УДК 004.8

А.Г. ЧУХРАЙ

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

МЕТОД НЕЧЕТКОГО ПОИСКА ОБЪЕКТОВ

Разработан метод нечеткого поиска объектов, в основе которого лежат необходимые условия похожести в евклидовом пространстве, доказанные ранее для метрики В.И. Левенштейна. Метод основан на случайном выборе k-осей евклидова пространства среди исходных объектов, проецировании всех исходных объектов в k-мерное евклидово пространство, заполнении специальной хэш-структуры данных и быстром поиске объектов, похожих на искомый, на основе доказанных необходимых условий похожести объектов в евклидовом пространстве. Проведены экспериментальные исследования, которые показывают преимущества разработанного метода по быстродействию над одним из известных методов.

Ключевые слова: *нечеткий поиск, евклидово пространство, необходимые условия похожести.*

Введение

К настоящему времени для задач искусственного интеллекта разработаны и используются множество методов поиска „ближайшего соседа", поиска в диапазоне и кластеризации объектов [1 –3]. Тем не менее, вопрос разработки наиболее быстродействующих методов в условиях, когда расчеты расстояний между объектами все еще занимают основное время поиска, остается открытым.

Разработан и экспериментально апробирован метод, в основе которого лежат необходимые условия похожести, являющиеся обобщением условий, доказанных ранее для метрики В.И.Левенштейна и описанных в [4, 5]

1. Постановка задачи

В обобщенном виде постановка задачи 31 имеет следующий вид.

Пусть задано расстояние δ между объектами некоторого класса Cl, которое удовлетворяет условиям:

$$\begin{cases} \delta(X,Y) \ge 0 - \text{неотрицательность}; \\ \delta(X,X) = 0 - \text{свойство нуля}; \\ \delta(X,Y) = \delta(Y,X) - \text{симметричность}; \\ \delta(X,Z) \le \delta(X,Y) + \delta(Y,Z) - \text{неравенство треугольника}. \end{cases} \quad (1)$$

Пусть дан некий объект rt класса Cl и набор объектов ET=(et_1,et_2,...,et_n) этого же класса. Необходимо найти все et_i из набора ET такие, что расстояние δ между et_i и rt не больше некоторого заданного натурального числа λ.

Формально, требуется найти

$$ET_s = \{et_{s1}, et_{s2}, ..., et_{sl}\}$$

такое, что

$$\forall et_{si} \in ET_s \subseteq ET : \delta(et_{si}, rt) \le \lambda, \lambda \in N, l \le n .$$

2. Суть метода

Предлагаемый метод решения M1 состоит из двух шагов.

1 шаг. Из набора ET случайно выбираются k элементов $o_1, o_2, ..., o_k$, $(k \le n)$. Эти элементы в последующем ассоциируются с k осями k-мерного евклидова пространства E^k. После этого каждому элементу et_i из набора ET ставится в соответствие точка E^k, координаты которой равны расстояниям до осей, т.е.

$$P(et_i)_j = \delta(et_i, o_j), i = \overline{1,n}, j = \overline{1,k} .$$

2 шаг. Объекту rt также ставится в соответствие в E^k точка с координатами $P(rt)_j = \delta(rt, o_j)$, $j = \overline{1,k}$. На этом шаге расстояния рассчитываются лишь между rt и объектами, соответствующие точки которых расположены близко в E^k к точке $P(rt)$.

Для определения близко расположенных точек в евклидовом пространстве введем необходимые условия похожести объектов X, Y и Z класса Cl.

Утверждение 1. Для заданных объектов X, Y и Z класса Cl, расстояние между которыми δ удовлетворяет условиям (1), выполняется

$$\forall X, Y, Z \ \ \delta(X,Y) \ge |\delta(X,Z) - \delta(Z,Y)| .$$

Доказательство. Рассмотрим два неравенства треугольника:

1) $\delta(X,Z) \le \delta(X,Y) + \delta(Y,Z)$;

2) $\delta(Y,Z) \le \delta(Y,X) + \delta(X,Z)$.

