Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей*

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

Козлов Д. ВМиК МГУ ddk@cs.msu.su

Самусев С. ВМиК МГУ sam@lvk.cs.msu.su

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

Аннотация

В данной работе рассмотрены существующие методы автоматического извлечения метаданных и библиографических ссылок из текстов научных статей, описаны адаптация методов для работы с русскоязычными научными статьями и сравнение эффективности работы этих методов на русскоязычных и англоязычных научных статьях.

1 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. Тексты публикаций, как правило, размещают в форматах Postscript (PS), PDF, Microsoft Word Document (DOC) и, как правило, не снабжают метаинформацией (например, в виде BibTeX-описания)

Потребность в сохранении, накоплении и распространении результатов научных исследований привела к широкому распространению электронных архивов научных статей, например, CORR [3], NCSTRL [13]. Такие архивы пополняются вручную авторами, желающими разместить свои статьи. В то же время большое количество статей остается вне таких архивов. В связи с этим в работе [9] было предложено строить электронные библиотеки научных статей как вторичные информационные ресурсы: научные статьи в такую библиотеку помещаются автоматически в результате поиска в сети Интернет на домашних страницах авторов. На основе этого подхода построена широко известная библиотека CiteSeer.IST [4]. Одной из основных ее особенностей является технология Autonomous Citation Indexing (ACI) [10], которая предусматривает разбор текста статьи, извлечение из него метаданных (заглавия, авторов и т.п.) и библиографических ссылок и построение графа взаимного цитирования научных статей. Такой подход позволил существенно расширить возможности поиска научных статей в электронных библиотеках, предоставив исследователям возможность пользоваться не только традиционным поиском по ключевым словам, но и навигацией по библиографическим ссылкам.

2 Задача извлечения метаданных

Одним из основных моментов в технологии ACI является автоматическое извлечение метаданных и библиографических ссылок из текстов статей без участия человека. Задача состоит в том, чтобы из текста статьи, представленного в формате PDF (или, например, PS, DOC), извлечь подстроки, соответствующие каждому атрибуту метаданных в терминах стандарта Dublin Core, и библиографические ссылки.

Решение задачи, как правило, разбивается на два этапа:

– построение промежуточного текстового представления статьи, снабженного некоторой дополнительной разметкой, например, информацией о размере шрифтов;
– извлечение метаинформации и библиографических ссылок из промежуточного представления.

При этом следует отметить следующие особенности данной задачи:

– при преобразовании исходной статьи в текстовое промежуточное представление частично или полностью теряется информация, позволяющая человеку визуально отличать элементы метаинформации, такие как заглавие, авторов и т.д.;
– научная статья не содержит специализированной разметки элементов, относящихся к метаинформации;
– для англоязычных статей существует несколько общепринятых стандартов оформления. Поэтому большинство англоязычных статей очень схожи по структуре: обязательно

* Данная работа выполнена при частичной поддержке компании Яндекс в рамках гранта «Интернет-математика 2007».

присутствуют заглавие, авторы, организация, адрес, email, затем abstract (аннотация), introduction (введение), в конце есть conclusion (заключение) и references (список литературы). Также для написания англоязычных статей широко используются стандартные стили LaTeX;

– для русскоязычных статей нет общепринятых норм, и авторы оформляют статьи, руководствуясь исключительно своими пожеланиями (требования разных конференций также очень сильно различаются). К тому же в русском языке приняты разные обозначения заголовков разделов статьи (например, только библиография может называться «литература», «ссылки», «источники», «список литературы» и т.п.).

Для построения промежуточного представления статьи используются конверторы, преобразующие PDF (или PS, DOC) в текст. Например, в [9] использована модифицированная утилита ps2text из состава Ghostscript, а в рамках данной работы – модифицированный вариант pdf2text из состава Xpdf. При этом конверторы в обеих реализациях модифицированы таким образом, чтобы частично сохранять информацию о разметке: окончания строк, изменения шрифтов.

Методы решения второго этапа задачи являются объектом исследования данной работы. В разделе 3 кратко рассмотрены существующие наработки в этой области.

