ĐẠI HỌC QUỐC GIA HÀ NỘI
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Đinh Thị Thanh Loan**

# NGHIÊN CỨU KĨ THUẬT SO SÁNH TRUY VẤN ĐỂ GỢI Ý TÌM KIẾM THÔNG TIN CHO THANH THIẾU NIÊN VÀ THỬ NGHIỆM

**CHUYÊN NGÀNH: KỸ THUẬT PHẦN MỀM**
**MÃ SỐ: 60480103**

## LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

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

**Hà Nội - 2016**

# LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là công trình nghiên cứu của riêng cá nhân tôi, không sao chép của ai, do tôi tự nghiên cứu, đọc, dịch tài liệu, tổng hợp và thực hiện. Trong luận văn, việc sử dụng nội dung các công trình nghiên cứu của người khác đều được chỉ dẫn tường minh từ các tài liệu tham khảo. Các số liệu, chương trình phần mềm và những kết quả trong luận văn là trung thực và chưa được công bố trong bất kỳ một công trình nào khác.

Hà Nội, tháng 10 năm 2016

Học viên thực hiện

Đinh Thị Thanh Loan

# LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời biết ơn sâu sắc đến PGS.TS. Hà Quang Thụy người đã tận tình hướng dẫn, chỉ bảo, giúp đỡ em trong suốt quá trình làm luận văn.

Em cũng xin gửi lời cảm ơn đến các các thầy cô Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã truyền đạt những kiến thức và giúp đỡ em trong suốt quá trình học của mình.

Tôi cũng xin gửi lời cảm ơn tới Ban Giám Hiệu trường THPT Đại Mỗ đã tạo điều kiện thuận lợi cho tôi tham gia khóa học và trong suốt quá trình hoàn thành luận văn.

Và cuối cùng tôi xin gửi lời cảm ơn tới các đồng nghiệp, gia đình và bạn bè những người đã ủng hộ, động viên tạo mọi điều kiện giúp đỡ để tôi có được kết quả như ngày hôm nay.

Hà Nội, tháng 10 năm 2016

Học viên

Đinh Thị Thanh Loan

# MỤC LỤC

LỜI CAM ĐOAN ..... 2
LỜI CẢM ƠN ..... 3
MỤC LỤC ..... 4
CÁC HÌNH VẼ TRONG LUẬN VĂN ..... 6
CÁC BẢNG BIỂU TRONG LUẬN VĂN ..... 7
CHÚ THÍCH VIẾT TẮT THUẬT NGỮ TIẾNG ANH ..... 8
MỞ ĐẦU ..... 9
Chương 1. GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN ..... 10
1.1. Giới thiệu chung an toàn Internet đối với thanh thiếu niên ..... 10
1.1.1. Ảnh hưởng của Internet đối với giới trẻ ..... 10
1.1.2. Biện pháp an toàn Internet đối với thanh thiếu niên ..... 10
1.2. Gợi ý truy vấn cho thanh thiếu niên ..... 13
1.2.1. Ý nghĩa của gợi ý truy vấn cho thanh thiếu niên ..... 13
1.2.2. Gợi ý truy vấn cho thanh thiếu niên và một số bài toán liên quan ..... 14
1.2.3. Một số kỹ thuật gợi ý truy vấn cho thanh thiếu niên ..... 15
1.3. Bài toán gợi ý truy vấn bằng kỹ thuật so sánh câu truy vấn ..... 17
Chương 2. MỘT SỐ KỸ THUẬT GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN ..... 18
2.1. Gợi ý truy vấn bằng "đi ngẫu nhiên" ..... 18
2.1.1. Cách tiếp cận ..... 18
2.1.2. Xếp hạng thẻ ..... 18
2.1.3. Phương pháp ..... 18
2.1.4. Nhận xét ..... 23
2.2. Kỹ thuật gợi ý truy vấn bằng so sánh truy vấn (QS) ..... 23
2.2.1. Cách tiếp cận ..... 23
2.2.2. Nội dung phương pháp ..... 25
2.2.3. Nhận xét ..... 33

2.3. Tính tương tự của truy vấn ..... 33
2.3.1. Cách tiếp cận ..... 33
2.3.2. Các phương pháp tính độ tương tự ..... 34
2.4. Ý tưởng giải pháp gợi ý truy vấn cho thanh thiếu niên ..... 38
Chương 3. MỘT MÔ HÌNH GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN ..... 40
3.1. Giới thiệu ..... 40
3.2. Mô hình ..... 40
3.3. Các thành phần của mô hình ..... 41
Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ ..... 44
4.1. Đặt vấn đề ..... 44
4.2. Thi hành mô hình (Phần mềm và phần cứng) ..... 44
4.3. Dữ liệu và quá trình thực nghiệm ..... 44
4.3.1. Dữ liệu ..... 44
4.3.2. Quá trình thực hiện ..... 44
4.4. Kết quả thực nghiệm và đánh giá ..... 46
4.4.1. Giao diện chương trình tính độ tương tự ..... 46
4.4.2. Đánh giá ..... 47
4.4.3. Kết quả trả về từ máy tìm kiếm Google sau khi truy vấn ..... 48
KẾT LUẬN ..... 53
TÀI LIỆU THAM KHẢO ..... 54

# CÁC HÌNH VẼ TRONG LUẬN VĂN

**Ví dụ gợi ý truy vấn của công cụ tìm kiếm google.....14**
**Gợi ý trực quan và gợi ý dạng text .....15**
**Ví dụ gợi ý trực quan .....16**
**Ví dụ gợi ý dạng text .....16**
**Mô hình gợi ý truy vấn.....25**
**Mô hình đề xuất so sánh truy vấn.....41**
**Tìm kiếm với câu truy vấn 1.....48**
**Tìm kiếm với câu truy vấn 2.....49**
**Tìm kiếm với câu truy vấn tổng hợp 1 .....49**
**Tìm kiếm với câu truy vấn 3.....50**
**Tìm kiếm với câu truy vấn 4.....51**
**Tìm kiếm với câu truy tổng hợp 2.....52**

# CÁC BẢNG BIỂU TRONG LUẬN VĂN

**Bảng 2.1 Sắp xếp số truy vấn ứng viên** ..... **31**
**Bảng 2.2 Sắp xếp số gợi ý truy vấn** ..... **31**
**Bảng 4.1 Kết quả tính độ tương tự giữa các truy vấn** ..... **45**
**Bảng 4.2 Bảng phân loại đánh giá** ..... **48**

# CHÚ THÍCH VIẾT TẮT THUẬT NGỮ TIẾNG ANH

| | |
|---|---|
| OFSD | Online frequent sequence discovery |
| P2R | Page rank reviser |
| VSM | Vector space model |
| SBM | Standard boolean model |
| SE | Search engine |

# MỞ ĐẦU

Trong bối cảnh lượng thông tin trên Internet ngày càng lớn và cập nhật gần tức thời như hiện nay thì người dùng, đặc biệt là thanh thiếu niên, một công cụ tìm kiếm những thông tin họ cần một cách hiệu quả nhất ngày càng trở nên cần thiết. Thanh thiếu niên gặp khó khăn khi xây dựng các truy vấn tìm kiếm hoặc lựa chọn loại chủ đề có liên quan, bởi vì kiến thức miền của thanh thiếu niên ít hơn cũng như vốn từ vựng chưa thật phát triển so với người trưởng thành hơn. Một khía cạnh khác, thanh thiếu niên gặp khó khăn trong đánh giá sự liên quan của kết quả tìm kiếm. Để hỗ trợ thanh thiếu niên trong việc tìm kiếm thông tin thực sự hiệu quả, hệ thống tìm kiếm cần phải được thiết kế sao cho việc hỗ trợ phù hợp với nhu cầu và năng lực nhận thức của tuổi trẻ. Luận văn này đề cập tới các nội dung nghiên cứu về gợi ý truy vấn thông tin phù hợp với thanh thiếu niên và thử nghiệm ứng dụng các nghiên cứu đó.

Luận văn tập trung vào các phương pháp và kỹ thuật gợi ý truy vấn như random walk (nghĩa là "đi một cách ngẫu nhiên", sử dụng từ gốc tiếng Anh "random walk" trong luận văn) [1], so sánh câu truy vấn [7] sao cho phù hợp với nhu cầu và năng lực nhận thức của thanh thiếu niên, hỗ trợ họ tìm kiếm thông tin hiệu quả. Luận văn thực hiện khảo sát, nghiên cứu các phương pháp tính độ tương tự truy vấn trong hệ tìm kiếm như phương pháp dựa trên từ vựng [16] và phương pháp dựa trên nhật ký truy vấn [18]. Từ đó đưa ra hướng phát triển cho phương pháp tính độ tương tự truy vấn phù hợp để áp dụng thử nghiệm vào một hệ tìm kiếm cho thanh thiếu niên.

# Chương 1. GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN

## 1.1. Giới thiệu chung an toàn Internet đối với thanh thiếu niên

### 1.1.1. Ảnh hưởng của Internet đối với giới trẻ

Theo báo cáo khảo sát của LSE Research Online năm 2010 [8], tại 25 quốc gia châu Âu thì có đến 93% thanh thiếu niên sử dụng Internet mỗi năm và 60% lên mạng mỗi ngày, trong đó 80% thanh thiếu niên sử dụng Internet có độ tuổi từ 15-16. Cũng theo báo cáo, 85% thanh thiếu niên sử dụng Internet tại trường học, 83% sử dụng trò chơi, 62% đọc tin tức 62%, 16% dùng các website chia sẻ dữ liệu và 11% sử dụng blog. Khảo sát cũng tập trung vào các chủ đề như trấn lột, nội dung khiêu dâm, thông tin tình dục, giao lưu hẹn hò trực tuyến... là những chủ đề có khả năng ảnh hưởng gây hại đến thanh thiếu niên. Đối tượng thanh thiếu niên luôn có xu hướng thiếu kỹ năng và độ tự tin khi truy cập mạng Internet. Tuy nhiên, hầu hết trẻ từ 11-16 tuổi có thể ngăn chặn hoặc từ chối tới những người mà chúng không muốn liên lạc (64%) hoặc tìm lời khuyên an toàn trực tuyến (64%). Khoảng một nửa có thể thay đổi cài đặt riêng tư trên hồ sơ ở các trang mạng xã hội mà mình tham gia (56%), ngăn chặn thư rác (51%).

Sách trắng Công nghệ Thông tin Việt Nam năm 2014[1] cho biết, vào năm 2013, số người Việt Nam sử dụng Internet lên tới trên 33 triệu 191 nghìn người, chiếm tỷ lệ 37,00% dân số và doanh thu dịch vụ Internet đạt trên 965 triệu đô la Mỹ. Đối tượng sử dụng internet chủ yếu là giới trẻ với độ tuổi từ 15 đến 24, phần chủ yếu trong đó là các đối tượng thanh thiếu niên.

Ngày nay, với sự phát triển gia tăng đến cấp số nhân các dòng điện thoại thông minh và người sử dụng để truy cập Internet, mà phần lớn là thanh thiếu niên, thì nguy cơ độc hại đối với đối tượng này lại càng cao [10]. Ngoài việc tham gia vào các hoạt động xã hội, thể hiện bản thân, học tập và quản lý cuộc sống hàng ngày đã trở nên dễ dàng hơn thì nguy cơ tiếp xúc trực tiếp với các loại thông tin độc hại tạo ra những thách thức mới về an toàn trực tuyến cho trẻ em, chẳng hạn như mới nổi các rủi ro liên quan đến dịch vụ định vị theo dõi...

### 1.1.2. Biện pháp an toàn Internet đối với thanh thiếu niên

[1] http://www.mediafire.com/view/6606y4bt1yju6wc/SachTrang2014.pdf

Nghiên cứu của LSE Research Online [8] cũng đưa ra các lời khuyên về an toàn Internet đối với thanh thiếu niên: Nhà trường, đặc biệt là giáo viên, phụ huynh học sinh, chính phủ, chính quyền địa phương hoặc từ chính bạn bè của học sinh nếu có thể thì cần tham gia trực tiếp cùng với học sinh khai phá thông tin trên Internet. Cần hỗ trợ tìm kiếm thông tin và đưa ra những lời khuyên bổ ích cho học sinh của mình. Ngành giáo dục có một vai trò rất lớn đảm bảo an toàn internet tại trường học, vì vậy rất cần những nỗ lực và các biện pháp cụ thể từ ngành giáo dục. Khi mà nhu cầu sử dụng Internet ngày càng có tính chất riêng tư, cá nhân hóa, trong đó sự tham gia của bố mẹ và thầy cô giáo cũng trở nên khó khăn hơn, thì vai trò của cơ chế chính sách quản lý truy cập mạng của chính phủ phải được phát huy tích cực. Trong đó, vai trò các công ty, doanh nghiệp, ngành công nghiệp phát hành nội dung thông tin và truyền thông trên mạng Internet phải được làm rõ nhằm hạn chế những rủi ro mà trẻ em thường gặp phải.

Nếu có thể thành lập được một tổ chức, hiệp hội tại mỗi quốc gia về An toàn Internet cho thanh thiếu niên, lập ra đường dây nóng nhằm mục đích hỗ trợ giải đáp các thắc mắc về các vấn đề khi truy cập Internet của giới trẻ. Ở đó tất cả các vấn đề phát sinh trong quá trình sử dụng Internet của giới trẻ đều được giải đáp một cách kịp thời và nhanh chóng [9].

Tóm lại, Internet là một công cụ nghiên cứu tuyệt vời và cũng là một cách giao tiếp vui thích để liên lạc với bạn bè và gia đình. Nhưng làm việc trực tuyến trên Internet cũng có thể hiện một số nguy hiểm mà cần phải biết để phòng ngừa. Dưới đây là một số cách thức hoặc giải pháp được tổng hợp trong báo cáo tại Safer Internet Day[2] nên được thực hiện để tránh xa các rắc rối khi sử dụng các trang Web [8, 9, 11]:

- Không bao giờ gửi thông tin cá nhân của chúng ta, chẳng hạn như tên, địa chỉ, số điện thoại, hình ảnh hoặc tên trường chúng ta vào không gian ảo.

- Có thể gửi chuyển tiếp thư điện tử bằng cách nhắp chuột. Hãy nhớ rằng bất kỳ thông tin cá nhân nào mà chúng ta gửi đến cho người nào đó thì cũng có thể được gửi đến cho những người khác rất nhanh.

- Không bao giờ lập các kế hoạch gặp một "người bạn" trực tuyến tận mặt mà không kiểm tra trước với phụ huynh/người giám hộ của chúng ta. Nếu phụ

[2] www.saferinternetday.org

huynh/người giám hộ ĐỒNG Ý với ý kiến này, hãy dẫn phụ huynh/người giám hộ đi cùng và gặp người bạn đó tại một địa điểm công cộng. Hãy nhớ rằng bất kể người nào đó trực tuyến có vẻ vui tính và thân thiện, nhưng trong thực tế, họ có thể là người hoàn toàn khác.

- Hành vi trực tuyến của mỗi người là trách nhiệm của bản thân. Không quấy rối hoặc bạo hành và không trả lời khi có người nào khác cố ý tranh luận trực tuyến.

- Nếu chúng ta đương đầu với người nào hoặc cái gì đó trực tuyến làm cho chúng ta bực bội khó chịu, hãy nói cho một người lớn đáng tin cậy biết ngay lập tức! Người lớn này có thể xem xét thông tin trên màn hình và quyết định xem có nên báo cáo cho chính quyền hay không.

- Nhắc nhở con em thanh thiếu niên của chúng ta không tiết lộ thông tin cá nhân trực tuyến.

- Cùng nhau phác thảo một danh sách về những gì không nên chia sẻ, gồm cả tên, tuổi, trường học, số điện thoại và hình ảnh.

- Nói chuyện thường xuyên với con em thanh thiếu niên của chúng ta. Thảo luận với bạn bè trực tuyến của chúng khi chúng ta nói về những người bạn khác của chúng.

- Để máy tính trong một khu vực chung trong nhà. Làm như thế để giám sát việc sử dụng máy tính dễ dàng hơn.

- Yêu cầu con em thanh thiếu niên của chúng ta nói cho chúng ta ngay lập tức về bất kỳ trải nghiệm trực tuyến phiền phức nào, chúng ta phải bảo đảm rằng sẽ không giận dữ nếu chúng tâm sự với chúng ta.

