ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN THỊ TƯƠI

ỨNG DỤNG CÁC MÔ HÌNH CHỦ ĐỀ ẨN VÀO MÔ HÌNH PHÂN HẠNG LẠI DÒNG CẬP NHẬT TRÊN MẠNG XÃ HỘI TWITTER

Ngành: Hệ thống thông tin
Chuyên ngành: Hệ thống thông tin
Mã số: 60480104

TÓM TẮT LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. HÀ QUANG THỤY

Hà Nội - 2016

MỞ ĐẦU

Ngày nay, mạng xã hội phát triển mạnh mẽ mang những nhận xét, đánh giá, những thông tin phản ánh xã hội thực tới mỗi người, và ngày càng đi sâu vào cuộc sống của mỗi chúng ta. Chúng cung cấp nhiều thông tin cập nhật có tính thời gian thực có được từ kết nối trực tuyến của mọi người. Dòng các tin mới đến trang cá nhân của mỗi người dùng được gọi là dòng cập nhật của người dùng đó. Mặc dù dòng cập nhật đưa đến những thông tin mới, nhưng tồn tại một hạn chế là không ít người dùng đã phải dành khá nhiều thời gian với dòng cập nhật, vì có không ít tin mới trong dòng cập nhật mang lại thông tin không cần thiết cho họ. Nhiều người dùng rơi vào tình cảnh bị ngập trong dòng cập nhật mà không thể xử lý chúng một cách đầy đủ. Với mục đích giải quyết vấn đề này, giải pháp được quan tâm là sắp xếp các tin trong dòng cập nhật sao cho hợp lý nhất với mỗi người dùng. Liangjie Hong và cộng sự (2012) nêu bật vấn đề xếp hạng dòng cập nhật (gọi tắt là Xếp hạng dòng).

Bài toán xếp hạng dòng trong mạng xã hội được đặt ra để giải quyết vấn đề cập nhật tin cho mỗi người dùng, đưa ra danh sách các tin trong dòng cập nhật theo một thứ tự (theo "hạng") quan tâm của người dùng, như là một hình thức tư vấn cho người dùng đó. Với bài toán này, việc xếp hạng các tin trong dòng cập nhật cần căn cứ vào lịch sử hành vi của người dùng để tìm ra mối quan hệ giữa cá nhân người dùng đó với đối tượng xếp hạng, thậm chí cả quan hệ với người dùng khác.

Tương tự như các mạng xã hội khác, người dùng trên Twitter cũng đối mặt với lượng lớn các dòng cập nhật liên tục từ những người bạn của mình. Trong phạm vi luận văn, chúng tôi tập trung vào bài toán xếp hạng dòng trên mạng xã hội Twitter, và tiếp tục đề cập tới mô hình hệ thống xếp hạng dòng của mình [1]. Phương pháp phương pháp học tính hạng CRR [2] (Combined Regression and Ranking) được sử dụng.

Mô hình xếp hạng dòng sử dụng thuật toán học tính hạng – thuật toán dựa trên nền tảng học máy, nên việc xây dựng các tập dữ liệu huấn luyện là cần thiết. Chúng tôi đi tìm các yếu tố đặc trưng của tweet. Như đã phát biểu trong [1], yếu tố nội dung của tweet - một yếu tố cơ sở tất yếu cho quá trình học, được tìm ra dựa vào phương pháp phân cụm không giám sát, đó là mô hình chủ đề ẩn [3, 4]. Yếu tố nội dung được biểu diễn dưới

hình thức một tập các phân phối tweet theo chủ đề. Trong mô hình xếp hạng dòng, mô hình chủ đề ẩn LDA được sử dụng. Ngoài yếu tố nội dung, độ ảnh hưởng người dùng được nhận diện là một yếu tố quan trọng. Các cập nhật của người dùng có độ ảnh hưởng lớn thường được nhiều người theo dõi hơn [5, 6]. Dựa trên quan điểm này, chúng tôi nhận thấy các dòng cập nhật từ những người bạn có ảnh hưởng tới người dùng đang xét nên được tư vấn cho người dùng đó. Hay nói cách khác, độ ảnh hưởng người dùng (user influence) nên được tham gia vào quá trình học tính hạng. Do vậy, chúng tôi quyết định cải thiện mô hình tính hạng [1] với sự tham gia của đặc trưng độ ảnh hưởng người dùng. Trong [7], Fredik và cộng sự đã thực hiện tìm các người dùng có độ ảnh hưởng lớn trên mạng xã hội dựa vào khai phá luật kết hợp. Học theo phương pháp này, chúng tôi công thức hóa độ ảnh hưởng của người dùng qua số lượng luật kết hợp tìm được trên tập các tweet. Thuật toán khai phá luật kết hợp được sử dụng là thuật toán Apriori [8].

Khái quát lại, luận văn đề xuất phương pháp cải thiện mô hình tính hạng mà chúng tôi đã đề xuất trong [1] thành mô hình với cốt lõi là phương pháp học tính hạng, xây dựng đặc trưng nội dung dựa trên mô hình LDA, và xây dựng đặc trưng người dùng dựa trên luật kết hợp. Nội dung của luận văn chia thành các chương như sau:

Chương 1: Luận văn trình bày về các dòng cập nhật của mỗi người dùng trên mạng xã hội Twitter và phát biểu bài toán xếp hạng các dòng cập nhật đó. Đồng thời nêu lên hướng giải quyết và ý nghĩa của bài toán này.

Chương 2: Luận văn trình bày về các phương pháp mà mô hình đề xuất sẽ sử dụng: phương pháp học tính hạng, mô hình chủ đề ẩn và luật kết hợp.

Chương 3: Luận văn trình bày mô hình xếp hạng dòng và cách hoạt động của mô hình đó.

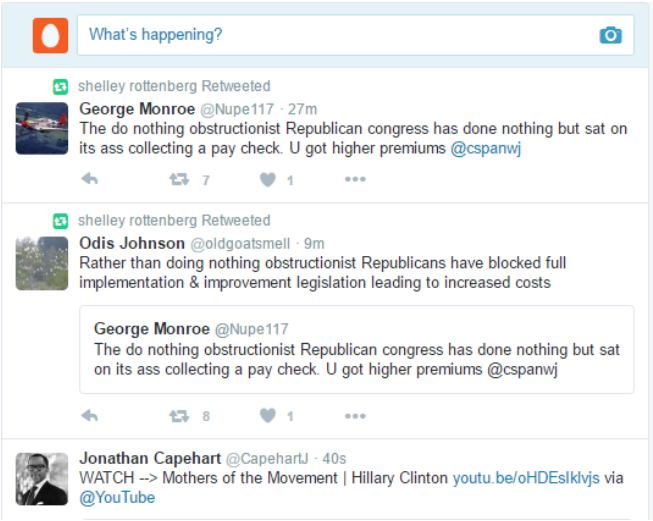
Chương 4: Luận văn trình bày thực nghiệm cho việc áp dụng mô hình xếp hạng trong chương 3 vào việc tính hạng tập các tweet của người dùng trên Twitter.

