

TẠI SAO BẠN KHÔNG HÀNH ĐỘNG ĐÚNG VỚI ĐỘ TUỔI CỦA MÌNH?

NHẬN BIẾT KHUÔN MẪU CỦA TRÉ 8-12 TUỔI

NGƯỜI TÌM KIẾM THEO HÀNH VI TÌM KIẾM CỦA HỌ

qua

Michael Xanh



Một luận án

nộp một phần để hoàn thành

của các yêu cầu về mức độ

Thạc sĩ Khoa học Máy tính

Đại học Boise State

Tháng 8 năm 2021



TRƯỜNG CAO ĐẲNG ĐẠI HỌC BANG BOISE

ỦY BAN QUỐC PHÒNG VÀ PHÊ DUYỆT ĐỌC CUỐI CÙNG

của luận án được nộp bởi

Michael Xanh

Tiêu đề luận văn: Tại sao bạn không hành động theo đúng độ tuổi của mình? Nhận biết những người tìm kiếm theo khuôn mẫu từ 8-12 tuổi qua hành vi tìm kiếm của họ

Ngày thi vấn đáp cuối kỳ: 14 tháng 8 năm 2021

Những cá nhân sau đây đã đọc và thảo luận luận án do sinh viên Michael nộp Green, và họ đã đánh giá bài thuyết trình và phản hồi các câu hỏi trong phần cuối kỳ thi vấn đáp. Họ thấy rằng sinh viên đã vượt qua kỳ thi vấn đáp cuối cùng.

Tiến sĩ Maria Soledad Pera

Chủ tịch, Ủy ban giám sát

Tiến sĩ Michael Ekstrand

Thành viên, Ủy ban giám sát

Casey Kennington, Tiến sĩ

Thành viên, Ủy ban giám sát

Sự chấp thuận đọc cuối cùng của luận án đã được chấp thuận bởi Maria Soledad Pera, Tiến sĩ, Chủ tịch Ủy ban giám sát. Luận án đã được Hội đồng sau đại học chấp thuận

Trường cao đẳng.

Dành tặng Keith Baderman và Marino Francioni.

## LỜI CẢM ƠN

Tác giả xin bày tỏ lòng biết ơn tới Tiến sĩ Maria Soledad Pera, Shane Panter và Luke Hindman. Niềm đam mê và sự khích lệ của bạn đã liên tục được truyền cảm hứng. Tác giả cũng muốn cảm ơn Ashlee Milton và Garrett Allen vì sự hỗ trợ của họ. Làm cho quá trình dễ dàng hơn bằng cách chia sẻ gánh nặng, cảm xúc và khoa học. Hơn nữa, tác giả muốn bày tỏ lòng biết ơn tới Boise State Khoa Khoa học máy tính của trường đại học đã tài trợ cho nghiên cứu này. Cảm ơn Bạn.

## TÓM TẮT

Các công cụ tìm kiếm trực tuyến dành cho trẻ em được biết đến là có thể lọc các tài nguyên đã truy xuất dựa trên độ phức tạp của trang và cung cấp chức năng chuyên biệt nhằm giải quyết các khoảng trống trong tìm kiếm trình độ đọc viết theo độ tuổi hoặc trình độ của người dùng. Tuy nhiên, không phải mọi người tìm kiếm đều được nhóm theo các định danh này hiển thị cùng một mức độ hiểu văn bản hoặc yêu cầu cùng một hỗ trợ tìm kiếm. Hơn nữa, các công cụ tìm kiếm này thường dựa vào phản hồi trực tiếp để xác định những định danh này. Sự phụ thuộc vào việc tự nhận dạng này có thể khiến người dùng vô tình trình bày sai lệch về bản thân. Do đó, chúng tôi tìm cách nhận ra người dùng từ kỹ năng dựa trên các tín hiệu thay vì sử dụng các mã định danh tuổi hoặc lớp, vì kỹ năng quyết định phù hợp viện trợ và tài nguyên. Do đó, trong luận án này chúng tôi đề xuất một chiến lược tự động nhận dạng người dùng ngay lập tức bằng cách phân tích hành vi tìm kiếm được tìm thấy trong các phiên tìm kiếm. Trong cụ thể, nỗ lực của chúng tôi tập trung vào việc nhận dạng người tìm kiếm theo khuôn mẫu từ 8 đến 12 tuổi, người mà chúng tôi đặt ra thể hiện các kỹ năng được xác định theo các giai đoạn phát triển có tác động mạnh mẽ về sự phát triển ngôn ngữ (giai đoạn hoạt động cụ thể của Piaget) và tìm kiếm khả năng đọc viết (mức độ đầu tiên của năng lực số). Chiến lược này phân tích văn bản do người dùng tạo ra được trích xuất từ các truy vấn và các mẫu tương tác tìm kiếm để suy ra các tính năng được sử dụng bởi một bộ phân loại rừng ngẫu nhiên để xác định xem người dùng có là một phần của phân khúc cụ thể này của người tìm kiếm. Các kết quả từ luận án này đặt ra nền tảng cho phép các công cụ tìm kiếm nhận dạng người dùng dựa trên kỹ năng tìm kiếm của họ và cung cấp cái nhìn sâu sắc hơn về hành vi tìm kiếm của thanh thiếu niên.

## MỤC LỤC

TÓM TẮT . . . . .	
DANH SÁCH BẢNG . . . . .	ix
DANH SÁCH HÌNH ẢNH . . . . .	xi
1 Giới thiệu . . . . .	1
1.1 Phát biểu luận đề. . . . .	7
2 Bối cảnh và công trình liên quan. . . . .	8
2.1 Nhật ký truy vấn và phiên tìm kiếm . . . . .	8
2.2 Các lý thuyết phát triển. . . . .	9
2.3 Xác định loại người dùng . . . . .	10
3 Phương pháp . . . . .	12
3.1 Tính năng dựa trên văn bản. . . . .	12
3.1.1 Đặc điểm từ vựng . . . . .	13
3.1.2 Từ vựng . . . . .	15
3.1.3 Chính tả và dấu câu . . . . .	16
3.1.4 Cú pháp . . . . .	18
3.2 Tính năng dựa trên phiên . . . . .	19
3.2.1 Tương tác dựa trên truy vấn. . . . .	21
3.2.2 Tương tác dựa trên nhấp chuột . . . . .	22

3.3 Phân loại . . . . .	22
4 Thiết lập thử nghiệm . . . . .	24
4.1 Bộ dữ liệu . . . . .	24
4.1.1 Bộ dữ liệu Phiên có nhấp chuột. . . . .	24
4.1.2 Phiên truy vấn đơn. . . . .	27
4.2 Đường cơ sở . . . . .	27
4.3 Số liệu . . . . .	30
4.4 Chuẩn bị thí nghiệm. . . . .	31
4.4.1 Điều chỉnh siêu tham số. . . . .	31
4.4.2 Phân chia dữ liệu để thử nghiệm và đào tạo. . . . .	32
4.5 Xác thực kết quả của chúng tôi . . . . .	34
5 Kết quả và Phân tích . . . . .	35
5.1 Tính hiệu quả của tính năng . . . . .	35
5.2 So sánh với đường cơ sở. . . . .	39
5.3 Tác động của thời lượng buổi học đến hiệu quả. . . . .	43
5.4 Thảo luận . . . . .	45
6 Kết luận, Hạn chế và Công việc Tương lai . . . . .	48
TÀI LIỆU THAM KHẢO . . . . .	53



## DANH SÁCH CÁC BẢNG

3.1 Đặc điểm từ vựng, trong đó từ phức tạp (SWtypes), từ vựng (Slex)	
và các loại động từ (SVtypes), được định nghĩa là các từ hoặc loại không tìm thấy	
trong “2.000 từ thường gặp nhất của British National Corpus” [44],	
Nsim và Ncom được định nghĩa là những từ có ít hơn 3 âm tiết và những từ có 3	
âm tiết hoặc nhiều hơn, tương ứng. . . . .	14
3.2 Đặc điểm từ vựng, với inter là danh sách các câu hỏi sau đây	
từ: “ai”, “cái gì”, “khi nào”, “ở đâu”, “tại sao”, “như thế nào”, “là”, “là”,	
“có thể”, “có thể”, “nên”, và “sẽ”, và tất cả các biến được xác định bởi	
được chứa trong truy vấn dưới dạng các giá trị boolean. . . . .	17
3.3 Các tính năng chính tả và dấu câu, trong đó Nc biểu thị danh sách các đề xuất	
sửa lỗi đánh máy được đề xuất, LD là viết tắt của Levenhstein Distance	
(được đo ở cấp độ ký tự) và Qpunct được tính bằng cách đầu tiên en-	
đảm bảo rằng truy vấn không chứa tiền tố/hậu tố tìm kiếm (như đã định nghĩa	
trong Bảng 3.2), sau đó tính toán giá trị boolean dựa trên việc liệu	
truy vấn chứa dấu câu trong danh sách sau: .!?, . . . . .	17
3.4 Các đặc điểm của phần lời nói được định nghĩa bởi phần lời nói NLTK	
người gán thẻ. . . . .	20
3.5 Các tính năng cấp D, được tính toán bằng trình phân tích cấp D	
có thể được tìm thấy tại [5]. . . . .	21

3.6 Các tính năng tương tác dựa trên truy vấn trong đó $Setq$ là tập hợp tất cả các truy vấn,	
$Q_i$ là truy vấn thứ $i$ trong một phiên và $TimestampQ_i$ là dấu thời gian	
cho truy vấn thứ $i$ . Khoảng cách Levenhstein giữa các truy vấn được đo trên một	
cấp độ nhân vật. . . . .	22
3.7 Các tính năng nhấp chuột trong đó $C_i$ là lần nhấp thứ $i$ trong phiên và $TimestampC_i$	
là dấu thời gian của lần nhấp thứ $i$ trong phiên. . . . .	23
4.1 Nguồn dữ liệu được sử dụng để tạo tập dữ liệu. . . . .	25
4.2 Mô tả các tập dữ liệu được sử dụng trong các thí nghiệm của chúng tôi. . . . .	26
5.1 Kết quả từ nghiên cứu cắt bỏ trên các Phiên có Nhấp chuột. * biểu thị	
ý nghĩa thống kê của một tập hợp tính năng nhất định liên quan đến $RYSe(p \leq .05)$ . . . . .	36
5.2 Kết quả từ nghiên cứu cắt bỏ trên Phiên truy vấn đơn. * chỉ ra	
ý nghĩa thống kê của một tập hợp tính năng nhất định liên quan đến $RYSe(p \leq .05)$ . . . . .	38
5.3 Kết quả đánh giá hiệu suất của chúng tôi khi so sánh $RYSe$ với	
đường cơ sở trên Phiên có Nhấp chuột. * biểu thị ý nghĩa thống kê	
của một đường cơ sở nhất định liên quan đến $RYSe(p \leq .05)$ . . . . .	41
5.4 Kết quả đánh giá hiệu suất của chúng tôi khi so sánh $RYSe$ với	
đường cơ sở trên Phiên truy vấn đơn. * biểu thị ý nghĩa thống kê	
của một đường cơ sở nhất định liên quan đến $RYSe(p \leq .05)$ . . . . .	42
5.5 Đánh giá hiệu suất của $RYSe$ (kết quả bên trái) so với	
Bộ phân loại đa tính năng (kết quả bên phải) trên các phiên có nhiều tính năng khác nhau	
chiều dài. * biểu thị ý nghĩa thống kê ( $p \leq .05$ ) liên quan đến $RYSe$	
kết quả trên các phiên có cùng độ dài. . . . .	45

DANH SÁCH CÁC HÌNH ẢNH

1.1 Ví dụ trực quan về sự khác biệt giữa dữ liệu được tạo ra bởi Twitter  
sử dụng (xem trong Hình 1.1a), chứa nhiều tương tác văn bản,  
dữ liệu hồ sơ và hình ảnh; so với dữ liệu được tạo ra bởi các tương tác với SE  
(xem trong Hình 1.1b), chứa các truy vấn, nhấp chuột và dấu thời gian. . . . . 3

4.1 Hình ảnh làm nổi bật cách chúng ta tạo ra quá trình đào tạo/điều chỉnh/kiểm tra rời rạc của mình  
tập hợp con để điều chỉnh siêu tham số từ Phiên có nhấp chuột. . . . . 32

4.2 Hình ảnh làm nổi bật cách chúng tôi tạo ra các phân chia dữ liệu đào tạo/kiểm tra của mình,  
chứa thông tin về cách kết hợp các phần đào tạo và  
những phần nào của mỗi tập dữ liệu được thử nghiệm. . . . . 33

## CHƯƠNG 1

### GIỚI THIỆU

Trẻ em từ lớp K-81 thường xuyên sử dụng Công cụ tìm kiếm (SE) [28, 63, 64] cụ thể được thiết kế cho họ để hoàn thành bài tập ở trường cũng như đáp ứng sự tò mò [6, 27, 59]. Tuy nhiên, những người tìm kiếm trẻ tuổi này có thể gặp phải những khó khăn khi hoàn thành các tìm kiếm thành công [30], một phần là do sự cố khi xây dựng các truy vấn [21], hiểu kết quả thu được [10] hoặc điều hướng các trang kết quả SE [34]. Gốc của những khó khăn này thường liên quan đến kỹ năng của người dùng, hoặc thiếu kỹ năng, trong việc đọc, viết, và khả năng tìm kiếm [34]. Người dùng có cùng một bộ kỹ năng thường gặp phải các vấn đề tương tự [7]. SE dành cho trẻ em cung cấp hỗ trợ thuật toán để giải quyết những vấn đề này cuộc đấu tranh [32] thường nhắm mục tiêu vào người dùng dựa trên độ tuổi hoặc cấp bậc [21, 27, 32]. Điều này bao gồm các chiến lược giúp dễ dàng xây dựng truy vấn [21], cung cấp các gợi ý truy vấn phù hợp chủ đề tìm kiếm của trẻ em [49], thu thập và cung cấp các nguồn tài nguyên có liên quan và dễ hiểu thông qua việc tuyển chọn nội dung [13], tức là các trang web được thu thập thủ công cho đối tượng được chọn; và sử dụng các giao diện thích ứng [32], tức là các giao diện cho phép nội dung được truy xuất được điều chỉnh cho người dùng dựa trên phản hồi rõ ràng của họ. Tuy nhiên, nhắm mục tiêu người dùng bằng phạm vi rộng của độ tuổi [8, 21] và các lớp [4] hoạt động theo giả định rằng những nhóm người dùng chia sẻ một bộ kỹ năng thống nhất. Tuy nhiên, chúng tôi biết rằng người dùng trong cùng độ tuổi hoặc cùng lớp có khả năng khác nhau về đọc, viết và tìm kiếm khả năng đọc viết [28]. Các nguồn lực và hỗ trợ được coi là phù hợp với người dùng được công nhận theo độ tuổi/lớp có thể

---

<sup>1</sup>Trong hệ thống trường học Hoa Kỳ, trẻ em từ mẫu giáo đến lớp 8 thường có độ tuổi từ 5 đến 14.

không đáp ứng được khả năng thực sự của người dùng, cung cấp nội dung có thể thấp hơn nhiều hoặc vượt quá trình độ hiểu biết của họ hoặc cung cấp các công cụ có thể gây ra sự thất vọng hơn là cho phép dễ dàng. Cung cấp cho SE một cách để nhận ra khả năng của người dùng của họ có thể cho phép các nền tảng này có một cách để đảm bảo rằng các nguồn lực phù hợp đang đến được người dùng phù hợp. Do đó, chúng tôi coi việc xác định cách những người tìm kiếm trẻ tuổi có thể được công nhận bởi các kỹ năng tiềm ẩn của họ, vì hiện tại không có chiến lược nào để thực hiện điều đó.

Một cách để nhận ra người dùng theo bộ kỹ năng của họ là xác định loại người dùng đó [8, 32]-một phương pháp xác định và tận dụng các đặc điểm chính của người dùng để để phân biệt chúng với nhau và xếp chúng vào các nhóm [46]. Việc nhóm là thường dựa trên những người dùng có chung đặc điểm, chẳng hạn như tuổi tác và giới tính [50, 67], tính cách [65] và trình độ chuyên môn [69]. Nghiên cứu trong các lĩnh vực bên ngoài của SE đã khám phá những cách tiếp cận thành công trong việc phát hiện trẻ em dựa trên về tương tác của họ trong: phòng trò chuyện [46, 67], trang web truyền thông xã hội [3] và trực tuyến bảng câu hỏi [55]. Thật không may, những tương tác giữa người dùng và hệ thống này khác rất nhiều so với tương tác người dùng-SE (như được thấy trong Hình 1.1), thường chứa các đoạn văn bản so với các truy vấn thừa thớt văn bản, ngăn chặn ứng dụng đa miền. Hơn nữa, các chiến lược này sử dụng dữ liệu lịch sử cụ thể của từng miền để tạo ra những sự khác biệt này (blog, trang mạng xã hội, phòng chat). Do các biện pháp nghiêm ngặt của COPPA (Trẻ em) Đạo luật bảo vệ quyền riêng tư trực tuyến của trẻ em), thiếu dữ liệu lịch sử ghi lại tương tác của người tìm kiếm trẻ với SE [26].

SE thường giải quyết tình trạng thiếu dữ liệu lịch sử này bằng cách dựa vào người dùng để xác định loại người dùng thông qua phản hồi trực tiếp. Thông thường, phản hồi này được rút ra từ hồ sơ người dùng có chứa thông tin người dùng có liên quan [65, 66, 67]. Tuy nhiên, SE dành cho trẻ em giải quyết tình trạng thiếu thông tin hồ sơ này bằng cách yêu cầu người dùng xác định loại của họ khi mục nhập [32, 4]. Thật không may, loại phản hồi trực tiếp này có thể không đáng tin cậy vì người dùng

có thể nói dối về danh tính của họ [25], có thể thiếu khả năng tìm kiếm để tự xác định đúng tức là vô tình nhấp vào độ tuổi hoặc lớp sai và không biết cách hoàn tác lựa chọn đó [8], hoặc người dùng hiện tại có thể thay đổi sau khi đăng nhập. Ngoài ra, trực tiếp phản hồi thường được sử dụng để nhận dạng độ tuổi hoặc cấp bậc của người dùng vì chúng có liên quan đến kỹ năng. Ngay cả khi người dùng có thể tự xác định chính xác độ tuổi hoặc lớp của mình bằng cách sử dụng trực tiếp phản hồi, chúng tôi vẫn còn vấn đề ban đầu của chúng tôi. Không có đảm bảo rằng người dùng sẽ “hành động theo độ tuổi của họ” [43](hoặc độ tuổi mà lớp của họ tương ứng) như sự phát triển, thay vì tuổi tác, quyết định cách trẻ em cư xử, cũng như xác định các kỹ năng tiềm ẩn của chúng [56].



(một) Twitter

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	www.prescriptionfortune.com	2006-03-12 12:31:06		
142	staple.com	2006-03-17 21:19:29		
142	staple.com	2006-03-17 21:19:45		
142	www.newyorklawyersite.com	2006-03-18 08:02:58		
142	www.newyorklawyersite.com	2006-03-18 08:03:09		
142	westchester.gov	2006-03-20 03:55:57	1	http://www.westchestergov.com
142	space.comhttp	2006-03-24 20:51:24		
142	dfdf	2006-03-24 22:23:07		
142	dfdf	2006-03-24 22:23:14		
142	vanlga.comh	2006-03-25 23:27:12		
142	www.collegeucla.edu	2006-04-03 21:12:14		
142	www.elaorg	2006-04-03 21:25:20		
142	207 ad2d 530	2006-04-08 01:31:04		
142	207 ad2d 530	2006-04-08 01:31:14	1	http://www.courts.state.ny.us
142	broadway.vera.org	2006-04-08 08:38:23		
142	broadway.vera.org	2006-04-08 08:38:31		
142	vera.org	2006-04-08 08:38:42	1	http://www.vera.org
142	broadway.vera.org	2006-04-08 08:39:30		
142	frankmellace.com	2006-04-09 02:19:24		
142	ucs.ljk.com	2006-04-09 02:20:44		
142	attorneyleslie.com	2006-04-13 00:25:27		
142	merit release appearance	2006-04-22 23:51:18		
142	www.bonsai.wbff.org	2006-05-06 08:49:34		

(b) Công cụ tìm kiếm

Hình 1.1: Ví dụ trực quan về sự khác biệt giữa dữ liệu được tạo ra bởi việc sử dụng Twitter (như trong Hình 1.1a), chứa nhiều tương tác văn bản, dữ liệu hồ sơ và hình ảnh; so với dữ liệu được tạo ra bởi các tương tác với SE (như trong Hình 1.1b), chứa các truy vấn, lượt nhấp và dấu thời gian.

