УДК 519.7:007.52

М.В. Головянко, Д.А. Плиско

Харьковский национальный университет радиоэлектроники, Харьков

ПОСТРОЕНИЕ РАСПРЕДЕЛЕННОЙ СИСТЕМЫ ОНТОЛОГИЙ НА БАЗЕ ТЕХНОЛОГИИ Peer-To-Peer (P2P)

В статье рассмотрены преимущества построения распределенного хранилища онтологий на основе Peer-To-Peer сетей, определены принципы функционирования таких сетей, рассмотрены технические аспекты обеспечения надежности и масштабируемости, а так же организация маршрутизации поисковых запросов.

Ключевые слова: *онтологии, P2P, хранилище, Semantic Web, поиск, маршрутизация.*

Введение

В настоящее время онтологии применяются во множестве различных приложений – для описания и категоризации сайтов, для описания товаров и их характеристик в интернет-магазинах, для формализации бизнес-процессов в системах автоматизации управления предприятиями. Целью создания описаний предметной области в виде онтологий является представление знаний о предметной области в форме, пригодной для машинной обработки.

Важной составляющей любой системы является техническая платформа. На сегодняшний день существуют несколько вариантов организации таких платформ, однако наиболее часто используется клиент-серверная технология.

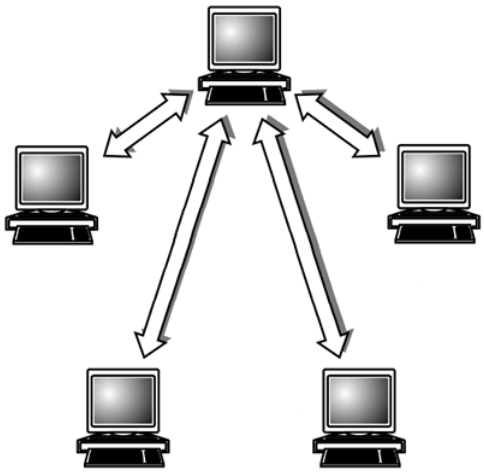

Рис. 1. Клиент-серверная архитектура

В этом случае предоставляемые ресурсы хранятся на одном узле сети, называемом «сервер», а доступ к ним осуществляется множеством пользователей – «клиентов». К преимуществам такого подхода можно отнести централизованное управление, контроль версий, наличие готовых технических и программных решений и большого числа специалистов. Основные недостатки – невысокое быстродействие, недоступность ресурса при сбоях, необходимость в мощном оборудовании и каналах связи, необходимость резервного копирования информации.

Частично недостатки клиент-серверной архитектуры решает технология Peer-To-Peer (P2P) сетей. Их особенность – каждый узел сети является одновременно и сервером, и клиентом, то есть предоставляет некоторую информацию и обращается к другим узлам. Координация так же выполняется всеми участниками одновременно, за счет распределенного хранения метаинформации всеми участниками сети.

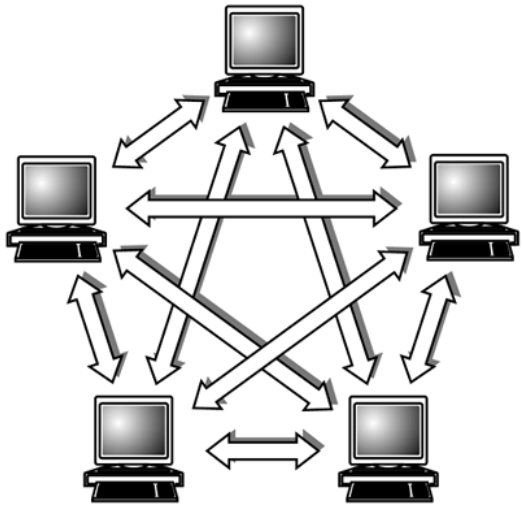

Рис. 2. Архитектура Peer-To-Peer

Сеть может иметь серверные узлы, однако они не предоставляют ресурсов, выполняя лишь координационные функции для ускорения поиска и авторизации. В качестве примеров можно привести популярные сети обмена файлами, такие как BitTorrent, Kademilia, сети интернет-телевидения Joost и Babelgum, интренет-телефон Skype. Данные технологии хорошо зарекомендовали себя и активно используются тысячами пользователей во всем мире. Одновременное хранение информации различными узлами позволяет обеспечить высокое быстродействие и надежность, а так же простоту развертывания и использования таких систем.

