ỨNG DỤNG KHAI PHÁ DỮ LIỆU NÂNG CAO DỊCH VỤ THƯ VIỆN SỐ

Nguyễn Thị Yên*

Tóm tắt: *Việc phát triển thư viện số đã trở thành một xu hướng mạnh mẽ trên thế giới và ở Việt Nam. Làm thế nào để sử dụng tài nguyên một cách hiệu quả để nâng cao chất lượng dịch vụ của thư viện số là vấn đề rất quan trọng. Bài viết mô tả các kỹ thuật khai phá dữ liệu, giới thiệu quy trình khai phá dữ liệu, nghiên cứu kỹ thuật phân cụm và luật kết hợp trong khai phá dữ liệu để phân tích dữ liệu người dùng, tạo ra hệ thống khuyến nghị nhằm nâng cao dịch vụ thư viện số. Các hồ sơ mượn sách của thư viện được kiểm tra và phân cụm theo một số đặc điểm của độc giả, sử dụng các quy tắc kết hợp làm kỹ thuật khai phá dữ liệu để khám phá những điểm tương đồng giữa sở thích của người dùng và hành vi mượn sách, xây dựng một dịch vụ giới thiệu cho người đọc để tìm kiếm sách từ Web và chủ động tìm kiếm những cuốn sách phù hợp nhất cho người đọc.*

Từ khóa: *Khai phá dữ liệu; Thư viện số; Phân cụm; Luật kết hợp.*

1. GIỚI THIỆU

Công nghệ ngày càng phát triển nhanh chóng, buộc con người và các lĩnh vực trong xã hội phải thay đổi, thích ứng, trong đó có hoạt động thư viện. Thư viện số ra đời là để nâng cấp chất lượng dịch vụ thư viện truyền thống bằng cách sử dụng tự động hóa thông tin và công nghệ mạng. Tuy nhiên, ngày nay nguồn thông tin trên internet ngày càng đa dạng và nguồn dữ liệu thông tin của thư viện số ngày càng tăng lên nhanh chóng.

* Thạc sĩ, Khoa Thông tin Thư viện, Đại học Văn hóa Hà Nội.

Trong nghiên cứu trước đây, hầu hết các nhà nghiên cứu đã phân tích nội dung của tài liệu số. Sau đó, họ cố gắng khám phá mối quan hệ giữa các tài liệu, cũng như giữa tài liệu và người dùng. Tuy nhiên, ngày càng có nhiều định dạng cho các ấn phẩm kỹ thuật số như âm thanh, video, hình ảnh,... Trong những trường hợp này, thật khó để phân tích các từ khóa hoặc nội dung của nó để tinh chỉnh thông tin đề xuất cho người dùng. Bài viết này trình bày cách thiết lập một hệ thống khuyến nghị dựa trên các phương pháp khai phá dữ liệu, đó là các quy tắc liên kết và phân cụm được áp dụng để khám phá những độc giả thích ứng với một cuốn sách. Đầu tiên, các hồ sơ mượn trong thư viện số được nhóm lại theo một số đặc điểm của độc giả. Cách tiếp cận được đề xuất sử dụng tính năng phân cụm tự động của Thuật toán phân cụm đàn kiến (Ant Colony Clustering Algorithm) để gom thành một nhóm người dùng có đặc điểm giống nhau. Sau đó, dựa trên độ hỗ trợ tối thiểu và độ tin cậy, liên kết các đối tượng để tạo ra các quy tắc đề xuất. Các quy tắc liên kết sẽ đánh giá sách nào mượn bởi độc giả trong cùng một cụm được sử dụng làm cơ sở giới thiệu cuốn sách tương tự. Cuối cùng, một hệ thống khuyến nghị trực tuyến tự động được đề xuất. Bài báo này không chỉ trình bày cách xây dựng một dịch vụ khuyến nghị cho người đọc trong tìm kiếm sách từ trang Web mà còn chủ động tìm sách phù hợp nhất cho người đọc.

2. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

2.1. Khai phá dữ liệu (KPDL) và khám phá tri thức

Khai phá dữ liệu *(Data mining)* là một khái niệm bao hàm nhiều kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn (các kho dữ liệu). Về bản chất, khai phá dữ liệu liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy trong kho dữ liệu lưu trữ [1],[4].

Khai phá dữ liệu là bước chính của quá trình khám phá tri thức trong cơ sở dữ liệu *(Knowledge Discovery in Database - KDD)*, quá trình này bao gồm các bước cơ bản sau:

- Xác định vùng đối tượng *(Determine area object)*: bước này có ý nghĩa quan trọng cho việc rút ra được các tri thức hữu ích và chọn

các phương pháp KPDL thích hợp sao cho phù hợp với mục đích ứng dụng và bản chất dữ liệu.

