УДК 681.51:57

МЕТОД БЫСТРОГО ПОИСКА «ПОХОЖИХ» КОРТЕЖЕЙ РЕЛЯЦИОННОГО ОТНОШЕНИЯ

А.Г. Чухрай
Национальный аэрокосмический университет им. Н.Е.Жуковского «ХАИ»

Изложен метод поиска «похожих» кортежей в реляционном отношении. Теоретически доказано соответствие точности результатов его применения необходимым требованиям. На основе данных о сотрудниках университета «ХАИ» проведены вычислительные эксперименты, продемонстрировавшие высокое быстродействие предложенного метода в случае небольшого порога «похожести».

* * *

Викладено метод пошуку «схожих» кортежів у реляційному відношенні. Теоретично доведено відповідність точності результатів його застосування необхідним вимогам. На основі даних про співробітників університету «ХАІ» проведено обчислювальні експерименти, які продемонстрували високу швидкодію запропонованого методу в разі невеликого порогу «схожості».

* * *

Method of searching "similar" tuples in the relation is presented. Correspondence accuracy of results of its using to necessary requirements is theoretically proved. On the basis of university "KHAI" employees' data computing experiments are conducted. Experiments demonstrated high speed of offered method in the event of the small "similarity" threshold.

Актуальность. Постоянное совершенствование возможностей компьютерных технологий привело к тому, что на современных предприятиях стали накапливаться и обрабатываться огромные объемы информации. Руководители многих предприятий осознали важность информационной поддержки принятия управленческих решений в различных аспектах хозяйственной деятельности. Вместе с тем качество решений насущных задач предприятий во многом зависит от актуальности и достоверности данных информационных систем (ИС). Известны экспертные оценки потерь предприятий, причина которых – некачественные данные ИС. Так, L. English в монографии [1] показывает, что из-за низкого качества данных предприятия ежегодно теряют от 10 до 25 % своего дохода.

Для повышения актуальности и достоверности данных ИС необходимы меры, направленные на применение результативных методов и алгоритмов очистки данных, построенных с учетом возможностей современных аппаратных средств, а также конкретной специфики ошибок, свойственных человеку-оператору. В настоящей работе рассматривается одна из центральных проблем очистки данных – поиск «похожих» кортежей реляционного отношения после поступления в буфер хранилища данных информации из различных источников [2-8]. В литературе данная проблема известна как слияние/очистка (merge/purge) [3,4]. Другими ее названиями являются «связывание записей» (record linkage), «семантическая интеграция» (semantic integration) или «идентификация объектов» (object identity) [5].

Постановка задачи. Пусть дано реляционное отношение, включающее в себя кортежи с некоторой предметно-ориентированной информацией. Требуется найти все пары «похожих» кортежей.

Прежде всего следует отметить, что авторы известных исследований в этой области опираются на различные критерии похожести «кортежей». В частности, в работах [3,4] в качестве такого критерия используется лексикографическая близость кортежей в упорядоченном отношении; исследование [6] построено на основе критерия равенства q-грамм (подстрок длины q) для похожих строк. Наряду с этим в качестве критерия похожести кортежей используется минимум расстояния редактирования между ними. Подобный выбор характерен для работ

[7,8] и обусловливается независимостью результатов поиска избыточных данных от конкретной специфики информации, его адекватностью наиболее распространенным ошибкам человека-оператора, а также возможностью относительно простого перехода от набора строк к отношению кортежей.

Таким образом, формальная постановка задачи выглядит следующим образом. Пусть дано отношение R арности h, включающее в себя n-кортежей вида $m_i = (pk_i, m_{i1},..., m_{ih}) \in R, i \in \{1,..,n\}$, где pk_i - первичный ключ в виде целого числа, уникально идентифицирующего кортеж m_i в отношении R; $m_{i1},...,m_{ih}$ - предметно-ориентированные компоненты кортежа m_i, представляющие собой строковые значения. Необходимо найти все пары кортежей (m_i, m_j), $pk_i > pk_j$, такие, что

$$\sum_{g=1}^{h} d(m_{ig}, m_{jg}) \le \lambda, \qquad (1)$$

где $d(m_{ig}, m_{jg})$ - минимальное расстояние редактирования между строками m_{ig}, m_{jg}, впервые предложенное В.И. Левенштейном в 1965 г. [9]; λ - некоторый заданный порог, $\lambda \in N$.

