**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**TRẦN MINH LUẬN**

**NGUYỄN THANH ANH TUYÊN**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG FRAMEWORK ĐÁNH GIÁ**

**CÁC THUẬT TOÁN KHUYẾN NGHỊ VIỆC LÀM**

**KỸ SƯ NGÀNH KỸ THUẬT PHẦN MỀM**

**TP. HỒ CHÍ MINH, 2016**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**TRẦN MINH LUẬN – 12520248**

**NGUYỄN THANH ANH TUYÊN - 12520492**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG FRAMEWORK ĐÁNH GIÁ**

**CÁC THUẬT TOÁN KHUYẾN NGHỊ VIỆC LÀM**

**KỸ SƯ NGÀNH KỸ THUẬT PHẦN MỀM**

**GIẢNG VIÊN HƯỚNG DẪN**

**TS. HUỲNH NGỌC TÍN**

**TP. HỒ CHÍ MINH, 2016**

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …………………… ngày ………………….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

* 1. …………………………………………. – Chủ tịch.
  2. …………………………………………. – Thư ký.
  3. …………………………………………. – Ủy viên.
  4. …………………………………………. – Ủy viên.

|  |  |
| --- | --- |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |
|  | ***TP. HCM, ngày…..tháng…..năm……..*** |

NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(CỦA CÁN BỘ HƯỚNG DẪN)

|  |  |  |
| --- | --- | --- |
| **Tên khóa luận:** | | |
| **XÂY DỰNG FRAMEWORK ĐÁNH GIÁ CÁC THUẬT TOÁN KHUYẾN NGHỊ VIỆC LÀM** | | |
| **Nhóm SV thực hiện:** | | **Cán bộ hướng dẫn:** |
| Trần Minh Luận | 12520248 | TS. Huỳnh Ngọc Tín |
| Nguyễn Thanh Anh Tuyên | 12520492 |  |

**Đánh giá Khóa luận**

1. Về cuốn báo cáo:

Số trang Số chương

Số bảng số liệu Số hình vẽ

Số tài liệu tham khảo Sản phẩm

Một số nhận xét về hình thức cuốn báo cáo:

1. Về nội dung nghiên cứu:

1. Về chương trình ứng dụng:

1. Về thái độ làm việc của sinh viên:

**Đánh giá chung:**

**Điểm từng sinh viên:**

Trần Minh Luận**:………../10**

Nguyễn Thanh Anh Tuyên**:………../10**

|  |  |  |
| --- | --- | --- |
|  | **Người nhận xét**  (Ký tên và ghi rõ họ tên) | |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** | |
|  | | ***TP. HCM, ngày…..tháng…..năm……..*** | |

NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(CỦA CÁN BỘ PHẢN BIỆN)

|  |  |  |
| --- | --- | --- |
| **Tên khóa luận:** | | |
| **XÂY DỰNG FRAMEWORK ĐÁNH GIÁ CÁC THUẬT TOÁN KHUYẾN NGHỊ VIỆC LÀM** | | |
| **Nhóm SV thực hiện:** | | **Cán bộ phản biện:** |
| Trần Minh Luận | 12520248 |  |
| Nguyễn Thanh Anh Tuyên | 12520492 |  |

**Đánh giá Khóa luận**

1. Về cuốn báo cáo:

Số trang Số chương

Số bảng số liệu Số hình vẽ

Số tài liệu tham khảo Sản phẩm

Một số nhận xét về hình thức cuốn báo cáo:

