СЕКЦИЯ 3
ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ

УДК 004.82

АВТОМАТИЗАЦИЯ АНАЛИЗА МНЕНИЙ ПОЛЬЗОВАТЕЛЕЙ ОБ УСЛУГАХ В СФЕРЕ ТУРИЗМА НА ОСНОВЕ ОТЗЫВОВ В СЕТИ ИНТЕРНЕТ

Бурдук Е.А. (*burduk_z2@bk.ru*)
Гомельский государственный технический университет имени П.О. Сухого, Республика Беларусь, Гомель

В статье изложено описание системы, которая осуществляет поиск отзывов об отелях и турфирмах и определяет эмоциональную оценку полученной информации путем анализа, позволяющего извлечь из текста эмоционально окрашенную лексику и определить эмоциональное отношение авторов к объекту исследования.

Ключевые слова: парсинг, классификатор, анализ тональности, система анализа мнений

1 Введение

В настоящее время количество публикуемых отзывов достигает нескольких десятков тысяч и сбор информации в интернете вручную становится трудоемкой, рутинной и отнимающей много времени работой. Также активное развитие социальных сетей, форумов и блогов увеличивает интерес к задаче автоматизированного анализа мнений пользователей сети Интернет по различным вопросам. В связи с этим возникает потребность в автоматическом сборе информации и автоматической оценке текста [Меньшиков и др., 2012].

Анализ тональности текста представляет класс методов компьютерной лингвистики и занимается изучением эмоций и мнений в текстах. Данный анализ способен помочь разобраться в законах, по которым живет

естественный язык, и научить компьютер воспринимать его на уровне, приближенном к человеческому уровню [Pang, 2008].

Актуальность задачи анализа текста заключается в возможности оценить отношение общества к тому или иному объекту, выраженному в тексте.

Тональность текста определяется тремя факторами:

- субъектом тональности;
- тональной оценкой по различной шкале;
- объектом тональности.

Под субъектом тональности понимают автора написанного текста. К объекту тональности следует относить то, о чем автор высказывает свое мнение. Тональная оценка представляет собой эмоциональное отношение автора к объекту тональности.

Анализ тональности текста можно разделить на несколько категорий:

- анализ тональности экспертами (ручной);
- автоматизированный анализ тональности.

2 Система анализа мнений

Система анализа мнений, структура представлена на рисунке 1, позволяет с помощью специальной программы-парсера осуществлять поиск отзывов об отелях и турфирмах и проводить анализ тональности текста, который дает возможность получить представление об эмоциональной окрашенности текста.

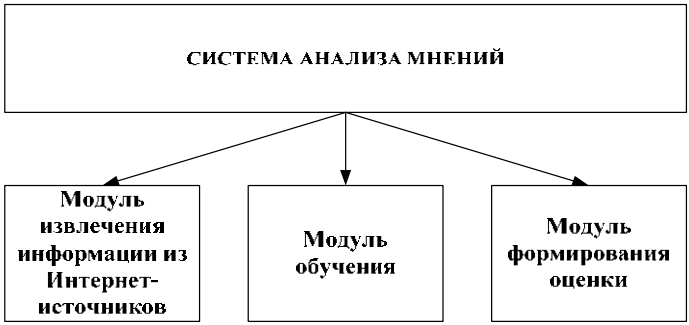

Рисунок 1 – Структурная схема системы анализа мнений

Модуль извлечения информации из Интернет-источников реализован при помощи парсинга и позволяет получить отзывы об отелях или турфирмах. Модуль обучения и формирования оценки реализованы на основе наивного байесовского классификатора. В модуле для обучения осуществляется обучение классификатора на основе обучающей коллекции данных.

3 Парсинг

Парсинг представляет собой обработку информации, расположенной на страницах сайтов и выделение из нее необходимых данных. Процесс парсинга выполняется специальной программой-парсером. Программа-парсер быстро изучит большое количество сайтов, аккуратно отделит нужную информацию от программного кода и безошибочно выберет нужную информацию. Парсер предоставляет информацию в определенном виде, который задается разработчиком программы.

Весь процесс парсинга можно разделить на несколько этапов:

1) получение исходного кода интернет страницы;

2) проведение анализа полученных данных, путем извлечения требуемой информации из кода разметки;

3) обработка и преобразование данных в необходимый формат для дальнейшего использования;

4) генерация результата и его вывод в файл или на экран – завершающий этап парсинга.

