Модель хранения статистики о навигационных выражениях для составных XML документов в контексте мультибаз данных

© Юрий Солдак

Санкт-Петербургский Государственный Университет
математико-механический факультет
ysoldak@acm.org

Аннотация

В работе предлагается структура (ДНВ-СД*) для хранения статистической информации о навигационных выражениях (XPath) в контексте XML мультибаз данных. Обсуждается использование метода обратной связи для обновления структуры и исследуются ее свойства.

1 Введение

В последнее время активно ведутся исследования в области XML баз данных. Много результатов получено, существуют реализации XML СУБД (в том числе коммерческие), будь то надстройки над реляционной моделью (MSSQl, Oracle и пр.) или написанные «с нуля» системы (eXist, Ipedo, MarkLogic, Sedna, X-Hive и пр.). При этом исследователи и разработчики, на наш взгляд, уделяют недостаточно внимания вопросам организации распределенных хранилищ XML данных, предпочитая обсуждать проблемы локального характера. Справедливости ради стоит отметить, что в локальном случае все еще существует много нерешенных проблем и разногласий.

Существует не так много работ в области распределенных хранилищ XML данных. Исследования либо ведутся в направлении распределенных реляционных систем, либо посвящены локальным XML хранилищам. Следует выделить две работы повлиявшие на данное исследование. В первой из них [1] предлагаются два подхода к оценке селективности простых навигационных выражений (XPath [6]) для больших объемов XML данных: (а) подход, основанный на использовании дерева навигационых выражений (ДНВ или Path Tree); (б) подход на основе использования марковских таблиц. Идея использования ДНВ (см. раздел 3.1) в качестве основы для статистической структуры взята из этой работы. В другой работе [3] представлена техника XPathLearner, основанная на использовании метода обратной связи для обновления статистических данных. В обеих работах обсуждается вопрос сбора, хранения и обновления статистических данных без учета фрагментации данных.

В нашей работе представлена структура, сохраняющаяя информацию о фрагментации данных и дающая возможность оценить не только селективность обработки навигационного выражения, но и стоимость выполнения этой операции.

Работа организована следующим образом. В разделе 2 дается определение составного XML документа и обозначается цель работы. Предлагаемая статистическая структура представлена в разделе 3. Затем следует обсуждение экспериментальных результатов (раздел 4). И, наконец, раздел 5 содержит выводы и обзор направлений дальнейших исследований.

2 Цели работы

Насколько нам известно, ни одна из существующих на данный момент XML СУБД не позволяет прозрачно работать с составными XML документами. Под составным документом мы понимаем документ, состоящий из некоторого количества частей распределенных между несколькими физическими (обычными) документами. Составной XML документ возникает из обычного путем добавления в последний конструкции XInclude [4] и/или XLink [5] (см. листинг 1). Составной документ может быть частью другого составного, таким образом глубина рекурсивного включения частей не ограничена. Единственным естественным ограничением на структуру является отсутствие циклических включений.

Работа с составными документами является прозрачной когда СУБД обеспечивает прозрачное выполнение запросов на таких документах. При этом прозрачная обработка навигационных выражений здесь играет ключевую роль. Прозрачность выполнения запросов и обработки выраже-

Труды 8 Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

```
<company
xmlns:xi="http://www.w3.org/2001/XInclude"
xmlns:xl="http://www.w3.org/2000/xlink"
>
<name>The very big company</name>
<staff>
<office id="main">
<person position="CEO" office="main">
<name>John Smith</name>
<email>jsmith@company.com</email>
</person>

<xi:include xi:href="/db/RnD.xml" xi:xpointer="element(/persons/person)" />
<xi:include xi:href="/db/QA.xml"  xi:xpointer="element(/persons/person)" />

<xl:link xl:type="simple" xl:href="/db/managers.xml#xpointer(//person)" />
</office>

<office id="o1">
<xi:include xi:href="http://o1.company.com/db/staff.xml"
xi:xpointer="element(/staff/person)" />
</office>

<office id="o2">
<xi:include xi:href="http://o2.company.com/hr.xml"
xi:xpointer="element(/staff/persons/person)" />
</office>

</staff>
</company>
```

Рис. 1. Пример составного XML документа

ний понимается здесь в классическом смысле, т.е. детали фрагментации данных полностью скрыты от пользователя, что дает ему возможность формулировать запросы и писать навигационные выражения в привычном виде, совершенно не заботясь о структуре включений документа.