3 Существующие методы извлечения метаданных

Для извлечения метаданных из текстов научных статей применяют метод, основанный на регулярных выражениях, и различные методы машинного обучения: на основе классификации с помощью метода опорных векторов (SVM) [6], скрытых Марковских моделей (HMM) [15], Марковских моделей максимальной энтропии [12], условных случайных полей [8]. В данной работе проводится экспериментальное исследование трех базовых методов: метода, основанного на регулярных выражениях, метода, основанного на скрытых Марковских моделях, метода, основанного на классификации.

3.1 Метод, основанный на регулярных выражениях

В работе [9] для извлечения метаданных и библиографических ссылок из англоязычных научных статей используется метод, основанный на регулярных выражениях.

Входными данными для метода является промежуточное представление статьи, содержащее помимо самого текста статьи дополнительную разметку в виде тэгов <E'n'> и <F+/-'n'>. Тэг <E'n'> обозначает конец строки в исходном файле, где 'n' - расстояние до правого края страницы. Тэг <F+/-'n'> означает изменение размера шрифта, где +/-'n' - это разница между текущим размером шрифта и размером шрифта предыдущего фрагмента.

Работа метода заключается в том, что для каждого элемента метаинформации строится свое регулярное выражение, описывающее данный элемент. Подстрока, соответствующая образцу поиска, заданному регулярным выражением, ищется по всему тексту. Как правило, образец содержит в себе специальные слова, при нахождении которых в строке можно сделать вывод либо о принадлежности всей строки или ее части к соответствующему классу, либо о принадлежности соседнего фрагмента к какому-либо классу. Так, например, при нахождении слова "University" в строке можно считать, что эта строка, скорее всего, относится к элементу "Affiliation", а при нахождении слова "References" можно сделать вывод, что далее, скорее всего, следует список литературы.

Также, применяется и обратное использование ключевых слов, т.е. для каждого класса составляются группы определенных специальных слов, которые показывают, что строка или ее часть, вероятно принадлежит какому-нибудь другому классу, и значит, не является искомым элементом метаинформации. В группу таких слов для элемента "Title" входят, например, 'University', 'Abstract', 'References' и т.п.

Для извлечения заголовка рассматриваемым методом также используется информация о том, что он должен располагаться в начале статьи и выравниваться по центру, а также то, что после него, скорее всего, происходит уменьшение размера шрифта, которое характеризуется тэгом <F-'n'>.

В случае поиска авторов учитывается тот факт, что обычно указание на них расположено после заголовка статьи, по центру и удовлетворяет одному из регулярных выражений, описывающих возможное представление списка имен авторов. Например, имя автора может содержать либо имя и фамилию, либо инициалы и фамилию. Предполагается, что в списке авторов не могут содержаться никакие другие символы, кроме букв, запятых, точек и дефисов.

Все, что располагается после слова ‘References’, считается списком литературы, если каждый абзац начинается с ‘n.’, с ‘[n]’, с ‘n)’, с ‘(n)’ или просто с ‘n’, где n - число. Библиографической ссылкой считается предложение в тексте, содержащее ‘[n]’ ,‘n.’, ‘n)’ или ‘(n)’.

Электронным адресом считается выражение соответствующее модели ‘слово@слово.домен’, где ‘слово’ – любая последовательность символов без пробелов и @, а ‘доменом’ может быть ru, com, fr и т.п.. А при обнаружении последовательности ‘www.слово.домен’ считается, что найден URL.

Метод, основанный на регулярных выражениях на практике используется в электронной библиотеке CiteSeer.IST.

Данный метод требует настройки вручную экспертом на особенности оформления статей. При большом разнообразии способов оформления этот метод теряет свою эффективность.

3.2 Метод, основанный на Марковских моделях

Для повышения точности извлечения в работе [11] в рамках проекта Cora было предложено использовать методы машинного обучения, основанные на скрытых Марковских моделях (HMM) [5], которые до этого широко применялись в задачах распознавания речи.

Входными данными для метода является промежуточное представление статьи, содержащее только символы перевода строк.