1. Về nội dung nghiên cứu:

1. Về chương trình ứng dụng:

1. Về thái độ làm việc của sinh viên:

**Đánh giá chung:**

**Điểm từng sinh viên:**

Trần Minh Luận**:………../10**

Nguyễn Thanh Anh Tuyên**:………../10**

|  |  |
| --- | --- |
|  | **Người nhận xét**  (Ký tên và ghi rõ họ tên) |

LỜI CẢM ƠN

|  |  |
| --- | --- |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |

ĐỀ CƯƠNG CHI TIẾT

|  |  |
| --- | --- |
| **TÊN ĐỀ TÀI: XÂY DỰNG FRAMEWORK ĐÁNH GIÁ CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ VIỆC LÀM.** | |
| **Cán bộ hướng dẫn:** TS. Huỳnh Ngọc Tín | |
| **Thời gian thực hiện:**Từ ngày 10/06/2016 đến ngày 16/01/2017 | |
| **Sinh viên thực hiện:**  Trần Minh Luận - 12520248  Nguyễn Thanh Anh Tuyên - 12520492 | |
| **Nội dung đề tài:***(Mô tả chi tiết mục tiêu, phạm vi, đối tượng, phương pháp thực hiện, kết quả mong đợi của đề tài)*   * **Mục tiêu đề tài:** * Xây dựng một framework sử dụng các độ đo đánh giá phổ biến để đánh giá độ chính xác của các phương pháp khuyến nghị việc làm. * **Phạm vi đề tài:** * Hệ thống sử dụng thư viện Mahout của Apache để xây dựng các thuật toán khuyến nghị gồm: lọc cộng tác (Collaborative filtering), lọc nội dung (Content-based) và phương pháp lai tuyến tính giữa lọc cộng tác và lọc nội dung (Simple hybrid). * Nguồn dữ liệu việc làm được thu thập tự động từ các website đăng tuyển trực tuyến: vietnamworks.com, careerlink.vn, itviec.com * Nguồn dữ liệu gán nhãn phục vụ cho việc đánh giá các thuật toán khuyến nghị được thu thập từ việc khảo sát người dùng. * Hệ thống sẽ cài đặt các độ đo đánh giá chuẩn phổ biến như: Precision, Recall, F-Measure, NDCG, RMSE, MAE, MRR. * **Đối tượng nghiên cứu:** * Các thuật toán khuyến nghị: lọc cộng tác, lọc nội dung, phương pháp lai. * Các độ đo đánh giá thuật toán: Precision, Recall, F-Measure, NDCG, RMSE, MAE, MRR. * **Phương pháp thực hiện gồm các bước chính:** * Thu thập dữ liệu việc làm và tiến hành khảo sát mức độ phù hợp của việc làm đối với người dùng * Xây dựng thuật toán khuyến nghị việc làm * Xây dựng các độ đo đánh giá các thuật toán khuyến nghị đã cài đặt * Tiến hành thực nghiệm, phân tích, giải thích các kết quả thực nghiệm. * **Kết quả mong đợi của đề tài:** * Hoàn thành framework phát triển trên nền web cho phép người dùng sử dụng chạy thực nghiệm đánh giá các thuật toán khuyến nghị việc làm. * Có bộ dữ liệu gán nhãn để chạy thực nghiệm * Báo cáo đề tài   **Kế hoạch thực hiện:***(Mô tả kế hoạch làm việc và phân công công việc cho từng sinh viên tham gia)*   * **Kế hoạch làm việc:** * Viết thuyết minh đề tài (13/10/2016 – 14/10/2016) * Khảo sát hiện trạng các nghiên cứu, các hệ thống liên quan. (6/2015 – 7/2015) * Xây dựng tập dữ liệu gán nhãn, phân tích đánh giá dữ liệu gán nhãn. (7/2015 – 6/2016) * Xây dựng kiến trúc framework. (6/2016 – 6/2016) * Tìm hiểu và cài đặt các thuật toán khuyến nghị. (6/2016 – 10/2016) * Tìm hiểu và cài đặt các độ đo đánh giá. (6/2016 – 10/2016) * Chạy thực nghiệm, phân tích, giải thích kết quả thực nghiệm. (10/2016 – 11/2016) * Viết báo cáo. (10/2016 – 12/2016) | |
| * **Phân công công việc:**  |  |  | | --- | --- | | Luận | Tuyên | | Viết thuyết minh đề tài | | | Khảo sát hiện trạng các nghiên cứu, các hệ thống liên quan |  | | Xây dựng tập dữ liệu gán nhãn, phân tích đánh giá dữ liệu gán nhãn | | | Xây dựng kiến trúc framework | | | Tìm hiểu và cài đặt thuật toán Content-based, Hybrid | Tìm hiểu và cài đặt thuật toán Collaborative filtering | |  | Tìm hiểu và cài đặt các độ đo đánh giá | | Chạy thực nghiệm, phân tích, giải thích kết quả thực nghiệm | | | Viết báo cáo | | | |
| **Xác nhận của CBHD**  (Ký tên và ghi rõ họ tên) | **TP. HCM, ngày….tháng …..năm…..**  **Sinh viên**  (Ký tên và ghi rõ họ tên) |

MỤC LỤC

[Chương 1. GIỚI THIỆU 3](#_Toc464464707)

[1.1. Dẫn nhập 3](#_Toc464464708)

[1.2. Mục tiêu đề tài 3](#_Toc464464709)

[1.3. Đối tượng nghiên cứu 3](#_Toc464464710)

[1.4. Phạm vi nghiên cứu 3](#_Toc464464711)

[1.5. Bố cục báo cáo 3](#_Toc464464712)

[Chương 2. TỔNG QUAN VỀ HỆ KHUYẾN NGHỊ VÀ CÁC PHƯƠNG PHÁP TIẾP CẬN 4](#_Toc464464713)

[2.1. Dẫn nhập 4](#_Toc464464714)

[2.2. Khái niệm hệ khuyến nghị 4](#_Toc464464715)

[2.3. Phát biểu bài toán khuyến nghị việc làm 4](#_Toc464464716)

[2.4. Các phương pháp tiếp cận cơ bản 4](#_Toc464464717)

[2.4.1. Collaborative filtering approach (Tiếp cận lọc cộng tác) 4](#_Toc464464718)

[2.4.2. Content based approach (Tiếp cận nội dung) 4](#_Toc464464719)

[2.4.3. Hybrid approach (Tiếp cận lai) 4](#_Toc464464720)

[2.5. Các phương pháp đánh giá hệ khuyến nghị 4](#_Toc464464721)

[2.6. Các độ đo đánh giá 4](#_Toc464464722)

[2.6.1. Độ đo Precision 4](#_Toc464464723)

[2.6.2. Độ đo Recall 4](#_Toc464464724)

[2.6.3. Độ đo F-Measure 4](#_Toc464464725)

[2.6.4. Độ đo RMSE (Root Mean Square Error) 4](#_Toc464464726)

