МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра общей информатики

Тян Алексей Юрьевич

Методы автоматизированного порождения поисковых эвристик по предметной области «информационная безопасность»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ
по направлению высшего профессионального образования

230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема диссертации утверждена распоряжением по НГУ №64 от «12» марта 2012г.
Тема диссертации скорректирована распоряжением по НГУ №538 от «14»декабря 2012г.

Научный руководитель:

к.ф.-м.н., доцент
Яхъяева Г.Э.

Новосибирск 2013

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

———————————————————————————————————————

Кафедра общей информатики

УТВЕРЖДАЮ

Зав. кафедрой Пальчунов Д. Е.

.....
(подпись, дата)

ЗАДАНИЕ
на магистерскую диссертацию

студент Тян Алексей Юрьевич

факультета ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Направление подготовки 230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Магистерская программа: Технология разработки программных систем

Тема: Методы автоматизированного порождения поисковых эвристик по предметной области «информационная безопасность»

Цели работы: Создание рубрики по предметной области «информационная безопасность» в виртуальном каталоге, автоматическое порождение эвристик для пар рубрика-ресурс

Руководитель

Яхъяева Г.Э.

к.ф.-м.н., доцент

.....

(подпись, дата)

Содержание

Введение **4**

1 Постановка задачи **5**

2 Обзор существующих решений **6**

2.1 Обзор существующих решений для поиска информации по предметной области информационная безопасность 6

2.2 Определения 7

2.3 Подробнее о поисковых системах 7

2.4 Подробнее об Интернет-каталогах 9

3 Виртуальный каталог **12**

3.1 Организация виртуального каталога 12

4 Рубрикатор **14**

4.1 Требования к рубрикатору 14

4.2 Существующие решения 15

4.3 Создание рубрикатора 15

5 Эвристики **16**

5.1 Система для порождения эвристик 16

5.2 Обучение 16

5.3 Автоматический подбор эвристик 16

5.3.1 Пример порождения конъюнкций 17

5.4 Оценка результатов 18

5.4.1 Описание работы системы 18

Заключение **22**

Литература **24**

Введение

Всё большое число людей использует Интернет для решения различных задач. С его помощью можно поддерживать связь с людьми, производить покупки, следить за чем либо, например за возведением зданий или ходом выборов. Последние лет 10 становятся всё популярнее и популярнее различного вида социальные сети. И люди заняты теперь тем, что называют генерацией контента. Но пока что, основной задачей Интернета остаётся предоставление человеку доступа к информации. Как уже было замечено информацией Всемирную сеть наполняют в промышленных масштабах, а по тому часто найти среди всего что есть то, что нужно не так то просто. В итоге появились инструменты позволяющие осуществлять поиск информации в Интернете.

В работе речь пойдёт о методах поиска информации в Интернете.

Сегодня для поиска информации как правило используют поисковые системы, реже Интернет-каталоги. И у тех и у других есть преимущества и недостатки, которые будут рассмотрены подробнее ниже. Что бы избавиться от их недостатков и объединить их преимущества был разработан виртуальный каталог, который является симбиозом первых двух. В статье рассмотрены средства облегчающие работу с виртуальным каталогом.

При поиске так же часто бывает нужно найти информацию в каком то конкретном виде, например в виде статьи или форума. В работе будет рассмотрена и эта проблема.

1 Постановка задачи

В виртуальном каталоге [7] каждой рубрике соответствует поисковая эвристика. Под поисковой эвристикой в работе подразумевается набор ключей. Они используются как запрос или часть запроса, который виртуальный каталог делает в поисковой системе.

Ранее было создано средство позволяющее автоматизировать составление поисковых эвристик [8], но оно не было рассчитано на составление эвристик для пар рубрика-тип ресурса. Так как информацию часто необходимо найти в определённом виде, то эвристики должны быть рассчитаны не только на поиск по выбранной рубрике, но и на поиск определённых видов ресурсов по выбранной рубрике. Возникает задача создания эвристик для каждой пары рубрика-тип ресурса.

