Министерство образования Республики Беларусь

Учреждение образования  
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет информационных технологий и управления

Кафедра информационных технологий автоматизированных систем

Лабораторная работа №2

по теме

**ТЕХНОЛОГИя анализа текста и извлечения**

**ключевых слов**

Студенты Шумигай В. В.

Группа 020601

Руководитель Ярмолик В. И.

СОДЕРЖАНИЕ

[1 Цель работы 3](#_Toc114800461)

[2 Законы Зипфа 3](#_Toc114800462)

[3 Анализ теста‑источника 3](#_Toc114800463)

[3.1 Выбор ключевых слов поискового запроса без использования программных средств 3](#_Toc114800464)

[3.2 Выбор ключевых слов поискового запроса с помощью аналитической системы *TextAnalyst* 5](#_Toc114800465)

[4 Анализ полученных результатов 6](#_Toc114800466)

[5 Вывод 7](#_Toc114800467)

[Приложение А 8](#_Toc114800468)

# Цель работы

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

# Законы 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.

Частота – количество вхождений слова в тексте.

Ранг – номер слова в общем списке обнаруженных слов, упорядоченном по убыванию частоты (слову с максимальной частотой присваивается ранг 1, следующему по частоте – ранг 2 и т.д.).

Слова с одинаковой частотой, имеют одинаковый ранг.

Если произвольно выбрать случайную позицию в тексте, вероятность нахождения заданного слова именно на этой позиции обратно пропорциональна его частоте. Эта вероятность равна отношению частоты слова к общему числу слов.

Первый закон Зипфа «ранг – частота»: вероятность обнаружения любого слова, умноженная на его ранг – постоянная величина (С).

Второй закон Зипфа: «количество – частота»: если построить график, где ось X отображает частоту слова, а ось Y – количество слов, входящих в текст с такой частотой, то для любого текста этот график будет одинаковым.

# Анализ теста­‑источника

## Выбор ключевых слов поискового запроса без использования программных средств

Таблица 1 – Выбор ключевых слов

|  |  |  |
| --- | --- | --- |
| **Слова** | **Частота** | **Ранг** |
| Рынок | 24 | 1 |
| CRM | 21 | 2 |
| Клиенты | 18 | 3 |
| Компаний | 17 | 4 |
| Конкуренция | 6 | 5 |
| Предприятие | 6 | 5 |
| Офис | 5 | 6 |
| Сегментация | 5 | 6 |

