Ключевые системы в случайных базах данных

О.В. Селезнев, Б. Тальхайм

1. Введение

Случайные базы данных (или *таблицы*) состоят из последовательности случайных векторов (наборов, записей) с общим дискретным распределением $\mathcal{P} = \{p(\mathbf{k})\}$. Теория баз данных исследует главным образом *отношения* (наборы в отношении не совпадают). Таблицы также широко используются в вычислительной математике и приложениях (например, в распознавании образов, в диагностике отказов в системах, для геологических прогнозов). Ключом называется группа столбцов (атрибутов), которая позволяет идентифицировать любой набор в базе данных. Будем называть ключ минимальным, если любые его собственные подмножества ключами *не* являются и кратчайшим, если его длина минимальна. Пусть A и B не пересекающиеся множества атрибутов. B функционально зависит от A, если нет наборов, совпадающих в столбцах A, но отличающихся в столбцах B. Многие теоретические и прикладные задачи, связанные с базами данных (например, оптимизация поиска данных, проектирование баз данных), существенно зависят от *сложности* базы данных, то есть, размера ключевых (минимальных ключевых) систем и систем функциональных зависимостей. До недавнего времени сложность изучалась главным образом в рамках традиционного подхода по наихудшему случаю. (см., например, [15], [12], [9]). Мы рассматриваем *среднюю сложность*, чтобы оценить некоторые статистические свойства сложности, которые в среднем можно ожидать или которые осуществляются с данной вероятностью (см. также соответствующий подход в вычислительной математике [16]).

При подходе по наихудшему случаю сложность определяется наихудшим случаем для данного класса баз данных. Тогда комбинаторные методы дают экспоненциальную оценку сложности по числу атрибутов. Такой подход ограничен достаточно узким классом моделей и возможных приложений, так как событие с наихудшим случаем очень маловероятно на практике. Более того, расхождение между математическими результатами и вычислительными опытом прикладников возрастает. При подходе по среднему случаю для класса баз данных вводится вероятностное распределение $\mathcal{P}^m$, и тогда затраты на поиск данных и сложность измеряются их математическими ожиданиями относительно $\mathcal{P}^m$, где m обозначает число наборов в базе данных. Такие вероятностные модели позволяют в частности применять методы статистического моделирования для исследования сложности.

Некоторые модели случайных баз данных рассматривались в [11], [8]. Различные задачи вычислительной математики (например, тесты и тупиковые тесты, [1]; длина бинарных деревьев для задач поиска, [6], [14]) аналогичны соответствующим задачам для ключей, минимальных и кратчайших ключей в случайных базах данных.

В работах [4], [13], [5] рассматривались равномерные и близкие к ним модели, когда когда распределение каждого набора определяется дискретным равномерным распределением. В настоящей работе расширяется и существенно обобщаются эти результаты для широкого класса стохастических моделей. Помимо равномерности, зависимости между атрибутами и конечности алфавитов атрибутов исследуются также другие обобщения и численные примеры. Для построения стохастических моделей ключевых систем рассматривается новый подход, который основывается на

(i) энтропии Реньи $h_{\mathcal{P}} = -\log_2(\sum_{\mathbf{k}} p(\mathbf{k})^2)$ (*теория информации*);

(ii) методе пуассоновской (Стейн-Чен) аппроксимации (*теория вероятностей*).

Получены асимптотики для параметров (например, энтропия, длина) наиболее вероятных минимальных ключей. Исследуется *где* сосредоточены распределения наиболее вероятных минимальных ключей и длины кратчайших ключей, и для *каких* значений параметров модели. Показано, что экспоненциальный размер ключевой системы, который предсказывается для наихудшего случая, маловероятен по крайней мере, ко-

гда n атрибутов независимы, имеют общее распределение и энтропию d, и $2/d\ \log_2 m = \mathrm{o}(n/\ln n)$ при $n, m \to \infty$.

Длина кратчайшего ключа в данном множестве атрибутов имеет различные интерпретации в вычислительной математике (например, высота бинарных деревьев для некоторых моделей). Показывается, что распределение этой величины может быть аппроксимировано с помощью стандартного распределения Гумбеля (двойного экспоненциального) с заданной точностью, что позволяет применять для исследования метод Монте-Карло. Основные асимптотические результаты для ключей и минимальных ключей оказываются близкими для таблиц и отношений.

1.1. Основные обозначения

Пусть все алфавиты D_i в случайной базе данных R состоят из целых неотрицательных чисел. В нашей стохастической модели предположим что все наборы – независимые и одинаково распределенные случайные векторы с заданным дискретным распределением $P\{t_j = \mathbf{k}\}$, $\mathbf{k} = \mathbf{k}(U) \in \prod_{i \in U} D_i$. Без потери общности можно предположить, что $\mathcal{P}$ является невырожденным, то есть, $0 < p(\mathbf{k}(A)) < 1$ для каждого множества атрибутов A.

Для демонстрации общих результатов, будут использоваться следующие важные примеры моделей случайных баз данных. Случайная база данных будет называться *равномерной* случайной базой данных, если все наборы имеют дискретное равномерное распределение $\mathcal{P}_u$. Обозначим через $H(A) := \sum_{i \in A} \log_2 |D_i|$ функцию *информации* для множества A и данного набора алфавитов $\{D_i, i \in A\}$. Тогда в равномерном случае $p(\mathbf{k}) = 2^{-H(U)}$. Если все атрибуты в наборе независимы и одинаково распределены с общим одномерным дискретным распределением $Q = \mathcal{P}(\{j\})$, R будет называться базой данных *Бернулли*, и *стандартной базой данных Бернулли*, если Q – дискретное равномерное. Заметим, что обычная бернуллиевская модель соответствует бинарной модели с $D_i = 2$ для всех алфавитов. Обозначим через $\mathcal{P}_{\mathcal{B}}$ распределение базы данных Бернулли.

Далее, $h_{\mathcal{P}}(A) := -\log_2(\sum_{\mathbf{k}(A)} p(\mathbf{k}(A)))^2$ будет называться **энтропией Реньи** распределения $\mathcal{P}(A)$, [10]. Величина $h_{\mathcal{P}}(A)$ может рассматриваться как мера количества информации, соответствующая распределению $\mathcal{P}(A)$ (или мера неопределенности для $t_j(A)$). Очевидно, что для

равномерной и бернуллиевской моделей $h_{\mathcal{P}_u}(A) = H(A)$ and $h_{\mathcal{P}_B}(A) = |A|h_Q$ соответственно. Сравнивая с энтропией Шеннона $H_{\mathcal{P}}(U) := -\sum_{\mathbf{k}} p(\mathbf{k}) \log_2 p(\mathbf{k})$, широко используемой в статистической механике, теории кодирования и передачи информации, получаем из неравенства Йенсена, что $0 < h_{\mathcal{P}}(A) \leq H_{\mathcal{P}}(A)$. Для измерения расстояния между двумя вероятностными распределениями μ_1 и μ_2 целочисленных случайных величин будет использоваться *полное расстояние по вариации* $d_{TV}(\mu_1, \mu_2)$,

$$d_{TV}(\mu_1, \mu_2) := \sup\{|\mu_1(C) - \mu_2(C)| : C \subset \{1, 2, \ldots\}\} = 1/2 \sum_{j \geq 0} |\mu_1(j) - \mu_2(j)|.$$

В основных полученных результатах мы рассматриваем последовательности множеств атрибутов $\{A_l\}_{l \geq 1}$ и случайных баз данных $\{R_l\}_{l \geq 1}$. Для оценки асимптотического поведения некоторых характеристик $\{A_l\}_{l \geq 1}$ будет предполагаться, что энтропия Реньи $h_{\mathcal{P}_l}(A_l)$ и/или число атрибутов m_l стремятся к бесконечности. Для удобства обозначений положим $a_l = h_{\mathcal{P}_l}(A_l)$ и будем опускать индекс l для параметров, если это понятно из контекста.

