

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ СЛОВ В ВЕКТОРНЫХ БАЗАХ ДАННЫХ

Рассматриваются методы векторного представления слов в векторных базах данных. Исследуются подходы к созданию и обработке векторных представлений, включая прогнозные (Word2Vec), контекстуальные (BERT) и графовые модели (GNN), а также классический латентно-семантический анализ (LSA) и глобальные векторные представления (GloVe). Основное внимание уделяется способности методов справляться с разнообразием задач обработки текстовой информации.

ВВЕДЕНИЕ

Векторные базы данных становятся ключевым инструментом для эффективного поиска и обработки семантически связанных данных, а выбор методов векторного представления слов (word embeddings) напрямую влияет на их производительность и точность. В последние годы появилось множество подходов — от классических (Word2Vec, GloVe) до современных контекстных моделей (BERT), — каждый из которых обладает своими преимуществами и ограничениями в задачах индексирования, поиска и кластеризации. Однако при интеграции в векторные базы данных эти методы демонстрируют разную эффективность в зависимости от масштаба данных, требований к скорости и точности. Исследование ключевых методов векторного представления, оценивание их применимости в реальных сценариях работы с векторными базами, и выявление оптимальных стратегий выбора эмбедингов для различных задач способствует разработке более эффективных и масштабируемых систем обработки информации, а также повышает точность поиска и анализа данных.

I. ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ СЛОВ

Векторное представление слов (word embeddings) — это способ преобразования текстовой информации в многомерные числовые векторы фиксированной размерности, которые сохраняют семантические и синтаксические связи между словами [1]. Такие представления позволяют учитывать смысловые связи между словами, их контекст и семантику, делая возможным применение сложных алгоритмов машинного обучения для обработки и анализа текстов. Современные методы, от классических до контекстно-зависимых, различаются по способу кодирования информации, что влияет на их применимость в различных задачах — поиске, классификации или кластеризации текстов.

II. ТРЕБОВАНИЯ К МЕТОДАМ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ СЛОВ В ВЕКТОРНЫХ БАЗАХ ДАННЫХ

При выборе подхода для интеграции в векторные базы данных ключевыми становятся несколько критериев: способность сохранять семантические и синтаксические отношения между лексическими единицами в компактном векторном пространстве, масштабируемость для обработки больших объемов текстовых данных, эффективность индексирования и поиска в многомерных пространствах, адаптивность к различным предметным областям и языковым особенностям, устойчивость к полисемии и другим лингвистическим явлениям, баланс между размерностью векторов и вычислительной сложностью операций, а также поддержка динамического обновления представлений без полного переобучения модели.

III. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ СЛОВ

Сравнительный анализ методов векторного представления слов, таких как LSA, Word2Vec, GloVe, BERT и графовые нейронные сети, позволяет выявить их применимость в задачах, связанных с векторными базами данных, где LSA подходит для выявления скрытых семантических связей, Word2Vec и GloVe эффективны для семантического поиска, BERT учитывает контекст, а графовые модели — структурные связи. Выбор метода, определяемый задачей, требованиями к точности, скорости и доступным ресурсам, критически важен для эффективной работы векторных баз данных.

1. Lychenko, N. Comparison of the effectiveness of word vector representation methods for determining the sentiment of texts / N. Lychenko, A. Sorokova // University of Central Florida – 2019. – Vol. 6. – P. 1-6.

Савоневская Маргарита Олеговна, студентка ФИТиУ БГУИР, go.to.mychannel745@gmail.com.

Столярова Виолетта Витальевна, студентка ФИТиУ БГУИР, violettastolarova@icloud.com.

Черникова Лолита Александровна, студентка ФИТиУ БГУИР, lolita190511@gmail.com.

Научный руководитель: Трофимович Алексей Фёдорович, старший преподаватель кафедры ИТАС БГУИР, trofimaif@bsuir.by.