Извлечение метаинформации и библиографических ссылок из текстов русскоязычных научных статей

Козлов Дмитрий Дмитриевич
Факультет вычислительной математики и кибернетики
МГУ им. М.В. Ломоносова
Лаборатория вычислительных комплексов
ddk@cs.msu.su

Постановка задачи

Задача: в автоматическом режиме из текста статьи, представленного в виде PDF-файла, извлечь метаданные и библиографические ссылки.

Использование интеллектуальных сетевых роботов для построения тематических коллекций*

Романова Е.В., Романов М.В., Некрестьянов И.С.
Санкт-Петербургский Государственный Университет, Санкт-Петербург.
emails: katya@tepkom.ru, rmv@sparc.spb.su, igor@meta.math.spbu.ru

Abstract

В работе рассматривается задача создания интеллектуального сетевого робота для сбора тематических коллекций. Для повышения производительности обнаружения тематических ресурсов используется специализированный алгоритм обхода сети, учитывающий информацию о тематической релевантности уже посещенных страниц. Робот также производит грубый отсев "мусора" среди посещенных документов, для того чтобы повысить качество рекомендаций.

1 Введение

В течение ряда лет вопросы создания и применения сетевых роботов привлекают все больше внимания [8, 10, 13, 11]. Сетевой робот или *Crawler* — это программа, которая, начиная с некоторой Интернет-страницы, рекурсивно обходит ресурсы Интернет, извлекая ссылки на новые ресурсы из получаемых документов.

Классической областью применения сетевых роботов является построение индексов Интернет-ресурсов для поисковых систем [14, 3, 5, 15]. Однако в последнее время сетевые роботы используются для выполнения множества других задач — сбора статистики, поиска определенных ресурсов сети (например, домашних страниц), проверки целостности существующих гипертекстовых ссылок, и т.п. Разработаны даже соответствующие правила "вежливого" поведения для сетевых роботов — Standard for Robot Exclusion и Rapid Fire Requests. Текущий вариант списка добровольно зарегистрированных роботов на странице `info.webcrawler.com` содержит более сотни позиций, а общее число существующих сетевых роботов по некоторым оценкам превышает десятки тысяч.

Большинство сетевых роботов посещают огромное количество Internet страниц, индексируя все полученные документы. Очевидно, что такой подход требует значительных сетевых и аппаратных ресурсов. Однако текущий объем доступной информации в Интернет оценивается в 6 терабайт и быстро растет, поэтому даже самый мощный сетевой робот не может посетить все Интернет-страницы.

Поскольку посещение всех Интернет-страниц не представляется возможным, то разумно посещать в первую очередь наиболее важные из них. Простейший критерий важности, используемый многими из современных сетевых роботов собирающими информацию для популярных поисковых систем, является глубина URL, т.е. количество промежуточных каталогов упоминающихся в URL между именем Интернет-узла и именем самого ресурса. Чем больше глубина, тем ниже важность соответствующего ресурса. Подобный подход позволяет быстро посетить стартовые и близкие к ним страницы на большом числе Интернет-узлов.

7 Заключение

В работе рассматривается задача создания интеллектуального сетевого робота для сбора тематических коллекций.

Описана базовая архитектура системы, структура тематического фильтра и методы оценки тематической релевантности документа. Использование дополнительной информации от клиента робота во время работы для уточнения тематического фильтра позволяет улучшать качество оценок в процессе работы. Описываемая стратегия обхода сети учитывает тематические оценки уже посещенных документов, что позволяет посетить тематически релевантные документы в первую очередь.

Предварительные результаты экспериментов показывают преимущество тематически-ориентированной стратегии обхода над другими стратегиями для сбора тематических коллекций. Все это подтверждает перспективность предлагаемого подхода.

Отметим, что проблема построения тематических коллекций не является специфичной для проекта OASIS и актуальна во многих других задачах информационного рынка, например, таких как построение тематических

Библиография

[1] I.J. Aalbersberg. Incremental relevance feedback. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–22, 1992.

*Эта работа была выполнена в рамках проекта Open Architecture Server for Information Search and Delivery (OASIS) и поддержана грантом Европейской комиссии (INCO Copernicus Programme Project PL 961116).

