BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

Lâm Tùng Giang

MỘT SỐ PHƯƠNG PHÁP PHỤC VỤ XẾP HẠNG CÁC TRANG WEB TRONG TÌM KIẾM XUYÊN NGỮ

Chuyên ngành : Khoa học máy tính
Mã số : 62 48 01 01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

ĐÀ NẴNG - 2017

Công trình được hoàn thành tại: Trường Đại học Bách khoa, Đại học Đà Nẵng

Cán bộ hướng dẫn khoa học:

- PGS.TS. **Võ Trung Hùng**
- PGS.TS. **Huỳnh Công Pháp**

Phản biện 1: GS. TS. Hoàng Văn Kiếm

Phản biện 2: PGS. TS. Lê Mạnh Thạnh

Phản biện 3: PGS. TS. Phan Huy Khánh

Luận án được bảo vệ trước Hội đồng chấm luận án cấp Đại học Đà nẵng họp tại Đại học Đà Nẵng vào hồi 14h00 giờ ngày 26 tháng 5 năm 2017

LỜI MỞ ĐẦU

Tìm kiếm web xuyên ngữ đặt ra nhiệm vụ từ nhu cầu thông tin của người dùng được trình bày ở một ngôn ngữ (ngôn ngữ nguồn) thực hiện việc xác định các trang web phù hợp được viết bằng một ngôn ngữ khác (ngôn ngữ đích). Xếp hạng trong tìm kiếm Web xuyên ngữ liên quan đến việc tạo lập kết quả khi thực hiện một câu truy vấn ở dạng một danh sách các tài liệu theo thứ tự phù hợp với nhu cầu truy vấn.

Nhằm thực hiện việc xếp hạng trong truy vấn thông tin nói chung và trong bài toán tìm kiếm Web xuyên ngữ nói riêng, cần giải quyết hai nhiệm vụ trọng tâm: Thứ nhất là nhiệm vụ dịch thuật, nhằm biểu diễn câu truy vấn và các tài liệu trong một không gian chung, cụ thể là trong cùng một ngôn ngữ. Thứ hai là nhiệm vụ xếp hạng, thông qua việc triển khai các giải pháp kỹ thuật, các thước đo nhằm đánh giá, so sánh mức độ phù hợp giữa các tài liệu và câu truy vấn.

Một số hạn chế của các giải pháp hiện tại bao gồm chất lượng dịch thuật thấp và sự lệ thuộc vào cặp ngôn ngữ. Với các hệ thống tìm kiếm liên quan tiếng Việt, các vấn đề về xử lý ngôn ngữ cũng như dịch thuật đã khiến hiệu quả xếp hạng kết quả tìm kiếm còn rất hạn chế. Bên cạnh đó, một hệ thống tìm kiếm Web cần có thiết kế riêng biệt so với một hệ thống truy vấn thông tin văn bản truyền thống nhằm khai thác cấu trúc đặc thù của các tài liệu HTML phục vụ quá trình xếp hạng. Từ các hạn chế đã nêu, phát sinh nhu cầu nghiên cứu nâng cao chất lượng dịch thuật cũng như nhu cầu nghiên cứu tăng hiệu quả xếp hạng thông qua việc khai thác đặc thù của các tài liệu HTML.

Xuất phát từ tình hình thực tiễn, đề tài *"Một số phương pháp phục vụ xếp hạng trang Web trong tìm kiếm xuyên ngữ"* được chọn làm đề tài nghiên cứu của luận án Tiến sĩ kỹ thuật nhằm đề xuất một mô hình hệ thống tìm kiếm Web xuyên ngữ và các giải pháp kỹ thuật được áp dụng tại các thành phần của mô hình nhằm nâng cao hiệu

quả xếp hạng danh sách kết quả tìm kiếm.

1. Mục tiêu, đối tượng và phạm vi nghiên cứu của luận án

Các mục tiêu cụ thể của luận án bao gồm: nghiên cứu và đề xuất các phương pháp phục vụ dịch thuật, bao gồm các kỹ thuật tiền xử lý câu truy vấn, dịch câu truy vấn và xử lý câu truy vấn ở ngôn ngữ đích cũng như nghiên cứu và đề xuất các phương pháp xếp hạng lại danh sách kết quả tìm kiếm trong truy vấn xuyên ngữ, chú trọng việc xếp hạng các trang Web. Thước đo hiệu quả chính được sử dụng là điểm MAP (Mean Average Precision).

2. Bố cục của luận án

Ngoài phần mở đầu và kết luận, luận án được tổ chức thành 5 chương với cấu trúc như sau:

Chương 1: *Tổng quan và đề xuất nghiên cứu*

Chương 2: *Dịch tự động phục vụ truy vấn xuyên ngữ*

Chương 3 : *Hỗ trợ dịch câu truy vấn*

Chương 4: *Xếp hạng lại*

Chương 5: *Hệ thống tìm kiếm Web xuyên ngữ Việt Anh*

3. Đóng góp của luận án

- Đề xuất được các phương pháp khử nhập nhằng mới trong mô-đun dịch câu truy vấn;
- Đề xuất được phương pháp tiền xử lý câu truy vấn;
- Đề xuất được các phương pháp cải tiến câu truy vấn tại ngôn ngữ đích;
- Đề xuất được các mô hình lân cận xuyên ngữ;
- Đề xuất được phương pháp học xếp hạng dựa trên lập trình di truyền.
- Thiết kế một mô hình tìm kiếm Web xuyên ngữ cho cặp ngôn ngữ Việt-Anh.

CHƯƠNG 1: TỔNG QUAN VÀ ĐỀ XUẤT NGHIÊN CỨU

1.1. Truy vấn thông tin

1.1.1. Khái niệm

1.1.2. Định nghĩa hình thức

1.1.3. Sơ đồ xử lý của hệ thống truy vấn thông tin

Các giải pháp truy vấn thông tin được chia thành 2 giai đoạn:
Giai đoạn I: Thu thập, xử lý, đánh chỉ mục, lưu trữ tài liệu.
Giai đoạn II: Truy vấn.

1.1.4. Các mô hình truy vấn thông tin truyền thống

Các mô hình truy vấn thông tin truyền thống phục vụ việc đánh chỉ mục bao gồm mô hình Boolean (Boolean model), mô hình không gian vec-tơ (Vector Space model), mô hình xác suất (Probabilistic model).

1.1.5. Khai thác quan hệ giữa các thuật ngữ

Mô hình chỉ mục ngữ nghĩa ngầm và mô hình lân cận xem xét mối quan hệ ngữ nghĩa giữa các thuật ngữ trong văn bản.

1.2. Đánh giá hệ thống truy vấn thông tin

1.3. Truy vấn thông tin xuyên ngữ

1.3.1. Khái niệm

Truy vấn thông tin xuyên ngữ giải quyết trường hợp khi tài liệu cần truy vấn được viết bằng ngôn ngữ khác với ngôn ngữ truy vấn.

1.3.2. Các hướng tiếp cận

Hai hướng tiếp cận chủ yếu trong CLIR là dịch câu truy vấn và dịch tài liệu.