Основной целью было автоматизация создания эвристик для пар рубрика-тип ресурса. Кроме того для апробации результата было предложено составить новую рубрику в виртуальном каталоге по предметной области «информационная безопасность». Таким образом получаются следующие задачи:

1. Анализ инструментов для поиска информации в интернете на основании формальных критериев оценки качества поиска.
2. Изучение основных принципов работы виртуального каталога.
3. Исследование методов автоматического порождения эвристик для виртуального каталога.
4. Создание рубрикатора для предметной области «информационная безопасность».
5. Реализация возможности порождения эвристик для пар рубрика-тип ресурса.
6. Порождение эвристик.

2 Обзор существующих решений

2.1 Обзор существующих решений для поиска информации по предметной области информационная безопасность

Есть большое количество сайтов, контент которых связан с тематикой информационной безопасности. Среди них можно выделить.

- Порталы. Например securitylab.ru. В нём есть несколько разделов: новости, уязвимости, блоги, статьи и софт. Таким образом на сайте публикуются новости так или иначе связанные с информационной безопасностью, есть статьи, переводы с данной тематикой, и большая база уязвимостей.

  В таких порталов много, в том числе и русскоязычных (http://resources.infosecinstitute.com/, http://mindfulsecurity.com/, http://searchsecurity.techtarget.com/). В основном на них публикуются новости, статьи, даются ссылки на блоги. Поиск если есть, то по самому сайту. В итоге найти можно только то, что было обработано людьми работающими над порталом. Как правило это очень не большая часть из всего что есть в сети.

- Журналы с тематикой информационная безопасность. Например http://searchsecurity.techtarget.com/resources и http://www.itsec.ru/insec-about.php. На таких сайтах лежат выпуски самих журналов, отдельные статьи, а так же может быть всё что есть на порталах.

И те и другие содержат много данных по информационной безопасности, но они плохо подходят для того, чтобы искать с их помощью информацию. Подобные ресурсы предназначены для того, чтобы пользователи получали недавно размещённый на сайте контент, например новости.

Обычно есть поиск по сайту, но редко можно этот поиск как то настраивать: отсортировать или например выбросить из результатов поиска новостные заметки.

Для поиска информации лучше всё таки использовать не тематические сайты, а те что созданы специально для поиска. Среди них выделают Интернет-каталоги и поисковые

системы. Для их описания и сравнения понадобится следующий список понятий.

2.2 Определения

- Релевантность - соответствие полученной информации информационному запросу [3].
- Пертинентность - соответствие полученной информации информационной потребности [3].

То есть релевантность оценивает то, на сколько выдача соответствует запросу пользователя, а пертинентность - на сколько результаты поиска удовлетворяют ожиданиям пользователя. Так же важно что для пользователя гораздо важнее, чтобы поисковый инструмент обладал хорошей пертинентностью. Помимо этого используются:

- Полнота поиска - доля релевантных документов в выборке по отношению ко всем релевантным документам коллекции.
- Точность - доля релевантных документов выборки по отношению ко всем документам в выборке.
- Поисковый шум - совокупность выданных при информационном поиске не релевантных документов [3].
- Коэффициент шума - отношение числа не релевантных документов в выдаче к общему числу документов в выдаче [3].

2.3 Подробнее о поисковых системах

Поисковая система — программно-аппаратный комплекс с веб-интерфейсом, предоставляющий возможность поиска информации в Интернете.
Работает примерно так. Сначала поисковые роботы («веб-паук», краулер) «сёрфят», то есть заходя на известные страницы и, по ссылкам там расположенным, начинают переходить на другие. При этом они скачивают содержимое страниц. Поисковые системы могут использовать несколько таких роботов, каждый из которых предназначен

для определённой задачи. Например в Яндексе есть робот который индексирует только картинки или rrs-ленты. Естественно есть основной, который ищет информацию для базы основного поиска. Есть быстрый, предназначенный для индексирования свежей информации, например новостей.

Полученную информацию поисковый робот отдаёт на индексацию. Программа обрабатывает информацию и сохраняет результат в базе данных. Цель обработки получить информацию о документе, по которой можно быстро понять что есть в документе, например какие слова или ключи в нём часто встречаются. Это нужно, чтобы при поиске не просматривать каждый документ полностью, а только считать индекс и по нему определить релевантен ли документ запросу [4].

