

Асимптотическое решение задачи о метрической близости для одного базового множества функций*

Е.С. Быченкова

В работе исследуется среднее время поиска в задачах о метрической близости на n -мерном булевом кубе для одного базового множества функций. Найдена верхняя оценка среднего времени для любых задач данного типа и показано существование такого класса задач, для которых верхняя оценка асимптотически совпадает с нижней. На основе этих результатов получена асимптотика функции Шеннона для сложности задачи о метрической близости.

1. Введение

В данной работе исследуется следующая задача поиска. Дано конечное множество вершин n -мерного булева куба B_2^n , которое в дальнейшем будем называть библиотекой. Элементы библиотеки назовем записями. Запрос на поиск задает некая вершина куба $\bar{x} = (x_1, \dots, x_n)$. Надо перечислить все записи из библиотеки, которые отличаются от запроса не более, чем в одной компоненте. Назовем эту задачу *задачей о метрической близости*.

Приведем интерпретацию данной задачи. Предположим, что мы осуществляем поиск в дескрипторных информационно-поисковых системах [1], то есть информационный массив состоит из документов, каждый документ описывается множеством дескрипторов

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (гранты 98-01-00130, 01-01-00748)

(ключевых слов), запрос задает некоторую совокупность дескрипторов, и необходимо перечислить в информационном массиве все документы, описание которых отличается не более, чем в одном дескрипторе. Занумеруем некоторым образом множество всех дескрипторов (пусть их n штук). Каждому документу сопоставим запись, представляющую собой булев вектор длины n , в i -й компоненте которого стоит 1 в том и только том случае, когда i -й дескриптор входит в описание данного документа. Запросы описываются аналогично, то есть в i -й компоненте запроса стоит 1, если i -й дескриптор входит в описание запроса. При поиске искомая запись должна отличаться не более, чем в одной компоненте от запроса. Таким образом, получаем задачу о метрической близости на булевом кубе.

В работе для некоторого фиксированного базового множества функций найдена верхняя оценка среднего времени поиска для любых задач о метрической близости. Показано, что для класса задач о метрической близости, у которых библиотека удовлетворяет следующим условиям:

- нет записей, компоненты которых только 1 или только 0,
- все записи попарно различаются более, чем в четырех компонентах,

верхняя оценка асимптотически совпадает с нижней при $n \rightarrow \infty$. Кроме того, для функции Шеннона, которая равна максимуму среднего времени поиска в задачах, размер библиотеки которых не больше k , получена асимптотика для некоторых k при $n \rightarrow \infty$.

Автор выражает благодарность Э.Э. Гасанову за постановку задачи и помощь в работе.

2. Основные понятия и формулировка результатов

Мы будем использовать терминологию и обозначения из работы [2], но поскольку в настоящей работе не используются переключатели, то здесь будет приведена несколько упрощенная версия понятия информационного графа.

Пусть X – множество запросов, причем на X определено вероятностное пространство $\langle X, \sigma, \mathbf{P} \rangle$, где σ – алгебра подмножеств множества X , \mathbf{P} – вероятностная мера на σ ; Y – множество записей (объектов поиска); ρ – бинарное отношение на $X \times Y$, называемое отношением поиска; пятерку $S = \langle X, Y, \rho, \sigma, \mathbf{P} \rangle$ будем называть *типом*; тройку $I = \langle X, V, \rho \rangle$, где V – некоторое конечное подмножество множества Y , в дальнейшем называемое *библиотекой*, будем называть задачей информационного поиска (ЗИП) типа $S = \langle X, Y, \rho, \sigma, \mathbf{P} \rangle$, и будем считать, что задача $I = \langle X, V, \rho \rangle$ содержательно состоит в перечислении для произвольно взятого запроса $x \in X$ всех тех и только тех записей из V , которые находятся в отношении ρ с запросом x , то есть удовлетворяют запросу x ; $O(y, \rho) = \{x \in X : x\rho y\}$ – тень записи $y \in Y$; F – множество символов одноместных предикатов, определенных на множестве X , называемое *базовым множеством*.

