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 mô tả văn bản. Chúng tôi đề xuất một mô hình chỉ mục mà có thể tự tạo ra Schema Label dựa trên nội dung của bảng chứa tập dữ liệu. Chúng tôi kết hợp các Schema Label 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 Schema Label 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 Schema Label 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 mọi 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 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. 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 vì 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) và 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 Schema Label (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 Schema Label 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** bảngtừ vựng của Schema Label có thể khác xa với các trường khác và câu truy vấn của người dùng. “LocationAbbr” viết tắt của “Location Abbreviation” (địa chỉ viết tắt? ...) không thường xuất hiện trong câu truy vấn của người dùng vì thế tập dữ liệu này ít có khả năng được ghi nhớ. Tuy nhiên chúng ta có thể cải thiện tập dữ liệu này bằng cách tạo ra Schema Label ví dụ như là “place” và “city” xuất hiện trong các tập dữ liệu tương tự khác, điều này có thể cung cấp tín hiệu soft-matching tốt hơn mà vẫn tôn trọng câu truy vấn của người dùng, và vì thế tăng khả năng được ghi nhớ.

Trong nghiên cứu này, tước tiên chúng tôi đề xuất một phương pháp mới cho việc tạo Schema Label. Chúng tôi nghiên cứu cách biểu diễn chức năng ẩn của Schema Label một cách tự động bằng cách đồng thời phân tích ma trận tương tác giữa dataset với schema label và ma trận tương tác giữa schema label với schema label. Sau đó chúng tôi sẽ cung cấp một framework để cải thiện việc truy hồi dữ liệu bằng cách tạo ra schema label mới để giải quyết vấn đề của schema label cũ là không được sử dụng hiệu quả bởi các công cụ tìm kiếm. Chúng tôi còn tạo ra một tiêu chuẩn chung dựa trên bộ dữ liệu của (U.S.) và sử dụng nó để chứng minh hiệu quả của framework trên cho việc truy xuất dữ liệu. Chúng tôi cũng xem xét thêm các tác vụ truy xuất web table để chứng mình rằng các thuộc tính được tạo từ schema labels có thể có hiệu quả trong việc giám sát xếp hạng

**2, Các công việc liên quan**

Việc tìm kiếm tập dữ liệu đã trở thành một lĩnh vực nghiên cứu mới với các thử thách mới. Chapman và các cộng sự đã chia việc tìm kiếm tập dữ liệu thành “basic” và “constructive”. Tìm kiếm theo phương pháp basic trả về một danh sách các tập dữ liệu hiện có dựa trên câu truy vấn của người dùng trong khi tìm kiếm theo phương pháp constructive tạo ra tập dữ liệu một cách nhanh chóng dựa trên nhu cầu và thắc mắc của người dùng. Google gần đây đã đưa ra một dịch vụ tìm kiếm tập dữ liệu, cũng như rất nhiều nguồn dữ liệu khác dịch vụ của họ cũng dựa trên metadata của tập dữ liệu và được chú thích trên trang web sử dụng tiêu chuẩn do schema.org đưa ra.

Các công việc khác dựa trên ứng dụng của Web tables cũng có liên quan tới công việc của chúng tôi. Cafarella và các cộng sự đã đề xuất một hệ thống WebTables khai thác các Web tables từ các trang web đứng đầu theo từ khoá tìm kiếm. Sekhavat và các cộng sự đưa ra một phương pháp xác xuất để tăng cường cơ sở kiến thức hiện có với các sự việc từ Web tables. Zhang và các cộng sự đã nghiên cứu tạo ra mô hình xác suất để trang bị cho các bảng tính khả năng hỗ trợ thông minh. Đặc biệt hơn nữa, khi đưa cho chúng một bảng, chúng sẽ đề xuất thêm các hàng và cột tiêu đề bằng cách tận dụng thông tin từ Webtables. Họ cũng phát triển tính năng so sánh ngữ nghĩa cho bảng truy vấn.

Các kĩ thuật được thiết kế cho việc phân tích Web table có thể được áp dụng cho việc tìm kiếm tập dữ liệu. Trong công việc của chúng tôi, mỗi tập dữ liệu đều được kết hợp với dữ liệu dạng bảng. Việc trích xuất các thông tin hữu ích từ các bảng như là mối quan hệ và tên của các thuộc tính có thể giúp ích cho việc truy xuất. Trabelsi và các cộng sự gần đây đã đưa ra một cách dán nhãn tuỳ chỉnh cho các cột dựa trên nhiều ngữ cảnh trong bảng truy xuất và họ nhận thấy sự xuất hiện của các ô có giá trị dạng số có thể hữu dụng. Zhang và các cộng sự đưa ra khái niệm ngữ nghĩa để miêu tả các câu truy vấn và các bảng cho việc xếp hạng các bảng entity-focused. Tuy nhiên việc tìm kiếm tập dữ liệu vốn dĩ đã khó hơn kể từ khi tập dữ liệu không cần phải tập trung vào thực thể.