Очевидно, что альтернативным подходом может быть сведение задачи поиска похожих кортежей к задаче поиска похожих строк путем конкатенации всех предметно-ориентированных компонентов каждого кортежа в строку. Тем не менее, в этом случае эффективность поиска будет заметно ниже в силу того, что для универсального алгоритма расчета расстояния Левенштейна, предложенного Вагнером и Фишером в 1974 г. [10], справедливо следующее утверждение.

Утверждение 1. Сложность вычисления расстояния Левенштейна для двух строк s_i^{con} и s_j^{con}, таких, что $s_i^{con} = m_{i1} \cdot m_{i2} \cdot \ldots \cdot m_{ih}$, $s_j^{con} = m_{j1} \cdot m_{j2} \cdot \ldots \cdot m_{jh}$, посредством алгоритма Вагнера-Фишера превышает сложность h-вычислений расстояний Левенштейна для компонентов кортежей m_i и m_j.

Доказательство данного утверждения следует из факта, что $(a_1 + a_2 + ... + a_h) \times (b_1 + b_2 + ... + b_h) > > (a_1 b_1 + a_2 b_2 + ... + a_h b_h)$, где $a_i, b_i \in N$.

Цель исследования. Используя наивный подход, мы должны для каждой пары кортежей проверить, выполняется ли условие (1). При этом сложность наивного подхода квадратичная - $O(n^2)$. Ясно, что при достаточно большом n алгоритм нахождения нужных пар становится неэффективным.

Таким образом, цель данного исследования представляет собой создание метода быстрого поиска похожих кортежей, который давал бы те же результаты, что и наивный подход.

Сущность метода быстрого поиска похожих кортежей. Предлагаемое решение состоит из двух шагов. На первом шаге выберем из отношения R случайным образом k кортежей $o_1, o_2, ..., o_k$, $(k < n)$, которые в дальнейшем будут ассоциироваться с осями k-мерного эвклидова пространства E^k. Далее, для каждого кортежа $m_i \in R$ определяется соответствующая точка k - мерного эвклидова пространства $P(m_i)$, координаты которой равны суммам расстояний Левенштейна до компонентов осей, т.е. $P(m_i)_j = \sum_{g=1}^{h} d(m_{ig}, o_{jg}), i = \overline{1,n}, j = \overline{1,k}$. На втором шаге расстояния Левенштейна подсчитываются только для геометрически компактно расположенных пар точек $P(m_i)$ и $P(m_j)$.

Введем ряд следующих утверждений.

Утверждение 1. $\forall m_i, m_j, m_x \in R$:

$$\sum_{g=1}^{h} d(m_{ig}, m_{jg}) \ge \left| \sum_{g=1}^{h} d(m_{ig}, m_{xg}) - \sum_{g=1}^{h} d(m_{xg}, m_{jg}) \right|.$$

Доказательство. Перегруппируем слагаемые в правой части неравенства:

$|\sum_{g=1}^{h} d(m_{ig},m_{xg}) - \sum_{g=1}^{h} d(m_{xg},m_{jg})| = |d(m_{i1},m_{x1}) -$

$-d(m_{x1},m_{j1}) + d(m_{i2},m_{x2}) - d(m_{x2},m_{j2}) + ... +$

$+d(m_{ih},m_{xh}) - d(m_{xh},m_{jh})|$. Согласно неравенству треугольника $|d(m_{i1},m_{x1}) - d(m_{x1},m_{j1})| \le$, $\le d(m_{i1},m_{j1})$, $|d(m_{i2},m_{x2}) - d(m_{x2},m_{j2})| \le$ $\le d(m_{i2},m_{j2}),...,|d(m_{ih},m_{xh}) - d(m_{xh},m_{jh})| \le$ $\le d(m_{ih},m_{jh})$. Отсюда $\sum_{g=1}^{h} d(m_{ig},m_{jg}) \ge$ $\ge |\sum_{g=1}^{h} d(m_{ig},m_{xg}) - \sum_{g=1}^{h} d(m_{xg},m_{jg})|$.

