

Создание электронной библиотеки русскоязычных научных статей

Васильев А.
ВМиК МГУ
vasil@lvk.cs.msu.su

Самусев С.
ВМиК МГУ
sam@lvk.cs.msu.su

Шамина О.
ВМиК МГУ
sincere@lvk.cs.msu.su

Козлов Д.
ВМиК МГУ
ddk@cs.msu.su

Аннотация

В работе изложены результаты проекта по созданию электронной библиотеки русскоязычных и англоязычных научных статей по тематике Computer Science. Основной особенностью библиотеки является автоматическое построение цитатного индекса, соединяющего русскоязычные и англоязычные статьи. Также в работе рассмотрены вопросы автоматического поиска научных статей в сети Интернет для пополнения библиотеки.

1. Введение

Постоянное увеличение числа научных публикаций приводит к тому, что ученым сложно отслеживать новые исследования, чтобы постоянно быть в контексте развития того или иного научного направления. Традиционные библиотеки также не могут справиться с увеличением числа публикаций и не предоставляют доступа к современным исследованиям. Распространение сети Интернет существенным образом улучшило доступность результатов научных исследований: технических отчетов, журнальных статей, материалов конференций. Сеть Интернет стала наиболее социально значимым источником научной литературы: она доступна, охватывает исследования из разных стран, предоставляет базовые средства поиска. Доступность в сети Интернет отечественных научных статей существенно ниже, чем зарубежных. 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.

Одним из наиболее эффективных подходов к организации поиска научной литературы является создание электронных библиотек научных публикаций с возможностью цитатного

индексирования работ. Цитатный индекс является мощным инструментом поиска [4,14], позволяющим

- отслеживать взаимосвязь научных публикаций;
- выяснять, какие есть публикации, развивающие ту или иную научную работу;
- выяснять, какие есть ранее опубликованные работы по заданной теме;
- выяснять, в каком контексте цитируется работа, например, это может быть критика работы, использование работы, развитие работы и т.п.;
- находить малоизвестные работы, опубликованные не в основных тематических журналах, а в материалах конференций, технических отчетах;
- существенно сократить время на подбор литературы при составлении обзоров существующих наработок.

Цитатный индекс также является общепризнанным средством анализа научной деятельности, позволяющий оценивать вклад отдельных работ, отдельных ученых, научных организаций, грантодателей [19, 10].

Построение и поддержание электронной библиотеки с цитатным индексированием вручную очень трудоемко: приходится решать задачи наполнения библиотеки новыми статьями, извлечения из статей библиографических ссылок и построения цитатного индекса. В работе [9] был предложен новый подход, позволяющий автоматизировать как процесс наполнения библиотеки, так и построения цитатного индекса. В рамках данного подхода наполнение библиотеки производится с помощью поиска публикаций, которые ранее уже были размещены в Интернет, например, статей с домашних страниц авторов или с сайтов конференций. Для построения цитатного индекса используется технология Autonomous Citation Indexing, в которой из текста англоязычной статьи, представленной в формате Postscript (PS), в полностью автоматическом режиме извлекаются метаданные (авторы, заголовки и т.д.) и библиографические ссылки и строится цитатный индекс. В результате существенно снижается трудоемкость наполнения библиотеки. На основе этого подхода построена широко известная

англоязычная библиотека CiteSeer.IST [5] по тематике Computer Science.

В рамках данной работы на основе наработок проекта CiteSeer.IST была создана система, позволяющая хранить и русскоязычные, и англоязычные статьи, автоматически строить по ним общий цитатный индекс. На основе этой системы был построен прототип электронной библиотеки русскоязычных и англоязычных научных статей по тематике Computer Science. Наполнение библиотеки основано на материалах прошедших отечественных конференций.

Созданная библиотека предоставляет следующие основные возможности:

- автоматическое извлечение метаданных и библиографических ссылок из помещаемых в библиотеку текстов научных статей;
- автоматическое построение единого цитатного индекса по русскоязычным и англоязычным статьям;
- поиск в библиотеке не только с помощью ключевых слов, но и с помощью навигации по библиографическим ссылкам, в том числе между русскоязычными и англоязычными статьями;
- поиск близких по содержанию работ на основе текстового содержания и цитатного индекса;
- возможность полуавтоматического помещения статей путем указания адреса страницы, на которой размещены ссылки на публикации;
- возможность ручной корректировки неправильно извлеченных метаданных.

Также в работе рассмотрены возможные подходы к автоматическому пополнению электронной библиотеки путем поиска научных статей в русскоязычном сегменте сети Интернет (Рунет).

2. Существующие подходы к накоплению и обеспечению доступности научных публикаций

В данном разделе на основе работы [1] кратко описаны основные зарубежные и отечественные подходы к накоплению научных публикаций на примере области Computer Science, рассмотрены основные тенденции эволюции этих подходов.