Обозначение	Значение
$U = 1, \ldots, n$	множество атрибутов
$\|S\|$	мощность конечного множества S
$R = R(m, n)$	случайная таблица с m наборами и n атрибутами
$\mathcal{R} = \mathcal{R}(m, n)$	случайное отношение, соответствующее R
$M := m(m-1)/2$	число различных пар наборов
$t_j(A)$, $A \subseteq U$ ($t_j = t_j(U)$)	часть j-ого набора
$\mathcal{P} = \mathcal{P}(U) = \{p(\mathbf{k})\}$	распределения набора
$a = h_{\mathcal{P}}(A)$	энтропия Реньи распределения $\mathcal{P}(A)$
$\lambda = \lambda(A)$	среднее число нарушений ключевого условия для A
A, B	множества атрибутов, $B \subseteq U \setminus A$
$R \models A$ ($\mathcal{R} \models A$)	A – ключ в R ($\mathcal{R}$)
$R \models_{\min} A$ ($\mathcal{R} \models_{\min} A$)	A – минимальный ключ в R ($\mathcal{R}$)
$A \to B$	функциональная зависимость между A и B
$P\{R \models A\}$ ($P\{R \models_{\min} A\}$)	ключевая (для минимального ключа) вероятность

Таблица 1. Используемые обозначения.

2. Ключи в случайных таблицах и отношениях

2.1. Случайные таблицы и отношения

В этом разделе мы получим приближение для ключевой вероятности для случайных таблиц и отношений и покажем, что ключи - наиболее вероятны среди множеств с наибольшей энтропией.

Для каждого натурального k, обозначим через $q_k = q_k(A) := \sum_{\mathbf{k}(A)} p(\mathbf{k}(A))^k$, $0 < q_k < 1$, то есть вероятность, что k различных наборов совпадают для $k \geq 2$. Из неравенства Йенсена получаем $q_2^2 \leq q_3$ и $f(x) = (\sum_{\mathbf{k}(A)} p(\mathbf{k}(A))^x)^{1/x}$ – возрастающая функция для $x > 0$. Следовательно, существует $\delta = \delta_{\mathcal{P}} > 0$, $0 < \delta \leq 1$, характеризующее *"недостаток равномерности"* в распределении $\mathcal{P}$. Причем

$$q_2^{2/3} \leq q_3^{1/3} = q_2^{1/2+\delta/6} < q_2^{1/2}, \tag{1}$$

где $\delta = 1$ тогда и только тогда, если $\mathcal{P}$ равномерное. Обозначим среднее число нарушений ключевого условия через $\lambda = \lambda(A) = M2^{-h_{\mathcal{P}}(A)}$ и $u = h_{\mathcal{P}}(U)$.

Теорема 1. *Пусть R - случайная таблица и $\mathcal{R}$ - соответствующее случайное отношение. Пусть $0 < c < \delta/2 \ln 2$ и $\gamma = \delta/2 - c/\ln 2 > 0$.*

(i) Если $0 < \lambda \leq ca$, то $P\{R \models A\} = e^{-\lambda}(1 + O(2^{-\gamma a}))$ при $a \to \infty$; если $ca < \lambda$, то $P\{R \models A\} = O(2^{-\bar{\gamma} a}))$ при $a \to \infty$ для некоторого положительного $\bar{\gamma}$;

(ii) $|P\{R \models A\} - e^{-\lambda}| \leq 8 \cdot 2^{-\delta a/2} \lambda^{1/2}$;

(iii) если $0 < \lambda \leq ca$ и $0 < \lambda(U) \leq cu$, то $P\{\mathcal{R} \models A\} = \exp\{-\lambda(1 - 2^{a-u})\}(1 + O(2^{-\gamma a}))$ при $a \to \infty$.

Замечание 1. (i) Заметим, что оценка в теореме 1(ii) выполняется *для всех* a и m. В равномерном случае, можно сравнить скорость сходимости, полученную из этой оценки, и соответствующую скорость из явной формулы для равномерного распределения в работе [13]б [5]. Если $0 < \lambda_1 < \lambda < \lambda_2 < \infty$, то $C_1/m \geq |P\{N(A) = 0\} - e^{-\lambda}| \geq C_2 e^{-\lambda} \lambda^2/m \geq C_3/m$ некоторых положительных C_1, C_2 и C_3. Итак, скорость сходимости в теореме 1(ii) – точная для таких λ в общем случае.

(ii) Заметим, что модели для таблиц и отношений эквивалентны $P\{R \models A\} \sim P\{\mathcal{R} \models A\}$, в тех случаях, когда энтропии $u - a \to \infty$. В то же время, для многих множеств атрибутов с ограниченным значением разности $u - a$, разница в асимптотике может быть существенной.

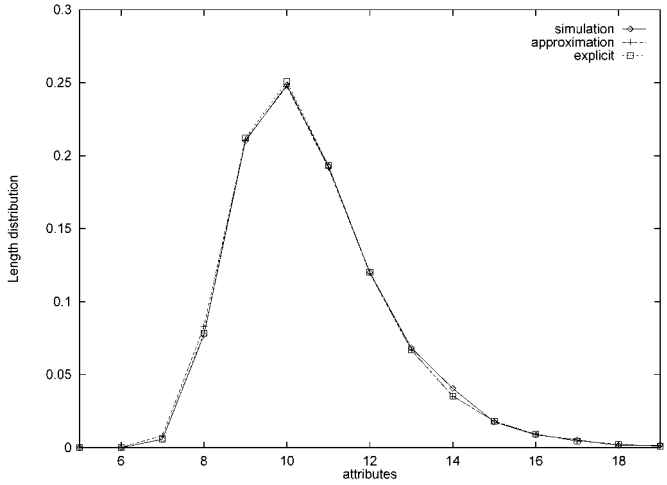

Рис. 1. Точность аппроксимации для ключевой вероятности (полигон частот). Стандартная база данных Бернулли $|D_i| = 2$, $m = 50$. Явная формула (*explicit*), Аппроксимация (*approximation*), Монте-Карло аппроксимация (*simulation*), $N_{sim} = 5000$.

