Лекция 7.

Задачи поиска с коротким ответом.
Теорема о существовании оптимального
древовидного информационного графа
для задач поиска с коротким ответом.

1 Задачи поиска с коротким ответом

Класс задач с коротким ответом — это совокупность задач поиска, в которых ответ на любой запрос содержит малое число записей. Здесь мы будем рассматривать, когда в ответе не более одной записи. Наиболее известными задачами из данного класса являются задача поиска идентичных объектов и задача поиска ближайшего объекта.

Результат, описываемый в данной лекции, относится к классу предикатных информационных графов (ПИГ). Напомним, что ПИГ — это такие ИГ, базовое множество которых не содержит переключателей, то есть ИГ, в которых имеются только предикатные ребра.

Напомним также, что ПИГ, различным листьям которого соответствуют различные записи, называется однозначным информационным графом (ОИГ), однозначный информационный граф, имеющий вид дерева, листья которого совпадают с концевыми вершинами дерева, называется информационным деревом (ИД).

ИД удобны и интересны тем, что структуры данных, им соответствующие, практичны и их гораздо проще реализовать на ЭВМ. Тогда как ПИГ обладают большими возможностями и охватывают более широкий класс алгоритмов. Поэтому представляет интерес выявление классов задач информационного поиска, для которых оптимальные (то есть с минимальной сложностью) ПИГ находятся в классе ИД.

Один из таких классов приводится в данном разделе. По сути это такой класс задач поиска, в которых мера множества запросов, содержащих в ответе задачи более одного элемента, равна 0.

Пусть нам даны множества запросов X, записей Y и отношение поиска ρ на $X \times Y$. Причем на множестве запросов задано вероятностное пространство $\langle X, \sigma, \mathbf{P} \rangle$.

Скажем, что ЗИП $I = \langle X, V, \rho \rangle$ *обладает A-свойством*, если

- для любой записи $y \in V$ $O(y, \rho) \in \sigma$ и $\mathbf{P}(O(y, \rho)) \neq 0$;
- для любых $y, y' \in V$, таких, что $y \neq y'$ $\mathbf{P}(O(y, \rho) \cap O(y', \rho)) = 0$.

Класс задач, обладающих A-свойством, мы и будем исследовать.

2 Теорема о существовании оптимального древовидного информационного графа для задач поиска с коротким ответом

В этом пункте мы докажем теорему о существовании древовидного оптимального графа для задач, обладающих A-свойством.

Скажем, что *вершина α графа схемно достижима из вершины β*, если из β в α существует ориентированная цепь.

Пусть β — вершина некоторого ИГ. Обозначим через V_β множество записей, соответствующих листьям, схемно достижимым из вершины β.

Скажем, что ИГ *обладает C-свойством*, если для любой вершины β ИГ, за исключением корня $\varphi_\beta = \bigvee\limits_{y \in V_\beta} \chi_{y,\rho}$.

Пусть $I = \langle X, V, \rho \rangle$ — некоторая ЗИП, где $V = \{y_1, y_2, \ldots, y_k\}$, тогда обозначим

$$F_0^I = \{\bigvee_{j=1}^{m} \chi_{y_{i_j},\rho} : m = \overline{1,k},\ 1 \leq i_1 < i_2 < \cdots < i_m \leq k\}.$$

Скажем, что ИД над базовым множеством $\mathcal{F} = \langle F_0^I, \emptyset \rangle$ *обладает D_I-свойством*, если оно решает ЗИП I, обладает C-свойством и у любой вершины ИД, не являющейся полюсом, полустепень исхода больше 1.

Обозначим через $\mathcal{D}^I$ множество всех ИД, обладающих D_I-свойством.

Справедлива следующая теорема.

Теорема 1. *Пусть $I = \langle X, V, \rho\rangle$ — ЗИП, $\mathcal{F} = \langle F, \emptyset\rangle$ — произвольное измеримое базовое множество, допустимое для I, U — произвольный ПИГ над базовым множеством $\mathcal{F}$, решающий ЗИП I. Тогда, если I обладает A-свойством, то существует ИД $D \in \mathcal{D}^I$, такое, что $T(D) \le T(U)$.*

Доказательство. Обозначим через

$$O'(y,\rho) = O(y,\rho) \setminus (\bigcup_{\substack{y'\in V\\ y'\neq y}} O(y',\rho)).$$

Понятно, что если I обладает A-свойством, то $\mathbf{P}(O'(y,\rho)) = \mathbf{P}(O(y,\rho))$.

Обозначим через F_1 следующее бесконечное множество предикатов

$$F_1 = \{f_A : N_{f_A} = A, A \in \sigma\}.$$

Отметим, что так как $\mathcal{F}$ измеримо, то $F \subseteq F_1$. Если I обладает A-свойством, то поскольку σ — алгебра, то $F_0^I \subseteq F_1$.