**3. Tăng cường xếp hạng Schema Label** (dài k dịch hết, bỏ qua)

Trong bài báo này, chúng tôi giới thiệu framework của schema label đã được tăng cường hiệu suất truy xuất thông tin. Như đã được thể hiện trong **hình 2,** frameworkcủa chúng tôi có hai giai đoạn: Trong giai đoạn đầu tiên chúng tôi huấn luyện một bộ sinh schema label với phương pháp được đưa ra trong **phần 3.1** và sử dụng chúng để tạo ra thêm các schema label cho toàn bộ tập dữ liệu. Trong giai đoạn hai chúng tôi sư dụng mô hình xếp hạng hỗn hợp để kết hợp điểm số của các schema labels và các trường khác để xếp hạng tập dữ liệu. Trong các phần sau chúng tôi sẽ trình bày chi tiết hai giai đoạn trên.

**3.1 Bộ sinh Schema Label**

Chúng tôi đưa ra cách tăng cường hiệu suất truy hồi dữ liệu bằng việc sử dụng các schema labels đã tạo vì chúng có thể bổ sung cho các schema label ban đầu và đặc biệt có giá trị khi xuất hiện trong tập dữ liệu.

Chúng tôi xử lý bộ sinh Schema Label giống như bài toán phân lớp.

Đặt L = {l1, l2, ..., lk} là các nhãn xuất hiện trong toàn bộ tập dữ liệu

D = {(xi, yi) |1 ≤ i ≤ n} là tập dữ liệu huấn luyện

Mỗi một mẫu huấn luyện (xi, yi), xi là một vector d chiều có thể được tính toán từ các dòng dữ liệu hoặc học được từ mà trận factorization được đưa ra ở phần sau.

yi là một vector k chiều với khi xi phù hợp với nhãn lj còn lại

Mục tiêu của chúng tôi là tìm ra một hàm có dạng P(l|xi), (l ∈ L). Để tạo ra m schema labels cho cột I chúng tôi chọn m nhãn Lm tốt nhất theo công thức:



Chúng tôi cũng tạo ra schema labels bằng việc lựa chọn ngẫu nhiên các ngưỡng θ:



Trong thực tế, chúng tôi có thể tạo ra m schema labels tốt nhất và lọc ra các kết quả với xác suất thấp hơn ngưỡng

Chen và các cộng sự đã đưa ra dự đoán schema labels dựa trên các thuộc tính được sắp xếp. Thay vì thiết kế thuộc tính được sắp xếp cho schema labels, chúng tôi xem xét việc nghiên cứu cách biểu diễn của chúng một cách tự động. Lấy cảm hứng từ việc kết hợp các phương pháp lọc trong hệ thống đề xuất, chúng tôi tạo mô hình với mỗi tập dữ liệu như là người dùng và với mỗi schema label như là một hạng mục. Sau đó một tập dữ liệu với schema label có thể được coi như một đánh giá tích cực giữa người dùng và hạng mục. Bằng cách khai thác số lần cùng xuất hiện của người dùng-hạng mục và hạng mục – hạng mục, chúng tôi có thể nghiên cứu sự xuất hiện ẩn của các schema label. Trong phần tiếp theo chúng tôi sẽ trình bày cách để xây dựng một ma trận ưu tiên trong từng trường hợp của bộ sinh schema label và cách để nghiên cứu các thuộc tính trong label features.

Xây dựng ma trận ưu tiên. Với m bảng dữ liệu và n schema labels khác nhau, chúng ta có thể xây dựng một tập dữ liệu cột ma trận ưu tiên Mm x n trong đó Mup là 1 nếu tập dữ liệu u chứa schema label p.

Dữ liệu hoá ma trận. MF chia M thành tích của Um x k và Pk x n

trong đó k < min (m, n).

UT = (α1, ..., αu ..., αm) với αu Rk là hệ số của vector tiềm ẩn của tập dữ liệu u.

PT = (β1, ..., βp ..., βn) với βp ∈ Rk là hệ số của vector tiềm ẩn của schema label p.

