ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BỘ MÔN KHOA HỌC VÀ KĨ THUẬT THÔNG TIN

ĐỒ ÁN MÔN MẠNG XÃ HỘI

ĐỀ TÀI:

TỐI ƯU HỆ THỐNG TÌM KIẾM WEB BẰNG VIỆC KHAI THÁC DỮ LIỆU MẠNG XÃ HỘI

|  |  |
| --- | --- |
| Giảng viên hướng dẫn:  Lớp:  Nhóm:  Sinh viên thực hiện là | Thái Bảo Trân  IS353.I21  6  Ngô Quốc Hải – 15520183  Võ Thành Tâm – 15520761  Lê Quốc Tiến – 15520883 |
|  |  |

Ngày 4 Tháng 06 năm 2018

**Lời mở đầu:**

Đồ án môn học dựa trên bài báo nghiên cứu, tạp chí khoa học tên “Tối ưu hệ thống tìm kiếm web bằng việc khai thác mạng xã hội ” do hai tác giả Nguyễn Thành Luân, Vũ Thanh Nguyên đến từ Trường Đại học Công nghệ Thông tin – ĐHQG TPHCM. Bài báo được đăng tải vào ngày 19-6-2017.

Dựa vào nội dung của bài báo, nhóm chia thành ba phần chính và phân công cho ba người như sau:

1. **Phương pháp Social Personalized Ranking (SoPRa): Lê Quốc Tiến.**
2. **Giải thuật Personalized Social Query Expansion (SoQuES): Ngô Quốc Hải.**
3. **Khai thác dữ liệu mạng xã hội Twitter + Kết quả thực nghiệm và đánh giá: Võ Thành Tâm.**

Ngoài ra còn các công trình khác được đề cập trong bài báo:

Năm 2006, P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita [2], nghiên cứu cách sử dụng chú thích cộng đồng trong Enterprise Search.

Năm 2007, Shenghua Bao, Xiaoyuan Wu, Ben Fei, Guirong Xue, Zhong Su, and Yong Yu trong [3] lần đầu tiên đề cập đến sự quan tâm của người dùng bằng cách xem xét đến các chú thích cộng đồng. Qua đó tác giả đã xây dựng giải thuật SocialSimRank và SocialPageRank. Độ đo này phản ánh một phần nào đó mối quan hệ giữa các từ khóa xuất hiện trong trang web đó.

Năm 2008, Ding Zhou và các cộng sự [4] đã nghiên cứu và sử dụng chú thích cộng đồng trong truy xuất thông tin (Information Retrieval) và đã mang lại kết quả khả quan. Noll and Meinel [5] đề xuất phương pháp tìm kiếm hướng người dùng, phương pháp đã khai thác chú thích của người dùng và các trang web để cải thiện hệ thống tìm kiếm web. Phương pháp tuy đơn giản nhưng mang lại hiểu quả cao. Xu et al. [6] đã xây dựng một framework tận dụng folksonomy để cải thiện kết quả tìm kiếm.

Năm 2010, Vallet et al. [7] đã sử dụng các thông tin liên quan đến người dùng và trang web cho tìm kiếm web theo hướng người dùng.

Năm 2011, Bouadjenek cùng các cộng sự của ông trong [8] đã đề xuất một phương pháp chuẩn hóa câu truy từ người dùng - SoQuES. Phương pháp này khai thác sự tương đồng về ngữ nghĩa giữ các chú thích trong câu truy vấn và mối quan tâm của người dùng thông qua thông tin của họ.

Năm 2013, M.R. Bouadjenek, H. Hacid, M. Bouzeghoub trong [9] đã đề xuất một phương pháp xếp hạng mới gọi là SoPRa, dựa trên personalized social ranking. Phương pháp này nghiên cứu việc sử dụng chú thích cộng đồng kết hợp khai thác mối quan tâm của người dùng để nâng cao hiệu quả tìm kiếm.

Năm 2015, M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan đã nâng cao hiệu quả của việc chuẩn hóa câu truy vấn bằng việc sử dụng từ điển WordNet và đã mang lại hiệu quả nhất định [10].

Bên cạnh đó, năm 2015, Khodaei cùng với các cộng sự [11] đã đề xuất một phương pháp nhằm cải tiến việc tìm kiếm theo hướng người dùng dựa trên cấu trúc và mối liên hệ của các thành phần trong mạng xã hội.

Hầu hết các hướng tiếp cận trên đều được thực hiện trong ngữ cảnh của folksonomies và có chung ý tưởng là độ quan trọng của một trang web (xếp hạng trang) được dựa trên hai yếu tố chính đó là độ tương đồng về nội dung và độ tương đồng về mối quan tâm của người dùng đối với trang web đó.

[1] M.R. Bouadjenek, A. Bennamane, H. Hacid, M. Bouzeghoub, “Social Networks and Information Retrieval, How Are They Converging? A Survey, a Taxonomy and an Analysis of Social Information Retrieval Approaches and Platforms,” Information Systems, Elsevier, 2016.

[2] Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita, “Using annotations in enterprise search,” In WWW ’06: Proceedings of the 15th international conference on World Wide Web, pp. 811–817, New York, NY, USA, 2006.

[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, Z. Su, “Optimizing web search using social annotations,” in: Proceedings of the 16th International Conference on World Wide Web, pp. 501–510, WWW '07, ACM, New York, NY, USA, 2007.