1.4. Các kỹ thuật xếp hạng lại

1.5. Xếp hạng trang Web

1.6. Các hạn chế và đề xuất nghiên cứu

1.6.1. Hạn chế

Các hạn chế chính trong các nghiên cứu bao gồm chất lượng dịch thuật và việc chưa khai thác đặc thù của tài liệu web khi xếp hạng.

1.6.2. Đề xuất nghiên cứu

Tác giả xác định 2 vấn đề cần thực hiện nghiên cứu bao gồm vấn đề *dịch thuật* nhằm tạo môi trường cho phép so sánh câu truy vấn và các tài liệu cần tìm kiếm và vấn đề *cải tiến chất lượng xếp hạng***,** đảm bảo hệ thống tìm kiếm được xây dựng phù hợp với loại tài

liệu lưu trữ và đạt hiệu năng cao dựa trên các thước đo đánh giá hệ thống đã trình bày. Từ đây, tác giả đề xuất xây dựng mô hình xếp hạng phục vụ tìm kiếm Web xuyên ngữ.

Các nội dung được tác giả thực hiện nghiên cứu bao gồm:

- Các kỹ thuật dịch tự động;
- Các kỹ thuật hỗ trợ dịch thuật bao gồm tiền xử lý câu truy vấn tại ngôn ngữ nguồn và tối ưu hóa câu truy vấn tại ngôn ngữ đích;
- Các phương pháp học xếp hạng;
- Xây dựng hệ thống tìm kiếm Web xuyên ngữ.

1.7. Tiểu kết chương

Tác giả xác định 2 vấn đề cần thực hiện nghiên cứu bao gồm vấn đề *dịch thuật* nhằm tạo môi trường cho phép so sánh câu truy vấn với các tài liệu cần tìm kiếm và vấn đề *cải tiến chất lượng xếp hạng*.

CHƯƠNG 2: DỊCH TỰ ĐỘNG PHỤC VỤ TRUY VẤN XUYÊN NGỮ

2.1. Các phương pháp dịch tự động

2.2. Khử nhập nhằng trong phương pháp sử dụng từ điển

Ba vấn đề chính có khả năng gây ảnh hưởng giảm hiệu năng của hệ thống bao gồm độ bao phủ của từ điển, việc phân đoạn câu truy vấn thành các phần có nghĩa và việc xác định bản dịch phù hợp.

2.3. Mô hình sử dụng từ điển máy

2.3.1. Các biến thể của công thức MI

2.3.1.1 Sử dụng tần xuất cùng xuất hiện của cặp từ

Công thức phổ biến tính giá trị *MI* thể hiện quan hệ cặp từ có dạng sau:

$$MI_{cooc} = \log\left(\frac{p(x,y)}{p(x) \times p(y)}\right) \quad (2.1)$$

trong đó, với $p(x,y)$ là xác suất hai từ x,y cùng xuất hiện trong cùng câu với khoảng cách không quá 5 từ, $p(x)$ và $p(y)$ là xác suất xuất hiện từ x và y trong kho ngữ liệu.

2.3.1.2 Sử dụng máy tìm kiếm

Với 2 từ *x* và *y*, các chuỗi *x,y* và *'x AND y'* được dùng như các câu truy vấn gửi tới máy tìm kiếm. Các giá trị *n*(*x*), *n*(*y*), *n*(*x,y*) tương ứng sẽ là số tài liệu chứa các chuỗi *x*, *y* và *x,y* cùng xuất hiện.

$$MI_{ir} = \frac{n(x,y)}{n(x) \times n(y)} \quad (2.2)$$

2.3.2. Thuật toán chọn bản dịch tốt nhất

Các thuật toán trong phần này được thực hiện khi câu truy vấn tiếng Việt q_v đã được phân tích thành một tập hợp $((v_1,L_1),(v_2,L_2), \ldots,(v_n,L_n))$ chứa các từ khóa tiếng Việt $v_1,..v_n$ và các danh sách bản dịch tương ứng $L_1,\ldots,L_n$, trong đó $L_i = (t_1, \ldots, t_{k_i})$ là danh sách chứa các bản dịch ứng viên của v_i.

2.3.2.1 Thuật toán sử dụng cohesion score

2.3.2.2 Thuật toán SMI

Mỗi bản dịch ứng viên $qtran_e$ biểu diễn dưới dạng $qtran_e = (e_1, \ldots, e_n)$, trong đó e_i được chọn từ danh sách L_i. Hàm *SMI* (Summary Mutual Information) được định nghĩa như sau

$$SMI(qtran_e) = \sum_{x,y \in qtran_e} MI(x,y) \quad (2.3)$$

Bản dịch ứng viên với giá trị *SMI* cao nhất được chọn là bản dịch tiếng Anh cho câu truy vấn tiếng Việt q_v ban đầu.

2.3.2.3 Thuật toán SQ chọn bản dịch một cách tuần tự

Đầu tiên, một danh sách các cặp bản dịch (t_i^k , t_{i+1}^j) của tất cả các cặp 2 cột liền kề (*i, i+1*) được tạo lập. Trong danh sách này, 2 cột tương ứng cặp bản dịch có giá trị hàm MI cao nhất là được chọn là cột *i0* và *i0+1*, tạo thành tập hợp *GoodColumns*. Sau đó bản dịch tốt nhất từ các cột liền kề với hai cột trên được xác định dựa trên giá trị của một hàm *cohesion score* trong công thức:

$$cohesion(t_i^k) = \sum_{c \in GoodColumns} MI(t_i^k, t_c^{best}) \quad (2.4)$$

Cột tương ứng bản dịch tốt nhất được bổ sung tập hợp *GoodColumns*. Quá trình trên tiếp tục cho đến khi mọi cột đều được kiểm tra. Tiếp theo, các bản dịch trong mỗi cột được sắp xếp lại.Kết quả, tương ứng với mỗi từ tiếng Việt, ta nhận được một danh sách các bản dịch tốt nhất.