Метаинформация:

- *Название:* Использование интеллектуальных сетевых роботов для построения тематических коллекций.
- *Авторы:* Романова Е.В., Романов М. В., Некрестьянов И.С.
- *Аннотация:* В работе рассматривается…

Библиографические ссылки:

1. *Автор:* Aalberg I.J. *Название:* Incremental relevance feedback. *Год:* 1992, *Издание:* ACM SIGIR…

Практическая ценность задачи

Рассматриваемая задача актуальна для построения графа взаимного цитирования.

Автоматическое построение графа взаимного цитирования состоит из двух задач:

- извлечение метаинформации и библиографических ссылок,
- сопоставление библиографических ссылок.

Рассматриваемая задача

Построение графа взаимного цитирования позволяет

- вычислять индекс научного цитирования,
- осуществлять поиск научных статей путем навигации по библиографическим ссылкам,
- применять методы тематического поиска научных статей, использующие структуру графа взаимного цитирования.

Особенности задачи

- Авторы не снабжают тексты статей метаинформацией в удобной для автоматического разбора форме => требуется извлечение метаинформации из текстов статей.
- Необходимо обработать большое количество статей (десятки-сотни тысяч) => ручная обработка невозможна.
- Нерегулярность структуры русскоязычных статей:

  - для русскоязычных статей нет общепринятых норм структурирования статей (для англоязычных статей такие нормы существуют);
  - в русскоязычных статьях нет единого стиля оформления статей и библиографических ссылок. Оформление статей существенно различается;
  - библиографические ссылки часто задаются неточно, с ошибками.

Особенности задачи (2)

Извлечение библиографических ссылок

Самусев С. Шамина О.
ВМиК МГУ
{sam,sincere}@lvk.cs.msu.su

Аннотация

В данной работе …

1 Введение

…

Литература

[1] Freitag D., McCallum A. Information extraction with HMMs and shrinkage. Proceedings of the AAAI-99 Workshop on Machine Learning for Informatino Extraction, 1999.

[2] …

5

Существующие подходы

Методы, применявшиеся для англоязычных статей

Методы, основанные на правилах:

- Метод, основанный на регулярных выражениях (Lawrence, 1999)
- Метод, основанный на шаблонах (Chowdhury, 1999)

Методы машинного обучения:

- Методы, основанные на вероятностных конечных автоматах:
  - Скрытые марковские модели (Freitag&McCallum, 1999).
  - Марковские модели максимальной энтропии (McCallum, 2000).
  - Условные случайные поля (Lafferty&McCallum, 2001).
- Метод, основанный на классификации SVM (C. Lee Giles, 2003).

Цель работы

Цель работы:

исследование применимости существующих методов, разработанных для англоязычных статей, для извлечения метаинформации и библиографических ссылок из текстов русскоязычных научных статей.

Методы, охваченные в данной работе:

- метод, основанный на регулярных выражениях.
- метод, основанный на скрытых марковских моделях.
- метод, основанный на классификации с помощью метода опорных векторов.

Этапы решения задачи

Этап 1: преобразование текста статьи в формате PDF в промежуточное текстовое представление с сохранением дополнительной разметки:

- окончаний строк,
- изменений размера шрифта,
- отступов строки от края страницы.

Этап 2: извлечение метаинформации и библиографических ссылок из промежуточного текстового представления с помощью одного из методов:

- метода, основанного на регулярных выражениях;
- метода, основанного на скрытых марковских моделях;
- метода, основанного на классификации.

Метод, основанный на регулярных выражениях

1. Из промежуточного представления текста статьи извлекается первая страница или текст до заголовка «Введение».
2. С помощью построенной вручную системы правил извлекается метаинформация. Пример правила:

   Если на предыдущем шаге список авторов найден не был, то в первых пяти строках текста ищется строка, которой соответствует максимальный размер шрифта. Выбранная строка рассматривается в качестве возможного заголовка на следующем шаге.
3. От конца статьи к началу осуществляется поиск заголовка «Литература» (с вариациями, например, «Список литературы» и т.п.)
4. С помощью вручную построенной системы правил разбираются библиографические ссылки.

Методы машинного обучения: предобработка

Упрощенный вариант метода Rule-Based Word Clustering (Giles, 2005):

- Слова в тексте статьи заменяются на признаки.
- Правила замены определяются по соответствию слова словарю или заданному в виде регулярного выражения шаблону.
- Слово заменяется на наиболее специфичный признак.

