TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

*

BÁO CÁO MÔN HỌC

CƠ SỞ DỮ LIỆU

Tên đề tài: Độ phức tạp tính toán
trong đa dạng hóa kết quả truy vấn

Mhóm sinh viên thực hiện : **Tạ Quang Tùng**
Phạm Minh Tâm

Giáo viên hướng dẫn : **Nguyễn Kim Anh**

HÀ NỘI
Ngày 2 tháng 12 năm 2017

Mục lục

Lời nói đầu

ABC

1 Cách tiếp cận tiên đề trong đa dạng hóa kết quả truy vấn

Trong chương này, chúng ta sẽ đề cập đến cách tiếp cận theo hướng tiên đề trong đa dạng hóa kết quả truy vấn [1]. Mục đích để trợ giúp cho việc lựa chọn hàm mục tiêu và đồng thời ràng buộc kết quả bài toán. Chúng ta sẽ xem xét những hàm đã được đề xuất thoả mãn những tính chất, đồng thời chỉ ra rằng không tồn tại hàm thỏa mãn tất cả những tính chất được đề xuất dưới đây.

1.1 Khái niệm cơ bản

Xét một tập hợp các bản ghi $U = \{u_1, u_2, ..., u_n\}$, trong đó $n \geq 2$ và giải sử tồn tại một tập hữu hạn tất cả các câu truy vấn Q. Với một câu truy vấn $q \in Q$ và một số nguyên k, chúng ta muốn nhận được kết quả là tập con $S_k \subset U$ của tập các bản ghi ban đầu (hay của cơ sở dữ liệu). Hàm tương thích của mỗi một bản ghi được xác định bởi hàm $w : U \times Q \to \mathbf{R}^+$, bản ghi càng phù hợp với câu truy vấn thì sẽ có giá trị càng cao. Mục tiêu đa dạng hóa kết quả có thể được hiểu đơn giản là việc các kết quả trả về không được tương tự nhau. Dưới dạng biểu thức, ta có thể định nghĩa hàm khoảng cách $d : U \times U \to \mathbf{R}^+$ giữa các bản ghi, ở đó giá trị càng nhỏ thể hiện rằng hai bản ghi càng tương tự nhau. Đồng thời ta muốn hàm khoảng cách phải có tính chất phân biệt: Với hai bản ghi bất kì $u, v \in U$, $d(u, v) = 0$ khi và chỉ khi $u = v$, và tính chất đối xứng: $d(u, v) = d(v, u)$. Tuy nhiên, không nhất thiết hàm khoảng cách phải tạo thành một metric.

Chúng ta ở đây chỉ quan tâm đến việc lựa chọn tập kết quả chứ không quan tâm đến vấn đề xếp hạng các bạn ghi. Nếu chúng ta đã có tập kết quả, chúng ta có thể sắp xếp kết quả cuối cùng theo thứ tự tương thích với câu truy vấn ban đầu.

Dưới dạng toán học, hàm lựa chọn tập kết quả $f : 2^U \times Q \times w \times d \to \mathbf{R}$ có thể được hiểu là gán mỗi một điểm số cho từng các tập con của U khi cho trước một câu truy vấn $q \in Q$, một hàm tương thích $w(.)$ và một hàm khoảng cách $d(.,.)$. Cố định $q, w(.), d(.,.)$ và một số nguyên $k \in \mathbf{Z}^+ (k \geq 2)$, mục tiêu là chọn một tập con $S_k \subseteq U$ của các bản ghi sao cho giá trị của hàm f là lớn nhất.

$$S_k^* = \underset{\substack{S_k \subseteq U \\ |S_k|=k}}{\operatorname{argmax}} f(S_k, q, w(.), d(.,.))$$

Trong đó tất cả các đối số khác S_k đều được cố định.