Vì ma trận ưu tiên tạo ra mô hình phản hồi ngầm nên MF được điều chỉnh theo hàm sau:

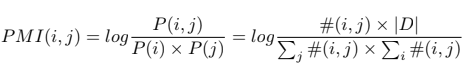


trong đó

cup là siêu tham số dùng để diều chỉnh sự cân bằng các giá trị 0 và khác 0 vì M là ma trận thưa (có rất nhiều giá trị 0)

λα và λβ là các tham số chính quy điều chỉnh tầm quan trọng của và

Dán nhãn. Gần đây các công nghệ nhúng (như word2vec….) rất có giá trị trong việc xử lý ngôn ngữ tự nhiên. Cho một tập các từ, việc biểu diễn sự liên tục ít chiều được gọi là dán nhãn từ, việc này có thể được học từ mỗi từ. Mô hình Word2vec’s skip-gram with negative sampling (SGNS) là mô hình đi tìm thừa số của ma trận từ - ngữ cảnh, trong đó các ô là các pointwise mutual information (PMI) của các cặp từ và ngữ cảnh tương ứng, thay đổi bởi một hằng số chung. Phương pháp PMI giữa từ I và ngữ cảnh j của nó được xác định theo công thức:



Trong đó:

# (i, j) là số lần xuất hiện của từ j trong context window của từ i

|D| là tổng số cặp từ - ngữ cảnh

Từ công thức trên ta có thể suy ra SPPMI của từ I và từ j là:



Trong đó:

k là số negative samples của SGSN

với một tập văn bản cho trước ma trận MSPPMI có thể được tính dựa trên Eq

Schema label tồn tại trong ngữ cảnh của schema label khác. Vì thế với một schema label cho trước, tất cả các schema label khác xuất hiện từ cùng bảng dữ liệu đề được coi như cùng ngữ cảnh. Với ma trận SSPMI đã được xây dựng từ cùng các schema labels, chúng tôi có thể phân tích chúng để nghiên cứu sự xuất hiện ẩn của các schema labels.

Joint Learning of Schema Label Representations. Các Schema label đại diện đã được nghiên cứu từ cách nắm bắt thông tin tương tác MF giữa tập dữ liệu và schema label, trong khi phương pháp word2vec giải thích mối quan hệ của các shcema label. Chúng tôi sử dụng mô hình CoFactor để kết hợp việc nghiên cứu sự xuất hiện của các schema label từ cả sự tương tác giữa tập dữ liệu – nhãn và nhãn -nhãn.

**4. Thu thập dữ liệu**

Ở đây chúng tôi mô tả chi tiết cách xây dựng phép đo mới cho việc truy hồi dữ liệu. Chúng tôi đã thu thập 2417 bộ dữ liệu từ chính phủ liên bang Hoa Kỳ trên trang Data.gov với rất nhiều chủ đề. Mỗi một bộ dữ liệu bao gồm một hoặc nhiều bảng với định dạng CSV và các metadata tương ứng. Mỗi một bảng CSV được coi là một tập dữ liệu đơn và chúng tôi sử dụng metadata để chú thích cho mỗi tập dữ liệu đó.

4.1 Tạo ra tác vụ và thu thập câu truy vấn

Chúng tôi tạo ra sáu tác vụ trong đó mỗi một tác vụ mô tả một thông tin rời rạc và cần được tìm thêm trong một hoặc nhiều tập dữ liệu. Với mỗi chuỗi thông tin rời rạc đó ta có một câu lệnh về thông tin cần thiết để tập dữ liệu đó được coi là có liên quan đến tác vụ. Chúng tôi còn xác nhận lại rằng có ít nhất một tập dữ liệu liên quan trong mỗi tác vụ được đưa ra. Tập dữ liệu được cung cấp công khai.

Chúng tôi sử dụng Amazon Mechanical Turk để có được những câu truy vấn đa dạng cho các tác vụ phía trên từ người dùng thực. Mọi ghi chú đều được thể hiện với mô tả tác vụ và cần được cung cấp một câu truy vấn cho mỗi tác vụ đã tạo. Để tránh ảnh hưởng từ thứ tự các tác vụ đến chất lượng ghi chú, chúng tôi đã xáo trộn các tác vụ cho từng ghi chú một cách ngẫu nhiên. Chúng tôi trả 1 dolar cho mỗi lần công việc ghi chú được hoàn thành và 20 câu truy vấn được thu thập sau mỗi tác vụ. Mỗi một câu truy vấn sau khi được thu thập được kiểm tra thủ công để loại bỏ các câu truy vấn không liên quan khỏi tập hợp kết quả.

4.2 Đánh giá mức độ liên quan

Với mỗi tác vụ và mỗi câu truy vấn được đề xuất, chúng tôi sử dụng hàm xếp hạng truyền thống để chấm điểm các trường đơn đại diện cho mỗi tập dữ liệu và thu thập 100 kết quả tốt nhất. Các mô hình xếp hạng đã được sử dụng là: BM25, TF-IDF, mô hình ngôn ngữ dựa trên phương pháp Jelinek-Mercer smoothing, và mô hình ngôn ngữ dựa trên phương pháp Dirichlet smoothing. Chúng tôi cũng sử dụng mỗi mô hình với hai đại diện khác nhau: Sự kết hợp tất cả các trường trong tập dữ liệu và sự kết hợp của tiêu đề và mô tả. Điều này dẫn đến 8 đường cơ sở cho kết quả.

