УДК 00.004.9

СЕМАНТИЧЕСКИЙ АНАЛИЗАТОР ТЕКСТОВЫХ ДОКУМЕНТОВ ТЕХНИЧЕСКОЙ НАПРАВЛЕННОСТИ

Ковалева А.П.

Научный руководитель – к.т.н., доцент Русанова О.А.

Сибирский федеральный университет

Современные технологии и темп их развития делают потенциально доступными огромные объемы информации, ставя тем самым новые проблемы – эффективной работы с такими объемами. В ситуации «информационной перегрузки» особенно актуальными становятся автоматические методы работы с большими объемами информации, в частности – методы получения сжатого представления текстовых документов – рефератов, или аннотаций.

Автоматическое реферирование текста – направление компьютерной лингвистики, широко применяемое в поисковых системах, где оно позволяет осуществлять интеллектуальный поиск, то есть не только учитывать ключевые слова, но и распознавать косвенное описание предмета поиска. Кроме того, данный метод используется в системах автоматической классификации документов на основе анализа содержания и в системах автоматического аннотирования.

Основные принципы и подходы для извлечения значимых предложений из текста на основе формальных параметров были сформулированы еще в конце 50-х – начале 60-х годов XX века. Эти методы включают в себя выделение ключевых слов текста, взвешивание предложений на основе весов входящих в него слов и дополнительных критериев: положения предложения в тексте (начало и конец документа, начало абзаца и т.д.), наличия «сигнальных фраз». В свою очередь, процедура выделения ключевых слов (или подсчета весов слов) может использовать частоту встречаемости слов в тексте и во всей коллекции, встречаемость слова в заголовках и так далее.

Большинство разработчиков подобных систем, считают, что в ближайшие 10 лет системы прикладной лингвистики изменят наш способ усвоения информации. Несмотря на то, что обоснованность данного утверждения остается под вопросом, невозможно отрицать, что автоматическое реферирование может существенно упростить обработку большого объема текстовой информации.

Выделяются следующие типы авторефератов:

1. Индикативные (предоставляют информацию о тексте, достаточную для принятия решения, есть ли необходимость обращаться к оригиналу);
2. Информативные (заменяют собой первоисточник, содержат фактическую информацию в сжатом виде);
3. Критические (не только передают основное содержание документа, но и дают ему оценку).

В нашем исследовании мы ставили перед собой целью разработать приложение, составляющее список индикативных авторефератов выбранных пользователей текстовых файлов. Нами были определены следующие допустимые форматы текстовых документов:

1. Техническая литература: *.txt, *.docx, *.odt;
2. Исходный код: *.cpp, *.cs, *.h, *.php и другие расширения исходных кодов программ;
3. Языки разметки: *.html, *.xml, *.xaml.

Разрабатываемая система представляет собой клиент-серверное приложение. Клиентская часть представляет собой программу с веб-интерфейсом. Пользователю предлагается выбрать список файлов, автоматические рефераты которых он хочет получить, максимальный размер автореферата, а так же выбрать путь, по которому сохранятся результаты работы программы. Серверная часть представляет с собой базу знаний, состоящих из правил, описанных на языке C#; базы данных, содержащей частотный словарь, базу синонимов и сильных связей; а так же и машину логического вывода, которая принимает решение и составляет автореферат.

В процессе анализа, прежде всего, определяется расширение файла, далее возможны два случая: определение содержание текста (если мы имеем дело с технической литературой или с языком разметки), либо определение назначения программного кода (если объектом анализа является исходный код). Анализ исходного кода также может включать в себя анализ семантики формального языка, однако в связи с несоответствием заявленному назначению программы – автореферирования – это выходит за рамки нашего исследования.

В том случае, если мы имеем дело с техническим текстом или языком разметки программа работает по следующему алгоритму:

1. Удаляются описания стилей, сценарии и все тэги, за исключением тех, которые определяют текст как заголовок, подзаголовок, либо выделенный текст (курсив, подчеркивание, жирный шрифт и т.д.). В случае если анализируемый текст имеет расширение *.txt, данный пункт пропускается ввиду отсутствия форматирования в данном формате.
2. Определяется частота встречаемости словоформ в тексте.
3. Исходя из таких параметров, как частотная характеристика слова, наличие его в названии файла, в заголовках и подзаголовках, частоты встречаемости в тексте, выделение его каким-либо способом определяется вес слова. Полным синонимам назначается одинаковый вес.
4. Вычисляются веса словосочетаний. При вычислении учитывается частота пар слов и такие морфологические шаблоны как согласованное прилагательное + существительное и существительное + существительное в родительном падеже.
5. Вычисляются веса предложений. Учитываемыми характеристиками являются: положение предложения в тексте (первые и последние предложения имеют больший вес), веса слов и словосочетаний в тексте, длина предложения; кроме того введен понижающий коэффициент для вопросительных предложений.
6. Определяется связь предложений между собой (наличие в предложении местоимений, связок, предлогов, союзов связывающих его с другим предложением).

7. Формирование автореферата.

a. Предложения сортируются в соответствии с вычисленным весом по убыванию.

b. Первое предложение помещается в реферат. Каждое следующее предложение берется из списка и сравнивается с предложениями реферата (в случае с большим количеством общих слов с предложениями реферата предложение отбрасывается). Также учитываются связи между предложениями.

c. Процесс повторяется, пока не будет отобрано заданное количество предложений.

d. Отобранные предложения выдаются в том порядке, в каком они находились в тексте.

В случае если анализируемый файл является исходным текстом программы, назначение программы определяется по названию файла и комментариям. Самым большим весом обладает комментарий перед основным текстом программы, так как в случае грамотного комментирования текста в нем содержится описание назначения программы. Если заданное количество предложений не достигнуто, отбираются комментарии, относящиеся к наибольшим по объему описаниям методов или классов. В автореферат также помещается информация о языке, на котором написана программа и является ли файл исходным кодом или библиотекой.

Результатом проделанной работы является клиент-серверное приложение, разработанное на платформе .NET, осуществляющее автоматическое реферирование текстовых файлов. Данная система может найти применение у технических специалистов, работающих с большими объемами информации, так как позволяет им сэкономить время, предоставляя информацию об исходном тексте достаточную, чтобы определить его релевантность задачам, стоящим перед пользователем.

Кроме того, возможна дальнейшая работа над данной темой. При работе с файлами, содержащими исходный код программы может быть реализован анализ семантики языка, на котором написана программа, итогом такого анализа будет являться вывод о корректности работы программы. В случае работу с технической литературой, в дальнейшем планируется ввести более интеллектуальные методы определения релевантности слов и предложений, углубленная работа с семантикой текста.