УДК 004.82

**Д. Е. Пальчунов** [1, 2], **Е. А. Ульянова** [2]

[1] Институт математики СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

[2] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

Email: [1] palch@math.nsc.ru; [2] airland4@gmail.com

# МЕТОДЫ АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ ПОИСКОВЫХ ЭВРИСТИК [*]

Статья посвящена разработке автоматических методов порождения поисковых эвристик для обеспечения релевантности работы виртуальных каталогов. Виртуальный каталог – это метапоисковая система, представляющая собой синтез Интернет-каталога и информационно-поисковой системы. Методы подбора поисковых эвристик для виртуальных каталогов основаны на использовании формул пропозициональной логики специального вида – дизъюнктивных нормальных форм. Автоматизация порождения эвристик осуществляется за счет извлечения информации из текстов естественного языка.

*Ключевые слова*: информационно-поисковые системы, виртуальный каталог, релевантность, пертинентность, онтология, эвристика.

## Введение

В настоящее время Интернет представляет собой один из самых больших источников разнообразной информации. В нем содержатся миллиарды документов, и их количество возрастает с каждым годом. При всем многообразии существующих в Интернете информационно-поисковых систем на поиск необходимых сведений тратится много сил и времени, особенно когда дело касается специфической информации [1–3].

Статья посвящена разработке автоматических методов порождения поисковых эвристик для обеспечения релевантности работы виртуальных каталогов. Виртуальный каталог [4; 5] – это метапоисковая система, представляющая собой синтез Интернет-каталога и информационно-поисковой системы. Методы подбора поисковых эвристик для виртуальных каталогов основаны на использовании формул пропозициональной логики специального вида – дизъюнктивных нормальных форм. Автоматизация порождения эвристик осуществляется за счет извлечения информации из текстов естественного языка.

## Оценки эффективности работы поисковых систем

Для того чтобы оценивать качество работы информационно-поисковых систем, рассмотрим следующие формальные критерии.

[*] Работа выполнена при поддрежке Федерального агентства по образованию, грант ГК-П-1008, а также гранта Междисциплинарного интеграционного проекта фундаментальных исследований СО РАН № 119.


© Д. Е. Пальчунов, Е. А. Ульянова, 2010

1. Пертинентность – степень соответствия результатов поиска информационной потребности пользователя. Пертинентность определяется субъективным восприятием человека.

Формальной пертинентностью информационной системы мы будем называть двухместную функцию, определенную на паре: первый аргумент – информационная потребность пользователя, которая была формализована в запросе, второй аргумент – упорядоченный список Интернет-ресурсов, который поисковая система выдала в ответ на этот запрос. Для простоты можно считать, что эта функция принимает значения в промежутке [0, 1]. Чем более результат выдачи соответствует информационной потребности пользователя, тем больше значение функции. Полное соответствие – 1, отсутствие соответствия – 0.

Пертинентность информационного поиска зависит от двух обстоятельств: насколько точно формальный запрос, составленный пользователем, соответствует его информационной потребности, и насколько хорошо результаты выдачи информационной системы соответствуют формальному запросу.

Таким образом, пертинентность разлагается в композицию (при формальном определении – в произведение) двух составляющих, которые мы будем называть адекватностью и релевантностью.

2. Адекватность – мера того, насколько запрос к поисковой системе соответствует информационной потребности пользователя.

*Формальной адекватностью* запроса мы будем называть двухместную функцию, определенную на паре: первый аргумент – информационная потребность пользователя, которая была формализована в запросе, второй аргумент – сам поисковый запрос. Также будем считать, что эта функция принимает значения в промежутке [0, 1].

3. Релевантность – степень соответствия формальному запросу набора полученных в результате поиска документов.

*Формальной релевантностью* информационной системы мы будем называть двухместную функцию, определенную на паре: первый аргумент – формальный запрос, второй аргумент – упорядоченный список Интернет-ресурсов, который поисковая система выдала в ответ на этот запрос. Функция принимает значения в промежутке [0, 1].

Для простоты будем считать, что формальная пертинентность равна формальной адекватности, умноженной на формальную релевантность.

В настоящий момент времени наиболее разработанной характеристикой информационного поиска является релевантность. Современные поисковые системы научились достигать достаточно высокого значения этого параметра. Числовое значение релевантности зависит от трех параметров – точности, полноты и ранжирования. *Ранжирование* – это правильность порядка, в котором представлен список результатов информационного поиска. *Точность* – это доля релевантных ресурсов среди всех ресурсов, присутствующих в выдаче, а *полнота* – доля релевантных ресурсов, присутствующих в выдаче, среди всех релевантных ресурсов, имеющихся в Интернете. Тем не менее с точки зрения пользователя главным критерием оценки результатов поиска информации является ее пертинентность.