Составные XML документы интересно рассматривать в контексте распределенных баз данных. В этом случае части одного составного документа могут храниться на разных узлах сети. В этих условиях особый интерес представляет информация о фрагментации обрабатываемого документа, стоимость соединения и пересылки данных с удаленных узлов, а также селективность навигационных выражений, выбирающих элементы на удаленных узлах хранилища. Обладая статистической информацией такого типа, оптимизатор может правильнее оценить стоимости выполнения планов запросов и, в конечном итоге, выбрать лучший из них.

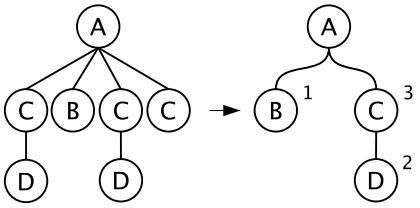

Рис. 2. Получение ДНВ из XML документа

Целью работы является разработка и исследование статистической структуры для хранения информации, необходимой для эффективной обработки навигационных выражений на составных XML документах в контексте мультибаз данных.

Контекст исследования сужен на класс мультибаз данных намеренно. Разрабатываемая структура не должна зависеть от тонкостей реализации узлов системы. Это дает надежду на то, что результаты исследований можно будет использовать при построении систем в условиях слабой связанности на основе глобальной сети Интернет. XML выбран в качестве языка представления и обмена данными также не случайно. Мы ожидаем, что в ближайшем будущем обмен текстовыми данными в Интернет будет происходить преимущественно в XML формате.

3 Статистическая структура

3.1 Дерево навигационных выражений

Основой нашей статистической структуры является Дерево Навигационных Выражений(ДНВ или Path Tree, рис. 2). ДНВ — это адаптация (для слабоструктурированных документов) широко известной структуры «Путеводитель по Данным» (DataGuide) [2]. В ДНВ каждый узел структуры однозначно задает навигационное выражение, состоящее из шагов в направлении child, и наоборот.

3.2 ДНВ для составных документов

При построении статистической структуры для составных документов, желательно сохранить информацию о точках включения частей документа. Понятно, что стоимость выполнения шагов навигационного выражения в пределах одного физического документа и соединяющих элементы документов, хранящихся на удаленных друг от друга серверах, может сильно отличаться. Структура должна давать возможность быстро и точно определять точки соединения частей документа и оценивать стоимость выполнения шага “сквозь” это соединение. Для решения этой задачи мы предлагаем снабжать дуги ДНВ весами. Так для дуг в пределах одного физического документа вес будет мал (например, равен 1), а для дуг, которые связывают элементы разных физических документов — больше (например, 10). Если же связываемые посредством дуги элементы принадлежат различным узлам мультибазы данных, то вес будет еще больше (например, 1000). Если первые два типа весов скорее константы (могут задаваться в файле настроек), то веса для меж-серверных дуг имеет смысл вычислять динамически. Реальные значения зависят от многих факторов, как то скорость соединения между узлами, загруженность магистрали и пр.

Полученную структуру предлагается называть ДНВСД (ДНВ для Составных Документов). Построение ее происходит следующим образом: сначала строится ДНВ для сериализованного составного документа, затем каждой дуге присваивается вес, в зависимости от состояния хранилища и целей создания статистической структуры.

Контекст мультибаз данных накладывает ограничения на возможности получения статистической информации о данных на удаленных узлах системы. Мы используем метод обратной связи (получение необходимой статистической информации из (частичного)результата(ов) обработки

запроса) для построения и обновления структуры (см. раздел 3.3). В этом случае, кроме дуг отец-сын (соответствующих направлению child в XPath) нам понадобилось описывать отношение предок-потомок между узлами структуры (направление descendant в XPath). Для этого мы вводим понятие «обобщенной» дуги и помечаем такие дуги знаком «*» на схемах. Таким образом, ДНВ* и ДНВСД* — это ДНВ и ДНВСД соответственно, в которых допускается наличие обобщенных дуг. ДНВ* можно рассматривать как «частичный» ДНВ, так как каждая обобщенная дуга, фактически, скрывает за собой набор обычных (точных) дуг.

На рис. 3 можно видеть пример структуры ДНВСД* для составного документа, распределенного на 4 узла мультибазы.

