Е. С. Кирьянов

РАЗБИЕНИЕ ТЕКСТА НА СЕМАНТИЧЕСКИЕ

Сложность структуры современного информационного общества постоянно растёт. В связи с этим, требования к эффективности алгоритмов обработки информации также увеличиваются. Одним из самых перспективных способов обработки информации могут быть семантические сети, которые предоставляют возможности по эффективному поиску и анализу данных как человеком так и программным обеспечением.

Введение

Современные системы дистанционного обучения используют базы знаний, обработку естественного языка, сложные механизмы поиска ответов и другие современные информационные технологии. Одной из задач, которые решают подобные системы, является представление текста семантической сетью. В частности, разбиение его на смысловые части. Смысловые части представляют собой законченные по смыслу и логике фрагменты текста, освещающие некоторый круг понятий. Они нужны при рассмотрении новых тем с целью фиксации на них внимания обучаемого, а также при контроле знаний.

I. Концепция семантических сетей

Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы. Таким образом, семантическая сеть является одним из способов представления знаний. В семантической сети роль вершин выполняют понятия базы знаний, а дуги задают отношения между ними. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений.

II. Систематизация на основе семантических отношений

Метод разбиения текста на смысловые страницы, а также применение семантических сетей в системах обработки текстов, характеризуется следующими основными задачами: – Изучение технологий построения семантических сетей; – Изучение и анализ имеющихся средств для работы с семантическими сетями; – Изучение применения семантических сетей в системах обработки текстов; – Рассмотрение методов кластеризации текстовых документов; – Сбор актуальной информации; – Разработка метода разбиения текста на смысловые страницы. Текст характеризуется набором наиболее часто используемых ключевых слов. По выбранным ключевым словам можно построить корелляционную матрицу, которая будет характеризовать степень связи между словами. В силу свойств корелляционной матрицы ее можно заменить вектором собственных чисел. Тогда задача разбиения упростится до определения близости двух векторов.

III. Выводы

Таким образом, результат применения разрабатываемого метода на текстовых базах знаний будет представлять собой законченные по смыслу и логике фрагменты текста, освещающие некоторый круг понятий.

1. JAVA и Интернет бизнес / О. В. Герман, Ю. О. Герман. – Минск : Бестпринт, 2010.- 384 с.

Кирьянов Егор Сергеевич, магистрант кафедры информационных технологий автоматизированных систем БГУИР, kiryanov.egor@gmail.com.

Научный руководитель: Герман Олег Витольдович, доцент кафедры информационных технологий автоматизированных систем БГУИР, кандидат технических наук, ovgerman@tut.by.