนั้น ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ที่มีประสิทธิภาพมากกว่าการทำนายโดยใช้การวัดระยะทางเพียง

อย่างเดียว ซึ่งค่าอันดับ p ที่ให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์น้อยที่สุดในการทดลองนี้ คือ $p = 125$ ซึ่งให้ค่าเฉลี่ยความผิดพลาดสัมบูรณ์ 0.80870

5.2 ข้อเสนอแนะ

1. ชุดข้อมูลที่ใช้ในการทดลองจาก Movielens อาจมีปัญหาความเบาบางของข้อมูล (Sparsity Problem) ถ้าจำนวนข้อมูลที่ผู้ใช้ให้ไว้กับระบบมีจำนวนน้อยเกินไป อาจส่งผลให้ประสิทธิภาพในการทำนายและแนะนำลดลง ดังนั้นหากนำการแทนค่าข้อมูลที่ขาดหายมาใช้ก่อนการทำนายคะแนน จะทำให้ปัญหาความเบาบางของ ข้อมูลลดน้อยลง และจะส่งผลให้ประสิทธิภาพในการทำนายเพิ่มขึ้น

2. ในการทดลองนี้ใช้เวลาในการคำนวณค่อนข้างนาน ซึ่งอาจเกิดจากความผิดพลาดในการออก แบบการทดลอง หรือการเลือกใช้เครื่องมือที่ไม่เหมาะสมกับการทดลอง ดังนั้นหากออกแบบการทดลองให้ดี หรือเลือกใช้เครื่องมือที่เหมาะสมกว่านี้ จะทำให้ใช้เวลาในการคำนวณน้อยลงอย่างแน่นอน

รายการอ้างอิง

[1] ชลัท ศรีวิเศษสม, ธีรเกียรติ พ่วงตามพงษ์, วิธวินท์ เทพศุภรังษิกุล(2551), "*การศึกษาทดลองตัววัดความเหมือนสำหรับการกรองด้วยความร่วมมือ*" , ปริญญานิพนธ์ วิทยาศาสตรบัณฑิต สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

[2] ภรัณยา อำมฤครัตน์, ดร.พยุง มีสัจ(2553), "*การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาท*" , งานวิจัยระดับบัณฑิตศึกษา วิทยาศาสตรมหาบัณฑิต มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

[3] สมชาย จำปาทอง, ศาสตรา วงศ์ธนวสุ, คำรณ สุนัติ, สิรภัทร เชี่ยวชาญวัฒนา(2005), "*อัลกอริทึมการแบ่งกลุ่มข้อมูลโดยใช้ฟัซซี่ซีมีนกับการวัดระยะทาง*" , บทความวิชาการโครงการประชุมวิชาการร่วมสาขาวิทยาการคอมพิวเตอร์และวิศวกรรมซอฟต์แวร์ประจำปี 2548 (JCSSE 2005) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl(2001), "*ItemBased Collaborative Filtering Recommendation Algorithms*", ACM

[5] Long-Sheng Chen, Chun-Chin Hsu and Yu-Shan Chang(2009), "*MDS: A Novel Method for Class Imbalance Learning*", ACM