XÂY DỰNG HỆ THỐNG THÔNG TIN TRA CỨU TỪ ĐIỂN CHUYÊN NGÀNH CÓ NGỮ CẢNH

BUILDING A INFORMATION SYSTEM FOR LOOKING UP CONTEXTUAL TECHNICAL DICTIONARY

Hồ Trung Thành, Trần Thị Ánh
Trường Đại học Kinh tế - Luật, ĐHQG HCM - *Email: thanhht@uel.edu.vn*
Nguyễn Khánh Hoà
Trường Đại học RMIT

(Bài báo nhận ngày 28 tháng 07 năm 2015, hoàn chỉnh sửa chữa ngày 12 tháng 09 năm 2015)

TÓM TẮT

Ngữ cảnh của từ điển chuyên ngành là rất quan trọng. Ngữ cảnh là một phần thông tin bằng văn bản giúp cho người tra từ hiểu rõ nội dung ý nghĩa của từ khoá nhằm giúp việc sử dụng từ đúng vào từng trường hợp cụ thể trong văn bản chuyên ngành, đặc biệt là trong học tập, nghiên cứu. Tuy nhiên, các hệ thống tra cứu từ hiện tại thường tập trung hỗ trợ tra cứu từ và giải thích từ mà chưa quan tâm đến ngữ cảnh của từ. Khi có được ngữ cảnh của từ, câu hỏi đặt ra là làm thế nào để có thể tìm kiếm được chính xác ngữ cảnh hoặc hiển thị kết quả tìm kiếm gợi ý có liên quan đến từ khoá trong kho dữ liệu văn bản ngữ cảnh? Trong bài báo này, chúng tôi đề xuất xây dựng phương pháp và mô hình tra từ điển chuyên ngành có ngữ cảnh trên cơ sở phân tích, đánh giá và lựa chọn giải thuật tối ưu trong các phương pháp so khớp văn bản. Sau đó, chúng tôi áp dụng giải thuật vào kỹ thuật tra từ của hệ thống. Tích hợp mô hình đề xuất trên hệ thống website và thực nghiệm trên 1500 từ chuyên ngành cùng với ngữ cảnh thuộc lĩnh vực Hệ thống thông tin quản lý và Thương mại điện tử. Hệ thống có thể hỗ trợ cùng lúc việc tra từ điển bằng tiếng Anh và tiếng Việt.

Từ khoá: *Giải thuật so khớp mẫu, hệ thống thông tin, từ điển, chuyên ngành, ngữ cảnh.*

ABSTRACT

The context of technical terms is very important. It is part of information in text which supports users in understanding the exact meaning of technical terms in particular specialized circumstances, especially in education and research. However, most of current dictionary systems only focus on the lookup function and the standard meaning of terms without considering related contexts. In this paper, we proposed the model for searching technical terms and context of terms based on analyzing, evaluating and choosing an optimal algorithm in pattern matching technique. Then, the model was integrated on a dictionary system and experimented on 1500 terms in the context of information system and electronic commerce. This dictionary system supports searching with technical terms both in Vietnamese and English.

Keywords: *Pattern - matching algorithm, information system, dictionary, technical term, context.*

1. GIỚI THIỆU

Dựa trên nền tảng phát triển Internet, hiện nay có rất nhiều công cụ hỗ trợ việc tra cứu nghĩa của từ tiếng Anh và nhiều ngôn ngữ khác. Tại Việt Nam, có thể dễ dàng tìm thấy nhiều sách từ điển Anh - Việt được xuất bản,

các phần mềm tra cứu như Lạc Việt[1], hay nhiều website hỗ trợ tra cứu online như: Vdict[2], Google translate[3], tratu.soha[4], oxford dictionaries.com[5], Dictionary.com[6]… Các website này có thể cung cấp đầy đủ từ điển mà người dùng cần. Đa phần các website đều có cấu trúc tương đối giống nhau với giao diện thân thiện, dễ sử dụng. Các hệ thống website này hỗ trợ tra cứu trên nhiều bộ từ điển như: Anh - Việt, Việt - Anh, Anh - Anh và một số ngôn ngữ khác như: Trung Quốc, Nhật, Pháp… Các hệ thống website từ điển hầu hết đều hỗ trợ dịch nghĩa của một từ và cả đoạn văn dài. Người dùng có thể truy cập vào hệ thống, thực hiện tra từ và hệ thống sẽ cung cấp một danh sách các nghĩa của từ, kể cả từ đồng nghĩa, từ liên quan,... Trong số các website tra từ điển tác giả đã khảo sát trên, trong đó một số website cho phép tra cứu từ chuyên ngành như thefreedictionary.com, whatis.techtarget.com, cambridge.org[7]… Các website này có công cụ lọc theo từng lĩnh vực chuyên ngành cụ thể. Trong đó, whatis.techtarget.com là một website hỗ trợ tra cứu và cho kết quả là định nghĩa các từ liên quan đến kỹ thuật và công nghệ; tudienthuoc.net[8] và ykhoanet.com là các website từ điển chuyên ngành về thuốc, y khoa; latin-phrases.co.uk/[9] là từ điển về câu thành ngữ; fetp, tratu.soha là các website hỗ trợ tra từ chuyên ngành về kinh tế; www.lawyerintl.com/law-dictionary[10] chuyên về lĩnh vực luật học; và một số website như tratu.soha, Vdict,.. hỗ trợ tra từ thuộc nhiều lĩnh vực như tin học, kinh tế, luật,… Tuy nhiên, hầu như vẫn chưa có hệ thống website nào hỗ trợ tra từ điển Anh - Việt, Việt - Anh thuộc chuyên ngành thương mại điện tử và hệ thống thông tin quản lý, đây là một trong những đóng góp của chúng tôi trong nghiên cứu này.