Работа метода состоит из этапа обучения, на котором на вход методу подается размеченная обучающая выборка и производится построение и настройка HMM и этапа распознавания, на котором настроенная HMM используется для извлечения метаинформации (для библиографических ссылок строится и используется отдельная HMM).

На этапе обучения осуществляется выбор структуры и настройка параметров HMM. Структура HMM может быть фиксированной, где состояния соответствуют элементам метаинформации, а может строиться на основе обучающей выборки, как это сделано в [15]. После определения структуры HMM, вероятности переходов между состояниями вычисляются как отношение количества переходов из состояния i в состояние j в обучающей выборке к общему числу переходов из состояния i в любое другое состояние. А вероятности порождения лексемы k состоянием j вычисляется как отношение количества лексем k, порожденных состоянием j в обучающей выборке, к общему числу всех лексем, порожденных состоянием j в этой же выборке. Затем производится настройка параметров HMM с помощью алгоритма Баума-Уэлша.

Метод, основанный на скрытых Марковских моделях применялся в проекте Cora.

Данный метод требует построения обучающей выборки, в которой размечены все метаданные, но в тоже время метод существенно легче адаптировать к новым стилям оформления статей – для этого требуется пополнение обучающей выборки и повторное обучение. Однако математический аппарат HMM использует предположения о независимости будущих состояний от прошлых, что не верно, например, для библиографии. Поэтому существует ряд развитий этого метода, снимающих данное предположение, в частности Марковские модели максимальной энтропии, условные случайные поля.

3.3 Метод, основанный на классификации

В работе [6] было предложено рассматривать задачу извлечения метаданных как задачу классификации строк статьи.

Входными данными для метода является промежуточное представление статьи, содержащее только символы перевода строк. Метод работает только с первой страницей статьи.

Метод аналогично предыдущему состоит из этапа обучения и этапа распознавания.

Для классификации используется метод опорных векторов. Каждому элементу метаинформации соответствует собственный класс. Многоклассовая классификация осуществляется путем построения бинарных классификаторов по стратегии «один против всех». Каждая строка может быть отнесена к одному или нескольким классам. Согласно [6], 95% строк относятся к одному классу, а строки, относящиеся к нескольким классам, как правило, содержат списки авторов или описания мест работы авторов.

Классификация строк на этапе распознавания производится в два шага: независимая классификация и контекстно-зависимая классификация. На первом шаге каждая строка классифицируется отдельно на основе содержащихся в ней слов. При построении вектора признаков из строки используется метод Rule-based Word Clustering [7], сопоставляющий каждому слову некоторый класс и существенно снижающий размерность задачи классификации. На втором этапе в вектор признаков добавляются контекстные признаки – позиция строки в тексте, классы соседних строк и др.

Для строк, относящихся к нескольким классам, производится разделение строки путем поиска оптимального положения разделителя в строке среди всех возможных положений. Возможными положениями разделителя считаются пробелы и знаки препинания. Для заданного положения разделителя каждая из частей строки классифицируется, а целевая функция при оптимизации линейно зависит от оценок классификаторов.

Авторы метода приводят результаты экспериментов, показывающие превосходство данного метода над методом, основанным на использовании HMM.

3.2 Методы для работы с русскоязычными статьями

Все описанные выше методы извлечения метаданных и библиографических ссылок из статей ориентированы на работу с англоязычными статьями. Единственной разработкой, ориентированной на работу с русскоязычными статьями является проект Российского индекса научного цитирования (РИНЦ), проводимый в научной электронной библиотеки eLibrary.ru, который, судя по [2], ориентирован не на автоматическую, а на полуавтоматическую работу с участием человека.

4. Адаптация методов для работы с русскоязычными научными статьями

В рамках данной работы была произведена адаптация для работы с русскоязычными научными статьями следующих методов: метода, основанного на регулярных выражениях, метода, основанного на скрытых Марковских моделях, метода, основанного на классификации.

В адаптированном методе, основанном на регулярных выражениях, семантика использования разметки была оставлена в соответствии с [9]. В остальном метод притерпел существенные изменения, поэтому ниже приведено описание алгоритмов извлечения метаданных и библиографических ссылок.

