МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА
Факультет вычислительной математики и кибернетики
Кафедра системного программирования

Выделение семантических связей в электронной библиотеке публикаций

Курсовая работа
Студента 1 курса магистратуры
Шабалова С.А.

Научный руководитель:
профессор кафедры СП ВМК МГУ,
заведующий отделом ВЦ РАН,
доктор физико-математических наук Серебряков В.А.

Москва-2019

Оглавление

1. Введение

В данной курсовой работе рассматривается проблема выделения семантических связей в электронной библиотеке публикаций.

Актуальность данной проблемы заключается в том, что в последние годы в сфере информационных технологий наметилась тенденция на работу с не просто некоторыми объемами данных, удобными для прочтения человеком, а с данными, пригодными для машинной обработки: автоматического анализа, синтеза выводов, преобразования как самих данных, так и сделанных на их основе заключений, и множества других применений. Подобные данные описывают семантическую структуру в виде графа, что позволяет использовать множество уже существующих алгоритмов для обработки. Технологии такие как XML, RDF и OWL, необходимые для перехода на описанный вид данных, разрабатываются начиная с конца 1990-х – начала 2000-х годов, поддерживаются консорциумом W3C, множеством компаний и на сегодняшний день достигли существенного уровня развития.

В процессе машинной обработки в данных могут быть выделены связи, облегчающие работу человека с ними, например, навигацию между документами. Но сама задача выделения этих связей нетривиальна для машинной обработки, если данные изначально не были приведены в пригодный формат.

2. Общая постановка задачи

2.1. Определения

Электронная библиотека публикаций – информационная система, ориентированная на действия (поиск, доступ и т.д.) с публикациями в цифровом формате.

Семантическая связь – смысловая связь. Термин *семантический* стал активно использоваться в контексте «семантический WEB» в противовес «несемантический WEB», основанному на гиперссылках. Фактически сегодня «семантической моделью WEB» имеется в виду использование RDF модели для представления информции.

Семантическая электронная библиотека – электронная библиотека, ориентированная на интеграцию в Semantic Web.

Единое Научное Информационное Пространство РАН (ЕНИП) - информационное пространство, ориентированного прежде всего на научного сотрудника РАН как потребителя, а с другой – ограниченного информацией, порождаемой и циркулирующей прежде всего в РАН.

2.2. Постановка задачи

Цель работы:

разработка методов выделения семантических связей между публикациями в электронной библиотеке в рамках ЕНИП

Этапы работы:

1. Предобработка исходных данных
2. Разработка модели хранения промежуточных данных
3. Разработка и реализация методов выделения семантических связей

3. Обзор существующих решений

Рассмотрим ряд работ, связанных с семантическими электронными библиотеками.

Публикации (2) и (3) посвящены семантическому структурированию контента научных электронных библиотек на примере научно-образовательной электронной библиотеки Соционет, основанной на технологии открытых архивов и содержащей информационные ресурсы социально-экономической тематики. Эта библиотека функционирует уже более десяти лет и приобрела в последние годы статус де-факто институциональной электронной библиотеки Отделения общественных наук РАН. Основная часть работы заключается в структурировании и классификации семантических связей между научными публикациями библиотеки с целью их последующего использования пользователями Соционет. При этом строгость формата научных публикаций облегчает выделение самих связей, т.к. могут быть указаны исследователи, их организации, ГРНТИ, ключевые слова, цитирования и т.д. Также в платформу Соционет была добавлена возможность добавления пользовательских семантических связей. Таким образом, эти публикации имеют большую ценность для этапа, следующего за выделением семантических связей.

Также была рассмотрена работы, посвященные проблеме автоматического выделения ключевых слов в тексте. Доклад (4) посвящен модификации алгоритма KEA (Key Extraction Algorithm), используемого для извлечения ключевых слов и словосочетаний. KEA широко известен своей эффективностью для извлечения ключевых слов и словосочетаний из англоязычных текстов. В этой публикации рассматривается применения данного алгоритма к текстам на русском языке, которое может быть полезно в процессе выделения семантических связей.

4. Исходные данные

4.1. Единое Научное Информационное Пространство

Единое Научное Информационное Пространство РАН (ЕНИП) – информационное пространство, ориентированного прежде всего на научного сотрудника РАН как потребителя, а с другой – ограниченного информацией, порождаемой и циркулирующей прежде всего в РАН. Предпосылки ее создания:

- Накопившийся огромный объем научной информации в электронном виде в различных отраслях науки;
- Осознанная потребность научных сотрудников в необходимости как поиска качественной информации, так и в выставлении собственной информации в сетью;
- Осознанная потребность научных сотрудников в необходимости приведения имеющихся у них накопившихся массивов унаследованной информации к каким-либо стандартам (желательно международным);
- Осознание административным уровнем управления наукой в РАН критической необходимости наведения информационного порядка в РАН как организации для сохранения возможности управления.

