1. Введение

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.

В данной курсовой работе рассматривается проблема выделения ключевых слов для дальнейшего выделения семантических связей.

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.

2.2. 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.

2.3. 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).

2.4. Постановка задачи

Цель работы:

разработка алгоритма выделения ключевых слов в публикациях электронной библиотеке в рамках ЕНИП

Этапы работы:

1. Предобработка исходных данных;
2. Разработка и реализация алгоритма выделения ключевых слов.

3. Обзор темы

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 алгоритму 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 была представлена модификация KEA для текстов на русском языке.

Авторы работы [7] исследуют существующие open-source решения для выделения ключевых слов: Google KEA, надстройку над ним Maui, lvsh, KeywordFinder и TextRank. К сожалению в результате качество алгоритмов было довольно низким, что было связано с малым размером тестового набора данных, недостатками методов и общей сложностью задачи.

4. Предобработка исходных данных

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.

Таким образом, исходные данные «Наследия» изначально не были готовы к какой-либо текстовой обработке, поэтому предобработка происходила следующим образом:

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

4.1. 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:

- Вход – исходный текст;
- Выход – язык документа.

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 ‘і’ в нижнем регистре, то относим к русскому языку в новом правописании;

   2.2. Иначе относим к руссому языку в петровском правописание

   Сложность алгоритма – O(N), где N – длина текста.

Этот алгоритм работает на исходных данных, где в названиях оглавлений встречаются римские цифры, обозначающие номер главы, поэтому был выделено условие про букву ‘і’ 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.

4.2. 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.

4.3. Нормализация

Процесс нормализации происходил следующим образом:

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

Алгоритм удаления стоп-слов, пунктуации, отдельно стоящих букв и чисел:

- Вход – исходная строка;
- Выход – строка без стоп-слов, пунктуации, отдельно стоящих букв и чисел, записанных как арабскими, так и римскими цифрами;

1. Удалить все символы пунктуации из строки;
2. Разбить строку по пробельным символам на отдельные слова;
3. Для каждого отдельного слова: если слово записано только арабскими цифрами или является числом, записанном римскими цифрами или слово принадлежит списку стоп-слов и отдельно стоящих букв, то удалить его из списка отдельных слов;
4. Объединить оставшиеся отдельные слова в строку, в качестве разделителя используется пробел.

Сложность алгоритма – O(N*M), где N – количество слов в строке, M – длина списка стоп-слов и отдельных букв.

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.

5. Выделение ключевых слов

5.1. Разработка алгоритма

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.

Таким образом, алгоритм выделения ключевых слов следующий:

- Вход – исходные тексты, набор заголовков статей Википедии, список стоп-слов и отдельных букв;
- Выход – набор пар исходный текст – список ключевых слов.

1. Удалить из исходных текстов и набора заголовков статей Википедии стоп-слова, пунктуацию, числа, привести каждое из слов к нормальной форме (алгоритм описан в разделе 4.3);
2. Найти в нормализованных текстах полные совпадения нормализованных заголовков, при этом каждое слово по отдельности должно совпадать с каждым;
3. Вычислить меру TF-IDF, где термин – нормализованный заголовок, а документ – нормализованный исходный текст;
4. Вычислить пороговое значение следующим образом: отсортировать значения TF-IDF для каждого исходного текста по убыванию и взять N-ое с начала, где N равно 5;
5. Сохранить в качестве ключевых слов те совпадения, для которых TF-IDF выше порогового значения (определяется для каждого текста отдельно).

Сложность алгоритма: O(N*M*K) + O(N*M) + O(N*M) + O(N*M*logM) + O(N*M) = O(N*M*(K + logM)), где N – количество исходных текстов, M – самое большое количество слов в документе, K – количество стоп-слов и отдельно стоящих букв. Сложность вычислена с учетом выбранного в разделе 5.2 алгоритма для пункта 2 описанного алгоритма.

5.2. Детали реализации

В наборе заголовков статей Википедии около 3 миллионов слов и словосочетаний, который необходимо было найти в текстах для 18 тысяч нормализованных текстов. Фактически это задача полнотекстового поиска с дополнительным требованием на учет границ слов.

Одним из наиболее распространенных средств для решения задачи полнотекстового поиска является алгоритм Ахо-Корасик, но он не удовлетворяет требованию на границы слов.

Также была найден алгоритм FlashText, описанный в статье [5], который удовлетворяет всем требованиям и показала высокую производительность для данной задачи. При его разработке автор взял за основу алгоритм Ахо-Корасик. 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.

Таким образом, для реализации был выбран алгоритм FlashText. С учетом выбранного алгоритма сложность п.2 алгоритма выделения ключевых слов – O(N*M), где N – количество исходных документов, M – самая большая длина документа.

6. Результаты

Тестирование алгоритма проводилось на наборе статей научной электронной библиотеки «Киберленинка» (https://cyberleninka.ru/), в которых были выделены ключевые слова (https://github.com/mannefedov/ru_kw_eval_datasets). В наборе было 4072 статьи на различные научные темы. Оценивание происходило по стандартным мерам Precision и Recall.

Алгоритм	Precision	Recall
Wiki + tf-idf	10.14%	7.46%

Отметим, что разработанный алгоритм показал низкое качество. Это можно связать с несколькими причинами:

- Используемые в методе признаки слабо представительны и недостаточны для решения задачи выделения ключевых слов;
- Общая сложность задачи: разметка ключевых слов производится субъективными асессорами;
- Особенность тестового набора статей: выделенные ключевые слова в тестовой выборке часто состоят из 2 и более слов и более точно отражают содержание статьи, но подобных понятий в Википедии довольно мало.

7. Заключение

В ходе работы была проведена большая работа по подготовке данных книжных ресурсов электронной библиотеки «Научное наследие России», разработан алгоритм выделения ключевых слов на основе фиксированного набора словарных строк. К сожалению, качество разработанного алгоритма неудовлетворительно.

Дальнейшая работа:

- Прододжить разработку алгоритма выделения ключевых слов;

- В рамках дополнительного профиля ЕНИП разработать модель хранения связей документ-термин и термин-понятие ГРНТИ.
- Разработка алгоритма выделения семантических связей.