[4] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C. Lee Giles, “Exploring social annotations for information retrieval,” In WWW ’08: Proceeding of the 17th international conference on World Wide Web, pp. 715– 724, New York, NY, USA, 2008.

[5] M.G. Noll, C. Meinel, “Web search personalization via social bookmarking and tagging,” in: ISWC'07 and ASWC'07, 2007.

[6] S. Xu, S. Bao, B. Fei, Z. Su, Y. Yu, “Exploring folksonomy for personalized search,” In: SIGIR, 2008.

[7] D. Vallet, I. Cantador, and J. M. Jose, “Personalizing web search with folksonomy based user and document profiles,” In ECIR, 2010.

[8] M.R. Bouadjenek, H. Hacid, M. Bouzeghoub, Johann Daigremont, “Personalized social query expansion using social bookmarking systems,” in Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, July 25-29, 2011.

[9] M.R. Bouadjenek, H. Hacid, M. Bouzeghoub, “SoPRa: a new social personalized ranking function for improving web search,” in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2013.

[10] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, “Query expansion via WordNet for effective code search,” in Proceedings of IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering, pp. 545-549, 2015.

[11] Khodaei, Ali, Sina Sohangir, and Cyrus Shahabi, “Personalization of Web Search Using Social Signals,” Recommendation and Search in Social Networks, Springer International Publishing, pp. 139-163, 2015.

[12] Twitter (2016). [Online]. Available: <https://twitter.com>

[13] Statista Inc (2016, Oct 1). Twitter Statistics & Facts. [Online]. Available: <https://www.statista.com/topics/737/twitter>.

**1. Giải thuật Personalized Social Query Expansion (SoQuES)**

**1.1. Đặt vấn đề:**

Từ khi, Web 1.0 bị thay thế bởi Web 2.0:

- Lượng người dùng trở nên ngày một nhiều hơn.

- Lượng dữ liệu trở nên khổng lồ.

- Nhu cầu tìm kiếm, truy cập ngày càng trở nên cấp thiết hơn bao giờ hết.

Mặc dù, công cụ tìm kiếm ra đời để đáp ứng cho người dùng, nhưng với số lượng thông tin khổng lồ hiện nay thì việc tìm kiếm thông tin cũng trở nên khó khăn cho người dùng:

- Người dùng không thực sự biết rõ những gì mình đang tìm kiếm cho đến khi tìm thấy nó.

- Nếu biết thì người dùng cũng không biết dùng câu truy vấn nào cho phù hợp với nhu cầu.

Do đó, ta cần có giải pháp hỗ trợ người dùng có được trải nghiệm “lướt web” tốt nhất.

**1.2. Giải pháp:**

Có nhiều cách để hỗ trợ trong việc tìm kiếm của người dùng, và việc chuẩn hóa câu truy vấn bằng việc mở rộng nó (query expansion) là một trong những giải pháp cơ bản nhất. Phương pháp này làm phong phú thêm cho câu truy vấn ban đầu của người dùng bằng các thông tin bổ sung có thể liên quan tới câu truy vấn ban đầu để hệ thống có thể đề xuất các kết quả phù hợp đáp ứng tốt hơn nhu cầu của người sử dụng.

**1.3. Phương pháp mở rộng câu truy vấn (Query Expansion) của Bouadjenek và các đồng nghiệp :**

Năm 2011, Bouadjenek cùng các đồng sự của ông đã đề xuất phương pháp chuẩn hóa câu truy vấn từ người dùng – SoQues. Phương pháp này khai thác sự tương đồng về ngữ nghĩa giữa các chú thích trong câu truy vấn và mối quan tâm của người dùng thông qua thông tin của họ.

Hướng tiếp cận “Social Bookmarking Systems” hay còn gọi là “folksomonies”, thường được mô hình hóa dưới dạng “Tripatite Social Graph” với 3 loại thực thể:

- Users.

- Documents.

- Tags.

Từ đây, ta có thể mở rộng câu truy vấn thông qua việc kết hợp thông tin ngữ nghĩa và xã hội để giảm bớt một số giới hạn của các phương pháp tiếp cận hiện hành, ví dụ như: việc cá nhân hóa.

**a. Vấn đề đặt ra ở đây:**

Một người dùng u đưa ra câu truy vấn Q = {t1 , t2, …, tm}, làm sao để cung cấp cho mỗi ti ϵ Q một danh sách thứ hạng liên quan {ti1 ,ti2 , …, tik}. Từ đó, khoảng cách giữa sự mong đợi của người dùng và kết quả trả về từ hệ thống tìm kiếm được giảm thiểu.

Ta cần một hàm xếp hạng để tính sự liên kết ngữ nghĩa giữa ti và tij (1 ≤ j ≤ k), và sự liên kết xã hội giữa tij và u, biểu thị sự tương đồng giữa người dùng đưa ra câu truy vấn và những người dùng khác trong hệ thống.

**b. Mục tiêu cần đạt được:**

Chuyển đổi câu truy vấn Q thành câu truy vấn Q’ với điều kiên:

- Q nhất định phải bao gồm Q’ (Q ⊂ Q’).

- Các kết quả tìm được của Q có trong các kết quả của Q’.

- Các kết quả thu được với Q’ cho kết quả chính xác và không làm giảm sự hài lòng của khách hàng.

