Лекция 12.

Задачи поиска на конечных
частично-упорядоченных множествах данных.
Задача включающего поиска.

1 Задачи поиска на конечных частично-упорядоченных множествах данных

Задачи поиска, в которых отношение поиска, является отношением частичного порядка, встречаются практически во всех системах баз данных и в зависимости от интерпретации носят название включающего поиска, дескрипторного поиска, поиска по ключевым словам, задачи о доминировании и т. п.

В данном разделе мы приведем некоторые свойства отношений частичного порядка, полезные в том случае, когда эти отношения используются в качестве отношений поиска. Мы будем рассматривать типы задач поиска, в которых в качестве отношения поиска выступает отношение частичного порядка, а именно рассмотрим следующий тип $S_{part} = \langle X, X, \succeq \rangle$, где X — некоторое конечное множество, $\succeq$ — некоторое *отношение частичного порядка* на $X \times X$, под которым будем понимать отношение, для любых $x, y, z \in X$ удовлетворяющее условиям

- рефлексивности $x \succeq x$;
- транзитивности $(x \succeq y)\&(y \succeq z) \to (x \succeq z)$;
- антисимметричности $(x \succeq y)\&(y \succeq x) \to (x = y)$.

Пусть $a \in X$, K_a — функция, действующая из X в $\{0, 1\}$ такая, что $N_{K_a} = \{x \in X : x \succeq a\}$. Отметим, что $K_a(x)$ есть характеристическая функция записи a. Пусть $\mathcal{K} = \{K_a(x) : a \in X\}$, $\mathcal{M} = \{\bigvee_{i=1}^{n} K_{a_i}(x) : a_i \in X, n \in \mathbf{N}\}$, где $\mathbf{N}$ — множество натуральных чисел.

Справедливы следующие свойства.

Свойство 1. *Если* $a, b \in X$ *и* $K_b(a) = 1$, *то* $N_{K_a} \subseteq N_{K_b}$.

Доказательство. Так как $K_b(a) = 1$, то $a \succeq b$. Для любого $c \in N_{K_a}$ справедливо $c \succeq a \succeq b$. Отсюда следует, что $c \succeq b$ и, следовательно, $c \in N_{K_b}$. Что и требовалось доказать.

Свойство 2. *Если* $a \in X$, $f \in \mathcal{M}$ *и* $f(a) = 1$, *то* $N_{K_a} \subseteq N_f$.

Доказательство. Пусть $f = \bigvee_{i=1}^{n} K_{a_i}$. Так как $f(a) = 1$, то существует K_{a_i} такое, что $K_{a_i}(a) = 1$, но тогда согласно свойству 1 $N_{K_a} \subseteq N_{K_{a_i}}$ и, следовательно, $N_{K_a} \subseteq N_f$. Что и требовалось доказать.

Свойство 3. *Пусть* $a \in X$ *и* $\bigvee_{i=1}^{n} f_i = K_a$, *где* $f_i \in \mathcal{M}$. *Тогда существует такое* $i \in \{\overline{1,n}\}$, *что* $f_i = K_a$.

Доказательство. Так как $K_a(a) = 1$, то существует такое i, что $f_i(a) = 1$. Тогда согласно свойству 2 имеем $N_{K_a} \subseteq N_{f_i}$, но из условия следует, что $N_{f_i} \subseteq N_{K_a}$, следовательно, $N_{K_a} = N_{f_i}$, то есть $K_a = f_i$. Что и требовалось доказать.

Свойство 4. *Если* $a \in X$, $f = \bigwedge_{i=1}^{n} f_i$, *где* $f_i \in \mathcal{M}$, *и* $f(a) = 1$, *то* $N_{K_a} \subseteq N_f$.

Доказательство. $N_f = \bigcap_{i=1}^{n} N_{f_i}$. Так как $a \in N_f$, то $a \in N_{f_i}$ для любого $i = 1, \ldots, n$. Отсюда согласно свойству 2 имеем $N_{K_a} \subseteq N_{f_i}$ для любого $i = 1, \ldots, n$. Следовательно, $N_{K_a} \subseteq \bigcap_{i=1}^{n} N_{f_i} = N_f$. Что и требовалось доказать.

Точку $a \in R \subseteq X$ назовем *нижней единицей множества* R, если не существует точки $b \in R \setminus \{a\}$ такой, что $a \succeq b$.

Свойство 5. *Если* $f = \bigwedge_{i=1}^{n} f_i$, *где* $f_i \in \mathcal{M}$, *то либо* $f \in \mathcal{M}$, *либо* $f \equiv 0$.

