

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	1
1 Cách tiếp cận tiên đề trong đa dạng hóa kết quả truy vấn	1
1.1 Khái niệm cơ bản	1
1.2 Những tiên đề trong đa dạng hóa kết quả truy vấn	2
1.3 Hàm mục tiêu và thuật toán	4

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 thỏa 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 $U = \{u_1, u_2, \dots, u_n\}$, trong đó $n \geq 2$ và giả 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 $w : U \times Q \rightarrow \mathbf{R}^+$, 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 $d : U \times U \rightarrow \mathbf{R}^+$ 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 kỳ $u, v \in U$, $d(u, v) = 0$ 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 bản 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 $f : 2^U \times Q \times w \times d \rightarrow \mathbf{R}$ 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 $w(\cdot)$ và một hàm khoảng cách $d(\cdot, \cdot)$. Cố định $q, w(\cdot), d(\cdot, \cdot)$ 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.

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

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 chúng 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 sử 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 đa 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 $q, w(\cdot), d(\cdot, \cdot), k, f$ và $S_k^* = \operatorname{argmax}_{S_k \subseteq U} f(S_k, q, w(\cdot), d(\cdot, \cdot))$. 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 $S_k^* = \operatorname{argmax}_{S_k \subseteq U} f(S_k, q, \alpha \cdot w(\cdot), \alpha \cdot d(\cdot, \cdot))$ 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 bản ghi không phải kết quả ít tương thích, ít có tính đa dạng hơn thì kết quả của bài toán vẫn không thay đổi. Một cách hình thức, với hai hàm bất kì $\alpha : U \rightarrow \mathbf{R}^+$ và $\beta : U \times U \rightarrow \mathbf{R}^+$, 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:

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

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 thỏa mãn $S_k^* \subset S_{k+1}^*$.

5. **Độc lập với phần tử không tương thích.** Tiên đề này nói rằng điểm số của tập hợp không bị ảnh hưởng bởi những bản ghi nằm ngoài tập đó. Cụ thể, với một tập S bất kì, hàm f tại S : $f(S)$ sẽ độc lập với những giá trị sau:

- $w(u)$ với mọi $u \notin S$.
- $d(u, v)$ với mọi $u, v \notin S$.

6. **Đơn điệu.** Tính đơn điệu yêu cầu việc thêm bản ghi vào một tập hợp bất kì không làm tăng điểm số của hàm mục tiêu đối với tập đó. Cố định $w(\cdot), d(\cdot, \cdot), f$ và $S \subseteq U$. Với mọi $x \in S$, ta có:

$$f(S \cup \{x\}) \geq f(S)$$

7. **Độ mạnh của tính tương thích.** Tính chất này đảm bảo rằng không có hàm f nào bỏ qua hàm tương thích. Một cách hình thức, chúng ta cố định $w(\cdot), d(\cdot, \cdot), f$ và S , những tính chất sau đây phải đúng với mọi giá trị $x \in S$:

- (a) Tồn tại số thực $\delta_0 > 0$ và $a_0 > 0$ để mà những điều kiện dưới đây được thoả mãn sau khi đã thực hiện chỉnh sửa sau: Sửa đổi giá trị của hàm tương thích trở thành hàm $w'(\cdot)$ sao cho $w'(\cdot)$ giống hệt $w(\cdot)$ ngoại trừ tại phần tử x , $w'(x) = a_0 > w(x)$. Khi đó, ta có:

$$f(S, w'(\cdot), d(\cdot, \cdot), k) = f(S, w(\cdot), d(\cdot, \cdot), k) + \delta_0$$

- (b) Nếu $f(S \setminus x) < f(S)$ thì tồn tại số thực $\delta_1 > 0$ và $a_1 > 0$ để mà những điều kiện sau vẫn đúng: chỉnh sửa hàm tương thích $w(\cdot)$ để đạt được một hàm mới $w'(\cdot)$ sao cho hàm $w'(\cdot)$ giống hệt hàm cũ, ngoại trừ tại phần tử x , $w'(x) = a_1 < w(x)$. Từ đó, ta có:

$$f(S, w'(\cdot), d(\cdot, \cdot), k) = f(S, w(\cdot), d(\cdot, \cdot), k) - \delta_1$$

8. **Độ mạnh của tính tương tự.** Tính chất này đảm bảo rằng không có hàm f nào bỏ qua hàm khoảng cách. Một cách hình thức, nếu cố định $w(\cdot), d(\cdot, \cdot), f$ và S , thì những tính chất sau đúng với mọi giá trị $x \in S$:

- (a) Tồn tại số thực $\delta_0 > 0$ và $a_0 > 0$ để mà những điều kiện dưới đây được thoả mãn sau khi đã thực hiện chỉnh sửa sau: tạo một hàm $d'(\cdot, \cdot)$ từ hàm $d(\cdot, \cdot)$, trong đó, ta tăng giá trị của $d(x, u)$ tại một số vị trí u cần thiết nào đó sao cho $\min_{u \in S} d(x, u) = b_0$. Từ đó, ta có:

$$f(S, w(\cdot), d'(\cdot, \cdot), k) = f(S, w(\cdot), d(\cdot, \cdot), k) + \delta_0$$

