PHÂN LOẠI NGƯỜI DÙNG WEB SỬ DỤNG KỸ THUẬT SO SÁNH CHUỖI

LƯU VĨNH TRUNG
Trường Đại học Mở Thành phố Hồ Chí Minh – trung.lv@ou.edu.vn

(Ngày nhận: 17/03/2017; Ngày nhận lại: 11/04/2017; Ngày duyệt đăng: 08/05/2017)

TÓM TẮT

Ngày nay cùng với sự phát triển của thương mại điện tử, nhu cầu tìm hiểu sở thích của người dùng để tối ưu hóa lợi nhuận ngày càng tăng. Sở thích được thể hiện qua hành vi của người dùng trong quá trình duyệt web hoặc các ứng dụng liên quan thương mại điện tử khác. Bài báo này trình bày cách tiếp cận sử dụng kỹ thuật so sánh chuỗi trên các phiên duyệt web để đánh giá sự tương tự trong hành vi người dùng và phân loại họ. Kết quả phân loại này có thể sử dụng để dự đoán hành vi người dùng web trong thời gian thực, và có những đề xuất duyệt web phù hợp với từng loại người dùng.

Từ khóa: Khai phá dữ liệu web; so sánh chuỗi; phân loại người dùng; thương mại điện tử.

Web user segmentation using sequence alignment

ABSTRACT

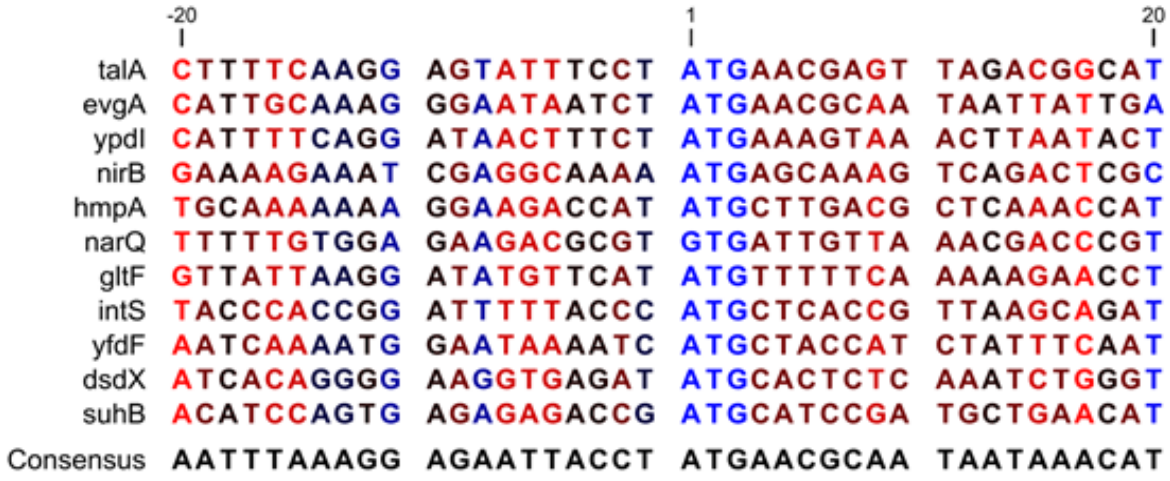
Nowadays, with the rapid advances in e-commerce, user interest understanding becomes more and more essential in order to benefit the business. Users reveal this kind of interest through their behavior during their sessions in e-commerce applications. In this paper, we present the approach using sequence alignment for web sessions to evaluate the user behavior similarity in order to segment them. The segmentation result is applicable for real-time web prediction and recommendation.

Keywords: Web mining; sequence alignment; user segmentation; e-commerce.

