УДК 681.3

Ю.В. Рогушина, И.Ю. Гришанова

РАЗРАБОТКА ПРИНЦИПОВ ПРЕДСТАВЛЕНИЯ ЭЛЕКТРОННЫХ ИЗДАНИЙ, ОБЕСПЕЧИВАЮЩИХ КОРРЕКТНУЮ ИНДЕКСАЦИЮ ПОИСКОВЫМИ СИСТЕМАМИ ИНТЕРНЕТА

Рассматриваются вопросы, связанные с преобразованием традиционного (бумажного) представления информации в научном журнале в Web-представление. Анализируется специфика представления информации в таких изданиях и обеспечение ее индексации поисковыми системами.

Введение

За последние годы Интернет превратился в одно из основных средств публикации информации. Создание электронных версий традиционных изданий позволяет обеспечить более широкий и быстрый доступ к содержащейся в них информации и упрощает поиск необходимых пользователю сведений. Очевидно, что наличие в библиотеке одного экземпляра журнала не может обеспечить доступ к нему одновременно всех желающих. При этом высока вероятность того, что содержащиеся в нем материалы интересны только небольшой части потенциальных читателей. В то же время электронные версии не замещают полностью бумажные издания, так как подавляющее большинство пользователей предпочитает при детальной работе с публикациями использовать именно традиционные бумажные версии изданий.

Сегодня известны и широко используются научными сообществами академические репозитории электронных публикаций, электронные библиотеки (в основном в сфере естественных наук). Например, база данных цитирования Nec ResearchIndex [1] предназначена для распространения научной литературы и обработки отзывов на публикации, а сервер DBLP [2] предоставляет библиографическую информацию о научных журналах и материалах конференций в области компьютерных наук.

Наличие электронных версий научных журналов позволяет не только читателям, но и авторам статей непосредственно через Интернет ознакомиться с требованиями к оформлению статей, составом редакционной коллегии и тематической направленностью журнала, а также переслать в редакцию по электронной почте файлы своих статей.

Анализ публикаций показывает, что сейчас широкое распространение получили различные стандарты (например, BibTex [3], SGML [4], XML [5]) и программные продукты для создания электронных библиотек, однако большинство из них сложны в эксплуатации, имеют проблемы с локализацией и чересчур требовательны к ресурсам.

На сегодня довольно много публикаций посвящается примерам разработки электронных версий научных изданий [6]. Однако большинство из них основное внимание уделяют контенту своих сайтов, а не общим принципам их разработки и не помогают другим разработчикам найти разумный компромисс между эффективностью функционирования сайта и его интеллектуализацией.

Постановка задачи

Для того чтобы пользователи могли обнаружить информационные ресурсы, соответствующие их информационным потребностям, необходимы усилия со стороны не только разработчиков информационно-поисковых систем, но и тех, кто публикует информацию.

В данной статье мы рассматриваем принципы, согласно которым следует разрабатывать электронные версии научных изданий, чтобы обеспечить, с одной стороны, удобный пользовательский интерфейс как для читателей, так и для авторов статей, а с другой — доступность предлагаемых информационных ресурсов для информационно-поисковых систем. Эти ресурсы имеют большой объем и хорошо структурированы, поэтому для их хранения используются локальные базы данных. Проблема заключается в том, что содержимое таких баз данных не индексируется информационно-поисковыми системами, и поэтому при разработке структуры такого сайта необходимо учитывать специфику индексации его элементов поисковыми системами, а также сопровождать его корректными метаописаниями.

Основные принципы разработки сайта научного журнала

Авторами статьи разработана электронная версия журнала "Проблемы программирования", которая использовалась для тестирования различных подходов к индексированию информационных ресурсов Интернет поисковыми системами (так как известны ее структура и информационное наполнение, то можно было сравнить реальные результаты индексирования с декларируемыми). Кроме того, мы ставили своей целью представить информацию таким образом, чтобы пользователи, для которых представляют интерес публикации журнала, могли найти сведения о них при помощи глобальных поисковых систем.

