Автоматический поиск научных статей в сети Интернет

© Шамина О. Козлов Д.

Факультет Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова
sincere@lvk.cs.msu.su, ddk@cs.msu.su

Аннотация

В данной работе рассмотрены два частных случая задачи автоматического пополнения электронных библиотек (ЭБ) научных статей:

- поиск статьи по заданной метаинформации;
- поиск домашних страниц ученых и извлечение публикаций из них.

В работе предложены новые методы для двух частных случаев задачи. Оба метода позволяют искать русскоязычные научные статьи, для которых большинство существующих методов неприменимо из-за особенностей русскоязычного сегмента сети Интернет. Предложенные методы комбинируют различные подходы, что делает их более универсальными. В работе приводятся результаты экспериментального исследования работы предложенных методов.

1. Введение

С каждым днем увеличивается количество авторов, журналов, учебных заведений и электронных архивов, которые позволяют любому человеку получить доступ к их научным статьям через Интернет. Для того чтобы было легче ориентироваться среди этой рассредоточенной информации, создаются электронные библиотеки научных статей. Чем больше статей включает в себя библиотека, тем полезнее она пользователям. Одной из важных проблем является своевременное добавление в библиотеку новых статей.

Пополнение электронной библиотеки научными статьями может осуществляться следующими способами:

1. Добавление новых научных статей в библиотеку вручную. Таким образом, пополнение может осуществляться пользователями ЭБ, либо самими авторами или изданиями, которые хотят разместить свои статьи в библиотеке.
2. Полуавтоматический поиск новых статей. Так, в [11] описана процедура Relevance Feedback.
3. Автоматический поиск научных статей в сети Интернет и их добавление в библиотеку. Этот метод был предложен в работе [5]. Он позволяет охватить большое многообразие статей, представленных в сети Интернет.

Далее в работе рассматривается именно автоматический поиск научных статей.

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;
- 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;
- часто домашняя страница ученого (или страница организации) предоставляет не тексты статей, а только библиографические ссылки;
- 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;
- 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 исполь-

зуется большое разнообразие слов для обозначения одних и тех же разделов (например, только библиография может называться «литература», «ссылки», «источники», «список литературы» и т.п.).

2. Задача автоматического пополнения ЭБ научных статей

Задача автоматического пополнения электронной библиотеки состоит в том, чтобы автоматически находить в Интернет доступные статьи, которых еще нет в ее базе, и добавлять их в библиотеку.

Были выделены следующие частные случаи автоматического пополнения ЭБ:

1. *Поиск текста статьи по заданным авторам и названию.*

   Надо найти статью, которой нет в библиотеке, если известно название этой статьи и ее авторы. Такая задача может возникнуть, например, когда пользователь сам ищет конкретную статью, задавая ее метаинформацию. Также, в уже найденной статье может содержаться библиографическая ссылка на другую статью, которой еще нет в библиотеке. В этом случае из библиографической ссылки можно извлекать название и авторов статьи, что приводит к рассматриваемой задаче.

2. *Поиск домашней страницы автора/конференции/ учреждения/издания, содержащей список публикаций, по заданному имени автора, названию конференции/учреждения/издания. И поиск новых статей на этой странице.*

   На домашних страницах ученых, сайтах различных конференций, учреждений и изданий часто содержатся ссылки на различные научные статьи. Задача состоит в том, чтобы найти такие страницы, распознать на них ссылки на тексты научных статей и добавить новые статьи в библиотеку. Такая задача возникает, когда есть какие-то статьи одного автора, либо принадлежащие к одной конференции/учреждению/изданию, и хочется найти и другие статьи этого автора/конференции/учреждения/издания.

3. *Поиск статей по заданным авторам и предметной области.*

   Пусть имеется статья определенной тематики. Можно предположить, что авторы этой статьи занимаются исследованиями в области, к которой относится рассматриваемая статья. Тогда следует попро-

бовать найти другие их статьи, задавая их имена и ключевые слова, описывающие предметную область.