1. Giới thiệu

Các chiến lược tiếp thị trên Internet dựa trên hành vi người dùng đang nhận được sự quan tâm ngày càng lớn của các doanh nghiệp thương mại điện tử. Hoạt động của các chiến lược dạng này dựa trên việc thích nghi ứng dụng thương mại điện tử với hành vi người dùng trong thời gian thực, khi họ đang truy cập ứng dụng. Để đạt được mục đích này, các công cụ tính toán nhanh sự tương tự giữa các phiên truy cập là thiết yếu, nhằm xác định người dùng thuộc nhóm tương ứng nào. Mức độ tương tự này được sử dụng để gom nhóm các phiên truy cập, và qua đó phân loại người dùng web (Cooley, R. và cộng sự, 1997). Phiên truy cập có thể được xem như chuỗi các sự kiện, nên để đơn giản hóa phần trình bày trong bài báo này, chúng tôi sử dụng chuỗi ký tự như A-B-C-D-E để đại diện cho chuỗi các trang web được thăm viếng trong phiên truy cập.

Kỹ thuật so sánh chuỗi đã được ứng dụng từ rất lâu trong Công nghệ Sinh học và các ngành liên quan, nhằm tìm ra những đoạn tương tự nhau giữa các chuỗi RNA, ADN hoặc protein (Hình 1). Hai hướng tiếp cận chính trong kỹ thuật này là so sánh toàn cục (global alignment) và so sánh cục bộ (local alignment) để đánh giá một cách toàn diện sự tương tự giữa các chuỗi. Hai thuật toán tiêu biểu và được áp dụng rộng rãi, lần lượt đại diện cho so sánh toàn cục và cục bộ là Needleman-Wunsh (Needleman, S.B. và cộng sự, 1970) và Smith-Waterman (Smith, T.F., 1981; Zahid, S.K., 2015).



Hình 1. So sánh các chuỗi trong Công nghệ Sinh học nhằm phát hiện mức độ tương tự

2. Phương pháp nghiên cứu

Như đã đề cập, so sánh toàn cục và so sánh cục bộ đánh giá mức độ tương tự của các chuỗi theo những cách khác nhau. Needleman-Wunsh (NW) có xu hướng tìm kiếm sự tương tự tổng quát trên suốt chiều dài của các chuỗi, vì vậy thuật toán này rất hiệu quả trên các chuỗi có chiều dài tương đương nhau (Hình 2). Smith-Waterman (SW), ngược lại, tập trung vào những vùng tương tự giữa hai chuỗi nên thích hợp với các chuỗi có chiều dài chênh lệch (Hình 3).

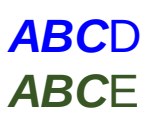
Hình 2. So sánh chuỗi toàn cục

Hình 3. So sánh chuỗi cục bộ

Trong bài báo này, để đánh giá mức độ tương tự giữa hai chuỗi cho từng thuật toán, chúng tôi dùng thang đo +1 cho cặp phần tử giống nhau và -1 cho cặp phần tử khác nhau khi so sánh chuỗi sử dụng NW. Với SW, thang đo tương ứng là +2 và -1 tương ứng, vì SW tập trung vào những vùng tương tự rời rạc giữa hai chuỗi. Với thang đo này, sự khác biệt trong cách so sánh chuỗi được thể hiện rõ trong các ví dụ sau (Hình 4, 5, 6, 7, 8):

Hình 4. So sánh hai chuỗi có độ dài chênh lệch có một phần tử tương tự, kết quả SW = 2

Hình 5. So sánh hai chuỗi trùng lặp, kết quả NW = 2



Hình 6. So sánh hai chuỗi có độ dài như nhau có các phần tử tương tự, kết quả NW = 2

Cặp chuỗi trong hình 4 có độ dài rất chênh lệch và chỉ có một phần tử chung. Trong khi hai cặp chuỗi trong hình 5 và 6 có độ dài tương đương và có nội dung trùng lặp hoặc nhiều phần tử tương tự. Tuy nhiên đánh giá về độ tương tự của của SW cho cặp chuỗi hình 4 và của NW cho hai cặp chuỗi hình 5 và 6 là giống nhau. Điều này cho thấy sự khác biệt của hai thuật toán trong đánh giá độ tương tự của các cặp chuỗi. Một ví dụ khác về sự khác biệt này được trình bày tại hình 7 và 8. Hai cặp chuỗi đều có điểm NW = 0, nhưng điểm SW của cặp chuỗi hình 7 (4) cao hơn cặp chuỗi hình 8 (3) vì độ liên tục của các phần tử tương tự trong hình 7 cao hơn.