2.3.3. Xây dựng câu truy vấn

2.3.3.1 Kết hợp 2 phương pháp gán trọng số thủ công

Câu truy vấn được tạo có dạng:

$$q = (t_1^1 w_1^1 OR\ t_1^2 w_1^2 \dots t_1^{m_1} w_1^{m_1})\ w_1\ AND \dots AND\ (t_n^1 w_n^1 OR\ t_n^2 w_n^2 \dots t_n^{m_n} w_n^{m_n}) w_n \quad (2.5)$$

2.3.3.2 Gán trọng số dựa trên kết quả quá trình khử nhập nhằng

Gọi $t_i^1, t_i^2, \dots t_i^{m_i}$ là các phương án dịch của v_i trong danh sách L_i với các trọng số tương ứng là $w_i^1, w_i^2, \dots w_i^{m_i}$. Khi đó, câu truy vấn có dạng:

$$q = (t_1^1 w_1^1 OR\ t_1^2 w_1^2 \dots t_1^{m_1} w_1^{m_1})\ AND \dots AND\ (t_n^1 w_n^1 OR\ t_n^2 w_n^2 \dots t_n^{m_n} w_n^{m_n}) \quad (2.6)$$

2.3.4. Áp dụng công thức SMI chọn bản dịch tốt nhất

Bảng 2.1: Kết quả thực nghiệm

STT	*Cấu hình*	*P@1*	*P@5*	*P@10*	*MAP*	*So sánh*
1	nMI	0.497	0.482	0.429	0.436	74.79%
2	SMI	0.511	0.488	0.447	0.446	76.50%
3	Dịch Google	0.489	0.535	0.505	0.499	85.59%
4	Dịch thủ công	0.605	0.605	0.563	0.583	100%

2.4. Thực nghiệm tạo bản dịch câu truy vấn có cấu trúc

Bảng 2.2: So sánh P@k và MAP các cấu hình

	Cấu hình	*P@1*	*P@5*	*P@10*	*MAP*	*Tỷ lệ*

1	*top_one_ch*	0.64	0.48	0.444	0.275	71.24%
2	*top_one_sq*	0.52	0.472	0.46	0.291	75.39%
3	*top_three_ch*	0.68	0.528	0.524	0.316	81.87%
4	*top_three_sq*	0.64	0.552	0.532	0.323	84.55%
5	*top_three_all*	0.76	0.576	0.54	0.364	94.30%
6	*Google*	0.64	0.568	0.536	0.349	90.41%
7	*Baseline*	0.76	0.648	0.696	0.386	100%

2.5. Tiểu kết chương

Chương 2 trình bày nghiên cứu của tác giả liên quan các kỹ thuật dịch tự động phục vụ truy vấn xuyên ngữ. Đề xuất của tác giả trình bày trong chương là các phương án *dịch câu truy vấn bằng từ điển: Phương pháp thứ nhất* định nghĩa hàm Summary Mutual Information nhằm chọn một phương án dịch tốt nhất cho mỗi từ khóa trong câu truy vấn. *Phương pháp thứ hai* dựa trên một thuật toán chọn bản dịch cho các từ khóa truy vấn một cách tuần tự.

Việc sử dụng công thức SMI cho kết quả tốt hơn phương pháp sử dụng thuật toán *Greedy*, tuy nhiên vẫn không tốt bằng máy dịch Google. Phương pháp chọn bản dịch một cách tuần tự SQ cho kết quả vượt trội máy dịch Google. Điều kiện để triển khai thuật toán là máy tìm kiếm phải hỗ trợ câu truy vấn có cấu trúc.

CHƯƠNG 3: CÁC KỸ THUẬT HỖ TRỢ DỊCH CÂU TRUY VẤN

3.1. Phân đoạn câu truy vấn

3.1.1. Sử dụng công cụ *vnTagger*

3.1.2. Thuật toán WLQS

Thuật toán WLQS (Word-length-based Query Segmentation) - do tác giả đề xuất - thực hiện việc phân đoạn câu truy vấn dựa trên độ dài từ khóa. Việc đề xuất thuật toán trên cơ sở của giả thuyết: nếu một từ đa âm (compound word) tồn tại trong từ điển và chứa các từ bên trong khác, bản dịch của từ có xu hướng tốt hơn việc kết hợp bản dịch của các từ bên trong.

3.1.3. Kết hợp WLQS và công cụ *vnTagger*

Nhằm nâng cao hiệu quả của thuật toán WLQS cũng như khai thác các ưu điểm của bộ công cụ vnTagger, một thuật toán phân đoạn, bóc tách từ khóa từ câu truy vấn được xây dựng trên cơ sở kết hợp các ưu điểm của hai thành phần. Thuật toán bóc tách từ khóa từ câu truy vấn tiếng Việt gồm 5 bước: tìm từ trong từ điển, gán nhãn từ, loại bỏ các từ chứa trong từ khác, loại bỏ các từ chồng chéo, bổ sung lại các từ còn sót.

3.2. Điều chỉnh câu truy vấn ở ngôn ngữ đích

3.2.1. Phản hồi ẩn trong truy vấn xuyên ngữ

Trong truy vấn xuyên ngữ, PRF được áp dụng ở các giai đoạn khác nhau: trước hoặc sau quá trình dịch thuật hoặc kết hợp sử dụng trong cả 2 giai đoạn với mục tiêu nâng cao hiệu quả truy vấn

3.2.2. Điều chỉnh câu truy vấn có cấu trúc ở ngôn ngữ đích

Với tập hợp các tài liệu trả về từ câu truy vấn ban đầu, trọng số của các thuật ngữ chứa trong câu truy vấn được tính lại để xây dựng lại câu truy vấn mới với dạng

$$q' = (t_1^1 w_1^1 OR\ t_1^2 w_1^2 \dots t_1^{m_1} w_1^{m_1})\ AND$$
$$\dots AND\ (t_n^1 w_n^1 OR\ t_n^2 w_n^2 \dots t_n^{m_n} w_n^{m_n})$$

Để mở rộng câu truy vấn, xem xét 4 công thức khác nhau phục vụ việc tính toán trọng số mới cho các thuật ngữ:

Công thức FW1:

$$w(\mathrm{t_j}) = \frac{\lambda}{|D_r|} \times \sum_{d_i \in D_r} wd_i^j \tag{3.1}$$

Công thức FW2, kết hợp trọng số *tf-idf* cục bộ và trọng số *idf* của các từ khóa:

$$w(\mathrm{t_j}) = \frac{\lambda}{|D_r|} \times \sum_{d_i \in D_r} wd_i^j \times \log(\frac{N+1}{N_{t_i}+1}) \tag{3.2}$$

Ở đây, N là tổng số tài liệu trong kho tài liệu, N_t là số tài liệu chứa thuật ngữ t, λ là tham số điều chỉnh.

Với thuật ngữ t_j và từ khóa q_k, $mi(t_j,q_k)$ là số lần cùng xuất hiện của hai từ với khoảng cách không quá 3 ký tự. Công thức FW3:

$$w(\mathrm{t_j}) = \lambda \times \sum_{q_k \in q} mi(t_j, q_k) \quad (3.3)$$

Công thức FW4:

$$w(t_j) = \lambda \times \sum_{q_k \in q} mi(t_j, q_k) \times log(\frac{N+1}{N_{q_k}+1}) \quad (3.4)$$

Bằng cách thêm p thuật ngữ với trọng số cao nhất, câu truy vấn cuối cùng có dạng như sau:

$$\begin{aligned} q_{final} = q' AND(expaned\ terms) \ = \\ = (w_1^1 OR\ t_1^2 w_1^2 \dots t_1^{m_1} w_1^{m_1}) \\ AND \dots AND\ (t_n^1 w_n^1 OR\ t_n^2 w_n^2 \dots t_n^{m_n} w_n^{m_n}) \\ AND\ e_1 w_1 \dots e_p w_p \end{aligned} \quad (3.5)$$

Trong đó $t_i^1, t_i^2, \dots t_i^{m_i}$ là các phương án dịch của v_i trong danh sách L_i với các trọng số tương ứng là $w_i^1, w_i^2, \dots w_i^{m_i}$; $e_1, \dots, e_p$ là các thuật ngữ mở rộng với các trọng số tương ứng là $w_1, \dots, w_p$.

