**Лабораторная  работа № 2**

**технология анализа текста и извлечения ключевых слов**

                                          1. **Цель работы**

          Практическое освоение технологии анализа текста, извлечения ключевых слов  и  профессионального  поиска информации.

                                                   2. **Общие сведения**

           2.1.   **Основные принципы выбора ключевых слов**

Знание общих принципов функционирования поисковых средств и умение грамотно составить запрос поисковой машине необходимые, но недостаточные условия успешного поиска требуемой информации. Надо еще и уметь правильно выбирать ключевые слова поиска. Это особенно ощутимо при поиске в незнакомой предметной области, поскольку выбор ключевые слов, которые должны максимально соответствовать тематическому направлению, затруднен именно незнанием специфики предметной области (заметим, что это штатная ситуация для поисковой машины). Выбор  ключевых слов в данном случае может осуществить специалист узкого профиля, но труд его дорог и малопроизводителен, или специальные программные средства, основанные на применении законов Зипфа. Джордж Зипф установил, что все тексты подчиняются  общим закономерностям, и   сформулировал в 1946—49 гг. несколько законов, которые нашли  применение  в технологии поиска информации.

Для ознакомления с положениями первого закона Зипфа введем, следуя  Зипфу, необходимые терминологические определения. Рассмотрим некоторый произвольный текст. Выпишем все различающиеся слова данного текста в виде  сОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img1.jpg, сОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img2.jpg,…,сОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img3.jpg,…,сОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img4.jpg, где  сОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img5.jpg- i-ое слово, не совпадающее ни с каким другим словом в данном множестве. Для каждого из этих слов подсчитаем  количество  его  повторов  в тексте.  В  результате получим      fОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img6.jpg,  fОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img7.jpg,…,fОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img8.jpg,…,fОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img9.jpg, где fОписание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img10.jpg- количество повторений i-го слова в тексте, названное Зипфом **частотой слова** (в данном случае i-го слова). Далее Зипф, располагая слова в порядке убывания их частот, поставил им в соответствие числа натурального ряда, назвав эти числа **рангами слов (R):** слову с максимальной частотой присваивается ранг 1,  следующему по частоте — ранг 2 и т.д.  При этом если несколько разных слов имеют одинаковые частоты, то они объединяются в один блок. Наконец Зипф ввел понятие **вероятности встречи слова** как отношение частоты слова к общему количеству слов в тексте (в математической статистике такое отношение называется частотой события; а в справочной системе поисковых машин — относительной  частотой).

  Первый закон Зипфа  утверждает, что произведение частоты встречи слова в тексте (или вероятности встречи слова по Зипфу) на его ранг есть величина приблизительно постоянная для  любых текстов определенного языка, т.е.  имеет место   C = f Описание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img11.jpgR Описание: Описание: E:\ЭРУД\ЭРУД_ОинфТ\Практика\content\lb2\img12.jpgconst.

Комментарий по поводу круглых скобок. Если левую и правую часть этого равенства разделить на общее количество слов в тексте, то равенство не нарушится, но его правая часть будет не что иное, как произведение вероятности встречи  слова  на его ранг.

На рис.1 приведена зависимость  частоты слов f от его ранга R. Кривая представляет собой гиперболу  f = C / R.

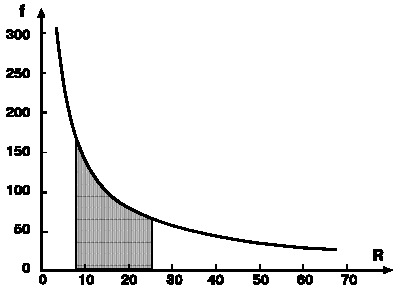


Рис. 1. График зависимости частоты слова f от его ранга R.

