TÌM KIẾM CHUYÊN GIA VỚI GIẢI THUẬT MÁY HỌC C4.4-*k*NN

Văn Thị Xuân Hồng[1] *và Đỗ Thanh Nghị*[2]

ABSTRACT

In this paper, we investigate a learning to rank model called C4.4-kNN for searching experts. This model is based on the bag of words model and also uses the C4.4 algorithm (well-known as a good ranking algorithm) and the k nearest neighbors algorithm (considered as the simplest instance-based learning). In addition, the model also takes into account user-relevance-feedback to improve ranking tasks. The numerical test results on the French speaking data mining conference (EGC) showed that our C4.4-kNN is better than kNN for the assignment task. C4.4-kNN proposes appropriate program committee members for a given paper abstract after a few of clickthrough experts.

Keywords: Experts search, Learning to rank, Bag of words, k nearest neighbors, C4.4 machine learning algorithm

Title: Searching Experts with C4.4-kNN Machine Learning Algorithm

TÓM TẮT

Trong bài viết này chúng tôi đưa ra hướng tiếp cận học xếp hạng cho vấn đề tìm kiếm chuyên gia. Cơ sở dữ liệu chuyên gia được tạo ra từ các tóm tắt bài báo của các chuyên gia trong những năm gần đây. Sau khi tiền xử lý và biểu diễn theo mô hình túi từ. Chúng tôi đã đề xuất tiếp cận học xếp hạng C4.4-kNN dựa trên cây quyết định C4.4 kết hợp với thuật toán k láng giềng kNN có sử dụng phản hồi kết quả của người dùng. Kết quả thực nghiệm từ 87 chuyên gia của hội đồng xét duyệt bài báo của hội thảo khai mỏ dữ liệu cho thấy cách tiếp cận của chúng tôi C4.4-kNN tìm được các chuyên gia để xét duyệt bài báo phù hợp hơn so với chỉ sử dụng giải thuật kNN. Chúng tôi cũng thử nghiệm trên mô hình RF-C4.4-kNN dựa trên rừng cây quyết định C4.4 và kNN cho kết quả tốt hơn so với chỉ sử dụng một cây quyết định như C4.4-kNN.

Từ khóa: Tìm kiếm chuyên gia, học để xếp hạng, mô hình túi từ, k láng giềng, máy học cây quyết định C4.4

1 GIỚI THIỆU

Trong thực tiễn, vấn đề thường đặt ra với nhiều cộng đồng khoa học, các hội thảo, các ban chương trình hay nhóm chuyên gia của một lĩnh vực nào đó, là làm sao để tìm kiếm một hay những chuyên gia liên quan đến chuyên ngành hẹp để có thể đánh giá một đề tài, một dự án, một bài báo một cách có hiệu quả. Ví dụ như ở một hội thảo chuyên ngành khai mỏ dữ liệu, chúng ta đã có các thành viên trong ban chương trình (được gọi là các chuyên gia của nhiều chuyên ngành hẹp của hội thảo về máy học, phân tích dữ liệu, …). Khi có một bài báo gửi đến hội thảo, làm sao ban tổ chức hội thảo có thể chuyển bài báo này đến chuyên gia nào trong ban chương trình để có thể nhận được đánh giá chuẩn xác về bài báo. Hay một sở khoa học công nghệ nhận được một dự án đề xuất, làm sao để gửi dự án đó đến chuyên gia có thể thẩm định tốt về đề xuất. Vấn đề này có thể được giải quyết theo tìm

[1] Trung tâm Công nghệ Phần mềm, Khoa CNTT&TT, Trường Đại học Cần Thơ
[2] Bộ môn Khoa Học Máy Tính, Khoa CNTT&TT, Trường Đại học Cần Thơ

kiếm thông tin (Manning *et al*., 2009). Một nghiên cứu liên quan đến vấn đề tìm kiếm chuyên gia dựa trên phương pháp hiển thị trực quan cũng được tìm thấy trong (Fortuna et al., 2005). Thời gian gần đây, các nghiên cứu được đề cập trong (Agarwal *et al*., 2005), (Radlinski & Joachims, 2007), (Liu, 2009) đưa ra nhiều mô hình máy học xếp hạng có sử dụng phản hồi từ người sử dụng nhằm cải thiện được độ chính xác cho tìm kiếm thông tin.

