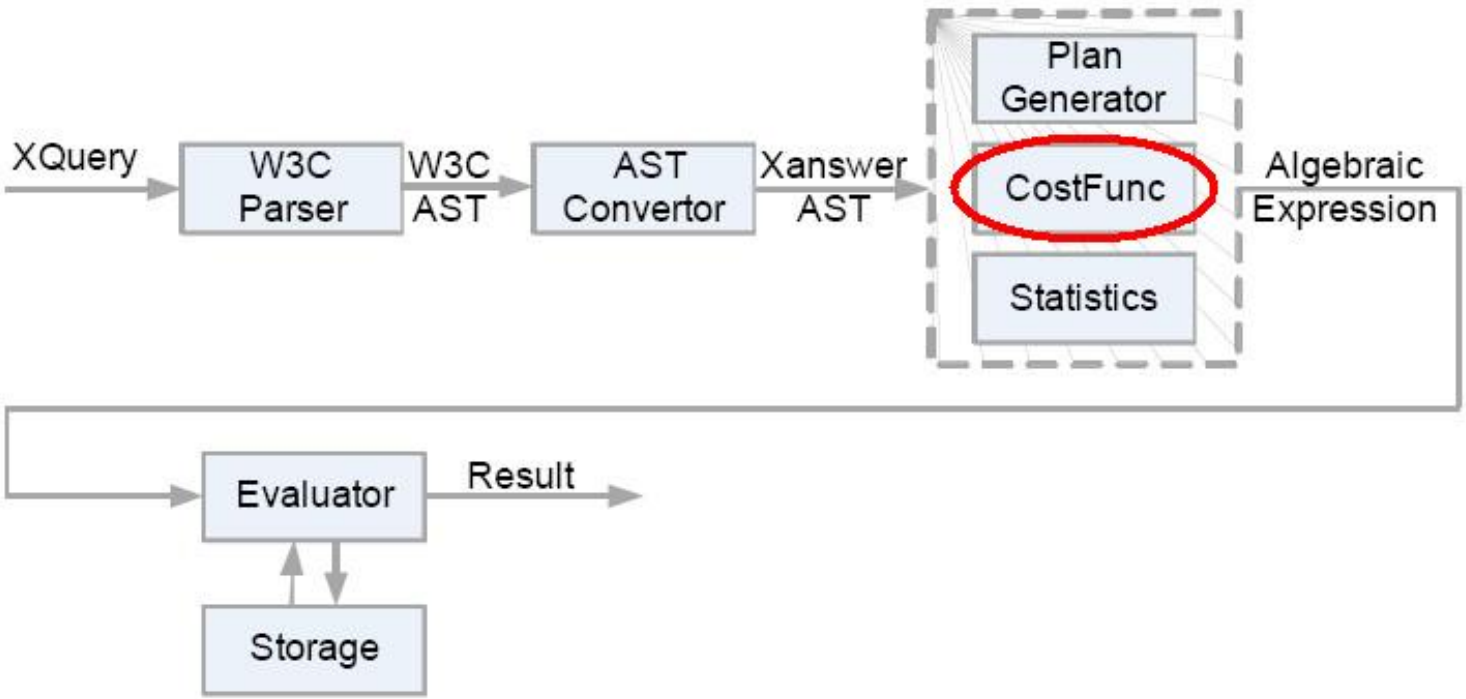
Стоимостная модель оптимизации вычисления XPath-выражений

