Алгоритм приближенного поиска ближайшего цифрового массива на множестве пирамидальных представлений данных

Ланге М.М. , Ганебных С.Н., Ланге А.М.

Федеральный исследовательский центр "Информатика и управление" РАН
Вычислительный центр им. А.А. Дородницына РАН

17 - я Всероссийская конференция "Математические методы распознавания образов"
(ММРО - 17, Россия, г. Светлогорск, 19 - 25 сентября 2015 г.)

2. Содержание доклада

- Проблема приближенного поиска ближайшего соседа
- Источник и представление данных
- Мера и критерий поиска на множестве представлений
- Решаемая задача
- Структура базы данных и алгоритм поиска
- Асимптотика сложности алгоритма поиска
- Экспериментальные оценки точности и сложности алгоритма
- Результаты и выводы

3. Проблема приближенного поиска ближайшего соседа (Approximate Nearest Neighbor Search)

Множество векторов $\mathbf{S} \subset \mathbf{E}^d$ мощности $\|\mathbf{S}\| = n$ в пространстве размерности d

Ближайший сосед $\mathbf{x}_{NN} \in \mathbf{S}$ вектора $\mathbf{x} \in \mathbf{E}^d$ по заданной мере различия $D(\mathbf{x}, \mathbf{x}')$:

$$\mathbf{x}_{NN} = \arg\min_{\mathbf{x}' \in \mathbf{S}} D(\mathbf{x}, \mathbf{x}')$$

Аппроксимация ближайшего соседа $\mathbf{x}_{AN} \in \mathbf{S}$ вектора $\mathbf{x} \in \mathbf{E}^d$ с точностью $\varepsilon \geq 0$:

$$D(\mathbf{x}, \mathbf{x}_{AN}) \leq (1+\varepsilon) D(\mathbf{x}, \mathbf{x}_{NN})$$

Вычислительная сложность приближенного поиска (поиска аппроксимации $\mathbf{x}_{AN}$):

$$C = O(c_{\varepsilon,d} \log n), \quad \text{где} \quad c_{\varepsilon,d} = d\lceil 1 + 6d/\varepsilon \rceil^d$$

****

S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu.
An optimal algorithm for approximate nearest neighbor searching in fixed dimensions.
Journal of the ACM, 45(6), 1998.

A. Andoni and P. Indyk. Near – optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM, 2008, *vol*. 51, *no*.1, *pp*.117 – 122.

4. Источник и представление данных

Множество массивов $\mathbf{X} = \{\mathbf{x}\}$, заданных m – мерными кубами элементов с ребром N:

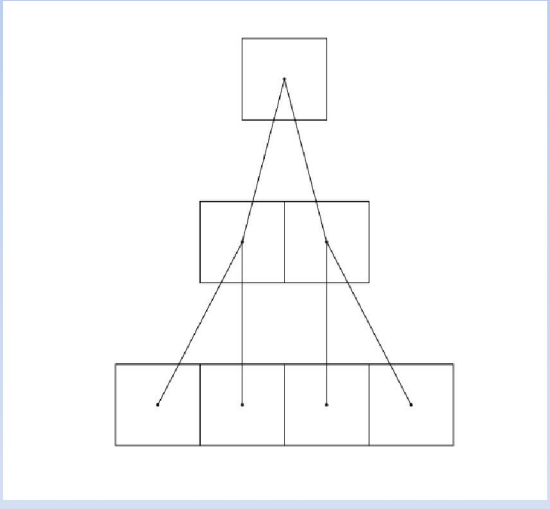
$$\|\mathbf{x}\| = N^m,\ N = 2^L\ ,\ L > 1$$

Многоуровневое пирамидальное представление m – мерных массивов:

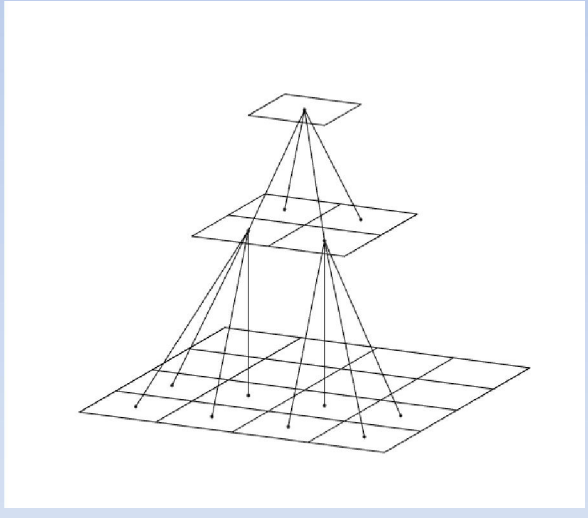
$$\mathbf{X} \to \mathbf{X}^L, \text{ где } \mathbf{x} \to X^L = (\mathbf{x}_0,\ldots,\mathbf{x}_l,\ldots\mathbf{x}_L) \in \mathbf{X}^L\ ,\ \ \|\mathbf{x}_l\| = 2^{lm}, l = 0,\ldots,L$$

Примеры пирамидальных представлений глубины $L = 2$: $X^2 = (\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2)$

$m = 1$



$m = 2$



$\mathbf{x}_0 \quad X^0 = \mathbf{x}_0$

$\mathbf{x}_1 \quad X^1 = (\mathbf{x}_0, \mathbf{x}_1)$

$\mathbf{x}_2 \quad X^2 = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2)$

Индексы элементов пирамидального представления:

$$\mathbf{x}_l = \{z_{\mathbf{k}_l^m}\},\ \mathbf{k}_l^m = (k_{l1},\ldots,k_{lq},\ldots,k_{lm})\ ,\ \ k_{lq} = 1,\ldots,2^l,\ q = 1,\ldots,m$$

5. Критерий поиска на множестве представлений

Мера различия порядка $l = 0,...,L$ для пары массивов $\mathbf{x} \to \mathbf{x}_l = \{z_{\mathbf{k}_l^m}\}$, $\hat{\mathbf{x}} \to \hat{\mathbf{x}}_l = \{\hat{z}_{\mathbf{k}_l^m}\}$:

$$D_l(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2^{lm}} \sum_{k_{l1}}^{2^l} \dots \sum_{k_{lm}}^{2^l} \left| z_{k_{l1},\dots k_{lm}} - \hat{z}_{k_{l1},\dots k_{lm}} \right|$$

Средневзвешенная мера порядка $l = 1,...,L$:

$$D_l^*(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=1}^{l} w_i D_i(\mathbf{x}, \hat{\mathbf{x}}) \ , \ w_i = \frac{\log 2^{im}}{\sum_{i=1}^{l} \log 2^{im}}$$

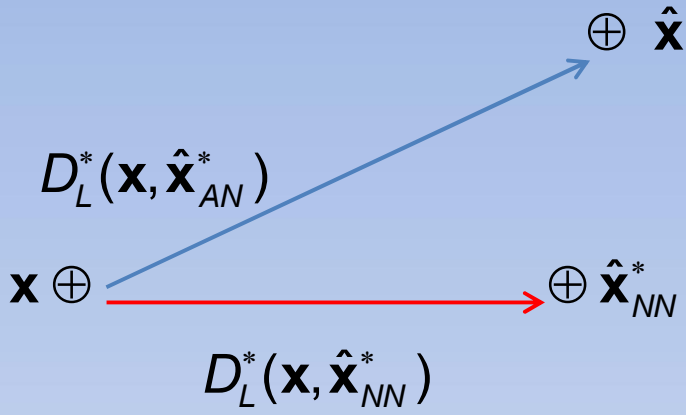
Правило приближенного поиска ближайшего соседа на заданном множестве $\hat{\mathbf{X}} \subset \mathbf{X}$ по мере наибольшего порядка L :

$$\hat{\mathbf{x}}^* = \arg\min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}^*} D_L^*(\mathbf{x}, \hat{\mathbf{x}})$$

где $\hat{\mathbf{X}}^* \subset \hat{\mathbf{X}}$ - подмножество, отбираемое стратегией поиска

6. Решаемая задача

Точность приближенного поиска (дефект подмножества $\hat{\mathbf{X}}^* \subset \hat{\mathbf{X}}$)



$$D^*_L(\mathbf{x}, \hat{\mathbf{x}}^*_{AN}) = \min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}^*} D^*_L(\mathbf{x}, \hat{\mathbf{x}})$$

$$D^*_L(\mathbf{x}, \hat{\mathbf{x}}^*_{NN}) = \min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}} D^*_L(\mathbf{x}, \hat{\mathbf{x}})$$

$$D^*_L(\mathbf{x}, \hat{\mathbf{x}}^*_{AN}) \geq D^*_L(\mathbf{x}, \hat{\mathbf{x}}^*_{NN})$$