Sau đó, các cặp tác vụ – tập dữ liệu liên quan đã được ghi chú lại để sử dụng cho dịch vụ cung cấp cộng đồng được cung cấp bởi Figure Eight. Chúng tôi không ghi chú lại các cặp câu truy vấn – dữ liệu vì mục tiêu của việc truy xuất dữ liệu là tìm ra các tập dữ liệu liên quan cho nhu cầu thông tin thực sự.

Các ghi chú được thể hiện ngay trong tiêu đề, mô tả, liên kết đến bảng dữ liệu của tác vụ. Mỗi một cặp tác vụ-dữ liệu được đánh giá dựa trên 4 mức: 0 (sai chủ đề), 1 (nghèo nàn), 2 (tốt), 3 (xuất sắc). Mỗi một người ghi chú được trả 10 cents cho mỗi lần đánh giá cặp tác vụ-dữ liệu

Mỗi một cặp tác vụ-dữ liệu đơn được đánh giá bởi 3 người và chúng tôi căn cứ vào phần đông biểu quyết để gắn nhãn cho chúng. Nếu không đạt được thoả thuận trên chúng tôi sẽ tính điểm số trung bình để đánh giá nhãn của chúng. Số liệu thống kê ghi chú được thể hiện trong bảng 1

5. Đánh giá

5.1 Số liệu đánh giá

Chúng tôi đánh giá hiệu xuất truy hồi dữ liệu thông qua số liệu: Dộ chính xác tại k và Normalized Discounted Cumulative Gain (NDCG) tại k. Để kiểm tra độ hiệu quả giữa các mô hình khác nhau chúng tôi sử dụng phương pháp t-test với p = 0.01

5.2 Đường cơ sở

Đầu tiên chúng tôi trình bày về đường cơ sở của các phương pháp truy xuất

Phương pháp Single-Field Document Ranking (SDR). Tập dữ liệu được coi như một tài liệu duy nhất. Chúng tôi sử dụng mô hình xếp hạng BM25 để đánh giá sự liên quan giữa tiêu đề và mô tả, nội dung văn bản của bảng dữ liệu và sự liên kết của tất cả chúng. Bằng cách so sánh ba kết quả này chúng tôi có thể nghiên cứu về mức độ quan trọng của các trường trong việc truy hồi thông tin. Các tham số được lựa chọn bằng phương pháp grid search

Phương pháp Multifield Document Ranking (MDR). Bằng việc đặt wl = 0, Eq đã chia nhỏ thành các mô hình Mixture of Language. Phương pháp xếp hạng BM25 cũng được sử dụng ở đây: scoretext () để có sự so sánh công bằng với các phương pháp khác. Để điều chỉnh weight của các trường chúng tôi sử dụng coordinate ascent. Cuối cùng các thông số làm mịn được tối ưu giống như phương pháp SDR

5.3 Kết quả thực nghiệm

Trong phần này chúng tôi sẽ giải quyết các câu hỏi sau đây:

Q1: Bảng nội dung dữ liệu có giúp ích trong việc truy hồi thông tin không?

Q2: Bộ sinh schema label có giúp ích trong việc truy hồi thông tin không?

Q3: Trường nào là quan trọng nhất trong tác vụ truy xuất thông tin?

Đầu tiên chúng tôi thu được các thuộc tính của schema label như đã nói trong phần 3.1 và hệ số ẩn được đặt là 40. Sau đó chúng tôi huấn luyện mô hình Random Forest với những gì đã học được từ các thuộc tính của schema label. Thư viện scikit-learn triển khai thuật toán Random Forest với các tham số mặc định ngoại trừ số lượng cây được sửa là 25. Trong thực tế, chúng tôi có thể chọn rất nhiều thuật toán phân lớp khác. Với mỗi cột chúng tôi chọn 10 schema label tốt nhất và bỏ đi các kết quả có xác xuất thấp hơn 0.5. Với mỗi tập dữ liệu chúng tôi lập chỉ mục các schema label như một trường bổ sung. Bảng 2 đã ghi lại NDCG và độ chính xác tại k của các mô hình khác nhau. Lưu ý rằng với Schema Label Mixed Ranking (SLMR), chúng tôi huấn luyện ba mô hình khác nhau và các trọng số weight của các trường đã sử dụng phải khác 0 để nghiên cứu các câu hỏi được đề xuất. Trọng số của các trường được sử dụng để biểu diễn tài liệu nhiều trường cũng được đặt khác 0 khi tối ưu hóa các tham số.

Từ kết quả của Single-Field Document Ranking, chúng tôi có thể thấy rằng việc chỉ sử dụng bảng dữ liệu để việc đánh giá xếp hạng cho hiệu suất kém nhất. Tuy nhiên việc chấm điểm theo sự kết hợp giữa tiêu đề và mô tả đạt kết quả tốt nhất, điều này chỉ ra rằng tiêu đề và môt tả quan trọng hơn bảng dữ liệu trong việc xếp hạng dữ liệu. (Q3) Việc xử lý tất cả các trường của tập dữ liệu như một tập dữ liệu đơn trường cung cấp hiệu xuất tốt hơn hai mô hình trước đó. Đây là kết quả được mong đợi vì độ dài của bảng dữ liệu thường lớn hơn tiêu đề và mô tả vì thế chiếm ưu thế trong việc biểu diễn bảng.

