ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
РАБОТЫ СО ЗНАНИЯМИ:
ОБНАРУЖЕНИЕ, ПОИСК, УПРАВЛЕНИЕ
(Вычислительные системы)
2008 год Выпуск 175

УДК 51:004.023

АВТОМАТИЗАЦИЯ ПОИСКА НАУЧНЫХ ПУБЛИКАЦИЙ В СЕТИ ИНТЕРНЕТ

А.А. Крушинская[1]

Введение и постановка задачи

Задача поиска научно–технической информации, находящейся в свободном поиске, является актуальной и важной. Быстрый рост количества документов в сети Интернет, активная деятельность по продвижению сайтов значительно осложняют задачу поиска информации в сети Интернет. Попытка поиска научной информации общеизвестными поисковыми системами приводит к результатам с большим количеством неподходящих документов. Поиск полнотекстовых научных публикаций: книг, статей, научных работ и обзоров является особенно сложной задачей, поскольку полный текст публикации, как правило, не выкладывается в свободный доступ или недоступен поисковой системе. Чаще всего названию работы сопутствует лишь краткая аннотация, что не является желаемым результатом поиска.

В данной работе была поставлена задача поиска научных публикаций и некоторой другой научной информации в сети Интернет. Требуется находить полные тексты статей, книг, научных работ и др. из области «Математика». Решение задачи подразумевает следующее:

[1]Новосибирский государственный университет

- поиск и описание справочных систем по математике:
  - сплайн–библиотек с доступом к полному тексту книг и статей,
  - научных порталов,
  - сайтов журналов и изданий,
  - математических систем, ориентированных на математические публикации (или более широкий набор наук, включающий математику),
  - специализированных поисковых систем;
- создание механизма поиска научной информации в составленном каталоге ресурсов и в сети Интернет вообще;
- механизм поиска должен предоставлять возможность поиска по типу информации, т.е. пользователь должен иметь возможность выбрать, к какому типу будут относиться результаты (например, статьи или книги);
- результатом поиска должны становиться документы, непосредственно относящиеся к выбранному типу информации, должно быть как можно меньше «неподходящих» документов.

В работе были заданы типы информации, которые пользователь может выбрать для поиска:

- статьи (публикации с полными текстами);
- организации (информация с сайтов организаций);
- диссертационные советы (информация о диссертационных советах);
- сайты жургалов (статьи с сайтов журналов, печатные издания);
- электронные издания (публикации в таких изданиях приравниваются к публикациям в научных печатных журналах);
- форумы (обсуждения);
- научные сообщества (информация, публикуемая в сообществах);
- конференции (информация о конфиренциях, тезисы);
- научные школы;
- электронные библиотеки (полные тексты книг, статей);
- персональные страницы.

Специализированные информационно–поисковые системы

В настоящее время существуют специализированные информационно–поисковые системы, которые занимаются поиском научно–технической информации. Они опираются на те или иные источники научной информации и позволяют находить полные тексты научных публикаций как платные, так и свободные; однако, они не являются общеизвестными, и лишь единицы позволяют искать русскоязычные публикации. Кроме поисковых систем существует множество интернет–ресурсов (научных порталов, библиотек, сайтов издательств и журналов), которые в виде каталога или списков предлагают научные книги, статьи и другое, причем, с возможностью свободно получать полный текст; но подобные ресурсы являются еще менее известными, чем специализированные поисковые системы. При этом даже если пользователь знает адреса этих ресурсов, то вручную провести поиск по ним довольно долгая и утомительная работа.

Рассмотрим некоторые наиболее популярные поисковые решения для научного поиска в русском секторе Интернета.

Google Scholar является службой поисковой системы Google, производит поиск среди статей, книг, обзоров научной литературы. Включает статьи крупных научных издательств, архивы препринтов, публикации на сайтах университетов, научных обществ и других научных организаций. Ищет статьи в том числе и на русском языке, кроме того в разделе Advanced search можно задать область поиска «Engineering, Computer Science, and Mathematics». Однако после такого уточнения области поиска, результат поиска по русскоязычным запросам мал.

Scirus — универсальная научная поисковая система. Осуществляет полнотекстовой поиск по статьям журналов большинства крупных иностранных издательств (порядка 17 млн. статей), по статьям в крупных архивах статей и препринтов, по научным ресурсам Internet (более 250 млн. проиндексированных страниц). Многократно признавалась лучшей специализированной поисковой системой. Позволяет выбирать какого типа информацию искать: статьи, конфернеции, книги, патенты и др. Система дает очень хорошие результаты при поиске на английском

языке; нельзя сказасть, что на русском она дает плохие результаты, но ее русская база, видимо, не столь обширна.

Windows Live Academic — бета–версия научной поисковой системы от Microsoft. Предназначена для поиска научных статей как в открытых источниках, так и в архивах изданий с платным доступом. Не предлагает никакого разделения по типам информации, результаты поиска по русскоязычным запросам ненамного превосходят результаты простого поиска в Google.

