ПОДХОДЫ К ПОСТРОЕНИЮ И РЕАЛИЗАЦИЯ СПЕЦИАЛИЗИРОВАННОЙ МЕТАПОИСКОВОЙ МАШИНЫ PROTHES*

П. И. Браславский, А. С. Шишкин
Институт машиноведения УрО РАН, Екатеринбург, Россия
e-mail: pb@imach.uran.ru, whoarym@sigmakom.ru

In this paper we introduce ProThes, a pilot metasearch engine for focused Web information retrieval. The paper describes basic approaches to system implementation: thesaurus format, query language, graphical user interface, as well as ProThes' architecture, key components, used standards, and development platform.

Введение

Стремительный рост объемов Интернета является мощным стимулом совершенствования средств поиска информации в сети. Современные машины поиска (МП) Интернета демонстрируют высокую производительность, широкий охват ресурсов, высокие темпы развития. Бизнес-модель коммерческих МП ставит целью обслуживание максимально большого количества пользователей. Отсюда вытекает ориентация МП на среднего пользователя, что подразумевает однозначную интерпретацию запросов, сильные предположения о предпочтениях и поведении пользователей. Это не всегда идет на пользу "искателям" со специфическими запросами и требованиями. Поэтому в последнее время можно наблюдать тенденцию к персонализации универсальных средств поиска и развитию специализированных поисковых сервисов. Так, многие поисковики предоставляют возможность настройки внешнего вида страницы поиска (см., например, http://www.alltheweb.com/help/alchemist/index, http://my.yahoo.com, http://my.msn.com). Ссылки на специализированные поисковые машины (поиск медицинской и правовой информации, поиск по детским ресурсам, по товарным предложениям и т. д.) можно найти по адресам http://searchenginewatch.com/links/, http://directory.google.com/Top/Computers/Internet/Searching/Search_Engines/Specialized/. В литературе описаны методы фокусированного (тематического) индексирования ресурсов [1]. Еще один подход к специализации поиска демонстрирует система Eurekster (www.eurekster.com) — она ранжирует результаты МП AlltheWeb (www.alltheweb.com) с учетом предпочтений участников определенного сообщества. Кроме того, предлагаются методы персонализации поиска на стороне клиента [2], основанные на анализе активности (поведения) пользователя.

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 03-07-90342).

Предлагаемая нами специализированная метапоисковая машина (МПМ) позволяет устранить дисбаланс между универсальностью МП Интернета и специфичностью информационных потребностей группы пользователей. Специфичность МПМ придают тезаурус предметной области и предпочтения, влияющие на ранжирование результатов. Оба компонента независимы от ядра МПМ, что делает возможной настройку на различные предметные области.

Метапоиск — распространенная техника информационного поиска, которая основана на опросе многих информационно-поисковых систем (в противоположность созданию собственного индекса). Метапоисковая машина предоставляет единый интерфейс ко многим МП и может служить средством повышения полноты откликов (хотя полнота не является критичным параметром Интернет-поиска в общем случае, она может быть важным критерием при узком специализированном поиске). Кроме того, метапоиск используется для доступа к так называемому "глубинному Вебу" [3] и сверхточному поиску документов [4].

Первая версия системы ProThes была представлена на конференции Elpub-2003 [5]. Описание текущей версии, которая существенно отличается от первой, можно найти в [6].

В настоящей работе более подробно рассмотрены архитектура и средства реализации системы, описаны принципы, на которых построена МПМ ProThes: формат тезауруса, язык запросов, графический интерфейс пользователя, методы слияния и ранжирования результатов поиска. Описаны архитектура и основные компоненты ProThes, а также стандарты и инструментальные средства, которые были использованы при разработке. Говорится об ограничениях подхода и приводятся соображения по дальнейшему развитию системы.

1. Подходы к построению метапоисковой машины

1.1. Тезаурус

При нашем подходе тезаурус — это основа для процедур формирования запросов к машинам поиска Интернета. При разработке формата тезауруса мы руководствовались следующими принципами:

— тезаурус является автономным компонентом, т.е. не привязан к конкретной МП;

— тезаурус описывает терминологию узкой предметной области;

— основной элемент тезауруса — концепция (а не отдельный термин);

— концепции тезауруса связаны отношениями, семантика которых может быть различной (набор типов отношений не фиксируется).