- Xem xét việc sàng lọc hoặc giám sát phần mềm. Trong khi chúng có thể giúp chúng ta kiểm soát việc sử dụng trực tuyến của con em thanh thiếu niên chúng ta, các thiết bị này không thay thế được sự liên quan và giám sát của phụ huynh.

- Con em thanh thiếu niên của chúng ta có thể dùng máy tính tại trường học, thư viện công và nhà bạn bè.

Và để trợ giúp cho việc truy cập Internet an toàn, việc tìm kiếm thông tin trên mạng tốt hơn thì rất cần thiết đưa ra được giải pháp truy vấn an toàn cho các đối tượng là thanh thiếu niên trong việc tìm kiếm và khai phá thông tin trên mạng Internet.

## 1.2. Gợi ý truy vấn cho thanh thiếu niên

### 1.2.1. Ý nghĩa của gợi ý truy vấn cho thanh thiếu niên

Đối tượng thanh thiếu niên thường xuyên sử dụng công cụ tìm kiếm như là điểm khởi đầu trong việc tìm kiếm các thông tin của họ. Thật không may, kinh nghiệm tìm kiếm của họ có thể bị ảnh hưởng tiêu cực bởi họ thiếu năng lực xây dựng các truy vấn tìm kiếm phù hợp [5].

Gần đây, các nhà nghiên cứu và phát triển cũng đã có quan tâm đến phân tích hành vi tìm kiếm của các nhóm người dùng khác nhau, đặc biệt là nhóm người dùng trẻ em. Đối tượng này gặp khó khăn lớn trong việc thao tác, định hướng tìm kiếm thông tin [7]. Vì vậy, việc đưa ra được giải pháp gợi ý tìm kiếm có ý nghĩa hết sức to lớn cho các đối tượng thanh thiếu niên.

Gợi ý truy vấn nói chung là một phần tích hợp của công cụ tìm kiếm web. Các công cụ tìm kiếm hiện nay đã cung cấp khá tốt cho mọi đối tượng người sử dụng.


*Hình 1.1 Ví dụ gợi ý truy vấn "game" của công cụ tìm kiếm google*

Tuy nhiên, với lượng kết quả trả về có thể là rất lớn, việc tìm được kết quả của người dùng là khá khó khăn nếu không có biện pháp sắp xếp kết quả, lọc trả về tối ưu cho mỗi đối tượng sử dụng [2].

*Mục tiêu chính của một công cụ tìm kiếm là để lấy kết quả liên quan của một truy vấn với kết quả chính xác nhất có thể. Mặc dù mục tiêu này chủ yếu phụ thuộc vào các thuật toán xếp hạng của công cụ tìm kiếm và chất lượng của các truy vấn được gửi cũng là quan trọng* [6].

Việc có quá nhiều kết quả trả về một phần cũng vì câu truy vấn người dùng đưa vào là khá mơ hồ và không rõ nghĩa. Do đó, việc đưa ra những câu gợi ý truy vấn cho người dùng cho các đối tượng khác nhau, đặc biệt là trẻ em, cũng là một bài toán thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu nhằm xây dựng được một công cụ tìm kiếm thông tin cho người trẻ giải quyết được những khó khăn như trên một cách toàn diện nhất có thể [4].

### 1.2.2. Gợi ý truy vấn cho thanh thiếu niên và một số bài toán liên quan

Mặc dù đã có một số công cụ tìm kiếm được thiết kế đặc biệt dành riêng cho đối tượng là thanh thiếu niên chẳng hạn như safe-searchkids.com, kidsclick.org, và kidrex.org, nhưng đa số trong đó là không tích hợp kỹ thuật tìm kiếm gợi ý dành riêng cho thanh thiếu niên [5].

Từ những những khó khăn khi chủ thể tìm kiếm là thanh thiếu niên nêu trên như: không biết diễn đạt câu truy vấn, không biết lựa chọn từ khóa hoặc không biết cách xử lý tình huống kết quả trả về quá dài v.v… và từ đó đã hình thành các bài toán về gợi ý truy vấn.

Gợi ý truy vấn thường được định nghĩa là để "tìm kiếm một số truy vấn liên quan cho truy vấn ban đầu do người dùng phát hành" [7]. Ví dụ, khi người dùng truy vấn các vấn đề "hãng hàng không vietnamairline", công cụ tìm kiếm sẽ đề nghị tìm kiếm những thuật ngữ như "vé máy bay", "vé máy bay trực tuyến", "phòng hãng hàng không vietnamairline", vv.

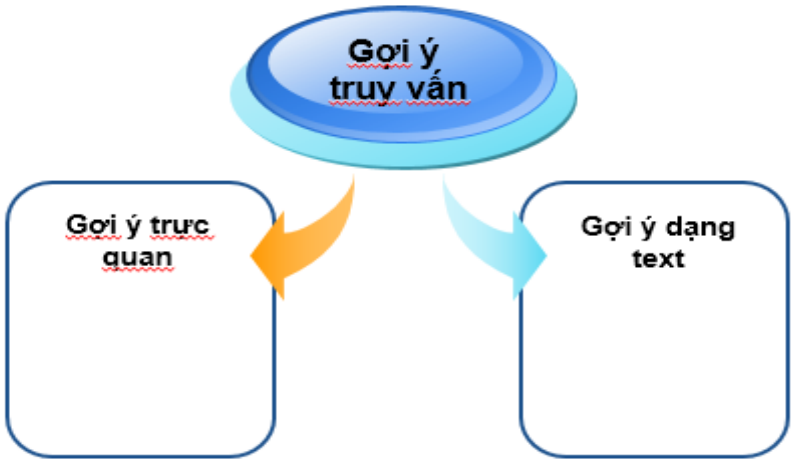
Các bài toán liên quan đến gợi ý truy vấn có thể kể đến như: Tìm ngữ cảnh truy vấn (*Query Context*), lưu trữ truy vấn vào *QueryLog*, xếp thứ hạng kết quả truy vấn, thứ hạng thẻ tìm kiếm truy vấn,… Ngữ cảnh truy vấn có thể được hiểu đơn giản là hợp của hai tập mở rộng truy vấn (query extensions) và tương tự truy vấn (query adjacent) [4], mà tập mở rộng thường nối tiếp sau truy vấn đang xét, và tập tương tự truy vấn thường xuất hiện trước hoặc sau truy vấn đang xét. Ngữ cảnh truy vấn thường cung cấp những gợi ý quan trọng về ý đồ tìm kiếm của người dùng.

Một cách sơ bộ, bài toán của luận văn này là đưa ra giải pháp trợ giúp được người dùng trẻ tuổi trong việc diễn đạt ý đồ tìm kiếm và đưa ra những câu truy vấn gần với ý đồ tìm kiếm.

### 1.2.3. Một số kỹ thuật gợi ý truy vấn cho thanh thiếu niên

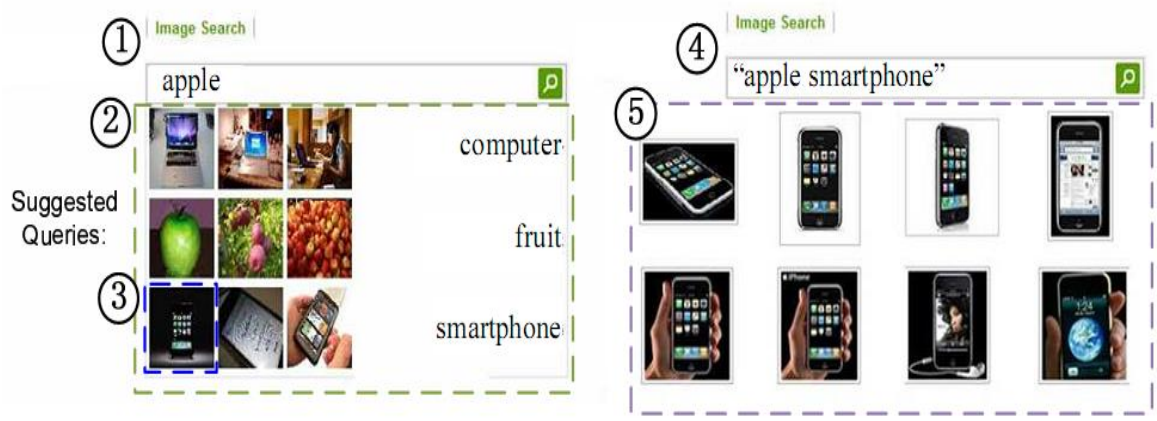
Đa số giải pháp kỹ thuật gợi ý truy vấn có thể hoạt động trên nhiều loại dữ liệu như khai phá dữ liệu nhật ký truy vấn, nội dung tài liệu hoặc sử dụng bối cảnh.

Có hai thể hiện chính gợi ý truy vấn cho thanh thiếu niên, đó là gợi ý trực quan và gợi ý dạng văn bản (text) [4]:




*Hình 1.2 Gợi ý trực quan và gợi ý dạng text*

- Gợi ý trực quan tức là dùng các hình ảnh trực quan để thể hiện các gợi ý khi tìm kiếm




*Hình 1.3 Ví dụ gợi ý trực quan*

- Gợi ý dạng văn bản là đưa ra một danh sách các từ liên quan để người dùng có thể tự tìm kiếm

Searches related to **tiger**

| | |
|---|---|
| tiger **pictures** | **white** tiger |
| **bengal** tiger | tiger **habitat** |
| tiger **photos** | **onitsuka** tiger |
| **animal** tiger | **lion** |

*Hình 1.4 Ví dụ gợi ý dạng text*

Các kỹ thuật gợi ý truy vấn có thể áp dụng truy vấn cho thanh thiếu niên tập trung vào khai phá nhật ký truy vấn QueryLog. QueryLog được định nghĩa là nơi lưu trữ dữ liệu về hành vi của người dùng trong quá khứ. Với đặc thù của hệ thống tìm kiếm là nặc danh, bất cứ ai cũng có thể sử dụng mà không cần xác thực. Tuy nhiên, hệ thống vẫn cho phép cấp phát một mã số cho từng phiên làm việc của những người dùng khác nhau. Điều này cho phép xác định được các hành vi của một người dùng trong một phiên. Phiên làm việc ở đây được hiểu là một lần sử dụng của người dùng từ lúc truy cập hệ thống đến lúc thoát khỏi hệ thống. QueryLog là tập các bản ghi, mà về phổ biến, bao gồm các trường thông tin sau:

- SessionID: mã của phiên làm việc

- TimeStamp: nhãn thời gian ghi lại thời điểm xảy ra sự kiện
- Query: câu truy vấn của người dùng
- TopN: N tài liệu đầu tiên trong tập kết quả
- UrlClicked: tài liệu được nhấn bởi người sử dụng
- QuerySegmented: các từ khóa trong câu truy vấn

Các kỹ thuật dùng để gợi ý truy vấn như: Kỹ thuật phân cụm truy vấn, thống kê, hướng ngữ cảnh, so sánh câu truy vấn, hay sử dụng các phương pháp học giám sát để xếp hạng truy vấn gợi ý cho thanh thiếu niên [5]. Các kỹ thuật gần đây hơn tập trung khai phá khái niệm phiên, tính chất chuỗi truy vấn để phát hiện và gợi ý các truy vấn hướng ngữ cảnh. Tuy nhiên, không nhiều các công trình nghiên cứu khai phá tiêu chí về kỹ thuật so sánh truy vấn...

### 1.3. Bài toán gợi ý truy vấn bằng kỹ thuật so sánh câu truy vấn

*Các vấn đề gợi ý truy vấn (Query suggestion: QS) nên chỉ nghĩ đơn giản như là "một chuỗi của so sánh hai câu truy vấn"* [7]. Truy vấn đầu tiên trong chuỗi chính là truy vấn ban đầu đã được phát ra bởi người tìm kiếm (người sử dụng). Các truy vấn thứ hai gọi là "truy vấn ứng viên" là các truy vấn được đề nghị (gợi ý) cho người sử dụng, thường đặt ở phần cuối của danh sách tìm kiếm. Việc so sánh các truy vấn có thể phụ thuộc vào một số tính năng như sự giống nhau thuật ngữ, nhật ký truy vấn, vv. Cách tiếp cận so sánh truy vấn này làm cho đơn giản hóa vấn đề gợi ý truy vấn, đơn giản quá trình theo dõi, mở rộng và có thể gỡ lỗi.

Một tập hợp các truy vấn ứng cử viên (đề nghị) query $q_i$ được so sánh với một tập các ứng viên query $q_c$. Cuối cùng, các truy vấn ứng cử viên có thể được sắp xếp dựa trên thứ hạng/trọng số của chúng và $n$ ứng cử viên truy vấn tốp đầu có thể được trình bày cho người dùng như gợi ý truy vấn.

# Chương 2. MỘT SỐ KỸ THUẬT GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN

## 2.1. Gợi ý truy vấn bằng "đi ngẫu nhiên"

### 2.1.1. Cách tiếp cận

Theo S. D. Torres và cộng sự [1], trong kỹ thuật đi ngẫu nhiên (random walk), một phương pháp gợi ý truy vấn để giúp trẻ em dễ dàng tìm các từ khóa liên quan sử dụng kỹ thuật random walk. Phương pháp gợi ý truy vấn này dựa trên các thẻ (Tag) từ vựng từ một hệ thống đánh dấu Delicious (Delicious- là một trang web internet được thiết kế để cho phép truy cập vào bất kỳ trang web nào mà người dùng đánh dấu) liên quan các kết quả truy vấn web và các tài nguyên web nhìn thấy trước đây dành cho trẻ em.

Các thẻ liên quan thường xuyên hơn đến URL tập trung vào trẻ em với các chủ đề là ứng cử viên tốt hơn để xây dựng đề xuất truy vấn cho trẻ em. Ví dụ: Hãy xem xét truy vấn về xe ô-tô. Theo đề xuất gợi ý truy vấn phổ biến của Google, các khía cạnh liên quan đến truy vấn này có thể là cho thuê xe hơi, xe ô tô để bán, sử dụng xe hơi, xe ô tô mới hay hình ảnh xe hơi... Trong khi khía cạnh định hướng để đáp ứng nhu cầu thông tin trẻ em cần thay vào đó bao gồm các khía cạnh như trò chơi xe hơi, đồ chơi xe hơi, phim về xe hơi, hình ảnh xe hơi... Hệ thống này xếp hạng các thẻ cao hơn và cung cấp các gợi ý tập trung hơn vào nội dung dành riêng cho các đối tượng được phân loại.

### 2.1.2. Xếp hạng thẻ

Xếp hạng thẻ hoặc từ khóa gần đây đã nhận được nhiều sự quan tâm chú ý cho sự phát triển chia sẻ của xã hội. Đã có những phương pháp để ước tính đến trọng số liên quan giữa thẻ và hình ảnh dựa trên phương pháp dự đoán xác xuất. Phương pháp random walk được biểu diễn trên một đồ thị hai chiều bao gồm thẻ và tài nguyên web (url) [1]. Vấn đề quan trọng của cấu trúc đồ thị của phương pháp này là khai phá các đặc điểm tài nguyên web nhắm vào trẻ em.

### 2.1.3. Phương pháp

Phần này mô tả các kịch bản kỹ thuật truy vấn mở rộng và phương pháp random walk được S. D. Torres và cộng sự đề xuất [1] về gợi ý truy vấn sử dụng các thẻ từ mạng xã hội.

#### 2.1.3.1. Kịch bản

Phương pháp dùng lại một dịch vụ tìm kiếm cho trẻ em của công cụ tìm kiếm tiên tiến nhất để cung cấp nội dung cho trẻ. Trong hệ thống này, các truy vấn của người dùng được gửi đến công cụ tìm kiếm để lấy các từ khóa, các trích đoạn, các chủ đề kết quả trên web. Những từ khóa này đại diện cho các chủ đề có thể liên quan đến truy vấn của người dùng. Nhiệm vụ phương pháp là tạo ra những từ khóa và xếp hạng chúng để xây dựng gợi ý truy vấn. Lưu ý rằng trong kịch bản này không có quyền truy cập để tìm kiếm nhật ký truy vấn được sử dụng rộng rãi cho các gợi ý truy vấn trước đây. Hơn thế nữa mối quan tâm ngày càng tăng về tính riêng tư và các đặc trưng đối tượng mục tiêu của phương pháp này là trẻ em, cần tránh tạo tình huống theo dõi thông tin người dùng.