Для реализации парсинга была выбрана библиотека AngleSharp, которая представляет собой быстрый парсер с удобным API. API построен на базе официальной спецификации по JavaScript HTML DOM. DOM описывает структуру веб-станицы в виде древовидного представления и предоставляет возможность получить доступ к отдельным элементам веб-станицы.

4 Эмоциональная оценка текста

Для осуществления задачи эмоциональной оценки текста был использован подход на основе машинного обучения с учителем [Heerschop, 2011]. Данный подход является одним из наиболее распространенных подходов, применяемых в исследованиях. В основе этого подхода лежит обучение машинного классификатора на предварительно собранной коллекции текстов, каждому из которых заранее указывается правильный тип тональности. После чего полученная модель используется для анализа новых документов.

В качестве алгоритма классификации можно выбрать наивный байесовский классификатор, являющийся одним из самых простых в тестировании. В основе наивного байесовского классификатора лежит теорема Байеса, которая позволяет определить вероятность события при условии, что произошло другое взаимозависимое событие:

$$P(c|d) = \frac{P(d|c) \cdot P(c)}{P(d)}, \quad (1)$$

где $P(c|d)$ – вероятность, что документа d принадлежит классу c; $P(d|c)$ – вероятность встретить документ d среди всех документов класса c; $P(c)$ – безусловная вероятность встретить документ класса c в обучающей

выборке документов; $P(d)$ – безусловная вероятность встретить документ класса *d* в обучающей выборке документов.

Целью классификации является отнесение документа к какому-то классу. Для определения наиболее вероятного класса классификатор использует оценку апостериорного максимума [Осокин, 2014]. Другими словами, наиболее вероятностный класс, которому принадлежит документ тот, при котором условная вероятность принадлежности документа необходимому классу максимальна.

$P(d)$ является константой, поэтому его можно опустить, и оценка апостериорного максимума будет вычисляться:

$$C_{map}=argmax_{c\in C}\frac{P(d|c)\cdot P(c)}{P(d)}, \quad (2)$$

где C_{map} – оценку апостериорного максимума.

Байесовский классификатор использует предположение об условной независимости слов. Под независимостью слов понимают предположение о том, что разные слова в тексте появляются независимо друг от друга и позиция, на которой располагаются эти слова, не имеет значения [Narayanan, 2013]. Благодаря данному допущению модель классификатора и получила название «наивная».

Как следствие, совместная модель вероятности представлена в следующем виде:

$$P(d|c) \approx P(\omega_1|c)P(\omega_2|c) \dots P(\omega_n|c) = \prod_{i=1}^{n} P(\omega_i|c), \quad (3)$$

где $P(\omega_i|c)$ – условные вероятности всех слов, входящих в документ.

Оценка вероятностей $P(c)$ и $P(\omega_i|c)$ осуществляется на обучающей выборке. Вероятность класса можно оценить следующим образом

$$P(c) = \frac{D_c}{D}, \quad (4)$$

где D_c – количество документов, принадлежащих классу *c; D* – общее количество документов в обучающей выборке.

Оценка вероятности слова в классе определяется как:

$$P(\omega_i|c) = \frac{W_{ic}}{\Sigma_{i'\in V} W_{i'c}}, \quad (5)$$

где W_{ic} – описывает, сколько раз слово встречается в файлах класса *c*; $W_{i'c}$ – количество слов во всех документах класса *c*.

Если при классификации встретиться слово, которого нет в наборе обучающей выборке, то вероятность $P(\omega_i|c)$ будет равна нулю, и документ с этим словом невозможно будет отнести ни к одному из классов. В связи с этим необходимо внести поправки во все оценки вероятностей, что бы ни одна из вероятностей не была равна нулю. Для решения этой проблемы используется аддитивное сглаживание (сглаживание Лапласа).

Идея сглаживания заключается в том, что к частоте каждого слова прибавляется единица:

$$P(\omega_i|c) = \frac{W_{ic}+1}{\sum_{i' \in V}(W_{i'c}+1)} = \frac{W_{ic}+1}{|V|+\sum_{i' \in V} W_{i'c}}. \quad (6)$$

Логически данный подход смещает оценку вероятностей в сторону менее вероятных исходов. Таким образом, слова, которых нет на этапе обучения модели, получают пусть маленькую, но не нулевую вероятность.

