УДК 004.652.5

# ПОСТРОЕНИЕ БАЗ ДАННЫХ СО СВОБОДНОЙ СТРУКТУРОЙ (МОДЕЛЬ RDF), ПРИМЕНЕНИЕ ИХ В МОБИЛЬНЫХ УСТРОЙСТВАХ

*Меренков А.В[1], Бродт А.[2], Митчанг Б.[2]*
*[1]Донецкий национальный технический университет*
*[2]Штутгартский университет IPVS AS*

В настоящее время происходит настоящий информационный бум, возрастает роль высоких технологий в нашей повседневной жизни, что влечет за собой рост информационной нагрузки, как на человека, так и на всю компьютерную отрасль. Этому способствует увеличение объемов информации, усложнение её структуры, способов хранения, обработки и обмена.

В настоящее время важной задачей является разработка алгоритмов, методов и технологий хранения, обработки и обмена информацией. Все эти методы должны соответствовать современным реалиям. Таким образом, они должны удовлетворять следующим критериям: быть достаточно открытыми; простыми в использовании для простого пользователя; легко переносимыми; ускоряющими и упрощающими обмен информацией; иметь возможность параллельной обработки.

Одной из таких моделей является модель RDF(Resource Description Framework). В ходе научного сотрудничества со Штутгартским Университетом, была принята технология RDF-3x engine, которая разрабатывается под руководством Т.Ноймана (Институт М.Планка) [1].

## Описание модели RDF(Resource Description Framework) и возможностей RDF-3x Engine

RDF - это разработанная консорциумом W3C[2] модель для описания ресурсов, в особенности - метаданных о ресурсах. Эта

модель является форматом представления данных со свободной структурой, которая набирает силу в Semantic-Web Corporation, естественных науках, а также разработках на базе платформы Web 2.0.

В основе этой модели лежит идея об использовании специального вида утверждений, высказываемых о ресурсе. Каждое утверждение имеет вид «субъект — предикат — объект» (eng. Subject — Predicate — Object или SPO) и в терминологии RDF называется триплетом(eng. Triple). Например, информация о Факультете компьютерных наук и технологий может выглядеть следующим образом, где первое значение в строке — субъект, второе — предикат, а третье — объект:

(id1; <имеетНазвание>; «Факультет компьютерных наук и технологий»),

(id1; <тип>; «факультет»),

(id1; <основан>; «1967»),

(id1; <возглавляем>; «Аноприенко А.Я.»),

(id1; <имеетКафедру>; id2),

(id1; <имеетКафедру>; id3),

(id1; <имеетКафедру>; id4),

(id2; <имеетНазвание>; «Кафедра ЭВМ»),

(id2; <тип>; «кафедра»),

(id2; <имеетНазвание>; «Кафедра ПМИ»),

(id2; <тип>; «кафедра»)....

Одной из главных целей RDF является предоставление утверждений одинаково в машинном и удобном для распознавания человеком виде. Из примера видно, что в таком виде информацию можно легко прочитать как человеку, так и обработать машине. Однако это не единственный синтаксис для представление RDF-данных. Самые распространенные: RDF/XML, триплеты (или Нотация 3) и графовая модель. Для машинной обработки существуют также специальные языки запросов: например, RQL, RDQL, SPARQL[3].

Можно обратить внимание, что, хотя имена предикатов,

такие как «основан» или «имеетНазвание» напоминают атрибуты, как таковой схемы базы данных не существует; в той же базе данных могут содержаться триплеты с предикатами «основан_В_Году». По сравнению с моделью «сущность - отношение», как атрибуты сущности так и её отношения к другим сущностям представлены предикатами. Таким образом, всё множество RDF триплетов вместе можно рассматривать как большой граф, т.е. для его обработки, хранения и изменения можно использовать теорию графов, а также уже известные алгоритмы.

Язык запросов SPARQL является стандартным для поиска в RDF хранилищах. Он поддерживает союзы(конъюнкции и дизъюнкции) тройных структур, что аналогично запросам на выборку-объединение(select-join) в реляционной модели. Например, мы можем получить список кафедр ФКНТ следующим запросом:

```
Select ?title where{
?id1  < имеетНазвание > “Факультет компьютерных наук и технологий”;
<имеетКафедру> ?id2.
?id2 <тип> ”кафедра”; < имеетНазвание > ?title. }
```

Здесь каждая конъюнкция обозначается точкой и соответствует объединению(join).Точка с запятой означает дизъюнкцию, т.е пересечение. В SPARQL предикаты также могут быть переменными или регулярным выражением, что позволяет выполнять запросы не зависящие от схемы.

Исследуемая разработка RDF-3x Engine является реализацией SPARQL, которая достигает отличных результатов. Основными особенностями RDF-3x являются:

- это общее решение для хранения и индексирования RDF триплетов, что полностью исключает необходимость для настроек на физическом уровне;
- мощный, но в тоже время простой обработчик запросов, который использует быстродействующие операции

объединений(merge joins) максимально возможным способом;
- оптимизатор запросов, который определяет оптимальный порядок объединений, используя модель стоимости операций, основанную на статистических данных для всех операций объединения;
- открытый код и переносимость проекта, что позволяет использовать проект в различных ОС, проект исследовался под системами Windows, Linux(как общего назначений – Debian, Ubuntu – так и специального назначения для мобильных устройств – Maemo Linux v5.2 на базе устройства Nokia N800).

**Использование баз данных(RDF-3x Engine) в мобильных устройствах**

В последние года имеет место тенденция перехода многих технологий и проектов со стационарных компьютеров на мобильные устройства, будь то ноутбуки, мобильные аппараты или коммуникаторы. Но этот переход связан со многими особенностями и сложностями, так как различаются архитектуры устройств, оборудование и так далее. Таким образом, переход на мобильные устройства требует исследований и решения возникающих проблем.