▪ Chuẩn bị dữ liệu *(Data preparation)*: Giai đoạn này có thể chia thành 3 bước:

- Chọn lọc dữ liệu *(Data selection):* Trong bước này, nó chỉ đơn giản là loại bỏ một số dữ liệu dư thừa hoặc không liên quan và trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses).

- Tiền xử lý dữ liệu *(Data preprocessing)*: Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,...), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu,...), rời rạc hoá dữ liệu (dựa vào histograms, entropy, phân khoảng,…). Phần lớn các cơ sở dữ liệu đều ít nhiều mang tính không nhất quán. Vì vậy khi gom dữ liệu rất có thể mắc một số lỗi như dữ liệu không đầy đủ, chặt chẽ và không lôgic (bị trùng lặp, giá trị bị sai lệch,...). Do đó cần phải được "tiền xử lý" trước khi khai phá dữ liệu nếu không sẽ gây nên những kết quả sai lệch nghiêm trọng.

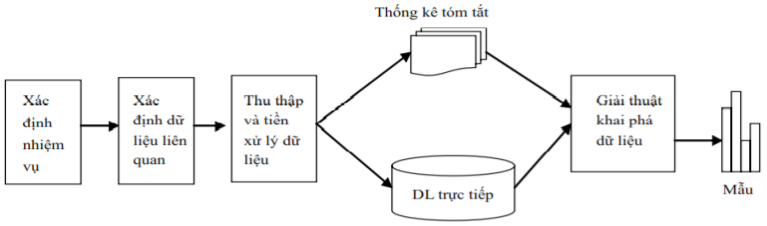
- Chuyển đổi dữ liệu *(Data conversion)*: Trong giai đoạn này, dữ liệu sẽ được chuyển đổi về dạng thuận tiện để tiến hành các thuật toán khám phá dữ liệu.

▪ Khai phá dữ liệu *(Data mining)*: Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khám phá tri thức, áp dụng các kỹ thuật khai phá (phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.

▪ Đánh giá và biểu diễn tri thức *(Knowledge representation & Evaluation)*: Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai phá ở bước trên biểu diễn theo dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,… Đồng thời, bước này cũng đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

KPDL là một giai đoạn trong quá trình khám phá tri thức. Về bản chất nó là giai đoạn duy nhất tìm ra thông tin mới, thông tin tiềm ẩn có trong cơ sở dữ liệu chủ yếu phục vụ cho mô tả và dự đoán. Dự đoán là thực hiện việc suy luận trên dữ liệu để đưa ra các dự báo nhằm phân tích tập dữ liệu huấn luyện và tạo ra một mô hình cho phép dự đoán các mẫu, mô hình mới chưa biết. Mô tả dữ liệu là tổng kết hoặc diễn tả những đặc điểm chung của những thuộc tính dữ liệu trong kho dữ liệu mà con người có thể hiểu được. Quá trình KPDL bao gồm các bước sau:

- Xác định nhiệm vụ: xác định chính xác các vấn đề cần giải quyết
- Xác định các dữ liệu liên quan: dùng để xây dựng các giải pháp
- Thu thập và tiền xử lý dữ liệu: thu thập các dữ liệu liên quan và tiền xử lý chúng sao cho thuật toán KPDL có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), …
- Thuật toán KPDL: lựa chọn thuật toán KPDL và thực hiện việc KPDL để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định, luật sản xuất, … tương ứng với ý nghĩa của nó. Đặc điểm của mẫu phải là mới (ít nhất là đối với hệ thống đó). Độ mới có thể được đo tương ứng với độ thay đổi trong dữ liệu (bằng cách so sánh các giá trị hiện tại với các giá trị trước đó hoặc các giá trị mong muốn), hoặc bằng tri thức (mối liên hệ giữa phương pháp tìm mới và phương pháp cũ như thế nào). Thường thì độ mới của mẫu được đánh giá bằng một hàm logic hoặc một hàm đo độ mới, độ bất ngờ của mẫu. Ngoài ra, mẫu còn phải có khả năng sử dụng tiềm tàng. Các mẫu này sau khi được xử lý và diễn giải phải dẫn đến những hành động có ích nào đó được đánh giá bằng một hàm lợi ích. Mẫu khai thác được phải có giá trị đối với các dữ liệu mới với độ chính xác nào đó.



Hình 1. Quá trình khai phá dữ liệu

Kỹ thuật KPDL thực chất là phương pháp không hoàn toàn mới. Nó là sự kế thừa, kết hợp và mở rộng của các kỹ thuật cơ bản đã được nghiên cứu từ trước như máy học, nhận dạng, thống kê (hồi quy, xếp loại, phân cụm), các mô hình đồ thị, các mạng Bayes, trí tuệ nhân tạo, thu thập tri thức hệ chuyên gia, v.v… Tuy nhiên, với sự kết hợp tài tình của KPDL, kỹ thuật này có ưu thế hơn hẳn các phương pháp trước đó, đem lại nhiều triển vọng trong việc ứng dụng phát triển nghiên cứu khoa học.