Принципы разработки конкретного сайта определяются путем анализа ряда параметров:

1. Определение целевой аудитории сайта. Перед тем, как создавать сайт, необходимо определить его целевую аудиторию. Можно предположить, что посетителями сайта — электронной версии журнала "Проблемы программирования" (http://www.progproblems.org.ua) являются научные сотрудники, аспиранты и т.д., специализирующиеся в области ИТ. Из этого следует, что они имеют достаточные навыки в обращении с достаточно сложным программным обеспечением. Поэтому на сайте можно использовать развитые языки запросов, поисковые онтологии, подсистему регистрации пользователей и не сопровождать их подробными подсказками и средствами защиты от случайных ошибок.

2. Требования к защите информации. Так как информация на сайте не является конфиденциальной (она уже опубликована в журнале), то нет необходимости в сложных средствах защиты ее от несанкционированного доступа. Посетители сайта могут получить информацию о статьях, опубликованных в журнале (http://www.progproblems.org.ua/archiv.php), о правилах подачи публикаций и об Институте программных систем НАНУ, который выпускает данный журнал. Кроме того, на сайте представлена дополнительная информация, ориентированная на ту же целевую аудиторию — о научных конференциях ИПС НАНУ, о научных и учебных изданиях, подготовленных сотрудниками института (http://www.progproblems.org.ua/news.php), и т. д.

3. Цель разработки сайта. Цель разработки сайта — не получение прибыли, а привлечение внимания к публикациям журнала именно тех людей, которые специализируются в научных исследованиях, связанных с разработкой программных систем. Поэтому необходимо четко описать тематическую направленность сайта в виде, удобном для пользователей, и в формате, который могут автоматически обрабатывать поисковые механизмы Интернета.

4. Техническая база целевой аудитории сайта. Дизайн сайта разрабатывался с учетом достаточно слабой технической базы НАН Украины, поэтому в нем отсутствуют громоздкие графические объекты, не несущие смыслового наполнения. Рисунки сопровождаются текстовым описанием в атрибуте ALT, что позволяет поисковым системам индексировать и их.

При разработке сайта не использовались фреймы и ролики Flash, которые плохо индексируются поисковыми роботами и не позволяют пользователям сохранять необходимую им информацию. На сайте отсутствуют рекламные баннеры, что ускоряет его загрузку и не отвлекает пользователей.

5. Особенности методов индексации современных поисковых систем. Многие поисковые системы не индексируют страницы сайта глубже 3-го уровня. Поэтому разработанный нами сайт имеет 3 уровня вложенности. Первая страница содержит основные ключевые фразы и выражения. Все страницы сайта содержат ссылки на главную страницу, что повышает рейтинг этой страницы при ее индексации поисковыми системами.

6. Создание метаданных, описывающих контент сайта и обеспечивающих его автоматическую интерпретацию его семантики. На сайте представлена онтология предметной области журнала. Она содержит ключевые слова и словосочетания, при наличии которых в запросах поисковые системы должны находить сайт. Метаописания информационных ресурсов позволяют их авторам явным образом сообщить определенные сведения, характеризующие эти ресурсы. Метаданные могут сохраняться и обновляться независимо от ресурсов.

Сегодня наиболее популярны стандарт описания информационных ресурсов RDF (Resource Description Framework) [7], разработанный консорциумом W3C в рамках проекта Semantic Web, и набор элементов для создания метаданных "Dublin Core Metadata Elements" [8]. Все страницы сайта содержат тэги заголовка и метатэги, а также синтаксически правильные RDF-описания, использующие элементы набора Dublin Core. В титульных фразах страниц используются слова, часто встречающиеся в поисковых запросах пользователей, ответом на которые могут являться эти страницы.

7. Семантическая насыщенность html-кода. Размер каждой страницы относительно невелик, так как поисковые системы плохо обрабатывают большие документы, а пользователям удобнее работать с хорошо структурированной информацией. Логическое выделение заголовков производится не стилями font, а тэгами заголовков <H>. Для расшифровки аббревиатур используется тэг <acronym>, а наиболее важная информация выделена тэгами логического выделения.

