Journal of educational equipment: Applied research, Volume 2, Issue 297 (September 2023)
ISSN 1859 - 0810

Kỹ thuật và ứng dụng của khai thác văn bản

Nguyễn Văn Quang*

**ThS. Công nghệ thông tin, Trường Đại học Hải phòng*
Received: 30/7/2023; Accepted: 7/8/2023; Published: 14/8/2023

Abstract: *Text mining has become an exciting research field as it tries to discover valuable information ji·om unstructured texls. The unstructured texts which contain vast amount of information cannot simply be used for further processing by computers. Therefore, exact processing methods, algorithms and techniques are vital in order to extract this valuable information which is completed by using text mining. In this paper, we have discussed general idea of text mining and comparison of its techniques. In addition, we briefly discuss a number of text mining applications which are used presently and in future.*

Keywords: *Retrieval, Extraction, Categorization, Clustering, Summarization.*

1. Đặt vấn đề

Khai thác văn bản (KTVB) đã trở thành vùng nghiên cứu quan trọng. Một số lượng rất lớn thông tin được lưu trữ ở những nơi khác nhau trong cấu trúc phi cấu trúc. Khoảng 80% dữ liệu của thế giới ở dạng văn bản phi cấu trúc [l]. Văn bản phi cấu trúc này không thể được máy tính dễ dàng sử dụng để xử lý thêm. Vì vậy cần có một số kỹ thuật hữu ích để trích xuất một số thông tin quý giá từ văn bản phi cấu trúc. Những thông tin này sau đó được lưu trữ ở định dạng cơ sở dữ liệu văn bản chứa các trường có cấu trúc và một số trường không có cấu trúc. Văn bản có thể được lưu trữ trong thư, cuộc trò chuyện, SMS, bài báo, tạp chí, đánh giá sản phẩm và hồ sơ tổ chức [2]. Hầu hết các tổ chức, khu vực chính phủ, Có nhiều tên gọi khác nhau để KTVB như khai thác dữ liệu văn bản, khám phá tri thức [4] từ cơ sở dữ liệu văn bản, phân tích văn bản thông minh đề cập đến việc trích xuất hoặc truy xuất thông tin có giá trị từ văn bản phi cấu trúc. Nó có thể được xem như một phần mở rộng của khai phá dữ liệu hoặc khám phá tri thức từ cơ sở dữ liệu (có cấu trúc). KTVB phát hiện ra các mẫu thông tin mới từ dữ liệu văn bản vốn là thông tin bí mật hoặc chưa được xác định trước đó bằng cách trích xuất nó bằng các kỹ thuật khác nhau. KTVB là một lĩnh vực đa ngành, liên quan đến việc truy xuất thông tin, phân tích văn bản, trích xuất thông tin, phân loại, phân cụm, trực quan hóa, khai thác dữ liệu và học máy.

Có năm bước khai phá văn bản cơ bản như sau:

Các bước khai phá văn bản:

a) Thu thập thông tin từ dữ liệu phi cấu trúc.

b) Chuyển đổi thông tin nhận được này thành dữ liệu có cấu trúc

c) Xác định mẫu từ dữ liệu có cấu trúc

d) Phân tích mẫu

e) Trích xuất các thông tin có giá trị và lưu trữ vào cơ sở dữ liệu. Công nghệ KTVB cơ bản

2. Nội dung nghiên cứu

Truy xuất thông tin: Hệ thống truy xuất thông tin (IR) nổi tiếng nhất là các công cụ tìm kiếm của Google nhận dạng các tài liệu trên World Wide Web có liên quan đến một tập hợp các từ nhất định. Nó được đo lường như một phần mở rộng cho việc truy xuất tài liệu trong đó các tài liệu được trả về được xử lý để trích xuất thông tin hữu ích quan trọng cho người dùng [3]. Do đó, việc truy xuất tài liệu được theo sau bởi giai đoạn tóm tắt văn bản tập trung vào truy vấn do người dùng đặt ra hoặc giai đoạn trích xuất thông tin. IR theo nghĩa rộng hơn đề cập đến toàn bộ phạm vi xử lý thông tin, từ truy xuất thông tin đến truy xuất tri thức. Đây là một lĩnh vực nghiên cứu tương đối cũ, nơi những nỗ lực lập chỉ mục tự động đầu tiên được thực hiện vào năm 1975. Nó ngày càng thu hút được sự chú ý với sự phát triển của World Wide Web và nhu cầu về các công cụ tìm kiếm đẳng cấp.