В ЕНИП определено несколько уровней поддержки схем:

- *Минимальная* – необходимый разумный минимум, минимально достаточный для обмена метаданными, поддержки взаимосвязей ресурсов;
- *Базовая* – объем достаточный для эффективной работы «дилетантов» в конкретной предметной области;
- *Расширенная* – объем достаточный для основной работы «специалистов» предметной подобласти;
- *Специализированная* – объем, существенно ориентированный на специалистов предметной области, используется только в рамках подпространства, включающего специализированные системы.

Также в ЕНИП выделен основной профиль, включающий общеприменимые и первоочередные предметные области, и специфицирован механизм расширения стандарта дополнительными специализированными профилями, ориентированными на использование в специализированных научных сообществах.

4.2. Научное наследие России

Исследование в рамках данной работы проводится на примере электронной библиотека «Научное наследие России» (далее – «Наследие»), которая разрабатывается в рамках одноименной программы Президиума РАН с целью обеспечения сохранности и предоставления публичного доступа к научным трудам известных российских и зарубежных ученых и исследователей, работавших на территории России. В настоящее время на сайте библиотеки (http://e-heritage.ru) доступно порядка 28 тысяч публикаций, разделенные на рубрики на базе классификатора ГРНТИ.

Второй важной задачей этой библиотеки является интеграция существующих библиотечных ресурсов в ЕНИП. На данный момент все доступные на «Наследии» публикации имеют метаданные в формате RDF, совместимые с ЕНИП. Пример такого описания в приложении А. Также для большей части библиотечных ресурсов доступны оглавления (приложение В).

4.3. Используемые данные

Из-за технических сложностей работа проводилась на отдельной версии «Наследия», ранее доступной на сайте http://dev.e-heritage.ru, где хранилось только 18 тысяч публикаций. После полной загрузки выяснилось, что:

- У 6.6 тысяч публикаций отсутствуют оглавления;
- 1.5 тысячи публикаций имеют оглавление на русском языке в петровском правописании;
- Около 700 публикации имеют оглавление на западно-европейский языках.

  Примеры 2 и 3 пунктов в приложении С.

Эти проблемы вызвали сильное усложнение предобработки данных и замедлили ход работы.

5. Модель хранения данных

В ходе работы было рассмотрено 2 способа хранения промежуточных данных: JSON и реляционная модель. Была выбрана реляционная модель с использованием СУБД SQLite, т.к.:

- Все данные хранятся в одном файле; при использовании JSON по ходу работы было бы необходимо либо добавлять новые файлы, либо постепенно усложнять структуру единственного JSON-файла;
- Возможности языка SQL позволяют оперативно отслеживать обновленные промежуточные данные.

Модель БД описана в приложении D.

6. Построение решения

Одним из методов построения семантических связей является выделение ключевых слов и словосочетаний в исходных текстах с последующей проставкой связей между ресурсами. Задача выделения ключевых слов хорошо исследована, но обычно рассматривается применение к полноценным текстам, а исходные данные этой работы состоят из метаданных публикации и оглавления. Их отличительной чертой является повышенная плотность ключевых слов и словосочетаний в тексте.

Детализируем решение:

1. Предобработка исходных данных с учетом требования на дальнейшее выделение ключевых слов выбранным методом;
2. Выделение ключевых слов, их ранжирование;
3. Разработка методов простановки связей между документами

7. Предобработка исходных данных

В связи с тем, что исходные данные «Наследия» изначально не были готовы к какой-либо текстовой обработке из-за проблем, описанных в разделе 4.3, предобработка происходила следующим образом:

1. Классификация оглавлений публикаций по их языку;
2. Автоматический перевод оглавлений по необходимости;
3. Удаление стоп-слов, нормализация.

7.1. Классификация по языку

В процессе классификации оглавлений публикации были разделены на следующие категории:

- русский язык, новое правописание;
- русский язык, петровское правописание;
- один из западно-европейских языков.

Русский язык в петровском правописании был выделен отдельно, т.к. популярные средства машинного перевода его не поддерживают, в отличие от западно-европейских языков. Отметим, что в RDF описаниях публикации на петровском правописании язык было отмечено, что они на русском языке (isir:publicationLanguage).