Из первого неравенства следует
$\delta(X,Z)-\delta(Y,Z)\le\delta(X,Y)$.
Из второго –
$\delta(Y,Z)-\delta(X,Z)\le\delta(Y,X)$.

Объединяя оба следствия и, воспользовавшись свойством симметричности, получаем систему неравенств

$$\begin{cases}\delta(X,Y)\ge\delta(X,Z)-\delta(Z,Y),\\ \delta(X,Y)\ge\delta(Z,Y)-\delta(X,Z)\end{cases}$$

или $\delta(X,Y)\ge|\delta(X,Z)-\delta(Z,Y)|$, что и требовалось доказать.

Утверждение 2. Для заданных объектов et_i и et_j класса Cl, расстояние между которыми δ удовлетворяет условиям (1) и не превышает некоторый порог λ, точки $P(et_i)$ и $P(et_j)$ пространства E^k, соответствующие исходным объектам по методу M1, удалены в E^k друг от друга на расстояние не более чем $\lambda\sqrt{k}$, т.е.

$$\forall i\forall j\ne i\ \delta(et_i,et_j)\le\lambda:$$
$$\rho(P(et_i),P(et_j))\le\lambda\sqrt{k}.$$

Доказательство. По определению метрика пространства E^k

$$\rho(P(et_i),P(et_j))=$$
$$=\sqrt{(P(et_i)_1-P(et_j)_1)^2+...+(P(et_i)_k-P(et_j)_k)^2}.$$

Согласно утверждению 1

$$|\delta(et_i,o_1)-\delta(et_j,o_1)|\le\delta(et_i,et_j),\ldots,$$
$$|\delta(et_i,o_k)-\delta(et_j,o_k)|\le\delta(et_i,et_j).$$

Отсюда, транзитивно,

$|P(et_i)_1-P(et_j)_1|\le\lambda$, …, $|P(et_i)_k-P(et_j)_k|\le\lambda$,

а значит,

$$\sqrt{(P(et_i)_1-P(et_j)_1)^2+...+(P(et_i)_k-P(et_j)_k)^2}\le$$
$$\le\sqrt{\lambda^2k}=\lambda\sqrt{k},$$

что и требовалось доказать.

Утверждение 3. Для заданных объектов et_i и et_j класса Cl, расстояние между которыми δ удовлетворяет условиям (1) и не превышает некоторый порог λ, согласно методу M1 точка $P(et_j)$ размещается в E^k в пределах гиперкуба с центром в точке $P(et_i)$ и стороной длиной 2λ.

Доказательство. Согласно утверждению 1 получаем следующую систему неравенств.

$$\begin{cases}|\delta(et_i,o_1)-\delta(et_j,o_1)|\le\delta(et_i,et_j);\\ |\delta(et_i,o_2)-\delta(et_j,o_2)|\le\delta(et_i,et_j);\\ \ldots\\ |\delta(et_i,o_k)-\delta(et_j,o_k)|\le\delta(et_i,et_j).\end{cases}\quad(2)$$

Отсюда,

$$\begin{cases}P(et_j)_1\ge P(et_i)_1-\lambda;\\ P(et_j)_1\le P(et_i)_1+\lambda;\\ P(et_j)_2\ge P(et_i)_2-\lambda;\\ P(et_j)_2\le P(et_i)_2+\lambda;\\ \ldots\\ P(et_j)_k\ge P(et_i)_k-\lambda;\\ P(et_j)_k\le P(et_i)_k+\lambda\end{cases}\quad(3)$$

Геометрический смысл системы неравенств (3) представляет собой гиперкуб с центром в точке $P(et_i)=(\delta(et_i,o_1),\delta(et_i,o_2),...,\delta(et_i,o_k))$ и стороной длиной 2λ, что и требовалось доказать.