Khai thác thông tin: Mục tiêu của phương pháp trích xuất thông tin (IE) là trích xuất thông tin hữu ích từ văn bản. Nó xác định việc trích xuất các thực thể, sự kiện và mối quan hệ từ văn bản bán cấu trúc hoặc không cấu trúc. Hầu hết các thông tin hữu ích như tên của người, địa điểm và tổ chức đều được trích xuất mà không hiểu đúng văn bản [4]. IE quan tâm đến việc trích xuất thông tin ngữ nghĩa từ văn bản. IE có thể được mô tả như là việc xây dựng một hình ảnh có cấu trúc gồm các thông tin có liên quan được chọn lọc từ văn bản.

Phân loại: Phân loại văn bản là một loại hình học tập "có giám sát", trong đó các danh mục được biết trước và được tiến hành chắc chắn cho từng tài liệu đào tạo. Sau đó, mục đích sử dụng chính của nó là lập chỉ mục tài liệu khoa học thông qua các từ được kiểm soát. Chỉ đến những năm 1990, lĩnh vực này mới phát triển đầy đủ với sự sẵn có của số lượng tài liệu văn bản ở dạng kỹ thuật số ngày càng tăng và yêu cầu

sắp xếp chúng để sử dụng dễ dàng hơn [5]. Phân loại là việc gán các tài liệu ngôn ngữ thông thường cho tập hợp các chủ đề được xác định trước theo nội dung của chúng. Nó là tập hợp các tài liệu văn bản, là quá trình tìm kiếm chủ đề hoặc các chủ đề chính xác cho từng tài liệu. Ngày nay, việc phân loại văn bản tự động được áp dụng trong nhiều bối cảnh khác nhau, từ lập chỉ mục văn bản tự động hoặc bán tự động cổ điển đến phân phối quảng cáo được cá nhân hóa, lọc thư rác và phân loại trang Web theo danh mục phân cấp, tạo siêu dữ liệu tự động và phát hiện thể loại văn bản, theo dõi chủ đề và nhiều người khác. Việc học phân loại văn bản tự động bắt đầu vào đầu những năm 1960. Đây là một chủ đề nóng trong lĩnh vực nghiên cứu học máy ngày nay.

Phân cụm

Phân cụm là một trong những chủ đề thú vị và quan trọng nhất trong khai phá văn bản. Mục đích của nó là tìm ra cấu trúc nội tại của thông tin và sắp xếp chúng thành các nhóm nhỏ quan trọng để nghiên cứu và phân tích sâu hơn. Đó là một quá trình không giám sát thông qua đó các đối tượng được phân loại thành các nhóm gọi là cụm. Vấn đề là nhóm bộ sưu tập không có nhãn đã cho thành các cụm có ý nghĩa mà không có bất kỳ thông tin nào trước đó. Bất kỳ nhãn nào liên quan đến đối tượng đều chỉ được lấy từ dữ liệu. Ví dụ: phân cụm tài liệu hỗ trợ truy xuất bằng cách tạo liên kết giữa các tài liệu liên quan, điều này cho phép truy xuất các tài liệu liên quan sau khi một trong các tài liệu được coi là có liên quan đến truy vấn.

Phân cụm rất hữu ích trong nhiều lĩnh vực ứng dụng như sinh học, khai thác dữ liệu, nhận dạng mẫu, truy xuất tài liệu, phân đoạn hình ảnh, phân loại mẫu, bảo mật, kinh doanh thông minh và tìm kiếm trên Web. Phân tích cụm có thể được sử dụng như một công cụ KTVB độc lập để đạt được phân phối dữ liệu hoặc làm bước xử lý trước cho các thuật toán KTVB khác hoạt động trên các cụm được phát hiện.

Tóm tắt văn bản là một thách thức cũ trong KTVB nhưng rất cần sự quan tâm của các nhà nghiên cứu trong các lĩnh vực trí tuệ tính toán, kiến thức máy và xử lý ngôn ngữ tự nhiên. Tóm tắt văn bản là quá trình tự động tạo phiên bản nén của một văn bản nhất định để cung cấp thông tin hữu ích cho người dùng. Ở tổ chức, công ty lớn, người nghiên cứu không có thời gian đọc hết tài liệu nên họ tóm tắt tài liệu và đánh dấu phần tóm tắt bằng những điểm chính [4]. Tóm tắt là văn bản được tạo từ một hoặc nhiều văn bản chứa một phần thông tin quan trọng, được giảm độ dài và giữ nguyên ý nghĩa tổng thể như trong văn bản gốc. Tóm tắt văn bản bao gồm nhiều phương pháp khác nhau sử dụng phân loại văn bản, chẳng hạn như mạng lưới thần kinh, cây quyết định, biểu đồ ngữ nghĩa, mô hình hồi quy, logic mờ và trí tuệ bầy đàn. Tuy nhiên, tất cả các phương pháp này đều có một vấn đề chung, đó là chất lượng phát triển của các bộ phân loại rất khác nhau và phụ thuộc nhiều vào loại văn bản được tóm tắt.