3.3. Thực nghiệm

Kết quả thực nghiệm cho thấy việc kết hợp áp dụng thuật toán đề xuất để xác định lại trọng số từ khóa truy vấn và mở rộng câu truy vấn giúp tăng độ chính xác và độ bao phủ cho hệ thống.

3.4. Tiểu kết chương

Các đóng góp của tác giả được trình bày ở chương 3 bao gồm: Thuật toán thực hiện việc phân đoạn câu truy vấn, được thực hiện ở bước tiền xử lý câu truy vấn thông qua việc kết hợp thuật toán phân đoạn dựa trên độ dài từ khóa và công cụ *vnTagger* và các kỹ thuật điều chỉnh câu truy vấn ở ngôn ngữ đích dựa trên việc sử dụng phản hồi ẩn nhằm tính lại trọng số của các từ khóa truy vấn và mở rộng câu truy vấn.

CHƯƠNG 4: XẾP HẠNG LẠI

4.1. Ứng dụng lập trình di truyền phục vụ học xếp hạng

4.1.1. Mô hình ứng dụng lập trình di truyền

Tác giả sử dụng bộ dữ liệu đánh giá OHSUMED để đánh giá việc học xếp hạng dựa trên lập trình di truyền. Mỗi cá thể (gene)

được xác định là một hàm $f(q,d)$ đo mức độ phù hợp của văn bản so với câu truy vấn, với các phương án như sau:

Phương án 1: Hàm tuyến tính sử dụng 45 thuộc tính:

$$TF - AF = a_1 \times f_1 + a_2 \times f_2 + \cdots + a_{45} \times f_{45} \quad (4.1)$$

Phương án 2: Hàm tuyến tính, chỉ sử dụng một số thuộc tính chọn lọc ngẫu nhiên:

$$TF - RF = a_{i1} \times f_{i1} + a_{i2} \times f_{i2} + \cdots + a_{in} \times f_{in} \quad (4.2)$$

Phương án 3: Áp dụng hàm số lên các thuộc tính. Giới hạn sử dụng các hàm số x, $1/x$, $sin(x)$, $log(x)$, và $1/(1+e^x)$.

$$TF - FF = a_1 \times h_1(f_1) + a_2 \times h_2(f_2) + \cdots + a_{45} \times h_{45}(f_{45}) \quad (4.3)$$

Phương án 4: Tạo dựng hàm *TF-GF* với cấu trúc hình cây tương tự phương pháp của Yeh và các đồng sự, giữ lại đánh giá các hàm phi tuyến tính.

Trong các công thức, a_i là các tham số, f_i là giá trị thuộc tính của văn bản, h_i là hàm số.Các hàm lượng giá (fitness function) tương ứng với giá trị MAP.

4.1.2. Xây dựng công cụ và kết quả thực nghiệm

4.1.3. Đánh giá

Các bảng so sánh cho thấy các phương án *TF-AF*, *TF-RF* cho kết quả tốt. Các giá trị MAP, NDCG@k và P@k vượt trội hơn hẳn so với giá trị tương ứng của các phương pháp *Regression*, *RankSVM* và *RankBoost*, tương đương và có phần nhỉnh hơn so với các phương pháp *ListNet* và *FRank*. Phương pháp *TF-GF* cho kết quả không cao. Kết quả này cho thấy việc sử dụng các hàm tuyến tính phục vụ xếp hạng đảm bảo tính hiệu quả.

4.2. Đề xuất các mô hình lân cận

Tác giả đề xuất các mô hình lân cận (proximity models), áp dụng trong bối cảnh truy vấn xuyên ngữ.

4.2.1. Mô hình CL-Büttcher

4.2.2. Mô hình xếp hạng CL-Rasolofo

4.2.3. Mô hình xếp hạng CL-HighDensity

4.2.13. Thực nghiệm ứng dụng mô hình lân cận xuyên ngữ

Các hàm xếp hạng sau được sử dụng để kiểm tra và so sánh:

$$s_{CL-Buttcher}(d,q) = score_{solr}(d,q) + score_{okapi}(d,q) + 10 \times score_{CL-Buttcher}(d,q) \quad (4.4)$$

$$s_{CL-Rasolofo}(d,q) = score_{solr}(d,q) + score_{okapi}(d,q) + 10 \times score_{CL-Rasolofo}(d,q) \quad (4.5)$$

$$s_{CL-HighDensity}(d,q) = score_{solr}(d,q) + score_{okapi}(d,q) + 5 \times score_{CL-HighDensity}(d,q) \quad (4.6)$$

Bảng 3.1: Điểm MAP của các cấu hình thực nghiệm

	Origin	***CL-Buttcher***	***CL-Rasolofo***	***CL-HighDensity***
top_three_ch	0.350	0.352	0.372	0.365
top_three_sq	0.370	0.375	0.397	0.389
top_three_all	0.380	0.386	0.403	0.397
Join-all	0.351	0.357	0.376	0.374
Flat	0.262	0.271	0.310	0.299
Google	0.372			
Baseline	0.381			

Bảng 3.2: Mức độ tăng hiệu quả khi áp dụng mô hình lân cận

	CL-Butcher	***CL-Rasolofo***	***CL-HighDensity***
top_three_ch	0.57%	6.29%	4.29%
top_three_sq	1.35%	7.30%	5.14%
top_three_all	1.58%	6.05%	4.47%
Join-all	1.71%	7.12%	6.55%
Flat	3.44%	18.32%	14.12%

4.3. Học xếp hạng trang Web

4.3.1. Các mô hình học xếp hạng

Hai mô hình học xếp hạng dựa trên lập trình di truyền được đề xuất nhằm *"học"* hàm xếp hạng dưới dạng tổ hợp tuyến tính của

ác hàm xếp hạng cơ sở. Mô hình thứ nhất sử dụng dữ liệu huấn luyện chứa điểm số gán cho các thành phần trong các tài liệu HTML và nhãn xác định tài liệu có phù hợp hay không so với câu truy vấn. Mô hình thứ hai chỉ sử dụng điểm số gán cho các thành phần trong các tài liệu HTML, sau đó so sánh thứ tự xếp hạng của các hàm ứng viên so với các hàm xếp hạng cơ sở.

4.3.2. Cá thể

Với một tập *n* hàm xếp hạng cơ sở $F_0, F_1,\ldots,F_n$, mỗi cá thể được xem xét có dạng một hàm tuyến tính *f* kết hợp các hàm xếp hạng cơ sở:

$$f(d) = \sum_{i=0}^{n} \alpha_i \times F_i(d) \qquad (4.7)$$

Với α_i là các số thực, *d* là tài liệu cần gán điểm. Mục đích của chúng ta là xác định hàm *f* cho kết quả xếp hạng tốt nhất.

