

ĐÁNH GIÁ TẦN SỐ SỬ DỤNG CHỮ VIẾT TẮT TIẾNG VIỆT TRÊN INTERNET

ASSESSMENT OF USE FREQUENCY OF VIETNAMESE ABBREVIATIONS ON THE INTERNET

Phan Huy Khánh², Nguyễn Nho Túy¹, Đặng Huy Hòa²

¹VNPT Đà Nẵng, tuynn.dng@vnpt.vn; hoadh89@gmail.com

²Trường Đại học Bách khoa, Đại học Đà Nẵng; khanhph29@gmail.com

Tóm tắt - Thời đại bùng nổ thông tin trên Internet, chữ viết tắt (CVT) tiếng Việt được sử dụng rộng rãi trong nhiều lĩnh vực. Tuy nhiên, việc đánh giá chỉ số là tần số sử dụng, tần suất xuất hiện CVT gặp nhiều khó khăn do tính đa dạng sử dụng, tính nhập nhằng, đa nghĩa của CVT; thiếu phương tiện thống kê, đánh giá và hầu như chưa có nghiên cứu chuyên sâu vấn đề này. Trên cơ sở kho dữ liệu CVT chúng tôi đã xây dựng được, trong bài báo này, chúng tôi đề xuất một giải pháp mới đánh giá tần số sử dụng CVT trên Internet; qua đó lượng hóa chỉ số tần suất xuất hiện trong những cụm CVT giống nhau, nhằm cung cấp một góc nhìn, một cách đánh giá mức độ sử dụng CVT tiếng Việt, phục vụ nhiều mục đích sử dụng trong xử lý ngôn ngữ, trong khai thác dữ liệu CVT, tư vấn doanh nghiệp...

Từ khóa - chữ viết tắt; từ viết tắt; tần số chữ viết tắt; tần suất chữ viết tắt; chỉ số đánh giá xuất hiện chữ viết tắt.

1. Đặt vấn đề

Hiện nay, các ứng dụng khai thác dịch vụ trên Internet ngày càng phong phú, đa tạp, biến đổi nhanh chóng, làm thay đổi chiều sâu cuộc sống. Điện thoại, Email, nhắn tin, lướt web... gần như là hoạt động không thể thiếu của mỗi người. Việc sử dụng bàn phím, gõ văn bản tiếng Việt có dấu nhiều khi gây trở ngại cho người sử dụng (NSD). Chính sự bất cập này dẫn đến hiện tượng viết tắt, nói tắt. Theo đà phát triển của công nghệ, viết tắt, nói tắt trở thành trào lưu phổ biến trên Internet trong mọi lĩnh vực, chủ đề, lớp NSD, nhất là giới trẻ.

Tuy nhiên, tính đa tạp, không nhất quán, lẫn lộn tiếng Anh, tiếng Việt, thiếu quy chuẩn, không có định hướng... dẫn đến chữ viết tắt sử dụng rất tùy tiện, nhập nhằng. Nhu cầu chuẩn hóa, thống nhất hóa trong tiến trình phát triển ngôn ngữ nói chung, và tiếng Việt nói riêng, trong đó có CVT, là nhu cầu thực tế, là nhiệm vụ đặt ra của chuyên ngành xử lý ngôn ngữ tự nhiên, xử lý tiếng Việt. Việc triển khai xây dựng các ứng dụng khai thác thông tin trên Internet trong nhiều lĩnh vực luôn có nhu cầu đánh giá các tiêu chí ứng dụng xử lý thông tin khác nhau. Chẳng hạn, trong lĩnh vực thương mại: CVT ngày càng sử dụng rộng rãi, nhất là thương hiệu, nhãn hiệu bởi nó dễ nhớ, dễ đọc, tạo ra sự khác biệt của sản phẩm, giúp khách hàng nhận biết và lựa chọn, từ đó giúp phát triển, quảng bá sản phẩm của doanh nghiệp. Do đó, nhu cầu thống kê (đếm) tần số, tần suất xuất hiện, sử dụng các CVT là rất cần thiết và có ý nghĩa ứng dụng trong thực tiễn.

Chúng ta nhận thấy CVT được sử dụng trong các kho ngữ liệu (Corpus) hay dữ liệu lớn khổng lồ (Big Data): Internet, sách báo, phim ảnh, giao tiếp... Rõ ràng, không thể có giải pháp hoàn hảo cho việc thống kê tần số sử dụng CVT trong nguồn dữ liệu lớn đó. Tuy nhiên, việc xác định một phạm vi giới hạn nào đó để đưa ra một phương pháp thống kê, đánh giá là hoàn toàn khả thi.

Abstract - In this era of information explosion on the Internet, Vietnamese abbreviations are widely used in many fields. However, the evaluation of index which is the use frequency, appearance frequency of abbreviations is difficult due to the variety, the ambiguity of abbreviations; lack of means of statistical evaluation and because there has been almost no in-depth study of this issue. On the basis of our data warehouse of abbreviations we have built, in this paper, we propose a new evaluation of the use frequency of abbreviations on the Internet and thereby quantify the frequency indicators that appear in the same abbreviation clusters to provide a perspective, a way of assessing the use of Vietnamese abbreviations to serve many purposes in language processing, in abbreviation data mining, business counseling

Key words - abbreviation; Acronyms; frequency of abbreviations; indicators to assess abbreviations.