4. *Поиск статей по заданной предметной области.*

   В различных тематических сообществах, блогах, на форумах, посвященных каким-либо научным областям, часто делятся ссылками на различные статьи заданной тематики. Задача состоит в том, чтобы найти такие страницы, извлечь из них ссылки на публикации и загрузить новые статьи в библиотеку.

5. *Периодический поиск новых статей в известных источниках.*

   На домашних страницах ученых, сайтах конференций/ учреждений/изданий время от времени появляются новые статьи, которые тоже нужно добавлять в библиотеку. Поэтому необходимо периодически производить поиск обновлений в уже известных источниках научных статей.

В следующем разделе описаны существующие методы, которые применяются для англоязычных статей в ходе решения поставленной задачи. Однако русскоязычные научные статьи имеют свою специфику, описанную во введении, которую необходимо учитывать. В разделе 4 и 5 предложены решения для первого и второго (соответственно) частных случаев задачи для русскоязычных научных статей. В разделе 6 приводится описание экспериментального исследования разработанных методов.

3. Существующие методы автоматического поиска научных статей в сети Интернет

В этом разделе описаны существующие методы поиска статей по ключевым словам (3.1-3.4) и методы поиска домашних страниц ученых (3.2, 3.5-3.6).

3.1 Метод, применяемый в системе CiteSeer

В CiteSeer [5] 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 словам (СПКС) для того, чтобы найти страницы с большой вероятностью содержащие ссылки на научные публикации. Для этого запрос пользователя к CiteSeer дополняется ключевыми словами, характери-

зующими 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 [6].

3.2 Метод, использующий HPSearch и Mops

В работе [4] 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 на основании различных характеристик домашних страниц, после чего страницы с наибольшим весом загружаются, ранжируются еще раз и результат (несколько страниц-кандидатов) сохраняется в базе данных.

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 содержания URL.

3.3 Метод, использующий тематического поискового робота

В работе [10] предложен метод поиска научных статей с помощью тематического поискового робота. На первом шаге используется репозитарий 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 заносятся 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 страницы, найденные на предыдущем шаге, кроме того, роботу также задается список разрешенных для посещения доменов. Робот использует очередь с 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 архивов WebDB и JAIR методом было найдено около 80% статей.

3.4 Метод поиска по заданной библиографической ссылке

В работе [8] 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 сопоставление библиографической ссылки и ссылки на статью в рамках найденной страницы. Этот процесс подразделяется на два этапа:

1) поиск на странице библиографической ссылки, указывающей на искомую статью;

2) поиск PDF или PS файла, содержащего статью, отвечающую библиографической ссылке.

Так как среди различных элементов метаинформации заглавие статьи наиболее вероятно имеет только один вариант написания, именно оно ищется для начала на первом этапе. Далее в окрестности найденного названия ищутся заданные авторы статьи. Если они найдены, то они объединяются вместе с названием в блок библиографической ссылки, если не найдены, то блоком считается одно название.

В окрестности выделенного блока библиографической ссылки может быть несколько ссылок на PDF/PS-файлы. На втором этапе определяется, какая же из них – верная. Для этого вводится понятие “расстояния”. В качестве меры расстояния может выступать, например, количество байт или слов от блока библиографической ссылки до исходящей ссылки на файл. Далее выбирается ближайшая ссылка на файл.

3.5 Метод, основанный на машинном обучении

В работе [9] рассматривается использование машинного обучения для решения задачи поиска домашней страницы по запросу пользователя. Работа метода подразделяется на три этапа:

1. Поиск страниц релевантных запросу пользователя.

2. Построение дерева решений для определения, какие из найденных страниц являются домашними страницами, а какие – нет.

3. Применение модели логистической регрессии для совмещения различных оценок и предсказания, какая из домашних страниц наиболее релевантна пользовательскому запросу.

На первом этапе для нахождения страниц, релевантных пользовательскому запросу, используется обычная поисковая система. Однако поиск ведется не по всей странице, а по ее отдельным частям. Так в работе [9] было показано, что наиболее хорошие результаты получаются при использовании либо заглавия документа, либо анкерного текста, либо аннотации. Таким образом, для каждой части выдается своя оценка релевантности запросу. Далее выбираются несколько наиболее релевантных результатов и передаются на второй этап.