Примеры признаков:

- :email: - по соответствию регулярному выражению
- :country: - название страны, определяется по словарю
- :dictWord: - словарное слово
- :Cap1DictWord: - словарное слово, написанное с заглавной буквы
- :mayName: - слово из словаря имен

Методы машинного обучения: предобработка (2)

Использование интеллектуальных сетевых роботов для построения тематических коллекций

Романова Е.В., Некрестьянов И.С.
Санкт-Петербургский Государственный Университет, Санкт-Петербург.
emails: katya@tepkom.ru, igor@meta.math.spbu.ru

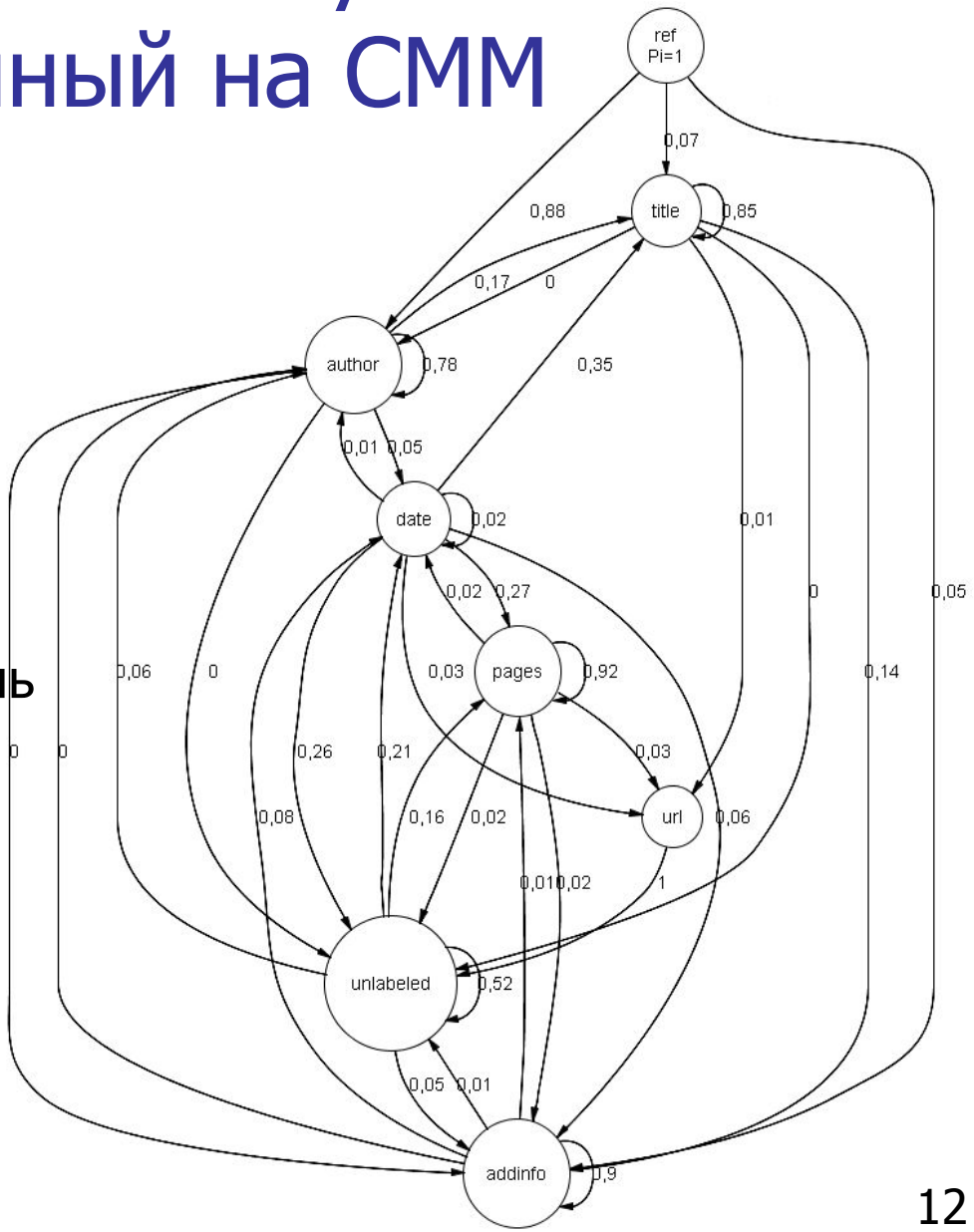
Abstract:
В работе рассматривается задача создания ...