#### 2.1.3.2. Mô hình Random walk hướng tới nội dung cho trẻ em

Mô hình Random walk sử dụng một đồ thị hai phía là gồm các nút nguồn tài nguyên web (tức là, url) và các nút thẻ (Tag). Một số nghiên cứu gợi ý truy vấn dựa trên xếp hạng thẻ sử dụng phương pháp Random walk cho hệ thống gợi ý nhưng chỉ sử dụng đồ thị chỉ gồm các thẻ (Tag) [1, 4].

Việc xem các URL (nguồn tài nguyên web tin cậy) như là các nút là một việc rất hữu ích trong phương pháp này, nó là yếu tố nguồn gốc theo xu hướng random walk phù hợp hơn cho các đối tượng mục tiêu. Kết hợp thường xuyên hơn giữa các thẻ với URL với mục tiêu nhắm vào đối tượng nhất định người sử dụng (ví dụ trẻ em) sẽ được thường xuyên làm nổi bật hơn trên các thẻ để mô tả các url thích hợp cho người sử dụng khác (ví dụ như người lớn).

Lưu ý rằng sẽ không dễ dàng để trình biểu diễn tin trong trường hợp biểu đồ chỉ bao gồm những nút thẻ(Tag), hơn nữa biểu diễn cho biểu đồ này cho phép thêm một tiêu chuẩn để đánh giá nguồn gốc của một url như thế nào là tin cậy hay đáng tin cậy (ví dụ, dựa trên nguồn hoặc độ phổ biến của nó).

Trong kỹ thuật này, các biểu đồ được thể hiện nhờ một tập các đánh dấu (bookmarks). Cụ thể, đánh dấu các url được biết đến là phù hợp cho trẻ em để tạo ra tập bao gồm các url và các thẻ. Biểu đồ chính thức được định nghĩa là:

**Định nghĩa 1**. (đồ thị hai chiều) một đồ thị hai chiều của các url và các thẻ [1]:

$$G = (U,T,E = \{(u,t)|(u,t) \in U \times T\}) \quad (2.1)$$

Trong đó $U=\{u1, u2,..un\}$ là một tập các URL mô tả bởi các Tag

*T={t1,t2,..tn}* và E là tập cạnh trên đồ thị.

Xác xuất chuyển đổi được định nghĩa như sau:

$$P_{fw}(i|j) = \begin{cases} (1-\alpha)\dfrac{c(i,j)}{\sum_{k:(j,k)\in E} \mathrm{c(j,k)}} & if\ i \neq j \\ \alpha & for\ i = j \end{cases} \qquad (2.2)$$

Gọi *c(i; j)* tượng trưng cho số lần một từ khóa mà *i* đã được sử dụng để mô tả một nguồn tài nguyên web *j* và chính số hạng đó là xác suất chuyển đổi được sử dụng để làm chậm (giảm) truyền tin của trọng số. Trong kỹ thuật random walk này sẽ sử dụng công thức này như danh giới (điểm dừng).

Kỹ thuật random walk sử dụng khoảng cách Kullback-Leibler (KL) trong đo lường thông tin. Khoảng cách Kullback-Leibler (hoặc entropy tương đối) là một cách so sánh hai phân bố: phân bố "thật" p(x) và một phân bố bất kì q(x). Nó được định nghĩa như sau:

$$D_{\mathrm{KL}}(p(\mathrm{X})\|q(\mathrm{X})) = \sum_{z\in X} -p(x) log\ q(x) - (-p(x)\log p(x)) = \sum_{z\in X} p(x)\log\frac{p(x)}{q(x)} \qquad (2.3)$$

Mặc dù đôi khi được gọi như một "khoảng cách metric", tuy nhiên, khoảng cách Kullback-Leibler không phải là một metric do nó không đối xứng và không thỏa mãn bất đẳng thức tam giác.

Bằng trực giác, độ đo này cho phép một cách thức minh bạch để nâng cấp các thẻ có một kỳ vọng lớn hơn sẽ xuất hiện trong bộ tập các nội dung cho trẻ em (mô hình tiền sảnh) hơn trong cho nội dung văn bản cho đối tượng trưởng thành (mô hình nền). Phương trình 2.4 và 2.5 phản ánh chức năng chuyển đổi mới.

$$P_{fwKL}(i|j) = p(i)\log\frac{p(i)}{g(j)}P_{fw}(i|j) \qquad (2.4)$$

$$PbwKL(i|j) = \begin{cases} (1-a)\dfrac{P_{fwKL}(j|i)}{\sum k:(i,k)\in E\ P_{fwKL}(j|i)} & if\ i \neq j \\ \alpha & if\ i = j \end{cases} \qquad (2.5)$$

Trong đó *p(i)* là xác suất của một thẻ (hoặc *url*) để xuất hiện trong các bộ sưu tập của các nguồn tài nguyên cho trẻ em và *g(j)* là xác suất của *i* xuất hiện trong bộ tập nguồn tài nguyên chung. Kỹ thuật đã bình thường hóa khoảng cách Kullback-Leibler (KL) nằm giữa 0 và 1 trong đề xuất mô hình random walk. Việc bình thường hóa được thực hiện bằng cách sử dụng khoảng cách lớn nhất và nhỏ nhất theo từng điểm KL trong tập theo cách sau đây:

$$Kl_n(p||q) = kl(p||q) - minKL/(maxKL - minKL) \quad (2.6)$$

Ta cũng thấy rằng việc sử dụng một tiêu chuẩn thống nhất cho quá trình chuyển đổi của các url vào thẻ đã cải thiện hiệu suất của random walk. Bằng trực giác, điều này xảy ra bởi vì các tiêu chuẩn quá trình chuyển đổi của các url đến các thẻ dẫn đến xu hướng thúc đẩy độ phổ biến của thẻ (Tag) nhất, tuy nhiên tập trung của kỹ thuật là phổ biến các thẻ mà định hướng nhiều theo trẻ em, mà không nhất thiết phải là phổ biến nhất cho một url nào. Do đó, một sự thống nhất bình thường hóa làm nổi bật các trọng số KL giới thiệu trong phương trình 2.4 và 2.5. Sử dụng quan sát này, công thức bình thường hóa lại xác suất được viết như sau:

$$PfwN(i|j) = \begin{cases} (1-a)\frac{c(i|j)}{\sum k:(j,k)\in E\, c(j|k)} & if\ i \neq j, j \in T \\ (1-a)\frac{PfwKL(j|i)}{\sum i:(j,i)\in E\, P\, fw(j|i)} & if\ i \neq j, j \in U \\ \alpha & if\ i = j \end{cases} \quad (2.7)$$

Từ phương trình 2.4, chúng ta cần phải ước tính xác suất của các thẻ và url trong hai phần chính. những xác suất được ước tính dựa trên một tập hợp của trang đánh dấu Delicious đại diện cho lợi ích của các nhóm mục tiêu.

Phương pháp xác định một mục đánh dấu trang làm một bộ chứa một URL và một thẻ(Tag), trong đó mô tả các URL: $b=<b_i,t_i>$ trong đó $b_i \in B$ và $t_i \in T$, tập hợp của các url và các thẻ tương ứng. Các đánh dấu được định nghĩa như là một túi của N đánh dấu $B=\{b_1,b_2,..b_n\}$

Kỹ thuật này sử dụng một bộ các chứa chỉ mục tin cậy và url định hướng cho một đối tượng mục tiêu cụ thể (tức là trẻ em).

**Định nghĩa 2**. (Đánh dấu dành cho trẻ em) Túi đựng các đánh dấu bao gồm các url đáng tin cậy và định hướng cho một đối tượng mục tiêu được định nghĩa là [1]:

$$B_k = \{b_1, b_2, \ldots, b_N | proj_{url}(b_i) \in U_k\} \quad (2.8)$$

Trong đó $U_k$ là tập các nguồn url. Việc đánh giá xác suất chuyển đổi mô tả trong Phương trình 2.4 được đánh giá sử dụng tối đa khả năng đánh giá (MLE- Ước lượng hợp lý cực đại, gọi tắt từ Maximum-Likelihood Estimation là một kỹ thuật trong thống kê dùng để ước lượng giá trị tham số của một mô hình xác suất dựa trên những dữ liệu có được) sử dụng $B_k$ cho mô hình mặt trước (bên ngoài) và $B$ cho các mô hình nền (bên trong)

$$P(t) = \frac{CfB_k(t)}{|T|}, \; p(u) = \frac{CfB_k(u)}{|U|}$$

$$g(t) = \frac{CfB(t)}{|T|}, \; g(u) = \frac{CfB(u)}{|U|} \quad (2.9)$$

Trong đó | T | và | U | là kích thước của thẻ (Tag) và các url trong bộ sưu tập $B_k$

2.1.3.3. Biểu diễn truy vấn

Các truy vấn được biểu diễn như là một nút đơn trong đồ thị và chúng ta định nghĩa một xác suất chuyển đổi riêng từ các nút truy vấn đến các nút thẻ của đồ thị. Chúng ta không tính đến xác xuất chuyển đổi từ các truy vấn đến các nút url vì truy vấn của người dùng được biểu diễn như một túi đựng thẻ (Tag). Các truy vấn được biểu diễn là cấu tạo từ chính các truy vấn và các thẻ được tìm thấy trong các tiêu đề và trích đoạn xếp hạng đầu của kết quả tìm kiếm. Các truy vấn cũng có thể được xem như là một tài liệu cấu thành với các thẻ được tìm thấy trong các kết quả trên web và truy vấn. Chúng ta định nghĩa chính thức tập truy vấn

**Định nghĩa 3**. (Query) Một truy vấn q có chiều dài l được đại diện là chuỗi các từ (w1,w2,..wn) [1]

**Định nghĩa 4**. (tập Tag của một truy vấn) Tập Tag của một truy vấn *q* bao gồm các thẻ *m* trích ra từ một hệ thống (trang) xã hội đánh dấu *S*, trong đó có liên quan đến kết quả top đầu của web truy vấn q: $Q=\{t_1,t_2,..t_m\}$ [1]

Biểu diễn này là thuận tiện vì gợi ý truy vấn này thường có thể đạt được ngay lập tức được lấy trực tiếp từ các từ khóa xuất hiện trong các đoạn của các kết quả web. Ví dụ sử dụng 10 nghìn truy vấn từ nhật ký truy vấn AOL (AOL là viết tắt của America Online, là một công ty cung cấp dịch vụ Internet toàn cầu có trụ sở tại Hoa Kỳ) thấy rằng giao điểm giữa các từ khóa được tạo ra từ các

đoạn / tiêu đề và bảng từ vựng của các biểu diễn lại truy vấn (và cũng có mặt như các thẻ trong Delicious) là 65%. Sử dụng biểu diễn truy vấn này, chúng ta xác định các quá trình chuyển đổi xác suất p (t | Q) là:

$$P(t|Q) = \frac{p(Q|t)p(t)}{p(Q)}$$

$$P(t|Q) \propto p(t)p(Q|t)$$

$$P(t|Q) \propto p(t)\prod_{i=1}^{|Q|} p(q_i|t) \quad (2.10)$$

Vế bên tay phải là thẻ ứng viên *t* trong tập và vế thứ hai mô tả các khả năng của *t* xảy ra đồng thời giữa các thẻ trong truy vấn và tập. Những xác suất này được ước lượng sử dụng MLE trong một cấu tạo tương tự như trong 2.9.

$$p(q_i|t) = \frac{cf(q_i,t) + \mu\, p(q_i)}{|T| + \mu} \quad (2.11)$$

Trong đó p ($q_i$) là xác suất trước của $q_i$ và μ là *Dirichti* tham số làm mịn.

### 2.1.4. Nhận xét

Kỹ thuật này đẩy các thẻ trong random walk sử dụng thường xuyên hơn để mô tả các nguồn tài nguyên cho trẻ em và làm nổi bật hơn với một mô hình nền của các nguồn tài nguyên web nhằm vào các tài nguyên công cộng nói chung.

Phương pháp này tập trung thường xuyên hơn đến các liên kết URL và các thẻ (Tag) dành cho các chủ đề trẻ em, đưa ra các ứng viên tốt hơn cho trẻ em khi xây dựng truy vấn cho trẻ.

## 2.2. Kỹ thuật gợi ý truy vấn bằng so sánh truy vấn (QS)

### 2.2.1. Cách tiếp cận

Theo I. B. Vidinli và cộng sự [7], gợi ý truy vấn thường được định nghĩa là "tìm kiếm một số truy vấn liên quan tới truy vấn do người dùng phát hành ban đầu". Ví dụ, khi người dùng đặt ra truy vấn "hãng hàng không Mỹ", công cụ tìm kiếm sẽ đề nghị tìm kiếm những thuật ngữ như "vé máy bay", "vé máy bay trực tuyến", "đại lý hãng hàng không Mỹ" v.v. Theo một cách tiếp cận đơn giản và thiết thực, I. B. Vidinli và cộng sự khuyến nghị bài toán gợi ý truy vấn có thể được đơn giản hóa như sau:

Bài toán gợi ý truy vấn nên nghĩ một cách đơn giản như là "một loạt các so sánh hai câu truy vấn". Truy vấn đầu tiên trong việc so sánh là “truy vấn ban đầu” do người tìm kiếm (người sử dụng) đưa ra. Truy vấn thứ hai là "truy vấn ứng viên" được đề nghị cho người sử dụng, thường được để lựa chọn. Việc so sánh các truy vấn có thể phụ thuộc vào một số đặc trưng như câu từ tương quan, nhật ký truy vấn, vv .

Với cách tiếp cận này, bài toán so sánh câu truy vấn trong thực tế rất đơn giản và quá trình theo dõi là đơn giản, dễ mở rộng và gỡ lỗi.

Một tập các truy vấn ứng viên (đề nghị) $q_c$ được xác định cho một truy vấn ban đầu được so sánh với truy vấn ban đầu $q_i$. Cuối cùng, các truy vấn ứng viên có thể được sắp xếp dựa trên thứ hạng / điểm số và các truy vấn top $n$ ứng viên có thể được trình bày cho người dùng như một truy vấn đề nghị.

Cách tiếp cận này có những ưu điểm [7]:

- Bài toán gợi ý truy vấn rõ ràng là được giảm nhẹ tới mức "so sánh hai truy vấn", truy vấn gốc và ứng cử viên;
- Hai truy vấn có thể được so sánh với các phương pháp đơn giản;
- Có thể dễ dàng kết hợp nhiều phương pháp so sánh truy vấn;
- Rất dễ dàng theo dõi, gỡ lỗi và phát triển các phương pháp mới dựa trên cách tiếp cận này.

Với kỹ thuật này, người ta chỉ cần quan tâm đến việc so sánh hai truy vấn. Hình 2.1 là mô hình gợi ý truy vấn bằng kỹ thuật so sánh truy vấn bao gồm một số bước như sau [7]:

- Chọn / tìm các truy vấn ứng viên

- Điều khiển chung

- Sắp xếp các truy vấn ứng cử viên với một/hoặc nhiều thuật toán (bước quan trọng)

- Điều khiển cuối

Màu sắc khác nhau chỉ dẫn mức độ bước là chính hay phụ. Trong các bước trên thì giai đoạn tìm kiếm / lựa chọn ứng viên truy vấn là bước quan trọng đầu tiên của mô hình gợi ý truy vấn. Trong bước này, mục đích là để tìm ứng viên cho truy vấn đề nghị. Để lựa chọn các truy vấn ứng viên có thể được chọn

từ một tập các câu truy vấn trước hoặc không phát sinh trong các bản ghi truy vấn.


*Hình 2.1 Mô hình gợi ý truy vấn*

Tuy nhiên, sự so sánh không nhất thiết phải đề cập đến sự giống nhau hoặc mối liên hệ của hai truy vấn nhưng nó cũng có thể định lượng các khía cạnh khác nhau của các truy vấn được so sánh. Ví dụ, người ta có thể kiểm tra tính chính xác hoặc sự giống nhau của các truy vấn cho mục đích đa dạng hóa.