Chúng tôi nghi ngờ rằng sự khác biệt trong dữ liệu có sẵn giữa các miền không cho phép ứng dụng một đối một của các chiến lược đã đề cập trước đó để xác định loại người dùng và đặt câu hỏi về độ tin cậy của phản hồi trực tiếp. Hơn nữa, xem xét sức mạnh những hạn chế liên quan đến việc tiếp cận dữ liệu lịch sử của những người tìm kiếm trẻ tuổi, chúng tôi phải xem xét các chiến lược thay thế liên quan đến việc xác định kỹ năng của họ. Một trong những các giải pháp thay thế là tận dụng thông tin suy ra về người dùng dựa trên tương tác của họ

với một hệ thống [69, 70]. Đối với SE, những tương tác này là ngầm định (chẳng hạn như thời gian phiên bắt đầu) hoặc rõ ràng (chẳng hạn như văn bản từ chính truy vấn); và cùng nhau bao gồm hành vi tìm kiếm của người dùng. Thông thường, những tương tác này được lưu trữ trong truy vấn nhật ký; mỗi mục nhập thường chứa một ID người dùng, một truy vấn và một dấu thời gian (một ví dụ có thể được nhìn thấy trong Hình 1.1b). Các mục nhập sau đó có thể được nhóm vào các phiên<sup>2</sup>, cung cấp một khối văn bản cần được phân tích và cho phép suy ra hành vi tìm kiếm (như số lượng truy vấn trong một phiên hoặc thời gian giữa mỗi truy vấn) không thể ước lượng được từ một truy vấn duy nhất [28].

Thông tin suy luận này có thể được sử dụng bởi các chiến lược lập hồ sơ tác giả [46, 67] hoặc chiến lược nhận dạng chuyên gia [69, 71, 72] để xác định loại người dùng. Cái trước là quá trình phân tích một văn bản với mục đích rõ ràng là xác định các thuộc tính của một tác giả. Các nhà nghiên cứu đã sử dụng các chiến lược lập hồ sơ tác giả để nhóm người dùng theo loại như: nam hay nữ [55], hướng ngoại hay hướng nội [65], và thậm chí là thanh thiếu niên hay người lớn [46]. Các chiến lược này phụ thuộc vào lượng lớn văn bản, chẳng hạn như bài đăng trên Facebook [65] hoặc mục nhập blog [50]. Thật không may, các phiên tìm kiếm hiếm khi đạt đến giới hạn dưới là 100 từ ngữ cần thiết để xác định hiệu quả loại người dùng. Chiến lược xác định chuyên gia tập trung về việc phân tích hành vi tìm kiếm để xác định trình độ chuyên môn của người dùng trong một lĩnh vực cụ thể miền [69, 71, 72]. Chuyên môn có thể được thiết lập bằng cách kiểm tra các tính năng liên quan đến cả để ngầm định hành vi tìm kiếm, chẳng hạn như độ dài phiên và số lượng truy vấn trung bình mỗi phiên, cũng như hành vi tìm kiếm rõ ràng, chẳng hạn như độ dài truy vấn trung bình hoặc sử dụng của các thuật ngữ kỹ thuật. Vì trẻ em gặp khó khăn trong lĩnh vực sử dụng SE [6, 30], phát triển khả năng tìm kiếm có thể cung cấp một cơ hội duy nhất để xác định sự thiếu hụt của họ của chuyên môn. Tuy nhiên, những ví dụ trước đây về nhận dạng chuyên gia được thiết kế

---

<sup>2</sup>Phiên tìm kiếm có thể được định nghĩa là một tập hợp các truy vấn do người dùng tạo ra và được nhóm lại theo mục tiêu tích lũy là đáp ứng nhu cầu thông tin hoặc đạt đến một ngưỡng thời gian nhất định.

để công nhận các chuyên gia trong các lĩnh vực cụ thể, không phải để thiết lập khả năng tìm kiếm hiểu biết. Do những hạn chế đã đề cập ở trên của chiến lược lập hồ sơ tác giả và xác định chuyên gia, chúng tôi nghi ngờ rằng không có loại chiến lược nào có thể được áp dụng trực tiếp để nhận ra các kỹ năng của SE tự động. Tuy nhiên, cả hai sẽ thông báo lựa chọn tính năng của chúng tôi một khi chúng tôi rõ ràng xác định loại người dùng thể hiện sự thiếu hiểu biết về tìm kiếm và cũng hiển thị hành vi văn bản tương quan với các kỹ năng văn học được thể hiện trong các truy vấn của họ.

Một phương pháp phổ biến để nhận ra người dùng thiếu dữ liệu lịch sử được gọi là sự rập khuôn. Phương pháp này tìm cách nhóm người dùng dựa trên các đặc điểm chung rằng tất cả họ đều chia sẻ và được mong đợi ở họ. Để chúng ta có thể xác định một khuôn mẫu có trong thuật ngữ bao trùm “trẻ em” để nhận ra, chúng ta cần hiểu các đặc điểm xác định các nhóm phụ trong dân số này, đặc biệt là các thuộc tính liên quan đến kỹ năng. Sự thành thạo của trẻ em trong một số kỹ năng nhất định có liên quan đến các giai đoạn phát triển, như được diễn đạt bởi các lý thuyết tâm lý học hiện đại. Một số các giai đoạn này được sắp xếp theo chương trình giảng dạy và tương ứng với những gì được dạy ở trường, chẳng hạn như đọc, viết và đánh máy [41]. Những người khác giải quyết vấn đề phát triển nhận thức của bộ não của họ [56], thay đổi cách họ suy nghĩ, cho phép suy nghĩ trừu tượng và tư duy bên lề. Các giai đoạn này có thể phụ thuộc vào nhau. Ví dụ: việc xây dựng từ khóa để tìm kiếm đòi hỏi phải suy nghĩ trừu tượng (giảm tìm kiếm xuống các bộ phận cấu thành), nhưng cũng đòi hỏi khả năng tìm kiếm (biết rằng một truy vấn từ khóa sẽ lấy lại kết quả). Các lý thuyết phát triển bao gồm các giai đoạn khác nhau có bộ kỹ năng được xác định rõ ràng và mô tả trong từng giai đoạn.

Vì khả năng tìm kiếm của trẻ em phụ thuộc vào khả năng tạo truy vấn của chúng, cả về mặt hiểu biết tìm kiếm [28] cũng như phát triển ngôn ngữ [31], trong luận án này chúng tôi đề xuất một chiến lược phân loại được gọi là RYSe (Nhận dạng trẻ) Người tìm kiếm), nhằm mục đích xác định người tìm kiếm trung bình từ 8-12 tuổi dựa trên



hành vi tìm kiếm. Khuôn mẫu này được chọn vì nó bao gồm các giai đoạn phát triển quyết định sự phát triển ngôn ngữ (giai đoạn thứ ba của Piaget: hoạt động cụ thể giai đoạn [56]), cũng như khả năng tìm kiếm thông tin (mức độ đầu tiên của năng lực số [15]). Để làm rõ, chúng tôi tìm cách nhận ra những người dùng thể hiện các đặc điểm được xác định bởi các giai đoạn này nhưng vì các giai đoạn phát triển này phức tạp và nhiều sắc thái nên chúng tôi không tìm cách giải thích rõ ràng gắn nhãn người dùng là thuộc về các giai đoạn này. Do đó, sử dụng khuôn mẫu. Hơn nữa, chúng tôi thiết lập phạm vi để nhận dạng người dùng dựa trên cách họ sử dụng tiếng Anh Mỹ.

Là cấp độ thứ ba và đầu tiên của năng lực số của Piaget, định nghĩa các kỹ năng có mối tương quan để phát triển ngôn ngữ và tìm kiếm khả năng đọc viết, chúng tôi lấy cảm hứng từ hồ sơ tác giả và các chiến lược nhận dạng chuyên gia khi lựa chọn các tính năng nhận dạng những kỹ năng này. Trẻ em trong giai đoạn hoạt động cụ thể thể hiện sự lú loát trong việc viết lách, điều này mang lại họ có khả năng tạo ra các truy vấn [18] nhưng cũng sẽ gặp khó khăn với tư duy trừu tượng. Như vậy, chúng tôi điều tra các tính năng liên quan đến việc tạo truy vấn như từ điển và từ vựng lary, đồng thời cũng điều tra các tính năng liên quan đến sự khác biệt giữa tự nhiên ngôn ngữ và truy vấn từ khóa như cú pháp. Chúng tôi cũng xem xét các đặc điểm duy nhất đến việc trẻ em thiếu chuyên môn trong lĩnh vực sử dụng SE. Mức độ đầu tiên của kỹ thuật số năng lực nêu rằng người dùng có thể, với sự hỗ trợ: xác định nhu cầu thông tin, tìm và truy cập dữ liệu thông qua các tìm kiếm đơn giản và điều hướng giữa các nguồn này của dữ liệu. Bất kỳ khó khăn nào trong việc hoàn thành bất kỳ nhiệm vụ nào trong số này sẽ phản ánh trong cả hai hành vi tìm kiếm ngầm định và rõ ràng, cho phép chúng tôi nhận ra những người dùng đang gặp khó khăn với khả năng tìm kiếm [6]. Do đó, chúng tôi xem xét các toán tử tìm kiếm, việc xây dựng lại truy vấn và các tính năng tạm thời để xác định xem người dùng có ở cấp độ đầu tiên của năng lực số hay không. Tất cả các tính năng này sau đó được sử dụng trong Bộ phân loại Rừng ngẫu nhiên [45]. Vì Bộ phân loại rừng lấy mẫu ngẫu nhiên các đặc điểm cho mỗi cây, nó có thể được điền vào với các cây đặc biệt trong việc nhận dạng người dùng dựa trên các tập hợp con cụ thể

của các tính năng chúng tôi điều tra [48].

Bằng cách xác định hành vi tìm kiếm trung bình của trẻ em từ 8-12 tuổi, chúng tôi cho thấy trẻ em có thể được nhận dạng bằng cách thể hiện các kỹ năng được xác định bởi các giai đoạn phát triển. Thông qua việc công nhận một khuôn mẫu của người tìm kiếm trẻ dựa trên các kỹ năng của họ, cánh cửa mở ra để xác định các khuôn mẫu phi truyền thống khác dựa trên hành vi tìm kiếm của họ như: người tìm kiếm dưới 8 tuổi, thanh thiếu niên trong chung, cũng như người già. RYSe cũng cung cấp SE cho trẻ em và các chiến lược được thiết kế để hỗ trợ những người tìm kiếm trẻ tuổi có khả năng nhận ra kỹ năng của người dùng cần hỗ trợ. Điều này có thể cho phép SE cung cấp các nguồn lực phù hợp với kỹ năng cho những người dùng này. Hơn nữa, kết quả từ luận án này cung cấp cái nhìn sâu sắc về hành vi tìm kiếm của những đứa trẻ.

## 1.1 Luận đề

Chúng tôi đưa ra giả thuyết rằng người tìm kiếm trung bình từ 8-12 tuổi có thể được nhận dạng bằng kỳ thi hành vi tìm kiếm trong suy ra từ tương tác giữa người dùng và công cụ tìm kiếm để hiển thị tìm kiếm sự phát triển về ngôn ngữ và khả năng đọc viết có liên quan đến các giai đoạn phát triển chúng tôi mong đợi khuôn mẫu này sẽ ở trong: giai đoạn thứ ba của Piaget và ở cấp độ đầu tiên của kỹ thuật số năng lực.

## CHƯƠNG 2

### BỐI CẢNH VÀ CÔNG VIỆC LIÊN QUAN

Trong khi xác định người dùng trên 13 tuổi từ văn bản trực tuyến được tìm thấy trong blog và trò chuyện phòng đã là trọng tâm nghiên cứu kể từ giữa những năm 2000 [46, 67, 69], nhận ra người tìm kiếm khuôn mẫu 8-12 tuổi từ các tương tác của họ với SE vẫn là một khả năng chưa được công nhận mà RYSe tìm cách khám phá. Trong phần này, chúng tôi thảo luận về các khái niệm cung cấp thông tin cơ bản và định hướng cho việc thiết kế RYSe.

### 2.1 Nhật ký truy vấn và phiên tìm kiếm

Tương tác tìm kiếm được định nghĩa là tương tác rõ ràng mà người dùng có với SE, chẳng hạn như gửi truy vấn hoặc nhấp vào kết quả. Những tương tác này có thể được lưu trữ trong danh sách chi tiết các mục nhập được gọi là nhật ký truy vấn [23, 39, 61, 69, 70]. Các nhật ký này không chỉ chứa các truy vấn (một chuỗi chứa văn bản đầu vào), nhưng cũng có thể chứa dấu thời gian (phản ánh ngày và giờ), ID người dùng (hoặc ID phiên, phân định một người dùng/phiên từ khác), truy vấn (có chứa văn bản được sử dụng để thực hiện tìm kiếm), URL đã nhấp (nếu có) được chọn, vì không phải tất cả các truy vấn đều mang lại kết quả được nhấp vào) và giá trị thứ tự của URL được nhấp vào (nếu có bất kỳ mục nào được chọn). Các mục sau đó có thể được nhóm vào các phiên tìm kiếm, được định nghĩa là “chuỗi các trang được một người dùng duy nhất truy cập tại một trang web duy nhất trong một khoảng thời gian xác định của thời gian” [58]. Mặc dù có nhiều cách tiếp cận để xác định những gì cấu thành nên một

phiên tìm kiếm [29], chúng tôi tập trung vào các chiến lược dựa trên thời gian vì nghiên cứu chứng minh rằng các chiến lược dựa trên thời gian phù hợp chính xác với các mô hình hoạt động của người dùng [33].

## 2.2 Các lý thuyết phát triển

Người ta suy đoán rằng các tương tác SE của người dùng có thể chỉ ra tình trạng hiện tại của họ giai đoạn giáo dục và phát triển [23, 25, 69]. Có rất nhiều lý thuyết tồn tại với thể hiện mục đích xác định và phân loại các giai đoạn này: 4 giai đoạn của Piaget sự phát triển nhận thức [56], tập trung vào các kỹ năng nhận thức cho phép ngôn ngữ tiếp thu; và 5 giai đoạn biết chữ của Cooper và Kiger [18] cũng như 4 giai đoạn biết chữ của Kroll phát triển viết [41], cả hai đều mô tả các bộ kỹ năng được sử dụng trong việc hiểu và thể hiện ngôn ngữ. Chúng ta thấy 5 giai đoạn của việc biết chữ và 4 giai đoạn của việc viết như bao gồm trong 4 giai đoạn phát triển nhận thức, như sự phát triển nhận thức thông báo sự phát triển ngôn ngữ. 4 giai đoạn phát triển của chữ viết quyết định một khả năng của cá nhân để diễn đạt bản thân bằng ngôn ngữ viết, trong khi 5 giai đoạn của khả năng đọc viết xác định khả năng đọc của họ mô tả các kỹ năng cần thiết để thể hiện ngôn ngữ. Hơn nữa, có 8 cấp độ năng lực số, xác định khả năng tìm kiếm của một người, như được thiết lập bởi Carretero et al. [15]. Năng lực số là duy nhất trong số các giai đoạn đã đề cập ở trên vì các kỹ năng được lý thuyết này chỉ định có thể có thể có được ở bất kỳ độ tuổi nào.

Tất cả các giai đoạn này có thể tác động đến việc sử dụng SE, cho dù đó là trừu tượng hóa để xây dựng từ khóa (được xác định bởi sự phát triển nhận thức), đọc và viết khả năng (được xác định bởi sự phát triển về khả năng đọc viết và viết), hoặc khả năng tìm kiếm khả năng đọc viết (được xác định theo năng lực số). Hơn nữa, các giai đoạn phát triển không chỉ có thể chồng chéo với nhau, nhưng cũng có chương trình giảng dạy phù hợp với những kỹ năng đang phát triển này

bộ. Ví dụ, Sáng kiến Tiêu chuẩn Nhà nước Lỗi Chung (một sáng kiến giáo dục của Mỹ chương trình giảng dạy quy định từ vựng theo trình độ lớp) [38], và Độ tuổi tiếp thu (một lý thuyết tâm lý ngôn ngữ xác định độ tuổi phổ biến mà từ ngữ được học) [35] tương quan với các giai đoạn phát triển khả năng viết và đọc viết.

Các giai đoạn phát triển của Piaget (và sau đó là Kroll, Cooper và Kiger), cũng như chương trình giảng dạy có liên quan, nhấn mạnh cách các truy vấn do người dùng tạo ra có thể cung cấp những manh mối riêng biệt cho các giai đoạn phát triển mà người dùng có thể đang trải qua. Điều tương tự cũng có thể nói về năng lực số, vì các kỹ năng được xác định có thể được nhìn thấy trong màn hình của khả năng tìm kiếm hiểu biết. Tuy nhiên, các giai đoạn trong các lý thuyết đã đề cập trước đó của phát triển được giới hạn mềm. Để xác định xem người dùng có hiển thị tìm kiếm hay không hành vi liên quan đến các giai đoạn phát triển nhất định, một loại người dùng được xác định rõ ràng phải được thiết lập. Điều này cho phép chúng ta xác định những giai đoạn phát triển nào người dùng được mong đợi sẽ thể hiện các đặc điểm khi ở trong.

## 2.3 Xác định loại người dùng

Một phương pháp để giải quyết việc xác định loại người dùng mơ hồ được gọi là sự rập khuôn [16]. Có nguồn gốc từ phạm vi khuyến nghị, cơ sở của điều này cách tiếp cận dựa trên việc nhóm người dùng với những người khác dựa trên các đặc điểm chung tất cả đều chia sẻ. Những đặc điểm này có thể liên quan đến “nhân khẩu học, địa lý hoặc thông tin tâm lý” [16]. Làm như vậy cho phép người dùng được nhận dạng dựa trên đặc điểm chung. Do đó, việc phân khúc người dùng dựa trên khuôn mẫu của họ cho phép việc xác định loại người dùng đó. Vì “trẻ em” là một loại người dùng rộng, nên việc thiết lập một khuôn mẫu có trong phỏ này cho phép chúng ta nêu rõ các kỹ năng được xác định các giai đoạn phát triển mà chúng ta muốn nhận biết.

Lĩnh vực phân tích tác giả, còn được gọi là phong cách học, quan tâm đến việc quy kết quyền tác giả đối với các văn bản có nguồn gốc không rõ. Bằng cách phân tích văn bản do người dùng tạo ra để xác định đặc điểm, có thể xác định các thuộc tính của một tác giả như tuổi tác và giới tính [46, 50, 67] và tính cách [65]. Các nhà nghiên cứu đã đi xa như để xác định những kẻ săn mồi trực tuyến [12, 37, 54] và kẻ bắt nạt [40]. Hồ sơ tác giả hiện có các chiến lược thường dựa vào các bộ phân loại nhị phân được đào tạo trên các đặc điểm văn bản được suy ra từ văn bản do người dùng tạo ra được tìm thấy trong các nguồn như blog, nhật ký trò chuyện và trang web truyền thông xã hội [46, 67, 68]. Tuy nhiên, theo hiểu biết của chúng tôi, không có chiến lược lập hồ sơ tác giả nào trong số này xác định loại người dùng tìm kiếm dựa trên các truy vấn tìm kiếm. Với điều kiện là các phiên tìm kiếm của người dùng được ghi lại trong nhật ký truy vấn chỉ có một phần nhỏ văn bản do người dùng tạo ra các chiến lược đã đề cập ở trên đòi hỏi, chúng tôi đưa ra giả thuyết rằng một ứng dụng một trong những điều này các chiến lược sẽ không hiệu quả trong lĩnh vực SE. Hạn chế này truyền cảm hứng cho chúng tôi xem xét một loạt các đặc điểm văn bản rộng hơn những đặc điểm được tìm thấy trong các phần đã đề cập ở trên ví dụ, đồng thời cũng thúc đẩy việc xem xét các tính năng phi văn bản sẽ được sử dụng bởi RYSe.

Có những chiến lược tận dụng nhiều hơn là chỉ các tính năng văn bản để nhận dạng loại người dùng. Các chiến lược này chứng minh các phương pháp nhận dạng dựa vào về việc bổ sung các tính năng văn bản bằng thông tin độc đáo cho môi trường SE, chẳng hạn như các trang web đã truy cập hoặc thời gian dành cho việc tìm kiếm; để xác định người dùng theo độ tuổi và giới tính [23], chuyên môn về lĩnh vực [69] hoặc kiến thức thu được [70]. Đặc biệt lưu ý là chuyên gia nhận dạng, được thể hiện để xác định các chuyên gia trong nhiều lĩnh vực khác nhau, chúng tôi đưa ra giả thuyết có thể được tận dụng để nhận ra chuyên môn trong các bộ kỹ năng có liên quan trực tiếp đến các giai đoạn phát triển [69]. Trong khi các chiến lược xác định chuyên gia cho thấy rằng nó là có thể nhận ra chuyên môn, hiện tại không có chiến lược nào để xác định chuyên môn trong lĩnh vực tìm kiếm kiến thức, đặc biệt là ở trẻ em.

