ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**TRẦN THỊ TUYẾT VÂN**

**NGHIÊN CỨU PHƯƠNG PHÁP ĐỒNG HUẤN LUYỆN CHO HỆ THỐNG TƯ VẤN DỰA TRÊN LỌC CỘNG TÁC**

**ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ**

**NGÀNH HỆ THỐNG THÔNG TIN**

**NGƯỜI HƯỚNG DẪN KHOA HỌC:**

**TS. HỒ BẢO QUỐC**

**Thành phố Hồ Chí Minh – 2012**

**MỤC LỤC**

[1. ĐẶT VẤN ĐỀ 3](#_Toc335509214)

[1.1 Lý do chọn đề tài 3](#_Toc335509215)

[1.2 Mục tiêu nghiên cứu 4](#_Toc335509216)

[2. NGHIÊN CỨU HIỆN TRẠNG 5](#_Toc335509217)

[2.1 Hệ thống tư vấn lọc cộng tác và những thách thức 5](#_Toc335509218)

[2.2 Vấn đề dữ liệu thưa và hướng giải quyết hiện có 6](#_Toc335509219)

[3. CÁCH TIẾP CẬN DỰ KIẾN 8](#_Toc335509220)

[4. BỐ CỤC DỰ KIẾN CỦA BÀI LUẬN VĂN 10](#_Toc335509221)

[5. KẾ HOẠCH THỰC HIỆN 11](#_Toc335509222)

[6. TÀI LIỆU THAM KHẢO 12](#_Toc335509223)

# ĐẶT VẤN ĐỀ

## Lý do chọn đề tài

Công nghệ thông tin phát triển mạnh đã mang lại rất nhiều lợi ích cho chúng ta trên mọi lĩnh đặc biệt là lĩnh vực thương mại điện tử. Hàng loạt website ra đời với nhiều mục đích khác nhau. Thông tin ngày càng trở nên quá tải, nó làm cho con người cảm thấy choáng ngợp, không biết thông tin nào là cần thiết đối với mình. Mặc dù các bộ máy tìm kiếm khổng lồ phát triển như Google, Bing, … đã giúp con người tìm kiếm thông tin và hạn chế một số lớn thông tin không liên quan nhưng kết quả trả về quá nhiều, con người khó có thể kiểm tra hết, và thực tế nó vẫn chưa mang lại hiệu quả trong việc chọn ra những thông tin phù hợp riêng cho người dùng. Vấn đề đặt ra là cần một hệ thống hiệu quả tối ưu hơn cả bộ máy tìm kiếm, nó có thể tự động chọn lọc ra những thông tin gần với nhu cầu từng người dùng. Hệ thống tư vấn ra đời để giải quyết được vấn đề đó. Hệ thống tư vấn sẽ dự đoán và đưa ra những gợi ý hay những lời tư vấn giúp con người dễ dàng chọn “sản phẩm” mà họ yêu thích dựa vào những sở thích hay hành vi của họ trong quá khứ. Nó không chỉ mang lại người dùng những tiện ích trong chọn lựa thông tin mà còn giúp các doanh nghiệp hoạt động trong lĩnh vực thương mại điện tử phát triển hiệu quả và bền vững hơn . Một số website nổi tiếng như Amazon, Youtube, Ebay, Facebook … đã tích hợp hệ thống tư vấn mang lại hiệu quả rất cao.