Скажем, что ПИГ над базовым множеством $\langle F_1, \emptyset\rangle$ *обладает E-свойством*, если он решает ЗИП I и для любого листа ПИГ полустепень исхода этого листа равна 0.

Покажем, что существует ОИГ U_0, обладающий E-свойством, такой, что $T(U_0) \le T(U)$.

Пусть $C = (\alpha_1,\alpha_2)(\alpha_2,\alpha_3)\cdots(\alpha_{r-1},\alpha_r)$ — цепь в ПИГ, где α_1 — корень ПИГ. Множество ребер, исходящих из вершин $\alpha_1, \alpha_2, \dots, \alpha_{r-1}$, назовем *следом цепи C*, а множество ребер, исходящих из вершин $\alpha_2, \dots, \alpha_{r-1}$ — *усеченным следом цепи C*. Соответственно число $n = \sum_{i=2}^{r-1} \psi_{\alpha_i}$ будет *мощностью усеченного следа цепи C*.

Пусть $V = \{y_1, y_2, \dots, y_k\}$.

Рассмотрим сначала запись y_1.

Обозначим через $\mathcal{C}_{y_1} = \{C_1, C_2, \dots, C_m\}$ — множество цепей, ведущих из корня в листья множества $L_U(y_1)$. Пусть $f_1, \dots, f_m$ функции проводимости цепей $C_1, \dots, C_m$ соответственно. Так как U решает I, то согласно критерию допустимости информационных графов $\bigvee_{i=1}^{m} f_i = \chi_{y_i,\rho}$, или $\bigcup_{i=1}^{m} N_{f_i} = O(y_i,\rho)$.

Выберем в $\mathcal{C}_{y_1}$ подмножество цепей, такое, что характеристические множества их функций проводимости образуют тупиковое покрытие $O'(y_1, \rho)$. Без ограничения общности можно считать, что это первые s цепей ($s \leq m$), то есть $\bigcup\limits_{i=1}^{s} N_{f_i} \supseteq O'(y_1, \rho)$, но удаление любого множества из объединения в левой части нарушает данное соотношение.

Пусть $N_i' \subseteq N_{f_i}(i = \overline{1, s})$, такие, что $N_i' \cap N_j' = \emptyset$, если $i \neq j$ ($i, j \in \{\overline{1, s}\}$) и $\bigcup\limits_{i=1}^{s} N_i' = O'(y_1, \rho)$.

Понятно, что такие N_i' ($i = \overline{1, s}$) можно подобрать.

Обозначим через n_i мощность усеченного следа цепи C_i ($i = \overline{1, m}$).

Пусть c ребро ПИГ, через $[c]$ будем обозначать его нагрузку, то есть предикат, приписанный этому ребру.

Для каждого ребра c графа U заменим его нагрузку $[c]$ на $f_{N_{[c]} \setminus O'(y_1, \rho)}$. Так как $N_{[c]} \in \sigma$ и $O'(y_1, \rho) \in \sigma$, то $f_{N_{[c]} \setminus O'(y_1, \rho)} \in F_1$.

После такой операции функции фильтра листьев, не принадлежащих $L_U(y_1)$, не изменятся, а дизъюнкция функций фильтра листьев из $L_U(y_1)$ станет равной $f_{O(y_1, \rho) \setminus O'(y_1, \rho)}$, при этом сложность графа уменьшится по крайней мере на $\sum\limits_{i=1}^{s} n_i \cdot \mathbf{P}(N_i')$.

Пусть $n_j = \min\limits_{1 \leq i \leq s} n_i$.

Пусть цепь C_j ведет в некоторый лист $\alpha_1 \in L_U(y_1)$. Все листья из $L_U(y_1)$, отличные от α_1, объявим обычными вершинами и уберем приписанную им нагрузку y_1. После этой операции множество $L_U(y_1)$ будет состоять только из одного листа α_1.

Для каждого ребра c цепи C_j заменим его нагрузку $[c]$ на $[c] \vee \chi_{y_1, \rho}$. После этой операции функция фильтра листа α_1 станет равной $\varphi_{\alpha_1} = \chi_{y_1, \rho}$. Таким образом, полученный граф снова решает задачу I.

В результате последней операции сложность графа увеличится на $n_j \cdot \mathbf{P}(O(y_1, \rho))$.

Заметим, что

$$\sum_{i=1}^{s} n_i \cdot \mathbf{P}(N_i') \geq n_j \cdot \sum_{i=1}^{s} \mathbf{P}(N_i') = n_j \cdot \mathbf{P}(O'(y_1, \rho)) = n_j \cdot \mathbf{P}(O(y_1, \rho)).$$

Отсюда следует, что полученный граф по сложности не больше чем $T(U)$.