Bằng việc so sánh kết quả xếp hạng của tài liệu đơn trường và đa trường chúng ta có thể thấy sự kết hợp điểm số của bảng dữ liệu, tiêu đề và mô tả có thể cải thiện NDCG@k. Thông qua việc tăng k lên thì NDCG@k giảm, xếp hạng các tài liệu đơn trường được cải thiện tốt hơn. Ngược lại với Precision@5, Precision@10, xếp hạng các tài liệu đơn trường cao hơn các tài liệu đa trường mặc dù sự khác biệt là rất nhỏ. Vì vậy với câu hỏi Q1, dưới thiết lập xếp hạng các tài liệu đa trường nội dung của bảng dữ liệu có thể cải thiện NDCG nhưng không cải thiện độ chính xác của kết quả truy hồi dữ liệu.

Ngoài việc đánh giá các bảng dữ liệu, schema label mixed của chúng tôi đã đạt được điểm NDCG cao nhất trên tất cả giới hạn về thứ hạng. Điều đó chỉ ra rằng các schema label đã tạo có thể hữu ích trong việc cải thiện điểm NDCG của kết quả truy xuất dữ liệu (Q2). Mặc dù Precision@20 của xếp hạng tài liệu đa trường cao hơn những mô hình mà chúng tôi đưa ra, nhưng chênh lệch không quá 0.4% (p value > 0.9). Quan trọng hơn là mô hình của chúng tôi vượt trội hơn 21.3% so với mô hình Precision@5 ( và 20.1% so với Precision@10 ( và còn có đường cơ sở tốt nhất (p\_value < 0.01). Dù các bảng dữ liệu có được chấm điểm hay không thì phương pháp Precision@k vẫn không có sự khác biệt đáng kể đối với xếp hạng schema label maxed. Vì vậy dưới thiếp lập của xếp hạng schema label mixed các bảng dữ liệu có rất ít tác dụng trong trường hợp này (Q1). Một lý do nữa có thể là các bảng dữ liệu được thu thập từ data.gov có quá nhiều giá trị dạng số và chúng hiếm khi phù hợp với câu truy vấn của người dùng

Nếu mô hình xếp hạng schema label mixed chỉ chấm điểm trên tiêu đề và mô tả (wl=0) thì nó sẽ giống với mô hình xếp hạng đơn trường. Vì thế chúng ta có thể gộp kết quả ở dòng 1 và dòng 5 trong bảng 2. Với các schema label đã tạo, mô hình xếp hạng có thể có hiệu quả tốt hơn trong tác vụ truy hồi dữ liệu (Q2)

**5.4 Schema label tạo ra tìm kiếm nâng cao cho Web Tables**

Tác vụ tìm kiếm dữ liệu tương tự với tìm kiếm Web table vì chúng cùng sử dụng bảng thành phần để thể hiện dữ liệu. Khác nhau duy nhất đó là số lượng lớn các Web table là các thực thể tập trung và chúng chứa rất nhiều các thực thể được đặt tên liên kết với kiến thức cơ bản. Tuy nhiên tập dữ liệu của chúng tôi được thu thập từ trang gov.data có nhiều thực thể hữu ích cho bảng thành phần. Vì vậy rất nhiều phương pháp được thiết kế cho việc xếp hạng web table không thể áp dụng với tác vụ tìm kiếm dữ liệu. Phương phápThe semantic table retrieval (STR) được đưa ra bởi Zhang và Balog dựa trên các thuộc tính từ hiểu biết cơ bản (bag of entities) và chúng không có sẵn trong quá trình tìm kiếm dữ liệu. Tuy nhiên schema label được tạo ra dựa trên phương pháp trên có thể được áp dụng vào việc tìm kiếm bảng. Do vậy chúng tôi đã thực hiện các thử nghiệm để có thể thấy hiệu suất phương pháp của chúng tôi trong quá trình tìm kiếm bảng.