2.2. Một số kỹ thuật khai phá dữ liệu

Hiện nay có rất nhiều các kỹ thuật KPDL khác nhau, tuy nhiên chúng được phân thành 2 nhóm chính:

- Kỹ thuật KPDL dự đoán:

Sử dụng một số biến hoặc trường trong cơ sở dữ liệu để đoán ra các giá trị không biết hoặc sẽ có của các biến chú ý khác, sử dụng các dự đoán dựa vào các suy diễn trên dữ liệu hiện tại. Các kỹ thuật này bao gồm: phân lớp *(classification)*, hồi quy *(regression)*, … Là quá trình xếp một đối tượng vào một trong những lớp đã biết trước (VD: Phân lớp các bệnh nhân theo dữ liệu hồ sơ bệnh án, …). Kỹ thuật này thường sử dụng một số kỹ thuật của học máy như cây quyết định *(decision tree)*, mạng noron nhân tạo *(neural network)*, …

- Kỹ thuật KPDL mô tả:

Tập trung vào việc tìm kiếm các mẫu mà con người có thể hiểu được để mô tả dữ liệu, mô tả các tính chất hoặc các đặc tính chung của dữ liệu trong cơ sở dữ liệu hiện có. Các kỹ thuật này bao gồm: phân

cụm *(clustering)*, khái quát hóa *(summerization)*, phát hiện và thay đổi độ lệch *(Evolution and deviation analyst)*, mô hình hóa sự phụ thuộc, phân tích luật kết hợp (*Association Rule*)…

Phân cụm: là mô tả chung việc tìm ra tập xác định các nhóm hay loại để mô tả dữ liệu. Các nhóm có thể tách riêng, phân cấp, hoặc chồng lên nhau.

Khái quát hóa: bao gồm các phương thức để tìm kiếm một mô tả cho một tập con dữ liệu.

Mô hình hóa sự phụ thuộc: bao gồm việc tìm kiếm một mô hình để mô tả sự phụ thuộc giữa các biến. Các mô hình phụ thuộc tồn tại có hai mức: mức cấu trúc của mô hình xác định các biến nào là phụ thuộc cục bộ với nhau, và mức định lượng của mô hình xác định các phụ thuộc theo một quy tắc nào đó.

Phát hiện và thay đổi độ lệch: tập trung vào khai thác những thay đổi đáng kể nhất trong dữ liệu từ các giá trị chuẩn hoặc được đo trước.

Luật kết hợp: mô tả mối quan hệ kết hợp giữa các thuộc tính khác nhau.

Bài báo này nghiên cứu cách áp dụng luật kết hợp và thuật toán phân cụm để trích xuất cùng sở thích của độc giả và giới thiệu sách cho họ. Những điều này được giải thích ngắn gọn như sau:

2.2.1. Luật kết hợp (Association Rule - AR)

Thuật toán Apriori được đề xuất bởi Agrawal và Srikant (1994), và là một thuật toán nổi tiếng trong vùng khai phá luật kết hợp. Trong lĩnh vực Data mining, mục đích của luật kết hợp là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như sau:

Cho cơ sở dữ liệu gồm các giao dịch T là tập các giao dịch t1, t2, …, tn.

T = {t1, t2, …, tn}. T gọi là cơ sở dữ liệu giao dịch (Transaction Database)

Mỗi giao dịch ti bao gồm tập các đối tượng I (gọi là itemset)

I = {i1, i2, …, im}. Một itemset gồm k items gọi là k-itemset

Mục đích của luật kết hợp là tìm ra sự kết hợp *(association)* hay tương quan *(correlation)* giữa các items. Những luật kết hợp này có dạng X → Y

Hai tiêu chí rất quan trọng trong việc đo lường luật kết hợp đó là độ hỗ trợ *(support)* và độ tin cậy *(confidence)*.

Công thức tính độ hỗ trợ và độ tin cậy của luật kết hợp X→Y [1]:

$$Suport\ (X \rightarrow Y) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

$$Confidence\ (X \rightarrow Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

Trong đó:

n(X): Số giao dịch chứa X

N: Tổng số giao dịch

Các luật kết hợp có độ hỗ trợ và độ tin cậy lớn hơn hoặc bằng độ hỗ trợ tối thiểu *(min_sup)* và độ tin cậy tối thiểu *(min_conf)* gọi là các luật mạnh, min_sup và min_conf gọi là các giá trị ngưỡng *(threshold)* được xác định trước khi sinh các luật kết hợp [1].

