BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TRẦN ANH HUY

NGHIÊN CỨU KỸ THUẬT TRỘN KẾT QUẢ TÌM KIẾM WEBSITE

Chuyên ngành: **Khoa học máy tính**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. VÕ TRUNG HÙNG

NCS. LÂM TÙNG GIANG

Phản biện 1: **PGS.TSKH. TRẦN QUỐC CHIẾN**

Phản biện 2: **TS. HOÀNG THỊ LAN GIAO**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 18 tháng 5 năm 2013.

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại Học Đà Nẵng

MỞ ĐẦU

1. TÍNH CẤP THIẾT CỦA ĐỀ TÀI

Trong thời đại ngày nay, thông tin là nhu cầu thiết yếu đối với mọi người và mọi lĩnh vực. Mỗi phút trôi qua có hàng ngàn trang web được đưa lên Internet nhằm làm giàu nguồn tài nguyên vô tận này. Tuy nhiên việc khai thác nguồn thông tin khổng lồ này chưa được triệt để, ngay cả bộ máy tìm kiếm lớn nhất là Google vẫn chưa đáp ứng được nhu cầu tìm kiếm đa dạng của người sử dụng. Một trong các nỗ lực cải thiện kết quả tìm kiếm là việc thực hiện trộn kết quả của nhiều máy tìm kiếm.

Việc trộn kết quả tìm kiếm từ nguồn dữ liệu của các máy tìm kiếm khác nhau, sẽ cho tăng cường độ chính xác hoặc độ bao phủ của kết quả tìm kiếm. Vì lý do này, tôi đã quyết định chọn đề tài: *"Nghiên cứu kỹ thuật trộn kết quả tìm kiếm Website"* dưới sự hướng dẫn trực tiếp của PGS.TS. Võ Trung Hùng và sự hỗ trợ của ThS. Lâm Tùng Giang.

2. MỤC TIÊU NGHIÊN CỨU

Mục tiêu của đề tài là cải thiện chất lượng của dịch vụ tìm kiếm. Đề tài tập trung nghiên cứu các kỹ thuật và các giải thuật trộn kết quả tìm kiếm Website trên Internet. Và xây dựng thực nghiệm chương trình tìm kiếm Website có sử dụng các kỹ thuật trộn kết quả tìm kiếm Website đã nghiên cứu.

3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng nghiên cứu:

+ Đối tượng nghiên cứu là các kỹ thuật trộn kết quả tìm kiếm Website và các công cụ, kỹ thuật, giải thuật sử dụng trong các máy tìm kiếm.

Phạm vi nghiên cứu:

+ Tìm kiếm các Website trên Internet và trộn kết quả tìm kiếm Website trên cơ sở kết quả trả về từ các máy tìm kiếm có sẵn như: Google, Yahoo, Bing, …
+ Cài đặt giao diện người dùng.

4. PHƯƠNG PHÁP NGHIÊN CỨU

Phương pháp lý thuyết:

+ Để thực hiện nghiên cứu, chúng tôi thu thập, chọn lọc, đánh giá, phân tích và tổng hợp các tài liệu liên quan đến lĩnh vực tìm kiếm Website. Tìm hiểu tư liệu về hoạt động của các bộ máy tìm kiếm Website hiện có.
+ Chúng tôi sẽ nghiên cứu, đánh giá các kỹ thuật trộn kết quả tìm kiếm Website nhằm áp dụng triển khai vào ứng dụng.

Phương pháp thực nghiệm:

+ Bằng phương pháp thực nghiệm, chúng tôi lựa chọn hướng giải quyết nhằm đáp ứng được nhu cầu tìm kiếm đa dạng của người dùng.

+ Thực nghiệm trên các công cụ hỗ trợ xây dựng máy tìm kiếm Website.

+ Dựa trên thực trạng các bộ máy tìm kiếm hiện có để xây dựng ứng dụng tìm kiếm Website có sử dụng kỹ thuật trộn kết quả tìm kiếm Website đã nghiên cứu.

5. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN CỦA ĐỀ TÀI

Ý nghĩa khoa học:

+ Đề xuất giải pháp ứng dụng các kỹ thuật xếp hạng, bóc tách thông tin trang Web, kỹ thuật trộn kết quả tìm kiếm Website. Giải pháp này có thể cho tăng cường độ chính xác hoặc độ bao phủ của kết quả tìm kiếm Website.

Ý nghĩa thực tiễn:

+ Ứng dụng nhằm trợ giúp đáp ứng được nhu cầu tìm kiếm cho người sử dụng tìm kiếm thông tin trên Internet.