На втором этапе для каждой страницы строится вектор, содержащий следующие признаки:

• URL length: количество слэшей в URL;

• In link: количество входящих ссылок на страницу;

• In link normalized by homepage: количество входящих ссылок, нормализованное по страницам;

• In link from outer domain: количество входящих ссылок с других доменов;

• In link from same domain: количество входящих ссылок с того же домена;

• Out link: количество исходящих ссылок на странице;

• Out link normalized by homepage: количество исходящих ссылок, нормализованное по страницам;

• Out link to outer domain: количество исходящих ссылок, указывающих на другие домены;

• Out link to same domain: количество исходящих ссылок, указывающих на тот же самый домен;

• Keyword: заканчивается ли URL ключевым словом; ключевые слова: “home”, “homepage”, “index”, “default”, “main”;

• Slash: заканчивается ли URL слэшом “/”;

• Result: домашняя ли это страница.

На этапе обучения на основе построенных векторов строится дерево решений. На этапе обработки страниц это дерево применяется для проверки того, является ли она домашней страницей.

На третьем этапе для получения окончательного результата применяется модель логистической регрессии. Это модель, использующая несколько факторов (переменных) для предсказания вероятности события. В нашем случае, в качестве факторов выступают результаты, выданные на первом этапе (для заголовка, анкерного текста и аннотации), а также информация о ссылке и длине URL, а в качестве события – то, что документ релевантен запросу. При этом рассматриваются только те страницы, которые успешно прошли проверку на втором этапе.

Авторы метода приводят следующие результаты применения машинного обучения для поиска домашних страниц:

- для 66% запросов верная домашняя страница была представлена на первом месте;

- для 84% верная домашняя страница была найдена в десятке первых документов.

3.6 Метод, использующий общую поисковую стратегию

В работе [2] рассматривается применение техники “impact transformation”, описанной в работе [3], к решению задачи поиска домашних страниц. В этом методе страница с самой высокой релевантностью запросу считается искомой домашней страницей.

Техника “impact transformation” состоит в том, что степень соответствия документа запросу считается по формуле:

$$S_{d,q} = \sum_{t \in q \cap d} w_{d,t} \cdot w_{q,t} \ ,$$

где $w_{d,t}$ и $w_{q,t}$ – это “влияние” терма на документ и запрос соответственно. Они зависят от частоты встречаемости терма в документе и в запросе, от количества документов, в которых встречается терм и от длины документа.

Модификация этой техники для решения задачи нахождения домашних страниц представляет собой добавление слагаемого C к частоте встречаемости терма в документе. Значение C различается в зависимости от полей документа, в которых встречается терм. Таким образом:

C = 1, если терм встречается в тексте документа;

C = 8, если терм встречается в анкерном тексте исходящих ссылок или входящих ссылок с других доменов;

C = 4, если терм встречается в анкерном тексте входящих ссылок с того же самого домена;

C = 2, если терм входит в заголовок, ключевые слова или описание документа.

Такой метод продемонстрировал не очень хорошие результаты (меньше 80% найденных домашних страниц в первой десятке) и, таким образом, он пока проигрывает другим методам. Однако авторы считают важным то, что этот метод показал, как можно комбинировать различные задачи поиска, путем добавления различных слагаемых. А также доказал возможность построения поисковой системы, которая сможет сама, оптимизировать поиск, путем настройки дополнительных параметров, если будет известно, что именно ей надо искать.

3.7 Вероятностный метод

В работе [7] рассмотрен метод поиска домашних страниц, в котором для поиска используется комбинирование различной информации из самого текста документа и информации, представленной в структуре документа.

Сначала формируется база данных документов, найденных по запросу пользователя. Далее из документов удаляются стоп-слова, а у остальных

остаются только основы. Затем для каждого документа строится языковая модель.