Характерно, что все значимые (ключевые) для текста слова размещаются в средней части графика (выделенная область рис. 1). При этом часто встречающиеся слова (ранг от 1 до 5), как правило, являются вспомогательными (это предлоги,  частицы, местоимения и т.д.). Такие слова называются «стоп-словами»; поисковая машина при простом поиске их по умолчанию игнорируют, так как эти слова являются "шумом", помехой, которая затрудняет поиск. Редко встречающиеся слова (правая часть графика) также не имеют существенного  значения при  анализе текста.

Важно уметь правильно выбрать диапазон значений ранга для извлечения ключевых слов. Если он будет слишком широким, то значимые слова  затеряются среди  второстепенных;  если же установить очень узкий диапазон,  то  возможна потеря существенных  ключевых слов.

Законы Зипфа используются при создании на поисковых серверах базы данных, в которой хранится индексированная информация; при этом учитывается целый ряд факторов, таких как вес слова, его местоположение  в документе,  морфологические особенности и др. Они же используются и для оценки релевантности (степени соответствия) документов в процессе поиска. Релевантность изменяется от 0 до 1, в зависимости от того, какое количество слов поискового выражения содержится в найденном документе, а также веса каждого из таких слов. Уточненные законы Зипфа используются также в алгоритмах автоматического распознавания текста программ-экстракторов, которые осуществляют семантический анализ текстов и извлекают ключевые слова и выражения.

* 1. **2.2. Программы-экстракторы**

        Возможность извлечения ключевых слов из текстовых материалов имеется и в текстовом редакторе MS Word (меню «**Сервис**» > «**Автореферат**»**),** однако использование этой возможности дает   неудовлетворительные результаты. Из программ-экстракторов, которых в настоящее время не так много,  одними из лучших являются   **TextAnalyst   и RCO Fact Extractor.**

**RCO Fact Extractor —** персональная аналитическая система,   предназначенная для автоматического анализа текста и выделения фактов, связанных с интересующими объектами. RCO Fact Extractor  позволяет найти в тексте описания фактов заданного типа и извлечь связанную с ними  информацию. Основная сфера применения этого программного средства —  аналитические задачи из области компьютерной разведки, требующие высокоточного отбора информации по заданным смысловым критериям, например, автоматизированное составление досье на целевые персоны или организации. Программа имеет простой интерфейс и хорошую справочную систему. К сожалению Demo-RCO Fact Extractor работает только со  встроенной коллекцией документов;  цена лицензии 2000$.

Персональная система автоматического анализа текста**TextAnalyst** предназначена для анализа содержания текстов, смыслового поиска информации и формирования электронных архивов. TextAnalyst  предоставляет пользователю следующие возможности:

* анализ содержания текста с автоматическим формированием семантической сети с гиперссылками - получения смыслового портрета текста в терминах основных понятий и их смысловых связей;
* анализ содержания текста с автоматическим формированием тематического древа с гиперссылками - выявления семантической структуры текста в виде иерархии тем и подтем;
* смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста;
* автоматическое реферирование текста - формирования его смыслового портрета в терминах наиболее информативных фраз;
* кластеризация информации - анализ распределения материала текстов по тематическим классам;
* автоматическая индексация текста с преобразованием в гипертекст;
* ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;
* автоматическое/автоматизированное формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

Как видно из перечисления, программа обладает большими возможностями, которые достаточно полно описаны во встроенном учебнике по TextAnalys, дополнительную информацию можно в  справочной системе этой программы. При выполнении лабораторной работы следует иметь ввиду, что установленная в лаборатории Demo-версия TextAnalys не позволят сохранять результаты анализа, так что их следует копировать отдельно.

         TextAnalys позволяет осуществлять эффективную семантическую обработку текстов с извлечением ключевых слов и выражений. Вот один из примеров использования этой программы. Допустим, мы располагаем текстом заинтересовавшей нас работы и хотим отыскать подобные ей  публикации. Для достижения этой цели, выберем с использованием  TextAnalys ключевые слова, составим поисковый запрос и введем его в поисковую машину. Достоинство такого подхода в том, что TextAnalys позволяет найти ключевые слова, адекватно отображающие интересующую нас тематику, а это главное условие успешного поиска.  По-видимому, эта программа может быть использована и для выявления нарушения авторских прав.

