ным физическим воздействиям // Молодежный инновационный форум Приволжского федерального округа. – Ульяновск: УлГТУ, 2009. С.183. 9. Гречихин Л.И. Наночастицы и нанотехнологии. Мн., 2008. 10. Витязь П.А., Подлозный Э.Д., Гречихин Л.И. Нанотехнология производства упрочненного бетона с различной матрийей композита //Вести Национальной академии наук Беларуси, 2010, № 1, С. 5-13. 11. Радциг А.А., Смирнов Б.М. Справочник по атомной и молекулярной физике. М., 1980. 12. Гречихин Л.И., Иващенко С.А. //Весці НАН Беларусі. Сер. фіз. – тэх. навук, 2002. № 4. С. 11 - 17. 13. Скребцов А.М. //Известия вузов. Черная металлургия. 2009. № 2. С. 28 - 31.

УДК 004.912

Романюк Г.Э., Мардас Д.В., Журавский А.О.

ПРОГРАММА ПОДСЧЕТА ЧАСТОТЫ ВХОЖДЕНИЯ СЛОВ В ТЕКСТЕ НА БЕЛОРУССКОМ И ПОЛЬСКОМ ЯЗЫКАХ

Белорусский национальный технический университет
Минск, Беларусь

Сегодня параллельно с бурным развитием технологий, общества, маркетинговых программ увеличивается объем информационных потоков. Успешное ведение бизнеса невозможно без непрерывного контроля за рыночной средой. С помощью получения, обработки информации и ее управления менеджеры компаний могут узнать об изменении потребностей покупателей, новых шагах конкурентов, состоянии каналов сбыта.

Несмотря на широкое использование мультимедиа, текст остается одним из основных видов информации в большинстве электронных хранилищ. Разработка эффективных подходов к обработке текстов с целью фильтрации, формирования смыслового портрета, навигации по базе текстов является одним из наиболее актуальных направлений современных информационных технологий.

В связи с развитием информационных ресурсов сети Интернет документальное информационное пространство развилось до такого уровня, который требует новых подходов. Рост объемов информации и скорости ее распределения фактически породил понятие информационных потоков. Вместе с тем, математический аппарат и инструментальные средства уже не всегда способны адекватно отражать ситуацию, речь идет не столько об анализе конечных массивов документов, сколько о навигации в документальных информационных потоках [1].

Большое значение приобретает такое понятие как "преобразование информации в знания". Этому в значительной мере способствуют чисто прикладные успехи в машинной обработке потоков данных, содержащих документы, не только составленные на разных языках, но и относящихся к различным социокультурным контекстам. Понятно, что в таком случае обработка потока данных (т. е. информации в чистом виде), какой бы она ни была, не предполагает активного использования содержания документов. Практика показывает, что информация может вполне успешно обрабатываться вне зависимости от того, какой смысл в нее заложен. В связи с этим возникает интерес к подходам, основанным на статистической обработке текста.

Наиболее широко используемым способом поиска нужной информации в Интернете является метод с использованием поисковых систем, но в то же время он является и наиболее сложным. Его широкая распространенность обусловлена тем, что поисковые системы содержат в себе индексы громадного количества сайтов и при правильно сформированном запросе можно сразу же получить ссылки на интересующие ресурсы. Сложность метода состоит в том, что для того, чтобы результат был качественным, необходимо уметь выбрать наиболее подходящие поисковые системы, правильно формулировать запросы к ним, учитывать их особенности и функциональные возможности.

Двоякая характеристика данного метода связана с тем, что проведение эффективного поиска требует одновременного решения двух противоположных задач: увеличении охвата с целью извлечения максимального количества значимой информации и уменьшении охвата с целью миними-

зации шумовой информации. Нетрудно увидеть, что одновременно осуществить и то и другое довольно сложно, хотя найти оптимальное соотношение все-таки возможно [2].

Для эффективного использования поисковых серверов, прежде всего необходим список ключевых слов, организованный с учетом семантических отношений между ними, то есть тезаурус.

Одним из подходов к составлению тезауруса может стать использование законов Ципфа. Рассмотрим их более подробно.