Классификация происходила по следующему алгоритму:

1. Если в тексте нет букв из обычного русского алфавита, то относим к западно-европейским языкам
2. Если в тексте есть буквы из обычного русского алфавита, то:

   2.1. Если в тексте нет букв ‘ѣ’, ‘ѵ’, ‘ѳ’ в любом регистре и нет буквы ‘i’ в нижнем регистре, то относим к русскому языку в новом правописании;

   2.2. Иначе относим к руссому языку в петровском правописание

Этот алгоритм работает на исходных данных, где в названиях оглавлений встречаются римские цифры, обозначающие номер главы, поэтому был выделено условие про букву ‘i’ в нижнем регистре. Также она не встречается в качестве первой буквы слова в начале предложения, поэтому ложно отрицательных результатов нет. Отметим недостаток, что данный классификатор определяет белорусский и украинский языки как русский в петровском

правописании, но тексты на этих языках не были обнаружены при поверхностном изучении исходных данных.

7.2. Перевод оглавлений

Для перевода текстов на западно-европейских языках было использовано API Яндекс.Переводчик. Он был выбран т.к. имеет достаточно большой для данной задачи лимит запросов, не требующий оплаты, простой в использовании и имеет высокую производительность.

Для перевода русского языка из петровского правописания в новое был использован бесплатный сервис «Славеница» (http://slavenica.com/). Других средств для автоматического решения этой задачи в свободном доступе не было найдено. К сожалению, «Славеница» не имеет API, процесс перевода производится на их сервере, а запросы, передаваемые на сервер, обфусцированы и хранят переводимый текст в неизвестном формате. Процесс кодирования в этот формат производится также обфусцированным кодом страницы сайта, поэтому задачи перевода происходила при помощи популярного инструмента Selenium, используемого для автоматизации действий веб-браузера. Такой подход имеет крайне низкую производительность, но он был наиболее прост в реализации. Другие рассматриваемые подходы:

- Полноценный реверс-инжиниринг кода страницы «Славеницы» для автоматизации запросов с серверу переводчика;
- Реализации собственного переводчика из петровского правописания в новое.

7.3. Нормализация

Процесс нормализации происходил следующим образом:

1. Перевод текста в нижний регистр;
2. Удаление стоп-слов по списку, пунктуации, отдельно стоящих букв, чисел записанных как арабскими, так и римскими цифрами;
3. Приведение каждого слова в нормальную форму при помощи библиотеки pymorphy.

Процедура нормализации едина для всех текстов, которые используются в дальнейшем процессе выделения ключевых слов. Нормализованные тексты сохраняются в БД для дальнейших экспериментов. Отметим, что немалая часть оглавлений содержат только номера

глав и синоним слова «глава», поэтому в изначальный список стоп-слов были добавлены найденные в текстах при поверхностном изучении подобные синонимы.

8. Выделение ключевых слов

Изучив все данные, которые были получены для отдельно взятого книжного ресурса, было решено обрабатывать название публикации, оглавление и понятие рубрикатора ГРНТИ, к которому отнесен ресурс.

Изначальный подход к выделению ключевых слов заключался в выделении всех одиночных слов и ранжирование их по мере TF-IDF с последующей фильтрацией. Результаты такого подхода оказались неудовлетворительны, т.к. то, что было отмечено как ключевые слова, никак не соотносилось со смыслом исходных текстов. Отметим, что подобный результат ожидался изначально, но этот подход был проверен из-за повышенной плотности ключевых слов в названиях публикации и оглавлениях.

Возникло предположение, что необходимо привязать некоторый фиксированный набор понятий, слова и словосочетания из которого можно будет искать в исходных данных. В качестве такого набора был выбран набор заголовков статей русскоязычной Википедии. Заголовки, используемые в качестве искомых ключевых слов, относятся к широкому спектру тем, также как и публикации из исходных данных для данной работы. Отметим, что в Википедии есть статьи как для более общих понятий, так и более узких, поэтому ранжирование найденных ключевых слов все равно необходимо, т.к. более общие понятия могут быть также не соотноситься со смыслом исходных текстов. Также в качестве набора словосочетаний для поиска ключевых слов могут быть использованы и иные источники, например тезаурусы.

В наборе заголовков статей Википедии около 3 миллионов слов и словосочетаний, который необходимо было найти в текстах для 18 тысяч нормализованных текстов. Необходимо было решение, которое:

- Оптимизирует процесс поиска фиксированного набора строк (особая структура данных как, например, в алгоритме Ахо-Корасик);
- Учитывает границы слов.