## CHƯƠNG 3

### PHƯƠNG PHÁP

RYSe nhận dạng người tìm kiếm khuôn mẫu từ 8-12 tuổi thông qua hành vi tìm kiếm của họ bằng cách phân tích các khía cạnh khác nhau của tương tác người dùng-SE. Chúng tôi tập trung vào việc nhận ra các kỹ năng dự kiến sẽ được hiển thị bởi người tìm kiếm trong giai đoạn thứ ba của Piaget (P3), tương quan đến sự phát triển ngôn ngữ, cũng như ở cấp độ đầu tiên của năng lực số, tương quan với khả năng tìm kiếm hiểu biết (DC1). Những kỹ năng này được suy ra từ các tính năng được tìm thấy trong văn bản do người dùng tạo ra (được mô tả trong phần 3.1); cũng như các tương tác tìm kiếm của họ (được mô tả trong phần 3.2).

#### 3.1 Tính năng dựa trên văn bản

Như đã đề cập trước đó, nhiều chiến lược để lập hồ sơ tác giả chứng minh rằng văn bản do người dùng tạo có thể được sử dụng để nhận dạng loại người dùng từ nhật ký trò chuyện, blog và các trang mạng xã hội [46, 67, 68]. Chúng tôi lấy cảm hứng từ các chiến lược này, phân tích văn bản do người dùng tạo ra, tức là truy vấn tìm kiếm, để nhận ra hành vi phổ biến đối với loại người dùng của chúng tôi. Phân tích này tập trung vào bốn nhóm tính năng: Từ vựng, Cú pháp, Lỗi chính tả và từ vựng. Các tính năng này được tính toán trên một truy vấn theo truy vấn cơ sở; đối với các phiên bao gồm nhiều hơn một truy vấn, chúng tôi tính giá trị trung bình cho mỗi tính năng trên mỗi truy vấn trong phiên tương ứng.

### 3.1.1 Đặc điểm từ vựng

Các đặc điểm từ vựng cho phép chúng tôi đánh giá mức độ tinh vi và tính đa dạng của vốn từ vựng của người dùng khi họ diễn đạt bằng tiếng Anh Mỹ viết. Xác định những điều này khả năng đòi hỏi chúng ta phải kiểm tra sự phong phú về từ vựng của văn bản do các cá nhân tạo ra [47].

Sự phong phú về từ vựng bao gồm ba phần chính: mật độ từ vựng, phương tiện từ vựng sự tinh vi và sự thay đổi từ vựng. Mật độ từ vựng kiểm tra tỷ lệ của “danh từ, tính từ, động từ (trừ động từ khiếm khuyết, động từ trợ động, “be” và “have”), và trạng từ có gốc là tính từ, bao gồm cả những trạng từ có thể hoạt động như một tính từ và trạng từ (ví dụ, “nhanh”) và những trạng từ được hình thành bằng cách thêm hậu tố -ly vào tính từ gốc (ví dụ, “đặc biệt)” [47] cho các từ trong một mẫu văn bản. Trong trường hợp của chúng tôi, điều này tỷ lệ từ vựng so với tổng số từ trên mỗi truy vấn cho thấy khả năng diễn đạt bị hạn chế ở một cá nhân vẫn đang học một ngôn ngữ. Sự tinh tế về từ vựng được định nghĩa là “tỷ lệ các từ tương đối bất thường hoặc nâng cao trong văn bản của người học” [52], chứng minh sự quen thuộc của người dùng với một ngôn ngữ. Biến thể từ vựng “đề cập đến phạm vi vốn từ vựng của người học được thể hiện qua cách sử dụng ngôn ngữ của người đó” [47] và không chỉ thể hiện sự nắm bắt ngôn ngữ mà còn giúp làm nổi bật sự khác biệt giữa các truy vấn từ khóa và ngôn ngữ tự nhiên.

Ngoài sự phong phú về từ vựng, chúng tôi cũng xem xét các đặc điểm từ vựng của các truy vấn đã được sử dụng để xác định chuyên môn về lĩnh vực [69] và làm nổi bật thêm sự khác biệt giữa ngôn ngữ tự nhiên và truy vấn từ khóa. Những tính năng này rất quan trọng, vì không phải tất cả các khác biệt về từ vựng giữa hai loại truy vấn đó đều có thể được nắm bắt trong sự phong phú về từ vựng. Một lời giải thích chi tiết về cách tính toán từng từ vựng liên quan Các tính năng có thể được tìm thấy trong Bảng 3.1.



Tên tính năng	Tính năng tính toán	Loại tính năng P3 DC1
Mật độ từ vựng	Nlex/N	Tỉ trọng
Sự tinh tế từ vựng-I	Slex/Nlex	Sự tinh tế
Sự tinh tế từ vựng-II	Kiểu SW/Kiểu N	Sự tinh tế
Động từ tinh tế-I	SVtypes/V	Sự tinh tế
Đã sửa VS1	SVtypes/ $\sqrt{2V}$	Sự tinh tế
Động từ tinh vi-II	SV loại/V	Sự tinh tế
Số lượng các từ khác nhau	T	Đa dạng
Tỷ lệ loại-mã thông báo (TTR)	T/T	Đa dạng
TTR đã sửa	Đ/ (2Đ)	Đa dạng
Gốc TTR	T/ (T)	Đa dạng
Biến thể từ vựng	Tlex/Nlex	Đa dạng
Biến thể động từ-I	Tverb/Nverb	Đa dạng
VV-I vuông	T <sup>2</sup> động từ/Động từ	Đa dạng
Đã sửa VV-I	Tverb/ (Động từ	Đa dạng
Biến thể động từ-II	Tverb/Nlex	Đa dạng
Biến thể danh từ	Danh từ/Nlex	Đa dạng
Biến thể tính từ	Tadj/Nlex	Đa dạng
Tổng số từ	N	Đặc điểm
Số lượng ký tự	Nchar	Đặc điểm
Độ dài từ trung bình	Nchar/Nwords	Đặc điểm
Số lượng âm tiết	Nsyl	Đặc điểm
Trung bình âm tiết mỗi từ	Nsyl/Nwords	Đặc điểm
Từ dễ	Nsim	Đặc điểm
Từ phức tạp	Ncom	Đặc điểm
Số âm tiết tối đa cho mỗi từ	ArgM ax(Nsyl)	Đặc điểm
Số âm tiết tối thiểu cho mỗi từ	ArgM trong(Nsyl)	Đặc điểm

Bảng 3.1: Đặc điểm từ vựng, trong đó từ phức tạp (SVtypes), từ vựng (Slex) và các loại động từ (SVtypes), được định nghĩa là các từ hoặc loại không tìm thấy trong “2.000 những từ thường gặp nhất của British National Corpus” [44], Nsim và Ncom được định nghĩa dưới dạng những từ có ít hơn 3 âm tiết và những từ có 3 âm tiết trở lên.

### 3.1.2 Từ vựng

Từ vựng mà người tìm kiếm sử dụng khi xây dựng truy vấn của họ cung cấp cho chúng ta những hiểu biết sâu sắc vào trạng thái phát triển của người dùng, cả về khả năng tận dụng của họ

ngôn ngữ và truy vấn thủ công. Do đó, chúng tôi kiểm tra từ vựng được tìm thấy trong các truy vấn từ nhiều góc nhìn khác nhau để xác định những tính năng nào cần kiểm tra. Ở mỗi cấp độ

học sinh trình độ được mong đợi biết một số từ vựng nhất định. Điều này dẫn chúng ta đến việc xem xét các tính năng đếm số lần xuất hiện của các thuật ngữ được xác định bởi Tiêu chuẩn chung của tiểu bang

Sáng kiến [38], có thể tìm thấy trong danh sách Từ vựng cốt lõi chung <sup>1</sup>. Những cái này

tiêu chuẩn quy định từ vựng cần được dạy và học ở một cấp lớp cụ thể.

Hơn nữa, người dùng cũng được mong đợi biết một số từ nhất định ở độ tuổi nhất định theo quy định

theo Độ tuổi tiếp thu (AoA) [35], một biến tâm lý ngôn ngữ quyết định

độ tuổi trung bình mà các cá nhân được mong đợi để học một số từ nhất định. Xếp hạng AoA cũng

được sử dụng để xác định mức độ phức tạp của truy vấn, có liên quan đến chuyên môn tìm kiếm

[70]. Vì vậy, chúng tôi xem xét các tính năng sử dụng tập dữ liệu AoA 50K từ 2. Chúng tôi cũng

biết rằng trẻ em có xu hướng sử dụng vốn từ vựng khác với người lớn khi tìm kiếm [49],

do đó chúng tôi đếm số lần xuất hiện của các từ trên mỗi truy vấn được tìm thấy trong các từ phổ biến nhất

được tìm thấy trong các trang web dành cho trẻ em từ "Tìm kiếm bộ phim Seven hoặc Sven từ

Bộ dữ liệu 3 của Movie Frozen?" . Để kiểm tra thêm từ vựng được sử dụng bởi khuôn mẫu của chúng tôi

trong các tìm kiếm của họ, chúng tôi trích xuất 250 từ hàng đầu được tìm thấy trong các truy vấn được tạo ra bởi

người tìm kiếm theo khuôn mẫu 8-12 tuổi, cũng như 250 từ hàng đầu được người dùng sử dụng

không phải là khuôn mẫu của chúng tôi; và tính toán các tính năng liên quan đến hai danh sách này. Ngưỡng

của 250 được thiết lập sau khi xác định vị trí đường cong phân phối của Zipf phẳng trên

danh sách các từ trong truy vấn của khuôn mẫu của chúng tôi. Chúng tôi tiếp tục trích xuất 50 từ hàng đầu

---

<sup>1</sup>Danh sách từ vựng có tại <https://www.flocabulary.com/wordlists/>.

<sup>2</sup>Bạn có thể tìm thấy tập dữ liệu này tại [http://crr.ugent.be/papers/AoA\\_51715\\_words.zip](http://crr.ugent.be/papers/AoA_51715_words.zip).

<sup>3</sup>Bộ dữ liệu này có thể được tìm thấy tại [https://scholarworks.boisestate.edu/cs\\_scripts/5/](https://scholarworks.boisestate.edu/cs_scripts/5/)

bi-grams dành cho những người dùng không phải là khuôn mẫu của chúng tôi; dựa trên cùng một tiền đề và sử dụng Phân phối Zipf để xác định ngưỡng. Mở rộng trên cơ sở kiểm tra từ vựng của các truy vấn cho từ vựng thường thấy trong các truy vấn, chúng tôi thực hiện một cách tiếp cận chi tiết hơn để nhận ra vốn từ vựng được sử dụng và không được sử dụng bởi chúng tôi khuôn mẫu. Điều này liên quan đến việc tính toán các giá trị TF-IDF của các từ được sử dụng bởi tất cả người dùng, từ được sử dụng bởi người tìm kiếm khuôn mẫu 8-12 tuổi của chúng tôi, cũng như các từ được người dùng sử dụng đó không phải là khuôn mẫu của chúng tôi.

Việc sử dụng từ ngữ cụ thể của miền đã được sử dụng để xác định chuyên môn của miền đã chọn các lĩnh vực như y học và khoa học máy tính [69]. Vì vậy, chúng tôi kiểm tra từ vựng chứa các toán tử tìm kiếm và tiền tố/hậu tố url đóng vai trò là chỉ báo của năng lực kỹ thuật số [14]. Bất kỳ người dùng nào sử dụng các tiền tố/hậu tố này trong các truy vấn của họ thể hiện trình độ chuyên môn tìm kiếm vượt xa DC1. Để phân biệt rõ hơn giữa truy vấn ngôn ngữ tự nhiên và truy vấn từ khóa, nhưng cũng làm nổi bật chuyên môn tìm kiếm, chúng tôi sử dụng danh sách từ dừng NLTK [11] để đếm số lần xuất hiện của từ dừng trên mỗi truy vấn. Người dùng thành thạo sử dụng SE sẽ tránh sử dụng các từ dừng. Cuối cùng, chúng tôi kiểm tra các truy vấn để xác định xem chúng có chứa từ nghi vấn hay không vì những từ này là thường thấy trong các truy vấn từ khóa. Một giải thích chi tiết về cách tính toán từng của các tính năng liên quan đến từ vựng có thể được tìm thấy trong Bảng 3.2

### 3.1.3 Chính tả và dấu câu

Người dùng khuôn mẫu 8-12 tuổi mắc lỗi chính tả. Điều này không chỉ do kỹ năng viết và đánh máy hạn chế nhưng ngày càng phát triển của họ [31], nhưng cũng có thể xuất hiện từ cuộc đấu tranh của họ để sử dụng chức năng "bạn đang tìm kiếm" và tự động sửa lỗi của một số SE [34]. Điều này thúc đẩy chúng tôi xem xét các tính năng khám phá các loại khác nhau của lỗi chính tả được tìm thấy trong truy vấn.

Tên tính năng	Công thức P3 DC1
Tỷ lệ các từ tìm thấy trong danh sách từ vựng cốt lõi	Ncore/N
Tỷ lệ các từ không tìm thấy trong danh sách từ vựng cốt lõi	NnotCore/N
lỗi Xếp hạng AoA thấp	AoAmin
nhất Xếp hạng AoA cao	AoAmax
nhất Độ phức tạp của truy vấn (xếp hạng AoA trung bình)	AoAavg
Tỷ lệ từ có xếp hạng AoA dưới 13 Tỷ lệ từ tìm thấy trong tập dữ liệu "Seven or Sven" Số từ tìm thấy trong top 250 từ khuôn	(NAoA < 13,00)/N Nsven/N
mẫu Tỷ lệ số từ tìm thấy trong top 250 từ khuôn mẫu Tỷ lệ số từ không tìm thấy trong top 250 từ khuôn mẫu Số từ tìm thấy trong top 250 từ không khuôn	N250 N250/N 1
mẫu Tỷ lệ số từ tìm thấy trong top 250 từ không khuôn mẫu Tỷ lệ số từ không tìm thấy trong top 250 từ không khuôn mẫu Số từ tìm thấy trong top 50 bi-gram từ	(N250/N) N250ns
khuôn mẫu Tỷ lệ số từ tìm thấy trong top 50 bi-gram từ khuôn mẫu Tỷ lệ số từ không tìm thấy trong top 50 bi-gram từ Số từ tìm thấy trong top 50 bi-gram từ không khuôn mẫu Tỷ lệ số từ tìm thấy trong top 50 bi-gram từ không khuôn mẫu Tỷ lệ số từ tìm thấy trong top 50 bi-gram từ không khuôn mẫu Tỷ lệ số từ không tìm thấy trong top 50 bi-gram từ không khuôn mẫu TF-IDF của truy vấn dựa trên vốn từ vựng của tất cả người dùng TF-IDF của truy vấn dựa trên vốn từ vựng của những người không theo khuôn mẫu Số lượng từ dừng Chứa www Chứa .com Chứa .net Chứa .org Chứa .edu Chứa .gov Chứa http Chứa AND Chứa OR Chứa ""	N250ns /N 1 (N250ns /N) N250bi N50bi/N 1 (N50bi/N) N50bins N50bins /N 1 (N50bins /N) TF IDFall TF IDFs TF IDFns Nstop www Q com Q net Q org Q edu Q gov Q http Q AND Q OR ""
Chứa từ nghi vấn	Q Q inter Q

Bảng 3.2: Các đặc điểm từ vựng, với inter là danh sách các câu hỏi sau đây từ: “ai”, “cái gì”, “khi nào”, “ở đâu”, “tại sao”, “như thế nào”, “là”, “có thể”, “có thể”, “should”, và “would”, và tất cả các biến được xác định bằng cách được chứa trong truy vấn dưới dạng giá trị boolean.

Tính năng	Công thức P3 DC1
Số lỗi chính tả	không có
Số lỗi chính tả được tìm thấy trong tập dữ liệu KidSpell	không có tôi=1 Nsi KidSpellT ypos
Số lỗi chính tả sai một	không có tôi=1 NcNsi j=1 Ncj LD(Nsi, Ncj ) = 1)
Chứa từ viết hoa	Từ U Q
Có chứa dấu chấm câu	Qđiểm

Bảng 3.3: Các tính năng chính tả và dấu câu, trong đó Nc biểu thị danh sách các đề xuất sửa lỗi đánh máy, LD là viết tắt của Levenhstein Distance (được đo trên một ký tự mức độ), và Qpunct được tính toán bằng cách đầu tiên đảm bảo rằng truy vấn không chứa tìm kiếm tiền tố/hậu tố (như được định nghĩa trong Bảng 3.2), sau đó tính toán giá trị boolean dựa trên truy vấn có chứa dấu câu trong danh sách sau không: .!?, .

Chúng tôi khám phá tổng số lỗi đánh máy trên mỗi truy vấn. Chúng tôi cũng so sánh từng lỗi chính tả từ được tìm thấy với danh sách các từ được biên soạn từ tập dữ liệu KidSpell [20] bao gồm lỗi đánh máy do trẻ em thực hiện. Làm như vậy cho phép chúng tôi nhận ra các truy vấn có chứa lỗi đánh máy được tạo ra bởi những người dùng có độ tuổi tương tự như trẻ em từ 8-12 tuổi người tìm kiếm. Cuối cùng, chúng tôi so sánh tất cả các từ viết sai chính tả với danh sách các đề xuất sửa lỗi, kiểm tra xem có bất kỳ lỗi đánh máy nào có khoảng cách Levenshtein<sup>4</sup> bằng một (được gọi là sai lệch một lỗi đánh máy). Loại lỗi đánh máy này bao gồm nhiều lỗi chính tả khác nhau do trẻ em mắc phải [20].