В качестве формата представления тезауруса выбран язык XML, формат тезауруса описывается в виде XML Scheme. Подробно формат представления тезауруса описан в [7]; актуальная версия формата находится по адресу http://imach.uran.ru/pb/thesaurus/. Описание автоматических процедур формирования запросов с помощью тезауруса и предварительные оценки их эффективности можно найти в [8].

К настоящему моменту разработаны пробные тезаурусы предметных областей "Автоматический оптический контроль печатных плат" и "Технология редких металлов".

1.2. Язык запросов

Одна из проблем при разработке МПМ — определение единого языка запросов, который должен использоваться для обращения к нескольким машинам поиска. В нашем случае мы

пошли по пути наименьшего сопротивления и остановились на булевых запросах, которые представлены практически во всех МП. Пользователь может формировать запросы с логическими связками AND, OR, ANDNOT. Кроме того, поддерживается поиск по точным фразам.

Поскольку одной из особенностей системы является графический интерфейс пользователя, естественным решением было представить формируемый запрос в графической форме вместо традиционного текстового представления (см. также [9] о представлении булевых запросов в виде диаграмм Венна). Графический интерфейс для задания запросов имеет следующие преимущества:

— простота и наглядность представления;

— удобство работы;

— синтаксический контроль при составлении запросов.

Запрос представляется в виде дерева, где в качестве узлов выступают логические операции, а в качестве листьев — термины (рис. 1). Иерархическая структура позволяет избежать неоднозначностей интерпретации запросов различными поисковыми машинами. Например, запрос `mommy AND daddy OR son` Google интерпретирует как `mommy AND (daddy OR son)`, в то время как Яндекс этот запрос интерпретирует как `(mommy AND daddy) OR son`.

Важно ограничение на длину запроса, устанавливаемое поисковыми машинами. Например, у Яндекса максимальная длина запроса ограничена 255 символами (следует учитывать, что, поскольку запрос задается в формате XML, некоторые символы необходимо представлять в виде esacape-последовательностей: например, оператор `&&` следует представить в виде escape-последовательности `&&`, что ведет к увеличению длины запроса). Поэтому длинные запросы, сформированные пользователем, разбиваются по вхождениям OR на несколько “подзапросов”, которые по отдельности передаются на МП. В длинном запросе, использующем только операторы AND, последние термины отсекаются.

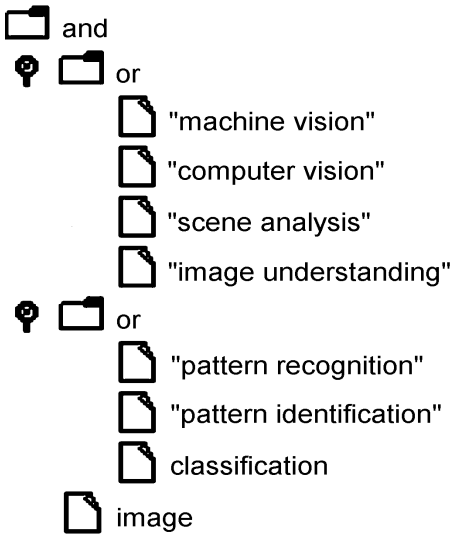

Рис. 1. Пример структуры запроса.

1.3. Графический интерфейс пользователя

Визуализация и просмотр тезауруса. Графический интерфейс пользователя системы ProThes представлен на рис. 2. Он состоит из списка терминов, отсортированных в алфавитном порядке, области визуализации тезауруса и области построения запроса.

Пользователь может выбрать термин из списка двойным щелчком мыши или нажатием клавиши Enter. Соответствующая концепция отобразится справа в области визуализации тезауруса вместе с концепциями-соседями и относящимися к ним терминами. Пользователь имеет возможность масштабировать и передвигать визуализированную часть тезауруса. Всплывающие подсказки предоставляют дополнительную информацию: определение концепции, пример использования термина и т. д.

Щелкая мышью на соседней концепции, мы делаем ее “центральной” — так мы можем просмотреть сеть связанных концепций всего тезауруса. Кнопки Back и Forward позволяют пользователю передвигаться по цепочке просмотренных концепций.

Также существует возможность включить языковой фильтр для терминов: отображать все термины или отображать термины только одного из языков.

Построение запросов. Пользователь системы может строить запрос как дерево с узлами в виде операций AND, OR, ANDNOT. Он может добавлять новые узлы дерева (как операторы, так и термины), удалять их и редактировать (изменять тип оператора, редактировать термин). Пользователь может выбирать для запроса термины тезауруса и вводить произвольные. Щелчок правой кнопкой мыши по термину добавляет сам термин, а по концепции — все термины данной концепции (по умолчанию объединяются по OR).