Текст, расположенный внутри гиперссылки, считается поисковыми системами важным и индексируется. Поэтому на страницах сайта широко используются тэги гипер-

ссылки, которые сопровождаются атрибутом title с "всплывающей" подсказкой.

8. Выбор имени сайта. Наиболее важными элементами контента сайта поисковые системы считают заголовок страницы и URL, особенно имя домена. То, что сайт относится к зоне ORG, повышает его рейтинг у поисковых систем, так как отображаемая на таких сайтах информация обычно достоверна и корректно оформлена.

9. Релевантность текста. Каждая поисковая система по-своему определяет вес документа по отношению к запросу пользователя. На оценку поисковой системы влияют различные факторы, начиная с имени домена и заканчивая общим количеством слов запроса в документе. Необходимо, чтобы в документе явным образом присутствовали слова, задекларированные в его метаописании, так как в настоящее время поисковые системы не используют онтологический подход при анализе страниц, а просто проверяют наличие ключевых слов.

10. Популярность сайта. Популярность сайта определяется не только количеством ссылок на него, найденных поисковой машиной, но и релевантностью этих ссылок. При разработке научного сайта важно придерживаться общепринятой терминологии, при этом рекомендуется не включать по возможности в заголовки и метаописания слова, часто используемые в других областях с иными значениями.

11. Позиционирование сайта. Уровень конкуренции понижается по мере смещения от общего к частному. Можно добиться более высокого результата и дольше удерживать его, сузив тематику сайта. Для этого рекомендуется включать в метаописание научного издания не только его название, но и основные тематические рубрики, подразделы, названия специальных выпусков и т.д.

Специфика разработки электронных версий научных изданий

Следует выделить специфику создания электронных версий именно научных журналов, связанную с особенностями представленной в них информации.

1. Научные статьи имеют такой обязательный атрибут (кроме фамилий авторов и названия), как *аннотация*. Это устраняет необходимость реферирования статей, так как автор статьи сам выделяет наиболее существенную информацию о предложенных им материалах, что, разумеется, обеспечивает более высокое качество отражения смысла статьи, чем автоматическое реферирование.

2. В научных статьях наряду с текстовой информацией часто содержатся *рисунки, таблицы, графики* и *формулы*. В то же время такие мультимедийные элементы, как звуковые и видеоролики, в них отсутствуют (в связи с невозможностью их включения в бумажный вариант издания). В то время как представление изображений и таблиц в формате HTML не вызывает практически никаких проблем, то преобразование формул в графические изображения — крайне трудоемкая процедура. Анализ материалов, опубликованных в журнале “Проблемы программирования“, показывает, что научная статья может содержать от 1–2 до нескольких десятков или сотен формул. Кроме того, формулы, преобразованные в изображения, менее читабельны и не могут редактироваться. Таким образом, преобразование научных статей в формат HTML представляется нецелесообразным. Поэтому статьи, помещенные на сайте, представлены в формате Microsoft Word или в формате *.rtf (что соответствует требованиям редакции к тек-