Была найдена библиотека FlashText, использующий одноименный алгоритм, описанный в статье (5), которая удовлетворяет всем требованиям и показала высокую производительность для данной задачи. Также она разработана для выделение самого длинного совпадения, т.е. если в списке ключевых слов есть два отдельных слова и словосочетание из них, а в тексте, подаваемой ей на вход, будет именно словосочетание из этих двух слов, то на выходе будет только словосочетание, без отдельных ключевых слов.

В итоге было получен набор соответствий между статьями и набором ключевых слов и словосочетаний. Затем для них была посчитана мера TF-IDF, по ней наборы были отранжированы и отфильтрованы наименее значимые. Также был составлен обратный индекс на основе полученных данных для дайнейших экспериментов.

9. Простановка связей

После этапа выделения ключевых слов добавлены связи документ-термин>и понятие ГРНТИ-термин, где термин – заголовок статьи из Википедии. Также, из RDF описания выводятся связи документ-понятие ГРНТИ. Следовательно, можно вывести связи документ-документ на основе их общей связи с одним термином. Подобные связи легко выводятся машиной логического вывода после ее применения к онтологии, составленной из набора RDF описаний книжных ресурсов и выделенных связей с терминами. Напомним, что данная работа происходит в рамках ЕНИП, и в нем уже есть класс термина aux:VocabularyTerm и класс рубрики ГРНТИ sci:GRNTI. Поэтому возникает задача: в рамках дополнительного профиля ЕНИП разработать модель хранения связей документ-термин и термин-понятие ГРНТИ. Решение этой задачи не входит в данную работу.

10. Заключение

В ходе работы была проведена большая работа по подготовке данных книжных ресурсов электронной библиотеки «Научное наследие России», выделены ключевые слова и словосочетания на основе фиксированного набора словарных строк, позволяющие выделить семантические связи между документами.

Дальнейшая работа:

- В рамках дополнительного профиля ЕНИП разработать модель хранения связей документ-термин и термин-понятие ГРНТИ;
- Улучшить метод выделения ключевых словосочетаний, например, другими фиксированными наборами терминов, или применить специализированные алгоритмы, которые обычно применяют к полноценным текстам.

11. Список литературы

1. Серебряков В. А. Что такое семантическая цифровая библиотека // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Дубна, 2014 – Р. 21 – 25.
2. Когаловский М. Р., Паринов С. И., Классификацияи использование семантических связей между информационными объектами в научных электронных библиотеках // Информатика и её применение. – 2012 – Том 6, выпуск 3 – С. 32-42.
3. Когаловский М. Р., Паринов С. И. Семантическое структурирование контента научных электронных библиотек на основе онтологий // Современные технологии интеграции информационных ресурсов: сборник научных трудов. – 2011 – Выпуск 2.
4. Митрофанова О. А., Соколова Е. В. Автоматическое извлечение ключевых слов и словосочетаний из русскоязычных текстов с помощью алгоритма KEA // Компьютерная лингвистика и вычислительные онтологии. – 2017 – Выпуск 1.
5. Vikash Singh, Replace or Retrieve Keywords In Documents At Scale [Электронный ресурс] – Режим доступа: https://arxiv.org/pdf/1711.00046, свободный.
6. Интеграция метаданных Единого Научного Информационного Пространства РАН / А.А.Бездушный, А.Н.Бездушный, В.А.Серебряков, В.И.Филиппов. - М.: ВЦ РАН, 2006.