Утверждение 2. $\forall i \forall j \ne i \sum_{g=1}^{h} d(m_{ig},m_{jg}) \le \lambda$: $\rho(P(m_i),P(m_j)) \le \lambda\sqrt{k}$.

Доказательство. По определению метрика пространства E^k $\rho(P(m_i),P(m_j)) =$ $= \sqrt{(P(m_i)_1 - P(m_j)_1)^2 + ... + (P(m_i)_k - P(m_j)_k)^2}$. Согласно утверждению 1 $|\sum_{g=1}^{h} d(m_{ig},o_{1g}) - \sum_{g=1}^{h} d(m_{jg},o_{1g})| \le \sum_{g=1}^{h} d(m_{ig},m_{jg}),...,$ $|\sum_{g=1}^{h} d(m_{ig},o_{kg}) - \sum_{g=1}^{h} d(m_{jg},o_{kg})| \le \sum_{g=1}^{h} d(m_{ig},m_{jg})$, т.е. $|P(m_i)_1 - P(m_j)_1| \le \lambda$, $|P(m_i)_k - P(m_j)_k| \le \lambda$, $\sqrt{(P(m_i)_1 - P(m_j)_1)^2 + ... + (P(m_i)_k - P(m_j)_k)^2} \le$ $\le \sqrt{\lambda^2 k} = \lambda\sqrt{k}$, что и требовалось доказать.

Утверждение 3. Если $m_i, m_j \in R$, кортежи строк такие, что выполняется $\sum_{g=1}^{h} d(m_{ig},m_{jg}) \le \lambda$, то точка $P(m_j)$ размещается в E^k в пределах гиперкуба с центром в точке $P(m_i)$ и стороной 2λ.

Доказательство. Согласно утверждению 1

$$(2)\quad \begin{cases} |\sum_{g=1}^{h} d(m_{ig},o_{1g}) - \sum_{g=1}^{h} d(m_{jg},o_{1g})| \le \sum_{g=1}^{h} d(m_{ig},m_{jg}), \\ |\sum_{g=1}^{h} d(m_{ig},o_{2g}) - \sum_{g=1}^{h} d(m_{jg},o_{2g})| \le \sum_{g=1}^{h} d(m_{ig},m_{jg}), \\ \dots \\ |\sum_{g=1}^{h} d(m_{ig},o_{kg}) - \sum_{g=1}^{h} d(m_{jg},o_{kg})| \le \sum_{g=1}^{h} d(m_{ig},m_{jg}). \end{cases}$$

Геометрическим смыслом представленной системы неравенств (2) является гиперкуб с центром в точке $P(m_i) = (\sum_{g=1}^{h} d(m_{ig},o_{1g}), \sum_{g=1}^{h} d(m_{ig},o_{2g}),..., \sum_{g=1}^{h} d(m_{ig},o_{kg}))$ и стороной 2λ.

Утверждение 4. Если $m_i, m_j \in R$, кортежи строк такие, что выполняется $\sum_{g=1}^{h} d(m_{ig},m_{jg}) \le \lambda$, то абсолютное значение разности расстояний от точек $P(m_i)$ и $P(m_j)$ до начала координат в E^k не превышает $\lambda\sqrt{k}$, т.е. $|\rho(P(m_i),0) - \rho(P(m_j),0)| \le \lambda\sqrt{k}$.