Hình 7. So sánh hai chuỗi có độ dài tương đương có các phần tử tương tự theo thứ tự, kết quả SW = 4, NW = 0

Hình 8. So sánh hai chuỗi có độ dài tương đương có các phần tử tương tự theo thứ tự, kết quả SW = 3, NW = 0

Cặp chuỗi để so sánh có độ dài càng khác biệt, NW càng cho thấy sự không phù hợp của thuật giải này trong việc đánh giá độ tương tự. Như trình bày tại Bảng 1, NW đánh giá cặp (ABC, BCD) có độ tương tự thấp hơn (ABC, ABCDEFGHIJKLMNO). Do đó, NW cần được kết hợp với thuật giải khác tập trung vào sự tương tự cục bộ để có được kết quả tối ưu và phù hợp với ngữ cảnh của các phiên truy cập trên web.

Bảng 1

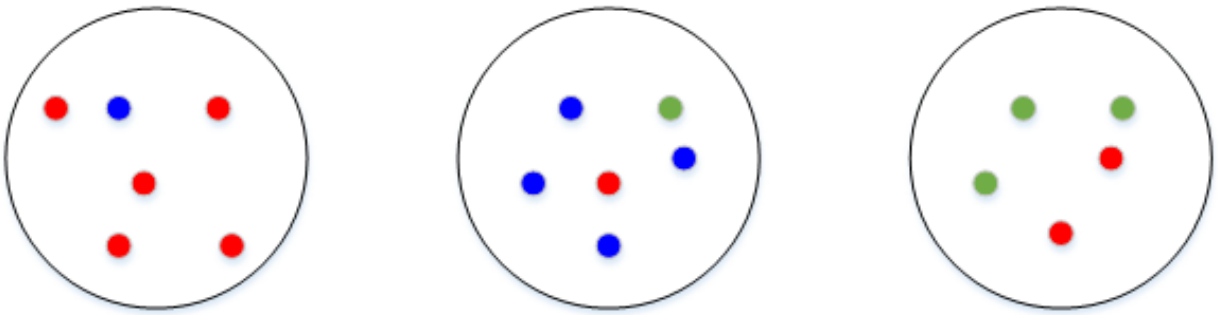
Độ tương tự đo bởi NW trên một số cặp chuỗi có độ dài khác biệt nhau

	ABCDEFGHIJKLMNO	**ABC**	**BCD**	**ABCDPFQHRJSLTNU**	**AAAAAAAAABCD**
ABCDEFGHIJKLMNO		3.0	3.0	3.0	-10.2
ABC	3.0		0	3.0	2.99
BCD	3.0	0		3.0	2.99
ABCDPFQHRJSLTNU	3.0	3.0	3.0		-10.2
AAAAAAAAABCD	-10.2	2.99	2.99	-10.2	

Chúng tôi đề xuất một sự kết hợp giữa NW và SW trong việc đánh giá sự tương tự giữa các cặp chuỗi đại diện cho phiên truy cập web của người dùng. Để chứng minh cho ưu điểm của sự kết hợp NW và SW thay vì ứng dụng riêng lẻ, chúng tôi đưa ra kết quả về độ tinh khiết (purity) của cụm (cluster) trong ba trường hợp:

1. Ứng dụng NW
2. Ứng dụng SW
3. Ứng dụng kết hợp NW và SW.

Độ tinh khiết của cụm cho thấy hiệu quả của thuật toán phân cụm. Thuật toán càng hiệu quả, các phần tử của cụm càng đồng nhất, độ tinh khiết của cụm càng cao. Hình 9 minh họa độ tinh khiết của ba cụm, với các phần tử đồng nhất có màu giống nhau.



