УДК 004.738.52

О.С. Волковский, Е.Р. Ковылин

СЕМАНТИЧЕСКИЙ АНАЛИЗ СОДЕРЖИМОГО WEB-ПРИЛОЖЕНИЙ

Анотація. Розглянуто систему семантичного аналізу текстових даних (без опори на знання), яка дозволяє квазиреферувати web -сторінки. Отримані результати можуть бути застосовані для скорочення часу пошуку та аналізу даних в мережі інтернет.

Введение. С развитием компьютерных и, в частности, web технологий можно наблюдать значительное расширение информационного пространства сети интернет. На данный момент в сети можно наблюдать огромное число тематических сайтов и электронных документов самой разной направленности и их количество растет экспоненциально. В связи с этим возникает ряд вопросов связанных с обработкой, анализом и оптимизацией работы с текстовыми данными в сети. Одним из них является высокая вариативность подачи информации на одинаковую тематику, что приводит к увеличению времени поиска и анализа найденных данных. Поскольку наиболее естественной формой общения и подачи информации для человека является естественный язык, то основными инструментами для решения этой проблемы становятся методы математической лингвистики, позволяющие программе оперировать с массивами данных, описанных на естественном человеческом языке. В этой статье рассматривается система, позволяющая минимизировать время обработки текстовой информации пользователем в интернете, путем составления информативного автореферата страницы на основе семантического анализа данных.

Постановка проблемы. В контексте этой работы, под семантическим анализом мы будем понимать процесс автоматического реферирования – составление небольшого текста-реферата, способного донести до пользователя основную идею исходного документа. Решение данной задачи влечет за собой возникновение ряда научных вопросов – от понимания естественного языка как объекта математического

моделирования, до выбора оптимальных алгоритмов для семантического анализа.

Лингвистические проблемы автореферирования заключаются в формировании необходимых знаний о языке. Это связано с тем, что естественный язык, как объект моделирования, обладает рядом свойств, затрудняющих семантический анализ. Наиболее очевидной проблемой является смешанность правил естественного языка. Точность и полнота работы алгоритмов, напрямую зависит от того, насколько шаблонны правила, заложенные в них. Однако, для любого естественного языка на фоне общих правил, существует огромное количество исключений и частных случаев, причем практически во всех областях лингвистики. Это приводит к отклонению от заданного шаблона, что значит получение некорректных результатов на выходе. С другой стороны, появляется вопрос о недостаточной формализации языка – отсутствия абсолютно полного набора правил, способных точно описать структуру естественного языка, которая будет единственной и непротиворечивой в общем случае. Указанные свойства также приводят к еще одной проблеме – отсутствию универсальных тестовых наборов данных. Под этим подразумевается, что в следствии смешанности и недостаточной формализации языка, для алгоритма невозможно подобрать тестовый набор данных, обработав который, мы сможем достаточно полно оценить результаты его работы.

Одним из этапов построения автоматического реферата является взвешивание частей исходного текста по заранее заданным критериям. Это необходимо для последующего включения наиболее «тяжелых» предложений в реферат. Кроме того, необходимо проводить статистическую и стилистическую корректировки текста, как исходного, так и получаемого. Необходимость построения и выбора оптимальных алгоритмов приводит к алгоритмическим проблемам семантического анализа – выбор наилучшей математической модели представления естественного языка, построение эффективных алгоритмов предварительной обработки текста, обмена данными между частями системы, алгоритмов взвешивания и конечной сборки реферата.

Анализ последних разработок и публикаций. В течении последних лет появилось множество материалов по теме интеллектуальной обработки текста. Термин «автоматическое реферирование» разделился на два направления – квазиреферирование, под которым

 ISSN 1562-9945

понимается поиск и выделение наиболее информативных фрагментов текста и генерация рефератов – составление новых текстов на основе исходного. Разработанный алгоритм относится к направлению квазиреферирование, поскольку основной целью работы является оптимальное сокращение объема исходного текста, а не генерация нового.

Большая часть алгоритмов реферирования отталкивается от алгоритма Г. Луна [1]. Для современных алгоритмов характерно сочетание классического подхода Луна со статистическими методами реферирования и методиками, основанными на знаниях. Так, в работе [2] приводится метод реферирования, основанный на машинном обучении, а в работе [3] предлагается подсчитывать вес предложения на основе связей с предложениями, стоящими слева и справа от него, используя специальный словарь слов-связей.