Множество мобильных устройств становятся все более функциональными, усложняются операционные системы, под управлением которых они работают. Появилась тенденция переноса Linux систем на мобильные устройства. Это дает возможность использовать огромный потенциал этой системы применительно мобильных устройств, обеспечивается открытость и переносимость программ, многозадачность и так далее. Таким образом, появляется возможность проектировать различные приложения, которые могут быть намного сложнее, чем было доступнее раньше. Это могут быть и базы данных в том числе. Например: база данных на основе RDF

модели, которая будет хранить всю информацию о пользователе, контактах, звонках, картинках или видео, встречах и так далее, при этом не нужно создавать или менять структуру базы данных каждый раз, при добавлении новой информации, возможность обмена между пользователями и устройствами и многое другое.

Еще одной из особенностей мобильных устройств является использование флэш-памяти, что вносит некоторые особенности. Этот тип памяти имеет меньшую скорость чтения, чем у жестких дисков, но у флэш-памяти отсутствуют движущиеся детали, это повышает надежность, а также позволяет делать выборку информации по непоследовательным адресам без потери скорости чтения (что происходит при использовании жестких дисков, из-за перемещения головок по поверхности). Это может быть полезно при чтении индексов баз данных, B-деревьев(B-trees). По этому нужно исследовать такую возможность, для этого нужно определить оптимальный размер страницы памяти, также время кэширования страниц в основную память.

Еще одна область исследований, это место флэш-памяти в системной архитектуре. Через несколько лет флэш-память будет использоваться для ликвидации разрыва между традиционной основной памятью и традиционными дисковыми устройствами во многих операционных системах, файловых системах и системах баз данных[4]. Флэш-память можно использовать для расширения основной памяти или постоянного (персистентного) хранилища.

Еще одна проблема при разработке приложений для мобильных устройств (в том числе и баз данных) – это относительно малый объем оперативной памяти и системных буферов. Эта проблема была исследована на примере RDF-3x Engine. В этом проекте реализована возможность буферизации файла базы данных в оперативную память. Эта возможность реализована на системном уровне, то есть оперативная система сама кэширует (отображает) различные участки файла в ОЗУ, что, несомненно, ускоряет работу. Однако если на стационарных компьютерах вопрос о размере ОЗУ не стоит так остро, то на мобильных устройствах объем ОЗУ

ограничен и четко фиксирован.

По этой проблеме были проведены исследования на мобильном устройстве(Nokia N800[5]). Устройство обладает 128МБ ОЗУ, что достаточно мало, учитывая объем занимаемый в ОЗУ операционной системой. По этому была произведена группа экспериментов, тестирующих как саму архитектуру и систему, так и взаимодействие с базой данных на базе RDF-3x Engine:

1. На скорость обмена между ОЗУ и флэш-памятью. Была использована утилита SysBench, проверены различные режимы чтения и записи;
2. Определен оптимальный размер обмениваемой страницы. Для этого также использовалась утилита SysBench. Был создан тестовый файл размером 30 МБайт, он отображался в память, и тестировалась скорость обмена целой страницы. Тест показал, что оптимальной страницей, удовлетворяющей правилу 5 минут( если в записи в памяти нет обращений больше 5 минут, то её целесообразнее хранить на постоянном носителе), является страница размером 2КБайта.
3. определен средний процент попадания в кэш (буфер) при работе базы данных. RDF-3x Engine использует системные функции для отображения файл с физического диска в память(CreateFileMapping – для Windows, mmap - для Linux). Результаты показали, что при первом запуске процент попадания в кэш составляет 27-30%, при повторном запуске 45-50%. Это доказывает целесообразность использования системных функций для кэширования информации в основную память.

Таким образом, исследуемая система RDF-3x Engine показала отличные результаты. Тестирование происходило на базе знаний YAGO. Она содержит более чем 2 миллионов объектов или сущностей (информация о людях, организациях, городах и так далее). База знаний содержит 20 миллионов фактов относящихся к этим объектам. Сама база знаний является частью проекта YAGO-

NAGA Института им. М.Планка[6].

В итоге можно сделать вывод, что модель RDF баз данных имеет некоторые достоинства по сравнению с реляционными базами данных, но также имеются недостатки и проблемы (например - длинные пути союзных операций замедляют работу и другие), которые требуют исследований и усовершенствования модели. Но можно уверенно сказать, что технология может получить дальнейшее развитие, так как необходимость обмена и хранения информации в современных компьютерных сетях растет с каждым днем.

**Литература**

[1] T. Neumann, G. Weikum, RDF3X:a RISCstyle Engine for RDF[Электронный ресурс]: New Zealand,2008. - Режим доступа: http://www.vldb.org/pvldb/1/1453927.pdf

[2] About W3C[Электронный ресурс]: - Режим доступа: http://www.w3.org/Consortium/

[3] SPARQL Query Language for RDF[Электронный ресурс]: 2008. – Режим доступа: http://www.w3.org/TR/rdf-sparql-query/

[4] Goetz Graefe. The Five-minute Rule: 20 Years Later and How Flash Memory Changes the Rules[Электронный ресурс]:, ACM QUEUE, July/August 2008. – Режим доступа: http://queue.acm.org/detail.cfm?id=1413264

[5] Nokia N800 Internet Tablet[Электронный ресурс]. – Режим доступа: http://www.nokia.de/produkte/mobiltelefone/nokia-n800/funktionen

[6] The YAGO-NAGA Project: Harvesting, Searching, and Ranking Knowledge from the Web[Электронный ресурс]: - Режим доступа: http://www.mpi-inf.mpg.de/yago-naga