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 Độ phức tạp trong vấn đề đa dạng hóa kết quả tìm kiếm

Đa dạng hóa kết quả try vấn là một vấn đề tối ưu hóa đa điều kiện cho việc xếp hạng kết quả truy vấn. Cho bộ dữ liệu D, một truy vấn Q và một số nguyên k, yêu cầu đặt ra là tìm một tập có k phần tử từ kết quả truy vấn Q(D) sao cho các bộ trả về có sự liên quan nhất có thể tới câu truy vấn và đồng thời trả về kết quả đa dạng nhất cho mỗi bộ. Tập con của Q(D) được sắp xếp bởi một hàm mục tiêu được định nghĩa bởi tính liên quan của các kết quả với câu truy vấn và tính đa dạng của tập kết quả. Đa dạng hóa kết quả truy vấn có rất nhiều ứng dụng trong cơ sở dữ liệu, trích xuất thông tin và các hoạt động tìm kiếm. Có 3 vấn đề liên quan đến đa dạng hóa kết quả truy vấn được định nghĩa đó là:

- Xác định xem liệu có tồn tại tập k phần tử thỏa mãn một ràng buộc về khía cạnh độ liên quan và độ đa dạng đối với truy vấn (QRD).
- Xếp hạng một tập k phần tử cho trước (DRP).
- Đếm xem có bao nhiêu tập k phần tử thỏa mãn điều kiện cho trước (RDC).

Chúng ta nghiên cứu vấn đề trên với nhiều ngôn ngữ truy vấn cho 3 hàm mục tiêu *max-sum diversification, max-min diversification, mono-objective formulation*. Với mỗi vấn đề ta sẽ dùng các hàm mục tiêu khác nhau để truy vấn. Với mỗi trường hợp, các thông số sẽ được thay đổi để nghiên cứu về độ phức tạp của bài toán và tìm yếu tố ảnh hưởng tới độ phức tạp của bài toán.

1.1 Các vấn đề liên quan đến đa dạng hóa kết quả truy vấn

1.1.1 QRD (The query result diversification problem)

Cho một truy vấn Q, tập cơ sở dữ liệu D, một số nguyên k, một cận B, hàm mục tiêu F.Khi đó ta nói một bộ $U \subseteq Q(D)$ là một tập ứng viên nếu nó có k phần tử $|U| = k$. Một tập ứng viên U là thỏa mãn cho bài toán (Q,D,k,F,B) nếu $F(U) \geq B$.
Vấn đề $QRD(\mathcal{L}_Q, F(.))$ được trình bày như sau:
INPUT: Một cơ sở dữ liệu D, một truy vấn $Q \in \mathcal{L}_Q$, một hàm mục tiêu F(.) một số thực B và một số nguyên $k \geq 1$.
OUTPUT: Trả lời câu hỏi liệu có tồn tại một tập thỏa mãn bàn toán (Q,D,k,F,B) nêu trên không?

1.1.2 DRP (The diversity ranking problem)

Với một tập U được cho trước bởi người dùng hoặc hệ thống, chúng ta sẽ xác định xem liệu tập U đó có đạt được mục tiêu đa dạng hóa kết quả từ đó đánh giá kết quả có thỏa mãn với người dùng không.Đây là một vấn đề quyết định khác có thể gọi là xếp hạng tập kết quả cho trước. Ta xem xét một tập U và một số r. Ta nói hạng của U là r, ký hiệu $rank(U) = r$ nếu tồn tại một mảng S có r-1 phần tử các tập ứng viên cho vấn đề (Q,D,k) sao cho:

- Tất cả các tập thuộc mảng $s_i \in S$ đều thỏa mãn $F(s_i) > F(U)$
- Bất kỳ một tập $s_k \notin S$ đều có $F(U) \geq F(s_k)$

Vấn đề $DRP(\mathcal{L}_Q, F(.))$ được trình bày như sau:
INPUT: Một cơ sở dữ liệu D, một truy vấn $Q \in \mathcal{L}_Q$, một hàm mục tiêu F(.) một bộ U nằm trong thuộc Q(D) với |U| = k.
OUTPUT: Trả lời câu hỏi liệu $rank(U) \leq r$?