2.2.2. Phân cụm dữ liệu (Clustering)

Kỹ thuật phân cụm hoạt động bằng cách xác định các nhóm người dùng có sở thích giống nhau và phân chia các nhóm có sở thích rất khác nhau. Phân cụm dữ liệu là qui trình tìm cách nhóm các đối tượng đã cho vào các cụm *(clusters)*, sao cho các đối tượng trong cùng 1 cụm càng giống nhau *(similar)* càng tốt và các đối tượng khác cụm thì càng khác nhau *(dissimilar)* càng tốt [5]. Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Có rất nhiều kỹ thuật phân cụm, như phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ... Tuy nhiên, không có tiêu chí nào được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của bài toán phân cụm [5].

Thuật toán K-Means thường được sử dụng để tiến hành phân cụm vì nó có thể phân cụm một cách nhanh chóng. Tuy nhiên, Thuật

toán K-Means có nhược điểm khó xác định số cụm mà không gian dữ liệu có mà chỉ phù hợp ra các cụm hình cầu, ngoài ra nó nhạy cảm với nhiễu và những mẫu cá biệt. Bài viết trình bày sự kết hợp thuật toán tối ưu đàn kiến với phân cụm dữ liệu để có thể có được giải pháp tối ưu hóa toàn cục. Cách tiếp cận này làm giảm bớt những nhược điểm khiến Thuật toán K-Means dễ rơi vào tình huống khó xử khi giải pháp tối ưu hóa cục bộ có sai sót [3].

2.3. Thuật toán tối ưu đàn kiến (Ant Colony Optimization – ACO)

ACO là một phương pháp nghiên cứu lấy cảm hứng từ việc mô phỏng hành vi của đàn kiến trong tự nhiên nhằm mục tiêu giải quyết các bài toán tối ưu phức tạp. Được giới thiệu lần đầu tiên vào năm 1991 bởi A. Colorni và M. Dorigo

Trong tự nhiên, kiến thật có khả năng tìm ra con đường ngắn nhất từ nguồn thức ăn đến tổ của chúng và chúng giao tiếp với những con khác bằng cách khai thác thông tin về vết mùi (pheromone). Trên đường đi, mỗi con kiến để lại một chất hóa học pheromone gọi là vết mùi dùng để đánh dấu đường đi. Bằng cách cảm nhận vết mùi, kiến có thể lần theo đường đi đến nguồn thức ăn được các con kiến khác khám phá theo phương thức chọn ngẫu nhiên có định hướng theo nồng độ vết mùi để xác định đường đi ngắn nhất từ tổ đến nguồn thức ăn. Vết mùi này sẽ bay hơi dần và mất đi theo thời gian, nhưng nó cũng có thể được củng cố nếu những con kiến khác tiếp tục đi trên con đường đó lần nữa. Dần dần, các con kiến theo sau sẽ lựa chọn đường đi với lượng mùi dày đặc hơn, và chúng sẽ làm gia tăng hơn nữa nồng độ mùi trên những đường đi được yêu thích hơn. Các đường đi với nồng độ mùi ít hơn sẽ bị loại bỏ và cuối cùng, tất cả đàn kiến sẽ cùng kéo về một đường đi mà có khuynh hướng trở thành đường đi ngắn nhất từ tổ đến nguồn thức ăn của chúng (Dorigo và Gambardella, 1996). Ý tưởng từ đàn kiến tự nhiên được chuyển sang kiến nhân tạo. Kiến nhân tạo có bộ nhớ riêng, có khả năng ghi nhớ các đỉnh đã thăm trong hành trình và tính được độ dài đường đi nó chọn. Ngoài ra, kiến có thể trao đổi thông tin với nhau, thực hiện tính toán cần thiết, cập nhật mùi…

Thuật toán ACO có thể được tóm tắt như sau:

- Các dấu vết ảo được tích lũy trên các đoạn đường đi.

- Đường đi được lựa chọn bằng cách lựa chọn ngẫu nhiên dựa trên lượng dấu vết hiện tại trên các đoạn đường từ nút bắt đầu đi.

- Các con kiến đến điểm tiếp theo, lựa chọn đường đi tiếp sau đó.

- Tiếp tục cho đến khi đến nút bắt đầu.

- Mỗi hành trình kết thúc là một giải pháp.

- Hành trình sẽ được phân tích để tối ưu.