Chương 1. DÒNG CẬP NHẬT TRÊN MẠNG XÃ HỘI TWITTER VÀ BÀI TOÁN XẾP HẠNG DÒNG

1.1. Mạng xã hội Twitter và dòng cập nhật trên Twitter

Twitter là dịch vụ mạng xã hội ra đời năm 2006, một trang micro-blog được phát triển bởi Twitter Inc, cung cấp một dịch vụ mạng miễn phí cho phép người dùng sử dụng gửi và nhận các tin nhắn (tweet), và đã trở thành một hiện tượng phổ biến toàn cầu. Tính đến tháng 12 năm 2012, số lượng thành viên của Twitter lên tới gần 500 triệu người dùng [9].

Dòng cập nhật trên mạng xã hội Twitter được hiểu là dòng cập nhật của mỗi người dùng. Người dùng A following B, thì A được gọi là follower của B, và B được gọi là followee của A. Khi các followee đăng các thông điệp, các thông điệp này sẽ được hiển thị trên timelines của follower [10]. Khi số lượng followee là lớn thì lượng dòng cập nhật đến trang của follower có thể lên tới hàng trăm tweet. Cheng Li và cộng sự [10] cũng chỉ ra rằng một khi số lượng dòng cập nhật là lớn, các cập nhật mới sẽ hiển thị trên đầu, thay thế các cập nhật cũ. Như vậy bất kì người dùng nào cũng có thể rơi vào tình cảnh bị tràn ngập thông tin và dễ bỏ qua những tin cần thiết với bản thân họ. Giải pháp xếp hạng dòng cập nhật của mỗi người dùng được đưa ra để giải quyết vấn đề này.



Hình 1.1. Minh họa dòng cập nhật trên Twitter

1.2. Bài toán xếp hạng dòng cập nhật

Bài toán xếp hạng dòng cập nhật là bài toán sắp xếp các cập nhật đến trang của mỗi người dùng. Trước khi phát biểu về bài toán này trên mạng xã hội Twitter, chúng tôi đưa ra một số định nghĩa để tường minh hơn về bài toán.

1.2.1. Một số định nghĩa

- Dòng trên mạng xã hội Twitter được hiểu là dòng cập nhật của người dùng. Mỗi người dùng có các thông điệp mới (các cập nhật) đăng bởi các bạn bè trên trang của họ, đó là dòng cập nhật của họ.
- Xếp hạng dòng trên mạng xã hội Twitter cơ bản là xếp hạng các thông điệp mới của mỗi người dùng trên mạng xã hội này.

1.2.2. Bài toán xếp hạng dòng cập nhật

Bài toán xếp hạng dòng trên mạng xã hội Twitter là bài toán sắp xếp các tweet xuất hiện trong mỗi trang người dùng theo mức độ quan tâm của người dùng đó.

Ta có:

- Tập các người dùng trên mạng xã hội Twitter là $U = \{u_i\}, i = 1, N$
- Tập các người dùng mà u_i following là $U_i = \{u_{i'}\}, i' = 1, n$ $(i \neq i')$
- Tập các tweet hiển thị trên trang nhà (home) của u_i là $T_{u_i} = \{t_{u_i j}\}$. Đây là tập hợp các tweet do các người dùng trong tập U_i đăng lên Twitter.

Nhiệm vụ của bài toán là sắp thứ tự các tweet t_k theo mức độ quan tâm của người dùng u_i. Bài toán được phát biểu như sau:

Input: Các tweet mới đưa lên trên trang của người dùng u_i.

Output: Danh sách các tweet đó theo thứ tự giảm dần mức độ quan tâm của người dùng u_i.

1.3. Hướng tiếp cận giải quyết bài toán

Để giải quyết một bài toán xếp hạng các dòng cập nhật hay các tweet mới đến của mỗi người dùng, hoàn toàn có thể áp dụng phương pháp xếp hạng đã được nghiên cứu trước đó dù bài toán này không có câu truy vấn.

Một trong các hướng giải quyết gần đây là kĩ thuật học máy để học hàm xếp hạng tự động như học xếp hạng [11]. Trong [12], Liangjie và cộng sự cũng đề cập tới một mô hình giải bài toán xếp hạng cập nhật trên mạng xã hội LinkedIn, có liên quan tới phương pháp học tính hạng. Trong [1], chúng tôi nghiên cứu và áp dụng phương pháp của Liangjie và cộng sự cùng mô hình chủ đề ẩn được sử dụng để làm giàu đặc trưng dữ liệu vào bài toán trên. Trong luận văn, chúng tôi nâng cao hệ thống xếp hạng của mình bằng cách áp dụng độ ảnh hưởng của user (user influence) vào làm giàu đặc trưng vì độ ảnh hưởng của người dùng được đánh giá là rất hữu ích trong hệ tư vấn… [5, 6]. Do vậy, đây sẽ là một đặc trưng quan trọng góp phần vào nâng cao hệ thống xếp hạng. Đặc trưng này được tìm ra dựa vào luật kết hợp [7].

1.4. Ý nghĩa của bài toán xếp hạng dòng

Kết quả của bài toán xếp hạng dòng là sự tư vấn cho người dùng, giúp họ nhanh chóng hơn trong việc nắm bắt các thông tin mình quan tâm và tiết kiệm thời gian cho bản thân. Mặt khác, sự tư vấn cho người dùng có kết quả tốt sẽ mang lại sự yêu thích của người dùng với mạng xã hội và số lượng người tham gia mạng sẽ tăng lên đáng kể.

1.5. Tóm tắt chương 1

Luận văn đã trình bày tổng quan về mạng xã hội Twitter và nội dung liên quan tới dòng cập nhật. Luận văn cũng đã nêu lên được vấn đề bất lợi cho người dùng khi bị tràn ngập thông tin và phát biểu được bài toán xếp hạng các dòng cập nhật cùng hướng tiếp cận để giải quyết bài toán. Ngoài ra, luận văn cũng đã nêu lên ý nghĩa của bài toán này.