Hình 9. Purity = 5/6 Purity = 4/6 Purity = 3/5

3. Kết quả

Như đề xuất trong phần trước, chúng tôi thực nghiệm các ứng dụng riêng lẻ và kết hợp của NW và SW trên dữ liệu người dùng được trích xuất từ website http://www.campus-fonderie.uha.fr/. Dịch vụ được triển khai phía back-end của trang web này cho phép thu thập dữ liệu của các phiên truy cập, như các trang được thăm viếng, thời gian, thời điểm… tương ứng, và trả về log file với các định dạng như .csv, .txt… Log file được làm sạch (cleaning) để loại trừ dữ liệu bị lỗi/không hợp lệ trước khi áp dụng các thuật toán clustering phân cụm. Log file bao gồm nhiều phiên truy cập, mỗi phiên chứa ít nhất một trang web được viếng thăm, sau đây là ví dụ rút gọn của một phiên truy cập được ghi nhận trên log file:

Mã phiên truy cập	URLs
000001	http://www.campus-fonderie.uha.fr/fr/droit/
000001	http://www.campus-fonderie.uha.fr/fr/economie-et-societe/
000001	http://www.campus-fonderie.uha.fr/fr/management/
000001	http://www.campus-fonderie.uha.fr/fr/management-interculturel/

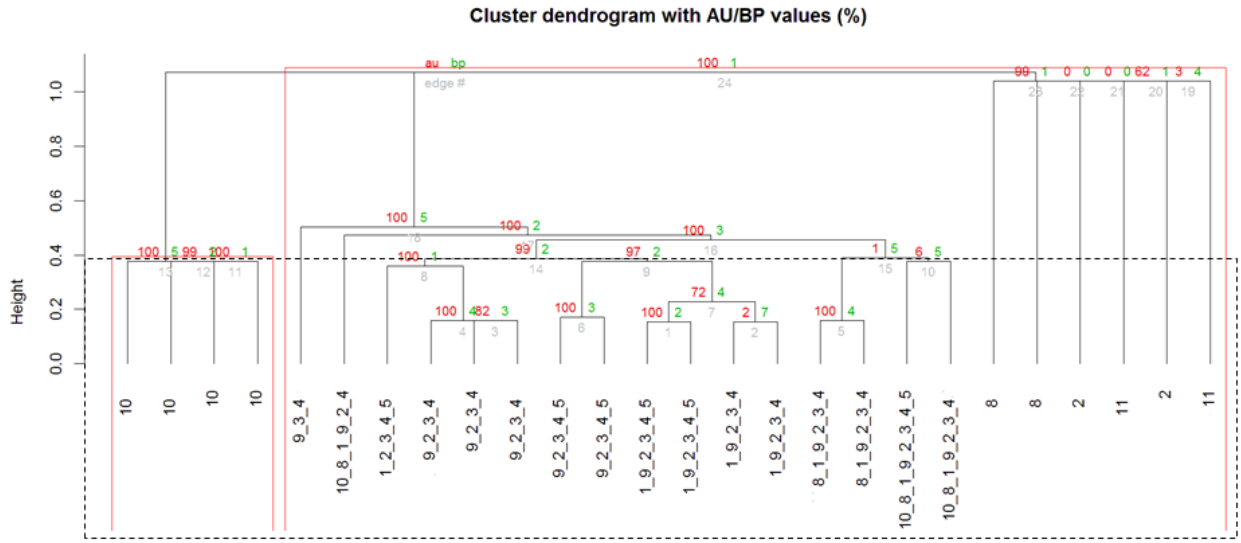
Để tăng hiệu quả của thuật toán clustering và số lượng URL có thể xử lý, các URL sẽ được đại diện bằng các chữ số. Ví dụ, phiên truy cập 000001 trên bao gồm 4 URL 1_2_3_4. Kết quả độ tinh khiết của cụm, sau khi ứng dụng riêng lẻ và kết hợp NW và SW trên log file gồm 2000 phiên truy cập, được trình bày tại Bảng 2:

Bảng 2

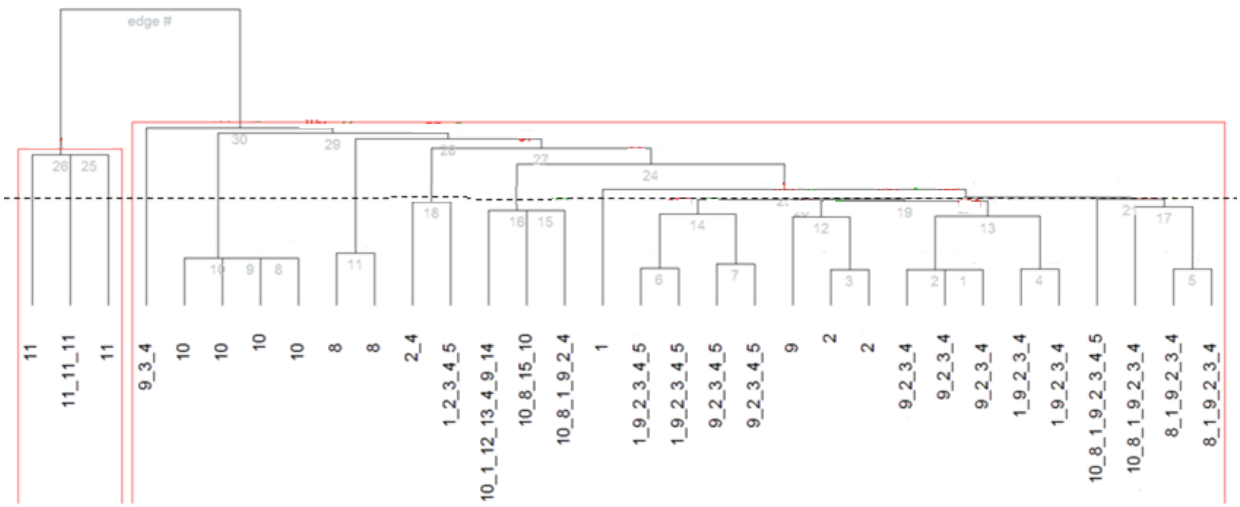
Kết quả độ tinh khiết của cụm qua các ứng dụng riêng lẻ và kết hợp NW và SW

	Điểm NW > ¼ độ dài chuỗi dài hơn	**Điểm SW gấp đôi độ dài chuỗi ngắn hơn**	**Điểm NW > ¼ độ dài chuỗi dài hơn *và* điểm SW gấp đôi độ dài chuỗi ngắn hơn**
Độ tinh khiết của cụm	**81%**	**63%**	**92%**

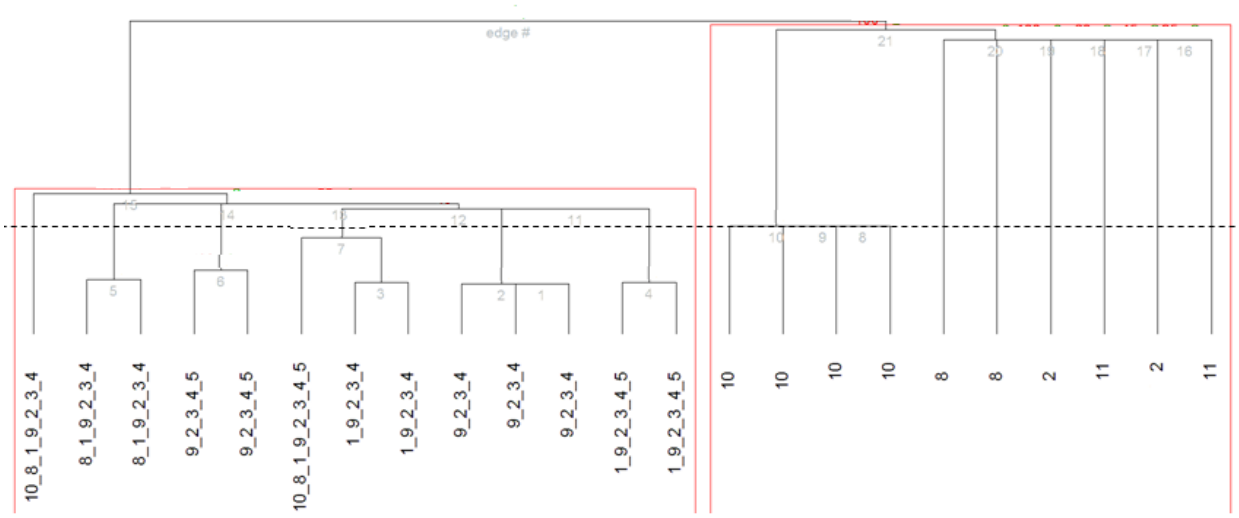
Hình 10, 11, 12 lần lượt minh họa kết quả phân cụm bằng NW, SW và kết hợp NW và SW trên dữ liệu gồm 32 phiên truy cập đại diện. Sau khi áp dụng NW và SW riêng lẻ và kết hợp như bộ lọc, số phiên truy cập tương ứng trên hình 10, 11, 12 lần lượt là 26, 32 và 23. Việc áp dụng NW khiến các phiên truy cập tương tự nhau một cách toàn cục, nhưng 10_8_1_ 9_ 2_ 4 hoặc 1_ 2_ 3_ 4_ 5 là ví dụ về sự không tương tự cục bộ so với các phiên truy cập khác. Ngược lại, SW khiến 10_ 1_ 12_ 13_ 4_ 9_ 14, 9_3_4, 11_11_11, 10_8_15_10, 10_8_1_9_2_4 là những phiên truy cập không có sự tương tự toàn cục so với các phiên còn lại, xuất hiện trong các phân cụm. Còn sự kết hợp giữa NW và SW tối ưu hơn trong việc gom nhóm các phiên truy cập, số phiên ít hơn nhưng chọn lọc được các phiên tương tự nhau về toàn cục cũng như cục bộ.



