УДК 519.8

О СЛУЧАЙНОМ ВЫБОРЕ ЗНАЧЕНИЙ В АЛГОРИТМАХ ПО ВОССТАНОВЛЕНИЮ НАРУШЕНИЙ ФУНКЦИОНАЛЬНЫХ ЗАВИСИМОСТЕЙ

Борисов Д.В., студент, E-mail: zdvighkov@mail.ru
МГТУ МИРЭА, Москва, Россия

Аннотация. В статье рассмотрены проблемы нарушения целостности данных. Проанализированы существующие алгоритмы по восстановлению нарушений функциональных зависимостей. Выявлена и обоснована низкая вероятность правильного восстановления данных, а так же необходимость создания дополнительных ограничений, влияющих на выбор значений для восстановления нарушенных кортежей.
Ключевые слова: нарушение целостности, функциональная зависимость, кортежи, схема отношений.

ON A RANDOM SELECTION OF VALUES IN THE ALGORITHMS FOR RESTORATION OF FUNCTIONAL DEPENDENCY VIOLATIONS

Borisov D.V., student, E-mail: zdvighkov@mail.ru
MSTU MIREA, Moscow, Russia

Abstract. The article considers the problem of data breaches. Analyzed existing algorithms for restoration of functional dependency violations. Identified and justified the low probability of correct recovery of the data, as well as the need to create additional constraints that affect the choice of values for the recovery of damaged tuples.
Keywords: integrity, functional dependence, tuples, the pattern of relationships.

В последнее время наблюдается значительное повышение спроса на вычислительную технику, в том числе на смартфоны, планшеты, компьютеры. Это влечет за собой увеличение объемов обрабатываемой и хранимой информации [5].

Во многих случаях целостность данных может быть нарушена. Например, при извлечении данных из интернета, при заполнении пользователями данных в различных приложениях, интеграции данных из различных источников. В связи с этим необходимо создание алгоритмов, устраняющих подобные нарушения. Для того что бы упростить этот процесс, зависимости разбиваются на следующие типы:

- функциональная зависимость [1];
- условная функциональная зависимость [1];
- функциональная зависимость от включения [2];
- многозначная функциональная зависимость [4];
- транзитивная функциональная зависимость [4].

Такой подход позволяет более детально рассмотреть каждый из них.

Самым распространенным из всех типов зависимостей является функциональная зависимость. Задача восстановления нарушений функциональных зависимостей (*FDs*) сложна. На сегодняшний день не удалось решить все проблемы, связанные с ней, в полном объеме. В работе “Sampling from repairs of conditional functional dependency violations” [1] были приведены алгоритмы по устранению нарушений одной и нескольких функциональных зависимостей.

В ходе изучения статьи и реализации одного из алгоритмов, восстанавливающих нарушения функциональной зависимости, был выявлен их недостаток: все представленные алгоритмы восстанавливают нарушенные ячейки в кортежах случайным образом.

Рассмотрим предложенный в данной статье алгоритм по восстановлению нарушений одной *FDs*. Кортежи *t* проходят через программу, которая распознает нарушения *FDs* и присваивает нарушенным ячейкам значение *null*. Таким образом, происходит предварительный подсчет нарушенных кортежей и создание промежуточных таблиц, которые содержат нарушенные кортежи. После работы алгоритма каждый кортеж, у которого один из атрибутов содержит в себе ячейку со значением *null*, подается на вход следующему блоку алгоритма. Этот блок восстанавливает ячейки таким образом, что *FDs* не нарушены.

Кортежи, которые были отмечены алгоритмом как нарушенные, подвергаются восстановлению путем выбора случайного значения из множества допустимых значений *Dom(A)* в случае, если значение *null* содержится в атрибуте, который определяет функциональную зависимость(*LHS*). Если же значение *null* содержится в функционально зависимом атрибуте(*RHS*), тогда переменная, которая будет участвовать в восстановлении, выбирается из множества значений всех нарушенных кортежей$\{(t_i, RHS)\}$, у которых атрибуты определяющие *FDs* равны. Таким образом, происходит восстановление всех нарушений.

Реализованный алгоритм выбирает значения для кортежей случайным образом, соответственно восстанавливая *LHS* или *RHS*. Пространство выборки для *LHS* постоянно, так как значения выбираются из домена атрибута – *Dom(A)\ADom(A)*, где *ADom(A)* является множеством значений, которые уже были использованы в других кортежах. Значения для *RHS* выбирается из множества $\{(t_i, RHS)\}$. Таким образом, при одной *FDs* множество $\{(t_i, RHS)\}$ содержит 2 и более значений, из которых будет производиться выборка.

Допустим, что $\{(t_i, RHS)\}$ не больше 2. В силу случайного выбора переменной, правильное восстановление *RHS* будет выбрано с вероятностью $P=1/2$. При

восстановлении *LHS* вероятность получения правильного значения равна $P=1/(Dom(A)\backslash ADom(A))$. Если область допустимых значений *Dom(A)* $[1\ldots10^5]$, и восстановление происходит путем случайного выбора значений, то вероятность получить правильное восстановление очень мала.

Функциональная зависимость $F: X \rightarrow Y$ согласно теории множеств не нарушена. Каждому значению атрибутов кортежа отношения из *X* соответствует не более одного значения атрибутов того же кортежа отношения из *Y*, но этого недостаточно, так как неизбежно возникнут смысловые нарушения между данными [3].

Таким образом, полученные результаты показывают невысокую эффективность выбора значений, участвующих в восстановлении нарушенных *FDs* для *LHS*. Существующий алгоритм нуждается в создании дополнительных ограничений, для выбора значений, участвующих в восстановлении, как для *LHS*, так и для *RHS*.

Список литературы

1. Beskales G., Golab L., Galiullin A. Sampling from repairs of conditional functional dependency violations // The VLDB Journal - 2014. №23. С. 103-128.

2. Fan W., Geerts F., Jia X. Conditional Dependencies: A Principled Approach to Improving Data Quality//University of Edinburgh and Bell Laboratories. 2009.

3. Функциональные зависимости //http://cribs.me/bazy-dannykh /funktsionalnye - zavisimosti// 2014г.

4. Понятие функциональной, транзитивной и многозначной зависимости. // http://e-educ.ru/bd21.html // 2014г.

5. ИТ-рынок России//http://www.tadviser.ru/index.php/Статья:ИТ-рынок_России//2014