Темы исследовательских работ, Александр Болховитянов

1. Методы онлайн определения дублирующего контента.

Тут нас интересуют алгоритмы, позволяющие происзодить онлайн склейку дублкатов документов на потоке в 10-ки тысяч документов в секунду.

2. Многофакторные методы определения дубликатов веб-документов.

Когда мы говорим о дубликатах веб-документов, то в первую очередь считаем, что у таких документов совпадает контент (текстовый). Однако, часто встречаются случаи, когда на страницах мало текста, он очень похож, но страницы не дубликаты, потому что на них есть разные картинки или видео. Поэтому и возникается задача: как научить существующие алгоритмы использовать сигнал не только от текстовго сравнения, но и любой другой (картинки, видео).

3. Метрики качества обнаружения дубликатов веб-документов.

Ни одна система не обходится без метрик и проблема, которая возникает в рассматриваемой области заключаетс в том, что не очевидны воппросы: как выбирать поток для разметки асессорами? как учитывать качество кластеризации?

4. Сегментация и выделение основного контента веб-документов.

На практике для решения задач информационного поиска требуется иметь дело не просто с html, а со значимой частью документа, получающейся после отбрасывания навигационных блоков, блоков рекламы и т.д. Тут требуется провести детальный анализ существующих решений, понять чем они плохи и как можно улучшить имеющиеся результаты.

5. Методы автоматического выделения незначащих параметров.

Существуют параметры в урлах, различные значения которых не влияют на отдаваемый сервером контент. Требуется проанализировать существующие решшения и предложить собственные подходы к решению данной задачи.

6. Методика выявления многоязычных сайтов.

В Интернете существуют сайты, которые отдают конент на разных языках, если их скачивают из разных регионов. Требуется разработать систему, которая могла бы определять такие сайты.

7. Компьютерная лингвистика в задачах анализа контента.

Тут широкий простор для исследований: например, построение графов синтаксических зависимостей и их применение для поиска похожих текстов. Учет морфологии в задачах поиска дублирующегося контента и т.д.