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).

   то строка удаляется и поиск происходит снова (переход к пункту 3).
5) На данном этапе возможны следующие варианты:
   а) Если заголовок статьи не был найден, т.е. ни одна из строк не прошла проверку, то следует переход к шагу 6.
   б) Если заголовок был найден, но список авторов найден не был: следует 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.)
6) 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 метаинформации 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 модели «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 следующее за списком авторов.
4) Год издания извлекается по шаблону.

Для адаптации методов машинного обучения [5,6] была построена упрощенная и адаптированная для русского языка реализация Rule-based Word Clustering сокращения пространства признаков. Целью метода является преобразование описанного выше промежуточного представления статьи таким

образом, чтобы избавиться от большого разнообразия возможных слов, заменив слова на признаки, число которых невелико (десятки-сотни). Признаки заранее формируются на основе той роли, которую слова могут играть в структуре научной статьи. Кластеризация слов осуществляется на основе так называемых доменных баз данных (например, базы данных русских имен, базы данных названий городов) и орфографических свойств слов (регистра букв, наличия характерной структуры, например, aaa@bbb.ru , и т.п.). В рамках адаптированного метода использовалась следующие признаки:

Таблица 1. Признаки, основанные на свойствах слов.

Признак	Объяснение
:email:	Использование соответствия регулярному выражению
:url:	Использование соответствия регулярному выражению
:singleCap:	Заглавная буква, как «M» или «M.».
:postcode:	Почтовый индекс
:abstract:	Аннотация
:intro:	Введение
:month:	Назване месяца
:prep:	Предлоги
:degree:	Слово или выражение из списка слов домена степеней
:pubnum:	Слово или выражение из списка слов домена номеров публикаций
:notenum:	Слово или выражение из списка слов домена замечаний
:affi:	Слово или выражение из списка слов домена организаций
:addr:	Слово или выражение из списка слов домена адресов
:city:	Слово из списка названий городов
:state:	Слово из списка названий штатов
:country:	Слово из списка названий стран
:mayName:	Слово из одного из трех списков имен
:Cap1DictWord:	Словарное слово, начинающееся с заглавной буквы (аналогичные признаки для несловарных слов и слов целиком написанных заглавными буквами)
:DictWord:	Словарное слово, начинающееся с маленькой буквы
:NonDictWord:	Несловарное слово, начинающееся с маленькой буквы
:Dig[X]:	Число из X цифр

Поиск признака для слова осуществляется от наиболее специфичного к общему. Например, слово из списка имен может быть также словарным словом и т.п. Для слов, относящихся к нескольким признакам используется оригинальная форма слова. Для слов, состоящих из сложных последовательностей букв и цифр могут формироваться сложные признаки, например, «:Dig[2]:CapWord::Dig[3]:».

В созданной реализации используются следующие русскоязычные доменные базы данных:

- словарь русского языка ispell;
- словарь русских имен;
- словарь русских фамилий;
- список городов России;
- список месяцев и их аббревиатур;
- список слов, встречающихся в классе адреса;
- список слов, встречающихся в классе организаций;
- список предлогов;
- список слов, обозначающих класс аннотации;
- список слов, обозначающих класс введения;
- список слов, обозначающих класс ключевых слов.

В качестве дополнения к признакам, основанным на словах могут также использоваться статистические признаки, отражающие относительное количество слов того или иного класса в строке, например, процент имен в строке и т.п. Такие признаки также учитываются при использовании метода Rule-Based Word Clustering с методом, основанным на классификации.

Метод, основанный на скрытых Марковских моделях применялся совместно с построенной реализацией Rule-based Word Clustering, в то время как в работе [5] нет упоминания о том, какое используется пространство признаков. В методе, основанном на скрытых Марковских моделях, использовалась фиксированная структура модели, состояния которой соответствовали элементам метаинформации.