Trên cơ sở lý thuyết xác suất, thống kê [8], chúng tôi vận dụng và đưa ra thống kê tần số sử dụng CVT bằng cách kế thừa nguồn dữ liệu đã có, chẳng hạn như kế thừa dữ liệu CVT trên hệ thống tìm kiếm chuyên nghiệp như Google, Yahoo... Từ đó, tìm cách tính toán, đánh giá tần số sử dụng, tần suất xuất hiện của CVT trên Internet, đưa ra một chỉ số tham khảo mức độ sử dụng CVT, làm cơ sở tư vấn, đánh giá riêng cho một số chuyên ngành.

Nội dung bài báo như sau: trước tiên, trình bày những khái niệm và thuật ngữ về tần số, tần suất; đồng thời điểm qua vài nét về CSDL CVT chúng tôi xây dựng được. Phần tiếp theo trình bày một giải pháp mới đánh giá tần số sử dụng CVT trong phạm vi thông tin CVT xuất hiện trên Internet; qua đó, đưa ra chỉ số tần suất xuất hiện CVT trong những cụm CVT giống nhau, lượng hóa chỉ số đánh giá mức độ sử dụng CVT tiếng Việt tại thời điểm thu thập dữ liệu; đồng thời có những khuyến nghị ứng dụng chỉ số này trong xử lý nhập nhằng, khai thác dữ liệu CVT, trong sử dụng CVT cho nhãn hiệu, thương hiệu. Phần cuối cùng là kết luận.

2. Vấn đề tần số sử dụng và tần suất xuất hiện chữ viết tắt

2.1. Một vài nghiên cứu về tần số CVT

Hiện nay, chưa có nhiều công trình, bài báo liên quan đến vấn đề này, đặc biệt là về tiếng Việt (Kinh). Một số công trình nghiên cứu thu thập và xây dựng cơ sở dữ liệu CVT tiếng Việt [2], [6] đã thống kê được các CVT thông dụng, nhưng chưa thống kê được tần số xuất hiện. Hội Ngôn ngữ học Việt Nam đưa ra bản “Danh sách chữ viết tắt xếp theo tần số” [5] (năm 2002), ghi nhận được 1.151 đơn vị CVT, có thống kê tần số xuất hiện, nhưng hầu như không đưa ra các CVT trùng nhau và tần suất xuất hiện của chúng. Ở đây, không nói rõ căn cứ, tiêu chí thống kê nào để đưa ra tần số sử dụng CVT. Với cách làm thủ công, thiếu công cụ quản lý, cập nhật thường xuyên thì cách đánh giá

tần số sử dụng là gặp nhiều khó khăn, bất cập; không theo kịp sự phát triển CVT trong thực tế và xu hướng sử dụng. Sau hơn 10 năm, CVT có nhiều biến đổi, chắc chắn dữ liệu thống kê tần số này đã lỗi thời.

Một nghiên cứu về thực trạng của hiện tượng viết tắt từ ngữ [1], bằng cách khảo sát từ liệu thực tế trên 10 tờ báo với phương pháp thống kê thủ công đã đưa ra một số kết luận: Từ liệu thống kê cho biết tần số sử dụng của từng dạng viết tắt, từ đó cho thấy xu hướng (theo tần số) cố định hóa các dạng tắt, gợi ý cho việc đưa ra các giải pháp chuẩn hóa chữ viết tắt trên văn bản.

Với mục đích mô tả, phân tích đối chiếu viết tắt trong tiếng Anh và tiếng Việt đối với phương tiện giao tiếp - thư điện tử [11], một nghiên cứu đã tìm ra những nét tương đồng và khác biệt của những CVT trong tiếng Anh đối chiếu với tiếng Việt, từ đó có thể dự đoán những khó khăn của người đọc cũng như để giúp cho người sử dụng ngôn ngữ. Nghiên cứu này có chỉ ra Bảng thống kê tần suất xuất hiện các đặc điểm ngữ pháp của cụm từ và câu viết tắt, nhưng chỉ giới hạn trong một lĩnh vực hẹp là thư điện tử trong tiếng Anh và tiếng Việt; không có sự tính toán hay công bố tần số sử dụng, tần suất xuất hiện các CVT nói chung.

Các tập Corpus cũng có thể cho phép đánh giá tần số sử dụng CVT tương đối tập trung. Chẳng hạn, sử dụng tập Corpus *VNTQcorpus(big).txt* - kế thừa kết quả nghiên cứu [10], tập này có kích thước ~240 Mb, số lượng câu khoảng 1.750.000 câu, trích xuất từ khoảng 13.000 bài báo trên Internet. Bảng kỹ thuật tìm kiếm thông thường trên tệp văn bản (TXT), chúng tôi chỉ tìm thấy 2 chữ viết tắt UBND (Ủy ban nhân dân), không tìm thấy chữ viết tắt KCS (Kiểm tra chất lượng sản phẩm), hay HTX (Hợp tác xã). Điều này chứng tỏ nếu dựa trên các tập Corpus, rất khó có thể đưa ra tần số sử dụng CVT cũng như tần suất xuất hiện nó một cách tiệm cận với thực tế vốn có.

2.2. Đề xuất giải pháp mới đánh giá tần số CVT, tần suất CVT tiếng Việt

Chúng tôi đưa ra khái niệm mới về tần số sử dụng, tần suất xuất hiện chữ viết tắt tiếng Việt trên Internet.

2.2.1. Khái niệm tần số sử dụng, tần suất chữ viết tắt tiếng Việt trên Internet

Ta gọi một bảng số liệu các CVT là bảng có N chữ viết tắt, thực chất là bảng CSDL chữ viết tắt, nhưng chỉ có k CVT khác nhau là $x_1, x_2, \dots, x_k$.

