Khả năng truy xuất dữ liệu của công cụ tìm kiếm rất quan trọng đối với việc chia sẻ và tái sử dụng dữ liệu. Các công cụ tìm kiếm hiện nay thường dựa vào việc so sánh các câu truy vấn với phần mô tả của tập dữ liệu. Tuy nhiên người dùng thì không đủ kiến thức chuyên ngành để viết các câu truy vấn phù hợp với văn bản mô tả. Chúng tôi đề xuất một mô hình chỉ mục mà có thể tự tạo ra lược đồ chỉ mục dựa trên nội dung của bảng chứa tập dữ liệu. Chúng tôi kết hợp các lược đồ chỉ mục vào mô hình xếp hạng hỗn hợp, chúng không chỉ xem xét mối liên hệ giữa câu truy vấn và metadata của tập dữ liệu mà còn xem xét cả sự tương đồng giữa câu truy vấn và lược đồ chỉ mục đã được tạo. Để đánh giá phương pháp này trên tập dữ liệu thực tế, chúng tôi tạo ra một tiêu chuẩn mới dành riêng cho việc truy xuất dữ liệu. Qua các lần thử nghiệm đã cho thấy rằng cách tiếp cận của chúng tôi có thể cải thiện rõ rệt độ chính xác và điểm số NDCG của tác vụ truy xuất dữ liệu khi được so sánh với các phương pháp cơ sở. Chúng tôi cũng thử nghiệm trên tập hợp các bảng Wikipedia để thấy rằng các tính năng được tạo ra từ lược đồ chỉ mục cũng có thể cải thiện tác vụ truy xuất bảng web theo cách không giám sát và có giám sát

Từ khoá: Dataset search · Table retrieval · Text normalization Data fusion

1 giới thiệu

Việc truy xuất dữ liệu đang ngày càng nhận được nhiều sự quan tâm của nhiều người từ nhiều lĩnh vực khác nhau dựa trên tập dữ liệu trong công việc của họ.

Có rất nhiều nguồn dữ liệu với mục đích chia sẻ và quản lý dữ liệu hiệu quả như là

data.gov, datahub và data. world. Hầu các hết nguồn dữ liệu này đều sử dụng CKAN như là lớp truy cập dữ liệu. Tuy nhiên có hai vấn đề của các công cụ tìm kiếm tập dữ liệu sử dụng cơ sở hạ tầng đó là: Thứ nhất là hiệu suất xếp hạng dựa vào chất lượng metadata của tập dữ liệu, trong khi đó nhiều tập dữ liệu thiếu metadata chất lượng cao. Thứ hai là thông tin trong metadata có thể không đủ đáp ứng được nhu cầu thông tin của người dùng hoặc hỗ trợ họ giải quyết công việc. Một người dùng có thể không biết cấu tạo của tập dữ liệu tiềm năng, hoặc các thẻ dữ liệu mà nhà xuất bản cung cấp cùng với tập dữ liệu. Các thông tin như vậy khó có thể được sử dụng trong xếp hạng tập dữ liệu.

Trong bài báo này, chúng tôi tập trung vào vấn đề truy xuất dữ liệu mà ở đó nội dung tập dữ liệu được lưu trữ ở dạng bảng, dữ liệu dạng bảng được sử dụng rộng rãi và rất dễ để đọc và viết. Như đã được minh hoạ ở **Lược đồ 1**, một tập dữ liệu bao gồm một bảng dữ liệu (nội dung tập dữ liệu) mà metadata. Một bảng dữ liệu thường có một hàng tiêu đề, theo sau đó là một hoặc nhiều hàng dữ liệu. Hàng tiêu đề bao gồm một tập các lược đồ chỉ mục (tên thuộc tính) mà giá trị thực chất của chúng được lưu trong các hàng dữ liệu. Metadata thường bao gồm tên và mô tả của tập dữ liệu.

Các lược đồ chỉ mục đại diện cho các khái niệm cấp cao không được sử dụng nếu chúng tôi trực tiếp sử dụng chúng để chấm điểm bằng câu truy vấn của người dùng. Hãy xem các ví dụ trong **lược đồ 1** từ vựng của lược đồ chỉ mục có thể khác xa với các trường khác và câu truy vấn của người dùng. “LocationAbbr” viết tắt cho “Location Abbreviation” không chắc xuất hiện trong câu truy vấn của người dùng vì thế tập dữ liệu này ít được gọi lại. Tuy nhiên chúng ta có thể cải thiện tập dữ liệu này bằng cách tạo ra lược đồ chỉ mục như là “place” và “city” xuất hiện trong

Vì thế chúng ta có hàm tính điểm cho tập dữ liệu như sau:

trong đó

là sự kết hợp giữa tiêu đề và mô tả

là dữ liệu của bảng

là các nhãn của biểu đồ đã tạo

Mỗi một trường đều có trọng số Weight tương ứng

Trong khi và có cùng hàm hàm tính điểm là thì có hàm tính điểm là

Với và ta có thể sử dụng hàm tính điểm tiêu chuẩn cho tài liệu bình thường, còn trong nghiên cứu ta dùng hàm BM25 thay cho

Do sự tồn tại một lượng lớn các từ không phải từ điển trong nhãn của lược đồ nên chúng có thể sẽ vượt ra ngoài vốn từ của các từ đã được nhúng