**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

## Mục tiêu đề tài

## Đối tượng nghiên cứu

## Phạm vi nghiên cứu

## Bố cục báo cáo

# 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 [1].

Theo Ricci và cộng sự trong Recommender Systems Handbook [2], 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 [3], Adomavicius và Tuzhilin [4], Bobadilla và cộng sự [5]. 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 [4] [6]

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 [4] [6]

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 [6]

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 [4], 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 [6]

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 [6] [3]. 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 những người dùng đồng sở thích với với là với , , là không gian người dùng của hệ thống.

#### Lọc dựa trên đối tượng khuyến nghị

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

### Hybrid approach (Tiếp cận lai)

## 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

## Xác định yêu cầu

## Phân tích yêu cầu

## Kiến trúc hệ thống

## 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