+ Hỗ trợ cho người dùng tìm kiếm, thu thập được thông tin cần tìm nhất để sử dụng cho mục đích của mình.

6. BỐ CỤC LUẬN VĂN

Toàn bộ luận văn được chia làm ba chương được tóm tắt nội dung như sau:

MỞ ĐẦU

Phần này giới thiệu về nhu cầu cần thiết để thực hiện đề tài, xác định mục tiêu, nhiệm vụ, đối tượng nghiên cứu, phương pháp nghiên cứu, cơ sở nghiên cứu và kết quả mong muốn đạt được.

CHƯƠNG 1 - CƠ SỞ LÝ THUYẾT

Chương này trình bày tổng quan về cơ sở lý thuyết máy tìm kiếm và các kỹ thuật ứng dụng trong phương pháp trộn kết quả tìm kiếm website

CHƯƠNG 2 - CÁC KỸ THUẬT TRỘN KẾT QUẢ TÌM KIẾM WEBSITE

Trong chương này, nêu giải pháp trộn kết quả tìm kiếm website. Chúng tôi tiến hành phân tích các kỹ thuật trộn kết quả tìm kiếm website. Qua các phân tích đánh giá các mô hình để xác định mô hình trộn kết quả tìm kiếm cho việc cài đặt thử nghiệm.

CHƯƠNG 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Phân tích các chức năng của hệ thống, thiết kế kiến trúc hệ thống và thực hiện xây dựng ứng dụng theo kỹ thuật trộn kết quả tìm kiếm website đã phân tích, sau đó thử nghiệm và đánh giá kết quả đạt được của chương trình.

CHƯƠNG 1 - CƠ SỞ LÝ THUYẾT

Chương này trình bày về cơ sở lý thuyết liên quan đến đề tài, làm nền tảng để nghiên cứu các kỹ thuật trộn kết quả tìm kiếm Website và xây dựng hệ thống tìm kiếm liên hợp (meta search engine). Gồm các nội dung sau:

+ Tìm hiểu về tìm kiếm thông tin.
+ Giới thiệu về khái niệm, các thuật ngữ cơ bản trong tìm kiếm.
+ Tập trung tìm hiểu về hệ thống tìm kiếm liên hợp (meta search engine).
+ Giới thiệu các kỹ thuật bóc tách thông tin và kỹ thuật xếp hạng.

1.1. CÁC KHÁI NIỆM CƠ BẢN TRONG TÌM KIẾM THÔNG TIN

1.1.1. Tài liệu

1.1.2. Thuật ngữ

1.1.3. Chỉ mục và chỉ mục ngược

1.1.4. Tần suất, độ xuất hiện, trọng số

1.1.5. Truy vấn

1.1.6. Sự phù hợp

1.2. TÌM KIẾM THÔNG TIN

1.2.1. Tổng quan về tìm kiếm thông tin và hệ thống tìm kiếm thông tin

1.2.2. Cách thức hoạt động của một hệ thống tìm kiếm thông tin

1.2.3. Các bộ phận cấu thành và nguyên lý hoạt động của hệ thống tìm kiếm thông tin

1.2.4. Mục tiêu của hệ thống tìm kiếm thông tin

Mục tiêu chính của hệ truy tìm thông tin (IR) là truy tìm những văn bản trong tập văn bản của hệ thống liên quan đến thông tin mà người sử dụng hệ thống cần. Những thông tin được người dùng đưa vào hệ thống bởi các câu truy vấn (query). Những tài liệu – văn bản "liên quan" (relevant) với câu truy vấn sẽ được hệ thống trả về. Như vậy, mục đích của hệ IR là để tự động quy trình kiểm tra tài liệu bằng cách tính độ đo tương quan giữa câu truy vấn và tài liệu.

1.2.5. Các tiêu chí đánh giá hiệu quả tìm kiếm thông tin

Có rất nhiều cách đo lường khác nhau cho việc đánh giá mức độ xử lý trả về kết quả của một hệ thống tìm kiếm thông tin. Các cách đo lường đều đòi hỏi một tập tài liệu và một câu truy vấn trên tập tài liệu đó, giả sử rằng mỗi tài liệu có thể liên quan hoặc không liên quan đến câu truy vấn. Để đánh giá hiệu quả của hệ truy tìm thông tin có thể dựa theo các tiêu chuẩn sau:

+ Dựa trên hai độ đo: "độ chính xác" (precision) và "độ bao phủ" (recall).
+ Độ chính xác (Precision): được đo bởi tỉ lệ của tài liệu trả về chính xác trên tổng các tài liệu nhận được

$$Độ\ chính\ xác = \frac{\{Tài\ liệu\ liên\ quan\} \cap \{Tài\ liệu\ nhận\ được\}}{\{Tài\ liệu\ nhận\ được\}}$$