### 2.2.2. Nội dung phương pháp

#### 2.2.2.1. Mô hình so sánh truy vấn

Trong phần này trình bày mô hình Query suggestion (QS) đơn giản mà có thể được mở rộng bằng cách gắn vào các thuật toán QS mới. Qua thiết lập một mô hình rõ ràng, quá trình QS và các vấn đề được đơn giản hóa. Phương pháp và thuật toán khác nhau có thể gắn vào mô hình này, làm cho nó có thể kết hợp các phương pháp khác nhau để thực hiện các phép so sánh *and/or* [7].

Mô hình này bao gồm hai bước chính: *select & sort*. Một số bước tương đối đơn giản và nhỏ cũng có thể được bổ sung bao gồm trong quá trình để cải

thiện độ chính xác; vì vậy mô hình này thêm các bước *post-select* (điều khiển chung), *post-sort* (điều khiển cuối cùng).

Mô hình (như đã được thể hiện trong Hình 2.1) bao gồm các bước sau đây:

1. Chọn / tìm các truy vấn ứng viên (bước quan trọng)
2. Điều khiển chung (tùy chọn, bước tương đối nhỏ)
3. Sắp xếp các truy vấn ứng cử viên với một/hoặc nhiều thuật toán (bước quan trọng)
4. Điều khiển cuối
    a) Khái quát hóa, đa dạng hóa (tùy chọn, bước tương đối nhỏ)
    b) Sắp xếp lại, xử lý sau (tùy chọn, bước tương đối nhỏ)

Những ý tưởng cơ bản các bước thực hiện trong mô hình được mô tả như sau:

- Lựa chọn các truy vấn ứng viên có thể được thực hiện trong một bước riêng biệt, bằng cách sử dụng thuật toán duyệt theo chiều dọc, ngang của đồ thị - Depth First Search (DFS) hoặc Breadth First Search (BFS) vv. Đây là một bước hoàn toàn khác nhau và riêng biệt những bước khác. Mục đích là để "tìm, khám phá" lời gợi ý truy vấn có thể (sau đây gọi là truy vấn ứng viên đề nghị). Trong trường hợp chung nhất, tất cả các truy vấn đầu vào có thể được là ứng viên truy vấn. Nếu chúng ta có cấu hình đủ mạnh để xử lý, chúng tôi có thể sử dụng trường hợp chung này, nơi tất cả các truy vấn được coi là truy vấn ứng viên đề nghị.
- Điều khiển chung (mặc dù không phải là một bước quan trọng và cũng không bắt buộc) có thể được sử dụng để loại bỏ một số truy vấn vô ích từ các truy vấn ứng viên. Xóa truy vấn rất ngắn 1-2 ký tự, các truy vấn rất dài hoặc truy vấn gõ sai là những ví dụ điều khiển.
- Sắp xếp các truy vấn ứng viên là bước quan trọng tiếp theo. Thuật toán QS hiện có của mô hình đề cập đến hoặc bất kỳ thuật toán phân loại để sắp xếp truy vấn ứng viên có thể được sử dụng trong bước này. Kỹ thuật đã sử dụng kết hợp nhiều phương pháp sắp xếp trong mô hình này.

- Truy vấn tổng quát hoặc đa dạng hóa các thủ tục (bước không lớn, ít nhất, tại thời điểm này) có thể được áp dụng sau khi giai đoạn phân loại để tinh chỉnh các đề nghị trước khi hiển thị cho người dùng. Truy vấn tổng quát lựa chọn tổng quát hơn hình thức truy vấn ban đầu ví dụ như: đề nghị "cell structure " hay "cell" cho người dùng gửi truy vấn "mitochondria ". Đa dạng hóa kỹ thuật cũng có thể được sử dụng ở bước này để không nên được hiển thị cho người sử dụng như truy vấn đề nghị rất giống nhau. Ví dụ, truy vấn "phép nhân với số hữu tỉ", "phép nhân của số hữu tỉ" và "biểu thức phép nhân số hữu tỉ" không nên được hiển thị với nhau cho người sử dụng.
- Một trong những khía cạnh quan trọng nhất của mô hình này là để phá vỡ các vấn đề truy vấn đề nghị thành mảnh rời. Khái niệm này còn được gọi là "Tách mối quan tâm- Separation of Concerns" được ưa chuộng trong nhiều lĩnh vực. Sử dụng mô hình này, người ta có thể đóng góp cho vấn đề QS bằng cách đề xuất thuật toán mới (ví dụ, thuật toán truy vấn lựa chọn ứng viên) cho một bước cụ thể mà không cần phải xử lý các bước khác.

Phần tiếp theo cung cấp thêm thông tin chi tiết của từng bước trong mô hình này.

2.2.2.2. Pha lựa chọn

Giai đoạn lựa chọn là bước quan trọng đầu tiên của mô hình đề xuất gợi ý truy vấn này. Trong bước này, mục đích là để tìm ứng viên cho truy vấn đề nghị. Truy vấn ứng viên hoặc có thể được lựa chọn từ một tập các câu truy vấn trước hoặc không phát sinh trong các bản ghi truy vấn. Trong nghiên cứu này, chúng ta tập trung vào các phương pháp gợi ý truy vấn sử dụng các bản ghi truy vấn. Ứng viên truy vấn có thể được lấy từ các bản ghi truy vấn bằng cách duyệt qua đồ thị truy vấn bằng cách Click sử dụng DFS hoặc BFS. Trong trường hợp chung nhất, tất cả truy vấn đầu vào hoặc tất cả các truy vấn có thể được là truy vấn ứng viên, mặc dù điều này đòi hỏi bộ xử lý cao.

Thực nghiệm ban đầu của nhóm tác giả [7] cho thấy rằng các truy vấn ứng viên tìm thấy sử dụng DFS dường như đi lạc hướng khỏi chủ đề của truy vấn ban đầu. Tuy nhiên, phương pháp này được sử dụng thuật toán Hitting Time.

Mặt khác tìm kiếm theo chiều rộng (BFS) có vẻ phù hợp hơn cho việc tìm kiếm các truy vấn liên quan từ đồ thị ngược lại so với trường hợp DFS. Vì lý do

này, nhóm tác giả sử dụng và thử nghiệm với BFS như một "lựa chọn thuật toán truy vấn" và thấy nó phù hợp hơn/hữu ích cho gợi ý truy vấn, ít nhất là đăng nhập truy vấn

Sau khi loại bỏ các truy vấn ứng viên dựa trên các tiêu chí này, cuối cùng chúng ta áp dụng một ngưỡng tần số nhấp chuột mà truy vấn ứng viên với số lượng rất thấp cũng được lọc.

Tất cả các cơ chế lọc được áp dụng cho công cụ tìm kiếm chung. Tuy nhiên, mục đích cuối cùng của việc khai phá tính năng giáo dục và chỉ áp dụng cho công cụ tìm kiếm theo chiều dọc tập trung vào tài liệu giáo dục, đó là trường hợp mục tiêu của luận văn này. Bằng trực giác, sau bước này là hệ thống gợi ý truy vấn sẽ đề nghị các truy vấn liên quan đến quá trình tương tự như quá trình truy vấn ban đầu. Lưu ý rằng thông tin yêu cầu của truy vấn ban đầu và truy vấn ứng viên đã được biết hoặc dự đoán.

Danh sách các điều khiển có thể được mở rộng bằng cách thử nghiệm với mô hình/thuật toán QS hoặc dựa trên phạm vi cụ thể chúng ta đang tìm kiếm. Công việc về sau như kiểm tra chính tả và sửa chữa có thể được thực hiện trong giai đoạn này.

#### 2.2.2.3. Pha sắp xếp

Phần này mô tả các giai đoạn sắp xếp mô hình. Mục đích duy nhất của bước quan trọng này là sắp ứng viên truy vấn dựa trên một số biện pháp như sự đồng dạng với truy vấn ban đầu hoặc đồng xảy ra với các truy vấn ban đầu trong cùng một phiên truy vấn, vv. Truy vấn ứng viên có thể được sắp xếp dựa trên các khía cạnh khác nhau như vậy. Mô hình cung cấp một mô-đun cơ chế để sắp xếp khác của các truy vấn ứng viên có thể kết hợp cho độ chính xác cao hơn trong đề xuất truy vấn.

Lưu ý rằng mô hình này cho phép cơ chế mới xếp ứng viên truy vấn được sử dụng một cách đơn giản.

Chi tiết hóa bước sắp xếp trong mô hình như sau. Đặt $q_i$ biểu thị các truy vấn ban đầu cung cấp bởi người sử dụng. Trong ví dụ sau đây truy vấn ban đầu là "american airlines". Đặt $Cq_i$ là tập các truy vấn ứng viên cho $q_i$ trong một vector hình như hình dưới đây (để đơn giản, chỉ có 4 ứng cử viên được hiển thị). Giả sử có *N* ứng viên khác nhau sử dụng phương pháp có sẵn trong mô hình và mỗi ứng viên được sắp xếp bằng cách tính điểm trọng số. Chúng ta biểu thị vector điểm này là *Vj (J* từ 1 đến *N)* và hiển thị ví dụ dưới đây.

Truy vấn ứng viên : Candidate queries $Cq_i$ *được tính:*

$$Candidate\ queries(Cq_i) = \begin{bmatrix} \text{american airlines} \\ \text{airline tickets} \\ \text{airline phone} \\ \text{airline reservation} \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix}$$

$$V_1 = \begin{bmatrix} \text{Score of } q_1 \text{ wrt } q_i \\ \text{Score of } q_2 \text{ wrt } q_i \\ \text{Score of } q_3 \text{ wrt } q_i \\ \text{Score of } q_4 \text{ wrt } q_i \end{bmatrix} = \begin{bmatrix} Sq_1 \\ Sq_2 \\ Sq_3 \\ Sq_4 \end{bmatrix}$$

Mô hình này kết hợp thuật toán sắp xếp ứng viên xếp hàng khác nhau. Điều này có thể được thực hiện bằng bất kỳ phương pháp kết hợp nào. Việc tổng hợp thuật toán sắp xếp có thể thấy giống như sự kết hợp của kết quả công cụ tìm kiếm trong một công cụ tìm kiếm siêu dữ liệu). Kỹ thuật này đã cố gắng để cải thiện hiệu suất truy vấn đề nghị bằng cách kết hợp nhiều thuật toán sắp xếp. Sau đây luận văn trình bày **Phương pháp ghép (Aggregation methods).**

Phương pháp ghép có thể sắp xếp được ít nhất trong hai loại; phương pháp dựa trên điểm và dựa trên thứ hạng [7]. Phương pháp tiếp cận khác cũng có thể được đề nghị.

*1) Dựa trên điểm*

Chúng ta cho rằng, tổng hợp các thuật toán sắp *sorting/ordering* khác nhau có thể được thực hiện bằng tổng trọng số của vectơ điểm, như trong các công thức sau đây [7]:

Final score vector = $k_1$.Norm($V_1$) + $k_2$.Norm($V_2$) +…+ $k_n$.Norm($V_n$)

$$=\sum_{i=1}^{number\ of\ \ algorithms} k_i.\,\text{Norm}(V_i) \qquad (2.12)$$

Trong đó, hàm chuẩn hóa Norm (V) được tính toán như sau:

$$Norm(V) = \left\{\frac{x}{max(V)} \middle| \forall x\ \in V\right\} \qquad (2.13)$$

Mỗi bậc trong công thức (2.12) được tính trọng số từ $k_1$ tới $k_n$

Lưu ý rằng mỗi điểm vector $V_j$ có thể có điểm trong dao động khác nhau vì vậy chúng cần phải được bình thường hóa lại (tỉ lệ lại vào phạm vi 0-1) trước khi kết hợp. Cuối cùng, thu được một vector điểm số.

Bằng cách tỉ lệ lại các vecto, *điểm số cuối cùng là không quá bị ảnh hưởng bởi một thuật toán duy nhất* (thậm chí nếu nó có cao/sai giá trị); thay vào đó, thuật toán đơn ảnh hưởng đến điểm số cuối cùng lên đến một mức hợp lý nhất định. Điều này cũng rất hữu ích để ngăn chặn sai sót, sai lầm hoặc ảnh hưởng do một số thuật toán. Các thuật toán khác nhau đạt được tỷ lệ thành công khác nhau trên các truy vấn đầu vào khác nhau. Cũng giống như trường hợp tìm kiếm siêu dữ liệu, phương pháp đầu vào có thể hoặc không cùng đạt được tốc độ, do đó, sự kết hợp có trọng số của thuật toán có thể hiển thị hiệu suất tốt hơn.

Tiếp theo, một thuật toán sắp xếp ứng viên (đặc biệt là những ứng viên phụ thuộc vào tần số truy vấn, số kết quả) được sử dụng nhằm tạo một vài điểm số rất cao với vô số điểm số rất thấp. Công thức (2.14) tính vector điểm cuối cùng liên quan tới lịch sử chuyển đổi như biểu diễn dưới đây.

$$\text{Final score vector} = k_1.\text{Norm}(\log(V_1)) + k_2.\text{Norm}(\log(V_2)) + \ldots + k_n.\text{Norm}(\log(V_n)) = \sum_{i=1}^{\text{number of algorithms}} k_i.\,\text{Norm}(V_i) \quad (2.14)$$

Các hệ số $k_1, k_2, \ldots$ được sử dụng để tăng cấp các thuật toán được biết hoặc quan sát để có hiệu quả tốt hơn trong kết quả cuối cùng.

Đối với mục đích thử nghiệm, hệ số của một số thuật toán có thể được thực hiện bằng không. Nó cũng có thể hữu ích để thử các hệ số khác nhau (điều chỉnh tham số) để kiểm tra tổng hiệu suất QS. Đây là loại tính toán mô-đun an toàn và cho phép chúng ta sử dụng các thuật toán sắp xếp khác nhau được phát triển bởi các nhà nghiên cứu *and / or của* các nhà phát triển khác nhau.

Các tác giả [7] cải tiến công thức tính toán vector điểm như sau*:*

$$\textit{Finar score vector for 1st group of QS} = FSV_1 = \sum_{i=1}^{number\ of\ \ algorithms} k_{1i}.\,\text{Norm}(V_i)$$

$$\textit{Finar score vector for 2nd group of QS} = FSV_2 = \sum_{i=1}^{number\ of\ \ algorithms} k_{2i}.\,\text{Norm}(V_i) \quad (2.15)$$

Trong đó $FSV_1$ sẽ được sử dụng để lựa chọn 8-10 gợi ý truy vấn đầu tiên và $FSV_2$ sẽ được sử dụng để lựa chọn còn lại 2-6 đề nghị (tùy thuộc vào số lượng

đề xuất truy vấn cuối cùng, có thể là một tổng số 10 hoặc 15). Nếu có truy vấn đề nghị chồng chéo, chúng phải được thống nhất trong tập kết quả cuối cùng. Điều này có thể được xem như là một phương pháp kết hợp của hai phương pháp lai khác nhau, nơi chúng có –hệ số *k* khác nhau cho mỗi thuật toán. Trong bài báo này, thử nghiệm với các phiên bản trước đó của công thức và để lại sau này như một công việc tương lai.

Borda count ranking of candidates for a method

| Ranking | Candidates | Formula | points |
|---|---|---|---|
| 1st | Query 1 | n | 5 |
| 2nd | Query 2 | n-1 | 4 |
| 3rd | Query 3 | n-2 | 3 |
| 4th | Query 4 | n-3 | 2 |
| 5th | Query 5 | n-4 | 1 |

*Bảng 2.1 Sắp xếp số truy vấn ứng viên*

Sample query suggestions for rank aggregatinon

| Rank | Query suggestion | Scores (using some method) |
|---|---|---|
| 1 | American airline | 0.9 |
| 2 | Airline tickets | 0.2 |
| 3 | Airline bus | 0.1 |

*Bảng 2.2 Sắp xếp số gợi ý truy vấn*

*2) Dựa trên thứ hạng*

Trong dòng phương pháp tập hợp này, các thứ hạng của một tập kết quả của một thuật toán (trong trường hợp của chúng ta là ứng viên gợi ý truy vấn) được sử dụng trọng số để xếp tổng thể, khi nhiều thuật toán được kết hợp [7].

*Phương pháp Borda Count.* Phương pháp Borda Count kết hợp kết quả của các phương pháp khác nhau bằng cách tính mỗi mục tương ứng với vị trí của mục đó trong kết quả thiết lập của từng phương pháp con. Các công thức và các điểm được cung cấp bởi một phương pháp con được minh họa trong Bảng

1 . Các điểm cho mỗi phương pháp được tính toán sử dụng ở bảng này, sau đó tất cả các điểm truy vấn ứng viên được tổng lại và xếp hạng cuối cùng là thu được bằng cách sắp xếp truy vấn ứng viên với điểm trung bình giảm. Phương pháp này được ghi nhận như là lai-2-Borda Count hoặc tương tự.