:Cap1DictWord: :DictWord: :DictWord: :DictWord: :DictWord: :DictWord: :DictWord: :DictWord:
:mayName: :singleCap: :singleCap:, :mayName: :singleCap: :singleCap: :city: :Cap1DictWord: :affi: :city:.
:DictWord: : :email:, :email:, :email:

:abstract:
:singleCap: :DictWord: :DictWord: :DictWord: :DictWord: …

Методы машинного обучения: метод, основанный на СММ

- Состояния соответствуют элементам метаинформации.
- Наблюдаемая цепочка – последовательность признаков после предобработки.
- В режиме распознавания модель по заданной наблюдаемой последовательности восстанавливает цепочку состояний, т.е. каждому признаку сопоставляет класс метаинформации.



Методы машинного обучения: метод , основанный на классификации

Задача извлечения метаинформации рассматривается как задача классификации строк статьи:

- Для каждого класса метаинформации (Title, Author, Affiliation, Address, Email, Date и т.д.) строится бинарный классификатор, использующий метод опорных векторов и стратегию «один против всех».
- Контекстно-независимая классификация:
  Строка представляется в виде набора признаков, основанных на свойствах слов (признаки, получены в результате предобработки).
- Каждая строка классифицируется всеми классификаторами.

Методы машинного обучения: метод, основанный на классификации (2)

Осуществляется второй шаг классификации - контекстно-зависимая классификация:

- Строка представляется в виде **расширенного** набора признаков: добавляются метки классов соседних строк и признаки, основанные на свойствах строки (ее номер, количество слов того или иного типа и т.д.).
- Для каждого класса метаинформации строятся контекстно-зависимые классификаторы и производится второй шаг классификации.

Методы машинного обучения: метод, основанный на классификации (3)

95% строк принадлежат к одному классу, остальные – к нескольким (4,5% - к двум, 0,5% - к трем и более).

Разделение строк, относящихся к нескольким классам:

- Поиск оптимальной границы (пробела или знака препинания), разделяющей строку на две части, каждая из которых относится к одному классу:

  max((P1 – P2)* (N2 – N1)), где

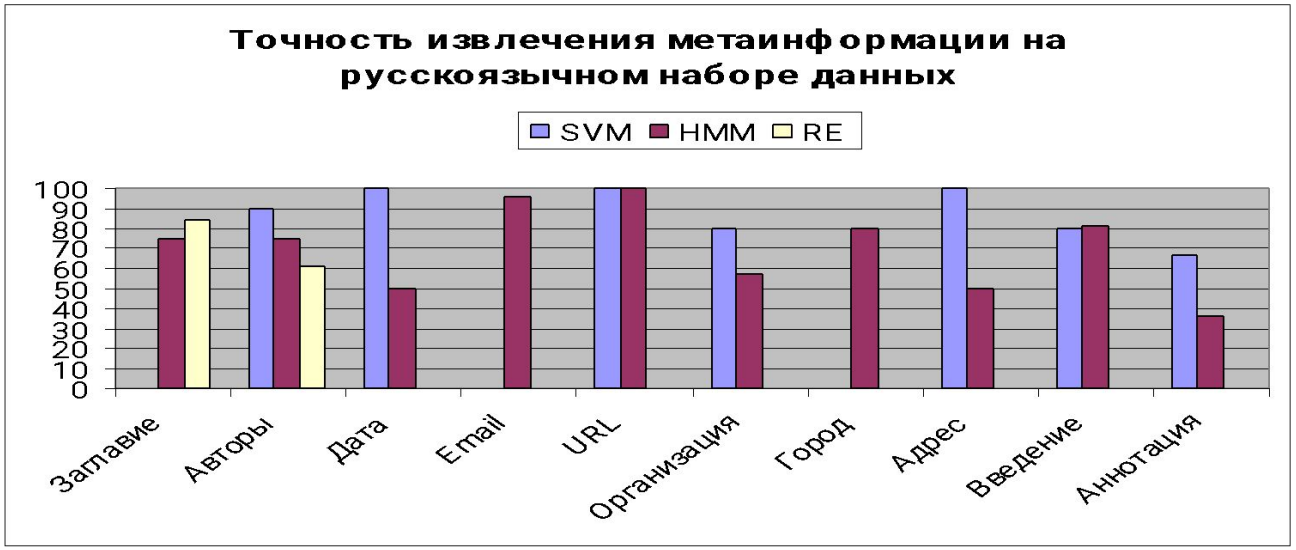
  P1 – оценка части P классификатором 1;
  P2 – оценка части P классификатором 2;
  N1 – оценка части N классификатором 1;
  N2 – оценка части N классификатором 2;