+ Độ bao phủ (Recall): được đo bởi tỉ lệ của tài liệu trả về chính xác trên tổng các tài liệu có liên quan

1.2.6. Mô hình xếp hạng áp dụng cho các phương pháp trộn kết quả tìm kiếm

a. Mô hình xác suất – Probabilistic model

b. Mô hình không gian vector – Vector Space Model VSM

c. Đánh giá theo kết quả thử nghiệm trên hai mô hình VSM và mô hình xác suất

d. Mô tả kiến trúc hệ IR được tính điểm theo mô hình VSM (VSM – IR)

1.2.7. Đặc tả các bước xây dựng hệ VSM – IR

1.3. HOẠT ĐỘNG CỦA MÁY TÌM KIẾM LIÊN HỢP

1.3.1. Máy tìm kiếm liên hợp

1.3.2. Đánh giá về máy tìm kiếm liên hợp

1.3.3. Các bước xây dựng một máy tìm kiếm liên hợp

a. Chọn các máy tìm kiếm nguồn

b. Xử lý kết quả trả về từ máy tìm kiếm nguồn

1.4. KỸ THUẬT BÓC TÁCH DỮ LIỆU TRONG .NET

Để triển khai xây dựng ứng dụng máy tìm kiếm liên hợp – meta search engine – chúng tôi phải sử dụng kết quả tìm kiếm trả về từ các máy tìm kiếm thành phần như: Google, Yahoo, Bing,… Do các giới hạn về kinh phí và kỹ thuật phổ biến, chúng tôi sử dụng phương pháp bóc tách dữ liệu để lấy kết quả trả về từ các máy tìm kiếm.

1.4.1. Bóc dữ liệu của một trang Web

Để bóc tách được nội dung HTML của một trang Web bất kì thì chúng tôi sử dụng lớp `WebRequest` để tạo yêu cầu, lớp `WebReponse` để nhận đáp ứng từ Webserver và một số dạng Reader (`StreamReader` đối với dữ liệu `Html` hoặc `Text` hoặc `BinaryReader` đối với dữ liệu nhị phân) để phân tích các đáp ứng đó.

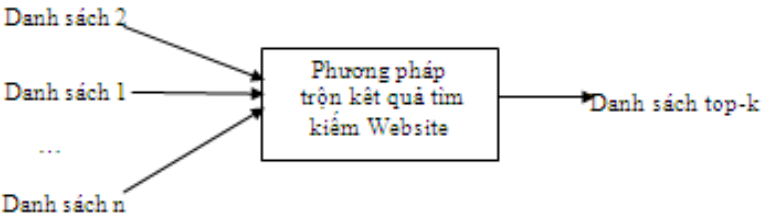
1.4.2. Giới thiệu về Regular Expression

CHƯƠNG 2 - CÁC KỸ THUẬT TRỘN KẾT QUẢ TÌM KIẾM WEBSITE

Trong chương này, luận văn tập trung thực hiện các công việc sau:

+ Phân tích mô hình trộn kết quả tìm kiếm Website và trình bày các phương pháp đã nghiên cứu: Phương pháp Borda, Phương pháp Top D, Phương pháp đếm tham chiếu và phương pháp đánh giá lại tuần tự (SRR).
+ Đưa ra giải pháp để xây dựng hệ thống tìm kiếm liên hợp (meta search engine)

2.1. MÔ HÌNH TRỘN KẾT QUẢ TÌM KIẾM WEBSITE



Mô hình trộn kết quả tìm kiếm Website

Thuật toán trộn kết quả tìm kiếm Website trong hình trên nhận đầu vào là `n` danh sách `top-k`. Mỗi danh sách là kết quả trả về của một máy tìm kiếm thông tin (như Google, Yahoo,…). Các kỹ thuật trộn kết quả tìm kiếm Website sẽ tính điểm cho mỗi tài liệu trong n danh sách `top-k` tài liệu hoặc tính điểm cho từng danh sách `top-k` và trả về một danh sách `top-k` tài liệu có điểm cao nhất.

2.2. PHƯƠNG PHÁP ĐẾM BORDA

2.2.1. Thuật toán

Bước 1: Tính điểm cho tất cả các tài liệu d_{ij} trong n danh sách top-k đầu vào.

Bước 2: Phát sinh danh sách top-k tài liệu có điểm cao nhất.