Точность приближенного поиска предъявляемого массива $\mathbf{x}$:

$$\varepsilon_{\mathbf{x}}(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) = \frac{\min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}^*} D^*_L(\mathbf{x}, \hat{\mathbf{x}})}{\min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}} D^*_L(\mathbf{x}, \hat{\mathbf{x}})} - 1$$

Средняя по множеству из M массивов точность приближенного поиска:

$$\varepsilon(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) = \frac{1}{M} \sum_{k=1}^{M} \varepsilon_{\mathbf{x}_k}(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}})$$

Выбор подмножества $\hat{\mathbf{X}}^* \subset \hat{\mathbf{X}}$ при заданной сложности алгоритма поиска $C^* > 0$

$$\varepsilon(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) \to \min_{\hat{\mathbf{X}}^* : C(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) \leq C^*}$$

7. База данных и алгоритм поиска

Структура базы данных на представлениях множества массивов $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} \to (\hat{\mathbf{X}}^1,\ldots,\hat{\mathbf{X}}^l,\ldots,\hat{\mathbf{X}}^L),\ \left\|\hat{\mathbf{X}}\right\| = \left\|\hat{\mathbf{X}}^l\right\| = n,\ l = 1,\ldots,L$$

Стратегия поиска массива в базе данных:

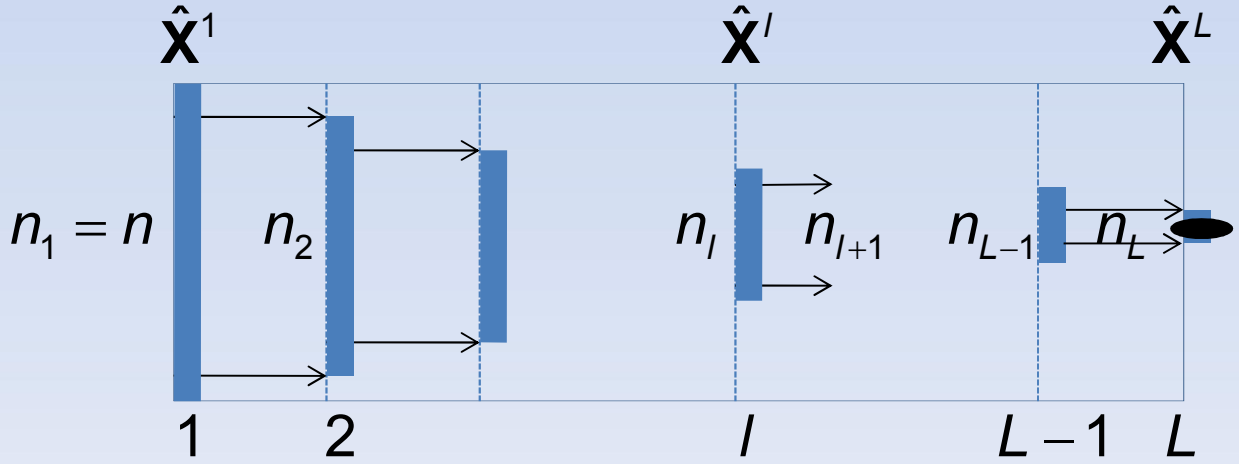
число просматриваемых массивов на последовательных уровнях

$$n_l = \left\lfloor n2^{-\alpha m(l-1)} \right\rfloor,\ l = 1,\ldots,L;$$

коэффициент сужения зоны поиска

$$\alpha = \frac{1}{(L-1)m}\log_2\left(\frac{n}{n^*}\right),\ n^* = n_L = \left\|\hat{\mathbf{X}}^*\right\|$$

Блок - схема алгоритма поиска :



8. Асимптотика сложности алгоритма поиска

Сложность алгоритма поиска в базе данных глубины $L = \log_2 N$:

Затраты на вычисление меры:

$$C_{n^*}^{msr} = \sum_{l=1}^{L} n_l 2^{ml} \le n2^m \sum_{l=1}^{L} 2^{m(1-\alpha)(l-1)}, \quad 1 \le n^* \le n$$