Доказательство. $N_f = \bigcap_{i=1}^{n} N_{f_i}$. Допустим, что $\bigcap_{i=1}^{n} N_{f_i} \neq \emptyset$, то есть $f \not\equiv 0$. Просмотрим все точки множества N_f (а их конечное число) и для каждой проверим, является ли она нижней единицей? Пусть $M = \{a_1, \ldots, a_m\}$ — множество всех нижних единиц множества N_f, которое мы в результате получили. Легко видеть, что для любой точки $a \in N_f$ либо $a \in M$, либо существует $a_i \in M$ такое, что $a \succeq a_i$. Из того, что $a \succeq a_i$ следует $K_{a_i}(a) = 1 \Rightarrow a \in N_{K_{a_i}}$. Следовательно, $N_f \subseteq \bigcup_{i=1}^{m} N_{K_{a_i}}$. С другой стороны, согласно свойству 4 (используем то, что $f(a_i) = 1$) для любого $i \in \{1, \ldots, m\}$ имеем $N_{K_{a_i}} \subseteq N_f$. Следовательно, $N_f = \bigcup_{i=1}^{m} N_{K_{a_i}}$ и, значит, $f \in \mathcal{M}$, что и требовалось доказать.

Пусть базовое множество имеет вид $\mathcal{F} = \langle \mathcal{M}, \emptyset \rangle$. Пусть U — ПИГ над базовым множеством $\mathcal{F}$. Подмножество $\{\beta_1, \ldots, \beta_m\}$ вершин ПИГ U назовем *характерным*, если существует такая запись $a \in X$, что $\bigvee_{i=1}^{m} \varphi_{\beta_i} = K_a$.

Пусть U — ПИГ над базовым множеством $\mathcal{F} = \langle \mathcal{M}, \emptyset \rangle$. Пусть $\mathcal{B} = \{\beta_1, \ldots, \beta_m\}$ — характерное подмножество вершин ПИГ U такое, что $\bigvee_{i=1}^{m} \varphi_{\beta_i} = K_a$. Если в ПИГ U существует такая цепь, ведущая из корня в какую-либо вершину множества $\mathcal{B}$, что проводимость этой цепи равна K_a, то эту цепь назовем *главной цепью характерного множества* $\mathcal{B}$.

Теорема 1 (о существовании главных цепей). *Пусть U — произвольный ПИГ над базовым множеством $\mathcal{F} = \langle \mathcal{M}, \emptyset \rangle$. Пусть $\mathcal{B}$ — произвольное характерное подмножество вершин ПИГ U. Тогда в ПИГ U существует главная цепь множества $\mathcal{B}$.*

Доказательство. Рассмотрим некоторую цепь C графа U. Обозначим через f_C проводимость цепи C. Пусть C такая цепь, что $f_C \not\equiv 0$. Так как f_C равна конъюнкции нагрузок ребер цепи C, то согласно свойству 5 $f_C \in \mathcal{M}$. Пусть $\mathcal{C}_\mathcal{B}$ - множество цепей, ведущих из корня в вершины множества $\mathcal{B}$. Так как $\mathcal{B}$ — характерное множество, то существует такая запись $a \in X$, что $\bigvee_{\beta \in \mathcal{B}} \varphi_\beta = \bigvee_{C \in \mathcal{C}_\mathcal{B}} f_C = K_a$. Отсюда согласно свойству 3

существует цепь $C' \in \mathcal{C}_{\mathcal{B}}$ такая, что $f_{C'} = K_a$. Эта цепь C' и есть главная цепь множества $\mathcal{B}$. Что и требовалось доказать.

Пусть U — ПИГ над базовым множеством $\mathcal{F} = \langle \mathcal{M}, \emptyset \rangle$, решающий некоторую ЗИП $I = \langle X, V, \succeq \rangle$ типа S_{part}. Пусть $y \in V$. Если в графе U существует такая цепь, ведущая из корня в какую-либо вершину множества $L_U(y)$, что проводимость этой цепи равна $\chi_{y,\succeq}$, то эту цепь назовем *главной цепью записи* y.