Пример 1. Чтобы продемонстрировать точность аппроксимации в теореме 1, мы рассмотрим случай стандартной базы данных Бернулли ($|D_i| = 2$, $m = 50$). Для ключевой вероятности мы используем явную формулу, приближение теоремы 1(i) и сравним эти результаты с результатами, полученными моделированием по методу Монте-Карло для соответствующей стохастической модели ($m = 50, N_{sim} = 5000$). На рис. 1 изображен полигон частот $p(r) := P\{R \models A, |A| = r\} - P\{R \models A, |A| = r - 1\}$ для трех методов. Все эти три метода дают очень близкие результаты для ключевой вероятности. Рис. 2(a) иллюстрирует влияние возрастания размера алфавита. При этом существенно сокращается дисперсия, и сдвигаются границы наиболее вероятных значений к началу координат. График 2(b) демонстрирует сдвиговый эффект для

ключевой вероятности, когда число наборов в базе данных возрастает.

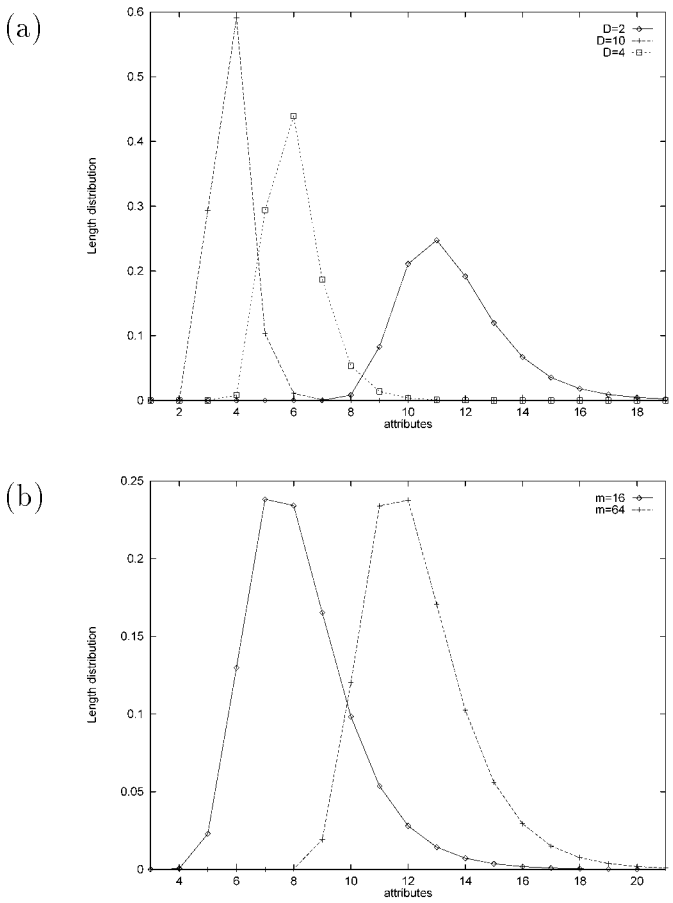

Рис. 2. Изменение ключевой вероятности (полигон частот). Стандартная база данных Бернулли (a) $|D_i| = 2$, $|D_i| = 4$ и $|D_i| = 10$, $m = 50$ (b) $|D_i| = 2$, $m = 16$ и $m = 64$, $i = 1, \ldots, 20$.

2.2. Функциональные зависимости в случайных базах данных

Исследуем вероятности функциональной зависимости $P\{A \to B\}$, применяя тот же самый подход. Пусть A и $B \subseteq U \setminus A$ – множества

атрибутов в R. Обозначим $q_{kf} = q_{kf}(A,B) := \sum_{\mathbf{k}(A\cup B)}(p(\mathbf{k}(A)) - p(\mathbf{k}(A\cup B)))^{k-1}p(\mathbf{k}(A\cup B)$, и следовательно $q_{2f}(A,B) = q_2(A) - q_2(A\cup B)$. Тогда из неравенства Йенсена следует, что $q_{2f}^2 \le q_{3f}$, используя аналогичные рассуждения, как и для ключа, можно получить

$$q_{3f}^{1/3} = q_{2f}^{1/2+\delta_f/6} < q_{2f}^{1/2}, \quad 0 < \delta_f \le 1. \tag{2}$$

Обозначим среднее число нарушений условия функциональной зависимости через $\lambda_f = M2^{-a_f}$, где $a_f = -\log_2 q_{2f}$. Например, для равномерных баз данных и моделей Бернулли, $\lambda_f = M2^{-a}(1-2^{-b})$, где $a = h_{\mathcal{P}}(A)$, $b = h_{\mathcal{P}}(B)$. Следующая теорема показывает, что наиболее вероятны функциональные зависимости в тех случаях, когда значения λ_f достаточно малы: в силу ключевых свойств множества атрибутов A и/или в силу стохастической зависимости между $t(A)$ и $t(B)$.

Теорема 2. *Пусть R – случайная таблица. Пусть $0 < c < \delta_f/2\ln 2$ и $\gamma = \delta_f/2\ln 2 - c > 0$.*

(i) Если $0 < \lambda_f \le ca_f$, то $P\{A\to B\} = e^{-\lambda_f}(1+O(2^{-\gamma a_f}))$ as $a_f\to\infty$; если $ca_f < \lambda_f$, то $P\{A\to B\} = O(2^{-\bar\gamma a_f})$ при $a_f\to\infty$ для некоторого положительного $\bar\gamma$;

(ii) $|P\{A\to B\} - e^{-\lambda_f}| \le 8\cdot 2^{-\delta_f a_f/2}\lambda_f^{1/2}$.

3. Минимальные ключи

3.1. Вероятность минимального ключа для таблиц и отношений

Приведем сначала основные асимптотические результаты для вероятности минимального ключа.

Пусть A – данное множество атрибутов и $r = |A|$. Обозначим через $\rho_k := q_{2k}/q_2 - 1 > 0$, $k = 1,2,\dots,r$, $\rho_{\max} := \max_k \rho_k$, $\rho_{\min} := \min_k \rho_k$ и $\rho_\cdot := \sum_{k=1}^r \rho_k$, $\kappa := \log_2(\sum_{k,l=1}^r q_{3kl}/q_3)$, где $q_{2k} := P\{t_1(A_k) = t_2(A_k)\}$ and $q_{3kl} := P\{t_1(A_k) = t_2(A_k), t_1(A_l) = t_3(A_l)\}$, $A_k = A\setminus\{k\}$, $k,l = 1,2,\dots,r$,

Теорема 3. *Пусть R – случайная таблица и $\mathcal{R}$ – соответствующее случайное отношение Пусть $0 < c < \delta a/2$, $\kappa = o(a)$ и $r = O(a)$ при*

$a \to \infty$. Если $0 < c_0/\rho_{\min} \le \lambda < ca$ и $0 < \lambda(U) < cu$, то существует $\gamma > 0$ такое, что достаточно большого $c_0 > 0$

$$P\{R \models_{\min} A\} = e^{-\lambda} \prod_{k=1}^{r} (1 - e^{-\rho_k \lambda})(1 + O(2^{-\gamma a}));$$

$$P\{\mathcal{R} \models_{\min} A\} = e^{-\lambda(1-2^{a-u})} \prod_{k=1}^{r} (1 - e^{-\rho_k \lambda})(1 + O(2^{-\gamma a})) \text{ при } a \to \infty;$$

если $0 < \lambda < c_0/\rho_{\min}$ или $ca < \lambda$, то для некоторого $\bar{\gamma} > 0$, $P\{R \models_{\min} A\} = O(2^{-\bar{\gamma} a})$ при $a \to \infty$.