Доказательство. Согласно неравенству треугольника $|\rho(P(m_i),0) - \rho(P(m_j),0)| \le \rho(P(m_i),P(m_j))$. С другой стороны, согласно утверждению 2 $\rho(P(m_i),P(m_j)) \le \lambda\sqrt{k}$. Отсюда транзитивно $|\rho(P(m_i),0) - \rho(P(m_j),0)| \le \lambda\sqrt{k}$.

Основываясь на приведенных выше теоретических утверждениях, сформулируем суть предлагаемого метода в виде последовательности действий.

1. На первом шаге из отношения R случайно выбираются k кортежей $O = \{o_1,o_2,...,o_k\}$, $(k < n)$, которые представляются в виде осей k-мерного эвклидова пространства E^k. Далее, каждому кортежу $m_i \in R$ ставится в соответствие точка k-мерного эвклидова пространства $P(m_i)$, координаты которой равны суммам расстояний Левенштейна до компонентов осей, т.е.

$P(m_i)_j = \sum_{g=1}^{h} d(m_{ig}, o_{jg}), i = \overline{1,n}, j = \overline{1,k}$. Здесь также вычисляются расстояния в E^k от точек $P(m_i)$ до начала координат – $\rho(P(m_i),0) = \sqrt{P(m_i)_1^2 + P(m_i)_2^2 + ... + P(m_i)_k^2}$. Кроме того, формируется матрица D - распределения расстояний в E^k от точек $P(m_i)$ до начала координат. Для этого введем множество $\Psi = \{[\rho(P(m_1),0)],[\rho(P(m_2),0)],...,[\rho(P(m_n),0)]\} = \{\psi_1,\psi_2,...,\psi_z\}, z \le n$, где $[\rho(P(m_i),0)]$ обозначает целую часть от $\rho(P(m_i),0)$. Поставим в соответствие каждому $\psi_i \in \Psi$ множество $IND_i = \{ind_{i1}, ind_{i2},..., ind_{iw}\}$, такое, что $\forall j \in \{1,...,w\}\, ind_{ij} \in \{1,...,n\}, \rho(P(m_{ind_{ij}}),0) = \psi_i$. Теперь перейдем непосредственно к построению матрицы D, размерность которой равна $(\max(\Psi) - \min(\Psi) + 1) \times (\max\{|IND_1|,...,|IND_z|\})$. Воспользовавшись вспомогательными множествами IND_i, присвоим $D_{\psi i j} = ind_{ij}$. Таким образом, ψ_i-я строка матрицы D содержит индексы исходных кортежей, для которых целая часть расстояния в E^k до начала координат равна ψ_i.

2. На втором шаге для каждого кортежа находятся «похожие» кортежи следующим образом. Для каждого m_i находится строка матрицы D с индексом $[\rho(P(m_i),0)]$. После этого, согласно утверждению 4, просматриваются ближайшие к ней строки матрицы D с индексами из множества $\Psi 1 = \{ [\rho(P(m_i),0)] - [\lambda\sqrt{k}] - 1,..., [\rho(P(m_i),0)],..., [\rho(P(m_i),0)] + [\lambda\sqrt{k}] + 1\} = \{\psi 1_1,...,\psi 1_v\}$, причем $\Psi 1 \subset \Psi, v \le z$. Затем при просмотре элементов $\psi 1_t$-й строки, т.е. $D_{\psi 1_t j}$, происходит дальнейшее "отсеивание кандидатов" в похожие кортежи путем применения утверждения 3, т.е. проверки условия: лежит ли точка $P(m_{D_{\psi 1_t j}})$ в гиперкубе, построенном с центром в точке $P(m_i)$ и стороной 2λ. Наконец, если $P(m_{D_{\psi 1_t j}})$ находится в пределах заданного гиперкуба, то вычисляется расстояние Левенштейна между кортежами m_i и $m_{D_{\psi 1_t j}}$.

