TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
TẠP CHÍ KHOA HỌC
KHOA HỌC TỰ NHIÊN VÀ CÔNG NGHỆ
ISSN: 1859-3100 Tập 14, Số 6 (2017): 138-145
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
NATURAL SCIENCES AND TECHNOLOGY
Vol. 14, No. 6 (2017): 138-145
Email: tapchikhoahoc@hcmue.edu.vn; Website: http://tckh.hcmue.edu.vn

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

Nguyễn Thành Luân*, Vũ Thanh Nguyên
Trường Đại học Công nghệ Thông tin - ĐHQG TPHCM
Ngày Tòa soạn nhận được bài: 31-12-2016; ngày phản biện đánh giá: 19-01-2017; ngày chấp nhận đăng: 19-6-2017

TÓM TẮT

Với sự bùng nổ thông tin như hiện nay, thì vấn đề tìm kiếm thông tin cho người dùng vẫn đang còn nhiều thách thức. Chính vì vậy, mục tiêu của nghiên cứu này là (1) khai thác chú thích cộng đồng từ mạng xã hội Twitter, (2) chuẩn hóa câu truy vấn theo hướng người dùng, (3) kết hợp sử dụng giải thuật SoPRa để xếp hạng kết quả tìm kiếm, (4) xây dựng hệ thống tìm kiếm hỗ trợ người dùng tìm kiếm một cách nhanh chóng và hiệu quả.

Từ khóa: chú thích xã hội, mạng xã hội, tìm kiếm thông tin, tối ưu truy vấn, xếp hạng trang web.

ABSTRACT

Improving Web Search By Exploiting Social Data

With the booming of information nowadays, the issue of searching for information for users is facing many challenges. Therefore, the study aims at: (1) exploiting social annotation from Twitter, (2) standardizing query following a user-orientated approach, (3) utilizing SoPRa to perform ranking of search results, (4) developing a search system to facilitate users to search information quickly and effectively.

Keywords: social annotation, web ranking, query optimization, information search.

1. Giới thiệu

Hiện nay, Internet đang phát triển một cách mạnh mẽ, đi sâu vào mọi lĩnh vực của cuộc sống và đã trở thành một kênh thông tin quan trọng trong cuộc sống của con người. Các website phát triển ngày càng nhiều và ngày càng đa dạng về cấu trúc lẫn nội dung trang web. Vì vậy, không có gì ngạc nhiên khi lượng thông tin quá tải, hỗn độn, rối rắm thường làm sai lệch các thông tin mà người dùng muốn tìm kiếm cũng như khi duyệt web. Chính vì lẽ đó mà các hệ thống tìm kiếm (Search Engine) được xây dựng như là một công cụ để giúp người dùng tìm và chọn được các thông tin phù hợp với mình.

Theo một nghiên cứu mới nhất từ [1], hiện có 3 hướng cải tiến chính đó là: (i) chuẩn hóa câu truy vấn, bao gồm việc thêm hoặc bớt các từ khóa cho câu truy vấn, (ii), sắp xếp lại kết quả tìm kiếm dựa trên ngữ cảnh hoặc thông tin của người dùng, (iii) cải tiến mô hình tìm kiếm thông tin.

* Email: *thanhluan.uit@gmail.com*

Với sự phát triển của công nghệ Web 2.0, nhiều hệ thông web hỗ trợ người dùng đánh dấu, chia sẻ cũng như bình luận các tài nguyên mà họ quan tâm. Đặc biệt, các hệ thống này cho phép người sử dụng web tổ chức và chia sẻ trực tuyến các trang web mà họ quan tâm bằng cách sử dụng các chú thích cộng đồng. Các chú thích này thường là những tóm lược của các trang web tương ứng. Vậy làm cách nào để có thể tận dụng tốt lợi ích của các chú thích cộng đồng này vào công cụ tìm kiếm. Trong nghiên cứu này, chúng tôi sẽ kết hợp 2 hướng cải tiến đó là chuẩn hóa câu truy vấn và xếp hạng lại kết quả tìm kiếm theo hướng người dùng dựa trên chú thích cộng đồng, để từ đó xây dựng một hệ thống tìm kiếm hiệu quả.

2. Các công trình liên quan

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.

3. Phương pháp Social Personalized Ranking (SoPRa)

Trong phần này, chúng tôi sẽ trình bày chi tiết về phương pháp SoPRa – một phương pháp xếp hạng trang web theo hướng người dùng. Cách tiếp cận của phương pháp là khai thác chú thích cộng đồng trong ngữ cảnh folksonomies.