Этот результат позволяет описать поведение максимальной величины вероятности минимального ключа $P_{\max}(a)$ и установить соответствующие *оптимальные* значения параметров модели. Пусть $p_{\min} := (\rho_\cdot + 1)^{-1/\rho_{\min}}$, $p_{\max} := (\rho_\cdot + 1)^{-1/\rho_{\max}} < 1$ и $P(a) := P\{R \models_{\min} A\}$. Если $\max(1, \ln(1 + \rho_\cdot))/\rho_{\min} = \mathrm{o}(a)$, то

$$p_{\min} \exp\{-1/\rho_{\min}\}(1 + \mathrm{o}(1)) \le P_{\max}(a) \le p_{\max} \exp\{-\rho_\cdot/(\rho_{\max}(1 + \rho_\cdot))\}(1 + \mathrm{o}(1)) \tag{3}$$

при $a \to \infty$. Далее, $P(a) = P_{\max}(a)$ тогда и только тогда, если $a_{\max} \sim 2 \log_2 m - \log_2(1/\bar{\rho} \ln(\rho_\cdot + 1)) - 1 + \mathrm{o}(1/m)$ при $a, m \to \infty$. Таким образом, существует определенное соотношение между величиной энтропией множеств – наиболее вероятных кандидатов в ключи, и энтропией самой системы наборов как множества M пар элементов ($\log_2 M \sim 2 \log_2 m - 1$ при $m \to \infty$). Можно предложить, что это свойство справедливо и для более общего класса моделей.

Пример 2.

(i) Пусть R – бернуллиевская база данных с энтропией атрибута $d := h_Q$. Пусть $\rho_k = 2^d - 1$, $\rho_\cdot = r(2^d - 1) \ge rd = a$, и $\kappa \le 2 \log_2 r + 2d = \mathrm{o}(rd)$. Следовательно, предположения теоремы 3 выполнены если, например, $\ln r/r = \mathrm{o}(d^2)$ при $a \to \infty$. Получаем $P\{R \models_{\min} A\} = e^{-\lambda}(1 - e^{-\rho\lambda})^r(1 + \mathrm{O}(2^{-\gamma a}))$ при $a \to \infty$. Пусть $0 < d_1 < d < d_2 < \infty$; тогда $P_{\max}(a) \sim e^{-1/\rho}(1 + r\rho)^{-1/\rho}$,

$$r_{\max} \sim 1/d\,(2 \log_2 m - \log_2 \ln \log_2 m - 1 + \log_2(2^d - 1) + \mathrm{O}(1/\ln \ln m)) \text{ при } a, m \to \infty.$$

Отсюда следует, что экспоненциальный размер для постановки в среднем маловероятен, если $2/d \log_2 m = \mathrm{o}(n/\ln)$ при $n, m \to \infty$.

(ii) Пусть R – равномерная база данных и положим $D_{\max} := \max_k |D_k|$ and $D_{\min} := \min_k |D_k|$. Тогда $\rho_k = |D_k| - 1$, $\rho_\cdot = \sum_{k=1}^{r}(|D_k| - 1) \geq r$, и $\kappa = \log_2(\sum_{k,l} |D_k||D_l|) = 2\log_2(\sum_k |D_k|)$. Следовательно, предположения теоремы 3 выполнены если, например,

$$\log_2 D_{\max} / \log_2 D_{\min} = \mathrm{o}(r)$$

и $D_{\min}/D_{\max} \ln r \to \infty$ при $a \to \infty$. Тогда получаем $P\{R \models_{\min} A\} = e^{-\lambda}\prod_{k=1}^{r}(1 - e^{-(|D_k|-1)\lambda})(1 + \mathrm{O}(2^{-\gamma a}))$ при $a \to \infty$, и также для $\log_2 D_{\min} - 1 \leq C \leq \log_2 D_{\max} - 1$, $a = \log_2(\prod_{k=1}^{r} |D_k|)$,

$$a_{\max} \sim (2\log_2 m - \log_2 \ln(\sum_{k=1}^{r} |D_k| - r + 1) + C + \mathrm{O}(1/\ln\ln m)) \text{ при } a, m \to \infty.$$

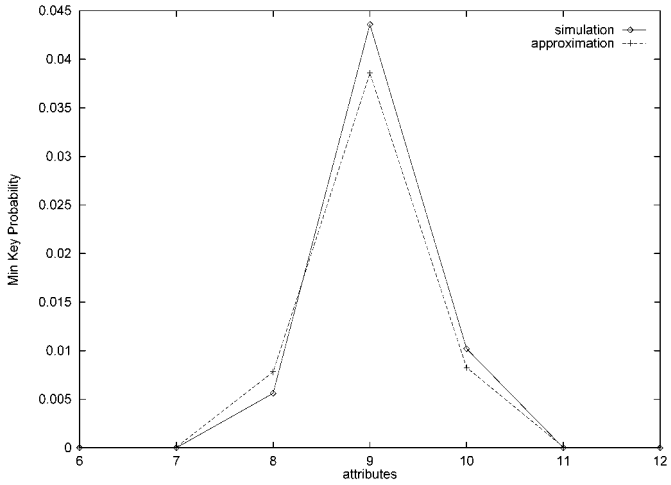

Рис. 3. Точность аппроксимации для вероятности минимального ключа. Стандартная база данных Бернулли $|D_i| = 2$, $m = 50$. Аппроксимация (*approximation*), Монте-Карло аппроксимация (*simulation*), $N_{sim} = 5000$.

Пример 3. В этом примере демонстрируется точность приближения в теореме 3 для вероятности минимального ключа в случае стандартной базы данных Бернулли ($|D_i| = 2$, $m = 50$). Используется приближение теоремы 3(i), которое сравнивается с результатами моделирования по методу Монте-Карло для соответствующей модели ($m = 50, N_{sim} =$

5000). Рис. 3, $p_{\min}(r) := P\{R \models_{\min} A, |A| = r\}$, показывает, что оба метода дают достаточно близкие результаты. Рис. 4(a) иллюстрирует влияние роста размера алфавита. При этом границы наиболее вероятных значений сдвигаются к началу координат и возрастает величина $P_{\max}(r)$. На графике 4(b) показан сдвиг вероятности минимального ключа, когда число наборов возрастает. В этом случае $P_{\max}(r)$ убывает.