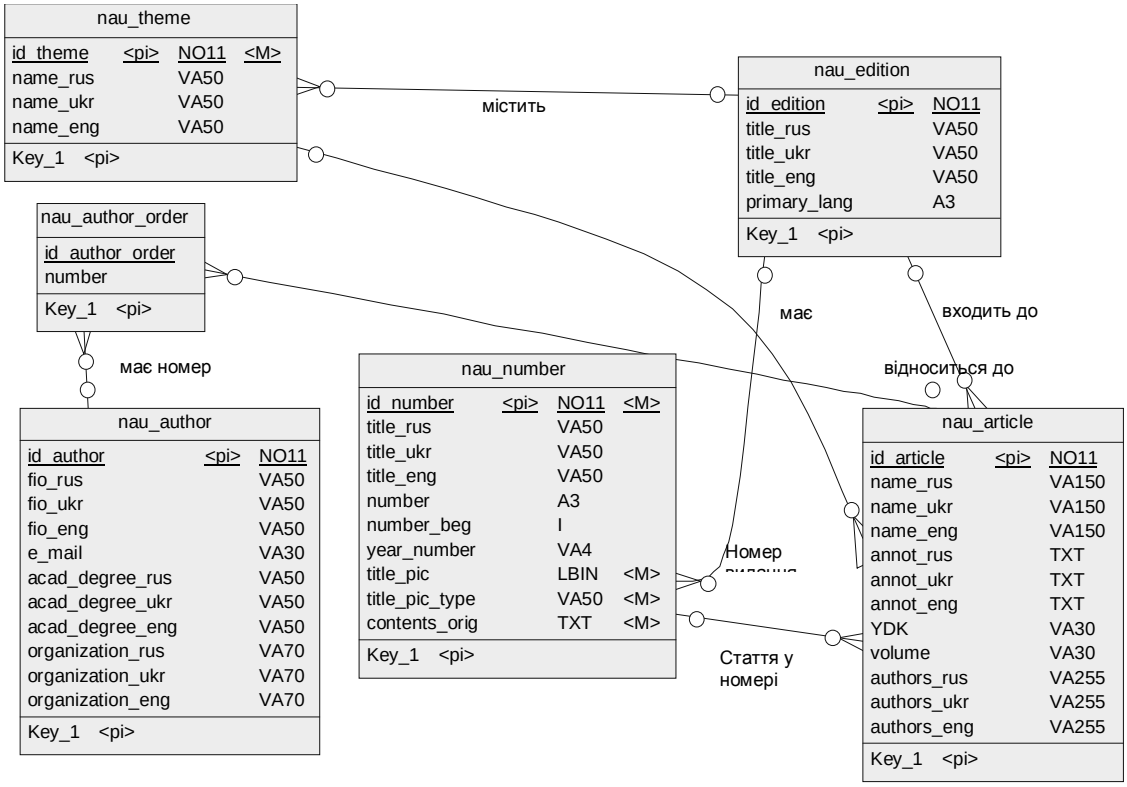

Рис. 1. Концептуальная модель БД сайта "Проблемы программирования"

стовому файлу, предоставляемому авторами), а список литературы может быть оформлен в виде гиперссылок на ресурсы Интернета.

3. Каждая из опубликованных в научном журнале статей обычно бывает явным образом отнесена к одной из *тематических рубрик*. Название этих рубрик практически не меняется от номера к номеру и может быть положено в основу классификации статей. Однако в связи с ограниченностью количества таких рубрик некоторые статьи могут классифицироваться неоднозначно (в том смысле, что при публикации статья была отнесена к одному разделу, а читатель ожидает обнаружить ее в другом). Несмотря на это, подобная классификация достаточно полезна при поиске.

Таким образом, можно выделить следующие атрибуты статьи, публикуемой в научном журнале:

- Фамилии авторов;
- Название статьи;
- Аннотация;
- Год публикации;
- Номер журнала, в котором была опубликована статья;
- Тематический раздел;
- Имя файла, в котором хранится текст статьи.

Эта информация содержится в базе данных. Ее использование позволяет осуществлять поисковые запросы пользователей. В данном случае поисковые средства ориентированы только на текстовую информацию.

Концептуальная модель БД приведена на рис. 1.

Так как база данных не индексируется поисковыми системами, необходимо включить в информационное наполнение сайта альтернативное представление хранящихся в ней сведений, которое поддается индексации. Для этого мы дополнили главную страницу сайта ссылкой на архив. Архив представляет собой перечень всех публикаций журнала, структурированный по годам.

Форма представления информационного запроса пользователя

Чтобы четче охарактеризовать возможности, предоставляемые поисковой машиной сайта, введем ряд формализмов.

1. Существуют поля $P_1,...,P_n$, по которым осуществляется поиск

(в данном случае, например, “Авторы“, “Название“ и т.д.).

Для осуществления поиска пользователь задает в своем запросе их значения $p_1,...,p_n$, причем хоть одно из полей не пусто.