Рассмотрим существующие поисковые системы, используя введенные критерии оценки эффективности их работы.

*Поисковые системы* с большим количеством проиндексированных документов (в идеале – совпадающим со всем Интернетом) – Google, Яндекс, Rambler и др. Такими системами достигается высокая релевантность поиска [6]. Однако, несмотря на то, что результаты поиска в таких системах будут формально соответствовать поисковому запросу, ими не достигается пертинентность: реальные поисковые потребности пользователей, набравших один и тот же поисковый запрос – последовательность из нескольких слов, могут быть совершенно различными. Поэтому выдача, соответствующая поисковой потребности одного пользователя, может совершенно не соответствовать потребности другого.

*Интернет-каталоги* – сайты, представляющие собой организованные по тематическому принципу коллекции ссылок на другие сайты. Поисковые каталоги, получившие наибольшую популярность, – это Yahoo, Open Directory, Яндекс-каталог, Апорт и др. За счет предварительной обработки документов модераторами, поисковые каталоги в большой степени мо-

гут обеспечить пертинентность выдачи. Но при этом Интернет-каталоги абсолютно не гарантируют пользователям полноту выдаваемой информации.

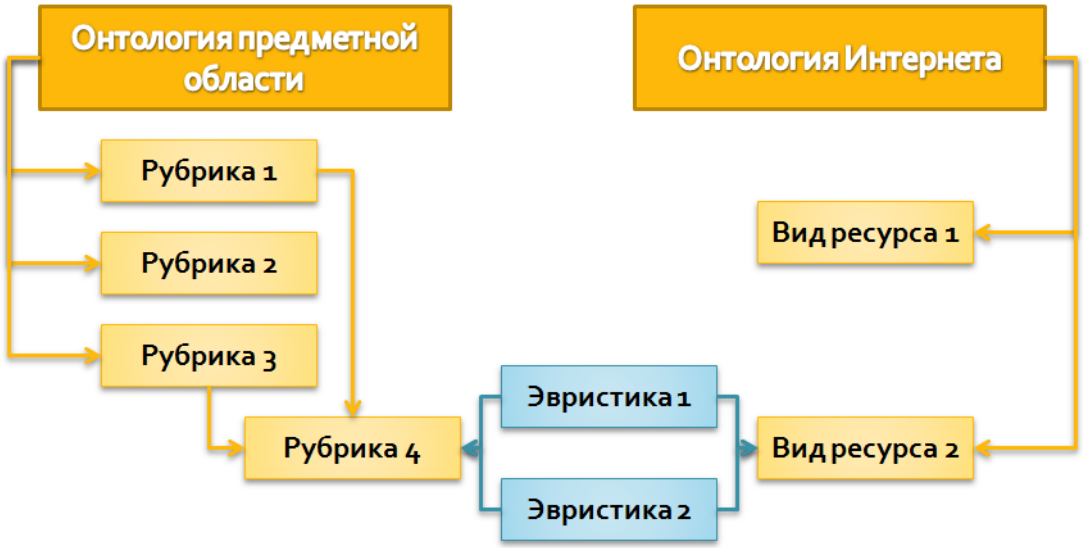
*Виртуальный каталог* – это метапоисковая система, цель которой – объединить в себе основные достоинства поисковых систем и Интернет-каталогов [4; 5]. Интерфейс виртуального каталога похож на обычный Интернет-каталог. Однако поиск информации в виртуальном каталоге происходит по всему Интернету, а не по своей базе данных. Принцип работы виртуального каталога основан на трех составляющих.

1. Онтология предметной области, содержащая, в частности, набор рубрик и ассоциации к ним [5; 7–9]. Рубрика соответствует определенному разделу предметной области. Рубрики связаны между собой отношением «общее-частное».

2. Онтология Интернета, которая задает классификацию видов Интернет-ресурсов [5]. Видами Интернет-ресурсов являются: статьи, конференции, персональные страницы, тематические сайты, форумы и т. д.

3. Эвристики – наборы ключевых терминов, сужающие область поиска до выбранной рубрики и вида ресурса [5].

Для релевантной отработки формального запроса, специфицированного пользователем, необходима функция порождения эвристик. Это двухместная функция, ее первый аргумент – выбранная рубрика каталога, второй – вид Интернет-ресурсов. Каждой паре аргументов (рубрика, вид ресурса) соответствует набор эвристик, из которых составляется расширенный запрос к поисковой системе, например к системам Google и Яндекс (см. рисунок). При правильно определенном наборе эвристик пользователь должен получить именно ту информацию, которая ему необходима.




