Quá Trình Khai Thác Dữ Liệu Trong Thông Tin - Thư Viện

Khai phá dữ liệu trong thư viện số

Dựa trên những đặc trưng của kỹ thuật khai phá dữ liệu và tổ chức thư viện số, bài viết trình bày các khả năng khai thác của kỹ thuật khai phá dữ liệu trong thư viện số sau đó đề xuất các áp dụng của kỹ thuật khai phá dữ liệu trong thư viện số trên khía cạnh: Cải thiện tốc độ; nâng cấp chất lượng dịch vụ thông tin của thư viện; hỗ trợ các quyết định của thư viện; dịch vụ thông tin cá nhân; tự động hóa xử lý thông tin; hỗ trợ các công việc khác nhằm nâng cao hiệu quả sử dụng cũng như khai thác thư viện số.

1. Khai phá dữ liệu và thư viện số

1.1 Khai phá dữ liệu Khai phá dữ liệu

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

Khai phá dữ liệu là lĩnh vực nghiên cứu trong khoa học máy tính nói chung, trong trí tuệ nhân tạo, xử lý tri thức thông minh nói riêng. Khai phá dữ liệu là bước chính của quy trình khai phá tri thức trong CSDL (Knowledge Discovery in Database- KDD).

1.2 Thư viện số

Thư viện số là một nguồn tài nguyên thông tin số vô cùng to lớn trong đó có các phương tiện truyền thông với nhiều nội dung đa dạng khác nhau. Nó còn là một hệ thống thông tin kỹ thuật số được ra đời bởi sự hỗ trợ của nhiều công nghệ cao và hiện đại, là thế hệ tiếp theo của phương thức quản lý tài nguyên thông tin Internet, là một loại cơ chế dịch vụ thuận tiện cung cấp thông tin cho độc giả.

Quá trình khai phá dữ liệu có thể được chia thành 3 giai đoạn: giai đoạn chuẩn bị dữ liệu, giai đoạn khai phá tri thức và giai đoạn trình bày và thể hiện kết quả [4, 5]. Quá trình khai phá tri thức được lặp đi lặp lại với sự tham gia của người sử dụng. Có sự khác biệt nhất định giữa khai phá dữ liệu và khai phá tri thức.

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

2. Khả năng khai phá dữ liệu trong thư viện số

2.1 Khai phá cấu trúc thư viện số

Thư viện số được thiết kế trên cấu trúc các trang web, nó sử dụng các ngôn ngữ thiết kế web cùng với các siêu liên kết để tổ chức thông tin. Trên cơ sở đó, thông qua các siêu liên kết và tổ chức của trang, các kết nối, các thư mục, nội dung mà chúng liên kết đến chúng ta có thể khám phá ra các kiến thức mới và bổ ích. Các kỹ thuật khai phá trang web (web mining) được khai thác một cách triệt để để thu được các thông tin mới và ý nghĩa nhất.

2.2 Khai phá người sử dụng thư viện số

Khi người sử dụng khai thác tài nguyên trên thư viện số, một phiên giao dịch sẽ ghi lại tất cả các lần người sử dụng trình duyệt web theo thời gian để hình thành cơ sở dữ liệu giao dịch,

kết quả là chúng ta có thể thu thập và lưu trữ lại các thông tin của người sử dụng như là các bộ sưu tập đặc biệt thông qua chế độ duyệt web, từ đó sử dụng các kỹ thuật để khai phá thông tin.

Sử dụng các thuật toán khai phá luật kết hợp để tìm các giao dịch tập hợp có tần số truy cập vượt quámột ngưỡng nhất định, sau đó sử dụng kết quả này để phân loại dữ liệu.

Sử dụng khai phá web để có được mô hình chuỗi các truy cập của người dùng trước đó và thực hiện truyền các trang người dùng có thể đọc theo dự đoán.

2.3 Khai phá nội dung trong thư viện số

Dựa trên nội dung các trang web, nội dung có trong thư viện số bao gồm: văn bản có cấu trúc, văn bản phi cấu trúc, các loại văn bản, các bảng, dữ liệu đa phương tiện, âm thanh, ảnh. Có thể khai phá nội dung từ thư viện kỹ thuật số thông qua các hình thức sau:

Sử dụng kỹ thuật tóm tắt văn bản để khai phá các tóm tắt (abstract) từ các file dữ liệu. Đây là phần nội dung quan trọng và là trọng tâm của mỗi tài liệu, nó phản ảnh nội dung chính của tài liệu đó.

- Phân loại văn bản: Tự động phân loại văn bản trên cơ sở tài liệu người dùng, kết quả phân loại sẽ phục vụ các tìm kiếm và khai thác của người sử dụng.
- Phân cụm là kỹ thuật được sử dụng để nhóm các tài liệu tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng cùng cụm là tương đồng. Kết quả phân cụm sẽ giúp cho người sử dụng xác định được các tài liệu tương tự hay có cùng nhóm nội dung.
- Dự đoán và đánh giá đó là tìm ra những thông tin mới, những quyết định mới từ những dữ liệu đã có thông qua quá trình quan sát và xử lý.