*Trọng số phương thức Borda Count (weighted borda fuse).* Đây là một phiên bản sửa đổi nhỏ của phương pháp Borda Count, trong đó mỗi phương pháp là trọng số riêng bởi một hệ số). Chúng ta cũng có thể thử nghiệm phương pháp này để đo lường hiệu quả của mỗi từ mỗi phương pháp con. Phương pháp này được ghi nhận như là lai-3-Weighted Borda Count hoặc tương tự.

*Phương thức trọng số biểu quyết* . Đây là một phương pháp khá đơn giản mà một truy vấn ứng cử viên được cho một điểm/phiếu (tương đương với trọng lượng của phương thức con) khi nó tồn tại trong các kết quả top n của một phương pháp con. Ví dụ, nếu một truy vấn tồn tại trong kết quả *top-n* 4 phương pháp khác nhau, nó được đưa ra 4 điểm (giả định cùng một trọng số). Chúng ta sử dụng trên 30 kết quả của phương pháp con để thực hiện phương pháp thứ hạng tập hợp đơn giản này.

*3) Lựa chọn các điểm và xếp hạng tập hợp các phương pháp.*

Mục đích của việc này không phải là để so sánh /đánh giá thứ hạng phương pháp tập hợp. Nhưng chúng ta cố gắng để lựa chọn và sử dụng thích hợp nhất trong mô hình đối với các lĩnh vực gợi ý truy vấn.

Ta có thể rút ra kinh nghiệm việc thực hiện điểm số dựa vào tập thứ hạng trong các thí nghiệm ban đầu. Ưu điểm của điểm số dựa trên tập hợp cũng được ghi nhận trong tài liệu, Biểu mẫu của điểm số đầu ra được xếp hạng dựa trong thiết lập cài đặt tìm kiếm siêu dữ liệu.

Ta có thể đề xuất số điểm dựa trên tập hợp hơn là thứ hạng dựa trên tập hợp. Phương pháp của chúng ta sử dụng "điểm" của các danh sách được sắp xếp thay vì vị trí xếp hạng là tốt hơn, nơi điểm sẵn có hoặc do tính toán.

Sử dụng "thứ hạng, vị trí của đề xuất" có thể được áp dụng thành công khi không có "điểm" của các gợi ý truy vấn. Nếu có những điểm số, nó sẽ là tốt hơn để sử dụng chúng, như chúng sẽ đạt được tính nhạy cảm hơn, đặc biệt là khi chúng ta cố gắng kết hợp kết quả của các phương pháp khác nhau. Điểm truy vấn cho thấy có thể được nghĩ như là một chiều của thông tin.

Hãy xét xem các danh sách xếp hạng như trong Bảng 2. Nếu chúng ta chỉ dùng thứ hạng này để đề nghị, các giá trị thứ hạng của mục thứ hai gần đóng lại đến mục đầu tiên. Tuy nhiên, nếu chúng ta có thể sử dụng điểm số (nếu chúng có sẵn hoặc tính toán), thì giá trị xắp xếp của mục thứ hai là thấp hơn nhiều so với mục đầu tiên. Tình huống này không có ảnh hưởng khi chúng ta chỉ có một phương pháp, kể từ mục thứ 2 là không thay đổi cho dù chúng ta dùng thứ hạng hay điểm; Tuy vậy, nếu chúng ta có nhiều hơn một phương pháp để kết hợp / tổng hợp (đặc biệt là khi chúng ta có nhiều phương pháp / tính năng) nên giá trị điểm số đề xuất truy vấn sẽ trở nên rất quan trọng, rất hiệu quả trong việc sắp xếp chính thức. Đó là lý do tại sao chúng ta giới thiệu ý tưởng "So sánh các truy vấn" và sử dụng điểm số để gợi ý. Sử dụng kỹ thuật này chúng ta đạt được tỷ lệ thành công cao hơn khi sử dụng điểm số trên thứ hạng tập hợp.

#### 2.2.3. Nhận xét

Kỹ thuật này xác định lại và làm giảm các vấn đề trong "Query Suggestion (QS)"

Kỹ thuật này đề xuất một module, mở rộng mô hình đề xuất truy vấn để các phương pháp mới với nhiều thuật toán QS dễ dàng đưa vào.

Kỹ thuật này đánh giá hiệu năng của dữ liệu Click dựa trên kỹ thuật QS đề xuất cho mục đích chung công cụ tìm kiếm tài liệu, trên nhật ký công cụ tìm kiếm giáo dục thực tế.

Kỹ thuật này đề xuất thuật toán QS mới khai phá các tính năng truy vấn chung (truy vấn, phiên làm việc, tính năng người dùng) và công cụ tìm kiếm giáo dục (thuộc tính trường, lớp).

Chúng ta cũng đề xuất các thuật toán lai ghép cho phép kết hợp một số kỹ thuật QS cho hiệu quả cao hơn. Các thuật toán này được tích hợp trong mô hình đề cập ở trên.

### 2.3. Tính tương tự của truy vấn

#### 2.3.1. Cách tiếp cận

Để đưa ra được các truy vấn ứng viên, các gợi ý truy vấn cho truy vấn ban đầu, bài toán tính độ tương tự giữa các truy vấn (query similarity) được đưa ra để giải quyết vấn đề này.

Khi sử dụng hệ thống tìm kiếm, người dùng sẽ nhập vào câu truy vấn và yêu cầu máy tìm kiếm trả về tập các tài liệu liên quan. Tuy nhiên, máy tìm kiếm

thông thường dựa vào các từ ngữ của truy vấn mà trả về các tài liệu với nội dung khác nhau, cụ thể là:

- Máy tìm kiếm hiển thị kết quả với nội dung liên quan tới chính xác các từ ngữ thuộc truy vấn. Ví dụ: nếu ta đưa vào truy vấn "Học toán trực tuyến" vào máy tìm kiếm thì các kết quả sẽ hiển thị ra các trang web có chứa chính xác cụm từ "Học toán trực tuyến" hoặc có từ "học toán" "trực tuyến" "học" "toán" sẽ được hiển thị.

- Máy tìm kiếm hiển thị kết quả với nội dung các từ ngữ liên quan đến truy vấn và các từ đồng nghĩa với truy vấn. Ví dụ: nếu ta đưa vào máy tìm kiếm câu truy vấn "decease" thì máy tìm kiếm có thể đưa ra được các kết quả liên quan đến từ khóa "decease" hoặc từ "die", "death", "demise", "dying", "fate" … là các từ đồng nghĩa của "decease" (nghĩa là "chết" trong tiếng Việt)

- Máy tìm kiếm hiển thị các kết quả có liên quan đến các lĩnh vực khác nhau liên quan đến truy vấn. Ví dụ: Người dùng đưa vào từ khóa Apple thì máy tìm kiếm sẽ hiển thị các tài liệu liên quan đến máy tính apple hoặc apple fruit.

Để máy tìm kiếm có thể hiển thị kết quả phù hợp với mục đích người dùng, cần tìm ra các câu truy vấn mới mà theo người dùng những câu truy vấn này có cùng ý nghĩa (tương tự) với câu truy vấn hiện thời để máy tìm kiếm có thể tự động viết lại truy vấn của người dùng, tiến hành tìm kiếm và đưa ra được kết quả tốt hơn. Đấy là nội dung của bài toán tính độ tương tự câu truy vấn.

Ví dụ: Người dùng đưa vào truy vấn: Lê Hồng Phong thì người ta cũng muốn có những kết quả liên quan đến Hà Huy Tập hoặc Tổng bí thư giai đoạn 1935-1936. Như vậy, máy tìm kiếm cần viết lại truy vấn Lê Hồng Phong thành Tổng bí thư Lê Hồng Phong, Hà Huy Tập.

### 2.3.2. Các phương pháp tính độ tương tự

#### 2.3.2.1. Tính độ tương tự dựa trên từ vựng

Để tính độ tương tự giữa hai truy vấn dựa trên từ vựng, người ta sử dụng phương pháp biểu diễn truy vấn đơn giản nhất là dựa trên chính những từ ngữ nội tại của truy vấn – *"surface representation"*.

Chúng ta xác định một số tiêu chuẩn sau để tính toán tính phù hợp giữa các truy vấn [16]:

- *Chính xác*: *Q* và *S* là hai tập hoàn toàn tương đương.

- Ví dụ: hai câu truy vấn
  - *q = "Thanh là học sinh giỏi"*
  - *s = "Thanh là học sinh giỏi"*

- *Cụm từ: S* là một phần trong *Q*.
  - Ví dụ: Hai câu truy vấn
    - *q = "Thanh là học sinh giỏi"*
    - *s = "học sinh giỏi"*

- *Tập con:* các từ nằm trong *S* nằm hoàn toàn trong *Q* nhưng sắp xếp không đồng nhất.
  - Ví dụ: Hai câu truy vấn
    - *Q = "Thanh là học sinh giỏi"*
    - *S = "Thanh học giỏi"*

Độ tương tự giữa hai câu truy vấn q và s có thể tính được bằng một trong các công thức sau [17]:

- ***Độ đo kết hợp***

$$sim(q,s) = |Q \cap S| \quad (2.16)$$

- ***Độ đo Dice***

$$sim(q,s) = 2 \times \frac{|Q \cap S|}{|Q| + |S|} \quad (2.17)$$

- ***Độ đo Jaccard***

$$sim(q,s) = \frac{|Q \cap S|}{|Q \cup S|} \quad (2.18)$$

- ***Độ đo Overlap***

$$sim(q,s) = \frac{|Q \cap S|}{\min(|Q|, |S|} \quad (2.19)$$

- ***Độ đo Cosin***

$$sim(q,s) = \frac{|Q \cap S|}{\sqrt{|Q| \times |S|}} \quad (2.20)$$

Cách tính độ tương tự giữa các câu truy vấn theo phương pháp này đưa ra kết quả là một số từ 0 đến 1. Hai câu truy vấn được coi là tương tự hoặc không tương tự chỉ dựa trên việc tính toán xem chúng có chung từ hoặc cụm từ hay không.

**Ưu, nhược điểm**

- **Ưu điểm**
  - Cách biểu diễn truy vấn đơn giản
  - Tính toán độ tương tự giữa các truy vấn đơn giản
  - Độ chính xác cao
- **Nhược điểm**
  - Độ hồi tưởng thấp

2.3.2.2. Tính độ tương tự dựa trên nhật ký truy vấn

Nhật ký người sử dụng (Userlog) là những dữ liệu được lưu lại khi người dùng truy vấn trên máy tìm kiếm và lựa chọn các kết quả mà máy tìm kiếm trả về. Việc tính độ tương tự dựa trên userlog chỉ đặc trưng cho tính độ tương tự giữa các câu truy vấn.

Lịch sử truy vấn – query logs là những truy vấn tự nhiên, là hoạt động trực tiếp của người dùng, mô phỏng những nhu cầu thực tế của họ. Tất nhiên, với một lượng nhỏ dữ liệu các trang web mà người dùng chọn mở với mỗi câu truy vấn thì không thể đưa ra được kết quả tính độ tương tự giữa các câu truy vấn đó với độ chính xác cao. Tuy nhiên, với một lượng lớn dữ liệu được ghi lại từ máy chủ của một máy tìm kiếm lớn, có độ tin cậy cao thì việc tính toán độ tương tự truy vấn dựa vào kết quả mà người dùng chọn mở các văn bản với mỗi câu truy vấn là có thể tin tưởng được [18]. Google là một máy tìm kiếm có thể nói là lớn nhất, phổ biến nhất tại Việt Nam cũng như trên thế giới. Nó được tín nhiệm sử dụng do tính tin cậy của các kết quả trả về cũng như các trang web mà người dùng lựa chọn mở sau khi người dùng đưa vào máy tìm kiếm một câu truy vấn. Vì vậy, luận văn sử dụng dữ liệu kết quả người dùng chọn mở các văn bản

khi tiến hành truy vấn trên máy tìm kiếm Google, nói cách khác, luận văn sử dụng dữ liệu lưu vết truy vấn của máy tìm kiếm.

Chúng ta xét mối liên kết giữa truy vấn của người dùng trên máy tìm kiếm và những trang web mà người dùng lựa chọn mở [18]. Có các phương thức tính toán

- Với hai câu truy vấn khác nhau mà người dùng chọn mở cùng một liên kết URL thì hai câu truy vấn này là gần nhau – nghĩa là có độ tương tự cao.

- Nếu một tập các URL thường được chọn cho cùng một truy vấn thì nội dung các từ ngữ trong tài liệu có liên quan đến các từ ngữ trong truy vấn.

- Ngoài công thức tính độ tương tự chỉ dựa vào các liên kết chính xác chung của tài liệu khi tiến hành tìm kiếm với các câu truy vấn, người ta còn sử dụng một số thông tin khác như dựa vào các kết quả trùng nhau – tức là các người dùng khác nhau cùng đưa vào một câu truy vấn và cùng chọn mở các tài liệu giống nhau hoặc người ta cũng có thể sử dụng các miền domain chung giữa hai liên kết tài liệu để tạo ra mối liên kết giữa hai câu truy vấn mà người dùng lựa chọn liên kết là tương tác để làm tăng độ chính xác của tính tương tự giữa hai truy vấn [17]

Phương pháp tính độ tương tự cho các câu truy vấn bằng userlog với phương thức 1 đã liệt kê phía trên được luận văn sử dụng.

Kết quả được trả về khi lưu lịch sử truy vấn của người dùng đặt tại máy tìm kiếm khác nhau thường có cấu trúc khác nhau, đôi khi nó còn chứa tiêu đề, tóm tắt hay thư mục mà tài liệu thuộc về, ... tuy nhiên ta sẽ đưa chúng về dạng như sau:

$$session :=< query\ text > [clicked\ \ document]^{*} \qquad (2.21)$$

Trong đó:

- *Session:* lượt truy vấn của người dùng
- *Query text:* câu truy vấn được biểu diễn dưới dạng văn bản

- *Clicked URL:* các tài liệu được người dùng chọn mở

Đặt $U(Q_j)$ biểu diễn tập các các liên kết được người dùng lựa chọn khi thực hiện tìm kiếm với truy vấn $Q_j$

$$U(Q_j) = \{u_1, u_2, ..., u_i\} \qquad (2.22)$$

Trong đó:

- $u_i$: liên kết tài liệu thứ i mà người dùng lựa chọn khi thực hiện truy vấn $Q_j$

Đặt $R_{ij}$ là tập các liên kết mà người dùng lựa chọn trùng nhau khi đưa vào hai truy vấn $Q_i$ và $Q_j$, ta định nghĩa

$$R_{ij} = \{u : u \in (U(Q_i) \cap U(Q_j))\} \qquad (2.23)$$

Trong đó:

- $u$: liên kết thuộc cả hai tài tập liên kết mà người dùng lựa chọn khi đưa vào máy tìm kiếm hai truy vấn $Q_i$ và $Q_j$

Theo [17] ta có định nghĩa

Định nghĩa: Một truy vấn $Q_i$ là gần với truy vấn $Q_j$ nếu $N(R_{ij})>0$. Trong đó $N(R_{ij})$ là lựu lượng của tập $R_{ij}$.

Công thức tính độ tương tự giữa hai câu truy vấn dựa vào lưu vết truy vấn của máy tìm kiếm của máy tìm kiếm có thể được định nghĩa [17]

$$sim - result(Q_i, Q_j) = \frac{N(R_{ij})}{Max(N(U(Q_i)), N(U(Qj)))} \qquad (2.24)$$

Trong đó:

*sim-result($Q_i$, $Q_j$):* Độ tương tự giữa hai câu truy vấn $Q_i$ *và* $Q_j$

*N(U($Q_i$)):* Số lượng liên kết tài liệu được người dùng nhấn vào khi thực hiện tìm kiếm với truy vấn $Q_i$

## 2.4. Ý tưởng giải pháp gợi ý truy vấn cho thanh thiếu niên

Như trong phần 1.3 đã phân tích việc lựa chọn các ứng viên là bước quan trọng nhất trong các mô hình đề xuất truy vấn khi sử dụng phương pháp so sánh truy vấn (QS). Khi cần chọn một ứng viên truy vấn nào đó ta sẽ sử dụng kỹ thuật so sánh truy vấn để lựa trọn. Trong kỹ thuật so sánh ta sẽ sử dụng phương pháp

tính độ tương tự giữa các truy vấn (query similarity). Một phương pháp đo độ tương tự giữa hai truy vấn có độ chính xác cao rất hữu ích cho các ứng dụng giúp hỗ trợ người dùng trong việc tìm kiếm cũng như giúp máy tìm kiếm đưa ra được những câu trả lời đúng với mục đích người hỏi hơn. Nên trong luận văn này tôi chọn phương pháp tính độ tương tự của truy vấn để áp dụng cho mô hình gợi ý truy vấn cho đối tượng thanh thiếu niên.