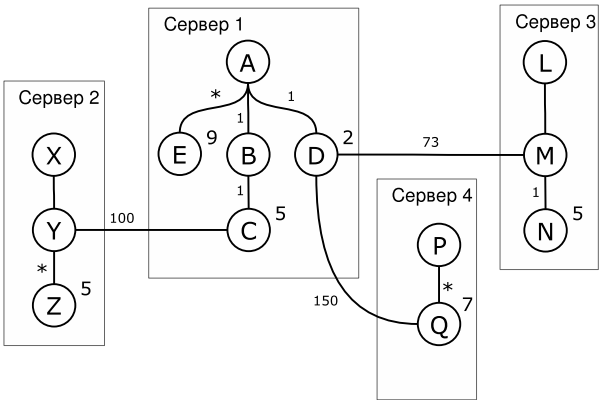

Рис. 3. ДНВСД*

Далее в статье обсуждаются вопросы, связанные с построением и поддержкой актуальности ДНВСД*. Так как ДНВСД* по-сути является объединением двух других понятий — ДНВ* и ДНВСД, то при обсуждении обобщенных дуг мы будем говорить о ДНВ*, а в случае обсуждения весов дуг — о ДНВСД.

3.3 Построение ДНВСД* методом обратной связи

Метод обратной связи позволяет строить/обновлять ДНВСД* используя результаты выполнения пользовательских запросов. Работа системы обновления состоит из двух последовательных этапов: 1) получение статистической информации из пользовательского запроса и 2) обновление структуры на основе полученной информации. Этап получения информации, в свою очередь, состоит из выделения навигационных выражений из тела запроса и извлечения соответствующих статистических данных (например, селективности) в процессе его выполнения.

Теоретически, использование такого подхода может привести к построению структуры с информацией обо всех возможных навигационных выражениях для документа. Практически же, объем собранной статистической информации сильно зависит от набора запросов, используемых в приложении. Кроме того, можно ожидать, что значения селективности чаще будут определены для листьев и реже для промежуточных узлов.

По способу обработки все навигационные выражения можно разделить на локальные или распределенные. В первую группу попадают выражения, не проходящие ни через одну точку соединения физических частей составного документа. Кроме этого выражения могут быть либо точными (все шаги в направлении child), либо обобщенными (хотя бы один шаг в направлении descendant). Комбинации этих характеристик дают 4 типа навигационных выражений. Рассмотрим особенности их обработки.

Локальные точные выражения обрабатываются тривиально: если ветви, соответствующей выражению, нет в структуре — она добавляется, в противном случае обновляется значение селективности в конечном узле. Веса всех дуг ветви будут иметь минимальные значения.

3.3.1 Обработка распределенных выражений

Допустим, для того, чтобы выполнить очередной шаг точного навигационного выражения требуется получить подключаемую часть составного документа. Обработчик должен запомнить момент наступления этого события для того, чтобы после получения необходимого фрагмента иметь возможность вычислить вес дуги, соответствующей обрабатываемому шагу.

Пусть подключаемый фрагмент принадлежит документу, расположенному на удаленном узле мультибазы. В этом случае обработчику будет необходимо сформировать подзапрос на языке XQuery к удаленному узлу. Подзапрос может быть сформирован таким образом, чтобы получить (а) результат выполнения оставшейся части исходного навигационного выражения либо (б) весь подключаемый фрагмент локально. В первом случае обработка оставшейся части выражения делегируется удаленному узлу, во втором случае с удаленного узла выгружается весь фрагмент и дальнейшая обработка выражения происходит локально. У каждого из подходов есть свои преимущества и недостатки. Выбор наиболее подходящего из них является примером задачи оптимизатора, в решении которой может помочь разрабатываемая структура.

В том случае, если подключаемый фрагмент находится на локальном узле мультибазы, обработчику не обязательно формировать XQuery подзапрос — ожидается, что в этом случае реализацию можно сделать эффективнее.

После того, как результат обработки навигационного выражения получен, структура обновляется аналогично случаю обработки локального точного выражения, за исключением того, что вес дуги, соединяющей узлы ДНВСД, соотвествующие разным физическим документам, вычисляет-

ся динамически.

Если при обработке шага в направлении descendant понадобилось обработать подключаемый фрагмент, то такому шагу в ДНВСД* будет соответствовать не одна обобщенная дуга, а цепочка из обобщенной, точной и снова обобщенной дуг. Здесь точная дуга будет хранить вычисленный вес и явно задавать точку связи фрагментов составного документа.