Bên cạnh đó, các website hỗ trợ tra từ hiện tại chỉ dừng lại ở mức độ giải thích nghĩa của từ hay định nghĩa từ mà chưa quan tâm đến ngữ cảnh của từ chuyên ngành giúp hiểu rõ cách sử dụng từ trong trường hợp cụ thể. Để hiểu rõ được từ chuyên ngành, chúng tôi đề xuất xây dựng một hệ thống dữ liệu ngữ cảnh văn bản tương ứng với từng từ chuyên ngành. Tuy nhiên, việc xây dựng hệ thống dữ liệu ngữ cảnh của từ sẽ làm hạn chế tốc độ xử lý trên hệ thống website tra từ điển vì phải tìm kiếm trên hệ thống dữ liệu ngữ cảnh để trả lời kết quả cho yêu cầu tìm kiếm từ người dùng. Vì thế, ngoài những yếu tố ảnh hưởng đến kết quả tìm kiếm như phần cứng, băng thông, thiết kế… yêu cầu đặt ra của một hệ thống tìm kiếm là tốc độ xử lý và sự chính xác. Chúng tôi quan tâm đến việc xử lý bên trong hệ thống để có kết quả chính xác hơn. Để xử lý trên bộ ngữ liệu tiếng Việt (gồm từ chuyên ngành và ngữ cảnh của từ) và quá trình tìm kiếm, chúng tôi phải tìm ra sự liên kết giữa các từ dựa trên các ngữ cảnh khác nhau. Vì thế, giải thuật tìm kiếm là một trong những yếu tố quan trọng để đáp ứng yêu cầu về tốc độ trong tra cứu từ. Trong nghiên cứu này, dựa trên cơ chế tìm kiếm từ có ngữ cảnh hay nói cách khác là tìm kiếm từ trong văn bản, chúng tôi lựa chọn bài toán so sánh mẫu để giải quyết yêu cầu tìm kiếm đặt ra. Các giải thuật tìm kiếm như KMP [5], Naïve [7], Rubin – Karp [12] được chúng tôi khảo sát và so sánh để tìm ra giải thuật phù hợp nhất trong việc giải quyết yêu cầu. Chi tiết của việc khảo sát và so sánh các thuật giải sẽ được trình bày trong phần 2.

[1] http://tratu.coviet.vn/
[2] http://vdict.com/
[3] https://translate.google.com.vn/?hl=vi&tab=wT&authuser=0
[4] http://tratu.soha.vn/
[5] http://www.oxforddictionaries.com/
[6] http://dictionary.reference.com/
[7] http://dictionary.cambridge.org/
[8] http://www.tudienthuoc.net
[9] www.lawyerintl.com/law-dictionary
[10] http://www.lawyerintl.com/law-dictionary/

Mục tiêu tiếp theo trong hệ thống tìm kiếm từ điển chuyên ngành của chúng tôi là hỗ trợ chức năng tìm các từ chuyên ngành có liên quan. Để đạt được mục tiêu này, ngoài phương pháp sử dụng giải thuật, có thể sử dụng câu truy vấn thông thường như SQL. Tuy nhiên việc phải so sánh từ chuyên ngành với số lượng ngữ cảnh lớn sẽ dẫn đến tốc độ xử lý chậm trong quá trình tìm kiếm [5][11][15]. Việc trả về các dữ liệu không cần thiết (nếu không tìm thấy từ khóa trong ngữ cảnh) cũng là một nguyên nhân khiến những câu truy vấn chậm [5]. Chính vì vậy, trước khi đưa dữ liệu vào quá trình tìm kiếm, chúng tôi phải thực hiện trước việc lọc những từ trong stopwords[11] để cải thiện tốc độ tìm kiếm. Phần 2 của bài báo sẽ trình bày về các nghiên cứu liên quan. Trong phần 2, các giải thuật sẽ được so sánh và giải thuật phù hợp sẽ được chọn. Trong phần 3, chúng tôi đề xuất mô hình và phương pháp tra từ có ngữ cảnh. Phần 4 sẽ trình bày việc thử nghiệm và thảo luận kết quả. Cuối cùng là kết luận và hướng phát triển nghiên cứu.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng tôi tập trung khảo sát các kỹ thuật và phương pháp liên quan đến tìm kiếm và so sánh mẫu từ trong văn bản tiếng Việt. Từ đó, chúng tôi lựa chọn kỹ thuật phù hợp áp dụng và mô hình đề suất cho hệ thống tra từ chuyên ngành có ngữ cảnh.

Phương pháp so sánh chuỗi là phương pháp tìm kiếm tất cả các lần xuất hiện của một chuỗi mẫu (pattern) trong một chuỗi khác [1], [2], [8], [15], [17], [20]. Quá trình so sánh chuỗi là hoạt động diễn ra rất thường xuyên trong các chương trình chỉnh sửa văn bản, các trình duyệt web, các bộ máy tìm kiếm, và các hệ thống gợi ý trên các trang thương mại điện tử [9][16]. Trong nghiên cứu này, chúng tôi khảo sát các giải thuật so sánh chuỗi cho bài toán của hệ thống từ điển chuyên ngành có ngữ cảnh.

Bài toán đặt ra, với một bộ dữ liệu từ điển có số lượng từ khóa lớn, kèm theo đó là ngữ cảnh trong từng trường hợp sử dụng từ, làm sao để xác định mối liên hệ giữa các từ với nhau? Ngoài ra, việc tra từ có ngữ cảnh đòi hỏi phương pháp tra từ phải làm việc trên một lượng dữ liệu lớn là văn bản (ngữ cảnh). Vậy làm sao để có thể tra từ nhanh và trả về nghĩa và ngữ cảnh của từ tìm kiếm chính xác? Để giải quyết các vấn đề trên, trong phần này chúng tôi tập trung khảo sát các phương pháp, giải thuật so khớp mẫu với ba giải thuật để từ đó chọn ra một phương pháp hỗ trợ tốt cho việc xây dựng hệ thống từ điển chuyên ngành có ngữ cảnh. Cụ thể, chúng tôi đã khảo sát các giải thuật Naive [7], giải thuật Rabin - Karp [3], [12] và giải thuật Knuth – Morris - Pratt (KMP)[5] dựa trên một mô tả bài toán sau:

"Cho mẫu P có độ dài M và văn bản S có độ dài N trên cùng bảng chữ A. Tìm một (hoặc tất cả) các lần xuất hiện của mẫu P trong S". Với việc xuất hiện một bài toán so sánh mẫu như trên, giải thuật nào là phù hợp để giải bài toán với thời gian tìm kiếm có giới hạn?