Ta có thể có rất nhiều lựa chọn hàm mục tiêu f với các hàm tương thích và hàm khoảng cách cho trước. Những hàm đó có sự đánh đổi giữa tính tương thích và tính

tương tự theo những cách khác nhau. Do đó chung ta cần chỉ định ra các tiêu chuẩn để lựa chọn hàm mục tiêu tốt trong vô số các hàm mục tiêu đó. Cách tiếp cận toán học được sự dụng phổ biến trong trường hợp này là đưa ra một hệ tiên đề mà được mong đợi trong các hệ thống hỗ trợ đa dạng hóa. Từ đó cung cấp một cơ sở để so sánh sự khác biệt giữa các hàm mục tiêu được chọn.

1.2 Những tiên đề trong đang dạng hóa kết quả truy vấn

Chúng ta đề xuất hàm f thỏa mãn tập các tiên đề dưới đây. Mỗi một tiên đề đều khá là trực quan đối với vấn đề đa dạng hóa kết quả. Thêm vào đó, chúng ta sẽ chỉ ra rằng bất kì một tập con thực sự của những tiên đề này là "cực đại", có nghĩa là không tồn tại hàm mục tiêu nào thỏa mãn tất cả những tiên đề dưới đây. Từ đó cung cấp một phương pháp tự nhiên cho việc lựa chọn giữa các hàm mục tiêu, khi mà một số tính chất là thiết yếu cho một hệ thống nào đó.

Cố định $q, w(.), d(.,), k$, f và $S_k^* = \text{argmax}_{S_k \subseteq U} f(S_k, q, w(.), d(.,.))$. Ta có các tiên đề sau:

1. **Bất biến theo tỉ lệ**: Tính chất này chỉ định rằng hàm mục tiêu không được phép bị ảnh hưởng khi mà thay đổi đầu vào theo cùng một tỉ lệ. Một cách hình thức, xét tập tối ưu S_k^*, chúng ta muốn f vẫn thỏa mãn $S_k^* = \text{argmax}_{S_k \subseteq U} f(S_k, q, \alpha \cdot w(.), \alpha \cdot d(.,.))$ với bất kì một giá trị dương của α.

2. **Nhất quán**: Tính nhất quán nói rằng nếu làm cho những bản ghi trong tập kết quả càng tương thích và càng có tính đa dạng hơn và đồng thời làm cho các bạn ghi không phải kết quả ít tương thích, ít có tích đa dạng hơn thì kết quả của bài toán vẫn không thay đổi. Một cách hình thức, với hai hàm bất kì $\alpha : U \to \mathbf{R}^+$ và $\beta : U \times U \to \mathbf{R}^+$, chúng ta thay đổi hàm tương thích và hàm khoảng cách như sau:

$$w(u) = \begin{cases} w(u) + \alpha(u) & u \in S_k^* \\ w(u) - \alpha(u) & \text{Trường hợp còn lại} \end{cases}$$

$$d(u,v) = \begin{cases} d(u,v) + \beta(u,v) & u \in S_k^* \\ d(u,v) - \beta(u,v) & \text{Trường hợp còn lại} \end{cases}$$

   Thì S_k^* vẫn là tập tối ưu của hàm mục tiêu f.

3. **Phong phú**. Tính phong phú nói rằng ta có thể đạt được bất kì một tập nào đó là kết quả, nếu như lựa chọn đúng hàm tương thích và hàm khoảng cách. Một cách hình thức:

$$\forall k \geq 2, \exists w(.), \exists d(.,.), !\exists S_k^* = \underset{S_k \subseteq U}{\text{argmax}}\, f(S_k, q, w(.), d(.,.))$$

4. **Ổn định**. Tính ổn định quy định những hàm mà kết quả bài toán không thay đổi tùy ý với kích thước của tập kết quả. Hàm f phải thoải mãn $S_k^* \subset S_{k+1}^*$.

Tài liệu tham khảo

[1] diversification.pdf. http://theory.stanford.edu/~aneeshs/papers/diversification.pdf.