Thuật toán tính điểm của các tài liệu d_{ij} trong n danh sách top-k, mỗi tài liệu d_{ij} được tính điểm dựa vào vị trí của tài liệu này trong n danh sách L_i (Điểm của tài liệu d có vị trí thứ j trong danh sách L_i là $s(d_i) = k - j$. Tổng điểm của d là s(d)). Thuật toán tính điểm của các tài liệu d như sau:

```
For each dij
{
        s(di)=0;
        for(j=1;j<=n;j++)
        {
                for(t=1;t<=k;t++)
                        If(di=Ljt)
                            s(di)=s(di)+k+1-t;
        }
}
```

Từ điểm của các tài liệu, thuật toán xây dựng danh sách L_0 gồm k tài liệu có điểm cao nhất.

Thuật toán đếm Borda có thể được sử dụng để kết hợp những danh sách phân hạng trả về bởi các hệ thống tìm kiếm, không cần quan tâm đến điểm phân hạng từng tài liệu(điểm do hệ thống tìm

kiếm R_i đánh giá) và không cần dữ liệu huấn luyện, thuật toán kết hợp đơn giản và hiệu quả. Hơn nữa, thực nghiệm cho thấy hiệu suất đếm Borda khá tốt.

2.2.2. Công thức tính điểm D-WISE

2.3. PHƯƠNG PHÁP TOPD

Phương pháp TopD sẽ sử dụng sự tương đồng giữa câu query và document xếp trên cùng được trả về từ máy tìm kiếm j (gọi là - d_{ij}) để ước tính S_j.

Nhìn chung, nguyên nhân là do văn bản được xếp ở vị trí trên cùng (ở vị trí cao nhất) sẽ là cái liên quan nhiều nhất đối với truy vấn của người dùng dựa trên sắp xếp dữ liệu của máy tìm kiếm. Nội dung của văn bản sẽ phản ánh máy tìm kiếm hữu ích như thế nào đối với query của người dùng. Việc tìm kiếm văn bản được xếp trên cùng từ máy tìm kiếm thành phần sẽ gây ra sự trì hoãn 1 thời gian trong quá trình trộn. Nhưng chúng tôi có thể bỏ qua sự trì hoãn này bởi vì chỉ có 1 văn bản duy nhất sẽ được tìm ra từ mỗi máy tìm kiếm thành phần được sử dụng cho mỗi truy vấn.

Đối với hàm đồng dạng, chúng tôi thử cả hàm Cosine và hàm BM25.

Đối với hàm Cosine, trọng số liên quan đến mỗi từ trong Q và sự tương đồng giữa câu query với document xếp trên cùng (d_{ij}) là trọng số tf (chúng tôi thử tính $tf*idf$ và các kết quả thu được cũng tương tự như thế).

Sử dụng hàm BM25 thì sự đồng dạng giữa truy vấn Q và d_{ij} là tổng trọng số Okapi của mỗi từ thuộc câu truy vấn T. Công thức đó là:

$$\sum_{T \in Q} w * \frac{(k_1+1)*tf}{k+tf} * \frac{(k_3+1)qtf}{k_3+qtf}$$

Với $w = \log \frac{N-n+0.5}{n+0.5}$ và $K = k_1 * ((1-b)+b * \frac{dl}{avgdl})$

Trong đó:

- tf: là tần số xuất hiện của mỗi từ thuộc câu truy vấn T trong văn bản được xử lý.
- qtf: là tần số của T trong truy vấn.
- N: số văn bản.
- n: số văn bản chữa T.
- dl: độ dài văn bản.
- avgdl: độ dài trung bình của tất cả các văn bản có trong bộ sưu tập.
- k1, k3, b : hằng số.
- k1=1,2, k3=1,000, b=0,75.

Bởi vì N, n và avgdl là những giá trị chưa biết nên chúng tôi sử dụng 2 số phép tính xấp xỉ để tính chúng.

Đối với avgdl, chúng tôi sử dụng độ dài trung bình của các văn bản mà chúng tôi tập hợp được trong thử nghiệm của chúng tôi và Avgdl=1424,5 từ, chúng tôi sử dụng kích thước của Google để

ước tính N=8,058,044,651. Chúng tôi xem mỗi từ thuộc truy vấn T trong các truy vấn chúng tôi đang thử nghiệm là 1 truy vấn riêng lẽ đối với Google và số văn bản được trả về mà chúng tôi tìm được sẽ là giá trị của n.

Như đề cập trước đó, chúng tôi sử dụng công thức 1 để tính toán thì những điểm số xếp hạng của những kết quả được xếp trên cùng từ tất cả các máy tìm kiếm thành phần là bằng 1. Chúng tôi giải quyết vấn đề này bằng cách tính 1 điểm số xếp hạng được điều chỉnh ars_{ij} bằng cách nhân điểm số xếp hạng theo công thức 1: rs_{ij} với S_j ($ars_{ij}=rs_{ij}*S_j$).