Затраты на сортировку значений меры (вычисление порядковых статистик):

$$C_{n^*}^{srt} = \begin{cases} O\left(\sum_{l=1}^{L} n_l\right) = O(nL) & , \ 1 \le n^* < n \\ n-1 & , \ n^* = n \end{cases}$$

Суммарные вычислительные затраты:

$$C_{n^*} = C_{n^*}^{msr} + C_{n^*}^{srt} = \begin{cases} O(n \log N) & , \ n^* \le n2^m / N^m \\ \Omega(nN^m) & , \ n^* = n \end{cases}$$

Утверждение. Для источника массивов с параметрами $m \ge 1, N \to \infty, n \to \infty$ и алгоритма поиска с параметром $n^* \le n2^m / N^m$ справедлива оценка

$$\frac{C_{n^*}}{C_n} = O\left(\frac{\log N}{N^m}\right)$$

9. Схема экспертимента

Источник изображений рукописных цифр : $m = 2, N = 32, \|\mathbf{X}\| = 60000$

$$\begin{cases} \|\hat{\mathbf{X}}\| = n = 50000 - \text{число массивов в базе данных} \\ \|\mathbf{X} \setminus \hat{\mathbf{X}}\| = M = 10000 - \text{число предъявляемых массивов} \end{cases}$$

Параметры алгоритма поиска :

$n / n^* = 2^k$, $k = 0,1,\ldots,10$; $n^* = \|\hat{\mathbf{X}}^*\|$

Оценки точности поиска по M предъявляемым объектам :

средний дефект поиска $$\varepsilon(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{\min\limits_{\hat{\mathbf{x}}\in\hat{\mathbf{X}}^*} D_L^*(\mathbf{x}_k, \hat{\mathbf{x}})}{\min\limits_{\hat{\mathbf{x}}\in\hat{\mathbf{X}}} D_L^*(\mathbf{x}_k, \hat{\mathbf{x}})} - 1\right)$$

стандартное отклонение $$\sigma(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}}) = \left(\frac{1}{M}\sum_{k=1}^{M}\left(\frac{\min\limits_{\hat{\mathbf{x}}\in\hat{\mathbf{X}}^*} D_L^*(\mathbf{x}_k, \hat{\mathbf{x}})}{\min\limits_{\hat{\mathbf{x}}\in\hat{\mathbf{X}}} D_L^*(\mathbf{x}_k, \hat{\mathbf{x}})} - 1\right)^2 - \varepsilon^2(\hat{\mathbf{X}}^* \mid \hat{\mathbf{X}})\right)^{1/2}$$

Численные оценки вычислительной сложности поиска :

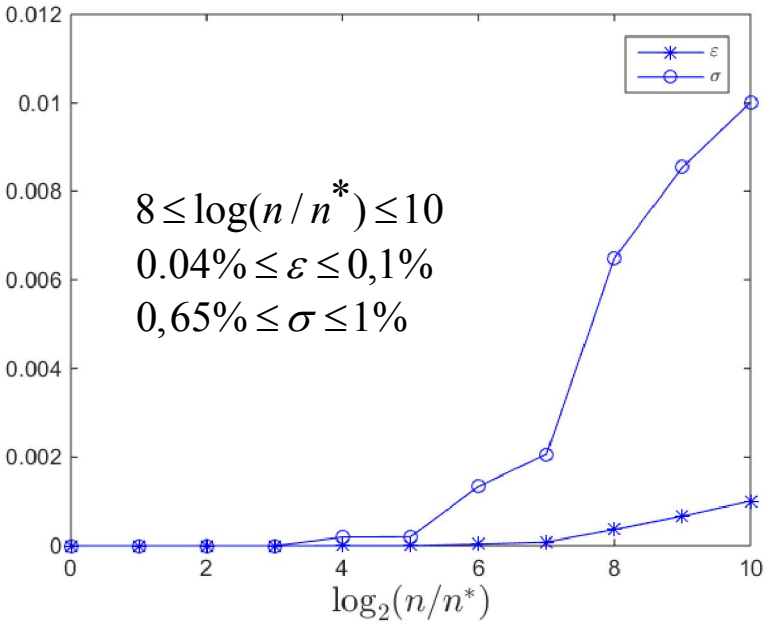
сложность вычисления меры $$C_{n^*}^{msr} = \sum_{l=1}^{L}\left\lfloor n\left(\frac{n^*}{n}\right)^{\frac{l-1}{L-1}}\right\rfloor 2^{ml}$$