3.3.2 Обработка локальных обобщенных выражений

Все рассматриваемые в этом разделе выражения являются локальными, поэтому рассматриваемой структурой является ДНВ*.

При обработке локальных обобщенных выражений возникают две проблемы: дублирование данных и распределение селективности между несколькими узлами структуры.

Наличие обобщенных дуг может привести к дублированию данных в структуре. Известно, что один и тот же набор узлов XML документа часто можно выбрать при помощи различных навигационных выражений. Будем называть такие выражения эквивалентными. Нас интересуют эквивалентные выражения с различной степенью точности. Наиболее точным будет выражение у которого все шаги заданы направлением child, наименее точным — выражение состоящее из единственного шага в направлении descendant. По такому же принципу определяется точность соответствующих ветвей ДНВ*. При обновлении структуры мы следуем принципу наибольшей точности, т.е. менее точные ветви всегда заменяются более точными. Понятно, что замена обобщенной дуги на цепочку обычных дуг не всегда будет эквивалентной операцией. Здесь мы сознательно делаем допущение об эквивалентности с целью сохранить контроль над размером и поведением структуры в ущерб точности.

Вторая упомянутая проблема возникает при обработке навигационного выражения, которому соответствует несколько узлов ДНВ*. Пусть $\dot{S}$ и S — реальная и приближенная (хранимая в ДНВ*) селективности обобщенного выражения соответственно. $\dot{S}$ получена из результата обработки запроса, а $S = \sum_{i=1}^{n} s_i$, где s_i — приближенные селективности для узлов ДНВ*, соответствующих обобщенному выражению, $i = 1 : n$ — номер узла, а n — число узлов.

Не умаляя общности будем считать, что $D > 0$, где $D = \dot{S} - S$. Возникает проблема распределения остатка селективности D между n узлами. Предположив, что изменения селективностей узлов равновероятны (в смысле частоты изменений и величин на которые они изменяются), можно предложить простейший способ распределения D (*равноправный*):

$$s_i := s_i + D/n$$

Ясно, что после распределения остатка реальная и приближенная селективности обобщенного выражения будут равны. При этом могут пострадать приближенные селективности для узлов.

3.4 Устойчивость ДНВ*

Определение ДНВ* *точна* тогда и только тогда, когда в каждом её узле содержится реальное значение селективности соответствующего навигационного выражения.

Для определения степени точности структуры мы используем следующую формулу:

$$A = \sum_{i=1}^{n} |d_i|$$

где n — количество узлов в ДНВ*, $d_i = \dot{s}_i - s_i$ — разница между реальным и приближенным значением селективности для узла с номером i. Согласно определению, ДНВ* точна тогда и только тогда, когда $A = 0$. Понятно, что разбиение множества узлов ДНВ* на набор дизъюнктных множеств и вычисление A для каждого из них с последующим суммированием является эквивалентым способом определения степени точности ДНВ*.

*Устойчивость ДНВ** означает, что структура стремится быть точной независимо от вида и величины первоначального возмущения значений в ее узлах.

Предложение 3.1 *Пусть есть ДНВ* у которой некоторые узлы содержат неверные значения селективности ($A \neq 0$). Предположим, что:*

- *исходный документ не изменяется,*
- *распределение остатка происходит с использованием равноправного подхода,*
- *вероятность выполнения точного обновления любого из узлов ДНВ* больше 0.*

Тогда $A \downarrow 0$.

Доказательство Ясно, что все типы обновлений, кроме распределения остатка, либо уменьшают A, либо оставляют его наизменным. $A = 0$ достигается, например, после выполнения навигационных выражений точно соответствующих каждому из узлов ДНВ*. То, что для всех узлов такие выражения выполняются, гарантируется предположением о вероятностях в условии предложения. Осталось показать, что те части ДНВ*, которые состоят из набора узлов, соотвествующих обобщенным выражениям, также стремятся быть точными.

Пусть есть обобщенное выражение и I — набор индексов тех узлов, которые этому выражению соответствуют. $I_0 \subseteq I$ — индексы узлов, в которых хранятся точные значения селективностей ($\forall i \in I_0 \quad d_i = 0$). Символом A будем обозначать степень точности исследуемой части ДНВ*.

