Лабораторная работа №3
«Технологии обработки, автоматизированного реферирования и аннотирования текстов на естественном языке»

1. Цель работы

Изучить методы и средства автоматизированного реферирования и аннотирования текстов на естественном языке, а также получить навыки работы с подобными системами.

2. Подготовка к работе

Изучить основные понятия технологии обработки, автоматизированного реферирования и аннотирования текстов на естественном языке [1].

Установить систему автоматического анализа текста *TextAnalyst v*2.0, воспользовавшись ссылкой *http://www.analyst.ru* или распаковав архив *ta201rus_eval.zip* в папке «8 - Дополнительные материалы» и запустив для инсталляции программы файл *setup.exe*. Системные требования *TextAnalyst v*2.0: операционная система *Microsoft Windows* 7 (32 *bit*)*, Microsoft Windows Vista* (32 *bit*)*, Microsoft Windows XP.*

3. Лабораторное задание

1. Выполнить автоматическое реферирование текстов различной тематики (хорошо структурированных технических текстов, художественной литературы, прозы, стихотворений и т.д.) с помощью системы *TextAnalyst v*2.0.
2. Построить сеть понятий и тематическую структуру для заданного текста. Произвести автоматическое реферирование текста с разными коэффициентами сжатия. Привести примеры смыслового поиска по тексту и работы со словарями.
3. В отчет по лабораторной работе включить исходный текст для реферирования, выводы по работе с системой автоматического реферирования и другую дополнительную информацию.
4. Подготовить отчет для защиты лабораторной работы №3.

4. Требования

- Подготовить несколько текстов различной тематики в формате *txt* не менее 10-20 страниц.
- В отчет дополнительно необходимо включить следующую информацию: сеть понятий; тематическую структуру текста; результаты реферирования текста с различными коэффициентами сжатия; результаты смыслового поиска; сформированный гипертекстовый документ; результаты работа со словарями.

5. Методические указания

Лабораторная работа выполняется с помощью системы *TextAnalyst*. Кратко рассмотрим возможности *TextAnalyst v*2.0.

Сеть понятий

Сеть понятий - это множество терминов из текстов - слов и словосочетаний, связанных между собой по смыслу. В сеть включены не все термины из текста, а лишь наиболее значимые, несущие основную смысловую нагрузку. Аналогичным образом представлены и смысловые связи между понятиями. Поэтому, с одной стороны сеть достаточно полно описывает смысл текстов, а с другой - позволяет отбросить несущественную информацию и представить содержание в сжатом виде. Также собирается информация по смысловым связям каждого понятия – в виде списка всех связанных с ним в тексте понятий, дополненного предложениями, в которых отражаются данные связи.

Таким образом, можно сразу увидеть всю информацию по каждому понятию (рис. 1).

