**Trừu tượng:** Tác giả đã đề xuất một mô hình tạo schema labels mới giúp tạo ra các schema labels có thể dựa trên tập dữ liệu bảng nội dung. Chúng tôi kết hợp các schema labels đã tạo thành một hỗn hợp mô hình xếp hạng không chỉ xem xét mức độ liên quan giữa truy vấn và siêu dữ liệu mà còn là sự giống nhau giữa truy vấn và tạo nhãn lược đồ.

Tạo một điểm chuẩn mới dành riêng cho nhiệm vụ truy xuất tập dữ liệu.

Các thử nghiệm cho thấy rằng cách tiếp cận của chúng tôi có thể cải thiện hiệu quả độ chính xác và điểm NDCG của nhiệm vụ truy xuất tập dữ liệu so với phương pháp cơ sở. Chúng tôi cũng thử nghiệm trên một bộ sưu tập các bảng Wikipedia để cho thấy rằng các tính năng được tạo từ schema labels cũng có thể cải thiện nhiệm vụ truy xuất bảng web không được giám sát và giám sát

Trong bài báo này, chúng tôi tập trung vào vấn đề truy xuất tập dữ liệu trong đó nội dung tập dữ liệu ở dạng bảng, vì dữ liệu dạng bảng được sử dụng rộng rãi và dễ đọc và dễ ghi. một tập dữ liệu bao gồm một bảng dữ liệu (nội dung tập dữ liệu) và siêu dữ liệu. Một bảng dữ liệu thường có một hàng tiêu đề, theo sau là một hoặc nhiều hàng dữ liệu. Hàng tiêu đề bao gồm danh sách các schema labels (tên thuộc tính) có giá trị thực được lưu trữ trong các hàng dữ liệu. Siêu dữ liệu thường bao gồm tiêu đề và mô tả của tập dữ liệu.

**Giới thiệu:** Các nhãn lược đồ, đại diện cho các khái niệm cấp cao, không được sử dụng tối đa nếu chúng tôi trực tiếp chấm điểm chúng bằng một truy vấn của người dùng. từ vựng của các schema labels có thể rất khác so với các trường và truy vấn của người dùng khác.

Có 2 vấn đề của 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 vấn đề.
* 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 sử dụng để phân hạng dữ liệu.

**Xếp hạng nâng cao nhãn lược đồ:** nâng cao tập dữ liệu này bằng cách tạo các schema labels xuất hiện trong các tập dữ liệu tương tự khác, có thể cung cấp tín hiệu đối sánh mềm tốt hơn đối với truy vấn của người dùng và do đó tăng cơ hội có thể được thu hồi.

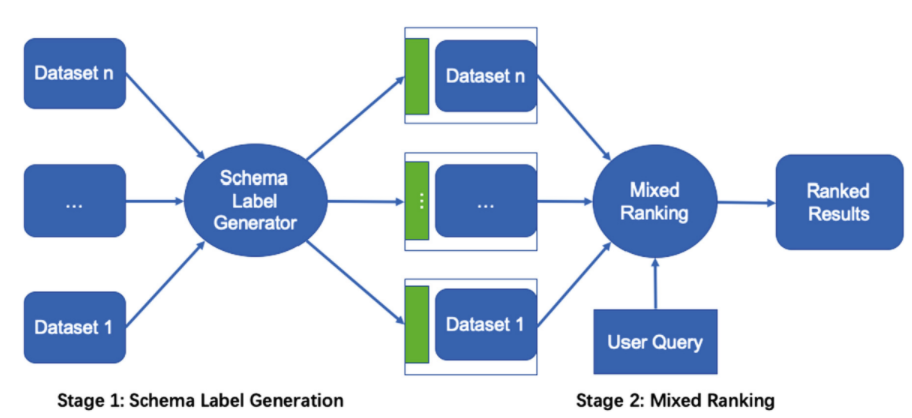
* trước tiên chúng tôi đề xuất một phương pháp mới để tạo nhãn lược đồ. Chúng tôi tìm hiểu các biểu diễn tính năng tiềm ẩn của schema labels một cách tự động bằng cách cùng nhau phân tách ma trận tương tác dataset-schema label và ma trận tương tác schema label-schema label.
* Sau đó, chúng tôi đề xuất một khuôn khổ để tăng cường truy xuất tập dữ liệu bằng cách tạo schema labels để giải quyết vấn đề schema labels không được sử dụng hiệu quả bởi các công cụ tìm kiếm tập dữ liệu hiện có.

tạo một điểm chuẩn công khai mới và sử dụng nó để chứng minh tính hiệu quả của khuôn khổ được đề xuất của chúng tôi để truy xuất bộ dữ liệu.