2.4. Vấn đề sử dụng khai phá dữ liệu trong thư viện số

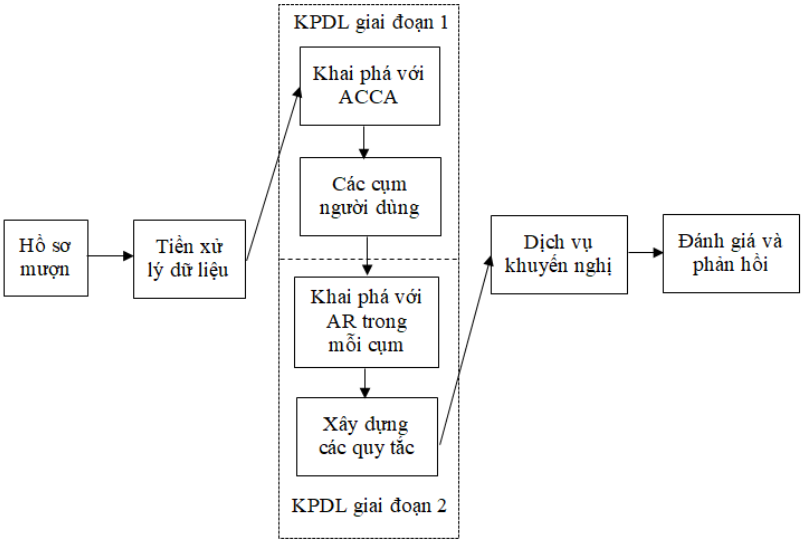
Trong nghiên cứu của mình, Borgman cho rằng thư viện số là một tập hợp các tài nguyên và các kỹ thuật liên quan để tạo, tìm kiếm và sử dụng thông tin [2]. Do sự phổ biến của thương mại điện tử và xu hướng cá nhân hóa, kỹ thuật khai phá dữ liệu cũng được sử dụng rộng rãi để phân tích hành vi của người dùng. Điều này là để xác định sở thích cá nhân và cung cấp thông tin sản phẩm nhằm nâng cao mức tiêu thụ (Agrawal và cộng sự, 1993). Áp dụng các kỹ thuật khai phá dữ liệu trong dịch vụ thư viện số cũng được coi là xu hướng vì nó có thể tự động lọc ra thông tin hữu ích theo hồ sơ người dùng và chức năng phân tích thống kê. Ví dụ: lọc ra các chủ đề phổ biến từ lịch sử mượn có thể giúp thúc đẩy lưu thông sách trong thư viện. Thư viện số cũng có thể sử dụng khai phá dữ liệu để phân tích thống kê cung cấp thông tin về sách, báo, chủ đề và các dịch vụ cá nhân khác nhằm thúc đẩy lưu thông. Thư viện số trong tương lai chắc chắn sẽ phát triển nhanh chóng. Việc áp dụng công nghệ khai phá dữ liệu trên các nguồn thông tin rộng lớn sẽ là một sự lựa chọn lớn của các công cụ khai phá tri thức và các thuật toán, cá nhân hoá dịch vụ thư viện số trở thành một phần không thể thiếu trong xây dựng hỗ trợ kỹ thuật cho thư viện số.

3. PHƯƠNG PHÁP LUẬN

Khi người dùng truy cập thư viện số, họ thường nhập các từ khóa thích hợp và sử dụng chức năng "Tìm kiếm" để khám phá thông tin

họ muốn. Tuy nhiên, không phải lúc nào kết quả tìm kiếm cũng khiến người dùng hài lòng. Trước đây, đã có một số nghiên cứu về tìm kiếm theo từ khóa. Tuy nhiên, các từ khóa được cung cấp bởi các tác giả tài liệu (nhà xuất bản hoặc thủ thư) [8], và không nhất thiết phải phản ánh kỳ vọng ngữ nghĩa của người dùng. Do đó, có một số nghiên cứu sâu hơn đã cố gắng xây dựng một số khuyến nghị cho người dùng để hỗ trợ tìm kiếm từ khóa. Năm 1999, Luis dẫn đầu một dự án có tên là "Active Recommendation Project" (ARP) tại Phòng thí nghiệm Quốc gia Los Alamos. Dự án này phát triển nghiên cứu về hệ thống khuyến nghị cho cơ sở dữ liệu lớn và Web trên toàn thế giới (www), thích ứng với mong đợi của người dùng [7]. Tiếp sau đó có Heylighen và Bollen (2002) đề xuất hệ thống khuyến nghị dựa trên các thuật toán Hebbian [6].

