application server) дает возможность наряду с достаточно ограниченными интерфейсными возможностями WEB-браузера использовать богатый набор интерфейсных средств Oracle Forms для разработки модуля выборок и запросов и служебных модулей системы.

Разработка специализированного модуля обмена данными и стандартов обмена на базе XML-TEI должна обеспечить возможность обмена данными с приложениями, не использующими непосредственного соединения с базой данных «Манускрипт».

Список литературы

Баранов и др. 2003 — *Баранов, В. А.* Электронные издания древних письменных памятников и технология создания полнотекстовых баз данных / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Круг идей: электронные ресурсы исторической информатики : тр. VIII конф. Ассоциации «История и компьютер» / под ред. Л. И. Бородкина, В. Н. Владимирова. — М. ; Барнаул : Изд-во Алт. ун-та, 2003. — С. 234–270.

Баранов и др. 2004 — *Баранов, В. А.* Информационно-поисковая система «Манускрипт»: новые возможности электронного издания древнерусских рукописей / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Информационный бюллетень Ассоциации «История и компьютер». № 32 : Материалы IX конф. АИК (апр. 2004 г.) — Москва ; Томск : Изд-во Том. ун-та, 2004. — С. 7–9.

Baranov et al. 2004 — Victor Baranov, Andrey Votintsev, Roman Gnutikov, Aleksey Mironov, Sergey Oshchepkov, Vitaliy Romanenko. Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases // EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond). Conference Proceedings. — University College London. Institute of Archaeology. Principal Editor: James Hemsley. — London, 2004. — 11.1–11.8.

Об одном методе статистической фильтрации текстовой информации

С. В. Моченов, А. М. Бледнов, Ю. А. Луговских
Ижевский государственный технический университет, Россия

The paper presents a review of the statistic methods of analysis of texts in natural languages and shows the possibilities of filtration of the text information on the basis of one of the most informative statistic text characteristics. There is developed a method of filtration of the texts in the Russian language that can be applied to the solution of various tasks of text processing like the decrease of the volume of textual information, writing essays, determination of the semantic component of the text units etc.

Введение

Проблеме анализа и синтеза текстовых документов посвящено значительное количество работ [Математические 1977; Фоменко 1980; Носовский и др. 1989]. Среди них значительное место занимают работы, связанные с задачами автоматического анализа полнотекстовых документов, автоматической классификацией и идентификацией тем документов, автоматическим реферированием, выявлением смысловых связей и др.

Статистические методы достаточно хорошо зарекомендовали себя при построении поисковых систем, выделении ключевых слов и словосочетаний и т. п.

В то же время при решении задач анализа и синтеза текстовой информации, возникающих при построении информационных систем, в частности при формировании профессиональных баз знаний, требуется привлечение алгоритмически более сложных процедур проведения синтаксического и семантического анализа.

В данной статье проводится обзор статистических методов анализа текстов на естественных языках, а также показаны возможности фильтрации текстовой информации на основе одной из наиболее информативных статистических характеристик текста.

1. Краткий обзор статистических методов исследования текстовой информации

Широкое распространение статистические методы нашли в гуманитарных областях знаний, в частности, в истории и литературе [Математические 1977]. Например, метод статистического исследования текста был применен для выявления авторства литературного произведения «Тихий Дон» М. В. Шолохова [Фоменко 1985]. Большинство этих методов используют одни и те же статистические характеристики текста и основываются на понятии «авторского инварианта» [Фоменко 1983].

При этом под авторским инвариантом понимается некоторое множество статистических количественных показателей, с помощью которых можно однозначно охарактеризовать произведения одного автора или небольшого числа «близких авторов». В то же время предполагается, что количественные показатели могут принимать существенно разные значения для произведений разных групп авторов.

Однако многообразие грамматических структур, участвующих в формировании литературных текстов, сильно затрудняет поиски таких инвариантов. Установлено, что при написании литературного произведения существенную роль играют как сознательные, так подсознательные факторы, которые определяют общий стиль произведения. Например, частота употребления редких и иностранных слов может служить некоторым показателем стиля автора и его эрудиции.

Очевидно, что количественная оценка индивидуальных отличительных особенностей произведения того или иного автора — весьма нетривиальная задача.

При выборе тех или иных статистических характеристик, используемых в качестве инвариантных, необходимо, чтобы они удовлетворяли следующим требованиям:

– *интегральности* (обобщение по совокупности группы показателей);

– *постоянству* (для определенного автора на группе произведений);

– *диапазону* изменений (для групп авторов).

Такое сочетание всех трех перечисленных условий позволяет говорить о наличии некоторого авторского инварианта.

В качестве количественных характеристик текстов предлагаются следующие:

– длина предложений (среднее число слов в предложении, подсчитанное для каждой выборки);

– длина слов (среднее количество слогов в слове, подсчитанное для каждой выборки);

– частота употребления служебных слов — предлогов, союзов, частиц;

– частота употребления существительных;

– частота употребления глаголов;

– частота употребления прилагательных;

– частота употребления предлога «в»;

– частота употребления частицы «не»;

– количество служебных слов в предложении (среднее число союзов, предлогов и частиц в предложении для каждой выборки).

Следует отметить, что в указанных работах [Харин] приведенные статистические характеристики использовались лишь для оценки принадлежности того или иного литературного произведения конкретному автору и для выявления интегральных числовых особенностей различных групп авторов. В то же время при решении задач анализа важным является сокращение объема исходной текстовой информации без потери ее семантической составляющей.