Đầu tiên chúng tôi tạo ra các schema label cho tập các bảng được chia sẻ bởi Zhang và Balog sử dụng phuoeng pháp được đưa ra trong phần 3.1. Sau đó chúng tôi thêm 5 thuộc tính vào các tính năng đã được đề xuất của họ dựa trên schema label. Mỗi một thuộc tính có thể là một loại tương đồng giữa câu truy vấn và schema label. Bốn thuộc tính được tính toán sử dụng phép đo đưa ra bởi Zhang và Balog (một thuộc tính hợp nhất trước, 3 thuộc tính hợp nhất sau) và thuộc tính cuối cùng là phủ định của Word Mover’s Distance. Cuối cùng, giống như Zhang và Balog chúng tôi sử dụng thuật toán Random Forest để thực hiện hồi quy và kết quả là trung bình sau 5 lần chạy 5-fold cross-validation được thể hiện trong bảng 3

Chúng ta có thể thấy các thuộc tính của schema label không thể vượt quá STR, nhưng kết hợp chúng lại sẽ có kết quả tốt hơn. Tuy vậy bằng việc tính toán các thuộc tính được chuần hoá theo phương pháp chấm điểm Gini chúng tôi nhận thấy với mỗi STR và các thuộc tính của schema label, phép đo dựa trên WMD có ý nghĩa rất lớn trong các tính năng semantic. Điều này chứng tỏ các schema label cũng có thể có giá trị với tác vụ truy xuất bảng.

Đáng chú ý trong tập các bảng này, rất nhiều bảng thiếu nội dung nhưng lại có mô tả rất chi tiết, điều này có thể không công bằng khi tạo schema label. Trong khi đó với việc tìm kiếm tập dữ liệu, mỗi một bảng đều có giá trị nhưng lại thiếu mô tả chi tiết. Chúng tôi tin rằng phương pháp tạo ra schema label của chúng tôi có thể tốt hơn STR trong viễn cảnh mô tả cung cấp ít thông tin hơn bản thân các bảng.

Chúng tôi cũng đưa ra kết quả xếp hạng không giám sát với Eq trong bảng 4. Không như Zhang và Balog, chúng tôi xem xét từng tiêu đề và các phần viết hoa như là đối với một trường đơn để giảm số lượng siêu tham số (field weights). Kết quả phía trên cho thấy rằng các nhãn được tạo hiệu quả hơn các nhãn gốc trong việc xếp hạng bảng. Không quá ngạc nhiên vì các nhãn được tạo thường không chỉ bao gồm các nhãn gốc mà còn có các nhãn bổ sung để mang lại lợi ích cho mô hình xếp hạng. Chúng tôi cũng nhận thấy rằng việc thêm các trường trong bảng dữ liệu sẽ đạt được kết quả tốt hơn so với việc thêm, điều này trái với kết quả xếp hạng tập dữ liệu. Đây cũng là điều đã được dự đoán vì WikiTables tập trung vào các thực thể và có rất nhiều thông tin dạng văn bản trong khi bảng dữ liệu từ data.gov có rất nhiều giá trị số.

**6. Tổng kết**

Trong bài báo này, chúng tôi đã đưa ra một framework cho việc xếp hạng nâng cao schema label trong việc truy hồi dữ liệu. Framework có 2 phần: trong phần đầu bộ sinh schema label đã được huấn luyện để tạo ra thêm các schema label cho mỗi cột của tập dữ liệu; trong phần thứ hai khi đưa ra một câu truy vấn, tập dữ liệu sẽ được xếp hạng dựa trên các trường của chính nó cùng với các schema label đã được tạo. Quá trình tạo schema label dựa trên nhiệm vụ phân loại nhiều nhãn trong đó mỗi cột của tập dữ liệu có liên quan đến nhiều schema label. Thay vì sử dụng các thuộc tính được quản lý thủ công, chúng tôi nghiên cứu sự xuất hiện của các tính năng ẩn bằng mô hình CoFactor, trong đó các tương tác giữa dataset-schema label và schema label-schema label được ghi lại. Với mô hình xếp hạng schema label mixed, điểm xếp hạng truyền thống cho các trường dạng chữ (tiêu đề, mô tả, dòng dữ liệu) và điểm word embedding-based cho việc tạo schema label có thể được dùng để xếp hạng tập dữ liệu.

Chúng tôi đã tạo một chuẩn mới để đánh giá hiệu suất của việc truy xuất tập dữ liệu. Kết quả thử nghiệm cho thấy framework được chúng tôi đề xuất có thể cải thiện hiệu quả hiệu suất tác vụ truy hồi dữ liệu. Nó đã đạt được NDCG cao nhất trên tất cả các xếp hạng rút gọn so với các phương pháp cơ sở. Chúng tôi cũng áp dụng phương pháp này vào việc truy xuất Web table tương tự như tìm kiếm tập dữ liệu và nhận thấy các thuộc tính được tạo ra từ schema label có thể giúp ích trong việc xếp hạng có giám sát.

Tài liệu này dựa trên công trình được hỗ trợ bởi Quỹ Khoa học Quốc gia theo Khoản tài trợ số IIS-1816325.

Tóm tắt

Bài toán

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 mô tả văn bản.

Hạn chế

Việc truy xuất dữ liệu đang ngày càng nhận được nhiều sự quan tâm của mọi 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 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.