Первое направление представлено независимыми электронными архивами, например, CORR [6], которые пополняются самими авторами, учебными организациями и т.п. В таких электронных библиотеках (ЭБ) бесплатно предоставляются полные тексты статей и предоставляется возможность поиска по ключевым словам. Существует коммерческий вариант этого направления – ЭБ профессиональных ассоциаций, например, ACM Digital Library [3] и IEEE Computer Society Digital Library [8]. Под эгидой этих профессиональных ассоциаций проходит большинство зарубежных конференций по

Computer Science, а полные тексты статей распространяются на коммерческой основе.

Вторым направлением являются библиографические базы данных, например, широко известная Science Citation Index [19], содержащая библиографические описания научных статей и цитатный индекс. SCI ориентирована на предоставление библиографической информации и цитатного индекса, и не содержит сами тексты статей.

Третьим подходом является построение библиотек путем поиска научных статей в сети Интернет и цитатного индексирования. На основе цитатного индексирования построены ЭБ CiteSeer.IST и REXA [18]. Они бесплатно предоставляют полные тексты статей, свободно доступных в сети Интернет, и обеспечивают поиск с помощью цитатного индекса. Похожий подход использует Google Scholar.

Важной тенденцией развития зарубежных ЭБ является декоммерциализация, направленная на повышение доступности статей. Исследование [14] показало, что статьи бесплатно доступные в сети Интернет чаще цитируются. По объему бесплатные ЭБ близки к библиотекам ACM и IEEE [17].

Развитие электронных библиотек в России во многом повторяет зарубежный опыт с заметным отставанием (создан аналог CORR – открытый электронный архив OREL при РГБ, планируется создание аналога SCI – Российского индекса научного цитирования). В настоящее время в России накопилось уже достаточно большое количество научных статей, доступных в сети Интернет, откуда можно сделать предположение о перспективности построения ЭБ, построенной по аналогии с CiteSeer.

3. Специфика русскоязычных статей

Специфика русскоязычных научных статей по сравнению с англоязычными состоит в следующем:

- в отличие от англоязычных статей, в которых существуют общепринятые нормы структурирования и оформления статьи, для русскоязычных статей нет таких норм, и авторы структурируют и оформляют статьи руководствуясь исключительно своими пожеланиями (требования разных конференций и журналов также очень сильно различаются);
- в отличие от английского языка, где разделы статьи имеют традиционные названия (Abstract, Introduction и т.п.) в русском языке используется большое разнообразие слов для обозначения одних и тех же разделов (например, только библиография может называться «литература», «ссылки», «источники», «список литературы» и т.п.);
- в России еще не сложилось Интернет-ориентированной культуры когда каждый

ученый ведет свою домашнюю страничку с публикациями;

- в России научные статьи часто размещаются в Интернет в различных форматах, например, Word, PDF, PS, HTML, тогда как зарубежные статьи традиционно размещаются в PS и PDF.

4. Система для хранения и поиска научных статей

В основе построенной электронной библиотеки лежит информационная система хранения и поиска научных статей, созданная на основе CiteSeer. Работа системы логически может быть разделена на два основных цикла: добавление статей в систему и поиск статей в системе.

Добавление статей в систему происходит следующим образом:

- статьи в систему помещаются двумя возможными путями: через web-интерфейс полуавтоматического помещения статей, через систему автоматического поиска статей;
- из помещенных в систему статей автоматически извлекаются метаданные и библиографические ссылки;
- извлеченная из статьи метаинформация заносится в реляционную базу данных;
- для каждой помещаемой в систему статьи и библиографической ссылки производится определение, соответствует ли эта ссылка уже имеющейся статье или нет, если не соответствует, то в базу данных заносится новая запись, содержащая метаинформацию, извлеченную из библиографической ссылки;
- производится обновление цитатного индекса;
- производится обновление инвертированного индекса поискового модуля;
- производится обновление оценок значимости документов.

Поиск статей в системе осуществляется следующим образом:

- пользователь посредством web-интерфейса формулирует запрос к системе в виде набора ключевых слов. Целью поиска могут быть документы или цитаты;
- поисковый модуль возвращает отсортированный набор документов (или цитат);
- по ссылке из результатов поиска выдается страница со сводной информацией о статье, включающей в себя заголовок статьи, список авторов, год издания, частоту цитируемости, источник, откуда была получена статья, ссылку на локальную копию статьи, начало аннотации, примеры контекста, в которых цитируется данная статья.
- пользователь может перемещаться по библиографическим ссылкам между всеми статьями, например, посмотреть статьи,

цитирующие заданную, посмотреть контекст, в котором цитируется статья и т.д.

4.1 Добавление статей в систему

Добавление статей в систему осуществляется в полуавтоматическом режиме через web-интерфейс, в котором пользователь задает URL страницы с публикациями (например, страницы с материалами конференции или персональной домашней страницы). Система осуществляет извлечение всех ссылок на PDF-документы, скачивает сами документы и пытается преобразовать документы в промежуточное текстовое представление с помощью модифицированного варианта программы pdftotext из проекта xpdf [22]. Модифицированный вариант позволяет сохранить в текстовом файле дополнительную информацию об окончаниях строк, абзацев (точнее отступах от края), изменении размера шрифта.