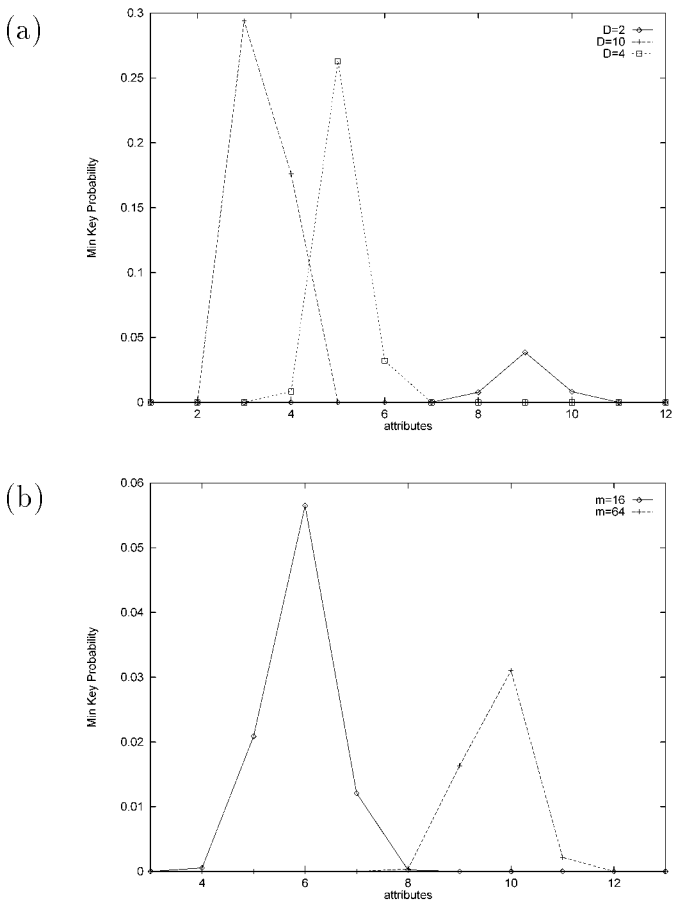

Рис. 4. Изменение вероятности минимального ключа. Стандартрая база данных Бернулли (a) $|D_i| = 2$, $|D_i| = 4$ и $|D_i| = 10$, $m = 50$ (b) $|D_i| = 2$, $m = 16$ и $m = 64$, $i = 1, \ldots, 20$.

3.2. Кратчайшие ключи

Пусть $\nu_m = \nu_m(A)$ – длина кратчайших ключей в A, $|A| = r$, то есть целочисленная случайная величина, которая равна длине минимального ключевого подмножества B в A. Заметим, что A является ключем тогда и только тогда, если $\nu_m(A) \leq r$, и следовательно, для исследования распределения ν_m теорема 1 может быть применена теорема 1. Мы предполагаем, что $a(r)$ является возрастающей функцией и $a(r) \to \infty$, если $r \to \infty$ для натуральных r. Проинтерполируем $a(v)$ с помощью ломаной на каждом интервале $v \in [r, r-1]$, $r \geq 1$. Пусть X – случайная величина с распределением Гумбеля (двойным экспоненциальным) $P\{X \leq x\} = \exp\{-e^{-x}\}$, и $a(X_m) = X/\ln 2 + 2\log_2 m - 1$. Из теоремы 1 следует, что равномерно по $r \geq 1$ и $m \geq 1$

$$P\{\nu_m \leq r\} - P\{X_m \leq r\} = \mathrm{O}(m2^{-(1+\delta)a/2}). \tag{4}$$

Замечание 1. Двойное экспоненциальное распределение часто встречается как предельное для распределений экстремумов независимых и слабо зависимых случайных величин (см., например, [7]). Такая экстремальная структура распределения в нашем случае достаточно естественна. Обозначим через $H_m = \max\{C_{ij} : 1 \leq k < l \leq m\} + 1$, где $C_{ij} = k$, если $t_i(A)$ и $t_j(A)$ совпадают при k атрибутах, но отличаются при любом $(k+1)$-наборе атрибутов (то есть, *наибольшая* длина цепочки с совпадающими атрибутами в наборах $t_i(A)$ и $t_j(A)$) (ср. *линейный блок для цифровых деревьев* в задачах поиска, [14]). Высоты цифровых деревьев были введены [6] для построения модели сохранения информации в компьютере. По определению $\nu_m \leq r$ тогда и только тогда, если $H_m \leq r$. Такая интерпретация позволяет применить по крайней мере некоторые полученные результаты для кратчайших ключей и к исследованию вероятностных свойств высот цифровых деревьев.

В качестве примера общего подхода рассмотрим более подробно модель Бернулли. Тогда $a(r) = rd$. Обозначим через $\hat{X}_m = [X_m] = [1/d(X/\ln 2 + 2\log_2 m - 1)]$, где $[b]$ – целая часть b. Тогда $\hat{X}_m$ является целочисленной случайной величиной с распределением $\mathcal{L}(\hat{X}_m)$ и функцией распределения $F(r)$. Для этой модели оценим $d_{TV}(\mathcal{L}(\nu_m), \mathcal{L}(\hat{X}_m))$ и моменты ν_m.

Предложение 1. *Пусть R - база данных Бернулли и δ определена в (1), $0 < \delta \le 1$.*

(i) $d_{TV}(\mathcal{L}(\nu_m), \mathcal{L}(\hat{X}_m)) = O((\ln m)^{(1+\delta)/2} m^{-\delta})$ *при* $m \to \infty$;

(ii) Для каждого $k \ge 1$, $\mathrm{E}\nu_m^k = \mathrm{E}\hat{X}_m^k + O((\ln m)^{(2k+1+\delta)/2} m^{-\delta})$ *при* $m \to \infty$.

В частности, для среднего $\mathrm{E}\nu_m$ и дисперсии $\mathrm{Var}(\nu_m)$,

$$0 \le \mathrm{E}\hat{X}_m - (2\log_2 m - m_g/\ln 2 - 1)/d \le 1, \quad |\mathrm{Var}(\hat{X}_m)^{1/2} - \sigma_g/(d\ln 2)| \le 1,$$

где $m_g = \mathrm{E}X = 0.5772\ldots$, $\sigma_g^2 = \mathrm{Var}(X) = \pi^2/6$. Заметим, что значения длин кратчайших (минимальных) ключей концентрируются около $r_1 = 2/d\,\log_2 m$. Сравнивая этот результат с длиной наиболее вероятного минимального ключа

$$r_2 = 1/d\,(2\log_2 m - \log_2(\ln(2/d\,\log_2 m)))(1 + \mathrm{o}(1)) < r_1,$$

можно сделать вывод, что множества минимальных и кратчайших минимальных ключей в наиболее вероятной части практически совпадают.

4. Доказательства

Приведем один из результатов метода пуассоновской аппроксимации (метод Стейн-Чена), который будет использоваться в дальнейшем. Пусть $\{C_\alpha\}_{\alpha\in\Gamma}$ – система событий, I_α – индикатор события C_α и $N := \sum_{\alpha\in\Gamma} I_\alpha$ с распределением $\mathcal{L}(N)$. Метод Стейн-Чена дает возможность построить приближенное пуассоновское распределение для $\mathcal{L}(N)$. Обозначим через $\Gamma_\alpha = \Gamma \setminus \{\alpha\}$. Предположим, что Γ_α разделено на два множества $\bar{\Gamma}_\alpha$и $\hat{\Gamma}_\alpha$ так, что I_α и $\{I_\beta\}_{\beta\in\hat{\Gamma}_\alpha}$ независимы и $\bar{\Gamma}_\alpha := \Gamma \setminus \hat{\Gamma}_\alpha$, $\bar{m} = |\bar{\Gamma}_\alpha|$. Обозначим через λ среднее значение N. Следущее предложение следует непосредственно из следствия 2.C.5., [2].