Hiện nay có rất nhiều phương pháp để xây dựng hệ thống tư vấn nhưng có thể phân ra theo hai hướng là dựa vào nội dung (content-based) và lọc cộng tác (collaborative filtering). Hệ thống tư vấn dựa vào nội dung đưa ra những kết quả tư vấn là những sản phẩm có nhiều đặc trưng tương tự với những sản phẩm mà họ đã thích trong quá khứ[6]. Ví dụ như người dùng thích phim khoa học viễn tưởng thì họ có thể sẽ thích phim Avatar. Hướng tiếp cận này thực hiện hiệu quả trên các đối tượng dữ liệu biểu diễn dưới dạng văn bản nhưng rất khó thực hiện trên các dạng thông tin đa phương tiện (ví dụ: Phim, nhạc, ảnh …). Bên cạnh đó nó còn phụ thuộc nhiều vào phương pháp trích xuất nội dung đặc trưng của sản phẩm và đó cũng là một trong những hạn chế của phương pháp. Khác với dựa vào nội dung, hệ thống tư vấn lọc cộng tác khai thác thông tin về những hành động trong quá khứ hoặc ý kiến của cộng đồng người dùng để dự đoán những sản phẩm mà người dùng hiện tại có thể thích[9]. Hướng tiếp cận này được sử dụng phổ biến hơn bởi sự đơn giản và không phụ thuộc vào nội dung đặc trưng của sản phẩm mà người dùng đã đánh giá. Lọc cộng tác sử dụng đánh giá của người dùng lên những sản phẩm thường được biểu diễn bởi một ma trận hai chiều user-item để dự đoán và đưa ra kết quả tư vấn. Tuy nhiên hướng tiếp cận này cũng có những khó khăn trong quá trình xây dựng và phát triển. Một trong những thách thức lớn nhất là vấn đề dữ liệu thưa (sparse data), còn được là missing-data, gây ảnh hưởng rất nhiều đến hiệu quả tư vấn. Nhiều công trình nghiên cứu đã đưa ra những giải pháp khác để giải quyết vấn đề này (trình bày trong phần 2), bước đầu đã có những hiệu quả đáng kể, cải thiện phần nào vấn đề dữ liệu thưa nhưng vẫn còn tồn tại những khuyết điểm riêng. Vì thế, hiện tại vẫn cần nhiều nghiên cứu để có những giải pháp tối ưu hơn nữa. Tiếp tục phát triển lĩnh vực này, đề tài luận văn nghiên cứu và đề xuất một phương pháp mới với mong muốn để cải thiện sự thưa thớt dữ liệu hiệu quả hơn một số hướng giải quyết trước đó, mang hiệu quả tư vấn cao hơn cách tiếp cận lọc cộng tác truyền thống.

## Mục tiêu nghiên cứu

Vấn đề dữ liệu thưa là một rào cản lớn cho sự phát triển hệ thống tư vấn lọc cộng tác đối với các hệ thống thương mại lớn. Mục tiêu của đề tài là nghiên cứu hướng tiếp cận của hệ thống tư vấn lọc cộng tác và những thách thức lớn của nó. Đặc biệt là vấn đề dữ liệu thưa, phân tích và đánh giá các hướng giải quyết đã có để xây dựng một phương pháp mới giải quyết vấn này một cách hiệu quả và tối ưu hơn.

# NGHIÊN CỨU HIỆN TRẠNG

## Hệ thống tư vấn lọc cộng tác và những thách thức

Lọc cộng tác là một kỹ thuật mạnh đã được sử dụng trong nhiều hệ thống tư vấn với những thành công đáng kể. Kỹ thuật này sử dụng cơ sở dữ liệu sở thích của người dùng đối với các sản phẩm để dự đoán những sản phẩm mà họ có thể thích[11]. Ý tưởng chính của hệ thống tư vấn lọc cộng tác là đánh giá của người dùng u lên sản phẩm mới i sẽ tương tự với đánh giá của người dùng v lên sản phẩm i nếu hai người dùng u và v có cùng sở thích với nhau, có nghĩa là u và v đã cùng có đánh giá lên một số sản phẩm nào đó. Tương tự, người dùng u đánh giá lên sản phẩm i và j giống nhau nếu những người dùng khác cũng đánh giá giống nhau trên hai sản phẩm i và j [7]. Hệ thống tư vấn lọc cộng tác có thể chia theo hai hướng chính: Memory-based và model-based[18]. Thuật toán Memory-based (Neighborhood-based[7]) sử dụng toàn bộ cơ sở dữ liệu người dùng để dự đoán. Ngược lại, thuật toán Model-based (matrix factorization[8]) sử dụng một phần cơ sở dữ liệu người dùng để ước lượng hoặc học theo một mô hình sau đó mới dự đoán.