Рис 1 – График зависимости частоты слов от ранга

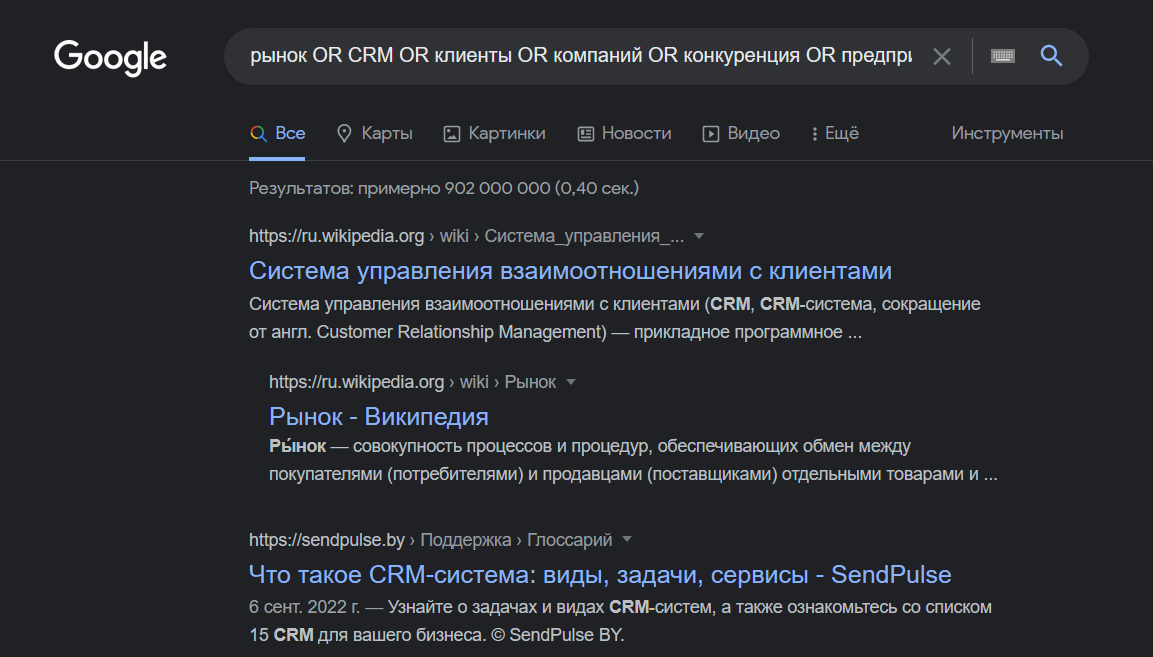


Рис 2 – Результаты поиска в Google

## Выбор ключевых слов поискового запроса с помощью аналитической системы *TextAnalyst*

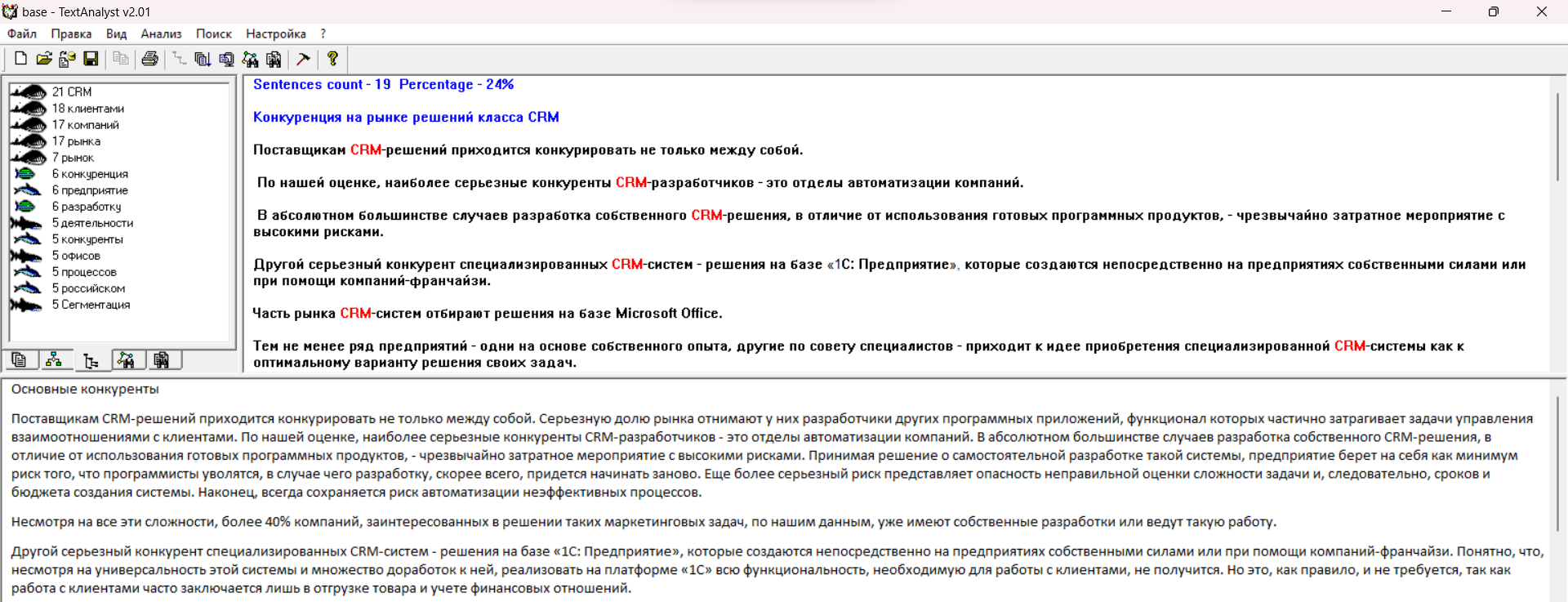
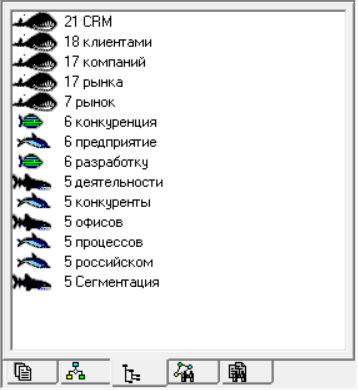


