Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.383

Фарафонов
Максим Павлович

Система управления содержимым для вики-ресурсов

АВТОРЕФЕРАТ
на соискание академической степени
магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

Научный руководитель
Егорова Н.Г.
к.т.н., доцент

Минск 2020

КРАТКОЕ ВВЕДЕНИЕ

Структуризация информации это одна из больших задач современного веба. Поисковые движки стараются как можно лучше категоризировать сайты, трекеры пытаются выделить привычки и принадлежность пользователей по их поведению. В рамках обработки текста существует много задач и решений в области обработки естественного языка.

С другой стороны, люди используют большое количество систем, первоначально заточенных под сбор информации. Это и корпоративные хранилища знаний, и личные ежедневники. Выделение правильной структуры данных для таких систем это связанная, но сильно отличающаяся задача.

Дипломная работа автора была посвящена техническим аспектам реализации вики-подобного движка. Обсуждались существующие реализации, их основной функционал, способности к стилизации, интерфейс, структура. Это бытовые аспекты, касающиеся каждодневного использования. В данной работе автору хотелось сделать шаг назад, взглянуть более широко и задать, скорее, философские вопросы:

- Какие плюсы и минусы есть у вики-формата?
- Какие фундаментальные проблемы стоят перед пользователями?
- Какие есть трудности в технической поддержке таких систем?
- Каков потенциал накопленных знаний?

Помимо проблем интерес представляют улучшения простого стандартного формата вики. У компании Atlassian есть система управления знаниями под названием Confluence. Существует шутка, что Confluence это кладбище ретроспектив. Это очень близко к правде, потому что часто люди записывают большое количество информации, которое потом нигде не используется. Хорошо это или плохо? Может ли организация движка помочь с организационными проблемами?

Одно из наиболее значительных направлений в этой области - семантические вики. В своей сути это расширение концепции вики для добавления на страницы вычисляемого содержимого. Представляет интерес исследовать имеющиеся решения в этой области, их возможности и ограничения.

Целью данной работы является анализ задач, проблем и зон роста вики-движков как класса. На его основе будут выдвинуты гипотезы для решения проблем и реализации улучшений, а также разработаны прототипы ПО.

Тема тесно связана со специальностью: обработка больших данных это одна из тем, которые будут затронуты в этой работе. На определенном масштабе объем данных требует качественного перехода к другим методам обработки.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является анализ концептуальной модели управления знаниями и разработка ПО для решения найденных проблем.

Для достижения поставленной цели необходимо решить следующие задачи:

1. провести исследование предметной области, изучить существующие продукты;
2. проанализировать модель и найти зоны роста;
3. найти способы развития систем управления знаниями.

Объектом исследования являются вики-движки и другие системы управления знаниями.

Предметом являются проблемные места и зоны роста систем управления знаниями.

Основной *гипотезой*, положенной в основу диссертационной работы, является несовершенство существующих решений и технологий в сфере систем управлния знаниями. Предполагается, что при должно анализе предметной области можно найти пространство для роста.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. Г. Егоровой заключается в формулировке целей и задач исследования.

Публикации

По теме диссертации публикаций не было.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников и приложений. В первой главе представлен анализ предметной области, определены понятия и рассмотрены существующие продукты. Вторая глава посвящена рассуждению на тему прооблем классической модели вики-движка. В третьей главе предложены решения описанных проблем, описаны существующие технологии и дана их оценка. В четвертой главе предложена практическая реализация прототипа нейросетевого суммаризатора текста, а также продемонстрирована работа с графовой моделью данных

Общий объем работы составляет 47 страниц, из которых основного текста 30 страниц, 8 рисунков на 8 страницах, список использованных источников из 26 наименований на 2 страницах и 3 приложения на 8 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой** главе проводится исследование предметной области, в которой работает разрабатываемый проект, разъяснение терминологии, исследование литературы по теме с целью построения разностороннего взгляда на решаемую проблему. Рассматриваются имеющиеся аналоги, отмечаются их положительные и, если применимо, отрицательные стороны. На основе полученных данных проводится анализ применимости продукта и ему подобных.

Из значимых продуктов следует отметить следующие:

- MediaWiki: программный продукт, на котором построена Wikipedia и многие другие ресурсы
- RoamResearch: современный инструмент для организации знаний в виде графа
- NLTK: популярная библиотека для работы с естественным языком

Во **второй** главе описывается, с какими задачами и проблемами сталкивается на практике классический вики-движок, аналогичный реализованному в моей дипломной работе. Предполагается, что в движке присутствует возможность многопользовательского редактирования, долгосрочного хранения данных и некий язык разметки.

Одной из основных задач выделена возможность машинной обработки данных. С ростом объема информации возникает потребность выполнения запросов на данных без ручного анализа.