Nếu 1 văn bản được truy xuất từ nhiều máy tìm kiếm thành phần thì chúng tôi tính điểm số xếp hạng cuối cùng của nó bằng cách cộng tất cả các điểm số xếp hạng được điều chỉnh ars_{ij} lại với nhau.

2.4. PHƯƠNG PHÁP ĐÁNH GIÁ LẠI TUẦN TỰ

2.4.1. Hoạt động của hệ thống xếp hạng lại tuần tự

Bước 1. Máy chủ trung tâm Sc tiếp nhận các danh sách $L_1, .., L_m$ từ nguồn tìm kiếm gốc $S_1, \ldots, S_m$.

Bước 2. Máy chủ Sc thực hiện "xếp hạng sơ bộ" các tài liệu trong mỗi danh sách L_i dựa trên công thức tính điểm TSS trình bày ở trên.

Bước 3. Máy chủ Sc thực hiện việc tuần tự tải các tài liệu từ các nguồn theo thuật toán TDR (Top - Down Re-Ranking): tải dần các tài liệu theo thứ tự xếp hạng từ trên xuống

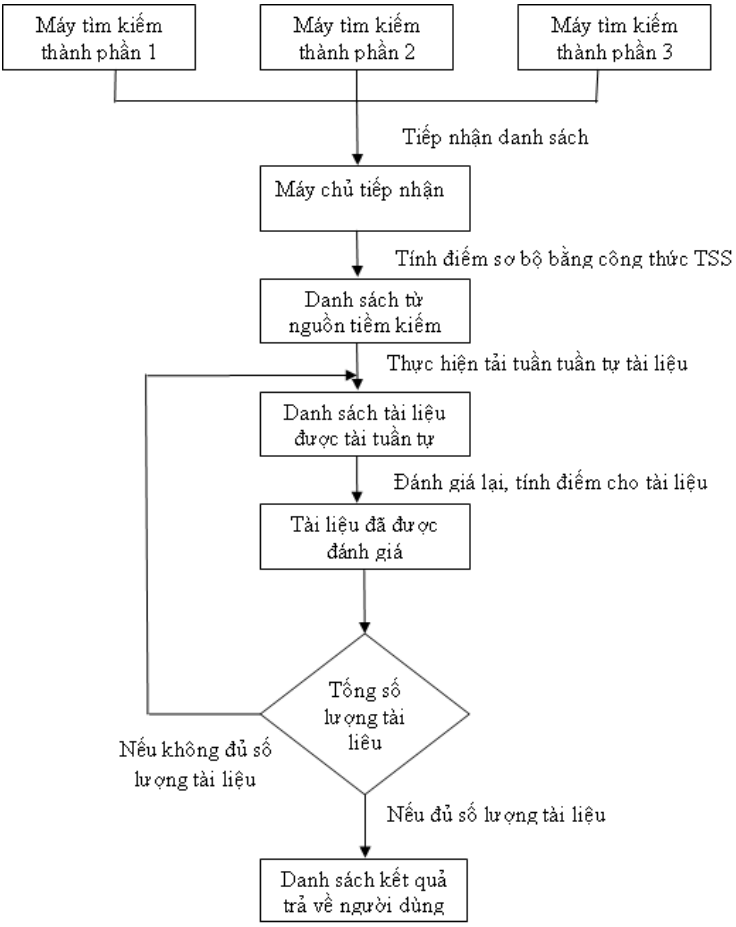
Bước 4. Thực hiện việc đánh giá lại tài liệu tải về, bao gồm tạo chỉ mục cho tài liệu bằng công cụ đánh chỉ mục tại máy chủ Sc, tính điểm của tài liệu tương ứng câu truy vấn.

Bước 5. Điều chỉnh danh sách ứng viên, bổ sung tài liệu tốt nhất vào danh sách kết quả cuối cùng, tải và xem xét tài liệu tiếp theo từ cùng nguồn với tài liệu vừa được bổ sung.

Vòng lặp các bước 3 đến 5 được thực hiện cho đến khi danh sách kết quả cuối có đủ N tài liệu. Tổng cộng có N+M bước được thực hiện (M: số nguồn tìm kiếm, N: số tài liệu từ mỗi nguồn) - nhỏ hơn nhiều so với NxM nếu như phải tải tất cả các tài liệu về

Bước 6. Máy chủ Sc hoàn chỉnh danh sách kết quả cuối cùng, trả về cho người sử dụng.