Число, показывающее сколько раз встречается слово в тексте, называется частотой вхождения слова. Если расположить частоты по мере убывания и пронумеровать, то порядковый номер частоты называется рангом частоты. Вероятность обнаружения слова в тексте равно отношению частоты вхождения слова к числу слов в тексте:

$$C = \frac{f \cdot r}{n},$$

где f — частота вхождения слов, r — ранг частоты, n — число слов.

Это значит, что график зависимости ранга от частоты представляет из себя равностороннюю гиперболу (рис. 1).

Ципф также установил, что зависимость количества слов с данной частотой от частоты постоянна для всех текстов в пределах одного языка и также является гиперболой [3].

Исследование вышеуказанных зависимостей для различных текстов показали, что наиболее значимые слова текста лежат в средней части диаграммы, так как слова с максимальной частотой, как правило, являются предлогами, частицами, местоимениями, в английском языке — артиклями (так называемые «стоп-слова»), а редко встречающиеся слова в большинстве случаев не имеют решающего значения. Таким образом, данная особенность может помочь правильно выбрать ключевые слова для проведения поиска информации.

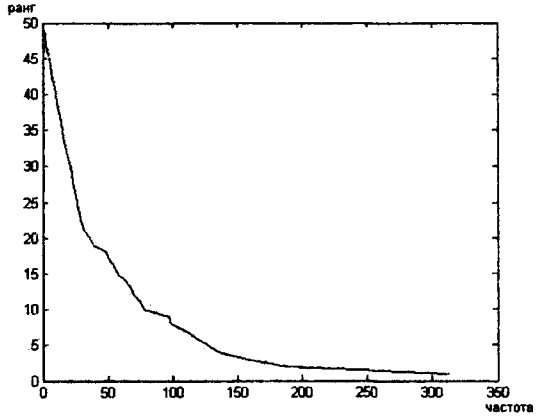

Рис. 1. График зависимости ранга от частоты вхождения слова

Процедура оптимального выбора ключевых слов, основанная на применении законов Ципфа, заключается в следующем: берут любой текст-источник, близкий к искомой теме, то есть «образец», и анализируют его, выделяя значимые слова. В качестве текста-источника может служить книга, статья, web-страница, любой другой документ. Анализ текста производится в следующем порядке:

- «стоп-слова» удаляются из текста;
- вычисляется частота вхождения каждого слова и составляется список, в котором слова расположены в порядке убывания их частоты;
- выбирается диапазон частот, лежащий в середине списка, и из него отбираются слова, наиболее полно соответствующие смыслу текста;
- составляется запрос к поисковой машине в форме перечисления отобранных таким образом ключевых слов, связанных логическим оператором OR(ИЛИ) Запрос в таком виде позволяет обнаружить тексты, в которых встречается хотя бы одно из перечисленных слов.

Число документов, полученных в результате поиска по этому запросу, может быть огромно. Однако, благодаря ранжированию документов, то есть расположению их в порядке убывания частоты вхождения в документ слов запроса, применяемому в большинстве поисковых машин, на первых страницах найденных ресурсов практически все документы должны оказаться релевантными.

Ципф определил, что если умножить вероятность обнаружения слова в тексте на ранг частоты, то получившаяся величина приблизительно постоянна для всех текстов на одном языке. Так, например, для английских текстов константа Зипфа равна приблизительно 0,1. Для русского и украинского языков коэффициенты Ципфа составляют приблизительно 0,06 - 0,07.

Для белорусского и польского языков константа Ципфа рассчитана не была, и поэтому ее расчет представляет несомненный интерес.

Но если для русского и других языков создан ряд программ для подсчета количества слов в тексте (например, Wordstat), то для белорусского и польского языков таких программ не существовало (по крайней мере, в печати о них сообщений не поступало).

Поэтому актуальной представлялась задача разработки программы для подсчета количества слов в тексте на белорусском и польском языках. Такая программа была разработана на кафедре «Интеллектуальные системы» БНТУ студентами группы 103616 Мардасом Дмитрием Васильевичем и Журавским Алексеем Олеговичем под руководством Романюк Галины Эдуардовны.

Для упрощения проведения анализа законов Ципфа для текста на белорусском и польском языках была написана программа POL-BEL с помощью объектно-ориентированного языка программирования C#. Данная программа выполняет следующие операции:

- открывает текст на польском либо на белорусском языке;
- автоматически распознает открытый текст;
- проводит анализ открытого текста. То есть подсчитывает частоту вхождения слова в текст, подсчитывает общее количество слов в тексте, производит поиск и отображение слова, введенного с клавиатуры пользователем;
- сохраняет результат проведения анализа в текстовый документ с расширением «.TXT».