Tần số sử dụng CVT trên Internet là giá trị số biểu thị kết quả trả về khi thực hiện tìm kiếm trên hệ thống tìm kiếm chuyên nghiệp trên Internet trong phạm vi nào đó tại một thời điểm nhất định. Chữ viết tắt x_i xuất hiện n_i lần khi tìm thấy trên một chương trình tìm kiếm chuyên nghiệp, với $1 \leq i \leq k$, ta nói n_i là *tần số sử dụng* của chữ viết tắt x_i .

Phạm vi, tiêu chí tìm kiếm tần số sử dụng: Công cụ tìm kiếm đều có chức năng tìm kiếm mở rộng, chuyên sâu, theo tiêu chí nào đó (tên miền, gần đúng, chính xác...). Dựa vào đặc điểm này, có thể đánh giá *tần số sử dụng CVT* trong một phạm vi nhất định. Ví dụ, công cụ tìm kiếm cho phép tìm CVT trong phạm vi tên miền.vn

Phạm vi không giới hạn trên Internet: Phạm vi CVT sử dụng trên hệ thống tìm kiếm chuyên nghiệp với bất kỳ tên miền

nào, ở mọi quốc gia, với mọi ngôn ngữ, ký hiệu là N_ALL .

Phạm vi giới hạn: Đây là tiêu chí tìm kiếm CVT có giới hạn quốc gia, tên miền trên Internet. Có 4 phạm vi giới hạn được định nghĩa:

N_VN1 : Phạm vi CVT sử dụng trên hệ thống tìm kiếm chuyên nghiệp với tiêu chí tìm kiếm bất kỳ tên miền nào nằm quốc gia Việt Nam, với mọi ngôn ngữ, ký hiệu là N_VN1 .

N_VN2 : Phạm vi CVT sử dụng trên hệ thống tìm kiếm chuyên nghiệp với tiêu chí tìm kiếm trong tên miền.vn, trong quốc gia Việt Nam, và với ngôn ngữ tiếng Việt, ký hiệu là N_VN2 .

N_VN3 : Phạm vi CVT sử dụng trên hệ thống tìm kiếm chuyên nghiệp với tiêu chí tìm kiếm trong tên miền.com, trong quốc gia Việt Nam, và với ngôn ngữ tiếng Việt, ký hiệu là N_VN3 .

N_VN4 : Phạm vi CVT sử dụng trên hệ thống tìm kiếm chuyên nghiệp với tiêu chí tìm kiếm trong tên miền.org, trong quốc gia Việt Nam, và với ngôn ngữ tiếng Việt, ký hiệu là N_VN4 .

Tiêu chí tìm kiếm tần số sử dụng trên Internet: Để có kết quả tần số sử dụng có giá trị chính xác; chúng tôi định nghĩa tiêu chí *tìm kiếm chính xác*, tức là là bọc chuỗi từ khóa cần tìm trong dấu ngoặc kép ("*Từ khóa*").

Nhận xét: Giá trị tần số sử dụng này không phải là duy nhất, luôn biến động do sự cập nhật thông tin liên tục từ Internet. Số lần sử dụng là trùng lặp, mỗi sự xuất hiện hay trích dẫn trên Internet đều có thể được xem như một lần sử dụng và giá trị tần số sử dụng được tăng lên 1 đơn vị.

Tần suất xuất hiện chữ viết tắt tiếng Việt trên Internet: Tần suất xuất hiện CVT là tỷ số giữa tần số sử dụng CVT trên môi trường Internet và tổng các tần số sử dụng của các CVT trong bảng thống kê tần số sử dụng CVT đang xem xét trong một đơn vị thời gian, một phạm vi thống kê nào đó.

Hay nói cách khác, với một bảng số liệu thống kê các chữ viết tắt có N giá trị nhưng chỉ có k giá trị khác nhau $x_1, x_2, \dots, x_k$.

Giá trị chữ viết tắt x_i xuất hiện n_i lần ($1 \leq i \leq k$), n_i là tần số của chữ viết tắt x_i

Tỉ số $f_i = \frac{n_i}{N}$ được gọi là tần suất xuất hiện của chữ viết tắt x_i . với:

$$n_1 + n_2 + \dots + n_k = N; f_1 + f_2 + \dots + f_k = 1$$

Hoặc: Tỉ số $f_i = \frac{n_i}{N} * 100\%$ được gọi là tần suất xuất hiện của chữ viết tắt x_i . với:

$$n_1 + n_2 + \dots + n_k = N; f_1 + f_2 + \dots + f_k = 100$$

Ví dụ: Với cách định nghĩa trên, trong phạm vi chỉ định tại một thời điểm, *tần số sử dụng CVT* và *tần suất xuất hiện CVT* được mô tả trong Bảng 1 dưới đây:

Bảng 1: Tính toán tần suất xuất hiện CVT trong một phạm vi chỉ định

STT	Chữ viết tắt	Tần số (trong phạm vi)	Tần suất (%)
1	Phát triển (PT)	n_1	f_1

2	Phạm trù (PT)	n_2	f_2
3	Phương thức (PT)	n_3	f_3
4	Phát thanh (PT)	n_4	f_4
5	Phương trình (PT)	n_5	f_5
6	Phương Thanh (PT)	n_6	f_6
	Tổng cộng	N	100%
7	PT	n_7	

Với bảng có tổng N giá trị sử dụng của k (k=6) CVT khác nhau, tần số sử dụng CVT thứ i là các giá trị n_i tương ứng. Tần suất xuất hiện CVT “PT” trong phạm vi CVT sử dụng đã chỉ định trên Internet theo nghĩa “Phát triển” là f_1 , theo nghĩa “Phương trình” là f_5 , theo nghĩa “Phương Thanh” là f_6 , ...