Hệ thống tư vấn Neighborhood-based được chia thành hai hướng[7]: Hệ thống tư vấn dựa trên người dùng (User-based Recommendation) và hệ thống tư vấn dựa trên sản phẩm (Item-based Recommendation). Hệ thống tư vấn dựa trên người dùng (MovieLens, Bellcore video, Ringo, ...) dự đoán đánh giá người dùng u lên sản phẩm i sử dụng đánh giá lên sản phẩm i của những người dùng khác có sở thích giống u. Hai người dùng u và v gọi là cùng sở thích khi cả u và v cùng đánh giá lên một số sản phẩm nào đó. Còn hệ thống tư vấn dựa trên sản phẩm (Amazon) thì dự đoán đánh giá của người dùng lên sản phẩm i dựa vào đánh giá của họ lên những sản phẩm tương tự i. Những sản phẩm tương tự i có thể hiểu là những sản phẩm có cùng đánh giá của một nhóm người dùng. Để đo độ giống nhau giữa người dùng với người dùng hoặc giữa sản phẩm với sản phẩm có thể dùng độ đo Cosine, Pearson Correlation (PC), Adjusted Cosine (AC), Mean Squared Difference (MSD),…[7]. Đối với hai cách tiếp cận này, độ chính xác và hiệu quả của hệ thống phụ thuộc vào số người dùng và số sản phẩm. Nếu số sản phẩm tư vấn ít hơn nhiều so với số người dùng hệ thống ta nên chọn Item-based, ngược lại chọn User-based. Về tính ổn định, nó phụ thuộc vào sự thay đổi số lượng của sản phẩm và người dùng, nếu hệ thống có số lượng sản phẩm thường xuyên thay đổi hơn số lượng người dùng thì ta nên chọn User-based. Về tính trong suốt thì Item-based là sự lựa chọn tốt hơn User-based. Ngược lại User-based có khả năng dự đoán sản phẩm đa dạng hơn Item-based. Tùy vào mục tiêu của hệ thống mà ta chọn hướng tiếp cận phù hợp.

Trong lĩnh vực thương mại điện tử, xây dựng một hệ thống tư vấn đưa ra kết quả tư vấn nhanh và chính xác sẽ mang lại tiện ích cho khách hàng và mang lại lợi nhuận cho doanh nghiệp. Đối với hệ thống lọc cộng tác, chất lượng và hiệu quả của kết quả tư vấn phụ thuộc vào cách giải quyết những thách thức sau đây[11]: Dữ liệu thưa (data sparity), khả năng mở rộng (Scalability), tính đồng nghĩa (Synonymy), Gray Sheep, Shilling Attacks. Trong số các thách thức này thì giải quyết vấn đề dữ liệu thưa là quan trọng hàng đầu khi xây dựng hệ thống tư vấn bằng lọc cộng tác. Nội dung này được trình bày rõ trong phần sau.

## Vấn đề dữ liệu thưa và hướng giải quyết hiện có

Trên thực tế, các hệ thống tư vấn thương mại điện tử có số lượng người dùng và số lượng sản phẩm rất lớn và ngày càng mở rộng. Do vậy ma trận đánh giá user-item rất lớn. Tuy nhiên, số lượng người dùng đánh giá lên các sản phẩm lại rất hạn chế, vì thế dữ liệu đánh giá rất ít và vấn đề dữ liệu thưa được đặt ra.

Vấn đề này thường gặp trong một số trường hợp, đặc biệt là khi có một người dùng mới sử dụng hệ thống, họ chưa đánh giá một sản phẩm nào, hoặc khi có một sản phẩm mới được thêm vào hệ thống, chúng chưa được người dùng nào đánh giá. Trường hợp này còn gọi là vấn đề *cold-start* [10] (hay còn gọi là vấn đề người dùng mới và sản phẩm mới). Nó thật sự rất khó khăn khi tìm người cùng sở thích hay những sản phẩm tương tự. Sản phẩm mới sẽ không được tư vấn cho đến khi có người dùng đánh giá và người dùng mới cũng không có được tư vấn vì họ chưa đánh giá hoặc mua sản phẩm nào trong quá khứ. Khi người dùng có đánh giá một số sản phẩm nhưng số lượng đánh giá này rất ít so với số sản phẩm có trong hệ thống thì một vấn đề khác xuất hiện là vấn đề *độ bao phủ bị giảm* (reduced coverage). Độ bao phủ là tỉ lệ phần trăm của sản phẩm mà hệ thống có thể tư vấn cho người dùng. Độ bao phủ giảm tức số lượng sản phẩm được tư vấn sẽ rất hạn chế. Ngoài ra khi dữ liệu thưa không thể xác định những người có cùng sở thích với nhau nếu như họ không cùng đánh giá trên một số sản phẩm. Vấn đề này gọi là vấn đề *tính bắc cầu của láng giềng (neighbor transitivity)*. Sự thưa thớt dữ liệu ảnh hưởng rất nhiều đến hiệu quả của hệ thống tư vấn.