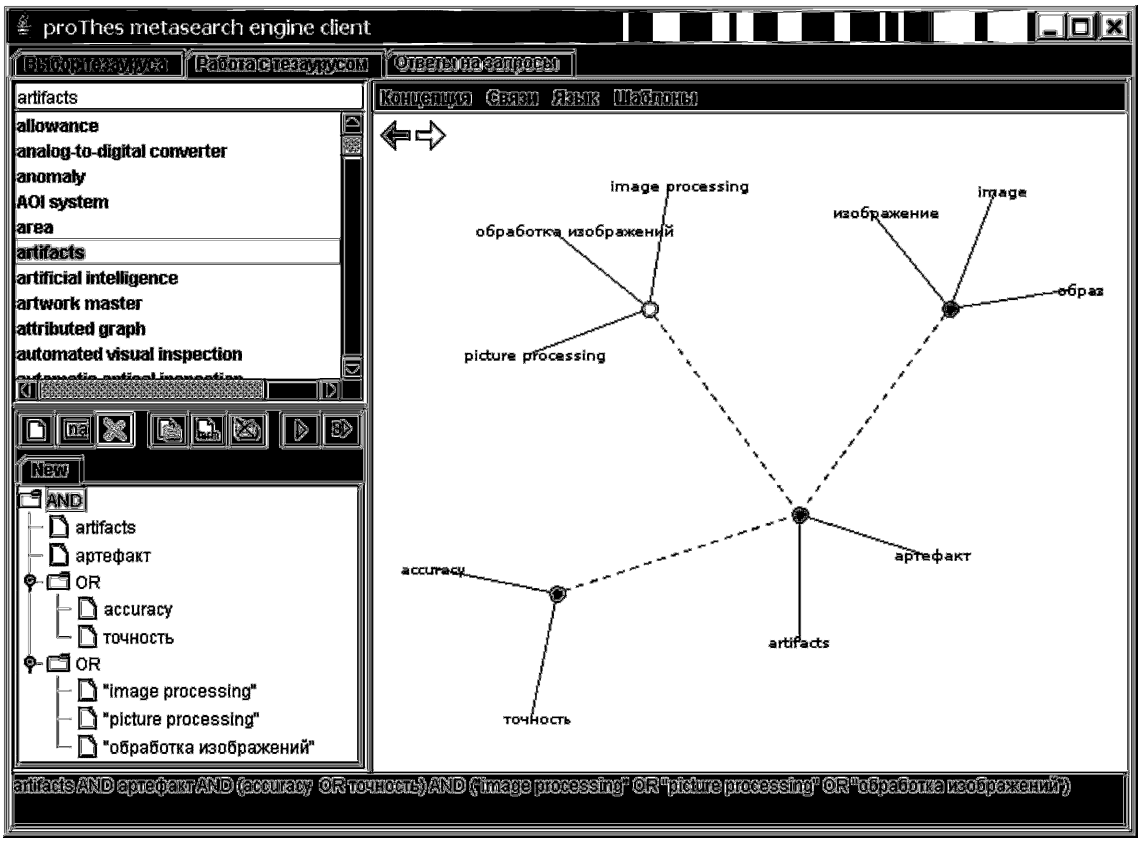

Рис. 2. Графический интерфейс пользователя.

Добавление происходит к текущему узлу запроса. Дерево запроса находится в нижнем левом углу окна пользовательского интерфейса.

Второй подход к построению запросов — это построение запросов на основе шаблонов. В шаблоне пользователь может определить следующие параметры:

— глубину расширения;

— используемые элементы терминологического входа (варианты, сокращения, однокоренные слова);

— набор типов связей, участвующих в построении запроса;

— оператор (OR, AND или ANDNOT);

— языковые параметры (построение одного запроса или нескольких — по одному для каждого языка).

Пользователь может применить шаблон к текущей концепции и при необходимости отредактировать автоматически построенный запрос перед отправкой на сервер. Ответ сервера будет отображен на вкладке “Результаты поиска”.

1.4. Слияние и ранжирование результатов

Обнаружение дубликатов в ProThes происходит по набору простых признаков: URL, имя и размер файла, заголовок документа. При ранжировании объединенного списка откликов мы располагаем лишь ограниченной информацией, содержащейся в отклике. Поэтому исходным параметром для ранжирования является исходная позиция документа в отклике, которая обычно вычисляется с использованием глобальной информации (например, ссылочной структуры Web). Правила ранжирования определяются с помощью XML-файла, в котором можно задавать “очки” для таких параметров, как домен, поисковая машина, формат документа.