Составляющие виртуального каталога

За счет представления формального запроса в виде композиции – подраздел предметной области и вид Интернет-ресурса – достигается адекватность формального запроса. За счет правильного подбора эвристик мы имеем релевантность отработки формального запроса. В результате имеется возможность достичь высокой пертинентности работы виртуального каталога.

- Полнота достигается за счет использования информационно-поисковой системы (например, Google, Яндекс).

- Адекватность достигается за счет выбора пользователем интересующей его рубрики – раздела предметной области и вида Интернет-ресурсов.
- Релевантность достигается за счет правильного и точного подбора эвристик, соответствующих паре – выбранная рубрика и выбранный вид Интернет-ресурсов.

Таким образом, цель виртуального каталога – обеспечить не только полноту и релевантность поиска, но и его наиболее важный параметр – пертинентность.

**Методы автоматического порождения эвристик**

Правильность и точность подбора эвристик определяет степень пертинентности результатов поиска в виртуальном каталоге. Но пертинентность – достаточно субъективный критерий. Поэтому процесс построения эвристик должен происходить с участием экспертов [10]. При этом необходимо учитывать, что набор эвристик должен быть подобран для каждой пары (рубрика, вид Интернет-ресурсов), а количество таких пар может быть достаточно велико. Виртуальный каталог может содержать несколько сотен рубрик, и если при этом имеется возможность выбора одного хотя бы из десяти видов Интернет-ресурсов, то количество пар будет составлять уже несколько тысяч.

Это делает практически невозможным ручной подбор эвристик. Необходимы методы их автоматического порождения, эксперту необходим инструмент, осуществляющий помощь в подборе эвристик. Таким инструментом является экспертная система автоматического подбора поисковых эвристик, принципы работы которой мы изложим далее. Первоначально происходит обучение экспертной системы, после чего система на основании логических методов автоматически строит искомый набор эвристик.

*Обучение экспертной системы* происходит независимо для каждой пары (рубрика, вид ресурсов). Введем следующие определения.

*Релевантный текст* – это текст, соответствующий выбранной паре (рубрика, вид ресурсов).

*Нерелевантный текст* – это текст, не соответствующий выбранной паре (рубрика, вид ресурсов).

Процесс обучения экспертной системы основан на наборах релевантных и нерелевантных текстов (относительно данной пары – рубрика, вид ресурсов). Первоначальные наборы таких текстов строятся следующим образом. Эксперт выбирает рубрику и вид ресурсов и вводит такой запрос к поисковой системе, который, на его взгляд, является наиболее адекватным. Система отображает результаты поиска по введенному запросу. Эксперт выбирает из результатов поиска наиболее релевантные и наиболее нерелевантные тексты; предполагается, что таких текстов будет от половины до двух третей общего числа текстов, выданных системой. После этого запускается автоматический подбор эвристик.

*Логические методы автоматического подбора эвристик.* Автоматический подбор эвристик основан на построении по обучающей выборке текстов – наборам релевантных и нерелевантных текстов – специальной формулы логики высказываний. Главной задачей выбора данной формулы является отделение множества релевантных текстов от множества нерелевантных текстов. Формулу, разделяющую эти наборы текстов, мы строим в виде дизъюнктивной нормальной формы (ДНФ). В качестве пропозициональных переменных, входящих в ДНФ, мы будем использовать высказывания следующего типа: «Данная последовательность символов встречается в данном тексте». Причем мы рассматриваем произвольные последовательности символов, включающие буквы, цифры, знаки препинания, пробелы и служебные символы.

*Определение*. Элементарной конъюнкцией называется конъюнкция пропозициональных переменных и их отрицаний. Дизъюнктивной нормальной формой (ДНФ) называется дизъюнкция элементарных конъюнкций.

*Теорема* [11]. Любая формула логики высказываний эквивалентна некоторой дизъюнктивной нормальной форме.

*Следствие*. Любое конечное множество формул логики высказываний эквивалентно некоторой ДНФ.

Использование ДНФ с нашей точки зрения является исключительно удобным по двум причинам. Во-первых, как следует из Теоремы и Следствия, с помощью ДНФ мы можем сформулировать все, что может быть выражено конечным множеством формул бесквантор-ной логики.