Chương 2. CÁC PHƯƠNG PHÁP HỌC XẾP HẠNG, MÔ HÌNH CHỦ ĐỀ ẨN VÀ LUẬT KẾT HỢP

2.1. Một số nội dung cơ bản về Xếp hạng dòng

2.1.1. Giới thiệu

Xếp hạng dòng chính là một loại Xếp hạng đối tượng (Tweet). Công việc thiết yếu là sắp xếp các đối tượng tweet của mỗi người dùng theo sự

giảm dần mức độ quan tâm của mỗi người dùng đó. Để xếp hạng các đối tượng, ta cần xác định hàm tính giá trị thứ hạng, gọi là *hàm tính hạng*. Mỗi đối tượng gồm có các đặc trưng là những chi tiết của bản thân đối tượng đó. Hàm tính hạng là sự kết hợp của các đặc trưng này.

2.1.2. Học xếp hạng

Học xếp hạng là một loại học máy giám sát hoặc bán giám sát, trong đó mục tiêu là để tự động xây dựng một mô hình xếp hạng từ dữ liệu huấn luyện là tập dữ liệu đã có xếp hạng đúng.

Như đã đề cập trong [1], các thuật toán học xếp hạng đều có hai nhiệm vụ chính: (1) xây dựng hàm tính hạng, (2) tính toán thứ hạng của đối tượng mới. Các nhiệm vụ có đầu vào và đầu ra khác nhau, cụ thể như sau:

- *Xây dựng hàm tính hạng*
  - Đầu vào: Tập các đối tượng có sẵn thứ tự đúng và các đặc trưng
  - Đầu ra: Hàm tính hạng
- *Tính toán thứ hạng đối tượng mới*
  - Đầu vào: Tập đối tượng mới và hàm tính hạng
  - Đầu ra:Thứ hạng của mỗi đối tượng

2.1.3. Các phương pháp học xếp hạng điển hình

2.1.3.1. Phương pháp SVM-rank

Xếp hạng SVM (SVM-rank) [13] là một ứng dụng của máy véc-tơ hỗ trợ (Support vector machine) được sử dụng để giải quyết bài toán xếp hạng bằng việc sử dụng thuật toán học giám sát SVM. SVM-rank được Joachims công bố năm 2002 với mục đích cải thiện hiệu suất của các công cụ tìm kiếm trên Internet. SVM-rank là thuật toán học xếp hạng theo hướng tiếp cận pairwise.

Nhiều phương pháp dựa vào tối ưu SVM như [14]…Trong [2], Sculley đưa ra thuật toán CRR là sự kết hợp xếp hạng dựa trên SVM-rank với hồi quy.

2.1.3.2. Phương pháp CRR

D.Sculley [2] đưa ra đưa ra phương pháp kết hợp cho hiệu quả tốt ở cả hồi quy và xếp hạng. Tư tưởng chính của phương pháp này là xây dựng mô hình tính hạng dựa trên mô hình hồi quy tuyến tính và mô hình tính

hạng pairwise (sử dụng SVM-rank). Thuật toán D.Sculley đưa ra gọi là thuật toán CRR, được trình bày như **Error! Reference source not found.Error! Reference source not found.**.

Cho trước: α, λ, dữ liệu huấn luyện D và số lần lặp t.

$w_o \leftarrow \emptyset$
$for\ i = 1\ to\ t$
    lấy ngẫu nhiên số z từ [0,1]
    $if\ z < \alpha\ then$
        $(x, y, q) \leftarrow RandomExample(D)$
    $else$
        $\big((a, y_a, q), (b, y_b, q)\big) \leftarrow RandomCandidatePair(P)$
        $x \leftarrow (a - b)$
        $y \leftarrow t(y_a - y_b)$
    $end\ if$
    $\eta_i \leftarrow \frac{1}{i\lambda}$
    $w_i \leftarrow StochasticGradientStep\big(w_{i-1}, x, y, \lambda, \eta_i\big)$
$end\ for$
$return\ w_t$

Hình 2.1. Thuật toán CRR [2]

Thuật toán thuần cho việc tối ưu sự kết hợp sẽ liệt kê đầy đủ tập các cặp ứng viên P. Số thành phần thuộc P là bình phương số thành phần thuộc D hay $|P|=|D|^2$ nên khó thực hiện ở tập dữ liệu lớn. Joachims [14] đã đưa ra phương thức cho độ phức tạp O(|D|log|D|). Thuật toán đưa ra phương thức tối ưu sự kết hợp hồi quy và xếp hạng sử dụng phương pháp *Stochastic gradient descent* [2]. Phương pháp này giúp tối thiểu hàm mục tiêu, vấn đề xuất hiện trong học mô hình. Phương thức StochasticGradientStep trả ra kết quả khác nhau với các hàm sai số khác nhau. Chẳng hạn, với square loss, $y \in R$, phương thức này trả ra $\big(1 - \eta_i\lambda\big)w_{i-1} + \eta_i x(y - (w_{i-1}, x))$

Với logistic loss, giả sử $y \in \{0,1\}$, phương thức trả ra

$$\big(1 - \eta_i\lambda\big)w_{i-1} + \eta_i x\left(y - \frac{1}{1 + e^{-(w_{i-1}, x)}}\right)$$

Như vậy, mô hình w được trả ra là mô hình học tính hạng.

2.1.4. Phương pháp đánh giá xếp hạng dòng

Liangije và cộng sự [12] đã phân tích và lựa chọn các thước đo phổ biến dựa trên xếp hạng trong thu hồi thông tin (Information Retrieval). Đó là độ chính xác mức k (Precision@K – P@K) và độ chính xác trung bình (Mean Average Precision – MAP).

❖ **Độ chính xác mức K:** *P@K*

Độ chính xác xếp hạng ở mức K - Precision@K (P @K): độ chính xác của K đối tượng đầu bảng xếp hạng. Xác định số đối tượng đúng ở K vị trí đầu tiên của xếp hạng và gọi là Match@K, và độ chính xác mức K:

$$\mathrm{P@K} = \frac{\mathrm{Match@K}}{\mathrm{K}}$$

❖ **Độ chính xác trung bình: MAP**

Độ chính xác trung bình là giá trị trung bình của các P@K tại các mức K có đối tượng đúng. Gọi I(K) là hàm xác định đối tượng ở vị trí hạng K nếu đúng I(K) =1 và ngược lại I(K) = 0. Độ chính xác trung bình:

$$AP = \frac{\sum_{K=1}^{n} P@K \times I(K)}{\sum_{j=1}^{n} I(j)}$$

Với n là số đối tượng được xét.