Переобозначим цепь C_j на C'_1.

Проделаем выше описанную процедуру для всех остальных записей y_i, $i = \overline{2,k}$, и для каждой записи y_i получим цепь C'_i, ведущую из корня в некоторый лист α_i (причем α_i будет единственным листом в $L_U(y_i)$) и имеющую проводимость $\chi_{y_i,\rho}$.

Удалим из полученного графа все ребра, не принадлежащие ни одной из цепей $C'_1, \dots, C'_k$. Граф U', получающийся после этого удаления, будет решать задачу I и иметь сложность, не превышающую $T(U)$.

Граф U' является ОИГ, так как каждой записи соответствует ровно один лист. Покажем, что в графе U' полустепень исхода любого листа равна 0.

Предположим, что это не так. Так как граф U' состоит только из ребер, принадлежащих цепям $C'_1, \dots, C'_k$, то, значит, некоторая цепь C'_j ($j \in \{\overline{1,k}\}$) проходит через некоторый лист α_m, где $m \in \{\overline{1,k}\}$ и $m \neq j$. Так как граф U' решает ЗИП I, то $\varphi_{\alpha_m} = \chi_{y_m,\rho}$. Следовательно, проводимость f_j цепи C'_j такая, что $N_{f_j} \subseteq O(y_m, \rho)$. Но этого не может быть, так как $N_{f_j} = O(y_j, \rho)$ и $\mathbf{P}(O(y_j, \rho) \cap O(y_m, \rho)) = 0$, и $\mathbf{P}(O(y_j, \rho)) \neq 0$.

Таким образом, мы доказали, что U' обладает E-свойством и, значит, его можно взять в качестве искомого графа U_0.

Теперь в графе U_0 изменим нагрузку всех ребер следующим образом. Пусть c произвольное ребро графа, такое, что оно принадлежит цепям $C'_{i_1}, \dots, C'_{i_m}$, тогда в качестве нагрузки этого ребра возьмем $\bigvee\limits_{j=1}^{m} \chi_{y_{i_j},\rho} \in F_0^I$. В частности нагрузка ребра, ведущего в некоторый лист α_i, будет равна $\chi_{y_i,\rho}$.

Поскольку эта замена не меняет проводимости ни одной из цепей $C'_1, \dots, C'_k$, то функционирование графа не меняется.

Граф, полученный после осуществления такой замены нагрузки для всех ребер, обозначим через U_1. $T(U_1) \leq T(U_0)$, так как в графе U_0 для каждого ребра, принадлежащего C'_i ($i \in \{\overline{1,k}\}$), конъюнкция его нагрузки и функции $\chi_{y_i,\rho}$ равна $\chi_{y_i,\rho}$.

ОИГ U_1 является графом над F_0^I, $T(U_1) \leq T(U)$, U_1 обладает E-свойством, причем в каждый лист графа U_1 ведет единственное ребро, и еще, если β вершина графа U_1, отличная от корня, через которую проходят некоторые цепи $C'_{i_1}, \ldots, C'_{i_s}$, то $\varphi_\beta = \bigvee\limits_{j=1}^{s} \chi_{y_{i_j},\rho}$. Причем так как через эту вершину β не проходит других цепей, ведущих в листья, то $V_\beta = \{y_{i_1}, \ldots, y_{i_s}\}$, откуда следует, что ОИГ U_1 обладает C-свойством.

Предположим, что в U_1 есть вершины с полустепенью захода более 1. Пусть этих вершин m штук.

Рассмотрим произвольную вершину β с полустепенью захода, превышающей 1.

Пусть через нее проходит s цепей $C'_{i_1}, \ldots, C'_{i_s}$. Согласно замечанию $\varphi_\beta = \bigvee\limits_{j=1}^{s} \chi_{y_{i_j},\rho}$.

Отметим, что ребра, входящие в вершину β не принадлежат никаким другим цепям, кроме цепей $C'_{i_1}, \ldots, C'_{i_s}$.

Обозначим через C''_{i_j} и C'''_{i_j} части цепи C'_{i_j} $(j \in \{\overline{1,s}\}$ соответственно от корня до вершины β и от вершины β до листа α_{i_j}, а через n'_{i_j} — мощность усеченного следа цепи C''_{i_j}. Обозначим через G_β подграф графа U_1, состоящий из цепей $C''_{i_1}, \ldots, C''_{i_s}$, а через $O_\beta = \bigcup\limits_{j=1}^{s} O'(y_{i_j}, \rho)$.