Понятие *информационного графа* (ИГ) над базовым множеством F , определяется следующим образом. Берется конечная многополюсная ориентированная сеть. В ней выбирается некоторый полюс, который называется корнем. Остальные полюса называются листьями и им приписываются записи из Y , причем разным листьям могут быть приписаны одинаковые записи. Ребрам приписываются предикаты из множества F . Таким образом нагруженную многополюсную ориентированную сеть называем ИГ над базовым множеством F .

Функционирование ИГ определяется следующим образом. Скажем, что: ребро проводит запрос $x \in X$, если предикат, приписанный этому ребру, принимает значение 1 на запросе x ; ориентированная цепочка ребер проводит запрос $x \in X$, если каждое ребро цепочки проводит запрос x ; запрос $x \in X$ проходит в вершину β ИГ, если существует ориентированная цепочка, ведущая из корня в вершину β , которая проводит запрос x ; запись y , приписанная листу α , попадает в ответ ИГ на запрос $x \in X$, если запрос x проходит в лист α . Ответом ИГ U на запрос x назовем множество записей, попавших в ответ ИГ на запрос x , и обозначим его $\mathcal{J}_U(x)$. Эту функцию $\mathcal{J}_U(x)$ будем считать результатом функционирования ИГ U .

Пусть нам дана ЗИП $I = \langle X, V, \rho \rangle$. Скажем, что ИГ U разрешает ЗИП $I = \langle X, V, \rho \rangle$, если $\mathcal{J}_U(x) = \{y \in V : x\rho y\}$.

Введем понятие *сложности ИГ*. Пусть β – некоторая вершина ИГ. Предикат, определенный на множестве запросов, который при-

нимает значение 1 на запросе x , если запрос проходит в вершину β , и 0 – в противном случае, назовем функцией фильтра вершины β и обозначим $\varphi_\beta(x)$.

Сложностью ИГ U на запросе $x \in X$ назовем число $T(U, x) = \sum_{\beta \in \mathcal{R}} \psi_\beta \cdot \varphi_\beta(x)$, где \mathcal{R} – множество вершин ИГ U , ψ_β – количество ребер, исходящих из вершины β .

Скажем, что базовое множество F измеримое, если каждая функция из F измерима (относительно алгебры σ). Далее всюду будем предполагать, что базовое множество измеримое. В этом случае для любого ИГ U над F функция $T(U, x)$ как функция от x измерима.

Сложностью ИГ U назовем математическое ожидание величины $T(U, x)$, то есть число $T(U) = \mathbf{M}_x T(U, x)$. Сложностью задачи I при базовом множестве F назовем число $T(I, F) = \inf \{T(U) : U \in \mathcal{U}(I, F)\}$, где $\mathcal{U}(I, F)$ – множество всех ИГ над базовым множеством F , разрешающих ЗИП I .

Введем функцию $R(\bar{x}, \bar{y}) = \sum_{i=1}^n |x_i - y_i|$ – расстояние по Хэммингу между точками $\bar{x} = (x_1, \dots, x_n)$ и $\bar{y} = (y_1, \dots, y_n) \in B_2^n$.

Опишем формально тип, соответствующий типу задач о метрической близости на n -мерном булевом кубе. Будем рассматривать следующий тип $S_m = \langle B_2^n, B_2^n, \rho_m, \sigma, \mathbf{P} \rangle$, где ρ_m – отношение поиска на $B_2^n \times B_2^n$, определяемое следующим соотношением $\bar{x} \rho_m \bar{y} \Leftrightarrow R(\bar{x}, \bar{y}) \leq 1$, σ – алгебра подмножеств B_2^n , представляющая собой множество всех подмножеств B_2^n , \mathbf{P} – равномерная вероятностная мера на σ , то есть такая мера, что для любого $x \in B_2^n$ $\mathbf{P}(x) = 1/2^n$ и $\forall A \subseteq B_2^n$ $\mathbf{P}(A) = |A|/2^n$.