2.4.2. Mô tả hệ thống đánh giá lại tuần tự dùng để trộn kết quả tìm kiếm Website



Mô tả hoạt động của phương pháp đánh giá lại tuần tự

Để thực hiện việc đánh giá, đánh chỉ mục và tính điểm, câu truy vấn và các tài liệu sẽ được biểu diễn dưới dạng véc-tơ `((t1,n2), (t2,n2),..))` trong đó t_i là các từ trong văn bản, n_i là số lần xuất hiện của từ t_i.

2.5. PHƯƠNG PHÁP ĐẾM THAM CHIẾU

2.5.1. Thuật toán cơ sở

Mỗi L_i:

- d_{ij} được gọi là tài liệu ban đầu; $d_{l,p} = d_{ij}$, $l \neq i$ thì $d_{l,p}$ là tài liệu tham chiếu.

- $o(d_{ij})$: là số các tài liệu tham chiếu của d_{ij} trong $m-1$ danh sách L_l $(l \neq i)$.

- $S_i(k)$: tổng số đếm tài liệu tham chiếu của danh sách L_i:

$$S_i(k) = \sum_{j=1}^{k} o(d_{ij}).$$

Thuật toán tính điểm của $S_i(k)$:

```
For(i=1;i<=m;i++)
{
        for(j=1;j<=k;j++)
      {
              o(dij)=0
              for(t=1;t<=m;t++)
              {
                    If(t≠i) then
                    If dij in Lt then
o(dij)=o(dij)+1
```

```
                    }
            Si(k)=Si(k)+o(dij)
            }
    }
```

Trong thuật toán cơ sở, $o(d_{ij})$ được tính bằng cách cộng thêm 1 khi xuất hiện một tài liệu tham chiếu của d_{ij}, mà không quan tâm đến vị trí của các tài liệu tham chiếu. Ví dụ, tài liệu ban đầu $d_{1,1}$ có hai tài liệu tham chiếu là $d_{2,n}$ và $d_{3,n}$; tài liệu ban đầu $d_{2,1}$ có hai tài liệu tham chiếu là $d_{3,1}$, $d_{4,1}$. Mặc dù $d_{1,1}$ xuất hiện ở cuối danh sách của L_2 và L_3 và $d_{2,1}$ xuất hiện ở đầu danh sách L_3 và L_4, $o(d_{1,1}) = o(d_{2,1}) = 2$. Thông tin về vị trí của một tài liệu trong danh sách là rất quan trọng trong các thuật toán phân hạng nhưng thuật toán này đã bỏ qua nó.

2.5.2. Một số thuật toán đếm tham chiếu trọng số

2.5.3. Một số kết quả thực nghiệm

CHƯƠNG 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Trong chương trước, luận văn đã trình bày tổng quan về tìm kiếm thông tin trên Website và đánh giá các kỹ thuật trộn kết quả tìm kiếm Website và đưa ra giải pháp xây dựng ứng dụng hệ thống tìm kiếm liên hợp (meta search engine).

Chương này sẽ trình bày vấn đề thiết kế, cài đặt và thử nghiệm ứng dụng. Từ đó đánh giá kết quả đạt được của hệ thống tìm kiếm liên hợp.

Để thực hiện việc xây dựng hệ thống Máy tìm kiếm liên hợp - Meta Search Engine chúng tôi cần phải nắm bắt được các yêu cầu chức năng và các yêu cầu phi chức năng đối với hệ thống và chúng tôi cần biết được kỹ thuật bóc tách dữ liệu web và xử lý các kết quả trả về từ các máy tìm kiếm nguồn và sau đó áp dụng phương pháp đánh giá lại tuần tự (SRR) để trộn kết quả tìm kiếm Website và hiển thị kết quả. Để thực hiện được việc xây dựng máy tìm kiếm liên hợp đáp ứng được các yêu cầu thì chúng tôi cần có một tài liệu đặc tả yêu cầu để quá trình thực hiện được tốt hơn.

3.1. XÂY DỰNG MÁY TÌM KIẾM LIÊN HỢP

3.1.1. Phân tích và đặc tả yêu cầu

a) Mô tả chung

b) Yêu cầu chức năng

c) Yêu cầu phi chức năng

d) Tính an toàn, bảo mật, hiệu suất

3.1.2. Tài liệu đặc tả

3.1.3. Thiết kế

a) Thiết kế kiến trúc

b) Thiết kế giao diện

3.1.4. Mã hoá

3.1.5. Áp dụng kỹ thuật trộn kết quả tìm kiếm Website

a) Tổng hợp danh sách

b) Cách tổ chức dữ liệu

c) Ma trận biểu diễn tập văn bản

d) Truy vấn trong văn bản

3.2. ĐÁNH GIÁ KẾT QUẢ THỬ NGHIỆM CHƯƠNG TRÌNH

3.2.1. Tạo lập danh sách chuẩn thực tế

3.2.2. Đánh giá kết quả thử nghiệm

Tương ứng với mỗi câu truy vấn, kết quả tìm kiếm từ 5 nguồn thông tin được trộn, kết suất 10 hoặc 20 tài liệu phù hợp nhất với câu truy vấn. Kết quả được so sánh với danh sách chuẩn thực tế đã được tạo lập.