Алексей Богатов

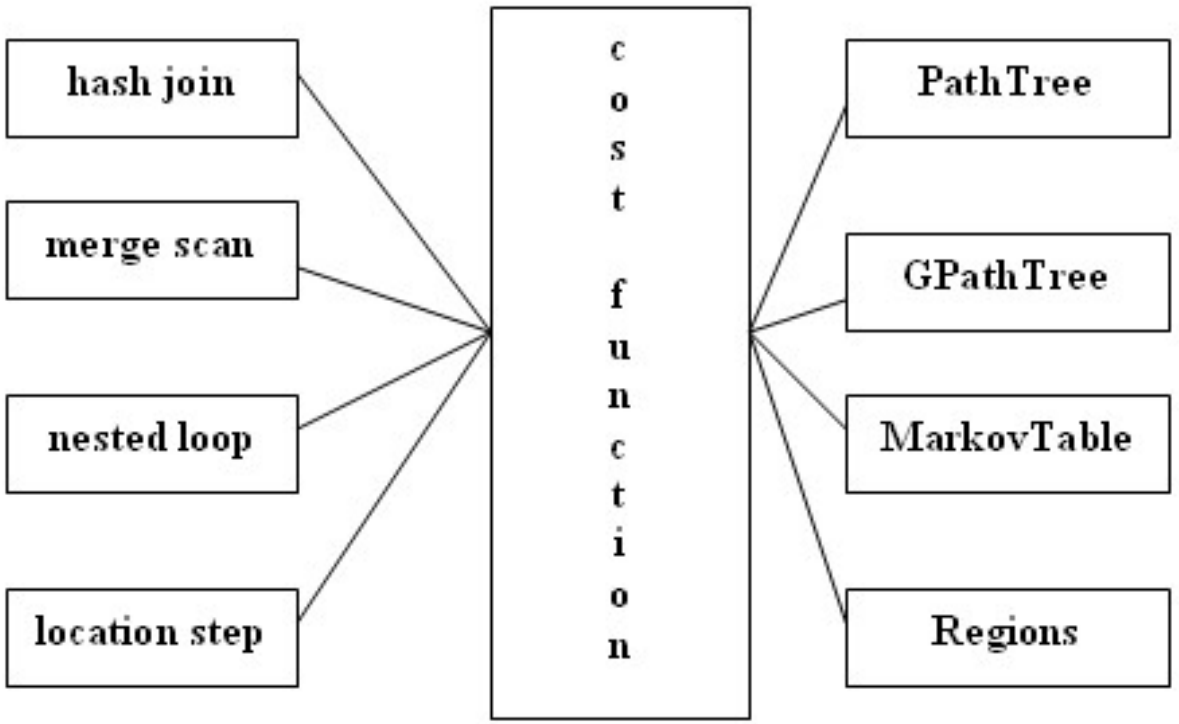
Научный руководитель: д. ф.-м. н. Б. А. Новиков

14 июня 2007 г.

Как устроен XAnswer



Место функции стоимости в проекте

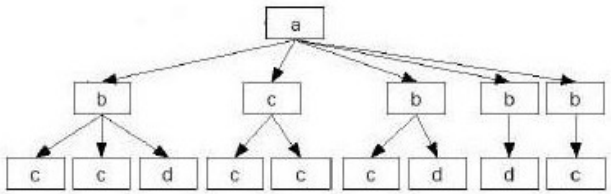


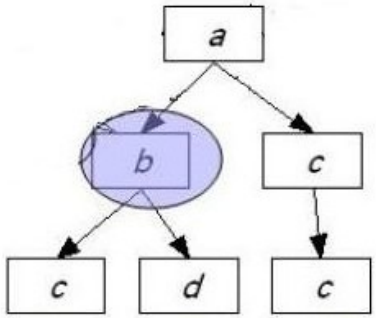
Постановка задачи

- Реализовать недостающие физические операции
- Реализовать статистическую структуру MarkovTable
- **Разработать метод оценки стоимости**

Статистические структуры

- PathTree





- MarkovTable

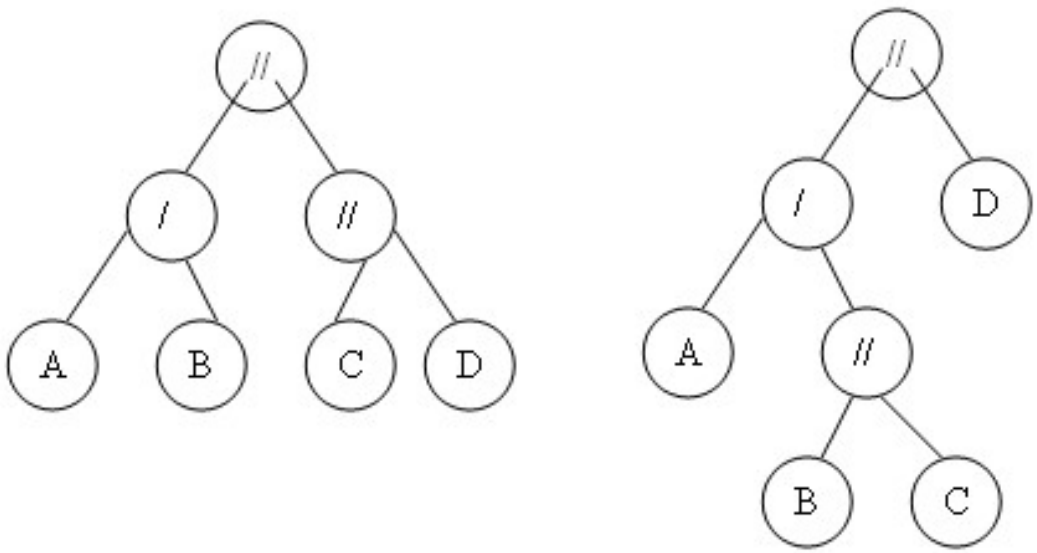
  Храним селективность путей до длины m (обычно 2 или 3).