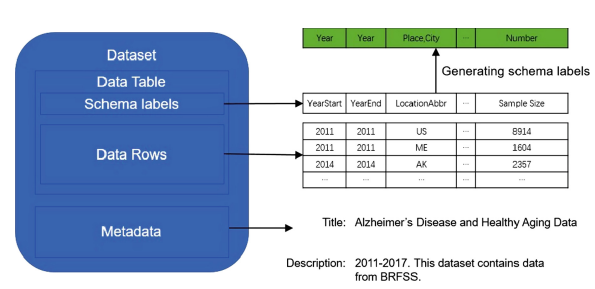
Một người dùng có thể không biết cấu tạo của tập dữ liệu tiềm năng, hoặc các thẻ dữ liệu mà nhà xuất bản cung cấp cùng với tập dữ liệu. Các thông tin như vậy khó có thể được sử dụng trong xếp hạng tập dữ liệu

Giải pháp

Chúng tôi đề xuất một mô hình chỉ mục mà có thể tự tạo ra Schema Label dựa trên nội dung của bảng chứa tập dữ liệu. Chúng tôi kết hợp các Schema Label 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 Schema Label đã được tạo

Trọng tâm bài toán:

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 vì dữ liệu dạng bảng được sử dụng rộng rãi và rất dễ để đọc và viết.



Cấu trúc của tập dữ liệu

Các công việc liên quan

Một số nghiên cứu đi trước:

* Chapman và các cộng sự đã chia việc tìm kiếm tập dữ liệu thành “basic” và “constructive”.
* Google gần đây đã đưa ra một dịch vụ tìm kiếm tập dữ liệu dựa trên metadata của tập dữ liệu
* Cafarella và các cộng sự đã đề xuất một hệ thống WebTables khai thác các Web tables từ các trang web đứng đầu theo từ khoá tìm kiếm
* Sekhavat và các cộng sự đưa ra một phương pháp xác xuất để tăng cường cơ sở kiến thức hiện có với các sự việc từ Web tables.
* Zhang và các cộng sự đã nghiên cứu tạo ra mô hình xác suất để trang bị cho các bảng tính khả năng hỗ trợ thông minh
* Trabelsi và các cộng sự gần đây đã đưa ra một cách dán nhãn tuỳ chỉnh cho các cột dựa trên nhiều ngữ cảnh trong bảng truy xuất và họ nhận thấy sự xuất hiện của các ô có giá trị dạng số có thể hữu dụng
* Zhang và các cộng sự đưa ra khái niệm ngữ nghĩa để miêu tả các câu truy vấn và các bảng cho việc xếp hạng các bảng entity-focused.

Trong nghiên cứu của chúng tôi mỗi tập dữ liệu đều được kết hợp với dữ liệu dạng bảng. Việc trích xuất các thông tin hữu ích từ các bảng như là mối quan hệ và tên của các thuộc tính có thể giúp ích cho việc truy xuất

3 tăng cường …

4 thu thập dữ liệu

Chúng tôi đã thu thập 2417 bộ dữ liệu từ chính phủ liên bang Hoa Kỳ trên trang Data.gov với rất nhiều chủ đề. Mỗi một bộ dữ liệu bao gồm một hoặc nhiều bảng với định dạng CSV và các metadata tương ứng

Từ dữ liệu trên tạo ra tác vụ và các câu truy vấn

Chúng tôi tạo ra sáu tác vụ trong đó mỗi một tác vụ mô tả một thông tin rời rạc và cần được tìm thêm trong một hoặc nhiều tập dữ liệu. Với mỗi chuỗi thông tin rời rạc đó ta có một câu lệnh về thông tin cần thiết để tập dữ liệu đó được coi là có liên quan đến tác vụ.

Chúng tôi sử dụng Amazon Mechanical Turk để có được những câu truy vấn đa dạng cho các tác vụ phía trên từ người dùng thực.

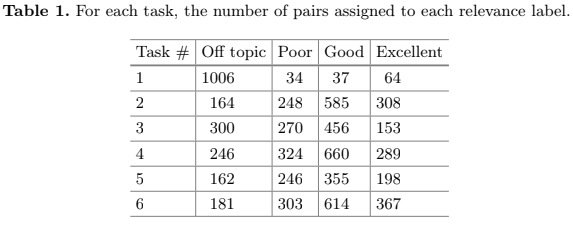
Đánh giá độ liên quan

Với mỗi tác vụ và mỗi câu truy vấn được đề xuất, chúng tôi sử dụng hàm xếp hạng truyền thống để chấm điểm các trường đơn đại diện cho mỗi tập dữ liệu và thu thập 100 kết quả tốt nhất. Các mô hình xếp hạng đã được sử dụng là: BM25, TF-IDF, mô hình ngôn ngữ dựa trên phương pháp Jelinek-Mercer smoothing, và mô hình ngôn ngữ dựa trên phương pháp Dirichlet smoothing

Sau đó, các cặp tác vụ – tập dữ liệu liên quan đã được ghi chú lại để sử dụng cho dịch vụ cung cấp cộng đồng được cung cấp bởi Figure Eight

Các ghi chú được thể hiện ngay trong tiêu đề, mô tả, liên kết đến bảng dữ liệu của tác vụ. Mỗi một cặp tác vụ-dữ liệu được đánh giá dựa trên 4 mức: 0 (sai chủ đề), 1 (nghèo nàn), 2 (tốt), 3 (xuất sắc).