Во-вторых, сама дизъюнктивная нормальная форма в данном контексте является совершенно естественной. Каждая элементарная конъюнкция, входящая в ДНФ, говорит о том, что если определенные последовательности символов встречаются в тексте, а другие – не встречаются, то данный текст является релевантным поисковому запросу, т. е. каждая элементарная конъюнкция должна быть ложна на всем множестве нерелевантных текстов (и, естественно предполагать, истинна на некоторых релевантных текстах). Дизъюнкция элементарных конъюнкций, т. е. сама ДНФ, говорит, что каждый релевантный текст удовлетворяет условию хотя бы одной элементарной конъюнкции. Таким образом, ДНФ разбивает процесс определения релевантности текстов на несколько случаев: первый случай описывается первой элементарной конъюнкцией, второй – второй конъюнкцией, и т. д., последний случай – последней элементарной конъюнкцией.

При этом каждая элементарная конъюнкция отвечает за релевантность отобранных документов, а весь набор элементарных конъюнкций – их дизъюнкция – отвечает за полноту, т. е. за то, что все релевантные документы будут найдены. Как было сказано, смыслом пропозициональной переменной в ДНФ является утверждение о том, что последовательность символов встречается в данном тексте. Следует заметить, что, как правило, невозможно найти единственную элементарную конъюнкцию, которая бы разделяла два множества текстов. Именно поэтому нам необходима их дизъюнкция – ДНФ.

*Алгоритм построения ДНФ*

1. Построение множества релевантных лексем и множества нерелевантных лексем. На первом этапе алгоритма построения ДНФ система формирует два набора лексем:
   - лексемы, встречающиеся в релевантных текстах (*relevantLexemSet*);
   - лексемы, встречающиеся в нерелевантных текстах (*IrrelevanteLexemSet*).

В качестве лексем используются части речи в нормальной форме.

2. Построение конъюнкций. Для описания алгоритма построения конъюнкций введем следующие обозначения:
   - конъюнкция *Con* – это некоторый набор лексем *Lexem[i]* (*Con = Lexem[1] & Lexem[2] & … Lexem[m]*);
   - конъюнкция *Con* истинна на некотором множестве лексем *LexemSet*, если все лексемы конъюнкции {*Lexem[i]*} содержатся во множестве лексем *LexemSet* (т. е. *Con* истинна на множестве *LexemSet*, если для любого *i*, такого что $Lexem[i] \in Con$, выполняется $Lexem[i] \in LexemSet$);
   - конъюнкция *Con* ложна на множестве лексем *LexemSet*, если хотя бы одна лексема из множества лексем конъюнкции {*Lexem[i]*} не содержится во множестве *LexemSet*.

*Алгоритм построения элементарных конъюнкций*

- Пусть *relevantConjunction* – искомое множество конъюнкций. Первоначально оно пусто.
- Программа начинает составлять все возможные уникальные конъюнкции *Con[m]* размера от одного слова до максимального количества слов, заданного экспертом, из множества *relevantLexemSet*. При этом программа проверяет, истинность полученной конъюнкции на множестве *IrrelevantLexemSet*. Если *Con[m]* ложна на множестве *IrrelevantLexemSet*, то

*Con[m]* добавляется в *relevantConjunction*. В противном случае такая конъюнкция отбрасывается.

В конце работы алгоритма система получает искомую ДНФ, истинную на множестве релевантных текстов и ложную на множестве нерелевантных текстов. Следует отметить, что описанная процедура может быть повторена экспертом при необходимости любое количество раз. Таким образом, эксперт может за несколько шагов найти именно тот набор эвристик, который будет наиболее эффективно улучшать поиск для пары (рубрика, вид ресурсов).

Механизм информационного поиска организован следующим образом.

1. Система отправляет информационно-поисковой системе (Google, Яндекс) группу запросов. Каждый запрос – эвристика из набора.
2. Система получает результаты поиска от каждого запроса, ранжирует их и отображает эксперту.

*Алгоритм ранжирования*

1. Первичная сортировка текстов происходит по тому же принципу, как были отсортированы тексты информационно-поисковой системой (Google, Яндекс).
2. Вторичная сортировка происходит по убыванию частоты встречаемости эвристик, входящих в текст.

Следует отметить, что эксперт обладает возможностью просмотреть не только обобщенный список найденных документов, но и результаты поиска для каждой эвристики по отдельности. Таким образом, применение экспертной системы для автоматического порождения эвристик позволит эксперту за несколько шагов подобрать искомый набор эвристик, который обеспечит высокую пертинентность результатов поиска.

