OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822 + 81’322.2

ОБ ОДНОМ МЕТОДЕ ПРИМЕНЕНИЯ ОБОБЩЕННОЙ ОНТОЛОГИИ ДЛЯ АНАЛИЗА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ СООБЩЕНИЙ ПОЛЬЗОВАТЕЛЕЙ СЕТИ ИНТЕРНЕТ

Деева Н.В., Вишневский С.Я.

Гродненский государственный университет имени Янки Купалы,
г. Гродно, Республика Беларусь

nvdeeva@gmail.com

svishnevskij@gmail.com

В данной статье предложен один метод применения обобщенной онтологии для анализа открытых текстовых отзывов в сети Интернет для поддержки принятия решения пользователя по выбору наиболее предпочтительного варианта. Описан алгоритм определения списка критериев анализа отзывов «на лету», а также алгоритм получения эмоциональной оценки базовых слов по критерию и визуализация полученных результатов.

Ключевые слова: естественно-языковые сообщения, обобщенная онтология, эмоциональная оценка словосочетания

Введение

В современном мире все большую зависимость от знаний всемирного интернет сообщества приобретают различные сферы жизнедеятельности человека. Такую зависимость в целом определяет переход в начале 21 века подавляющего большинства web-ресурсов на концепцию формирования контента и взаимодействия с пользователем Web 2.0, а в последние 5 лет стремительный переход на Web 3.0. Web 2.0 характеризует пользователя не как пассивного потребителя информации, а как активного участника создания и верифицирования интернет-контента. В концепции Web 3.0 особый акцент сделан на семантику создаваемой и хранимой информации, извлечению знаний из массивов данных.

Все чаще пользователи высказывают свое мнение о покупке, услуге, сервисе и т.д. в виде комментариев, отзывов или статей, тем самым, расширяя не только количественную, но и географическую область обмена впечатлениями. Как правило, такого рода информацию размещают на официальных сайтах производителя в разделе комментариев и на форумах, в социальных сетях, а также на специализированных площадках для обмена отзывами по различным направлениям (туризм, шоппинг, медицинские услуги и др.). Очевидна актуальность извлечения данных из контента упомянутых выше ресурсов и их дальнейшая обработка с целью создания среды для поддержки принятия решения пользователя в части выбора услуги или товара. Задача обработки текстового контента открытого ресурса сводится к трем принципиально различным подзадачам: формирование базы небольших текстовых сообщений, то есть разбор страницы ресурса и извлечение фрагментов текста из нее; подзадача обработки текстов на естественном языке и визуализация полученных результатов.

В данной статье главный акцент сделан на обработку уже готового набора текстовых сообщений на естественном языке, а также построения на основе полученных данных визуального представления проанализированного контента, технические вопросы по разбору страниц web-ресурса опущены, так как в контексте исследования не представляют научного интереса.

1. Постановка задачи

В данном исследовании в качестве источника информации используется ресурс по обмену отзывами о курортных отелях, размещенный по адресу http://otzyv.ru/. Ресурс позволяет оставлять отзывы и оценки по отелям даже незарегистрированным пользователям, тем самым предоставляя возможность оставлять анонимные

комментарии, упрощая интерфейс обмена информацией между желающими отдохнуть. Справедливости ради необходимо отметить, что именно эта открытость для пользователя не гарантирует объективность и достоверность предоставленной в комментариях информации, предлагая доступный механизм манипулирования информацией в конкурентной среде. Однако тема ложных или намеренно сгенерированных отзывов не является целью данного исследования, и мы будем исходить из предположения, что вся информация, приведенная в отзывах пользователей достоверна.

Итак, предложенный в качестве источника данных открытый ресурс позволяет пользователю определить независимую оценку отеля, и проанализировать пользовательские предпочтения, но только лишь в ручном режиме обработки данных.

Свои оценки пользователи оставляют в виде текстовых сообщений на естественном языке, а также выставляют баллы по нескольким категориям, например, «Питание», «Обслуживание» и т.д. На рисунке 1 приведен фрагмент пользовательского отзыва.


Noche de luna 15.05.14 14:50:35
отели Египта / Макади Бей / Sunwing Waterworld Makadi 5* Оценка отелю: 4+
Туроператор: Библио-Глобус Оценка оператору: 5
Время отдыха: май 2014
Достоинства:
хорошее море для деток
хорошая территория и номера
отличный аквапарк
Недостатки:
не все хорошо с сервисом
проблемы с пакетированными соками
Оценки отелю:
Расположение: 5
Территория: 5
Обслуживание: 3+
Питание: 4
Развлечения: 4
Для детей: 4
Отзыв:
Ну вот мы и вернулись. Сразу отпишусь, пока впечатления свежи.
Завтраки. Несколько видов хлопьев, йогурты натуральные и фруктовые в упаковке (увы, желатина в них много). Рис с молоком, овсянка на воде, блины. Яйца вареные, скрембл, омлет или глазунью пожарят прямо для вас. Сосиски, фасоль, вареные бобы, пюре (плохое) или вареная обжаренная картошка (отличная). Несколько видов сладкой выпечки, много вкусного хлеба и СЕКРЕТ – теплые круассаны ищите в кастрюльке с крышкой. Сыр, неплохая фета, местные колбасы, оливки и свежие овощи и фрукты.