Trong bài toán trên, giả sử ta có tập văn bản S'= [S, S_1, S_2…S_n], lúc này bài toán sẽ được thực hiện đối với mỗi cặp [P,S] [P,S_1] [P,S_2]… Trong trường hợp độ dài N của văn bản S_x là rất lớn và tập S' có n phần tử con (n rất lớn) thì thời gian tìm kiếm sẽ rất tốn kém. Do đó, việc tìm hiểu một giải thuật để giải quyết vấn đề là cần thiết. Dựa vào việc phân tích, thiết kế, xây dựng bộ dữ liệu, chúng tôi có một số nhận xét sau:

Độ dài N của văn bản S_x (phần tử của tập ngữ cảnh) là không quá lớn.

Tập S' gồm khoảng 1000 phần tử (và có thể phát triển nhiều hơn).

[11] Stopwords là những từ, cụm từ phổ biến hay nói chung chung không có ý nghĩa trong kết quả tìm kiếm

2.1. Giải thuật Naive

Đây là giải thuật cơ bản và đơn giản nhất, sử dụng nguyên lý vét cạn. Giải thuật Naive [7] kiểm tra tất cả các khả năng của chuỗi mẫu P[1..m] nằm trong chuỗi S[1..n] bằng cách duyệt từ đầu tới cuối chuỗi S.

Giải thuật 1. Naive Algorithm [7]

NAIVE-STRING-MATCHER(S, P)

1. *n = S.length*
2. *m = P.length*
3. *for s = 0 to n-m do*
4. *j = 1*
5. *while (j <= m AND S[s+j] == P[j]) do*
6. *j = j+1*
7. *if (j > m)*
8. *"Tìm thấy mẫu với độ dịch chuyển s"*

Nhận xét: Vòng lặp while bên trong chạy tối đa m lần, vòng lặp for bên ngoài chạy tối đa n-m+1 lần. Do vậy, thời gian chạy của giải thuật này là S(n) = O((n-m+1)*m) = O(n*m). Rõ ràng, giải thuật này không hiệu quả vì bỏ qua mọi thông tin hữu ích có được trong quá trình so sánh chuỗi tại từng giá trị của S.

2.2. Giải thuật Rabin - Karp

Giải thuật này do Rabin và Karp đề xuất trong [3][12]. Giải thuật với độ phức tạp O(m) để tiền xử lý các dữ liệu nhập, và thời gian chạy tệ nhất là O((n-m+1)m). Mặc dù vậy, trung bình các trường hợp đều tiêu tốn thời gian ít hơn.

Ta nhận thấy rằng mỗi chuỗi S có thể số hóa thành một số. Ví dụ S = {0,1,2..,9}, S = "1234" thì ta sử dụng hàm digit(S) = 1,234. Gọi p là giá trị số hóa của P, hay nói cách khác p là giá trị thập phân tương ứng của P. Gọi t_s là giá trị thập phân tương ứng của T[s+1,…,s+m] , s<n-m+1. Ta nhận thấy rằng t_S =p khi và chỉ khi P = T[s+1,…,s+m].

Mặt khác, ta có thể tính p và t_0 theo 2 công thức:

$$p = P[m] * 10^0 + P[m-1]*10^1 + \ldots. + P[1]*10^{m-1} \quad (1)$$

$$t_0 = T[1]*10^{m-1} + T[2]*10^{m-2} + \ldots. + T[m]*10^0 \quad (2)$$

Hai công thức trên cho cùng tiêu tốn thời gian là O(m). Sau khi tính t_0, việc tính các $t_1,t_2,..t_{n-m-1}$ trở nên đơn giản hơn và chỉ tiêu tốn O(1) cho mỗi t_i. Ta tính các $t_1,t_2,\ldots t_{n-m-1}$ lần lượt theo công thức sau:

$$t_i = 10*(t_{i-1} - 10^{m-1}*T[i]) + T[i+m] \quad (3)$$

Sau khi tính được các giá trị của p và t_i, bài toán trở nên đơn giản khi được qui về bài toán "tìm một số trong một mảng số nguyên". Điều này có nghĩa là tìm sự xuất hiện của p trong t_i .

Vì vậy để tính được tất cả các giá trị p và t_i, hay nói cách khác là tìm được chuỗi P trong T sẽ có độ phức tạp là O(m) + O(n-m-1). Và điều này cũng cho thấy giải thuật Rabin - Karp với thời gian tiêu tốn là O(m) cho tiền xử lí và O(n-m-1) để so sánh chuỗi.

Nhận xét:

Quá trình tiền xử lý với giải thuật tiêu tốn O(m) thời gian, với 1 vòng lặp là for i =1 to n.

Quá trình so sánh trong trường hợp tốt nhất là p'! = t'_i với mọi i thì việc so sánh với thời gian tiêu tốn là O(n-m). Tuy nhiên, trong trường hợp xấu nhất khi p' = t'_I thì việc so sánh phải thực hiện thêm lệnh kiểm tra P[1..m] và T[i+1,i+m], điều này có thời gian tiêu tốn là O(m). Như vậy, độ phức tạp của Rabin - Karp là O((n-m+1)*m).

