|  |  |
| --- | --- |
|  |  |
| **HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**  **---------------------------------------** | |
|  | |
|  | |
| **LÊ TRUNG HIẾU** | |
|  | |
|  | |
|  | |
| **ĐỀ CƯƠNG**  **LUẬN VĂN THẠC SĨ KỸ THUẬT** | |
|  | |
|  | |
| HÀ NỘI - 2016 | |
|  |  |
| **HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**  **---------------------------------------** | |
|  | |
|  | |
| **LÊ TRUNG HIẾU** | |
|  | |
|  | |
|  | |
| **DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET**  **DỰA TRÊN LỊCH SỬ TRUY CẬP** | |
| CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN | |
| MÃ SỐ: 60.48.01.04  **ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ KỸ THUẬT**  **NGƯỜI HƯỚNG DẪN KHOA HỌC**  **PGS.TS. TỪ MINH PHƯƠNG** | |
| HÀ NỘI - 2016 | |

**I. MỞ ĐẦU**

1. **Lý do chọn đề tài**

Ngày nay, người ta thường dành một lượng lớn thời gian trong ngày để truy cập internet. Internet được người dùng sử dụng cho việc tìm kiếm thông tin, đọc tin tức, mua sắm, chơi trò chơi vv. Và các nhà quảng cáo không thể bỏ lỡ cơ hội để tiếp thị trực tuyến đến với khách hàng của họ nhằm cung cấp các dịch vụ phù hợp với nhu cầu của tổ chức, cá nhân sử dụng mạng internet. Nhưng người dùng thường phải đối mặt với số lượng lớn các thông tin không mong muốn, điều quan trọng là họ luôn cố gắng thoát ra khỏi và bỏ qua những thông tin không cần thiết. Tình trạng quá tải thông tin không đến đích này dẫn đến sự sụt giảm đáng kể trong hiệu quả tiếp thị trực tuyến.

chỗ này em phải mô tả ngắn gọn bài toán của em là gì

Từ đó việc xác định giới tính người dùng được coi là một phương pháp để đưa ra các số liệu thông kê, kế hoạch quảng cáo và cung cấp các dịch vụ đến người dùng trên mạng internet. (Phương pháp nhắm mục tiêu theo hành vi là một phương pháp hiệu quả để giúp các nhà quảng cáo nhắm mục tiêu đến người dùng một cách thích hợp khi theo dõi hành vi sử dụng internet của họ.) Đây chính là một trong các kỹ thuật nhắm mục tiêu theo hành vi mang lại hiệu quả, lợi ích cao, vì vậy đề tài luận văn này của em sẽ tập trung vào phương pháp dự đoán giới tính người dùng internet dựa trên lịch sử truy cập sử dụng kỹ thuật học máy SVM (Support Vector Machine).

1. **Mục đích nghiên cứu**

Mục đích nghiên cứu là tìm hiểu và thử nghiệm được phương pháp dự đoán giới tính người dùng dựa trên lịch sử truy cập của học bằng cách sử dụng kỹ thuật học máy. Sau đó em chi tiết ra thành các mục tiêu cụ thể hơn như tìm hiểu được bài toán và các phương pháp đã có, hiểu được về học máy và SVM, áp dụng và thử nghiệm.

Bằng nhận thức khi sử dụng internet ta nhận thấy việc đưa thông tin tiếp thị, tin tức phù hợp đến một người dùng nào đó thật hữu ích khi mà người đó quan tâm hoặc sẵn sàng để truy cập dịch vụ. Với kết quả và dữ liệu thu được, ta có thể dự đoán một người dùng truy cập vào mạng internet có thể là nam hay nữ. Cùng một loạt các kỹ thuật học máy cụ thể kỹ thuật học máy SVM chúng ta có thể đưa ra tỉ lệ tương đối chính xác về vấn đề dự đoán giới tính người dùng. Dựa vào đó, chúng ta đưa ra các số liệu cụ thể về sự khác biệt về mặt giới tính của người truy cập internet. Đây là cũng chính là cách tiếp cận của các cá nhân, tổ chức và công ty cho việc nhắm mục tiêu thích hợp đến các hành vi người dùng.