Bài viết trình bày phương pháp xây dựng dịch vụ khuyến nghị trong thư viện số bằng cách khai phá dữ liệu hai giai đoạn thông qua phân tích hành vi truy cập của độc giả. Trước khi tiến hành khai thác dữ liệu, nguồn dữ liệu (hồ sơ mượn trong thư viện) cần được xử lý trước. Tính đầy đủ của dữ liệu nguồn là một trong những chìa khóa cho sự thành công của việc khai phá dữ liệu. Các nhiệm vụ chính trong tiền xử lý dữ liệu bao gồm làm sạch dữ liệu, tích hợp dữ liệu và chuyển đổi dữ liệu. Để đảm bảo mức độ tinh khiết của dữ liệu, cần phải xác định các ngoại lệ và làm mịn dữ liệu nhiễu. Sau khi dữ liệu được làm sạch. Giai đoạn đầu sử dụng thuật toán phân cụm đàn kiến làm phương pháp khai phá dữ liệu và tách người dùng thành một số cụm tùy thuộc vào lịch sử truy cập của họ. Những người dùng có cùng sở thích và hành vi được gom trong cùng một cụm. Giai đoạn thứ hai, sử dụng luật kết hợp làm phương pháp khai phá dữ liệu và phát hiện ra liên kết giữa các sở thích và hành vi truy cập của người dùng. Sau đó, xây dựng các quy tắc cho dịch vụ khuyến nghị. Quá trình được thể hiện ở hình 2.



Hình 2: Quá trình KPDL hai giai đoạn

Sau đây là mô tả chi tiết về các phương pháp khai phá dữ liệu:

3.1. Thuật toán phân cụm đàn kiến (Ant Colony Clustering Algorithm - ACCA)

Các nguyên tắc cơ bản của thuật toán rất đơn giản: kiến được mô hình như các tác nhân di chuyển ngẫu nhiên trong môi trường của chúng, một lưới vuông với các điều kiện tuần hoàn. Các mục dữ liệu nằm rải rác trong môi trường này có thể được các tác nhân nhặt, vận chuyển và thả ra. Các hoạt động nhặt và thả bị sai lệch bởi sự giống nhau và mật độ của các mục dữ liệu trong khu vực lân cận của kiến: kiến có khả năng nhặt các mục dữ liệu bị khác với những dữ liệu còn lại và chúng có xu hướng thả những dữ liệu này ở vùng dữ liệu tương tự lân cận. Bằng cách này, việc phân nhóm và sắp xếp các phần tử được thu thập trên lưới. Trong quá trình phân cụm dựa trên nguyên tắc tìm kiếm thức ăn của kiến, dữ liệu được phân nhóm được coi là kiến có các đặc tính khác nhau và trung tâm phân nhóm được coi là nguồn thức ăn cần tìm kiếm. Do đó, quá trình phân cụm dữ liệu có thể được coi là quá trình kiến tìm kiếm nguồn thức ăn. Trong mỗi chu trình tìm kiếm, những con kiến sẽ tính toán xác suất chuyển tiếp (liên quan đến lượng thông tin đến tâm cụm) và thông tin heuristic để quyết định vị trí chuyển tiếp tiếp theo.

Ý tưởng: Bước đầu tiên là khởi tạo các tham số, và một nhóm kiến nhân tạo. Mỗi con kiến xây dựng cụm riêng. Khi kiến đã hoàn thành các cụm của chúng, phương sai của mỗi cụm(CV_{intra}) được tính toán. Phần trăm các nút xa nhất được chọn để được nhóm lại thành cụm có khoảng cách đến tâm $O_{center}(M)$ là ngắn nhất. Nếu phương sai mới (CV'_{intra}) nhỏ hơn CV_{intra}, thì các nút trong cụm mới cập nhật gần giống nhau hơn so với cụm trước. Trong khi áp dụng các cụm mới, con kiến sẽ cập nhật lượng pheromone trên các hành trình của nó (áp dụng quy tắc cập nhật cục bộ). Sau khi tất cả các con kiến đã tạo ra giải pháp, giải pháp tốt nhất sẽ được cập nhật cho toàn bộ hệ thống (bằng cách áp dụng quy tắc cập nhật toàn cục) cho lần lặp hiện tại. Quá trình kết thúc sau các lần lặp được xác định trước. Thuật toán phân cụm đàn kiến hoàn chỉnh được tóm tắt dưới đây:

Các ký hiệu:

NC: số lượng các cụm;

M: tổng số kiến;

M_k: tập hợp M được thực hiện bởi kiến k;

$p_k(r, s)$: xác suất mà con kiến k chọn để di chuyển từ nút r tới nút s;

$\tau(r, u)$: lượng pheromone trên cạnh (r, u);

$\overline{p}_k$: giá trị trung bình của $p_k(r, s)$ đối với tập $\in M_k$;

η *(r, u):* hàm heuristic được tính bằng nghịch đảo của khoảng cách giữa các nút r và u;

β: tham số cân nhắc tầm quan trọng tương đối của pheromone;

q: một giá trị được chọn ngẫu nhiên với xác suất đồng nhất trong [0,1];

q_0: tham số xác định tầm quan trọng tương đối của khai thác so với thăm dò ($0 \ q_0 \leq 1$);