Để giải quyết cho bài toán tìm kiếm chuyên gia, chúng tôi đề xuất mô hình theo hướng tiếp cận học để xếp hạng. Trước tiên, một cơ sở dữ liệu chuyên gia được tạo thành từ mô tả về chuyên ngành, chuyên môn, các lý lịch khoa học, tóm tắt bài báo khoa học của các chuyên gia. Chúng tôi sử dụng mô hình túi từ để biểu diễn cơ sở dữ liệu chuyên gia thuận lợi cho quá trình tìm kiếm. Sau đó, khi có một tóm tắt bài báo, hay dự án được yêu cầu, hệ thống trước hết sẽ sử dụng phương pháp tìm kiếm k láng giềng (*k*NN (Fix & Hodges, 1952)) để đưa ra các chuyên gia gần với yêu cầu. Sau đó người sử dụng có thể xác định những câu trả lời nào là gần giống với yêu cầu nhất từ các kết quả trả về. Hệ thống sẽ bắt đầu quá trình học có giám sát của cây quyết định cho xếp hạng C4.4 (Provost & Domingos, 2003) với lớp dương (+1) là các kết quả vừa được người sử dụng xác nhận và lớp âm (-1) là các dữ liệu còn lại. Tiến trình cứ lặp lại cho đến khi nào người sử dụng thấy kết quả tìm kiếm phù hợp với yêu cầu. Kết quả thực nghiệm từ 87 chuyên gia của hội đồng xét duyệt bài báo của hội thảo khai mỏ dữ liệu EGC của khối pháp ngữ cho thấy cách tiếp cận của chúng tôi C4.4-*k*NN tìm được các chuyên gia để xét duyệt bài báo phù hợp hơn so với chỉ sử dụng giải thuật *k*NN (chỉ với khoảng 3 lần lặp). Chúng tôi cũng thử nghiệm trên mô hình RF-C4.4-*k*NN dựa trên rừng cây quyết định C4.4 và *k*NN cho kết quả tốt hơn so với chỉ sử dụng một cây quyết định như C4.4-*k*NN.

Phần tiếp theo của bài viết được tổ chức như sau. Phần 2 sẽ trình bày toàn bộ tiếp cận học xếp hạng C4.4-kNN cho tìm kiếm chuyên gia. Phần 3 trình bày các kết quả thực nghiệm trước khi kết luận và hướng phát triển.

2 TIẾP CẬN HỌC XẾP HẠNG C4.4-*k*NN CHO TÌM KIẾM CHUYÊN GIA

Trong tiếp cận học xếp hạng C4.4-kNN mà chúng tôi đề xuất, trước tiên cần phải tạo tập dữ liệu chuyên gia. Trước tiên, chúng tôi sưu tập các tóm tắt bài báo từ thư viện trực tuyến **DBLP** của các chuyên gia thuộc ban chương trình của hội thảo khai mỏ dữ liệu EGC của khối pháp ngữ. Chúng tôi sử dụng các tóm tắt bài báo của 87 chuyên gia (theo đề xuất của (Fortuna *et al*., 2005)), các bài báo của mỗi chuyên gia được xem là thông tin về lĩnh vực nghiên cứu của chuyên gia đó. Cơ sở dữ liệu bao gồm các đoạn văn bản phi cấu trúc, chúng tôi cần biểu diễn thành dạng bảng có cấu trúc để có thể thực hiện việc tìm kiếm chuyên gia.

2.1 Biểu diễn cơ sở dữ liệu chuyên gia với mô hình túi từ

Trong các ứng dụng về phân loại văn bản hay tìm kiếm thông tin, các dữ liệu phi cấu trúc có thể được chuyển về dạng có cấu trúc nhờ vào áp dụng mô hình túi từ. Bước tiền xử lý bao gồm phân tích từ vựng và tách các từ trong nội dung của các văn bản (tóm tắt bài báo). Sau đó chọn tập hợp các từ mà có thể dùng để tìm kiếm. Tiếp theo, tóm tắt bài báo của các một chuyên gia được biểu diễn bằng một véctơ tần số của các từ trong tóm tắt đó. Véctơ này được xem như một phần tử trong tập

dữ liệu, để làm được điều này, chúng tôi sử dụng thư viện Bow (McCallum, 1998) để tách từ và chuyển dữ liệu về với dạng bảng, gồm hai bước sau:

- Xây dựng mô hình tách từ của các tóm tắt bài báo của chuyên gia. Ở bước này chúng ta thu được mô hình gồm có 9441 từ đã bỏ qua các từ có ít ý nghĩa trong các tóm tắt, chẳng hạn như mạo từ, giới từ.