[2.6.5. Độ đo NDCG (Normalized Discounted Cumulative Gain) 4](#_Toc464464727)

[2.6.6. Độ đo MAE (Mean Average Precision) 4](#_Toc464464728)

[2.6.7. Độ đo MRR (Mean Reciprocal Rank) 4](#_Toc464464729)

[2.7. Khó khăn và thách thức 4](#_Toc464464730)

[2.8. Kết chương 4](#_Toc464464731)

[Chương 3. PHÂN TÍCH THIẾT KẾ HỆ THỐNG 5](#_Toc464464732)

[3.1. Dẫn nhập 5](#_Toc464464733)

[3.2. Xác định yêu cầu 5](#_Toc464464734)

[3.3. Phân tích yêu cầu 5](#_Toc464464735)

[3.4. Kiến trúc hệ thống 5](#_Toc464464736)

[3.5. Thiết kế dữ liệu 5](#_Toc464464737)

[3.6. Thiết kế giao diện 5](#_Toc464464738)

[3.7. Thiết kế xử lý 5](#_Toc464464739)

[3.8. Kết chương 5](#_Toc464464740)

[Chương 4. THỰC NGHIỆM VÀ NHẬN ĐỊNH 6](#_Toc464464741)

[4.1. Dẫn nhập 6](#_Toc464464742)

[4.2. Môi trường thực nghiệm 6](#_Toc464464743)

[4.3. Dữ liệu thực nghiệm 6](#_Toc464464744)

[4.4. Cách tiến hành thực nghiệm 6](#_Toc464464745)

[4.5. Kết quả thực nghiệm 6](#_Toc464464746)

[4.6. Nhận định về kết quả thực nghiệm 6](#_Toc464464747)

[4.7. Kết chương 6](#_Toc464464748)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 7](#_Toc464464749)

[5.1. Các kết quả đạt được 7](#_Toc464464750)

[5.2. Giá trị thực tiễn của khóa luận 7](#_Toc464464751)

[5.3. Hướng phát triển 7](#_Toc464464752)

DANH MỤC HÌNH VẼ

[Hình 1.1: Tên hình 1 3](#_Toc367742554)

DANH MỤC BẢNG

[Bảng 1.1: Tên bảng 1 3](#_Toc367742567)

[Bảng 2.1: Tên bảng 1 4](#_Toc367742568)

DANH MỤC TỪ VIẾT TẮT

TÓM TẮT KHÓA LUẬN

MỞ ĐẦU

# GIỚI THIỆU

## Dẫn nhập

Tìm kiếm việc làm là một nhu cầu của tất cả mọi người. Với sự trợ giúp của các công cụ đăng tuyển trực tuyến nên các ứng viên có thể tìm kiếm các thông tin ứng tuyển nhanh và dễ dàng hơn. Tuy vậy, các công cụ này thường chứa rất nhiều thông tin việc dẫn đến quá tải. Theo thống kê ngày 22/11/2016 trang tuyển dụng Vietnamworks hiện có 8,116 đăng tuyển [1], trang careerbuider hiện có 8,895 đăng tuyển [2],... với đăng tuyển nhiều như vậy các hệ thống tìm kiếm thông thường sẽ cho ra rất nhiều kết quả khác nhau và gây ra nhiễu cho người dùng. Vì vậy xây dựng một hệ thống để thay thế hoặc hỗ trợ các hệ thống tìm kiếm thông thường sẽ mang lại nhiều lợi ích hơn cho ứng viên và nhà tuyển dụng.

Xây dựng hệ thống khuyến nghị [3] việc làm có thể giúp cho người tìm việc giảm được thời gian tìm kiếm và đem lại hiệu quả cao hơn các hệ thống tìm kiếm thông thường. Đây là đề tài được nhiều khoa học gia trên thế giới quan tâm và có một số website đăng tuyển đã đưa hệ thống khuyến nghị vào sử dụng như Linkedin, Indeed,… và ở Việt Nam cũng đã có một số hệ thống khuyến nghị tương tự. Để có thể tìm ra phương pháp khuyến nghị tốt luận văn này hướng đến xây dựng framework đánh giá các thuật toán khuyến nghị khác nhau, phục vụ cho mục đích nghiên cứu khoa học về xây dựng hệ thống khuyến nghị việc làm.

## Mục tiêu đề tài

Mục tiêu của đề tài hướng đến tìm hiểu và xây dựng một framework dựa trên các phương pháp khuyến nghị và đánh giá có sẵn phục vụ cho việc nghiên cứu hệ thống khuyến nghị. Cụ thể gồm:

* Xây dựng dataset về việc làm, hồ sơ người dùng, dữ liệu gán nhãn.
* Tìm hiểu và cài đặt các thuật toán khuyến nghị sau:
  + Lọc cộng tác.
  + Lọc nội dung.
  + Phương pháp lai đơn giản.
* Tìm hiểu và cài đặt một số độ đo:
  + Precision and recall
  + NDCG
  + RMSE
  + F-Measure
  + MAP
  + MRR
* Xây dựng framework đánh giá các thuật toán khuyến nghị.

## Đối tượng nghiên cứu

* Đối tượng:
  + Các phương pháp khuyến nghị và các độ đo chuẩn.
* Phạm vi:
  + Lĩnh vực tìm kiếm việc làm.