Giải thuật Knutt – Morris - Pratt được trình bày trong các phần sau tỏ ra tốt hơn nhiều so với Naïve và Rabin - Karp Algorithm vì tận dụng các thông tin hữu ích khi tìm kiếm.

2.3. Giải thuật Knuth – Morris - Pratt (KMP)

Giải thuật KMP [5] với độ phức tạp tuyến tính này được Knuth, Morris và Pratt phát hiện ra nhờ việc phân tích chặt chẽ giải thuật Naive[8]. Giả sử ta muốn tìm chuỗi mẫu P[1..m] trong S[1..n], đến một lúc nào đó thì ta sẽ có P[i] != S[j].

Xét về độ phức tạp và chạy thực tế với 4 mẫu thử ngẫu nhiên cho 3 giải thuật (Naïve, Rabin - Karp và KMP), nhận thấy KMP có những ưu điểm vượt trội so với 2 giải thuật còn lại là Naive và Rabin - Karp, kết quả thử nghiệm thực tế cũng nói lên rằng, dù kích cỡ Text (T) và từ khóa (Paragraph - P) có khác nhau giữa các mẫu thử thì giải thuật KMP luôn đạt hiệu suất trung bình tốt nhất.

Trong quá trình tiền xử lý chuỗi P, mỗi p[i], $1 \le i \le m$, lưu lại độ dài của biên rộng nhất của P[1..i]. Vì chuỗi rỗng không có biên nên ta gán: p[0] = -1. Giả sử các giá trị p[0], …, p[i] đã biết, giá trị p[i+1] sẽ được tính bằng cách kiểm tra xem biên của chuỗi P[1..i] có thể được mở rộng bằng ký tự P[i+1] hay không. Ta sử dụng biến k lưu trữ các p[i]. Nếu P[i+1] = P[k] thì khi đó ta gán p[i+1] = k+1, ngược lại ta xét k = p[k] và quay lại các bước so sánh P[i+1] với P[k] ở trên.

Giải thuật so khớp chuỗi KMP - Matcher được trình bày trong đoạn mã giả sau đây. Giải thuật này gọi tới giải thuật tiền xử lý Compute - Prefix - Function để tính p.

Giải thuật 2. KMP - Matcher [5]

KMP - Matcher*(S, P)*
1. *n* ← *length*[*S*]
2. *m* ← *length*[*P*]
3. *π* ← Compute – Prefix - Function*(P)*
4. *q* ← 0 *//Số lượng ký tự trùng nhau*
5. for *i* ← 1 to *n //Duyệt chuỗi S từ trái qua phải*
6. do while *q* > 0 and *P*[*q* + 1] ≠ *S* [*i*]
7. do *q* ← *π*[*q*] *//Ký tự không trùng nhau*
8. if *P*[*q* + 1] = *S* [*i*]
9. then *q* ← *q* + 1 *//Ký tự trùng nhau*
10. if *q* = *m //Nếu đã kiểm tra toàn bộ chuỗi P*
11. then print "Mẫu xuất hiện với độ dịch chuyển" *i – m*
12. *q* ← *π*[*q*] *//Tìm ký tự trùng nhau tiếp theo*

Giải thuật 3. Compute – Prefix - Function[5]

Compute - Prefix - Function(P)
1. m ← length*[P]*
2. π[1] ← 0
3. k ← *0*
4. for q ← *2 to* m *do*
5. while k > *0 and* P*[*k + *1]* ≠ P*[*q*]*
6. do k ← π*[*k*]*
7. if P*[*k + *1]* = P*[*q*]*
8. then k ← k + *1*
9. π*[*q*]* ← k
10. return π

Nhận xét: Độ phức tạp của giải thuật tiền xử lý Compute – Prefix - Function là O(m) bởi vì vòng lặp while bên trong sẽ không bao giờ thực hiện quá m lần. Tương tự, giải thuật tìm kiếm KMP - Matcher cũng chỉ có độ phức tạp là O(n).

2.4. Đánh giá các giải thuật

Sau khi phân tích các giải thuật trên, cần đánh giá và lựa chọn giải thuật phù hợp với yêu cầu đặt ra một cách tổng quát như sau:

Bảng 1. Kết quả đánh giá các giải thuật

Tên giải thuật	Thực hiện tiền xử	Độ phức tạp
Naïve	No	O((n-m+1)*m) = O(n*m)
Rubin-Karp	Yes	O((n-m+1)*m)
KMP	Yes	O(n)

Xét về độ phức tạp nhận thấy rằng, KMP có độ phúc tạp thấp hơn với 2 giải thuật còn lại. Thông qua những phân tích và so sánh trên, trong nghiên cứu này chúng tôi chọn giải thuật KMP để làm cơ sở giải quyết bài toán đã đặt ra cho việc tìm kiếm từ điển và ngữ cảnh từ.

3. ĐỀ XUẤT MÔ HÌNH KHAI THÁC NGỮ CẢNH VÀ TÌM TỪ GỢI Ý

Trong phần này, chúng tôi trình bày bài toán về khai thác ngữ cảnh của từ điển. Chúng ta xét bài toán cụ thể sau, giả sử ta có từ khóa cần tìm "tiếp thị", và có 2 ngữ cảnh có liên quan đến từ "tiếp thị" như sau :

Ngữ cảnh (1) có chứa cụm từ "tiếp thị trực tiếp": "Thị trường sản phẩm và dịch vụ Dell là các doanh nghiệp từ nhỏ đến trung bình và người tiêu dùng chính qua các kênh quảng cáo trên truyền hình và Internet, qua các phương tiện truyền thông in ấn, và bằng cách gởi các ấn phẩm tiếp thị trực tiếp, như các mẩu quảng cáo, catalog, và các bản tin khách hàng. Tại các địa điểm nhất định, chúng cũng được đưa vào cửa hàng Dell hay những ki-ốt nằm trong trung tâm mua sắm. Điều đó cho phép khách hàng có thể xem sản phẩm và mua hàng trực tuyến với sự trợ giúp của một chuyên gia Dell".