1.1.3 RDC (The result diversity counting problem)

Cho một mục tiêu B chúng ta muốn biết có bao nhiêu tập U thỏa mãn mục tiêu trên.
Vấn đề $RDC(\mathcal{L}_Q, F(.))$ được trình bày như sau:
INPUT: Một cơ sở dữ liệu D, một truy vấn $Q \in \mathcal{L}_Q$, một hàm mục tiêu F(.) và một số nguyên k như ở trong vấn đề $QRD(\mathcal{L}_Q, F(.))$
OUTPUT: Trả lời câu hỏi có bao nhiêu bộ thỏa mãn bàn toán (Q,D,k,F,B) ?

1.2 Độ phức tạp của bài toán

Chương này sẽ tìm hiểu về độ phức tạo của các vấn đề $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$, $RDC(\mathcal{L}_Q, F(.))$ sử dụng các ngôn ngữ CQ,UCQ,$\exists FO^+$ và FO với các hàm mục tiêu F(.) là $F_{MS}(.)$ (max-sum diversification), $F_{MM}(.)$ (max-min diversification) và $F_{mono}(.)$ (mono-objective formulation).
Với mỗi trường hợp ta nghiên cứu:

1. Độ phức tạp hỗn hợp khi cả truy vấn và dữ liệu D là thay đổi.
2. Độ phức tạp dữ liệu khi truy vấn Q được định nghĩa và cố định, dữ liệu D thay đổi, độ phức tạp được đánh giá trên một truy vấn cố định đối với nhiều loại dữ liệu đầu vào.

1.2.1 Độ phức tạp hỗn hợp

QRD (*The query result diversification problem*)

1. Khi hàm mục tiêu là max-sum hoặc max-min ngôn ngữ truy vấn có ảnh hưởng tới độ phức tạp của thuật toán, cụ thể: là NP-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$, nhưng trở thành PSPACE-complete khi $\mathcal{L}_Q$ là FO. Trong khi sự hiện diện của UCQ và $\exists FO^+$ không làm cho thuật toán phức tạp hơn so với khi dùng ngôn ngữ CQ, thì sự hiện diện của FO làm phức tạp hóa thuật toán.

2. Khi hàm mục tiêu là mono-objective thuật toán trở nên phức tạp hơn khi $\mathcal{L}_Q$ là CQ,UCQ, $\exists FO^+$ với độ phức tạp là PSPACE-complete giống như khi dùng ngôn ngữ FO.

Định lý 1: Cho vấn đề $QRD(\mathcal{L}_Q, F(.))$, khi hàm mục tiêu là max-sum hoặc max-min, độ phức tạp hỗn hợp là:

- *NP-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$*
- *PSPACE-complete khi $\mathcal{L}_Q$ là FO*

Còn khi hàm mục tiêu là mono-objective độ phức tạp hỗn hợp là:

- *PSPACE-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hay FO.*

DRP (*The diversity ranking problem*)

1. Giống như $QRD(\mathcal{L}_Q, F(.))$ khi F là max-min objective hoặc max-sum objective sự khác nhau của các ngôn ngữ ảnh hưởng đến độ phức tạp của thuật toán. Khi F là mono-objective thay đổi ngôn ngữ không ảnh hưởng tới độ phức tạp

2. Khi hàm mục tiêu là max-sum hoặc max-min $DRP(\mathcal{L}_Q, F(.))$ là coNP-complete cho ngôn ngữ CQ, UCQ.

Định lý 2: Cho vấn đề $DRP(\mathcal{L}_Q, F(.))$, khi hàm mục tiêu là max-sum hoặc max-min, độ phức tạp hỗn hợp là:

- *coNP-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$*
- *PSPACE-complete khi $\mathcal{L}_Q$ là FO*

Còn khi hàm mục tiêu là mono-objective độ phức tạp hỗn hợp là:

- *PSPACE-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hay FO*

RDC (*The result diversity counting problem*) Khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$, vấn đề trở nên khó đối với ngôn ngữ FO hơn là với CQ, UCQ, $\exists FO^+$. Ngược lại $F_mono(.)$ tác động dến sự phức tạp của thuật toán nhiều hơn so với ảnh hưởng của ngôn ngữ $\mathcal{L}_Q$. Điều này tương tự như các vấn đề $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$ đã nêu ở trên. Các kết quả được xác minh bởi sự giảm bớt. *Định lý 3: Cho vấn đề $RDC(\mathcal{L}_Q, F(.))$, khi hàm mục tiêu là max- sum hoặc max-min, độ phức tạp hỗn hợp là:*

- *#.NP-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$*
- *#.PSPACE-complete khi $\mathcal{L}_Q$ là FO*

Còn khi hàm mục tiêu là mono-objective độ phức tạp hỗn hợp là:

- *#.PSPACE-complete khi $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hay FO*

1.2.2 Độ phức tạp dữ liệu

QRD (*The query result diversification problem*) Khi F là $F_{MS}(.)$ hoăc $F_{MM}(.)$, cố định truy vấn Q không làm thay đổi độ phức tạp của vấn đề $QRD(\mathcal{L}_Q, F(.))$ cho ngôn ngữ CQ, UCQ, $\exists FO^+$, vấn đề vẫn có độ phức tạp là NP-comlete. Ngược lại khi $\mathcal{L}_Q$ là FO và F là $F_{MS}(.)$ hoặc $F_{MM}(.)$ vấn đề đơn giản hơn với độ phức tạp là NP-comlete. Khi F là $F_{mono}(.)$ vấn đề cũng trở nên dễ giải quyết hơn. *Định lý 4: Cho vấn đề $QRD(\mathcal{L}_Q, F(.))$, độ phức tạp dữ liệu là:*

- *NP-complete khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$*
- *Trong PTIME khi F là $F_{mono}(.)$*

với $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hoặc FO

DRP (*The diversity ranking problem*) Giống như vấn đề $QRD(\mathcal{L}_Q, F(.))$, cố định truy vấn Q làm cho $DRP(\mathcal{L}_Q, F(.))$ đơn giản hơn khi:

1. Hàm mục tiêu F là $F_{mono}(.)$
2. Khi ngôn ngữ truy vấn $\mathcal{L}_Q$ là FO, và hàm mục tiêu F là $F_{MS}(.)$ hoặc $F_{MM}(.)$

Độ phức tạp dữ liệu vẫn tương tự như độ phức tạp hỗn hợp của khi LQ là CQ, UCQ hoặc $\exists FO^+$, và khi F là $F_MS(.)$ hoặc $F_{MM}(.)$.
Định lý 5: Cho vấn đề $DRP(\mathcal{L}_Q, F(.))$, độ phức tạp dữ liệu là:

- *coNP-complete khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$*
- *Trong PTIME khi F là $F_{mono}(.)$*

với $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hoặc FO
RDC (*The result diversity counting problem*) *Định lý 6: Cho vấn đề $RDC(\mathcal{L}_Q, F(.))$, độ phức tạp dữ liệu là:*

- *#P-complete dưới sụ giảm thiểu khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$*
- *#P-complete dưới sự tối giản đa thức Turing khi F là $F_{mono}(.)$*

với $\mathcal{L}_Q$ là CQ, UCQ, $\exists FO^+$ hoặc FO
Tổng kết Từ những kết quả trên chúng ta thấy;

1. Cả ngôn ngữ truy vấn và hàm mục tiêu đều có tác động đến độ phức tạp hỗn hợp của những vấn đề trên. Cụ thể hơn:
   - #P-complete khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$, những vấn đề này đối với ngôn ngữ truy vấn FO có độ phức tạp phức tạp lớn hơn so với các ngôn ngữ CQ, UCQ và $\exists FO^+$.
   - #P-complete Đối với $F_{mono}(.)$ hàm mục tiêu là yếu tố chính chi phối sự phức tạp.
2. Độ phức tạp của dữ liệu không liên quan tới ngôn ngữ truy vấn.

1.3 Một số trường hợp đặc biệt

Truy vấn nhận dạng

Hệ quả 1: Với truy vấn định danh, khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$, cả độ phức tạp hỗn hợp và độ phức tạp dữ liệu là:

- *NP-complete cho vấn đề $QRD(\mathcal{L}_Q, F(.))$*
- *coNP-complete cho vấn đề $DRP(\mathcal{L}_Q, F(.))$ và*
- *#P-complete cho vấn đề $RDC(\mathcal{L}_Q, F(.))$ dưới sụ giảm thiểu*

Khi F là $F_mono(.)$ cả độ phức tạp hỗn hợp và độ phức tạp dữ liệu là:

- *trong thời gian PTIME cho vấn đề $QRD(\mathcal{L}_Q, F(.))$*
- *trong thời gian PTIME cho vấn đề $DRP(\mathcal{L}_Q, F(.))$*
- *#P-complete cho vấn đề $RDC(\mathcal{L}_Q, F(.))$ dưới sự tối giản đa thức Turing*