MAP là độ chính xác trung bình trên N xếp hạng. (N truy vấn, mỗi truy vấn có một thứ tự xếp hạng kết quả tương ứng). MAP được tính như sau:

$$MAP = \frac{\sum_{i=1}^{N} AP_i}{N}$$

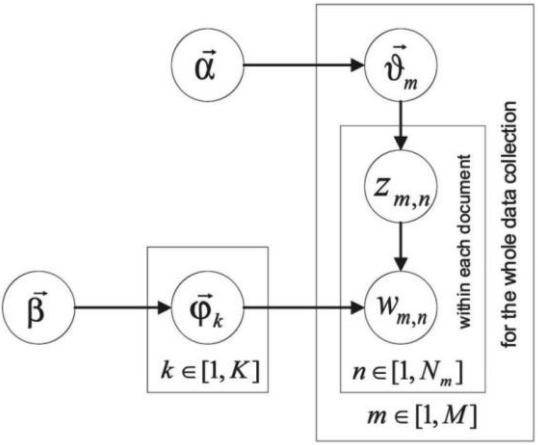
2.2. Mô hình chủ đề ẩn

2.2.1. Giới thiệu

Mô hình chủ đề ẩn [3] là mô hình xác suất phân phối các chủ đề ẩn trên mỗi tài liệu. Chúng được xây dựng dựa trên ý tưởng rằng mỗi tài liệu có một xác suất phân phối vào các chủ đề, và mỗi chủ đề là sự phân phối kết hợp giữa các từ khóa. Hay nói cách khác, ý tưởng cơ bản là dựa trên việc coi tài liệu là sự pha trộn của các chủ đề. Biểu diễn các từ và tài liệu dưới dạng phân phối xác suất có lợi ích rất lớn so với không gian vector thông thường.

2.2.2. Phương pháp mô hình chủ đề ẩn

LDA là một mô hình Bayes phân cấp 3 mức (mức kho ngữ liệu, mức tài liệu và mức từ ngữ). Mỗi tài liệu trong tập hợp được coi là một hỗn hợp xác định trên tập cơ bản các chủ đề. Mỗi chủ đề là một hỗn hợp không xác định trên tập cơ bản các xác suất chủ đề. Về khía cạnh mô hình hóa văn bản, các xác suất chủ đề là một biểu diễn cụ thể, rõ ràng cho một tài liệu. Dưới đây, luận văn sẽ trình bày những nét cơ bản về mô hình sinh trong LDA.



Hình 2.2. Mô hình biểu diễn của LDA [17]

Cho trước tập M tài liệu D = {d_1, d_2…d_M}, trong đó tài liệu thứ m gồm N_m từ, từ w_i được rút ra từ tập các thuật ngữ {t_1, t_2…t_V), V là số các thuật ngữ. Quá trình sinh trong mô hình LDA diễn ra như Hình 2.2

Ước lượng tham số cho mô hình LDA bằng tối ưu hóa một cách trực tiếp và chính xác xác suất của toàn bộ tập dữ liệu là khó có thể thực hiện. Một giải pháp đã được đề ra là sử dụng phương pháp ước lượng xấp xỉ như phương pháp biến phân [3] và lấy mẫu Gibbs [15]. Lấy mẫu Gibbs được xem là một thuật toán nhanh, đơn giản và hiệu quả để huấn luyện LDA.

Trong luận văn, chúng tôi sử dụng phân phối topic của mỗi tài liệu được tìm ra từ LDA để làm đặc trưng nội dung cho việc xây dựng tập huấn luyện cho quá trình học của phương pháp học xếp hạng.

2.3. Luật kết hợp

2.3.1. Giới thiệu

Luật kếp hợp (Association Rule - AR) là lớp các quy tắc quan trọng trong khai phá dữ liệu, được Agarwal giới thiệu năm 1993 [16]. Mục đích của khai phá luật kết hợp là tìm ra các mối quan hệ đồng xảy ra giữa các đối tượng trong khối lượng lớn dữ liệu. Luật kết hợp không chỉ ứng dụng rộng rãi trong phân tích dữ liệu thị trường [8], mà còn được ứng dụng trong tìm những người dùng có độ ảnh hưởng lớn tới các người dùng khác trên mạng xã hội [7].

Các khái niệm cơ bản của luật kết hợp được tóm tắt như dưới đây.

Cho tập các giao dịch (transaction) $T = \{t_1, t_2, \dots, t_n\}$, và tập các đối tượng (item) $I = \{i_1, i_2, \dots, i_m\}$. Mỗi giao dịch t_i là tập các item $t_i \subseteq I$.

Những luật kết hợp này có dạng $X \to Y, với\ X \subseteq I, Y \subseteq I, và\ X \cap Y = \emptyset$

X (hoặc Y) là một nhóm các item và được gọi là itemset. Một itemset gồm k item gọi là k-itemset.

Để đo lường luật kết hợp, độ hỗ trợ (support) và độ tin cậy (confidence) là 2 tham số được sử dụng. $support = \frac{(X \cup Y).count}{n}$ $confidence = \frac{(X \cup Y).count}{X.count}$

Trong đó: n là tổng số giao dịch.

$(X \cup Y).count$ là số giao dịch có X và Y

$X.count$ là số giao dịch có X

Mục tiêu: Với cơ sở dữ liệu giao dịch T, khai phá luật kết hợp là tìm các luật kết hợp trong T thỏa mãn 2 tiêu chí *minimum support (minsup)* và *minimum confidence (minconf)*. Nói cách khác, cần tìm các luật kết hợp AR sao cho $support(AR) \geq minsup$ và $confidence(AR) \geq minconf$.

2.3.2. Thuật toán Apriori

2.3.2.1. Tạo các tập phổ biến

Thuật toán Apriori tìm tất cả frequent itemset bằng cách sử dụng frequent k-itemset để tìm frequent (k+1)-itemset, cho đến khi không có frequent (k+n)-itemset được tìm thấy.