Интерфейс программы представлен на рис. 2.

Рассмотрим подробнее элементы управления и поля ввода и вывода информации написанной программы.

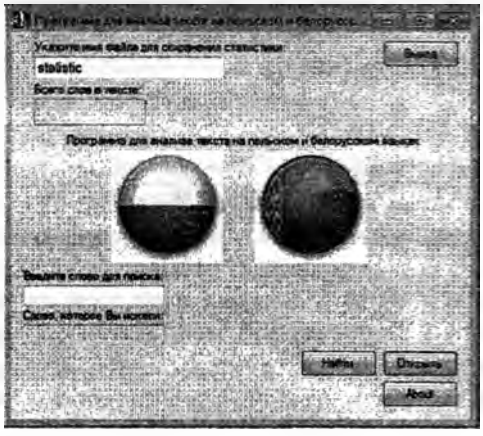

Рис. 2. Пользовательский интерфейс программы POL-BEL

Поле ввода под именем «Укажите имя файла для сохранения статистики» (рис. 2), предназначено для того, чтобы пользователь ввел имя файла, в котором будет сохранен результат анализа текста, либо если пользователь не внесет свое имя файла, то программа автоматически сохранит данный файл под именем «statistic» с расширением «.TXT».

Следующим управляющим элементом программы является кнопка под именем «Открыть» (рис. 2). Данная кнопка необходима для того, чтобы пользователь указал путь к файлу с расширением «.TXT», «.DOC», который хранит текст на польском или белорусском языке.

После того как пользователь открыл указанный им файл, программа автоматически производит анализ текста. А именно:

- автоматически распознает на каком языке написан текст (польский или белорусский язык);
- подсчитывает общее количество слов в тексте;
- подсчитывает частоту вхождения каждого слова в тексте;
- при необходимости осуществляет поиск интересующего пользователя слова и вывод его в поле под именем «Слово, которое Вы искали»;
- сохраняет результат проведения анализа текста.

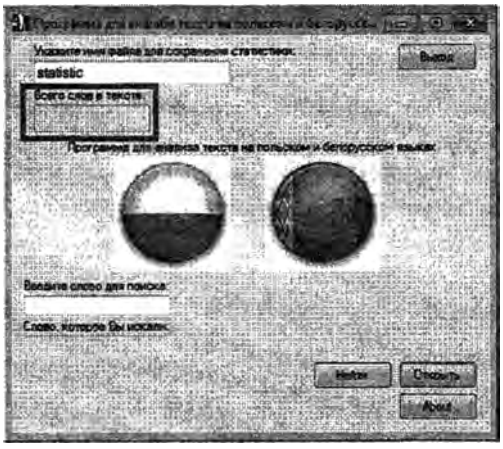

Рис. 3. Пользовательский интерфейс

Рис. 4. Окно About

На рис.3 прямоугольником выделено поле под именем «Всего слов в тексте», которое отображает сколько всего слов содержит данный текст.

В нижней части окна программы имеется два поля (рис. 3):

- поле ввода «Введите слово для поиска», в которое пользователь при необходимости может ввести интересующее его слово, которое содержится в данном тексте.
- поле вывода «Слово, которое Вы искали», в котором отображается слово, введенное пользователем в поле под именем «Введите слово для поиска» и его частота. Если пользователь ввел слово, которого нет в тексте, то данное поле останется пустым.

Для того, чтобы программа произвела поиск слова, введенного пользователем с клавиатуры, необходимо нажать кнопку «Найти» (рис. 3).