Khi $\lambda = 0$

Tiếp theo, chúng ta nghiên cứu tác động của hàm liên quan đến tính đa dạng đối với sự phức tạp của thuật toán. Khi $\lambda = 0$ chỉ có hàm tính sự đa dạng $\delta_{rel}(.,.)$ được sử dụng trong hàm mục tiêu F(.). *Hệ quả 2: Với $\lambda = 0$, khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$, độ phức tạp hỗn hợp của những vấn đề $QRD(\mathcal{L}_Q, F(.))$, $DRC(\mathcal{L}_Q, F(.))$ và $RDC(\mathcal{L}_Q, F(.))$ giống với độ phức tạp ta đã đề cập đến trong định lý 1,2 và 3 nhưng độ phức tạp dữ liệu là:*

- *trong thời gian PTIME cho vấn đề $QRD(\mathcal{L}_Q, F(.))$*
- *trong thời gian PTIME cho vấn đề $DRP(\mathcal{L}_Q, F(.))$ và*
- *#P-complete dưới sự tối giản đa thức Turing cho vấn đề $RDC(\mathcal{L}_Q, F(.))$ khi F(.) là $F_{MS}(.)$, trong FP khi F(.) là $F_{MM}(.)$*

Khi F là $F_{mono}(.)$ độ phức tạp hỗn hợp là:

- *NP-complete cho vấn đề $QRD(\mathcal{L}_Q, F(.))$ khi $\mathcal{L}_Q$ là CQ, UCQ hoặc $\exists FO^+$ và PSPACE-complete khi $\mathcal{L}_Q$ là FO.*
- *coNP-complete cho vấn đề $DRP(\mathcal{L}_Q, F(.))$ khi $\mathcal{L}_Q$ là CQ, UCQ hoặc $\exists FO^+$ và PSPACE-complete khi $\mathcal{L}_Q$ là FO.*
- *#.NP-complete cho vấn đề $RDC(\mathcal{L}_Q, F(.))$ khi $\mathcal{L}_Q$ là CQ, UCQ hoặc $\exists FO^+$ và #.PSPACE-complete khi $\mathcal{L}_Q$ là FO.*

Độ phức tạp dữ liệu giống như trong các định lý 4,5,6.
Khi $\lambda = 1$
Không giống như kết quả ở hệ quả 2, việc bỏ hàm tính sự liên quan $\delta_{rel}(.,.)$ khỏi F $(\cdot)$ không làm cho thuật toán dễ dàng hơn. Thực tế, cả độ phức tạp kết hợp và độ phức tạp dữ liệu của $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$ và $RDC(\mathcal{L}_Q, F(.))$ vẫn không đổi. Điều này chứng minh hàm tính sự đa dạng $\delta_{rel}(.,.)$ chi phối sự phức tạp của những vấn đề này.
Định lý 7: Khi $\lambda = 1$ độ phức tạp hỗn hợp giống như trong định lý 1,2 và 3 độ phức tạp dữ liệu giống như trong định lý 4,5,6 và không đổi với các vấn đề $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$, $RDC(\mathcal{L}_Q, F(.))$ khi hàm mục tiêu là $F_{MS}(.)$,$F_{MM}(.)$,$F_{mono}(.)$ với các ngôn ngữ CQ, UCQ, $\exists FO^+$ và FO
Khi k là hằng số
Hệ quả 3: Với k cố định

- *Độ phức tạp hỗn hợp như trong định lý 1,2,3 với các vấn đề $QRD(\mathcal{L}_Q, F(.))$,$DRP(\mathcal{L}_Q, F(.))$ và $RDC(\mathcal{L}_Q, F(.))$*
- *Độ phức tạp dữ liệu là :*
  - *PTIME với vấn đề $QRD(\mathcal{L}_Q, F(.))$*
  - *PTIME với vấn đề $DRP(\mathcal{L}_Q, F(.))$*
  - *FP với vấn đề $RDC(\mathcal{L}_Q, F(.))$*

không quan trọng hàm mục tiêu là $F_{MS}(.)$,$F_{MM}(.)$,$F_{mono}(.)$ và ngôn ngữ truy vấn là CQ, UCQ, $\exists FO^+$ và FO.
Tổng kết Từ những kết quả trên chúng ta thấy;

- Ảnh hưởng của ngôn ngữ $\mathcal{L}_Q$: Khi F là $F_{MS}(.)$ hoặc $F_{MM}(.)$, các vấn đề $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$ và $RDC(\mathcal{L}_Q, F(.))$ độ phức tạp cho ngôn ngữ FO cao hơn so với các ngôn ngữ CQ, UCQ và $\exists FO^+$.Khi F là $F_{mono}(.)$, các giới hạn về độ phức tạp kết hợp của các vấn đề này vẫn không thay đổi khi LQ thay đổi (Theorems 1, 2 và 3). Ngược lại, đối với lớp các truy vấn nhận dạng, tất cả những vấn đề này trở nên đơn giản hơn. Ngôn ngữ truy vấn $\mathcal{L}_Q$ không ảnh hưởng đến độ phức tạp của các vấn đề này.