S: một biến ngẫu nhiên được chọn theo $p_k(r,s)$

α: tham số bay hơi pheromone của cập nhật toàn cục ($0 < \alpha < 1$)

p: tham số bay hơi pheromone của cập nhật cục bộ ($0 < p < 1$)

τ: mức pheromone ban đầu;

$O_{center}(M)$: tâm của tất cả các nút trong M;

γ: một tham số là phần trăm các nút xa nhất được chọn để tập hợp lại;

CV_{intra}: phương sai trong cụm;

CV'_{intra}: phương sai trong cụm sau khi gom cụm

Tóm tắt thuật toán:

Đầu vào: n nút

Đầu ra: xác định số các cụm

Bước 0: Khởi tạo các tham số, bao gồm số lượng kiến m, tham số q_0, β, tham số bay hơi pheromone α, p phần trăm các nút xa nhất được chọn để tập hợp lại.

Bước 1: Đặt ngẫu nhiên m con kiến vào các nút

Bước 2: Nhóm các nút thành các cụm. Một con kiến k tại nút r chọn nút s để di chuyển dọc theo các nút không thuộc bộ nhớ làm việc của nó M_k. Quy tắc chuyển đổi trạng thái được áp dụng theo công thức xác suất sau:

$$s = \begin{cases} \underset{u \in M_k}{argmax}\{[\tau(r,u)]\cdot[\eta(r,u)]^{\beta}\} & \text{nếu } q < q_0 \\ S & \end{cases} \quad (1)$$

Trong đó S là biến ngẫu nhiên được chọn theo phân phối xác suất cho trong phương trình (2), ưu tiên các cạnh ngắn hơn và có mức độ pheromone cao hơn:

$$p_k(r,s) = \begin{cases} \dfrac{[\tau(r,s)]\cdot[\eta(r,s)]^{\beta}}{\sum_{u \in M_k}[\tau(r,u)]\cdot[\eta(r,u)]^{\beta}} & \text{nếu } u \notin M_k \\ 0 & \end{cases} \quad (2)$$

Nếu $p_k(r,s) \geq \overline{p_k}$ kiến k sẽ nhặt nút s

Bước 3: Tính tâm $O_{center}(M)$ và CV_{intra} trong mỗi cụm, các nút trong được chọn để gom lại nhóm gần nhất.

Bước 4: Tính CV'_{intra}. Nếu $CV'_{intra} > CV_{intra}$, sau khi một con kiến đi qua một cạnh, lượng pheromone của cạnh đó tăng lên, việc cập nhật vệt mùi trên cạnh áp dụng theo công thức:

$$\tau(r,s) = (1-p)\cdot\tau(r,s) + CV_{inter}^{-1} \quad (3)$$

Bước 5: cập nhật vệt mùi toàn cục, khi tất cả kiến đã xây dựng hành trình của mình, vệt mùi toàn hệ thống được cập nhật theo công thức:

$$\tau(r, s) = (1 - \alpha) \cdot \tau(r, s) + CV^{-1} \quad (4)$$

Trong đó CV là tổng của các CV_{intra} nhỏ nhất trong các CV_{intra}

Bước 6. Quá trình được lặp lại cho đến khi thỏa mãn điều kiện cuối cùng.

3.2. Luật kết hợp

Giai đoạn thứ hai của phương pháp khai phá dữ liệu là tìm ra các mẫu chung của các mối quan hệ trong mỗi cụm bằng luật kết hợp. Trước khi thực hiện, dữ liệu phải được tích hợp. Bài báo này trình bày cách sử dụng thuật toán Apriori để khai phá luật kết hợp. Có hai bước để khai phá luật kết hợp:

Bước 1: Tìm tất cả tập các mục lớn

(1) Độ hỗ trợ của tập các mục lớn phải lớn hơn độ hỗ trợ tối thiểu do người dùng xác định.

$$Support(AB) = p(AB)$$

(2) Nếu có k mục trong một tập lớn, thì chúng ta gọi nó là tập k mục lớn.