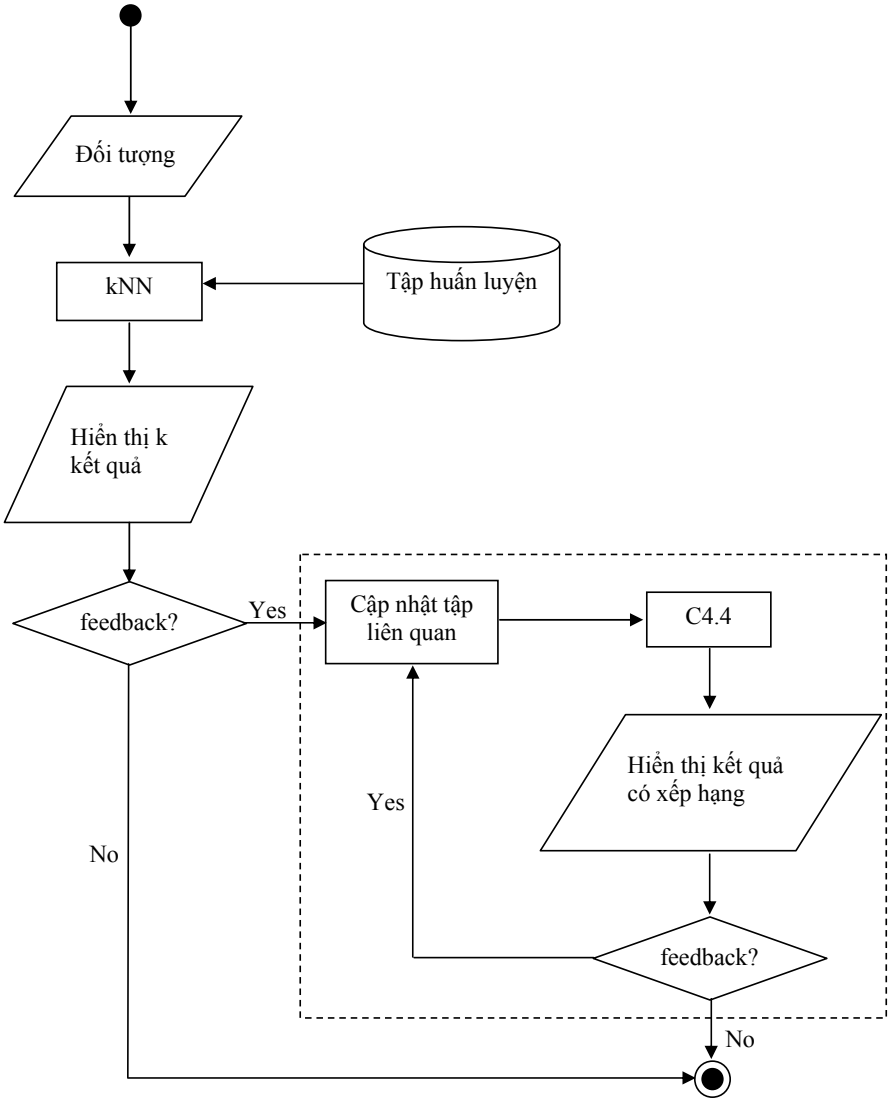
- Dựa trên mô hình tách từ của Bow vừa xây dựng, chúng tôi biểu diễn tóm tắt bài báo của chuyên gia về mô hình túi từ bằng cách tính tần số xuất hiện của các từ đưa về một bảng dữ liệu. Với mô hình túi từ, chúng tôi thu được bảng dữ liệu có 87 dòng (mỗi dòng tương ứng với một chuyên gia) và 9441 thuộc tính (mỗi thuộc tính tương ứng với một từ, giá trị mỗi thuộc tính là tần số xuất hiện của từ trong tóm tắt bài báo của chuyên gia).