4.3.2.1 Hàm mục tiêu

Hàm mục tiêu (fitness function) xác định mức độ thích nghi của mỗi cá thể. Hàm mục tiêu được sử dụng trong mô hình học xếp hạng có giám sát được đề xuất là giá trị MAP

Thuật toán 4.1: tính độ phù hợp (có giám sát)

```
Input: Hàm ứng viên f, tập các câu truy vấn Q
Output: mức độ phù hợp của hàm f
begin
n = 0; sap = 0;
for each câu truy vấn q do
        n+=1;
        tính điểm mỗi tài liệu bởi hàm xếp hạng f;
        ap = độ chính xác trung bình cho hàm xếp hạng f;
        sap += ap;
map = sap/n
return map
```

Trong mô hình học xếp hạng không giám sát, gọi *r*(*i*,*d*,*q*) là

thứ hạng của tài liệu *d* trong danh sách kết quả tìm kiếm bằng câu truy vấn *q*, sử dụng hàm xếp hạng F_i; *rf*(*d*,*q*) là thứ hạng của tài liệu *d* trong danh sách kết quả tìm kiếm bằng câu truy vấn *q*, sử dụng hàm xếp hạng *f*; thuật toán được trình bày như sau:

Thuật toán 4.2: tính độ phù hợp (không giám sát)

```
Input: Hàm ứng viên f, tập các câu truy vấn Q
Output: mức độ phù hợp của hàm f
begin
s_fit = 0;
for each câu truy vấn q do
        tính điểm mỗi tài liệu bởi hàm xếp hạng f;
        D = tập hợp 200 tài liệu đứng đầu;
        for each tài liệu d in D do
                k+=1;d_fit = 0;
                for i=0 to n do
                        d_fit +=distance(i,k,q)
                s_fit += d_fit
return s_fit
```

Tác giả thực nghiệm 3 phương án của hàm *distance*(*i*,*k*,*q*) được sử sụng trong thuật toán 4.2 như sau:

Bảng 4.3: Các phương án hàm distance

Phương án	***distance*(*i*,*k*,*q*)**
1	*abs*(*r*(*i*,*d*,*q*)-*rf*(*d*,*q*))
2	*abs*(*r*(*i*,*d*,*q*)-*rf*(*d*,*q*))/*log*(*k*+1)
3	(*r*(*i*,*d*,*q*)-*rf*(*d*,*q*))/ *k*

4.3.2.2 Quá trình huấn luyện

4.3.3. Môi trường thực nghiệm

4.3.4. Cấu hình thực nghiệm

Thuật toán đề xuất được kiểm tra với các cấu hình sau:
Cấu hình ***SQ***: sử dụng bản dịch có cấu trúc.
Cấu hình ***SC***: kết quả học xếp hạng có giám sát.

Các cấu hình ***UC1, UC2, UC3***: kết quả học xếp hạng không giám sát, tương ứng với 3 cấu hình hàm mục tiêu định nghĩa tại Bảng 4.3.

4.3.5. Kết quả thực nghiệm

Bảng 4.4: Kết quả thực nghiệm

Cấu hình	***Giá trị MAP***
Baseline	0.3742
Google	0.3548
SQ	0.4307
SC	0.4640
UC1	0.4284
UC2	0.4394
UC3	0.4585

4.4. Tiểu kết chương

Từ câu truy vấn ở ngôn ngữ nguồn, việc áp dụng các kỹ thuật trình bày tại chương 2 và chương 3 cho phép tạo lập và hiệu chỉnh một câu truy vấn có cấu trúc tại ngôn ngữ đích. Chương 4 kế thừa các kết quả của các chương này và trình bày các đề xuất kỹ thuật của tác giả phục vụ xếp hạng lại kết quả tìm kiếm. Các đóng góp của tác giả được trình bày trong chương 4 bao gồm:

- Đề xuất bóc tách và đánh chỉ mục các thành phần nội dung trong trang web trong máy tìm kiếm nhằm định nghĩa tập hợp các hàm xếp hạng cơ sở;
- Định nghĩa các mô hình lân cận xuyên ngữ *CL-Buttcher*, *CL-Rasolofo* và *CL-HighDensity* áp dụng trong tìm kiếm xuyên ngữ nhằm tìm kiếm các hàm xếp hạng cơ sở mới;
- Đề xuất mô hình học xếp hạng trong một hệ thống tìm kiếm Web xuyên ngữ, trong đó hàm xếp hạng cuối cùng được xây dựng dưới dạng một tổ hợp tuyến tính các hàm xếp hạng cơ sở.

Kết quả thực nghiệm cho thấy việc áp dụng học xếp hạng giúp tăng hiệu quả của hệ thống (đo bằng độ đo MAP).

CHƯƠNG 5: HỆ THỐNG TÌM KIẾM WEB XUYÊN NGỮ VIỆT-ANH

Chương 5 trình bày chi tiết thiết kế hệ thống tìm kiếm Web xuyên ngữ Việt-Anh và các kết quả thực nghiệm nhằm đánh giá ảnh hưởng của việc áp dụng các giải pháp kỹ thuật đề xuất trong luận án cũng như so sánh hiệu năng với các giải pháp kỹ thuật khác.

5.1. Thiết kế hệ thống

5.1.1. Các thành phần hệ thống

Các thành phần chính của hệ thống bao gồm *tiền xử lý câu truy vấn, dịch câu truy vấn, điều chỉnh câu truy vấn, tìm kiếm tiếng Anh* và *xếp hạng lại*; tương ứng với kết quả nghiên cứu trình bày tại các chương 2, 3 và 4.

5.1.2. Dữ liệu từ điển

5.1.3. Dữ liệu đánh chỉ mục

5.2. Phương pháp thực nghiệm

5.3. Thực nghiệm các giải pháp dịch câu truy vấn

5.3.1. Cấu hình thực nghiệm

Bảng 5.1: Các cấu hình đánh giá các giải pháp dịch câu truy vấn

Cấu hình	***Diễn giải***
Baseline	Các câu truy vấn được dịch thủ công
Google	Các câu truy vấn được dịch bằng cách sử dụng máy dịch Google
nMI	Sử dụng thuật toán khử nhập nhằng *greedy*
SMI	Sử dụng thuật toán khử nhập nhằng SMI
Top_one_all	Sử dụng thuật toán chọn bản dịch một cách tuần tự, kết xuất chỉ một bản dịch tốt nhất cho mỗi từ khóa, tạo lập câu truy vấn có cấu trúc.
Top_three_all	Sử dụng thuật toán chọn bản dịch một cách tuần tự, kết xuất 3 bản dịch tốt nhất cho mỗi từ khóa, tạo lập câu truy vấn có cấu trúc.
Top_three_weight	Sử dụng thuật toán chọn bản dịch một cách tuần, kết xuất 3 bản dịch tốt nhất cho mỗi từ

	khóa,tạo lập câu truy vấn có cấu trúc với trọng số xác định trong quá trình khử nhập nhằng.
Top-Three_flat	Sử dụng thuật toán chọn bản dịch một cách tuần tự, tạo lập câu truy vấn có cấu trúc bằng cách lập nhóm các bản dịch của từng từ khóa bằng toán tử OR, nối các nhóm bằng toán tử AND
Join-All	Lập nhóm các bản dịch kết xuất từ từ điển của từng từ khóa bằng toán tử OR, sau đó nối các nhóm bằng toán tử AND