* 1. **2.3. Последовательность действий  при выборе ключевых слов**

**поискового запроса**

При выборе ключевых слов поискового запроса без использования программных средств необходимо выполнить следующее:

1. Удалить (или вычеркнуть) из выбранного текста все стоп-слова и  вычислить частоту вхождения каждого из оставшихся слов.  Заметим, что так работает поисковая машина. В учебных целях стоп-слова можно и не вычеркивать.
2. Выписать слова в порядке убывания их частоты вхождения f и присвоить словам ранги (см. п. 2.1).  В учебных целях рекомендуется  построить график зависимости частоты слова f от его ранга R.
3. Выбрать, руководствуясь здравым смыслом, диапазон значений ранга слов. При этом следует помнить, что при слишком широком диапазоне значимые слова затеряются среди  второстепенных слов, а при очень узком значимые слова  могут просто  потеряться.
4. Сформировать, используя выбранный диапазон ранга слов, список ключевых слов.  Достаточно   взять 10-20  слов.
5. Составить поисковый запрос,  используя  логический оператор  «ИЛИ» и соблюдая порядок следования ключевых слов сформированного списка. Напомним  (см. Лаб. раб. № 1), по умолчанию поисковая машина использует логический оператор «И»**,** так что если  оператор «ИЛИ**»** в поисковом запросе не указан, то при значительном количестве ключевых слов в запросе результаты поиска могут оказаться нулевыми.

На сформированный таким образом запрос поисковая машина может выдать несколько сотен миллионов страниц, но, поскольку поисковая   машина, как правило, ранжирует результаты поиска, то на первых страницах окажутся наиболее релевантные документы.

**3. Порядок выполнения работы**

         3.1. Изучите теоретический материал, изложенный в пункте 2 этой  работы.

       3.2. Включите закрепленный за Вами компьютер и ознакомьтесь с индивидуальным  заданием (папка ИНД\_ЗАДАНИЯ; откройте подпапку, последняя цифра которой совпадает с последней цифрой Вашей учебной группы, и выберите номер, соответствующей  Вашему порядковому номеру в списке группы).

         3.3. Составьте, используя  заданный текст-источник (Ваше индивидуальное задание), поисковый запрос по методике,  приведенной в п. 2.3.

        3.4. Осуществите  поиск  документов, используя сформированный согласно п. 3.3 поисковый запрос, и проанализируйте полученные результаты.

         3.5. Ознакомьтесь по встроенному учебнику с аналитической системой  TextAnalyst и постройте поисковый запрос с использованием  этой системы  по томуже тексту-источнику.

         3.6. Выполните то же, что и в п. 3.4, но по поисковому запросу, сформированному в п. 3.5.

                 3.7.  Сравните результаты,   полученные  в п.3.4 и  п.3.6.

          3.7. Оформите отчет и защитите лабораторную работу.

         3.8. Выключите компьютер и приведите в порядок рабочее  место.

**4. Содержание отчета**

          4.1.  Цель работы.

          4.2.  Краткую характеристику использованных методов построения   поискового запроса.

 4.3. Частоты вхождения и ранги слов текста-источника. График зависимости частоты слова f от его ранга R.

          4.4. Описание результатов поиска документов.

          4.5.  Анализ полученных результатов.

          4.6.  Выводы по работе.

**5. Контрольные вопросы**

            1.  Что такое  «стоп-слова»?

            2.  Как  определяется  ранг слова?

            3.  Что такое  программы-экстракторв?

           4.   Из каких соображений выбирается диапазон ранга слов?

            5. Какова  последовательность действий  при выборе ключевых слов поискового запроса?