Hơn nữa, vì người tìm kiếm theo khuôn mẫu 8-12 tuổi có xu hướng điều trị truy vấn dưới dạng câu. Chúng tôi nghi ngờ điều này bao gồm các lỗi thường được sửa bằng cách gợi ý truy vấn, sửa lỗi "ý bạn là" và tự động hoàn thành; chẳng hạn như loại bỏ dấu câu và chuyển truy vấn sang chữ thường [34]. Như vậy, chúng ta thấy những điều này như là sự thể hiện của việc thiếu hiểu biết về tìm kiếm. Do đó, chúng tôi kiểm tra các truy vấn cho dấu câu dựa trên câu cũng như các từ viết hoa. Một lời giải thích chi tiết về cách tính toán từng tính năng chính tả và dấu câu có thể được tìm thấy trong Bảng

### 3.3

#### 3.1.4 Cú pháp

Có một số cách mà ngôn ngữ tự nhiên và truy vấn từ khóa có thể khác nhau. Một trong những cách đó là trong cú pháp của chúng. Truy vấn từ khóa có xu hướng là chuỗi danh từ, trong khi các truy vấn ngôn ngữ tự nhiên giống với câu hơn, chứa các mạo từ, cụm từ giới từ; thậm chí cả tính từ và trạng từ. Để điều tra sự khác biệt này chúng ta tính toán tỷ lệ số lượng từ loại so với số lượng từ trong truy vấn cho mỗi thể phần lời nói được hiển thị trong Bảng 3.4. Tuy nhiên, có một số

---

<sup>4</sup>Được xác định là số lượng tối thiểu các chỉnh sửa ký tự đơn (chèn, xóa hoặc thay thế) cần phải thay đổi một từ thành từ khác

hàng trăm hoán vị của các bi-gram phần lời nói, và hàng ngàn cho một phần lời nói tri-gram. Để xác định bi-gram và tri-gram nào để đếm trong truy vấn, trước tiên chúng tôi tính toán 10 bi-gram xảy ra hàng đầu được tìm thấy trong các truy vấn được tạo bởi người tìm kiếm khuôn mẫu 8-12 tuổi, cũng như 5 bộ ba hàng đầu được tìm thấy trong cùng một truy vấn. Chúng tôi xem tất cả các đặc điểm của các phần lời nói này có liên quan đến DC1, vì chúng tôi đang sử dụng những tính năng này để làm sáng tỏ sự khác biệt giữa ngôn ngữ tự nhiên và các truy vấn từ khóa, không làm nổi bật khả năng phát triển ngôn ngữ của người dùng. A giải thích chi tiết về cách tính toán từng tính năng liên quan đến phần lời nói có thể được tìm thấy trong Bảng 3.4

Phân tích cấp độ D là một quá trình thiết lập mức độ phức tạp về mặt phát triển của cú pháp trong câu dựa trên cây cú pháp của câu. Phân tích này được thực hiện bằng cách đầu tiên gắn thẻ các phần của bài phát biểu cho mỗi từ trong câu, sau đó sử dụng ngữ pháp không ngữ cảnh xác suất để phân tích các thẻ đó thành cây cú pháp. Đầu ra của phép phân tích này được máy phân tích Cấp độ D sử dụng để xác định Cấp độ D của một câu. Mỗi câu được đánh giá từ 0-7, trong đó 0 là cú pháp đơn giản và 7 là một câu phức tạp về mặt cú pháp [19]. Các truy vấn ngôn ngữ tự nhiên sẽ có độ chính xác cao hơn D-Levels hơn các truy vấn từ khóa, vì các truy vấn từ khóa về mặt cú pháp đơn giản hơn truy vấn ngôn ngữ tự nhiên hoặc thậm chí có thể chứa các cây cú pháp không thể phân tích cú pháp, cung cấp một sự phân biệt rõ ràng giữa hai loại truy vấn này. Một lời giải thích chi tiết về cách các tính năng này được tính toán có thể được tìm thấy trong Bảng 3.5

## 3.2 Tính năng dựa trên phiên

Các phiên tìm kiếm chứa thông tin ngoài các tính năng văn bản được tìm thấy trên một truy vấn theo truy vấn cơ sở. Vì những tương tác tìm kiếm này có thể cung cấp cái nhìn sâu sắc về khả năng tìm kiếm của người dùng

Tên tính năng	Công thức PoS Loại P3 DC1
Tỷ lệ của các liên từ phổ hợp	Ký tự Ncc/N
Tỷ lệ các chữ số đếm	Ký tự Ncd/N
Tỷ lệ của các yếu tố quyết định	Ký tự Ndt/N
Tỷ lệ của sự tồn tại	Nex/N Đơn vị
Tỷ lệ từ nước ngoài	Đơn vị Nfw/N
Tỷ lệ của liên từ giới từ/liên từ phụ thuộc	Ký tự Nin/N
Tỷ lệ tính từ	Chữ cái Njj/N
Tỷ lệ tính từ so sánh	Njjr/N Đơn vị
Tỷ lệ tính từ so sánh nhất	Njjs/N Unigram
Tỷ lệ đánh dấu danh sách	Đơn vị Nls/N
Tỷ lệ các phương thức	Đơn vị Nmd/N
Tỷ lệ danh từ	Ký tự Nnn/N
Tỷ lệ danh từ số nhiều	Đơn vị Nnnp/N
Tỷ lệ danh từ riêng	Đơn vị Nnnps/N
Tỷ lệ danh từ riêng số nhiều	Ký tự Nnns/N
Tỷ lệ của các yếu tố quyết định trước	Đơn vị Npdt/N
Tỷ lệ kết thúc sở hữu	Ký tự Npos/N
Tỷ lệ đại từ nhân xưng	Đơn vị Nprp/N
Tỷ lệ đại từ sở hữu	Nrb/N Đơn vị
Tỷ lệ của trạng từ	Số hiệu Nrbr/N
Tỷ lệ của trạng từ so sánh	Nrbs/N Unigram
Tỷ lệ của trạng từ so sánh nhất	Nrp/N Đơn vị
Tỷ lệ các hạt	Nsym/N Đơn vị
Tỷ lệ của	Ký tự Nto/N
Tỷ lệ của các câu cảm thán	Nuh/N Unigram
Tỷ lệ của động từ cơ sở	Nvb/N Đơn vị
Tỷ lệ động từ thì quá khứ	Nvbd/N Đơn vị
Tỷ lệ của động từ phân từ hiện tại	Nvbg/N Đơn vị
Tỷ lệ của động từ phân từ quá khứ	Ký tự Nvbn/N
Tỷ lệ của động từ hiện tại số ít	Đơn vị Nvbp/N
Tỷ lệ động từ ngôi thứ ba số ít	Nvzb/N Đơn vị
Tỷ lệ của các yếu tố quyết định	Unigram Nwdt/N
Tỷ lệ đại từ	Nwp/N Đơn vị
Tỷ lệ của các động từ wh	Nwrb/N Đơn vị
Tỷ lệ của cụm danh từ	Không có không có Bigram
Tỷ lệ của cụm danh từ tính từ so sánh	Njj pn/N Bigram
Tỷ lệ của cụm danh từ số nhiều	Nnnnn/N Bigram
Tỷ lệ của cụm động từ cơ sở	Nto yb/N Bigram
Tỷ lệ của cụm danh từ số nhiều tính từ so sánh	Njj- nns/N Bigram
Tỷ lệ tính từ so sánh với cụm từ	Njj đến/N Bigram
Tỷ lệ cụm danh từ giới từ	Nnn trong/N Bigram
Tỷ lệ cụm danh từ giới từ số nhiều	Nnns trong/N Bigram
Tỷ lệ cụm danh từ giới từ	Nin pn/N Bigram
Tỷ lệ của cụm danh từ xác định	Ndt pn/N Bigram
Tỷ lệ của cụm danh từ tính từ so sánh	Njj pn nn/N Trigram
Tỷ lệ của danh từ danh từ cụm danh từ	Nnn pn nq/N Bát quái
Tỷ lệ của cụm từ tính từ so sánh với động từ	Njj thành vb/N Bát quái
Tỷ lệ của danh từ danh từ số nhiều cụm danh từ	Nnn pn nqs/N Bát quái
Tỷ lệ của cụm danh từ động từ	Nto yb pn/N Trigram

Bảng 3.4: Các đặc điểm của phần lời nói được định nghĩa bởi phần lời nói NLTK

người gắn thẻ.

Tên tính năng	Công thức P3 DC1
Số câu D-Level 1 trong truy vấn	Nd11
Số câu D-Level 2 trong truy vấn	Nd12
Số câu D-Level 3 trong truy vấn	Nd13
Số câu D-Level 4 trong truy vấn	Nd14
Số câu D-Level 5 trong truy vấn	Nd15
Số câu D-Level 6 trong truy vấn	Nd16
Số câu D-Level 7 trong truy vấn	Nd17i
Mức D trung bình của truy vấn	$\frac{7}{=1} \frac{N(d1i)}{\text{Đã gửi}}$

Bảng 3.5: Các tính năng cấp D, được tính toán bằng trình phân tích cấp D  
có thể được tìm thấy tại [5].

[31], chúng tôi lấy cảm hứng từ các ví dụ về nhận dạng chuyên gia kiểm tra phiên  
thông tin và nhận ra các chuyên gia trong lĩnh vực như Y học và Máy tính  
Khoa học [69, 70] để định hình cách tiếp cận của chúng tôi trong việc nhận ra chuyên môn trong lĩnh vực  
của khả năng tìm kiếm. Các tính năng này được tính toán ở cấp độ phiên. Tuy nhiên, chúng tôi  
tính toán khoảng cách Levenhstein cho các tính năng ở cấp độ truy vấn và cho tất cả các phiên dài hơn  
giá trị này được tính trung bình trong suốt phiên.

3.2.1 Tương tác dựa trên truy vấn

Có một số đặc điểm chung của người tìm kiếm khuôn mẫu từ 8-12 tuổi  
tách biệt các tương tác dựa trên truy vấn của họ khỏi những người dùng khác. Việc thiếu kiến thức của họ  
trong cách thức hoạt động của công cụ tìm kiếm có thể khiến chúng lặp lại cùng một truy vấn nhiều lần  
với hy vọng có được kết quả khác nhau, hoặc thậm chí nhấn nút tìm kiếm nhiều lần [9].  
thermore, chúng tôi biết rằng các chuyên gia tìm kiếm thể hiện sự tinh tế trong việc tạo truy vấn của họ,  
thường thêm hoặc xóa một từ duy nhất để làm rõ kết quả của họ [70], trong khi không  
các chuyên gia có thể gặp khó khăn với việc tái cấu trúc truy vấn. Chúng tôi cũng thấy người dùng thiếu kinh nghiệm có thể  
gõ chậm hơn, nhanh hơn và mất nhiều thời gian hơn để tạo truy vấn [60], trong khi người dùng có kinh nghiệm có thể  
dành ít thời gian hơn để xem xét kết quả và thậm chí xây dựng lại truy vấn nhiều lần trước khi  
nhấp vào một liên kết. Được thông báo bởi kiến thức này, chúng tôi điều tra các tính năng được tìm thấy trong



Bảng 3.6.

Tên tính năng	Công thức P3 DC1
Số lượng truy vấn trong một phiên	<small>Số lượng</small>
Số lượng truy vấn duy nhất	$N_{set\_q}$
Nếu tất cả các truy vấn đều là cùng một truy vấn	$N_{set\_q} = 1$
Số lượng truy vấn lặp lại	$N_{repQ}$
Tỷ lệ truy vấn so với nhấp chuột	<small>Số lượng/Số lượng</small>
Khoảng cách Levenhstein giữa các truy vấn	$\frac{\sum_{i=1}^{N_q} \text{Khoảng cách Levenhstein}(Q_i, Q_{i+1})}{N_q}$
Thời gian giữa các truy vấn	$\frac{\sum_{i=1}^{N_q} \text{TimeStamp}(Q_{i+1}) - \text{TimeStamp}(Q_i)}{N_q}$

Bảng 3.6: Các tính năng tương tác dựa trên truy vấn trong đó  $Set_q$  là tập hợp tất cả các truy vấn,  $Q_i$  là truy vấn thứ  $i$  trong một phiên và  $\text{TimeStamp}(Q_i)$  là dấu thời gian cho truy vấn thứ  $i$ . Khoảng cách Levenhstein giữa các truy vấn được đo theo cấp độ ký tự.

3.2.2 Tương tác dựa trên nhấp chuột

Việc thiếu chuyên môn tìm kiếm cũng có thể biểu hiện ở cách người tìm kiếm tương tác với công cụ tìm kiếm kết quả được SE thu thập để trả lời các truy vấn của họ. trẻ em từ 8-12 tuổi theo khuôn mẫu người tìm kiếm được biết là nhấp nhiều lần vào cùng một kết quả khi trang không xuất hiện ngay lập tức, có xu hướng thích nhấp vào kết quả đầu tiên xuất hiện và truy cập lại các trang web mà họ đã nhấp vào [9]. Vì vậy, chúng tôi điều tra các tính năng dựa trên cú nhấp chuột được tìm thấy trong Bảng 3.7.

3.3 Phân loại

Chúng tôi đồng thời xem xét từng tính năng được mô tả trong Phần 3.1 và Phần

3.2.1 khi phân tích phiên tìm kiếm. Các tính năng này đóng vai trò là bằng chứng trong việc xác định người dùng có thể hiện các kỹ năng được mong đợi từ trẻ em 8-12 tuổi hay không người tìm kiếm cũ trong hành vi tìm kiếm của họ là người tìm kiếm khuôn mẫu từ 8-12 tuổi. Chúng tôi

Tên tính năng	Công thức P3 DC1
Số lần nhấp chuột	$N_c$
Số lần nhấp chuột duy nhất	Tập $c_{\_}$
Nếu tất cả các nhấp chuột đều là cùng một trang web	Tập $c_{\_} = 1$
Số lần nhấp lặp lại	$N_{repC}$
Vị trí nhấp chuột trung bình	$\frac{N_c \sum_{i=1}^{N_c} ClickVị trí (C_i)}{N_c}$
Thời gian giữa các lần nhấp	$\frac{\sum_{i=1}^{N_c} T imeStamp(C_{i+1}) - T imeStamp(C_i)}{N_c - 1}$

Bảng 3.7: Các tính năng nhấp chuột trong đó  $C_i$  là lần nhấp thứ  $i$  trong phiên và  $T imeStampC_i$  là dấu thời gian của lần nhấp thứ  $i$  trong phiên.

coi nhiệm vụ nhận dạng khuôn mẫu này là một vấn đề phân loại và do đó sử dụng các tính năng này làm đầu vào cho Bộ phân loại Rừng ngẫu nhiên [45]. Chúng tôi đã chọn điều này phân loại vì nhiều lý do. Lý do đầu tiên là cách mà phân loại này thực hiện lựa chọn tính năng của nó. Như đã thấy trong [48], bộ phân loại rừng ngẫu nhiên được ghi chú vì khả năng tiềm tàng của nó trong việc xây dựng các cây tương quan với các tập hợp tính năng cụ thể. Tính năng lấy mẫu phụ này được coi là một cách để xây dựng các cây chuyên biệt có thể nhận dạng người dùng dựa trên các tập hợp con tính năng này. Hơn nữa, khoảng cách được biểu diễn bởi bộ tính năng không tuyến tính. Về một đường thẳng rõ ràng ở giữa những con số này là một cách tiếp cận kém khi xem xét đến sự khác biệt tiềm ẩn của kỹ năng người dùng trong khuôn mẫu Người tìm kiếm từ 8-12 tuổi. Do đó, chúng tôi đã chọn một bộ phân loại phi tuyến tính.

## CHƯƠNG 4

### THIẾT LẬP THỬ NGHIỆM

Trong chương này chúng tôi mô tả các thành phần chính cần thiết để thực hiện thí nghiệm cho phép chúng ta xác định hiệu quả của RYSe.

#### 4.1 Bộ dữ liệu

Do lo ngại về quyền riêng tư và bảo vệ trẻ em [26], không có các phiên tìm kiếm công khai được tạo ra bởi người tìm kiếm theo khuôn mẫu từ 8-12 tuổi. Với suy nghĩ này, chúng tôi tận dụng các nguồn dữ liệu hiện có (chi tiết trong đó là được tóm tắt trong Bảng 4.1) để xây dựng các tập dữ liệu cho phép phát triển và đánh giá RYSe: Phiên có nhấp chuột (như mô tả trong Phần 4.1.1) và Đơn Phiên truy vấn (như mô tả trong Phần 4.1.2).

##### 4.1.1 Bộ dữ liệu Phiên có nhấp chuột

Chúng tôi sử dụng hai nguồn dữ liệu để xây dựng Phiên với Nhấp chuột. Đầu tiên là TREC

Nhật ký truy vấn theo dõi phiên từ năm 2011-2014 [2], chứa các phiên tìm kiếm được tạo ra bởi người lớn. Thứ hai là nhật ký truy vấn AOL [36], chưa được dán nhãn, tức là, các phiên tìm kiếm thuộc về các phân khúc người dùng khác nhau chưa được xác định. Để dán nhãn nguồn dữ liệu này để chúng ta có thể thiết lập sự thật cơ bản cần thiết cho nhận ra người tìm kiếm khuôn mẫu 8-12 tuổi, chúng tôi dựa vào một số quy tắc. Chúng tôi

Tên	Nguồn	Sự miêu tả
Nhật ký truy vấn AOL	[53]	36 triệu truy vấn. Mỗi mục nhập chứa một dấu thời gian, ID người dùng, truy vấn, URL đã nhấp, và vị trí của URL được nhấp vào.
Các truy vấn từ “Looking cho Phim Bầy hoặc Sven từ bộ phim Đông lạnh? Một Đa-quan điểm Chiến lược cho Đề xuất truy vấn dành cho trẻ em”	[49]	602 truy vấn truy vấn riêng lẻ. Không có phiên dữ liệu. 301 truy vấn được tạo ra bởi người dùng từ lớp K đến lớp 9 (tuổi từ 6 đến 13), 301 được tạo ra bởi người lớn.
Phiên hợp TREC 2011-2014 Theo dõi	[2]	Khoảng 1800 phiên tìm kiếm có độ dài được xác định trước với số lượng khác nhau dữ liệu phiên do người lớn tạo ra.

Bảng 4.1: Nguồn dữ liệu được sử dụng để tạo tập dữ liệu.

bắt đầu bằng cách xác định các phiên tìm kiếm của người dùng với ngưỡng lên đến một giờ [33]. Chúng tôi sau đó kiểm tra mọi phiên có chứa cú nhấp chuột vào trang web được chỉ định là “dành cho người dùng trong độ tuổi từ 8-12” [23]1 . Nếu một phiên chỉ truy xuất các trang web được thiết kế cho 8-12 tuổi, sau đó nó được dán nhãn là một phiên được tạo ra bởi khuôn mẫu 8-12 tuổi người tìm kiếm cũ. Nếu một phiên không chỉ truy xuất các trang web được thiết kế cho những người dùng này, sau đó chúng tôi xem xét thời lượng của phiên đó. Thời lượng phiên trung bình cho trẻ em là khoảng 3,75 phút, trong khi các buổi học của người lớn thường là 8,35 phút [24]. Chúng tôi xem xét một phiên tìm kiếm có chứa một cú nhấp chuột vào một trang web được thiết kế cho trẻ em và thời lượng phiên gần 3,75 phút hơn là 8,35 phút như chỉ định của phiên tìm kiếm được tạo ra bởi người tìm kiếm theo khuôn mẫu từ 8-12 tuổi.

Để xây dựng Phiên với Nhấp chuột, chúng tôi sử dụng tất cả các phiên TREC với ít nhất một cú nhấp chuột, sau đó lấy mẫu các bản ghi truy vấn AOL được gắn nhãn duy trì 80/20 (không tỷ lệ khuôn mẫu/khuôn mẫu), vì điều này phù hợp với sự phân bố của người dùng SE [1].

1Danh mục này được xác định bằng cách sử dụng danh sách các trang web được phân loại theo nội dung.

	Khuôn mẫu Phiên hợp Với Nhấp chuột	Không- Khuôn mẫu Phiên hợp Với Nhấp chuột	Khuôn mẫu Truy vấn đơn Phiên hợp	Không- Khuôn mẫu Truy vấn đơn Phiên hợp
# Độc nhất Phiên hợp	7.980	31.920	301	1.204
Trung bình # Truy vấn Mỗi Phiên hợp	1,87	2,95	1	1
Trung bình # Nhấp chuột mỗi Phiên hợp	2.31	2,86	-	-
Trung bình # Từ Mỗi truy vấn	2,36	2,33	-	-
Thời lượng trung bình của phiên (phút)	.95	12.15	-	-
% Độc nhất Truy vấn	41,1%	64,8%	100%	100%

Bảng 4.2: Mô tả các tập dữ liệu được sử dụng trong các thí nghiệm của chúng tôi.

nhật, chúng tôi chỉ xem xét các phiên nhật ký truy vấn AOL có chứa ít nhất một cú nhấp chuột, vì các phiên thuộc về người tìm kiếm theo khuôn mẫu từ 8-12 tuổi được gắn nhãn theo cách này hành vi và không làm như vậy sẽ làm lệch dữ liệu. Với điều kiện là chúng ta sử dụng một số quy tắc cho nhãn phiên, cũng như 2 nguồn dữ liệu khác nhau, Phiên có nhấp chuột kết quả nắm bắt nhiều hành vi tìm kiếm khác nhau thể hiện khuôn mẫu của chúng tôi, đồng thời cũng rút ra trong các phiên hợp mà chúng tôi có thể chắc chắn rằng không phải do nhóm mục tiêu của chúng tôi tạo ra.

#### 4.1.2 Phiên truy vấn đơn

Chúng tôi cũng tạo một tập dữ liệu các truy vấn đơn lẻ không cần nhấp chuột nhằm mục đích mô phỏng

sự khởi đầu của một phiên tìm kiếm. Để xây dựng tập dữ liệu này, trước tiên chúng tôi trích xuất tất cả các truy vấn

được tạo ra bởi người dùng giữa các lớp K-8 (tuổi từ 6-13) từ dữ liệu đào tạo

được tìm thấy trong “Bạn đang tìm kiếm bộ phim Seven hay Sven từ bộ phim Frozen?”

Nguồn dữ liệu “Chiến lược đa quan điểm để đề xuất các truy vấn cho trẻ em” (như

được mô tả trong Bảng 4.1), gắn nhãn những người dùng này là người tìm kiếm theo khuôn mẫu từ 8-12 tuổi

vì đây là những câu hỏi được hình thành bởi khuôn mẫu của trẻ em từ 8-12 tuổi. Sau đó, chúng tôi lấy mẫu

các truy vấn đơn lẻ được rút ra từ nhật ký truy vấn TREC Session Track từ năm 2011-2014 [2],

vì chúng tôi có thể chắc chắn rằng những truy vấn này được tạo ra bởi những người dùng không phải là của chúng tôi

khuôn mẫu. Lưu ý rằng khi xây dựng Phiên truy vấn đơn, chúng tôi duy trì nguyên tắc 80/20

(tỷ lệ không theo khuôn mẫu/khuôn mẫu) được đề cập trong Mục 4.1.1, để chính xác

đại diện cho các phân phối sử dụng SE điển hình. Bộ dữ liệu này cho phép chúng tôi đánh giá mức độ tốt

RYSe có thể nhận ra khuôn mẫu của chúng ta ngay từ đầu phiên làm việc.