So sánh các kỹ thuật KTVB: KTVB sử dụng nhiều kỹ thuật khác nhau đóng một vai trò quan trọng. Các kỹ thuật khác nhau. Kỹ thuật truy xuất thông tin sử dụng văn bản phi cấu trúc trong đó nó có thể truy xuất thông tin có giá trị trong khi thông tin trích xuất sẽ trích xuất thông tin từ cơ sở dữ liệu có cấu trúc. Kỹ thuật Tóm tắt được sử dụng để tóm tắt tài liệu giúp giảm độ dài và giữ nguyên ý nghĩa. Việc phân loại là quá trình được giám sát và sử dụng các tài liệu được thiết lập trước theo nội dung của chúng. Tính đáp ứng và tính linh hoạt của hệ thống hậu phối hợp ngăn cản việc thiết lập các mối quan hệ có ý nghĩa một cách hiệu quả vì một danh mục được tạo ra bởi cá nhân chứ không phải hệ thống. Trong khi việc phân cụm được sử dụng để tìm các cấu trúc nội tại trong thông tin và sắp xếp chúng thành các nhóm nhỏ liên quan để nghiên cứu và phân tích thêm. Đó là một quá trình không giám sát thông qua đó các đối tượng được phân loại thành các nhóm gọi là cụm. Phân cụm đang xử lý dữ liệu nhiều chiều, tìm kiếm mẫu thú vị liên quan đến dữ liệu. Một đặc điểm khác là nó là một nhóm các loại dữ liệu tương tự nhau và mối quan hệ giữa chúng.

Ứng dụng KTVB bao gồm việc khám phá các mô hình và xu hướng trong các tạp chí và kỷ yếu từ khối lượng lớn các bài báo là một nhiệm vụ thiết yếu trong lĩnh vực nghiên cứu [l]. Vấn đề quan trọng đối với các nhà xuất bản nắm giữ cơ sở dữ liệu thông tin lớn cần lập chỉ mục để truy xuất. Điều này đặc biệt đúng trong các ngành khoa học trong đó thông tin rất cụ thể thường được chứa trong văn bản. Công cụ KTVB này được áp dụng để khám phá các xu hướng về các chủ đề khác nhau tồn tại trong quá trình tố tụng và cho thấy chúng thay đổi như thế nào theo thời gian. Nó cũng được sử dụng như theo dõi chủ đề. Do đó, các sáng kiến đã được thực hiện như đề xuất của Nature về Giao diện KTVB mở (OTMI) và Định nghĩa loại tài liệu xuất bản tạp chí chung (DTD) của Viện Y tế Quốc gia sẽ cung cấp tín hiệu ngữ nghĩa cho máy để trả lời các truy vấn cụ thể có trong văn bản mà không xóa rào cản của nhà xuất bản đối với sự tiếp cận của công chúng. Công việc nghiên cứu đã phát triển trong lĩnh vực tin sinh học, nơi tài liệu y sinh đã trở thành một lĩnh vực ứng dụng nghiên cứu quan trọng để KTVB. Vào năm 2005, cuốn sách giáo khoa đầu tiên về KTVB

y sinh đã xuất hiện, trong đó báo cáo rằng ngành công nghiệp đã gợi ý rằng 90% mục tiêu về ma túy đều bắt nguồn từ tài liệu. Động lực cho công việc này chủ yếu đến từ các nhà sinh học, những người nhận thấy mình phải đối mặt với sự gia tăng lớn về số lượng ấn phẩm trong lĩnh vực của họ, việc theo kịp các tài liệu liên quan là điều gần như không thể đối với nhiều nhà khoa học. Mục tiêu của việc KTVB trong lĩnh vực này là cho phép các nhà nghiên cứu y sinh trích xuất kiến thức từ các tài liệu y sinh nhằm tạo điều kiện cho sự đổi mới mới theo cách hiệu quả hơn. Một ứng dụng KTVB trực tuyến trong tài liệu y sinh là sự kết hợp KTVB y sinh với trực quan hóa mạng như một dịch vụ Internet. Nhận dạng thực thể sinh học nhằm mục đích xác định và phân loại các thuật ngữ kỹ thuật trong lĩnh vực sinh học phân tử tương ứng với các trường hợp khái niệm được các nhà sinh học quan tâm. Nhận dạng thực thể ngày càng trở nên quan trọng với sự gia tăng lớn về kết quả được báo cáo do các phương pháp thử nghiệm có năng suất cao. Nó có thể được sử dụng trong một số tác vụ truy cập thông tin cấp cao hơn như trích xuất quan hệ, tóm tắt và trả lời câu hỏi.

