В. Г. Васильев (Москва, ИПИ РАН). **Комплексная обработка результатов классификации текстов.**

Во многих организациях для автоматизации обработки потоков текстовых данных используются средства автоматической классификации текстов. Например, с их помощью осуществляется подготовка различных обзоров и дайджестов путем отбора текстов по интересующим тематикам. В докладе рассматривается комплексное использование различных методов кластерного анализа для автоматизированной подготовки итоговых обобщенных документов. В частности, предлагается метод тематического упорядочения текстов, а также подходы к выявлению «почти дубликатов» и основных тем в подборке документов.

Метод тематического упорядочения осуществляет нахождение такой последовательности номеров документов, чтобы сумма расстояний между соседними документами была минимальная. Это соответствует задаче коммивояжера, которая является NP-полной. Для ее эффективного решения был разработан приближенный алгоритм, основанный на рекурсивной процедуре обработки результатов иерархического кластерного анализа. Показывается, что построенный алгоритм имеет квадратичную сложность, а также приводятся результаты экспериментов по оценке его качества работы.

Задача выявления «почти дубликатов» сводится к нахождению таких компактных групп документов, в которых без существенного потери информации можно было бы оставить только одного представителя. Для этого предлагается использовать стандартные алгоритмы иерархического кластерного анализа со специальными мерами близости, ориентированными на различные варианты размещения похожей информации в сравниваемых документах.

Задача выявления основных тем сводится к нахождению небольшого числа крупных, наиболее представительных групп документов. Для ее решения предлагается использовать метод вероятностного кластерного анализа на основе смеси вероятностных анализаторов главных компонент [1], дополненный итерационной процедурой очистки документов от избыточной информации.

В заключительной части доклада рассматривается комплексная процедура обработки результатов классификации с использованием предлагаемых методов, приводятся примеры автоматически сформированных обзоров, отмечаются особенности программной реализации [2].

Работа выполнена при поддержке гранта Президента МК-12.2008.10.

СПИСОК ЛИТЕРАТУРЫ

1. *Кривенко М. П., Васильев В. Г.* Модели кластерной структуры массивов текстовых данных. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 3, с. 743.
2. *Васильев В. Г.* Комплексная технология автоматической классификации текстов. — Компьютерная лингвистика и интеллектуальные технологии. М.: РГГУ, 2008, в. 7(14), с. 83–90.