Mã giả tạo các tập phổ biến của thuật toán thể hiện trong **Error! Reference source not found.**Hình 2.3 và Hình 2.4

2.3.2.2. Tạo luật kết hợp

Sử dụng các frequent itemset để tạo tất cả các luật kết hợp. Mã giả tạo các luật kết hợp thể hiện trong Hình 2.5

2.4. Nhận xét và ý tưởng

Ý tưởng cốt lõi của hệ thống xếp hạng là sử dụng phương pháp *học tính hạng* để *xây dựng mô hình tính hạng* cho các dòng cập nhật của mỗi người dùng trên mạng xã hội Twitter. Ở giai đoạn xác định các đặc trưng xây dựng mô hình tính hạng, *mô hình chủ đề ẩn* được sử dụng trong hệ thống để *bổ sung các đặc trưng* liên quan đến nội dung và *khai phá luật kết hợp* giữa các người dùng để bổ sung *đặc trưng độ ảnh hưởng người dùng* cho các tweet

```
Algorithm Apriori(T)
    1  C_1 ← init − pass(T);                          //the first pass over T
2   F_1 ← {f|f ∈ C_1, f.count/n ≥ minsup};   //n is the no of transactions in T
3   for (k = 2; F_{k−1} ≠ ∅; k++) do          //subsequent passes over T
4       C_k ← candidate − gen(F_{k−1});
5       for each transaction t ∈ T do          //scan the data once
6           for each candidate c ∈ C_k do
7           if c is contained in t then
8           c.count++;
9           endfor
10      endfor
11  F_k ← {c ∈ C_k| c.count/n ≥ minsup}
12  endfor
13  return F ← U_k F_k;
```

Hình 2.3. Thuật toán Apriori tạo các frequent itemset [8]

```
Function candidate − gen(F_{k−1})
1  C_1 ← ∅;                                   //initialize the set of candidates
2  forall f_1, f_2 ∈ F_{k−1}                  //find all pairs of frequent itemsets
3      with f_1 = {i_1, ..., i_{k−2}, i_{k−1}}   //that differ only in the last item
4      and f_2 = {i_1, ..., i_{k−2}, i'_{k−1}}
5      and i_{k−1} < i'_{k−1} do              //according to the lexicographic order
6          c ← {i_1, ..., i_{k−2}, i_{k−1}, i'_{k−1}}   //join the two itemsets f_1 and f_2
7          C_k ← C_k ∪ {c};                   //add the new itemset c to the candidates
8          for each (k − 1)-subset s of c do
9              if (s ∉ F_{k−1}) then
10                 delete c from C_k;         //delete c from the candidates
11         endfor
12 endfor
13 return C_k;                                //return the generated candidates
```

Hình 2.4. Hàm candidate-gen [8]

```
ALgorithm genRules(F)                        //F is the set of all frequent itemsets
1 for each frequent k − itemset f_k in F, k ≥ 2 do
2  output every 1 − item consequent rule of f_k with confidence ≥
   minconf and support ← f_k.count/n //n is the total number of transactions in T
3  H_1 ← {consequents of all 1-item consequent rules derived from f_k above};
4  ap − genRules{f_k, H_1};
5  endfor

Procedure ap − genRules{f_1, H_m}      //H_m is the set of m-item consequents
1  if (k > m + 1) AND (H_m ≠ ∅) then
2    H_{m+1} ← candidate − gen(H_m);
3    for each h_{m+1} in H_{m+1} do
4      conf ← f_k.count/(f_k − h_{m+1}).count;
5      if (conf ≥ minconf) then
6          output the rule (f_k − h_{m+1}) → h_{m+1} with confidence = conf and
           support = f_k.count/n;      //n is the total number of transactions in T
       else
8          delete h_{m+1} from H_{m+1};
9    endfor
10   ap − genRules(f_k, H_{m+1});
11 endif
```

Hình 2.5. Thuật toán sinh luật kết hợp [8]

2.5. Tóm tắt chương 2

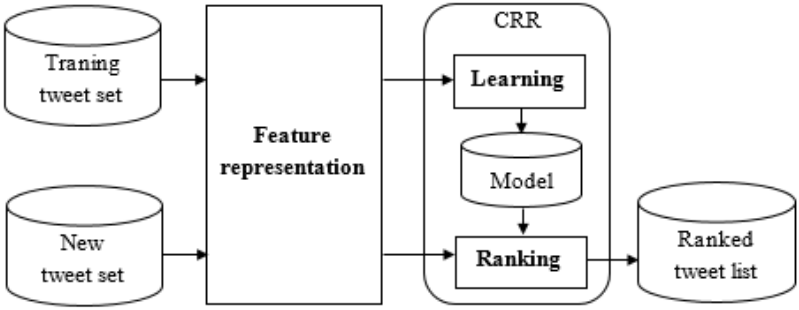
Trong chương 2, luận văn đã trình bày cơ sở nền tảng về học tính hạng, phương pháp xếp hạng CRR, mô hình chủ đề ẩn LDA và thuật toán Apriori khai phá luật kết hợp. Chúng tôi cũng trình bày sơ lược được ý tưởng của mình về mô hình xếp hạng dòng.

Chương 3. MÔ HÌNH XẾP HẠNG DÒNG CẬP NHẬT TRÊN TWITTER

3.1. Phương pháp đề xuất

Như đã được đề cập trong [1], mô hình hệ thống xếp hạng dòng cập nhật bao gồm hai pha chính: học tính hạng (learning) và xếp hạng (ranking)

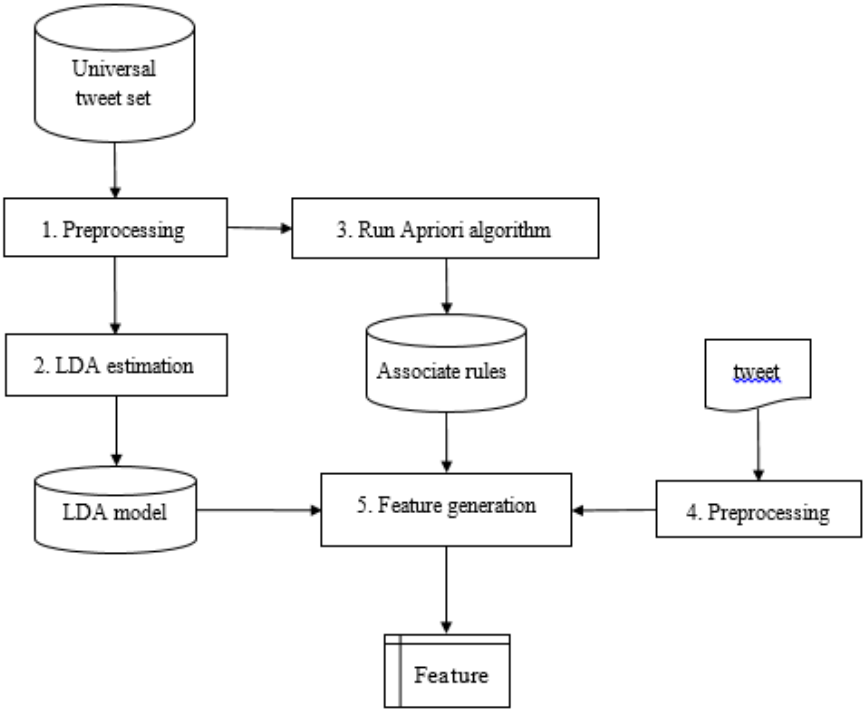
- Learning: Tìm ra ***mô hình tính hạng*** theo sự quan tâm của người dùng dựa vào nội dung tweet và độ ảnh hưởng của người gửi.
- Ranking: Sử dụng các kết quả của pha learning để tính hạng cho các tweet mới. Từ đó, thực hiện ***xếp hạng các tweet mới***