Следствие 1. *Пусть $I = \langle X, V, \succeq \rangle$ — ЗИП типа S_{part}. Пусть U — ПИГ над базовым множеством $\mathcal{F} = \langle \mathcal{M}, \emptyset \rangle$, решающий ЗИП I. Тогда для любой записи $y \in V$ в графе U существует главная цепь записи y.*

Доказательство. В силу свойства рефлексивности отношения $\succeq O(y, \succeq) \neq \emptyset$. Так как ПИГ U решает ЗИП I, то согласно критерию допустимости информационных графов $\bigvee\limits_{\alpha \in L_U(y)} \varphi_\alpha = \chi_{y,\succeq} = K_y$. Следовательно, множество $L_U(y)$ есть характерное множество, и согласно теореме 1 в графе U существует главная цепь множества $L_U(y)$. Что и требовалось доказать.

2 Включающий поиск

Среди задач поиска, в которых в качестве отношения поиска выступают отношения частичного порядка, наверное, наиболее распространенными являются задачи включающего поиска (set-inclusion search problem).

Приведем наиболее распространенную интерпретацию задачи включающего поиска.

Предположим, что мы осуществляем поиск в дескрипторных автоматизированных информационно-поисковых системах, то есть информационный массив состоит из документов, каждый документ описывается множеством дескрипторов (ключевых слов), запрос задает некоторую совокупность дескрипторов, и необходимо перечислить в информационном массиве все документы, содержащие в своем описании все дескрипторы, входящие в запрос. Занумеруем некоторым образом множество всех дескрипторов (*тезаурус*). Каждому документу сопоставим запись, представляющую собой булев вектор длины, равной мощности тезауруса, в i-ой компоненте которого стоит 0 (ноль) в том и только в том случае, когда i-ый дескриптор входит в описание данного документа. Запросы

будем описывать аналогичными векторами, то есть в i-ой компоненте запроса будет стоять 0, если i-ый дескриптор входит в запрос. Теперь если в качестве отношения поиска взять отношение $\overset{b}{\succeq}$, определяемое следующим соотношением

$$(x_1, \ldots, x_n) \overset{b}{\succeq} (y_1, \ldots, y_n) \Longleftrightarrow x_i \geq y_i,\ i = \overline{1, n},\ x_i, y_i \in \{0, 1\},$$

то получившаяся задача включающего поиска будет полностью соответствовать исходной. В самом деле, каждая компонента запроса должна быть не меньше соответствующей компоненты записи, или если в i-ой компоненте запроса стоит 0 (то есть i-ый дескриптор входит в запрос), то в i-ой компоненте записи тоже должен стоять 0 (то есть запись должна содержать i-ый дескриптор).

В общем случае включающий поиск встречается всегда, когда элементы информационного массива задаются множеством свойств (в частности, в дескрипторных автоматизированных информационно-поисковых системах — свойствами наличия дескриптора в описании документа) и необходимо перечислить в этом массиве элементы с определенным набором свойств, задаваемым запросом.

Хотя задачи включающего поиска имеют широкое применение, исследование оценок сложности включающего поиска не проводилось.

Для этих задач удалось получить нижнюю оценку в два раза лучшую, чем мощностная нижняя оценка. Кроме того, было доказано существование задач, для которых эта нижняя оценка асимптотически не улучшаема. Эти и некоторые другие результаты приводятся в данном разделе.

Рассмотрим следующий тип задач поиска:

$$S_{bool} = \langle B^n, B^n, \overset{b}{\succeq}, \sigma, \mathbf{P} \rangle,$$

где B^n — *единичный n-мерный куб*, то есть

$$B^n = \{(\alpha_1, \ldots, \alpha_n) :\ \alpha_i \in \{0, 1\}\ (i = \overline{1, n})\};$$

$\overset{b}{\succeq}$ — отношение поиска на $B^n \times B^n$, определяемое следующим соотношением $(x_1, \ldots, x_n) \overset{b}{\succeq} (y_1, \ldots, y_n) \Longleftrightarrow x_i \geq y_i,\ i = \overline{1, n}$; σ — алгебра подмножеств B^n, представляющая собой множество всех подмножеств B^n;

$\mathbf{P}$ — равномерная вероятностная мера на σ, то есть такая мера, что для любого $x \in B^n$ $\mathbf{P}(x) = 1/2^n$ и любого $A \subseteq B^n$ $\mathbf{P}(A) = |A|/2^n$.

Задачи поиска, принадлежащие данному типу, есть разновидность задач, именуемых в литературе задачами включающего поиска, поэтому тип S_{bool} мы назовем *типом включающего поиска*, а задачи, принадлежащие этому типу, — *задачами включающего поиска.*

Гранью единичного n-мерного куба B^n называется множество

$$\{(\alpha_1, \ldots, \alpha_n) \in B^n : \alpha_{i_1} = \sigma_1, \ldots, \alpha_{i_k} = \sigma_k\},$$

при этом число $n - k$ называется *размерностью* этой грани.