Ngữ cảnh (2) có chứa cụm từ "tiếp thị tương tác": "Một tính năng quan trọng của tính tương tác truyền thông tiếp thị là chúng có thể được thiết kế cho cá nhân, không giống như các phương tiện truyền thông truyền thống, nơi cùng một thông điệp có xu hướng được phát sóng đến tất cả mọi người. Quá trình thiết kế riêng cũng được gọi là cá nhân hóa và là một khía cạnh quan trọng của việc đạt được quản lý quan hệ khách hàng trực tuyến".

Hai nội dung ngữ cảnh trên liên quan đến từ khóa "tiếp thị", tuy nhiên lại theo từng ngữ cảnh và ý nghĩa khác nhau thuộc 2 lĩnh vực: "tiếp thị trực tiếp", và "tiếp thị tương tác". Vậy, cơ sở chúng tôi đưa ra cách giải quyết bài toán trên là dựa vào ngữ cảnh từ, từ đó người dùng có thể hiểu hơn về cách sử dụng từ khóa cần tìm trong từng trường hợp, và gia đình từ (word family) của từ khóa đó.

Một ý nghĩa quan trọng nữa trong mô hình đề xuất (xem giải thuật 4) là khai thác ngữ cảnh để tìm ra từ gợi ý giúp cho người tra từ mở rộng thêm kiến thức liên quan đến từ đã tra cứu.

Giải thuật 4. Khai thác ngữ cảnh từ điển chuyên ngành và tìm từ gợi ý.

Đầu vào: từ khóa và ngữ cảnh của từ.

Đầu ra: thông tin liên quan từ khoá, tập những từ khóa được gợi ý có chứa từ khóa đầu vào và ngữ cảnh từ

Xử lý: Bài toán tra từ điển có ngữ cảnh và tìm từ gợi ý được chuyển về một dạng toán kinh điển trong giải thuật so sánh mẫu.

Ý tưởng, nếu người dùng nhập từ khóa gồm hai chữ hoặc nhiều hơn hai chữ, quy trình xử lý như sau:

Bước 1: Lọc những từ không có ý nghĩa, sau đó tìm kiếm theo những từ còn lại. Việc này sẽ xảy ra trường hợp có thể có nhiều từ có ý nghĩa.

Bước 2: Thực hiện tách từ tìm kiếm

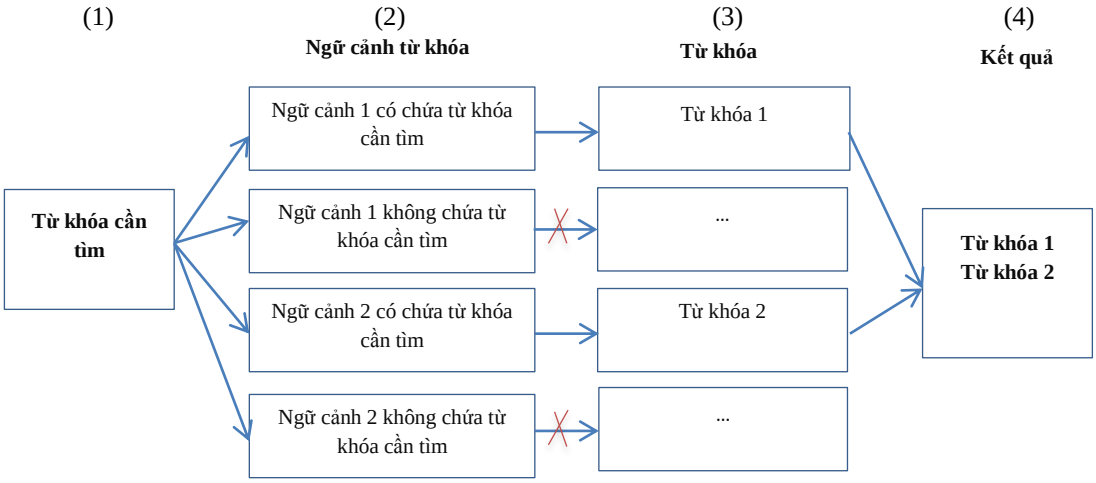
Bước 2.1: Nếu trong một ngữ cảnh có tất cả những từ được tách sẽ trả về kết quả từ gợi ý hay gọi là từ tương đương. Qua bước 2.3.

Bước 2.2: Nếu những từ được tách không xuất hiện trong ngữ cảnh, thì kết quả sẽ không trả về từ gợi ý hay từ tương đương. Quan bước 3.

Bước 2.3: Bổ sung danh sách từ gợi ý có liên quan và hiển thị ngữ cảnh của từ.

Bước 3: Kết thúc tìm kiếm từ

Hình 2 trình bày quy trình khai thác ngữ cảnh và tìm kiếm từ gợi ý, gồm bốn thành phần: (1) trình bày về từ khoá cần tìm, (2) trình bày về nội dung ngữ cảnh có liên quan đến từ chuyên ngành cần tìm, (3) gồm các từ chuyên ngành có chứa từ khoá cần tìm và (4) trình bày kết quả tìm kiếm các từ liên quan đến từ khoá và có chứa từ khoá cần tìm.



Hình 2. Quy trình khai thác ngữ cảnh và tìm kiếm từ gợi ý

Trên thực tế bài toán so sánh mẫu nhằm xác định vị trí của một "từ" hoặc "cụm từ" trong một đoạn văn bản cho trước. Với cách xác định này, chúng tôi sẽ sử dụng giải thuật để xác định ngữ cảnh nào có chứa từ khóa, từ đó đưa ra từ gợi ý. Bên cạnh đó, dạng phổ biến nhất của bài toán so khớp chuỗi như sau: cho trước nguồn tìm kiếm là một tập D các văn bản. Cho một q - một từ, hoặc một cụm từ, tìm tất cả các văn bản thuộc D mà có chứa q. Để thực hiện bài toán, hệ thống phải kiểm tra văn bản xem q có là một cụm từ thuộc các văn bản thuộc tập D hay không và đưa ra các văn bản gợi ý.