Qua bước tiền xử lý dữ liệu, cơ sở dữ liệu chuyên gia được biểu diễn về dạng bảng thuận lợi cho quá trình tìm kiếm với tiếp cận C4.4-*k*NN.

2.2 Mô hình học xếp hạng C4.4-*k*NN

Cơ chế hoạt động của mô hình được mô tả như sau. Khi có tóm tắt bài báo yêu cầu được đánh giá hệ thống sẽ dùng thuật toán k láng giềng (*k*NN) để đưa ra *k* chuyên gia đầu tiên có khoảng cách gần với tóm tắt của bài báo cần xét duyệt. Tiếp đến, hệ thống sẽ nhận phản hồi từ người sử dụng về tính liên quan của các kết quả này. Người sử dụng chỉ cần xác nhận các chuyên gia nào trong số *k* chuyên gia trả về là gần với chuyên môn của bài báo. Hệ thống bắt đầu thực hiện bước lặp học cho xếp hạng với thuật toán C4.4 (được mô tả trong phần tiếp theo). Các kết quả được người sử dụng xác nhận được gán nhãn (+1) hay lớp dương và các kết quả không được chọn sẽ được gán nhãn là (-1) hay lớp âm. Lúc này hệ thống sẽ cập nhật lại tập mẫu huấn luyện. Thực hiện việc học xếp hạng dựa trên ước lượng xác suất của thuật toán C4.4. Nếu người sử dụng chưa thấy hài lòng với kết quả thì họ tiếp tục phản hồi để hệ thống cập nhật lại tập huấn luyện và học xếp hạng để cho ra cải thiện kết quả xếp hạng tốt hơn sau khi học. Tiến trình học cứ tiếp tục cho đến khi người dùng cảm thấy hài lòng.

Lưu đồ của hệ thống học xếp hạng dựa trên mô hình đề xuất C4.4-*k*NN để giải quyết vấn đề tìm kiếm thông tin được thể hiện ở hình 1.



Hình 1: Mô hình học xếp hạng C4.4-*k*NN cho tìm kiếm chuyên gia

2.3 Ước lượng xác suất trên cây quyết định C4.4

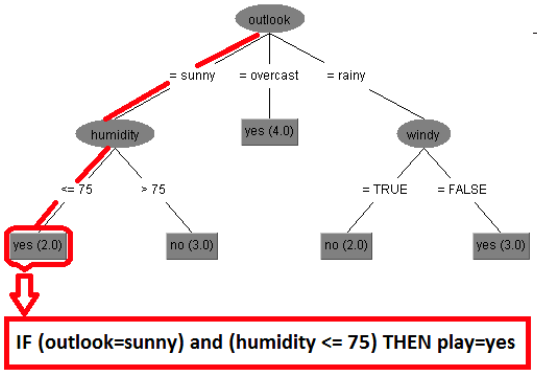
Mô hình cây quyết định C4.5 của (Quinlan, 1993) được biết đến như là giải thuật quan trọng của khai mỏ dữ liệu. Mô hình học của cây quyết định đơn giản, nhanh, cũng cho kết quả tốt. Điều đặc biệt quan trọng là giải thuật cây quyết định C4.4 (Provost & Domingos, 2003) cho ước lượng xác suất (xếp hạng) tốt hơn Bayes thơ ngây (Good, 1965), SVM (Vapnik, 1995) hay *k*NN (Fix & Hodges, 1952). Giải thuật cây quyết định có thể xử lý được cả kiểu dữ liệu rời rạc và liên tục. Chính vì lý do đó, chúng tôi đề nghị chọn mô hình cây quyết định C4.4 trong trong hệ thống học xếp hạng cho tìm kiếm chuyên gia.

Mô hình cây quyết định có cấu trúc dạng cây mà ở đó:

- Nút lá được gán nhãn tương ứng với lớp của dữ liệu,

- Nút trong được tích hợp với điều kiện kiểm tra để rẽ nhánh.