Hiện nay, nhiều giải pháp đã được đề xuất để giải quyết vấn đề dữ liệu thưa. Một giải pháp đã được sử dụng thành công là kỹ thuật giảm số chiều như SVD (Singular Value Decomposition)[1][7]. Kỹ thuật này sẽ loại bỏ những người dùng và sản phẩm không quan trọng để giảm số chiều của ma trận đánh giá user-item, sau đó tính độ tương đồng của user hoặc item trên ma trận đã rút gọn đó và đưa ra dự đoán. Một số kỹ thuật khác[12] như kỹ thuật thống kê PCA(Principle Component Analysis) và kỹ thuật truy tìm thông tin LSI (Latent Semantic Indexing). Tuy nhiên, khi giảm số chiều thì nhiều thông tin hữu có thể cũng bị mất và chất lượng của hệ thống tư vấn bị giảm.

Một hướng giải quyết khác là kết hợp lọc cộng tác với hướng tiếp cận dựa trên nội dung[15]. Phương pháp dựa trên nội dung có thể hỗ trợ lọc cộng tác trong tính độ tương đồng của sản phẩm từ nội dung thông tin của sản phẩm và cho kết quả dự đoán chính xác hơn. Tuy nhiên, hướng giải quyết này lại gặp khó khăn của phương pháp dựa trên nội dung là giới hạn của phân tích nội dung. Ngoài ra, lọc cộng tác còn có thể kết hợp với phương pháp dựa trên đồ thị[7][16] để giải quyết vấn đề dữ liệu thưa nhưng bị hạn chế ở tính trong suốt, không đưa lời giải thích cho kết quả tư vấn. Ba tác giả Nguyễn Duy Phương, Lê Quang Thắng và Từ Minh Phương [5] đã đề xuất một mô hình kết hợp giữa lọc cộng tác và lọc nội dung dựa trên biểu diễn đồ thị các mối quan hệ giữa người dùng và sản phẩm, giữa sản phẩm và đặc trưng nội dung sản phẩm để xác định nên mối quan hệ giữa người dùng và đặc trưng nội dung ưa thích. Bằng cách biểu diễn này, mô hình trên sẽ khắc phục được hạn chế của lọc cộng tác trong trường hợp ít dữ liệu thông qua việc xác định các đặc trưng nội dung có ý nghĩa quyết định đến thói quen và sở thích người dùng. Gần đây, Ying He và các tác giả[3] đã đề xuất một hệ thống lọc cộng tác lai dựa vào mô hình Bayesian. Ý tưởng của phương pháp này là xây dựng mô hình item bởi các item tương đồng và mô hình Bayesian, sau đó dự đoán cho item chưa được đánh giá và làm đầy ma trận thưa user-item.

Theo những hướng nghiên cứu gần đây, để giải quyết vấn đề dữ liệu thưa, các tác giả[2] đã khai thác kết hợp độ tương đồng của user và độ tương đồng của item thay vì dùng các cách tiếp cận user-based hay item-based riêng lẻ truyền thống. Ngoài ra, Gilad Katz và các đồng nghiệp[4] đã sử dụng nguồn dữ liệu từ Wikipedia để tính độ tương đồng của các item. Các tác giả tạo một độ đo sự tương đồng bằng cách kết hợp đánh giá của người dùng và dữ liệu tổng hợp từ Wikipedia và sử dụng độ đo này để dự đoán đánh giá cho những item chưa được đánh giá, sau đó thêm chúng vào ma trận đánh giá user-item để giảm vấn đề dữ liệu thưa.

# CÁCH TIẾP CẬN DỰ KIẾN

Như đã nêu trên, hệ thống tư vấn lọc cộng tác sử dụng dữ liệu đánh giá được biễu diễn bằng ma trận 2 chiều user-item như hình sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | user1 | user2 | user3 | user4 | user5 | ...  User/item đã gán nhãn |
| Item1 | 1 | **?** | 5 | **?** | 5 | ... |
| Item2 | **?** | 3 | **?** | **?** | **?** | ...  User/item chưa gán nhãn |
| Item3 | 5 | **?** | **?** | 2 | **?** | … |
| Item4 | **?** | **?** | 1 | **?** | **?** | … |
| Item5 | **?** | 4 | **?** | **?** | 3 | … |
| … | **…** | … | **…** | **…** | … | … |

Những item được user đánh giá có thể được gọi là item đã gán nhãn. Tương tự những user có đánh giá lên item gọi là user đã gán nhãn. Các trường hợp user chưa đánh giá và item chưa được đánh giá gọi là user chưa gán nhãn và item chưa gán nhãn.