Утверждение 4. Для заданных объектов et_i и et_j класса Cl, расстояние между которыми δ удовлетворяет условиям (1) и не превышает некоторый порог λ, согласно методу M1 абсолютное значение разности расстояний от точек $P(et_i)$ и $P(et_j)$ до начала координат в E^k не превышает $\lambda\sqrt{k}$, т.е. $|\rho(P(et_i),0)-\rho(P(et_j),0)|\le\lambda\sqrt{k}$.

Доказательство. Согласно утверждению 1:

$|\rho(P(et_i),0)-\rho(P(et_j),0)|\le\rho(P(et_i),P(et_j))$.

С другой стороны, согласно утверждению 2

$$\rho(P(et_i),P(et_j))\le\lambda\sqrt{k}.$$

Отсюда

$|\rho(P(et_i),0)-\rho(P(et_j),0)|\le\rho(P(et_i),P(et_j))\le\lambda\sqrt{k}$ и,

следовательно, $|\rho(P(et_i),0)-\rho(P(et_j),0)|\le\lambda\sqrt{k}$, что и требовалось доказать.

Утверждение 5. Пусть $u,w\in R$ и $u,w>0$. Тогда из $[u]\le w$ следует $[u]\le[w]$, где $[u],[w]$ – целые части чисел u и w соответственно.

Доказательство. Рассмотрим два случая:

1) $[u]\le w$ и $[u]=[w]$;

2) $[u]\le w$ и $[u]<[w]$.

В обоих случаях следует $[u]\le[w]$.

Также очевидно, что третий случай, когда $[u]\le w$ и $[u]>[w]$ – не выполним, что и требовалось доказать.

Утверждение 6. Пусть $u,v,w \in R$ и $u,v,w > 0$. Тогда из $|u-v| \le w$ следует $|[u]-[v]| \le [w]+1$, где $[u],[v],[w]$ – целые части чисел u, v и w соответственно.

Доказательство. Рассмотрим случай, когда $u \ge v$. Тогда из $u \le w+v$ следует $[u] \le w+v$, поскольку $u \ge [u]$, а также $[u] \le w+[v]+1$, так как $v < [v]+1$. Согласно утверждению 5 и тому факту, что любое положительное число больше любого неположительного, из $[u]-[v] \le w+1$ следует $[u]-[v] \le [w]+1$. Для случая $v > u$, который рассматривается аналогично, получаем $[v]-[u] \le [w]+1$. Обобщая оба случая, имеем $|[u]-[v]| \le [w]+1$, что и требовалось доказать.

Определение 1. Размером объекта et_i класса Cl назовем количество составляющих его элементов. Введем также обозначение размера – $\overline{et_i}$. Например, если et_i – строка «дом», то $\overline{et_i}$ =3 – длина строки; если et_i – дерево, представленное на рис. 1, то $\overline{et_i}$ =7 – количество его вершин.

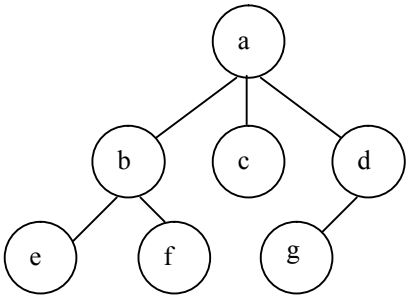

Рис. 1. Пример дерева как объекта класса Cl

Утверждение 7. Если абсолютное значение разности размеров объектов et_i и et_j класса Cl, расстояние между которыми δ удовлетворяет условиям (1), больше λ, то и расстояние между этими объектами больше λ, т.е.

$$(\left|\overline{et_i}-\overline{et_j}\right| > \lambda) \Rightarrow (\delta(et_i, et_j) > \lambda).$$

Доказательство данного утверждения очевидно и следует из факта, что для того чтобы превратить объект et_i в объект et_j или наоборот необходимо выполнить как минимум $\lambda+1$ операцию удаления составляющих объект элементов.

Рассмотрим теперь более подробно суть метода M1 с учетом доказанных необходимых условий похожести объектов.