**1.4. Giải thuật SoQuES**:

**a. Phương pháp:**

Nguyên văn:

“To provide ‘social’ and ‘personalized’ expansion of a query term ‘t’ with related term ‘tj’. we propose to take into account two main features:”

i.

- “The Similarity between ‘t’ and ‘tj’“.

→ Sự tương đồng

- “The Semantic strength between ‘t’ and ‘tj’”.

→ Mức độ ngữ nghĩa

ii.

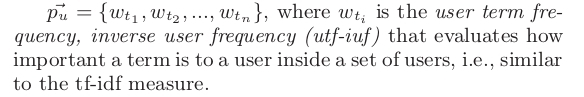
- “The Similarity between ‘tj’ and the user profile expressing the extent to which a tag ‘tj’ is likely to be interesting to the considered user”.

→ Mối tương quan giữa hồ sơ cá nhân khách hàng (thông tin khách hàng về sở thích hay mối quan tâm,…) và thẻ ‘tj’.

**b. Chi tiết:**

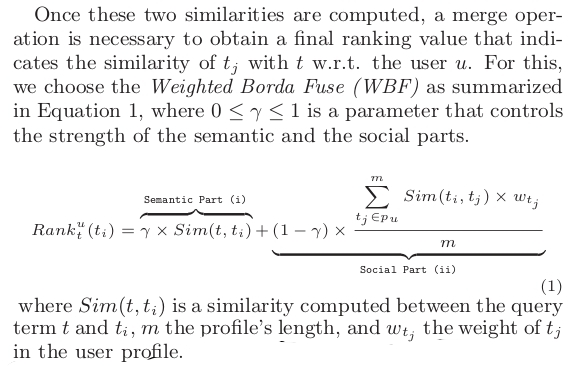
- “User profile → A weighted vector pu = {wt1 , wt2 ,…, wtm}, wti is the utf-iuf (user term frequency, inverse user frequency) that evaluates how important a term is to a user inside a set of users i.e similar to the tf-idf measure”. (Pic1)

→ Thông tin người dùng được biểu diễn bằng Vector trọng số pu với các wti là sự thể hiện cho các mức độ quan trọng của thẻ ti với người dùng trong một nhóm người dùng.



- “Once these two similarities are computed, a merge operation is necessary to obtain a final ranking value that indicates the similarity of ‘tj’ with ‘ti’. For this we use WBF (Weighted Borda Fuse).” (Pic Equation1)

→ Kết quả từ i. và ii. sẽ được hợp nhất để được một giá trị cuối cùng bằng việc sử dụng WBF.



>> Sim(t, ti) : (viết tắt của similarity) là sự tương đồng giữa tag t và ti .

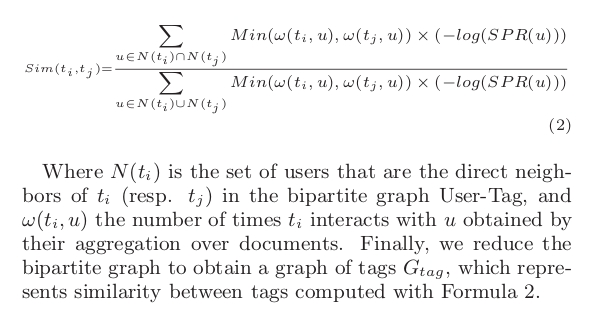
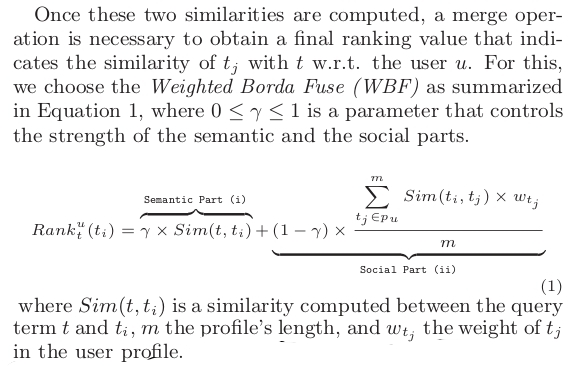
>> m : chiều dài của thông tin người dùng .

>> wt**j** : trọng số cuả tj trong thông tin nguời dùng.

- “In order to consider specific constraints in our approach, we propose a similarity measure that takes into account the credibility of entities in a folksonomy based on their popularity obtained using SPR.”

→ Xem xét các ràng buộc cụ thể bằng phép đo SPR (Social Page Rank) (PDF ThamKhaoSPR.pdf).

- “Use User-Tag graph, to compute the similarity between tags (inspired from Jaccard Similarity) using Formula2 (Pic Formula2)”.

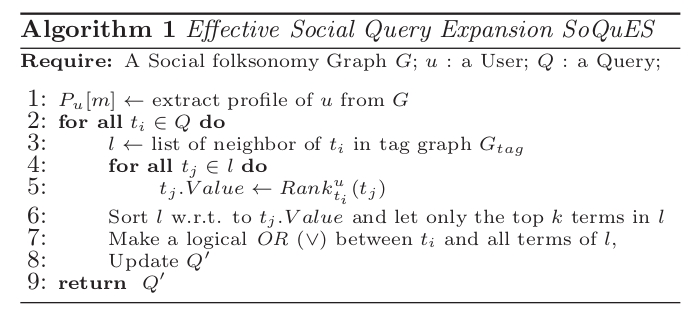
 → Tính toán sự tương đồng giữa các tags.