В поставленной задаче требуется не просто находить научные публикации и информацию, но также разделять ее по типам ресурсов. Одним из типов являются «Электронные библиотеки». При выборе этого типа пользователь задает, что поиск по его запросу следует проводить в электронных библиотеках, и результатом этого поиска должны стать электронные книги, а точнее, полные тексты книг. Этот тип ресурсов следует рассмотреть подробнее, поскольку в этом направлении ведутся работы и создаются различные системы для поиска книг, в том числе научной литературы. Были найдены некоторые решения для поиска полных текстов книг научного содержания по электронным библиотекам.

«Поиск книг» (www.poiskknig.ru). Поиск проводится по ранее составленному индексу по библиотекам. В поиске участвуют следующие источники: lib.mexmat.ru (с сайта нельзя скачивать книги, поисковик их имеет в своей базе, возможно ранее выкачаны с сайта), lib.homelinux.org (не открывается), lib.org.by (белорусская библиотека, не очень большая), sci–lib.com, mccme.ru, math.ru, ihtik.lin.ru (математический раздел закрыт), osnovanija.narod.ru (персональная библиотека), lib.ru (художественная литература). Поиск не ориентирован на математическую литературу.

Google Books (books.google.ru). Поиск по русскоязычным запросам неудовлетворителен, в основном находятся краткие статейки из энциклопедий (часто на транслите), книги практически не находит, хотя для английского языка хорошо находит и книги.

Книжная поисковая система **eBdb** (www.ebdb.ru). Проект eBdb является попыткой создания специализированной поисковой системы в области электронных книг. Поиск в этой системе

дает действительно хорошие результаты с учетом, что пользователь ищет математическую литературу и на русском языке.

Поиск научно–технической информации при помощи виртуального каталога. Подбор эвристик

Для поиска публикаций можно было бы проиндексировать все сайты библиотек, журналов, изданий и др., составив индекс имеющихся публикаций (названия, авторы и др.), и далее проводить поиск по этому индексу, как это сейчас делается в специализированных поисковых системах. Но в виду того, что сайты могут обновляться, изменяться, перестать работать или могут появиться новые, этот индекс придется периодически обновлять, что требует времени и трафика, кроме того невозможно учесть все такие ресурсы.

Поэтому решено было использовать для поиска внутри этих сайтов информационно–поисковую систему Google. При достаточно точном запросе Google выдает малое число результатов, и все они соответствуют запросу, или же, если не найден ни один докумет, удовлетворяющий запросу, то мусор, состоящий из документов, удовлетворяющих только части запроса, не отображается. Таким образом, если достаточно точно задать дополнение к пользовательскому запросу [3,4], то среди результатов поиска будут страницы, содержащие ключевые слова и относящиеся к нужной тематике и типу. То есть необходимо найти поисковые эвристики, которыми лучше всего дополнять запросы пользователя. Эвристика — это некоторая конструкция из слов и символов, которая наилучшим образом уточняет запрос, т.е. дает лучшие результаты поиска по уточненому запросу.

Кратко метод можно описать следующим образом (на примере поиска книги).

На входе есть:

Ключевые слова, введенные пользователем (часть названия книги, имя автора или еще что–то).

xml–файл со списком библиотечных эвристик.

Обработка:

Разбор xml–файла.

Для каждой эвристики из xml–файла отправление скомпонованного запроса в Google, получение от Google результатов поиска по нему.

Объединение результатов по всем запросам, подготовка итогового представления.

На выходе (*то, что получает пользователь*):

Список книг, удовлетворяющий введенным ключевым словам, полный текст этих книг доступен из соответствующих библиотек.

Эвристики находились для каждого типа ресурсов (вида информации) отдельно. Для этого рассматривались научные ресурсы Интернета, на которых или с помощью которых можно найти требуемый вид информации. Для каждого типа ресурсов может быть по несколько эвристик, как узкоспециализированных, т.е. для проведения поиска на сайте какого–то конкретного ресурса, так и широкого охвата, для проведения общего поиска в Google или Yandex. В состав эвристик входят ключевые слова, операторы языков запросов поисковых систем и интернет–адреса этих ресурсов.

Так как список ресурсов может изменяться, то жестко записывать эвристики для него в программную реализацию метода неудобно. Необходим отдельный элемент, который будет содержать эвристики для этого списка и который будет погружаться в систему поиска без необходимости изменения самой системы. Список эвристик ресурсов решено было сделать в виде xml–файла.

Результаты работы встроены как функционал в метапоисковую систему Metasearch, которая разрабатывается в Институте математики им. С.Л.Соболева.