Из построенного промежуточного представления статьи осуществляется извлечение метаинформации (авторы, заголовок, год издания и т.д.) и библиографических ссылок (библиографическая ссылка разбирается на авторов, заголовок, год издания и все остальное). Для этого в существующей реализации используется разработанный в рамках данного проекта алгоритм, основанный на применении механизма регулярных выражений.

Извлечение заголовка и авторов статьи, происходит следующим образом:

- 1) Из первой страницы текста извлекается текст предшествующий разделу «Введение» или «Аннотация». Если такого раздела найдено не было, берутся первые 2000 символов.
- 2) В первых пяти строках извлеченного текста ищется подстрока похожая на список авторов. Для этого используются разработанные шаблоны имен, учитывающие различные способы написания.
- 3) Далее возможно два варианта:
 - а) Список авторов был найден. В этом случае из строк расположенных непосредственно до или после найденного списка выбирается та, которой соответствует больший размер шрифта. Выбранная строка рассматривается в качестве возможного заголовка на шаге 4.
 - б) Список авторов найден не был. Тогда в первых пяти строках извлеченного на первом шаге текста ищется строка, которой соответствует максимальный размер шрифта. Выбранная строка рассматривается в качестве возможного заголовка на шаге 4.
- 4) В том случае если выполнено одно из следующих условий:
 - а) Выбранная строка определена как название журнала, конференции, института, издательства и т.д. Для этого используется составленный вручную список ключевых

слов: «материалы конференции», «труды конференции», «научный журнал», «издательство», «сборник трудов» и т.д.

- б) Длина выбранной строки превышает установленную максимальную длину заголовка (200 символов).
- в) Длина выбранной строки меньше установленной минимальной длины заголовка (15 символов).

строка удаляется и поиск происходит снова (переход к пункту 3).

- 5) На данном этапе возможны следующие варианты:
 - а) Заголовок статьи не был найден, т.е. ни одна из строк не прошла проверку: переход к шагу 6.
 - б) Заголовок был найден, но список авторов найден не был: происходит попытка угадать имя автора в тексте, который следует за заголовком. Для этого используется список слов, которые часто встречаются рядом с именем автора в научных статьях («факультет», «институт», «лаборатория» и т.д.)
- б) В том случае если список авторов был найден, происходит его разбор и извлечение имени каждого автора.

Для извлечения аннотации в тексте ищется соответствующий раздел. Если такого раздела найдено не было, то берутся первые 1000 символов введения. Если заголовок «Введение» также отсутствует, то, в качестве аннотации выбираются первые 4 строки текста.

Извлечение списка использованной литературы происходит по следующему алгоритму:

- 1) Из текста статьи по соответствию ключевым словам («Список литературы», «Библиография», «Источники» и т.д.) извлекается раздел соответствующий списку литературы. Если такого раздела найдено не было, то процесс извлечения завершается.
- 2) Из списка литературы извлекаются отдельные библиографические ссылки. Для этого используются принятые правила составления списка: квадратные скобки, нумерация и т.д.
 - а) Далее в каждой ссылке ищется список авторов (предполагается, что имена авторов предшествуют названию). Поиск имен происходит посредством разработанных шаблонов, в которых используются инициалы, поэтому для облегчения извлечения, производится нормализация имен авторов до И.О. Фамилия или И. Фамилия, а авторы перечисляются через запятую. На этом же этапе проверяется, наличие списка авторов в ссылке. Так, часто можно встретить ссылки на «Большой академический словарь», «Большую советскую энциклопедию» и т.д., авторы, для которых не указываются. В

этом случае список авторов, помечается пустым.

- 3) Далее происходит поиск названия статьи. В качестве названия выбирается предложение идущее после списка авторов.
- 4) Год издания извлекается по шаблону.

Следует отметить, что для извлечения метаинформации из англоязычных статей применяются также методы, основанные на применении алгоритмов классификации (SVM) [11], скрытые Марковские модели [20] и условные случайные поля (Conditional Random Fields) [13]. В рамках настоящей работы были проведены адаптация и экспериментальное исследование трех методов: основанного на регулярных выражениях, основанного на скрытых Марковских моделях и основанного на классификации с помощью SVM. Подробное изложение методов и экспериментального исследования приведено в работе [2]. Исследование показало, что при извлечении метаданных лучшую точность показал метод, основанный на классификации, а при извлечении библиографических ссылок – метод, основанный на скрытых Марковских моделях. В рамках дальнейшей работы планируется расширить исследование [2] методом, основанным на условных случайных полях.

Для сопоставления извлеченных из статьи метаинформации и библиографических ссылок с уже существующими в базе данных записями был использован алгоритм, предложенный авторами CiteSeer, который основан на нормализации библиографической ссылки и пословного сравнения заголовков и списков авторов. Этот алгоритм работает достаточно хорошо для практического использования. В тоже время существуют более точные алгоритмы, основанные на методах машинного обучения. Их исследование планируется провести в будущем. В системе предусмотрена возможность ручного изменения неверно извлеченной из статьи метаинформации.