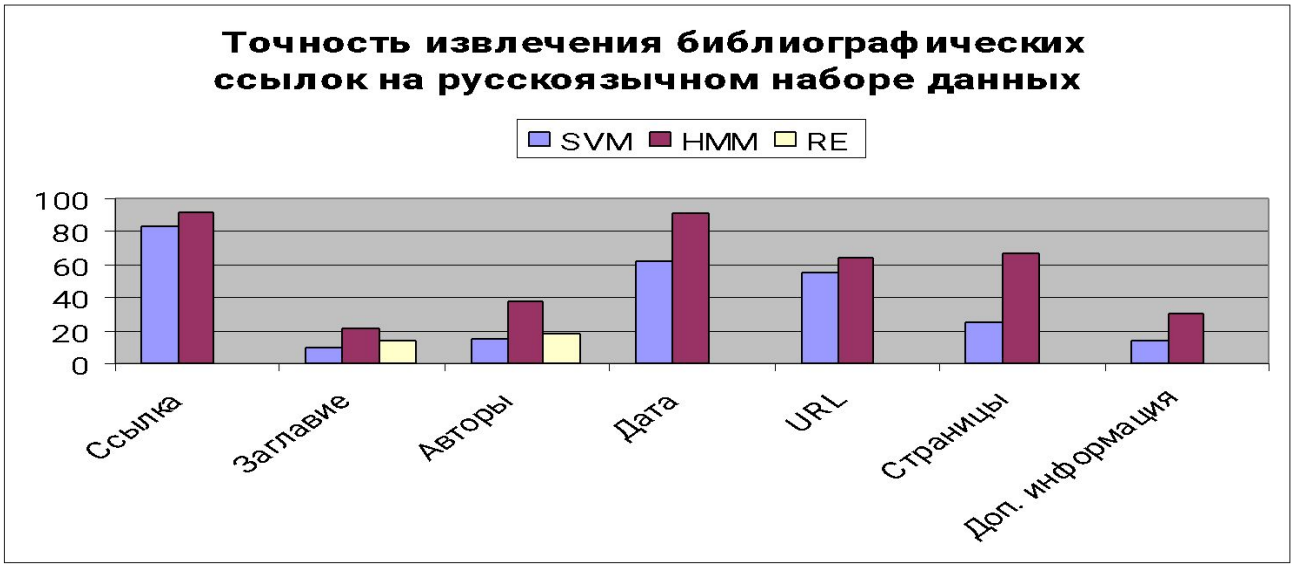
- Случай трех и более классов сводится к последовательному применению метода для двух классов.

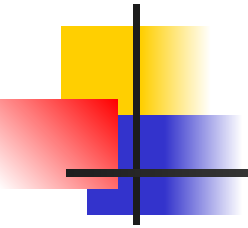
Экспериментальное исследование

- Цель: сравнение точности методов.
- Наборы данных:
  - англоязычный (McCallum, 935 заголовков, 500 библиографических ссылок).
  - русскоязычный (материалы конференций и семинаров ММРО, РОМИП, Диалог, Интернет-математика, публикации с graphics.cs.msu.su, 180 заголовков, 1000 библиографических ссылок).
- Четыре варианта оценки: извлечено правильно (1), извлечено не все (0), извлечено лишнее (0), не извлечено (0).

Экспериментальное исследование







Выводы

- Экспериментальное исследование показало, что все три метода обеспечивают точность порядка 70-80%, что является пригодным для практического использования.
- Результаты на русскоязычных данных существенно хуже, чем на англоязычных.
- Метод, основанный на скрытых марковских моделях наиболее успешно работает для извлечения библиографических ссылок. Следовательно, возможно совместно применять несколько методов с учетом их специализации.

Планы дальнейшего развития

- Повышение точности рассмотренных методов машинного обучения за счет учета разметки.
- Использование условных случайных полей для устранения недостатков метода скрытых марковских моделей.
- Повышение точности за счет совместного использования нескольких методов.
- Автоматическое обнаружение возможных ошибок извлечения для передачи на ручную обработку.

http://lvk.cs.msu.su

Спасибо за внимание

Козлов Дмитрий Дмитриевич

Факультет вычислительной математики и кибернетики
МГУ им. М.В. Ломоносова

Лаборатория вычислительных комплексов

ddk@cs.msu.su