1. На первом шаге после случайного выбора из набора ET k объектов-осей и вычисления координат точек $P(et_i)$ также вычисляем расстояния в E^k от точек $P(et_i)$ до начала координат: $\rho(P(et_i),0) = \sqrt{P(et_i)_1{}^2 + ... + P(et_i)_k{}^2}$. Кроме того, образуем хэш-структуру данных D распределения расстояний в E^k от точек $P(et_i)$ до начала координат. Для этого введем множество

$$\Psi = \{[\rho(P(et_1),0)],[\rho(P(et_2),0)],...,[\rho(P(et_n),0)]\} = \{\psi_1,\psi_2,...,\psi_z\}, z \le n,$$

где $[\rho(P(et_i),0)]$ обозначает целую часть от $\rho(P(et_i),0)$.

Поставим в соответствие каждому $\psi_i \in \Psi$ множество целых чисел (индексов исходных объектов с расстоянием до начала координат равным ψ_i), т.е. $IND_i = \{ind_{i1}, ind_{i2}, ..., ind_{iw}\}$. При этом выполняется следующее условие

$$\forall q \in \{1,...,w\}\, ind_{iq} \in \{1,...,n\}, \rho(P(et_{ind_{iq}}),0) = \psi_i.$$

Теперь перейдем непосредственно к построению хэш-структуры D. Воспользовавшись вспомогательными множествами IND_i, присвоим $D_{\psi_i q} = ind_{iq}$. Таким образом, строка с индексом ψ_i D содержит индексы исходных объектов, для которых целая часть расстояния в E^k до начала координат равна ψ_i.

2. Для искомого объекта rt находим строку D с индексом $[\rho(P(rt),0)]$. После этого согласно утверждениям 4 и 6, просматриваем ближайшие к ней строки D с индексами из множества

$$\Psi 1 = \{[\rho(P(rt),0)]-[\lambda\sqrt{k}]-1,\ [\rho(P(rt),0)]-[\lambda\sqrt{k}],...,$$
$$[\rho(P(rt),0)]-1,\ [\rho(P(rt),0)],\ [\rho(P(rt),0)]+1,...,$$
$$[\rho(P(rt),0)]+[\lambda\sqrt{k}],\ [\rho(P(rt),0)]+[\lambda\sqrt{k}]+1\} =$$
$$= \{\psi 1_1, \psi 1_2, ..., \psi 1_v\},$$ причем $\Psi 1 \subset \Psi$, $v \le z$.

Затем при просмотре элементов строки D с индексом $\psi 1_t$, т.е. $D_{\psi 1_{tq}}$, производим дальнейшее отсеивание "кандидатов" в похожие объекты: во-первых, путем проверки условия, сформулированного в утверждении 7, а во-вторых, проверяя условие из утверждения 3: лежит ли точка $P(et_{D_{\psi 1_{tq}}})$ в гиперкубе, построенном с центром в точке P(rt) и стороной 2λ. Наконец, если $P(et_{D_{\psi 1_{tq}}})$ находится в пределах заданного гиперкуба, то вычисляем расстояние между объектами rt и $et_{D_{\psi 1_{tq}}}$, т.е. $\delta(rt, et_{D_{\psi 1_{tq}}})$.

3. Инстанцирование метода

Рассмотрим первуй случай, когда объекты класса Cl – m-арные упорядоченные деревья, $m \in N$.

Тогда постановка задачи будет выглядеть следующим образом.

Пусть дано дерево rt и набор деревьев $ET=(et_1,et_2,...,et_n)$. Необходимо найти все деревья et_i из набора ET такие, что расстояние между et_i и rt не больше некоторого заданного натурального числа λ.

В этом случае возможен выбор одной из нескольких метрик δ. В качестве альтернатив могут быть рассмотрены такие метрики, как метрика, используемая в работе [6] и метрика из работ [7, 8].