Языковая модель определяет вероятности распределения всех слов в сформированной базе документов. Эти вероятности интерпретируются как вероятности порождения слова, и документы, в свою очередь, ранжируются по вероятности порождения термов запроса.

Наилучшие результаты (почти 90%) авторы получили при работе метода с учетом типа URL и при использовании специально подобранных весовых коэффициентах.

4. Метод поиска статьи по заданной метаинформации

На вход методу подаются элементы метаинформации, описывающие научную статью (название и авторы). Требуется найти в сети Интернет статью, отвечающую заданной метаинформации.

Процесс поиска подразделяется на четыре этапа:

1. Поиск с ограничением на формат файла.
2. Проверка, не найдена ли искомая статья.
3. Поиск с помощью тематического поискового робота.
4. Сопоставление библиографических ссылок и ссылок на файлы.

Далее каждый этап рассматривается более подробно.

4.1 Поиск с ограничением на формат файла

На этом этапе запрос, содержащий название статьи и список авторов, посылается поисковым системам (Яндекс и Google) с ограничением поиска только по файлам в формате PDF/PS (именно в этих форматах в основном хранятся научные публикации).

Далее для десяти первых результатов, до тех пор, пока не будет найдена искомая статья, выполняются следующие действия:

- документ, соответствующий рассматриваемому URL, сохраняется в файл;
- происходит конвертация полученного PDF или PS файла в текстовый формат;
- из текстового файла извлекается метаинформация с помощью метода, основанного на SVM [1];
- производится проверка, не найдена ли искомая статья (см. раздел 4.2).

4.2 Проверка, не найдена ли искомая статья

Для проверки того, является ли найденный документ искомой статьей, проводится:

1. Сравнение извлеченной метаинформации с заданной. Заглавие сравнивается полностью, без учета знаков препинания. Заданные авторы (инициалы не учитываются), ищутся по отдельности среди извлеченных из документа. Чем больше совпадений найдено, тем больше очков выдается в качестве результата сравнения.

2. Оценивается, насколько найденный документ похож на научную статью. Для этого учитываются следующие факторы:
   - научные статьи обычно содержат список литературы;
   - научные статьи обычно имеют больше двух и меньше тридцати страниц.

В случае если суммарная оценка превышает заданный порог, процесс поиска считается успешно завершенным.

4.3 Поиск с помощью тематического поискового робота

Если на первом этапе искомая статья не была найдена, то осуществляется поиск с помощью тематического поискового робота.

Первым шагом является формирование начального множества страниц, с которых робот начинает поиск. Для этого поисковым системам посылается тот же запрос, однако теперь поиск производится только среди HTML-страниц. Предполагается, что одна из таких страниц либо сама может содержать ссылку на файл с искомой статьей, либо может привести к странице, содержащей такую ссылку.

Далее для каждого из первых десяти результатов поиска определяется приоритет. Для этого составляются два списка ключевых слов (английский и русский). В них входят слова, которые часто встречаются на страницах, содержащих ссылки на публикации (например, research, publication, статьи, конференция и т.п.).

В списки также добавляются фамилии авторов искомой статьи. Если URL или заголовок страницы содержит ключевое слово, то ей присваивается высокий приоритет, в противном случае – низкий.

Затем на основе полученного приоритета страницы распределяются в две очереди, которые подаются роботу. Робот рассматривает только HTML-страницы. Также задается ограничение на глубину поиска: робот не ходит дальше третьего уровня.

Алгоритм поиска статьи поисковым роботом:

До тех пор, пока искомая статья не найдена, либо обе очереди не опустеют, выполняется следующее:

- загружается страница из очереди с высоким приоритетом (если эта очередь пуста, то рассматривается очередь с низким приоритетом);
- на странице ищется библиографическая ссылка на искомую статью и отвечающая ей ссылка на PDF или PS файл (см. раздел 4.4 “Сопоставление библиографических ссылок и ссылок на файлы”);
- если искомая статья не найдена или страница вообще не содержит библиографической ссылки на нее, то:
  - из страницы извлекаются все исходящие URL;
  - для каждого нового URL определяется приоритет, в соответствии с которым, он добавляется в нужную очередь.