Хранение онтологий

Онтология – формальное явное описание понятий в рассматриваемой предметной области – классов (иногда называемых понятиями или концептами), свойств каждого понятия, описывающих различные характеристики и атрибуты понятий – слотов (иногда называемых ролями или свойствами), и ограничений, наложенных на слоты – фацетов или ограничений ролей, а так же индивидуальных экземпляров классов. Иерархия классов без описания их экземпляров образует таксономию описываемой предметной области.

На практике такие базы знаний как правило представляют собой документы на языке OWL (Ontology Web Language). Данный язык разрабатывался консорциумом W3C в рамках концепции Semantic Web, то есть изначально в язык закладывались принципы функционирования современных информационных технологий, а именно использование XML-синтаксиса и универсальных идентификаторов ресурсов URI. Благодаря этому, во-первых, существует возможность повторного использования уже разработанных онтологий, во-вторых – разделение концептуально целых онтологий на различные документы. С технической точки зрения, онтологию можно представить в виде отдельных элементов, однозначно определяемых их идентификаторами, состоящими из URI документа и внутреннего идентификатора ресурса.

При использовании клиент-серверной технологии, онтология будет полностью доступна при доступности всех онтологий, на которые она ссылается. Отказ одного из серверов или загруженность канала может сделать невозможным доступ к хранимым документам и, как следствие, может создать затруднения для работы систем использующих онтологию. Для решения данной проблемы предлагается использовать технологию ячеистых («пиринговых», P2P) сетей. В такой сети онтология будет храниться одновременно на нескольких узлах, что позволит обеспечить резервирование информации теми узлами, которые ее используют.

Распределенное хранилище на основе P2P сети

Для распределенного хранения онтологий на основе P2P-сети необходимо обеспечить возможность поиска идентификаторов узлов, содержащих необходимую информацию, определить их физические адреса, а так же обеспечить возможность одновременного получения информации от различных источников в многопоточном режиме, что может быть необходимо при получении больших объемов данных. Кроме того, необходимо обеспечить работоспособность сети без использования координационных серверов.

Для организации таких сетей как правило применяется мера «близости» значений некоторых хеш-функций. Характерной особенностью таких функций является то, что на вход они принимают набор байт произвольной длины, а на выходе функции поток байт имеет всегда один и тот же размер. Хеш-функции активно применяются в криптографии, например, для организации цифровой подписи, а так же в информационных технологиях, например, для контроля ошибок при передачи файла по сети (контрольные суммы файла). Кроме того, хеш-функции обладают устойчивостью к коллизиям, то есть незначительные изменения значений входных параметров значительно изменяют выходное значение (хеш-строку). Наиболее часто применяются известные алгоритмы: SHA1, MD4, MD5.

Таким образом, применяя хеш-функцию к URI любого ресурса онтологии можно получить идентификатор ресурса в виде строки байт постоянной длины, а применяя ее к содержимому проверять соответствие содержимого на разных узлах и контролировать ошибки передачи информации.

Для идентификации узлов так же применяется значение хеш-функции от случайной строки (благодаря свойствам хеш-функций вероятность совпадения крайне мала) которая называется глобальным идентификатором пользователя (GUID). Важно, чтобы в рамках одной сети применялся один и тот же алгоритм. Это обеспечит возможность сравнения значений хеш-функции между собой.

Для сравнения могут применяться различные меры длины. Как правило, используется расстояние Хэмминга (линейное расстояние) – количество различающихся бит в соответствующих позициях. Таким образом, чем меньше различаются соответствующие биты в хеш-строках, тем «ближе» эти строки друг к другу. Данная метрика удовлетворяет всем необходимым условиям (неотрицательность, коммутативность, транзитивность), а так же является простой в вычислении, что позволяет использовать ее для сравнения больших массивов данных.

Именно на этих принципах строиться маршрутизация запросов в P2P-сетях. Каждый узел сети хранит информацию о «соседях» - узлах, чей GUID имеет минимальное расстояние по отношению к собственному. Таким образом, отсутствует необходимость в построении полносвязной сети. Каждый узел может получить физический адрес необходимого ему узла, «опрашивая» те узлы, чей GUID имеет минимальную метрику к искомому. Узел, получив такой запрос, либо отвечает на него, либо передает одному из своих «соседей», наиболее близкому к искомому. Если же в списке отсутствует узел с меньшей метрикой до искомого чем у текущего – это означает отсутствие такого участника в сети.