Với mỗi câu truy vấn, độ chính xác trung bình (Average Precision - AP) được tính để xác định chất lượng kết quả. Giá trị bình quân độ chính xác trung bình (Mean Average Precision - MAP) được tính bằng trung bình giá trị AP sau khi thực hiện 10 câu truy

vấn.

Với hai nhóm thử nghiệm, giá trị MAP của các phương pháp trộn kết quả tìm kiếm, tương ứng các trường hợp kết xuất 10, 20 tài liệu phù hợp nhất trong các thử nghiệm được trình bày ở bảng sau:

So sánh giá trị MAP (Mean Average Precision) của các phương pháp trộn kết quả tìm kiếm Website

Thử nghiệm	Top D	Borda	Borda có trọng số	Đếm tham chiếu	**Đánh giá lại tuần tự (SRR)**
Q_1-10	35,23%	31,57%	36,57%	42,27%	**67,44%**
Q_1-20	37,24%	30,73%	34,94%	44,35%	**68,00%**
Q_2-10	45,77%	41,60%	32,33%	57,92%	**73,71%**
Q_2-20	48,98%	42,50%	38,56%	58,78%	**72,33%**

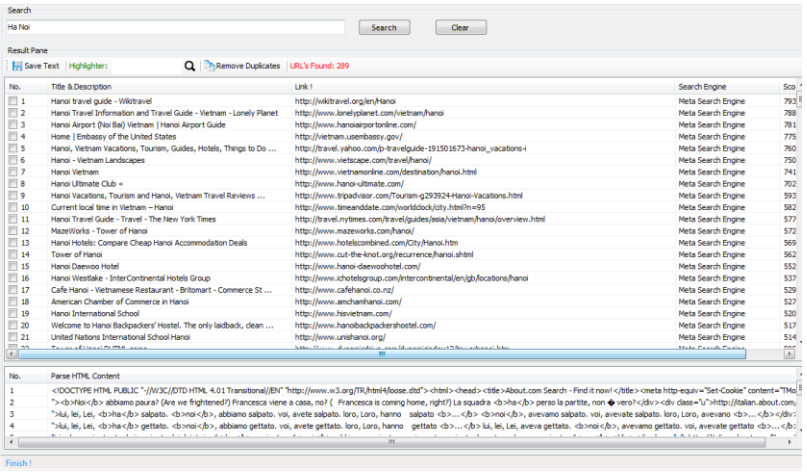
Trong cả hai trường hợp kết suất 10 hoặc 20 tài liệu đứng đầu, độ chính xác bình quân MAP tương ứng với các phương pháp khác nhau có giá trị khá ổn định so với độ dài câu truy vấn. Việc sử dụng phương pháp đánh giá lại tuần tự đạt hiệu quả khoảng 70% tốt hơn nhiều so với các phương pháp trộn kết quả tìm kiếm Website khác. Phương pháp Borda dựa trên số điểm gốc của tài liệu hay thứ tự xếp hạng khi thực nghiệm đảm bảo tốc độ nhanh và hiệu suất khá tốt. Thuật toán Borda cài đặt đơn giản, nhưng kết quả trả về của nó có độ chính xác trung bình thường vượt qua các máy tìm kiếm đầu vào, với

phương pháp Borda có trọng số thì chắc chắn cải thiện được độ chính xác của kết quả tìm kiếm. Phương pháp Top D trộn kết quả tìm kiếm Website dựa trên thông tin do các máy chủ trả về, phương pháp này sử dụng sự tương đồng giữa câu truy vấn và với tiêu đề bằng cách sử dụng công thức Cosine cho hiệu quả thấp nhất. Có thể nhận thấy khi câu truy vấn dài hơn, độ chính xác của các phương pháp cũng tăng lên, đặc biệt khi sử dụng phương pháp đánh giá lại tuần tự (Sequential Re-Ranking – SRR).

3.2.3. Xác định mô hình cài đặt thử nghiệm cho chương trình

Qua các phân tích đánh giá, đề tài xác định kỹ thuật trộn kết quả tìm kiếm Website cho việc cài đặt thử nghiệm là kỹ thuật trộn kết quả tìm kiếm pheo phương pháp đánh giá lại tuần tự (Sequential Re-Ranking – SRR).