Если обе очереди опустошаются, либо время, отведенное на поиск, истекает, а статья до этого момента не найдена, то процесс поиска считается завершенным неуспешно.

4.4 Сопоставление библиографических ссылок и ссылок на файлы

Когда поисковым роботом найдена страница, которая может содержать ссылку на статью, возникает задача найти на ней нужное библиографическое описание статьи. Следует отметить, что среди различных элементов метаинформации заглавие статьи наиболее вероятно имеет только один вариант написания. Поэтому поиск производится по следующему алгоритму:

- заданное название статьи ищется на странице;
-в окрестности найденного названия ищутся заданные авторы статьи;
- если они найдены, то они объединяются вместе с названием в блок библиографической ссылки;
- если не найдены, то блоком считается одно название;
- если не найдено даже название, то считается, что на этой странице ссылки на искомую статью не найдено.

После определения блока библиографической ссылки необходимо найти отвечающую ей ссылку на PDF/PS – файл, если такая существует. Для определения верной ссылки вводится понятие “расстояния”. В качестве меры расстояния выступает количество байт от блока библиографической ссылки до исходящей ссылки на файл. Если внутри блока есть ссылки, то выбираются они, в противном случае – выбираются две ближайшие с разных сторон к блоку ссылки на файлы. Выбранные ссылки проходят проверку, описанную в разделе 4.2.

5. Метод поиска домашних страниц и извлечения публикаций

На вход методу подаются фамилия, имя (отчество) ученого. Требуется найти в сети Интернет его домашнюю страницу и извлечь ссылки на публикации.

Работа метода подразделяется на три этапа:

1. Поиск домашней страницы.
2. Поиск страницы с публикациями.
3. Извлечение ссылок на публикации.

Далее каждый этап рассматривается более подробно.

5.1 Поиск домашней страницы

Поисковым системам (Яндекс и Google) посылается запрос, содержащий заданные данные ученого и ключевые слова – “homepage” и “домашняя страница”. Далее первые 10 результатов, выданных системами, проходят проверку, найдена ли искомая домашняя страница. Проверка состоит из двух этапов:

1. Первичная оценка. На этом этапе оценивается, насколько найденная страница похожа на домашнюю страницу, без рассмотрения ее содержания. Для этого используются следующие характеристики домашних страниц:
   - URL заканчивается символом ‘/’;
   - URL заканчивается строкой “.html”, либо “.htm”;
   - URL содержит в своей последней части “index”, “main”, “default”, “home”, либо “homepage”;
   - URL содержит ‘~’, “/people”, либо “/users”;
   - URL содержит не больше трех слешей;
   - URL содержит слово из заданных данных ученого.
2. Содержательная оценка. На этом этапе оценивается, насколько страница соответствует искомой. Для этого используются следующие признаки:
   - заголовки страницы содержат слова запроса, либо слова из списка ключевых слов для домашних страниц;
   - страница имеет размер меньше 10000 байт;
   - в тексте страницы встречаются слова из списка ключевых слов для домашних страниц;
   - на странице встречается слово “publications”, “papers”, “публикации”, либо “статьи”.

Страницы с хорошими результатами, полученными при первичной оценке, проходят вторую часть проверки – содержательную оценку. В результате, страницы, получившие наилучшие суммарные оценки, считаются найденными вариантами.

5.2 Поиск страницы с публикациями

На этом этапе в окрестностях найденной домашней страницы ученого производится поиск страницы, содержащей список публикаций. В качестве кандидатов рассматриваются все страницы, найденные по исходящим ссылкам с домашней страницы, анкерный текст или URL которых, содержит одно из следующих ключевых слов: publications, papers, articles, bibliography, публикации, статьи, библиография, труды, доклады, тезисы, а также сама домашняя страница.

