1. Автоматизация процесса извлечения метаданных из слабоструктурированных документов.

2.Автоматизация процесса получения метаданных документа с использованием удаленных описаний.

3.Автоматическое извлечение из текстов ключевых слов.

Выполнили: Копылова Юлия,
Лагуткина Яна, Ковалев Егор,
Васильев Николай, Леонтьев Артем

# 1. Автоматизация процесса извлечения метаданных из слабоструктурированных документов.

- **Метаданные** – это структурированные, кодированные данные, которые описывают характеристики объектов-носителей информации, способствующие идентификации, обнаружению, оценке и управлению этими объектами
- **Слабоструктурированные документы(данные) -** это форма структурированных данных, не соответствующая строгой структуре таблиц и отношений в моделях реляционных баз данных.
- **XML** и другие языки разметки, **email** и сообщения в форматах **EDI** —примеры слабоструктурированных данных.

Первоначальный этап научно-информационного процесса с участием электронных документов включает в себя их сбор и каталогизацию, сводящуюся к извлечению из документов их метаданных

Полноценное удовлетворение информационных потребностей пользователя возможно лишь при каталогизации отдельных документов, в частности, статей.

Трудоёмкость процесса извлечения метаданных из документов приводит к необходимости его частичной автоматизации. Основные сложности при решении этой задачи состоят в разработке алгоритма, позволяющего в автоматизированном режиме извлекать из слабоструктурированного документа основные элементы его библиографического описания.

Алгоритм, основанный на типичном для интеллектуальных информационных систем человеко-машинном взаимодействии, сводится к выполнению последовательных операций:

- Создание шаблона для обрабатываемого сайта;
- Создание списка адресов, где расположены документы;
- Обработка документов, включая возможное извлечение метаданных из удалённых библиографических источников;
- Поддержание актуальности информации