Hình 3.1. Mô hình ranking [1]

Theo [5, 6], độ ảnh hưởng của người dùng được đánh giá là rất hữu ích trong hệ tư vấn, tuyên truyền thông tin…Vì vậy, độ ảnh hưởng của người dùng rất có thể nâng cao hiệu quả cho hệ thống xếp hạng dòng cập nhật [1]. Luận văn tập trung nâng cao mô hình này ở bước biểu diễn đặc trưng (feature representation). Ngoài việc sử dụng các đặc trưng cho tweet như cũ, chúng tôi sử dụng **độ ảnh hưởng của người dùng** vào làm giàu đặc trưng cho hệ thống phân hạng. Thuật toán Apriori [8] được sử dụng

để tìm các luật kết hợp cho tập người dùng liên quan. Hình 3.2 thể hiện bước biểu diễn đặc trưng sau khi đã thay đổi mô hình.



Hình 3.2. Bước biểu diễn đặc trưng (Feature representation)

Bước tiền xử lý dữ liệu (preprocessing) thực hiện nhiệm vụ sau:

- Tách từ (word segmentation): xử lý loại bỏ các dấu cách nếu thừa, tách các từ ghép như won't thành will not…
- Loại bỏ tên người dùng vì nó không bổ sung nghĩa cho nội dung của tweet (bắt đầu bằng kí tự @)
- Loại bỏ từ dừng[1] – những từ không có ý nghĩa.
- Loại bỏ các kí tự đặc biệt, như là kí tự "#" – kí tự được sử dụng để đánh dấu hash tag (cách thức cho phép người dùng đánh dấu các từ khóa mà mình quan tâm để dễ dàng truy cập sau này)

[1] Những từ phổ biến và không có nghĩa. Danh sách các stopword lấy tại http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

- Thực hiện tạo đầu vào cho ước lượng LDA và thuật toán Apriori.

3.2. Đặc trưng và điểm số quan tâm của tweet

3.2.1. Điểm số quan tâm của tweet

Như đã được đề cập trong [1], xét $T_{u_i} = \{t_{u_ij}\}, j = [1,..]$ là tập các dòng cập nhật - tweet của người dùng u_i. Trong đó gồm có tập các tweet mà u_i quan tâm (interesting tweet) (T_{u_i1}) và tập các tweet mà u_i không quan tâm (T_{u_i2}). Gọi Urw_{u_i} là tập người bạn được u_i retweet và Ure_{u_i} là tập người bạn được u_i reply. Với mỗi tweet t_{u_ij}, (j là số thứ tự của tweet trong tập các tweet đang xét của người dùng u_i), thực hiện tính các điểm sau:

$$Score_{rw}(t_{u_ij}) = \begin{cases} 1, & t_{u_ij} \in Urw_{u_j} \\ 0, & trường\ hợp\ còn\ lại \end{cases}$$

$$Score_{re}\left(t_{u_ij}\right) = \begin{cases} 1, & t_{uij} \in Ure_{u_i} \\ 0, & trường\ hợp\ còn\ lại \end{cases}$$

$$Score_{fv}(t_{u_ij}) = \begin{cases} 1, & t_{u_ij} là\ favourite \\ 0, & trường\ hợp\ còn\ lại \end{cases}$$

Điểm của tweet t_{u_ij} là tổng điểm của 3 điểm trên. Nếu điểm của tweet lớn hơn 0 thì đó là interesting tweet.

$$l(t) = \begin{cases} 1,2\ or\ 3 & t \in T_{u_i1} \\ 0, & t \in T_{u_i2} \end{cases}$$

3.2.2. Đặc trưng của tweet

1. Đặc trưng tác giả gửi tweet

Điểm của tác giả đăng tweet được tính theo số following và follower của tác giả đó: $author(u) = \frac{i(u)}{i(u)+o(u)}$

Trong đó, i(u) là số người theo dõi của u (follower) và o(u) là số người u theo dõi (following).

2. Đặc trưng nội dung

Trên cơ sở [1], luận văn sử dụng tập phân phối xác suất của các chủ đề trên mỗi tài liệu là thành phần của tập đặc trưng nội dung.

Giả sử chúng ta xác định được K topic từ tập dữ liệu học. Với mỗi tweet *t*, luận văn tính các xác suất để tài liệu d thuộc vào topic *i* là *pt(i)*, với *i=1,...,k.*

Từ đó xác định được tập đặc trưng nội dung từ mô hình chủ đề ẩn LDA là:

$$T = [pt_1, pt_2 \dots pt_k]$$

3. Đặc trưng Retweet

Đặc trưng Retweet được tính điểm như sau:

$$Rw(t_{uj}) = \begin{cases} 1, & t_{uj} \text{ được } retweet \\ 0, & trường\ hợp\ còn\ lại \end{cases}$$

4. Đặc trưng reply

Tương tự với đặc trưng retweet, đặc trưng reply cũng được tính dựa theo công thức như sau:

$$Re(t_{uj}) = \begin{cases} 1, & t_{uj} \text{ là } tweet\ reply \\ 0, & trường\ hợp\ còn\ lại \end{cases}$$

5. Đặc trưng hash tag

Hash tag là đặc trưng liên quan tới nội dung của tweet. Đặc trưng này được tính như sau: $htag(t) = \begin{cases} 1, & t\ chứa\ hashtag \\ 0, & trường\ hợp\ còn\ lại \end{cases}$

6. Đặc trưng URL

URL cũng là một đặc trưng liên quan tới nội dung của tweet. Đặc trưng này được tính như sau: $url(t) = \begin{cases} 1, & t\ chứa\ URL \\ 0, & trường\ hợp\ còn\ lại \end{cases}$

7. Đặc trưng độ ảnh hưởng người dùng

Xét cơ sở dữ liệu giao dịch là tập các tweet T = {t_1, t_2…t_n}

Tập các item là tập các người dùng U = {u_1, u_2… u_m}.