Приложение А. Описание публикации.

```
<?xml version="1.0" encoding="Windows-1251"?>
<rdf:RDF xmlns:isp="http://umeta.ru/namespaces/platform/ixsp"
    xmlns:xsp-request="http://apache.org/xsp/request/2.0"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:isir="http://isir.ras.ru/namespace/"
    xmlns:aux="http://umeta.ru/namespaces/blocks/auxiliary/"
    xmlns:media="http://umeta.ru/namespaces/heritage/media/"
    xmlns:kernel="http://umeta.ru/namespaces/platform/kernel/" lang="ru">
    <isir:Publication>
        <kernel:id>42040361</kernel:id>
        <isir:mainTitle>
            <rdf:value>История Земли</rdf:value>
        </isir:mainTitle>
        <isir:author>
            <kernel:id>42040359</kernel:id>
            <isir:personName>
                <isir:first>Карл Фридрих</isir:first>
                <isir:last>Мор</isir:last>
                <dc:language>ru</dc:language>
            </isir:personName>
        </isir:author>
        <isir:pages>527</isir:pages>
        <isir:publicationLanguage>
            <isir:name>
                <isir:full>русский</isir:full>
            </isir:name>
        </isir:publicationLanguage>
        <isir:publishedPlace>
            <rdf:value>М.</rdf:value>
        </isir:publishedPlace>
        <isir:publishingHouse>
            <rdf:value>Изд. А.И. Глазунова</rdf:value>
        </isir:publishingHouse>
        <dcterms:issued>
            <isir:dateValue>1868-01-01T00:00:00.000</isir:dateValue>
            <rdf:value>1868-01-01</rdf:value>
        </dcterms:issued>
        <dc:identifier>
            <rdf:type>http://isir.ras.ru/namespace/JSCCIdentifier</rdf:type>
            <rdf:value>10015406</rdf:value>
        </dc:identifier>
        <msc>10015406</msc>
        <isir:fullText>
            <viewUrl>http://books.e-heritage.ru/book/</viewUrl>
```

```
                    <aux:href>http://nasledie.enip.ras.ru/books/10015406/book.elb</aux:href>
                    <dc:format>
                        <pcv:code xmlns:pcv="http://prismstandard.org/namespaces/1.2/pcv/">text/xml</
pcv:code>
                    </dc:format>
                </isir:fullText>
                <isir:biblioDescription>
                    <isir:textContent>Мор, К.Ф. История Земли : Геология на новых основаниях / Соч.
Фридриха Мора ; пер. с нем. П.И. Шульгин. - М. : Изд. А.И. Глазунова, 1868. - XVI, 510
с.</isir:textContent>
                    <aux:href>/files/download?id=46952373</aux:href>
                    <dc:format>
                        <pcv:code xmlns:pcv="http://prismstandard.org/namespaces/1.2/pcv/">text/html</
pcv:code>
                    </dc:format>
                </isir:biblioDescription>
                <aux:isGatheredInto>
                    <kernel:id>42034099</kernel:id>
                    <dc:title>
                        <rdf:value>38.19.00 Геолого-геофизические исследования глубинного строения
Земли</rdf:value>
                        <dc:language>ru</dc:language>
                    </dc:title>
                </aux:isGatheredInto>
            </isir:Publication>
        </rdf:RDF>
```

Приложение В. Примеры оглавления.

```
<book>
  <msc>10079734</msc>
  <description>Феофан Затворник, епископ Тамбовский и Шацкий [Говоров]  Собрание писем святителя Феофана. Вып. 1. – 1898</description>
  <firstPage>1</firstPage>
  <lastPage>261</lastPage>
  <content>
    <el sPage="3" lPage="4" level="0">Вместо предисловия</el>
    <el sPage="5" lPage="8" level="0">1. Милость Божья буди с вами, Достопочтеннейший о. Архимандрит!</el>
    <el sPage="8" lPage="9" level="0">2. Ваше Высокопреподобие, Достопочтеннейший о. Архимандрит</el>
    <el sPage="9" lPage="10" level="0">3. Милость Божья буди с Вами, Достопочтеннейший о. Архимандрит</el>
    <el sPage="10" lPage="10" level="0">4. Милость Божья буди с Вами! Бог благословит ваше послушание</el>
    <el sPage="10" lPage="11" level="0">5. Бог в помощь! Спасайтесь! N... N.</el>
    <el sPage="11" lPage="12" level="0">6. Бог в помощь! Спасайтесь!</el>
    <el sPage="13" lPage="13" level="0">7. Милость Божия буди с вами. N... N...</el>
    <el sPage="14" lPage="14" level="0">8. Милость Божия буди с вами!</el>
    <el sPage="15" lPage="15" level="0">9. Благослови Вас Господи!</el>
    <el sPage="15" lPage="16" level="0">10. Милость Божья буди с Вами!</el>
    <el sPage="16" lPage="17" level="0">11. Милость Божья буди с Вами! Батюшка мой о. N...</el>
    <el sPage="17" lPage="19" level="0">12. Бог в помощь! Спасайтесь!</el>
.....
  </content>
</book>

<?xml version="1.0" encoding="utf-8"?>
<book>
  <msc>10005998</msc>
  <description>Максимилиан Александрович Волошин. О Репине Москва. Книгоиздательство 'Оле-Лукойе'. 1913</description>
  <firstPage>1</firstPage>
  <lastPage>158</lastPage>
  <content>
    <el sPage="5" lPage="9" level="0">Предисловіе.</el>
    <el sPage="11" lPage="22" level="0">О смыслѣ катастрофы, постигшей картину Рѣпина.</el>
    <el sPage="23" lPage="84" level="0">О художественной цѣнности пострадавшей картины Рѣпина.</el>
    <el sPage="85" lPage="137" level="0">Психологія лжи.</el>
    <el sPage="139" lPage="158" level="0">ПРИЛОЖЕНІЕ: А. Ландцертъ о картинѣ Рѣпина.</el>
  </content>
</book>
```