## Bố cục của luận văn

Luận văn được bố cục theo những chương sau:

Chương 1: Giới thiệu về đề tài, nội dung nghiên cứu.

Chương 2: Trình bày định nghĩa về hệ thống khuyến nghị, các phương pháp khuyến nghị, các độ đo chuẩn. Các khó khăn và thách thức khi xây dựng hệ thống khuyến nghị.

Chương 3: Nội dung chương này sẽ trình bày về framework được xây dựng.

Chương 4: Chương này sẽ trình bày các thực nghiệm trên tập dữ liệu do nhóm tự xây dựng, cách thức tiến hành thực nghiệm và các giải thích cho kết quả thực nghiệm.

Chương 5: Ở chương cuối luận văn sẽ trình bày tóm gọn những kết quả của nhóm cũng như hướng phát triển tiếp theo.

# TỔNG QUAN VỀ HỆ KHUYẾN NGHỊ VÀ CÁC PHƯƠNG PHÁP TIẾP CẬN

## Dẫn nhập

Dựa trên mục tiêu, nội dung và đối tượng nghiên cứu, chương này sẽ trình bày một cách tổng quát về hệ khuyến nghị, các phưong pháp tiếp cận và ưu điểm, nhược điểm của từng phương pháp tiếp cận truyền thống, cùng với đó là các phương pháp để đánh giá một thuật toán khuyến nghị.

## Khái niệm hệ khuyến nghị

Hệ khuyến nghị tiếng Anh là Recommender System hay Recommendation System, là một hệ thống con của hệ thống lọc thông tin dùng để tiên đoán giá trị đánh giá hoặc yêu thích mà một người dùng dành cho một đối tượng [4].

Theo Ricci và cộng sự trong Recommender Systems Handbook [5], hệ khuyến nghị là những công cụ phần mềm và các kỹ thuật đưa ra các gợi ý về các đối tượng có thể được sử dụng bởi người dùng. Những gợi ý được đưa ra nhằm hỗ trợ người dùng của hệ thống đưa ra các quyết định chẳng hạn nên mua gì, nên nghe nhạc gì hay nên đọc tin tức gì. Hệ khuyến nghị đã chứng tỏ giá trị ý nghĩa đối với những người dùng online, giúp họ đương đầu với việc quá tải thông tin và trở thành một trong những công cụ mạnh mẽ và phổ biến nhất trong thương mại điện tử.

Theo quan điểm của luận văn, hệ khuyến nghị là một hệ thống thông minh với các thuật toán có khả năng tiên đoán được những đối tượng người dùng có thể quan tâm, nhằm mục đích giúp người dùng nhanh chóng tiếp cận được những gì mà người dùng cần.

## Phát biểu bài toán khuyến nghị

Hiện nay có nhiều nghiên cứu liên quan đã phát biểu về bài toán khuyến nghị và các khái niệm liên quan như Jannach và cộng sự [6], Adomavicius và Tuzhilin [7], Bobadilla và cộng sự [8]. này sẽ dựa và các nghiên cứu liên quan trên để trình bày một số khái niệm liên quan và phát biểu bài toán khuyến nghị.

1. Không gian người dùng [7] [9]

Không gian người dùng là tập tất cả các người dùng được quan sát bởi hệ thống để thực hiện khuyến nghị, ký hiệu:

1. Không gian đối tượng khuyến nghị [7] [9]

Không gian đối tượng khuyến nghị là tất cả các đối tượng có khả năng sẽ được hệ thống khuyến nghị cho người dùng. Tùy vào mục tiêu của mỗi hệ thống mà đối tượng khuyến nghị sẽ là gì, chẳng hạn như phim, các bài báo, nhạc… Với hệ thống mà khóa luận đang phát triển thì các đối tượng khuyến nghị chính là những việc làm đã đăng tuyển. Ký hiệu:

1. Hàm hữu ích [9]

Hàm hữu ích là một phép ánh xạ: , dùng để xác định giá trị hữu ích của đối tượng khuyến nghị với người dùng . Với là tập các giá trị hữu ích có giá trị trong một khoảng cụ thể và được sắp xếp theo thứ tự giảm dần.

**Phát biểu bài toán khuyến nghị**

Cho trước:

* Không gian người dùng
* Không gian đối tượng khuyến nghị

Tìm hàm hữu ích , xác định giá trị hữu ích của đối tượng khuyến nghị đối với người dùng . Với mỗi người dùng hệ khuyến nghị cần trả về một danh sách các đối tượng khuyến nghị có giá trị hữu ích cao nhất được sắp xếp theo thứ tự giảm dần.

Việc làm thế nào tìm được hàm hữu ích hiện nay đã có rất nhiều các phương pháp, hướng tiếp cận khác nhau. Phần tiếp theo sẽ trình bày về các phương pháp tiến cận truyền thống phổ biến nhất.

## Các phương pháp tiếp cận cơ bản

Theo Adomavicius và Tuzhilin [7], hệ khuyến nghị thường được chia thành 3 hướng tiếp cận sau dựa trên cách thức thực hiện khuyến nghị gồm: Collaborative filtering approach (tiếp cận lọc cộng tác), content based approach (tiếp cận nội dung) và hybrid (tiếp cận lai).

