

Введение

Актуальность работы

В настоящее время человек постоянно сталкивается с проблемой обработки огромного количества информации, поступающей из различных источников. Рост объемов электронной информации носит экспоненциальный характер, что делает проблему острее с каждым годом. Автоматическая обработка естественных текстов занимает важное место в аналитико-синтетической работе, поскольку позволяет экономить временные и кадровые ресурсы.

Одним из актуальных направлений является автоматическое реферирование текстов. Развитие программных и аппаратных средств для реализации алгоритмов реферирования и повышение качества самих алгоритмов поднимают данную область на высокий уровень и позволяют использовать ее методы во многих отраслях человеческой деятельности. Решение этой задачи позволит увеличить эффективность работы поисковых систем, систематизировать рабочую документацию предприятия для управления документооборотом, структурировать и уменьшить объем веб-сайтов за счет размещения рефератов, защитить авторские права.

Для обработки больших объемов текстовой информации наиболее подходят методы экстрагирования – извлечения предложений из исходного текста без изменения их структуры. Данные методы отличаются простотой реализации и высокой скоростью работы, поскольку используют статистическую информацию о тексте. Первые работы в этом направлении появились еще в 60х годах прошлого века, однако значительного прогресса не наблюдается, поскольку число статистически важных параметров текста ограничено. Поэтому необходимо привлечение знаний и алгоритмов из других направлений обработки естественного текста для нахождения дополнительной информации о структуре текста. Исследования в области поиска алгоритмов продолжаются, таким образом, проблема обработки большого количества информации является актуальной с точки зрения развития методик реферирования текстов.

Объект и предмет исследования

Объектом исследования являются тексты различной природы, в частности, тексты новостных электронных изданий и их характеристики.

Предмет исследования - модели, основанные на количественных статистических характеристиках текста, и методы реферирования текстовых данных с использованием этих моделей.

Целью работы является исследование статистической модели текстовых данных на основе онтологического подхода и экспериментальное

подтверждение ее применимости для решения задачи автоматизированного реферирования текстовых данных различной природы.

Задачи исследования

Для достижения поставленной цели были поставлены следующие задачи:

- анализ существующих моделей и методов, применяемых для реферирования текстов;
- исследование онтологического подхода с обоснованием возможности его применения при построении рефератов текстовых документов;
- разработка методики реферирования текстов на основе рассмотренных подходов;
- экспериментальная проверка эффективности разработанной методики

Методы исследования

Для решения поставленных задач применялись методы системного анализа, теории информации и методы статистической обработки экспериментальных данных.

Научная новизна

Предложена модель текста в виде статистических характеристик текстовых данных на основе онтологического подхода, с помощью которой могут быть описаны тексты различной природы.

Достоверность и обоснованность результатов исследования обеспечены корректным использованием современного математического аппарата, использованием результатов исследований в области автоматической обработки текстов, в частности, автоматизированного реферирования текстовых данных, и подтверждены результатами экспериментальных исследований на примере конкретных текстов.

Практическая значимость работы

Полученные теоретические и практические результаты в ходе написания выпускной квалификационной работы наглядно отражают важность этапа текстового моделирования в задаче автоматизированного реферирования. Реализован алгоритм построения модели текста на основе онтологического подхода, применимый к произвольным текстовым данным. Показана применимость данного подхода для решения задачи реферирования текстов.

Заключение

В результате проведённого исследования, были выполнены все поставленные задачи:

- доказана актуальность применения алгоритмов реферирования для решения задачи автоматической обработки текстовых данных;
- был обоснован выбор статистического алгоритма реферирования;
- была предложена модель представления текстовых данных с учетом онтологии;
- построен статистический алгоритм реферирования на базе предложенной модели;
- была произведена оценка сложности разработанного алгоритма;
- выполнена программная реализация модифицированного статистического алгоритма;
- произведено тестирование работы программы на реальных данных;
- по результатам экспериментов был выполнен сравнительный анализ разработанного алгоритма на основе онтологического подхода и алгоритмов, не учитывающих онтологию.

Перспективным является использование разработанного статистического алгоритма в совокупности с позиционными и индикаторными методами реферирования текстов. Это позволит наряду со статистическими данными получить дополнительную информацию о структуре документа и учитывать вспомогательные лексические единицы при обработке текстовых данных на естественном языке.