Из существующих систем, направленных на семантическую обработку текста, особенно стоит отметить продукт Intelligent Miner [5], со стоимостью пакета от 18 до 60 тыс. долларов, разработанный фирмой IBM как набор пяти утилит, предоставляющий инструментарий для text mining.

Постановка задачи. Разработать алгоритм автоматического реферирования web-страниц. Реализовать систему на основе построенного алгоритма и проанализировать результаты ее работы на базе тестового набора данных.

Основная часть. Структуру программы условно можно разделить на четыре части: блок синтаксического анализа, блок-стеммер, блок семантического анализа, блок сборки реферата. Логику работы программы можно увидеть на рис. 1. Синтаксический блок включает в себя методы для выделения предложений и слов из текста, методы удаления стоп-слов – незначимых и неинформативных слов в предложении, методы проставления тэгов – специальных дескрипторов для семантического анализа, методы для работы со вспомогательными словарями. Блок-стеммер представляет собой набор методов, необходимых для проведения специальных операций над каждым информативным словом в тексте. Это операция лемматизации – приведение слова к словарной форме, операция стемминга по алгоритму Портера – выделения основы слова, операция нахождения наибольшей общей основы слова. Блок семантического анализа необходим для взвешивания предложений по разработанному алгоритму и подсчета количе-

ства предложений в конечном реферате. Блок сборки реферата включает методы для парсинга web-страницы, стилистической корректировки текста, сборки конечного реферата.

Рисунок 1 - Архитектура системы реферирования

Рассмотрим алгоритм работы программы подробнее, и поясним назначение и смысл некоторых операций.

1. В первую очередь производится операция парсинга web-страницы. На этом шаге текстовая информативная часть отделяется от неинформативного текстового web-кода.
2. Из полученного текста происходит выделение предложений и слов. При этом учитывается, что в предложении могут быть сокращения, многоточия, лишние пробелы.
3. Следующей задачей становится тегирование – расстановка специальных дескрипторов, которые необходимы для последующей семантической обработки текста и сборки реферата. Это необходимо для стилистической корректировки текста.
4. Тегированный текст проходит через операцию лемматизации – разновидности морфологического анализа, которая заключается в привидении словоформы к исходной словарной форме. В результате лемматизации от словоформы отбрасываются флективные окончания и возвращается основная форма слова, что позволяет получить более точные результаты при семантическом анализе.
5. Над каждым предложением происходит операция удаления стоп-слов. Под стоп-словом мы будем понимать слово, анализи-

 ISSN 1562-9945

ровать которое нет смысла в силу его низкой информативности. Все стоп-слова занесены в специальный словарь.

6. Следующий шаг – это операция стемминга по алгоритму Портера над каждым словом. Стемминг – это процесс нахождения основы слова (которая не всегда может совпадать с морфологической основой). Предварительная операция лемматизации позволяет сократить количество ошибок, появление которых обусловлено не совершенностью алгоритма Портера. Алгоритм включает в себя четыре шага. На первом шаге отсекается максимальный формообразующий суффикс, на втором — буква «и», на третьем — словообразующий суффикс, на четвертом — суффиксы превосходных форм, «ь» и одна из двух «н».
7. Говоря об алгоритмах отсечения окончаний, можно отметить, что зачастую слово обрезается больше чем это необходимо. Кроме того, алгоритмы неустойчивы к выпадению и замене букв в корне и суффиксах при словообразовании. Леманнизация хоть и сокращает число таких ошибок, однако не позволяет добиться точности, необходимой для семантического анализа. Поэтому над каждой парой слов, уже пропущенных через лемматизатор и стеммер, проводится операция нахождения максимальной наибольшей части слова. Найти общую часть для пары слов не трудно, сложность этого шага заключается в оценки пригодности найденной общей части. Для того, чтобы понять, насколько найденная общая часть адекватна с точки зрения морфологии, мы используем механизм, основанный на оценке длин слов. По формуле:

$$EQ = Q \cdot \max(wl_i, wl_j), \quad (1)$$

где Q – коэффициент эквивалентности, $max(wl_i,wl_j)$ - функция нахождения максимальной из длин пары слов (wl_i,wl_j) мы находим величину EQ. Если значение EQ меньше, чем длинна найденной общей части, то она пригодна для семантического анализа, и каждое слово из пары заменяется на найденную общую часть. Коэффициент Q указывает, насколько процентов должна совпадать длинна найденной части с максимальной длинной слова из пары слов. Экспериментально было установлено, что оптимальное значение Q равно 0,4. Таким образом,

мы находим корпуса слов для конкретного текста, которые могут быть использованы для взвешивания.