Рис 3 – Результат работы системы *TextAnalyst*

Таблица 2 – Выбор ключевых слов системой *TextAnalyst*

|  |  |  |
| --- | --- | --- |
| **Слова** | **Частота** | **Ранг** |
| CRM | 21 | 1 |
| Клиентами | 18 | 2 |
| Компаний | 17 | 3 |
| Рынка | 17 | 3 |
| Рынок | 7 | 4 |
| Конкуренция | 6 | 5 |
| Предприятие | 6 | 5 |
| Разработку | 6 | 5 |
| Деятельности | 5 | 6 |
| Конкуренты | 5 | 6 |
| Офисов | 5 | 6 |
| Процессов | 5 | 6 |
| Российском | 5 | 6 |
| Сегментация | 5 | 6 |

Рис 4 – График зависимости частоты от ранга

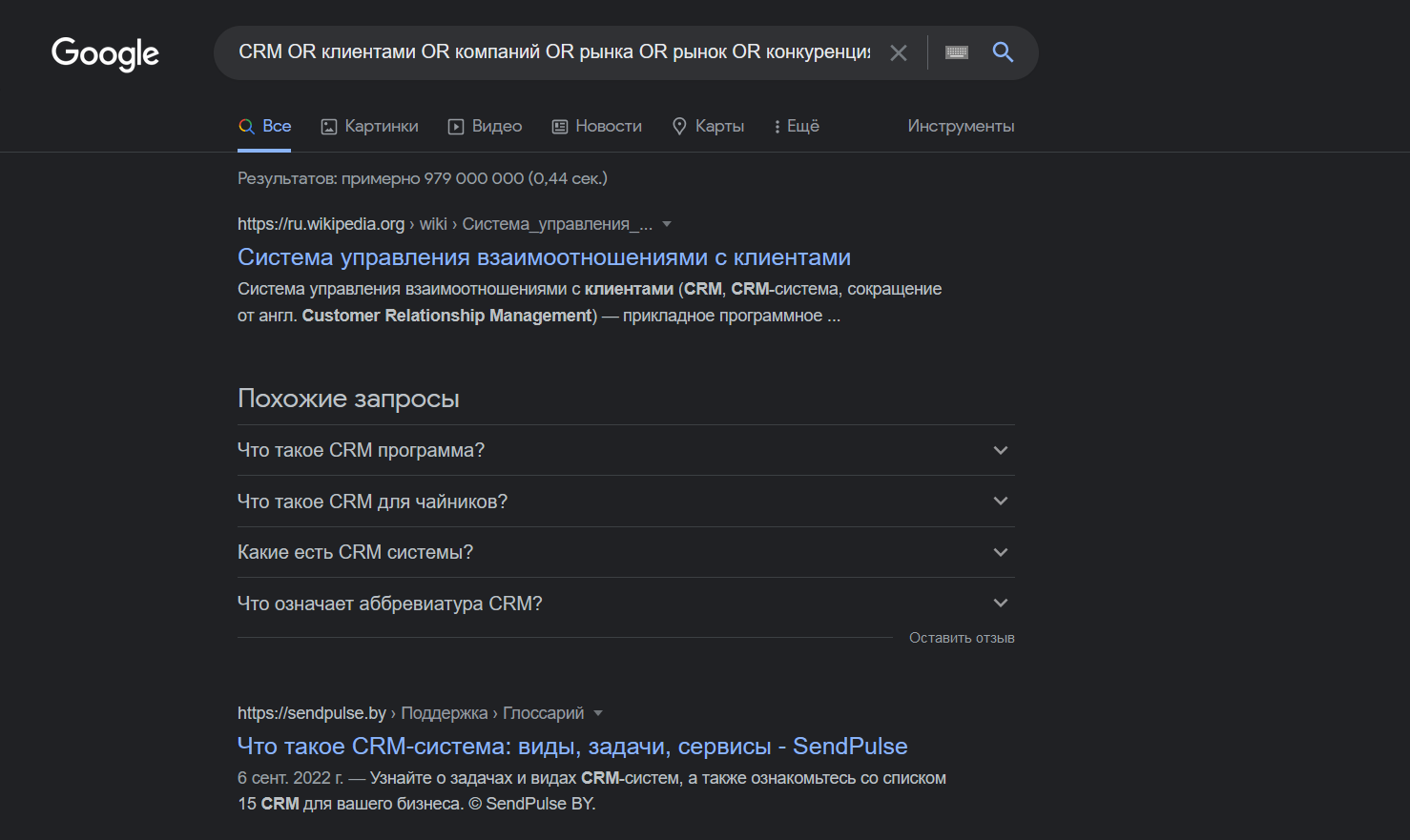


Рис 5 – Результаты поиска в Google

# Анализ полученных результатов

По результатам запроса, составленного из ключевых слов, которые были выбраны вручную из исходного текста, получили примерно те же документы, что и по запросу, составленному на основании слов, которые были выбраны аналитической системой *TestAnalyst*. Однако по выбранным ключевым словам видно, что программа не на столько универсальна, она не может учитывать всё разнообразие форм слов. Ручной же разбор текста позволяет более точно и правильно найти, и определить все формы слов.