>> N(ti) : nhóm những users kề với ti (tương ứng với tj) trong đồ thị 2 bên User-Tag.

>> ω(ti, u) : số lần titương tác với u bằng cách tập hợp các tài liệu của chúng.

**c. Algorithms:**

Pic Algorithms



- Explain:

Graph G ; u : a user ; Q : a Query

> Line 1: Getting the user's profile as explained above

-> Gán thông tin người dùng cho Pu[m]

> Line 2: The purpose is to enrich each term ti of Q with related terms

-> Khởi tạo vòng lặp với tất cả tag ti thuộc Q

> line 3: Then, the objective is to get all the neighboring tags tj of ti in the tag graph Gtag

-> gán danh sách các tag tj kề với tag ti trong Gtag bằng danh sách l

> line 4: Compute for each tj with loop

-> Khởi tạo vòng lặp với từng tj

> line 5: the ranking value of tj that indicates its similarity with ti

-> Gán giá trị xếp hạng của tj bằng độ đo sự tương đồng của tj với ti của user u.

> line 6: The neighborhood list has to be sorted according to the value of Rank[tj] and keep only the top tags

-> Sort danh sách l sau đó lọc ra giá trị rank tj lớn nhất và giữ lại.

> line 7: ti and its remaining neighbors must be linked with the OR (∨) logical connector

-> Sử dụng phép: "ti" OR "danh sách l"

> line 8: Updates Q'

-> Cập nhật

> line 9: Return Q'

-> Trả về giá trị Q'

- Example:

Q = t1 ∧ t2 ∧ ... ∧ tm sau khi mở rộng sẽ trở thành:

Q' = (t1 ∨ t**11** ∨ t12 ∨ ... ∨ t1l) ∧ (t**2** ∨ t21 ∨ t22 ∨ ... ∨ t2k) ∧ ... ∧ (tm ∨ tm1 ∨ tm2 ∨ ... ∨ tmr)

**d. Ứng dụng:**

Thông thường, Search Engine sẽ trả về các kết quả dựa trên những gi người dùng nhập.

VD:

-Người dùng nhập “eat” thông thường bộ máy tìm kiếm sẽ trả về những trang web có từ khóa là “eat”. Cải tiến hơn một chút, bộ máy tìm kiếm trả về các giá trị liên quan đến từ khóa “eat” như “eats”, “eaten”, “ate”, …

Ngày nay, các bộ máy tìm kiếm như google khi nhập từ khóa,hệ thống sẽ trả về các kết quả dựa theo xu hướng hiện hành và dựa trên việc xếp hạng các trang web (đó là đối người dùng không đăng nhập.

Giả sử ta lập một tài khoản google và lướt web, hệ thống sẽ tự động lưu lại lịch sử truy cập. Trên các trang mạng xã hội hiện nay như G+, Facebook, Twitter,… mọi tương tác như Thumbup, like, comment,… đều được khai thác bởi hệ thống tìm kiếm và từ đó cá nhân mỗi người sẽ được hình thành một hồ sơ cá nhân.

Về mặt kỹ thuật, ngoại trừ theo dõi người dùng thông qua các hoạt động của các tài khoản đã đằng ký, người ta thường yêu cầu người dùng cho phép sử dụng cookies, cache,…

VD:

Giả sử bạn đăng ký một tài khoản đằng nhập mua hàng trên Lazada, với bất cứ món hàng nào bạn đã xem hay mua, bạn sẽ thấy chúng xuất hiện trên các banner, ads quảng cáo xuất hiện trên trang facebook của bạn.

Một ví dụ khác:

Chắc chắn một điều là người dùng facebook nào cũng đã từng thấy các quảng cáo, lời mời tham gia các group hay “like” các page nào đó. Giả sử bạn “like” một trang chuyên về xe máy, những gợi ý tiếp theo của bạn sẽ hiện thị những trang liên quan đến xe máy.

Điều này không hề ngẫu nhiên, mà đều là mục đích của các doanh nghiệp.

Nhìn chung, việc sử dụng thông tin, hồ sơ người dùng gặp khá nhiều tranh cãi. Nhìn về mặt tốt, điều này giúp cho người dùng có thể tìm được những gì mình muốn mà không cần biết chính xác nó, chỉ bằng sử dụng các từ khóa liên quan. Thông tin tìm kiếm không hiển thị một cách bừa bãi mà tập trung dựa trên mối quan tâm của khách hàng. Tất cả những điều trên khiến cho trải nghiệm lướt web của người dùng thật sự tốt. Bên cạnh đó, việc theo dõi hoạt động của người dùng được cho là xâm phạm quyền riêng tư của họ. Không những thế, người dùng còn có nguy cơ để lộ thông tin cá nhân khi chấp nhận sử dụng các dịch vụ, điều khoản từ các công ty lớn, mặc dù những công ty đó vẫn thường hay “tuyên bố đảm bảo” cho người dùng.

**2. Khai thác dữ liệu dữ liệu mạng xã hội Twitter**

**Giới thiệu**

Sự bùng nổ ngày càng mạnh mẽ của mạng xã hội mở ra nhiều cơ hội cho các tổ chức, cá nhân thu thập, tìm kiếm thông tin cũng như xử lý chúng và nhiều bài toán được đặt ra để khai thác nguồn thông tin dồi dào từ mạng xã hội.