8. Получив набор корпусов слов, можно приступать к взвешиванию предложений. Вес каждого предложения рассчитывается по формуле:

$$P = \sum_{i=0}^{N} FQ_i + H \cdot \frac{WC}{SC}, \qquad (2)$$

где: FQ_i – частота слова в тексте, N – число слов в предложении, H – количество слов-слов подсказок в предложении, WC – количество слов в тексте, SC – количество предложений в тексте.

9. В результате, мы получаем набор пар предложение и его вес. Последней задачей становится сборка конченого реферата. Для этого, мы подсчитываем величину сжатия исходного теста, которая определит количество предложений в конечном реферате: Из множества найденных предложений, мы выбираем предложения с наибольшим весом, и включаем их в реферат, в порядке их следования в тексте. Из набора убираются предложения, которые помечены тегом связки с предыдущим предложением, которое не попало в реферат. Это делается для устранения смысловых разрывов в реферате. Результаты реферирования приведены в табл. 1.

Таблица 1

Исходный текст и его автореферат

Исходный текст	Автореферат
Программирование это процесс создания компьютерных программ. В узком смысле (так называемое кодирование) под программированием понимается написание инструкций (программ) на конкретном языке программирования (часто по уже имеющемуся алгоритму, методу решения поставленной задачи). Соответственно, люди, которые этим занимаются, называются программистами (на профессиональном жаргоне кодерами), а	Программирование это процесс создания компьютерных программ. В узком смысле (так называемое кодирование) под программированием понимается написание инструкций (программ) на конкретном языке программирования (часто по уже имеющемуся алго-

 ISSN 1562-9945

те, кто разрабатывают алгоритмы алгоритмистами, специалистами предметной области, математиками. В более широком смысле под программированием понимают весь спектр деятельности, связанный с созданием и поддержанием в рабочем состоянии программ программного обеспечения ЭВМ. Иначе это называется программная инженерия (инженерия ПО). Сюда входят анализ и постановка задачи, проектирование программы, построение алгоритмов, разработка структур данных, написание текстов программ, отладка и тестирование программы (испытания программы), документирование, настройка (конфигурирование), доработка и сопровождение. Программирование для ЭВМ основывается на использовании языков программирования, на которых записывается программа. Чтобы программа могла быть понята и исполнена ЭВМ, требуется специальный инструмент транслятор. В настоящее время активно используются интегрированные среды разработки, включающие в свой состав также редактор для ввода и редактирования текстов программ, отладчики для поиска и устранения ошибок, трансляторы с различных языков программирования, компоновщики для сборки программы из нескольких модулей и другие служебные модули.	ритму, методу решения поставленной задачи). В более широком смысле под программированием понимают весь спектр деятельности, связанный с созданием и поддержанием в рабочем состоянии программ программного обеспечения ЭВМ. Иначе это называется программная инженерия (инженерия ПО). В настоящее время активно используются интегрированные среды разработки, включающие в свой состав также редактор для ввода и редактирования текстов программ, отладчики для поиска и устранения ошибок, трансляторы с различных языков программирования, компоновщики для сборки программы из нескольких модулей и другие служебные модули.

Выводы. Проведенные исследования позволили разработать систему автоматического реферирования содержимого web приложений на естественном человеческом языке. Семантические алгоритмы, которые легли в основу системы позволяют проводить эффективный синтаксический, морфологический и частотный анализ. Работа системы показывает, что алгоритмы семантического анализа без опоры на знания могут давать достаточно хорошие результаты, применение которых в рамках задачи оптимизации поиска информации в интернете весьма уместно. Следующим шагом может стать оптимизация алгоритмов синтаксического анализа, что приведет к улучшению качества получаемого реферата.

ЛИТЕРАТУРА

1. Luhn H. The automatic creation of literature abstracts. // In IBM Journal of Research and Development, Vol. 2(2), pp. 159–165, 1958.
2. Браславский П.И., Густелев В. Система автоматического реферирования новостных сообщений на основе машинного обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Девятой Всероссийской научной конференции RCDL '2007 (Переславль–Залесский, Россия, 15–18 октября 2007 г.). – Переславль-Залесский: Изд-во «Университет гор. Переславля», 2007. – С. 142–147.
3. Яцко В.А. Симметричное реферирование: теоретические основы и методика // НТИ. Сер. 2, №5, 2002. – С. 18–28.
4. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP - БХВ-Петербург, 2007. – 384с.
5. IBM Intelligent Miner for Text Version [Электронный ресурс]. Режим доступа: URL: http://www-01.ibm.com/

 ISSN 1562-9945