## 4.2 Đường cơ sở

Để ngữ cảnh hóa hiệu quả của chiến lược của chúng tôi, chúng tôi so sánh và đối chiếu

hiệu suất của RYSe với hiệu suất của các đường cơ sở phù hợp. Vì hiện tại không có

tồn tại một phương pháp đã được thiết lập để xác định người tìm kiếm khuôn mẫu từ 8-12 tuổi

dựa trên các tương tác SE của họ, chúng ta phải điều chỉnh các phương pháp nhận dạng người dùng có

đã được sử dụng để nhận dạng hoặc gắn nhãn người dùng tương tự như người tìm kiếm khuôn mẫu từ 8-12 tuổi

trong các lĩnh vực khác nhau, chẳng hạn như phòng trò chuyện và các trang web truyền thông xã hội. Chúng tôi mô tả những

đường cơ sở bên dưới.

Phân loại đa số Chúng tôi bắt đầu với Phân loại đa số, một đường cơ sở ngây thơ

gắn nhãn mọi phiên là dành cho lớp người dùng đa số. Chiến lược này được cho là mô phỏng mô hình của SE lớn giả định người dùng không có hồ sơ người dùng hoặc chưa đăng nhập để trở thành người dùng truyền thống.

Bộ phân loại dựa trên quy tắc Đường cơ sở này thiết lập một tập hợp các quy tắc được sử dụng để

phân biệt một loại người dùng với loại người dùng khác. Chúng tôi áp dụng phương pháp nhận dạng người dùng được tìm thấy trong [23], trong đó gắn nhãn các phiên là thuộc về các loại người dùng khác nhau dựa trên trên các trang web họ nhấp vào, vào một bộ phân loại dựa trên quy tắc. Chúng tôi làm như vậy bằng cách xác định tất cả các phiên có thời lượng dưới nửa giờ và gắn thẻ tất cả các phiên đó tạo ra một cú nhấp chuột vào trang web được chỉ định cho người dùng trong độ tuổi từ 8-12 thuộc về loại người dùng của chúng tôi. Bộ phân loại này đóng vai trò là đường cơ sở phù hợp vì nó sử dụng một chiến lược là một bước tiến từ Phân loại đa số, nhưng không tận dụng các phương pháp nhận dạng tinh vi hơn như máy học.

Bộ phân loại dựa trên văn bản Là một ví dụ về phân loại học máy, chúng tôi

xem xét phương pháp nhận dạng người dùng được đề xuất bởi Tam và Martell [67], rằng xác định độ tuổi của người dùng phòng chat dựa trên nhật ký trò chuyện. Người dùng được phân loại như là thanh thiếu niên (13-19), ở độ tuổi 20 (20-29), 30 (30-39), 40 (40-49), 50 (50-59), hoặc được nhóm lại với nhau như Người lớn (20-59). Tam và Martell [67] xác định rằng một Máy vectơ hỗ trợ (SVM) được đào tạo trên tri-gram gram là bộ phân loại hiệu quả nhất trong việc nhận dạng thanh thiếu niên. Do đó, chúng tôi áp dụng người dùng này phương pháp nhận dạng bằng cách sử dụng túi tri-gram của mô hình SVM được đào tạo trên văn bản của mỗi truy vấn của phiên được nối lại như đầu vào. Bộ phân loại này phục vụ như một đường cơ sở phù hợp vì nó đại diện cho một trong những nỗ lực đầu tiên để xác định trực tuyến tuổi của người dùng dựa trên văn bản họ tạo ra bằng máy học.

Bộ phân loại đa tính năng Bộ phân loại này tận dụng nhiều hơn là chỉ từ ngữ và ký tự

ter token, vì nó cũng sử dụng các phần của bài phát biểu và các tính năng nội dung để nhận dạng người dùng theo độ tuổi. Một phương pháp nhận dạng người dùng sử dụng loại phân loại này là được trình bày bởi Santosh et al. [62], những người nhận ra người dùng trong các độ tuổi sau dựa trên các blog được viết bằng tiếng Anh và tiếng Tây Ban Nha: (10 tuổi: 13-17, 20 tuổi: 23-27 và 30 tuổi: 33-47). Bằng cách sử dụng cây quyết định của bộ phân loại, SVM được đào tạo trên túi từ và các phần của bài phát biểu n-gram cũng như các bộ phân loại hồi quy logistic được đào tạo trên Mô hình chủ đề phân bố Dirichlet tiềm ẩn; phương pháp nhận dạng người dùng này có thể để xác định độ tuổi của người dùng. Chúng tôi áp dụng phương pháp này bằng cách sử dụng các mô hình tương tự được đào tạo trên cùng các tính năng, sử dụng tất cả các truy vấn trong một phiên được nối lại cùng nhau như là đầu vào. Chúng tôi đã chọn điều này làm cơ sở vì nó biểu thị một bước tiến tới việc chỉ sử dụng n-gram từ và ký tự.

Bộ phân loại đa mô hình Bộ phân loại này khác với Bộ phân loại đa tính năng ở

chuyển trọng tâm từ dựa trên tính năng sang dựa trên mô hình. Thay vì sử dụng nhiều tính năng khác nhau trên các mô hình tương tự, bộ phân loại này thay vào đó sử dụng một nhiều mô hình có các tính năng tương tự. Nemati [51] cung cấp cho chúng tôi một người dùng phương pháp phát hiện sử dụng loại phân loại này, nhận dạng độ tuổi của người dùng (15-19, 20-24, 25+) và giới tính dựa trên một số tài liệu được thu thập từ nhiều trang web truyền thông xã hội (Twitter, Facebook và các trang khác). Nemati's mô hình tổng hợp bao gồm 4 bộ phân loại duy nhất được đào tạo trên người dùng tạo ra văn bản: Hồi quy logistic, Bayes ngây thơ, Mạng Perceptron đa lớp và Tăng cường Gradient. Chúng tôi áp dụng cách tiếp cận này bằng cách sử dụng các mô hình tương tự trên các tính năng tương tự, nhưng thay vì đào tạo trên các tài liệu truyền thông xã hội, chúng tôi sử dụng các truy vấn được nối tiếp theo phiên làm văn bản đầu vào. Trong khi các độ tuổi Nemati [51] nhận ra là hơi cũ hơn loại người dùng của chúng tôi, đó là ví dụ duy nhất chúng tôi tìm thấy



sử dụng mạng nơ-ron để nhận dạng người dùng dưới 18 tuổi từ tài khoản trực tuyến của họ tạo ra văn bản và sử dụng nó nhờ vào sự khác biệt này.

### 4.3 Số liệu

Để định lượng hiệu suất chiến lược của chúng tôi, cũng như so sánh và đối chiếu với

Đối với hiệu suất của các phương pháp tiếp cận cơ bản khác, chúng tôi sử dụng một số số liệu.

Đầu tiên là Độ chính xác (được tính như trong Phương trình 4.1), vì đây là số liệu cần thiết

để đánh giá hiệu suất của bộ phân loại nhị phân nói chung [57]. Tuy nhiên, Độ chính xác

có thể trình bày sai kết quả vì "nó cung cấp ước tính quá lạc quan về bộ phân loại

khả năng trên lớp đa số" [17]. Do đó, chúng tôi xem xét Tỷ lệ dương tính thực (TPR)

(như được thấy trong Phương trình 4.2) và Tỷ lệ âm tính thực sự (TNR) (như được thấy trong Phương trình 4.3).

TPR cho phép chúng ta đánh giá khả năng của một chiến lược trong việc nhận ra khuôn mẫu 8-12 tuổi

người tìm kiếm cũ tương tác với SE, trong khi TNR cho phép chúng tôi đánh giá ngược lại; làm thế nào

chiến lược hiệu quả là nhận ra đúng người dùng không phải là khuôn mẫu của chúng ta. Cả hai

rất quan trọng vì SE muốn cung cấp hỗ trợ và nguồn lực phù hợp cho người dùng

những người thuộc nhóm mục tiêu của chúng tôi, đồng thời đảm bảo rằng những người dùng không thuộc nhóm đó

phần khúc cũng được cung cấp các dịch vụ phù hợp với kỹ năng của họ. Bằng cách sử dụng TPR,

TNR và Độ chính xác, chúng ta được cấp một sự hiểu biết đa chiều về hiệu quả

RYSe cũng như các đường cơ sở của chúng tôi.

$$\text{Độ chính xác} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

trong đó TP (True Positive) là số lượng người dùng được xác định chính xác là

người tìm kiếm khuôn mẫu 8-12 tuổi, TN (True Negative) là số lượng người dùng

được xác định chính xác là không phải là người tìm kiếm khuôn mẫu 8-12 tuổi, FP (Sai

Tích cực) là số lượng người dùng được xác định là có độ tuổi từ 8-12 người tìm kiếm năm tuổi, nhưng thực tế không phải vậy, và FN (False Negative) là số người dùng được xác định không phải là người tìm kiếm theo khuôn mẫu từ 8-12 tuổi, nhưng thực tế là.

$$TPR = \frac{TP}{TP + FN} \quad (4.2)$$

trong đó TP và FN được định nghĩa như trong Phương trình 4.1.

$$TNR = \frac{TN}{TN + FP} \quad (4.3)$$

trong đó TN và FP được định nghĩa như trong Phương trình 4.1.

## 4.4 Chuẩn bị thí nghiệm

Để thực hiện các thí nghiệm của mình, trước tiên chúng ta phải chuẩn bị dữ liệu để điều chỉnh, đào tạo và thử nghiệm.

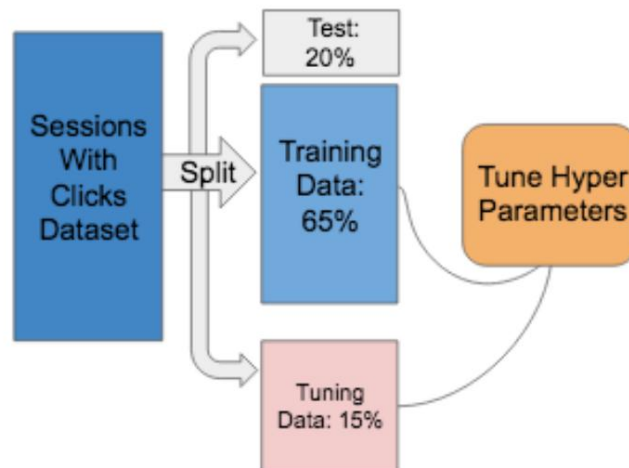
### 4.4.1 Điều chỉnh siêu tham số

Trước khi thực hiện các thí nghiệm của chúng tôi, chúng tôi điều chỉnh các siêu tham số của mô hình của chúng tôi. Điều này quá trình đầu tiên bao gồm việc chia tập dữ liệu Phiên với Nhấp chuột của chúng tôi thành một chuỗi rời rạc tập con điều chỉnh/điều chỉnh/kiểm tra tỷ lệ 65/15/20 (như được thấy trong Hình 4.1). Sau đó, chúng tôi thực hiện tìm kiếm lưới sử dụng các phân tách đào tạo và điều chỉnh từ tập hợp con này trên các phần sau các tham số, với dữ liệu điều chỉnh bị loại bỏ sau đó:

- Số lượng người ước tính
- Trọng lượng lớp (không có hoặc cân bằng)

- Tiêu chí (tập chất Gini hoặc lợi ích thông tin)
- Mẫu bootstrap (sử dụng mẫu bootstrap hoặc toàn bộ dữ liệu để xây dựng cây)

Chúng tôi chọn tập hợp các siêu tham số mang lại Độ chính xác và TPR cao nhất trên phần chia tách điều chỉnh. Kết quả điều chỉnh của chúng tôi là một Bộ phân loại Rừng ngẫu nhiên với 450 người ước tính, sử dụng một lớp trọng số không cân bằng, thông tin đạt được như là tiêu chí của chúng tôi, với không có mẫu bootstrap.



Hình 4.1: Hình làm nổi bật cách chúng tôi tạo tập hợp con đào tạo/điều chỉnh/kiểm tra rời rạc để điều chỉnh siêu tham số từ Phiên có nhấp chuột.

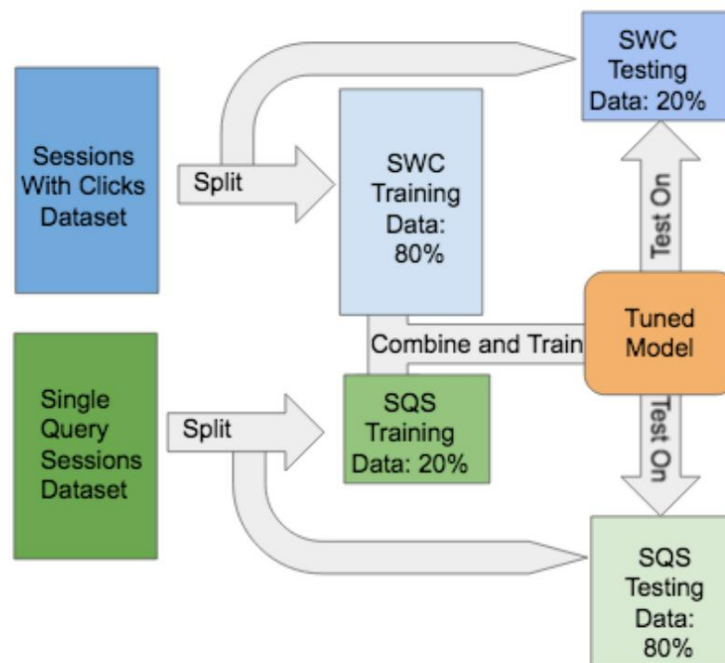
#### 4.4.2 Phân chia dữ liệu của chúng tôi để thử nghiệm và đào tạo

Tất cả các thí nghiệm được trình bày chi tiết trong Chương 5 sẽ yêu cầu chúng tôi đào tạo và kiểm tra cả RYSe

cũng như các đường cơ sở của chúng tôi. Việc tạo ra các phân chia thử nghiệm và đào tạo của chúng tôi đòi hỏi một

mô tả cho bản chất độc đáo của dữ liệu của chúng tôi. Phiên với nhấp chuột được chia 80/20

(đào tạo/kiểm tra) và Phiên truy vấn đơn được chia 20/80 (đào tạo/kiểm tra). RYSe và các Đường cơ sở sau đó đào tạo trên dữ liệu đào tạo từ cả hai tập dữ liệu, sau đó kiểm tra riêng biệt trên cả hai bộ dữ liệu thử nghiệm. Lý do đằng sau việc kết hợp đào tạo dữ liệu là do bản chất của các phiên mà Phiên truy vấn đơn mô phỏng. Những các phiên khác biệt đáng kể so với các phiên được tìm thấy trong Phiên có nhấp chuột, vì vậy RYSe và các đường cơ sở tương ứng (như được thấy trong Phần 4.2) yêu cầu một số mẫu đào tạo Phiên truy vấn đơn để kiểm tra đúng cách trên tập dữ liệu này. Quá trình này chúng tôi có vừa được mô tả có thể được thấy rõ hơn trong Hình 4.2.



Hình 4.2: Hình làm nổi bật cách chúng tôi tạo phân chia dữ liệu đào tạo/kiểm tra, chứa thông tin về cách kết hợp các phân chia đào tạo và những phần nào của mỗi tập dữ liệu được thử nghiệm.

#### 4.5 Xác thực kết quả của chúng tôi

Tất cả các kết quả thử nghiệm được báo cáo trong Chương 5 đều là kết quả của phép lai 5 lần xác thực qua các lần chia tách thử nghiệm và đào tạo của Phiên có nhấp chuột và Truy vấn đơn Các phiên hợp. Vì chúng tôi đang báo cáo về các giá trị nhị phân như kết quả chứ không phải là giá trị nhỏ, ý nghĩa của những kết quả này được thiết lập bằng cách sử dụng thử nghiệm McNemar [42] sử dụng ngưỡng giá trị  $p$  là .05.

## CHƯƠNG 5

### KẾT QUẢ VÀ PHÂN TÍCH

Chúng tôi tiến hành một loạt các thí nghiệm để xác định mức độ hiệu quả của RYSe nhận ra người tìm kiếm khuôn mẫu từ 8-12 tuổi từ hành vi tìm kiếm của họ. Làm do đó cho phép chúng ta thiết lập tính hợp lệ của tiền đề của chúng ta, cũng như làm nổi bật những điểm mạnh và điểm yếu của RYSe.

#### 5.1 Tính năng hiệu quả

Trong phần này, chúng tôi mô tả nghiên cứu cắt bỏ mà chúng tôi đã tiến hành để giới thiệu tính hợp lệ của thiết kế cho RYSe. Đặc biệt, nghiên cứu này cho phép chúng tôi xác định hiệu quả của từng bộ tính năng khi nhận ra khuôn mẫu của chúng ta và cho phép chúng ta hiểu vai trò của từng bộ tính năng trong sự công nhận này. Bằng cách phân tích những kết quả sau đó chúng tôi được cấp một cái nhìn đa chiều về điểm mạnh của chiến lược của chúng tôi, đồng thời cũng đã chỉ ra những lĩnh vực mà chúng ta có thể cải thiện. Như đã mô tả trong Chương 3, chúng tôi nhóm các tính năng dựa trên loại dữ liệu được xem xét: các tính năng dựa trên văn bản hoặc dựa trên phiên các tính năng. Các tính năng này sau đó được mô tả như là chỉ dẫn của các kỹ năng liên quan đến P3 hoặc DC1 (với một số tính năng là chỉ số cho cả hai). Chúng tôi thực hiện cắt bỏ này nghiên cứu về từng tập tính năng (tính năng văn bản, tính năng dựa trên phiên, tính năng liên quan đến P3 và các tính năng liên quan đến DC1) trên cả Phiên có Nhấp chuột và Truy vấn Đơn Các phiên riêng biệt. Bằng cách tiến hành nghiên cứu cắt bỏ này trên 2 tập dữ liệu khác nhau, chúng tôi có thể

miêu tả nhiều góc nhìn khác nhau để chứng minh thêm tính khả thi và hiệu quả của các bộ tính năng được xác định của chúng tôi cũng như RYSe nói chung.

Bộ tính năng	Độ chính xác TNR TPR
RYSe	0,947 0,988 0,783 0,942
Tính năng văn bản	0,985 0,768 0,788 0,940
Tính năng dựa trên phiên*	0,178
Các tính năng liên quan đến P3*	0,946 0,982 0,802 0,879
Các tính năng liên quan đến DC1*	0,976 0,493

Bảng 5.1: Kết quả từ nghiên cứu cắt bỏ trên các Phiên có Nhấp chuột. \* biểu thị ý nghĩa thống kê của một tập hợp tính năng nhất định liên quan đến RYSe ( $p \leq .05$ ).

Khi xem xét kết quả của nghiên cứu cắt bỏ được thực hiện trên Sessions With Nhấp chuột (như được thấy trong Bảng 5.1), chúng tôi nhận thấy rằng các tính năng văn bản nhận ra phần lớn người dùng là, và không phải là, khuôn mẫu của chúng tôi. Bộ tính năng này nhận dạng gần đúng 3 trong số 4 người dùng thuộc nhóm mục tiêu của chúng tôi và 99 trong số 100 người dùng không. Sau đó chúng tôi nhận thấy sự thiếu ý nghĩa thống kê giữa RYSe và các đặc điểm văn bản khiến chúng ta lý luận rằng các đặc điểm văn bản có một tín hiệu mạnh khi nói đến việc nhận ra khuôn mẫu của chúng ta. Như đã thấy trong Chương 3, phần lớn các tính năng của chúng tôi là văn bản, có thể góp phần vào sức mạnh của tín hiệu này. những quan sát được thực hiện liên quan đến các tính năng văn bản của chúng tôi cũng khiến chúng tôi tự hỏi liệu nhóm mục tiêu của chúng tôi có thể được nhận biết chỉ thông qua các đặc điểm văn bản.

