



СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ



ГУМАНИТАРНЫЙ
ИНСТИТУТ

Принципы семантической организации и метаданных текстовых файлов

Балцевич Артемий
Бочанова Олеся
Безруких Алиса
Васильева Дарья
Гинтер Милена
Манцаева Санжирма

ГФ25-01Б



Основные принципы семантической организации

Структурная
иерархия и
онтологическое
моделирование

Данные структурируются в
древовидные схемы,
облегчающие базовую
навигацию.

Для решения сложной
взаимосвязанности данных
используют онологии.

Применение: позволяет
компьютерам
автоматически
анализировать и предлагать
рекомендации.

Классификация,
категоризация и
фасетный анализ

Традиционная
классификация
ограниченна. Фасеты
позволяют гибкую
фильтрацию по множеству
критериев одновременно.

Машинное обучение
способно автоматически
выявлять тематику
документов, значительно
ускоряя процесс
классификации.

Семантическое
связывание

Современные системы
связывают объекты не только
физически, но и
семантически. Используются
уникальные идентификаторы
и стандарты записи связей.
Таким образом формируется
глобальная сеть знаний,
позволяющая отвечать на
комплексные запросы.



Основные принципы формирования и стандартизации метаданных

Стандарты и схемы метаданных

Метаданные необходимы для унификации описания ресурсов и совместимости между системами.

Примеры стандартных схем:

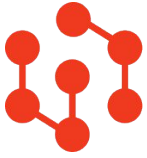
Dublin Core
Schema.org
MODS

Автоматическое извлечение и генерация метаданных

Ручное заполнение неэффективно при больших объемах данных. Программные инструменты способны автоматически извлекать и создавать метаданные. Обучение моделей осуществляется на огромных массивах примеров, повышая качество автоматического анализа.

Гибридные (полуавтоматические) подходы

Современные системы связывают объекты не только физически, но и семантически. Используются уникальные идентификаторы и стандарты записи связей. Таким образом формируется глобальная сеть знаний, позволяющая отвечать на комплексные запросы.



Продвинутые методы семантической обработки текста

Семантический анализ и распределённые представления слов

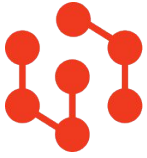
Каждое слово представляется числовой последовательностью (Word2Vec, GloVe), позволяя выявить сходства и различия между словами.

Подобные методы используются для поисковых запросов и рекомендаций товаров.

Нейросетевые архитектуры для обработки текста

Трансформеры, сверточные и рекуррентные нейросети превосходят старые методы в анализе текста.

Эти технологии применяются для перевода, Q&A и анализа эмоциональной окраски текста.



Заключение

Современные методы семантической организации и метаданных обеспечивают быструю и точную обработку больших объемов информации, существенно улучшая эффективность управления данными в организациях и системах любого масштаба.