Эксперименты. В качестве объекта исследования была выбрана реальная база данных о штатных сотрудниках университета «ХАИ», состоящая из 3043 записей и включающая в себя такие поля, как «Фамилия», «Имя», «Отчество», «Подразделение», «Должность», средняя длина значений которых равна 7.67, 6.32, 10.07, 12.27 и 16.37 соответственно. Все эксперименты проводились на персональном компьютере с процессором CELERON 566 МГц и 196 Мб ОЗУ. Операционная система – Windows NT 4 Server, компилятор – Borland Delphi 6. В качестве допустимого порога λ в случае наивного и быстрого поиска была выбрана величина, равная четырем.

В результате поиска и в том, и в другом случае были обнаружены 73 похожих кортежа. Выявлены такие похожие кортежи, как "Абєлєнцева Ганна Петрівна НТБ зав. відділом", "Абеленцева Ганна Петрівна НТБ зав. відділом" (d=2); "Агапова Оксана Олександрівна НТБ бібліотекар 1 кат", "Агапова Оксана Олександрівна НТБ бібліограф 1 кат" (d=4); "Невський Анатолій Євгенович кафедра 407 уч.майстер", "Невський Анатолій Євгенович кафедра 407 майстер" (d=3) и другие. Следует отметить, что автоматическое исправление похожих кортежей не целесообразно, поскольку похожие кортежи могут относиться к разным сотрудникам: например, "Бодрова Дар`я Валентинівна кафедра 302 технік 2 кат ", "Бодрова Марія Валентинівна кафедра 302 технік 1 кат " (d=3); "Колесник Володимир Петрович кафедра 402 доцент, к. н. ", "Олейник Володимир Петрович кафедра 502 доцент, к. н. " (d=4).

В таблице приведены усредненные результаты вычислительных экспериментов на базе наивного (н) и предложенного подхода (п) при $k = 10$.

N	500	1000	1500	2000	2500	3043
(н) t, с	12	62	127	200	313	500
(п) t, с	0,6	1,1	2,0	2,9	4,1	5,2

Временные характеристики обоих алгоритмов представлены графически на рисунке в виде графика зависимости времени выполнения алгоритма от общего количества кортежей в базе данных.

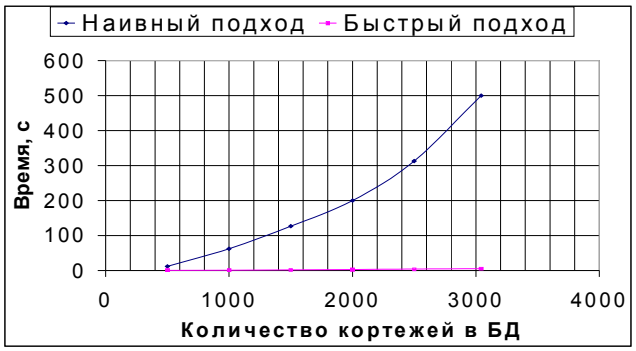

Рис. 1. Графики зависимости времени выполнения наивного и предложенного алгоритмов от общего количества кортежей в базе данных

Как показано на рисунке, результаты вычислительных экспериментов демонстрируют значительное превосходство в быстродействии предложенного метода над наивным подходом, которое на объеме базы данных, равном 3043 записям, достигает порядка 100 раз.

Сравнение с другими подходами. В публикациях [3,4] проблема поиска похожих кортежей рассматривается в контексте слияния записей из нескольких неоднородных гетерогенных источников данных. Основная идея описанного подхода заключается в конкатенации источников данных в один список размером N, последующей сортировке списка и сканировании отсортированного списка посредством окна размером $W, W \in \{2,...,N\}$. При этом каждая новая запись сравнивается с записями в окне относительно похожести исходя из некоторого заданного критерия. Недостатком данного подхода является то, что две похожие записи могут быть не «захвачены» окном небольшого размера. Так, например, расстояние Левенштейна между строками '*акт*', '*пакт*'; '*Дарья*', '*Марья*'; '*ВАЗ*', '*МАЗ*' – 1, но они будут располагаться далеко друг от друга в отсортированном списке. В то же время подход, представленный в настоящем описании, лишен данного недостатка, поскольку не основывается на упорядочении исходных записей.