Для каждого ребра c подграфа G_β заменим его нагрузку $[c]$ на $f_{N_{[c]} \setminus O_\beta}$. Не трудно заметить, что после этой операции проводимости цепей C'_{i_m}, таких, что $m \notin \{\overline{1,s}\}$, не изменится, более того нагрузка ребер из этих цепей, как и прежде, будет принадлежать F_0^I. При этом сложность графа уменьшится по крайней мере на $\sum\limits_{i=1}^{s} n'_{i_j} \cdot \mathbf{P}(O'(y_{i_j}, \rho))$.

Пусть $n'_{i_l} = \min\limits_{1 \le j \le s} n'_{i_j}$.

Для каждого ребра c цепи C''_{i_l} $(j = \overline{1,s})$ заменим его нагрузку $[c]$ на $[c] \vee \bigvee\limits_{j=1}^{s} \chi_{y_{i_j},\rho}$. Эта замена увеличит сложность графа на $n'_{i_j} \cdot \mathbf{P}(O(y_{i_j}, \rho))$, но поскольку это не больше чем $\sum\limits_{j=1}^{s} n'_{i_j} \cdot \mathbf{P}(O'(y_{i_j}, \rho))$, то полученный граф по сложности не превышает $T(U_1)$.

Объявим цепью C'_{i_j} цепь, составленную из C''_{i_l} и C'''_{i_j} $(j = \overline{1,s})$. Нетрудно заметить, что проводимость новообъявленных цепей C'_{i_j} по-прежнему равна $\chi_{y_{i_j},\rho}$.

Теперь удалим все ребра, не принадлежащие ни одной из цепей C'_j ($j = \overline{1,k}$). В частности мы удалим все ребра, входящие в вершину β, кроме ребра, принадлежащего C''_{i_l}, в силу сделанного выше замечания о ребрах, входящих в β.

Полученный таким образом граф обозначим U_2. Мы получили, что $T(U_2) \leq T(U_1)$, U_2 решает задачу I, и число вершин с полустепенью захода более 1 в U_2 по крайней мере на 1 меньше чем в U_1, поскольку теперь в вершину β ведет единственное ребро, а новых вершин с полустепенью захода более 1 образоваться не могло.

Нетрудно заметить, что $\widehat{T}(U_2) \leq \widehat{T}(U_1)$. В самом деле, для любого $x \in X \setminus O_\beta$ $T(U_2, x) = T(U_1, x)$, а для любого $j \in \{\overline{1,s}\}$ и для любого $x \in O(y_{i_j}, \rho)$ $T(U_2, x) = T(U_1, x) - n'_{i_j} + n'_{i_l} \leq T(U_1, x)$.

Применяя вышеописанную процедуру к графу U_2, мы получим граф U_3 с еще меньшим числом вершин с полустепенью захода более 1.

Применив данную процедуру нужное количество раз, мы получим некий граф U_r ($r \leq m+1$), в котором полустепень захода любой вершины равна 1.

Поскольку плюс к этому полустепень исхода любого листа U_r равна 0, то U_r является информационным деревом. По построению для любой некорневой вершины графа U_r выполняется условие $\varphi_\beta = \bigvee\limits_{y \in V_\beta} \chi_{y,\rho}$. Предположим, что в U_r есть неполюсная вершина β, полустепень исхода которой равна 1. Тогда нагрузка ребер, входящего в β и исходящего из β, одинакова. Следовательно, ребро, исходящее из β, можно удалить без ущерба для функционирования и сложности. Эту операцию повторим для каждой неполюсной вершины, полустепень исхода которой равна 1. После этого U_r будет ИД, обладающим D_I-свойством. Поскольку $T(U_r) \leq T(U)$ и $\widehat{T}(U_r) \leq \widehat{T}(U)$ по построению, то U_r можно взять в качестве искомого дерева D.

Тем самым теорема доказана.

Следствие 1. *Если ЗИП I обладает A-свойством $\mathcal{F} = \langle F, \emptyset \rangle$ — измеримое базовое множество, допустимое для I, такое, что $F_0^I \subseteq F$, то существует оптимальный ПИГ U для ЗИП I, принадлежащий классу $\mathcal{D}^I$.*

Доказательство. Так как $F_0^I \subseteq F$, то $\mathcal{D}^I \subseteq \mathcal{U}(I, \mathcal{F})$, то есть оптимальные ИГ, если они существуют, надо искать согласно теореме 1 в классе $\mathcal{D}^I$. Чтобы показать существование оптимального ИГ, достаточно заметить, что $\mathcal{D}^I$ — конечное множество. В самом деле, ИГ из $\mathcal{D}^I$ — это деревья с k концевыми вершинами (здесь k — количество записей в библиотеке задачи I), нагрузка ребер которых берется из конечного множества F_0^I, значит, $\mathcal{D}^I$ — конечное множество.

Тем самым следствие доказано.