Bảng 3.1. Minh họa cơ sở giao dịch tìm luật kết hợp giữa các người dùng

Giao dịch (Tweet)	Item (User)
t1	User 1, User 2, User 3
t2	User 1, User 4
t3	User 4, User 5

t4	User 1, User 2, User 4
t5	User 1,User 2,User 6, User 4, User 3
…	…
tm	User 2, User 3, User 6

Để thực hiện tính độ ảnh hưởng của người dùng cho tweet t của user u_k, ta thực hiện các bước sau:

Bước 1: Tìm tập luật kết hợp.

Với user đang xét u_k, thực hiện tìm các luật kết hợp có dạng $\{u_i, \ldots u_j\} \rightarrow \{u_k\}$ thỏa mãn độ support >= minsup và conf >= minconf với minsup, minconf cho trước. Tập luật kết hợp thỏa mãn: A = $\{a_1, a_2 \ldots\}$

Bước 2: Tìm tập user tham gia vào tweet t

Với tweet t, ta tìm tập các user tham gia vào tweet này qua các hoạt động thích, retweet, reply: U(t) = $\{u_1, u_2 \ldots u_t\}$ với $u_t \neq u_k$

Bước 3: Xác định độ ảnh hưởng qua số lượng luật kết hợp phù hợp

Gọi n(t) là số lượng các luật kết hợp trong A thỏa mãn có sự tham gia của các user trong U(t). Độ ảnh hưởng người dùng tới user u_k được tính như sau: $influ(t) = n(t)$

3.3. Tóm tắt chương 3

Trong chương 3, luận văn đã cụ thể hóa mô hình xếp hạng với các công việc cần làm trong mỗi giai đoạn. Ngoài ra, chương này cũng trình bày cách tính điểm cho tweet (nhãn tweet) và các đặc trưng để xây dựng tập dữ liệu huấn luyện.

Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chúng tôi tiến hành thực nghiệm dựa trên mô hình ở chương ba với dữ liệu tweet là các dòng cập nhật của một người dùng trên Twitter. Việc lựa chọn người dùng là hoàn toàn ngẫu nhiên. Bắt đầu với việc xây dựng tập dữ liệu huấn luyện và tập dữ liệu kiểm tra dựa trên công cụ JGibbLDA (cài đặt của mô hình chủ đề ẩn LDA) và các chương trình tự xây dựng. Sau đó, thực hiện quá trình học xếp hạng với chương trình mã nguồn mở chạy thuật toán CRR.

Chúng tôi thực hiện hai thí nghiệm: (1) sử dụng mô hình LDA và sử dụng đặc trưng độ ảnh hưởng người dùng dựa trên luật kết hợp, (2) sử dụng mô hình LDA nhưng không sử dụng đặc trưng độ ảnh hưởng người

dùng dựa trên luật kết hợp. Dựa vào kết quả thực nghiệm, chúng tôi tiến hành đánh giá, so sánh, nhận xét, và rút ra kết luận.

4.1. Môi trường thực nghiệm

4.1.1. Cấu hình phần cứng

Bảng 4.1. Cấu hình máy tính thực nghiệm

Thành phần	Chỉ số
CPU	Intel Core i3-2330M 2.2Ghz
RAM	4GB
HDD	500GB
OS	Ubuntu 11.10 (32bit) Window 8 (hỗ trợ tính dữ liệu)

4.1.2. Công cụ phần mềm

Bảng 4.2. Danh sách các phần mềm sử dụng trong thực nghiệm

STT	Tên phầm mềm	Tác giả	Nguồn
1	Eclipse-SDK-3.7.0		http://www.eclipse.org/dowloads
2	Mã nguồn mở thuật toán CRR: sofia-ml	D.Sculley	http://code.google.com/p/sofia-ml
3	JGibbLDA	Xuan-Hieu Phan và Cam-Tu Nguyen	http://jgibblda.sourceforge.net/
4	MS-Excel trong bộ MS-Office 2013	Microsoft	http://www.microsoft.com
5	Stopword	Nguyễn Thị Tươi	Tự xây dựng với ngôn ngữ java
6	Apriori	Nguyễn Thị Tươi	Tự xây dựng với ngôn ngữ java

4.2. Dữ liệu thực nghiệm

Trong thực nghiệm, chúng tôi sử dụng các dòng tweet của người dùng có tên Jon Bowzer Bauman (@JonBowzerBauman). minh họa về người dùng này trên Twitter.

Dữ liệu thực nghiệm được stream trong thời gian tháng 10 năm 2016, bao gồm hơn 6400 dòng cập nhật đến trang của người dùng này.

Hình 4.1. Minh họa người dùng được sử dụng trong thực nghiệm

4.3. Thực nghiệm

Chúng tôi thực hiện hai thí nghiệm sau với mục đích làm rõ vai trò của việc sử dụng luật kết hợp bổ sung đặc trưng độ ảnh hưởng của người dùng cho tweet trong xếp hạng dòng:

- Thí nghiệm 1 (TN1): Thực hiện xây dựng mô hình tính hạng có sử dụng mô hình LDA và **sử dụng đặc trưng độ ảnh hưởng người dùng** dựa trên luật kết hợp
- Thí nghiệm 2 (TN2): Thực hiện xây dựng mô hình tính hạng có sử dụng mô hình LDA **nhưng không sử dụng đặc trưng độ ảnh hưởng người dùng** dựa trên luật kết hợp

Với thí nghiệm 1, chúng tôi tiến hành các công việc sau:

(1) Thu thập và tiền xử lý dữ liệu.
(2) Xây dựng mô hình chủ để ẩn và đặc trưng nội dung
(3) Tìm tập luật kết hợp và xây dựng đặc trưng độ ảnh hưởng người dùng
(4) Tính các giá trị cho các đặc trưng còn lại của tweet.
(5) Xây dựng dữ liệu huấn luyện và dữ liệu kiểm tra.
(6) Học tính hạng từ dữ liệu huấn luyện
(7) Sử dụng mô hình tính hạng cho dữ liệu kiểm tra và đánh giá.