Отличия двух вышеуказанных метрик заключаются в наборе допустимых операций редактирования деревьев: в каждой метрике допустимыми являются операции переименования, удаления и вставки узлов дерева, причем в первой метрике последние две операции могут применяться только к листьям деревьев, т.е. к вершинам, не имеющим потомков, а во второй метрике эти две операции могут применяться к любым узлам.

Опишем более подробно суть операций вставки и удаления узлов деревьев для второй метрики.

Операция вставки делает некоторые или все потомки узла-родителя для вставляемого узла потомками вставляемого узла. Для операции удаления характерно следующее: все потомки удаляемого узла становятся потомками его родительского узла.

Рассмотрим ряд примеров.

Пример 1. Пусть заданы два упорядоченных бинарных дерева X и Y, изображенные на рис. 2 и 3.

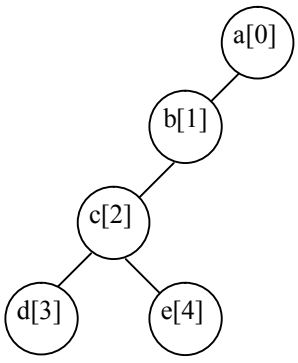

Рис. 2. Упорядоченное бинарное дерево X

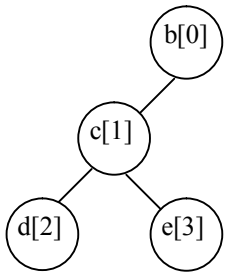

Рис. 3. Упорядоченное бинарное дерево Y

Тогда, расстояние редактирования между X и Y по метрике Selkow [6] равно 6. Минимальный набор операций редактирования для преобразования X в Y будет включать такие операции как:

1) замена наименования «a» узла с индексом 0 на наименование «b»;

2) вставка левым дочерним узлом к узлу с индексом 1 узла «d»;

3) замена наименования «b» узла с индексом 1 на наименование «c»;

4) удаление узла «e» с индексом 4 как потомка узла с индексом 2;

5) удаление узла «d» с индексом 3 как потомка узла с индексом 2;

6) замена наименования «c» узла с индексом 2 на наименование «e».

Для метрики, используемой в работах Tai, Zhang и Shasha [7,8] расстояние редактирования между X и Y будет равно 1, а для преобразования из X в Y необходимо совершить операцию удаления узла «a» с индексом 0.

Пример 2.

Пусть заданы два упорядоченных тернарных дерева X1 и Y1, изображенные на рис. 4 и 5.

Тогда по метрике Selkow [6] расстояние редактирования между X1 и Y1 равно 6.

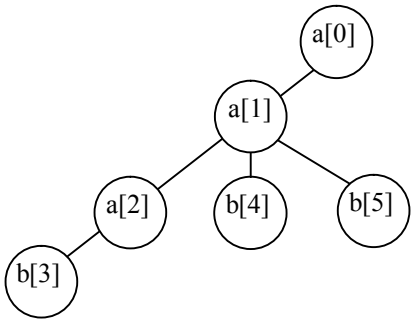

Рис. 4. Упорядоченное тернарное дерево X1

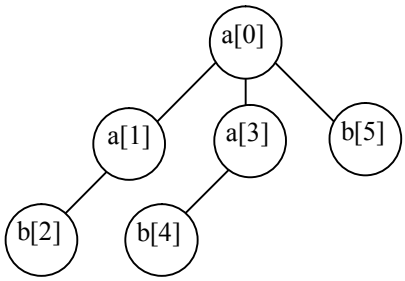

Рис. 5. Упорядоченное
тернарное дерево Y1

Минимальный набор операций редактирования для преобразования X1 в Y1 – это следующие операции:

1) вставка правым дочерним узлом к узлу с индексом 0 узла «b»;

2) вставка левым дочерним узлом к узлу с индексом 0 узла «a»;

3) вставка левым дочерним узлом к вновь вставленному узлу «a» узла «b»;

4) удаление узла «b» с индексом 4 как потомка узла с индексом 1;

5) удаление узла «b» с индексом 3 как потомка узла с индексом 2;

6) удаление узла «a» с индексом 2 как потомка узла с индексом 1.

