BÀI TẬP LÝ THUYẾT 4

Vũ Cao Nguyên – 18600187

Bài tập 1:

• John cần thiết kế hệ thống tư vấn cho một cửa hàng sách trực tuyến mới khai trương gần đây. Cửa hàng có hơn 1 triệu tựa sách nhưng cơ sở dữ liệu đánh giá mới chỉ có 10,000 đánh giá.

• Chiến lược nào sẽ giúp John có được hệ thống tư vấn tốt, content-based recommendation, user-based collaborative filtering, hay item-based collaboration filtering? Giải thích.

Trả lời: Chiến lược user-based collaborative filtering phù hợp cho John bởi vì cửa hàng mới khai trương(nên content-based re không phù hợp) , số lượng người dùng < số lượng mục(nên item-based cf chưa cần thiết để sử dụng) và user-based cf sẽ giúp sản phẩm đa dạng hơn trong việc đề xuất => tăng lượt đánh giá của các sản phẩm hơn.

• Một khách hàng đã đánh giá 5/5 sao cho cả hai cuốn sách "Linear Algebra" and "Differential Equations". Quyển sách nào sau đây ít có khả năng được giới thiệu nhất theo hệ thống tư vấn trên? Giải thích.

a) "Operating Systems"

b) "Convex Optimization"

c) "Harry Potter: The Goblet of Fire"

d) Không thể xác định được vì còn tùy thuộc vào đánh giá của những người dùng khác.

Trả lời: Đáp án (**d)** bởi vì user-based collaborative filtering được tính toán tùy thuộc vào xếp hạng của mặt hàng đó bởi những người dùng tương tự khác.

Bài tập 2:

• Bảng bên thể hiện độ yêu thích (1 – 5) của ba nhân vật, Mark Zuckerberg, Bill Gates, và Guido van Rossum, đối với bốn công nghệ, PHP, Spark, Microsoft.NET và Python.

	php	Spark	Microsoft .NET	
	4.5	4.0	1.5	4.5
	3.0	1.0	4.0	2.0
	4.5		2.0	5.0

• Áp dụng lọc cộng tác theo người dùng với k-NN (k = 1)

• Xác định độ tương tự về sở thích giữa Guido van Rossum với các nhân vật còn lại.

sim(Guido van Rossum, Mark Zuckerberg) = 4.587285

sim(Guido van Rossum, Bill Gates) = 0.972302

• Dự đoán điểm yêu thích của Guido van Rossum đối với Spark.

p(Guido van Rossum,spark) = 3.877

Bài tập 3:

• Bảng bên thể hiện độ yêu thích (1 – 5) của ba nhân vật, Mark Zuckerberg, Bill Gates, và Guido van Rossum, đối với bốn công nghệ, PHP, Spark, Microsoft.NET và Python.

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	?	5
Mary	?	?	4	3	4
Tom	1	1	5	4	?

• Áp dụng lọc cộng tác theo hạng mục với k-NN (k = 1)

• Xác định độ tương tự điểm đánh giá giữa Spark và các sản phẩm khác

sim(php,spark) = -0.27075

sim(ms.net,spark) = -0.75761

sim(py,spark) = 0.691905

• Dự đoán điểm yêu thích của Guido van Rossum đối với Spark.

P(Guido van Rossum,Spark) = 5

Bài tập 4:

• Bảng bên cạnh thể hiện tập dữ liệu đánh giá của 4 người dùng đối với 5 sản phẩm. Thang điểm đánh giá từ 1 (nhỏ nhất) đến 5 (lớn nhất). Dấu ? nghĩa là người dùng chưa xem hoặc chưa đánh giá sản phẩm này

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	?	5
Mary	?	?	4	3	4
Tom	1	1	5	4	?

• **Dự đoán điểm đánh giá của Tom đối với Book 5 bằng lọc cộng tác theo người dùng với k-NN (k = 1).**

sim(Tom,Mary) = -0.0314

sim(Tom,Geotge) = -0.98117

sim(Tom,Alice) = 0.869228

→ p(Tom,Book5) = 1.46826

• **Tương tự, dự doán bằng lọc cộng tác theo hạng mục với k=NN (k = 1)**

sim(Book5,Book1) = 1

sim(Book5,Book2) = 1

sim(Book5,Book3) = -0.96192

sim(Book5,Book4) = -1

→ P(Tom,Book5) = 1.5