Ví dụ mô hình cây quyết định trong hình 2 được xây dựng từ việc học trên tập dữ liệu **weather** để dự báo chơi hay không chơi golf (**yes** hay **no**) dựa trên các thuộc tính **outlook**, **temperature**, **humidity** và **windy**. Mô hình rất dễ hiểu bởi vì chúng ta có thể rút trích luật quyết định tương ứng với nút lá có dạng **IF-THEN** được tạo ra từ việc thực hiện **AND** trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.



outlook	temp.	hum.	windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .

Hình 2: Cây quyết định cho tập dữ liệu weather

Giải thuật C4.4 xây dựng cây quyết định không cắt tỉa nhằm nâng cao độ chính xác và sử dụng ước lượng Laplace để làm mịn ước lượng xác suất ở nút lá của cây.

Chẳng hạn, xét nút lá *90%* dữ liệu thuộc về lớp dương. Một mẫu bất kỳ rơi vào nút lá này sẽ được gán với xác suất là 0.9 thuộc về lớp dương. Vấn đề tiềm ẩn với phương pháp ước lượng xác suất là nếu một lá bao gồm 5 mẫu và tất cả đều là lớp dương thì xác suất ước lượng sẽ là 1.0. Trong khi đó 5 mẫu không đủ để khẳng định mạnh như thế. Vấn đề này có thể giải quyết bằng việc làm mịn ước lượng xác suất để giá trị ít cực đại hơn.

Giả sử có k mẫu của lớp tại nút lá, N là tổng số các mẫu tại nút lá, C là tổng số lớp. Ước lượng Laplace được tính bằng công thức $\frac{k+1}{N+C}$.

3 KẾT QUẢ THỰC NGHIỆM

3.1 Mô tả thực nghiệm

Để kiểm tra hiệu quả của hệ thống tìm kiếm chuyên gia, chúng tôi cài đặt chương trình bằng ngôn ngữ TCL/TK, có sử dụng rainbow (McCallum, 1998) để biểu diễn dữ liệu và tìm kiếm các chuyên gia theo mô hình túi từ và *k*NN. Chúng tôi viết mã chương trình cho giải thuật C4.4 dựa trên nguồn C4.5 (Quinlan, 1993). Sau đó, hệ thống được vận hành trên hệ điều hành Linux (Ubuntu10.04).

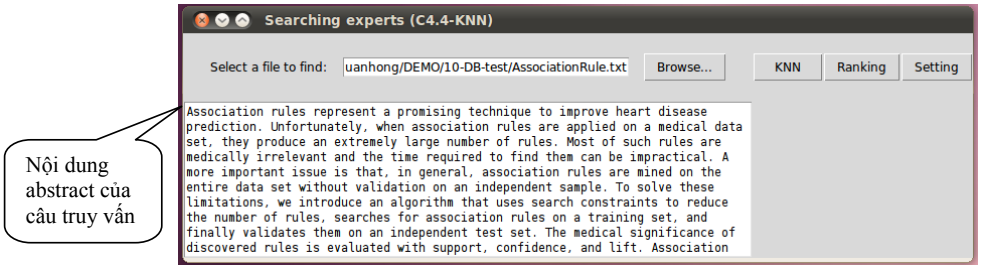
Như đã trình bày ở phần 2, chúng tôi tạo được cơ sở dữ liệu gồm 87 chuyên gia thuộc ban chương trình của hội thảo khai mỏ dữ liệu EGC khối pháp ngữ. Chúng tôi tiến hành kiểm thử 10 tóm tắt bài báo của 10 tác giả khác nhau làm các câu truy vấn cho hệ thốngdược mô tả như ở bảng 1.