Однако, если тест будет большим, то обработать такой текст вручную будет сложно, а в некоторых случаях невозможно. Тогда используют аналитические системы, выполняющие анализ текстов любого объема и сложности, возможно, с меньшей, но достаточной точностью.

# Вывод

По итогу работы было выявлено, что для обработки небольших текстов лучше использовать ручной разбор, он будет эффективней и запрос чаще выдаст большее количество документов. В свою очередь аналитические системы являются удобным инструментом ля обработки больших текстов, хоть и с меньшей точностью.



**(справочное)**

**Текст для выбора ключевых слов поискового запроса**

Конкуренция на рынке решений класса CRM

*Основные конкуренты*

Поставщикам CRM-решений приходится конкурировать не только между собой. Серьезную долю рынка отнимают у них разработчики других программных приложений, функционал которых частично затрагивает задачи управления взаимоотношениями с клиентами. По нашей оценке, наиболее серьезные конкуренты CRM-разработчиков – это отделы автоматизации компаний. В абсолютном большинстве случаев разработка собственного CRM-решения, в отличие от использования готовых программных продуктов, – чрезвычайно затратное мероприятие с высокими рисками. Принимая решение о самостоятельной разработке такой системы, предприятие берет на себя как минимум риск того, что программисты уволятся, в случае чего разработку, скорее всего, придется начинать заново. Еще более серьезный риск представляет опасность неправильной оценки сложности задачи и, следовательно, сроков и бюджета создания системы. Наконец, всегда сохраняется риск автоматизации неэффективных процессов.

Несмотря на все эти сложности, более 40% компаний, заинтересованных в решении таких маркетинговых задач, по нашим данным, уже имеют собственные разработки или ведут такую работу.

Другой серьезный конкурент специализированных CRM-систем – решения на базе «1С: Предприятие», которые создаются непосредственно на предприятиях собственными силами или при помощи компаний-франчайзи. Понятно, что, несмотря на универсальность этой системы и множество доработок к ней, реализовать на платформе «1С» всю функциональность, необходимую для работы с клиентами, не получится. Но это, как правило, и не требуется, так как работа с клиентами часто заключается лишь в отгрузке товара и учете финансовых отношений.

Часть рынка CRM-систем отбирают решения на базе Microsoft Оffice. Отмечу, что Microsoft выступает конкурентом почти всех специализированных под определенный класс задач приложений. Особо стоит выделить Outlook как удобный менеджер контактов, функциональности которого вполне достаточно для постановки процесса учета клиентов в компании.

Тем не менее ряд предприятий – одни на основе собственного опыта, другие по совету специалистов – приходит к идее приобретения специализированной CRM-системы как к оптимальному варианту решения своих задач.

*Сегментация рынка CRM*

Как показывает наша практика, среди CRM-решений еще нет отраслевой конкуренции (широко распространенной в случае других программных приложений), т. е. пока ни один из разработчиков не позиционирует свое решение на определенные отраслевые сегменты.

Точно так же не наблюдается и конкуренции торговых марок, когда за клиента борются функционально похожие решения и конкуренция идет на уровне известности бренда.

Нельзя также сказать, чтобы наблюдалась конкуренция между решениями по степени развитости какого-либо функционального модуля. На текущий момент клиенты не выбирают систему на основании, например, того, что в ней хорошо проработаны модули «Маркетинг» или «Сервисное обслуживание». Как правило, руководителей предприятия интересует в первую очередь автоматизация работы отдела сбыта, а потом уже все дополнительные функции, которые могут быть реализованы на основе исходно выбранного решения.

Зато, чрезвычайно распространена на российском рынке конкуренция за одного клиента между решениями разного класса и разной стоимости. В этом случае клиенту предлагаются абсолютно разные решения, и основным конкурентным преимуществом той или иной марки в итоге часто оказывается профессионализм консультантов по продаже.