5.3.2. Kết quả thực nghiệm

Bảng 5.2: So sánh các giải pháp dịch câu truy vấn

Cấu hình	***P@5***	***P@10***	***P@20***	***MAP***	***So sánh***
Baseline	0.636	0.562	0.514	0.3838	100%
Google	0.616	0.54	0.507	0.3743	97,52%
nMI	0.5	0.464	0.418	0.269	70,09%
SMI	0.496	0.478	0.427	0.2862	74,57%
Top_one_all	0.56	0.526	0.451	0.3245	84,55%
Top_three_all	0.64	0.582	0.52	0.3924	102,24%
Top_three_weight	0.64	0.592	0.52	0.3988	103,91%
Top-Three_flat	0.592	0.556	0.499	0.3737	97,37%
Join-All	0.612	0.574	0.509	0.3865	100,70%

5.3.3. Đánh giá

Giữa các phương pháp chỉ sử dụng một bản dịch tốt nhất cho mỗi từ khóa ở ngôn ngữ nguồn, cấu hình *SMI* cho kết quả tốt hơn so với cấu hình *nMI*, cấu hình *Top_one_all* sử dụng câu truy vấn có cấu trúc với trọng số cho kết quả tốt nhất.

Việc sử dụng các câu truy vấn có cấu trúc cho kết quả tốt hơn. Giữa các cấu hình sử dụng 3 bản dịch tốt nhất cho mỗi từ khóa, cấu hình *Top_three_weight* cho kết quả tốt nhất.

5.4. Thực nghiệm điều chỉnh câu truy vấn

5.4.1. Cấu hình thực nghiệm

Bảng 5.3: Cấu hình đánh giá kết quả điều chỉnh câu truy vấn

Cấu hình	*Diễn giải*
Baseline	
FW2_Top_three_all	Sử dụng thuật toán dịch câu truy vấn *Top_three_all*. Thực hiện điều chỉnh câu truy vấn.
FW2_Top_three_weight_A	Sử dụng thuật toán dịch câu truy vấn *Top_three_weight và* mở rộng câu truy vấn. Thực hiện việc tính lại trọng số từ khóa truy vấn.
FW2_Top_three_weight_B	Sử dụng thuật toán dịch câu truy vấn *Top_three_weight* và mở rộng câu truy vấn. Không thực hiện việc tính lại trọng số từ khóa truy vấn.
Top-Three_flat	Sử dụng thuật toán dịch câu truy vấn *Top-Three_flat* và mở rộng câu truy vấn.

5.4.2. Kết quả thực nghiệm

Bảng 5.4: So sánh các giải pháp điều chỉnh câu truy vấn

Cấu hình	*P@5*	*P@10*	*P@20*	*MAP*
Baseline	0.636	0.562	0.514	0.3838
FW2_Top_three_all	0.640	0.586	0.522	0.4261
FW2_Top_three_weight_A	0.644	0.586	0.522	0.4192
FW2_Top_three_weight_B	0.660	0.594	0.535	0.4312
FW2_Top-Three_flat	0.652	0.586	0.520	0.4220

5.4.3. Đánh giá

Bảng kết quả cho thấy việc áp dụng kỹ thuật điều chỉnh câu truy vấn giúp tăng hiệu quả của hệ thống với kết quả tốt nhất tương ứng với cấu hình kiểm thử *FW2_Top_three_weight_B*, tiếp theo là cấu hình *FW2_Top_three_all*.

5.5. Thực nghiệm xếp hạng lại

Các phương pháp học máy áp dụng lập trình di truyền do tác giả đề xuất được đánh giá và so sánh với một số phương pháp học xếp hạng khác được triển khai bằng công cụ RankLib.

5.5.1. Cấu hình thực nghiệm

Bảng 5.5: Cấu hình thực nghiệm học xếp hạng

Cấu hình	***Diễn giải***
SC-1	Áp dụng học xếp hạng, có sử dụng dữ liệu huấn luyện. Sử dụng phương án dịch câu truy vấn *FW2_Top_three_all*
UC3-1	Áp dụng học xếp hạng, không sử dụng dữ liệu huấn luyện. Sử dụng phương án dịch câu truy vấn *FW2_Top_three_all*
SC-2	Áp dụng học xếp hạng, có sử dụng dữ liệu huấn luyện. Sử dụng phương án dịch câu truy vấn *FW2_Top_three_weight_B*
UC3-2	Áp dụng học xếp hạng, không sử dụng dữ liệu huấn luyện. Sử dụng phương án dịch câu truy vấn *FW2_Top_three_weight_B*
MART	Sử dụng RankLib với phương pháp MART
Coordinate Ascent	Sử dụng RankLib với phương pháp *Coordinate Ascent*
Random Forests	Sử dụng RankLib với phương pháp *Random Forests*

5.5.2. Kết quả thực nghiệm

Điểm MAP trung bình cao nhất thuộc về 2 cấu hình học máy có huấn luyện SC-1 và SC-2. Các điểm này cao hơn điểm MAP trung bình tương ứng các thuật toán *MART*, *Coordinate Ascent* và *Random Forests* được triển khai với công cụ RankLib. Các cấu hình UC3-1 và UC3-2 cho kết quả điểm MAP trung bình tương ứng là 0.456 và 0.464.

5.5.3. Đánh giá

Bảng kết quả cho thấy hiệu quả của phương pháp học xếp hạng dựa trên lập trình di truyền. Các cấu hình học máy có giám sát *SC-1* và *SC-2* cho kết quả tương ứng là 0.476 và 0.484, bằng 123,96% và 126,09% so với phương án sử dụng bản dịch thủ công.