Перед классификацией данных необходимо провести предварительную обработку данных, которую можно разделить на четыре этапа:

1) токенизация файла;
2) стемминг;
3) нормализации лексем;
4) игнорирование распространенных терминов.

Токенизация представляет собой процесс выделения из текста отдельных слов, чисел и знаков пунктуации. На этом этапе была убрана вся пунктуация из текста, удалены переносы слов.

Стемминг представляет собой процесс нахождения основы слова для заданного исходного слова. Цель стемминга – приведение слов, имеющих одинаковую основу к единой форме, что в свою очередь приводит к уменьшению размерности задачи. В системе анализа мнений при реализации стемминга был использован алгоритм усечения окончаний в словах с использованием регулярных выражений.

Нормализация лексем представляет собой процесс приведения лексем к канонической форме. Этот этап необходим для того, чтобы устранить различия между последовательностями символов, которые являются эквивалентными. В работе используется правила отображения, которые удаляют символы, относящиеся к цифрам или буквам. Эти правила позволяют возникать классам эквивалентности неявно, что делает термины, которые в результате применения этих правил становятся идентичными, относятся к одному и тому же классу эквивалентности.

Распространенные слова, не представляющие ценности для решаемой задачи необходимо игнорировать, поэтому это еще одним способом обработки текста является игнорирование распространенных терминов.

Система анализа мнений позволяет на выбор получить сведения, как о турфирме, так и об отелях мира. Для оценки отеля необходимо заполнить информацию о стране и названии отеля, а для оценки турфирмы – информацию о городе и названии турфирмы. Тональность текста оценивается по шкале от –10 до 10. Результат работы системы анализа мнений показан на рисунке 2.

Рисунок 2 – Результат работы системы анализа мнений

5 Заключение

В ходе проведенной работы, описанной в данной статье, была описана система анализа мнений, каждый отзыв был оценен по шкале от –10 до 10 и пользователю был представлен результат оценки в виде круговой диаграммы, отображавшей соотношений отрицательных и положительных отзывов. Система позволяет пользователям получать требуемую информацию быстро, не затрачивая на это свое время.

Система анализа мнений необходима для определения отношения пользователей сети Интернет к отелям стран мира и турфирмам многих городов Беларуси и России. Система позволяет пользователю без труда получить полную информацию об интересующем его отеле или турфирме. Быстрая работа классификатора позволяет быстро и точно проводить оценку отзывов.

Список литературы

[Меньшиков и др., 2012] Меньшиков И. Л., Кудрявцев А.Г. Обзор систем анализа тональности текста на русском языке // Молодой ученый. – 2012. – № 12. – С. 140 – 143.

[Осокин, 2014] Осокин В.В. Анализ тональности русскоязычного текста // Интеллектуальные системы. Теория и приложения. 2014. – № 3. – С. 163 – 172.

[Heerschop, 2011] Heerschop B. Polarity analysis of texts using discourse structure // Proceedings of the 20th ACM international conference on Information and knowledge management. – 2011. – pp. 1061 – 1070.

[Narayanan, 2013] Narayanan, V. Fast and accurate sentiment classification using an enhanced Naive Bayes model // Intelligent Data Engineering and Automated Learning. 2013. – Vol. 8206. – P. 194 – 199.

[Pang, 2008] Pang, B. Opinion Mining and Sentiment Analysis // Philadelphia: Now Publishers Inc, 2008. P. 35-80.

AUTOMATION OF ANALYSIS OF USERS 'OPINIONS ON SERVICES IN THE SPHERE OF TOURISM BASED ON REFERENCES IN THE INTERNET NETWORK

Burduk E.A. (*burduk_z2@bk.ru*)
Gomel State Technical University named after P.O. Sukhoi, Republic of Belarus, Gomel

The article contains a description of the system that searches for hotel and travel agency reviews and determines the emotional evaluation of the information obtained by analysis that allows you to extract emotionally colored vocabulary from the text and determine the emotional attitude of the authors to the object of research.

Keywords: Parsing, classifier, sentiment analysis, the system of the analysis of opinions