Dưới đây là kết quả chạy thử chương trình máy tìm kiếm liên hợp - Meta Search Engine được trộn kết quả tìm kiếm Website bằng phương pháp đánh giá lại tuần tự.



Nhận xét: Chương trình được chạy thử nghiệm với từ khóa "Hà Nội", kết quả trả về cho thấy đã tăng độ bao phủ của kết quả tìm kiếm và có cải thiện được chất lượng dịch vụ tìm kiếm. Tuy nhiên, việc bóc tách dữ liệu từ máy các máy tìm kiếm nguồn thực hiện chậm. Thời gian cho quá trình bóc tách với từ khóa trên là: 12.2139 seconds cho 50 kết quả tìm thấy. Để cải thiện điều này, trong tương lai chúng tôi sẽ đề xuất thêm phần kinh phí và một số cơ chế trả kết quả tìm kiếm (các hàm api cho phép cung cấp dữ liệu từ các máy tìm kiếm) từ các máy tìm kiếm như: Google, Yahoo, Bing,.. để được hỗ trợ lấy danh sách kết quả tìm kiếm từ các bộ máy tìm kiếm được cải thiện hơn, rút ngắn thời gian cho việc tìm kiếm của người sử dụng.

KẾT LUẬN

1. Kết quả đạt được của luận văn

Về mặt lý thuyết, luận văn đã trình bày được những kiến thức về lĩnh vực tìm kiếm thông tin trên Website, tập trung nghiên cứu các phương pháp trộn kết quả tìm kiếm Website theo các phương pháp khác nhau như: phương pháp Borda, phương pháp Borda có trong số, phương pháp đếm tham chiếu, phương pháp đánh giá lại tuần tự. và cách vận dụng công cụ hỗ trợ để phát triển một hệ thống tìm kiếm liên hợp (meta search engine).

Về mặt thực nghiệm, chúng tôi đã xây dựng hệ thống tìm kiếm liên hợp và đánh giá việc áp dụng các phương pháp trộn các kết quả tìm kiếm thông tin từ nhiều nguồn bằng cách tính lại điểm dựa trên thông tin cơ bản trả về từ máy tìm kiếm gốc, đồng thời đề xuất phương pháp xếp hạng lại tuần tự, thông qua việc tải dần các tài liệu tốt nhất để tạo lập danh sách kết quả cuối cùng. Chúng tôi rút ra được một sô kết luận từ việc thực nghiệm các phương pháp:

+ Sử dụng các thông tin bản (thứ tự xếp hạng, tiêu đề, trích yếu) là chưa đủ để thực hiện việc so sánh, xếp hạng lại. Các phương pháp đã áp dụng trong nhóm này cho độ chính xác bình quân thấp hơn 50% với các câu truy vấn chứa 1 từ khóa. Tuy nhiên, việc kết hợp nhiều yếu tố thông tin cơ bản có thể cho chất lượng tốt hơn.

+ Việc tải thêm một số tài liệu trong phương pháp đánh giá lại tuần tự phục vụ đánh giá, xếp hạng lại cải thiện đáng kể chất lượng xếp hạng. Phương pháp này đã cân đối hai yếu tố tốc độ và chất lượng.

Hệ thống tìm kiếm liên hợp có giao diện thân thiện, dễ sử dụng và chức năng được thiết kế logic giúp người dùng nhanh chóng thích nghi.

Chương trình có tính thực tiễn cao, dành cho đa dạng người sử dụng, hữu ích cho nhiều người.

Chương trình luôn hoạt động ổn định và cho ra kết quả chính xác. Và khi có xung đột thì hệ thống có chiến lược giải quyết xung đột để quản lý thứ tự thực hiện các luật xung đột này.

Ngoài ra, kiến trúc của hệ thống tìm kiếm liên hợp mang tính "mở", cho phép dễ dàng cập nhật và mở rộng để chương trình có thể được nâng cấp một cách đơn giản.

2. Hướng phát triển

Mở rộng chương trình để hỗ trợ đầy đủ chức năng hơn và có thể hỗ trợ giúp người dùng tìm kiếm thông tin thuận tiện hơn nữa, đặc biệt là phát triển xây dựng cho phù hợp để người dùng nước ngoài có thể sử dụng thuận tiện hơn theo ngôn ngữ của họ, như vậy chương trình sẽ đa dạng và phong phú hơn.

Phát triển chương trình chạy trên nền Web để dễ dàng phổ biến cho nhiều người sử dụng hơn thông qua môi trường mạng Internet.