- (b) Nếu $f(S \setminus x) < f(S)$ thì tồn tại số thực $\delta_1 > 0$ và $a_1 > 0$ để mà những điều kiện sau vẫn đúng: chỉnh sửa hàm khoảng cách $d(\cdot, \cdot)$ bằng cách tăng giá trị $d(x, u)$ tại một số vị trí u cần thiết nào đó để đảm bảo rằng $\max_{u \in S} d(x, u) = b_1$. Gọi hàm được tạo ra là $d'(\cdot, \cdot)$. Từ đó, ta có:

$$f(S, w(\cdot), d'(\cdot, \cdot), k) = f(S, w(\cdot), d(\cdot, \cdot), k) - \delta_1$$

Từ những tiên đề này, một câu hỏi được đặt ra là làm thế nào để mô tả một tập các hàm f thỏa mãn những tiên đề này. Đáng ngạc nhiên là không thể tìm được hàm thỏa mãn được tất cả những tiên đề cùng một lúc.

Định lý: Không hàm f thỏa mãn tất cả những tiên đề đã được nêu ở trên.

Định lý này ngụ ý rằng bất kì một tập con thực sự của tập các tiên đề trên là tối đa. Kết quả này cho phép chúng ta mô tả một cách tự nhiên một tập các hàm đa dạng hóa, và lựa chọn một hàm cụ thể thỏa mãn tập con các tiên đề mà chúng ta mong muốn.

1.3 Hàm mục tiêu và thuật toán

Từ kết quả không thể của định lý trên, chúng ta chỉ có thể hi vọng những hàm đa dạng hóa có thể thỏa mãn một tập con của các tiên đề. Chú ý rằng số lượng các hàm thỏa mãn có thể khá lớn. Hơn nữa, đề xuất một hàm mục tiêu có thể không hữu dụng nếu không thể tìm được một thuật toán để tối ưu hàm mục tiêu đã chọn. Trong phần này, chúng ta sẽ xem xét ba hàm mục tiêu cụ thể, và đồng thời cung cấp thuật toán tối ưu hàm mục tiêu đó.

Hàm đa dạng hóa tổng lớn nhất (max-sum diversification)

Một hàm mục tiêu thỏa mãn đồng thời hai tiêu chuẩn (tương thích và đa dạng), biểu diễn dưới dạng tổng của hàm tương thích và hàm khoảng cách. Cụ thể được định nghĩa như sau:

$$f(S) = (k-1) \sum_{u \in S} w(u) + 2\lambda \sum_{u,v \in S} d(u,v) \quad (1)$$

Ở đó $|S| = k$, và $\lambda > 0$ là tham số xác định sự đánh đổi giữa tính tương thích và tính đa dạng. Để ý rằng chúng ta cần nhân thêm giá trị ở tổng bên trái để cân bằng hóa vì tổng bên phải có $\frac{k(k-1)}{2}$ phần tử trong khi đó tổng bên trái chỉ có k phần tử.

Nhận xét 1. Hàm mục tiêu thỏa mãn phương trình 1 thỏa mãn tất cả các tiên đề ngoại trừ tiên đề về tính ổn định.

Hàm mục tiêu này có thể được chuyển về hàm mục tiêu phân tán cơ sở (facility dispersion), được biết đến là bài toán phân tán tổng lớn nhất (max-sum dispersion problem). Bài toán phân tán tổng lớn nhất là bài toán phân tán mà mục tiêu là tối đa hóa tổng của tất cả các cặp khoảng cách giữa những điểm trong một tập S . Trong trường hợp này, nếu ta định nghĩa hàm khoảng cách:

$$d'(u,v) = w(u) + w(v) + 2\lambda d(u,v) \quad (2)$$

Dễ thấy, $d'(\cdot, \cdot)$ là một metric nếu như $d(\cdot, \cdot)$ cũng là một metric. Hơn nữa, với các giá trị $S \subseteq U$ ($|S| = k$), ta có:

$$\sum_{u,v \in S} d'(u,v) = (k-1) \sum_{u \in S} w(u) + 2\lambda \sum_{u,v \in S} d(u,v)$$

Từ đó, phương trình 1 có thể được viết lại thành:

$$f(S) = \sum_{u,v \in S} d'(u,v)$$

Đồng thời, đó cũng là mục tiêu của bài toán phân tán tổng lớn nhất đã được mô tả ở trên. Từ đó ta có thể giải được bài toán về hàm mục tiêu đa dạng hóa tổng lớn nhất. Bài toán tối đa hóa giá trị hàm mục tiêu của phương trình 1 là NP khó, nhưng tồn tại một thuật giải xấp xỉ cho bài toán. Trong trường hợp là metric, chúng ta có thể sử dụng thuật toán 1 để giải bài toán đã đặt ra.