Если рассматривать в качестве базового множества множество

$$F_1 = \{f_{\bar{\alpha}, 1}, \bar{\alpha} \in B_2^n\}, \quad (1)$$

где

$$f_{\bar{\alpha}, r}(\bar{x}) = \begin{cases} 1, & \text{если } R(\bar{\alpha}, \bar{x}) \leq r \\ 0, & \text{иначе,} \end{cases} \quad (2)$$

то согласно теореме 5 из [3] при этом базовом множестве существуют задачи типа S_m , для которых перебор является оптимальным алгоритмом. Поэтому мы расширим это множество множеством

$$F_2 = \{f_{\bar{0}, r}, f_{\bar{1}, r}, r = 0, \dots, n\}, \quad (3)$$

где $\bar{0} = (0, \dots, 0) \in B_2^n$, $\bar{1} = (1, \dots, 1) \in B_2^n$, и всюду далее будем рассматривать следующее базовое множество

$$F = F_1 \cup F_2. \tag{4}$$

Весом записи \bar{y} назовем величину $\|\bar{y}\| = R(\bar{y}, \bar{0})$.

Будем говорить, что \bar{y} принадлежит s -му слою куба, если $\|\bar{y}\| = s$.

Через $]a[$ будем обозначать наименьшее целое, не меньшее a .

Справедливы следующие теоремы.

Теорема 1 (верхняя оценка сложности). *Если F – базовое множество, определяемое соотношениями (1)–(4), то для любой ЗИП $I = \langle B_2^n, V, \rho_m \rangle$ типа S_m справедливо*

$$T(I, F) \leq 2]\log_2(n+1)[+ \sum_{\bar{y} \in V} \frac{C_n^{\|\bar{y}\|-1} + C_n^{\|\bar{y}\|} + C_n^{\|\bar{y}\|+1}}{2^n}.$$

Теорема 2 (нижняя оценка сложности). *Пусть ЗИП $I = \langle B_2^n, V, \rho_m \rangle$ типа S_m такая, что $V = \{\bar{y}_1, \dots, \bar{y}_k : \forall i, j \in \{1, \dots, k\} (i \neq j \rightarrow R(\bar{y}_i, \bar{y}_j) > 4) \ \& \ \forall i \in \{1, \dots, k\} (R(\bar{0}, \bar{y}_i) \neq 0 \ \& \ R(\bar{0}, \bar{y}_i) \neq n)\}$. F – базовое множество, определяемое соотношениями (1)–(4). Тогда справедливо*

$$T(I, F) \geq \sum_{i=1}^k \frac{C_n^{\|\bar{y}_i\|-1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\|+1}}{2^n}.$$

Следствие 1. *Если $\frac{kn^2}{2^n \log_2 n} \rightarrow \infty$ при $n \rightarrow \infty$, то для любой ЗИП I из теоремы 2 при $n \rightarrow \infty$ справедливо*

$$T(I, F) \sim \sum_{i=1}^k \frac{C_n^{\|\bar{y}_i\|-1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\|+1}}{2^n}.$$

Пусть $T(k, F) = \max_{|V| \leq k} T(\langle B_2^n, V, \rho_m \rangle, F)$ – функция Шеннона.

Теорема 3. *F – базовое множество, определяемое соотношениями (1)–(4), $\frac{k}{\sqrt{n \log_2 n}} \rightarrow \infty$ и $k \lesssim 2^{n+6} \sqrt{\frac{2}{\pi n^9}}$ при $n \rightarrow \infty$, то*

$$T(k, F) \sim 3k \sqrt{\frac{2}{\pi n}}$$

при $n \rightarrow \infty$.

3. Верхняя оценка сложности

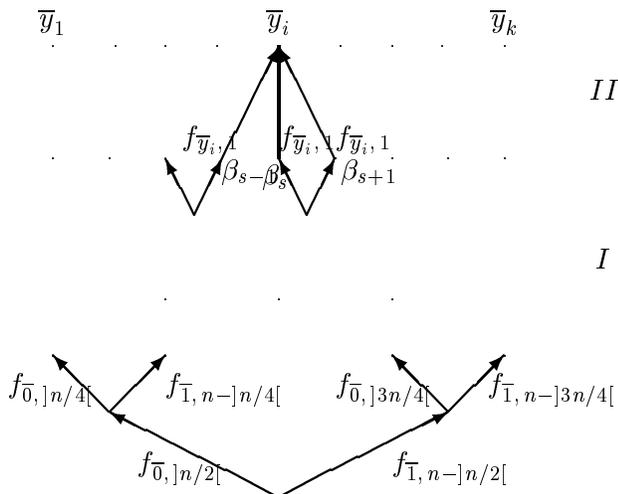


Рис. 1. Граф U

Сложностью ребра (α, β) ИГ назовем число $\mathbf{P}(\{\bar{x} \in B_2^n : \varphi_\alpha(x) = 1\})$. Легко показывается, что сложность ИГ есть сумма сложностей его ребер (см., например, [3]).