Khai thác dữ liệu là một trong các bài toán khai thác thông tin nằm trong ứng dụng của xử lý ngôn ngữ tự nhiên, giúp thu thập được những thông tin mong muốn về quan điểm của dữ liệu. Bài toán phân tích các đánh giá cho một chủ đề nhất định, hoặc sự kiện, sản phẩm để tự động phân loại đánh giá theo hướng tích cực, tiêu cực của quan điểm. Bài toán còn bao gồm nội dung tổng hợp quan điểm từ các tài liệu quan điểm thu được

Với sự phát triển nhanh chóng và mạnh mẽ, mạng xã hội Twitter đã trở thành một nguồn cung cấp nhiều thông tin quan điểm cho những người nghiên cứu về lĩnh vực này. Đã có nhiều công trình nghiên cứu liên quan đến bài toán khai phá quan điểm theo các phương pháp khác nhau từ nhiều nguồn dữ liệu khác nhau, nhưng Twitter là một nguồn dữ liệu phổ biến.

Đánh giá về quan điểm của bài viết trong Twitter vào một trong hai lớp Tích cực hoặc Tiêu cực mang lại nhiều ý nghĩa cho nhiều lĩnh vực như kinh tế, quảng cáo v.v... Một trong những cách tiếp cận là sử dụng học máy thống kê. Trong cách tiếp cận này, biểu diễn đặc trưng đóng vai trò quan trọng, ảnh hưởng trực tiếp tới độ chính xác của bộ phân lớp. Thông qua tìm hiểu, phân tích các phương pháp khai phá quan điểm trên tập dữ liệu Twitter, chúng tôi tập trung tới ba phương pháp biểu diễn đặc trưng: N-gram[7], Độ đặc trưng dựa trên tâm (Center-base similarity – CBS) [6] và Logcount ratio[9] để thực hiện đề tài chuyên mục: “Khai thác dữ liệu mạng xã hội Twitter”.

Nội dung gồm :

**2.1 Giới thiệu về khai thác dữ liệu mạng xã hội**

**2.2 Tổng quan về khai thác dữ liệu mạng xã hội Twitter**

**2.3 Một số phương pháp khai thác dữ liệu**

**2.4 Phân loại ý kiến khi khai khác trên Twitter**

**2.5 Kết quả dữ liệu thực nghiệm và đánh giá**

**2.1 Giới thiệu về khác thác dữ liệu mạng xã hội**

Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính..

Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.

Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến.

Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

**2.1.1 Một vài ví dụ trong khai thác dữ liệu**

Bài toán xác định quan điểm

Bài toán phân loại quan điểm

Bài toán khai phá quan điểm theo đặc trưng, khía cạnh

Bài toán tóm tắt quan điểm

**2.1.2 Một số khó khăn trong quá trình khai thác dữ liệu**

Phong cách viết của mỗi người là khác nhau

Sự phụ thuộc vào bối cảnh và thời gian

Tồn tại nhiều quan điểm mâu thuẫn trong cùng một tài liệu

**2.2 Tổng quan về khai thác dữ liệu mạng xã hội Twitter**

**2.2.1 Mạng xã hội Twitter là gì?**

Twitter là một dịch vụ mạng xã hội trực tuyến miễn phí cho phép người sử dụng đọc, nhắn và cập nhật các mẩu tin nhỏ gọi là tweets, một dạng tiểu blog.

Những mẩu tweet được giới hạn tối đa 280 ký tự được lan truyền nhanh chóng trong phạm vi nhóm bạn của người nhắn hoặc có thể được trưng rộng rãi cho mọi người.

**2.2.2 Tại sao phải chọn Twitter?**

Mạng xã hội Twitter ngày càng lớn mạnh, cùng với đó, việc thể hiện quan điểm trên mạng xã hội này cũng càng phổ biến. Việc khai thác dữ liệu trên mạng xã hội này có nhiều ý nghĩa, giúp ta thấy quan điểm của một người về một vấn đề nào đó.

**2.2.3 khó khăn và thách thức khi khai thác dữ liệu**

Việc phân loại quan điểm với dữ liệu Twitter gặp khá nhiều khó khăn và thách thức.

Các tweet có đặc điểm riêng khác so với các dạng văn bản truyền thống hay dữ liệu trên các forum, mạng xã hội khác như Facebook.

Số ký tự tối đa nhỏ, từ viết tắt, từ lóng, từ sai chính tả, hơn nữa, chất lượng và độ tin cậy thấp.

Nhiều đặc điểm của dữ 4 liệu Twitter làm giảm hiệu quả khai phá quan điểm dựa trên những kỹ thuật xử lý ngôn ngữ tự nhiên truyền thống.

**2.3 Một số phương pháp khai thác dữ liệu**

Các phương pháp ***cây quyết định***

* + - **Cây hồi quy** (*Regression tree*) ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)
    - **Cây phân loại** (*Classification tree*), nếu *y* là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

Phương pháp ***K-Nearest Neighbor***

* + - K-Nearest Neighbor được mô tả như sau:
    - Xác định giá trị tham số K (số láng giềng gần nhất)
    - Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng các Euclidean)
    - Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point
    - Lấy tất cả các lớp của K láng giềng gần nhất đã xác định
    - Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point

Thuật toán ***SVM (Support Vector Machine)***

là một khái niệm trong [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) và [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) cho một tập hợp các phương pháp [học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t) liên quan đến nhau để [phân loại](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_(h%E1%BB%8Dc_m%C3%A1y)&action=edit&redlink=1) và [phân tích hồi quy](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy).