На наш взгляд, рынок заказчиков, которые заинтересованы в использовании CRM-решений, можно сегментировать по двум основным критериям: развитость технологии работы с клиентами и размер рынка, на котором работает компания.

*Развитость технологии*

Мы предполагаем, что чем более развита технология работы с клиентами, тем она лучше документирована как совокупность процессов и тем более прозрачна для управления.

Хорошо проработанная технология может появиться как в результате эволюционного развития из более примитивных способов деятельности (обычно это происходит под давлением усиливающейся конкуренции на рынке, на котором работает данная компания), так и вследствие определенных правил игры, принятых на данном отраслевом рынке. Примером такого регламентированного рыночного сегмента может служить рынок финансовых услуг, требующий серьезной проверки клиента по многим показателям в процессе работы с ним.

С развитием технологии продаж у компании возрастает потребность в более детальном описании бизнес-процессов работы с клиентами (что улучшает возможность контроля и автоматизации этой деятельности), а затем в интеграции с другими системами предприятия, начиная с учетных и заканчивая полномасштабными ERP-решениями.

*Размер рынка*

Если рынок, на котором работает компания, относительно невелик, то для организации эффективной деятельности вполне достаточно одного офиса, в котором будут сосредоточены сотрудники, отвечающие за работу с клиентами. Аналогичным представляется случай, когда компания работает на нескольких географически независимых рынках и клиентура нескольких офисов или представительств не пересекается.

С ростом объема рынка компании требуется несколько офисов для обслуживания клиентов. В этом случае одним из требований к CRM-решению будет необходимость синхронизации распределенных баз данных. Заметим, что, если деятельность торговых или сервисных представителей автоматизирована, они могут синхронизировать свои данные с основной базой по приезде в офис.

На следующей стадии роста объемов деятельности компания начинает испытывать потребность в автоматизации каждого торгового представителя, с тем чтобы он мог в любой момент получить доступ к единой базе данных компании. Иными словами, компании уже не хватает нескольких стационарных отделений, и необходимо сделать каждого агента «мобильным офисом». Все это требует технологий Web-доступа.

Таким образом, всех потенциальных клиентов CRM-решений можно представить на двумерном графике, предполагая, что величина круга соответствует числу клиентов, заинтересованных в соразмерном решении.

Горизонтальная ось графика отражает экстенсивное развитие компании с точки зрения использования CRM-решений, а вертикальная – интенсивное развитие, предполагающее повышение качества работы с рынком.

Если не рассматривать CRM-решения, ориентированные на интеграцию с ERP, то можно дополнить этот график довольно интересными данными, показывающими, какие квадранты рынка CRM на сегодня заняты. Мы рассматриваем здесь только известные на российском рынке CRM-решения.

Понятно, что система Sales Expert займет в обозримом будущем всю нижнюю строчку сегментной таблицы, поскольку это естественное направление ее развития, а системы SalesLogix и Siebel, скорее всего, упрочат свои позиции на российском рынке, пользуясь накопленным мировым опытом продвижения. Можно предположить, что позиции этих трех систем не подвергнутся атаке конкурентов, так как создание еще одного недорогого CRM-бренда экономически нерентабельно, равно как и вложения в разработку сложных CRM-систем.

*Рынок CRM в России*

Если добавить данные о стоимости соответствующих решений, получится интересная картина. Выясняется, что стоимость решений принципиально не изменяется при движении по горизонтальной оси, так как «горизонтальное» нарастание функциональности идет за счет доработок/настроек одного и того же решения. Напротив, изменения по вертикали являются принципиальными, и между CRM-решениями для технологий продаж разного уровня сложности (от самых простых до интегрированных) сохраняется огромный ценовой разрыв. Можно прогнозировать, что в ближайшее время основная битва за CRM-клиентов начнется как раз на рынке относительно дешевых, так называемых средних решений с функциональностью, аналогичной западным системам, и стоимостью не дороже 1000 долл. за рабочее место (с учетом стоимости внедрения).

При этом абсолютно непонятно, кто из игроков рынка может появиться в этом сегменте. Возможно, это будут пока хранящиеся в тайне разработки российских компаний, но не исключено, что на рынок придут локализованные западные решения. Этим «пришельцам» будет легко и сложно одновременно: ведь, с одной стороны, они попадут на уже разогретый CRM-рынок, а с другой – им придется быстро учиться тем правилам игры за клиентов, которые описаны в этой статье и уже неоднократно обкатаны работающими на рынке компаниями.