быстрая агрегация большого объема данных. Серьёзными требованиями к современным веб-приложениям являются быстрый отклик и максимальная производительность при работе с большим объемом информации.

Современные веб-приложения обрабатывают большое количество запросов и действий пользователей. Такого рода информация имеет ценность с точки зрения поведенческого анализа субъекта.

Реляционные СУБД стали стандартом при проектировании слоя хранения данных в современных приложениях. Современные реляционные СУБД соответствуют требованиям ACID: Atomicity — Атомарность, Consistency — Согласованность, Isolation — Изолированность, Durability — Надежность. В основу реляционной модели заложены жесткие схемы, строго регламентирующие структуру и связи сущностей. Реляционные базы данных обеспечивают надёжное хранение данных, атомарность крупных операций и постоянную согласованность. Однако для достижения подобного поведения используются различные механизмы, которые накладывают пенальти на производительность и ограничивают возможности масштабирования.

С приходом огромных массивов информации и распределенных систем стало ясно, что обеспечить для них одновременно транзакционность набора операций и получить высокую доступность и быстрый отклик — невозможно.

Эта идея была положена в основу CAP теоремы, которая гласит о том, что реализация распределенных вычислений не может достичь одновременно трёх свойств: Consistency (Согласованность), Availability (Доступность), PartitionTolerance (Терпимость к разделению)

Напрактикесуществуютсистемытипа CA (Availability, Consistency), CP (Consistency, Partition tolerance), AP (Availability, Partition tolerance).

С лавинообразным ростом количества пользователей и информации увеличились требования к производительности хранилищ данных. Необходимость обрабатывать большое количество информации за разумное время столкнулась с проблемой вертикальной масштабируемости баз данных.

Выходом из ситуации является горизонтальное масштабирование, когда несколько независимых серверов соединяются между собой, и каждый владеет своей репликой или частью данных, и обрабатывает только часть запросов. В такой архитектуре для повышения мощности хранилища достаточной мерой является добавление нового сервера в кластер. Процедурами шардинга, репликации, обеспечением отказоустойчивости, перераспределения данных в случае добавления дополнительного сервера в таких системах занимается СУБД.

Общие характеристики NoSQLдокументоориентированных СУБД:

1. Представление данных в виде агрегатов. Документоориентированные хранилища пропагандируют агрегирование и встраивание данных в документ, чтобы оперировать с этими сущностями как с целостными объектами.
2. В большинстве своём это opensource решения, они бесплатны.
3. Возможность автоматически распределять данные между серверами.
4. Использование памяти, прозрачное кэширование - содержимое коллекций активно кэшируется для выборки.

Вопрос применения документоориентированных СУБД в современных веб-приложениях зависит от конечной аудитории пользователей и планируемой нагрузки. Очевидно, что упрощение процесса разработки и закладка в возможность горизонтального масштабирования также является весомым аргументом при выборе СУБД. Современные NoSQL СУБД могут стать очень эффективным инструментом для хранения данных современных веб-приложений.

Список использованных источников:

1. BASE: An ACID alternative - Dan Pritchett, http://queue.acm.org/detail.cfm?id=1394128
2. Why NoSQL? - Couchbase, http://www.couchbase.com/why-nosql/nosql-database
3. http://ru.wikipedia.org/wiki/ACID

ОЦЕНКА ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ ЧАСТОТНО-КОНТЕКСТНОГО АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Потараев В. В.

Серебряная Л. В. – канд. техн. наук, доцент

В современных информационных системах содержится огромное количество текстовой информации. Существуют различные методы анализа текстовой информации. Рассмотрим эффективность применения частотно-контекстной классификации при решении задачи выбора текста, наиболее полно отражающего некоторую тему.

Предположим, необходимо выбрать один наиболее содержательный текст из множества текстов по данному вопросу. Очевидно, что в этом случае можно использовать классификацию текстовых данных.