Hiện nay, mặc dù đã có nhiều sự tồn tại của một số nghiên cứu về việc nhắm mục tiêu theo hành vi như … tuy nhiên những nghiên cứu này vẫn chưa được bám sát được mục tiêu, mức độ nghiên cứu vẫn chỉ trong phạm vi hẹp, đặc biệt một vấn đề quan trọng là dữ liệu đầu vào cho các phương pháp gần như bị bỏ qua hoàn toàn trong các tài liệu. Nghiên cứu này nhằm mục đích thực hiện một tổng quan về các phương pháp trước đó và hiệu suất của các mô hình dự đoán đạt được để so sánh sự khác biệt về độ chính xác của các mô hình từ đó đưa ra phân tích và thảo luận với mô hình dự đoán giới tính của người dùng internet dựa trên dữ liệu lịch sử truy cập nhằm đưa ra các tỉ lệ, độ chính xác và cung cấp các kiến thức khoa học hiện tại khi dự đoán giới tính và khả năng hiệu quả khi sử dụng đến kỹ thuật học máy.

1. **Đối tượng và phạm vi nghiên cứu**

* Đối tượng: Kỹ thuật học máy SVM, dữ liệu lịch sử truy cập của người dùng internet.
* Phạm vi nghiên cứu: Xác định giới tính người dùng internet sử dụng lịch sử truy cập.

1. **Phương pháp nghiên cứu**

* Tìm hiểu các phương pháp dự đoán giới tính hiện này đang có để xác định những điểm mạnh và hạn chế của các phương pháp đó.
* Tìm hiểu các kỹ thuật học máy hiện nay, nhưng ưu nhược điểm của các phương pháp học máy đó trong việc dự đoán giới tính.
* Tìm kiểu chi tiết kỹ thuật học máy SVM và ứng dụng hiện này của phương pháp này.
* Tìm hiểu các đặc trưng lịch sử truy cập của người dùng trong phạm vi nghiên cứu. Xác định những yếu tốt ảnh hưởng đến giới tính người dùng và các cách để lấy dữ liệu để nghiên cứu.
* Cách phân tích, chuẩn hóa dữ liệu sử dụng kỹ thuật học máy SVM và đưa ra kết quả bằng phần mềm Weka.

**II. NỘI DUNG**

**Chương 1: TỔNG QUAN VỀ DỮ LIỆU LỊCH SỬ TRUY CẬP & CÁC ĐẶC TRƯNG**

Giới thiệu chương: Giới thiệu về dữ liệu truy cập của người dùng internet thông qua thống kê, các khái niệm và đặc trưng trong tập dữ liệu này, bao gồm các mối quan hệ giữa các trang thông tin và người dùng mạng, những hành vi của người dùng khi truy cập internet, cách thức truy cập, tìm kiếm thông tin. Giới thiệu những phương pháp nhắm mục tiêu theo hành vi hiện nay được áp dụng cho người dùng internet và những hạn chế của các phương pháp này.

Nội dung chương 1 sẽ bố cục theo các mục sau:

* 1. Bài toán xác định giới tính và ứng dụng.
  2. Các dạng dữ liệu lịch sử có thể dùng
  3. Các phương pháp xác định giới tính đã có.
     1. Phương pháp 1
     2. Phương pháp 2.
     3. ...
  4. Kết luận chương

**Chương 2: KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG VIỆC DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG INTERNET**

Giới thiệu chương: Trình bày tổng quan về kỹ thuật học máy, một số kỹ thuật học máy và tập trung vào kỹ thuật được sử dụng trong luận văn là kỹ thuật học máy SVM.

Dựa vào những đặc trưng việc truy cập thông tin của người dùng internet, đưa ra phương pháp dự đoán giới tính áp dụng kỹ thuật học máy và xếp hạng tỉ lệ độ chính xác nhằm tăng hiệu quả dự đoán so với các phương pháp đang tồn tại.

Nội dung chương 2 sẽ bố cục theo các mục sau:

* 1. Phương pháp học máy SVM.
  2. Giới thiệu về dữ liệu sẽ sử dụng
  3. Các đặc trưng sẽ dùng trong phân lớp
     1. Dạng đặc trưng 1
     2. Dạng đặc trưng 2
     3. ....
  4. Kết luận chương

**Chương 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ**

Giới thiệu chương: đề tài của em sử dụng dữ liệu PAKDD 2015 là dữ liệu của FPT chứ có phải của VCcorp đâu nhỉ ??? Sử dụng dữ liệu có sẵn được cung cấp bởi VCcorp (http://vccorp.vn), thực hiện xây dựng bộ dữ liệu từ dữ liệu thực tế chưa chuẩn hóa hiện có của VCcorp cho một số lượng người dùng, sử dụng kỹ thuật học máy SVM ở chương 2 và phần mềm Weka để đưa ra tỉ lệ, độ chính xác của phương pháp dự đoán giới tính dựa trên lịch sử truy cập. Đánh giá kết quả so với các phương pháp dự đoán khác, và so sánh với cách làm việc hiện tại trong việc dự đoán giới tính.