Метод, основанный на классификации был адаптирован для извлечения не только метаданных, но и библиографических ссылок. Извлечение библиографических ссылок происходит следующим образом:

- все строки оцениваются классификаторами атрибутов списка литературы;
- происходит объединение строк, относящихся к одному полю списка литературы, в одну строку;
- в каждой строке, соответствующей одному полю списка литературы, с помощью классификаторов для обработки строк, содержащих несколько классов, выделяются отдельные атрибуты, предсказанные для нее. Для извлечения отдельных атрибутов используется алгоритм разделения многоклассовых строк, применявшийся в [6].

5. Экспериментальное исследование методов

В рамках данной работы было произведено

экспериментальное исследование методов, целью которого было сравнение точности извлечения метаинформации и библиографических ссылок из текстов русскоязычных и англоязычных научных статей.

Для проведения экспериментов было использовано два набора данных:

- англоязычный набор, подготовленный в рамках проекта Cora, содержащий 935 заголовков научных статей и 500 библиографических ссылок;
- русскоязычный набор данных, содержащий около 180 русскоязычных статей с более чем 1000 библиографических ссылок.

Русскоязычный набор данных был построен из материалов отечественных конференций и сборников разных лет: РОМИП, Диалог, Математические методы распознавания образов, Интернет-Математика, публикации с сайта graphics.cs.msu.su. Случайным образом было выбрано по 10 статей из каждой конференции (и каждого года). Все статьи, входящие в набор данных, были преобразованы в текстовый формат (с сохранением дополнительной разметки с помощью модифицированного варианта программы pdf2text из состава Xpdf [1]) и размечены специальными тэгами, выделяющими элементы метаинформации. Данная разметка была предложена еще в рамках проекта Cora при построении англоязычного набора данных. Пример размеченного фрагмента документа приведен ниже:

<title> Эффективное использование C++ </title>
<author> Скотт Майерс </author>
<mail> ec++@awl.com </mail>
<url> http://www.awl.com/cp/ec++.html </url>

Эта разметка использовалась при обучении методов и для проверки автоматического сравнения правильности извлечения метаинформации.

Набор данных был разделен на 2 непересекающихся множества: обучающую выборку и тестирующую выборку. Было проведены эксперименты по определению корректности работы методов и определению точности работы методов. Проверка на корректность выполнялась путем тестирования работы метода на тех же данных, на которых происходило обучение. Оценка точности осуществлялась на данных, не входящих в обучающую выборку.

В Таблицах 2 и 3 приведены результаты экспериментов по определению корректности методов. В Таблицах 4-6 представлены результаты экспериментов по определению точности методов на русскоязычном и англоязычном наборах данных.

Проведенное экспериментальное исследование показало, что описанные методы применимы для извлечения метаинформации и библиографических ссылок из текстов научных статей. В то же время у всех методов есть тенденция не точно определять границу того или иного элемента метаинформации: часто добавляются лишние слова в заголовок. Для решения этой проблемы в методах машинного обучения необходимо более полно учитывать дополнительную разметку в документе, как это делается в методе, основанном на регулярных выражениях. Часто в статьях встречаются имена и фамилии авторов, не описанные в доменных базах данных. В этом случае строка авторов извлекается неверно. Также часто неверно классифицируются аббревиатуры. Эти проблемы могут быть решены расширением существующих доменных баз данных. В статьях встречается способ одновременной записи нескольких электронных адресов в виде: {kozachuk shabanov dobrin}@mail.com, такая форма записи извлекается неверно. Примечания, расположенные в начале статьи, содержащие информацию о названии конференции или сборника, часто классифицируются методами машинного обучения как заглавие статьи. Для устранения этого недостатка требуется учет размеров шрифтов. Все три метода неправильно работают в случае, когда на вход подается текст диссертации или автореферата, содержащий титульный лист.

6. Заключение

В рамках данной работы была произведена адаптация трех методов извлечения метаданных для работы с русскоязычными статьями и проведено экспериментальное исследование точности работы методов. Полученные результаты показали, что точность работы методов на русскоязычных научных статьях существенно хуже, чем на англоязычных. Такой результат вполне предсказуем: ведь для отечественных статей не существует общепринятых норм оформления, а каждое издание предлагает авторам свой вариант. В тоже время возможности по оптимизации методов далеко не исчерпаны. В частности, методы машинного обучения практически не используют дополнительную разметку, использование которой позволяет методу, основанному на регулярных выражениях, получать хорошие результаты.