Bảng 1: Danh sách 10 bài báo được lấy tóm tắt làm câu truy vấn

ID	Tiêu đề câu truy vấn
1	Ordonez C. (2006), "Association rule discovery with the train and test approach for heart disease prediction", *Information Technology in Biomedicine*, vol. 10(2), pp. 334-343.
2	Chaiyaratana N., Zalzala A.M.S. (1997), "Recent developments in evolutionary and genetic algorithms: theory and applications", *Genetic Algorithms in Engineering Systems: Innovations and Applications 1997, GALESIA 97*, pp: 270-277.
3	Gao B. J. , Ester M. (2006), "Cluster Description Formats, Problems and Algorithms", *SDM 2006.*
4	Silva A., Lechevallier Y., Carvalho F. (2007), "Analyzing Distance Measures for Symbolic Data Based on Fuzzy Clustering", *Intelligent Systems Design and Applications 2007, ISDA 2007*, pp: 109-114.
5	Abdelhalim A., Traore I. (2009), "A New Method for Learning Decision Trees from Rules", *Machine Learning and Applications 2009, ICMLA '09,* pp: 693-698.
6	Chen B., Hoberock L.L. (1996), "A fuzzy neural network architecture for fuzzy control and classification", *Neural Networks 1996.*, pp: 1168-1173 vol.2.
7	Abidin S.Z.Z., Idris N.M., Husain A.H. (2010), "Extraction and classification of unstructured data in WebPages for structured multimedia database via XML", *Information Retrieval & Knowledge Management, (CAMP), 2010*, pp: 44-49.
8	Ponmary Pushpa Latha D., Raj D.J.P., Sharmila D.J.S. (2007), "Generation of unified data structure and data warehouse for protein data banks", *Conference on Computational Intelligence and Multimedia Applications, 2007*, vol. 2, pp: 3-7.
9	Lamson B.G., Dimsdale B. (1996), "A natural language information Retrieval system", Proceedings of the IEEE , 54(12), pp: 1636-1640.
10	Kovalerchuk B. (2001), "Visualization and Decision-Making using Structural Information". *Conference on Imaging Science, Systems, and Technology (CISST'2001, June 25-28, 2001)*, Las Vegas, pp. 478-484.

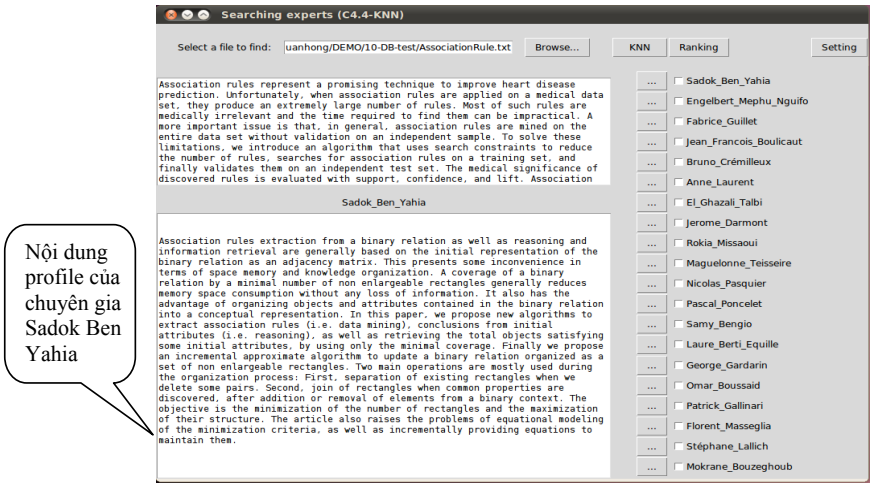
*** Đánh giá thực hiện lần lượt trên 10 câu truy vấn với các bước:**

Bước 1: Chọn câu truy vấn theo chủ đề trong tập kiểm thử gồm 10 tóm tắt bằng cách nhấn nút "Browse…"



Hình 3: Khi chọn câu truy vấn có ID = 1

Bước 2: Chọn nút "kNN" cho ra *k* kết quả gần nhất với câu truy vấn.



Hình 4: Hiển thị 20 kết quả gần với câu truy vấn có ID = 1 (theo *k*NN)

