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

Ngành: Công nghệ thông tin
Chuyên ngành: Kỹ thuật phần mềm
Mã số: 60480103

TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2016

Tóm tắt luận văn

Đề tài luận văn: Nghiên cứu kĩ thuật so sánh truy vấn để gợi ý tìm kiếm thông tin cho thanh thiếu niên và thử nghiệm.

Mục đích: 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.

Cấu trúc luận văn:

Cấu trúc của luận văn bao gồm các phần sau:

Mở đầu.

Chương 1: Gợi ý truy vấn cho thanh thiế niên.

Chương 2: Một số kĩ thuật gợi ý truy vấn cho thanh thiếu niên.

Chương 3: Một mô hình gợi ý truy vấn cho thanh thiếu niên.

Chương 4: Thực nghiệm và đánh giá.

Kết luận.

MỞ ĐẦU

Trong bối cảnh, lượng thông tin trên Internet ngày càng lớn và cập nhật kịp thời như hiện nay thì người dùng và đặc biệt là thanh thiếu niên càng cần một công cụ để tìm kiếm những thông tin họ cần một cách hiệu quả nhất. 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 người lớn cũng như vốn từ vựng kém phát triển hơn. Một khía cạnh khác, thanh thiếu niên gặp khó khăn trong đánh giá sự liên quan của kết quả tìm kiếm. Hỗ trợ thanh thiếu niên trong việc tìm kiếm thông tin hiệu quả, hệ thống tìm kiếm cần phải được thiết kế sao cho hỗ trợ phù hợp với nhu cầu của trẻ và kỹ năng nhận thức. 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 [1], so sánh câu truy vấn [6] sao cho phù hợp với nhu cầu và kỹ năng nhận thức của thanh thiếu niên hỗ trợ chúng 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 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. Trong đó 85% sử dụng Internet trong trường học, trò chơi 83%, tin tức 62%, các website chia sẻ dữ liệu 16% và blog 11%. Các nội dung khảo sát tập trung vào các vấn đề như bóc lột, nội dung khiêu dâm, thông tin tình dục, các ảnh hưởng của giao lưu hẹn hò trực tuyến... là những nội dung có khả năng ảnh hưởng gây hại đến thanh thiếu niên [8].

Theo số liệu của Trung tâm Internet Việt Nam (2012), ngày 19/11/1997 là ngày đầu tiên Việt Nam gia nhập vào mạng internet toàn cầu. Sau 15 năm tính tới tháng 10/2012, số người sử dụng internet đã lên tới 31,1 triệu người, chiếm tỷ lệ 35,49% dân số. Việt Nam đứng thứ 18/20 quốc gia có số người sử dụng Internet lớn nhất thế giới,đứng thứ 8 trong khu vực Châu Á và đứng vị trí thứ 3 ở khu vực Đông Nam Á.

Mặt khác, internet là phương tiện tiếp cận thông tin được sử dụng phổ biến ở Việt Nam. Theo kết quả nghiên cứu về thị trường internet Việt Nam năm 2011, internet đã vượt qua báo, tạp chí và radio để trở thành phương tiện tiếp cận thông tin phổ biến thứ hai, chỉ sau Tivi.

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 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 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

Cũng theo LSE Research Online 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 tham gia trực tiếp cùng với học sinh khai thác thông tin trên Internet. Giúp đỡ họ tìm kiếm thông tin và đưa ra những lời khuyên cho học sinh của mình.

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.

Tóm lại Internet có thể là một công cụ nghiên cứu tuyệt vời và là một cách vui thích để liên lạc với bạn bè và gia đình. Nhưng trực tuyến cũng có thể hiện một số nguy hiểm mà chúng ta cần biết. Dưới đây là một số cách hoặc giải pháp được tổng hợp trong báo cáo tại Safer Internet Day[1] mà chúng ta có thể tránh xa rắc rối trong khi sử dụng trang Web:

- 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 nhấp 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 huynh/người giám hộ ĐỒNG Ý với ý kiến này, hãy dẫn cậu/cô ta theo và gặp nhau tại một nơi 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 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.