### Collaborative filtering approach (Tiếp cận lọc cộng tác)

Collaborative filtering approach (tiếp cận lọc cộng tác) hay viết tắt là CF, là phương pháp tiếp cận dựa trên dữ liệu hành vi, sở thích trong quá khứ của người dùng – ma trận đánh giá, ý tưởng cơ bản của CF là nếu người dùng có cùng sở thích ở quá khứ thì cũng sẽ có cùng sở thích trong tương lai.

1. Ma trận đánh giá [9]

Cho không gian người dùng và không gian các đối tượng khuyến nghị . Ma trận kích thước chứa các giá trị đánh giá với . Những giá trị đánh giá thể hiện mức độ hữu ích của đối tượng với người dùng (hay ). Giá trị có thể là nguyên hay thực trong một khoảng cố định nào đó tùy thuộc vào bài toán cụ thể. Thông thường giá trị đánh giá trong các hệ thống ứng dụng phổ biến thường từ 1 (không hữu ích) đến 5 (rất hữu ích). Nếu một người dùng chưa thể hiện đánh giá với một đối tượng thì và cần được tính toán (dấu ? trong hình 2.1)



Hình 2.1. Dấu ? là các giá trị cần tiên đoán trong ma trận đánh giá.

Tiếp cận lọc cộng tác được xem là hướng tiếp cận phổ biến nhất và thành công nhất để xây dựng hệ khuyến nghị cho các hệ thống thương mại điện tử [9] [6]. Với rất nhiều các nghiên cứu, cài đặt và thực nghiệm đã thực hiện trước đó, CF được chia thành 2 nhóm chính gồm: CF dựa trên bộ nhớ và CF dựa trên mô hình. Trong phạm vi của đề tài, khóa luận tiến hành tìm hiểu cài đặt và thực nghiệm với các thuật toán CF dựa trên bộ nhớ.

CF dựa trên bộ nhớ dùng các kỹ thuật của thống kê để tính toán tìm người dùng hoặc đối tượng khuyến nghị tương tự từ thông tin trong ma trận đánh giá. Tiếp cận CF dựa trên bộ nhớ tìm cách ước lượng giá trị của hàm hữu ích , thể hiện cho mức độ hữu ích của đối tượng khuyến nghị đối với người dùng . Dựa trên ma trận đánh giá, CF sẽ khuyến nghị cho người dùng các đối tượng tương tự với đối tượng mà người dùng đã đánh giá trước đó (Item based) hoặc khuyến nghị các đối tượng mà những người dùng có đồng sở thích với người dùng đã đánh giá (User based).

Tiếp theo sẽ trình bày chi tiết về 2 phương pháp lọc dựa trên người dùng (User based) và lọc dựa trên đối tượng khuyến nghị (Item based).

#### Lọc dựa trên người dùng

1. Những người dùng đồng sở thích

Những người dùng đồng sở thích là những người có sở thích hay những đánh giá trong quá khứ tương tự với người dùng trên cùng những đối tượng khuyến nghị từ ma trận đánh giá. Ký hiệu là những người dùng đồng sở thích với , trong đó , , là không gian người dùng của hệ thống.

**Ý tưởng chính:** ý tưởng thực hiện của phương pháp lọc dựa trên người dùng gồm 3 bước chính: đầu tiên xác định danh sách những người dùng có cùng sở thích với người dùng hay ; Bước 2 tiến hành ước lượng giá trị hàm hữu ích của đối tượng khuyến nghị với người dùng bằng cách tổng hợp giá trị đánh giá của đối với . Bước 3 là thực hiện khuyến nghị dựa trên giá trị hàm hữu ích ước lượng được [9] [6].

**Các bước thực hiện:** các tác giả Owen và cộng sự [10] đã tiến hành cài đặt thuật toán với các bước có thể tóm gọn theo mã giả như sau:

Trong đó các phương pháp để tính độ tương tự (), làm thế nào để xác định được neighborhood và tổng hợp có trọng số giá trị đánh giá sẽ được trình bày tiếp theo.

Tính độ tương tự ()

Để tính độ tương tự giữa các người dùng, Owen và cộng sự [10] đã tiến hành cài đặt và thực nghiệm với các công thức tính độ tương tự của:

*Pearson correlation coefficient* [11]

Trong đó:

* + - N là số đối tượng khuyến nghị p mà u và cùng đánh giá
    - với là giá trị đánh giá của u cho p
    - là tổng các giá trị đánh giá của u
    - là tổng các giá trị đánh giá của
    - là tổng của bình phương các giá trị đánh giá của u
    - là tổng của bình phương các giá trị đánh giá của

***Tanimoto coefficient*** [12]

Trong đó:

* a là số đối tượng khuyến nghị u đã đánh giá
* b là số đối tượng khuyến nghị đã đánh giá
* c là số đối tượng khuyến nghị cả u và đã đánh giá

***Euclidean distance*** [13][14]

Kết quả không thuộc một khoảng cố định nên sẽ được normalized thành:

Trong đó:

* , là giá trị đánh giá của u, trên cùng 1 đối tượng khuyến nghị
* n là số đối tượng khuyến nghị mà u và cùng đánh giá
* với là tổng số đối tượng khuyến nghị