Khi kiểm tra các tính năng dựa trên phiên, chúng ta thấy TPR chỉ ra khoảng 1 trong 5 người dùng là khuôn mẫu của chúng tôi đã được công nhận thành công, điều này rõ ràng là ít hơn hơn TPR của các tính năng văn bản. Chúng tôi cũng thấy rằng các tính năng dựa trên phiên nhận ra 94% người dùng không thuộc nhóm mục tiêu của chúng tôi. TNR tương đối cao của các tính năng dựa trên phiên chứng minh rằng người dùng thể hiện khả năng tìm kiếm tốt hơn có thể được xác định. Những kết quả này cũng có vẻ chỉ ra rằng trong khi các tính năng văn bản

có thể có tín hiệu mạnh hơn tín hiệu của các tính năng dựa trên phiên, các tính năng dựa trên phiên tính đến số lượng người dùng không được nhận dạng chỉ từ văn bản (dựa trên việc so sánh các tính năng dựa trên phiên và các tính năng dựa trên văn bản để đưa ra kết quả cho RYSe). Dựa trên phiên các tính năng cũng tận dụng thông tin mà các tính năng văn bản không thể, chẳng hạn như dữ liệu nhấp chuột và mối quan hệ giữa các truy vấn.

Khi chúng tôi kiểm tra các tính năng liên quan đến P3, chúng tôi thấy bộ tính năng này đạt được mức cao hơn TPR hơn RYSe. Điều này chứng minh rằng khuôn mẫu mà chúng ta muốn nhận ra hiển thị hành vi tìm kiếm cho phép chúng tôi xác định màn trình diễn các kỹ năng mà chúng tôi mong đợi họ sở hữu dựa trên các giai đoạn phát triển mà họ đang trải qua, ít nhất là trên tập dữ liệu này. Chúng tôi thấy một sự sụt giảm nhẹ trong TNR trong các tính năng liên quan đến P3 khi so sánh với RYSe. Chỉ sử dụng các tính năng liên quan đến P3 để nhận dạng khuôn mẫu của chúng ta có thể đáng để đánh đổi khỏi việc xác định nhầm người dùng không phải là khuôn mẫu của chúng tôi. Khi xem xét kết quả của các tính năng liên quan đến DC1, chúng ta thấy một TNR tương đương với các bộ tính năng khác, và một TPR chứng minh rằng mọi người dùng khác thuộc nhóm mục tiêu của chúng tôi có thể được nhận ra bằng cách thể hiện các kỹ năng liên quan đến DC1. Vì TPR này thấp hơn là các tính năng liên quan đến P3, chúng tôi đặt câu hỏi về vai trò của các tính năng liên quan đến DC1 chơi trong việc nhận ra khuôn mẫu của chúng ta. Để hiểu đầy đủ hơn những đặc điểm này hiệu quả của bộ vai trò, điều quan trọng là phải xem xét kết quả từ việc thực hiện của chúng tôi nghiên cứu cắt bỏ trên các Phiên truy vấn đơn.

Khi xem Bảng 5.2, chúng ta thấy các tính năng liên quan đến DC1 hiển thị mức cao nhất TPR của tất cả các bộ tính năng. Việc thiếu ý nghĩa thống kê giữa RYSe và Bộ tính năng DC1 làm nổi bật sức mạnh của tín hiệu mà bộ tính năng này cung cấp. Trong khi kết quả liên quan đến P3 trên Phiên có nhấp chuột chứng minh rằng mục tiêu của chúng tôi nhóm có thể được nhận ra bằng cách thể hiện kỹ năng của họ trong các phiên có nhiều dữ liệu hơn, kết quả của bộ tính năng DC1 trên Phiên truy vấn đơn thể hiện những ý tưởng tương tự.



Bộ tính năng	Độ chính xác TNR TPR
RYSe	0,853 0,978 0,463 0,817
Đặc điểm văn bản*	0,991 0,156 0,800 1,000
Tính năng dựa trên phiên*	0,000 0,816 0,987 0,156
Các tính năng liên quan đến P3*	0,868 0,978 0,464
Các tính năng liên quan đến DC1	

Bảng 5.2: Kết quả từ nghiên cứu cắt bỏ trên Phiên truy vấn đơn. \* biểu thị ý nghĩa thống kê của một tập tính năng nhất định liên quan đến RYSe ( $p \leq .05$ ).

Chúng tôi cho rằng các đặc điểm văn bản cũng như các đặc điểm liên quan đến P3 đóng vai trò quan trọng vai trò trong việc thực hiện phân loại trên các Phiên có Nhấp chuột do giả định rằng các phiên có nhiều hơn một truy vấn có tín hiệu văn bản mạnh hơn. Tuy nhiên, khi chỉ đưa ra một truy vấn, các đặc điểm văn bản và các đặc điểm liên quan đến P3 có vẻ yếu hơn tín hiệu. Chúng tôi cũng nhận thấy sự thiếu ý nghĩa thống kê giữa các tính năng liên quan đến P3 và các tính năng văn bản khi thử nghiệm trên Phiên truy vấn đơn, điều này chứng minh cách một tín hiệu mạnh mẽ mà các tính năng liên quan đến P3 có liên quan đến các tính năng văn bản trong bối cảnh này. Hơn nữa, trong bối cảnh này, các đặc điểm văn bản và các đặc điểm liên quan đến P3 bị vượt trội hơn bởi các tính năng liên quan đến DC1. Như một lưu ý bên lề, phiên dựa trên bộ tính năng thực hiện như chúng tôi mong đợi, chuyển thành bộ phân loại đa số do thiếu thông tin phiên cho người dùng trong Phiên truy vấn đơn.

Hiệu suất khác nhau của các bộ tính năng khác nhau trên hai tập dữ liệu cho thấy tất cả đều là cần thiết vì mỗi phương pháp đều nhận ra hành vi của người dùng từ những góc nhìn bổ sung. Các tính năng liên quan đến DC1 hỗ trợ phát hiện phiên sớm, trong khi các tính năng P3 hỗ trợ phát hiện người dùng sâu hơn trong phiên của họ. Tuy nhiên, trong khi tính năng dựa trên phiên bộ có TPR thấp chúng ta có thể suy đoán từ kết quả tổng thể rằng việc kết hợp tính năng này thiết lập với bộ tính năng văn bản mang lại kết quả tổng thể cao hơn.

Tham khảo Bảng 5.1, chúng ta thấy rằng RYSe đạt được Độ chính xác xấp xỉ

95%, có thể nhận dạng thành công khoảng 80% người dùng thuộc nhóm mục tiêu của chúng tôi và khoảng 99% người dùng không làm như vậy. Điều này cho thấy rằng phần lớn những người tìm kiếm theo khuôn mẫu 8-12 tuổi từ Phiên của chúng tôi với Nhấp chuột có thể được công nhận bằng cách thể hiện các kỹ năng mà chúng ta tìm kiếm trong hành vi tìm kiếm của họ. Hơn nữa, với TNR của RYSe là khoảng 99%, chúng tôi thấy rằng những người dùng không thuộc nhóm mục tiêu của chúng tôi có thể được phát hiện một cách đáng tin cậy bằng các đặc điểm chỉ ra kỹ năng. Tuy nhiên, những người dùng này cũng giống như những người dùng không thuộc nhóm mục tiêu của chúng tôi được biết là sử dụng truy vấn ngôn ngữ tự nhiên và có thể sử dụng ngôn ngữ tương tự như nhóm mục tiêu của chúng tôi (đơn giản từ), có thể giải thích cho TPR thấp hơn một chút của chúng tôi (với mô hình của chúng tôi bị lỗi phía TN). Chúng tôi cũng điều tra kết quả của RYSe trên Phiên truy vấn đơn (Bảng 5.4), một tập dữ liệu có nghĩa là mô phỏng sự bắt đầu của một phiên tìm kiếm. Chúng ta thấy RYSe với Độ chính xác 87,5%, nhận dạng 99% người dùng không phải là khuôn mẫu của chúng tôi và 40% đó là. TPR này cho thấy rằng ngay cả với một truy vấn và không có thông tin phiên, RYSe vẫn có khả năng nhận dạng 2 trong số 5 người dùng là trẻ em từ 8-12 tuổi người tìm kiếm. TNR của thử nghiệm trên Phiên truy vấn đơn phù hợp với kết quả của chúng tôi kiểm tra Phiên có Nhấp chuột; có thể nhận ra người dùng không thuộc về nhóm mục tiêu của chúng tôi từ một truy vấn duy nhất. Để có được sự hiểu biết sâu sắc hơn về những kết quả này có thể có ý nghĩa gì và xác minh những quan sát của chúng tôi, chúng tôi cần đặt chúng vào bối cảnh thích hợp.

## 5.2 So sánh với các đường cơ sở

Để ngữ cảnh hóa hiệu suất tổng thể của RYSe, chúng tôi so sánh RYSe với các đường cơ sở được điều chỉnh để thực hiện trong phạm vi nhận dạng khuôn mẫu của chúng ta. Điều này thí nghiệm cho phép chúng ta không chỉ xác định RYSe vượt trội ở đâu mà còn xác định các lĩnh vực

để cải thiện tiềm năng. Đối với thí nghiệm này, chúng tôi đào tạo tất cả các đường cơ sở đã điều chỉnh của mình về các truy vấn nối tiếp của các phiên được tìm thấy trong dữ liệu đào tạo tổng hợp của chúng tôi (như được trình bày chi tiết trong Phần 4.4.2), sau đó kiểm tra riêng trên các Phiên có Nhấp chuột và Phiên truy vấn đơn. Hơn nữa, chúng tôi cũng thực hiện các thử nghiệm McNemar trên kết quả của các đường cơ sở của chúng tôi, xác định ý nghĩa thống kê của kết quả của họ không chỉ trong liên quan đến RYSe mà còn liên quan đến nhau. Trừ khi có quy định khác, kết quả được báo cáo là có ý nghĩa thống kê ( $p \leq .05$ ).

Khi chúng ta lần đầu tiên xem Bảng 5.3, chúng ta nhận thấy rằng RYSe vượt trội hơn đáng kể tất cả các đường cơ sở của chúng tôi về Độ chính xác. Bộ phân loại Đa số thực hiện như mong đợi, nhận ra phần lớn một cách hoàn hảo trong khi không nhận ra bất kỳ người dùng nào là của chúng tôi khuôn mẫu. Bộ phân loại dựa trên quy tắc đạt được độ chính xác khoảng 90%, với TNR là 100% và TPR là 52%. Đường cơ sở này nhận ra gần như nhiều người dùng không phải là khuôn mẫu của chúng tôi như là người phân loại Đa số, đồng thời cũng có thể để nhận ra những người dùng thuộc nhóm mục tiêu của chúng tôi. Tuy nhiên, khi chúng tôi tham khảo đối với Bảng 5.4, những hạn chế của bộ phân loại dựa trên quy tắc trở nên rõ ràng khi cơ sở nhận dạng chính xác cùng những người dùng như bộ phân loại Đa số. Sự thất bại này của Bộ phân loại dựa trên quy tắc để nhận dạng bất kỳ người tìm kiếm khuôn mẫu nào từ 8-12 tuổi Phiên truy vấn đơn làm nổi bật những thiếu sót khi dựa vào phương pháp phân loại này để xác định khuôn mẫu của chúng tôi. Bộ phân loại dựa trên văn bản hiệu quả hơn Bộ phân loại đa số, nhưng kém hiệu quả hơn bộ phân loại dựa trên quy tắc, trên các phiên có Nhấp chuột. Tuy nhiên, trình phân loại dựa trên văn bản vượt trội hơn cả hai trình phân loại trên Truy vấn đơn Phiên. Kết quả của bộ phân loại dựa trên văn bản trên cả hai tập dữ liệu này chứng minh rằng khuôn mẫu của chúng ta có thể được nhận ra chỉ từ văn bản của họ, mặc dù ít hơn đáng kể người dùng được nhận ra khi so sánh bộ phân loại này với RYSe. Tuy nhiên, có hai nhiều bộ phân loại tận dụng văn bản để xác định người tìm kiếm theo khuôn mẫu từ 8-12 tuổi,

bộ phân loại Đa tính năng cũng như bộ phân loại Đa mô hình. Bộ phân loại Đa tính năng bộ phân loại có TPR cao hơn khi so sánh với bộ phân loại Đa mô hình, nhưng Bộ phân loại đa mô hình có TNR cao hơn. Xu hướng này tiếp tục khi kiểm tra cả hai những mô hình này có hiệu quả khi thử nghiệm trên các Phiên truy vấn đơn. Tính năng đa dạng bộ phân loại có TPR cao hơn trong khi bộ phân loại Đa mô hình có TNR cao hơn. Hiệu suất từ cả hai mô hình này chứng minh rằng có sự đánh đổi khi nhận ra người dùng nào là, hoặc không phải, khuôn mẫu của chúng tôi từ văn bản của họ. Bộ phân loại đa tính năng đổi TPR cao hơn lấy TNR thấp hơn, trong khi Bộ phân loại đa mô hình phân loại thực hiện ngược lại. Tuy nhiên, RYSe có TPR tương đương với Multi-Feature phân loại và TNR tương đương với phân loại Đa mô hình. Cho dù văn bản nhận dạng được thực hiện bằng cách sử dụng các tính năng như nội dung hoặc các loại thẻ lời nói (như Bộ phân loại đa tính năng) hoặc từ tổng hợp các mô hình được đào tạo bằng cách sử dụng túi từ và các tính năng tf-idf (giống như bộ phân loại Đa mô hình), không cách tiếp cận nào có thể so sánh được với Hiệu suất tổng thể của RYSe phát hiện người dùng dựa trên các tính năng được thiết kế để nhận dạng thể hiện kỹ năng.

Bộ tính năng	Độ chính xác TNR TPR
RYSe	0,948 0,988 0,783
Phân loại đa số*	0,801 1,000 0,000 0,904
Phân loại dựa trên quy tắc*	0,999 0,523 0,856 0,995
Phân loại dựa trên văn bản*	0,299 0,915 0,943 0,807
Phân loại đa tính năng*	0,935 0,990 0,712
Phân loại đa mô hình*	

Bảng 5.3: Kết quả đánh giá hiệu suất của chúng tôi khi so sánh RYSe với các giá trị cơ sở trên Phiên có nhấp chuột. \* biểu thị ý nghĩa thống kê của một giá trị cơ sở nhất định liên quan đến RYSe ( $p \leq .05$ ).

Như đã đề cập trong Phần 4.3, TPR phản ánh mục tiêu chính của chúng tôi là nhận ra người tìm kiếm khuôn mẫu 8-12 tuổi. Do đó, chúng tôi xem xét bộ phân loại Đa tính năng

Bộ tính năng	Độ chính xác TNR TPR
RYSe	0,875 0,978 0,463
Phân loại đa số*	0,800 1,000 0,000
Phân loại dựa trên quy tắc*	0,800 1,000 0,000
Phân loại dựa trên văn bản*	0,808 1,000 0,041
Phân loại đa tính năng*	0,833 0,987 0,221
Phân loại đa mô hình*	0,834 0,996 0,179

Bảng 5.4: Kết quả đánh giá hiệu suất của chúng tôi khi so sánh RYSe với đường cơ sở trên Phiên truy vấn đơn. \* biểu thị ý nghĩa thống kê của một đường cơ sở liên quan đến RYSe ( $p \leq .05$ ).

là đường cơ sở mạnh nhất của chúng tôi vì nó có TPR cao nhất trong số tất cả các đường cơ sở trên cả hai bộ dữ liệu. Tuy nhiên, khi chúng ta so sánh trực tiếp bộ phân loại Đa tính năng với RYSe, chúng ta quan sát rằng RYSe có TNR cao hơn bộ phân loại Đa tính năng khi thử nghiệm trên Phiên có Nhấp chuột. Chúng tôi thừa nhận rằng việc đánh đổi TNR thấp hơn cho TPR cao hơn sẽ có giá trị, miễn là kết quả của bộ phân loại vẫn có hiệu quả trên tất cả các tập dữ liệu. Khi xem TPR của bộ phân loại Đa tính năng trên Truy vấn đơn Các phiên này mạnh hơn đáng kể so với tất cả các phiên bản cơ sở, nhưng chỉ bằng một nửa so với RYSe. Hơn nữa, TPR cao hơn của bộ phân loại Đa tính năng trên các Phiên có Nhấp chuột đi kèm với sự đánh đổi là nhận ra sai khoảng 5% những điều không theo khuôn mẫu của chúng ta người dùng. Sự đánh đổi này có thể có những hậu quả không lường trước được khi áp dụng trong thực tế thiết lập thế giới, chẳng hạn như yêu cầu người dùng không phải là khuôn mẫu của chúng tôi xác nhận rằng trong thực tế họ không phải là những người tìm kiếm khuôn mẫu 8-12 tuổi và do đó một lần nữa dựa vào phản hồi trực tiếp. TPR tương đương và TNR thấp hơn của Đa tính năng bộ phân loại trên Phiên có Nhấp chuột, cùng với TPR Đa tính năng thấp hơn đáng kể bộ phân loại có trên Phiên truy vấn đơn nhấn mạnh cách RYSe, nhìn chung; vượt trội hơn Bộ phân loại đa tính năng.

### 5.3 Tác động của thời lượng phiên họp đến hiệu quả

Vì RYSe có mục đích tự động nhận dạng người dùng từ hành vi tìm kiếm của họ, chúng tôi mô phỏng bối cảnh tự động phát hiện người dùng trong phiên tìm kiếm. Điều này được thực hiện bằng cách kiểm tra kết quả của RYSe dựa trên độ dài phiên như đã xác định theo số lượng truy vấn. Chúng tôi điều tra cách RYSe và bộ phân loại Đa tính năng (được chọn vì nó khác biệt ở chỗ là đường cơ sở mạnh nhất) thực hiện trong việc xác định khuôn mẫu của chúng tôi bằng cách kiểm tra kết quả thử nghiệm trên các Phiên có Nhấp chuột được phân đoạn theo độ dài phiên (số lượng truy vấn) cũng như kết quả thử nghiệm trên Single Query Các phiên họp. Làm như vậy cho phép chúng tôi hiểu được lượng thông tin chúng tôi yêu cầu từ người dùng để nhận dạng thành công và cung cấp cho chúng ta cái nhìn sâu sắc so sánh hiệu suất của RYSe với bộ phân loại đa tính năng.

Khi xem Bảng 5.5, chúng ta thấy rằng RYSe có khả năng nhận dạng tốt hơn khuôn mẫu từ các phiên ngắn hơn. Khi xem TNR của RYSe, chúng ta thấy ngược lại; phiên càng dài thì khả năng RYSe nhận dạng người dùng càng cao tạo ra phiên đó không thuộc về khuôn mẫu của chúng tôi. Mặc dù TPR giảm dần và TNR tăng dần khi nhìn vào kết quả chung chúng ta thấy những xu hướng này được cân bằng. Điều này có thể là do thực tế là khoảng hai phần ba số phiên chúng tôi kiểm tra có độ dài phiên là hai hoặc ít hơn hai truy vấn, có nghĩa là kết quả từ những phiên đó có trọng lượng tổng thể lớn hơn. Hơn nữa, chúng tôi cũng hãy xem xét Bảng 5.4, như đã đề cập trước đó, cho thấy RYSe có khả năng nhận dạng 4 trong số 10 người dùng là khuôn mẫu của chúng tôi. Để ngữ cảnh hóa những kết quả này, chúng tôi sẽ so sánh chúng với đường cơ sở có hiệu suất cao nhất.

Khi kiểm tra các kết quả tìm thấy trong Bảng 5.4 và Bảng 5.5, chúng tôi quan sát thấy rằng RYSe có TPR cao hơn bộ phân loại Đa tính năng trên các phiên có 1 truy vấn

(có và không có nhấp chuột), TPR tương đương trên các phiên có 2-3 truy vấn với nhấp chuột và TPR thấp hơn đáng kể trong bất kỳ phiên nào dài hơn thế. Tuy nhiên, khi nhìn vào ma trận nhầm lẫn đằng sau các số liệu, điều này chuyển thành RYSe nhận ra khoảng 30 người dùng ít hơn (trong số 1500) thuộc nhóm mục tiêu của chúng tôi hơn bộ phân loại Đa tính năng, làm cho kết quả của chúng có thể so sánh được với sự khác biệt của chúng các ences được nhấn mạnh bởi sự chuẩn hóa của TPR. Chúng ta cũng thấy rằng TNR của RYSe cao hơn đáng kể so với bộ phân loại Đa tính năng trên các phiên có bất kỳ độ dài nào, gợi ý lại cuộc thảo luận về ma sát được thêm vào quá trình nhận dạng đã đề cập trong phần trước. Xu hướng tăng của TNR và xu hướng giảm của TPR đối với kết quả của RYSe cũng khiến chúng ta tự hỏi tác động của việc tổng hợp các tính năng trong một phiên có khả năng nhận dạng người tìm kiếm khuôn mẫu từ 8-12 tuổi. Nó có thể những người dùng không phải là người tìm kiếm theo khuôn mẫu 8-12 tuổi có thể có hành vi tìm kiếm tương đối đồng đều, đặc biệt là trong các phiên tìm kiếm dài hơn, trong khi người dùng những người có khuôn mẫu của chúng ta có thể có hành vi ít đồng nhất hơn, khó nắm bắt hơn khi được tổng hợp qua các phiên dài hơn. Kiểm tra các phiên này theo từng truy vấn thay vào đó, thay vì tổng hợp qua các phiên, có thể mang lại kết quả mạnh mẽ hơn. Phương pháp luận việc xem xét các tính năng truy vấn theo truy vấn có thể là một yếu tố góp phần cho phép Đường cơ sở Đa tính năng để duy trì TPR trên 70%. Hơn nữa, Đa tính năng cơ sở làm nổi bật cách thức khuôn mẫu của chúng ta có thể được nhận ra từ văn bản của họ, vì điều này kết quả của bộ phân loại có thể so sánh với RYSe khi xem xét các phiên chỉ có một truy vấn và nhấp chuột (độ dài phổ biến nhất của một phiên trong Phiên có nhấp chuột là cũng như trực tuyến [23]).