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 thác thông tin trên mạng Internet.

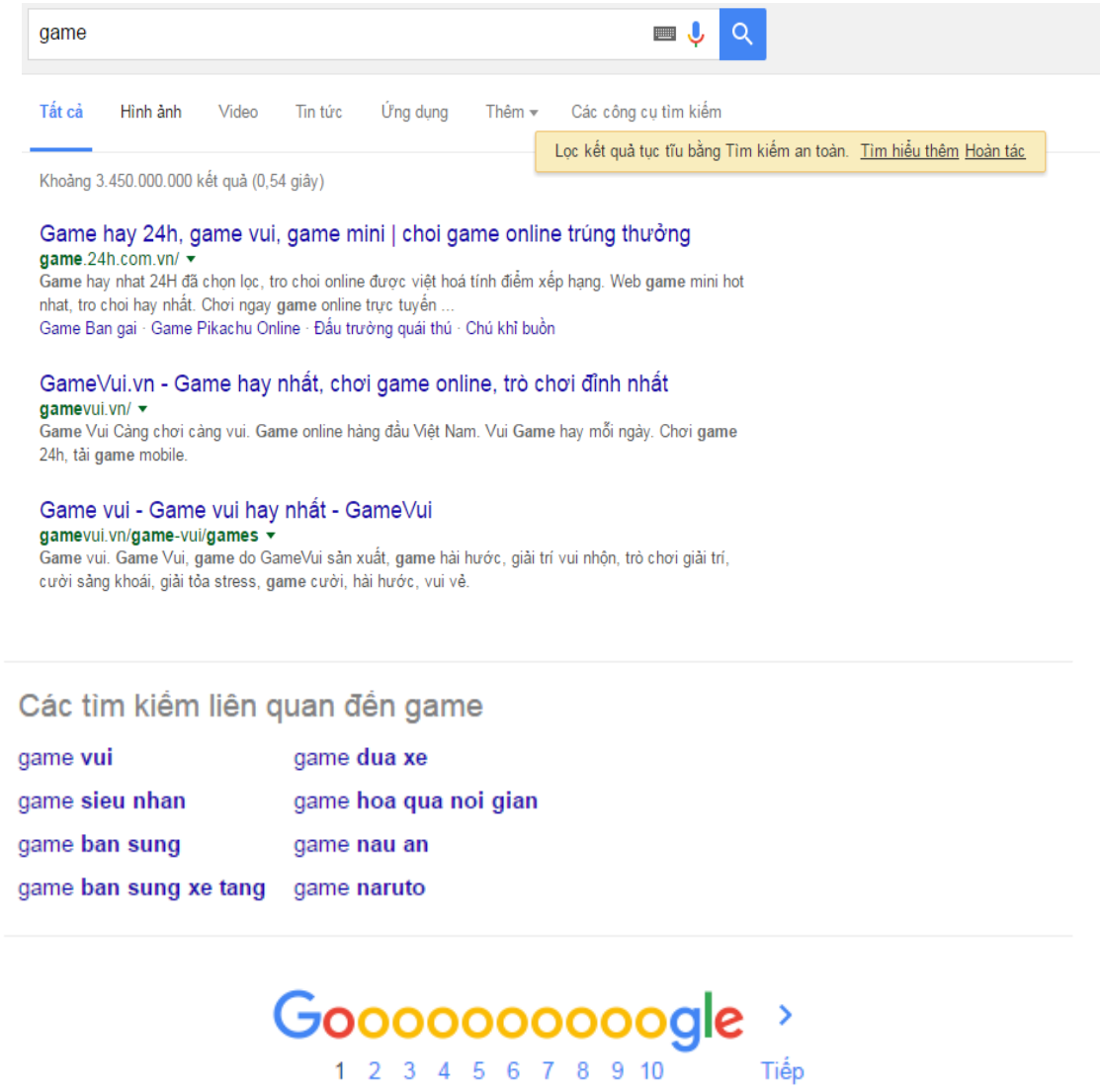
[1] www.saferinternetday.org

1.2. Gợi ý truy vấn cho thanh thiếu niên

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

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 trẻ em. Đối tượng này rất khó khăn trong việc thao tác, định hướng tìm kiếm thông tin [7]. Vì vậy 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 của công cụ tìm kiếm google

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 công cụ tìm kiếm, chất lượng của các truy vấn được gửi cũng là quan trọng.

Việc có quá nhiều kết quả trả về một phần cũng vì câu truy vấn người dùng đưa vào khá mơ hồ 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.

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

Từ những những khó khăn khi tìm kiếm của đối tượng là thanh thiếu niên nêu trên như: không biết diễn đạt câu truy vấn, không biết lựa chọn từ khóa hoặc kết quả trả về quá dài vv…nên 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 được phát hành bởi người dùng".