Предложение 2. *Пусть для каждого* $\alpha \in \Gamma$ *и* $\beta \in \bar{\Gamma}_\alpha$, $P\{I_\alpha = 1\} = q > 0$ *и* $EI_\alpha I_\beta \le s$; *тогда*

$$|P\{N = 0\} - e^{-\lambda}| \le d_{TV}(\mathcal{L}(N), \mathcal{P}(\lambda)) \le q\bar{m}\,(1 + 1/\bar{m} + s/q^2).$$

Доказательство теоремы 1. Используя введенные обозначения для базы данных, получаем $\Gamma = \{\alpha = (i,j) : 1 \le i < j \le m\}$ и $|\Gamma| = M = m(m-1)/2$. Для каждого α обозначим через $C_\alpha = C_{ij}$ событие $\{t_i(A) = t_j(A)$ и запишем $N = N(A) := \sum_{\alpha\in\Gamma} I_\alpha$. Далее, для *$\alpha$-зависимой* части $\bar{\Gamma}_\alpha = \{(i,k),(l,j) \in \Gamma : k \ne j,\ l \ne i\}$ и $|\bar{\Gamma}_\alpha| = 2(m-1)$. Индикаторные случайные величины I_α, $\alpha \in \Gamma$, одинаково распределены. Для каждого $\alpha \in \Gamma$ и $\beta \in \bar{\Gamma}_\alpha$, $P\{I_\alpha = 1\} =: q_2$ и $\mathrm{E}I_\alpha I_\beta = P\{t_1(A) = t_2(A) = t_3(A)\} =: q_3$. Тогда из предложения 2 и (1) следует, что

$$|P\{N=0\} - e^{-\lambda}| \le q_2\Big(2(m-1) + 1 + 2(m-1)q_3/q_2^2\Big) \le 4mq_3/q_2 = 4m2^{-(\delta+1)a/2} \tag{5}$$

так как $q_3 \ge q_2^2$. По определению $m \le 2\lambda^{1/2}2^{a/2}$, что и дает (ii).

Далее, для $\gamma > 0$, из предположений мы получаем $e^\lambda\lambda^{1/2}2^{-\delta a/2} =\le e^{-\gamma a}$. Для завершения доказательства для таблиц, заметим, что из (ii) вытекает

$$P\{R \models A\} \le \max(e^{-\lambda}, 8\exp\{1/2\ \ln\lambda - \delta a/2\ \ln 2\})$$

и следовательно, для $ca < \lambda < 2^{c_1 a}$, где $c_1 < \delta$, существует $\gamma_1 > 0$ такое, что $P\{R \models A\} = \mathrm{O}(2^{-\gamma_1 a})$ при $a \to \infty$. Пусть $2^{c_1 a} \le \lambda$. Утверждение (i)следует из

$$P\{N=0\} = 1 - P\{N \ge 1\} \le (\mathrm{E}N)^2/\mathrm{E}N^2 \le (1+\mathrm{O}(1/m))/(\lambda - 1/M) = \mathrm{O}(2^{-c_1 a}).$$

По определению единственное отличие между R и $\mathcal{R}$ состоит в том, что таблица *может* содержать совпадающие наборы. Очевидно, $N(U) \le N(A)$ и следовательно,

$$\begin{aligned} P\{\mathcal{R} \models A\} &= P\{N(A) = 0 / N(U) = 0\} = \\ &= P\{N(A) = 0\}/P\{N(U) = 0\} \ge P\{R \models A\}. \end{aligned} \tag{6}$$

Таким образом, примение теоремы 1(i) завершает доказательство. □

Доказательство теоремы 2. Доказательство аналогично доказательству теоремы 1. □

Доказательство теоремы 3. Пусть $(\Omega, \mathcal{F}, P)$ – основное вероятностное пространство. Для каждого $B \in \mathcal{F}$, обозначим через $\bar{B} := \Omega \setminus B$. Обозначим также через $\mathcal{A}_k = \{R \models A_k\}$, $k = 1, \ldots, r$, и $\mathcal{A} = \{R \models A\}$.

Из определения минимального ключа и формулы включений-исключений получаем

$$P\{R \models_{\min} A\} = P(\mathcal{A}) - \sum_{j=1}^{r} (-1)^{j-1} S_j, \quad e^{-\lambda} \prod_{k=1}^{r} (1 - e^{-\rho_k \lambda}) = e^{-\lambda} - \sum_{j=1}^{r} (-1)^{j-1} \hat{S}_j, \tag{7}$$

где $S_j := \sum_{k_1 < \ldots < k_j} P(\mathcal{A}_{k_1} \cdots \mathcal{A}_{k_j})$ и $\hat{S}_j := \sum_{k_1 < \ldots < k_j} e^{-(\sum_{i=1}^{j} \rho_{k_i} + 1)\lambda}$.

Применим сначала метод пуассоновской аппроксимации для оценивания вероятности $P(\mathcal{A}_{k_1} \cdots \mathcal{A}_{k_p})$, $p = 1, \ldots, r$; тогда утверждение теоремы будет следовать из (7) и неравенства Бонферрони. Без потери общности, пусть $k_1 = 1, \ldots, k_p = p$.

Лемма 1. *Если* $p \le r$, *то* $|P(\mathcal{A}_1 \ldots \mathcal{A}_p) - e^{-(\sum_{i=1}^{p} \rho_i + 1)\lambda}| \le 8 \cdot 2^{-\delta a/2 + \kappa} \lambda^{1/2}$.

Доказательство леммы 1. Пусть $E_p := \mathcal{A}_1 \cdots \mathcal{A}_p$. Для каждого $l = 1, 2, \ldots, r$, обозначим $D_{ij}^k := \{t_i(A_l) = t_j(A_l)\}$ и пусть $D_{ij} := \bigcup_{l=1}^{p} D_{ij}^l$. Тогда $E_p = \bigcap_{i<j} \bar{D}_{ij}$. По определению, $P(D_{12}^l) = q_{2l}$ и для каждого $l_1 \ne l_2$ и $k \ge 2$, $D_{12}^{l_1} D_{12}^{l_2} = \{t_1(A) = t_2(A)\}$, и следовательно $P(D_{12}^{l_1} \cdots D_{12}^{l_k}) = P(D_{12}^{l_1} D_{12}^{l_2}) = q_2$. Таким образом, из формулы включений-исключений вытекает

$$P(D_{12}) = \sum_{l=1}^{p} q_{2l} - \binom{p}{2} q_2 + \binom{p}{3} q_2 + \cdots + (-1)^{p-1} q_2 = (\sum_{l=1}^{p} \rho_l + 1) q_2. \tag{8}$$

Далее, по определению

$$P(D_{12} D_{13}) = P(\bigcup_{k,l=1}^{p} D_{12}^k D_{13}^l) \le \sum_{k,l=1}^{p} P(D_{12}^k D_{13}^l) \le \sum_{k,l=1}^{r} q_{3kl} = q_3 2^{\kappa}. \tag{9}$$

Применяя (8), (9) и предложение 2,

$$|P(E_p) - e^{-\lambda_p}| \le 4 \cdot 2^{\kappa} m q_3 / q_2,$$

где $\lambda_p := (\sum_{l=1}^{p} \rho_l + 1)\lambda$. Таким образом, утверждение следует, как и в теореме 1. □