Система работает с тремя областями: математика, патентология, катализ. Область выбирается при входе в систему и далее работа происходит уже в ней. Система Metasearch представляет собой виртуальный каталог [1,2]. Она имеет рубрикатор, но наполнение категорий не статическое, а динамическое. Когда пользователь выбирает какую–то рубрику, он может ввести ключевые

слова и выбрать вид информации, далее пользователю надо нажать на кнопку «Поиск». После этого произойдет наполнение выбранной категории документами (представляются в виде ссылок на другие сайты), которое происходит за счет проведения поиска в информационно–поисковых системах Google или Yandex по составному запросу. В состав запроса входят ключевые слова, эвристики для выбранной рубрики и вида информации.

Эвристики рубрик и эвристики типов ресурсов пишутся в два разных xml–файла. Формат файлов одинаков:

```
<Keywords>
<queres id="1" >
<query engine="Google" ></query>
<query engine="Yandex" ></query>
</queres>
...
</Keywords>
```

Здесь id — это универсальный идентификатор рубрики или типа ресурсов, теги query описывают эвристики для соотвествующей информационно–поисковой системы.

Каждый тип ресурсов имеет свой идентификатор от 1 до 11, по порядку в списке. Реализована функциональность по идентификатору выбранного пользователем типа ресурсов брать из xml–файла соответствующие эвристики. Далее эвристики типа компануются в эвристики рубрики. Новые эвристики отправляются в функцию, проводящую поиск с их использованием.

Поиск, проводимый с помощью системы Metasearch, т.е. с использованием поисковых эвристик для типов ресурсов, дает лучшие результаты, чем простой поиск общеизвестных информационно–поисковых систем. Прежде всего потому, что пользователь имеет возможность выбрать тип требуемой информации и получить результаты этого типа, вместо того, чтобы перебирать построчно результаты поиска в информационно–поисковой системе.

Рассмотрим некоторые типы ресурсов. При выборе типа «Электронные библиотеки» результатами поиска становятся электронные книги по математике; если пользователь знает авторов или название книги, то книга находится с большой вероятностью. Поиск в типе «Электронные издания» — это поиск статей в очень ограниченной группе журналов, соответствующей стандарту [5]; здесь, чтобы получить нужную информацию, надо поточнее формулировать ключевые слова запроса, так как обычные математические термины встречаются во многих статьях довольно часто. При выборе типа «Сайты журналов» поиск ведется среди описаний статей на сайтах журналов; здесь при поиске удобно использовать фамилии авторов статей и названия (если оно известно), тогда вероятность нахождения нужной статьи значительно увеличивается, причем не просто статьи, но и ее полного текста. При поиске по типу «Организации» находится различная информация, выкладываемая математическими организациями: институты, научные центры, факультеты, кафедры и т.д.

Заключение

Была поставлена задача поиска полнотекстовых публикаций из области математика и некоторой научной информации с разделением поиска по типам ресурсов.

В ходе решения задачи:

- проведен обзор имеющихся методов и средств поиска научной информации, решающих некоторые из частей поставленной задачи;
- собрана большая коллекция ресурсов (сайты библиотек, журналов, изданий, научных институтов и т.п., около 70 ресурсов), на которых выкладываются научные работы, публикации с полным текстом;
- на основе собранной коллекции создан метод поиска научных публикаций и некоторой другой научной информации с разделением поиска по типу ресурсов, коллекция разбита на категории ресурсов по типам;
- для отдельных категорий ресурсов найдены поисковые эвристики, дающие хорошие результаты поиска, эвристики записаны в виде xml–файлов для каждого типа ресурсов;

• проведен анадиз рузультатов, получаемых с использованием метода, на основе сравнения с результатами простого поиска в Google, показана эффективность метода;

• метод реализован как функционал в системе Matasearch.

Литература

1. ПАЛЬЧУНОВ Д.Е., СИДОРОВА Е.С. Виртуальный каталог./Тр. Всероссийской конф. "Знания–Онтология–Теория". Новосибирск,август 2007 г. – Новосибирск, 2007. – С.166-175.

2. ПАЛЬЧУНОВ Д.Е. Решение задачи поиска информации на основе онтологий// Бизнес–информатика. – 2008. – № 1. – С.3–13.

3. VLEZ B., WEISS R., SHELDON M.A., GIFFORD D.K. Fast and effective query refinement// Proceedings of the 20th annual international ACM SIGIR conf. on Research and development in information retrieval. – ACM Press. – 1997. – P.6–15.

4. KRAFT Reiner, ZIEN Jason. Mining Anchor Text for Query Refinement. Int. WWW Conf.Proceed.13th on WWW.IBM Labs. – ACM Association of Computing, 2004. – P.666–674.

5. Положение о порядке регистрации сетевых электронных научных изданий, публикации в которых учитываются при защите диссертационных работ, от 18 апреля 2002 г. [Электронный ресурс].

Поступила в редакцию
16 июня 2008 года