# Chương 3. MỘT MÔ HÌNH GỢI Ý TRUY VẤN CHO THANH THIẾU NIÊN

## 3.1. Giới thiệu

Như đã trình bày ở các chương trước, tính độ tương tự cho truy vấn là một trong những bài toán khó. Do đặc trưng của truy vấn thường ngắn và mang ý chủ quan của con người nên việc tính toán độ tương tự giữa các câu truy vấn chưa đạt được kết quả cao khi sử dụng các phương pháp tính độ tương tự văn bản truyền thống.

Userlog là những dữ liệu về lịch sử truy vấn của người dùng. Nó là những ví dụ thực tiễn nhất của quá trình người dùng thực hiện đưa truy vấn vào máy tìm kiếm và lựa chọn các tài liệu mà người dùng thấy phù hợp nhất. Vì vậy, userlog chính là nguồn dữ liệu rất có giá trị để so sánh, tính toán sự tương tự nhau giữa các truy vấn mà người dùng đưa vào dựa vào kết quả lựa chọn văn bản của người dùng. Tất nhiên, ta khó có thể tin vào kết quả thống kê của một lượng nhỏ dữ liệu lịch sử truy vấn nhưng với một lượng lớn userlog được sưu tập từ một máy tìm kiếm có uy tín như Google hay Yahoo, ... thì đây sẽ là khối dữ liệu mang lại kết quả tính độ tương tự truy vấn rất có hiệu quả. Thực tế, đã có rất nhiều bài báo sử dụng lưu vết truy vấn của máy tìm kiếm – userlog để tính toán độ tương tự giữa các câu truy vấn [16, 19, 21].

Từ bộ userlog được thu thập từ máy chủ của máy tìm kiếm Google, luận văn sẽ trình bày một phương pháp tính độ tương tự giữa các câu truy vấn dựa vào phương pháp sử dụng lưu vết truy vấn của máy tìm kiếm [16] và đề xuất mô hình tính toán như trình bày ở mục sau.

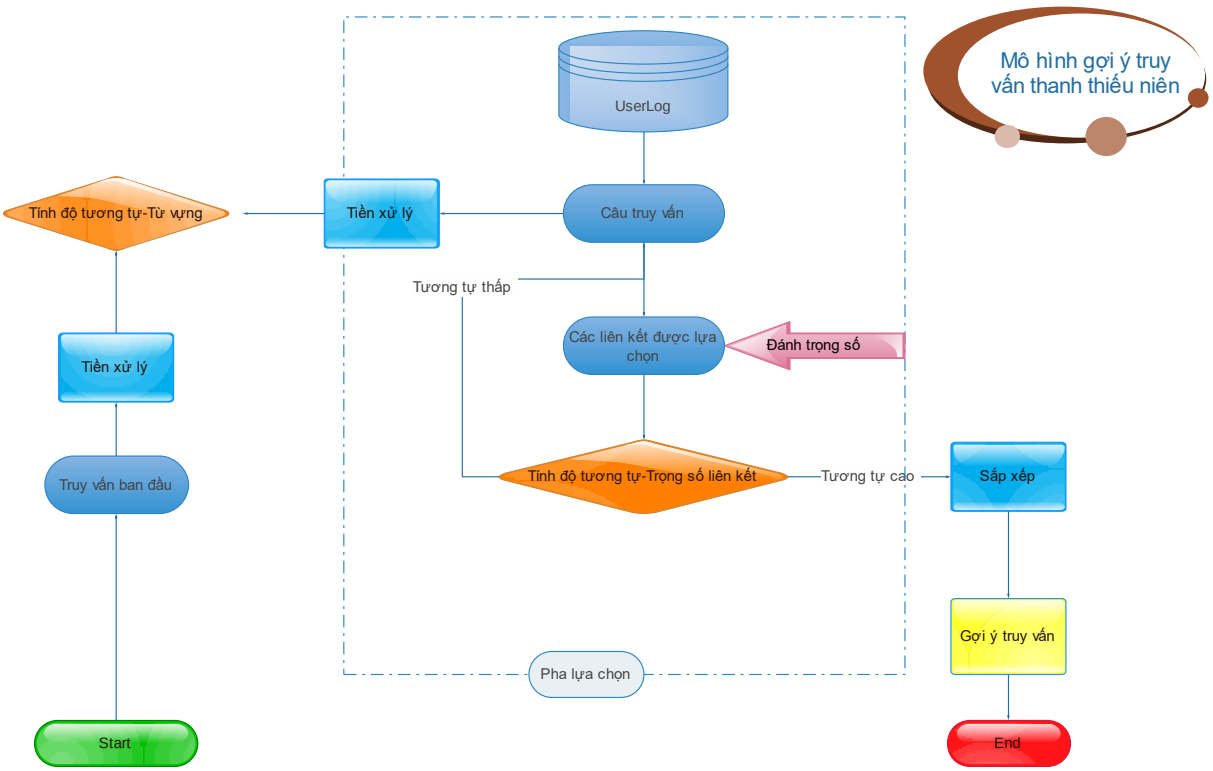
## 3.2. Mô hình

QueryLog là tập các bản ghi bao gồm các trường thông tin sau:

- SessionID: mã của phiên làm việc
- TimeStamp: nhãn thời gian ghi lại thời điểm xảy ra sự kiện
- Query: câu truy vấn của người dùng
- TopN: N tài liệu đầu tiên trong tập kết quả
- UrlClicked: liên kết được nhấn bởi người sử dụng
- QuerySegmented: các từ khóa trong câu truy vấn

Mô hình chú ý sử dụng hai thành phần có ý nghĩa là *câu truy vấn ban đầu* và *các liên kết được người dùng chọn mở* để sử dụng, tính độ tương tự giữa các

câu truy vấn. Mô hình này cải tiến từ mô hình trong luận văn [26]. Điểm mới của mô hình này là tính độ tương tự linh hoạt. Khi nào thì dùng cách tính độ tương tự theo từ vựng, khi nào thì dùng cách tính độ tương tự theo trọng số. Tức là phân rõ ra phần nào tính độ tương tự nào nhằm cải thiện hiệu năng hệ thống so với mô hình trước đây. Mô hình được thể hiện như sau:




*Hình 3.1 Mô hình đề xuất so sánh truy vấn dựa vào tính độ tương tự của các câu truy vấn*

## 3.3. Các thành phần của mô hình

**Các bước thực hiện mô hình:**

- *Bước 1:* Tiền xử lý câu truy vấn. Câu truy vấn đầu vào được tiền xử lý
  - Xử lý tiếng Việt, định dạng từ mã code thành dấu tiếng Việt
  - Liệt kê các liên kết tương ứng với cùng một câu truy vấn
- *Bước 2:* Lấy danh sách liên kết được chọn mở có cùng nội dung truy vấn.

- *Bước 3:* Đánh trọng số cho liên kết tương ứng với từng truy vấn
    - Đặt $\{x_1, x_2, x_3, ..., x_n\}$ là tập trọng số.
        - Trong đó, $x_i$ là trọng số của liên kết thứ *i* mà người dùng lựa chọn mở cho truy vấn.
        - $\{x_1, x_2, x_3, ..., x_n\}$ là bộ trọng số chung cho các link của tất cả các truy vấn trong bộ dữ liệu.
- *Bước 4:* Tính độ tương tự
    - So sánh các liên kết được lựa chọn, sử dụng 2 nguyên lý:
        - Sử dụng *"surface representation"* để so 2 câu truy vấn. Câu nào có độ chính xác cao nhất sẽ được chọn để tìm độ liên kết câu truy vấn đó so sánh ở nguyên lý 2
        - Những truy vấn nào được người dùng chọn mở cùng liên kết thì các truy vấn ấy tương tự nhau.
    - Sử dụng công thức

$$sim - result(Q_i, Q_j) = \frac{\sum (x_i + x_j)}{Max(N(U(Q_i)), N(U(Qj)))} \quad (3.1)$$

Trong đó:

- ✓ $Q_i$ và $Q_j$ là hai câu truy vấn
- ✓ *Sim-result*$(Q_i,Q_j)$: độ tương tự truy vấn giữa hai câu truy vấn $Q_i$ và $Q_j$
- ✓ $x_i$, $x_j$ trọng số của các link chung tương ứng thuộc hai câu truy vấn $Q_i$ và $Q_j$
- ✓ $N(U(Q_i))$: Số lượng liên kết được lựa chọn tương ứng với câu truy vấn $Q_i$

- ✓ *$N(U(Q_j))$:* Số lượng liên kết được lựa chọn tương ứng với câu truy vấn $Q_j$

Từ cơ sở lý thuyết đã trình bày ở các chương trước, luận văn tiến hành thực nghiệm tính độ tương tự của các câu truy vấn dựa trên phương pháp sử dụng lưu vết truy vấn của máy tìm kiếm đồng thời cũng đề xuất một số cải tiến trên phương pháp đã chọn.

Luận văn sử dụng dữ liệu lưu vết truy vấn tìm kiếm trên máy tìm kiếm Google, tiến hành cài đặt chương trình tính độ tương tự giữa các câu truy vấn theo mô hình đã đề xuất ở chương 3, gồm các chức năng chính: xử lý dữ liệu, tính độ tương tự truy vấn theo công thức (2.19) với cải tiến về việc đưa thêm giá trị trọng số cho các liên kết web, sau đó lựa chọn một số truy vấn điển hình để tiến hành đánh giá. Nội dung thực nghiệm được trình bày dưới đây.

# Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

## 4.1. Đặt vấn đề

Do hạn chế về mặt thời gian, việc thực nghiệm cả một mô hình gợi ý là rất phức tạp, nên trong luận văn sẽ thực nghiệm một phần trong luận văn đó là tính tương tự của 10 cặp truy vấn. Sau đó dùng công cụ tìm kiếm google tiếng Việt để chạy thử nghiệm truy vấn.

## 4.2. Thi hành mô hình (Phần mềm và phần cứng)

Cấu hình phần cứng

- CPU: Intel core 2 Duo T8300
- Cache: 2.4Ghz
- Ram: 4G
- Hệ điều hành: Window 7
- Bộ nhớ ngoài: 250Gb

Công cụ phần mềm sử dụng

- Visual Studio 2013
- Môi trường Net Framwork 4.1
- Ngôn ngữ lập trình C#

Phần mềm bài toán bao gồm các thành phần

- Chương trình xử lý gồm các mô đun: Tiền xử lý dữ liệu (loại bỏ truy vấn quá dài, quá ngắn, tập con của truy vấn ban đầu...);
- Mô đun tính toán độ tương tự giữa các câu truy vấn: *Similarity*, dùng để tính độ tương tự theo từ vựng.
- Mô đun hiển thị kết quả danh sách gợi ý truy vấn cho người dùng.

## 4.3. Dữ liệu và quá trình thực nghiệm

### 4.3.1. Dữ liệu

Dữ liệu này bao gồm Câu truy vấn thử nghiệm được kết hợp với việc sử dụng bộ userlog của một công cụ tìm kiếm trên trang CocCoc.vn là số lượt truy vấn của người dùng tại một số trường Trung học phổ thông ở Vệt Nam.

### 4.3.2. Quá trình thực hiện

❖ Tiền xử lý:

- Xóa bỏ các câu truy vấn không lành mạnh
- Lược bỏ các câu truy vấn với lựa chọn liên kết trùng nhau
- Định dạng lại dữ liệu về dạng: *"truy vấn" link1 link2 ...*

  → Kết quả: Được danh sách câu truy vấn với liên kết được lựa chọn tương ứng.

❖ Tính độ tương tự

Sử dụng mô hình tính độ tương tự đã trình bày ở trên, việc tính độ tương tự có hai công đoạn là tính độ tương tự của từ vựng đề tìm câu truy vấn trong lịch sử truy vấn có độ tương tự cao nhất. Sau đó sẽ lấy trọng số liên kết của câu truy vấn đó để so sánh với danh sách các trọng số của các câu truy vấn còn lại. Kết quả đạt được như sau

Sau quá trình thực nghiệm, luận văn thực hiện tính toán với những bộ trọng số $\{x_1, x_2, \ldots, x_n\}$ khác nhau thấy bộ trọng số {0.9, 0.85, 1, 1.05, 1, ..., 1} đạt kết quả tốt nhất. Kết quả thu được khi lấy ra 10 kết quả đầu tiên với ngưỡng đưa ra là 0.39.

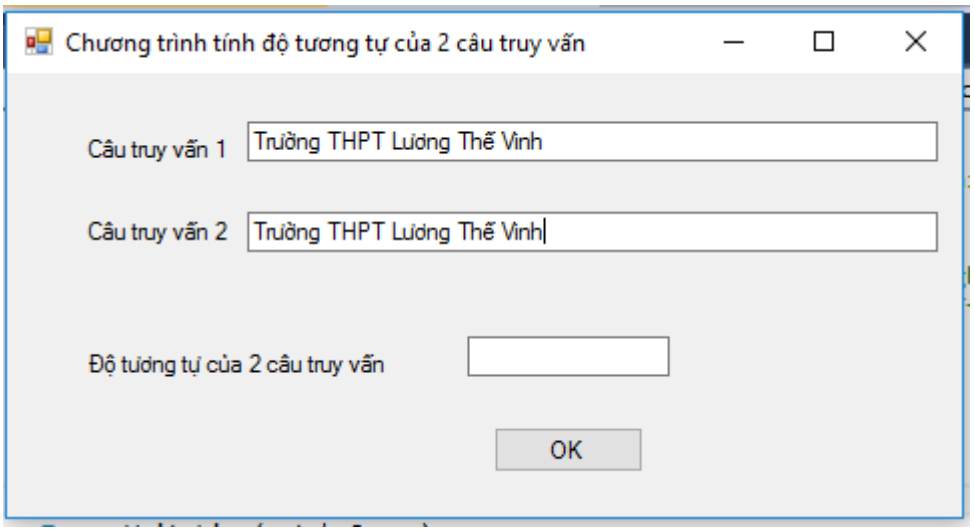
| STT | Truy vấn \| \| truy vấn | Độ tương tự |
|---|---|---|
| 1 | LTV \| \| Trường THPT Lương Thế Vinh | 0.95 |
| 2 | Toán \| \| Giải toán trên mạng | 0.85 |
| 3 | Nghe nhạc \| \| nhạc online | 0.9 |
| 4 | thi thpt 2016 \| \| Kỳ thi THPT Quốc gia năm 2016 | 0.9 |
| 5 | nghe nhạc online \| \|mp3 trực tuyến | 0.70000066 |
| 6 | Truyện tranh đẹp \| \| hình ảnh đẹp nhất | 0.38249998 |
| 7 | cách học văn hay \| \| nguyễn ngọc ngạn | 0.49249998 |
| 8 | tro choi trang diem \| \| game vui thoi trang | 0.41249998 |
| 9 | tro choi trang diem \| \| tro choi mien phi | 0.44249998 |
| 10 | Hoa học trò \| \| báo thanh niên | 0.39 |