Ранжирование результатов поиска в нашей системе производится по следующей формуле:

$$R = P_{\text{initial}} - a \cdot \operatorname{sign}(x+1) \cdot \ln\left(\operatorname{abs}(x+1)\right),$$

где P_{initial} — позиция ответа в отклике МП; a — положительная константа; x — дополнительные “очки”, набранные в соответствии с правилами ранжирования.

Пример XML-файла с параметрами ранжирования:

```
<?xml version="1.0" encoding="UTF-8"?> <rankingPreferences
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
      xsi:noNamespaceSchemaLocation="ranking.xsd">
   <extension rank="5">pdf</extension>
   <extension rank="10">ps</extension>
   <URL rank="2">http://www.a1dsn.com</URL>
   <URL rank="2">http://www.acae.com</URL>
   <URL rank="2">http://www.alphacinc.com</URL>
   <URL rank="2">http://www.appropriatetechcorp.com</URL>
   <URL rank="2">http://www.cadesign.net</URL>
   <URL rank="1">http://www.cadenergy.com</URL>
   <URL rank="2">http://www.capitaldesignservice.com</URL>
   <SE rank="1">yandex</SE>
   <SE rank="2">google</SE>
</rankingPreferences>
```

2. Архитектура системы

2.1. Общее описание

Система ProThes построена на основе архитектуры клиент-сервер. Для реализации серверной части системы была выбрана технология Java 2 Enterprise Edition. В качестве сервера приложений выступает Apache Tomcat (http://jakarta.apache.org), распространяемый по лицензии GPL. Клиентская часть реализована в виде *java applet*, что обеспечивает необходимую интерактивность и в то же время работу на различных платформах. Аналогичные технологии других производителей, к сожалению, являются платформозависимыми (например, технология Microsoft ActiveX). Обмен данными между клиентом и сервером происходит по протоколу SOAP, для чего используется свободно распространяемая библиотека Apache SOAP.

Серверная часть системы реализована в виде двух независимых web-служб (*Web services*): службы, отвечающей за работу клиента с тезаурусом, и службы, обеспечивающей обработку запросов (собственно функции метапоиска). Разделение на две службы произведено с целью повышения скорости работы системы и упрощения ее реализации.

2.2. Web-служба для работы с тезаурусом

Данная служба предоставляет клиенту интерфейс для работы с набором тезаурусов, хранящихся на сервере. Изначально тезаурус хранится в виде XML-документа. Поэтому вполне логично было бы использовать в качестве объектной модели тезауруса XML DOM. Однако при оценке производительности этого подхода, а также его реализации в условиях нашей задачи было принято решение отказаться от XML DOM в пользу самостоятельно разработанной объектной модели представления тезауруса.

Неэффективность использования XML DOM объясняется тем, что этот подход оптимизирован для иерархических структур данных. Тезаурус представляет собой сеть концепций с множеством связей. Использование XML DOM в данном случае повлекло бы за собой необходимость реализации большого количества алгоритмов выборки связанных элементов, что привело бы к усложнению логики работы и снижению производительности системы.

Оценка скорости работы созданной нами объектной модели тезауруса показала, что операции выборки подструктур тезауруса (наборов связанных концепций) выполняются быстро. Однако сериализация (представление информации в виде, удобном для передачи клиенту) этих данных занимает достаточно большое время (порядка нескольких секунд). В условиях большой загрузки сервера это может быть серьезным ограничением.

Дальнейшим шагом к повышению быстродействия сервера стало кэширование тезауруса уже в сериализованном виде. Это оправдано тем, что изменение тезаурусов на сервере происходит относительно редко. Реализация кэша позволила сократить время отклика системы в 7–8 раз (в зависимости от производительности компьютера). В идеале время жизни кэша тезауруса соответствует времени жизни сервера приложений, однако реализована возможность его динамической перестройки на случай изменений в наборе тезаурусов.

2.3. Web-служба обработки запросов

Система ProThes рассчитана на параллельную работу нескольких пользователей, поэтому необходимо обеспечить хранение данных об индивидуальных сеансах. Это послужило основной причиной выделения механизма метапоиска в отдельную web-службу, так как время жизни службы метапоиска (сеанс) не совпадает с временем жизни web-службы для работы с тезаурусом (равно времени жизни сервера приложений).