В дальнейшем поиск может производиться одним из трех возможных способов (очевидно, что первый и второй способы являются частными случаями третьего):

1) для всех непустых полей выполняется запрос $p_i \& ... \& p_k$;

2) для всех непустых полей выполняется запрос $p_i \vee ... \vee p_k$;

3) вводится $l_i, i = \overline{1,n}, l_i \in \{\&, \vee\}$.

Для всех непустых полей выполняется запрос вида $(p_i \vee ... p_k) \& ... \& p_m$.

2. Анализ информации, содержащейся внутри значений полей.

Если значение p_i некоторого поля P_i не пусто, то в общем случае его можно рассматривать как последовательность слов $a_{i_j}, i_j = \overline{1,i_k}$, разделенных символами-разделителями из множества E. E = {“.“, “,“, “ “, “-“, “!“, “&“, “$\vee$“}. Такой набор слов может интерпретироваться при поиске различными способами:

1) $p_i = a_{i_1} \& ... \& a_{i_k}$;

2) $p_i = a_{i_1} \vee ... \vee a_{i_k}$;

3) символы “&“ и “$\vee$“ используются явным образом при задании формулы вида

$$p_i = a_{i_{1_1}} \& ... \& a_{i_{1_m}} \vee ... \vee a_{i_{k_1}} \& ... \& a_{i_{k_m}} .$$

Очевидно, что, как и в п. 1, первый и второй способы интерпретации являются частными случаями третьего.

3. Достаточно часто при задании поискового запроса семантически значимыми являются не целые слова, а их части (т.е. в большинстве случаев окончание слова можно отбросить). Осуществить такое преобразование можно с помощью функции

$$f(a_i) = f(\langle r_i, s_i, t_i \rangle) = s_i .$$

Например,
f(“параллельных“) =
= “параллельн“. Такое преобразование полезно, если пользователь, например, не знает точного названия статьи, но знает слова, которые в нем были использованы.

Реализация такого преобразования требует использования баз данных, содержащих соответствующую лингвистическую информацию, и правила преобразования. Проблема усложняется тем, что статьи в “Проблемах программирования“ публикуются на трех языках — русском, украинском и английском. Поэтому мы рекомендуем пользователям выполнять это преобразование самостоятельно.

4. В некоторых случаях при поиске необходимо интерпретировать пробелы и другие символы из множества E не как разделители, а непосредственно как символы, которые следует найти. Кроме того, для пользователя может быть существенным порядок слов в значении поля. Например, пользователь ищет строку “реализации особенности“, но его не интересуют статьи, содержащие “особенности реализации“. В этом случае фрагмент текста заключается в кавычки и содержащийся между открывающей и закрывающей кавычками текст интерпретируется как одно слово.

5. Некоторые поисковые машины (глобальные и локальные) используют символы “*“ и “?“ для задания шаблонов поиска. Хотя в настоящее время необходимость в такой разновидности поиска на сайте электронной версии научного журнала не очевидна, целесообразно зарезервировать эти символы как специальные.

В настоящее время на сайте журнала “Проблемы программирования“ реализованы возможности, описанные в 1.1. и 2.2. Возможности, описанные в 1.2., 2.1. и 4., находятся в стадии разработки.

Обеспечение доступности информации, содержащейся в электронной версии научного журнала

В настоящее время доступ к информации, размещенной в глобальной сети Интернета, в подавляющем большинстве случаев обеспечивается при помощи поисковых машин. К сожалению, с работой поисковых машин связан ряд проблем, которые значительно снижают эффективность и релевантность поиска.

Обычно поисковая машина с помощью роботов индексирует страницы Интернета часто с учетом задаваемой авторами простой метаинформации. При этом индексируется лишь небольшая часть содержимого сети, причем делается это с опозданием. Но основная проблема заключается не в этом. Когда на начальном этапе развития Интернета ресурсы были в основном статическими и их было относительно немного, то поисковые машины справлялись с этой работой, хотя отставание наметилось сразу же. Но в настоящее время в связи с возрастанием объема и усложнением структуры информации на сайтах становится выгоднее хранить информацию не непосредственно на самих страницах, а отдельно в базах данных, используя Web лишь как универсальный интерфейс к этим базам.