Покажем, что $A^{k+m} \leq A^k$ если в моменты k и $k+m$ $(m > 1)$ происходила обработка обобщенного выражения, а между этими событиями обобщенное выражение не обрабатывалось.

Ясно, что

$$A^{k+m-1} \leq A^k$$

Пусть

$$D^{k+m-1} = 0$$

тогда

$$A^{k+m} = A^{k+m-1} \leq A^k$$

Если

$$D^{k+m-1} \neq 0$$

тогда

$$|d_i^{k+m}| \leq |d_i^{k+m-1}| + |\frac{D^{k+m-1}}{n}|$$

$$A^{k+m} \leq A^{k+m-1} + |D^{k+m-1}|$$

и учитывая, что

$$A^{k+m-1} = A^k - \sum_{i \in I_0} |d_i^k|$$

имеем

$$A^{k+m} \leq A^k - \sum_{i \in I_0} |d_i^k| + |D^{k+m-1}|$$

Осталось показать, что

$$|D^{k+m-1}| \leq \sum_{i \in I_0} |d_i^k|$$

Действительно

$$D^k = 0 \Rightarrow \sum_{i \in I \setminus I_0} d_i^k = -\sum_{i \in I_0} d_i^k$$

$$|D^{k+m-1}| = |\sum_{i \in I \setminus I_0} d_i^k| = |\sum_{i \in I_0} d_i^k| \leq \sum_{i \in I_0} |d_i^k|$$

Получаем

$$A^{k+m} \leq A^k$$

∎

Условие на неизменность документов в БД необходимо для строгого доказательства. В действительности достаточно, чтобы документы не менялись очень часто оставляя время для стабилизации структуры (см. раздел 4.1).

4 Эксперименты

С целью изучения свойств предлагаемой статистической структуры был проведен ряд экспериментов. Их описание и результаты содержатся в данном разделе. Так как действующего прототипа XML мультибазы на момент проведения экспериментов не было, изучить свойства ДНВСД не представлялось возможным. Поэтому экспериментальному исследованию подверглась другая базовая структура — ДНВ*. Далее в этом разделе обсуждаются свойства ДНВ* и вопросы, связанные с её точностью и скоростью стабилизации. Следует отметить, что все результаты, полученные для ДНВ*, также верны для ДНВСД*.

4.1 Скорость стабилизации

Важным свойством статистической структуры является точность содержащихся в ней данных. Поэтому большой интерес представляет исследование скорости и характера стабилизации, или, другими словами, скорости и характера уменьшения погрешности структуры. На рис. 4 приведены усредненные графики стабилизации частей ДНВ(СД)* для наборов из 2,3,4 и 5 навигационных выражений. Для каждого набора существует обобщенное навигационное выражение, при обработке которого может происходить распределение остатка. Значение погрешности на каждом шаге представлено в процентах от начальной.

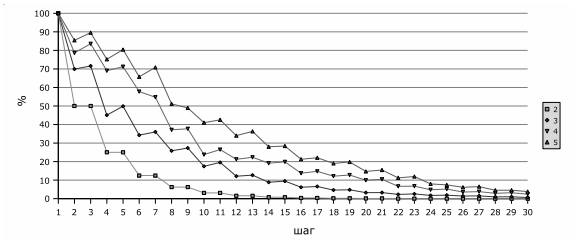

Рис. 4. Скорость стабилизации ДНВ(СД)*

Графики были получены при проведении следующего эксперимента. Имея набор навигационных выражений, строим статистческую структуру. Возмущая селективности навигационных выражений, добиваемся того, что статистическая информация устаревает. Затем, случайным образом выбирая точное навигационное выражение (вероятности выбора выражений равны), обновляем значение селективности для неё в статистической структуре. После этого соответствующий узел содержит точное значение селективности и погрешность уменьшается. После обработки точной ветви в эксперименте всегда выполнялось распределение остатка с использованием равноправного подхода (обработка обобщенного выражения). Эти шаги хорошо видны на графиках — значение погрешности остается неизменным, либо незначительно возрастает. Затем снова обрабатывалось точное выражение, выбранное случайным образом и т.д. Описанный сценарий обработки выражений является моделью поведения системы. В реальных условиях структура может стать точной очень быстро (за число шагов равное количеству точных выражений), может сохранять погрешность в случае, если на протяжении длительного времени обрабатывается только одно точное выражение. В конечном счете, все зависит от вероятностей появления выражений в обрабатываемых запросах.