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 2 tập query extensions và adjacent query[4], là tập các mở rộng thường nối tiếp sau truy vấn đang xét, và tập các 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.

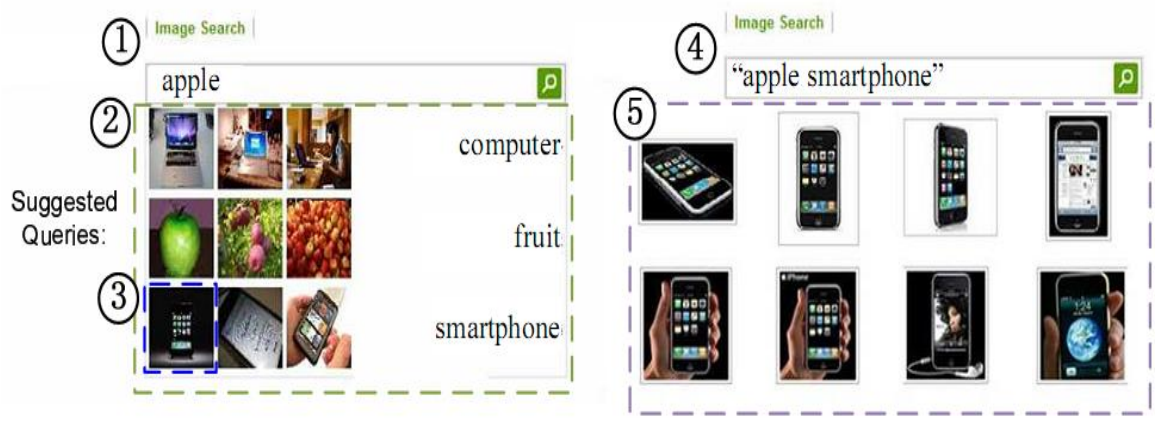
Bài toán của luận văn là đưa ra giải pháp: trợ giúp được người dùng trẻ tuổi trong việc diễn đạt ý đồ tìm kiếm và đưa ra những câu truy vấn gần với ý đồ tìm kiếm.

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

Có 2 thể hiện gợi ý truy vấn cho thanh thiếu niên đó là gợi ý trực quan và gợi ý dạng text:

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



Hinh 1.3 Ví dụ gợi ý trực quan

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

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 chung vào khai phá QueryLog. QueryLog được định nghĩa là tập các 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.

Các kỹ thuật dùng để gợi ý truy vấn như: Kỹ thuật phân cụm truy vấn, kỹ thuật thống kê, kỹ thuật hướng ngữ cảnh, so sánh câu truy vấn, hay sử dụng phương pháp học giám sát để xếp hạng truy vấn gợi ý cho thanh thiếu niên.