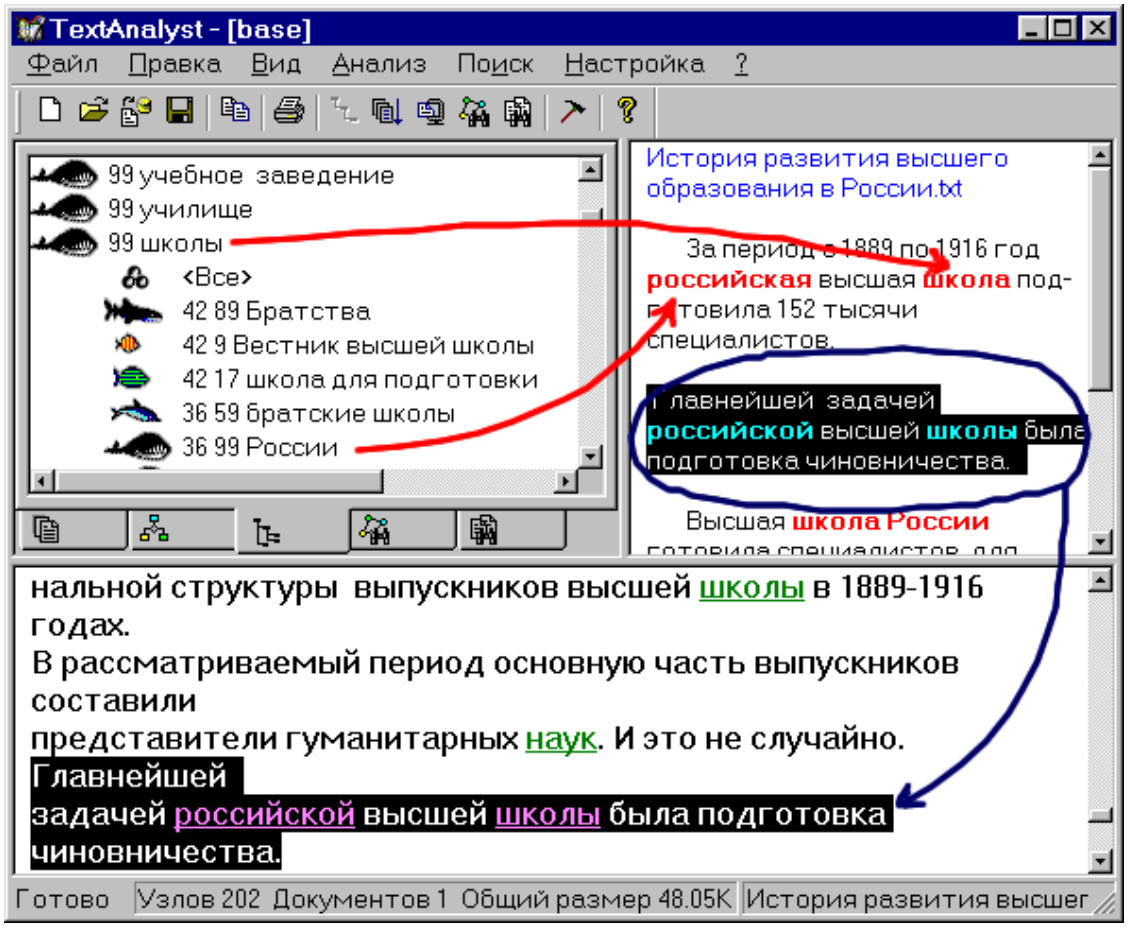

Рис. 1. Основное окно *TextAnalyst v*2.0.

Каждый элемент сети - понятие характеризуется числовой оценкой (весом). Связи между парами понятий, в свою очередь, также характеризуются весами. Эти оценки позволят сравнить относительный вклад различных понятий и их связей в семантику текста, выявить более или менее подробно проработанную в тексте тематику, задать способ сортировки информации, и наконец, позволят взглянуть на весь текстовый материал по пластам - смысловым срезам различной глубины.

Тематическую структуру текста

Тематическая структура описывает содержание анализируемых текстов в виде иерархии связанных тем. Все темы выражены в терминах исходных текстов и соответствуют узлам сети понятий. Представление тематической структуры является иерархическим. Тематическая структура, таким образом, имеет вид древа, в корне которого стоят главные темы, а в ветвях – их подтемы. Общий вид тематической структуры отражает смысловую структуру текстов. Так, если вся информация в текстах подчинена единой теме, структура будет иметь вид дерева с единственным корнем. Если же содержание текстов отражает несколько тем, то дерево распадается на целый "лес" независимых кустов, корни которых представляют главные темы, несвязанные друг с другом.

TextAnalyst дает возможность регулировать степень связности тематического дерева. Изменение порога по весу связей в сети понятий (разрыв более или менее сильных связей) изменяет вид дерева. В результате появляется возможность взглянуть на структуру текста в различных срезах, на разных уровнях глубины материала.

В остальном, с точки зрения интерфейса работа с тематической структурой полностью аналогична работе с семантической сетью.

Реферирование текста

Для самого быстрого ознакомления с содержанием текстов можно воспользоваться услугой автоматического реферирования, предоставляемой *TextAnalyst*. Формируемый реферат содержит список наиболее информативных предложений. Кроме того, все предложения реферата снабжены ссылками к соответствующим местам исходных текстов, что позволяет просмотреть контекст интересующего предложения. Подробность реферата можно легко настраивать, изменяя количество формирующих его предложений. При этом каждое предложение реферата характеризуется относительной степенью значимости во всем тексте.

Смысловой поиск

Функция **смыслового поиска** позволяет получить ответ на запрос, сформированный в виде фразы естественного языка, словосочетаний или же просто набора ключевых слов. При этом извлекаемая в ответ информация может не только иметь другую грамматическую форму, но и вообще не упоминаться в тексте запроса, однако имеет с ним смысловую связь.

Можно ввести запрос с клавиатуры, либо задать его участком текста, что реализует гипертекстовые ссылки.

Результаты ответа на запрос отображаются на экране в виде двух списков (рис. 2).