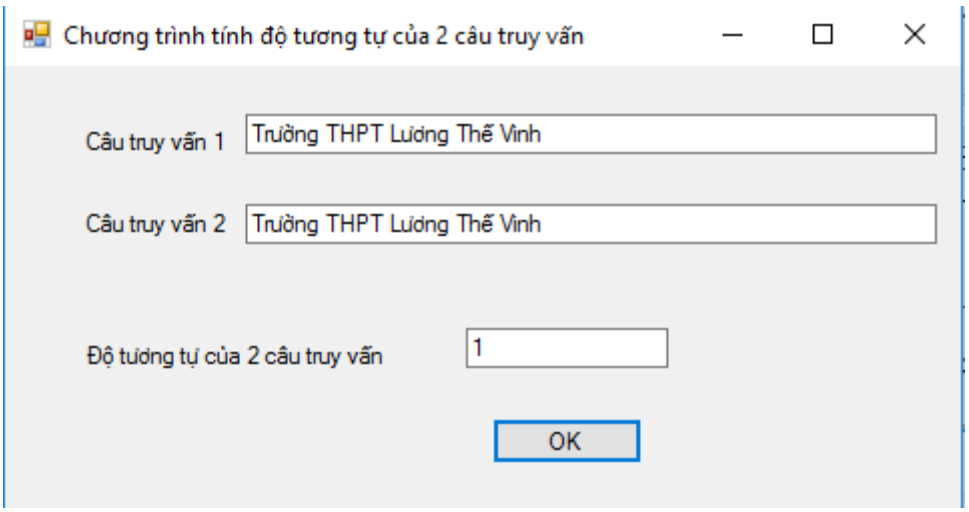
*Bảng 4.1 Kết quả tính độ tương tự giữa các truy vấn*

## 4.4. Kết quả thực nghiệm và đánh giá

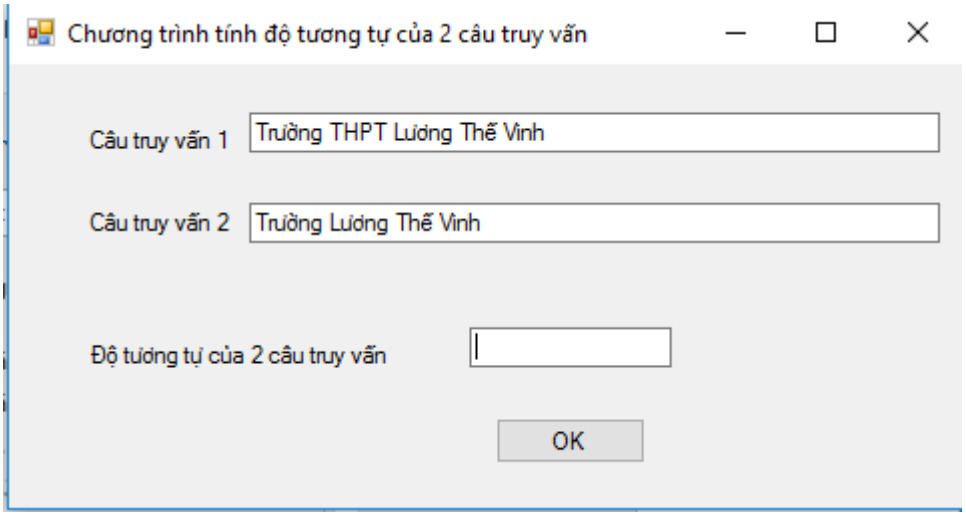
### 4.4.1. Giao diện chương trình tính độ tương tự

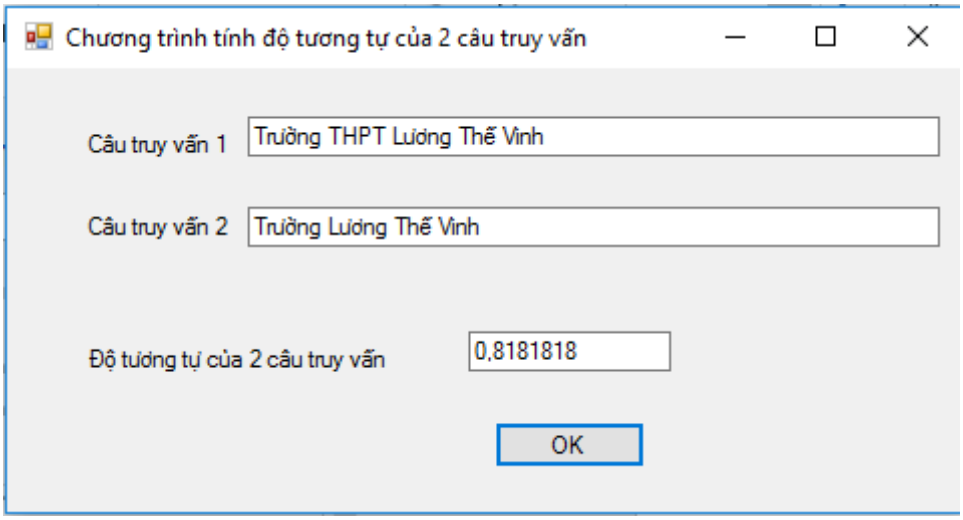
1- Chương trình so sánh 2 câu truy vấn: *Trường THPT Lương Thế Vinh* và *Trường THPT Lương Thế Vinh*




2- Chương trình so sanh 2 câu truy vấn *Trường THPT Lương Thế Vinh* và *Trường Lương Thế Vinh*




### 4.4.2. Đánh giá

Do việc tính toán định lượng để đánh giá tính chính xác của việc tính độ tương tự câu hỏi là khó khăn, nên người ta sử dụng phương pháp đánh giá định tính dựa vào con người/chuyên gia để đánh giá. Luận văn cũng sử dụng phương pháp đánh giá dựa vào con người để đánh giá việc tính toán độ tương tự giữa các câu truy vấn đã thực nghiệm.

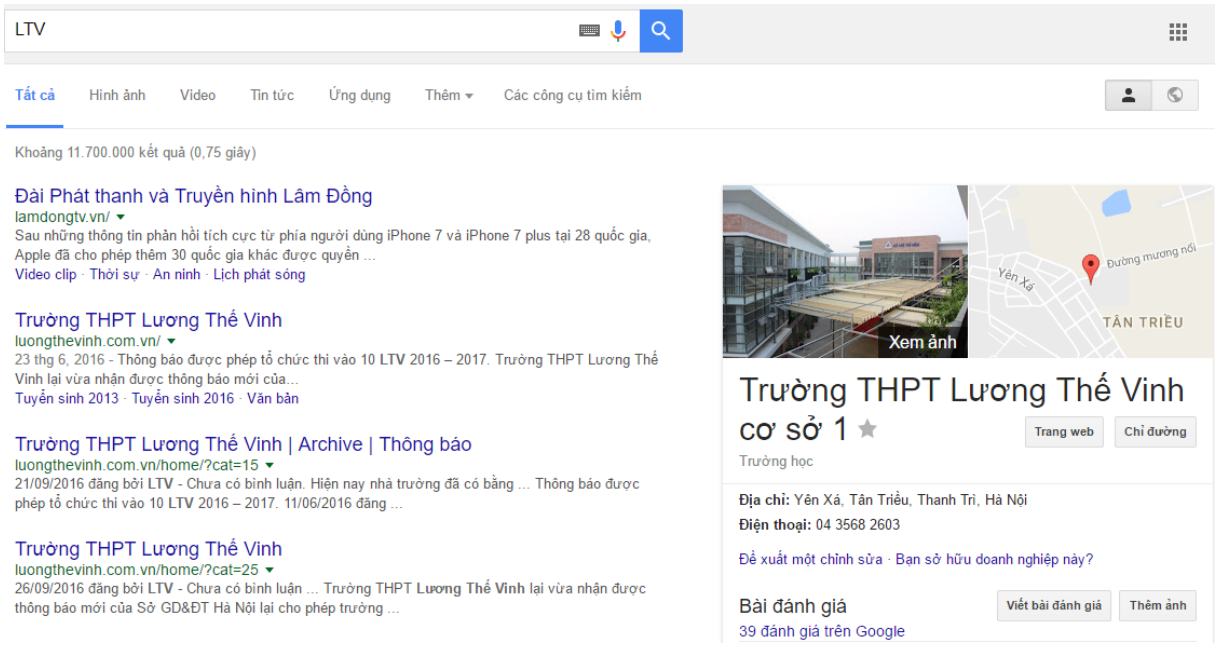
Sử dụng bảng đánh giá phân loại như sau:

| Phân loại | Mô tả | Ví dụ |
|---|---|---|
| Rất tốt | Hai câu truy vấn có tương đương về ngữ nghĩa | *LTV \|\| Trường THPT Lương Thế Vinh* |
| Tốt | Hai câu truy vấn đều có chung một mục đích truy vấn, mặc dù độ dài ngắn mô tả khác nhau. Người sử dụng muốn nói đến cùng một ý khi đưa vào truy vấn | *thi thpt 2016 \|\| Kỳ thi THPT Quốc gia năm 2016* |
| Khá tốt | Hai câu truy vấn có cùng mục đích truy vấn, nhưng sự liên quan là không rõ ràng | *nghe nhạc online \|\| mp3 trực tuyến* |
| Không tốt | Hai câu truy vấn không liên quan đến nhau | *Hoa học trò \|\| báo thanh niên* |

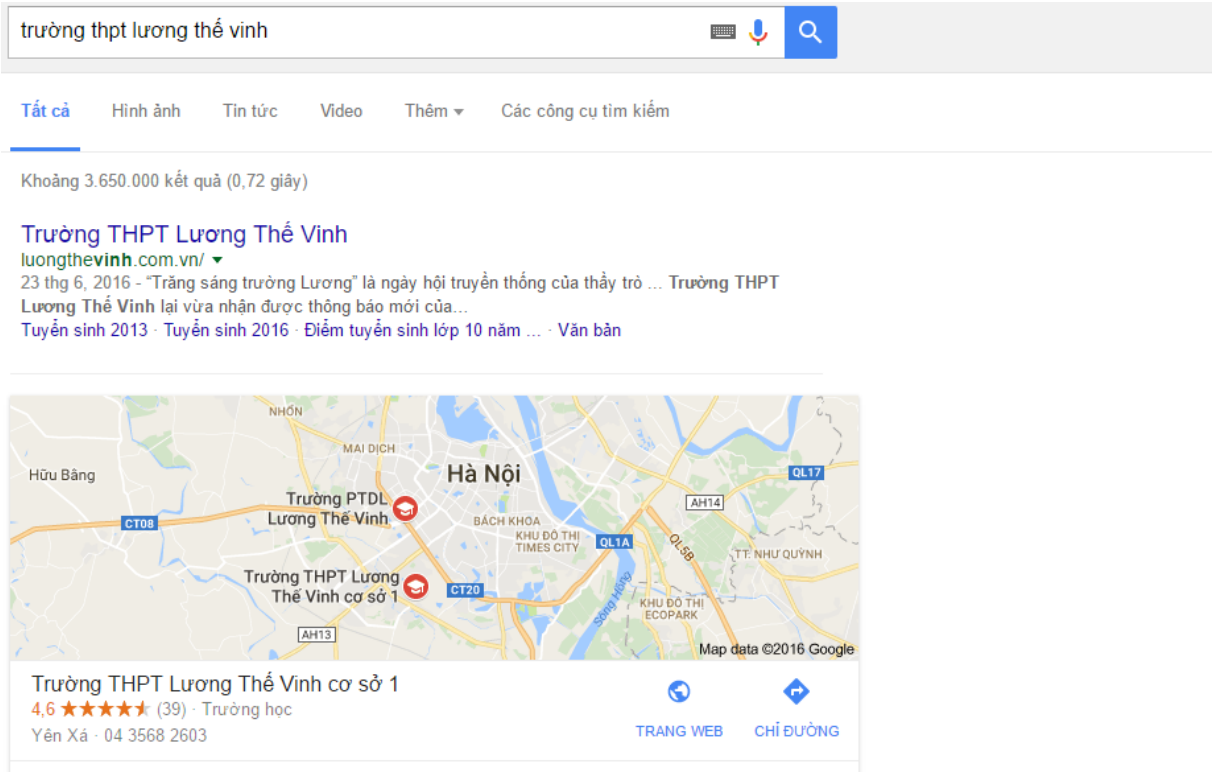
*Bảng 4.2 Bảng phân loại đánh giá*

### 4.4.3. Kết quả trả về từ máy tìm kiếm Google sau khi truy vấn

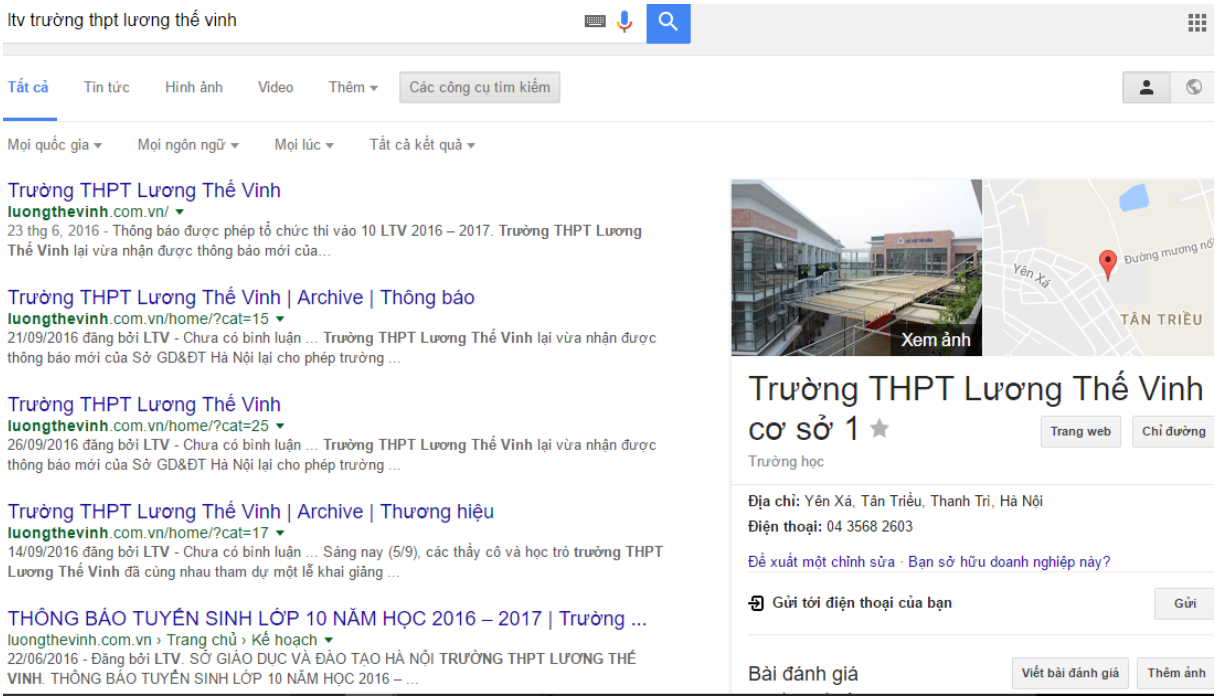
1. Với câu truy vấn: *LTV | | Trường THPT Lương Thế Vinh*