Для второй метрики расстояние редактирования между X1 и Y1 будет равно 2, а для преобразования из X1 в Y1 необходимо совершить такие операции:

1) удаление узла «a» с индексом 0;

2) вставка узла «a» как потомка узла «a» с индексом 1 и родителя узла «b» с индексом 4.

Как видно из приведенных примеров большая похожесть деревьев достигается при использовании второй метрики. В то же время критерии выбора метрики должны определяться исходя из конкретной практической задачи.

2 случай. Объекты класса Cl – суть строки. Для этого случая ряд задач и их решений с использованием расстояния Левенштейна – минимального количества операций редактирования строки для преобразования ее в другую строку, а также задачи и решения для поиска похожих наборов строк рассмотрены в работах [4,5].

4. Экспериментальные исследования метода M1

Экспериментальные исследования разработанного метода проводились для метрики Tai, Zhang и Shasha. Для сравнения результатов был выбран метод, описанный в работе [1]. Кроме того в метод этих же авторов было добавлено условие, сформулированное в утверждении 7, которое предназначено для отсечения ряда расчетов «дорогого» расстояния редактирования деревьев

Для случайно сгенерированных деревьев в каждом методе из общего времени выполнения второго этапа, поиска, было исключено время, затрачиваемое на расчет расстояний редактирования. Результаты экспериментальных исследований для исходного списка из 10, 100 и 1000 деревьев показаны на рис. 6 – 8.

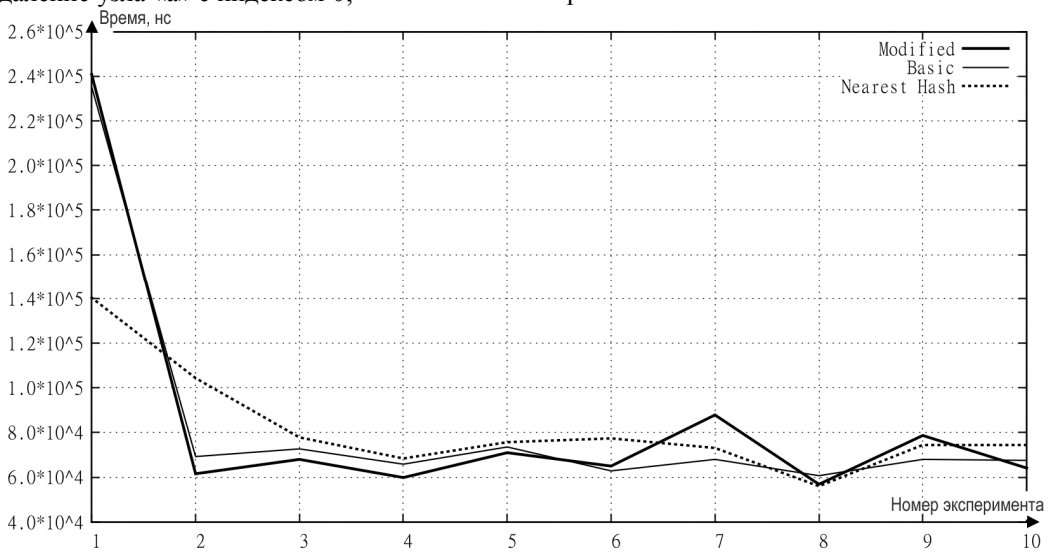

Рис. 6. Результаты экспериментальных исследований трех методов
для исходного списка из 10 деревьев

