**Фролов Дмитрий Сергеевич Агрегированное представление текстов для задач поиска в коллекциях текстовых документов**

ОГЛАВЛЕНИЕ ДИССЕРТАЦИИ

кандидат наук Фролов Дмитрий Сергеевич

0.8 Структура работы

1 Обзор литературы и существующих решений

1.1 Задача информационного поиска

1.1.1 Введение в задачу информационного поиска

1.1.2 Модели информационного поиска

1.1.3 Применение индексирования и ранжирования

1.1.4 Безындексные алгоритмы информационного поиска

1.2 Представление текста в задачах информационного поиска

1.2.1 Основные подходы

1.2.2 Признаковое описание документов

1.2.3 Агрегированное представление текстов с помощью аннотированных суффиксных деревьев (АСД) и других подходов

1.2.4 Вероятностное тематическое моделирование в коллекциях документов

1.2.5 Векторные представления слов и документов

1.3 Задача нечеткого поиска и методы ее решения

1.3.1 Признаковые подходы, алгоритмы Soundex, фрагментные подходы

1.3.2 Методы хэширования по сигнатуре

1.3.3 Метрические деревья, деревья Букхарда-Келлера и другие специализированные методы

1.4 Разведочный информационный поиск и анализ коллекций

1.4.1 Задача разведочного информационного поиска

1.4.2 Базовые методы разведочного поиска

1.4.3 Структуризация коллекций

1.4.4 Использование таксономий и понятие обобщения в разведочном поиске

1.4.5 Анализ коллекций научных публикаций

1.5 Популярные программные системы для информационного поиска

1.5.1 Lemur

1.5.2 ElasticSearch

1.5.3 Библиотека gensim

1.5.4 Другие популярные программные системы для информационного поиска

2 Разработка метода информационного поиска на основе

аннотированных суффиксных деревьев

2.1 Метод поиска АСДП и его оптимизация с помощью фрагментного индексирования

2.2 Экспериментальное сравнение АСДП с классическими методами информационного поиска

2.2.1 Сравнение качественных метрик

2.2.2 Сравнение производительности

2.3 Экспериментальное сравнение АСДП со специализированными методами нечеткого поиска

2.3.1 Сравниваемые методы и тестовые коллекции

2.3.2 Сравнение качества поиска

2.3.3 Сравнение скорости поиска

2.3.4 Результаты сравнения

2.3.5 Зависимость скорости поиска от длины строк для построения АСД в методе АСДП

2.4 Выводы на основе результатов экспериментов

2.5 Программная реализация распределенной поисковой системы, основанной на методе АСДП

3 Алгоритм оптимального обобщения нечеткого множества (ПарГеНМ) и его применение в задаче разведочного поиска

67

3.1 Описание задачи

3.2 Оптимальный подъем нечеткого тематического кластера в таксономии

3.2.1 Постановка задачи

3.2.2 Алгоритм ПарГеНМ

для поиска оптимального обобщения

3.2.3 Иллюстративные примеры

3.3 Применение алгоритма ПарГеНМ в задаче разведочного поиска

3.3.1 Структурирование коллекции научных публикаций с помощью таксономии предметной области

3.3.2 Подготовка коллекции научных публикаций

3.3.3 Таксономия DST

3.3.4 Вычисление степени релевантности между текстами

и темами таксономии

3.3.5 Определение и вычисление нечетких кластеров тем таксономии

3.3.6 Результаты подъема кластеров L, R, и C в таксономии DST

3.3.7 Выводы

3.4 Алгоритм ПарГеНМ как механизм анализа текстовых коллекций

3.5 Сравнение с результатами, полученными с помощью популярных подходов

3.5.1 Латентное размещение Дирихле

3.5.2 Иерархический кластер-анализ

4 Применение алгоритма ПарГеНМ для задачи расширения аудитории в рекламном таргетинге (Programmatic)

4.1 Модель интернет-рекламы Programmatic

4.2 Рекламный таргетинг в модели Programmatic

4.3 Применение алгоритма ПарГеНМ для обобщения пользовательских сегментов в интернет-рекламе

4.4 Оценка эффективности алгоритма обобщения пользовательских

сегментов

Заключение

Библиографический список использованной литературы

Список иллюстраций

Список таблиц

Приложение 1: Таксономия Науки о данных

согласно ЛСМ-СС8

Приложение 2: Темы, полученные с использованием

реализации ЬЮА из пакета gensim

Приложение 3: Тематические кластеры, полученные

с помощью метода ЦРОМЛ

Приложение 4: Программная реализация АСД с возможностью

вычисления степени вхождения строки в документ

Приложение 5: Программная реализация алгоритма ПарГеНМ

Приложение 6: Программная реализация отображения дерева

подъема нечеткого множества в таксономии

Введение