***LogLikelihood similarity*** [15]

Trong đó:

* k11 là những đối tượng khuyến nghị được u và đánh giá
* k12 là những đối tượng khuyến nghị được u đánh giá mà không đánh giá
* k21 là những đối tượng khuyến nghị được đánh giá mà u không đánh giá
* k22 là những đối tượng khuyến nghị không được cả u và đánh giá

Cách xác định TopN người dùng đồng sở thích [6] [10]

Việc xác định danh sách những người dùng đồng sở thích sẽ ảnh hưởng trực tiếp đến kết quả của hàm hữu ích, vì vậy ta cần cân nhắc để chọn danh sách TopN tốt nhất. Nếu chúng ta xem xét tất cả các người dùng đồng sở thích sẽ ảnh hưởng đến hiệu suất của việc tính toán và cũng sẽ ảnh hưởng tiêu cực đến kết quả tính toán bởi những người dùng có mức độ tương tự thấp.

Trong thực tiễn có 2 phương pháp được sử dụng phổ biến đó là lấy danh sách tất cả người dùng có độ tương tự lớn hơn một ngưỡng (threshold) nhất định hoặc lấy danh sách TopN cố định n người dùng có độ tương tự cao nhất.

Với cách thứ nhất ta có phải xem xét ngưỡng (threshold) ở mức nào là hợp lý, vì nếu chọn quá cao thì danh sách người dùng sẽ rất hạn chế (thậm chí không có), còn nếu chọn thấp quá thì danh sách này dường như không bị cắt giảm là bao so với ban đầu.

Với cách thứ 2 kích thước của tập người dùng đồng sở thích sẽ không thể bao phủ, giá trị n được chọn nếu quá cao sẽ không giới hạn được danh sách người dùng (tương tự khi chọn threshold thấp) điều này sẽ dẫn đến nhiễu dữ liệu tiên đoán, nhưng nếu chọn n quá thấp sẽ ảnh hưởng tiêu cực đến khả năng tiên đoán của hệ thống.

Phương pháp tổng hợp đánh giá

Sau khi đã tìm được danh sách người dùng có cùng sở thích và mức độ tương tự (giá trị ) ta cần tổng hợp lại để được kết quả cuối cùng của hàm hữu ích để thực hiện khuyến nghị.

Owen và cộng sự [10] đã nghiên cứu và cài đặt với công thức sau cho thư viện Mahout [16]:

Trong đó:

* + N là số người dùng có sở thích tương tự với u
  + M là số người dùng có đánh giá cho đối tượng p
  + là giá trị đánh giá của u với p trong ma trận đánh giá

#### Lọc dựa trên đối tượng khuyến nghị

Tương tự như lọc dựa trên người dùng, lọc dựa trên đối tượng khuyến nghị cũng với ý tưởng và các bước thực hiện khá giống nhau gồm 3 bước chính: đầu tiên xác định danh sách những đối tượng khuyến nghị tương tự nhất với đối tượng khuyến nghị ; Bước 2 tiến hành ước lượng giá trị hàm hữu ích của đối tượng khuyến nghị với người dùng bằng cách tổng hợp giá trị đánh giá của cho đối với . Bước 3 là thực hiện khuyến nghị dựa trên giá trị hàm hữu ích ước lượng được [9] [6].

### Tiếp cận nội dung (Content based approach)

Hướng tiếp cận dựa trên nội dung (Content based approach) [17] là một dạng đặc biệt của bài toán truy vấn và lọc thông tin [18]. Hướng tiếp cận này tìm cách khuyến nghị cho người dùng những đối tượng tương tự với những đối tượng mà người dùng quan tâm trong quá khứ [18]. Hay nói cách khác tiếp cận theo hướng nội dung trả lời cho câu hỏi sau của người sử dụng: “Hãy cho tôi biết những đối tượng tương tự với những đối tượng mà tôi thích” [18]. Các đối tượng ở đây có thể là các video ca nhạc, game, bài báo, việc làm, sản phẩm,… Ví dụ, một người thường xem những công việc liên quan đến những từ khóa như java, mysql, lập trình,… thì hệ thống đưa ra cho người đó những việc làm chứa những từ khóa liên quan mà người dùng chưa xem.

Để xây dựng một hệ thống khuyến nghị tiếp cận theo hướng nội dung chúng ta cần cung cấp cho hệ thống tập các thông tin sẵn có về đối tượng được khuyến nghị như thể loại, mô tả, tính chất, đặc điểm… và một tập dữ liệu mô tả sở thích của người dùng. Nhiệm vụ của hệ thống là “học” sở thích của người dùng và tìm ra những đối tượng có đặc điểm tương tự với sở thích của người dùng.

Độ tương tự của người dùng u và đối tượng i có thể tính bằng độ tương tự Cosine [19], khoảng cách Euclide [13], hoặc sử dụng Tanimoto [20] để tính toán. Thông thường các nghiên cứu sử dụng độ tương tự Cosine để tính toán. Công thức tính Cosine được biểu diễn như sau:

Trong đó A và B là các vector đặc trưng.