2.2.2. Giải pháp đánh giá tần số sử dụng trên Internet

Hiện nay, chúng tôi đã xây dựng CSDL CVT [4], đây được xem như bảng dữ liệu đầu vào cần đánh giá tần suất sử dụng CVT. Các bước đề xuất tìm kiếm cập nhật dữ liệu tính toán tần số sử dụng CVT trên Internet:

Bước 1: Chuẩn bị bảng dữ liệu CVT, thiết kế các trường lưu giữ kết quả tần số sử dụng.

Bước 2: Thiết lập phạm vi, giá trị cần tìm kiếm CVT trên hệ thống website tìm kiếm chuyên nghiệp.

Bước 3: Xây dựng chương trình tìm kiếm giả lập như thao tác tìm kiếm từ người sử dụng.

Bước 4: Thực hiện chạy chương trình tìm kiếm, ghi nhận thời gian, giá trị tìm kiếm CVT vào CSDL.

Bước 5: Phân tích dữ liệu, đánh giá, sắp xếp và lưu trữ, công bố.

Bước 6: Sau một chu kỳ (3 tháng hoặc 6), lặp lại các bước 4, 5.

Bước 7: So sánh dữ liệu, đánh giá sự biến động giá trị tần số sử dụng sau một chu kỳ thời gian.

2.2.3. Giải pháp đánh giá tần suất xuất hiện CVT tiếng Việt trên Internet

Sau khi có kết quả cập nhật dữ liệu tính toán tần số sử dụng CVT trên môi trường Internet, chúng tôi đưa ra cách đánh giá tần suất xuất hiện CVT tiếng Việt như sau:

Bước 1: Sắp xếp dữ liệu tần số sử dụng CVT.

Bước 2: Chọn lọc các CVT trùng lặp, đa nghĩa.

Bước 3: Lựa chọn phạm vi tần số xuất hiện để thực hiện tính toán tần suất xuất hiện CVT theo công thức:

$$f_i = \frac{n_i}{N} * 100\% \text{ Trong đó: } N: \text{ Tổng số các CVT giống}$$

nhau nhưng có ngữ nghĩa khác nhau, n_i là tần số sử dụng của chữ viết tắt x_i

Bước 4: Sắp xếp CVT theo giá trị f_i giảm dần, lưu trữ, nhận xét và công bố.

Bước 5: Sau một chu kỳ (3 hoặc 6 tháng), lặp lại các bước 2, 3, 4.

Bước 6: So sánh dữ liệu, đánh giá sự biến động giá trị tần suất xuất hiện sau một chu kỳ thời gian.

2.3. Đề xuất những ứng dụng về chỉ số tần số sử dụng CVT, tần suất xuất hiện CVT tiếng Việt

Với giải pháp đề xuất, chúng tôi đưa ra những ứng dụng

sau đây:

- Chỉ số tần suất xuất hiện: Tiếp cận xử lý nhập nhằng ngữ nghĩa CVT dựa trên chỉ số tần suất CVT. Các CVT có chỉ số tần suất f_i cao được ưu tiên gán ngữ nghĩa tương ứng, tính phổ dụng CVT càng cao, truyền thông quan tâm, có sức sống trong thực tiễn. CVT có tần suất xuất hiện ít và thấp thì ít phổ biến, cũng có thể là rất cá biệt.

- Sự gia tăng tần suất (sau khoảng thời gian đánh giá 3 hoặc 6 tháng) sẽ cho thấy bước phát triển và sử dụng CVT, đặc biệt đối với thương hiệu, nhãn hiệu; chứng tỏ sự lặp lại, sự quan tâm của truyền thông (qua Internet); lưu trữ, quan sát chỉ số này giúp cho việc tư vấn phát triển, nhận diện thương hiệu cho doanh nghiệp.

- Sự giảm tần suất CVT (sau khoảng thời gian đánh giá 3 hoặc 6 tháng) cho thấy việc phát triển thương hiệu có chiều hướng giảm, truyền thông ít quan tâm; có thể đưa ra khuyến nghị doanh nghiệp cần phải xem lại chiến lược truyền thông. Việc này có ý nghĩa ứng dụng trong tư vấn doanh nghiệp phát triển nhận dạng thương hiệu.

- Với mỗi phạm vi sử dụng, tần số xuất hiện CVT giúp tư vấn chọn lựa phương thức truyền thông để đạt hiệu quả mong đợi. Chẳng hạn CVT là thương hiệu xuất hiện nhiều nhất trên tên miền.com thì quảng bá trên web site có tên miền này sẽ hiệu quả hơn các tên miền khác.

- Bảng phép so sánh, đối chiếu, rút ra sự tương quan tần suất CVT với những chỉ số thị phần, xếp hạng thương hiệu. Sự tương quan giữa chúng cho phép dựa vào tần suất xuất hiện CVT đưa ra chỉ số lượng hóa trong phát triển thương hiệu, nhãn hiệu, thị phần của doanh nghiệp.

3. Triển khai thử nghiệm và đánh giá kết quả

Chúng tôi triển khai áp dụng giải pháp theo các bước đã nêu.

3.1. Xây dựng chương trình tìm kiếm tự động

Chúng tôi chọn hệ thống tìm kiếm chuyên nghiệp để khai thác dữ liệu chữ viết tắt là Yahoo.com. Sử dụng ngôn ngữ lập trình C# để triển khai xây dựng chương trình.