Теперь, применим неравенство Бонферрони к включениям-исключениям (7). Зададим целое $T = vr$, где $0 < v < 1/2$, и пусть

$$R_j := \max_{k_1 < \ldots < k_p} |P(\mathcal{A}_{k_1} \cdots \mathcal{A}_{k_p}) - e^{-(\sum_{i=1}^{p} \rho_{k_i} + 1)\lambda}|.$$

Для каждого $k_1 < \ldots < k_p$, $e^{-(\sum_{i=1}^{j}\rho_{k_i}+1)\lambda} \le e^{-(\rho_{\min} j+1)\lambda}$. Тогда,

$$\begin{aligned} |P\{R \models_{\min} A\} \;-\; e^{-\lambda}\prod_{k=1}^{r}(1-e^{-\rho_k\lambda})| &\le 2S_T + \sum_{j\le T}|S_j - \hat{S}_j| \\ &\le 2\binom{r}{T}e^{-(\rho_{\min}T+1)\lambda} + \sum_{j\le T}\binom{r}{j}R_j = e^{-\lambda}\prod_{k=1}^{r}(1-e^{-\rho_k\lambda})(I_1+I_2). \end{aligned}$$

Для $I_1 := 2\binom{r}{T}e^{-\rho_{\min}\lambda T}\prod_{k=1}^{r}(1-e^{-\rho_k\lambda})^{-1}$, из формулы Стирлинга находим

$$\binom{r}{T} = \exp\{f(v)r + g(v) - 1/2\ln r + \mathrm{O}(1)\}, \tag{10}$$

где $f(v) := -(v\ln v + (1-v)\ln(1-v)) = \mathrm{O}(v\ln v)$ при $v \to 0$, $0 < f(v) < \ln 2$, и $g(v) := -(\ln v + \ln(1-v))$. Далее, можем оценить

$$\prod_{k=1}^{r}(1-e^{-\rho_k\lambda})^{-1} \le (1-e^{-\rho_{\min}\lambda})^{-r}. \tag{11}$$

Будет использоваться следующее элементарное неравенство

$$-x/(1-x) \le \ln(1-x) \le -x \qquad \text{для } 0 < x < 1. \tag{12}$$

Следовательно, применяя (10) и (11), а также (12), получаем

$$I_1 \le C\exp\{rf(v) + r\ln(1-e^{-\rho_{\min}\lambda})^{-1} - \rho_{\min}\lambda v r\} \le e^{-\gamma_1 a}, \tag{13}$$

для некоторых $C, \gamma_1 > 0$, достаточно большого $\lambda > c_0/\rho_{\min} > 0$ и достаточно малого $v > 0$.

Для I_2, ипользуя лемму 1, получаем $\max_{j\le T} R_j \le 8\,2^{-\delta a/2+\kappa}\lambda^{1/2}$. Из предположений следует, $\kappa = \mathrm{o}(a)$ и $r = \mathrm{O}(a)$ при $a \to \infty$. Значит, для $c_0/\rho_{\min} < \lambda < ca$, мы получаем

$$\begin{aligned} I_2 \;&:=\; e^{\lambda}\prod_{k=1}^{r}(1-e^{-\rho_k\lambda})^{-1}\sum_{j\le T}\binom{r}{j}R_j \le e^{\lambda}(1-e^{-\rho_{\min}\lambda})^{-r}\binom{r}{T}\max_{j\le T}R_j \\ &\le\; C\exp\{rf(v) + \lambda + r\ln(1-e^{-\rho_{\min}\lambda})^{-1} + \kappa - \delta a/2\,\ln 2\} \le e^{-\gamma_2 a}, \end{aligned} \tag{14}$$

для некоторых $C > 0$ $\gamma_2 > 0$, достаточно большого $\rho_{\min}\lambda$ и достаточно малого v, как и выше. Наконец, утверждение (i) следует из (13) и (14).

Далее, следующее утверждение очевидно (ср. (7)) для каждого целого $T \le r$,

$$P\{R \models_{\min} A\} = P(\{R \models A\} \bigcap_{j=1}^{r} \overline{\{R \models A_j\}}) \le P(\{R \models A\} \bigcap_{j=1}^{T} \overline{\{R \models A_j\}}),$$

и следовательно, $P\{R \models_{\min} A\} \le P_a := P\{R \models A\} - \sum_{j=1}^{T} (-1)^{j-1} S_j$. Пусть $T = vr$, где $0 < v < 1$. Тогда, $P_a = \mathrm{O}(e^{-\gamma a})$ и утверждение (ii) следует из леммы 1, как и выше.

Очевидно, для отношения $\mathcal{R}$, $\{R \models_{\min} A\} \subseteq \{N(U) = 0\}$ и следовательно,

$$P\{\mathcal{R} \models_{\min} A\} = P\{R \models_{\min} A / N(U) = 0\} \ge P\{R \models_{\min} A\}. \qquad (15)$$

Применение теорем 1, 3 и (15) завершает доказательство теоремы. □

Доказательство оценки (3). Пусть $f(x) := x \prod_{k=1}^{r} (1 - x^{\rho_k})$, $0 < x < 1$. Тогда $f(x)$ непрерывная и вогнутая на $[0, 1]$, $f(0) = f(1) = 0$, и значит имеет единственную точку максимума $0 < x_0 < 1$. С помощью несложных вычислений x_0 может быть найдена из следующего уравнения

$$\sum_{k=1}^{r} \rho_k x_0^{\rho_k} / (1 - x_0^{\rho_k}) = 1, \qquad (16)$$

и следовательно, $\rho_\cdot / (x_0^{-\rho_{\min}} - 1) \ge 1 \ge \rho_\cdot / (x_0^{-\rho_{\max}} - 1)$. Отсюда вытекает, что $\rho_{\min} \le \bar{\rho} \le \rho_{\max}$ и $x_0 = (\rho_\cdot + 1)^{1/\bar{\rho}}$. Далее, применяя предположение, находим

$$\ln x_0 \ge \ln(1 + \rho_\cdot)/\rho_{\min} = \mathrm{o}(a). \qquad (17)$$

Из (12) и (16) следует, что

$$\ln \prod_{k=1}^{r} (1 - x^{\rho_k}) \ge -1/\rho_{\min} \sum_{k=1}^{r} \rho_k x_0^{\rho_k} / (1 - x_0^{\rho_k}) = -1/\rho_{\min} = \mathrm{o}(a) > -\gamma a. \qquad (18)$$

С другой стороны, из (12) и (16)

$$\ln \prod_{k=1}^{r} (1 - x^{\rho_k}) \le -(1 - x_0^{\rho_{\min}})/\rho_{\min} \sum_{k=1}^{r} \rho_k x_0^{\rho_k} / (1 - x_0^{\rho_k}) = -(1 - x_0^{\rho_{\min}})/\rho_{\max}. \qquad (19)$$

Итак, теорема 1(i), (17) и (18) позволяют получить, что для всех положительных λ и γ $P^{as}_{\max}(a) := x_0 \prod_{k=1}^{r}(1 - x_0^{\rho k}) \geq \exp\{\mathrm{o}(a)\} > e^{-\gamma a}$, и следовательно, $P_{\max}(a) = P^{as}_{\max}(a)(1 + \mathrm{o}(1))$, что завершает доказательство $\square$