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 việc so sánh là các 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 để 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 các truy vấn này làm cho đơn giản hóa vấn đề gợi ý truy vấn, đơn giản cho 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 các truy vấn top đâu n ứng cử viên 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 random walk

2.1.1. Cách tiếp cận

Trong kỹ thuật này, tôi trình bày một phương pháp đề nghị truy vấn để giúp trẻ em dễ dàng tìm các từ khóa liên quan sử dụng kỹ thuật random walk. Phương pháp đề nghị 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 Tag 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 xe truy vấn về xe Cars. Theo đề xuất truy vấn gợi ý phổ biến của Google, các khía cạnh liên quan đến truy vấn này có cho thuê xe hơi, xe ô tô để bán, sử dụng xe hơi, xe ô tô mới hay xe hơi hình ảnh... 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 xe, 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 (Tag)

Xếp hạng thẻ (Tab) 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). Vấn đề quan trọng của cấu trúc đồ thị của phương pháp này là khai thác các đặc điểm tài nguyên web nhắm vào trẻ em.

2.1.3. Phương pháp

Trong phần này mô tả các kịch bản của kỹ thuật truy vấn mở rộng và phương pháp đề xuất random walk về truy vấn khuyến nghị 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 gửi bởi 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 đề nghị truy vấn.

2.1.3.2. Random walk hướng tới nội dung cho trẻ em

Mô hình Random walk sử dụng một đồ thị hai chiều là nguồn tài nguyên web (tức là, url) và nút thẻ (Tag). Trước đây cũng đã có nghiên cứu 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).

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 phân khúc 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).

Trong kỹ thuật này biểu đồ được ra dử dụng 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. Tiền thân của phương pháp random walk được dựa trên mô hình đề xuất của Craswell và Szummer. 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:

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

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.

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:

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

Trong đó U_k là tập các nguồn url.

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).

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

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: $Q=\{t_1,t_2,..t_m\}$.

2.1.4. Nhận xét

Kỹ thuật này là một cách mới lạ để 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.

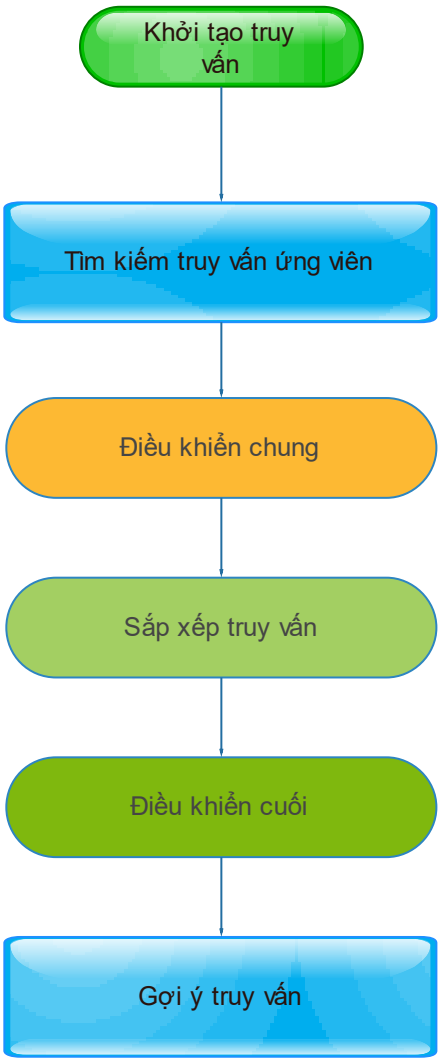
2.2. Kỹ thuật gợi ý truy vấn bằng so sánh truy vấn (QS)

2.2.1. Các tiếp cận

Truy vấn đề nghị (QS) thường được định nghĩa để "tìm kiếm một số truy vấn liên quan cho truy vấn được phát hành ban đầu của người dùng ".