Далее пользователь используя веб-интерфейс вводит запрос, и поисковая система по индексам подбирает страницы релевантные запросу, после чего выводит список, по её мнению релевантных ресурсов.

Для поисковых систем характерно:

1. Высокая релевантность поиска. Почти все современные поисковые системы выдают релевантный запросу список страниц. У большинства популярных поисковых систем принципы определения релевантности схожи, но конкретная реализация у каждого своя. Например все учитывают наличие слов из запроса в заголовке и тексте, но степень этого влияния не разглашается. Делается это помимо всего прочего для того, чтобы создателям сайтов было сложнее влиять на результаты поиска.

2. Полнота и актуальность. Под актуальностью подразумевается то, на сколько найденная информация новая, соответствует ли последним данным. Например информация о последнем номере какого-либо издания. Если её не обновить после выхода очередного номера, то информация устареет, станет не актуальной.

   Актуальность достигается за счёт того, что поисковые роботы не прекращают свою работу. Постоянно пополняя индекс новыми страницами. При этом для часто обновляемых ресурсов (новости, форумы) есть специальные быстрые поисковые

роботы, которые сканируют эти ресурсы гораздо чаще, чем основной робот.

В результате поиск осуществляется по большому числу страниц, кроме того новые страницы или недавно обновившиеся так же попадают в поле зрения поиска.

3. Низкий уровень поискового шума. Понятно, что создатели поисковых систем стремятся уменьшить поисковый шум. Одним из его видов является поисковый спам - сайты и страницы в Интернете, созданные с целью манипуляции результатами поиска в поисковых машинах. Делается это с помощью указывания в тегах популярных поисковых ключей, насыщением текста страницы такими ключами, созданием множества страниц ссылающихся друг на друга и т д. Нужно это для того, чтобы сайт заметили как можно больше пользователей и перешли на него, после чего им можно например показать рекламу.

   Сегодня поисковые системы достаточно успешно преодолевают эту проблему. По крайней мере наиболее популярные, такие как google и yandex, практически побороли поисковый спам. Борьба с поисковым спамом - одна из причин, по которым не разглашается то, как именно работают алгоритмы поисковых систем.

2.4 Подробнее об Интернет-каталогах

Интернет-каталоги — структурированный набор ссылок на сайты с кратким их описанием. Сайты внутри каталога разбиваются по темам. Каждая тема может иметь подтемы, которые ещё точнее специфицируют область поиска. Всё это организовано в виде дерева. Такое дерево называют классификатором или рубрикатором. Кроме того темы могут снабжаться кратким описанием. Ресурсы вносятся в каталог и разбиваются на разные темы модераторами.

Считается что среди инструментов для поиска информации в сети Интернет-каталоги были одними из первых. При поиске пользователь ходит по Интернет-каталогу и выбирает нужную ему тему. Выбрав рубрику пользователь получит список ссылок на Интернет-ресурсы.

Интернет-каталоги могут быть общими или специализированными.

В каталогах общего назначения содержатся ресурсы из разных областей знаний и сфер деятельности человека. К общим можно отнести http://www.dmoz.org, http://dir.yahoo.com - глобальные, http://yaca.yandex.ru/(яндекс каталог) http://fiftys.ru/ - русскоязычные.

В специализированных Интернет-каталогах представлены ссылки на ресурсы определённой тематики, области знаний. Примером специализированного Интернет-каталога может служить http://www.ru-ib.ru/.

Интерфейс Интернет-каталоги может быть удобен. Но у него есть ряд существенных недостатков.

1. Малое количество ссылок.

   Выделяют два способа наполнения каталога.

   Первый - когда работники Интернет-каталога сами добавляют ресурсы, либо ищут их, либо создатели веб сайтов отправляют им заявку на размещение. К таким можно отнести уже упомянутые http://www.dmoz.org и http://yaca.yandex.ru/. В этом случае, при условии добросовестного отбора ресурсов в каталоге будут сайты подходящие указанной в рубрике теме.