В дальнейшем планируется развитие функциональности рассмотренных методов для более полного учета дополнительной разметки и исследование развитий метода, основанного на скрытых Марковских моделях: Марковских моделей максимальной энтропии и условных случайных полей.

Литература

[1] Проект XPDF. http://www.foolabs.com/xpdf/
[2] Разработка РИНЦ. Проект. http://elibrary.ru/projects/citation/proposal.doc.
[3] Computing Research Repository. http://arxiv.org/corr/home.
[4] CiteSeer.IST

http://citeseer.ist.psu.edu.
[5] Freitag D., McCallum A. Information extraction with HMMs and shrinkage. Proceedings of the AAAI-99 Workshop on Machine Learning for Informatino Extraction, 1999.
[6] Giles L. et al. Automatic Document Metadata Extraction using Support Vector Machines, JCDL, 2003.
[7] Han H., Manavoglu E., Zha H., Tsioutsiouliklis K., Giles C.L., Zhang H.. Rule-based word clustering for document metadata extraction ACM Press, 2005.
[8] Lafferty J., Pereira F., McCallum A.. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, 2001.
[9] Lawrence S., Bollaker K., Giles L., Indexing and Retrieval of Scientific Literature, CIKM, 1999.
[10] Lawrence S., Giles L., Bollaker K., Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Vol 32, N 6, 1999.
[11] McCallum A. Information Extraction: Distilling Structured Data from Unstructured Text. ACM Queue, volume 3, 2005.
[12] McCallum A., Freitag D., Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation. ICML-2000
[13] National Computer Science Technical Report Library. http://ncstrl.org
[14] Rabiner L.. A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 1989.
[15] Seymore K, McCallum A., Rosenfeld R. Learning for Information Extraction Learning Hidden Markov Model Structure for Information Extraction AAAI'99 Workshop on Machine, 1999.

Automatic document metadata extraction from Russian scientific articles

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

Automatic document metadata extraction provides useful search mechanisms for digital libraries. In this paper three metadata extraction techniques are experimentally compared for metadata and bibliography extraction from Russian scientific articles.

Таблица 2. Корректность извлечения метаинформации на русскоязычном наборе данных

Оценка / Класс	Извлечено		Не извлечено		Извлечено лишнее		Извлечено не все	
	SVM	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*
Заглавие	89.3	91.9	0	2.2	10.7	1.1	0	5.7
Авторы	87.3	89.5	3.6	1.1	5.5	6.9	3.6	2.3
Дата	100	100	0	0	0	0	0	0
Email	91.9	84.8	0	1.5	2.7	7.5	8.1	7.5
URL	100	100	0	0	0	0	0	0
Организация	87.2	57.1	0	28.6	2.6	8.6	10.2	7.1
Город	94.4	74	0	7.4	0	14.8	5.6	3.7
Адрес	100	83.3	0	0	0	16.7	0	0
Введение	85.7	97.1	7.15	0	0	2.9	7.15	0
Аннотация	73.3	50	6.6	0	3.3	28.6	16.6	26.8

Таблица 3. Корректность извлечения метаинформации на англоязычном наборе данных.

Оценка / Класс	Извлечено		Не извлечено		Извлечено лишнее		Извлечено не все	
	SVM	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*
Заглавие	89.9	51.2	0	8.9	4	29.2	6.1	18.6
Авторы	92.4	65.8	0	5.2	1.5	26	6.1	7.3
Дата	89.1	82.8	0	2.2	8.1	12.7	4	2.2
Email	93.6	94.3	4.2	0.8	2.2	3.2	0	1.6
URL	75	85.7	16	3.6	9	3.6	0	7.1
Организация	88	50.2	0	6	7.2	28.4	4.8	25.7
Адрес	74.9	56.4	2	8	8.9	18.8	17.4	21.4
Введение	99.3	94.6	0.7	5.4	0	0	0	0
Аннотация	85.3	56.5	0	0	8.6	29	6.4	20.9
Телефон	85.8	79	0	5.2	0	10.5	14.2	5.2
Страницы	100	97.9	0	2.1	0	0	0	0
Примечание	84.6	18.9	3.8	26.3	0	16.8	11.6	51.1

Таблица 4. Точность извлечения метаинформации на русскоязычном наборе данных.