Доказательство теоремы 1. Пусть $I = \langle B_2^n, V, \rho_m \rangle$ – произвольная ЗИП типа S_m . Построим граф U , решающий задачу I и изображенный на рисунке 1. Он состоит из двух частей. I-ая часть представляет собой бинарное дерево высоты $\lceil \log_2(n+1) \rceil$ с $n+1$ -й висячей вершиной, такое, что если обозначить i висячую вершину через β_i ($i = 0, 1, \dots, n$), то

$$\varphi_{\beta_i}(\bar{x}) = \begin{cases} 1, & \text{если } \bar{x} \text{ принадлежит } i\text{-му слою куба} \\ 0, & \text{иначе} \end{cases}$$

Иными словами I-ая часть позволяет за $\lceil \log_2(n+1) \rceil$ шагов методом деления пополам находить слой, которому принадлежит запрос.

II-ая часть состоит из $3k$ ребер, где $k = |V|$, причем каждой записи $\bar{y}_i \in V$ ($i = 1, \dots, k$) сопоставлен лист, и в этот лист ведут три ребра из вершин β_{s-1} , β_s , и β_{s+1} , где $s = \|\bar{y}_i\|$, и каждому из этих

трех ребер приписан предикат $f_{\bar{y}_i, 1}$. То есть эти три ребра позволяют из трех нужных слоев выбирать запросы, которым удовлетворяет запись \bar{y}_i .

Сложность I-ой части равна

$$\sum_{i=1}^{\lceil \log_2(n+1) \rceil} 2 = 2 \lceil \log_2(n+1) \rceil,$$

сложность II-ой части:

$$\sum_{i=1}^k \frac{C_n^{\|\bar{y}_i\| - 1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\| + 1}}{2^n}.$$

Следовательно, сложность U

$$T(U) = 2 \lceil \log_2(n+1) \rceil + \sum_{i=1}^k \frac{C_n^{\|\bar{y}_i\| - 1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\| + 1}}{2^n}.$$

Теорема 1 доказана.

4. Нижняя оценка сложности

Для доказательства теоремы 2 потребуются следующие леммы.

Лемма 1. Пусть \bar{y}_1 и \bar{y}_2 такие записи, что $R(\bar{y}_1, \bar{y}_2) > 4$. Пусть U некоторый ИГ, в котором есть листья α_1 и α_2 , которым приписаны соответственно записи \bar{y}_1 и \bar{y}_2 . Пусть C_1 и C_2 пути в ИГ U , которые ведут из корня ИГ U в вершины α_1 и α_2 соответственно, и по которым соответственно проходят запросы $x_1 \in O(\bar{y}_1, \rho_m)$ и $x_2 \in O(\bar{y}_2, \rho_m)$. Тогда если пути C_1 и C_2 имеют реберное пересечение, то ни одному ребру из этого пересечения не приписан предикат типа $f_{\bar{y}, 1}$, где $\bar{y} \in B_2^n \setminus \{\bar{0}, \bar{1}\}$.

Доказательство. По каждому ребру из реберного пересечения путей C_1 и C_2 должны пройти и запрос x_1 , и запрос x_2 . Если некоторому ребру приписан предикат $f_{\bar{y}, 1}$, то это означает, что через это ребро могут пройти только запросы из шара радиуса 1 с центром в

точке \bar{y} . Из того, что $R(\bar{y}_1, \bar{y}_2) > 4$ следует, что $R(\bar{x}_1, \bar{x}_2) > 2$. Следовательно, запросы x_1 и x_2 не могут принадлежать одновременно ни одному шару радиуса 1. Откуда сразу следует утверждение леммы.