Thuật toán: Ghi nhận tần số sử dụng chữ viết tắt từ môi trường Internet

Input: File dữ liệu CVT, phạm vi tìm kiếm dữ liệu trên yahoo.com

Output: File dữ liệu tần số sử dụng CVT theo các tiêu chí.

Open CSDL làm việc trung gian
Thiết lập các tiêu chí, phạm vi tìm kiếm
Lựa chọn và Khởi động trình duyệt.

Repeat

Read CVT của file đầu vào

Truyền tham số trình duyệt tìm kiếm
giá trị CVT trong phạm vi, tiêu chí

Open tệp HTML sau khi có kết quả
tìm kiếm trả về

Read file HTML

Dùng Biểu thức chính quy tìm giá trị số

"<span.*>(.*) kết quả "

So khớp và Lưu giữ giá trị <Kết quả>

vào bảng theo phạm vi tương ứng

Lưu giữ ngày tìm kiếm

Until Không còn CVT trong file đầu vào.

Hiện thị Bảng kết quả

Lưu kết quả vào File

Kết thúc.

Chương trình nhận danh sách CVT trong CSDL đầu vào [2], khởi động trình duyệt, truyền tham số tìm kiếm gồm CVT, phạm vi, tiêu chí tìm kiếm. Tự động hóa tìm kiếm. Sau khi tìm kiếm xong, lấy kết quả tệp mã nguồn (source code dạng html) từ trang web, đọc kết quả tìm được và lưu vào CSDL. Chương trình gồm các lớp (class) sau:

a) **Class Utils**: chứa các hàm cần thiết để chạy chương trình. Có những hàm quan trọng sau:

Hàm **getSourceCode_FromSelenium**: Lấy source code dạng html từ trang web. Tham số đầu vào là đường link dẫn đến trang web đó. Hàm trả về kiểu string, là source code dạng html. Nếu lỗi trả về chuỗi "ERROR".

Hàm **getMatchedResult**: Tìm chuỗi chứa số kết quả CVT tìm thấy được nằm trong source code dạng file html. Sử dụng Regex (Regular Expression - Biểu thức chính quy) để tìm một chuỗi con trong một chuỗi. Pattern chuỗi con mà được sử dụng để tìm theo source mà yahoo.com trả về là:

"<span.*>(.*) kết quả "

Ngoài ra có những hàm khác liên quan đến những chức năng bổ sung trong chương trình như: đọc/ghi file CSV, đọc file xml, lấy ngày tháng,...

b) **Class PromptForBrowserSelection**: Dùng để hiển thị form, chọn trình duyệt khi bấm vào nút **Run** thực hiện chương trình, xuất ra tệp kết quả lưu trữ.

c) **Class Tieuchi**: chứa các thông số cho một trường hợp (tiêu chí) cụ thể để tạo thành một đường link tìm kiếm trên yahoo.com. Các tiêu chí này được mô tả và lưu trong một tệp xml (Data/tieuchi.xml). Thông số bao gồm:

1: Chế độ/tiêu chí: tìm chính xác cả cụm hay tìm theo từng từ khóa (tìm chính xác là bọc chuỗi từ khóa trong dấu ngoặc kép "...")

2: Tên miền: nếu tìm trong tất cả tên miền thì để trống

3: Quốc gia: nếu tìm trong tất cả quốc gia thì để trống

4: Ngôn ngữ: Nếu tìm với tất cả ngôn ngữ thì để trống

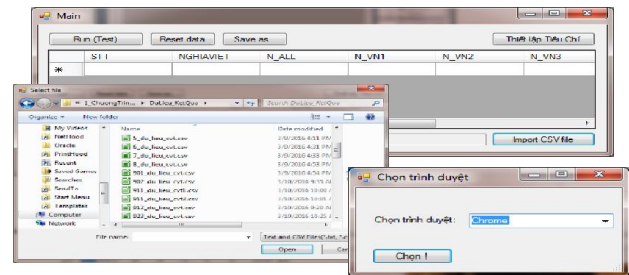
5: V_NAME: tên trường sẽ lưu số kết quả tìm lên lưới giao diện.

d) **Class TieuchiStore**: Chứa danh sách các trường hợp (tiêu chí) phạm vi tìm kiếm. Chương trình đọc tiêu chí từ file **Data/tieuchi.xml** (tệp này được tạo ra dựa theo những định nghĩa phạm vi tần số sử dụng, tiêu chí nêu trên: N_ALL, N_VN1, N_VN2, N_VN3, N_VN4).

3.2. Thực hiện chương trình và cập nhật dữ liệu tần số sử dụng CVT

- Thực hiện chạy chương trình. Tiêu chí tìm kiếm chính xác cả cụm từ (bọc chuỗi từ khóa trong dấu ngoặc kép "..."), trong các phạm vi tìm kiếm đã định nghĩa.

- Chương trình chính chạy trong khung cửa sổ Main (Hình 1), chọn nút lệnh **Import CSV file**, chọn file đầu vào; chọn lựa trình duyệt, sau đó kích chọn nút **Run** để thực hiện thu thập dữ liệu tần suất sử dụng CVT.