Trong phần tiếp theo, chúng tôi áp dụng mô hình, ý tưởng đề xuất trên và sử dụng các kỹ thuật để xây dựng hệ thống thông tin website hỗ trợ tra cứu từ điển chuyên ngành có ngữ cảnh.

4. THỬ NGHIỆM VÀ THẢO LUẬN KẾT QUẢ

4.1. Dữ liệu từ điển và ngữ cảnh từ

Mục tiêu đặt ra là xây dựng bộ từ điển chuyên ngành có ngữ cảnh trong lĩnh vực HTTTQL và TMĐT. Dữ liệu được xây dựng từ các giáo trình, sách và bài viết chính thức của từng chuyên ngành. Trong quá trình xây dựng bộ dữ liệu phục vụ cho việc tìm kiếm từ điển, chúng tôi sử dụng các "bản thuật ngữ" của các sách chuyên ngành HTTTQL và TMĐT [6], [10], [13], [18]. Tất cả sách trên được viết bằng tiếng Anh. Với mỗi từ tiếng Anh, "bản thuật ngữ" cũng cung cấp các giải thích ý nghĩa của từ bằng tiếng Anh. Tuy nhiên, để phục vụ nhu cầu tra từ theo nghĩa tiếng Việt, cần tiến hành dịch thuật những giải thích bằng tiếng Anh nêu trên sang tiếng Việt. Quá trình tìm hiểu ngữ nghĩa của từ, cụm từ và thuật ngữ chuyên ngành không chỉ dựa trên cơ sở ngữ nghĩa của

"bản thuật ngữ" cung cấp mà còn đòi hỏi phải có những hiểu biết nhất định về chuyên ngành ấy. Do đó, trong quá trình dịch thuật, bên cạnh tìm hiểu thông tin liên quan đến từ, cụm từ và thuật ngữ trong các giáo trình liên quan và trên Internet, chúng tôi cũng nhận được sự hỗ trợ rất lớn từ các chuyên gia về ngôn ngữ và các chuyên gia trong lĩnh vực chuyên ngành. Hiện tại, bộ dữ liệu được xây dựng với hơn 1500 từ cùng với nghĩa của từ và 1500 ngữ cảnh (bằng văn bản) tương ứng từng từ. Mỗi từ sẽ được cung cấp các thông tin cần thiết cho nhu cầu tra từ có ngữ cảnh như từ tiếng Anh, từ tiếng Việt, giải thích nghĩa bằng tiếng Anh, giải thích nghĩa bằng tiếng Việt, từ viết tắt và ngữ cảnh của từng từ. Hệ thống từ được chúng tôi tổ chức và quản lý trên Hệ quản trị CSDL SQL Server 2012.

Trong phần tiếp theo, chúng tôi sẽ trình bày kết quả xây dựng hệ thống thông tin website tìm kiếm từ điển chuyên ngành có ngữ cảnh.

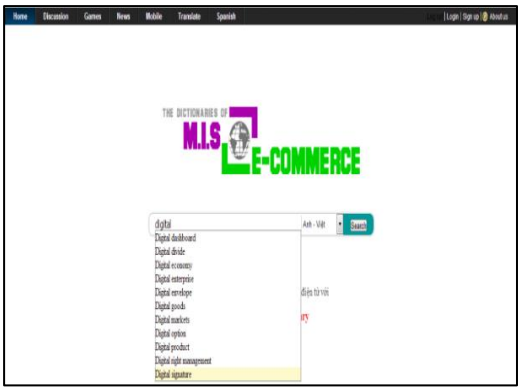
4.2. Hệ thống website từ điển chuyên ngành

Trong phần này, chúng tôi trình bày kết quả xây dựng hệ thống website theo từng bước thực hiện tra từ trên hệ thống.

Bước 1: Truy cập vào hệ thống website[12] để tra từ điển chuyên ngành…

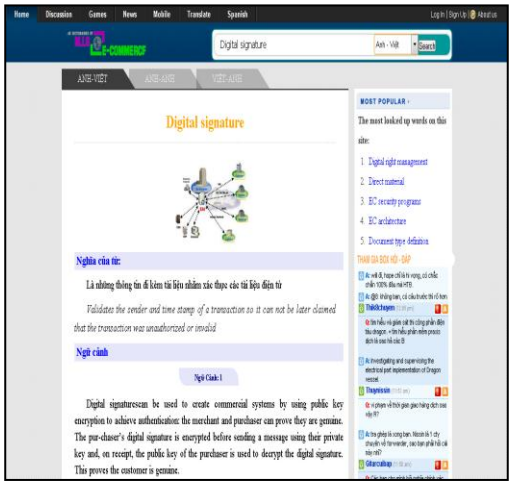
Bước 2: Người dùng nhập từ vào khung tìm kiếm (xem hình 3). Hệ thống xử lý để đưa ra các dự đoán từ khi người dùng chưa nhập hoàn chỉnh một từ nhằm giúp cho người dùng chọn nhanh từ cần tra.

[12] Website này chúng tôi đang trong quá trình hoàn thiện về hạ tầng kỹ thuật bảo mật cũng như kiểm định cơ sở dữ liệu từ điển. Bên cạnh đó, chúng tôi sẽ phát triển tiếp những ứng dụng tiện ích liên quan được tích hợp trên website và tiến hành đưa website công bố lên Internet.



Hình 3. Giao diện hệ thống tra từ điển chuyên ngành

Bước 3: Người dùng nhấn nút "Search" để tra cứu từ (hình 3). Hệ thống xử lý truy vấn từ CSDL và hiển thị kết quả tìm kiếm trên giao diện chi tiết. Trong quá trình xử lý truy vấn kết quả, hệ thống ứng dụng giải thuật tìm kiếm KMP (hình 4).



