С. Н. Гриднев, Г. В. Авакян (Москва, МИФИ). **Алгоритм поиска идентичных объектов в реляционной базе данных, основанный на анализе строковых данных.**

В настоящее время подавляющее большинство современных информационных систем хранит пользовательские данные и системные настройки в реляционных базах данных (РСУБД). Появляется необходимость разработки алгоритма поиска объектов-двойников путем анализа строковых данных на сходство с последующим совмещением результатов сравнения по всем анализируемым полям.

Для определения схожести объектов, характеризующихся набором текстовых полей, используется следующий алгоритм.

1) Для каждого текстового поля проводится оценка совпадения выбранным методом сравнения строк. В результате получается оценка совпадения, выраженная в процентах.

2) Далее выбирается один из подходов по выставлению весовых коэффициентов важности характеристик объектов. В случае отсутствия одного или нескольких параметров сравнения, производится перерасчет коэффициентов с исключением из расчета неиспользуемых полей.

3) Производится суммирование полученных результатов сравнения с использованием весовых коэффициентов. Результаты с наибольшей суммой сравнения признаются совпадающими.

Метод сравнения строк. Используемый нами метод распространяет обычный метод Shift-And на нахождение неточного вхождения образца в текст. Под «неточностью» понимается, что образец входит в текст либо точно, либо с малым числом несовпадений, вставок или пропусков символов. Например, образец «atcgaa» входит в текст «aatatccascaa» с двумя несовпадениями, начиная с четвертой позиции, а также он входит с четырьмя несовпадениями, начиная со второй позиции.

Алгоритмы выставления весовых коэффициентов

1. Метод *равновесных* коэффициентов. Данный метод позволяет произвести равномерное распределение весовых коэффициентов при условии, что поля объектов при сравнении являются равнозначными. Общая формула определения весового коэффициента имеет следующий вид: $W = (1/\mathrm{K_{сум}}) \cdot 100\%$, где W — значение весового коэффициента; $\mathrm{K_{сум}}$ — количество полей, по которым производится сравнение.

2. Метод *приоритетного распределения* коэффициентов. Метод применим в том случае, есть существуют приоритетные условия сравнения. Например, совпадение поля 1 имеет больший вес, чем совпадения поля 2, 3.

Пусть имеется N сравниваемых полей. При этом поле 1 является более важным в сравнении, чем поле 2. Поле 2 – чем поле 3 и т. д. Весовые коэффициенты будут вычисляться по следующему правилу: $W_i = ((N - i + 1)/(N^2 - \sum P) \cdot 100\%,\ P \in (1, 2, \ldots, N - 1)$.

3. Метод *выставления весов* в зависимости от длин сравниваемых строк. Основан на сравнении нормализованных средних длин строк: чем меньше средняя длина, тем больший вклад вносит данное поле в суммарный результат.

СПИСОК ЛИТЕРАТУРЫ

1. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология. СПб.: Невский Диалект, БХВ-Петербург, 2003, 656 с.
2. *Романовский И. В.* Дискретный анализ. 2-е изд., исправленное. СПб.: Невский диалект, 2000, 240 с.