*Hình 4.1 Tìm kiếm với câu truy vấn 1*

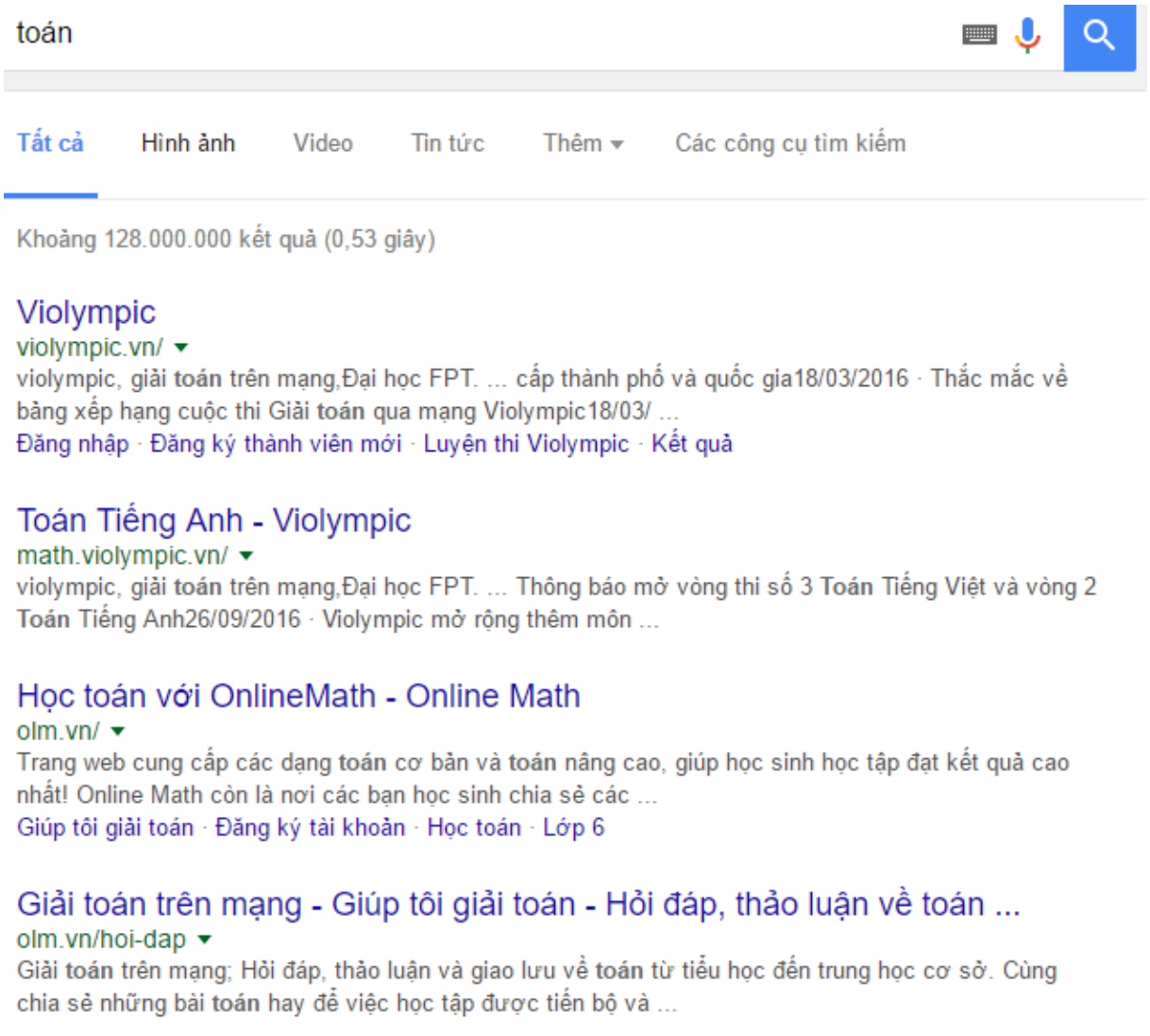



*Hình 4.2 Tìm kiếm với câu truy vấn 2*

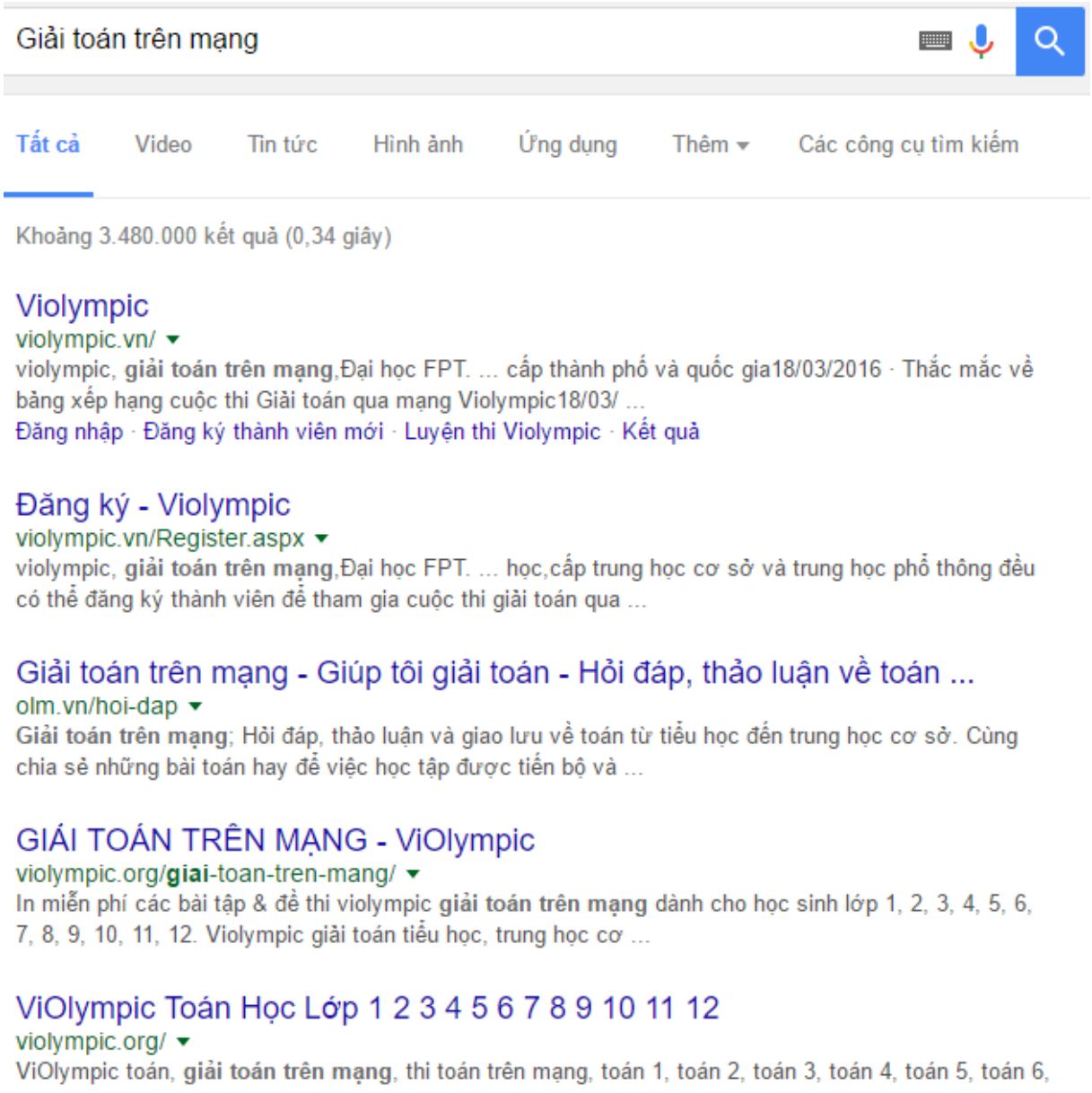



*Hình 4.3 Tìm kiếm với câu truy vấn tổng hợp 1*

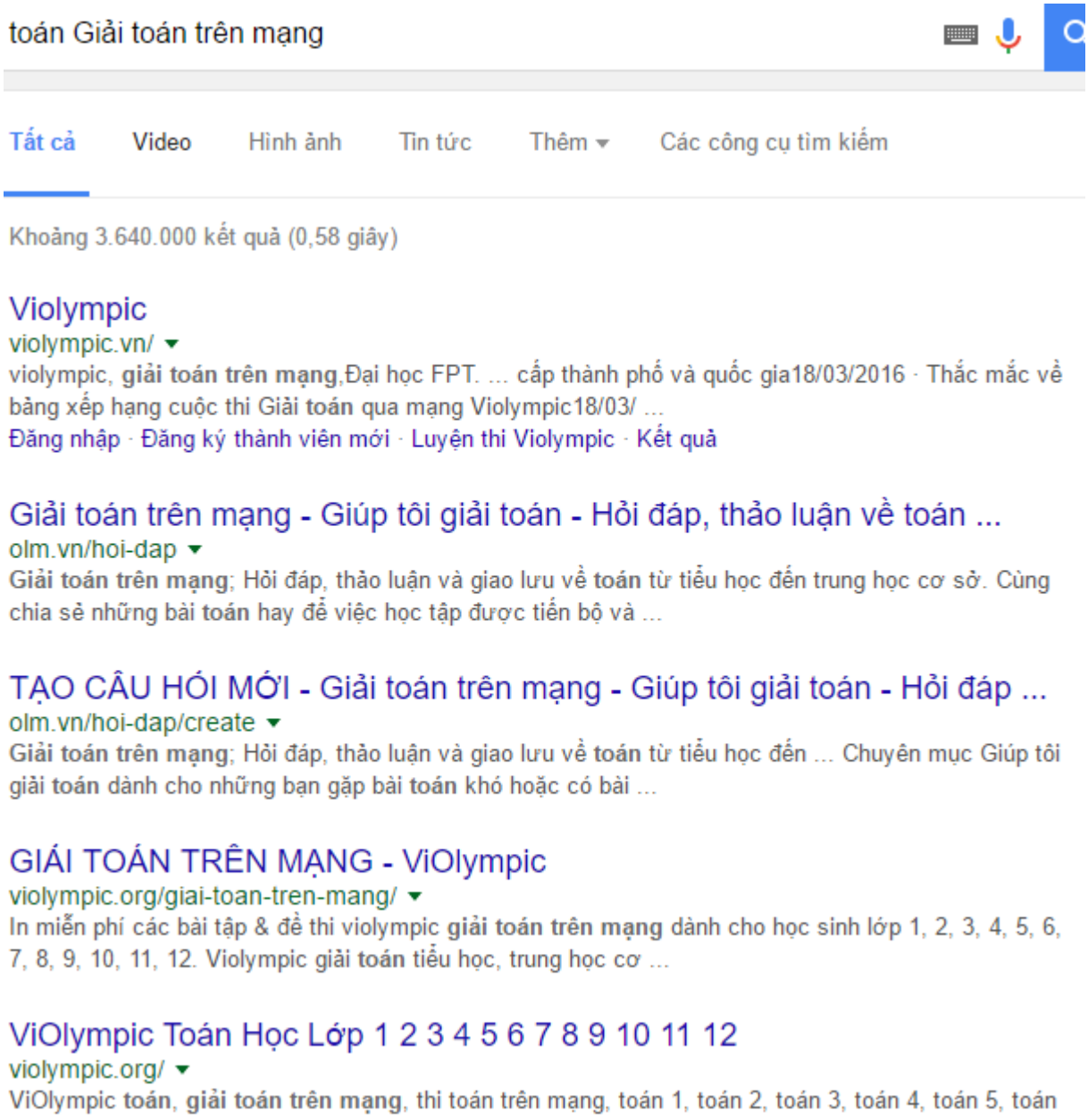
2. Với câu truy vấn: *Toán | | Giải toán trên mạng*




*Hình 4.4 Tìm kiếm với câu truy vấn 3*




*Hình 4.5 Tìm kiếm với câu truy vấn 4*




*Hình 4.6 Tìm kiếm với câu truy vấn tổng hợp 2*

# KẾT LUẬN

Luận văn định hướng chủ đề gợi ý truy vấn Internet đối với thanh thiếu niên. Kết quả chính của luận văn là:

- Cung cấp một khảo sát về gợi ý truy vấn tìm kiếm trên Internet đối với thanh thiếu niên.

- Trình bày hai kỹ thuật gợi ý truy vấn bước đi ngẫu nhiên (random walk) và kỹ thuật so sánh câu truy vấn. Nghiên cứu phương pháp thống kê và phương pháp sử dụng lưu vết truy vấn cho bài toán tính độ tương tự câu truy vấn trong máy tìm kiếm.

- Đề xuất một mô hình gợi ý truy vấn cho đối tượng thanh thiếu niên dựa trên việc kết hợp kỹ thuật gợi ý truy vấn so sánh và tính độ tương tự câu truy vấn sử dụng lưu vết truy vấn. Trong mô hình, luận văn đưa thêm giá trị trọng số cho các liên kết web để nâng cao độ chính xác của kết quả trả về.

- Xây dựng phần mềm thực nghiệm thi hành mô hình đề xuất, thực thi việc tính đoán độ tương tự của các câu truy vấn. Kết quả đánh giá định tính đối với 10 cặp câu truy vấn tương tự nhau đầu tiên cho kết quả trả về là phù hợp.

Do hạn chế về trình độ và thời gian, luận văn chưa tiến hành thử nghiệm trọn vẹn được mô hình đề xuất mà một số thành phần trong mô hình chỉ mới phân tích ở dạng định tính. Hơn nữa, mô hình trên đây chưa được tích hợp vào trang web của Trường THPT Đại Mỗ. Đấy là hướng nghiên cứu tiếp theo của luận văn.

# TÀI LIỆU THAM KHẢO


[1] Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, Pavel Serdyukov. *Query recommendation for children*. CIKM 2012: 2010-2014, 2012.

[2] Sergio Duarte Torres, Djoerd Hiemstra, Theo W. C. Huibers. *Vertical selection in the information domain of children*. JCDL 2013: 57-66, 2013.

[3] Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, Pavel Serdyukov. *Query recommendation in the information domain of children*. JASIST 65(7): 1368-1384, 2014.

[4] Sergio Raúl Duarte Torres. *Information Retrieval for Children: Search Behavior and Solutions*. PhD Thesis, University of Twentee, ...

[5] Meher T. Shaikh, Maria Soledad Pera, Yiu-Kai Ng. *Suggesting Simple and Comprehensive Queries to Elementary-Grade Children*. WI-IAT (1) 2015: 252-259.

[6] Shahrzad Karimi, Maria Soledad Pera. *Recommendations to Enhance Children Web Searches*. RecSys Posters 2015.

[7] I. Bahattin Vidinli, Rifat Ozcan. New query suggestion framework and algorithms: A case study for an educational search engine. Information Processing and Management, 2016.

[8] Livingstone, Sonia and Haddon, Leslie and Görzig, Anke and Ólafsson, Kjartan. *Risks and safety on the internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries*. EU Kids Online, Deliverable D4, 2011.

[9] Dinh, Thuy, Farrugia, Lorleen, O'Neill, Brian, Vandoninck, Sofie and Velicu, Anca (2016) *Internet safety helplines: exploratory study first findings.* Better Internet for Kids.

[10] Mascheroni, G. and Haddon, L. (2015). *Children, risks and the mobile internet*. In Y. Zheng (Ed.), Encyclopedia of Mobile Phone Behavior (pp.1409-1418). Hershey PA: IGI Global.

[11] https://www.betterinternetforkids.eu/

[12] Christopher D. Manning, Prabhakar Raghvan, Hinrich Schutze, An introduction to Information Retrieval, 2009.

[13] Manu Konchady, Building search applications – Lucene, LingPipe, and Gate, Mustru Publishing, 2008.

[14] Ziming Zhuang, Silviu Cucerzan, Q-rank: re-ranking search results using query logs.

[15] IR-models, http://en.wikipedia.org/wiki/Information_retrieval

[16] Donald Metzler, Susan T. Dumais, Christopher Meek (2007). Similarity Measures for Short Segments of Text, *ECIR 2007*: 16-27.

[17] Fu, L., Goh, H. L., Foo, S. B., & Na, J. C. (2003). Collaborative querying through a hybrid query clustering approach. *Conference on Asian Digital Libraries (6th:2003:Malaysia).*

[18] Ji-Rong Wen, Jian – Yun Nie, Hong-Jiang Zhang (2002), Query Clustering Using User Logs, *ACM Transactions on Information Systems*, Vol. 20, No. 1, January 2002.

[19] Ricardo Baeza-Yates, Carlos Hurtado, Marcelo Mendoza (2004), Query Recommendation Using Query Logs in Search Engines, *In Current Trends in Database Technology - EDBT 2004 Workshops*, Vol. 3268/2004 (18 November 2004), pp. 588-596.

[20] Siddharth Patwardhan (2003). Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. *MSc. Thesis*, University of Minnesota, Duluth, MN.

[21] Wen-tau Yih, Christopher Meek (2007). Improving Similarity Measures for Short Segments of Text. *Microsoft Research One Microsoft Way Redmo*nd, WA 98052, USA, 2007, pp 1489-1494.

[22] Wesley W. Chu, Guogen Zhang (1997). Associative query answering via query feature similarity, *Intelligent Information Systems (IIS '97)*: 405-409.

[24] Phan Xuân Hiếu. JGibbsLDA. http://gibbslda.sourceforge.net/, School of Information Sciences Tohoku University.

[25] http:// coccoc.com /users/home

[26] Nguyễn Thị Thu Chung. *Nghiên cứu, phát triển phương pháp tính độ tương tự truy vấn trong hệ tìm kiếm và ứng dụng thử nghiệm vào một hệ tìm kiếm thực thể tiếng Việt.* Luận văn Thạc sỹ, Trường Đại học Công nghệ, ĐHQGHN, 2011