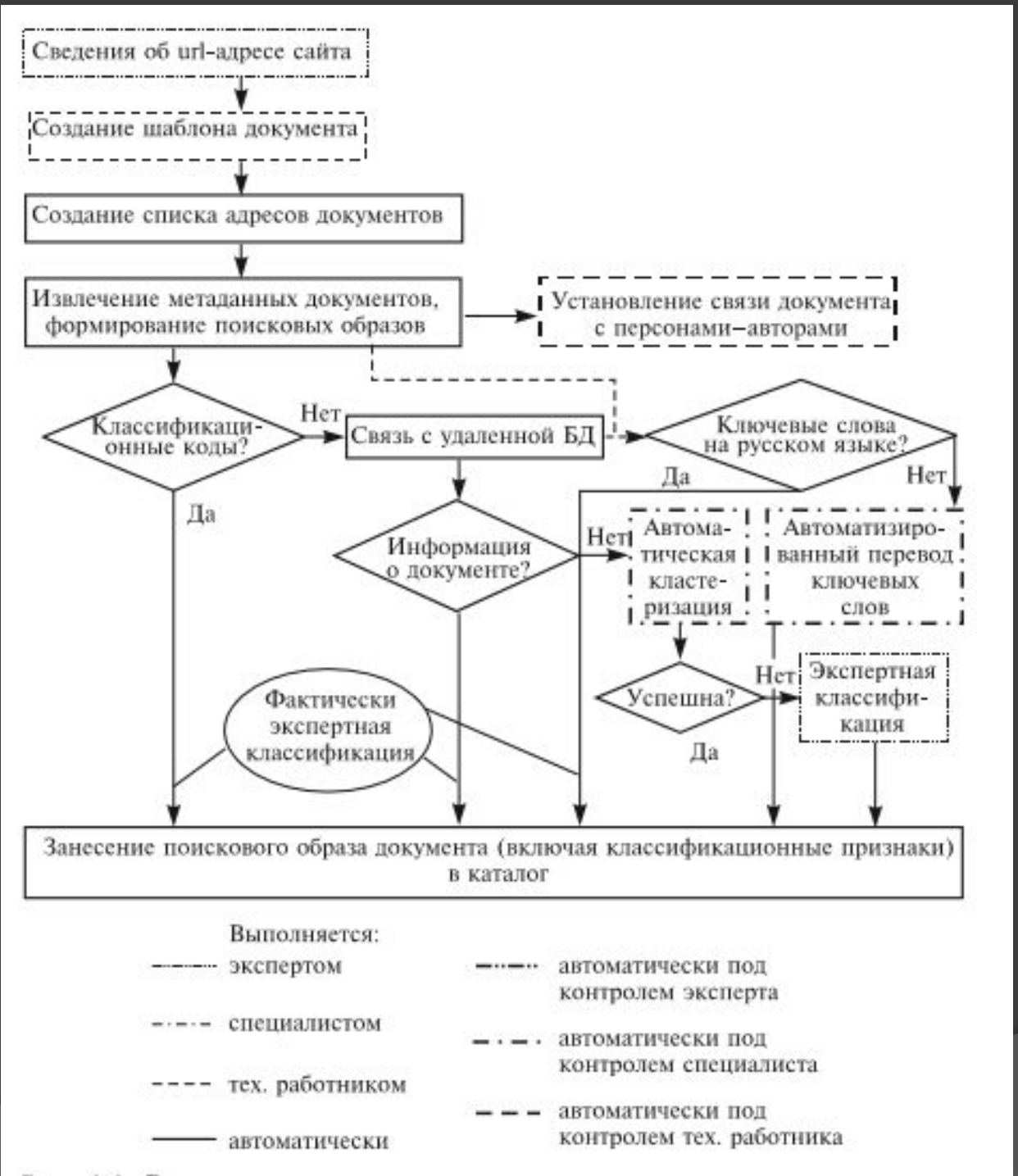



Схема алгоритма автоматизированного извлечения метаданных из слабоструктурированных документов

Шаблон документа необходим для автоматического выделения его основных метаданных. Пример – для документа – журнальной статьи метаданные включают в себя элементы описания: авторы, названия, ключевые слова, аннотация и т.д. В каждом конкретном случае шаблон создаётся сравнительно легко с использованием языка регулярных выражений в том или ином формате, но проблема в том, что разные сайты имеют сильно различающуюся структуру описания и представления документов. Поэтому необходимо создание алгоритма в **виде веб-приложения**, позволяющего пользователю, даже не владеющему языками обработки регулярных выражений, генерировать шаблоны для различных сайтов.

В общем случае пользователю-каталогизатору достаточно занести в поля формы теги, окружающие в HTML-коде статьи значения каждого из элементов метаданных одного документа обрабатываемого сайта, а также указать разделитель данных в случае множественности некоторого элемента метаданных, после чего создаётся и сохраняется шаблон веб-страницы документов сайта. Надёжность алгоритма особенно высока в случае автоматической вёрстки веб-страниц каталогизируемого сайта, когда их структура практически одинакова

Название шаблона

neva Конец

| Статическое | Название поля | Начало | Конец | Регулярное выражение |
|---|---|---|---|---|
| ☐ | name | <body>\s*<h2> | | [*) |
| ☐ | authon | <h2>\s*<h2> | <\h2> | ([*<]*) |
| ☐ | referat | <\h3> | <h3> | [*) |
| ☐ | link | <AHREF> | >FOF<\A> | ([*<]*) |

☐ Множественные данные

Разделитель данных

Сохранить Удалить

Пример создания шаблона документа.

# 2.Автоматизация процесса получения метаданных документа с использованием библиографических удаленных описаний.

Без классификационных признаков метаданных ценность каталожного описания документа минимальна, так как процесс поиска документа человеком или его обработка рассуждающей информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа.

Качественно решить задачу классификации может эксперт, поэтому прежде всего следует проверить, не внесена ли информация о полиграфической версии документа в ту или иную электронную библиографическую базу данных удалённого доступа, где документы классифицированы в соответствии с нужным классификатором. Исходя из особенностей описания документа, в той или иной конкретной базе, к ней формируется запрос, содержащий необходимые сведения о классифицируемом документе.

Полная репликация метаданных документа из библиографической базы не может служить эффективной заменой процессу непосредственного извлечения метаданных из интернет-документа, поскольку в большинстве случаев библиографические базы не содержат сведений об url-адресе электронной версии документа.

Процесс определения метаданных документа с использованием удаленной библиографической базы может быть частично автоматизирован.

Для обращения к этой базе данных с целью получения классификационных признаков документа автоматически формируется строка запроса к серверу библиографической базы, использующая в качестве параметров запроса уже извлеченные с веб-страницы журнала библиографические данные. При наличии сведений сервер выдает страницу с его описанием, на которой присутствуют, среди прочих библиографических данных, и классификационные признаки документа. Извлечение недостающих метаданных документа, производится по стандартному шаблону с помощью регулярных выражений.

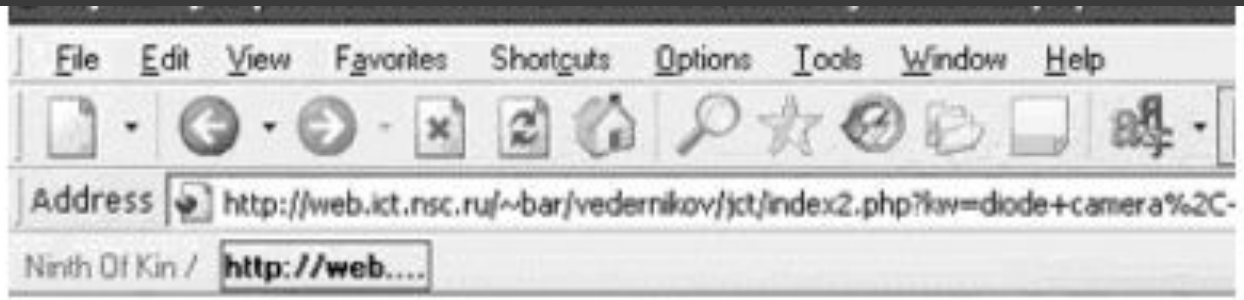
После получения ключевых слов документа из англоязычной библиографической базы данных может возникнуть проблема их перевода на русский язык. Прежде всего следует проверять наличие переводимого термина в англоязычной части тезауруса (онтологии) предметной области. Процесс перевода отсутствующих в тезаурусе терминов может быть частично автоматизирован с использованием словарей, доступных через Интернет. Также автоматически формируется строка к удаленному словарю с последующей обработкой результатов запроса.