Классификация подразумевает вычисление близости текста с другими текстами (представляющими классы) [1]. Текст, который наиболее близок с остальными текстами (классами) по содержанию, будет наиболее полно отражать смысл всех найденных статей [2].

Метод частотно-контекстного анализа текстовой информации отличается от частотного анализа тем, что он учитывает слова, используемые в тексте рядом с ключевыми.

Пример. Если информационный поток некоторого текста можно записать в виде F = (i3, i6, i7, i1, i2, i11, i9, i4, i10, i3, i5, i6, i7, i1, i8, i9, i4, i10, i5), то его структуру можно представить в виде графа (рис. 1):

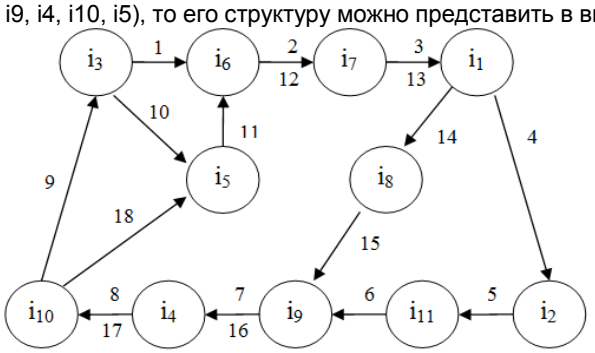

Рис. 1 – Структура, формируемая информационным потоком

Общая последовательность метода частотно-контекстного анализа выглядит следующим образом:

1) Моделирование текста и формирование его информационной структуры.

2) Выделение множества всех информационных элементов, ранжированных по их числу повторений в тексте.

3) Выделение множества ключевых элементов Sp.

4) Формирование уточняющего множества Ss на основе контекстного анализа информационных элементов множества Sp.

Предположим, что необходимо оценить эффективность метода частотно-контекстной классификации текстовой информации для решения задачи выбора текста, наиболее полно отражающего смысл некоторой темы (этот текст должен быть наиболее близок по содержанию к остальным текстам по теме).

Пусть есть множество текстов A_1, состоящее из текстов X_1, X_2, …, X_n. Пусть некоторым образом известно (например, согласно решению эксперта в предметной области), что текст, наиболее полно отражающий общую тему, – это текст Y. Некоторый метод вычисления близости текстов показывает, что суммарная близость с остальными текстами максимальна у текста X_{max}. Если $Y=X_{max}$, то текст выбран верно, иначе – неверно.

Проведя K подобных испытаний выбора текста, получим количество текстов R, классифицированных верно. Значение метрики, оценивающей эффективность применения метода, можно рассчитать по формуле:

$$V = R/K$$

Метрика принимает значения от 0 до 1.

Итак, была разработана метрика для оценки эффективности методов выбора текста, наиболее полно отражающего некоторую тему.

Список использованных источников:

1. Потараев В. В. Применение частотно-контекстной классификации текстовой информации при выборе текстов для изучения // Дистанционное обучение – образовательная среда XXI века: Материалы VIII международной научно-методической конференции – Минск, 2013 – с.340-341.
2. Тарасов, С.Д. Метод тематического связанного ранжирования для автоматического сводного реферирования новостных сообщений в задачах поддержки принятия управленческих решений/ С.Д. Тарасов // Вестник ВГУ.– 2010. №1. – С. 166–173.

ОБЕСПЕЧЕНИЕ АТРИБУТОВ ВЫСОКОЙ ДОСТУПНОСТИ И БЫСТРОГО ВОССТАНОВЛЕНИЯ ПОСЛЕ СБОЯ СИЛЬНО РАСПРЕДЕЛЕННЫХ СИСТЕМ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Базаревский В.Э., Базаревский Вл.Э.

Бранцевич П.Ю. − к. т. н., доцент

Рассматриваются архитектурные тактики, применимые для обеспечения следующих атрибутов качества сильно распределенных программных систем: высокая доступность, отказоустойчивость, быстрое восстановление