Hình 1. Thực hiện chương trình cập nhật tần số sử dụng CVT

- Giao diện chương trình khi thực hiện xong đưa ra kết quả các giá trị tìm kiếm được là các tần số CVT sử dụng theo từng phạm vi đã định nghĩa (các cột). Thực hiện **Save as** để lưu kết quả ra **File CSV** phục vụ cho lưu trữ và xử lý tính toán tần suất xuất hiện CVT (Hình 2):

STT	NGH/VN/ET	N_ALL	N_VN1	N_VN2	N_VN3	N_VN4	NGH/V_N
1	Phát triển (PT)	200	0	0	0	0	0
2	Phong phú (PF)	1	0	0	0	0	0
3	Phong phú (PF)	1	0	0	0	0	0
4	Phong phú (PF)	1	0	0	0	0	0
5	Phong phú (PF)	457	345	0	0	0	0
6	Phong phú (PF)	24	0	0	0	0	0
7	PT	307.000.000	345.000	0	0	0	0

Hình 2. Kết quả thực hiện chương trình cập nhật tần số CVT

- Kết quả hình 2 cho thấy, trong phạm vi N_VN1, chữ viết tắt PT với nghĩa "Phát triển" có tần số sử dụng định nghĩa 184 lần (thời điểm 24/03/2016), với nghĩa "Phong phú" có tần số sử dụng định nghĩa 184 lần..., chữ viết tắt PT có tần số sử dụng lên đến 345.000 lần.

- **Chú ý:** Do đó, các tệp đầu vào cần phải tách dữ liệu mức độ vừa phải (nên dưới 50 dòng) để Yahoo không từ chối phục vụ. Trong thực nghiệm giải pháp, để chạy chương trình số lượng lớn dữ liệu, cần phải sử dụng nhiều máy tính, trong nhiều khoảng thời gian để có kết quả như mong đợi.

- Tính chính xác của chương trình: So sánh với cách tìm kiếm thủ công của Yahoo.com (gõ chuỗi, tìm kiếm nâng cao tương ứng), kết quả tìm được tại cùng thời điểm là hoàn toàn khớp trùng nhau, vì thực chất thuật toán là sự tự động hóa cách tìm thủ công trên yahoo.com.

3.3. Kết quả thực hiện và đánh giá

Sử dụng nhiều máy tính chạy chương trình, trong nhiều khoảng thời gian. Dữ liệu đầu vào có 4.542 mẫu tin (với hơn 2.300 mẫu CVT thông dụng chọn lựa trong CSDL đã có [1]); tách thành gần 100 File dữ liệu đầu vào. Dữ liệu đầu ra với gần 100 file tương ứng, sau đó được ghép lại để có bảng dữ liệu tổng thể tần số sử dụng CVT.

- Sau khi có bảng dữ liệu tần số sử dụng CVT, chúng tôi thực hiện tính toán tần suất sử dụng theo các bước đã đề xuất trong giải pháp mục 2.2 trên đây. Trước mắt, kết quả đưa ra được hơn 30 bảng tần suất xuất hiện các CVT với những cụm chữ viết tắt giống nhau.

Ví dụ: Kết quả tính toán tần suất chữ viết tắt PT trong phạm vi N_VN1 tại 2 thời điểm khác nhau là ngày 15/03/2016 và 25/03/2016:

Bảng 2. Kết quả tính toán tần số, tần suất xuất hiện CVT

STT	Chữ viết tắt	Ngày 15/3/2016		Ngày 25/03/2016	
		N_VN1	Tần suất %	N_VN1	Tần suất %
1	Phương trình (PT)	382	63	385	63,1
2	Phát triển (PT)	182	30	184	30,2
3	Phương Thanh (PT)	16	2,9	17	2,8
4	Phát thanh (PT)	14	2,5	15	2,5
5	Phương thức (PT)	9	1,5	9	1,5
6	Phạm trù (PT)	0	0,0	0	0,0
	Tổng cộng	603	100	610	100
	PT	333.000		345.000	

Trong bảng trên, với 6 giá trị ngữ nghĩa khác nhau chữ viết tắt PT, ta có nhận xét:

- Tần số sử dụng gia tăng sau 10 ngày, chữ viết tắt PT tăng từ 333.000 lên 345.000, tức là tăng số lần sử dụng lên 12.000 lần; tần số chữ viết tắt gia tăng chứng tỏ sự sử dụng CVT ngày càng nhiều trên mạng Internet.

- Tần suất sử dụng của Phương trình (PT) tại thời điểm 15/03/2016 là 63,04%; Phát triển (PT) là 30,03%, Phương Thanh (PT) là 2,97%... Đến ngày 25/3/2016, tần suất biến đổi nhưng không đáng kể. Chữ viết tắt PT có tần suất sử dụng là 345.000 lần, tuy nhiên chữ viết tắt PT theo ngữ nghĩa “Phương trình” chiếm tần suất sử dụng cao nhất 63,11% với tần số sử dụng định nghĩa là 385 lần. Giá trị này cho thấy ngữ nghĩa “Phương trình” của chữ viết tắt PT là phổ biến hơn cả. Chỉ số cao nhất này chỉ dấu với một chữ viết tắt PT, ngữ nghĩa “Phương trình” là có xác suất xuất hiện cao nhất. Sau ngữ nghĩa “Phương trình”, theo thứ tự là “Phát triển”, “Phương Thanh”, “Phát Thanh”, “Phương thức”, và PT viết tắt cho “Phạm trù” hầu như không sử dụng. Nhờ chỉ số tần suất này, vấn đề xử lý nhập nhằng ngữ nghĩa CVT có một hướng tiếp cận giải quyết, tức là chọn ngữ nghĩa CVT theo tần suất sử dụng cao nhất.

- So sánh với phương pháp đánh giá tần số CVT theo Hội Ngôn ngữ học Việt Nam [5], [7] và tần số CVT xuất hiện trong Corpus, phương pháp này của chúng tôi có những ưu điểm nhất định. Dựa trên nguồn dữ liệu sẵn có của các hệ thống web site tìm kiếm chuyên nghiệp, chúng ta có thể trích lọc các thông tin hữu ích về tần số sử dụng và tần suất xuất hiện CVT, được cập nhật theo kịp sự phát triển CVT trong thực tiễn.