- Ảnh hưởng của hàm tính độ đa dạng và hàm tính độ liên quan: Độ phức tạp của bài toán xuất phát từ hàm tính độ đa dạng trong hàm mục tiêu.

- Ảnh hưởng của k: Khi k là hằng số cố định, độ phức tạp dữ liệu của $QRD(\mathcal{L}_Q, F(.))$, $DRP(\mathcal{L}_Q, F(.))$ và $RDC(\mathcal{L}_Q, F(.))$ trở nên đơn giản hơn, với hàm mục tiêu F là $F_{MS}(.)$ hoặc $F_{MM}(.)$

Bảng 1 Độ phức tạp hỗn hợp và độ phức tạp sữ liệu

Hàm mục tiêu	Ngôn ngữ	Vấn đề		
		$QRD(\mathcal{L}_Q, F(.))$	$DRP(\mathcal{L}_Q, F(.))$	$RDC(\mathcal{L}_Q, F(.))$
		Độ phức tạp hỗn hợp		
$F_{MS}(.)$ và $F_{MM}(.)$	CQ, UCQ và $\exists FO^+$	NP-complete	coNP-complete	#.NP-complete
	FO	PSPACE-complete	PSPACE-complete	#.PSPACE-complete
$F_{mono}(.)$	CQ, UCQ $\exists FO^+$, FO	PSPACE-complete	PSPACE-complete	#.PSPACE-complete
		Độ phức tạp dữ liệu		
$F_{MS}(.)$ và $F_{MM}(.)$	CQ, UCQ và $\exists FO^+$,FO	NP-complete	coNP-complete	#P-complete
$F_{mono}(.)$	CQ, UCQ $\exists FO^+$, FO	PTIME	PTIME	#P-complete

Bảng 2 Các trường hợp đặc biệt

Điều kiện	Độ phức tạp	Vấn đề		
		$QRD(\mathcal{L}_Q, F(.))$	$DRP(\mathcal{L}_Q, F(.))$	$RDC(\mathcal{L}_Q, F(.))$
Truy vấn định danh F là $F_{mono}(.)$	Độ phức tạp hỗn hợp	PTIME	PTIME	#P-complete(Turing)
$\lambda = 0$ F là $F_{MS}(.)$	Độ phức tạp dữ liệu	PTIME	PTIME	#P-complete(Turing)
$\lambda = 0$ F là $F_{MM}(.)$	Độ phức tạp dữ liệu	PTIME	PTIME	FP
$\lambda = 0$ F là $F_{mono}(.)$, $\mathcal{L}_Q$ là CQ, UCQ $\exists FO^+$	Độ phức tạp hỗn hợp	NP-complete	coNP-complete	#NP-complete
k là hằng số	Độ phức tạp dữ liệu	PTIME	PTIME	FP

2 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.

2.1 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 $w(\cdot)$ 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.

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

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 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.

2.2 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 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:

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

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 thoải mãn $S_k^* \subset S_{k+1}^*$.

5. **Độc lập với phần tử không tương thích**. Tiên đề này nói 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 đợi 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 thoải 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:

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

   (b) Nếu $f(S \setminus x) < f(S)$ thì tồi 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 thoải 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:

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

   (b) Nếu $f(S\backslash x) < f(S)$ thì tồi 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.

2.3 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 thoải 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.

2.3.1 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:

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

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:

$$d'(u, v) = w(u) + w(v) + 2\lambda d(u, v) \tag{2}$$

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 = ∅
for i ← 1 to ⌊k/2⌋ do
    Tìm (u, v) = argmax_{x,y∈U} d(x, y)
    Tập S = S ∪ {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
```

2.3.2 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:

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

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:

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

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 $S = \varnothing$
  Tìm $(u, v) = \text{argmax}_{x,y \in U}\, d(x, y)$ 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 $x = \text{argmax}_{x \in U \setminus S}\, d(x, S)$
    Tập $S = S \cup \{x\}$
  end while

2.3.3 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:

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

Trong đó

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

Giá trị tham số λ thể hiện sử 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.