Анализ тональности текста на русском языке при помощи графовых моделей

И. Л. Меньшиков

unkmas@gmail.com

УРФУ, Екатеринбург, Россия

Аннотация. Статья посвящена вопросу анализа тональности текста на русском языке при помощи графовых моделей. Описан и экспериментально исследован алгоритм решения данной задачи.

Ключевые слова: компьютерная лингвистика; обработка естественного языка; анализ тональности.

1 Введение

Количество генерируемого пользователями контента в интернете выросло экспоненциально за последнее десятилетие. Пользователи пишут на форумах, в блогах, оставляют комментарии на множестве страниц и пользуются социальными сетями. Согласно исследованиям Всероссийского центра изучения общественного мнения, количество россиян, регулярно (не реже раза в месяц) пользующихся интернетом выросло с 38% в 2010 г. до 55% в 2012 г. Число зарегистрированных в социальных сетях россиян за эти 2 года (с 2010 по 2012 гг.) также значительно возросло – с 53% до 82%.[1] Весь этот контент несет в себе огроное количество информации, которою мы регулярно получаем, анализируем и используем.

Для владельцев информационных ресурсов жизненно важно знать мнение пользователей — будь это оценка людьми нового про-

дукта в интернет магазине или отношение к свежей новости на новостном сайте. Для простого пользователя интернет-магазина будет интересна информация о том, насколько другим покупателям понравился или не понравился конкретный товар.[2] Однако, вся эта информация представляет собой большой объем текстовых данных. Для того, чтобы прочитать их и проанализировать требуется много времени. Для решения этой проблемы необходимы системы анализа тональности текста.

Анализ тональности текста — это класс методов, предназначенный для выявления эмоций в тексте. Он позволяет охарактеризовать текст по его эмоциональной окраске — положительный, отрицательный, нейтральный текст. Кроме того, возможно определение силы тональности, субъекта/объекта тональности и многих других характеристик текста.

2 Предпосылки

Исходными данными для этой работы послужило предположение о том, что не все слова в тексте равнозначны. Какие-то слова имеют больший вес, более значимы для данного текста. Какие-то слова — менее значимы. Очевидно, что более значимые слова будут оказывать более сильное влияние на общую тональность текста.

При этом слова, имеющие высокую силу тональности, могут оказывать большее влияние на тональность, нежели слова, имеющие больший вес, однако меньшую силу тональности.

3 Алгоритм

Анализ тональности происходит в несколько этапов:

1) Построение графа на основе текста
2) Ранжирование его вершин
3) Классификация найденных слов
4) Вычисление результата

Этапы 1 и 2 детально описаны в статье [3]. Этапы 3 и 4 описаны далее в данной статье.

4 Классификация слов

Для опредения классов слов и определения силы их тональностей используется тональный словарь.

В данной работе, для слов использовалось два класса:

— Положительное слово

— Отрицательное слово

Оценка силы тональности проводилось по шкале от 1 до 5 для каждого класса.

5 Вычисление результата

Для получения итогового результата необходимо получить две оценки: оценку положительной составлящей текста и оценку его негативной составляющей.

Для оценки положительной составляющей необходимо подсчитать сумму тональностей всех найдннх положительных терминов текста, с учетом их веса:

$$P = \sum_i TR(i) * Q(i), \tag{1}$$

где P — оценка положительнйо составляющей текста, TR(i) - вес слова, Q(i) - сила ег о тональности.

Аналогичным образом вычисляется значение отрицательной составляющей текста (N).

Для итоговой оценки тональности текста вычисляется отношение этих оценок:

$$T = P/N \tag{2}$$

Текст, в котором значение T близко к еденице, считается нейтральным. Текст, в котором значение T больше или меньше еденицы, считается положительным или, соответственно, отрицательным.

Кроме того, возможно разделение текстов на большее число классов — если T незначительно превосходит еденицу, текст считается слабо-положительным, если же он намного больше еденицы — сильно-положительным.

6 Результаты работы алгоритма

Для исследования работы алгоритма выполнена обработка ряда текстов, размеченных экспертами.

Было выбрано 40 текстов, каждый из которых был отмечен экспертами как положительный или отрицательный. Проверялось соответствие результатов, выданных системой результатм оценки экспертов. Система правильно обработала 66% текстов.

Для оценки результатов работы системы, тексты были разбиты на три группы, в соответствии с количеством найденных терминов.

Тексты, в которых обнаружено менее 30 терминов считаются мелкими, от 30 до 70 — средними и более 70 — крупными. Результаты приведены в Табл 1, где приняты следующие обозначения: C — количество обнаруженных в тексте терминов, A — отношение числа правильно проанализированных текстов данной группы к общему числу текстов данной группы, T — доля терминов, несущих эмоциональную окраску среди всех найденных терминов. Для получения величины T для расчетов брались средние значения среди группы, σ — стандартное отклонение T.

Табл. 1. Результаты работы системы

C	σ	T	A
< 30	18%	2,3%	75%
30 - 70	12%	5,8%	57%
> 60	14%	6,1%	71%

Как видно из вышестоящей таблицы, лучшие результаты система показала для коротких и крупных текстов, с небольшим снижением точности на средних.

В ряде случаев, система выдавала ошибку из-за отсутствия ряда эмоциионально окрашенных слов в тональном словаре. При использовании более полного тонального словаря точность работы системы повысится.

Однако, система не способна корректно обработать некоторые тексты. Например, отрицательный текст, написанный положительными словами с большим количеством отрицаний, не будет правильно обработан, так как слова-отрицания (“не”, “никогда” и т.д.) в данном алгоритме не учитываются.

7 Заключение

Описанный алгоритм показал хорошие результаты, которые могут быть улучшены путем. У алгоритма выявлены недостатки, трудные для устранения, однако точность работы системы можно улучшить.

Пути улучшения работы системы:

— расширение тонального словаря

— учет не только самих слов, но и взаимосвязей между ними

Список источников

1. РИФ+КИБ: Тренды Рунета-2012: всегда и везде быть в сети [Электронный ресурс]: Всероссийский центр изучения общественного мнения. – Режим доступа: http://wciom.ru/index.php?id=270&uid=112746 28.11.2012
2. Bo Pang, Lillian Lee: Opinion Mining and Sentiment Analysis // Journal Foundations and Trends in Information Retrieval. 2008. С. 1–135.
3. Усталов Д. А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей // Теория графов и приложения = Graphs theory and applications : материалы конференции. — Екатеринбург : Изд-во Урал. ун-та, 2012. — С. 62–69.