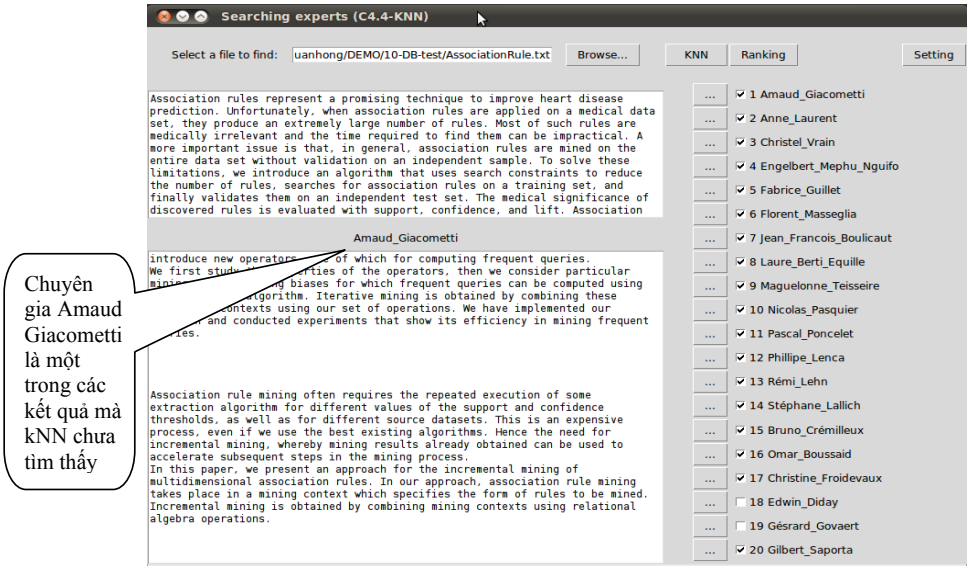
Bước 3: Dựa vào kết quả phán quyết tiến hành chọn những tác giả nào được gán nhãn là có liên quan từ danh sách *k* kết quả.

Bước 4: Sau khi phản hồi về kết quả, tiến hành quá trình học của thuật toán C4.4 để hệ thống cho ra danh sách *k* kết quả mới được xếp hạng. bằng cách nhấn vào nút "Ranking".

Bước 5: Từ danh sách *k* kết quả mới này, tiếp tục phản hồi bằng cách tra vào kết quả phán quyết tính liên quan của chuyên gia.

Bước 6: Lặp lại bước 4.

Bước 7: Lặp lại bước 5 cho đến khi danh sách *k* kết quả xếp hạng trả về từ hệ thống không có thay đổi. Hay mức độ tính liên quan của các kết quả trả về là tối đa.



Hình 5: Kết quả sau 3 lần lặp với câu truy vấn có ID = 1 (học C4.4)

3.2 Kết quả thực nghiệm

Bảng 2: So sánh kết quả của *k*NN và C4.4-*k*NN dựa trên Precision, Recall và F1

ID	Precision (%)		Recall (%)		F1 (%)	
	kNN	C4.4-kNN	kNN	C4.4-kNN	kNN	C4.4-kNN
1	70.00	90.00	60.87	78.26	65.12	83.72
2	45.00	75.00	50.00	83.33	47.37	78.95
3	55.00	100.00	42.31	76.92	47.83	86.96
4	45.00	100.00	36.00	80.00	40.00	88.89
5	45.00	70.00	47.37	73.68	46.15	71.79
6	50.00	85.00	55.56	94.44	52.63	89.47
7	65.00	100.00	43.33	66.67	52.00	80.00
8	50.00	75.00	58.82	88.24	54.05	81.08
9	40.00	95.00	28.57	67.86	33.33	79.17
10	35.00	100.00	28.00	80.00	31.11	88.89

Có nhiều phương pháp để tiến hành đánh giá hiệu suất của mô hình xếp hạng như Recall/Precision, F1, Precision@n, Precision trung bình (Mean Average Precision, MAP), độ lợi tích lũy giảm dần (Normalized Discounted Cumulative Gain, nDCG) (Sebastiani, 2002), (Liu, 2009). Chúng tôi tiến hành đo hiệu suất với hai phương diện là đánh giá chung và đánh giá thứ tự xếp hạng.

Tiêu chuẩn đánh giá chung

Ở bảng 2, thể hiện các chỉ số đánh giá dựa trên Precision, Recall và F_1-measure của giải thuật:

- *k*NN: cho ra *k* láng giềng gần với câu truy vấn nhất.
- C4.4-*k*NN: là mô hình đề xuất đã trình bày ở phần 2.

Từ kết quả ở bảng 2 khi so sánh trên ba tiêu chí Precision, Recall và F_1-measure cho thấy mô hình đề xuất C4.4-*k*NN cho kết quả tốt hơn nhiều so với *k*NN.