Ngoài ra, còn xem xét nhiệm vụ truy xuất bảng web và chứng minh rằng các tính năng được tạo từ nhãn giản đồ có thể có hiệu quả đối với xếp hạng được giám sát.

our framework has two stages:

* in the first stage, đào tạo một trình tạo schema labels với phương pháp được đề xuất trong Sect. 3.1 và sử dụng nó để tạo các schema labels bổ sung cho tất cả các tập dữ liệu;
* in the second stage, sử dụng mô hình xếp hạng hỗn hợp để kết hợp điểm số của schema labels các trường khác để xếp hạng tập dữ liệu.



**Tạo nhãn lược đồ:**

Chúng tôi đề xuất cải thiện việc tìm kiếm tập dữ liệu bằng cách sử dụng các schema labels đã tạo, vì chúng có thể bổ sung cho các schema labels ban đầu và đặc biệt có giá trị khi chúng không có trong tập dữ liệu.

đề xuất dự đoán các schema labels dựa trên các đặc điểm được sắp xếp của các giá trị dữ liệu. Thay vì thiết kế các tính năng được sắp xếp cho các nhãn giản đồ, chúng tôi xem xét việc tìm hiểu các biểu diễn của chúng theo cách tự động

Lấy cảm hứng từ các phương pháp lọc cộng tác trong hệ thống đề xuất, chúng tôi lập mô hình từng tập dữ liệu với tư cách là người dùng và từng nhãn giản đồ dưới dạng một mục. Khi đó, một tập dữ liệu có schema labels có thể được coi là phản hồi tích cực giữa người dùng và một mục.

Bằng cách khai thác số lần xuất hiện của user-item và số lần xuất hiện của item-item chúng ta có thể tìm hiểu mức dộ tiềm ẩn biểu diễn của nhãn lược đồ.

Trong phần sau, chúng tôi trình bày cách tạo ma trận ưu tiên trong ngữ cảnh tạo nhãn schema labels và cách tìm hiểu các tính năng của nhãn lược đồ.

**Xây dựng ma trận ưu tiên.** Với m bảng dữ liệu và n schema labels duy nhất, chúng ta có thể xây dựng ma trận ưu tiên cột dữ liệu Mm × n

**Dữ liệu hóa ma trận.** Phân tích nhân tử ma trận là một loại thuật toán lọc cộng tác được sử dụng trong các hệ thống khuyến nghị. Các thuật toán thừa số hóa ma trận hoạt động bằng cách phân rã ma trận tương tác giữa user-item thành tích của hai ma trận hình chữ nhật có kích thước thấp hơn.

Ý tưởng đằng sau việc phân tích nhân tử của ma trận là đại diện cho người dùng và các mục trong một không gian tiềm ẩn có chiều thấp hơn

**Nhúng nhãn**. ới một chuỗi các từ, bạn có thể học một biểu diễn liên tục số chiều thấp được gọi là nhúng từ có thể được học cho mỗi từ. mô hình Word2vec’s skip-gram với việc lấy mẫu phủ định (SGNS) tương đương với việc bao hàm ngầm định ma trận ngữ cảnh từ, có các ô là thông tin tương hỗ theo chiều kim (PMI) của các cặp từ và ngữ cảnh tương ứng, được dịch chuyển bởi một hằng số chung.

**Học chung về các biểu diễn nhãn lược đồ**. Biểu diễn schema labels học được từ MF nắm bắt thông tin tương tác giữa tập dữ liệu và schema labels. Nhóm tác giả sử dụng mô hình CoFactor để cùng tìm hiểu các biểu diễn schema labels từ cả tương tác dataset-label và tương tác label-label

**Tạo nhãn lược đồ**. Sau khi có được các biểu diễn schema labels đã học chung, chúng ta có thể sử dụng chúng làm các features để tạo nhãn lược đồ. Bất kỳ mô hình phân loại nhiều nhãn nào cũng có thể được sử dụng để huấn luyện trình tạo schema labels và trong bài báo này, chúng tôi chọn Rừng ngẫu nhiên.