Với thí nghiệm 2, chúng tôi không thực hiện công việc (3).

Thực hiện xử lý dữ liệu, chúng tôi thu được 5854 tweet. Chia tập tweet làm tập huấn luyện (5254 tweet) và tập kiểm tra (600 tweet).

Sử dụng hai tập dữ liệu này để tiến hành 2 thí nghiệm nêu trên.

4.4. Kết quả và Đánh giá

Sau khi thực nghiệm với hai thí nghiệm (1) và (2), chúng tôi thu được 2 hàm tính hạng. Sử dụng MS-Excel, chúng tôi đánh giá các mô hình của thí nghiệm trên, thể hiện trong các hình sau:

TN1

Dự đoán	Dự đoán	Đánh giá	P@k	P@k * I(k)
1.60704	2	0	0.421849	0
0.926122	1	0	0.421141	0
1.59794	2	0	0.420436	0
0.082801	0	1	0.421405	0.4214047
0.922971	1	0	0.420701	0
1.45476	1	0	0.42	0

P@50	0.64
P@100	0.55
P@600	0.42
Map	0.7634

TN2

Dự đoán	Dự đoán	Đánh giá	P@k	P@k * I(k)
1.87917	2	0	0.415126	0
1.13125	1	0	0.41443	0
1.87133	2	0	0.413735	0
0.017484	0	1	0.414716	0.4147157
1.1273	1	0	0.414023	0
1.74291	2	0	0.413333	0

P@50	0.62
P@100	0.51
P@600	0.41
Map	0.701

Hình 4.2. Đánh giá hai mô hình

Bảng dưới đây thể hiện sự so sánh giữa hai mô hình thu được:

Bảng 4.3. Bảng so sánh hai mô hình thu được

Mô hình	**MAP**
Mô hình 1	76,34%
Mô hình 2	70,1%

Mô hình 1 thu được ở thí nghiệm 1 và mô hình 2 thu được ở thí nghiệm 2. Các mô hình với độ chính xác mức K và độ chính xác trung bình Map được thể hiện trong bảng trên cho thấy mô hình 1 có độ chính xác cao hơn. Vì vậy, việc bổ sung them phần khai phá khoản mục thường xuyên trong luật kết hợp làm tăng chất lượng của các đặc trưng người dùng cho tweet, góp phần tăng độ chính xác của xếp hạng dòng trên mạng xã hội Twitter.

Kết luận và định hướng nghiên cứu tiếp theo

Qua tìm hiểu về luật kết hợp và dựa trên các kiến thức về học xếp hạng, mô hình chủ đề ẩn, luận văn đã thực hiện bổ sung phần khai phá khoản mục thường xuyên trong luật kết hợp nhằm tăng chất lượng của các đặc trưng cho mô hình xếp hạng dòng cập nhật trên mạng xã hội.

Luận văn đạt được các kết quả sau đây:

- Đề nghị mô hình xếp hạng dòng cập nhật cải tiến từ mô hình của chúng tôi [1] với bổ sung độ ảnh hướng người dùng được tính theo thuật toán Apriori.
- Xây dựng phần mềm thực nghiệm và kết quả thực nghiệm đối với hai phương án đạt MAP trên 0.70.

Tuy nhiên, do hạn chế về thời gian nên luận văn vẫn tồn tại những hạn chế như: dữ liệu và các đặc trưng sử dụng cho xếp hạng chưa được phong phú.

Tài liệu tham khảo

[1] Thi-Tuoi Nguyen, Tri-Thanh Nguyen and Quang-Thuy Ha, "Applying Hidden Topics in Ranking Social Update Streams on Twitter," no. RIVF 2013: 180-185 4, 2013.

[2] D.Sculley, "Combined Regression and Ranking," *KDD 2010*, pp. 979-988, 2010.

[3] D. Blei, A., Ng, and M. Jordan, "Latent Dirichlet Allocation," *In Journal of Machine Learning Research*, pp. 993-1022, January/2003.

[4] Thomas Hofmann, "Probabilistic Latent Semantic Analysis," *UAI 1999*, pp. 289-196, 1999.

[5] Chunjing Xiao, Yuxia Xue, Zheng Li and Xucheng L, "Measuring User Influence Based on Multiple Metrics on YouTube. PAAP," 2015, pp. 177-182.

[6] Fabián Riquelme and Pablo Gonzalez Cantergiani, "Measuring user influence on Twitter: A survey. Inf. Process. Manage. 52(5)," 2016, pp. 949-975.

[7] Fredrik Erlandsson, Piotr Bródka and Anton Borg, Finding Influential Users in Social Media Using Association Rule Learning, Entropy 18(5), 2016.

[8] Bing Liu (2007), "Chapter 2. Association Rules and Sequential Patterns," in *Web Data Mining, 2nd Edition: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 2011.

[9] Shea Bennet, "Twitter On Track For 500 Million Total Users By March, 250 Million Active Users By End Of 2012, http://www.mediabistro.com/alltwitter/twitter-active-total-users_b17655," 2012.

[10] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang and Sandeep Pandey, "Click-through Prediction for Advertising in Twitter Timeline," no. KDD 2015: 1959-1968.

[11] Tie-Yan Liu, "Learning to Rank for Information Retrieval," *Foundations and Trends in Information Retrieval 3(3)*, pp. 225-331, 2009.

[12] Liangjie Hong, Ron Bekkerman, Joseph Adler and Brian Davison, "Learning to rank social update streams," *SIGIR'12*, pp. 651-660, 2012.

[13] Joachims Thorsten, "Optimizing Search Engines using Clickthrough Data," *KDD'02*, pp. 133-142, 2002.

[14] Joachims Thorsten, "A support vector method for multivariate performance measures," *ICML 2005*, p. 377–384, 2005.

[15] Gregor Heinrich, "Parameter Estimation for Text Analysis," *Technical report*, University of Leipzig, 2005.

[16] Agarwal and D. Statistical Challenges in Online Adver, *In Tutorial given at ACM KDD-2009 conference*, 2009.

[17] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, Senior Member, IEEE and Quang-Thuy Ha, "A Hidden Topic-Based Framework toward Building Applications with Short Web Documents," vol. 23 NO. 7, July 2011.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC LIÊN QUAN ĐẾN LUẬN VĂN

Thi-Tuoi Nguyen, Tri-Thanh Nguyen and Quang-Thuy Ha, "Applying Hidden Topics in Ranking Social Update Streams on Twitter," no. RIVF 2013: 180-185 4, 2013.