# Создание русскоязычной библиотеки научных статей на факультете ВМиК МГУ

Д.Д. Козлов
Факультет Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова
ddk@cs.msu.su

## Введение

Повсеместное распространение сети Интернет существенным образом повлияло на доступность результатов научных исследований: технических отчетов, журнальных статей, материалов конференций. Так, многие отечественные и зарубежные конференции публикуют в сети Интернет сборники докладов, в зарубежных учебных заведениях принято размещать на домашних страницах авторов полные тексты публикаций. В результате сеть Интернет становится наиболее социально значимым источником научной литературы: она доступна, охватывает исследования из разных стран, предоставляет современные средства поиска.

Доступность зарубежных научных статей существенно выше, чем отечественных. Это связано с отсутствием у отечественных исследователей сложившейся Интернет-ориентированной культуры: в США, например, многие публикации можно найти в электронном виде на домашних страницах авторов, в России же редкие исследователи ведут такие домашние страницы. В результате отечественные исследователи предпочитают использовать и развивать более доступные зарубежные исследования, даже при наличии русскоязычных работ в той же области. Это, в свою очередь, приводит к потере наработок отечественных научных школ, снижению цитируемости отечественных работ.

Потребность в сохранении, накоплении и распространении результатов научных исследований востребована во многих научных школах. В данной работе рассмотрен зарубежный опыт накопления результатов научной деятельности и описано применение этого опыта на факультете ВМиК.

## Существующие подходы к накоплению результатов научных исследований

Общий объем информации в Интернет растет столь быстро, что традиционный поиск по ключевым словам не позволяет исследователям эффективно искать научную литературу. Для поиска научных статей в сети Интернет создаются специализированные электронные библиотеки научных статей (далее ЭБ), основанные на более эффективных методах, характерных именно для научной литературы [1], в частности, использующих библиографические ссылки. Рассмотрим основные виды таких библиотек за рубежом на примере области Computer Science.

Наиболее широко известна в научных кругах база данных Science Citation Index [2], содержащая библиографические описания научных статей и граф взаимного цитирования статей. SCI ориентирована на предоставление библиографической информации и индекса цитирования, и не содержат сами тексты статей, а доступ к ней предоставляется на коммерческой основе.

Другое направление развития – ЭБ профессиональных ассоциаций, например, ACM Digital Library [4] и IEEE Computer Society Digital Library [5]. Под эгидой этих профессиональных ассоциаций проходит большинство зарубежных конференций по

Computer Science, а полные тексты докладов распространяются ими на коммерческой основе.

Третье направление представлено независимыми электронными архивами, например, CORR [6], NCSTRL [7], которые пополняются самими авторами, учебными организациями и т.п. В таких ЭБ бесплатно предоставляются полные тексты статей, а граф цитирования обычно не строится.

Четвертым, стремительно набирающим популярность, направлением являются ЭБ, построенные как вторичные Интернет-ресурсы. Научные статьи в эти ЭБ помещаются в результате поиска в сети Интернет. Основоположником данного направления можно считать проект CiteSeer [8]. Данная библиотека содержит полные тексты статей, свободно доступных в сети Интернет, и поддерживает граф взаимного цитирования. Доступ к библиотеке осуществляется на некоммерческой основе. Близкие идеи использует Google Scolar.

Важной тенденцией развития ЭБ является декоммерциализация, направленная на повышение доступности статей. Исследования [12] показали, что статьи, бесплатно доступные в сети Интернет, чаще цитируются. По объему бесплатные ЭБ близки к библиотекам ACM и IEEE [13].

Согласно [9], SCI охватывает менее 15% отечественных публикаций. Создание Российского индекса научного тестирования (РИНЦ) планируется в 2007 году. В рамках научной электронной библиотеки (НЭБ) eLibrary.ru учебным и научным организациям предоставляется доступ к публикациям ведущих зарубежных издательств. В рамках РИНЦ разрабатываются методы полуавтоматического цитатного индексирования. В рамках РГБ им. Ленина была создана открытая электронная библиотека OREL и электронная библиотека диссертаций. Также в России очень большое количество некоммерческих электронных библиотек, поддерживаемых в основном учебными организациями. Эти библиотеки содержат очень маленькое количество статей и обеспечивают лишь средства поиска по ключевым словам.

Таким образом, развитие электронных библиотек в России во многом схоже с развитием зарубежных электронных библиотек, но существенно отстает от зарубежного опыта. В настоящее время в России накопилось уже достаточно большое количество научных статей, доступных через Интернет, откуда можно сделать предположение о перспективности построения ЭБ, построенной в виде вторичного Интернет-ресурса.