   Второй - когда пользователи, владельцы сайтов сами размещают ссылки в Интернет-каталоге на свои ресурсы. В этом случае количество сайтов в каталоге будет больше чем в предыдущем. Но при добавлении пользователи могут случайно или намерено добавить ресурс не в соответствующую рубрику. Что снизит релевантность и точность поиска.

   В обоих случаях Интернет-каталоги наполняют люди, и как бы они не работали они не смогут сравниться с поисковыми роботами поисковых систем по количеству обработанных сайтов.

2. Отсутствие актуальной информации. Опять таки из за того, что в Интернет-каталоги ресурсы добавляют люди, а в поисковые системы - роботы. Первые сильно проигрывают вторым в скорости. Кроме того если информация на сайте изменилась, поисковые роботы это заметят гораздо раньше людей.

Релевантность - как уже отмечалось есть разные «политики» по добавлению ресурсов в Интернет-каталог. И если строгий контроль над каждым добавляемым ресурсом, может дать высокую релевантность. То отпуск всё на совесть пользователей добавляющих ресурсы может привести к тому, что релевантность будет крайне низкой.

3 Виртуальный каталог

Виртуальный каталог - инструмент для поиска информации в Интернете. В основе лежит идея совместить преимущества Интернет-каталогов и поисковых систем. То есть ясность интерфейса каталога и полноту, актуальность и релевантность с которыми поисковые системы находят информацию [1].

Интерфейс виртуального каталога как и у обычного Интернет-каталога древовидный. Те же рубрики и подрубрики. Пользователь так же ходит по каталогу, выбирает нужную тему и получает список ссылок на ресурсы. Отличие состоит в том, что тут не хранятся ссылки на внешние ресурсы, как это делается в Интернет-каталогах. Вместо этого по названию рубрики, которую выбрал пользователь, формируется запрос и производится поиск в поисковой системе (google или yandex). Таким образом интернет каталог обладает всеми преимуществами поисковых систем, то есть релевантностью, полнотой и актуальностью. При этом имеет понятный интерфейс Интернет-каталога.

Интерфейс виртуально каталога помимо дерева рубрик содержит поле для запросов. С его помощью пользователь может уточнять запрос если рубрика не полностью соответствует тому, что он ищет. Другой особенностью виртуального каталога является то, что одна и та же подрубрика может быть в нескольких рубриках. Например теория чисел подрубрика рубрики «алгебра и логика» и рубрики «алгебра». Так же для того что бы получить результат (список ссылок) не обязательно выбирать один из разделов нижнего уровня (те у которых нет под разделов). При выборе рубрики имеющей подразделы запрос к поисковой системе будет так же составлен как и при выборе нижнего уровня.

3.1 Организация виртуального каталога

Как уже было сказано виртуальный каталог не хранит ссылки на ресурсы, а по названию рубрики выбранной пользователем формирует запрос к поисковой системе. Для этого каждой рубрике сопоставлено специальное слово, несколько слов или фраза,

называемых эвристиками. Под эвристикой подразумевается некоторое знание из предметной области, представленное в виде набора ключей. Причём эвристика находясь в запросе должна уточнять его таким образом, что бы найденные поисковой системой ресурсы были из выбранной предметной области. То есть эвристика должна точно определять ту область которая интересует пользователя.

4 Рубрикатор

Создание рубрики для какой-либо предметной области в виртуальном каталоге разбивается на два основных этапа: создание рубрикатора и порождение поисковых эвристик.

4.1 Требования к рубрикатору

Прежде чем составлять каталог, нужно понять что в нём должно быть представлено. Термин информационная безопасность могут использовать в разных смыслах, тут под ним будет пониматься механизм защиты, обеспечивающий:

- конфиденциальность
- целостность
- доступность

Выделяют следующие составляющие информационной безопасности:

1. Законодательная, нормативно-правовая и научная база.
2. Структура и задачи органов (подразделений), обеспечивающих безопасность ИТ.
3. Организационно-технические и режимные меры и методы (политика информационной безопасности).
4. Программно-технические способы и средства обеспечения информационной безопасности.

Поиск должен вестись по всем этим составляющим, то есть рубрикатор должен их охватывать.