Из графиков видно, что чем меньше выражений в наборе, тем быстрее стабилизируется соответствующая часть структуры. Также легко видеть, что погрешность уменьшается достаточно быстро и уже через 24 шага (погрешность для всех рассматриваемых наборов уменьшается более, чем в 10 раз) можно предположить, что остав-

шаяся погрешность не должна сильно влиять на качество выбора оптимального плана выполнения запроса. Вопрос о том, какую погрешность считать критичной для правильного выбора плана, на данный момент остается открытым и мы планируем ответить на него в следующих работах.

4.2 Поведение структуры в условиях изменяющихся исходных данных

Моделировались два вида изменения исходных данных: периодическое и постоянное. Под периодическим изменением мы понимаем ситуацию, когда селективности наблюдаемых навигационных выраженииий обязательно изменяются через равные промежутки времени, а между этими событиями селективности всегда остаются неизменными. Постоянное изменение данных подразумевает ситуацию, когда каждое навигационное выражение на каждом шаге может менять свою селективность с некоторой, заранее заданной, вероятностью. В обоих случаях использовалась модель вычисления новой селективности, описанная с следующем подразделе.

4.2.1 Вычисление новой селективности навигационного выражения

Опишем модель, примененную нами, для изменения селективности наблюдаемых навигационных выражений. Во-первых, мы полагаем, что небольшие изменения данных должны происходить чаще. Во-вторых, величина изменения ограничивается сверху. И, наконец, селективности могут как увеличиваться, так и уменьшаться. Ниже приведена формула вычисления (новой) селективности:

$$s_i = s_i + sign(p - 0.5) * Round(\frac{N}{N * p + 1}),$$

где s_i — селективность i-того навигационного выражения, $N = max_{i \in I}(\tilde{s_i}) * 10$ — максимально возможное изменение селективности, а $\tilde{s_i}$ — значение селективности i-того навигационного выражения на самом первом шаге, $p \in [0, 1)$ — равномерно распределенная по промежутку случайная величина.

4.2.2 Периодическое изменение исходных данных

Целью изучения этого случая является определение того, как структура будет себя вести при дефиците времени на стабилизацию.

На рисунке 5 показана типичная динамика относительной погрешности структуры при периодических возмущениях. Здесь число навигационных выражений равно 4, а возмущения происходят каждый 7 шаг, что хорошо видно по всплескам погрешности. Новые селективности для выражений вычислялись по формуле, приведенной

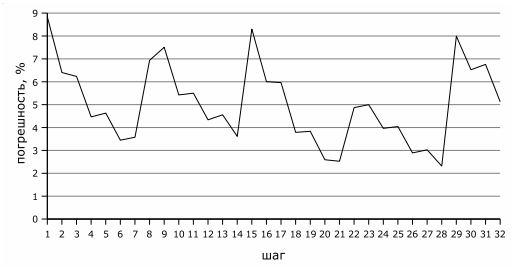

Рис. 5. Динамика погрешности для набора из 4х выражений и периодом изменения равным 7

выше. На каждом нечетном шаге структуре сообщалось о селективности обобщенного выражения, а на каждом четном — какого-либо из точных.

Рис. 6. Зависимость погрешности от частоты возмущений для набора из 4 выражений

Процесс увеличения средней величины погрешности при увеличении частоты возмущений показан на рисунке 6. Видно, что как только возмущения начинают происходить чаще, чем раз в 10 шагов структура стремительно теряет точность. Это объясняется тем, что структуре перестает хватать времени для стабилизации и хранимые селективности все больше отличаются от реальных значений.

4.2.3 Постоянное изменение исходных данных

Стремясь к улучшению модели поведения данных мы рассматриваем случай постоянного изменения селективности. На рисунке 7 представлены графики зависимости относительной погрешности структуры от вероятности изменения селективности навигационных выражений. Здесь можно видеть 4 графика для групп из 2,3,4 и 5 навигационных выражений. Графики получены следующим образом. Для каждого значения вероятности выполнялось по 30 запусков каждый длиной в 500 шагов. В начале каждого запуска строилось ДНВ* для исследуемого набора навигационных выражений. Затем, на каждом шаге, происходила попытка изменить селективность каждого из навигационных выражений в соответствии с вероятностью и по формуле, описанной выше. После этого, на нечетном шаге структуре сообщалось о селективности обобщенного выражения, а на каждом четном — какого-либо из точных. После обновления структуры вычислялось текущее значение погрешности. Затем, имея наборы зна-