**Оценка результатов поиска**

Рассмотрим с точки зрения пертинентности сравнительную таблицу результатов поиска с использованием разных поисковых систем:

| Рубрика | Количество текстов в выборке | Google, % | Яндекс, % | Виртуальный каталог, % |
|---|---|---|---|---|
| Вирус | 10 | 70 | 50 | 100 |
| Червь | 10 | | | 100 |
| Цифровая подпись | 10 | 80 | 80 | 100 |
| Троянский конь | 10 | 20 | 20 | 100 |
| Вирус | 30 | 30 | 35 | 100 |
| Червь | 30 | 50 | 40 | 100 |
| Цифровая подпись | 30 | 75 | 70 | 100 |
| Троянский конь | 30 | 20 | 20 | 100 |
| Вирус | 50 | 20 | 25 | 100 |
| Червь | 50 | 35 | 30 | 85 |
| Цифровая подпись | 50 | 50 | 50 | 80 |
| Троянский конь | 50 | 10 | 5 | 100 |

Оценка представляет собой процентное соотношение количества текстов, соответствующих тематике, от общего количества текстов. В данном случае мы не фиксировали вид Интернет-ресурсов.

Как видно из таблицы, пертинентность результатов поиска виртуального каталога при использовании полученного набора эвристик значительно превышает пертинентность поиска самих систем Google и Яндекс. Виртуальный каталог посылает поисковой системе несколько запросов, соответствующих разным элементарным конъюнкциям. Количество текстов, удовлетворяющих информационную потребность пользователя, становится значительно больше за счет того, что результаты поиска от разных эвристик сливаются в единый набор.

## Заключение

В работе предложены методы автоматизации подбора поисковых эвристик, основанные на использования формул пропозициональной логики специального вида – дизъюнктивных нормальных форм. Разработана программная система, предоставляющая средства подбора эвристик и сравнения их эффективности, а также средства пополнения и изменения набора эвристик для каждой пары – рубрика предметной области и вид Интернет-ресурсов. В результате использования полученных поисковых эвристик существенно повышается релевантность отработки поисковых запросов виртуальным каталогом.

В дальнейшем предполагается использовать методы анализа данных [12] для улучшения алгоритмов поиска подходящих дизъюнктивных нормальных форм.

## Список литературы

1. *Гультяев А. К.* Самое главное о... Поиск в Интернете. 2-е изд. СПб.: Питер, 2006. 144 с.

2. *Ландэ Д. В.* Поиск знаний в Internet. Киев: Изд. дом «Диалектика-Вильямс», 2003. 287 с.

3. *Браун М.* Методы поиска информации в Интернете. М.: Новый издательский дом, 2005. 136 с.

4. *Пальчунов Д. Е., Сидорова Е. С.* Виртуальный каталог // Тр. Всерос. конф. «Знания – Онтологии – Теории». Новосибирск, 2007. С. 166–175.

5. *Пальчунов Д. Е.* Решение задачи поиска информации на основе онтологий // Бизнес-информатика. 2008. № 1. С. 3–13.

6. *Гусев В. С.* Google – эффективный поиск информации в Интернет. Киев: Изд-во «Диалектика», 2006. 240 с.

7. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии // Философия науки. 2006. Т. 4 (31). С. 6–14.

8. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Ч. 2. Онтологии и формализация понятий // Философия науки. 2008. № 2 (37). С. 62–99.

9. *Pal'chunov D. E.* GABEK for Ontology Generation // Learning and Development in Organizations / Eds. Ph. Herdina, A. Oberprantacher, J. Zelger. Berlin; Wien, 2007. Vol. 2. P. 90–109.

10. *Бездольный А. М.* Экспериментальная машина подбора эвристик для Виртуального каталога // Материалы XLVI Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / Новосиб. гос. ун-т. Новосибирск, 2008, 193 с.

11. *Ершов Ю. Л., Палютин Е. А.* Математическая логика. М.: Наука, 2005. 360 с.

12. *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis. 2008. Vol. 18. No. 1. P. 1–6.

*Материал поступил в редколлегию 31.05.2010*

**D. E. Pal'chunov, E. A. Uljanova**

**METHODS FOR AUTHOMATIC GENERATION OF SEARCH HEURISTICS**

The paper is devoted to development of automated methods of search heuristics generation for ensure the relevance of the virtual catalogue. The virtual catalogue is a metasearch system which is a synthesis of Internet catalogues and search engines. The methods of search heuristics creation for the virtual catalogue are based on use of propositional logic formulas of a special kind – Disjunctive Normal Form (DNF). Automation of heuristics generation is based on information retrieval from natural language texts.

*Keywords*: search engines, virtual catalog, relevance, pertinence, ontology, heuristics.