Обозначим через Z_m m -й слой булевого куба.

Лемма 2. Пусть C – ориентированная цепочка ребер, ребрам которой приписаны только предикаты из множества F_2 , определяемого соотношением (3). Тогда если C проводит запрос \bar{x} , то она проводит все запросы из множества $Z_{\|\bar{x}\|}$.

Доказательство. Поскольку каждая из функций из множества F_2 является симметрической, то есть на наборах, принадлежащих одинаковым слоям принимает одинаковые значения, то и конъюнкция этих функций будет симметрической. Отсюда, учитывая, что функция проводимости цепочки есть конъюнкция проводимостей ребер цепочки, сразу получаем справедливость утверждения леммы.

Перейдем к доказательству теоремы 2. Пусть $V = \{\bar{y}_1, \dots, \bar{y}_k\}$ – библиотека, удовлетворяющая условиям теоремы. Сопоставим каждой записи $\bar{y}_i \in V$ три запроса $\bar{x}_i' \in Z_{\|\bar{y}_i\|-1} \cap O(\bar{y}_i, \rho_m)$, $\bar{x}_i'' \in Z_{\|\bar{y}_i\|} \cap O(\bar{y}_i, \rho_m)$, $\bar{x}_i''' \in Z_{\|\bar{y}_i\|+1} \cap O(\bar{y}_i, \rho_m)$. Из условия, что $R(\bar{y}_i, \bar{y}_j) > 4$ при $i \neq j$, следует, что запросы, сопоставленные разным записям, разные.

Рассмотрим произвольный ИГ U , разрешающий ЗИП $I = \langle B_2^n, V, \rho_m \rangle$.

Сопоставим каждой записи $\bar{y}_i \in V$ три ориентированные цепочки C_i', C_i'', C_i''' , ведущие из корня ИГ U в листья, которым приписана запись \bar{y}_i , проводящие соответственно запросы $\bar{x}_i', \bar{x}_i'', \bar{x}_i'''$. Какие-то две или даже все из этих трех цепочек могут совпадать.

Сопоставим каждой записи $\bar{y}_i \in V$ три ребра c_i', c_i'', c_i''' , таким образом, что c_i' (c_i'', c_i''') является первым в цепочке C_i' (C_i'', C_i'''), которое не принадлежит никаким другим цепочкам C_j', C_j'', C_j''' , где $j \neq i$. Понятно, что такие ребра обязательно существуют (в крайнем случае это ребра, непосредственно ведущие в листья). Какие-то два или даже все из этих трех ребер могут совпадать. По построению ясно, что ребра, соответствующие разным записям, разные.

Пусть $\beta_i', \beta_i'', \beta_i'''$ – начала ребер c_i', c_i'', c_i''' соответственно. Пусть P_i', P_i'', P_i''' – фрагменты цепочек C_i', C_i'', C_i''' соответственно, начинающиеся в корне ИГ U и заканчивающиеся в вершинах $\beta_i', \beta_i'', \beta_i'''$

соответственно (возможно пустые, если какие-то или все из вершин $\beta_i^I, \beta_i^{II}, \beta_i^{III}$ являются корнем ИГ U).

Если какие-то из вершин $\beta_i^I, \beta_i^{II}, \beta_i^{III}$ ($i = 1, 2, \dots, k$) являются корнем ИГ U , то соответствующие ребра $c_i^I, c_i^{II}, c_i^{III}$ имеют сложность 1.

Для непустых цепочек $P_i^I, P_i^{II}, P_i^{III}$ ($i = 1, 2, \dots, k$) согласно лемме 1 все ребра этих цепочек нагружены функциями из множества F_2 . Отсюда поскольку запрос \bar{x}_i^I ($\bar{x}_i^{II}, \bar{x}_i^{III}$) проходит в вершину β_i^I ($\beta_i^{II}, \beta_i^{III}$), то согласно лемме 2 все запросы из слоя $Z_{\|\bar{y}_i\|-1}$ ($Z_{\|\bar{y}_i\|}, Z_{\|\bar{y}_i\|+1}$) проходят в вершину β_i^I ($\beta_i^{II}, \beta_i^{III}$). Следовательно суммарная сложность ребер $c_i^I, c_i^{II}, c_i^{III}$ не меньше чем $(C_n^{\|\bar{y}_i\|-1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\|+1}) \cdot 2^{-n}$.