Рисунок 1 – Фрагмент отзыва пользователя, размещенного на открытом ресурсе http://www.otzyv.ru/

Безусловно, в разделе «Оценки отелю» пользователь достаточно быстро может получить некоторые цифры – первичные критерии фильтрации, не отражающие по сути ни причин, ни конкретных фактов, повлиявших на автора при их выставлении. Глубинный анализ потребует от пользователя прочтения всех отзывов авторов, участвующих в оценке. А в современных реалиях, количество отзывов по одному направлению может достигать тысяч, что делает процесс обработки в ручном режиме зачастую невыполнимым в приемлемые сроки.

Необходимо предложить решение задачи ускорения анализа большого числа отзывов. В данной статье предлагается формировать некоторое визуальное представление для каждого отзыва по заданному критерию. Такое представление должно при беглом просмотре отзыва пользователем акцентировать его внимание на положительных и отрицательных фактах по заданному пользователем критерию, например, просмотр темы «Питание». Целевой аудиторией данного решения в первую очередь являются рядовые пользователи, планирующие свой отдых, а также работники в сфере предоставления туристических услуг.

2. Алгоритм обработки пользовательских сообщений

2.1. Формирование списка критериев на базе обобщенной онтологии

С целью формирования списка критериев анализа необходимо определить наиболее популярные направления, раскрытые в отзывах пользователей. Для этого было принято решение получить набор наиболее частотных ссылок по тексту на концепты обобщенной онтологии. А в качестве обобщенной онтологии была выбрана классификация лексики Н.Ю. Шведовой, которая находится в открытом доступе на ресурсе http://slovari.ru/ [slovari, 2014]. На рисунке 2 приведен фрагмент классификации лексики с частично раскрытым концептом «Еда, питьё, кушанья: их компоненты».


Продукты питания, еда, напитки
Общие обозначения
Еда, питьё, кушанья; их компоненты
Общие обозначения
Хлеб, хлебные изделия. Мука, крупа; изделия из них
Молоко, молочные изделия. Яйца, яичные изделия. Жиры
Мясо. Рыба. Мясные и рыбные блюда, кушанья
Общие обозначения
Мясо млекопитающих, птиц; блюда из него
Само мясо млекопитающих (0052; 0302)
Части мясной туши (0052; 0303)

Рисунок 2 – Фрагмент классификации лексики Н.Ю. Шведовой

В качестве источника для анализа были выбраны около 20 000 уникальных отзывов, размещенных на открытом ресурсе www.otzyv.ru. Далее для формирования коллекции слов, претендующих на роль ссылок на концепты обобщенной онтологии, выполним морфологический анализ каждой найденной лексемы каждого сообщения, определим ее нормальную форму и добавим в ассоциативный массив-коллекцию слов, выполняя на данном этапе также подсчет частоты встречаемости данного слова во всем наборе сообщений. Кандидатами на роль ссылок будем считать слова, являющиеся именами существительными с признаком неодушевленности. На рисунке 3 представлен фрагмент xml-файла, содержащего полученную коллекцию слов.

```
<Word attr="1" count="77">САЛАТ</Word>
<Word attr="3" count="19">ОГУРЕЦ</Word>
<Word attr="3" count="28">ПОМИДОР</Word>
<Word attr="3" count="5">ПЕРЕЦ</Word>
<Word attr="1" count="10">КАПУСТА</Word>
<Word attr="1" count="8">ВИЛОК</Word>
<Word attr="3" count="3">БАКЛАЖАН</Word>
```

Рисунок 3 – Фрагмент xml-файла, содержащего коллекцию слов, претендующих на ссылки на концепты онтологии

Уровень морфологического анализа, на котором определяются морфологические характеристики

каждой лексемы, а также ее нормальная форма, реализован с использованием свободного COM-объекта, предоставленного группой разработчиков «Диалинг» [aot, 2014]. Данный программный инструментарий позволяет для заданной словоформы получить все возможные наборы морфологических характеристик, не решая тем самым проблему снятия омонимии и многозначности. В силу того, что на данном этапе строится только список критериев, то есть определяются основные темы массива сообщений, можно пренебречь заявленными выше проблемами и сформировать коллекцию для всех полученных слов, с условием их принадлежности к классу неодушевленных существительных.