SVM gồm 3 dạng:

1. SVM tuyến tính
2. SVM dạng ban đầu
3. SVM dạng đối ngẫu

Ngoài ra còn 1 số phương pháp khác…

Một số nghiên cứu liên quan:

* + - Đã có rất nhiều công trình nghiên cứu của các tác giả liên quan đến việc khai phá quan điểm từ nhiều nguồn dữ liệu với các phương pháp khác nhau ví dụ như :
      * Agarwal, Apoorv and Xie, Boyi and Vovsha, Ilia and Rambow, Owen and Passonneau, Rebecca (2011), Sentiment Analysis of Twitter. Columbia University. New York, NY 10027 USA.
      * Davidov, Dmitry and Tsur, Oren and Rappoport, Ari (2010). Enhanced sentiment learning using Twitter hashtags and smileys, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 241-249.

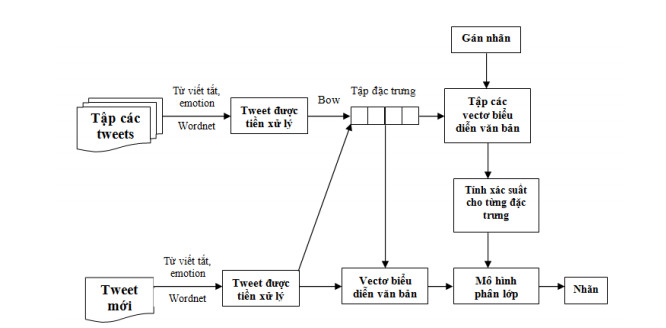
**2.4 Phân loại ý kiến khi khai khác trên Twitter**

Phân loại ý kiến trên Twitter là phân loại cho từng bình luận theo hướng quan điểm tích cực hay tiêu cực dựa trên nội dung bình luận. Trong bài viết này, chúng tôi đề xuất sử dụng mô hình túi từ và giải thuật máy học Multinomial Naïve Bayes để phân loại ý kiến.

Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, chúng tôi tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa, biểu diễn văn bản theo mô hình túi từ.

Giai đoạn tiền xử lý cho ra tập dữ liệu có số chiều lớn, nhưng trong đó đa số (khoảng 99%) các giá trị bằng 0. Để tiết kiệm bộ nhớ, chiến lược lưu trữ chỉ lưu những giá trị khác 0 (theo định dạng LibSVM). Cách lưu trữ này dẫn đến yêu cầu cài đặt lại giải thuật máy học Multinomial Naïve Bayes để có thể xử lý định dạng mới của tập dữ liệu.

Kết quả thực nghiệm trên các tập dữ liệu cho thấy bản cài đặt mới của giải thuật Multinomial Naïve Bayes (MNB) phân lớp hiệu quả, đơn giản và chính xác.



Hình: Sơ đồ phân lớp ý kiến với giải thuật MNB

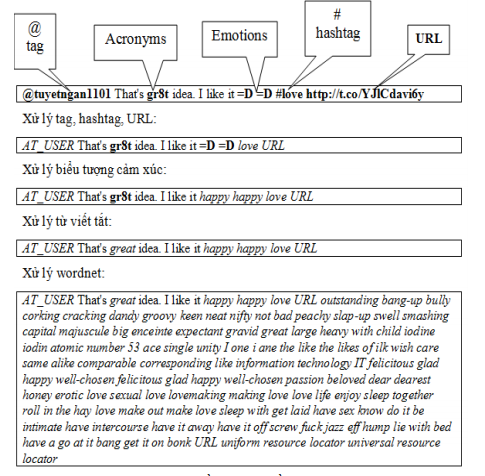
**2.4.1 Phương pháp tiền xử lý dữ liệu**

Dữ liệu của mạng Twitter rất phức tạp, phi cấu trúc, nhiễu. Để có thể phân lớp bằng giải thuật máy học, trước hết cần phải thực hiện các thao tác tiền xử lý.

Các tên tài khoản Twitter của người dùng được chuyển về dạng USER\_AT.

Địa chỉ trang web được chuyển về dạng URL.

Chúng tôi còn xử lý thêm các ký tự trùng lắp gần nhau, biểu tượng cảm xúc, từ viết tắt, tiếng lóng, mạng ngữ nghĩa.

 Hính: ví dụ về các bước tiền xử lý

**2.4.2 Biểu diễn dữ liệu**

Mô hình Bag of Words (BoW) là một mô hình được sử dụng phổ biến trong lĩnh vực phân loại văn bản. Mô hình này thường sử dụng để xử lý ngôn ngữ tự nhiên, được dùng để biểu diễn tài liệu, xem tài liệu là một tập hợp các từ (words) mà không quan tâm đến thứ tự cũng như cấu trúc cú pháp của chúng.