Mỗi một cặp tác vụ-dữ liệu đơn được đánh giá bởi 3 người và chúng tôi căn cứ vào phần đông biểu quyết để gắn nhãn cho chúng. Nếu không đạt được thoả thuận trên chúng tôi sẽ tính điểm số trung bình để đánh giá nhãn của chúng.



Đánh giá

Phương pháp

Chúng tôi đánh giá hiệu xuất truy hồi dữ liệu thông qua số liệu: Dộ chính xác tại k và Normalized Discounted Cumulative Gain (NDCG) tại k. Để kiểm tra độ hiệu quả giữa các mô hình khác nhau chúng tôi sử dụng phương pháp t-test với p = 0.01

* Phương pháp Single-Field Document Ranking (SDR). Tập dữ liệu được coi như một tài liệu duy nhất. Chúng tôi sử dụng mô hình xếp hạng BM25 để đánh giá sự liên quan giữa tiêu đề và mô tả, nội dung văn bản của bảng dữ liệu và sự liên kết của tất cả chúng.
* Phương pháp Multifield Document Ranking (MDR). Bằng việc đặt wl = 0, Eq đã chia nhỏ thành các mô hình Mixture of Language. Phương pháp xếp hạng BM25 cũng được sử dụng ở đây: scoretext () để có sự so sánh công bằng với các phương pháp khác.

Kết quả: trả lời 3 câu hỏi

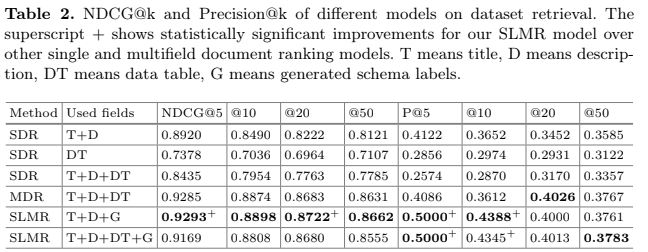
Q1: Bảng nội dung dữ liệu có giúp ích trong việc truy hồi thông tin không?

Q2: Bộ sinh schema label có giúp ích trong việc truy hồi thông tin không?

Q3: Trường nào là quan trọng nhất trong tác vụ truy xuất thông tin?

Đầu tiên chúng tôi thu được các thuộc tính của schema label như đã nói trong phần 3.1 và hệ số ẩn được đặt là 40. Sau đó chúng tôi huấn luyện mô hình Random Forest với những gì đã học được từ các thuộc tính của schema label.

Với mỗi cột chúng tôi chọn 10 schema label tốt nhất và bỏ đi các kết quả có xác xuất thấp hơn 0.5



Q3:

Từ kết quả của Single-Field Document Ranking, chúng tôi có thể thấy rằng việc chỉ sử dụng bảng dữ liệu để việc đánh giá xếp hạng cho hiệu suất kém nhất. Tuy nhiên việc chấm điểm theo sự kết hợp giữa tiêu đề và mô tả đạt kết quả tốt nhất, điều này chỉ ra rằng tiêu đề và môt tả quan trọng hơn bảng dữ liệu trong việc xếp hạng dữ liệu

Q1:

Bằng việc so sánh kết quả xếp hạng của tài liệu đơn trường và đa trường chúng ta có thể thấy sự kết hợp điểm số của bảng dữ liệu, tiêu đề và mô tả có thể cải thiện NDCG@k.

Thông qua việc tăng k lên thì NDCG@k giảm, xếp hạng các tài liệu đơn trường được cải thiện tốt hơn. Ngược lại với Precision@5, Precision@10, xếp hạng các tài liệu đơn trường cao hơn các tài liệu đa trường mặc dù sự khác biệt là rất nhỏ.

Vì vậy với câu hỏi Q1, dưới thiết lập xếp hạng các tài liệu đa trường nội dung của bảng dữ liệu có thể cải thiện NDCG nhưng không cải thiện độ chính xác của kết quả truy hồi dữ liệu.

Q2

Ngoài việc đánh giá các bảng dữ liệu, schema label mixed của chúng tôi đã đạt được điểm NDCG cao nhất trên tất cả giới hạn về thứ hạng. Điều đó chỉ ra rằng các schema label đã tạo có thể hữu ích trong việc cải thiện điểm NDCG của kết quả truy xuất dữ liệu

**5.4 Schema label tạo ra tìm kiếm nâng cao cho Web Tables**