Điểm MAP trung bình của các cấu hình *UC3-1* và *UC3-2* không sử dụng dữ liệu huấn luyện tương ứng là 0.456 và 0.464, tăng tương ứng 7% và 7,7% so với điểm MAP của các cấu hình *FW2_Top_three_all* và *FW2_Top_three_weight_B.*

5.6. Đánh giá hiệu quả việc áp dụng các kỹ thuật đề xuất

Bảng 5.6: Đánh giá việc áp dụng các kỹ thuật đề xuất

Cấu hình	***Diễn giải***	***MAP***
Baseline	Các câu truy vấn được dịch thủ công	0.384
Google	Các câu truy vấn được dịch bằng cách sử dụng máy dịch Google	0.374
Các phương pháp dịch câu truy vấn		
SMI	Sử dụng thuật toán khử nhập nhằng SMI	0.286
Top_one_all	Kết xuất chỉ một bản dịch tốt nhất cho mỗi từ khóa	0.325
Top_three_all	họn bản dịch một cách tuần tự, kết xuất 3 bản dịch tốt nhất cho mỗi từ khóa, tạo lập câu truy vấn có cấu trúc	0.392
Top_three_weight	Chọn bản dịch một cách tuần tự, tạo lập câu truy vấn có cấu trúc với trọng số được xác định trong quá trình khử nhập nhằng	0.399
Các phương pháp điều chỉnh câu truy vấn		
FW2_Top_three_all	Sử dụng thuật toán dịch câu truy	0.427

	vấn *Top_three_all.* Áp dụng công thức FW2.	
FW2_Top_three_weight_B	Sử dụng thuật toán dịch câu truy vấn *Top_three_weight.* Áp dụng công thức FW2. Không thực hiện việc tính lại trọng số từ khóa truy vấn	0.431
Áp dụng học máy		
UC3 FW2_Top_three_all	Sử dụng hàm xếp hạng tổng hợp là kết quả học xếp hạng không giám sát cho cấu hình *FW2_Top_three_all.*	0.456
SC FW2_Top_three_all	Sử dụng hàm xếp hạng tổng hợp là kết quả học xếp hạng có giám sát cho cấu hình *FW2_Top_three_all*	0.476
UC3 FW2_Top_three_weight	Sử dụng hàm xếp hạng tổng hợp là kết quả học xếp hạng không giám sát cho cấu hình *FW2_Top_three_weight.*	0.464
SC FW2_Top_three_weight	Sử dụng hàm xếp hạng tổng hợp là kết quả học xếp hạng có giám sát cho cấu hình *FW2_Top_three_weight.*	0.484

Các phương pháp dịch câu truy vấn chọn lựa một bản dịch tốt nhất cho mỗi từ khóa *SMI* và *Top_one_all* cho kết quả điểm MAP tương ứng là 0.286 và 0.325, bằng tương ứng 74,48% và 84,64% so với điểm MAP của cấu hình *baseline* sử dụng bản dịch thủ công.

Khi áp dụng việc chọn lựa 3 bản dịch tốt nhất cho mỗi từ khóa, sau đó điều chỉnh câu truy vấn rồi áp dụng học xếp hạng, kết quả điểm MAP tiếp tục được nâng cao.

5.7. Tiểu kết chương

Trong chương 5, một môi trường thực nghiệm thống nhất được xây dựng và sử dụng nhằm kiểm tra hiệu quả của việc áp dụng các đề xuất kỹ thuật của tác giả cũng như so sánh với một số các giải pháp kỹ thuật khác. Kết quả cho thấy, qua mỗi bước áp dụng kỹ thuật dịch thuật, điều chỉnh câu truy vấn và học xếp hạng, hiệu quả của hệ thống (đo bằng điểm MAP) đều được cải thiện.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

1.1. Tóm tắt nội dung luận án

Nội dung luận án trình bày kết quả nghiên cứu của tác giả về các phương pháp xếp hạng trong tìm kiếm các trang Web xuyên ngữ. Tác giả nghiên cứu cơ sở lý thuyết và các kết quả nghiên cứu về truy vấn thông tin, truy vấn thông tin xuyên ngữ và bài toán xếp hạng lại. Trên cơ sở sơ đồ xử lý của hệ thống truy vấn thông tin, tác giả đề xuất mô hình xếp hạng trang Web trong tìm kiếm xuyên ngữ và xác định các nội dung nghiên cứu.

Lần lượt trong các chương 2, 3 và 4, tác giả đi sâu nghiên cứu các thành phần kỹ thuật về xử lý câu truy vấn, bài toán dịch tự động và xếp hạng lại và đề xuất các giải pháp kỹ thuật áp dụng tại các thành phần này, phục vụ mục tiêu nâng cao hiệu quả xếp hạng các trang Web của mô hình xếp hạng trang Web trong tìm kiếm xuyên ngữ. Trong chương 5, tác giả triển khai việc kiểm tra, đánh giá các kỹ thuật đề xuất trong một môi trường thực nghiệm thống nhất. Kết quả thực nghiệm cho thấy: qua mỗi bước áp dụng kỹ thuật dịch thuật, điều chỉnh câu truy vấn và học xếp hạng, hiệu quả của hệ thống (đo bằng độ đo MAP) đều được cải thiện.

1.1. Các kết quả đạt được

Lý thuyết

Các kết quả lý thuyết do tác giả đề xuất bao gồm hai nhóm kỹ thuật được áp dụng tại các bước của mô hình tìm kiếm Web xuyên ngữ.

Nhóm thứ nhất bao gồm các đề xuất kỹ thuật phục vụ dịch thuật, bao gồm các kỹ thuật tiền xử lý câu truy vấn, dịch câu truy vấn và xử lý câu truy vấn ở ngôn ngữ đích. Cụ thể:

- Đề xuất một phương pháp tiền xử lý câu truy vấn tại ngôn ngữ nguồn. Một cách cụ thể, tác giả đã đề xuất thuật toán WLQS, được sử dụng cùng công cụ mã nguồn mở *vnTagger*, thực hiện việc phân đoạn câu truy vấn thành các cụm từ cần dịch, đi kèm với các danh sách bản dịch ứng viên.

- Đề xuất các phương pháp khử nhập nhằng trong mô-đun dịch thuật. Tác giả đã giới thiệu hàm Summary Mutual Information phục vụ việc chọn bản dịch tốt nhất cho mỗi từ khóa truy vấn và thuật toán chọn bản dịch một cách tuần tự nhằm xác định danh sách các bản dịch tốt nhất được xếp theo thứ tự cho mỗi từ khóa truy vấn.

- Đề xuất các phương pháp xây dựng câu truy vấn tại ngôn ngữ đích. Tác giả đầu tiên đề xuất phương án xây dựng câu truy vấn có cấu trúc tại ngôn ngữ đích dựa trên danh sách bản dịch của các từ khóa truy vấn. Tiếp theo, tác giả đề xuất việc sử dụng kỹ thuật phản hồi ẩn, kết hợp với việc áp dụng các công thức khác nhau cho việc tính toán trọng số thuật ngữ chứa trong các văn bản, nhằm xây dựng lại câu truy vấn có cấu trúc và mở rộng câu truy vấn.

Nhóm kỹ thuật thứ hai là các kỹ thuật phục vụ xếp hạng lại danh sách kết quả tìm kiếm trong truy vấn xuyên ngữ, chú trọng việc xếp hạng các trang Web. Cụ thể:

- Đề xuất các mô hình lân cận (proximity model) xuyên ngữ. Hai mô hình được xây dựng trên nền tảng của các mô hình lân cận đơn ngữ Büttcher và Rasolofo. Một mô hình khác được định nghĩa dựa trên việc xem xét các câu trong tài liệu chứa nhiều từ khóa truy vấn.

