

Учебное пособие написано в соответствии с образовательным стандартом для подготовки инженеров по специальности «Программное обеспечение вычислительной техники и автоматизированных систем» для учебной дисциплины «Теория языков программирования и методы трансляции» на основе многолетнего опыта преподавания данного курса авторами. Предлагаемые материалы могут быть использованы при изучении ряда вопросов для дисциплин «Лингвистическое и программное обеспечение САПР» (специальность «Системы автоматизации проектирования») и «Системное программное обеспечение» (специальность «Вычислительные машины, комплексы, системы и сети»). Пособие содержит разделы, связанные с автоматизированным анализом связанных текстов, информационным поиском в больших массивах документов. Этот материал полезен для магистров и аспирантов различных специальностей.

В книге описаны основы формальных языков и методов контекстно-свободных языков, рассмотрены компиляторы, ассемблеры, интерпретаторы; показана необходимость разработки языковых средств; проведен обзор процесса компиляции. Описаны основные части компилятора: лексический и синтаксический анализаторы, генератор кода, оптимизатор кода, анализ и исправление ошибок, синтаксически-ориентированный метод трансляции, порождающие грамматики Хомского. Рассмотрены методы распознавания автоматных, контекстно свободных языков, методы синтаксического анализа, перевода и генерации кода; основы методов обработки естественного языка.

В книге предложен анализ различных лингвистических теорий; описана модель представления семантики связанного текста в виде дискурсного графа; приведена обобщенная архитектура автоматизированной системы анализа ЕЯ-текста; определены области практического использования класса подобных систем.

Детально рассмотрен процесс структуризации документов внутри коллекции и расстановка соответствующих гиперссылок. Степень близости между документами реализуется латентным семантическим анализом, применимым к векторной модели пространства документов и термов. Рассмотрены глобальное и локальное взвешивания термов, входящих в матрицу «терм—документ». Для взвешивания использованы статистические меры и нормирование с помощью энтропии. По результатам анализа создано векторное пространство семантической близости документов.

Показан процесс выделения иерархии семантически связанных групп документов. Описан способ задания параметров, используемых при кластеризации, влияющих на характер получаемой структуры. Варьируя данным параметром, можно подобрать необходимую степень связности итогового гиперграфа, представляющего структуру сайта. Определена степень влияния существующих ссылок при уже существующей структуре. Кластеризация гипертекста произведена на основе синтезированного алгоритма. Структура гипертекста представлена в виде гиперграфа.

Предлагаемый материал будет полезен студентам младших курсов, старшекурсникам и аспирантам.