4.2 Существующие решения

Большие разделы по предметной области «информационная безопасность» имеют Интернет-каталоги http://www.dmoz.org (англоязычная версия) и http://dir.yahoo.com по 2580 и 567 ресурсов соответственно. Яндекс каталоги имеет 172 ресурса, а http://www.dmoz.org русскоязычный - 165.

В http://dir.yahoo.com в нужном разделе 31 рубрика. Они в основном относятся к ресурсам которые содержат информацию о программах и алгоритмах. Кроме того несколько разделов можно отнести к политике по информационной безопасности. Таким образом этот каталог охватывает разделы 4 и 3.

В www.dmoz.org так же имеет 31 рубрику, так же нет подрубрик. Разделы каталога описывают программы, уязвимости, организации, политики приватности. То есть так же охватывает 3 и 4 пункты.

4.3 Создание рубрикатора

Было решено, что существующие рубрикаторы не удовлетворяют условиям, по этому был сделан свой рубрикатор.

Основные подразделы рубрикатора: законодательство, криптография, управление доступом, программное обеспечение, стандарты, уязвимости, атаки, прочее.

В нём так же основное внимание уделено программам, алгоритмам и т. д. но кроме того учитываются такие вещи как законодательство. Деятельность и персоналии - подрубрики рубрики прочее описывают ответственность различных органов. Таким образом учтены все составляющие.

5 Эвристики

Для сравнения эвристики породили двумя способами, в ручную и при помощи экспертной системы.

5.1 Система для порождения эвристик

За основу экспертной системы была взята уже имеющаяся [8], сделанная раньше, в неё была добавленная возможность порождать эвристики для пар рубрика-ресурс.
Основной принцип работы состоит в следующем: происходит обучение системы, после чего она порождает эвристики на основе логических методов. Реализована система была на php с использованием MySQL.
Работу с системой можно разделить на 3 этапа:

1. Обучение.
2. Автоматическое построение эвристик.
3. Оценка результатов.

5.2 Обучение

Обучение состоит в том, что пользователь для выбранной пары ресурс-рубрика делает, кажущийся ему адекватным, запрос. Получает набор текстов. Из них выбирает релевантные и не релевантные. После чего сохраняет результат. Процедуру можно повторять. А сохранённые тексты удалять. Когда пользователь решит что набор текстов достаточно большой он переходит к следующему шагу.

5.3 Автоматический подбор эвристик

1. Анализ выбранных текстов. Релевантные тексты помещаются в множество $relevanceText[i]$, не релевантные в $irrelevanceText[i]$. В ходе анализа текстов получается набор лексем со статистикой встречаемости.

$relevanceTextLexem[i]$ - множество лексем из i-го релевантного текста.

$irrelevanceTextLexem[i]$ - множество лексем из i-го не релевантного текста.

2. Формирование множеств из лексем. Формируются 2 множества: $relevanceLexemSet$ и $irrelevanceLexemSet$. Множества уникальных лексем из релевантных и не релевантных текстов соответственно.

3. Формирование эвристик.

   Эвристики представляют собой конъюнкции из лексем.

   Считаем что конъюнкция истинна на множестве лексем, если все лексемы конъюнкции принадлежат множеству.

   Конъюнкция ложна на множестве лексем, если хотя бы одна из лексем конъюнкции не принадлежит множеству.

   Алгоритм формирования эвристик следующий.

   На первом шаге $relevConSet$ - искомое множество конъюнкций пусто.

   Далее начинают генерироваться все возможные конъюнкции размером начиная с одного до максимального, указанного пользователем. Конъюнкции формируются из лексем множества $relevanceLexemSet$, лексем из релевантных текстов.

   Каждая лексема проверяется на истинность на множестве $irrelevanceLexemSet$, лексем из не релевантных текстов. Если она оказывается ложной на этом множестве, то её добавляют в $relevConSet$.

5.3.1 Пример порождения конъюнкций

Пусть:

Множество $relevanceLexemSet = \{A,\ B,\ C,\ D\}$

Множество $irrelevanceLexemSet = \{B,\ D\}$

Максимальное количество лексем в конъюнкции равно двум.