Theo như Bouadjenek cùng các cộng sự [9], SoPRa xếp hạng trang web dựa trên 2 yếu tố chính đó là: (i) độ tương đồng giữa nội dung trang web với câu truy vấn, (ii) mức độ quan tâm của người dùng đối với các trang web.

Ở yếu tố đầu tiên, các tác giả cho rằng độ tương đồng giữa một trang web với một câu truy vấn dựa trên độ tương đồng về nội dung văn bản (textual matching score) và độ tương đồng về các yếu tố xã hội (social matching score). Trong đó, textual matching score thể hiện sự tương đồng giữa nội dung trang web với câu truy vấn. Còn social matching score thể hiện sự tương đồng giữa "social representation" với câu truy vấn. Với social representation được thể hiện thông qua các chú thích được dùng để đánh dấu trên trang web. Cuối cùng, độ đo của nhân tố đầu tiên được tính bằng cách kết hợp chúng bằng một hàm tuyến tính như sau:

$$Score(q, d) = \beta \times Cos(\overrightarrow{q}, \overrightarrow{s_d}) + (1 - \beta) \times Sim(\overrightarrow{q}, \overrightarrow{d}) \quad (1)$$

Trong đó, hệ số β chúng tôi chọn 0.5, $\overrightarrow{s_d}$ là vectơ đại diện cho social representation của trang web, $Sim(\overrightarrow{q}, \overrightarrow{d})$ biểu thị độ tương đồng về nội dung giữa trang web d với câu truy vấn q.

Ở yếu tố thứ 2, độ đo về mối quan tâm của người dùng (social interest score) đối với các trang web được tính bằng độ tương đồng về thông tin của người dùng với các chú thích của trang web (social representation of a document). Tiếp đến, chúng ta cộng độ đo về mối quan tâm của người dùng này với độ đo đã được tính ở công thức (1). Cuối cùng, công thức tính độ đo của một trang web d phù hợp với câu truy vấn q, được tìm kiếm bởi người dùng u thể hiện như sau:

$$Rank(d, q, u) = \alpha \times Cos(\overrightarrow{p_u}, \overrightarrow{s_d}) + (1 - \alpha) \times Score(q, d) \quad (2)$$

Tóm lại, phương pháp SoPRa xếp hạng trang web dựa trên: Độ tương đồng về nội dung văn bản của trang web với câu truy vấn; độ tương đồng về mặt social của trang web với câu truy vấn; và mức độ quan tâm của người dùng đối với trang web.

Bên cạnh đó thì thông tin người dung và "social representations" của trang web được tính toán dựa trên các chú thích xã hội mà liên kết với nó và được mô hình trong không

gian vectơ (Vector Space Model). Nếu chúng ta xem các trang web hoặc người dùng như những tài liệu và những chú thích như các từ, thì các thiết lập ở trên là đúng cho VSM. Một trong những điểm quan trọng trong VSM là trọng số của các từ. Và trong nghiên cứu này, trọng số của các chú thích xã hội được tính bằng phương pháp tf-idf (term frequency–inverse document frequency) như sau:

$$w = tf \times log\frac{N}{n_i} \quad (3)$$

Trong đó, tf là tần suất xuất hiện của từ đó trong một tài liệu (term frequency), N là tổng số tài liệu trong dataset và n_i là số lượng các tài liệu mà từ đó xuất hiện.

Phần tiếp theo, chúng tôi sẽ trình bày về giải thuật mở rộng câu truy vấn SoQuES.

4. Giải thuật Personalized Social Query Expansion (SoQuES)

Với lượng thông tin khổng lồ như hiện này thì việc tìm thông tin liên quan ngày càng trở nên khó khăn cho người dùng cuối bởi vì: (i) thông thường, người dùng ko thực sự biết rõ những gì mình đang tìm kiếm cho đến khi tìm thấy nó, (ii) nếu có biết thì người dùng cũng không biết dùng câu truy vấn nào cho phù hợp với nhu cầu.

Và việc chuẩn hóa câu truy vấn bằng việc mở rộng nó (query expansion) là một giải pháp tốt cho vấn đề trên. 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.

Trong nghiên cứu này, chúng tôi sử dụng phương pháp mở rộng câu truy vấn (query expansion) của Bouadjenek và các đồng nghiệp của ông đã đề xuất ở [8] để chuẩn hóa câu truy vấn cho hệ thống tìm kiếm.