Для определения ссылок на концепты обобщенной онтологии, построим xml-представление для основных концептов, доопределив их конкретными найденными экземплярами. На рисунке 4 представлен фрагмент xml-файла, содержащий конкретизацию концепта «Овощи. Бобовые. Бахчевые. Грибы». Название концептов соответствует классификации общеупотребимой лексики [Шведова, 1998].

```
<TreeNode xsi:type="Sheet" word_count="22">
  <name>Овощи. Бобовые. Бахчевые. Грибы</name>
  <Scheme>0049</Scheme>
  <Position>0311</Position>
  <WordList>
    <Word count="77">САЛАТ</Word>
    <Word count="23">КАРТОШКА</Word>
    <Word count="4">ГОРОХ</Word>
    <Word count="19">ОГУРЕЦ</Word>
    <Word count="28">ПОМИДОР</Word>
    <Word count="20">ЗЕЛЕНЬ</Word>
    <Word count="13">КАРТОФЕЛЬ</Word>
    <Word count="5">ПЕРЕЦ</Word>
    <Word count="10">КАПУСТА</Word>
    <Word count="18">АРБУЗ</Word>
    <Word count="6">ГРУША</Word>
    <Word count="3">БАКЛАЖАН</Word>
    <Word count="2">ФАСОЛЬ</Word>
    <Word count="14">ДЫНЯ</Word>
    <Word count="5">ГРИБ</Word>
    <Word count="2">ЛУК</Word>
    <Word count="7">КАБАЧОК</Word>
    <Word count="1">МОРКОВЬ</Word>
    <Word count="2">РЕПА</Word>
    <Word count="1">ТОМАТ</Word>
    <Word count="1">ТЫКВА</Word>
    <Word count="3">ШПИНАТ</Word>
  </WordList>
```

Рисунок 4 – Фрагмент xml-файла, содержащего найденные экземпляры концептов

Каждый концепт в данном xml-представлении содержит информацию о количестве различных слов, конкретизирующих его, а каждое слово хранит информацию о его частотности в массиве текстовых сообщений. Используя данные числовые характеристики, а также выбирая пороговую величину вхождения слова в массив и величину мощности, конкретизирующих концептов, можно настраивать величину и степень детализации списка критериев анализа пользовательских отзывов.

В данном исследовании на базе более 20000 отзывов по 1000 отелей было получено более 4000 слов в коллекции, 1500 из них были выделены в качестве экземпляров концептов. Так, например, коллекция уникальных слов по разделу «Питание» составила порядка 300 единиц.

2.2. Формирование эмоциональных оценок словосочетаний

На предыдущем этапе были построены коллекции слов-экземпляров для каждого критерия из полученного списка. В результате выделения популярных тем была частично решена проблема омонимии и многозначности, исходя из предположения, что верное значение слова (на классе неодушевленных существительных) с большей вероятностью будет экземпляром наиболее важного концепта, с одной стороны, и неверное значение(я) слова с меньшей вероятностью войдут в список конкретизирующих важный концепт, с другой стороны.

Для дальнейшего анализа отдельных отзывов, например, описывающих 1 отель, выполним поиск всех присутствующих слов-конкретизаций для заданного критерия и определим эмоциональную оценку для каждого из них, если представляется такая возможность.

Будем считать, что оценка эмоционального окраса экземпляра концепта может быть определена интерпретацией, связанного с ним, прилагательного.

Наиболее распространенным типом связи существительного и прилагательного в словосочетании является согласование. При этом виде связи прилагательное согласовано с существительным в числе, роде и падеже. При наличии данных морфологических признаков для участников вероятных словосочетаний, задача нахождения согласованных существительных и прилагательных может быть решена в автоматическом режиме с достаточно высоким коэффициентом достоверности.

Таким образом, на этапе морфологического анализа будем получать наборы морфологических характеристик не только для неодушевленных существительных, но и для прилагательных, находящихся в непосредственной близости с ними. В случае, если прилагательное и существительное согласуются, они образуют словосочетание и в дальнейшем прилагательное при необходимости может составить для существительного эмоциональную оценку.