Тогда:

Искомое множество $relevConSet = \{A,\ C,\ AB,\ AC,\ AD,\ BC,\ CD\}$, то есть мы отбросили конъюнкции $B,\ D,\ BD$, которые истинны на множестве $irrelevanceLexemSet$.

5.4 Оценка результатов

Последний этап. Пользователю выводится результат поиска. Для того, чтобы проверить результат, оценить его с точки зрения пертинетности, и решить остановиться на достигнутом или продолжить подбирать тексты.

Вывод результата устроен следующим образом. Система посылает поисковой системе группу запросов, представляющих собой полученные эвристики. Результат ранжируется и показывается пользователю.

Тексты ранжируются по убывании частоты с которой встречается эвристика, которой принадлежит этот текст.

5.4.1 Описание работы системы

Предположим, что нам необходимо подобрать эвристики для рубрики «вирус» предметной области «информационная безопасность».

1. Добавление текстов.

   Допустим был введён запрос virus. Тогда будет получен результат содержащий большое количество не пертинетных ссылок (Рис. 5.1).

2. Построение эвристик.

   На следующем шаге система сгенерирует набор эвристик из лексем

   - THE VIRUSES COMPUTER
   - TO VIRUS COMPUTER
   - A VIRUS COMPUTER
   - THE VIRUS
   - THE OF COMPUTER
   - THE VIRUS COMPUTER

- THE VIRUS COMPUTER
- VIRUS COMPUTER
- THE VIRUSES
- THE COMPUTER
- VIRUS
- THE TO VIRUS
- THE A VIRUS
- THE OF VIRUS
- THE A AND
- THE VIRUS AND
- VIRUSES

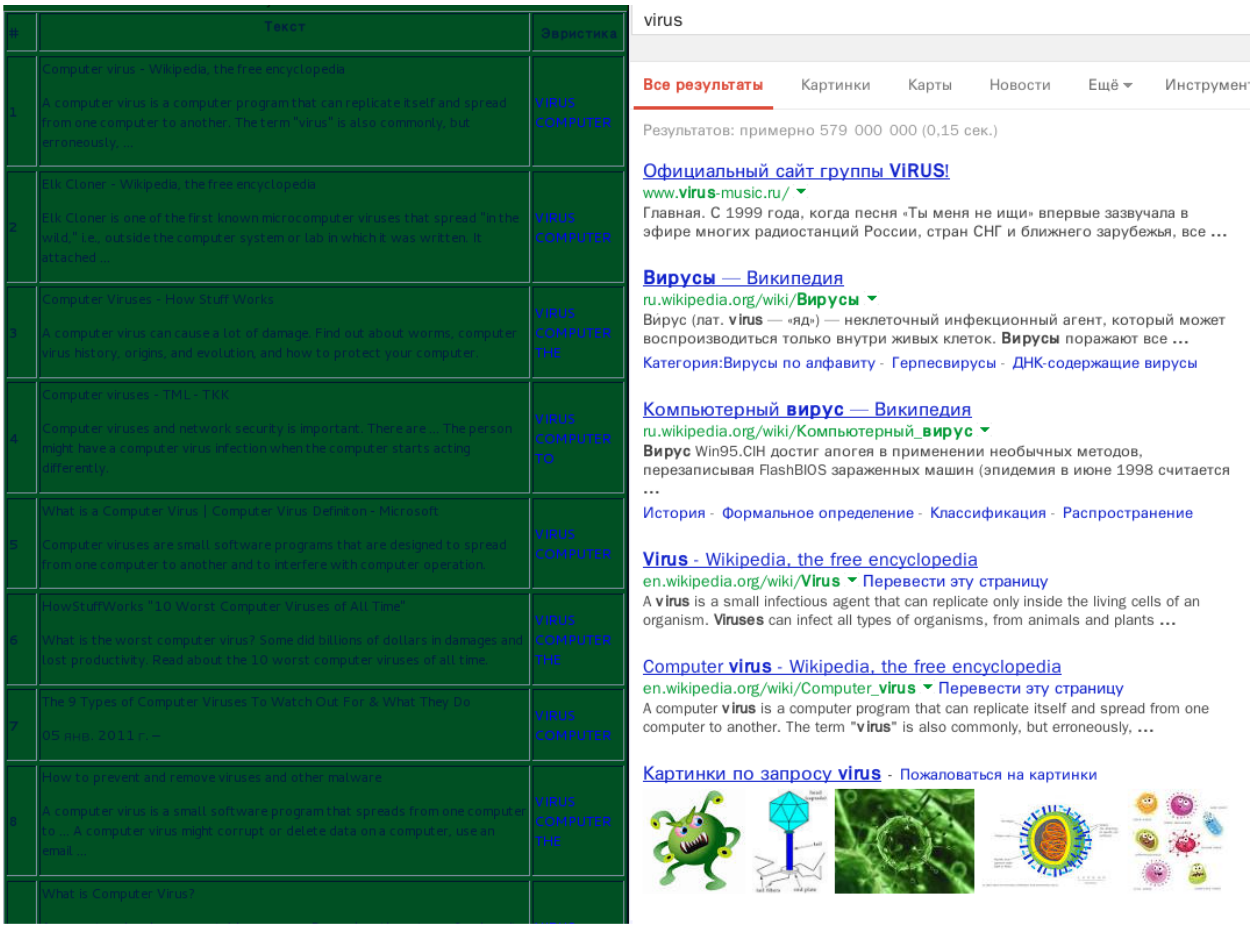
Эксперт может выбросить лишние, в результате останутся