**3.2 Mô hình xếp hạng hỗn hợp**

Dựa trên phương pháp tạo schema labels được đề xuất ở trên, ta tiến hành lập chỉ mục các schema labels đã tạo cho mỗi tập dữ liệu. Giờ đây, mỗi tập dữ liệu có các trường tiếp theo: siêu dữ liệu, hàng dữ liệu, nhãn giản đồ và schema labels đã tạo. Một cách đơn giản để xếp hạng tập dữ liệu là sử dụng các phương pháp xếp hạng truyền thống cho các tài liệu.

Đối với biểu diễn tài liệu trường đơn lẻ, tập dữ liệu được coi như một tài liệu duy nhất bằng cách nối văn bản từ tất cả các trường. Sau đó, các phương pháp truyền thống như BM25 có thể được sử dụng để chấm điểm tập dữ liệu. Đối với biểu diễn tài liệu đa trường, mỗi trường được tính điểm độc lập với truy vấn và một tổng có trọng số được sử dụng để xếp hạng.

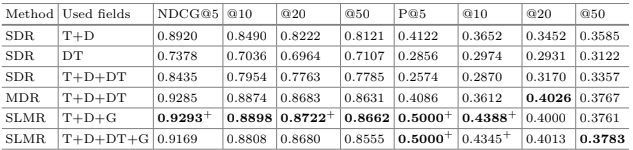
Trong mô hình Xếp hạng hỗn hợp schema labels (SLMR).Trọng tâm của công việc của chúng tôi là tìm hiểu cách các nhãn lược đồ, hàng dữ liệu và siêu dữ liệu khác có thể ảnh hưởng khác nhau đến hiệu suất truy xuất tập dữ liệu. Lưu ý rằng, để đơn giản, chúng tôi coi siêu dữ liệu khác (tiêu đề và mô tả) là một trường văn bản duy nhất, vì tiêu đề và mô tả là đồng nhất được so sánh với schema labels và hàng dữ liệu.

**4 Thu thập dữ liệu**

Ở Phần này, nhóm tác giả miêu tả cách xây dựng bộ benchmark mới cho việc truy hồi thông tin. Bộ dữ liệu đã thu thập được gồm 2417 tài nguyên từ Data.gov, gồm nhiều chủ đề khác nhau. Mỗi bảng CSV được coi là một tập dữ liệu duy nhất và sử dụng siêu dữ liệu cấp tài nguyên để chú thích từng tập dữ liệu.

**Đánh giá mức độ liên quan**

Đối với mỗi nhiệm vụ và mỗi truy vấn được đề xuất, chúng tôi đã sử dụng các hàm xếp hạng truyền thống để chấm điểm các đại diện trường đơn lẻ của mỗi tập dữ liệu và thu thập 100 kết quả hàng đầu. Các mô hình xếp hạng sau đã được sử dụng: **BM25, TF-IDF**, Mô hình ngôn ngữ dựa trên làm mịn Jelinek-Mercer và Mô hình ngôn ngữ với làm mịn Dirichlet. Chúng tôi cũng sử dụng mỗi mô hình với hai cách biểu diễn khác nhau: ghép nối tất cả các trường của tập dữ liệu và ghép nối tiêu đề và mô tả.



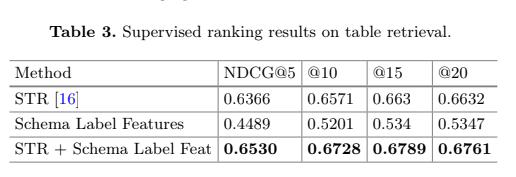
**Đánh giá**

Chúng tôi đánh giá hiệu suất truy xuất tập dữ liệu qua một loạt các chỉ số: Độ chính xác ở k và Thu nhập tích lũy chiết khấu được chuẩn hóa ở k. Để kiểm tra mức độ ý nghĩa của sự khác biệt giữa các hoạt động của mô hình, chúng tôi sử dụng phép t - test được ghép nối ở mức ý nghĩa p = 0,01.

**Kết quả thử nghiệm**

Từ kết quả xếp hạng document theo trường đơn, chúng ta có thể thấy rằng việc chỉ sử dụng bảng dữ liệu để xếp hạng dẫn đến hiệu suất kém nhất. Cho điểm dựa vào sự kết hợp giữa tiêu đề và mô tả đạt được kết quả tốt nhất, điều này cho thấy rằng tiêu đề và mô tả quan trọng hơn bảng dữ liệu để để xếp hạng tập dữ liệu (Q3). Việc coi tất cả các trường của tập dữ liệu như như một single-field document cung cấp hiệu suất giữa 2 mô hình trước đó. Kết quả này được mong đợi vì độ dài của bảng dữ liệu thường lớn hơn nhiều so với tiêu đề và mô tả, và do đó chiếm ưu thế trong việc biểu diễn bảng.

