ПРИБЛИЗИТЕЛЬНАЯ ОЦЕНКА ПОДОБИЯ ДЕРЕВЬЕВ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Лычковский А. В.

Волорова Н. А. – к-т. техн. наук, доцент

Сравнение данных является распространенной операцией в вычислительных системах. Большие объемы информации требуют эффективных алгоритмов их сравнения. Зачастую сравниваемые данные берутся из разных источников, в которых они представлены в разных форматах. Простое сопоставление таких данных не даст приемлемого результата, т.к. их структура различается. Однако, в большинстве случаев, информация о структуре также может быть использована в процессе сравнения.

Реальным примером является задача слияния баз данных с адресами. Эта задача не может быть решена простым сопоставлением строк, т.к. имена улиц могут отличаться в зависимости от соглашений принятых при их сохранении. Более того, имена улиц могут быть записаны на другом языке. Для решения этих проблем можно использовать иерархическую структуру адреса. На рисунке 1 представлены таблицы с именами улиц из разных источников, а также таблицы с адресами на этих улицах. Представление в виде дерева записей 30 и 91 из таблиц показано на рисунке 2. Наличие меры, которая оценивает схожесть деревьев, позволило бы решить поставленную задачу слияния баз данных с адресами.

SRD

id	*street*
30	Giuseppe-Cesare-Abba-Str.
120	Sebastian-Altmann-Str.
5220	Bozner-Boden-Str.
3000	Hermann-von-Gilm-Str.
3030	Pater-Reginaldo-Giuliani-Str.
3540	Italienallee
4440	Musterplatzl
7180	Raffaello-Sernesi-Galerie
7590	Telsergalerie
7620	Friedensplatz
7650	Turiner Str.
7740	Trienter Str.
7860	Triester Str.
8580	Walther-v.-d.-Vogelweide-Pl.
3930	Giannantonio-Manci-Str.
...	

SLR

id	*street*
91	CESARE ABBA STRASSE
74	S. ALTMANN STRASSE
33	BOZNER BODENWEG
109	GILMWEG
185	P. R. GIULIANI STR.
115	ITALIENSTRASSE
165	MUSTERPLATZ
207	SERNESIDURCHGANG
259	TELSERDURCHGANG
139	SIEGESPLATZ
266	TURINER STRASSE
262	TRIENTER STRASSE
263	TRIESTER STRASSE
285	WALTHERPLATZ
86	MANCISTRASSE
...	

RD

id	*num*	*entr*	*apt*	*resident*
30	1	-	1	Pichler
30	1	-	3	Rieder
30	2	A	-	Maier
30	2	B	1	Rossi
30	2	B	2	Woelk
30	2	B	3	Verdi
30	2	B	4	Verdi
30	2	C	-	Burger
30	3	-	-	Hofer
30	4	A	1	Tribus
30	4	A	2	Palermo
30	4	A	3	Palermo
30	4	B	-	Abel
30	4	C	-	Rossi
30	6	-	-	Spiro
120	3	A	1	Spiro
120	3	A	2	Barducci
120	3	A	3	Costanzi
120	3	A	4	Pichler
120	3	A	5	Spiro
120	3	A	6	Raifer
...				

LR

id	*num*	*entr*	*apt*	*owner*
91	1	-	1	Maier
91	1	-	2	Rossi
91	1	-	3	Sparber
91	2	A	-	Maier
91	2	B	1	Totti
91	2	B	2	Bracco
91	2	B	3	Mair
91	2	B	4	Lun
91	2	D	-	Tribus
91	3	-	-	Costanzi
91	4	A	-	Palermo
91	4	B	-	Abel
91	4	C	-	Rossi
91	6	-	-	Spiro
74	3	A	1	Spiro
74	3	A	2	Barducci
74	3	A	3	Costanzi
74	3	A	4	Spiro
74	3	A	6	Spiro
74	3	A	7	Hofer
74	4	-	-	Mueller
...				

Рисунок 1 – Исходные данные в разных источниках

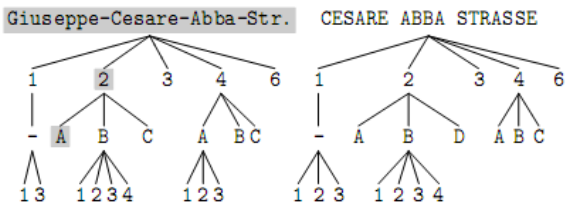

Рисунок 2 – Деревья с адресами для записей 30 и 91

Одна из хорошо известных мер – редакционное расстояние, которое определяется как минимальная по стоимости последовательность операций вставки узла, удаления узла, переименования метки, которая преобразует одно дерево в другое [1]. Существует алгоритм [2], вычисляющий редакционное расстояние за $O(n^2 \min^2(l,d))$ времени и $O(n^2)$ памяти для деревьев с n узлами, l листьями и глубиной d. Более ранние алгоритмы имеют сложность $O(n^2)$ и не могут быть использованы для больших деревьев.