4.1. Định nghĩa vấn đề

Cho một câu truy vấn $Q = \{t_1, t_2, ..., t_m\}$ được nhập bởi người dùng u, làm cách nào để cung cấp cho mỗi $t_i \in Q$ một danh sách xếp hạng các từ khóa liên quan đến nó $\{t_{i1}, t_{i2}, ..., t_{ik}\}$, như vậy khoảng cách giữa sự mong đợi của người dùng và kết quả trả về từ hệ thống được giảm thiểu. Mục tiêu ở đây là để chuyển đổi câu truy vấn Q thành câu truy vấn mới Q' sao cho: (i) Q là nhất thiết phải có trong Q', (ii) các kết quả của Q có trong những Q ', và (iii) các kết quả thu được với Q' nên tăng độ chính xác của các kết quả và không làm giảm sự hài lòng của người dùng. Phần tiếp theo là chi tiết về giải thuật SoQuES cho việc giải quyết vấn đề này.

4.2. Giải thuật SoQuES

Algorithm: Personalized Social Query Expansion (SoQuES)

Require: A social folksonomy Graph G; u: a User; Q: a Query;

1: **for all** $t_i \in Q$ do
2: L ← list of neighbor of t_i in tag graph G_{tag}

3: **for all** $t_j \in L$ **do**
4: t_j.Value ← $Rank_{t_i}^{u}(t_j)$
5: Sort L by t_j.Value and take top k terms in L
6: Make a logical OR (∨) between t_i and all terms of L
7: Update Q′
8: **return** Q′

Thông tin người dùng (user profile) được biểu diễn bằng một vectơ trọng số $\overrightarrow{p_u}$ = {w_{t_1} , w_{t_2} , ..., w_{t_n} }, trong đó w_{t_i} được tính bằng phương pháp tf-idf (term frequency - inverse document frequency) (dòng 1). Ở dòng 3, lấy tất cả các chú thích láng giềng t_j của t_i trong đồ thị chú thích G_{tag}. Sau đó, ở dòng 4 và 5, với mỗi t_j, tính độ tương đồng giữa chú thích t_i và t_j của người dùng u. $Rank_{t_i}^{u}(t_j)$ được tính toán như sau:

$$\text{Rank}_t^u(t_i) = \gamma \times \text{Sim}(t, t_i) + (1 - \gamma) \times \frac{1}{m}\sum_{t_j \in p_u}^{m} Sim(t_i, t_j) \times w_{t_j} \quad (4)$$

Trong đó, Sim(t, t_i) là độ tương đồng giữa từ khóa t và t_i, m là chiều dài của user profile và w_{t_j} là trọng số của t_j trong user profile. Chúng tôi sử dụng thuật giải SocialSimRank (SSR) [3] để tính độ tương đồ Sim(t_i, t_j).

Tiếp theo, sắp xếp danh sách chú thích ở dòng 3 dựa vào giá trị của $Rank_{t_i}^{u}(t_j)$ và chỉ giữ top k chú thích (dòng 6). Cuối cùng là kết hợp t_i với các từ trong danh sách được sắp xếp ở trên.

Ví dụ: Khi người dùng nhập vào câu truy vấn:
$Q = t_1 \wedge t_2 \wedge ... \wedge t_m$, nó sẽ được mở rộng để trở thành câu truy vấn mới:
$Q' = (t_1 \vee t_{11} \vee ... \vee t_{1l}) \wedge (t_2 \vee t_{21} \vee ... \vee t_{2k}) \wedge ... \wedge (t_m \vee t_{m1} \vee ... \vee t_{mr})$.

Trong phần này, chúng tôi vừa trình bày chi tiết các bước của giải thuật SoQuES.

Phần tiếp theo, chúng tôi sẽ nói về việc thu thập dữ liệu từ mạng xã hội Twitter.

5. Khai thác dữ liệu mạng xã hội Twitter

Twitter [12] 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. Theo số liệu của ngành truyền thông xã hội gần đây, Twitter hiện đang là một trong những mạng xã hội hàng đầu trên toàn thế giới dựa trên các thành viên hoạt động. Tính đến quý IV năm 2015, Twitter đã có 305 triệu người sử dụng hàng tháng hoạt động và hơn 500 triệu tweet mỗi ngày tạo ra [13].

Bên cạnh đó, Twitter cho phép chúng ta tương tác với dữ liệu tweets và các dữ liệu khác liên quan đến tweets thông qua Twitter APIs. Đặc biệt, chúng ta có thể thu thập dữ liệu tweets theo thời gian thực thông qua Twitter's Streaming API. Vì vậy, chúng tôi đã tiến hành khai thác dữ liệu từ đây để cung cấp dữ liệu cho hệ thống tìm kiếm của mình.