- THE VIRUSES COMPUTER
- TO VIRUS COMPUTER
- A VIRUS COMPUTER

3. Сравнение результатов (Рис. 5.2).

   Тут можно заметить что из пяти результатов выданных гуглом, нас могут заинтересовать два. В то вермя как все ссылки выданные системой ведут к информации о компьютерных вирусах. Так же нужно учесть что результат выданный системой - комбинация результатов нескольких запросов.

Рис. 5.1: Этап обучения.



#	Текст	Эвристика
1	Computer virus - Wikipedia, the free encyclopedia A computer virus is a computer program that can replicate itself and spread from one computer to another. The term "virus" is also commonly, but erroneously, ...	VIRUS COMPUTER
2	Elk Cloner - Wikipedia, the free encyclopedia Elk Cloner is one of the first known microcomputer viruses that spread "in the wild," i.e., outside the computer system or lab in which it was written. It attached ...	VIRUS COMPUTER
3	Computer Viruses - How Stuff Works A computer virus can cause a lot of damage. Find out about worms, computer virus history, origins, and evolution, and how to protect your computer.	VIRUS COMPUTER THE
4	Computer viruses - TML - TKK Computer viruses and network security is important. There are ... The person might have a computer virus infection when the computer starts acting differently.	VIRUS COMPUTER TO
5	What is a Computer Virus	Computer Virus Definition - Microsoft Computer viruses are small software programs that are designed to spread from one computer to another and to interfere with computer operation.	VIRUS COMPUTER
6	HowStuffWorks "10 Worst Computer Viruses of All Time" What is the worst computer virus? Some did billions of dollars in damages and lost productivity. Read about the 10 worst computer viruses of all time.	VIRUS COMPUTER THE
7	The 9 Types of Computer Viruses To Watch Out For & What They Do 05 янв. 2011 г. –	VIRUS COMPUTER
8	How to prevent and remove viruses and other malware A computer virus is a small software program that spreads from one computer to ... A computer virus might corrupt or delete data on a computer, use an email ...	VIRUS COMPUTER THE
	What is Computer Virus?	

Рис. 5.2: Сравнение результатов.

Заключение

Ниже приведены сравнительные таблицы результатов поиска с использованием разных поисковых систем с точки зрения пертинетности. Пертинентность считалась как отношение найденных ресурсов соответствующей тематики ко всем найденным ресурсам (рассматривались только первые десять результатов).

Таблица 1 содержит результаты поиска любых ресурсов.

Таблица 2 - результаты поиска статей.

Под эвристики1 подразумевается виртуальный каталог с эвристиками подобранными вручную, эвристики2 - сгенерированные автоматически.