Подсистема метапоиска реализуется в виде трех компонентов: диспетчер запросов, набор адаптеров МП, буфер ответов.

Диспетчер запросов обеспечивает распределение пришедшего от клиента запроса по зарегистрированным в системе адаптерам МП, которые указываются в конфигурационном файле.

Адаптер МП реализует стандартный интерфейс к универсальной МП Интернета. На него возлагаются следующие функции:

— разбиение запроса на подзапросы в соответствии с ограничениями на длину запроса МП;

— перевод запросов в форму, соответствующую синтаксису МП;

— отправка сформированных запросов к МП;

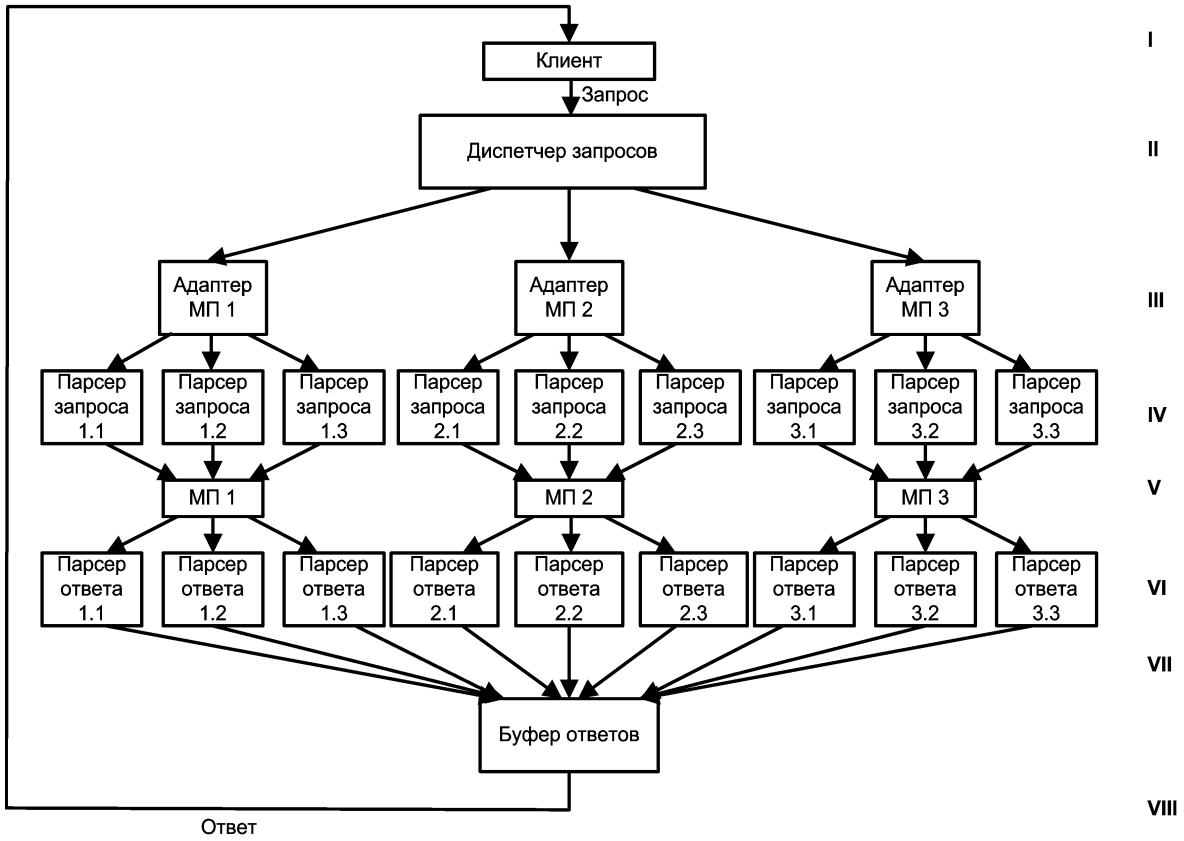

Рис. 3. Алгоритм работы службы метапоиска: I — поступление запроса от клиента во внутреннем виде; II — распределение запроса по адаптерам к МП; III — разбиение запросов и распределение по парсерам запросов; IV — парсинг запросов с учетом синтаксиса МП; V — обработка запроса МП; VI — обработка ответа МП; VI — помещение ответов в буфер, VIII — выборка ответов клиентом в диалоговом режиме.