6. Kết quả thực nghiệm và đánh giá

6.1. Dữ liệu chú thích cộng đồng

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: (1) lọc bỏ những tweets mà không thuộc ngôn ngữ tiếng Anh và không có chứa URL, (2) trích xuất những chú thích (hashtag) và địa chỉ web (URL) từ tweets, (3) loại bỏ những chú thích vô nghĩa như "!picspam", "atthissummer" dựa trên từ điển WordNet, (4) loại bỏ những địa chỉ trang web mà không thuộc ngôn ngữ tiếng Anh thông qua Apache Tika toolkit.

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

Bảng 1. *Dữ liệu từ Twitter*

Tweets	Người dùng	Chú thích	URL
2.520.358	365.939	162.987	745.286

6.2. Phương pháp đánh giá

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).

$$\text{MMAP} = \frac{\sum_{i=1}^{N_u} MAP_i}{N_u}$$

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

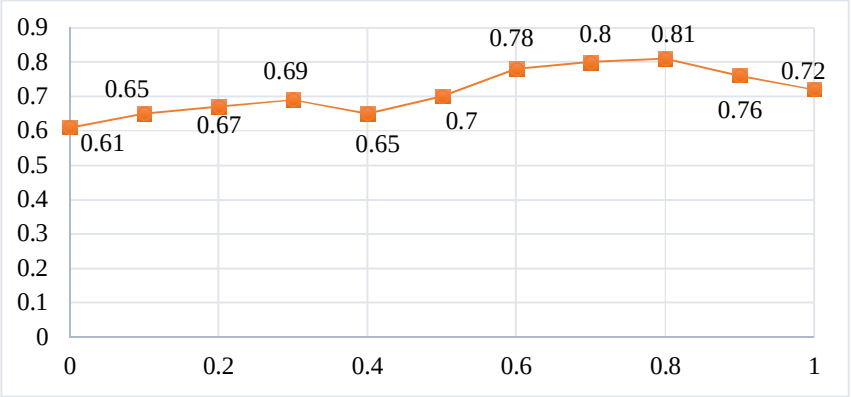
Chúng tôi chọn ngẫu nhiên 1000 cặp (u,t). Trong mỗi cặp, user u gửi câu truy vấn q = {t} đến hệ thống tìm kiếm. Tiếp theo, hệ thống sẽ tìm kiếm và xếp hạng kết quả tìm kiếm phù hợp với câu truy vấn. Cuối cùng, chúng tôi tính độ đo MAP cho 1000 câu truy vấn. Đánh giá một liên kết trả về có tương quan 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.

6.3. Kết quả đánh giá

Trong công thức xếp hạng trang web (2), chúng tôi thực nghiệm với $0 \leq \alpha \leq 1$ và $\beta = 0.5$.

Công thức (4), chúng tôi chọn $\gamma = 0.5$ 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.



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

Theo như hình 1, hiệu quả của hệ thống tìm kiếm đạt tốt nhất khi $\alpha \in [0.6, 0.8]$.

7. Kết luận

Trong bài báo này, chúng tôi đã nghiên cứu làm thế nào để khai thác và sử dụng các chú thích xã hội trong việc tìm kiếm thông tin. Chú thích xã hội không chỉ cung cấp nội dung mà còn là một bản tóm tắt, đồng thời chỉ ra sự phổ biến của các trang web. Thông qua đó cài đặt các thuật toán chuẩn hóa câu truy vấn và xếp hạng trang web ứng dụng trong công cụ tìm kiếm. Kết quả tìm kiếm khắc phục được sự cứng nhắc trong việc tìm kiếm chính xác các từ khóa trong câu truy vấn. Nghiên cứu đã tận dụng ưu điểm về thông tin cũng như mối quan tâm và sở thích của người dùng web để hỗ trợ người dùng tìm kiếm một cách nhanh chóng thông tin mà họ cần.

Kết quả thực nghiệm không chỉ chứng minh tính hiệu quả của công cụ tìm kiếm của chúng tôi, mà còn cho thấy mô hình ứng dụng chú thích cộng đồng vào trong công cụ tìm kiếm là một hướng nghiên cứu rất khả thi và có tiềm năng ứng dụng rất cao đối với các công cụ tìm kiếm.

TÀI LIỆU THAM KHẢO

[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