сложность сортировки $$C_{n^*}^{srt} = (n-1)[n^* = n] + \left((n^* - 1) + \sum_{l=1}^{L}\left\lfloor n\left(\frac{n^*}{n}\right)^{\frac{l-1}{L-1}}\right\rfloor \log_2\left\lfloor n\left(\frac{n^*}{n}\right)^{\frac{l-1}{L-1}}\right\rfloor\right)[n^* < n]$$

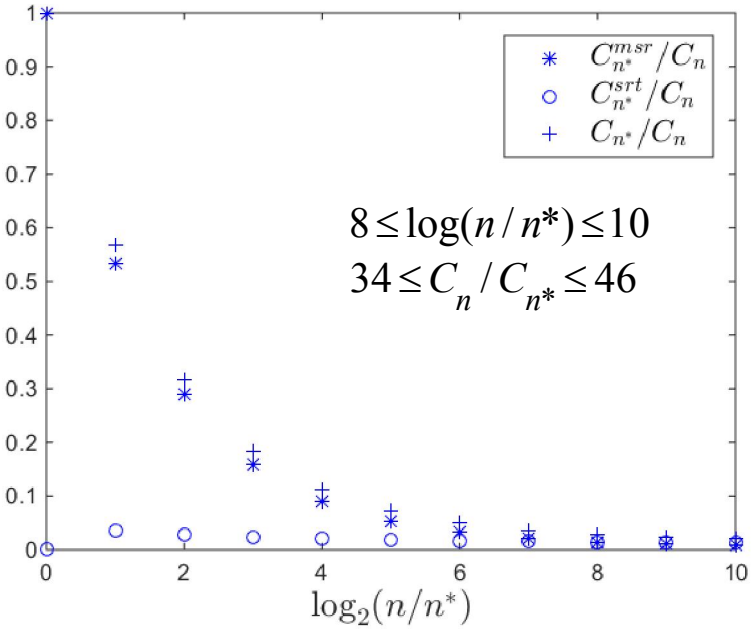
10. Численные оценки эффективности поиска

$\log(n/n^*)$	0	1	2	3	4	5	6	7	8	9	10
$C_{n^*}^{msr}/C_n$	0.9993	0.5325	0.2885	0.1597	0.0908	0.0535	0.0329	0.0214	0.0146	0.0107	0.0082
$C_{n^*}^{srt}/C_n$	0.0007	0.0356	0.0285	0.0237	0.0205	0.0182	0.0166	0.0154	0.0146	0.0139	0.0134
C_{n^*}/C_n	1.0000	0.5681	0.3170	0.1834	0.1113	0.0717	0.0496	0.0368	0.0292	0.0246	0.0216
ε	0	0	0	0	0.0000	0.0000	0.0000	0.0001	0.0004	0.0007	0.0010
σ	0	0	0	0	0.0002	0.0002	0.0013	0.0021	0.0065	0.0085	0.0100

средний дефект поиска ε
стандартноеотклонение σ



сложность приближенного поиска
относительно сложности перебора



11. Результаты и выводы

- **Для источника цифровых массивов, заданных m - мерными кубами из N^m из элементов, предложен иерархический алгоритм приближенного поиска ближайшего массива в множестве мощности n.**
- **Алгоритм ориентирован на ускорение поиска при большом линейном размере N в пространстве пирамидальных представлений массивов с многоуровневым разрешением за счет сужения зоны поиска на последовательных уровнях представления.**
- **При $m \geq 1$ и больших значениях n и N вычислительная сложность приближенного иерархического алгоритма и точного переборного алгоритма удовлетворяют оценкам $O(n \log N)$ и $\Omega(nN^m)$. Вычислительный выигрыш приближенного алгоритма относительно точного составляет $\Omega(N^m / \log N)$.**
- **Экспериментальные оценки точности и вычислительной сложности приближенного алгоритма поиска на множестве изображений рукописных цифр продемонстрировали средний дефект поиска порядка 0,1% при 40 - кратном выигрыше во времени по сравнению с точным переборным алгоритмом.**
- **Для понижения порядка роста вычислительной сложности поиска от мощности n базы данных планируется обобщение алгоритма на основе использования пирамидального представления массивов и решающего дерева, представляющего базу данных.**