Вся информация в таких базах данных доступна посетителям сайта (обычно при помощи локальной поисковой машины). Но глобальные информационно-поисковые системы не способны предложить ее пользователям в ответ на запрос, так как роботы не могут ее проиндексировать. Вследствие этого большинство пользователей не могут воспользоваться информацией этого сайта, так как не имеют сведений о ней.

Согласно [9], объем “глубинной“ части Web (Deep Web) во много раз больше “поверхностной“ (Surface Web), проиндексированной всеми традиционными поисковыми системами вместе взятыми. Это соотношение продолжает расти, поскольку тенденция к хранению информации в структурированных источниках очевидна и по крайней мере в ближайшие годы не изменится.

В связи с этим возникает потребность в разработке программного обеспечения, позволяющего обращаться за информацией к локальным поисковым машинам сайтов без непосредственного участия пользователя. Подобные функции способна предоставить мультиагентная информационно-поисковая система, в состав которой входят агенты информационных ресурсов, обеспечивающие интерфейс с локальными поисковыми системами различных сайтов, и агент-диспетчер, предоставляющий перечень таких сайтов [10]. Следует заметить, что создание интеллектуального поискового агента или мультиагентной системы, компетентной одновременно во многих предметных областях, является слишком сложной задачей, но эту проблему позволяют решить средства взаимодействия и обмена знаниями между агентами.

Агент информационного ресурса, получив запрос пользователя, преобразует его в формат, понятный локальной поисковой машине этого ресурса, а также дополняет (при необходимости) информацией, задающей контекст поиска. Для предоставления агенту информационного ресурса адекватных средств задания запроса, необходимо, чтобы локальная поисковая машина сайта поддерживала достаточно сложные запросы.

Программная реализация

С этой целью для сайта электронной версии журнала “Проблемы программирования” (рис. 2) были использованы серверный язык создания сценариев PHP [11] и система управления реляционными базами данных MySQL [12]. Сервер MySQL управляет доступом к данным, позволяя работать с ними одновременно нескольким пользователям, обеспечивает быстрый доступ к данным и гарантирует предоставление доступа только легитимным пользователям. И PHP, и MySQL являются продуктами с открытым кодом. Оба пакета (PHP и MySQL) доступны бесплатно, имеют высокую производительность и переносимость. Это и определило выбор инструментария.

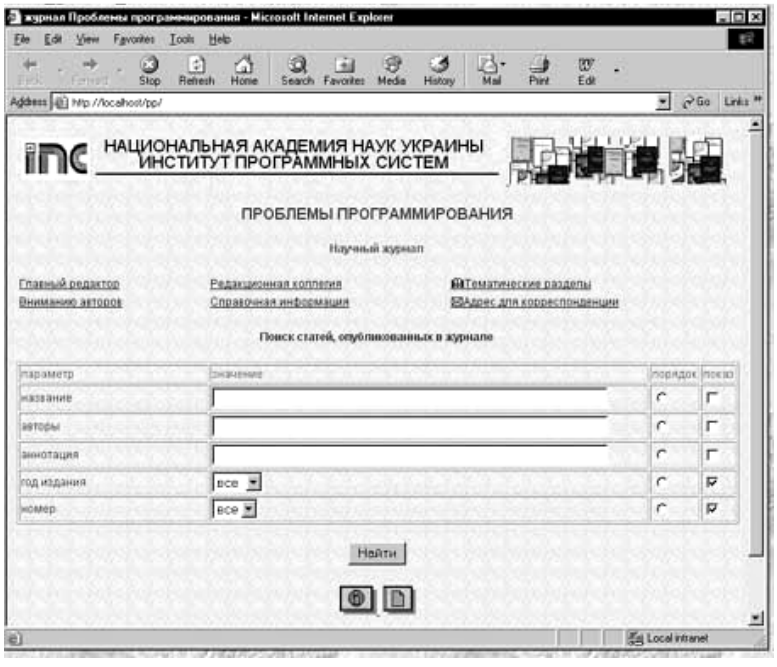

Рис. 2. Главная страница сайта электронной версии журнала “Проблемы программирования”