Структуризация может проводиться пользователями при вводе информации или автоматически системой. Второй вариант более удобен для пользователей, но сложен в реализации и будет выдавать много ошибок на большом масштабе.

Другой интересный вопрос это обеспечение хорошего UX, то есть пользовательского опыта от продукта. Конкретные шаги пока не описаны.

В **третьей** главе говорится о решениях и технологиях, которые помогают в поставленных задачах. Они не обязательно связаны, многие из них развивались параллельно с развитием концепции вики и веба самого.

Первая такая технология это графовые базы данных. Они предназначены для хранения взаимосвязей и навигации в них. Реляционные модели в таких случаях оказываются слишком громоздкими и медленными, а соответствующие запросы слабочитаемыми. Взаимосвязи в графовых базах данных являются объектами высшего порядка, в которых заключается основная ценность этих

баз данных. В графовых базах данных используются узлы для хранения сущностей данных и ребра для хранения взаимосвязей между сущностями. Ребро всегда имеет начальный узел, конечный узел, тип и направление. Ребра могут описывать взаимосвязи типа «родитель-потомок», действия, права владения и т. п. Ограничения на количество и тип взаимосвязей, которые может иметь узел, отсутствуют.

Второе важное решение, решающее вопрос машинной обработки, это семантическая паутина. Создатель современного веба Тим Бернерс-Ли выделял следующие принципы связанных данных в семантическом вебе:

1. использование URI для именования, чтобы идентификаторы были глобально уникальными;
2. использование URI в схеме HTML, чтобы по ним можно было удобно перейти;
3. при переходу по URI подавать информацию в стандартном формате (RDF, SPARQL), чтобы пользователи знали, как ее обработать;
4. при этом включить ссылки на другие ресурсы, чтобы можно было продолжить поиск.

Рассматриваются языки и технологии для реализации семантической паутины. Описывается популярный плагин Semantic MediaWiki. Обсуждаются проблемы семантической паутины.

Одним из способов улучшения пользовательского опыта является предоставления выдержек из текста, суммаризация статей. Описываются классификации подходов к задаче: по механизму работы (экстрактивные и абстрактивные) и по технологиям (статистические и нейросетевые). Описываются алгоритмы TextRank и BERT.

В **четвертой** главе рассматриваются практические аспекты реализации описанных технологий и выдвигаются оригинальные идеи по теме.

Семантический веб и возможность делать запросы к содержимому это хорошо, но что, если у нас нет размеченных данных? Конечно, можно использовать техники NLP для выделения информации. Но можно и взглянуть на это под другим углом: у нас есть HTML. Это тоже язык разметки со своей структурой. Можно делать запросы к самой структуре документа.

Описывается разбор HTML и запросы к модели при помощи SPARQL.

На основе предыдущей главы реализован прототип модели естественного языка с использованием BERT. Используется предобученная модель. Демонстрируется результат на двух задачах: восстановление скрытых слов и определение схожести фраз.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

В данной диссертации были описаны возможности классических викисистем: чего не хватает в этой базовой модели, какие проблемы из-за этого возникают и чего может не хватать пользователям.

В главе РЕШЕНИЯ был проведен анализ технологий, которые способствуют развитию вики-движков и развиваются вместе с ними. Рассмотрены пути решения существующих проблем.

В главе РЕАЛИЗАЦИЯ был продемонстрирован потенциал машинного обучения для задач NLP. Кроме того, приведены размышления на тему пользовательского опыта в системах с большими данными.

Машинное обучение, в частности глубокое обучение, хорошо подходит для задач NLP. В разрезе вики-движков его можно применить, например, для суммаризации текстов статей.

Основным научным результатом работы является анализ проблем вики-систем и конкретных их решений, в частности необходимости машинной обработки и обеспечения удобства пользователей. Сделан вывод, что для машинной обработки знаний интерес представляет семантическая паутина.

Семантическая паутина это технология с большим потенциалом в современном мире больших данных. В последнее время в ней не было серьезных развитий. Существует ряд проблем, которые ограничивают применение, например, чувствительность к некорректным данным.

Рекомендации по практическому использованию результатов

Основным практическим результатом работы является разработка прототипа нейросетевого суммаризатора текста. Кроме того, была продемонстрирована графовая модель данных на примере разборе HTML.

Полученные результаты можно применить как основу для проектирования новых систем управления знаниями. Прототипы ПО требуют доработки, но в целом могут быть использованы как реализация соответствующего функционала.

К дальнейшим шагам можно отнести:

- поиск актуальных сценариев применения семантического веба в современном мире;
- использование машинного обучения для автоматического выведения семантических связей;
- полноценная интеграция описанных идей и прототипов в ПО для вики-движка из дипломной работы.