Tuy nhiên, khi xem Bảng 5.4, RYSe nhận ra khoảng 46% người dùng ai là khuôn mẫu của chúng ta, so với Bộ phân loại đa mô hình nhận dạng khoảng 22% trong số những người dùng này. Những kết quả khác nhau của Bộ phân loại đa tính năng

chứng minh rằng phân loại văn bản không phải là lựa chọn hiệu quả nhất khi cố gắng để nhận ra người dùng mục tiêu của chúng tôi khi bắt đầu phiên của họ. Có lý do để lý giải rằng Càng phát hiện ra khuôn mẫu của chúng ta sớm trong một phiên, các nguồn lực và sự hỗ trợ phù hợp sẽ càng nhanh chóng có thể được cung cấp bởi nền tảng họ đang sử dụng. Lý luận này đặt RYSe ngoài Bộ phân loại đa mô hình, vì nó vượt trội hơn đáng kể so với đường cơ sở này trên Truy vấn đơn Phiên, một tập dữ liệu nhằm mô phỏng các phiên mới chỉ bằng một truy vấn và không cần nhấp chuột.

RYSe				Phân loại đa tính năng				
# Truy vấn	Acc	TNR	TPR	# Truy vấn	Acc	TNR	TPR	
1	0,946	0,982	0,844	0,935	1	0,936	0,979	0,817
	0,986	0,764	0,946	0,992	2*	0,923	0,962	0,794
2 3	0,731	0,941	0,994	0,613	3*	0,914	0,937	0,805
	0,954	0,996	0,616		4*	0,901	0,929	0,727
4 5					5*	0,902	0,913	0,811
6+	0,973	0,997	0,464	0,948	6+*	0,860	0,867	0,716
Tổng thể	0,988	0,783	Tổng thể*	0,915	0,943	0,807		

Bảng 5.5: Đánh giá hiệu suất của RYSe (kết quả bên trái) so với Bộ phân loại đa tính năng (kết quả bên phải) trên các phiên có độ dài khác nhau. \* biểu thị ý nghĩa thống kê ( $p \leq .05$ ) liên quan đến kết quả RYSe trong các phiên hợp của cùng một chiều dài.

5.4 Thảo luận

Trong chương này, chúng tôi đã trình bày một số thí nghiệm được tiến hành để xác định hiệu quả của RYSe phát hiện ra người tìm kiếm khuôn mẫu từ 8-12 tuổi từ màn hình hiển thị của kỹ năng suy ra từ hành vi tìm kiếm của họ. Những thí nghiệm này đã được thực hiện trên hai tập dữ liệu khác nhau; một tập mô phỏng nhiều phong cách và tương tác của người dùng bằng cách chủ yếu sử dụng nhân dựa trên quy tắc trên nguồn dữ liệu chưa được gắn nhãn (Phiên có Nhấp chuột), và cái còn lại dựa trên các truy vấn được tạo ra bởi người dùng được biết đến



để thuộc về nhóm mục tiêu của chúng tôi (Phiên truy vấn đơn). Với bản chất của những các tập dữ liệu, điều quan trọng là phải phân biệt hiệu suất dựa trên kết quả của mỗi thí nghiệm được thực hiện trên chúng. Các thí nghiệm được thực hiện trong Phần 5.1 trên Các phiên có nhấp chuột cho thấy khuôn mẫu của chúng ta có thể được nhận ra từ màn hình hiển thị kỹ năng chỉ ra người dùng trong P3, trong khi các thử nghiệm trên Phiên truy vấn đơn cho thấy rằng khuôn mẫu của chúng ta có thể được nhận ra từ màn trình diễn kỹ năng chỉ ra người dùng trong DC1. Những tập dữ liệu khác nhau này cũng làm nổi bật tầm quan trọng của việc xem xét các tính năng chỉ ra cả hai giai đoạn phát triển, cho thấy rằng chúng rất quan trọng trong việc nhận ra người tìm kiếm khuôn mẫu từ 8-12 tuổi trong nhiều phần khác nhau của phiên làm việc của người dùng này.

Mặc dù RYSe có thể so sánh với đường cơ sở mạnh nhất khi nhận ra khuôn mẫu trên dữ liệu mô phỏng, RYSe thực sự vượt trội hơn đáng kể so với đường cơ sở đó khi phát hiện ra người tìm kiếm khuôn mẫu 8-12 tuổi từ một truy vấn duy nhất. Với TPR của 46% khi thử nghiệm trên Phiên truy vấn đơn, so với Phiên truy vấn đa tính năng bộ phân loại ở mức 22%, chứng minh rằng RYSe có thể nhận ra khuôn mẫu của chúng ta từ màn hình của kỹ năng được tìm thấy trong một truy vấn duy nhất hiệu quả gấp đôi so với đường cơ sở gần nhất trên tập dữ liệu. Tầm quan trọng của hiệu quả RYSe khi thực hiện nhận dạng trên tập dữ liệu được khuếch đại thêm bởi tiền đề rằng phát hiện ra khuôn mẫu của chúng ta sớm phiên của họ rất quan trọng. Tầm quan trọng này có hai mặt. Đầu tiên, người dùng càng sớm có thể yêu cầu các công cụ và tài nguyên cụ thể có thể được nhận ra trong phiên của họ, sớm hơn một SE có thể đáp ứng những nhu cầu đó. Thứ hai, nếu người dùng cần những tài nguyên này chỉ được phát hiện sau một vài truy vấn, không có gì đảm bảo họ sẽ nhận được sự hỗ trợ và những người dùng này có thể chấm dứt phiên tìm kiếm của họ do thất vọng hoặc bối rối trước sự nhận biết này có thể xảy ra.

Mặc dù RYSe có hiệu quả hơn về tổng thể so với các đường cơ sở, nhưng điều quan trọng là làm nổi bật những gì có thể học được từ các đường cơ sở đã điều chỉnh này. Tìm kiếm các phiên được tạo

bởi người tìm kiếm khuôn mẫu 8-12 tuổi có thể cung cấp đủ văn bản để nhận ra chúng tôi  
khuôn mẫu, dù là thông qua túi mô hình từ ngữ hay phong cách và đặc điểm dựa trên nội dung.  
Như chúng tôi đã đề cập trước đó, các bộ phân loại Đa tính năng cho thấy triển vọng đặc biệt  
khi nhận ra người tìm kiếm khuôn mẫu 8-12 tuổi. Điều đáng chú ý về điều này  
phân loại là việc sử dụng mô hình chủ đề, điều này khiến chúng ta tự hỏi loại tính năng này như thế nào  
có thể được sử dụng để nhận ra các chỉ số về kỹ năng được thể hiện bởi khuôn mẫu của chúng ta. Một  
sự khác biệt là cách đường cơ sở này xem xét các tính năng dựa trên các truy vấn nối tiếp, thay vì  
hơn là tổng hợp các tính năng qua các phiên. Cả hai đều cần được điều tra thêm.

Cuối cùng, các thí nghiệm chúng tôi đã tiến hành cho thấy rằng khuôn mẫu của chúng tôi có thể  
được nhận biết bằng cách thể hiện kỹ năng cho thấy những giai đoạn phát triển nào  
người dùng có thể ở trong. Hơn nữa, những người dùng này có thể được nhận ra sớm trong các phiên của họ,  
vượt trội đáng kể so với tất cả các đường cơ sở trên các truy vấn mô phỏng sự khởi đầu của một  
phiên tìm kiếm (Phiên truy vấn đơn). Mặc dù vẫn còn chỗ để cải thiện, như  
một bằng chứng về khái niệm, các thí nghiệm này chứng minh rằng RYSe có thể nhận ra  
người tìm kiếm khuôn mẫu 8-12 tuổi bằng cách thể hiện kỹ năng chỉ ra các giai đoạn của  
sự phát triển mà chúng tôi mong đợi những người tìm kiếm này sẽ tham gia và sự công nhận này có thể xảy ra  
tự động.

## CHƯƠNG 6

### KẾT LUẬN, HẠN CHẾ VÀ CÔNG VIỆC TƯƠNG LAI

Trong luận án này chúng tôi giới thiệu RYSe, một chiến lược được thiết kế để tự động phát hiện người tìm kiếm khuôn mẫu 8-12 tuổi từ hành vi tìm kiếm của họ. RYSe thực hiện điều đó bằng cách kiểm tra các tính năng dựa trên văn bản và các tính năng dựa trên phiên, vì cả hai đều là chỉ số của các kỹ năng mà chúng tôi mong đợi người tìm kiếm khuôn mẫu từ 8-12 tuổi sẽ thể hiện dựa trên các giai đoạn phát triển mà họ có thể đang trải qua. Những đặc điểm này sau đó là được sử dụng bởi một bộ phân loại rừng ngẫu nhiên để nhận dạng khuôn mẫu của chúng tôi. Khi nhìn vào hiệu suất của RYSe, chúng ta thấy rằng khuôn mẫu của chúng ta thực sự có thể được xác định bằng màn hình của các kỹ năng được mong đợi từ một ai đó trong P3 và DC1. Sự công nhận này có thể xảy ra từ một phiên đang diễn ra (như được chứng minh bằng các thí nghiệm được thực hiện trên Phiên với Nhấp chuột) hoặc thậm chí khi bắt đầu một phiên (như được chứng minh bằng các thí nghiệm được thực hiện trên các Phiên truy vấn đơn). Hơn nữa, kết quả của các thí nghiệm được thực hiện trên hai tập dữ liệu này ngụ ý rằng RYSe có thể cung cấp cho SE một chiến lược nhận ra một người dùng dựa trên kỹ năng hiển thị (thay vì dựa vào độ tuổi hoặc lớp học) một cách tự động.

Trong khi thiết kế và phát triển RYSe, chúng tôi đã gặp phải một số hạn chế. Mặc dù bộ dữ liệu của chúng tôi hữu ích cho việc thực hiện các thí nghiệm cho phép chúng tôi thiết lập hiệu quả của RYSe, do bản chất của việc dán nhãn Phiên có Nhấp chuột, có một số các tính năng chúng tôi không thể sử dụng khi thực hiện nhận dạng này. Vì thời lượng phiên là được sử dụng để dán nhãn, chúng ta không thể sử dụng nó khi phân định khuôn mẫu của mình với người khác

người dùng. Hơn nữa, nhật ký truy vấn AOL đã được xử lý trước, với dấu chấm câu loại bỏ hầu hết các truy vấn và văn bản được chuyển thành chữ thường. Chúng tôi vẫn xem xét những điều này các tính năng, nhưng sự thừa thớt của chúng trong tập dữ liệu lớn nhất của chúng tôi có thể ảnh hưởng đến sức mạnh của báo hiệu rằng dấu câu và chữ viết hoa mang theo. Cuối cùng, tập dữ liệu Phiên với Nhấp chuột là giới hạn ở các phiên có ít nhất một lần nhấp, nghĩa là chỉ một số loại phiên nhất định được nắm bắt bằng phương pháp dán nhãn của chúng tôi. Chúng tôi cố gắng giảm thiểu những hạn chế này bằng cách cũng sử dụng Phiên truy vấn đơn cho mục đích đánh giá, trong đó có loại phiên cụ thể bị thiếu trong Phiên có nhấp chuột. Hơn nữa, Đơn Phiên truy vấn chứa các truy vấn mà chúng tôi chắc chắn sẽ được tạo ra bởi người dùng khuôn mẫu của chúng tôi. Tuy nhiên, chúng tôi cũng thừa nhận rằng tập dữ liệu Phiên truy vấn đơn hiển thị các hạn chế. Như được mô tả trong Phần 4.1.2, Phiên truy vấn đơn bao gồm của các mẫu truy vấn từ những người dùng thuộc nhóm mục tiêu của chúng tôi, cũng như những người dùng trên 18 tuổi. Tuy nhiên, điều đáng chú ý là các tương tác SE từ một phân khúc người dùng quan trọng không chỉ bị thiếu trong Phiên truy vấn đơn mà còn Phiên có nhấp chuột: khuôn mẫu điển hình của những người từ 13 đến 17 tuổi. Chúng tôi mong đợi những người dùng này có bộ kỹ năng gần nhất với khuôn mẫu của chúng ta và do đó có thể là thách thức nhất phân biệt với nhóm mục tiêu của chúng tôi. Thực tế là cả hai tập dữ liệu đều thiếu phân khúc này người dùng có thể tác động đến kết quả thử nghiệm trên các tập dữ liệu này vì hiện tại đã tồn tại điều này khoảng cách đáng kể trong các màn trình diễn kỹ năng dự kiến. Hơn nữa, cả hai tập dữ liệu đều thiếu độ chi tiết ngoài việc nhận ra người dùng có phải là khuôn mẫu của chúng tôi hay không, điều này có thể là được coi là một hạn chế khác. Trong khi sử dụng cả hai tập dữ liệu trong các thí nghiệm của chúng tôi giúp chúng ta phần nào giảm bớt những hạn chế mà chúng ta sẽ gặp phải khi chỉ sử dụng một trong hai, chúng tôi dự định sẽ giải quyết thêm những hạn chế này thông qua một loạt các nghiên cứu người dùng. Nghiên cứu người dùng đầu tiên mà chúng tôi đề xuất sẽ được thiết kế để tạo ra một tập dữ liệu mới. Chúng tôi dự định nắm bắt các tương tác giữa người dùng và SE từ nhiều hơn

phạm vi người dùng được phân loại: người tìm kiếm theo khuôn mẫu từ 8-12 tuổi, khuôn mẫu người tìm kiếm 13-17 tuổi và người dùng không thuộc bất kỳ khuôn mẫu nào trong số đó. Điều này sẽ giải quyết những hạn chế đã nêu ở trên và đồng thời mở ra cánh cửa tương lai nghiên cứu trong việc nhận ra những khuôn mẫu mới. Việc xây dựng bộ dữ liệu mới này cho phép chúng tôi áp dụng quá trình xử lý trước của chúng tôi, đảm bảo rằng thông tin có giá trị bị thiếu hoặc được sử dụng để dán nhãn, Phiên có nhấp chuột có thể được khám phá trong các thử nghiệm trong tương lai. Người dùng thứ hai nghiên cứu chúng tôi dự định thực hiện sẽ là đánh giá trực tuyến RYSe. Đánh giá này sẽ cho phép chúng tôi thực hiện phân tích và đánh giá sâu hơn so với ngoại tuyến đánh giá được trình bày trong công trình này đồng thời cũng cho phép chúng tôi phân tích hiệu quả của RYSe nhận dạng người tìm kiếm điển hình từ 8-12 tuổi trong môi trường thời gian thực.

Việc công nhận thành công người tìm kiếm khuôn mẫu từ 8-12 tuổi cũng đặt ra nền tảng để khám phá cách SE có thể điều chỉnh hỗ trợ thuật toán của họ dựa trên nhận ra người dùng bằng các kỹ năng tiềm ẩn của họ. SE dành cho trẻ em có thể sử dụng RYSe để cung cấp cung cấp tài nguyên và hỗ trợ dựa trên việc xác định bộ kỹ năng của người dùng được công nhận. Điều này có thể đòi hỏi phải quản lý các kết quả dễ hiểu đối với nhóm mục tiêu của chúng tôi, điều chỉnh truy vấn các gợi ý và đoạn trích dựa trên trình độ đọc viết dự kiến của người dùng và thậm chí sử dụng phép thuật các công cụ kiểm tra được thiết kế để nhận dạng lỗi chính tả thường gặp ở trẻ em từ 8-12 tuổi người tìm kiếm cũ. SE thương mại cũng có thể tận dụng RYSe để đạt được kết quả tương tự, trong khi cũng có khả năng sàng lọc nội dung không phù hợp với người dùng được công nhận là của chúng tôi khuôn mẫu cũng như thực thi các tiêu chuẩn COPPA. Cuối cùng, các nhà nghiên cứu khác có thể sử dụng RYSe như một công cụ nền tảng để nhận ra một phân khúc người tìm kiếm cụ thể khi nghiên cứu và thiết kế các công cụ giải quyết những khó khăn của người dùng, như khả năng nhận ra rằng định kiến của chúng ta đi đôi với việc cung cấp cho họ sự hỗ trợ họ cần.

Trong khi phạm vi của RYSe là người tìm kiếm khuôn mẫu từ 8-12 tuổi, trẻ em

một nhóm lớn các cá nhân, với những tài năng và khả năng độc đáo vượt ra ngoài phạm vi của chúng ta định kiến và áp đặt lên trẻ em những đặc điểm thể hiện ở các giai đoạn phát triển khác nhau. Vì RYSe chứng minh rằng khuôn mẫu của chúng ta có thể được nhận ra bằng cách thể hiện các kỹ năng liên quan đến các giai đoạn phát triển mà chúng tôi mong đợi những người dùng này sẽ đạt được, nó đặt nền tảng cho việc nhận ra các khuôn mẫu khác cũng có thể gặp khó khăn với SE sử dụng. Điều này bao gồm thanh thiếu niên (13-17), như đã đề cập trước đó; người dùng trẻ hơn chúng tôi khuôn mẫu (8 tuổi trở xuống) và người dùng trên 65 tuổi. Tất cả những khuôn mẫu này thể hiện các giai đoạn phát triển sẽ có tác động đến hành vi tìm kiếm.

Trong khi thực hiện thí nghiệm được trình bày chi tiết trong Phần 5.2, chúng tôi nhận ra rằng ngoài tổng hợp các giá trị tính năng qua các phiên cũng có tiềm năng sử dụng truy vấn bằng cách nhập dữ liệu dựa trên truy vấn để nhận dạng người tìm kiếm điển hình ở độ tuổi 8-12. Trong khi RYSe hiện cung cấp nền tảng để thực hiện phân loại nhằm nhận ra người tìm kiếm khuôn mẫu 8-12 tuổi, các hướng nghiên cứu trong tương lai liên quan đến xem xét các chiến lược có thể thực hiện phân loại dựa trên truy vấn theo đầu vào truy vấn, chẳng hạn như phân loại chuỗi thời gian. Có một số mô hình xem xét nhấp chuột thông qua dữ liệu để cung cấp khả năng cá nhân hóa có thể được điều chỉnh để nhận ra khuôn mẫu của chúng ta. Vì cá nhân hóa dựa trên việc nhận dạng đặc điểm của người dùng, có khả năng các chiến lược này có thể cung cấp một điểm khởi đầu cho khả năng mở rộng mô hình của chúng tôi để sử dụng dữ liệu chuỗi thời gian để nhận ra các chỉ số về kỹ năng và cải thiện hiệu quả chung của RYSe. Hơn nữa là những tính năng khác mà chúng ta cũng có thể xem xét để thể hiện lời hứa khi nhận ra người tìm kiếm theo khuôn mẫu 8-12 tuổi, chẳng hạn như mô hình chủ đề và những người vẫn còn chưa được khám phá, giống như các đặc điểm theo ngữ cảnh.