Hình 10. Kết quả phân cụm bằng hierarchical clustering khi điểm NW > ¼ độ dài chuỗi dài hơn



Hình 11. Kết quả phân cụm bằng hierarchical clustering khi điểm SW gấp đôi độ dài chuỗi ngắn hơn



Hình 12. Kết quả phân cụm bằng hierarchical clustering khi điểm NW > ¼ độ dài chuỗi dài hơn và điểm SW gấp đôi độ dài chuỗi ngắn hơn.

4. Kết luận

Kỹ thuật so sánh chuỗi được sử dụng phổ biến Công nghệ Sinh học, cũng được ứng dụng trong việc phân cụm các phiên truy cập web để tìm các nhóm người dùng tương tự nhau (Wang và cộng sự, 2002). Tuy nhiên, vì kỹ thuật so sánh chuỗi vốn không được tạo ra để sử dụng trên dữ liệu web, nó cần phải được phát triển tối ưu cho mục tiêu này. Cách tiếp cận của chúng tôi dựa trên sự kết hợp của hai kỹ thuật so sánh chuỗi toàn cục và cục bộ, mà đại diện là Needleman-Wunsh và Smith-Waterman, qua thực nghiệm đã chứng tỏ sự hiệu quả và thực tế khi làm việc trên dữ liệu các phiên truy cập của người dùng web.

Chúng tôi có kế hoạch phát triển một thang đo chính thức dựa trên sự kết hợp của hai kỹ thuật so sánh chuỗi toàn cục và cục bộ này, để việc phân cụm các phiên truy cập web, và qua đó tự động gom nhóm người dùng, được nhanh chóng và hiệu quả hơn với lượng dữ liệu ngày càng lớn từ các thiết bị sử dụng Internet phong phú hiện nay■

Tài liệu tham khảo

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Grouping web page references into transactions for mining world wide web browsing patterns. *IEEE Knowledge and Data Engineering Exchange Workshop Proceedings*, 2-9.

Needleman, S.B., & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

Smith, T.F., & Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.

Wang, W., & Zaiane, O.R. (2002). Clustering web sessions by sequence alignment. *Database and Expert Systems Applications Proceedings*, 394-398.

Zahid, S. K., Hasan, L., Khan, A. A., & Ullah, S. (2015). A novel structure of the Smith-Waterman Algorithm for efficient sequence alignment. *Digital Information, Networking, and Wireless Communications*, 6-9.