Có nhiều cách để biểu diễn nội dung của đối tượng được khuyến nghị như biểu diễn ở dạng cây, biểu diễn ở dạng văn có cấu trúc và biểu diễn văn bản phi cấu trúc. Thông thường người ta chọn phương pháp biểu diễn nội dung của đối tượng dưới dạng một vector đặc trưng [18]. Các bước để xây dựng hệ thống khuyến nghị dựa trên nội dung như sau:

* Bước 1: Biểu diễn nội dung của những đối tượng được khuyến nghị *i* dưới dạng một vector đặc trưng(feature vector) [19].
* Bước 2: Xây dựng và biểu diễn sở thích của người dùng *u* dưới dạng một vector đặc trưng.
* Bước 3: Tính toán độ tương tự cosine [20] của các đối tượng *i* thuộc tập đối đượng được khuyến nghị đối với người dùng *u*.
* Bước 4: Đưa ra *top N* những đối tượng có độ tương tự cao nhất.

Mã giả cho thuật toán:

Input: UserSet, JobSet, RatingSet

foreach(User in UserSet){

User = ModelingUserProfile(User)

}

foreach(Job in JobSet){

Job = ModelingItem(Job)

}

foreach(Job in JobSet){

JobVector = GetFeaturedVector(Job);

foreach(User in UserSet){

UserVector=GetFeaturedVector(User,GetUserRating(User, RatingSet));

Weight = Cosine(JobVector, UserVector);

If Weight is in TopN Then

AddToTopN(User,Job, Weight)

}

}

### Simple Hybrid approach (Tiếp cận lai đơn giản)

Có rất nhiều phương pháp khuyến nghị khác nhau. Mỗi phương pháp có một điểm mạnh và điểm yếu riêng. Phương pháp lai ra đời nhằm tận dụng các điểm mạnh của các phương pháp khuyến nghị khác nhau và hạn chế các điểm yếu [9]. Phương pháp lai kết hợp kết quả khuyến nghị của các phương pháp khuyến nghị khác theo một cách nào đó nhằm đưa ra một hệ thống khuyến nghị tốt hơn [18]. Có nhiều các lai khác nhau. Cách tiếp cận lai đơn giản nhất là sử dụng phương pháp lai có trọng số(Weighted) [18] [17]. Mỗi kết quả đầu ra của các phương pháp *i* sẽ được kết hợp tuyến tính với nhau theo một hệ số *βi*. Các hệ số *βi* này sẽ dần được hiệu chỉnh cho đến khi thu được kết quả tốt nhất [9] [22].

Công thức dự đoán cho phương pháp lai có trọng số [18]



Mã giả thuật toán khuyến nghị dựa trên phương pháp lai đơn giản

Input: UserSet, JobSet, RatingSet, Alpha

CFResult = RunCF(RatingSet);

CBResult = RunCB(UserSet,JobSet, RatingSet);

foreach(Job in JobSet)

{

foreach(User in UserSet)

{

CFScore = GetScore(CFResult, User, Job);

CBScore = GetScore(CBResult, User, Job);

HBScore = CFScore \* Alpha + (1 - Alpha) \* CBScore;

If(HBScore in topN) Then

AddToTopN(User,Job, HBScore)

}

}

## Các phương pháp đánh giá hệ khuyến nghị

## Các độ đo đánh giá

### Độ đo Precision

### Độ đo Recall

### Độ đo F-Measure

### Độ đo RMSE (Root Mean Square Error)

### Độ đo NDCG (Normalized Discounted Cumulative Gain)

### Độ đo MAE (Mean Average Precision)

### Độ đo MRR (Mean Reciprocal Rank)

## Khó khăn và thách thức

## Kết chương

# PHÂN TÍCH THIẾT KẾ HỆ THỐNG

## Dẫn nhập

Trong chương trước, nhóm đã trình bày về một số vấn đề về các phương pháp tiếp cận và đánh giá hệ thống khuyến nghị. Chương này, sẽ trình bày chi tiết về framework bao gồm kiến trúc, công nghệ, nền tảng.

## Xác định yêu cầu

### Danh sách các yêu cầu:

|  |  |  |
| --- | --- | --- |
| Mã Yêu cầu | Tên yêu cầu | Ghi chú |
| 1.1 | Upload dataset |  |
| 1.2 | Chạy các thuật toán khuyến nghị |  |
| 1.3 | Chạy đánh giá các thuật toán khuyến nghị |  |

### Đặc tả yêu cầu upload dataset

* Để bắt đầu sử dụng hệ framework. Người dùng phải cung dataset cho hệ thống. Dataset được upload bao gồm 03 file.
  + File Cv.txt: Chứa dữ liệu về người dùng bao gồm các thông tin về cv người dùng.
  + File Job.txt: Chứa dữ liệu về việc làm sẽ được khuyến nghị.
  + Score.txt: Chứa dữ liệu về sở thích của người dùng trong file Cv.txt đối với những việc làm trong file Job.txt
* Sau khi chọn file người dùng chọn chức năng upload để bắt đầu upload dữ liệu. Hệ thống sẽ lưu trữ dataset vào thư mục có dạng: mã người dùng/tên dataset và thông báo kết quả thực hiện cho người dùng.