Nội dụng chương 3 sẽ bố cục theo các mục sau:

* 1. Thu thập dữ liệu thử nghiệm từ lịch sử truy cập
  2. Các tiêu chuẩn đánh giá.
  3. Kết quả thực nghiệm
  4. Độ phức tạp và thời gian thực hiện phương pháp.
  5. Kết luận chương

**III. KẾT LUẬN**

Áp dụng kỹ thuật học máy SVM trong việc dự đoán giới tính người dùng internet dựa trên lịch sử truy cập mang lại hiệu quả và tính tối ưu cao, vì phương pháp này kết hợp được nhiều yếu tố, và đặc biệt nhắm mục tiêu theo hành vi của từng người dùng, đưa ra những cách tiếp cận của các cá nhân, tổ chức và công ty phù hợp với người dùng đó. Tuy nhiên phương pháp vẫn còn một số hạn chế cần cải thiện trong tương lai, và đề xuất các hướng đi tiếp theo nhằm nâng cao hơn nữa hiệu quả của việc dự đoán giới tính người dùng internet.

**IV. DANH MỤC CÁC TÀI LIỆU THAM KHẢO**

1. Do Viet Phuong and Tu Minh Phuong. “*Gender Prediction Using Browsing History”.* [KSE (1) 2013](http://dblp.uni-trier.de/db/conf/kse/kse2013-1.html#PhuongP13): 271-283.
2. Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z. (2007) “*Demographic prediction based onuser’s browsing behavior*”, Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada. [viewed 05.09.2016] Available from:  
   <http://wwwconference.org/www2007/papers/paper686.pdf>
3. Kabbur, S., Han, E.-H., Karypis, G. (2010) “*Content-based methods for predicting website demographic attributes*”, University of Minnesota Supercomputing Institute Research Report UMSI 2010/98 [viewed 06.09.2016]

Available from:  
<http://www.dtc.umn.edu/publications/reports/2010_01.pdf>

[4] Speltdoorn, S. (2010) “*Predicting demographic characteristics of web users using semisupervised classification techniques*” Master’s dissertation, Ghent University, Faculty of Economucs and Business Administration. [viewed 14.09.2016] Available from:  
<http://lib.ugent.be/fulltxt/RUG01/001/459/756/RUG01001459756_2011_0001_AC.pdf>

[5] Quanzeng You, Sumit Bhatia, Tong Sun, Jiebo Luo (2014) “*The eyes of the beholder: Gender prediction using images posted in Online Social Networks*”. Available from: <http://www.cs.rochester.edu/u/qyou/papers/gender_classification.pdf>

**V. DỰ KIẾN KẾ HOẠCH THỰC HIỆN**

|  |  |  |
| --- | --- | --- |
| **TT** | **Nội dung** | **Thời gian** |
| 1 | Tìm hiểu kỹ thuật học máy và kỹ thuật học máy SVM (Support Vector Machine). | 09/2016 – 10/2016 |
| 2 | Nghiên cứu các phương pháp dự đoán nhắm mục tiêu theo hành vi hiện tại, và xác định những mặt hạn chế của các phương pháp này | 10/2016 – 11/2016 |
| 3 | Nghiên cứu kỹ thuật học máy SVM và áp dụng trong việc dự đoán giới tính người dùng internet dựa trên lịch sử truy cập | 11/2016 – 12/2016 |
| 4 | Cài đặt thuật toán, xây dựng bộ dữ liệu | 12/2016 – 02/2017 |
| 5 | Thử nghiệm và đánh giá thuật toán | 02/2017 |

|  |  |
| --- | --- |
| **Ý KIẾN CỦA GIÁO VIÊN HƯỚNG DẪN**  *(Ký ghi rõ họ tên)* | **NGƯỜI LẬP ĐỀ CƯƠNG**  *(Ký ghi rõ họ tên)* |
| **DUYỆT CỦA TRƯỞNG TIỂU BAN CHẤM ĐỀ CƯƠNG**  *(Ký ghi rõ họ tên)* | |