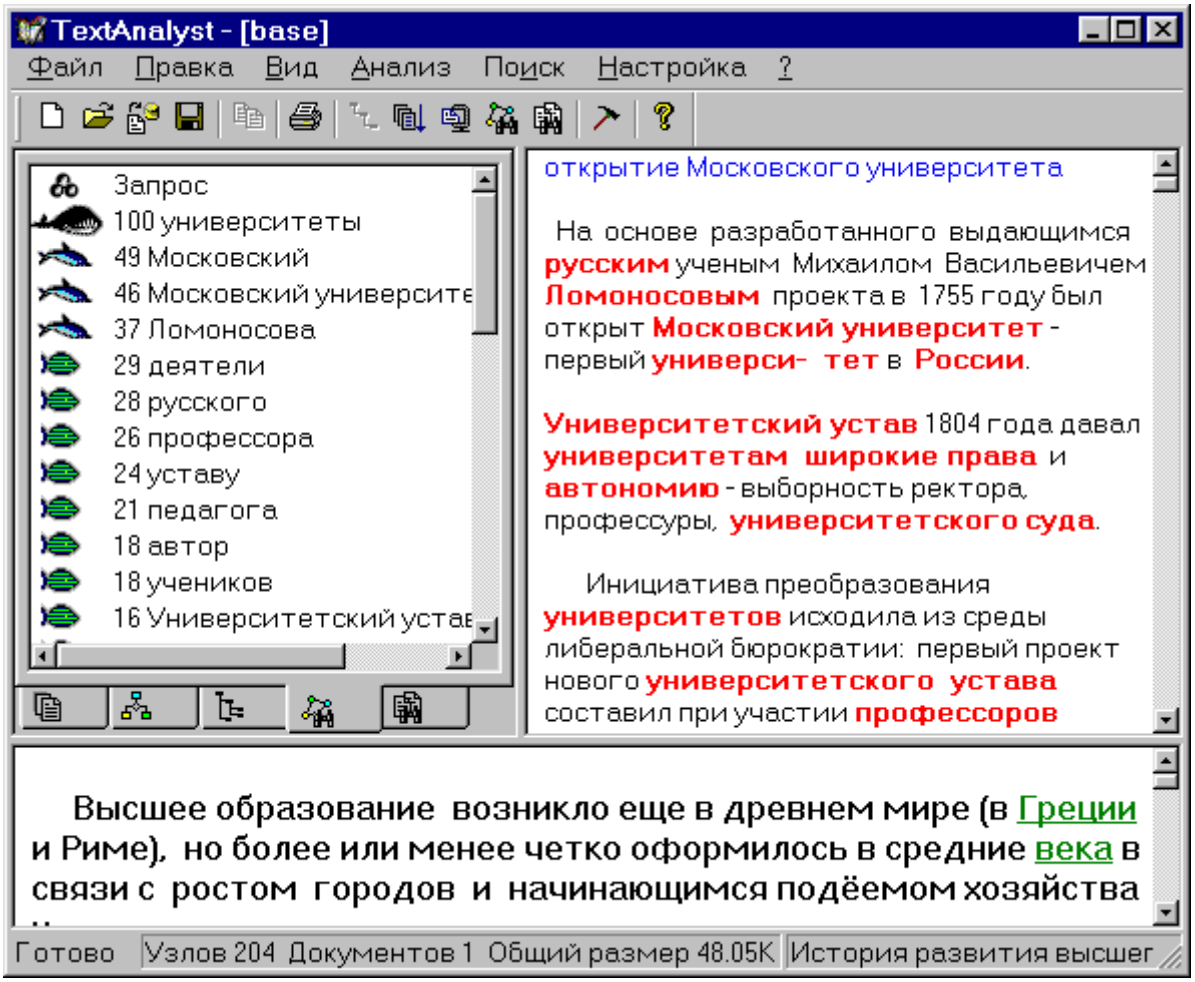

Рис. 2. Смысловой поиск в *TextAnalyst v*2.0.

Работа со словарями

При анализе текстов используется набор словарей, которые можно настраивать на конкретную предметную область.

В состав стандартной версии *TextAnalyst* входят два словаря общей лексики для русского и английского языков – “*NormalRussian.dic*” и “*NormalEnglish.dic*” соответственно. Возможно редактировать и пополнять

данные словари, а также создавать свои собственные, подгружая их из обычного текстового файла.

Любой словарь содержит четыре подсловаря (некоторые могут быть пустыми):

- Словарь удаляемых слов
- Словарь общей лекции
- Словарь слов–предпочтений пользователя
- Словарь слов-исключений.

Литература

1. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учеб. пособие. –М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.

Контрольные вопросы

1. Чем отличается реферат от аннотации?
2. Какие требования предъявляются к реферату?
3. Перечислите виды рефератов.
4. Перечислите технологии, реализуемые в *TextAnalyst* 2.0.