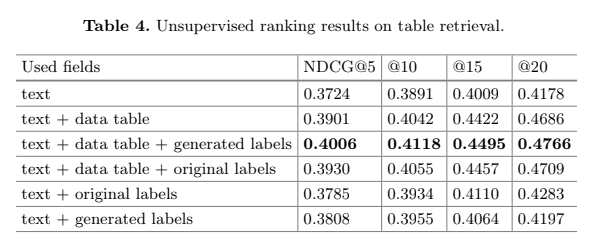
Bằng cách so sánh kết quả xếp hạng tài liệu trường đơn và tài liệu đa trường, chúng ta nhận thấy rằng sự kết hợp điểm số của bảng dữ liệu, tiêu đề và mô tả có thể cải thiện NDCG@k. Mặc dù NDCG@k giảm khi k tăng, mối quan hệ cải thiện so với xếp hạng single-field là đáng kể hơn. Ngược lại, đối với Precision@5, Precision@10, single-field document hoạt động tốt hơn multifield document ranking mặc dù sự khác biệt là nhỏ. Vì vậy, đối với Q1, dưới sự thiết lập của multifield document ranking, nội dung của bảng dữ liệu có thể giúp NDCG, nhưng không giúp ích cho Precision của của kết quả truy xuất tập dữ liệu. Với các schema labels đã tạo, Mô hình xếp hạng có thể có hiệu suất cao hơn đối với tác vụ truy xuất tập dữ liệu



**Tìm kiếm nâng cao tao Schema Label cho Web Table**

Chúng ta có thể thấy rằng các tính năng schema labels không thể vượt trội hơn STR. Nhưng kết hợp chúng sẽ dẫn đến sự cải thiện. Tuy nhiên, bằng cách tính toán mức độ quan trọng của đối tượng được chuẩn hóa được đo bằng điểm số Gini, chúng tôi thấy rằng đối với STR có các tính năng nhãn lược đồ, phép đo dựa trên WMD đóng góp nhiều nhất trong số tất cả các tính năng ngữ nghĩa. Do đó, nó chứng tỏ rằng các schema labels cũng có thể có giá trị cho nhiệm vụ truy xuất bảng.

 phương pháp tạo schema labels của chúng tôi có thể tốt hơn STR trong trường hợp mô tả văn bản cung cấp thông tin ít hữu ích hơn chính bảng.



**Phần kết luận**

Trong bài báo này, chúng tôi đã đề xuất một khung xếp hạng nâng cao schema labels (schema label enhanced ranking) để truy xuất tập dữ liệu. Khung có hai giai đoạn: trong giai đoạn đầu, một trình tạo schema labels được huấn luyện để tạo thêm các schema labels cho mỗi cột tập dữ liệu; trong giai đoạn thứ hai, với một truy vấn của người dùng, các tập dữ liệu được xếp hạng theo các trường ban đầu của chúng cùng với các schema labels đã tạo. Tạo schema labels được coi như một nhiệm vụ phân loại nhiều nhãn, trong đó mỗi cột của tập dữ liệu được liên kết với nhiều nhãn lược đồ. Thay vì sử dụng các tính năng được quản lý thủ công, chúng tôi học các biểu diễn tính năng tiềm ẩn của schema labels bằng mô hình CoFactor trong đó các tương tác schema label-schema label dữ liệu và tương tác dataset-schema label được ghi lại. Với mô hình xếp hạng hỗn hợp nhãn lược đồ, điểm xếp hạng truyền thống cho các trường văn bản (tiêu đề, mô tả, hàng dữ liệu) và điểm dựa trên nhúng từ cho các schema labels đã tạo có thể được sử dụng để xếp hạng tập dữ liệu.

Chúng tôi đã tạo một điểm chuẩn mới để đánh giá hiệu suất của việc truy xuất tập dữ liệu. Kết quả thử nghiệm cho thấy khung được đề xuất của chúng tôi có thể cải thiện hiệu quả hiệu suất của nhiệm vụ truy xuất tập dữ liệu. Nó đạt được NDCG cao nhất trên tất cả các mức cắt giảm xếp hạng so với tất cả các phương pháp cơ sở. Chúng tôi cũng áp dụng phương pháp của mình cho tác vụ truy xuất bảng web tương tự như tìm kiếm tập dữ liệu và nhận thấy rằng các tính năng được tạo từ nhãn giản đồ cũng có thể giúp ích trong supervised ranking.