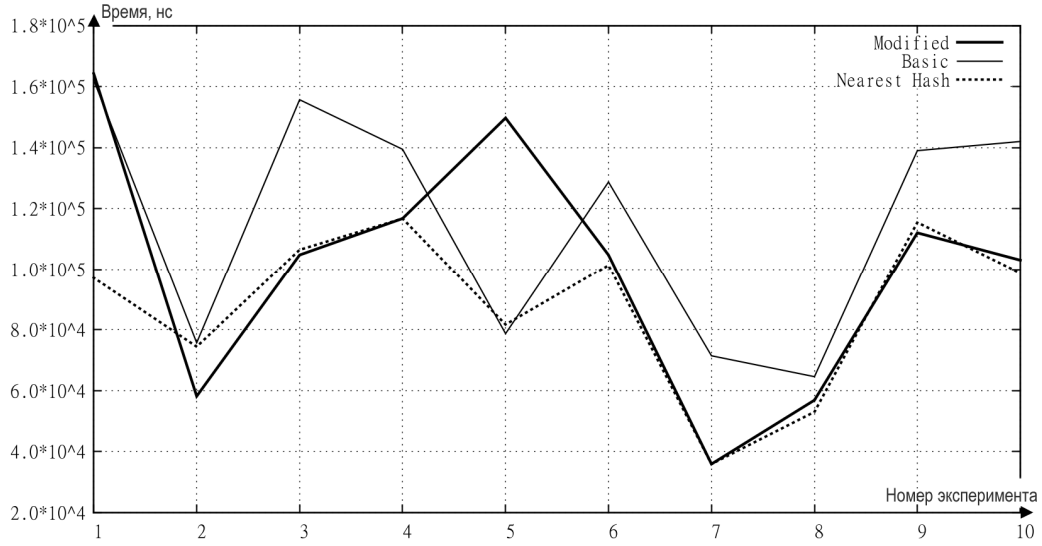

Рис. 7. Результаты экспериментальных исследований трех методов для исходного списка из 100 деревьев

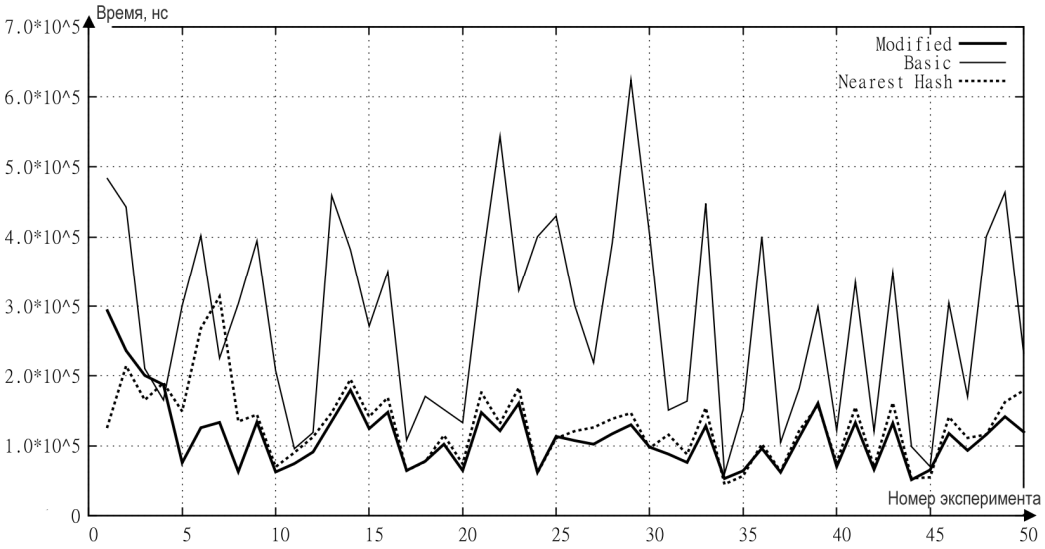

Рис. 8. Результаты экспериментальных исследований трех методов для исходного списка из 1000 деревьев

В эксперименте для 10 и 100 деревьев искомыми были первые 10 деревьев из исходного списка деревьев, а в эксперименте с 1000 – первые 50, что отмечено по оси абсцисс. Как видно из рисунков при первом поиске разработанный метод превосходит известный и модифицированный известный для первого эксперимента из 10 деревьев в 1,71 и в 1,67 раза; для второго эксперимента из 100 деревьев – в 1,66 и 1,69 раза и для третьего эксперимента из 1000 деревьев – в 3,80 и 2,30 раза.