Khi dữ liệu thưa thì số lượng dữ liệu chưa gán nhãn sẽ nhiều hơn rất nhiều so với dữ liệu đã gán nhãn, và trong phần lớn dữ liệu chưa gán nhãn ấy tiềm ẩn những thông tin hữu ích cần được khai thác. Một số hướng tiếp cận nêu trên đã giải quyết phần nào vấn đề dữ liệu thưa nhưng bên cạnh đó lại gặp những khuyết điểm riêng của nó. Vì thế đề tài đề xuất một phương pháp với mong muốn giải quyết vấn đề dữ liệu thưa vừa hạn chế những khuyết điểm của các hướng giải quyết trước đó. Vậy yêu cầu đặt ra cho phương pháp đề xuất là:

❶ Giải quyết tốt vấn đề dữ liệu thưa.

❷ Không mất thông tin hữu như phương pháp giảm số chiều.

❸ Không kết hợp lọc cộng tác với hướng tiếp cận khác, tránh thêm những khó khăn khác.

❹ Cho kết quả tư vấn tốt hơn lọc cộng tác truyền thống.

Qua những nghiên cứu ban đầu, đề tài xin đề xuất phương pháp Co-training cho hệ thống tư vấn lọc cộng tác vì nhận thấy phương pháp này có thể đáp ứng những yêu cầu trên.

Co-training là một phương pháp học bán giám sát sử dụng cả dữ liệu gán nhãn và dữ liệu chưa gán trong quá trình huấn luyện. Đặc biệt Co-training mang lại hiệu quả đối với bài toán dữ liệu chưa gán nhãn nhiều hơn dữ liệu đã gán nhãn[14], có thể khai thác những thông tin hữu ích từ dữ liệu chưa gán nhãn mà không mất thông tin. Theo giả thuyết, Co-training sử dụng tập đặc trưng có thể chia thành 2 tập con (còn được gọi là 2 khung nhìn), mỗi tập đặc trưng con đủ để huấn luyện một phân lớp tốt và 2 tập con này phải thỏa một điều kiện độc lập cho trước[13]. Khi áp dụng Co-training cho bài toán lọc cộng tác, sử dụng tập đặc trưng là ma trận đánh giá user-item, ta có thể chia thành 2 tập con tương ứng với 2 cách tiếp cận user-based và item-based. Hai quá trình user-based và item-based sẽ thực hiện song hành và hỗ trợ cho nhau trong việc học và gán nhãn cho dữ liệu chưa gán nhãn. Trong phương pháp này, sử dụng 2 hướng tiếp cận của lọc cộng tác, không sử dụng hướng tiếp cận khác nên vấn đề khó khăn không phát sinh thêm.

Qua những phân tích ban đầu, tôi tin rằng phương pháp đề xuất có khả thi trong việc giải quyết vấn đề dữ liệu thưa và các hạn chế của các hướng giải quyết trước đó.

# BỐ CỤC DỰ KIẾN CỦA BÀI LUẬN VĂN

Nội dung luận văn dự định gồm có 6 chương:

*Chương 1:* **Tổng quan**

Giới thiệu tổng quan về ngữ cảnh, mục tiêu, nội dung nghiên cứu và những kết quả luận văn đạt được.

*Chương 2:* **Hệ thống tư vấn**

Trình bày một số hướng tiếp cận của hệ thống tư vấn, ưu và khuyết điểm của từng phương pháp. Đặc biệt là phương pháp dựa trên lọc cộng tác, được trình bày rõ hơn.

*Chương 3:* **Phương pháp học bán giám sát đồng huấn luyện(co-training)**

Trình bày sơ lược học bán giám, các ứng dụng trong hệ thống tư vấn. Trình bày phương pháp co-training.

*Chương 4:* **Phương pháp co-training cho hệ thống tư vấn lọc cộng tác**

Đây là nội dung trọng tâm của đề tài. Trình bày phương pháp co-training cho hệ thống tư vấn lọc cộng tác.

*Chương 5:* **Thực nghiệm và phân tích kết quả**

Trình bày kết quả thử nghiệm phương pháp lọc cộng tác truyền thống(item-based và user-based) và phương pháp đề xuất trên bộ dữ liệu thực tế từ MovieLens. Qua đó, phân tích kết quả, so sánh và đánh giá các phương pháp dựa trên một số tiêu chí.