Tiêu chuẩn đánh giá thứ tự xếp hạng

Ở bảng 3, thể hiện các chỉ số dựa vào phép đo Precision@n của C4.4-*k*NN để đánh giá thứ tự xếp hạng của danh sách kết quả trả về ở top 5, top 10 và top 15. Từ kết quả ở bảng 3, với dòng in đậm cuối bảng là các giá trị trung bình, cho thấy các chỉ số Precision@5, Precision@10 và Precision@15 của C4.4-*k*NN đều cao với số lần lặp trung bình là 2.5 lần.

Bảng 3: Kết quả đánh giá C4.4-*k*NN dựa trên Precision@n tại top 5, 10 và 15

ID	Precision@5	Precision@10	Precision@15	Ghi chú (số lần lặp)
1	100.00	100.00	100.00	3
2	100.00	100.00	100.00	2
3	100.00	100.00	100.00	2
4	100.00	100.00	100.00	1
5	100.00	100.00	93.33	4
6	100.00	100.00	100.00	3
7	100.00	100.00	100.00	2
8	100.00	100.00	93.33	3
9	100.00	90.00	93.33	3
10	100.00	100.00	100.00	2
Avg	**100.00**	**99.00**	**98.00**	**2.5**

Với những kết quả đạt được như trên, có thể nói rằng hệ thống học xếp hạng dựa trên cây quyết định C4.4, C4.4-*k*NN, có thể ứng dụng hiệu quả cho hệ thống tìm kiếm chuyên gia.

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã trình bày hệ thống tìm kiếm chuyên gia với tiếp cận học xếp hạng C4.4-*k*NN. Cơ sở dữ liệu chuyên gia được tạo ra từ các tóm tắt bài báo của các chuyên gia trong những năm gần đây được tiền xử lý và biểu diễn theo mô hình túi từ. Chúng tôi đã đề xuất tiếp cận học xếp hạng C4.4-*k*NN dựa trên cây quyết định C4.4 kết hợp với thuật toán k láng giềng *k*NN có sử dụng phản hồi kết quả của người dùng. Kết quả thực nghiệm từ 87 chuyên gia của ban chương trình hội thảo khai mỏ dữ liệu EGC khối pháp ngữ cho thấy cách tiếp cận của chúng tôi C4.4-*k*NN tìm được các chuyên gia để xét duyệt bài báo phù hợp hơn so với chỉ sử dụng giải thuật *k*NN. Chúng tôi cũng thử nghiệm trên mô hình RF-C4.4-*k*NN dựa trên rừng cây quyết định C4.4 và *k*NN cho kết quả tốt hơn so với chỉ sử dụng một cây quyết định như C4.4-*k*NN.

Chúng tôi sẽ nghiên cứu thêm các mô hình cho phép hỗ trợ cho pha phản hồi từ phía người sử dụng để áp dụng được trong thực tế, chẳng hạn như: tìm kiếm chuyên gia để xét duyệt dự án, những chuyên gia có chuyên môn gần, hoặc tìm kiếm tài liệu học tập cho sinh viên.

TÀI LIỆU THAM KHẢO

Agarwal, S., Cortes, C. and Herbrich, R.: *Learning to Rank*. The workshop proceedings at NIPS'2005, 2005.

Fix, E. and Hodges, J.: Discriminatoiry Analysis: Small Sample Performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA, 1952.

Fortuna, B., Grobelnik, M. and Gunn, S.: PASCAL visualization challenge. 2005.

Good, I.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. *MIT Press*, 1965.

Liu, T-Y.: Learning to Rank for Information Retrieval. PO Box 1024 Hanover, MA 02339 USA, 2009.

McCallum, A.: Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. 1998. http://www-2.cs.cmu.edu/~mccallum/bow.

Manning, C. D., Raghavan, P. and Schutze, H.: *An Introduction to Information Retrieval*. Cambridge University Press Cambridge, 2009.

Provost, F. and Domingos, P.: Tree Induction for Probability-Based Ranking. *Machine Learning* 52(3):199-215, 2003.

Quinlan, J.: C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

Radlinski, F. and Joachims, T.: Active Exploration for Learning Rankings from Clickthrough Data. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2007.

Sebastiani, F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1):1-47, 2002.

Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.