Оценка / Класс	Извлечено			Не извлечено			Извлечено лишнее			Извлечено не все		
	SVM	*HMM*	*RE*	*SVM*	*HMM*	*RE*	*SVM*	*HMM*	*RE*	*SVM*	*HMM*	*RE*
Заглавие	88.2	75	84	1.9	6	5	5.8	11	11	3.9	12	9
Авторы	90	75	61	7.8	0	5	2.2	20	11	0	5	30
Дата	100	50	n/a	0	50	n/a	0	0	n/a	0	0	n/a
Email	97.3	96	n/a	2.7	0	n/a	0	0	n/a	0	4	n/a
URL	100	100	n/a	0	0	n/a	0	0	n/a	0	0	n/a
Организация	80	57	n/a	7.3	26	n/a	8	11	n/a	4.7	9	n/a
Город	66.6	80	n/a	22.2	0	n/a	11.1	20	n/a	0	5	n/a
Адрес	100	50	n/a	0	25	n/a	0	0	n/a	0	25	n/a
Введение	80	81	n/a	20	17	n/a	0	0	n/a	0	2	n/a
Аннотация	67	36	n/a	2.4	2	n/a	4.6	44	n/a	26	33	n/a

Таблица 5. Точность извлечения метаинформации на англоязычном наборе данных.

Оценка / Класс	Извлечено		Не извлечено		Извлечено лишнее		Извлечено не все	
	SVM	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*	*SVM*	*HMM*
Заглавие	72	53.8	6.9	8.1	10	27.8	12.2	15.6
Авторы	75.6	67.5	7	5.2	9.1	22.6	8.3	7.7
Дата	89.1	88.5	8	1.4	0	6.8	2.9	3.4
Email	88.2	93.2	3.6	0.4	2.7	1.3	6.3	5.1
URL	93.3	73.7	4.1	5.3	0	5.3	2.6	15.8
Организация	59	54.5	12	8.4	11.7	24.1	17.3	20.6
Адрес	73.2	58.8	12	5.6	7.5	19.9	7.3	23.1
Введение	95.2	92.3	4.7	7.7	0	0	0	0
Аннотация	74.3	50.7	0	0	10	33.3	15.7	21.4
Телефон	77.7	65.8	16.6	18.4	5.5	7.9	0	7.9
Страницы	100	98.6	0	1.4	0	0	0	0
Примечание	65	18.9	15	30.9	0	15.4	20	47.4

Таблица 6. Точность извлечения русскоязычной библиографии.

Оценка / Класс	Извлечено			Не извлечено			Извлечено лишнее			Извлечено не все		
	SVM	*HMM*	*RE*	*SVM*	*HMM*	*RE*	*SVM*	*HMM*	*RE*	*SVM*	*HMM*	*RE*
Ссылка	83.2	91.6	n/a	0	0	n/a	11.7	0	n/a	16.1	8.4	n/a
Заглавие	9.6	21.4	14	0	22.2	0	88.2	19.8	68	72.3	43.7	77
Авторы	14.9	37.5	18	0	9.8	0	62.9	42	73	81	13.4	59
Дата	61.3	90.8	n/a	3.8	4.6	n/a	33.8	4.6	n/a	29.2	0	n/a
URL	54.5	64.3	n/a	0	14.3	n/a	45.4	0	n/a	36.3	21.4	n/a
Страницы	25	66.7	n/a	25	16.7	n/a	55	13.3	n/a	65	0.3	n/a
Доп. информация	13.8	29.9	n/a	0	9.4	n/a	82.3	43.6	n/a	82.3	42.7	n/a