Подход, изложенный в [7], адаптирован к поиску похожих строк в двух разных источниках данных и также характеризуется наличием двух этапов: отображения исходных строк в k-мерное эвклидово пространство и поиска в нем похожих пар объектов. Вместе с тем сложность алгоритма для первого этапа, после которого расстояние между объектами многомерного эвклидова пространства сохраняется приблизительно равным расстоянию редактирования между исходными строками, - $O(5 \times k^2 \times N)$. В нашем подходе сложность первого этапа - $O(k \times N)$, причем k - величины одного и того же порядка в обоих случаях. На втором этапе описанного подхода для каждого из источников данных строится R - дерево. Далее для обхода построенных деревьев и поиска похожих пар объектов используется стратегия «поиска в глубину». Отличительной особенностью подхода, предложенного в настоящем описании, является 100%-ная точность результатов поиска, т.е. будут найдены все возможные пары похожих кортежей записей, в то время как для подхода, описанного в [7], необходим эмпирический выбор допустимого порога, который устанавливает степень похожести объектов в k-мерном эвклидовом пространстве и, таким образом, влияет на точность результатов поиска. Следующим недостатком сравниваемого подхода, в отличие от предложенного, является невозможность нахождения похожих кортежей внутри одного и того же источника.

Заключение

Таким образом, получен эффективный и точный метод поиска «похожих» кортежей в реляционном отношении. Доказаны условия, ограничивающие область поиска «похожих» кортежей. Результаты экспериментов позволяют судить о высоком быстродействии предложенного метода в случае небольшого порога "похожести", а сравнение с другими подходами – о точности производимых им результатов. В дальнейшем метод может быть исследован на наборе строк с неоднородной технико-экономической информацией.

Литература

1. English L. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. – New York, John Wiley & Sons, 1999. – 544 p.

2. Кулик А.С., Нечипорук Н.В., Чухрай А.Г. Выбор архитектуры баз данных информационной системы управления административно-финансовой деятельностью университета «ХАИ» // Авіаційно-космічна техніка та технологія. – Харків. – 2002. – Вип. 32.– С. 191-196.

3. Hernandez, M. A., Stolfo, J. S. The merge/purge problem for large database. In Proceedings of the ACM SIGMOD International Conference on Management of Data, May 1995. P. 127-138.

4. Hernandez, M. A., Stolfo, J. S. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem, Journal of Data Mining and Knowledge Discovery. Vol. 2, 1998. P. 9-37.

5. Maletic, J., Marcus, A. Data Cleansing: Beyond Integrity Analysis, Proceedings of The Conference on Information Quality, Massachusetts Institute of Technology, Boston, MA, 2000. P. 200-209.

6. Luis Gravano, Panagiotis G. Ipeirotis, H.V.Jagadish, Nick Koudas, S. Muthukrishnan, Divesh Srivastava. Approximate string joins in a database (almost) for free. In Proceedings of the 2001 VLDB Conference. P. 491-500.

7. Liang Jin, Chen Li, Sharad Mehrotra. Efficient Similarity String Joins in Large Data Sets, Technical Report, University of California, Irvine, Department of Information and Computer Science, Feb., 2002. - 26 p.

8. Monge A. Adaptive detection of approximately duplicate database record and the database integration approach to information discovery. Ph.D. thesis, University of California, San Diego, 1997. – 98 p.

9. Левенштейн В.И. «Двоичные коды с исправлением выпадений, вставок и замещений символов» Доклады Академии Наук СССР. Т. 163. - 1965. - №4, - С. 845-848.

10. Wagner, R.A., Fischer, M.J.: *The String-to-String Correction Problem*. // Journal of the Association for Computing Machinery. 1974. Vol. 21. No.1. – С. 168-173.

Поступила в редакцию: 05.04.03

Рецензенты: д-р техн. наук, доцент Соколов А.Ю., Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», г.Харьков; д-р техн. наук, профессор Петров Э.Г., ХНУРЭ, г. Харьков.