4.2 Поиск статей в системе

Для поиска научных статей внутри системы используются инвертированный индекс и цитатный индекс. Созданная реализация предоставляет следующие возможности поиска:

- Поиск статей по запросу в виде набора ключевых слов. Список сортируется по одному из критериев: дате поступления, оценкам значимости по Клейнбергу [12], популярности. По ссылкам из результатов поиска выдается страница со сводной информацией о статье (см. рисунок 1). Сводная информация включает в себя заголовок статьи, список авторов, год издания, частоту цитируемости, источник, откуда была получена статья, ссылку на локальную копию статьи, начало аннотации, примеры контекста, в которых цитируется

**Инициативный Проект Российского Семинара По Оценке
Методов Информационного Поиска (ромип)
(2003)** ([Редактировать](#)) ([Цитируется 4 раза](#))

П.И. Браславский М.В. Губин Б.В. Добров В.Ю. Добрынин И.Е.
Кураленок И.С. Некрестьянов Е.Ю. Павлова И.В. Сегалович

[Контекст](#)

Просмотреть или скачать:
dialog21.ru/Archi_slavskij_Gubin.pdf
PDF [Просмотр](#)

Источник: dialog21.ru/material...?id=56248
([Еще](#))
([Вести домашние страницы автора](#))

Аннотация: В последние годы был достигнут значительный прогресс как в теории информационного поиска, так и в создании промышленных информационно-поисковых систем. Непрерывная эволюция информационного пространства и применение методов поиска в новых контекстах определяет актуальность дальнейших исследований в области теории информационного поиска. Существует большое число задач, относимых к области информационного поиска: 1. 2. 3. 4. 5. 6. · традиционный поиск по коллекциям с фиксированным набором жанров... ([Редактировать](#))

Документ относится к группам:

1. [Диалог 2003](#)
([Изменить](#))

Контекст цитирования данного документа: [Еще](#)

...проведения семинара в 2004 году. 2. **Основы методологии РОМИП Принципы организации РОМИП уже неоднократно описывались нами ранее [1,4,5,6], поэтому мы лишь вкратце остановимся на них в рамках этой статьи.** Семинар РОМИП имеет циклическую природу. Для каждого годового цикла из...

... **проведение семинаров РОМИП, каждый из которых посвящен оценке эффективности решения одной или нескольких задач текстового поиска [1].** Список рассматриваемых задач определяется на основе обсуждения с участниками семинара и возможностей реализации этих проектов...

Документы, цитирующие данный: [Еще](#)

Результаты Первого Российского Семинара По Оценке Методов... - Шабанов (2004) ([Редактировать](#))

Рсо На Ромип 2003: Отчет Об Участии В Семинаре По... - В.В. Плешко А.Е... ([Редактировать](#))

Ромип 2003: Опыт Организации - Павлова (2003) ([Редактировать](#))

Предположительно похожие документы:

9.0%: Ромип2004: Отчет Организаторов - Некрестьянов (2004) ([Редактировать](#))

Документы, ссылающиеся на те же документы, что и данный (связанные документы): [Еще](#) [Все](#)

0.7: Российский Семинар По Оценке Методов Информационного Поиска... - Шабанов (2005) ([Редактировать](#))

0.5: Информационный Поиск В Коллекции Разнородных Документов - Губин (2005) ([Редактировать](#))

0.2: Пользовательская И Системная Сложность Данных - Уткин (2006) ([Редактировать](#))

Рисунок 1. Сводная информация о статье.

Браславский П.И., Губин М.В., Добров Б.В., Добрынин В.Ю., Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., Сегалович И.В., *Инициативный проект Российского семинара по оценке методов информационного поиска (РОМИП)*

[Просмотреть или скачать документ](#) [Резюме](#) [Связанные Документы](#)

Данный документ цитируется в следующих контекстах:

[Рсо На Ромип 2003: Отчет Об Участии В Семинаре По... - В.В. Плешко А.Е...](#) ([Редактировать](#))

...у российских исследователей до сих пор не имелось в распоряжении масштабных русскоязычных текстовых корпусов, на которых можно было бы получить достоверные оценки качества создаваемых систем. **Этот пробел призван устранить Российский семинар по Оценке Методов Информационного Поиска (РОМИП) [2,3].** Участникам семинара предлагалось принять участие в экспериментах (дорожках) по решению двух задач: поиск web страниц и тематическая классификация web сайтов. **Методика** оценки результатов, использованная организаторами семинара, является общепринятой для задач информационного поиска и описана

Браславский П.И., Губин М.В., Добров Б.В., Добрынин В.Ю., Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., Сегалович И.В. *Инициативный проект Российского семинара по Оценке Методов Информационного Поиска (РОМИП)*. Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003.