Из того, что разным записям сопоставлены разные ребра, и в силу произвольности ИГ U , следует, что

$$T(I, F) \geq \sum_{i=1}^k \frac{C_n^{\|\bar{y}_i\|-1} + C_n^{\|\bar{y}_i\|} + C_n^{\|\bar{y}_i\|+1}}{2^n}.$$

Тем самым теорема 2 доказана.

5. Асимптотика функции Шеннона

Доказательство теоремы 3. Рассмотрим функцию натурального аргумента

$$\Phi(m) = C_n^{m-1} + C_n^m + C_n^{m+1}.$$

Рассмотрим случай, когда n – четное число.

Понятно, что функция $\Phi(m)$ достигает максимума при $m = n/2$.

Поскольку

$$C_n^{\frac{n}{2}} \sim 2^n \sqrt{\frac{2}{\pi n}},$$

то согласно теореме 1 для любой ЗИП $I = \langle B_2^n, V, \rho_m \rangle$ ($|V| = k$) типа S_m справедливо

$$\begin{aligned} T(I, F) &\leq 2 \lceil \log_2(n+1) \rceil \left[+ \sum_{\bar{y} \in V} \frac{C_n^{\|\bar{y}\|-1} + C_n^{\|\bar{y}\|} + C_n^{\|\bar{y}\|+1}}{2^n} \right] \leq \\ &\leq 2 \lceil \log_2(n+1) \rceil \left[+ 2^{-n} 3k \Phi(n/2) \right] \lesssim 3k \sqrt{\frac{2}{\pi n}} \end{aligned}$$

при $\frac{k}{\sqrt{n} \log_2 n} \rightarrow \infty$ и $n \rightarrow \infty$. Следовательно при $\frac{k}{\sqrt{n} \log_2 n} \rightarrow \infty$ и $n \rightarrow \infty$

$$T(k, F) \lesssim 3k \sqrt{\frac{2}{\pi n}}.$$

Обозначим через $W_{\bar{y}}^r$ шар радиуса r с центром в \bar{y} в булевом кубе. Понятно, что если $\bar{y} \in Z_{n/2}$, то

$$l_n = |Z_{n/2} \cap W_{\bar{y}}^4| = 1 + C_{n/2}^1 \cdot C_{n/2}^1 + C_{n/2}^2 \cdot C_{n/2}^2 \sim n^4/64$$

при $n \rightarrow \infty$. Следовательно, в слое $Z_{n/2}$ может разместиться по крайней мере

$$C_n^{n/2}/l_n \sim 2^{n+6} \sqrt{\frac{2}{\pi n^9}}$$

записей, попарное расстояние между которыми больше чем 4. Откуда согласно теореме 2 при $k \lesssim 2^{n+6} \sqrt{\frac{2}{\pi n^9}}$ и $n \rightarrow \infty$

$$T(k, F) \gtrsim 2^{-n} 3k \Phi(n/2) \gtrsim 3k \sqrt{\frac{2}{\pi n}}.$$

Легко видеть, что при нечетном n асимптотическая картина сохраняется. Следовательно при $n \rightarrow \infty$, $\frac{k}{\sqrt{n} \log_2 n} \rightarrow \infty$ и $k \lesssim 2^{n+6} \sqrt{\frac{2}{\pi n^9}}$

$$T(k, F) \sim 3k \sqrt{\frac{2}{\pi n}}.$$

Тем самым теорема 3 доказана.

Список литературы

- [1] Черный А. И. Введение в теорию информационного поиска. М.: Наука, 1975.
- [2] Гасанов Э. Э. Информационно-графовая модель хранения и поиска данных // Интеллектуальные системы. 1998. Т. 3. Вып. 3–4. С. 163–192.
- [3] Гасанов Э. Э. Оптимальное решение базовых задач хранения и поиска в информационно-графовой модели данных. Дисс. на соискание уч. степени доктора физ.-мат. наук. М., 1999.