Bước 2: Sử dụng tập các mục lớn được tạo ở bước đầu tiên để tạo ra tất cả các quy tắc kết hợp:

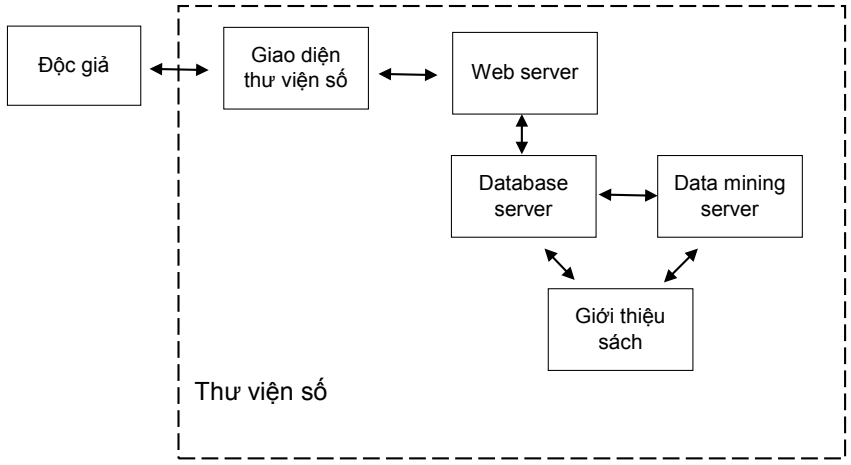
(1) Tính độ tin cậy:

$$Confidence(A \Rightarrow B) = (P|B) = support\_count(AB) \,/\, support\_count(A).$$

(2) Nếu độ tin cậy của quy tắc kết hợp lớn hơn độ tin cậy tối thiểu do người dùng xác định, thì nó mới hiệu quả.

Thuật toán kết thúc khi không có tập hợp mục nào có thể được xây dựng cho vòng tiếp theo.

Sau đó có thể giới thiệu sách dựa vào các quy tắc liên kết. Kiến trúc hệ thống khuyến nghị được hiển thị như hình 3.



Hình 3: Kiến trúc hệ thống khuyến nghị

4. KẾT LUẬN

Sự phát triển mạnh mẽ của công nghệ thông tin làm cho chức năng của các dịch vụ cá nhân trở nên quan trọng hơn trước. Thư viện số đã hình thành và ngày càng khẳng định giá trị của nó trong các cơ quan, tổ chức. Khai phá dữ liệu cung cấp và hỗ trợ kỹ thuật cho các tổ chức và quản lý các nguồn tài nguyên kỹ thuật số, thúc đẩy sự mở rộng chất lượng dịch vụ, và cùng một lúc làm cho phương pháp nghiên cứu các công nghệ khai phá dữ liệu phát triển cả về quy mô lẫn chiều sâu. Bài báo thảo luận về cách ứng dụng công nghệ khai phá dữ liệu để xây dựng các dịch vụ khuyến nghị cho người dùng dựa trên sở thích của người dùng. Bằng cách sử dụng Thuật toán phân cụm đàn kiến và các quy tắc kết hợp để thiết kế quy trình khai phá dữ liệu hai giai đoạn để tạo ra hệ thống khuyến nghị. Bài báo không chỉ trình bày cách xây dựng một cơ chế giới thiệu cho người đọc trong việc tìm kiếm sách từ trang Web mà còn đã chủ động tìm kiếm những cuốn sách phù hợp nhất cho độc giả. Từ đó, các nhà quản lý thư viện dự kiến sẽ mua những cuốn sách cốt lõi nhất và hấp dẫn để đáp ứng yêu cầu của độc giả cùng với việc quảng bá dịch vụ thư viện số.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn Đức Thuần (2013), *Nhập môn khai phá dữ liệu và quản trị tri thức*, NXB Thông tin và Truyền thông, 2013.

Tiếng Anh

2. Borgman, C.L. (1999*), "What are digital libraries? Competing visions"*, Information Processing & Management, Vol. 35, pp. 227-43.
3. Chen, A.P. and Chen, C.C. (2006), *"A new efficient approach for data clustering in electronic library using ant colony clustering algorithm"*, The Electronic Library, Vol. 24 No. 4, pp. 548-59.
4. Guo, Yike., & Grossman, R.L. (2002). *High Performance Data Mining Scaling Algorithms, Applications and Systems* (1st ed.). Springer US.
5. Han, Jiawei, & Kamber, Micheline (2006), *Data mining : concepts and techniques* (2nd ed.), Morgan Kaufmann, San Diego.
6. Heylighen, F. and Bollen, J. (2002), "Hebbian algorithms for a digital library recommendation system", Proceedings of International Conference on Parallel Processing Workshops, Vancouver, pp. 439-44.
7. Rocha, L.M. (1999), *"TalkMine and the Adaptive Recommendation Project, Proceedings of the Association for Computing Machinery (ACM) – Digital Libraries"*, University Of California, Berkeley, CA, pp. 242-3.
8. Rocha, L.M. and Bollen, J. (2001), "Biologically motivated distributed designs for adaptive knowledge management", in Segel, L. and Cohen, I. (Eds), *Design Principles for the Immune System and other Distributed Autonomous Systems*, Santa Fe Institute Series in the Sciences of Complexity, Oxford University Press, Oxford, pp. 305-34.