[Результаты Первого Российского Семинара По Оценке Методов... - Шабанов \(2004\)](#) **Самоцитирование (Добров Некрестьянов Сегалович)** ([Редактировать](#))

...недоверия участников, которые могут являться прямыми конкурентами вне РОМИП. **Поэтому,** задачей минимум в 2003 году было просто провести первый в РФ семинар, и эта задача была решена объединением доброй воли участников. **Концепция проведения семинара РОМИП была представлена в трудах Диалог 2003 [2], поэтому здесь мы лишь вкратце остановимся на общей методологии проведения и больше внимания уделим опыту и результатам полученным за 2003 год.** 1. **Общая** методология организации семинара РОМИП Семинар РОМИП имеет циклическую природу. **В** рамках годового цикла из множества реализуемых проектов

Браславский П.И., Губин М.В., Добров Б.В., Добрынин В.Ю., Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., Сегалович И.В., *Инициативный проект Российского семинара по оценке методов информационного поиска (РОМИП)*

Рисунок 2. Контексты, в которых цитируется статья.

данная статья, ссылки на статьи, цитирующие данную, ссылки на процитированные статьи, похожие статьи (как текстуально, так и по цитатам), статьи с того же сайта, что и данная статья.

- Поиск цитат статьи. Статья при этом задается набором ключевых слов. Если набору соответствуют несколько статей, то выводится список статей со ссылками на контексты, в которых статья цитируется. Пример контекста приведен на рисунке 2.
- Поиск статей на основе цитатного индекса: статьи, ссылающиеся на данную, статьи, на которые ссылается данная, статьи, имеющие общие библиографические ссылки с данной, статьи, встречающиеся вместе с данной в списке литературы
- Поиск похожих статей на основе текстуальной схожести, вычисляемой на основе TFIDF, и библиографической близости CCIDF (common citation \times inverse document frequency) [9].

Для построения инвертированного индекса использовалась реализация, созданная С. Лоренсом в рамках CiteSeer, доработанная для поддержки русского языка, и стеммер из проекта Snowball [21]. В дальнейшем планируется замена данной реализации на сопровождаемый (сообществом open source или какой-либо компанией) вариант, например, Yandex Server Free Edition или Apache Lucene.

5. Автоматический поиск научных статей в Рунет

Одним из важных вопросов эксплуатации электронной библиотеки научных статей является пополнение библиотеки новыми статьями. Без автоматического поиска новых статей развитие электронной библиотеки либо полностью зависит от пользователей, помещающих туда свои статьи, либо требует специально выделенного персонала, осуществляющего поиск и помещение новых статей в библиотеку. Содержание такого персонала, как правило, является слишком дорогостоящим, а практика развития библиотеки авторами статей зарекомендовала себя как малоэффективная (примером тому является библиотека OREL при РГБ, в которой авторы могут самостоятельно разместить текст диссертации, за более чем 5 лет существования она накопила около 2 тыс. диссертаций).

Альтернативным способом пополнения библиотек научных статей является поиск научных статей в сети Интернет, где их размещают сами авторы, организаторы конференций, учебные и научные заведения, а также другие электронные архивы и библиотеки. Далее рассмотрены существующие подходы к автоматическому поиску научных статей в сети Интернет.

5.1 Существующие подходы к автоматическому поиску статей

В CiteSeer поиск научных статей в сети Интернет основан на традиции американских ученых размещать свои публикации в формате PS на своих домашних страницах. Поиск осуществляется в два этапа. На первом этапе используются традиционные системы поиска по ключевым словам¹ (СПКС) для того, чтобы найти страницы с большой вероятностью содержащие ссылки на научные публикации. Для этого запрос пользователя к CiteSeer дополняется ключевыми словами, характеризующими тип документа, например, “publications”, “papers”, “postscript” и посылается СПКС. На втором этапе на страницах, которые нашла СПКС, осуществляется поиск всех ссылок на документы в формате PS (по расширению .ps, .pg.gz, .ps.Z) и загружаются найденные документы. В качестве СПКС используются AltaVista и метапоисковая машина Inquirus [15].

В работе [7] предложен подход к поиску новых статей по заданной предметной области на основе поиска домашних страниц исследователей. Работа метода состоит из трех шагов:

- поиск в библиографических базах данных, например, DBLP, имен ученых, работающих в данной предметной области (предметная область задается названиями журналов и конференций).
- поиск домашних страниц ученых с помощью системы HPSearch.
- поиск научных статей в окрестности найденных домашних страниц с помощью системы Mops.

На первом шаге список авторов строится путем выбора наиболее активных авторов в рамках заданной тематики (из тематик, представленных в DBLP). На втором шаге HPSearch сначала осуществляет поиск кандидатов на домашнюю страницу с использованием СПКС, выбирая несколько первых результатов, затем, осуществляя собственное ранжирование, после чего страницы с наибольшим весом загружаются, ранжируются еще раз и результат (несколько страниц-кандидатов) сохраняется в базе данных. В результате проведенного экспериментального исследования эффективности 500 возможных характеристик домашних страниц авторы выделили следующие:

- имеет смысл различать следующие части домашней страницы: заголовок, первый тэг header, другие тэги header, тексты ссылок и остальной текст;
- имя автора наиболее часто встречается в заголовке (title), первом тэге header или url;

¹ Термин «СПКС» используется как эквивалент англоязычному «search engines» (Google, Yandex, ...) ввиду того, что термин «поисковая система» в русскоязычных текстах используется в более широком смысле.