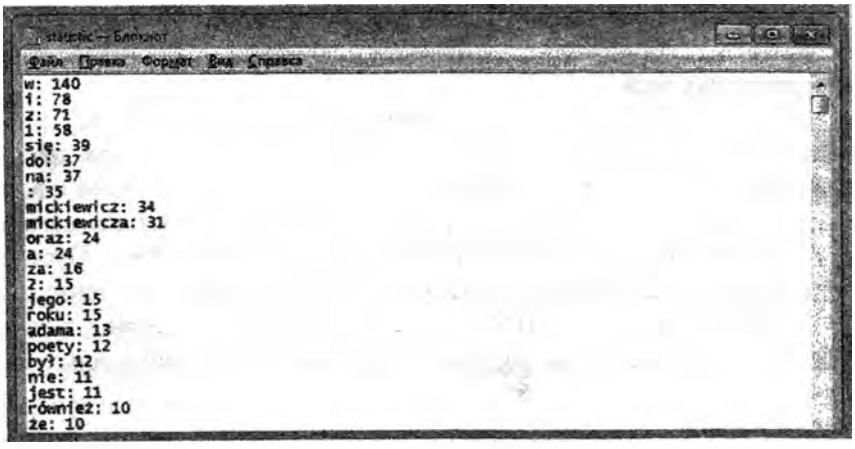

Рис. 5. Результат выполнения программы

На рис. 3 в верхнем правом углу находится кнопка под названием «Выход», при нажатии на которую происходит выход из программы POL-BEL, предварительно сохранив файл с выполненным анализом текста с расширением «.TXT».

На рис. 3 в нижнем правом углу прямоугольником выделена кнопка под названием «About», при нажатии на которую появляется окно About (рис. 4). В данном окне представлена краткая информация о разработчике программы POL-BEL, для чего данная программа предназначена и электронный почтовый адрес разработчика, на который можно присылать благодарности и замечания после использования программы POL-BEL.

Результат проведения анализа текста на польском языке «Adam Mickiewicz» представлен на рис. 5. Текст «Adam Mickiewicz» включает в себя 3096 слов или 20285 символов без учета знаков препинания, или 23287 символов с учетом знаков препинания и пробелов. В файле в колонку записано слово и частота вхождения данного слова в текст.

ЛИТЕРАТУРА

1. Ландэ Д.В., Литвин А.Б. Феномены современных информационных потоков // "Сети и бизнес". - 2001. - N 1. - С. 14-21 2. Успенский И.В. Интернет-маркетинг. Учебник.-СПб.: Издательство СПГУЭиФ.-2003.- 92с. 3. [Электронный ресурс] / Официальный сайт Wikipedia; Режим доступа: http://ru.wikipedia.org/wiki/Закон_Ципфа, - свободный. – Загл. С экрана. – Яз. Рус.- 2010

УДК 537.311.322

Соколова К.Г., Сунка В.Я., Трафимова Е.В.

ИЗМЕРЕНИЕ ТЕНЗОРА ПРОВОДИМОСТИ АНИЗОТРОПНЫХ ПОЛУПРОВОДНИКОВЫХ МАТЕРИАЛОВ

Белорусский национальный технический университет
Минск, Беларусь

Полупроводниковая кремниевая микроэлектроника достигла впечатляющих результатов и приступила к выпуску кристаллов с миллиардным количеством элементов в нем. Дальнейшее увеличение степени интеграции и плотности упаковки элементов на кристалле требует использования новых материалов, в. т. ч. и использование гетероструктурной полупроводниковой материалов, основанных и на анизотропных полупроводниках. Методика и оборудование для измерения электросопротивления изотропных полупроводниковых образцов различных геометрических форм и размеров, в т. ч. и микро - нано материалов хорошо разработаны [1], однако измерение составляющих тензора электросопротивления анизотропных полупроводниковых материалов остается не до конца решенной проблемой. Проанализируем известные методы измерения компонент тензора удельного сопротивления анизотропных полупроводниковых материалов с произвольной ориентацией кристаллографических осей и координатных осей.

Метод для объемных анизотропных полупроводниковых кристаллов [2] позволяет определять указанные компоненты путем двух измерений напряжения при постоянном токе через образец, требует изготовления одного образца и минимального числа точечных контактов к нему. Теоретическое обоснование метода выполнено на основе расчета электрического поля в анизотропном образце и компьютерного моделирования распределения потенциала и плотности тока. Пусть плоский образец диарсенида кадмия прямоугольной формы вырезан так, что главные оси тензора электропроводимости равные $\sigma_1 = 2600$ Ом$^{-1}$м$^{-1}$ и $\sigma_2 = 900$ Ом$^{-1}$м$^{-1}$, взаимно перпендикулярны, и составляют с границами образца угол θ, а толщина $d = 1$ мм образца значительно меньше его длины $a = 12$ мм и ширины $b = 4$ мм (рис. 1,а).