чений погрешности для каждой вероятности (для 30 запусков по 500 шагов), вычислялись их средние арифметические значения, представленные на графике. Как видно из рисунка, скорость паде-

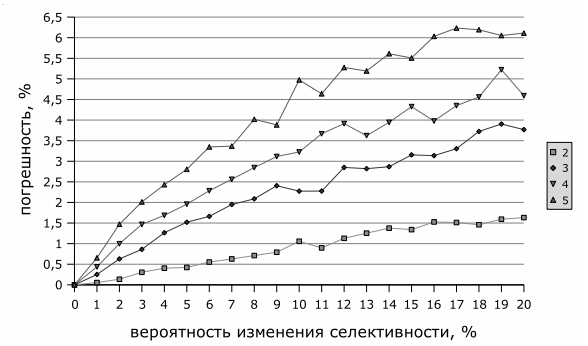

Рис. 7. Зависимость погрешности структуры от вероятности изменения селективности наблюдаемых навигационных выражений

ния точности постепенно замедляется с ростом вероятности изменений и достигает 6% для набора из 5 выражений при 20% вероятности изменений. Это, на наш взгляд, является достаточно хорошим показателем. Небольшие отклонения в значениях погрешности от легко видимой зависимости возникают случайным образом и не влияют на общую картину.

5 Заключение

В работе рассматривается вопрос об эффективном прозрачном выполнении запросов на составных XML документах в контексте мультибаз данных. Проблема оценки селективности навигационных выражений, а также стоимости их обработки является одной из ключевых для эффективного выполнения запросов в данном контексте. Предлагается расширяемая структура (ДНВСД*) для хранения необходимой статистической информации. Отличие предлагаемого подхода от близких решений заключается в сохранении информации о фрагментации составного документа, что позволяет точнее оценивать стоимость обработки навигационного выражения. Кроме того, структура адаптирована к обновлению с использованием метода обратной связи. Сформулирован критерий точности структуры и доказана ее устойчивость. Проведено экспериментальное исследование свойств структуры.

Дальнейшая работа может развиваться в нескольких направлениях: создание прототипа оптимизатора запросов для составных XML документов с использованием предлагаемой структуры хранения статистики; исследования свойств ДНВСД; поиск возможности хранить информацию о навигационных выражениях, содержащих шаги в направлениях following и preceding и/или предикатные конструкции.

6 Благодарности

Выражаю благодарность своему научному руководителю Новикову Борису Асеновичу за помощь в подготовке статьи.

Список литературы

[1] Ashraf Aboulnaga, Alaa R. Alameldeen, and Jeffrey F. Naughton. Estimating the selectivity of XML path expressions for internet scale applications. In *The VLDB Journal*, pages 591–600, 2001.

[2] Roy Goldman and Jennifer Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases*, pages 436–445. Morgan Kaufmann, 1997.

[3] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Parr. XPathLearner: An On-line Self-Tuning Markov Histogram for XML Path Selectivity Estimation. In *VLDB*, pages 442–453, 2002.

[4] XML Inclusions (XInclude) Version 1.0, 20 December 2004. W3C Recommendation.

[5] XML Linking Language (XLink) Version 1.0, 27 June 2001. W3C Recommendation.

[6] XML Path Language (XPath) Version 1.0, 16 November 1999. W3C Recommendation.

A model of statistical structure for path expressions on composite XML documents in multidatabase context

Yury Soldak

A new statistical structure based on PathTree notion and named PTCD* (PathTree for Composite Documents with support of descendant axis) is proposed in this paper. This structure is developed to store statictical information about XPath expressions in multidatabase context. A notion of Composite XML Document is discussed and multidatabase model based on this notion is proposed. The PTCD* can be used by an xml query plan optimizer in order to select best plan, since information about data fragmentation as well as path expression selectivities can easily be obtained from it. Suggested structure can be updated using feedback approach, so gracefully feets multidatabase node's communication abilities. It is shown, what structure tends to be accurate in terms of stored statistical information. Some PTCD* properties are studied on experimental phase of the research: structure size compared to corresponding PathTree size, speed and type of structure stabilization process (a process of becoming accurate after some initial disturbance of observing data).