Кроме того, узел хранит часть распределенной таблицы хешей (DHT, Distributed Hash Table). Эта

таблица формируется следующим образом: публикуя некоторый ресурс в сети, узел передает информацию о нем узлу, GUID которого наиболее близок к хешу URI ресурса. Каждый узел дописывает эту информацию в свою часть таблицы, отсортированную по расстоянию до собственного GUID и передает дальше, если ему известен более близкий узел чем он сам. Каждая локальная таблица ограничена по количеству записей, следовательно в ней будут оставаться информация о ресурсах с наиболее «близким» идентификатором, а остальные постепенно вытесняться. Таким образом, информация о местонахождении ресурса сохранится на нескольких узлах. Чем меньше метрика между идентификатором ресурса и GUID узла, тем больше вероятность, что данный узел имеет информацию о расположении ресурса. Так как записи DHT достаточно малы по объему (например, при использовании алгоритма SHA1 – 2 строки по 160 бит образуют запись длиной 40 байт), то локальные таблицы могут иметь достаточно большую длину, таким образом обеспечивается резервирование метаинформации. Даже в случае отключения нескольких узлов сети метаинформация остается доступной.

Поиск производится аналогично: запрос отправляется или передается наиболее «близкому» к ресурсу узлу. Если ближе узла не существует, а ресурс не найден – следовательно, в сети он гарантированно отсутствует. Кроме того, построенная цепочка маршрута будет оптимальной. Таким образом, определив GUID узла, у которого есть необходимый ресурс, аналогично находится его физический адрес.

Анализ эффективности и перспективы развития

Как уже упоминалось выше, основными преимуществами подхода являются высокая скорость получения ресурсов и надежность сети, что обеспечивается многократным автоматическим дублированием записей в распределенной таблице хешей. Так же важным достоинством является простота развертывания и масштабирования – для добавления новых узлов достаточно установить программное обеспечение и указать адрес любого работающего узла. Таким образом, не требуется специальное оборудование и подготовка персонала.

Основным недостатком данного подхода является децентрализация сети. Это создает трудности для организации безопасности (особенно в корпоративных сетях) и контроля версий предоставляемого ресурса. В настоящее время консорциумом W3C ведется разработка технологии P3P, которая должна обеспечивать надежность информации в распределенных одноранговых сетях, а так же распределенную систему авторизации пользователей, основанную на «подтверждении» аутентификационной информации сторонними узлами. Частично эта идея уже реализуется в некоторых сетях обмена файлами для защиты от «личинга» – потребления информации без предоставления своих ресурсов.

Также к недостаткам можно отнести невысокую скорость выполнения транзакционных запросов, так как выполнение таких запросов происходит с использованием множества узлов.

Решением проблем может стать использование координационных серверов, предназначенных для ускорения работы сети, но не являющихся необходимыми для поддержания работоспособности. В таком случае, в случае выхода их из строя сеть продолжит функционирование.

В целом, сети P2P являются мощным инструментом построения распределенных систем и могут применятся для построения распределенных хранилищ онтологий как корпоративного уровня, так и в сети Интернет.

Список литературы

1. Fu X.H. Distributed information search based on topic segments in structured peer-to-peer networks / X.H. Fu, B.Q. Feng // J. Xi'An JiaoTong Univ, 2005.

2. Gummadi, K.P. Measurement, modeling and analysis of a peer-to-peer filesharing workload / K.P. Gummadi, R.J. Dunn, S. Saroiu, S.D. Gribble, H.M. Levy, J. Zahorjan // ACM SIGOPS Operat. Syst. Rev., 37, 2004.

3. Perényi, M. Identification and analysis of peer-to-peer traffic / M. Perényi, T.D. Dang, A. Gefferth, S. Molnár // J. Commun., 1, 2006.

4. Singh, A. Apoidea: A decentralized peer-to-peer architecture for crawling the World Wide Web / A. Singh, M. Srivatsa, L. Liu, T. Miller. – Workshop on Distributed Information Retrieval, 2003.

Поступила в редколлегию 22.06.2010

Рецензент: д-р физ.-мат. наук А.А. Галуза, Национальный технический университет «ХПИ», Харьков.

ПОБУДОВА РОЗПОДІЛЕНОЇ СИСТЕМИ ОНТОЛОГІЙ НА БАЗІ ТЕХНОЛОГІЇ PEER-TO-PEER

Д.А. Пліско, М.В. Голов'янко

Розглядаються технічні аспекти побудови розподіленної системи онтології на основі мереж Peer-To-Peer.

Ключові слова: *онтологія, P2P, сховище, Semantic Web, пошук, маршрутизація.*

DISTRIBUTED ONTOLOGY-BASED SYSTEM DESIGN BY PEER-TO-PEER TECHNOLOGY

D.A. Plisko, M.V. Golovianko

Is considered technical aspects of distributed ontology-based systems design on Peer-To-Peer networks base.

Keywords: *ontology, P2P, depository, Semantic Web, search, routing.*