Các vấn đề QS nên chỉ nghĩ 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” của so sánh đã được đưa ra bởi người tìm kiếm (người sử dụng). Truy vấn thứ hai là "truy vấn ứng viên" được đề nghị cho người sử dụng, thường được để lựa chọn ở phần cuối của quá trình đề xuất. Việc so sánh các truy vấn có thể phụ thuộc vào một số tính năng như câu từ tương quan, nhật ký truy vấn, vv .

Với kỹ thuật này không gì hơn là người ta chỉ cần quan tâm đến việc so sánh hai truy vấn. Dưới đây là mô hình gợi ý truy vấn bằng kỹ thuật so sánh truy vấn bao gồm các bước:



Hình 2.1 Mô hình gợi ý truy vấn

- 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

2.2.2. Nội dung phương pháp

2.2.2.1. Mô hình so sánh truy vấn

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 (cũng được thể hiện trong hình. 1) Chứa 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)

2.2.2.2. Pha lựa chọn

Giai đoạn lựa chọn là bước quan trọng đầu tiên của mô hình đề xuất đề nghị 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 đề nghị 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.

2.2.2.3. Pha sắp xếp

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 như mô tả trong (Aslam & Montague, 2001). 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 tôi xin 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(Renda & Straccia, 2003). Phương pháp tiếp cận khác cũng có thể được đề nghị.

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 thác 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.

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"*.

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***

$$sim(q,s) = |Q \cap S| \quad (2.14)$$

- ***Độ đo Dice***

$$sim(q,s) = 2 \times \frac{|Q \cap S|}{|Q| + |S|} \quad (2.15)$$

- ***Độ đo Jaccard***

$$sim(q,s) = \frac{|Q \bigcap S|}{|Q \bigcup S|} \quad (2.16)$$

- ***Độ đo Overlap***

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

- ***Độ đo Cosin***

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

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.

2.3.2.2. Tính độ tương tự dựa trên nhật ký truy vấn (query logs)

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.

Luận văn sẽ trình bày phương pháp tính độ tương tự cho các câu truy vấn bằng userlog với phương thức 1 đã liệt kê phía trên.

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:

$$session :=< query\ text > [clicked\ \ document]^{*} \quad (2.19)$$

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

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

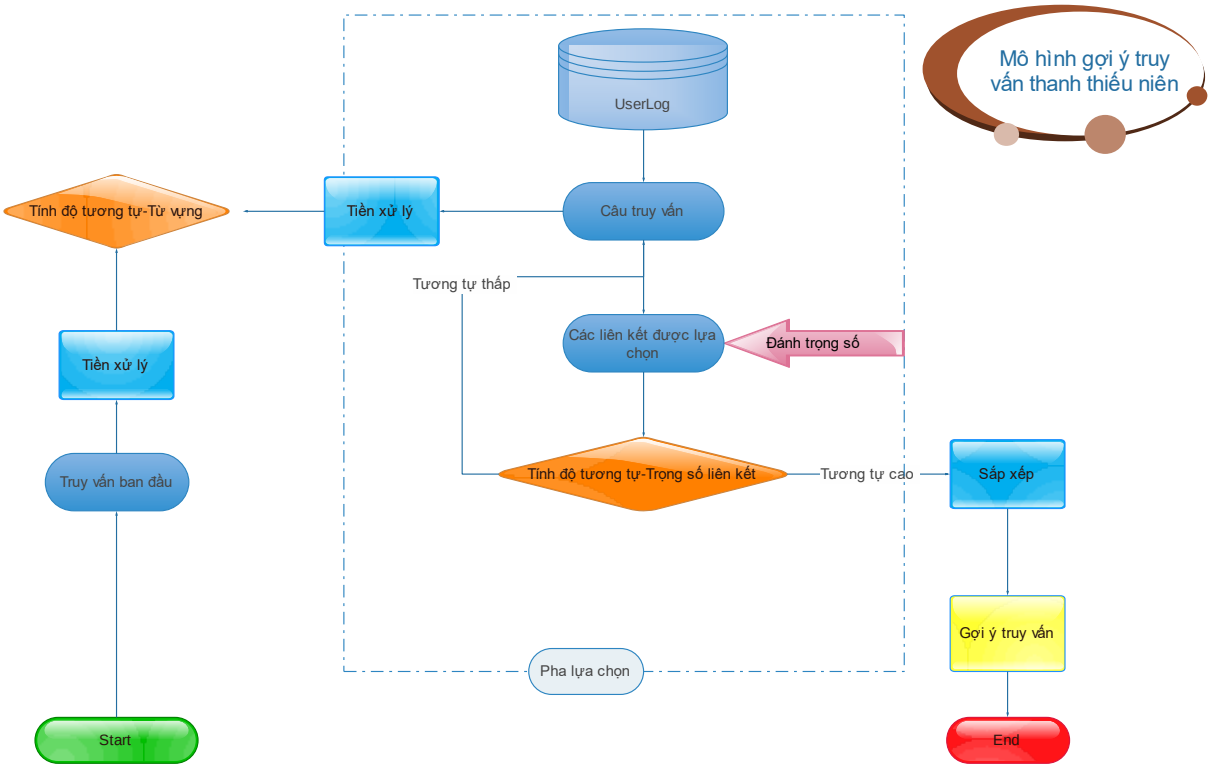
3.1. Giới thiệu

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.

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

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 "*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*" - Nguyễn Thị Thu Chung- 2011. 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
- *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
- Tính độ tương tự

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 và tính độ tương tự theo trọng số liên kết.
- Mô đun hiển thị kết quả danh sách gợi ý truy vấn cho người.

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:
- Tính độ tương tự

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.

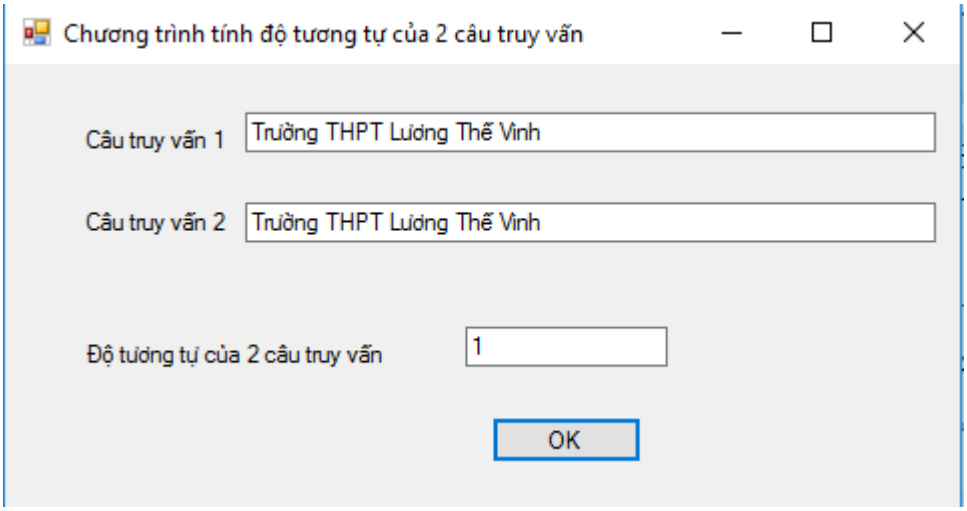
STT	Truy vấn | | truy vấn	Độ tương tự
1	LTV | | Trường THPT Lương Thế Vinh	0.95
2	Toán | | Giải toán trên mạng	0.85
3	Nghe nhạc | | nhạc online	0.9
4	thi thpt 2016 || Kỳ thi THPT Quốc gia năm 2016	0.9
5	nghe nhạc online| |mp3 trực tuyến	0.70000066
6	Truyện tranh đẹp | | hình ảnh đẹp nhất	0.38249998
7	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