Mặc dù người ta đã thừa nhận rằng các giai đoạn phát triển ảnh hưởng đến trẻ em hành vi tìm kiếm của dren [22, 31], nhận ra những người dùng này dựa trên việc thể hiện các kỹ năng

khao khát những giai đoạn này vẫn chưa được khám phá. Với RYSe, chúng tôi chứng minh rằng khi được phân loại theo khuôn mẫu, những người tìm kiếm trẻ tuổi có thể được nhận ra bằng cách thể hiện các kỹ năng được quyết định bởi các giai đoạn phát triển mà chúng ta mong đợi chúng sẽ trải qua. Khám phá này mở ra cánh cửa nghiên cứu được đề cập trong các đoạn trước như: nhận ra những khuôn mẫu, thiết kế các công cụ tập trung trực tiếp hơn vào việc hỗ trợ khuôn mẫu 8-12 người tìm kiếm năm tuổi và có ý nghĩa rằng SE dành cho trẻ em có thể điều chỉnh phản ứng của chúng đối với những người dùng này không dựa trên độ tuổi hoặc cấp bậc mà thay vào đó bằng cách tự động nhận dạng kỹ năng tìm kiếm của người trẻ. Hơn nữa, RYSe là bằng chứng về khái niệm, chứa đựng khả năng rõ ràng của việc cải thiện hơn nữa khả năng nhận ra chiến lược này người tìm kiếm theo khuôn mẫu 8-12 tuổi bằng cách xem xét phân loại chuỗi thời gian cũng như tính năng dựa trên nội dung.

## TÀI LIỆU THAM KHẢO

- [1] Máy tính và Truy cập Internet tại Hoa Kỳ: 2012. [//www.census.gov/data/tables/2012/demo/computer-internet/computer-use-2012.html](https://www.census.gov/data/tables/2012/demo/computer-internet/computer-use-2012.html), 2012.
- [2] Theo dõi phiên Trec 2010-2014. <https://trec.nist.gov/data/session.html>, 2014. (truy cập ngày 2020-18-09).
- [3] Twitter. <http://www.twitter.com/>, 2019.
- [4] Thời gian dành cho trẻ em. <https://www.timeforkids.com/>, 2020.
- [5] Trình phân tích cấp độ D. <http://www.personal.psu.edu/xx113/downloads/d-level.html>, 2021.
- [6] Oghenemaro Anuyah, Ashlee Milton, Michael Green và Maria Soledad Pera. Phân tích thực nghiệm về phản ứng của công cụ tìm kiếm đối với các truy vấn tìm kiếm trên web liên quan đến bối cảnh lớp học. Tạp chí Quản lý thông tin Aslib, 72(1):88-111, 2019.
- [7] Ion Madrazo Azpiazu, Nevena Dragovic và Maria Soledad Pera. Tìm kiếm, hiểu và học: Làm cho các nhiệm vụ khám phá thông tin hữu ích cho trẻ em và giáo viên. Trong Biên bản Hội thảo quốc tế lần thứ 2 về Tìm kiếm như Học tập, Đồng tổ chức với Hội nghị quốc tế lần thứ 39 của ACM SIGIR về Nghiên cứu và Phát triển trong Truy xuất thông tin. Hiệp hội Máy tính, Inc, 2016.
- [8] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera và Jerry Alan Fails. Tìm kiếm và học trực tuyến: Yum và các công cụ tìm kiếm khác dành cho trẻ em và giáo viên. Tạp chí Truy xuất thông tin, 20(5):524-545, 2017.
- [9] Dania Bilal. Trẻ em sử dụng công cụ tìm kiếm web yahooligans!. iii. hành vi nhận thức và thể chất trên các nhiệm vụ tìm kiếm hoàn toàn tự tạo. Tạp chí của Hiệp hội khoa học và công nghệ thông tin Hoa Kỳ, 53(13):1170-1183, 2002.



- [10] Dania Bilal và Li-Min Huang. Khả năng đọc và độ phức tạp của từ trong các đoạn trích serps và các trang web trên các truy vấn tìm kiếm của trẻ em: Google so với bing. Tạp chí Quản lý thông tin Aslib, 71(2):241-259, 2019.
- [11] Steven Bird, Ewan Klein và Edward Loper. Xử lý ngôn ngữ tự nhiên bằng Python: phân tích văn bản bằng bộ công cụ ngôn ngữ tự nhiên. " O'Reilly Media, Inc.", 2009.
- [12] Pamela J. Black, Melissa Wollis, Michael Woodworth và Jeffrey T. Hancock.  
Phân tích ngôn ngữ về các chiến lược dự dõ của tội phạm tình dục trẻ em trực tuyến: Ý nghĩa đối với sự hiểu biết của chúng ta về hành vi tình dục sẵn mồi trong một thế giới ngày càng được trung gian hóa bằng máy tính. Lạm dụng và bỏ bê trẻ em, 44:140-149, 2015.
- [13] Elana Broch. Công cụ tìm kiếm của trẻ em từ một quá trình tìm kiếm thông tin quan điểm. Nghiên cứu phương tiện truyền thông thư viện trường học, 3, 2000.
- [14] Elana Broch. Công cụ tìm kiếm của trẻ em từ một quá trình tìm kiếm thông tin quan điểm. Nghiên cứu phương tiện truyền thông thư viện trường học, 3, 2000.
- [15] Stephanie Carretero, Riina Vuorikari và Yves Punie. DigComp 2.1: Khung năng lực số cho công dân với tám cấp độ thành thạo và ví dụ sử dụng. Tháng 5 năm 2017.
- 、
- [16] Oscar Celma. Vấn đề đề xuất, trang 15-41. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [17] Davide Chicco và Giuseppe Jurman. Ưu điểm của hệ số tương quan matthews (mcc) so với điểm f1 và độ chính xác trong đánh giá phân loại nhị phân. BMC genomics, 21(1):1-13, 2020.
- [18] JD Cooper và ND Kiger. Đánh giá trình độ đọc viết: Giúp giáo viên lập kế hoạch Hướng dẫn. Công ty Houghton Mifflin, 2001.
- [19] Michael A Covington, Congzhou He, Cati Brown, Lorina Naci và John Brown.  
Câu đó phức tạp đến mức nào? một bản sửa đổi đề xuất của thang điểm cấp độ d của Rosenberg và Abbeduto. 2006.
- [20] Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright và Casey Kennington. Kidspell: Một trình kiểm tra chính tả ngữ âm, dựa trên quy tắc, hướng đến trẻ em. Trong Biên bản báo cáo của Hội nghị đánh giá và tài nguyên ngôn ngữ lần thứ 12, trang 6937-6946, 2020.
- [21] Nevena Dragovic, Ion Madrazo Azpiazu và Maria Soledad Pera. "is sven seven?": Một mô-đun ý định tìm kiếm dành cho trẻ em. Trong Biên bản Hội nghị quốc tế lần thứ 39 của ACM SIGIR về Nghiên cứu và Phát triển trong Truy xuất thông tin, trang 885-888, New York, NY, Hoa Kỳ, 2016. ACM.

- [22] Huizhong Duan và Bo-June Hsu. Sửa lỗi chính tả trực tuyến để hoàn thành truy vấn. Trong Biên bản báo cáo của hội nghị quốc tế lần thứ 20 về World wide web, trang 117-126, 2011.
- [23] Sergio Duarte Torres, Djoerd Hiemstra và Pavel Serdyukov. Phân tích các truy vấn nhằm tìm kiếm thông tin cho trẻ em. Trong Biên bản báo cáo của hội thảo chuyên đề thứ ba về Tương tác thông tin trong bối cảnh, trang 235-244. ACM, 2010.
- [24] Sergio Duarte Torres và Ingmar Weber. Trẻ em tìm kiếm gì và như thế nào trên web. Trong Biên bản báo cáo hội nghị quốc tế ACM lần thứ 20 về Quản lý thông tin và kiến thức, trang 393-402, 2011.
- [25] Sergio Duarte Torres, Ingmar Weber và Djoerd Hiemstra. Phân tích hành vi tìm kiếm và duyệt web của người dùng trẻ trên web. ACM Transactions on the Web (TWEB), 8(2):7, 2014.
- [26] Michael D Ekstrand. Những thách thức trong việc đánh giá các khuyến nghị cho trẻ em. Trong Hội thảo quốc tế về trẻ em và hệ thống khuyến nghị, 2017.
- [27] Jerry Alan Fails, Maria Soledad Pera, Oghenemaro Anuyah, Casey Kenning-ton, Katherine Landau Wright và William Bigirimana. Trợ giúp xây dựng câu hỏi cho trẻ em: Những gì có sẵn, khi nào cần trợ giúp –& trẻ em muốn gì. Trong Biên bản báo cáo Hội nghị quốc tế lần thứ 18 của ACM về Thiết kế tương tác và trẻ em, IDC '19, trang 109-120, New York, NY, Hoa Kỳ, 2019. Hiệp hội máy tính.
- [28] Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub và Hilary Hutchinson. Vai trò tìm kiếm của trẻ em tại nhà: Ý nghĩa đối với các nhà thiết kế, nhà nghiên cứu, nhà giáo dục và phụ huynh. Tạp chí của Hiệp hội Khoa học và Công nghệ Thông tin Hoa Kỳ, 63(3):558-573, 2012.
- [29] Daniel Gayo-Avello. Một cuộc khảo sát về các phương pháp phát hiện phiên trong nhật ký truy vấn và đề xuất cho đánh giá trong tương lai. Khoa học thông tin, 179(12):1822-1843, 2009.
- [30] Tatiana Gossen. Công cụ tìm kiếm dành cho trẻ em: giao diện người dùng tìm kiếm và hành vi tìm kiếm thông tin. Springer, 2016.
- [31] Tatiana Gossen, Julia Hempel và Andreas Nurnberger. Tìm nếu bạn có thể: nghiên cứu trường hợp về khả năng sử dụng của công cụ tìm kiếm dành cho người dùng trẻ. Máy tính cá nhân và phổ biến, 17(8):1593-1603, 2013.
- [32] Tatiana Gossen, Michael Kotzyba và Andreas Nurnberger. Triển lãm hành trình kiến thức: Hướng tới giao diện người dùng tìm kiếm thích ứng với độ tuổi. Trong Những tiến bộ trong truy xuất thông tin: Hội nghị châu Âu lần thứ 37 về nghiên cứu IR, trang 781-784, 03 2015.

- [33] Aaron Halfaker, Os Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage và Morten Warncke-Wang. Nhận dạng phiên người dùng dựa trên các quy luật mạnh mẽ trong thời gian tương tác. Trong Biên bản Hội nghị quốc tế lần thứ 24 về World Wide Web, trang 410-418, 2015.
- [34] Hyejung Han. Hành vi tìm kiếm sự trợ giúp của trẻ em và tác động của kiến thức chuyên môn trong việc sử dụng google và kids.gov: Các giai đoạn xây dựng truy vấn và đánh giá kết quả. Nghiên cứu khoa học thư viện và thông tin, 40(3-4):208-218, 2018.
- [35] Arturo E Hernandez và Ping Li. Độ tuổi tiếp thu: cơ chế thần kinh và tính toán của nó. Bản tin tâm lý, 133(4):638, 2007.
- [36] Jeff Huang. Nhật ký truy vấn Aol. [https://jeffhuang.com/search\\_query\\_logs.html](https://jeffhuang.com/search_query_logs.html), 2006.
- [37] Giacomo Inches và Fabio Crestani. Tổng quan về cuộc thi nhận dạng kẻ săn mồi tình dục quốc tế tại pan-2012. Trong CLEF (Ghi chú làm việc trực tuyến/phòng thí nghiệm/hội thảo), tập 30, 2012.
- [38] Sáng kiến Tiêu chuẩn Nhà nước cốt lõi chung. Sáng kiến tiêu chuẩn nhà nước cốt lõi chung tive. <http://www.corestandards.org/>, 2021.
- [39] Rosie Jones, Ravi Kumar, Bo Pang và Andrew Tomkins. "Tôi biết bạn đã làm gì vào mùa hè năm ngoái" - Nhật ký truy vấn và quyền riêng tư của người dùng. Trong Hội nghị quốc tế về quản lý thông tin và kiến thức, Biên bản báo cáo, trang 909-913, 2007.
- [40] April Kontostathis, Kelly Reynolds, Andy Garron và Lynne Edwards. Phát hiện bắt nạt trên mạng: Các thuật ngữ và kỹ thuật truy vấn. Trong Biên bản Hội nghị Khoa học Web ACM thường niên lần thứ 5, trang 195-204, New York, NY, Hoa Kỳ, 2013. ACM.
- [41] Barry M Kroll. Mối quan hệ phát triển giữa nói và viết. Khám phá mối quan hệ nói-viết: Kết nối và tương phản, trang 32-54, 1981.
- [42] Peter A Lachenbruch. Kiểm tra McNemar. Wiley StatsRef: Tài liệu tham khảo thống kê Trực tuyến, 2014.
- [43] Cheryl Laz. Hành động theo độ tuổi của bạn. Trong Diễn đàn xã hội học, tập 13, trang 85-113. Springer, 1998.
- [44] Geoffrey Leech, Paul Rayson, et al. Tần suất từ trong tiếng Anh viết và nói: Dựa trên British National Corpus. Routledge, 2014.

- [45] Andy Liaw, Matthew Wiener, et al. Phân loại và hồi quy bằng randomfor-est. *R news*, 2(3):18-22, 2002.
- [46] Jane Lin. Tự động lập hồ sơ tác giả của Nhật ký trò chuyện trực tuyến. Luận văn thạc sĩ, Trường Sau đại học Hải quân, Monterey, California, 2007.
- [47] Xiaofei Lu. Mối quan hệ giữa sự phong phú về từ vựng với chất lượng các bài tường thuật bằng miệng của người học ESL. *Tạp chí Ngôn ngữ Hiện đại*, 96(2):190-208, 2012.
- [48] Ion Madrazo. Hướng tới đánh giá khả năng đọc đa mục đích. Luận văn thạc sĩ, Đại học Boise State, Boise, Idaho, 2016.
- [49] Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah và Maria Soledad Pera. Bạn đang tìm bộ phim bảy hoặc bảy từ bộ phim Frozen? một chiến lược đa góc nhìn để đề xuất các truy vấn cho trẻ em.  
Trong *Biên bản Hội nghị năm 2018 về Tương tác và Truy xuất thông tin của con người*, trang 92-101, New York, NY, Hoa Kỳ, 2018. ACM.
- [50] James Marquardt, Sergio Davalos, Tacoma Washington, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, Martine De Cock và Tacoma Washington. Nhận dạng độ tuổi và giới tính trên phương tiện truyền thông xã hội. *Biên bản báo cáo của Phòng thí nghiệm đánh giá CLEF 2014*, trang 1129-1136, 2014.
- [51] Ali Nemati. Dự đoán giới tính và độ tuổi của hồ sơ tác giả đa ngôn ngữ dựa trên các bình luận. Trong *FIRE (Ghi chú làm việc)*, trang 232-239, 2018.
- [52] Felicity O'Dell, John Read, Michael McCarthy, et al. Đánh giá vốn từ vựng.  
Nhà xuất bản Đại học Cambridge, 2000.
- [53] Greg Pass, Abdur Chowdhury và Cayley Torgeson. Một bức ảnh tìm kiếm. Trong *Biên bản báo cáo của hội nghị quốc tế lần thứ nhất về Hệ thống thông tin có thể mở rộng*, trang 1-es, 2006.
- [54] Nick Pendar. Hướng tới việc phát hiện ra kẻ ấu dâm phân biệt nạn nhân với kẻ săn mồi trong các cuộc trò chuyện văn bản. Trong *Hội nghị quốc tế về máy tính ngữ nghĩa ICSC 2007*, trang 235-241, 2007.
- [55] Avar Pentel. Dự đoán tuổi và giới tính bằng động lực gõ phím và kiểu chuột. Trong *Ấn phẩm bổ sung của Hội nghị lần thứ 25 về Mô hình hóa, Thích ứng và Cá nhân hóa Người dùng*, trang 381-385. ACM, 2017.
- [56] J. Piaget. Nguồn gốc của trí thông minh ở trẻ em. Thư viện Norton, N202.  
Norton, 1963.

- [57] Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar và Mustafa Bilgic. Học tập chủ động: nghiên cứu thực nghiệm về các đường cơ sở chung. Khai thác dữ liệu và khám phá kiến thức, 31(2):287-313, 2017.
- [58] Pablo E. Román, Robert F. Dell, Juan D. Velásquez và Pablo S. Loyola. Xác định phiên người dùng từ nhật ký máy chủ web bằng lập trình số nguyên. Phân tích dữ liệu thông minh, 18(1):43-61, 2014.
- [59] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Field-house, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski và Carol Tenopir. Thế hệ Google: hành vi thông tin của nhà nghiên cứu trong tương lai. Trong biên bản Aslib, trang 290-310. Emerald Group Publishing Limited, 2008.
- [60] Sophie Rutter, Nigel Ford và Paul Clough. Trẻ em xây dựng lại truy vấn tìm kiếm của mình như thế nào?. Nghiên cứu thông tin: Tạp chí điện tử quốc tế, 20(1):n1, 2015.
- [61] Sara Salehi, Jia Tina Du và Helen Ashman. Khảo sát cá nhân hóa trong tìm kiếm trên web học thuật. Trong Biên bản báo cáo Hội nghị ACM lần thứ 26 về Siêu văn bản và Phương tiện truyền thông xã hội, trang 103-111. ACM, 2015.
- [62] K Santosh, Romil Bansal, Mihir Shekhar và Vasudeva Varma. Hồ sơ tác giả: Dự đoán tuổi và giới tính từ blog. Phòng thí nghiệm đánh giá và hội thảo CLEF 2013 - Bài báo ghi chú làm việc, 2013.
- [63] Katherine Schaeffer. Khám phá các công cụ nghiên cứu để nghiên cứu văn bản. <https://www.pewresearch.org/fact-tank/2019/08/23/most-us-teens-who-use-cellphones-do-it-to-pass-time-connect-with-others-learn-new-things/>, 2019. (truy cập ngày 09-09-2019).
- [64] Scholastic. Chào đón Internet vào lớp học của bạn. có tại: <https://www.scholastic.com/teachers/articles/teaching-content/coming-internet-your-classroom/>, 2018. (truy cập ngày 11 tháng 3 năm 2018).
- [65] Martin EP Seligman, David Stillwell, Lukasz Dziurzynski, H. Andrew Schwartz, Megha Agrawal, Achal Shah, Lyle H. Ungar, Stephanie M. Ramones, Margaret L. Kern, Michal Kosinski và Johannes C. Eichstaedt. Tính cách, giới tính và độ tuổi trong ngôn ngữ truyền thông xã hội: Phương pháp tiếp cận từ vũng mở. Thư viện công cộng khoa học ONE, 8(9):e73791, 2013.
- [66] Luke Sloan, Jeffrey Morgan, Pete Burnap và Matthew Williams. Ai là người dùng Twitter? trích xuất các đặc điểm nhân khẩu học về độ tuổi, nghề nghiệp và giai cấp xã hội từ siêu dữ liệu người dùng Twitter. Thư viện Khoa học Công cộng ONE, 10(3):1-20, 2015.

- [67] Jenny Tam và Craig H. Martell. Phát hiện tuổi trong trò chuyện. ICSC 2009 - Hội nghị quốc tế IEEE về máy tính ngữ nghĩa năm 2009, trang 33-39, 2009.
- [68] Edson RD Weren, Viviane Pereira Moreira và José Palazzo Moreira de Oliveira. Sử dụng các tính năng nội dung đơn giản cho sổ ghi chép nhiệm vụ lập hồ sơ tác giả cho pan tại clef 2013. Trong CLEF 2013 Phòng thí nghiệm đánh giá và Hội thảo - Bài báo ghi chú làm việc, 2013.
- [69] Ryen W. White, Susan T. Dumais và Jaime Teevan. Đặc trưng hóa ảnh hưởng của chuyên môn về lĩnh vực đối với hành vi tìm kiếm trên web. Trong Biên bản báo cáo Hội nghị quốc tế ACM lần thứ hai về tìm kiếm trên web và khai thác dữ liệu, WSDM '09, trang 132-141, New York, NY, Hoa Kỳ, 2009. ACM.
- [70] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes và Stefan Dietze. Dự đoán mức độ hiểu biết của người dùng trong các phiên tìm kiếm thông tin. Trong Hội nghị quốc tế lần thứ 41 của ACM SIGIR về Nghiên cứu và Phát triển trong Truy xuất thông tin, SIGIR 2018, trang 75-84. ACM, 6 2018.
- [71] Xiangmin Zhang, Michael Cole và Nicholas Belkin. Dự đoán kiến thức miền của người dùng từ hành vi tìm kiếm. Trong Biên bản báo cáo của hội nghị quốc tế ACM SIGIR lần thứ 34 về Nghiên cứu và phát triển trong Truy xuất thông tin, trang 1225-1226, 2011.
- [72] Xiangmin Zhang, Jingjing Liu, Michael Cole và Nicholas Belkin. Dự đoán kiến thức miền của người dùng trong việc truy xuất thông tin bằng cách sử dụng phân tích hồi quy bội của hành vi tìm kiếm. Tạp chí của Hiệp hội Khoa học và Công nghệ Thông tin, 66(5):980-1000, 2015.