- символ “~” или строки типа “/people”, “/users” часто встречаются в URL домашних страниц;
- на домашних страницах обычно есть ссылка на публикации;
- домашние страницы обычно имеют маленький размер (3-9 KB).

HPSearch постоянно обновляет базу данных статей, проверяя доступность URL. Экспериментальное исследование HPSearch, проведенное авторами, показало, что 84% домашних страниц из указанных в DBLP были найдены HPSearch. На третьем шаге система Mops осуществляет поиск по домашней странице автора всех файлов pdf, dvi, ps, которые затем делятся на научные статьи и другие материалы на основе структуры URL.

В работе [23] предложен метод поиска научных статей с помощью тематического поискового робота. На первом шаге используется репозиторий метаданных статей (им может быть электронная библиотека или библиографическая база данных, например, DBLP) для получения набора пар <автор, издание> с учетом возможных вариантов написания издания. На втором шаге осуществляется поиск домашних страниц: пары <автор, издание> посылаются СПКС, а результат поиска фильтруется для того чтобы убрать неверные варианты и однофамильцев. Затем найденные страницы ранжируются, а страницы с наибольшим весом заносится в базу данных домашних страниц. При фильтрации и ранжировании используются следующие эвристики:

- Удаление из списка тех страниц, которые заведомо не являются домашними:
 - URL или заголовок страницы соответствует сайту издательства или электронной библиотеке.
 - URL указывает не на .htm/.html файл.
- Удаление однофамильцев
 - Удаление из списка страниц принадлежащих тому же домену, что и найденная ранее домашняя страница другого автора.
 - Удаление страницы, если домен, в котором она находится уже найден ранее.
- Определение веса каждой страницы:
 - Высокий, в том случае если заголовок содержит имя автора и хотя бы одно из следующих слов: homepage, website, research, publication, papers.
 - Средний, если заголовок содержит хотя бы одно из следующих слов: homepage, website, research, publication, papers.
 - Низкий, во всех остальных случаях.

На третьем шаге осуществляется поиск научных статей с помощью тематического поискового робота. Начальными страницами для поиска являются найденные на предыдущем шаге, кроме того, роботу также задается список разрешенных

для посещения доменов. Робот использует очередь приоритетами (высокий, средний, низкий). При загрузке очередной страницы робот разбирает все исходящие ссылки и назначает каждой приоритет в зависимости от текста ссылки и приоритета родительской страницы. Приоритет текста ссылки определяется классификатором на основе вхождения ключевых слов (например, volume publication conference и т.п.) в текст ссылки.

Авторы сообщают, что на тестовых наборах данных из конференций WebDB и JAIR методом было найдено около 80% статей.

В работе [16] решается задача поиска научной статьи по заданной библиографической ссылке, что может также иметь применение для пополнения электронной библиотеки. Работа метода состоит из трех шагов: на первом шаге библиографическая ссылка посылается СПКС и отбираются первые 10 ссылок. На втором шаге с помощью поискового робота ищется страница, которая наиболее вероятно содержит ссылку на искомую статью. На третьем шаге осуществляется поиск заданной библиографической ссылки в рамках найденной страницы. Для этого на странице выделяются начало и конец каждой библиографической ссылки, вычисляется расстояние между ссылкой на PDF или PS файл и заголовком искомой статьи и выбирается ближайший PDF или PS файл.

Рассмотренные подходы к поиску новых научных статей основаны на традиции американских ученых размещать свои публикации на домашних страницах в сети Интернет в форматах PS или PDF. Большинство существующих подходов сводится к поиску домашней страницы заданного автора, а потом поиска публикаций на ней. На англоязычных данных этот подход весьма эффективен (обнаружение порядка 90% домашних страниц и 80% публикаций), при этом авторы исследования [23] отмечают, что две трети авторов, чьи домашние страницы не были найдены, работают не в США, а те американские авторы, домашние страницы которых не были найдены, в основном относятся к категории студентов, которые были соавторами публикаций, а потом не продолжили научную карьеру.

Существующие методы поиска домашних страниц, использующие для формирования начального набора страниц СПКС, зависят от наличия знаменитых однофамильцев, страницы про которых занимают первые строчки в СПКС. В таких случаях тематический краулер может получать неверное начальное приближение и не находить правильно домашнюю страницу. В этом случае важно отсекал популярных однофамильцев на ранних стадиях еще до работы краулера. Более устойчивыми к этой проблеме должны быть методы, основанные на передаче СПКС развернутого запроса, например, текста

библиографической ссылки или заголовка статьи.