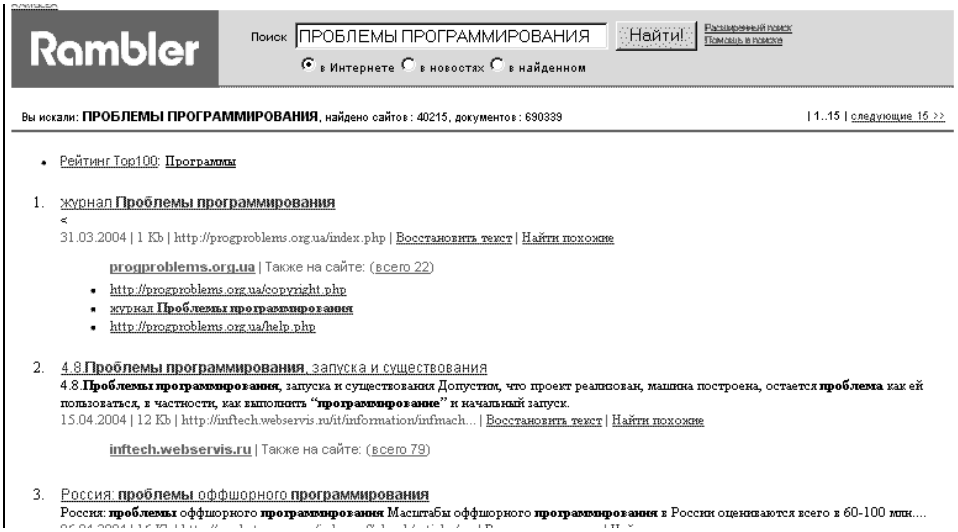

Рис. 3. Результаты поиска в **Rambler** по ключевым словам "проблемы программирования"

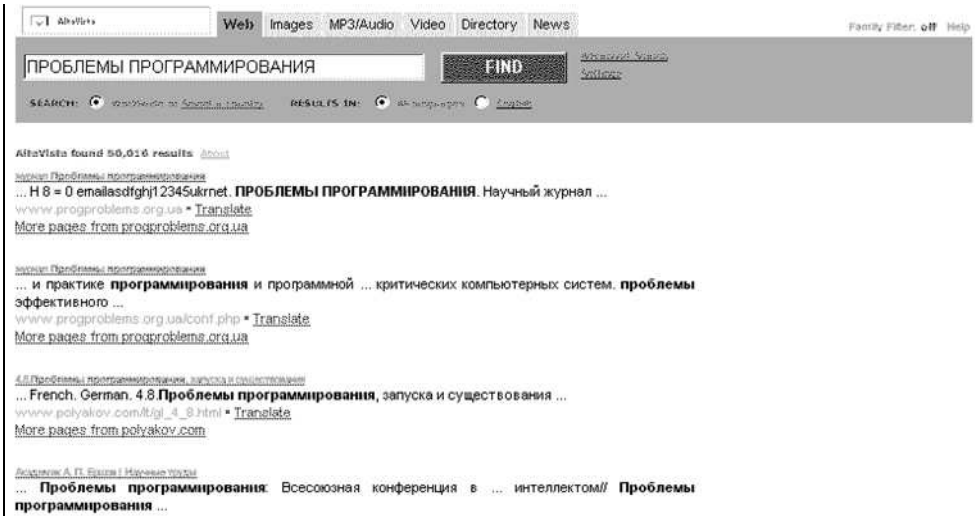

Рис. 4. Результаты поиска в **AltaVista** по словосочетанию "проблемы программирования"