- Đề xuất phương pháp xếp hạng lại kết quả tìm kiếm Web. Trên cơ sở sử dụng máy tìm kiếm *Solr*, tác giả phân tích các tập tin HTML thành các trường và tạo lập đa chỉ mục cho các tài liệu. Một danh sách các hàm xếp hạng được định nghĩa và được áp dụng như

các hàm xếp hạng cơ sở đối với các tài liệu trong danh sách kết quả tìm kiếm ban đầu. Cuối cùng, kỹ thuật học máy ứng dụng lập trình di truyền được áp dụng nhằm xây dựng hàm xếp hạng tổng hợp cho từ các hàm xếp hạng cơ sở để xếp hạng lại danh sách tài liệu.

Các đề xuất nêu trên được tích hợp như các thành phần trong mô hình tìm kiếm Web xuyên ngữ, đảm bảo việc hoàn thành kế hoạch nghiên cứu của tác giả .

Thực nghiệm

Các kết quả thực nghiệm được kiểm chứng và được trình bày tại các bài báo khoa học bao gồm:

- Kết quả thực nghiệm mô hình tìm kiếm áp dụng thuật toán phân đoạn WLQS và hàm Summary Mutual Information phục vụ việc khử nhập nhằng cho thấy hàm này có thể tạo kết quả tốt hơn so với việc áp dụng công thức nMI thường được sử dụng với cùng mục tiêu chọn bản dịch tốt nhất cho các từ khóa truy vấn.

- Kết quả thực nghiệm mô hình truy vấn xuyên ngữ kết hợp áp dụng thuật toán phân đoạn WLQS và công cụ *vnTagger* phục vụ phân đoạn câu truy vấn, quá trình chọn lọc các bản dịch tốt cho các từ khóa truy vấn dựa trên thuật toán chọn bản dịch một cách tuần tự tại bước khử nhập nhằng và xây dựng câu truy vấn có cấu trúc tại ngôn ngữ đích cho kết quả vượt trội so với việc sử dụng máy dịch *Google Translate*.

- Kết quả thực nghiệm việc áp dụng phản hồi ẩn để điều chỉnh và mở rộng câu truy vấn cho thấy kỹ thuật được đề xuất cho phép tăng hiệu quả của hệ thống truy vấn cả ở độ chính xác (precision) và độ bao phủ (recall).

- Trên cơ sở kết quả thực nghiệm việc học xếp hạng với bộ dữ liệu thực nghiệm truyền thống LETOR của Microsoft và kết quả thực nghiệm việc áp dụng mô hình lân cận trong truy vấn xuyên ngữ, tác giả đã tiến hành thực nghiệm hệ thống học xếp hạng phục vụ tìm kiếm Web xuyên ngữ, trên cơ sở áp dụng kỹ thuật học máy dựa trên lập trình di truyền và các hàm xếp hạng cơ sở được định nghĩa cho

các thành phần khác nhau của các tập tin HTML. Kết quả thực nghiệm, hệ thống đề xuất có hiệu năng tốt hơn (với độ đo MAP) so việc áp dụng dịch thủ công.

Tóm lại, việc áp dụng các kỹ thuật tại từng thành phần giúp từng bước nâng cao hiệu quả xếp hạng của hệ thống. Kết quả quan trọng của luận án là với việc áp dụng đồng thời các thành phần, chất lượng xếp hạng các trang Web trong tìm kiếm xuyên ngữ được nâng cao và vượt kết quả xếp hạng sử dụng phương pháp dịch thủ công tại các thực nghiệm đã tiến hành.

2. Hướng phát triển

Bên cạnh các kết quả đạt được, tác giả xác định hướng phát triển của luận án tập trung giải quyết các vấn đề sau:

- Các thuật toán xử lý câu truy vấn được trình bày trong luận án có thể rất nhạy cảm với loại ngôn ngữ, nội dung, kích thước câu truy vấn. Trong khuôn khổ giới hạn về thời gian, tác giả chỉ tập trung nghiên cứu mô hình tìm kiếm với câu truy vấn tiếng Việt và văn bản cần tìm kiếm tiếng Anh. Các câu truy vấn được chú trọng thực nghiệm là các câu truy vấn có độ dài trung bình, các trường hợp câu truy vấn ngắn và câu truy vấn dài chưa được xem xét. Hướng nghiên cứu tiếp theo là mở rộng, hoàn chỉnh việc đánh giá thực nghiệm với các cặp ngôn ngữ khác và với độ dài câu truy vấn khác nhau.

- Tối ưu hóa các thuật toán tiền xử lý câu truy vấn, khử nhập nhằng. Thời gian xử lý đối với các thuật toán xử lý câu truy vấn, khử nhập nhằng cần được cải thiện bằng cách tổ chức tốt hơn cấu trúc dữ liệu cũng như tối ưu hóa thuật toán.

- Nghiên cứu việc áp dụng các kỹ thuật học máy khác, xây dựng các tổ hợp hàm xếp hạng cơ sở khác. Hạn chế của học máy dựa trên lập trình di truyền là chi phí thời gian khá lớn. Bên cạnh đó, luận án mới tập trung xem xét một danh sách các hàm cơ sở hạn chế.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

[1] Giang L.T., Hùng V.T., "Các phương pháp xếp hạng lại trong trộn kết quả tìm kiếm". *Tạp chí Khoa học và Công nghệ các trường Đại học Kỹ thuật*, vol. 91, pp. 59–64, 2012.

[2] Giang L.T., Hùng V.T., "Ứng dụng lập trình di truyền trong học xếp hạng". *Tạp chí Khoa học và Công nghệ các trường Đại học Kỹ thuật*, vol. 92, pp. 58–63, 2013.

[3] Giang L.T., Hùng V.T., "Đánh giá thực nghiệm mô hình truy vấn thông tin đa ngữ". *In: Hội nghị quốc gia lần thứ VI Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin*, pp. 103–107, 2013.

[4] Giang L.T., Hung V.T., Phap H.C., "Building Evaluation Dataset in Vietnamese Information Retrieval". *Journal of Science and Technology Danang University*, vol. 12, no. 1, pp. 37–41, 2013.

[5] Giang L.T., Hung V.T., Phap H.C., "Experiments with query translation and re-ranking methods in Vietnamese-English bilingual information retrieval". *In: Proceedings of the Fourth Symposium on Information and Communication Technology*, pp. 118–122, 2013.

[6] Giang L.T., Hung V.T., Phap H.C., "Building Structured Query in Target Language for Vietnamese – English Cross Language Information Retrieval Systems". *International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 04, pp. 146–151, 2015.

[7] Giang L.T., Hung V.T., Phap H.C., "Improve Cross Language Information Retrieval with Pseudo-Relevance Feedback". *In: FAIR 2015*, pp. 315–320, 2015.

[8] Giang L.T., Hung V.T., Phap H.C., "Building proximity models for Cross Language Information Retrieval". *Issue on Information and Communication Technology- University of Danang*, vol. 1, no. 1, pp. 8–12, 2015.

[9] Giang L.T., Hùng V.T., Pháp H.C., "Áp dụng học máy dựa trên lập trình di truyền trong tìm kiếm Web xuyên ngữ". *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng*, vol. 1, no. 98, pp. 93-97, 2016.