3. Áp dụng kỹ thuật khai phá dữ liệu cho thư viện số

3.1 Nâng cao tốc độ

Nâng cao tốc độ trong mọi hoạt động của thư viện và dịch vụ người dùng là mục tiêu mà thư viện số hướng tới. Nâng cao tốc độ nhằm đáp ứng thời gian truy cập các thông tin cần thiết cho người sử dụng, đáp ứng khả năng trao đổi thông tin, truy xuất thông tin, khai thác các dịch vụ một cách hiệu quả nhất.

3.2 Nâng cấp chất lượng dịch vụ thông tin

Sử dụng các kỹ thuật để xây dựng thư viện phát triển theo hướng phần mềm thông minh, bao gồm dịch vụ truy vấn thông tin. Cải tiến công cụ phục hồi truyền thống thân thiện, dễ hiểu và tương tác theo kết quả. Tìm kiếm thông tin thông minh không chỉ hỗ trợ việc thu hồi khái niệm, tìm kiếm mờ, kết hợp thu hồi và phục hồi đa ngôn ngữ, mà còn có thể nhanh chóng sử dụng các thuật toán phân cụm, phân tích kết quả truy vấn, để thuận tiện cho việc lựa chọn của người sử dụng, và cùng một lúc xác định tìm kiếm thêm trên cơ sở này.

3.3 Hỗ trợ các quyết định của thư viện

Mức độ quản lý thấp là một trong những yếu tố cơ bản ảnh hưởng đến sự phát triển của các thư viện. Trước đây, việc đưa ra quyết định của thư viện chủ yếu dựa vào kinh nghiệm, điều này là chủ quan, một chiều, thiếu thông tin, và không thể đáp ứng yêu cầu của thời đại. Công nghệ khai phá dữ liệu có thể cung cấp thông tin bảo đảm cho việc ra quyết định của lãnh đạo quản lý thư viện, cụ thể: - Khai phá dữ liệu có thể cùng một lúc thu thập dữ liệu nội bộ và thông tin bên ngoài có liên quan đến hệ thống thông tin của thư viện, và sau khi xử lý, chuyển đổi, tạo thành các thông tin tập trung, thống nhất và có sẵn, để tránh việc đưa ra quyết định sai lầm do thiếu thông tin. - Sử dụng các công cụ hệ thống OLAP kho dữ liệu để

so sánh với việc tích hợp các dữ liệu đa chiều, xem xét và xác minh giả thiết của quyết định chính sách, để nâng cao tính khả thi và độ tin cậy của các quyết định, và sử dụng hợp lý các nguồn tài nguyên hạn chế, đồng thời tối ưu hóa phân bổ nguồn lực vào thư viện. - Sử dụng các công cụ khai phá dữ liệu để tìm ra một mô hình tiềm ẩn từ các dữ liệu lịch sử và dự báo tự động trên cơ sở của mô hình.

3.4 Cung cấp dịch vụ thông tin cá nhân

Việc áp dụng công nghệ khai phá dữ liệu làm cho các dịch vụ thông tin của thư viện hoạt động tốt hơn, giúp nâng cao hiệu quả của các dịch vụ thông tin và thư viện. Sử dụng công nghệ khai phá dữ liệu đối với CSDL duyệt web của người dùng để tìm mô hình sử dụng của người sử dụng và chủ động cung cấp dịch vụ cá nhân theo mô hình quan tâm của người dùng.

3.5 Hỗ trợ các công việc khác

a. Đối với bộ phận cung cấp tài nguyên

Bộ phận này có thể sử dụng các chức năng của khai phá dữ liệu để phân tích và sử dụng nguồn kinh phí một cách hiệu quả. Làm thế nào để việc sử dụng nguồn kinh phí hạn chế dành cho việc mua sách- đảm bảo về chất lượng và tính hợp lý của hệ thống tài nguyên thông tin của thư viện. Chính vì vậy, việc định vị chính xác nhu cầu độc giả là một yếu tố quan trọng để nâng cao tỷ lệ sử dụng các nguồn lực. Việc sử dụng phân nhóm khai phá dữ liệu và công nghệ phát hiện độ lệch và phương pháp câu hỏi của độc giả có thể cung cấp nền tảng cơ bản cho việc phân tích phân nhóm, phân tích kết quả khảo sát, và hiểu được nhu cầu của độc giả thông qua việc sử dụng sách, và thông tin phản hồi của độc giả, qua đó đưa ra quyết định phù hợp cho công tác bổ sung nguồn tài liệu.

b. Đối với bộ phận phục vụ: Sử dụng các phương pháp phân tích kết hợp khai phá dữ liệu để phân tích các dữ liệu mượn trả. Những cuốn sách có số lượng giao dịch lớn sẽ dành vị trí ưu tiên. Những người mượn thường xuyên hoặc những cuốn sách đã bị hư hỏng cần phải có hình thức phản hồi nhanh chóng cho bộ phận cung cấp tài nguyên để tăng số lượng hoặc thay đổi số lượng.