Кандидаты оцениваются следующим образом:

- за каждое найденное слово из списка ключевых слов, приведенного выше, в URL или заголовках страницы, страница получает очко;
- если на странице больше 30% исходящих ссылок представляют собой ссылки на PDF/PS-файлы, то страница получает очко;
- за каждое найденное слово из списка ключевых слов для страниц публикаций страница получает очко.

Страницы с оценкой больше заданного порога считаются страницами, содержащими список публикаций, и передаются на следующий этап для извлечения ссылок на статьи.

5.3 Извлечение ссылок на публикации

Для извлечения отдельных библиографических ссылок из страницы, ее содержание отделяется от HTML-разметки. При этом отдельные абзацы записываются в одну строку, а различные заголовки страницы удаляются. Затем строки представляются с помощью векторов признаков, содержащих свойства отдельных слов и свойства всей строки в целом. Далее, вектора признаков классифицируются с помощью метода опорных векторов (SVM). Классификатор обучен на наборе данных, содержащем библиографические ссылки и обычный текст, для определения, является ли рассматриваемая строка библиографической ссылкой или нет. Более подробно о том, как строятся вектора признаков, происходит обучение и распознавание, описано в [1].

Когда найдены строки, представляющие собой библиографические ссылки на статьи, в их окрестностях производится поиск ссылок на фай-

лы PDF/PS. Для этого содержимое страницы рассматривается уже полностью (вместе с разметкой). Сопоставление библиографической ссылки и ссылки на файл со статьей описано в разделе 4.4. Если файл найден, то происходит проверка того, найдена ли статья, отвечающая библиографической ссылке. Проверка происходит аналогично той, что описана в разделе 4.2, с той лишь разницей, что извлеченная метаинформация ищется в заданной библиографической ссылке. Возможным улучшением предложенного метода является разбор найденной библиографической ссылки на отдельные поля, представляющие собой метаинформацию искомой статьи. И тогда проверку уже можно будет осуществить по отдельным элементам метаинформации, как и в разделе 4.2.

6. Экспериментальное исследование методов

В рамках данной работы было проведено экспериментальное исследование методов, целью которого было выяснить, насколько успешно предложенные методы справляются с поставленной задачей.

6.1 Метод поиска статьи по заданной метаинформации

В качестве набора данных для экспериментального исследования использовался список публикаций сотрудников факультета ВМиК за 2004-2006 года. Этот список включает в себя статьи, которые могут быть как в электронном, так и в печатном виде. Задача состояла в том, чтобы найти все статьи из этого списка (по авторам и названию), которые доступны в сети Интернет.

Всего в наборе данных 939 статей:

- 311 статей за 2004 год;
- 324 статьи за 2005 год;
- 304 статьи за 2006 год.

Из них только для 15 статей указано, что они размещены в Интернет.

Осуществлялся поиск научной статьи, заданной названием и авторами, с помощью предложенного метода.

В случае если статья была найдена, вручную проверялось, соответствует ли она искомой. Если соответствовала, то считалось, что статья найдена верно, в противном случае – ошибочно.

Если статья не была найдена, то просматривались все загруженные во время поиска документы. Если вдруг среди них была обнаружена искомая статья, то считалось, что она не была найдена из-за ошибки распознавания (то есть статья не прошла проверку, описанную в разделе 4.2).

Затем для сравнения и оценки результатов производился поиск этой статьи с помощью системы Google Scholar. Рассматривались только первые десять результатов, выданные системой.

В итоге возможны четыре исхода:

Статья найдена верно (класс A):

- Google Scholar тоже ее нашел;
- Google Scholar ее не нашел.

Статья найдена ошибочно (класс B):

- Google Scholar нашел искомую статью;
- Google Scholar не нашел искомую статью.

Статья не найдена, но Google Scholar ее нашел (класс C):

- статья была найдена, но не было распознано, что это искомая статья;
- статья не была найдена.

Статья не найдена, и Google Scholar ее не нашел (класс D):

- статья была найдена, но не было распознано, что это искомая статья;
- статья не была найдена.

