УДК 004.75

РАСПРЕДЕЛЁННОЕ ХРАНИЛИЩЕ ДАННЫХ СЕМАНТИЧЕСКОЙ СЕТИ

Щербаков В.А. (Университет ИТМО)
Научный руководитель – старший преподаватель Цопа Е.А.
(Университет ИТМО)

В докладе представлен промежуточный итог реализации проекта распределённого хранилища данных семантической сети. В ходе сравнительного анализа были рассмотрены подходы к организации распределённых систем и выбран алгоритм, наиболее подходящий для реализации таковой в контексте семантической сети. Была изучена структура хранения данных в графовых БД. В итоге была разработана архитектура распределённого хранилища данных семантической сети, удовлетворяющая требованиям, предъявляемым предметной областью.

Введение.
В отличие от реляционной модели данных, данные в семантической сети описываются помеченным направленным графом. Структурными компонентами семантической сети являются концепты, свойства и атрибуты. Также перечисленные компоненты могут иметь конкретные экземпляры. К экземплярам также могут быть добавлены канонические формы лексем — леммы и глоссы — словарные определения понятия концепта.[1]

Ближайшими распространёнными аналогами семантических сетей являются словари и тезаурусы. В ходе выполнения поиска существующих реализаций были изучены следующие проекты:

- РуТез[2];
- Wiktionary[3];
- BabelNet[4];
- WordNet[5];
- YARN[6].

Также ближайшими структурными аналогами можно выделить графовые СУБД, в качестве представителя которых была выбрана СУБД Neo4j.

Основная часть.
На начальных этапах проекта были выделены следующие требования:

- Реализация должна быть выполнена на языке программирования C++.
- Модуль, реализующий доступ к хранилищу должен реализовывать обобщённый интерфейс для простоты замены механизма хранения.
- Для реализации интерфейса доступа к хранимым данным, предлагается реализовать простой планировщик выполнения запросов.
- Модуль хранения должен иметь поддержку MVCC транзакций.

Для построения распределённых систем применяется подход с использованием алгоритмов консенсуса. В частности, рассматривается получивший широкое распространение алгоритм консенсуса Raft[7].

Не всякая СУБД может быть эффективно использована для хранения данных семантической сети. Так как семантическая сеть это, де-факто, ориентированный граф, то для решения этой задачи в первую очередь были рассмотрены современные графовые СУБД, как наиболее близко соответствующие реальной структуре сети. В качестве примера графового хранилища была выбрана СУБД Neo4j как одно из наиболее распространённых сегодня решений для

построения высоконагруженных систем большого масштаба. Наибольший интерес в ходе исследования представляла структура хранения данных графа.

При реализации распределённой обработки данных в СУБД необходимым условием является обеспечение целостности данных, в частности гарантия продолжения существования и неизменность данных в рамках активного запроса вне зависимости от числа узлов задействованных в ходе обработки этого запроса. Одним из возможных решений этой проблемы являются MVCC транзакции, которые позволяют получать доступ ко множеству версий одного и того же набора данных. В качестве распространённого примера реализации MVCC транзакций была выбрана СУБД PostgreSQL.

В результате проведённого исследования были выбраны подходы, позволяющие решить ключевые проблемы, возникающие при построении распределённого хранилища. На следующем этапе была спроектирована архитектура программного продукта, реализующего эти подходы применительно к семантической сети.

Выводы.

В рамках первого этапа исследования были рассмотрены основные концепции, используемые при разработке распределённых систем хранения данных и исследовано применение этих концепций в архитектуре существующих программных систем. В результате проведённого анализа был сделан вывод о неприменимости готовых наработок для построения распределённого хранилища данных семантической сети, следствием чего стала необходимость разработки собственного решения.

На втором этапе исследования была разработана архитектура распределённого хранилища данных семантической сети, удовлетворяющая требованиям, предъявляемым предметной областью. В рамках предложенной архитектуры был переработан, дополнен и повторно использован ряд концепций, реализованных в существующих реляционных и графовых СУБД.

На следующих этапах исследования планируется доработать предложенную архитектуру и реализовать её требования в коде.

Список использованных источников:

1. Клименков С.В., Николаев В.В., Харитонова А.Е., Гаврилов А.В., Письмак А.Е., Покид А.В. Применение семантической сети для хранения слабоструктурированных данных // Инженерный вестник Дона [электронный журнал] -2020. -- № 2(62). -- С. 27
2. Алексеев А. А., Добров Б. В., Лукашевич Н. В. Лингвистическая онтология–тезаурус РуТез. – 2013.
3. Liebeck M., Conrad S. IWNLP: Inverse Wiktionary for Natural Language Processing // ACL (2). – 2015. – С. 414-418.
4. Navigli R., Ponzetto S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network //Artificial Intelligence. – 2012. – Т. 193. – С. 217-250.
5. Fellbaum C. WordNet //The encyclopedia of applied linguistics. – 2012.
6. YARN: Spinning-in-Progress / P. Braslavski, D. Ustalov, M. Mukhin, Y. Kiselev // Proceedings of the Eight Global Wordnet Conference. — Bucharest, Romania, 2016. — С. 58–65.
7. Ongaro D., Ousterhout J. In search of an understandable consensus algorithm // 2014 USENIX Annual Technical Conference (Usenix ATC 14). -- 2014. -- С. 305-319.