5.2 Особенности поиска научных статей в Рунет

Применение существующих англоязычных наработок с целью поиска в русскоязычном сегменте сети Интернет (Рунет) ограничено следующими особенностями Рунет:

- В Рунет отсутствуют такие качественные бесплатные источники библиографической информации как DBLP. При этом ЭБ сама является источником метаданных: имен авторов, учреждений, где работают авторы, адресов электронной почты авторов, названий изданий, конференций, библиографических ссылок на статьи, которые отсутствуют в библиотеке. Качество извлечения метаданных из уже найденных статей достаточно высокое, чтобы использовать метаданные для поиска вместо библиографических баз данных типа DBLP.
- Традиция американских ученых вести свои домашние странички с публикациями не так сильно распространена в России. В то время как существующие подходы используют поиск домашних страниц, в Рунет в качестве целей поиска могут также рассматриваться страницы конференций, страницы рабочих групп и организаций, электронные архивы и библиотеки, тематические профессиональные сайты.
- Часто домашняя страница ученого (или страница организации) предоставляет не тексты статей, а только библиографические ссылки. Эта информация также представляет интерес для наполнения ЭБ, так как является ценным источником библиографической информации, которая может быть использована для поиска текста статьи по библиографической ссылке (см. [23]).
- В Рунет многие публикации доступны не в формате PS, а в PDF, DOC, HTML. И если в США в формат PS используется в основном для научных публикаций, что активно используется в CiteSeer, то остальные форматы используются для публикаций совершенно разного назначения.
- В Рунет нет сервисов типа HPSearch для поиска домашних страниц исследователей.
- Для поиска статей по библиографической ссылке могут использоваться сервисы типа совсем недавно русифицированного Google Scholar. С покрытием СПКС не может соревноваться ни одна научная разработка.

5.3 Предлагаемый метод

В рамках задачи поиска новых статей можно выделить следующие частные случаи:

- поиск текста статьи по заданному библиографическому описанию;

- поиск домашней страницы автора/конференции/организации, содержащей список публикаций по заданному имени автора, названию конференции/организации.

В данной работе в качестве базы данных библиографической информации предлагается использовать, во-первых, саму ЭБ, так как много статей в ней представлены только библиографическими описаниями, во-вторых, домашние страницы ученых и организаций, содержащие не тексты публикаций, а библиографические ссылки. Для извлечения библиографических ссылок из HTML-страниц с описанием публикации в отличие от [23] предлагается использовать скрытые Марковские модели, успешно примененные в [2] к извлечению библиографических ссылок из PDF-документов.

Для поиска текста статьи по ее библиографическому описанию предлагается использовать подход из [23] в сочетании в разборе HTML-страниц с библиографией из [16], со следующими изменениями:

- на вход СПКС подается не библиографическое описание статьи, а список авторов, заглавие и ограничение на формат файла (PS, PDF). Это делается для того, чтобы не получить от СПКС статьи, ссылающиеся на искомую. Фильтрация по типу файла дает возможность найти сразу статью, а не HTML-страницы, ссылающиеся на нее;
- затем каждый из результатов по очереди проходит процедуру извлечения метаданных, и если извлеченная метаданная совпадает с требуемой, то процесс поиска считается законченным успешно;
- если среди первых 10 результатов СПКС статьи не найдено, то ослабляется ограничение на формат файла и осуществляется анализ HTML-страниц на наличие библиографической ссылки на требуемую статью;
- если найдена HTML-страница, содержащая библиографические ссылки, в том числе на искомую статью, то с помощью метода из [23] осуществляется поиск ближайшей ссылки на файл со статьей;
- если файл со статьей найден и извлеченная метаданная совпадает с требуемой, то процесс поиска считается законченным успешно;
- в противном случае процесс поиска завершается неуспешно.

Описанный процесс поиска может также иметь дополнительный положительный эффект: в случае, если найдена статья и страница с библиографическими ссылками, одна из которых ведет к искомой статье, то, вероятно, что это домашняя страница ученого/организации/конференции и с нее с большой вероятностью можно получить список

библиографических ссылок или новые статьи.

Для поиска домашней страницы ученого/организации/конференции, содержащей список публикаций, предлагается использовать метод [23], со следующими модификациями: в качестве запроса СПКС задается метаинформация из уже существующих статей: имена авторов, названия конференций, названия организаций, заголовки статей. Например, для поиска домашней страницы автора могут задаваться имя автора, названия известных его публикаций и дополнительные ключевые слова типа «публикации», «домашняя страница». Использование тематического краулера предлагается ограничить скачиванием только страниц, непосредственно достижимых по ссылке с найденной СПКС страницы и в пределах того же домена.

Предложенный метод существенно отличается от применяемого в CiteSeer: он не привязан к поиску только домашних страниц, а ориентирован также на поиск страниц организаций и конференций; в нем не используется запрос пользователя (например, название предметной области – «нейронные сети прямого распространения»), который может породить достаточно большой шум в результатах СПКС; в нем производится анализ результата поиска путем разбора документа и извлечения метаинформации, а не скачивание всех подряд документов в заданном формате; в нем в ходе поиска производится разбор статьи и извлечение метаинформации; в нем производится разбор страницы с библиографическими ссылками. В настоящее время ведется экспериментальное исследование предложенного метода.