Таблица 1: Для любого ресурса.

	google	**yandex**	**эвристики1**	**эвристики2**
инъекции	10%	10%	100%	100%
вирус	20%	10%	100%	90%
аутентификация	100%	80%	100%	100%
законодательство	0	0	80%	50%
троянский конь	40%	20%	90%	60%

Таблица 2: Тип ресурса - статья.

	google	**yandex**	**эвристики1**	**эвристики2**
инъекции	50%	60%	80%	60%
вирус	20%	40%	60%	30%
аутентификация	70%	60%	90%	90%
законодательство	0	0	40%	20%
троянский конь	30%	40%	50%	40%

Как видно из таблиц, результат полученный при помощи автоматически сгенерированных эвристик хоть и хуже чем результат полученный с помощью в ручную подобранных, но не значительно. И при этом всё равно лучше чем результат поисковых систем.

В результате работы над магистерской диссертацией:

1. Проведён анализ инструментов для поиска информации в интернете на основании формальных критериев оценки качества поиска.
2. Изучены основные принципы работы виртуального каталога.
3. Исследованы методы автоматического порождения эвристик для виртуального каталога.
4. Создан рубрикатор для предметной области «информационная безопасность».
5. Реализована возможность порождать эвристики для пар рубрика-тип ресурса.
6. Автоматически сгенерированы эвристки для предметной области «информационная безопасность».
7. Произведено сравнение результатов поиска.

Помимо этого были сделаны 2 публикации:

1. Тян А.Ю. Виртуальный каталог по предметной области информационная безопасность // Материалы 51 Международной Научной Студенческой Конференции «Студент и научно-технический прогресс»: Информационные технологии / Новосиб. гос. ун-т, Новосибирск, 2013, С. 125.
2. Тян А.Ю. Виртуальный каталог по предметной области «информационная безопасность» // Современные инновации в науке и технике. Материалы III Международной научно-практической конференции / Юго-Западный гос. ун-т, Курск, 2013, С. 185-192.

Литература

1. Пальчунов Д.Е., Сидорова Е.С. Виртуальный каталог. Труды Всероссийской конференции «Знания–Онтологии–Теории», Новосибирск, 2007, С. 166–175.

2. Dmitry E. Palchunov. Virtual catalog: the ontology-based technology for information retrieval. In: Lecture Notes in Artificial Intelligence 6581, Springer-Verlag Berlin Heidelberg, 2011, pp. 164–183.

3. ГОСТ 7.73-96. Система стандартов по информации, библиотечному и издательскому делу. Поиск и распространение информации. Термины и определения.— Взамен ГОСТ 7.27-80; Введ. 01.01.98 .— М.: Изд-во стандартов, 1997.— 13 с.

4. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze An Introduction to Information Retrieval. Cambridge University Press, 2009.

5. Пальчунов Д. Е. Решение задачи поиска информации на основе онтологии. Бизнес-Информатика .— М.: НИУ ВШЭ, 2008 .— С. 3-13.

6. ГОСТ Р 50922-2006. Защита информации. Основные термины и определения.— Взамен ГОСТ Р 50922-96; Введ. 02.01.2008 .— М.: Изд-во стандартов, 2007 .— 12 с.

7. MetaSearch [Электронный ресурс] – 2008.— Режим доступа: http://virtualcatalog.ru (дата обращения: 31.05.2013).

8. Пальчунов Д.Е., Ульянова Е.А. Методы автоматического порождения поисковых эвристик // Вестник НГУ, Серия: Информационные технологии .— т. 8 .— вып. 3 .— 2010 .— С. 5-12.

9. Бездольный А.М. Экспериментальная машина подбора эвристик для Виртуального каталога // Материалы XLVI Международной Научной Студенческой Конференции «Студент и научно-технический прогресс»: Информационные технологии/Новосиб. гос. ун-т.— Новосибирск, 2008.— С. 193.

10. Гусев В.С. Google – эффективный поиск информации в Интернет.— Киев: издательство Диалектика, 2006.— 240 с.

11. Ландэ Д.В. Поиск знаний в Internet.— Киев: Издательский дом «Диалектика Вильямс», 2005.— 272 с.

12. Холмогоров В. Поиск в Интернете и сервисы Яндекс.— Санкт-Петербург: Питер, 2006.— 128 с.