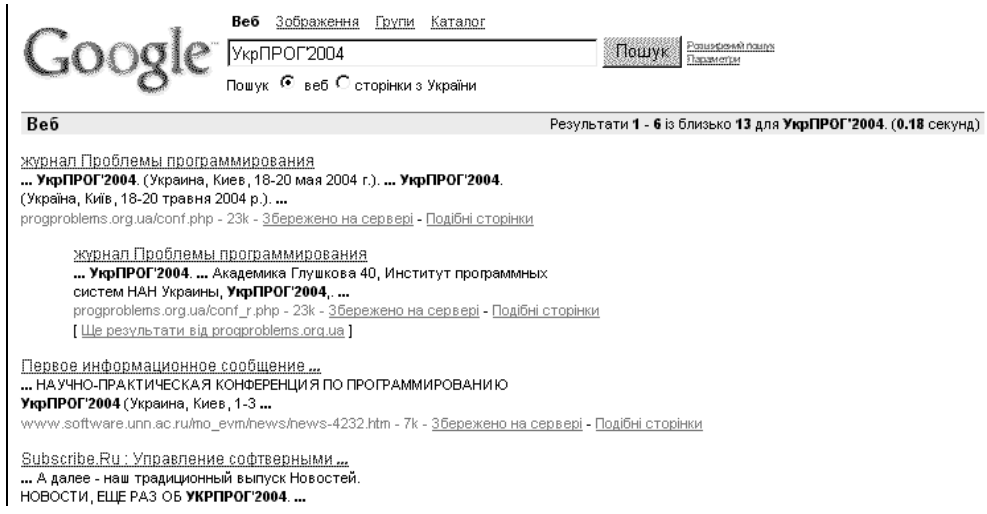

Рис. 5. Результаты поиска в **Google** по ключевым словам "УкрПРОГ`2004"

Результаты поиска информации в Интернете

О том, что предложенные принципы разработки сайта научного журнала оказались эффективными, свидетельствуют результаты выполнения поисковых запросов к ряду популярных поисковых систем.

Сайт занял первое место при поиске по ключевым словам "проблемы" и "программирование" в Rambler (www.rambler.ru) — из более чем 40 000 сайтов, содержащих почти 700 000 документов с такими ключевыми словами, и в AltaVista (www.altavista.com) — из более чем 50 000 результатов (рис. 3–5). Интересен тот факт, что сайт не был нами зарегистрирован ни в одной поисковой системе.

Выводы

Для того чтобы пользователи с помощью поисковых систем могли обнаружить структурированные не индексируемые "глубинные" информационные ресурсы Интернета, соответствующие их информационным потребностям, эти информационные ресурсы должны быть представлены в соответствии с определенными архитектурными принципами и сопровождаться корректными метаописаниями.

Апробация такого подхода к представлению информационных ресурсов при разработке электронной версии журнала "Проблемы программирования" показала их эффективность.

1. *Nec* ResearchIndex. — http://citeseer.nj.nec.com/.
2. *DBLP*. — http://dblp.uni-trier.de/.
3. *BibTex*. — http://www.ecst.csuchico.edu/-~jacobsd/bib/formats/bibtex.html.
4. *SGML*. ISO 8879. — http://www.iso.ch/cate/d16387.html.
5. *Extensible* Markup Language (XML) 1.0, W3C Recommendation 10.02.1998 — http://www.w3.org/TR/1998/REC-xml-19980210.
6. *Ermolayev V., Tolok V.* Academic E-editions in the Information Space of Ukraine. Novyj Kolegium. // Scientific and Information J. — 2002. — ¹ 3. — P. 38–45.
7. *RDF/XML* Syntax Specification. — http://www.w3.org/TR/rdf-syntax-grammar/.
8. *Dublin* Core Metadata Elements. — http://dublincore.org/.
9. *Жигалов В.* Как нам обустроить поиск в сети? // Открытые системы. — 2000. — ¹ 12. — С. 53–61.
10. *Рогушина Ю.В.* Разработка средств интеллектуализации поиска информации в Интернете // Пробл. программирования. — 2002. — № 1–2. — С. 378–385.
11. *PHP*. — http://www.php.net.
12. *MySQL*. — http://mysql.com/.

Получено ***10.08.04***

Об авторах

Рогушина Юлия Витальевна
канд. физ.-мат. наук, ст. науч. сотр.
Тел. (044) 268 4698

Гришанова Ирина Юрьевна
аспирант
Тел. (044) 234 4757

Место работы авторов:
Институт программных систем НАН Украины, Киев, просп. Акад. Глушкова, 40