**Создание электронной библиотеки на факультете ВМиК**

Целью проекта по созданию ЭБ на факультете ВМиК является накопление и обеспечение доступности результатов научных исследований, проводимых на факультете и за его пределами. В свою очередь, это предусматривает решение следующих задач:

- Формирование Интернет-ориентированной культуры у исследователей: чтобы, помещая ссылки на публикации в научный отчет, они помещали публикации в Интернет, делая их доступными.
- Создание достаточно большого русскоязычного корпуса статей по Computer Science и его популяризация.
- Соединение корпуса русскоязычных статей и корпуса англоязычных статей посредством цитатных ссылок.

В основу ЭБ на факультете ВМиК был положен проект CiteSeer. CiteSeer представляет собой набор методов и технологий для построения электронных библиотек [8], в частности им решаются следующие основные задачи:

1. Пополнение базы путем поиска научных статей в сети Интернет.
2. Полнотекстовое индексирование статей, представленных в форматах PS и PDF.
3. Автоматическое извлечение метаданных и библиографических ссылок из статей.
4. Автоматическое цитатное индексирование, позволяющее без участия человека стоить граф цитирования, аналогичный SCI.
5. Анализ графа цитирования, выявление наиболее значимых статей.

При построении ЭБ на ВМиК некоторые из решений были взяты из CiteSeer в готовом виде, другие – адаптированы под особенности русскоязычной части сети Интернет. Далее рассмотрено решение одной из самых интересных задач, возникших при адаптации CiteSeer – задачи извлечения метаданных и библиографических ссылок из статей.

**Извлечения метаданных и цитатных ссылок**

При помещении статьи ЭБ необходимо извлечь из нее метаданные и библиографические ссылки. Этот процесс производится в два этапа: преобразование статьи в специально размеченный текст и анализ этого текста с целью извлечения метаданных и библиографических ссылок.

Основной адаптацией первого этапа к русскоязычным статьям явился переход от формата PS к распространенному в России PDF. При этом оригинальный модуль преобразования из CiteSeer был заменен на новый, основанный на проекте xpdf [10]. В xpdf были внесены изменения для создания разметки, соответствующей оригинальному модулю CiteSeer, описывающей окончания строк, параграфов и изменения шрифтов.

Размеченный текст статьи используется для извлечения метаданных (заголовка, авторов, организаций, года издания, аннотации) и библиографических ссылок. Оригинальный метод извлечения метаданных CiteSeer использует регулярные выражения и основан на устоявшихся правилах оформления, характерных для англоязычных статей. Проблема с русскоязычными статьями заключается в том, что по сравнению с англоязычными статьями они очень плохо структурированы, допускают массу вариантов как структуры, так и оформления. В результате адаптированный для русского языка метод регулярных выражений дает существенно менее качественные результаты на русскоязычных статьях по сравнению с англоязычными. При тестировании на материалах восьми отечественных конференций заголовок статьи извлекается правильно в 80% случаев, а авторы статьи – в 40%.

В качестве альтернативы методу регулярных выражений был использован метод машинного обучения, рассматривающий процесс извлечения метаданных как задачу классификации [12]. Экспериментальные исследования этого метода на том же наборе данных показали результат около 90% правильных извлечений для заголовка и авторов и около 80% – для списка литературы. При этом для классификации требуется дополнительное обучение на размеченной выборке, в то время как для метода регулярных выражений этого не требуется.

**Заключение**

Создание ЭБ на факультете ВМиК направлено на накопление, повышение доступности и значимости результатов исследований отечественных научных школ. В рамках проекта

была успешно адаптирована и применена технология CiteSeer. В настоящее время ведутся исследования, направленные на адаптацию автоматического поиска научных статей в русскоязычной части сети Интернет.

Область применения данной технологии не ограничена Computer Science, на ее могут работать библиотеки и в других областях знаний.

**Литература**


1. Bates M., The design of browsing and berrypicking techniques for the online search interface. Online Review 13, 5, 1989.
2. Science Citation Index. http://scientific.thomson.com/products/sci/
3. Petricek V. Et al. A Comparison of On-line Computer Science Citation Databases. ECDL 2005.
4. ACM Digital Library. http://portal.acm.org/dl.cfm
5. IEEE Computer Society Digital Library. http://computer.org
6. CORR. http://arxiv.org/corr/home
7. NCSTRL. http://ncstrl.org
8. Lawrence S., Bollaker K., Giles L., Indexing and Retrieval of Scientific Literature, CIKM, 1999.
9. Разработка РИНЦ. Проект. http://elibrary.ru/projects/citation/proposal.doc
10. Проект XPDF. http://www.foolabs.com/xpdf/
11. Giles L. at al. Automatic Document Metadata Extraction using Support Vector Machines, JCDL, 2003.
12. Lawrence S., Giles L., Bollaker K., Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Vol 32, N 6, 1999.