Для оценки работы метода были введены следующие характеристики:

- precision (точность в классе найденных) – доля найденных верно статей среди найденных вообще (формула (1));
- accuracy (точность) – доля верных исходов среди всех возможных исходов (формула (2));
- recall (полнота) – доля найденных статей среди тех, которые можно было найти (формула (3)).

$$precision = \frac{|A|}{|A|+|B|}, \tag{1}$$

$$accuracy = \frac{|A|+|D|}{|A|+|B|+|C|+|D|}, \tag{2}$$

$$recall = \frac{|A|}{|A|+|C|}. \tag{3}$$

Таблица 1. Оценки, полученные в результате экспериментального исследования

Оценка / Год	**Precision**	**Accuracy**	**Recall**
2004	62,5%	99%	100%
2005	54%	97,8%	75%
2006	75%	98%	75%
Все	66,6%	98,2%	80,9%

В ходе экспериментального исследования работы метода были выявлены следующие проблемы, приводящие к ошибкам:

- некоторые PDF-файлы невозможно конвертировать в текстовый формат. Соответственно из них невозможно верно извлечь метаинформацию, чтобы проверить, найдена ли искомая статья;
- встречаются статьи одних и тех же авторов с очень похожими названиями, в результате чего на этапе проверки, не найдена ли искомая статья, одна из таких статей может быть ошибочно принята за другую;
- некоторые статьи встречаются исключительно в сборниках, представляющихся одним огромным файлом, что не позволяет их найти;
- ошибки и опечатки в названии статьи или ее авторах также могут привести к ошибкам поиска.

В целом, экспериментальное исследование показало, что предложенный метод справляется с задачей поиска статей по заданной метаинформации и в некоторых случаях даже превосходит Google Scholar.

Однако обнаружилось, что очень мало статей из набора данных размещено в Интернет, что не могло не сказаться на результатах исследования. Именно из-за этого показатель Accuracy столь высок, а Precision, в свою очередь, достаточно низок.

6.2 Метод поиска домашних страниц и извлечения публикаций

В качестве набора данных для экспериментального исследования использовался список, содержащий ФИО ученых МГУ им. М.В. Ломоносова. В этот список вошли первые 500 авторов по количеству опубликованных работ. Данные о них содержатся в Российском Индексе Научного Цитирования, построенном в рамках проекта Научной Электронной Библиотеки e-LIBRARY (www.elibrary.ru).

В рамках исследования осуществлялся поиск домашней страницы ученого по его имени. Далее, вручную оценивались выданные системой результаты. Определялось, являлась ли найденная страница домашней страницей заданного ученого. Если являлась, то считалось, что статья была найдена верно, и в ее окрестности производился поиск страницы со списком публикаций. Результаты этого поиска также проверялись вручную.

Для оценки полученных результатов была введена характеристика точности поиска, определяемая по формуле:

precision = количество страниц, найденных верно / общее количество найденных страниц

Таблица 2. Результаты тестирования метода поиска домашних страниц и страниц публикаций

	Верно	**Ошибочно**	**Всего**	**Precision**
Дом. стр.	84	40	124	68%
Стр. публ-ций	7	4	11	63%

В ходе экспериментального исследования работы метода были выявлены следующие проблемы:

1. В различных каталогах (например, каталогах ученых, списках сотрудников, списках лауреатов каких-либо премий и т.п.) часто содержатся страницы, посвященные ученым. Такие страницы во многом соответствуют характеристикам домашних страниц, и довольно часто встречаются в результатах поиска. С одной стороны, такие страницы могут оказаться столь же полезными, как и домашние. Например, на некоторых из них содержатся довольно подробные описания жизни и деятельности интересующих нас людей, и даже приводится список их публикаций, что нам собственно и нужно. В таких случаях логично отнести страницу к найденным верно (особенно, если настоящая домашняя страница у ученого отсутствует). Однако, с другой стороны, такие страницы могут содержать минимум информации. В этом случае их присутствие в результатах поиска логично относить к найденным ошибочно, т.к. они не приносят никакой пользы, однако могут сместить на дальние позиции более содержательные страницы, и даже настоящие искомые домашние страницы. Автоматическое нахождение границы между содержательными и несодержательными страницами такого рода представляет собой большую проблему.
2. Для ученых, у которых есть более знаменитые однофамильцы (а, возможно, еще и тезки), в результатах поиска могут ошибочно присутствовать домашние страницы этих однофамильцев. Особенно это актуально для людей с простыми и распространенными фамилиями. В случаях, когда совпадает только фамилия, можно попробовать искать только те страницы, на которых обязательно встречается нужное имя. В случаях же полного совпадения автоматически отделить чужие страницы будет очень сложно. Для начала можно попробовать дополнять запрос ключевыми словами научной области, в которой работает автор, чтобы отодви-