Hình 4. Giao diện chi tiết trình bày kết quả tra từ điển chuyên ngành

Bước 4: Người dùng xem kết quả trả về trên hình 4. Kết quả tìm kiếm thể hiện trên hình 4 bao gồm 2 nội dung: (1) nghĩa của từ được tra, (2) ngữ cảnh của từ được tra. Ngữ cảnh này được thể hiện ở cả hai ngôn ngữ tiếng Việt và tiếng Anh. Phần ngữ cảnh cũng giúp cho việc phân tích để tìm ra những từ điển bổ sung vào danh sách từ gợp ý có liên quan đến từ được ra.

Ngoài ra phần bên phải của hình 4, thể hiện danh sách các từ thường được người dùng tìm kiếm "Most popular", người dùng có thể chọn và xem chi tiết ngữ cảnh của từ đó trong danh sách đó. Hệ thống xử lý và trả kết quả chi tiết tra cứu liên quan đến từ, đồng thời gợi ý tiếp các từ liên quan được trình bày trong mô hình đề xuất trong phần 3 để thực hiện khai thác ngữ cảnh của từ và đưa ra các từ gợi ý liên quan đến từ cần tìm nhằm giúp người dùng hiểu rõ và rộng hơn ý nghĩa của từ đã tra. Trên hình 4 cũng thể hiện chức năng cho người dùng thảo luận. Mục đích chúng tôi xây dựng chức năng này nhằm giúp người dùng có thể trao đổi về từ điển, về các từ mới hoặc người dùng có thể đóng góp ý kiến cho nội dung liên quan đến từ điển và hệ thống tra từ.

Bên cạnh hỗ trợ tra từ điển chuyên ngành và tìm kiếm từ gợi ý được trình bày trong hình 3 và hình 4. Trên hệ thống website, chúng tôi còn xây dựng thêm các chức năng tiện ích khác nhằm giúp cho người dùng có thể tìm hiểu thêm những vấn đề, thông tin liên quan đến ngành nghề như trang tin tức, thảo luận (hình 5).

Hình 5. Giao diện trang tin tức về chuyên ngành HTTTQL và TMĐT

Ngoài ra, một tính năng quan trọng khác được xây dựng trên hệ thống website là chức năng giúp người dùng tham gia bổ sung mới từ chuyên ngành, chỉnh sửa nội dung liên quan đến từ chuyên ngành.

4.3. Thảo luận kết quả

Ngoài những yếu tố ảnh hưởng đến kết quả tìm kiếm như phần cứng, băng thông, thiết kế,… yêu cầu đặt ra của một công cụ tìm kiếm là tốc độ xử lý và sự chính xác. Vì vậy, việc xử lý bên trong để có thể cho ra kết quả tốt nhất là rất quan trọng. Để xử lý một tập ngữ liệu lớn và tìm sự liên kết giữa các từ dựa trên cơ sở ngữ cảnh, giải thuật được tính đến như một giải pháp đạt hiệu quả cao. Với cơ sở tìm kiếm dựa trên ngữ cảnh, bài toán so sánh mẫu được chọn để giải quyết yêu cầu đặt ra, giải thuật tìm kiếm như KMP được chúng tôi áp dụng vì những ưu điểm và sự linh hoạt trong tìm kiếm mà giải thuật mang lại. Dù chưa phải là lựa chọn tối ưu nhất nhưng việc áp dụng giải thuật KMP đã mang lại hiệu quả nhất định trong nghiên cứu. Bằng việc kết hợp với công cụ tìm kiếm tối ưu được trang bị từ Microsoft SQL Server là SQL Full Text Search hay còn được gọi là FTS [4][5], mỗi kỹ thuật được áp dụng nhằm tùy biến kết quả tìm kiếm, nhưng vẫn đáp ứng được những yêu cầu đặt ra trong tìm kiếm chính xác từ theo ngữ cảnh.

Việc đánh giá tính chính xác kết quả thực hiện của mô hình và hệ thống đề xuất, chúng tôi thực hiện theo phương pháp kiểm tra trực tiếp để đối chiếu từng dữ liệu kết quả với dữ liệu được lưu trữ trong hệ thống từ điển. Bên cạnh đó, chúng tôi đã kiểm tra kết quả bằng cách thực hiện các kỹ thuật truy vấn trực tiếp trên dữ liệu để so sánh với kết quả tìm kiếm trên hệ thống website. Kết quả cho độ chính xác 100% giữa kết quả thực hiện mô hình đề xuất trên hệ thống website so sánh với kiểm tra trực tiếp dữ liệu.

Tóm lại, dựa trên mô hình và phương pháp đề xuất, việc sử dụng SQL Full Text Search [11] để tìm kiếm từ khóa ban đầu, đồng thời

kết hợp giải thuật KMP để đưa ra từ gợi ý dựa trên cơ sở ngữ cảnh là ý tưởng chính trong việc xử lý tìm kiếm từ hoặc cụm từ trong hệ thống dữ liệu từ điển chuyên ngành Hệ thống thông tin quản lý và Thương mại điện tử và đã đạt hiệu quả tốt về tốc độ xử lý và tính chính xác dữ liệu.