Алгоритм включает в себя следующие этапы:

1. После получения списка ключевых слов для перевода программа разбивает этот список на отдельные словосочетания.
2. Каждое словосочетание из списка программа сначала пытается найти в англоязычной части тезауруса.
3. Если словосочетание в тезаурусе не нашлось, то делается запрос к удаленному словарю.
4. Если в удаленном словаре не удалось найти перевод словосочетания целиком, то оно разбивается на отдельные слова и для каждого из них выполняются вышеперечисленные действия. Потом из переводов отдельных слов вновь составляется словосочетание
5. После того как все словосочетания переведены, пользователю предлагается скорректировать переводы.
6. Ключевые слова заносятся в метаописание документа и при наличии русских соответствий в тезаурусе в его англоязычную часть.

Таким образом, происходит процесс обучения системы: чем больше слов и словосочетаний переведено, тем меньше программа обращается к удаленному словарю через Интернет, так как уже переведенные слова и словосочетания заносятся в тезаурус.

# Корректировка перевода ключевых слов




diode camera: не удалось найти перевод фразы целиком
diode: диод

camera: фотоаппарат

---

grid: решетка

---

probability function: вероятностная функция

---

Слова для занесения в базу:

диод, фотоаппарат, решетка, вероятностная функция

# 3.Автоматическое извлечение из текстов ключевых слов.

Важной задачей обработки текстовых документов, без решения которой практически невозможна автоматизация процесса извлечения из них информации и знаний, является координатное индексирование, т.е. извлечение из текстов ключевых слов (всех содержащихся в индексируемом тексте терминов, входящих в словарь онтологии данной предметной области).

Координатное индексирование документов может производиться автоматически, поскольку оно дает почти такие же результаты, как и ручное, но имеет перед ним ряд преимуществ:

- Обеспечивает единообразие индексирования, почти невозможное для человеческого интеллекта;
- Обходится, по крайней мере, в 3 раза дешевле.

Так как в русском языке имена существительные и прилагательные при склонении меняют свою форму, необходимо учитывать те случаи, когда слова, образующие термин, находятся не только в именительном, но и в косвенных падежах при разработке эффективного алгоритма автоматизации извлечения ключевых слов. В основу алгоритма положено использование двух индексов, содержащих триады «номер текста» - «позиция в тексте» - «номер слова из лексического словаря» и «номер термина» - «позиция в термине» - «номер слова из лексического словаря». Первый индекс встречается практически во всех информационно-поисковых системах, а второй носит оригинальный характер и позволяет резко повысить эффективность алгоритма. Индекс терминов и их список размещаются в хранилище данных программной библиотеки, реализующей алгоритм, и пополняется по мере изменения этого списка.

Алгоритм построения индекса терминов состоит из следующих этапов:

1. Разбиение термина на отдельные слова.
2. Создание предварительного индекса, содержащего триады «номер термина» «позиция слова в термине» - «слово в символьном представлении»
3. Добавление встретившихся неизвестных слов в лексический словарь библиотеки, где им присваиваются идентификационные номера
4. Переработка индекса в формат «номер термина» – «позиция в тексте» - «номер слова из лексического словаря»
5. Сбор статистики о длинах терминов для реализации поиска и идентификации составных терминов (т.е. терминов, состоящих более чем из одного слова)
6. Сбор статистики о количестве вхождений отдельных слов в термины для оптимизации поиска путем исключения из рассмотрения терминов, заведомо отсутствующих в тексте.

Алгоритм построения индексов текста аналогичен, но в нем отсутствует этап 3.

Заключительная стадия работы программной библиотеки – подсчет количества вхождений терминов в текст (тексты) . Её этапы:

1. Подсчет возможных комбинаций «текст» - «термин», основанный на статистике вхождения отдельных слов
2. Нахождение всех потенциально возможных мест вхождения каждого термина в текст на основе наличия хотя бы одного общего слова из лексического словаря. Позиция каждого потенциально возможного вхождения фиксируется
3. Рассмотрение каждого из возможных мест вхождений с точки зрения соответствия термину в целом.
4. Исключение учета вхождений, поглощаемых более длинными вхождениями.
5. Сбор статистики вхождений для каждой пары «текст» – «термин»