- Chúng tôi đã công bố trên www.chuviattat.com [12] về kết quả tần số, tần suất xuất hiện các CVT, dự kiến sẽ công bố giá trị tăng giảm tần suất sau một chu kỳ thời gian (3 - 6 tháng); sắp xếp giảm dần theo giá trị tần suất. Đưa ra cảnh báo các giá trị giảm (CVT là nhãn hiệu, thương hiệu). NSD có công cụ để tự đánh giá là các CVT mình đang sở hữu; chức năng tra cứu CVT sẽ bổ sung thêm giá trị tần suất xuất hiện và thời gian đánh giá. Sau một chu kỳ thời gian sẽ có bổ sung giá trị tăng/giảm tần suất xuất hiện.

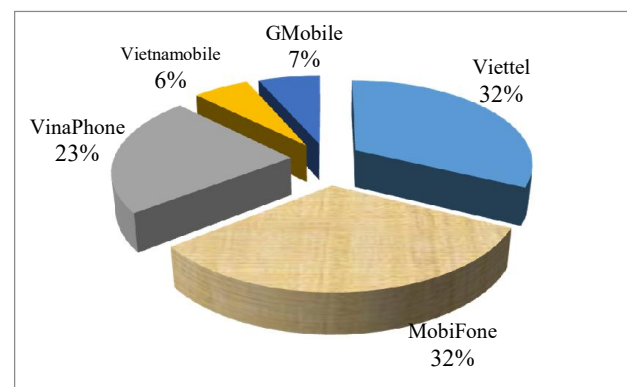
- Tương quan giữa thị phần dịch vụ và chỉ số tần suất xuất hiện chữ viết tắt: Bằng phép so sánh, đối chiếu, chúng tôi nhận thấy có sự tương quan tần suất CVT với những chỉ

số thị phần hay xếp hạng thương hiệu; chẳng hạn các ví dụ dưới đây:

- a) Tần suất xuất hiện 5 CVT (là thương hiệu nhà mạng di động) ngày 2/4/2016 so sánh với thị phần dịch vụ các nhà mạng (nguồn số liệu Bộ Thông tin và Truyền thông) năm 2015:

Bảng 3. Tương quan tần suất CVT và thị phần dịch vụ

STT	NGHIAVIET	N_VN1	Tần suất %	Thị phần %
1	Viettel	687.000	32,2	43,5
2	MobiFone	688.000	32,3	31,8
3	VinaPhone	489.000	22,9	17,5
4	Vietnamobile	127.000	6,0	4,1
5	GMobile	140.000	6,6	3,2
		2.131.000	100	100



Hình 3. Biểu đồ tần suất CVT là thương hiệu di động trên Internet

- b) Tần suất xuất hiện 8 cụm CVT (là thương hiệu thức ăn nhanh) ngày 02/04/2016 so sánh với thị phần thương hiệu bán luận trên mạng xã hội:

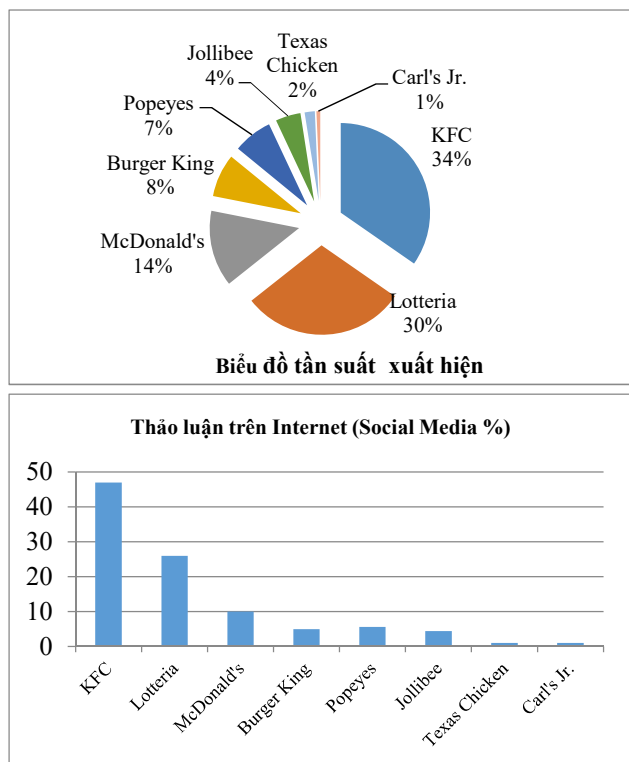
Bảng 4. Tương quan tần suất xuất hiện CVT và bán luận trên mạng xã hội

STT	NGHIAVIET	N_VN1	Tần suất %	Social Media %
1	KFC	146.000	34,7	47,0
2	Lotteria	125.000	29,7	26,0
3	McDonald's	57.900	13,8	10,0
4	Burger King	32.800	7,8	5,0
5	Popeyes	29.700	7,1	5,6
6	Jollibee	18.800	4,5	4,4
7	Texas Chicken	7.420	1,8	1,0
8	Carl's Jr.	2.760	0,7	1,0
		420.380	100	100

Nhận xét: Sự tương quan giữa chúng cho phép đưa ra cách tiếp cận: dựa vào tần suất xuất hiện CVT (thương hiệu, nhãn hiệu) có thể đưa ra chỉ số lượng hóa, sử dụng nó để dự đoán thị phần, mức độ phát triển thương hiệu, nhãn hiệu của doanh nghiệp.