Весом набора $(\alpha_1, \ldots, \alpha_n) \in B^n$ называют число его координат, равных 1. Множество вершин куба, имеющих вес k, называется *k-м слоем куба* B^n и обозначается B^n_k. *Номером набора* $\alpha = (\alpha_1, \ldots, \alpha_n) \in B^n$ назовем число $||\alpha|| = \sum_{i=1}^{n} 2^{n-i} \cdot \alpha_n$. Будем считать, что наборы в слое куба упорядочены в порядке убывания их номеров. Множество, состоящее из m первых наборов k-го слоя куба B^n, будем называть *начальным отрезком* длины m этого слоя.

Формула $x_{i_1}^{\sigma_1} \& x_{i_2}^{\sigma_2} \& \ldots \& x_{i_r}^{\sigma_r}$, где $\&$ - знак конъюнкции, $\sigma_k \in \{0, 1\}$, $x_{i_k}^0 = \overline{x}_{i_k}$, $x_{i_k}^1 = x_{i_k}$, $i_k \in \{1, 2, \ldots, n\}$, $k = 1, 2, \ldots, r$ ($r \geq 1$ и $n \geq 1$), называется *конъюнкцией над множеством переменных* $X^n = \{x_1, x_2, \ldots, x_n\}$, а число r — длиной этой конъюнкции. Если $x_{i_j} \neq x_{i_k}$ при $j \neq k$, то конъюнкция называется *элементарной*. Элементарная конъюнкция называется *монотонной*, если она не содержит отрицаний переменных. Множество элементарных монотонных конъюнкций от n переменных будем обозначать через $\mathcal{K}^n$. Функцию тождественная единица будем считать элементарной монотонной конъюнкцией длины 0, то есть будем считать, что она входит в множество $\mathcal{K}^n$, кроме того добавим тождественную единицу и в множество переменных X^n.

Будем говорить, что две элементарные монотонные конъюнкции *пересекаются по переменным*, если в формулах, описывающих эти конъюнкции, встречаются одинаковые переменные.

Функция алгебры логики $f(x_1, \ldots, x_n)$ называется *монотонной*, если для любых двух наборов α и β из B^n таких, что $\alpha \overset{b}{\succeq} \beta$, имеет место неравенство $f(\alpha) \geq f(\beta)$. Дизъюнкция элементарных монотонных конъюнкций есть монотонная функция. Множество монотонных булевых функций от n переменных будем обозначать через $\mathcal{M}^n$.

Рассмотрим произвольную запись $y \in B^n$. Нетрудно заметить, что если $\{i_1, \ldots, i_k\}$ есть множество номеров координат вектора y, которые равны 1, то характеристическая функция записи является элементарной монотонной конъюнкцией

$$\chi_{y,\overset{b}{\succeq}}(x_1, \ldots, x_n) = x_{i_1} \& x_{i_2} \& \ldots \& x_{i_k}.$$

Далее часто знак конъюнкции & в формулах мы будем опускать.

Таким образом, согласно критерию допустимости информационных графов граф, решающий некоторую задачу включающего поиска, представляет собой многополюсник, реализующий некоторую систему элементарных монотонных конъюнкций. Отсюда согласно критерию полноты базового множества для типа задач информационного поиска следует, что каждое из базовых множеств $\langle \mathcal{M}^n, \emptyset \rangle$, $\langle \mathcal{K}^n, \emptyset \rangle$ и $\langle X^n, \emptyset \rangle$ является полным для типа S_{bool}.

Следовательно, тип S_{bool} полностью удовлетворяет условиям теоремы о существовании оптимальных информационных графов, и, значит, для любой ЗИП I типа S_{bool} существует оптимальный ПИГ над любым из базовых множеств $\langle \mathcal{M}^n, \emptyset \rangle$, $\langle \mathcal{K}^n, \emptyset \rangle$ и $\langle X^n, \emptyset \rangle$.

Упражнения

1. Пусть $V = \{y_1, y_2, \ldots, y_k\} \subseteq B^n$ и число единиц в наборе y_i равно t_i $(i = 1, 2, \ldots, k)$. Приведите мощностную нижнюю оценку для задачи включающего поиска $I = \langle B^n, V, \overset{b}{\succeq} \rangle$.