— получение ответов от МП;
— перевод ответов МП во внутреннее унифицированное представление системы;
— помещение ответа в буфер.

Буфер ответов необходим для сокращения времени отклика сервера на запрос клиента: ответы, поступившие первыми, будут тут же затребованы клиентом, хотя, возможно, еще не все машины поиска вернули результат. По этой причине слияние и ранжирование результатов поиска производятся на клиенте (при поступлении новой порции ответов от сервера). Алгоритм работы web-службы метапоиска представлен на рис. 3.

На данный момент реализованы механизмы работы с машинами поиска Google (www.google.com) и Яндекс (www.yandex.ru), предоставляющими удобные программные интерфейсы: Google API (www.google.com/apis/) и Яндекс XML (xml.yandex.ru).

Заключение

Описаны подходы к построению и реализации метапоисковой системы ProThes. Предварительные эксперименты показывают, что ProThes может быть средством повышения эффективности поиска тематической информации с помощью универсальных машин поиска Интернета. Однако такой подход обнаруживает некоторые ограничения.

Препятствиями для активного развития системы являются лимит на количество запросов, задаваемых через специализированные API МП Google и Яндекс. Другой ограничивающий фактор — необходимость разработки тематических тезаурусов вручную. К ограничениям можно отнести также обеднение языка запросов МПМ по сравнению с языками запросов отдельных машин поиска (например, в МПМ нет возможности поиска в полях документа).

Загрузить клиентскую часть для работы с метапоисковой машиной, а также ознакомиться с подробным описанием интерфейса можно по адресу http://imach.uran.ru/prothes.

В дальнейшем наши усилия будут направлены на улучшение интерфейса пользователя и общего удобства использования, а также повышение скорости работы системы (в частности, за счет создания кэша запросов). Кроме того, мы планируем реализовать журналирование запросов с целью сбора статистической информации.

Список литературы

[1] Chakrabarti S., Berg M., Dom B. Focused crawling: a new approach to topic-specific Web resource discovery // Proc. of the 8th Intern. World Wide Web Conf., Toronto, Canada, 1999. http://www8.org/w8-papers/5a-search-query/crawling/index.html

[2] Budzik J., Hammond K.J. User interactions with everyday applications as context for just–in–time information access // Proc. of the Intern. Conf. on Intelligent User Interfaces, New Orleans, Louisiana, 2000. http://dent.infolab.nwu.edu/infolab/downloads/papers/paper10080.pdf

[3] Hamilton N. The mechanics of a deep net metasearch engine // Proc. of the 12th Intern. World Wide Web Conf., Budapest, Hungary, 2003. http://www2003.org/cdrom/papers/poster/p170/poster/poster.html

[4] Lawrence S., Giles C.L. Inquirus, the NECI meta search engine // Proc. of the 7th Intern. World Wide Web Conf., Brisbane, Australia, 1998. http://www7.scu.edu.au/programme/fullpapers/1906/com1906.htm

[5] Браславский П.И., Альшанский Г.А., Титов П.В. Формирование информационных запросов к машинам поиска Интернета на основе тезауруса: семантикоориентированный подход // Тр. 8-й Междунар. конф. по электронным публикациям “El-Pub 2003”. Новосибирск, 2003. http://www.ict.nsc.ru/ws/elpub2003/5964/

[6] Braslavski P., Shishkin A., Alshanski G. Meta-search, thesaurus, and gui for focused Web information retrieval // Digital Libraries: Advanced Methods and Technologies, Digital Collections: Proc. of the 6th National Rus. Research Conf. Pushchino, 2004. P. 135–140.

[7] Браславский П.И. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. “Диалог’2003”. 2003. С. 95–100.
http://www.dialog-21.ru/Archive/2003/Braslavskij.pdf

[8] Браславский П.И. Автоматические операции с запросами к машинам поиска Интернета на основе тезауруса: подходы и оценки // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. “Диалог’2004” (“Верхневолжский”, 2–7 июня, 2004). 2004. С. 79–84. http://www.dialog-21.ru/Archive/2004/Braslavskij.pdf

[9] Jones S., McInnes S., Staveley M.S. A graphical user interface for boolean query specification // Intern. J. on Digital Libraries Special Issue on User Interfaces for Digital Libraries. 1999. Vol. 2(2/3). P. 207–223.
http://www.cs.waikato.ac.nz/ stevej/Research/PAPERS/ijodlvquery.pdf

Поступила в редакцию 18 марта 2005 г.