$$S(A_1/\ldots/A_n) = S(A_1/\ldots/A_m) \times \prod_{i=1}^{n-m} \frac{S(A_{i+1}/\ldots/A_{i+m})}{S(A_{i+1}/\ldots/A_{i+m-1})}$$

$$S(A/B/C/D/E) = S(A/B/C) \cdot \frac{S(B/C/D)}{S(B/C)} \cdot \frac{S(C/D/E)}{S(C/D)}$$

Структурные соединения

A/B//C//D



Стоимость дерева есть стоимость операции в корне плюс сумма стоимостей поддеревьев.

Селективность

Марковские таблицы:

1. $S(X)$ храним
2. $S(X/Y)$ храним
3. $S(X//Y) = S(X/Y) + \sum_{Z_1} S(X/Z_1/Y) + \ldots + \sum_{Z_1,\ldots,Z_n} S(X/Z_1/\ldots/Z_n/Y)$, где n – настраиваемый параметр
4. $S(P_1/P_2) = S(P_1) \cdot S(P_2) \cdot \frac{S(A_1/A_2)}{S(A_1)S(A_2)}$, где $P_1 = \ldots/A_1,\ P_2 = A_2/\ldots$ (простейший вариант)
5. $S(P_1//P_2) = \sum_{i=1}^{n}(\sum_{Z_1,\ldots,Z_i} S(P_1/Z_1/\ldots/Z_n/P_2))$

Дерево навигационных выражений:
суммирование по соответствующим элементам, в случае descendant-or-self – обход поддеревьев

Реализация функции стоимости

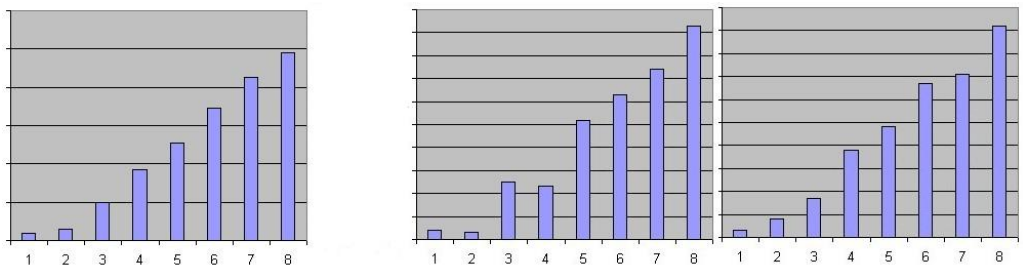
Учитывается лишь время выполнения операций чтения.
Мы считаем стоимость выполнения операции
зависящей только от вида операции и селективности операндов.

- ChildHash/ParHash: $n + m$
- DescHash/AncHash: оцениваем среднее число предков узла (используя сумму по уровням PathTree)
- Nested Loop: $n \cdot m$ для PC; AD – учет числа уровней
- Merge Scan: сортировка, если последовательность не отсортирована ($n \log n + m \log m$),
  дальше оцениваем средний размер поддерева элемента из внешней последовательности

Эксперименты

Качество функции стоимости:

- Генерируем физические планы выполнения XPath-выражения
- Выполняем планы, замеряем время
- Считаем стоимость выполнения плана с селективностью, вычисленной по MarkovTable или PathTree



Результат: приемлемая точность

Скорость вычисления стоимости

Результат: с использованием PathTree – точнее, но значительно медленнее, чем для MarkovTable