*Chương 6:* **Kết luận và hướng phát triển**

Tóm tắt những kết quả mà luận văn đã đạt được và xác định những hướng phát triển trong tương lai

# KẾ HOẠCH THỰC HIỆN

Dự kiến luận văn sẽ hoàn thành trong 8 tháng:

1. Nghiên cứu các hướng tiếp cận của hệ thống tư vấn, phân tích và đánh giá các hướng tiếp cập để tìm ra bài toán cần giải quyết cho đề tài.
2. Nghiên cứu các hướng giải quyết bài toán đã có, phân tích và đánh giá các phương pháp đó, đề xuất một phương pháp mới, xây dựng đề cương
3. Nghiên cứu các lý thuyết liên quan đến phương pháp đề xuất, vận dụng xây dựng mô hình cho phương pháp
4. Hoàn thiện thuật toán cho phương pháp mới.
5. Tiến hành xây dựng thực nghiệm
6. Tiếp tục xây dựng thực nghiệm, đưa ra phân tích, đánh giá, so sánh.
7. Hoàn thiện quyển báo cáo của luận văn
8. Kiểm tra và hoàn chỉnh luận văn.

# TÀI LIỆU THAM KHẢO

[1] Shien Ge and Xinyang Ge (2012), An SVD-based Collaborative Filtering approach to alleviate cold-start problems, 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1474 – 1477, IEEE.

[2] Hao Ma, Irwin King and Michael R.Lyu (2007). Effective Missing Data Prediction for Collaborative Filtering. SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 39-46.

[3] Ying He, Shaoyu Yang and Chenbin Jiao (2011), A Hybrid Collaborative Filtering Recommendation Algorithm for Solving the Data Sparsity, 2011 International Symposium on Computer Science and Society, pp. 118 – 121, IEEE

[4] Gilad Katz, Nir Ofek, Bracha Shapira, Lior Rokach and Guy Shani (2011),Using Wikipedia to Boost Collaborative Filtering Techniques, Proceeding RecSys '11 Proceedings of the fifth ACM conference on Recommender systems, pp. 285-288

[5]Nguyen Duy Phuong, Le Quang Thang and Tu Minh Phuong (2008), A Graph-Based Method for Combining Collaborative and Content-Based Filtering, Trends in Artificial Intelligence, pp. 859-869, Springer.

[6]Pasquale Lops, Marco de Gemmis and Giovanni Semeraro (2011), Content-based Recommender Systems: State of the Art and Trends, Recommender Systems Handbook, pp. 73-105, Springer.

[7]Christian Desrosiers and George Karypis (2011), A Comprehensive Survey of Neighborhood-based Recommendation Methods, Recommender Systems Handbook, pp. 107-144, Springer.

[8]Yehuda Koren and Robert Bell (2011). Advances in Collaborative Filtering. Recommender Systems Handbook, pp 145-186, Springer.

[9]Robin Burke (2002), Hybrid Recommender Systems: Survey and Experiments, User Modeling and User-Adapted Interaction, Volume 12, Number 4, pp. 331-370.

[10]J. Ben Schafer, Dan Frankowski, Jon Herlocker and Shilad Sen (2007), Collaborative Filtering Recommender Systems, The Adaptive Web, LNCS 4321, pp. 291 – 324, Springer-Verlag Berlin Heidelberg.

[11]Xiaoyuan Su and Taghi M. Khoshgoftaar (2009), A survey of collaborative filtering techniques, Advances in Artiﬁcial Intelligence, Volume 2009.

[12]Chrsistian Desrosiers and George Karypis (2008), Solving the Sparsity Problem: Collaborative Filtering via Indirect Similarities, University of Minnesota - Computer Science and Engineering, Technical Report 08-044.

[13]Avrim Blum and Tom Mitchell (1998), Combining labeled and unlabeled data with co-training, COLT' 98 Proceedings of the eleventh annual conference on Computational learning theory, pp. 92-100, ACM.

[14]Xiaojin Zhu, Semi-Supervised Learning Literature Survey(2008), Computer Sciences TR 1530, University of Wisconsin – Madison.

[15]Prem Melville, Raymond J. Mooney and Ramadass Nagarajan (2001), Content-Boosted Collaborative Filtering, Appears in Proceedings of the SIGIR-2001 Workshop on Recommender Systems, New Orleans, LA.

[16]Manos Papagelis, Dimitris Plexousakis and Themistoklis Kutsuras (2005), Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences, iTrust'05 Proceedings of the Third international conference on Trust Management, pp. 224-239, Springer.

[17] Beau Piccart, Jan Struyf and Hendrik Blockeel (2010). Alleviating the Sparsity Problem in Collaborative Filtering by using an Adapted Distance and a Graph-based Method. Proceedings of the Tenth SIAM International Conference on Data Mining, pp. 189-199.