Дальнейшие исследования будут нацелены на новые экспериментальные исследования полученного метода: сравнение с другими известными методами, выбор количества осей, обоснование выбора конкретных осей, а также на различные применения метода при решении прикладных задач.

Литература

1. Bustos, B. Pivot selection techniques for proximity searching in metric spaces [Text] / B. Bustos , G. Navarro , E.Chavez // Pattern Recognition Letters. – 2003. – Vol. 24, Issue 14. – 8 p.

2. Building a web-scale image similarity search system [Text] /M. Batko, F Falchi, C. Lucchese and others. Multimedia Tools and Applications. - 2010. - Vol. 47, № 3. – P.599 - 629.

3. Zezula, P. Similarity search : the metric space approach [Text] / P. Zezula, G. Amato, V. Dohnal, M. Batko. - New York: Springer, 2006. - 220 p.

4. Чухрай, А.Г. Метод быстрого поиска "похожих" кортежей реляционного отношения [Текст] / А.Г. Чухрай // Радіоелектронні і комп'ютерні системи. – 2003. – Вип. 2 (2).– С. 64 – 69.

5. Kulik, A. Similar strings detecting methods [Text] / A.Kulik, A. Chukhray, A. Zavgorodniy // Proceedings of the East-West Fuzzy Colloquium, Zittau, Germany, IPM. - 2005. – P. 38 – 47.

6. Selkow, S.M.. The tree-to-tree editing problem [Text] / S.M. Selkow // Information Processing Leters. – 1977. – Vol. 6, № 6. – P. 184–186.

7. Tai, K.C. The tree-to-tree correction problem [Text] / K.C. Tai // Journal of the ACM. – 1979. – Vol. 26, Issue 3. – P. 422 – 433.

8. Zhang, K. Simple fast algorithms for the editing distance between trees and related problems [Text] / K. Zhang, D. Shasha // Society for Industrial and Applied Mathematics Journal on Computing. – 1989. – Vol. 18, No. 6. – P. 1245 – 1262.

.

Поступила в редакцию 19.03.2012

Рецензент: д-р техн. наук, проф., зав. каф. информатики А.Ю. Соколов, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

МЕТОД НЕЧІТКОГО ПОШУКУ ОБ'ЄКТІВ

А.Г. Чухрай

Розроблено метод нечіткого пошуку об'єктів, в основі якого лежать необхідні умови схожості в евклідовому просторі, доведені раніше для метрики В.І. Левенштейна. Метод заснований на випадковому виборі k-осей евклідового простору серед вихідних об'єктів, проектуванні всіх вихідних об'єктів в k-мірний евклідовий простір, заповненні спеціальної хеш-структури даних і швидкому пошуку об'єктів, схожих на даний, на основі доведених необхідних умов схожості об'єктів в евклідовому просторі. Проведено експериментальні дослідження, які показують переваги розробленого методу за швидкодією над одним з відомих методів.

Ключові слова: нечіткий пошук, евклідовий простір, необхідні умови схожості.

METHOD OF SIMILAR OBJECTS SEARCH

A.G. Chukhray

The method was developed for similar search of objects, which is based on the necessary conditions of similarity in Euclidean space. These necessary conditions of similarity were proved earlier for the metric of V.I. Levenshtein,. The method is based on a random selection of k-axes of the Euclidean space within the original objects, projecting all the source objects in a k-dimensional Euclidean space, filling special hash data structure and fast search of objects, similar to the object, based on the proven necessary conditions for similarity of objects in Euclidean space . Experimental studies were shown the advantages of the developed method for speed over one of the known methods.

Key words: similar search, Euclidean space, necessary conditions of similarity

Чухрай Андрей Григорьевич – канд. техн. наук, доцент, докторант каф. 301, Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт», Харьков, Украина.