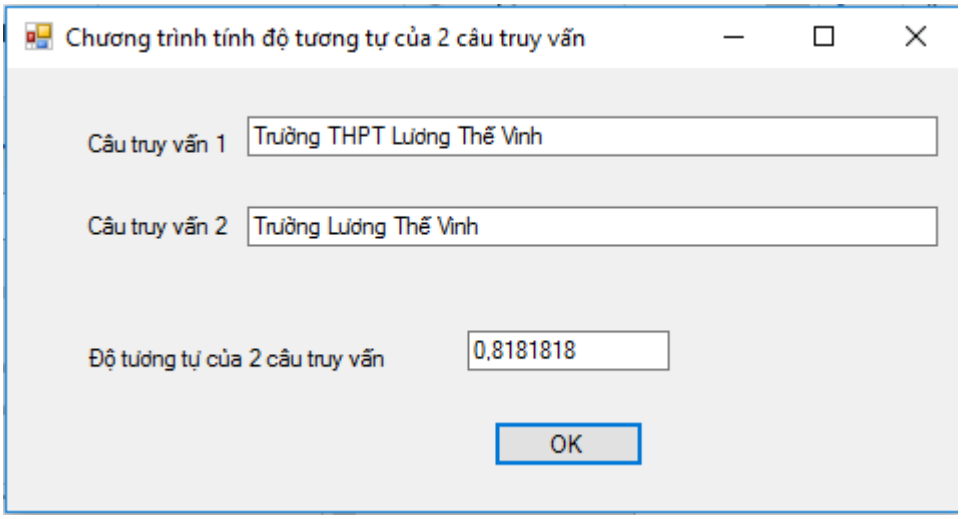
4.4. Kết quả thực nghiệm và đánh giá

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

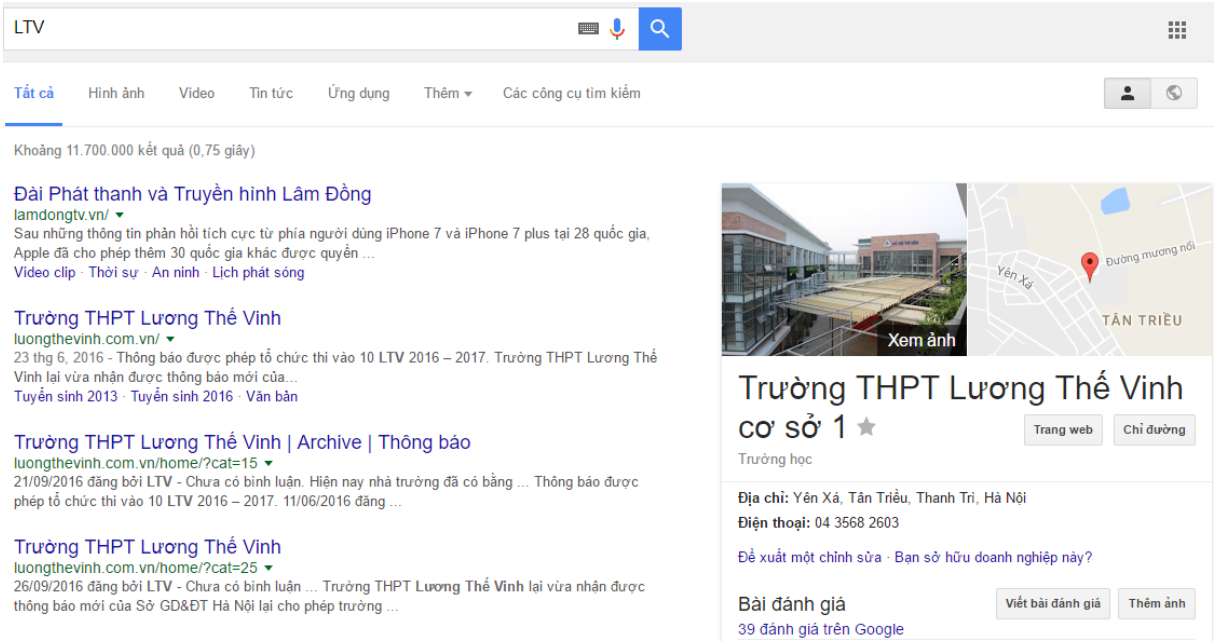
Sử dụng bảng đánh giá phân loại như sau:

Phân loại	Mô tả	Ví dụ
Rất tốt	Hai câu truy vấn có tương đương về ngữ nghĩa	*LTV* và *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*

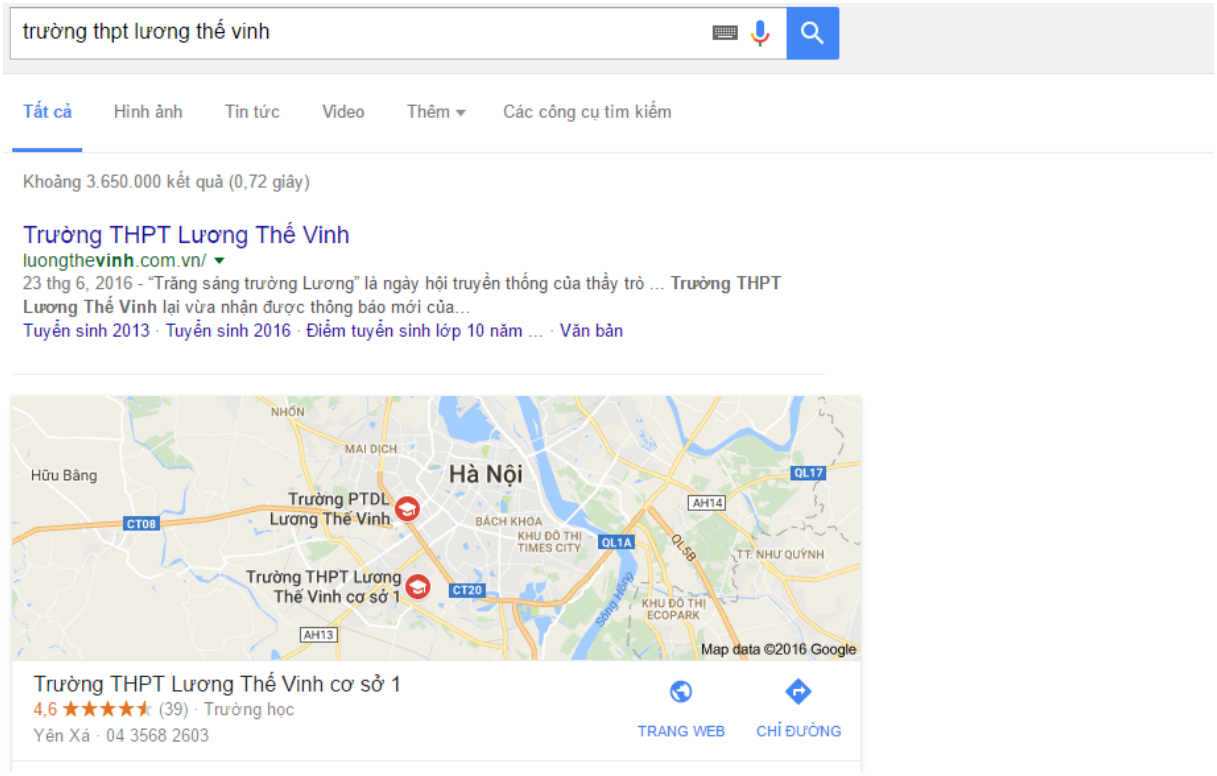
Bảng 4.2 Bảng phân loại đánh giá

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

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.