Một số hạn chế của giải pháp: còn phụ thuộc vào dữ liệu có được trên yahoo.com, chương trình tìm kiếm là sự mô phỏng tự động việc tìm kiếm thủ công. Nếu dữ liệu đầu vào lớn, cần nhiều thời gian, nhiều máy tính để chạy chương trình và ghi nhận dữ liệu. Dữ liệu trên Internet sử

dụng bảng mã tiếng Việt nhiều loại là rào cản cho dữ liệu tìm kiếm chưa đầy đủ, hoàn toàn chuẩn xác.



Hình 4. Biểu đồ thương hiệu thức ăn nhanh thảo luận trên Internet (theo nguồn số liệu trên Internet)

4. Kết luận

Giải pháp đánh giá tần số sử dụng, tần suất xuất hiện CVT tiếng Việt trên Internet đã đưa ra một cách tiếp cận, một phương pháp thống kê mới trong đánh giá, lượng hóa chỉ số sử dụng và sự phát triển CVT trong thực tiễn. Chỉ số tần số, tần suất giúp nhà nghiên cứu quan sát, lưu trữ, thống kê và phân tích các hiện tượng phát triển ngôn ngữ tiếng Việt nói chung và CVT nói riêng qua từng thời điểm.

Những kết quả đạt được trong quá trình triển khai thực hiện giải pháp CVT cho phép chúng tôi tiếp tục công việc nghiên cứu để từng bước hoàn thiện dần bảng tần số, tần suất CVT tiếng Việt trên Internet, lưu trữ và công bố định kỳ trên Internet. Trong đó, có những phân tích, đưa ra tư vấn riêng cho các CVT là thương hiệu, nhãn hiệu nhằm trợ

giúp dự báo thị phần sản phẩm, dịch vụ hay phát triển thương hiệu cho doanh nghiệp.

Giải pháp có thể tiếp tục được mở rộng áp dụng đánh giá tần số CVT cho các ngôn ngữ dân tộc thiểu số Việt Nam (Chăm, ÊĐê, Thái, Kh'mer...). Ý tưởng và kết quả bài báo cho phép đề xuất một phương pháp thống kê ngôn ngữ học trên Internet, để quan sát và phân tích các hiện tượng phát triển ngôn ngữ. Chẳng hạn như tìm những từ, cụm từ mới xuất hiện, có thống kê tần suất xuất hiện (theo định kỳ thời gian), lưu trữ phân tích; từ đó quan sát được mức độ quan tâm truyền thông và sử dụng chúng, giúp ta thấy được sự phát triển của từ vựng trong ngôn ngữ học...

Tài liệu tham khảo

- [1] Nguyễn Quang Hồng (2002), *Nói tắt và viết tắt trong tiếng Việt Hiện trạng và giải pháp*, Báo cáo tổng quan về đề tài cấp bộ 2001-2002.
- [2] Phan Huy Khánh, Nguyễn Nho Túy (2006), "Nghiên cứu xây dựng cơ sở dữ liệu chữ viết tắt cho dịch vụ 1080 Bưu điện Đà Nẵng", *Kỷ yếu Hội thảo Khoa học Quốc gia "Một số vấn đề chọn lọc của CNTT & Truyền thông"*, 2006.
- [3] Phan Huy Khánh, Nguyễn Nho Túy (2006), *Abbreviations Application in 108 VNPT Service Exploitation in Da Nang City*, IJISSET (International Journal of Innovative Science, Engineering & Technology), Vol. 3 Issue 1, January 2016, p.222-227
- [4] Phan Huy Khánh, Nguyễn Nho Túy (2016), "Developing database of Vietnamese abbreviations and some applications", *Kỷ yếu Hội thảo ICTCC 2016 - 2nd EAI International Conference on Nature of Computation and Communication*.
- [5] Hội Ngôn ngữ học Việt Nam (2002), *Danh sách chữ viết tắt xếp theo tần số*, Hà Nội 2002.
- [6] Huỳnh Công Pháp, Nguyễn Văn Huệ (2014), "Nghiên cứu thu thập và xây dựng cơ sở dữ liệu chữ viết tắt tiếng Việt", *Tạp chí Khoa học Công nghệ ĐHQĐ. Số 7 (80)*.
- [7] Hồ Thị Kiều Oanh, Lê Thị Bích Chi (2012), "Nghiên cứu những đặc trưng ngôn ngữ của hiện tượng viết tắt trong thư điện tử tiếng Anh đối chiếu với tiếng Việt", *Tuyển tập Báo cáo Hội nghị Sinh viên Nghiên cứu Khoa học lần thứ 8 Đại học Đà Nẵng*, năm 2012.
- [8] Lê Bá Long (2006), *Lý thuyết xác suất thống kê toán*, Giáo trình Học viện Bưu chính Viễn thông Việt Nam.
- [9] Manuel Zahariev (2004). *Acronyms*. Simon Fraser University, Jun 2004.
- [10] Lưu Tuấn Anh (2012). *Download dữ liệu các tập Corpus*, link: <http://viet.jnlp.org/download-du-lieu-tu-vung-corpus>
- [11] Lê Thị Bích Chi (2012). *Nghiên cứu đặc trưng ngôn ngữ của hiện tượng viết tắt trong thư điện tử tiếng Anh đối chiếu với tiếng Việt* Link: <http://www.udn.vn/app/webroot/svnckh2012/PDF/TB20-03.pdf>
- [12] Nguyễn Nho Túy (2015). *Web site thư viện chữ viết tắt*. Link: <http://www.chuviattat.com>

(BBT nhận bài: 14/04/2016, phân biện xong: 16/08/2016)