2.1. Phân tích bản quyền và hồ sơ khách hàng

Phân tích bản quyền đã phát triển thành một lĩnh vực ứng dụng rộng lớn trong những năm gần đây do số lượng đơn đăng ký bản quyền ngày càng tăng. Các kỹ thuật được giám sát và không giám sát được áp dụng để phân tích các tài liệu bản quyền và hỗ trợ các công ty cũng như cơ quan bản quyền ở một số quốc gia trong công việc của họ. Những thách thức trong phân tích bản quyền bao gồm độ dài của tài liệu, lớn hơn tài liệu thường được sử dụng trong phân loại văn bản và số lượng lớn tài liệu có sẵn trong kho ngữ liệu.

Các công ty sử dụng việc KTVB để rút ra sự xuất hiện và trường hợp của các thuật ngữ chính trong khối văn bản lớn như các bài báo, trang Web, diễn đàn khiếu nại. Phần mềm chuyển đổi các định dạng dữ liệu phi cấu trúc thành cấu trúc chủ đề và mạng ngữ nghĩa là những công cụ khoan thông tin quan trọng. Bằng cách nghiên cứu mạng ngữ nghĩa, người ta có thể tìm hiểu chất lượng chung của những lời phàn nàn, lý do phàn nàn. Nó cũng tìm thấy các từ phổ biến được sử dụng trong khiếu nại và mối quan hệ của chúng với các từ khác trong văn bản thông qua trọng số ngữ nghĩa.

2.2. Bảo mật mạng

Việc sử dụng công cụ KTVB trong lĩnh vực bảo mật đã trở thành một vấn đề quan trọng. Rất nhiều gói phần mềm KTVB được tiếp thị cho các ứng dụng bảo mật, đặc biệt là giám sát và phân tích các nguồn văn bản thuần túy trực tuyến như tin tức Internet, biogs, thư, v.v. vì mục đích bảo mật. Nó cũng tham gia vào việc nghiên cứu mã hóa/giải mã văn bản. Các cơ quan chính phủ đang đầu tư nguồn lực đáng kể vào việc giám sát tất cả các loại thông tin liên lạc, chẳng hạn như email, trò chuyện trực tuyến. Email được sử dụng trong nhiều hoạt động hợp pháp như trao đổi tin nhắn, tài liệu. Thật không may, nó cũng có thể bị lạm dụng, ví dụ như để phân phối thư rác không mong muốn, gửi tài liệu xúc phạm hoặc bắt nạt. Sự phát triển bùng nổ của email không được yêu cầu, thường được gọi là thư rác, trong những năm qua đã liên tục làm suy yếu khả năng sử dụng của e-mail. Một giải pháp được cung cấp bởi các bộ lọc chống thư rác. Hầu hết các bộ lọc có sẵn trên thị trường đều sử dụng danh sách đen và các quy tắc thủ công. Vì thời gian là rất quan trọng và xét đến quy mô của vấn đề, việc giám sát email hoặc trò chuyện trực tuyến một cách bình thường là không thể. Do đó, các công cụ KTVB tự động mang lại nhiều hứa hẹn trong lĩnh vực này.

3. Kết luận

KTVB thường đề cập đến quá trình trích xuất thông tin có giá trị từ văn bản phi cấu trúc. Trong cuộc khảo sát này một số kỹ thuật mmmg văn bản và ứng dụng của nó trong các lĩnh vực khác nhau đã được thảo luận. Một so sánh về KTVB khác nhau đã được hiển thị và có thể được nâng cao hơn nữa. Các thuật toán KTVB sẽ cung cấp cho chúng ta dữ liệu có cấu trúc và hữu ích, giúp giảm thời gian và chi phí. Thông tin ẩn trong các trang mạng xã hội, tin sinh học và bảo mật internet...được xác định bằng cách sử dụng KTVB là một thách thức lớn trong các lĩnh vực này. Sự tiến bộ của công nghệ web đã dẫn đến sự quan tâm to lớn đến việc phân loại tài liệu văn bản có chứa các liên kết hoặc thông tin khác.

Tài liệu tham khảo

[l] Vallikannu Ramanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.

[2] Vidya K A, G Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No 2, pp.613-622.

[3] R.Sagayam, S.Srinivasan, S.Roshini, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques". Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

[4] Vishal Gupta and Guruprit Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[5] http://www.cs.waikato.ac.nz/~ml/weka/