Được xây dựng nhằm hỗ trợ tra cứu thuật ngữ thuộc chuyên ngành HTTTQL và TMĐT, hệ thống CSDL dùng cho hệ thống website tra cứu từ điển chuyên ngành đã tương đối đáp ứng được các yêu cầu tra từ theo chuyên ngành HTTTQL và TMĐT, bao gồm tra từ theo các loại từ điển Anh - Việt, Việt - Anh, hoặc Anh - Anh, hoặc tất cả, tra từ viết tắt và hỗ trợ lưu trữ hình ảnh minh họa trực quan cho từ.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi tập trung nghiên cứu và giải quyết bốn vấn đề chính nhằm đạt được mục tiêu đặt ra là xây dựng hệ thống tra từ chuyên ngành có ngữ cảnh trong lĩnh vực Thương mại điện tử và Hệ thống thông tin quản lý, và đây cũng là 4 đóng góp của chúng tôi trong nghiên cứu, bao gồm: (1) xây dựng được trên 1500 từ điển chuyên ngành cùng với hệ thống dữ liệu ngữ cảnh của từng từ, (2) thiết kế index cho các cột thường xuyên được truy vấn, xử lý các câu truy vấn sao cho trả về kết quả mong muốn nhưng khả năng truy xuất dữ liệu tốt hơn đặc biệt là khi dữ liệu đã thật sự lớn, (3) phân tích và đánh giá các giải thuật để lựa chọn giải thuật KMP áp dụng trong phần xử lý trên hệ thống tìm kiếm, (4) xây dựng và triển khai hệ thống website tra từ chuyên ngành có ngữ cảnh thuộc lĩnh vực Hệ thống thông tin quản lý và Thương mại điện tử. Hiện tại, hệ thống website đang trong quá trình hoàn thiện về hạ tầng kỹ thuật bảo mật cũng như kiểm định cơ sở dữ liệu từ điển. Chúng tôi sẽ tiến hành sớm đưa website lên Internet để công bố rộng rãi đến người dùng.

Trong nghiên cứu tiếp theo, chúng tôi sẽ tiếp tục cải thiện hệ thống tra từ điển chuyên ngành để mở rộng lĩnh vực tra cứu và tốc độ xử lý bằng những giải thuật cải tiến có kết quả tìm kiếm nhanh hơn để hướng đến mở rộng cơ sở dữ liệu từ điển và ngữ cảnh của từ. Chúng tôi cũng sẽ bổ sung từ và ngữ cảnh của từ, đồng thời phát triển việc khai thác ngữ cảnh bằng cách thu thập dữ liệu từ Internet để có kết quả đa dạng hơn trong nhiều lĩnh vực ứng với từ chuyên ngành cụ thể. Bên cạnh đó, việc phát triển hệ thống website giúp truy cập ứng dụng trên điện thoại thông minh nhằm tạo điều kiện dễ dàng nhất cho người dùng khi cần tra cứu cũng sẽ được chúng tôi quan tâm.

TÀI LIỆU THAM KHẢO

[1]. Akinul Islam Jony, Analysis of Multiple String Pattern Matching Algorithms, *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, Vol. 3, No. 4, 2014, pp. 344-353 (2014).

[2]. Akhtar Rasool Amrita Tiwari, et al, String Matching Methodologies:A Comparative Analysis, *International Journal of Computer Science and Information Technologies*, Vol. 3 (2) ,3394-3397 (2012).

[3]. Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford [1990]. The Rabin-Karp algorithm. *Introduction to Algorithms (2nd ed.)*. Cambridge, Massachusetts: MIT Press. pp. 911-916. ISBN 978-0-262-03293-3 (2001).

[4]. D.E. KNUTH, J.H. MORRIS, V.R. PRATT, Fast Pattern Matching in Strings. *SIAM Journal of Computing 6*, 2, 323-350 (1977).

[5]. Dana Shapira, et al, Adapting the Knuth–Morris–Pratt algorithm for pattern matching in Huffman encoded texts, *Information Processing and Management* 42, 429-439 (2006).

[6]. Dave Chaffey, E-book: *E–Business and E–Commerce Management*, Prentice Hall, ISBN: 978-0273752011 (2010).

[7]. Domingos, Pedro; Pazzani, Michael, *On the optimality of the simple Bayesian classifier under zero-one loss*. Machine Learning 29: 103-137. 7 (1997).

[8]. D. Sunday, *Very Fast Substring Search Algorithm*, Comm. ACM, vol 33, issue 8, pp. 132-142 (1990).

[9]. Ellard, Daniel J. String Searching. *S-Q Course Book.* [Online] [Cited: 06 10, 2011.] *http://www.eecs.harvard.edu/~ellard/Q-97/HTML/root/root.html* (2011).

[10]. Jane P.Laudon & Kenneth C.Laudon, E-book: Essentials of Management Information Systems, PEARSON, ISBN: 978-0136025818 (2010).

[11]. http://msdn.microsoft.com/en-us/library/ms142571.aspx, Full-Text Search (SQL Server)

[12]. Karp, Richard M.; Rabin, Michael O. *Efficient randomized pattern-matching algorithms.* IBM Journal of Research and Development 31 (2), pp. 249-260 (March 1987).

[13]. Laudon, E-book: *E–Commerce*, Pearson Education, ISBN-13: 978-0135090787 (2010).

[14]. Michael Coles, Hilary Cotter, *Pro Full-Text Search in SQL Server 2008, Spinger-Verlag New York*, Inc (2009).

[15]. Muhammad, Rashid Bin. String Matching Agorithm. *Design and Analysis of Computer Algorithms.* [Online] Kent State University. [Cited: 06 20, 2011.] *http://www.personal.kent.edu/~rmuhamma/Algorithms/algorithm.html*.

[16]. R. Boyer, J. Moore, *A fast string searching algorithm*, Comm. ACM vol 20, pp. 762-772 (1977).

[17]. Siam J. Comput, et al, *Fast pattern matching in strings*, donald e. Knuth, Vol. 6, No. 2, (June 1977).

[18]. Ralph Stair & George Reynolds, E-book: *Principle of Information Systems*, Cengage Learning, ISBN: 0324665288 (2009).

[19]. Vidya SaiKrishna, et al, String Matching and its Applications in Diversified Fields, IJCSI *International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 1, January 2012 (2012).

[20]. Yanbing Liu et al, *A factor-searching-based multiple string matching algorithm for intrusion detection*, Communications (ICC), 2014 IEEE International Conference, pp. 653-658 (2014).