### Đặc tả yêu cầu chạy thực nghiệm thuật toán khuyến nghị

* Người dùng tạo chọn dataset cần thực nghiệm, chọn thuật toán cần chạy. Nếu tập dữ liệu chưa có thì người dùng có thể upload lên hệ thống. Sau đó người dùng bấm chạy. Hệ thống sẽ thực hiện chạy thuật toán khuyến nghị. Sau khi chạy xong. Hệ thống sẽ thông báo cho người dùng kết quả thực nghiệm và lưu trữ lại vào file output.txt.
* Cấu trúc file output gồm 3 cột: UserId, JobId, Score. Các cột ngăn cách nhau bằng dấu kí tự tab.

### Đặc tả chức năng chạy đánh giá thuật toán khuyến nghị

* Người dùng sẽ nhập những file output.txt sinh ra khi chạy thực nghiệm vào hệ thống. Hệ thống sẽ thực hiện tính toán và trả về kết quả kết quả so sánh các thuật toán khuyến nghị cho người dùng.

## Phân tích yêu cầu

### Biểu đồ phân rã chức năng hệ thống

Recomender evaluation system

Upload dataset

Run algorithms

Evaluate algorithms

### Sơ đồ tổng thể hệ thống



### Mô tả xử lý upload dataset

* Sơ đồ DFD của xử lý



* Mô tả xử lý: người dùng đưa dataset gồm 3 file job.txt, Cv.txt và Score.txt vào hệ thống. Hệ thống kiểm tra và lưu dataset lại.
* Các biểu mẫu liên quan:
  + File Job.txt: Định dạng file text, lưu trữ được kí tự unicode, các thuộc tính khác nhau được ngăn cách bằng dấu tab. Gồm các thuộc tính JobId,JobName, Category, JobRequirement, JobDescription, Tag, JobBenefit, Salary.
  + File Cv.txt: Định dạng file text, lưu trữ được kí tự unicode, các thuộc tính khác nhau được ngăn cách bằng dấu tab. Gồm các thuộc tính CvId, UserId, UserName, Education, Interested, ForeignLangue, ExpectedSalary.
  + File Score.txt: Định dạng file text, các thuộc tính khác nhau được ngăn cách bằng dấu tab. Gồm ba thuộc tính JobId, UserId, Score

### Mô tả xử lý chạy các thuật toán khuyến nghị

* Sơ đồ DFD của xử lý:



* Mô tả xử lý: Người dùng chọn thuật toán cần chạy thực nghiệm và tên của dataset. Hệ thống sẽ lấy dataset tương ứng từ kho User Dataset và chạy thực nghiệm dựa thuật toán mà người dùng chọn.
* Các biểu mẫu liên quan: Không có.

### Mô tả xử lý chạy đánh giá các thuật toán khuyến nghị

* Sơ đồ DFD của xử lý



* Mô tả xử lý: người dùng đưa vào tên các kết quả thực nghiệm cần so sánh. Hệ thống sẽ lấy từ kho result-set các kết quả thực nghiệm tương ứng. Sau đó hệ thống sẽ so sánh và trực quan hóa kết quả cho người dùng.
* Các biểu mẫu liên quan: Không có.

## Kiến trúc hệ thống

### Kiến trúc tổng quát

**SYSTEM DB**

**User DataSet**

**Spring MVC**

**Apache Mahout**

**Apache Lucene**

**RecAlg****EvaluationSystem**

**Dataset Analyzation**

**Runing & evaluation**

**System Interfaces**

**System core modules**

**System DB**

**Libraries**

### Các thành phần trong kiến trúc

* + System interfaces
* RecAlgComparisonSystem: Module này cung cấp cho người dùng chức năng quản lý dataset, xem thống kê dataset, quản lý các tiến trình thực thi các thuật toán khuyến nghị, xem kết quả và các đánh giá kết quả khuyến nghị.
  + System core modules
* Dataset analyzation: Đây là mô-đun phân tích dataset của người dùng.
* Experimental conduction & evaluation: chạy thực nghiệm các thuật toán khuyến nghị và đánh giá kết quả thực hiện.
  + Database
* System DB: database dùng để tổ chức lưu trữ thông tin người dùng, các tác vụ thực hiện của người dùng và các thông tin thiết lập chạy thuật toán.
* User dataset: hệ thống lưu trữ files dataset do người dùng upload cùng với đó là các file kết quả khuyến nghị, kết quả so sánh và kết quả thống kê.
  + Libraries
* Spring MVC: dùng hỗ trợ xây dựng website tương tác người dùng.
* Mahout: sử dụng các thư viện recommend có sẵn trong Mahout để thực hiện khuyến nghị.
* LUCENE: Dùng để tính toán tf-idf

## Thiết kế dữ liệu

## Thiết kế giao diện

## Thiết kế xử lý

## Kết chương

# THỰC NGHIỆM VÀ NHẬN ĐỊNH

## Dẫn nhập

## Môi trường thực nghiệm

## Dữ liệu thực nghiệm

## Cách tiến hành thực nghiệm

## Kết quả thực nghiệm

## Nhận định về kết quả thực nghiệm

## Kết chương

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Các kết quả đạt được

## Giá trị thực tiễn của khóa luận

## Hướng phát triển

TÀI LIỆU THAM KHẢO

Theo chuẩn IEEE