Ниже предлагается алгоритм фильтрации, основанный на использовании одной из выше описанных характеристик.

2. Алгоритм фильтрации

В работах [Моченов и др. 2004] при построении статистической информационной модели использовались тексты по экономической, экологической и правовой тематике. Все тексты были примерно одинакового объема (по 10 страниц машинописного текста формата А4, содержащих порядка 300 предложений каждый).

Предметом данных исследований является выявление обобщенных статистических характеристик распределения количества слов в предложениях на данной выборке с целью выполнения направленной фильтрации по всему объему текста. Направленная фильтрация предполагает решение следующих задач:

– уменьшение общего объема анализируемого текста при сохранении его семантической составляющей;

– автоматическое реферирование;

– выявление специфических фрагментов текста, удовлетворяющих заданным критериям;

– подготовка текста для дальнейших этапов анализа.

Выявленные статистические характеристики могут быть использованы и при решении задач синтеза текста.

В процессе исследований были выполнены эксперименты, направленные на изучение роли предложений разной длины. На рис. 1 показано типичное распределение для текстов по гуманитарной тематике.

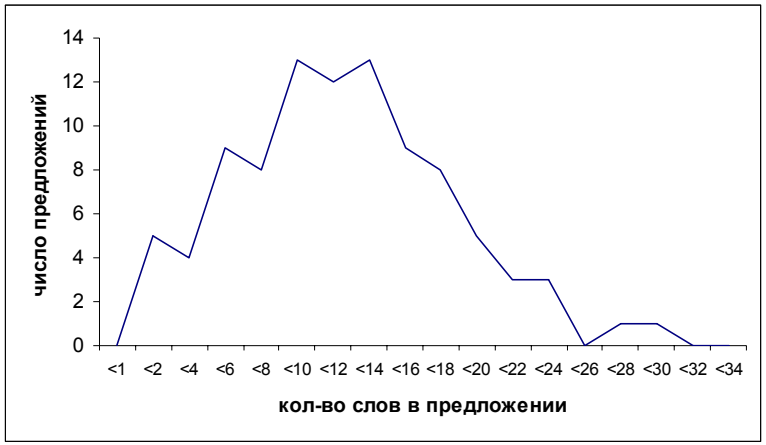

Рис. 1. Распределение предложений по количеству слов в предложении

На графике по оси ординат отложено количество предложений, а по оси абсцисс — интервалы групп по числу слов в предложении.

Исследования показали, что для данных текстов большая часть предложений приходится на предложения средней длины: от 10 до 16 слов. Короткие предложения представляют, как правило, либо заголовки, либо некоторые выводы, заключения, смысловой итог абзаца. Предложения большей длины (от 17 слов и более) либо уточняют некоторый смысл предыдущего предложения, либо являются «лирическим» отступлением автора от основной тематики на произвольную тему.

Направленная фильтрация предполагает использование статистических характеристик распределений, таких, как математическое ожидание, среднее квадратическое отклонение при проведении анализа текста.

Для целей направленной фильтрации могут быть использованы и другие обобщенные характеристики, о которых шла речь выше.

Заключение

Статистические методы анализа русскоязычных текстов могут с успехом применяться для решения разнообразных задач обработки текстов. С помощью организованной, управляемой фильтрации можно изменять объем текста, сохраняя его семантическую составляющую. Кроме того, управляемая фильтрация позволяет оперировать элементами текста, например, главой, страницей, абзацем.

Список литературы

Математические 1977 — Математические методы в историко-экономических и историко-культурных исследованиях. — М., 1977.

Математические 1985 — Математические методы и ЭВМ в исторических исследованиях. — М., 1985.

Моченов и др. 2005 — *Моченов, С. В.* Применение статистических методов при анализе текстовой информации / С. В. Моченов, А. М. Бледнов, Ю. А. Луговских. — Ижевск: НИЦ «Регулярная и хаотическая динамика», 2005.

Носовский и др. 1989 — *Носовский, Г. В.* Статистические дубликаты в упорядоченных списках с разбиением / Г. В. Носовский, А. Т. Фоменко // Вопросы кибернетики. Семиотические исследования. — М., 1989. — С. 138–148. — (Науч. совет по комплексной проблеме «Кибернетика», АН СССР).

Фоменко 1980 — *Фоменко, А. Т.* Некоторые статистические закономерности распределения плотности информации в текстах со шкалой / А. Т. Фоменко // Семиотика и информация. — Вып. 15. — М. : ВИНИТИ, 1980. — С. 99–124.

Фоменко 1983 — *Фоменко, А. Т.* Авторский инвариант русских литературных текстов // Методы количественного анализа текстов нарративных источников / А. Т. Фоменко. — М. : Институт истории СССР (АН СССР), 1983. — С. 86–109.

Фоменко 1985 — *Фоменко, А. Т.* Информативные функции и связанные с ними статистические закономерности / А. Т. Фоменко // Статистика. Вероятность. Экономика. — Т. 49. — М. : Наука, 1985. — С. 335–342.

Харин — *Харин, Н. П.* Исследование принципов семантического поиска текстовой информации на основе использования интеллектуальных и статистических методов / Н. П. Харин ; МАДИ. — М., 2002.