Giả sử dữ liệu có 15.000 tweets với 20.000 đặc trưng (từ vựng), thông thường mỗi tweet sẽ được lưu trữ như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chỉ mục | 1 | 2 | 3 | ... | 20.000 |
| Tần số  xuất hiện | 0 | 1 | 0 | … | 0 |

Đối với nghiên cứu này, chúng tôi đề xuất cách lưu trữ tiết kiệm bộ nhớ, tương tự như LibSVM [Chang & Lin, 2011], chỉ lưu những từ có tần số xuất hiện lớn hơn 0. Cách lưu trữ như sau:

**<label><index-1>:<value-1><index-2>:<value-2>…**

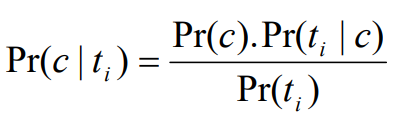
Trong đó:

* + - **<Lable>**là lớp ban đầu của tweet, 1 là tích cực, 0 là tiêu cực.
    - <**index-i>** chỉ mục của từ thứ i.
    - < **value-i**> tần số xuất hiện của từ i.

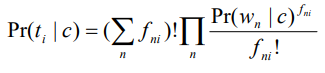
**2.4.3 phân loại ý kiến bằng giải thuật máy học**

Multinomial Naïve Bayes (MNB) là một mô hình đơn giản nhưng hoạt động rất tốt trong việc phân loại văn bản. [Lewis & Gale, 1994] đã đề xuất kết hợp mô hình túi từ và NB tạo ra giải thuật Multinomial Naïve Bayes. Cụ thể trong bài toán của chúng ta như sau:

Gọi C là tập hợp các lớp của văn bản (C có 2 phần tử +1 và -1). Gọi ti là một văn bản mới đến. Ta chọn xác suất để ti thuộc vào lớp ci lớn nhất. Xác suất này được tính bởi công thức:



* + Với c ∈ C
  + Pr(c) được tính bằng tổng số văn bản của lớp c chia cho tổng số văn bản của tất cả các lớp.
  + Khi tìm giá trị lớn nhất của Pr(c|ti) ta có thể bỏ qua tính Pr(ti) do không đổi khi so sánh.
  + Xác suất Pr(ti|c) được tính bằng công thức:



* fni là tần suất từ thứ n trong ti
* Pr(wn|c) là xác suất của từ thứ n khi cho trước lớp c.

**2.5 Kết quả dữ liệu thực nghiệm và đánh giá**

**2.5.1 Dữ liệu**

Để chuẩn bị dữ liệu cho việc thực nghiệm, chúng tôi xây dựng một module để tiến hành thu thập dữ liệu từ Twitter thông qua Twitter’s Streaming API.

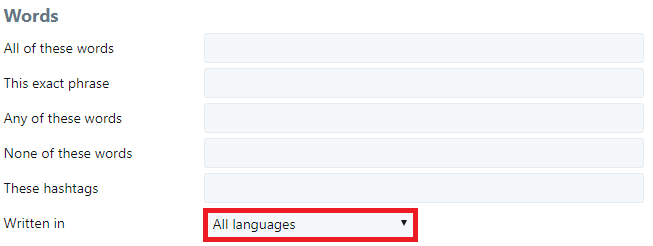
Sau khi thu thập dữ liệu từ Twitter, chúng tôi tiến hành chuẩn hóa dữ liệu bằng 4 thao tác sau trước khi sử dụng cho thực nghiệm:

Link thực hiện 3 thao tác (1)(2)(3): [**https://twitter.com/search-advanced**](https://twitter.com/search-advanced)

Link tham khảo Apache Tika toolkit:

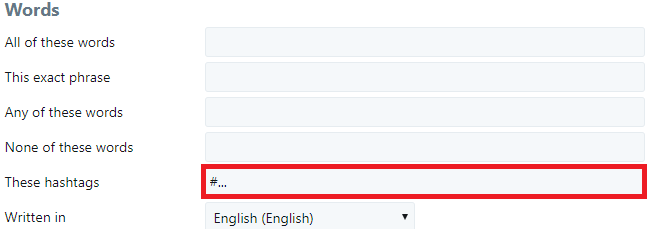
<https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html>

1. lọc bỏ những tweets mà không thuộc ngôn ngữ tiếng Anh và không có chứa URL

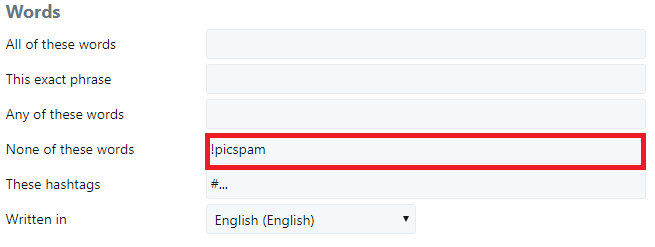
****

Ở phần All languages ta chọn English.

1. trích xuất những chú thích (hashtag) và địa chỉ web (URL) từ tweets



1. loại bỏ những chú thích vô nghĩa như “!picspam”, “atthissummer” dựa trên từ điển WordNet

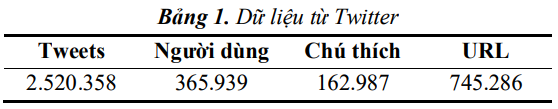


1. loại bỏ những địa chỉ trang web mà không thuộc ngôn ngữ tiếng Anh thông qua Apache Tika toolkit.