нуть в результатах поиска хотя бы тех однофамильцев, которые занимаются совсем другими вещами.
3. Список публикаций иногда встречается не на HTML-страницах, а в отдельно приложенных архивах или документах в различных форматах. В связи с этим в дальнейшем можно пробовать усовершенствовать метод, путем поиска публикаций в приложенных файлах.

В целом, экспериментальное исследование показало, что предложенный метод справляется с задачей поиска домашних страниц и страниц публикаций по заданной информации об ученых.

Было найдено 84 домашние страницы ученых из списка авторов, имеющих наибольшее количество статей. Однако, такая выборка, возможно, не достаточно объективно отражает процент ученых, ведущих домашние страницы. Было замечено, что в этом списке преобладающее большинство людей уже довольно преклонного возраста, что, в принципе, логично, т.к. чем дольше человек занимается научной деятельностью, тем больше у него публикаций. Ну а так, как ведение домашних страниц распространилось в России не очень давно, можно предположить, что более молодые ученые с большей вероятностью заводят домашние страницы. В связи с этим предполагается провести еще одно исследование, в котором в качестве набора данных будет использоваться список сотрудников факультета ВМиК, выбранных независимо от количества их публикаций.

7. Заключение

В рамках данной работы предложены решения для двух частных случаев задачи автоматического пополнения ЭБ научных статей:

1. Поиск статьи по заданной метаинформации.
2. Поиск домашних страниц ученых и извлечение публикаций из них.

Предложенные методы учитывают специфику русскоязычных научных статей и успешно применяются для их поиска. Было проведено экспериментальное исследование работы предложенных методов. Полученные результаты показали, что методы успешно справляются с поставленной задачей. Однако в ходе исследования были выявлены проблемы, решение которых представляет собой дальнейшее развитие методов и, скорее всего, позволит достичь более хороших результатов работы.

Литература

[1] Козлов Д., Самусев С., Шамина О. Создание электронной библиотеки русскоязычных научных статей. // Сборник работ стипендиатов гранта "Интернет-информатика 2007", Екатеринбург, Изд-во Уральского университета, 2007, С. 37-45.

[2] Anh, V. N. and Moffat, A. Homepage Finding and Topic Distillation using a Common Retrieval Strategy. In Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), 2002.

[3] N. Anh and A. Moffat. Impact transformation: Effective and efficient web retrieval. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. J¨arvelin, editors, Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–10, Tampere, Finland, August 2002. ACM Press, New York.

[4] 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.

[5] Lawrence S., Bollacker K., Giles C.L. Indexing and retrieval of scientific literature, Proceedings of the eighth international conference on Information and knowledge management, p.139-146, 1999.

[6] Lawrence S., Giles L. Inquirus, the NECI meta search engine // Proceedings of the seventh international conference on World Wide Web 7. 1998.

[7] Ogilvie, P. and Callan, J. Combining Structural Information and the Use of Priors in Mixed Named-Page and Homepage Finding. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), pp. 177-184, 2003.

[8] 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.

[9] Xi, W. and Fox, E. A. Machine Learning Approach for Homepage Finding Task. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001), pp. 686-698, 2001.

[10] Zhuang Z., Wagle R., Giles C.L.. What's There and What's Not? Focused Crawling for Missing Documents in Digital Libraries. JCDL 2005.

[11] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. ACM Press Series/Addison Wesley, New York, May 1999.