Thuật toán 1 Thuật toán cho bài toán phân tán tổng lớn nhất

Đầu vào: Tập U , giá trị nguyên k

Đầu ra: Tập S ($|S| = k$) sao cho giá trị của $f(S)$ là lớn nhất

Khởi tạo $S = \emptyset$

for $i \leftarrow 1$ **to** $\lfloor \frac{k}{2} \rfloor$ **do**

 Tìm $(u, v) = \operatorname{argmax}_{x,y \in U} d(x, y)$

 Tập $S = S \cup \{u, v\}$

 Xóa tất cả các cạnh mà gắn với u hoặc v

end for

if k là lẻ **then**

 Thêm một bản ghi bất kì vào S

end if

Hàm đa dạng hóa lớn nhất - nhỏ nhất (max-min diversification)

Một hàm mục tiêu thứ hai thỏa mãn đồng thời hai tiêu chuẩn (tương thích và đa dạng), biểu diễn dưới dạng giá trị nhỏ nhất của tổng hàm tương thích và hàm khoảng cách. Cụ thể được định nghĩa như sau:

$$f(S) = \min_{u \in S} w(u) + \lambda \min_{u,v \in S} d(u,v) \quad (3)$$

Ở đó $|S| = k$, và $\lambda > 0$ là tham số chỉ định sự đánh đổi giữa tính tương thích và tính đa dạng.

Nhận xét 2. Hàm mục tiêu đa dạng hóa được cho bởi phương trình 3 thỏa mãn tất cả những tiên đề ngoại trừ tiên đề về tính nhất quán và tính ổn định.

Như lần trước, hàm mục tiêu này tương ứng với một hàm mục tiêu phân tán cơ sở, trong trường hợp này là bài toán phân tán lớn nhất - nhỏ nhất (max-min dispersion problem). Hàm mục tiêu cho bài toán phân tán lớn nhất - nhỏ nhất là: $g(P) = \min_{v_i, v_j \in P} d(v_i, v_j)$, và có thể suy ra là tương đương với phương trình 3 bằng cách đặt khoảng cách:

$$d'(u, v) = \frac{1}{2}(w(u) + w(v)) + \lambda d(u, v) \quad (4)$$

Cũng tương tự, $d'(u, v)$ tạo thành một không gian metric. Hơn nữa, ta lại có:

$$\min_{u, v \in S} d'(u, v) = \min_{u \in S} w(u) + \lambda \min_{u, v \in S} d(u, v) = f(S)$$

Từ đó, ta có thể giải bài toán tìm cực đại hàm mục tiêu đa dạng hóa lớn nhất nhỏ nhất bằng việc giải bài toán phân tán lớn nhất - nhỏ nhất. Ta có thể sử dụng thuật toán 2 để giải bài toán đó. Đồng thời, bài toán đã đặt ra cũng là bài toán NP khó.

Thuật toán 2 Thuật toán cho bài toán phân tán lớn nhất - nhỏ nhất

Đầu vào: Tập U , giá trị nguyên k

Đầu ra: Tập S ($|S| = k$) sao cho giá trị của $f(S)$ là lớn nhất

Khởi tạo $S = \emptyset$

Tìm $(u, v) = \operatorname{argmax}_{x, y \in U} d(x, y)$ và tập $S = \{u, v\}$

for $x \in U$ **do**

 Định nghĩa $d(x, S) = \min_{u \in S} d(x, u)$

end for

while $|S| < k$ **do**

 Tìm $x \in U \setminus S$ sao cho $x = \operatorname{argmax}_{x \in U \setminus S} d(x, S)$

 Tập $S = S \cup \{x\}$

end while

Hàm đơn mục tiêu (mono-objective formulation)

Hàm mục tiêu cuối cùng không liên quan đến bài toán phân tán cơ bản nào cả và nó hợp giá trị tương thích và giá trị của hàm khoảng cách thành một giá trị đơn nhất cho mỗi bản ghi. Cụ thể được định nghĩa như sau:

$$f(S) = \sum_{u \in S} w'(u) \quad (5)$$

Trong đó

$$w'(u) = w(u) + \frac{\lambda}{|U| - 1} \sum_{u \in U} d(u, v)$$

Giá trị tham số λ thể hiện sự đánh đổi giữa tính tương thích và tính đa dạng. Một cách trực quan, giá trị $w'(u)$ tính toán sự quan trọng một cách toàn bộ mỗi một bản ghi u . Đặc trưng tiên đề của hàm mục tiêu này như sau:

Nhận xét 3. Hàm mục tiêu được cho trong phương trình 5 thỏa mãn tất cả các tiên đề trừ tiên đề về tính nhất quán.

Đồng thời nhận thấy rằng ta có thể tối ưu hàm mục tiêu một cách chính xác bằng việc tính toán giá trị $w'(u)$ cho mỗi bản ghi $u \in U$ và rồi lựa chọn những bản ghi trong top k giá trị để cho vào tập kết quả.

Tài liệu tham khảo

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