Для решения некоторых прикладных задач, например такой, как описана выше, требуется лишь примерная оценка схожести деревьев. Pq-грамм расстояние позволяет сделать такую оценку за время $O(n\log(n))$ и требует $O(n)$ памяти.

Ниже дается определение pq-грамм и мере основанной на этом понятии. Если сформулировать коротко, то pq-граммы дерева – это все поддеревья определенной формы. Для того чтобы быть уверенным, что каждый узел дерева присутствует хотя бы в одной pq-грамме, начальное дерево расширяется нулевыми узлами. Исходя из этого, pq-граммы определяются, как поддеревья расширенного дерева.

Определение 1(расширенное дерево) Пусть T дерево, а $p > 0, q > 0$ два целых числа. Расширенное дерево T^{pq} строится из T добавлением $p-1$ предка к корневому узлу, вставкой $q-1$ ребенка перед

первым и после последнего ребенка каждого не листового узла, а также добавлением q детей к каждому листу T. Все новые узлы, должны быть нулевыми, такими, что не встречаются в T.

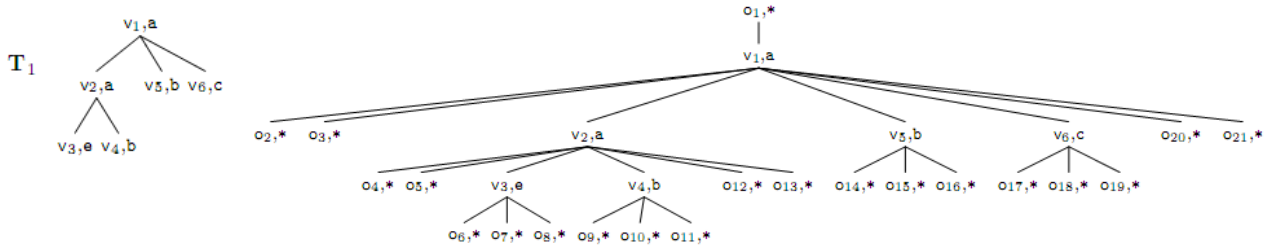

Рисунок 3 – Дерево T_1 и его расширенная версия $T_1^{2,3}$

Определение 2(pq-грамм шаблон) Для $p > 0, q > 0$, pq-грамм шаблон – дерево из основного узла с $p - 1$ предками и q детьми.

Определение 3(pq-грамм) Для $p > 0, q > 0$, pq-грамм дерева T - поддерево расширенного дерева T^{pq}, которое изоморфно pq-грамм шаблону.

Определение 4(кортеж меток) Пусть G – pq-грамма с узлами $V(G) = \{v_1, \dots, v_{p+q}\}$, где $v_i - i$ узел в начальной сортировке. Кортеж $l(G) = (l(v_1), \dots, l(v_{p+q}))$ называется кортежем меток дерева G.

Определение 5(pq-грамм профиль) Для $p > 0, q > 0$, pq-грамм профиль, $P^{p,q}(T)$, дерева T – набор кортежей меток $l(G_i)$ всех pq-грамм G_i дерева T.

Определение 6(pq-грамм расстояние) Для $p > 0, q > 0$, pq-грамм расстояние, $\Delta^{p,q}(T_1, T_2)$, для деревьев T_1, T_2 вычисляется по следующей формуле:

$$\Delta^{p,q}(T_1, T_2) = 1 - 2\frac{|P^{p,q}(T_1) \cap P^{p,q}(T_2)|}{|P^{p,q}(T_1) \cup P^{p,q}(T_2)|}$$

Pq-грамм расстояние равное единице, если два дерева не имеют одинаковых pq-грамм. Деревья с расстоянием равным нулю имеет одинаковые pq-грамм профили. Следует отметить, что нулевое расстояние не подразумевает эквивалентности деревьев. Pq-грамм расстояние может быть вычислено за $O(n \log(n))$ времени, нужное для вычисления количества общих pq-грамм в профилях размера $O(n)$. Существует оценка, которая показывает, что размер pq-грамм профиля равен $|P^{p,q}(T)| = 2l + qi - 1$, где l - количество листов, а q – количество узлов, не являющихся листами [3].

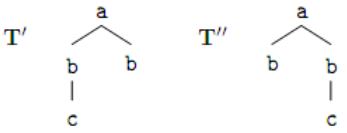

Рисунок 4 – Неравные деревья, имеющие одинаковый pq-грамм профиль

Рассмотренная мера может быть применена для приближенной оценки сходства деревьев. Такая оценка может понадобиться в задачах, пример которой приведен в начале статьи. Также эта оценка, благодаря скорости своей работы, может использоваться на предварительном этапе сравнения деревьев, при нахождении редакционного расстояния.

Список использованных источников:

1. K.-C.Tai. Tree to tree correction problem. Journal of the ACM (JACM), 1979
2. K.Zhang, D.Shasha. Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput. 1989.
3. N. Augsten, M.Bochlen, J.Gamper. Approximate matching of hierarchical data using pq-grams. VLDB. 2005