Доказательство (4). По определению $\lambda = M2^{-a(r)}$, и $|e^{-\lambda} - e^{-m^2 2^{-a(r)-1}}| = \mathrm{O}(m2^{-a(r)}) = \mathrm{O}(m2^{-(1+\delta)a/2})$ равномерно $r \geq 1$ и $m \geq 1$. Применяя теорему 1(i), получаем

$$P\{\nu_m \leq r\} - e^{-m^2 2^{-a(r)-1}} = \mathrm{O}(m2^{-(1+\delta)a/2}) \tag{20}$$

равномерно по $r \geq 1, m \geq 1$, откуда и следует утверждение. $\square$

Доказательство предложения 1. Пусть r_0 натуральное такое, что $m^{-2} \ln m \leq 2^{-r_0 d} \leq 4m^{-2} \ln m$. Тогда $r_0 \geq 1$, $r_0 = \mathrm{O}(\ln m)$, и

$$\begin{aligned} d_{TV}(\mathcal{L}(\nu_m), \mathcal{L}(\hat{X}_m)) &\leq \sum_{r=r_0}^{\infty} |P\{\nu_m \leq r\} - P\{\hat{X}_m \leq r\}| + P\{\hat{X}_m \leq r_0\} = \\ &= \mathrm{O}(m2^{-(1+\delta)r_0 d/2} + e^{-m^2 2^{-r_0 d-1}}) = \mathrm{O}\Big((\ln m)^{(1+\delta)/2} m^{-\delta} + 1/m\Big) \end{aligned}$$

при $m \to \infty$, и утверждение (i) следует.

Для среднего значения, заметим, что для целочисленных ν_m и $\hat{X}_m$

$$\begin{aligned} &|\mathrm{E}\nu_m - \mathrm{E}\hat{X}_m| \leq \\ &\leq \sum_{r=r_0}^{\infty} |P\{\nu_m \leq r\} - P\{\hat{X}_m \leq r\}| + \sum_{r<r_0} P\{\hat{X}_m \leq r_0\} + r_0 P\{\nu_m \leq r_0\} \leq \\ &\leq \mathrm{O}\Big((\ln m)^{(1+\delta)/2} m^{-\delta}\Big) + r_0(P\{\nu_m \leq r_0\} - P\{\hat{X}_m \leq r_0\}) + \mathrm{O}(r_0 P\{\hat{X}_m \leq r_0\}) = \\ &= \mathrm{O}\Big((\ln m)^{(1+\delta)/2} m^{-\delta} + (\ln m)^{(3+\delta)/2} m^{-\delta} + \ln m/m\Big) = \mathrm{O}\Big((\ln m)^{(3+\delta)/2} m^{-\delta}\Big) \end{aligned}$$

при $m \to \infty$, где r_0, как и при доказательстве (i). По определению, $0 \leq E\hat{X}_m - 1/d\,(m_g/\ln 2 + 2\log_2 m - 1) \leq 1$. Применяя аналогичные рассуждения для k-моментов $\mathrm{E}\nu_m^k$, $k \geq 2$,

$$|E\nu_m^k - E\hat{X}_m^k| = \mathrm{O}\Big((\ln m)^{(2k+1+\delta)/2} m^{-\delta}\Big) \text{ при } m \to \infty,$$

что завершает доказательство. $\square$

Заключение

Рассмотрен широкий класс стохастических моделей для баз данных. Свойства ключевых систем и функциональных зависимостей в стохастических базах данных исследованы для подхода в среднем. В сравнении с подходом по наихудшему случаю, экспоненциальный размер системы минимальных ключей для подхода в среднем маловероятен. Для нескольких стохастических моделей получено пуассоновское приближение основных характеристик наиболее вероятных ключевых кандидатов через соотвествующие энтропии Реньи. Предложенный общий метод анализа основывается на вероятностных и теоретико-информационых результатах. В качестве первого небходимого шага для дальнейшего *статистического* анализа, исследованы несколько *вероятностных* моделей для случайных таблиц и отношений. Общая схема для возможных дальнейших практических применений может быть следующей

(i) подбор соответствующей стохастической модели;

(ii) оценка параметров модели по базе данных;

(iii) аппроксимация и анализ характеристик ключевой системы.

Помимо рассмотренных дискретных распределений, могут использоваться также марковская модель для стохастических зависимостей и соответствующие многомерные нормальные приближения для дискретных распределений. Для оценки параметров нормальных распределений по базе данных могут быть применены хорошо известные методы (см., например, [3]).

Список литературы

[1] Андреев, А. (1984). Об асимптотическом поведении числа тупиковых тестов для почти всех таблиц. *Проблемы кибернетики* **41**, 117-142.

[2] Barbour, A. D., Holst, L. & Janson, S. (1992). *Poisson Approximation.* Clarendon Press: Oxford.

[3] Christodoulakis, S. (1983). Estimating record selectivities. *Inform. Syst.* **8**, 105-115.

[4] Demetrovics, J., Katona, G. O. H., Miklos, D., Seleznjev, O. & Thalheim, B. (1995). The average length of keys and functional dependencies

in (random) databases. In: *Proc. ICDT-95* (Gottlob, G., Vardi, M., eds). Lecture Notes in Computer Science **893**, Springer, 266-279.

[5] Demetrovics, J., Katona, G. O. H., Miklos, D., Seleznjev, O. & Thalheim, B. (1998). Asymptotic properties of keys and functional dependencies in random databases. *Theor. Comp. Science*, **190**, 151-166.

[6] Devroye, L. (1984). A probabilistic analysis of the height of tries иof the complexity of trie sort. *Acta Inform.* **21**, 229-237.

[7] Leadbetter, M. R., Lindgren, G. & Rootzén, H. (1986). *Extremes and related properties of random sequences and processes.* Springer-Verlag, New York. (Перевод: Лидбеттер, М. Р., Линдгрен, Г., Ротсен, Х. (1989). *Экстремумы случайных последовательностей и процессов.* Мир, Москва)

[8] Malvestuto, F.M. (1983). Theory of random observables in relational databases. *Inform. Syst.* **8**, 281-289.

[9] Mannila, H. & Räihä, K.-J. (1992). *The design of relational databases*, Addison-Wesley: Amsterdam.

[10] Rényi, A. (1970). *Probability Theory.* North-HollиPbl. Com.: Amsterdam.

[11] Pflug, G. (1986). *Stochastische Modelle in der Informatik.* Teubner Verlag: Stuttgart.

[12] Seleznjev, O. & Thalheim, B. (1988). On the number of minimal keys in relational databases over nonuniform domains. *Acta Cybern.* **8**, 267-271.

[13] Seleznjev, O. & Thalheim B. (1996). Random databases and keys. *Tech. Univ. Cottbus Research Report*, 1996:I-11, 1-15.

[14] Szpankowski, W. (1991). On the height of digital trees and related problems. *Algorithmica* **6**, 256-277.

[15] Thalheim, B. (1991). *Dependencies in Relational Databases.* Teubner Verlag: Leipzig.

[16] Traub, J. F., Wasilkowski, G. W. & Woźniakowski, H. (1988). *Information-based Complexity.* Academic Press: San Diego.