Bảng Tóm tắt Phương pháp Apache Tika toolkit.

|  |  |
| --- | --- |
| Công cụ sửa đổi và loại | Phương thức và mô tả |
| static void | [**addProfile**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#addProfile-java.lang.String-org.apache.tika.language.LanguageProfile-)([**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) language, [**LanguageProfile**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageProfile.html) profile)  Thêm 1 hồ sơ ngôn ngữ duy nhất |
| static void | [**clearProfiles**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#clearProfiles--)()  Xóa hồ sơ của ngôn ngữ hiện tại |
| static [**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) | [**getErrors**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#getErrors--)()  Trả về chuổi thông báo liên quan đến khởi tạo ngôn ngữ |
| [**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) | [**getLanguage**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#getLanguage--)()  Nhận ngôn ngữ được xác định |
| … | Chi tiết thao khảo thêm ở Link sau: <https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html> |

Bảng 1 dưới đây mô tả số liệu cụ thể của dữ liệu thu thập sau khi đã được chuẩn hóa:



**2.5.2 Phương pháp đánh giá**

**Sơ lược về MAP**

|  |  |
| --- | --- |
| MAP là gì? | MAP chỉ là mức trung bình của AP, hoặc độ chính xác trung bình, cho tất cả người dùng.  Ví dụ: Nếu chúng ta có 1000 người dùng, chúng tôi tính tổng số AP cho mỗi người dùng và chia tổng số cho 1000. Đây là MAP. |
| AP là gì? | Chúng tôi có thể đề xuất tối đa x mục cho mỗi người dùng  AP trả tiền để gửi tất cả các khuyến nghị x, bởi vì chúng tôi không bị phạt vì những dự đoán xấu  Các vấn đề trật tự, do đó, tốt hơn nên gửi thêm một số đề xuất nhất định, tiếp theo là các đề xuất chúng tôi ít chắc chắn hơn  Vì vậy, về cơ bản chúng tôi chọn x ứng cử viên tốt nhất (theo thứ tự) và đó là nó AP |
| Tài liệu tham khảo thêm | Nếu bạn cần hiểu thêm về Map có thể truy cấp vào đường link sau: **http://fastml.com/what-you-wanted-to-know-about-mean-average-precision**/ |

**Đánh giá thực nghiệm**

Ví dụ: Nếu chúng ta có 1000 người dùng, chúng tôi tính tổng số AP cho mỗi người dùng và chia tổng số cho 1000. Đây là MAP.

Ý tưởng của việc đánh giá được thực hiện như sau: Cho câu truy vấn q = {t} được nhập bởi người dùng u với từ khóa truy vấn t, kết quả tìm kiếm liên quan là những trang web được người dùng u với chú thích bằng từ khóa t.

Chúng tôi sử dụng độ đo Mean Average Precision để tiến hành thực nghiệm cho công cụ tìm kiếm. Cụ thể hơn, chúng tôi tính toán MAP cho mỗi người dùng và sau đó tính toán giá trị trung bình của tất cả các giá trị MAP (Mean MAP).



Trong đó, MAPi là giá trị độ đo MAP của người dung thứ i, Nu số lượng người dùng trong thực nghiệm.

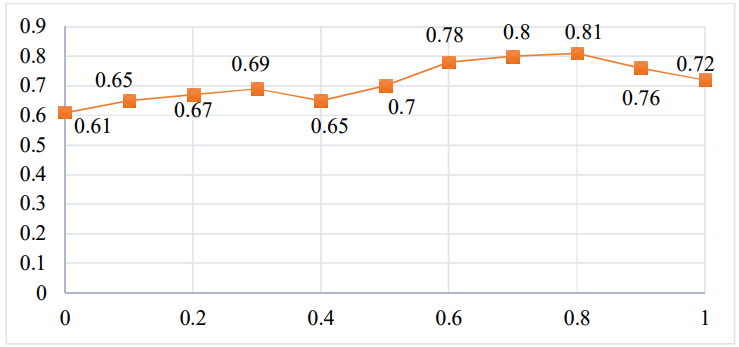
Chúng tôi chọn ngẫu nhiên 1000 cặp (u,t). Trong mỗi cặp, user u gửi câu truy vẫn q = {t} đến hệ thống tìm kiếm. Tiếp theo, hệ thống sẽ tìm kiếm và xếp hạng kết quả tìm kiếm phù hợp với câu truy vấn.

Cuối cùng, chúng tôi tính độ đo MAP cho 1000 câu truy vấn. Đánh giá một liên kết trả về có tương quan hay không tương quan tùy vào ý kiến chủ quan của người dùng, trong thực nghiệm chọn 10 kết quả đầu tiên để đánh giá.

**2.5.3 Kết quả đánh giá**

Trong công thức xếp hạng trang web (2), chúng tôi thực nghiệm với 0 ≤ α ≤ 1 và β = 0.5.

Công thức (4), chúng tôi chọn γ = 0.5 và kích thước câu truy vấn cho mỗi từ khóa là 5. Hình 1 dưới đây mô tả kết quả thực nghiệm của hệ thống tìm kiếm.



Giá trị α

**Hình 1.** Giá trị Mean Average Precision theo α

Theo như hình 1, hiệu quả của hệ thống tìm kiếm đạt tốt nhất khi α ∈ [0.6, 0.8].