6. Заключение

В рамках данной работы на основе наработок проекта CiteSeer была построена информационная система, позволяющая хранить и русскоязычные, и англоязычные статьи, автоматически строить по ним общий цитатный индекс. На основе этой системы был построен прототип электронной библиотеки русскоязычных и англоязычных научных статей по тематике Computer Science.

В дальнейшем предполагается провести ряд работ по развитию системы:

- экспериментальное исследование предложенного метода автоматического поиска научных статей в Рунет, обеспечение поиска не только новых статей, но и домашних страниц, адресов электронной почты и фотографий авторов, чтобы библиотеку можно было также рассматривать как каталог информации об отечественных ученых;

- исследование существующих методов сопоставления библиографических ссылок для более точного построения цитатного индекса;
- развитие методов извлечения метаинформации и библиографических ссылок: совершенствование учета разметки в методах машинного обучения, экспериментальное исследование применения условных случайных полей, поддержка различных форматов документов;
- интеграция ЭБ с CiteSeer.IST по протоколу OAI-PMH;
- создание средств тематического поиска:
 - предоставление возможности сформулировать запрос на одном языке и получить тематически близкие статьи на обоих языках (без перевода запроса);
 - кластеризация результатов поиска по различным критериям: тематике, авторам, году издания и т.п.;
 - автоматическая классификация научных статей для построения тематического каталога;
- внедрение средств персонализации и коллективных закладок, чтобы ученые могли строить свое представление библиотеки, размечать статьи тегами и обмениваться этой информацией по аналогии с сервисом del.icio.us;
- перевод системы на свободное ПО. Уход от трудносопровождаемого и не свободного кода компании NEC в CiteSeer.IST, чтобы код разрабатываемой системы мог иметь статус free software.

7. Литература

- [1] Козлов Д.Д. Создание русскоязычной библиотеки научных статей на факультете ВМиК МГУ. // Сборник трудов Всероссийской научной конференции "Научный сервис в сети Интернет" 2006
- [2] Васильев А., Козлов Д., Самусев С., Шамина О. Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей. // Сборник трудов конференции RCDL'2007, 2007.
- [3] ACM Digital Library [Electronic resource] <http://acm.org/dl/>.
- [4] Bates M., The design of browsing and berrypicking techniques for the online search interface. Online Review 13, 5, 1989.
- [5] CiteSeer.IST. [Electronic resource] <http://citeseer.ist.psu.edu>.
- [6] CORR. [Electronic resource] <http://arxiv.org/corr/home>.
- [7] Hoff G., Mundhenk, M. Finding scientific papers with homepage search and MOPS. In Proceedings of the Nineteenth Annual International Conference of Computer Documentation, Communicating in the New Millennium, pp. 201-207, 2001.
- [8] IEEE Computer Society Digital Library [Electronic resource] <http://computer.org>.

- [9] Giles C. L., Bollacker K. D., Lawrence S. CiteSeer: An Automatic Citation Indexing System. The Third ACM Conference on Digital Libraries, 1998.
- [10] Giles C.L., Council I. G. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. // Proceedings of the National Academy of Sciences, 2004
- [11] Giles L. et al. Automatic Document Metadata Extraction using Support Vector Machines, JCDL, 2003.
- [12] Kleinberg J. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, 1997.
- [13] Lafferty J., Pereira F., McCallum A.. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, 2001.
- [14] Lawrence S., Giles L., Bollacker K., Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Vol 32, N 6, 1999.
- [15] Lawrence S., Giles L. Inquirus, the NECI meta search engine // Proceedings of the seventh international conference on World Wide Web 7. 1998.
- [16] On B., Lee D. PaSE: Locating Online Copy of Scientific Documents Effectively. In Proceedings of the 7th International Conference of Asian Digital Libraries (ICADL), pp. 408-418, 2004.
- [17] Petricek V. Et al. A Comparison of On-line Computer Science Citation Databases. ECDL 2005.
- [18] Rexa Digital Library [Electronic resource] <http://rexa.org>.
- [19] Science Citation Index. [Electronic resource] <http://scientific.thomson.com/products/sci/>
- [20] Seymore K, McCallum A., Rosenfeld R. Learning for Information Extraction Learning Hidden Markov Model Structure for Information Extraction AAAI'99 Workshop on Machine, 1999.
- [21] Snowball Project. [Electronic resource] <http://snowball.tartarus.org/>.
- [22] XPDF Project. [Electronic resource] <http://www.foolabs.com/xpdf/>
- [23] Zhuang Z., Wagle R., Giles C.L.. What's There and What's Not? Focused Crawling for Missing Documents in Digital Libraries. JCDL 2005.

Digital library and Search engine for national-language research literature

Vasiliev S., Kozlov D., Samusev S., Shamina O.

This paper describes implementation of information system and digital library to store Russian and English papers together, search in both languages, extract metadata and citations from articles and interconnect Russian and English papers into common citation index. Also a method to automatically search for new research papers in Russian is described.