Однако не все прилагательные могут дать эмоциональную окраску существительному, будем рассматривать только класс качественных прилагательных. Причем по всему массиву отзывов найдем все потенциальные для оценивания прилагательные и составим из них список уникальных прилагательных. Затем в ручном режиме каждое прилагательное оценим по шкале от -1 до +1, с шагом 0.1, где 0 дает нейтральную окраску, в случае. если затруднительно оценить прилагательное вне контекста. Все «негативные»,

т.е. вызывающие негативные эмоции прилагательные оцениваются отрицательными значениями, абсолютная величина которых характеризует силу негативной оценки. Аналогично несущие позитивный смысл прилагательные оцениваются положительными числами. Так, например, в словосочетании «ужасное питание», прилагательное имеет оценку -1, а существительное «питание» получает максимально негативную эмоциональную оценку. На рисунке 5 представлен фрагмент xml-файла, содержащего оцененные прилагательные.

```
<adjective score="0.7" id="432">шустренький</adjective>
<adjective score="0.8" id="433">уважаемый</adjective>
<adjective score="-1" id="434">ужасный</adjective>
<adjective score="-0.8" id="435">громкий</adjective>
<adjective score="0.1" id="436">вечный</adjective>
<adjective score="-1" id="437">вонючий</adjective>
```

Рисунок 5 – Фрагмент xml-файла, содержащего прилагательные с эмоциональными оценками (где «score» - оценка)

По всему массиву отзывов было получено около 4000 словосочетаний, на базе которых был построен список из порядка 1000 уникальных прилагательных, для каждого из которых была назначена оценка. 453 прилагательных получили нейтральную оценку, что означает что более 50 % уникальных прилагательных получили положительную или отрицательную оценку по предложенной шкале.

2.3. Визуализация анализа текста отзыва

Последний этапом в решении поставленной задачи была определена подзадача визуализации проведенного, на предыдущих шагах, анализа. То есть необходимо предоставить пользователю визуальное представление отобранных по отелю отзывов с проведенным на этих отзывах анализом – поиском экземпляров концептов с их эмоциональной оценкой. Такое представление будем формировать с помощью цветовой схемы: зеленым выделяются положительные оценки, красным – отрицательные, желтым – нейтральные; причем интенсивность цвета показывает близость к максимальному по модулю значению. На рисунке 6 приведена демонстрация анализа и визуализации его результатов на одном фрагменте текста отзыва по критерию «Продукты питания, еда, напитки».

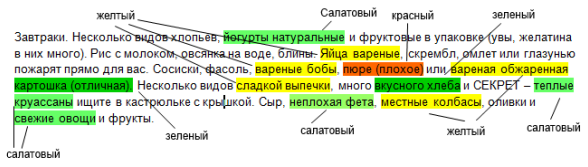

Рисунок 6 – Пример анализа отзыва (желтым выделена нейтральная оценка, зеленым - положительная, красным - отрицательная)

Предложенная визуализация анализа позволяет пользователю моментально обратить внимание на негативные оценки в отзыве, а цветовая схема позволит определить суммарную семантику оценки по заданному критерию, кроме того такой унифицированный способ визуализации позволяет по единой шкале проводить визуальную оценку разных отелей упрощая и ускоряя задачу выбора, что и было основной целью данного исследования.

Стоит тем не менее отметить, что в данном исследовании поиск словосочетаний ограничивался биграммами, нахождение триграмм и более сложных словосочетаний является темой другого исследования.

Заключение

В данной статье предложен метод использования обобщенной онтологии для анализа естественно-языковых текстов сети интернет, на примере обработки текстов туристических отзывов одного ресурса. Полученное в результате решение позволяет пользователю оперативно получать визуальное представление результатов анализа по сформированному «на лету» списку критериев, исключая из рассмотрения недостаточно представленные в описании. А также предложенная модель может быть применена на различных предметных областях, так как в основу метода положена обобщенная онтология, описывающая общеупотребительную лексику.

Библиографический список

[Шведова, 1998] Шведова, Н.Ю. 1. Толковый словарь, систематизированный по классам слов и значений / Российская академия наук. Ин-т рус. яз. им. В. В. Виноградова; Под общей ред. Н.Ю. Шведовой // М.: Азбуковник, 1998.

[slovari, 2014] Классификация лексики [Электронный ресурс]. – 2014. – Режим доступа: http://slovari.ru/default.aspx?s=0&p=2672 – Дата доступа: 13.04.2014.

[aot, 2014] Автоматическая обработка текстов [Электронный ресурс]. – 2014. – Режим доступа: http://aot.ru/. – Дата доступа: 1.02.2014.

ABOUT ONE METHOD OF USE OF GENERALIZED ONTOLOGY FOR THE ANALYSIS OF THE NATURAL LANGUAGE MESSAGES OF INTERNET USERS

Deeva N.V. , Vishneuski S.Y.

The Yanka Kupala Grodno State University, Grodno, Republic of Belarus

nvdeeva@gmail.com

svishnevskij@gmail.com

In this article we propose one method of using a generalized ontology for the analysis of shared text reviews on the Internet for the user decision support in the selection of the preferred option. There is also the description of an algorithm for determining the list of criteria for analyzing the feedback in real time, as well as an algorithm of obtaining emotional evaluation of basic words and visualization of the results.

Keywords: natural language messages; generalized ontology; emotional evaluation of collocation.