Анализ текстовых данных с применением конверсационного анализа

И.А. Рыцарев[1,2]

[1]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001
[2]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В данной работе предлагается алгоритм анализа текстовых данных на основе конверсационного анализа. В настоящее время, естественные языки динамично развиваются. В разговорный язык постоянно приходят новые смысловые единицы. В этих условиях, цепочки графов зависимостей смысловых единиц постоянно перестраиваются. В данной работе предлагается метод определения синонимов на основе конверсационного анализа. Предложенный метод был протестирован на данных, которые были собраны из социальных сетей.

1. Введение

В настоящее время социальные сети переживают бурный рост: каждый день их пользователи отправляют миллиарды сообщений и оставляют миллионы комментариев под интересующими их записями и постами. Их анализ имеет огромное значения во многих сферах бизнеса. К примеру, невозможно переоценить влияние интернет-маркетинга на продвижение товаров и услуг на рынке. Однако, для эффективного использования данных механизмов необходимо чётко понимать запросы пользователей. Источником такой информации как раз и могут служить материалы, публикуемые пользователями социальных сетей, а также формируемые в результате их обмена связи между пользователями и целые сообщества. [1] Таким образом, рассматриваемая в рамках данной работы задача определения близости текстовых еденицы в социальной сети Вконтакте с использованием технологии BigData является, несомненно, актуальной задачей, решение которой имеет также большое научное значение в сфере анализа данных.

2. Сбор данных с социальных сетей

Источником данных для исследования была выбрана социальная сеть Вконтакте. Это было сделано по следующим причинам:

- сеть предоставляет открытый доступ к своим данным (нет ограничения на доступ к данным сервера);
- социальная сеть является самой популярной социальной сетью в России, и пятой по популярности в мире;
- Вконтакте - это полноценная социальная сеть (в отличии от Twitter и Instagram, которые являются микроблогами), в которой реализована возможность создавать тематические сообщества, представляющие особый интерес для данной работы.

В рамках данного исследования был разработан собственный программный комплекс на языке программирования Python, содержащий модуль авторизации, модуль сбора данных, модуль фильтрации. Данный программный комплекс позволяет собирать данные и фильтровать их с целью выделения только необходимой информации [2, 3].

В рамках данного исследования при помощи разработанного программного комплекса было собрано более 5000 постов и более 170000 комментариев к ним из двух наиболее популярных сообществ города Самара (Подслушано Самара, Услышано Самара)..

3. Определение близости текстовых единиц на основе конверсационного анализа

Конверсационный анализ, т.е. изучение структур и формальных свойств языка, рассматриваемого в его социальном использовании, имеет отношение ко всем основным направлениям этно- методологических исследований.

Изначально конверсационный анализ предполагал изучение именно и только устной бытовой речи, более того, только разговоров между несколькими собеседниками. Г. Сакс, создатель метода, привлек внимание ученых к тому, что разговоры центральны дня социального мира.

Разговор прежде всего нуждается в организованности, подразумевает наличие порядка, который не надо постоянно вновь объяснять в ходе обмена репликами. Порядок также необходим для того, чтобы произносимое было понятно всем участникам беседы. В разговоре проявляется социальная, интерактивная компетентность людей, стремящихся объяснить свое поведение, а также проинтерпретировать поведение собеседников. Внутри локальных секвенций разговора, и только там, социальные институты окончательно «проговариваются в существование». В результате мельчайшие и па первый взгляд незначимые детали беседы оказываются на самом деле средством актуализации важнейших социальных институтов.

Целью конверсационалистов является описание *социальных практик и ожиданий,* на основе которых собеседники конструируют свое собственное поведение и интерпретируют поведение другого.

Конверсационный анализ дает преимущество случаям в противовес идеализации, неизбежно связанной, с точки зрения Гарфинкеля и Сакса, с любым теоретическим обобщением. По их мнению, идеализация мешает научному развитию, так как любая типология слабо связана с содержанием реальных случаев, на которых она якобы основывается. Сакс стремился развить такой метод анализа, который позволял бы оставаться на уровне первичных данных, сырого материала, специфических, единичных событий человеческого поведения. В противовес классической социологии, он утверждал, что детали любой спонтанной человеческой интеракции строго организованы - до степени, позволяющей их формальное описание.

Исходя из вышеизложенных предпосылок, особенности конверсационного анализа можно сформулировать следующим образом. Во-первых, этот метод следует за данными, т.е. анализ базируется на эмпирии без привлечения (по возможности) заранее сформулированных гипотез. Во-вторых, мельчайшие детали текста рассматриваются как аналитический ресурс, а не как помеха, которую надо отбросить. В-третьих, авторы метода убеждены, что порядок в организации деталей повседневной речи существует не только для исследователей, по и - прежде всего - для людей, конструирующих эту речь. [4]

Эта идея стала основой для исследования. Изначально было сделано предположение, что на большом наборе данных две текстовые единицы имеют похожие вектора дистанций употреблений V (вектор, который показывает, как две текстовые единицы соотносятся между собой в рамках данных, где, в качестве метрик, выступает индекс i (указывает дистанцию между единицами) и V_i (количество сочетаний между единицами, V_0 – общее количество употреблений двух текстовых единиц в пределах одного предложения).

Собранные данные из социальной сети Вконтакте были предобработаны, каждая текстовая единица была приведена к нормальной форме (для этого использовался пакет pymorphy2). Затем был произведен преданализ данных с целью извлечения необходимой статистики (WordCount, максимальная длина предложения). Следующим шагом стало составление матрицы дистанций.

Для расчета дистанций между двумя векторами использовалось косинусное расстояние:

$$distance = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Полученные результаты представлены на рисунке 1.

гендерный	половой	0.047420655584319626
петербургский	щербатый	0.05362810814737151
местный	чуждый	0.057190958417936644
пластический	ужасный	0.057190958417936644
булевый	переменный	0.07417990022744858
базовый	переносный	0.0871290708247231
базовый	системный	0.10433141049703976
дизбалансный	однообразный	0.10557280900008414
леденящий	неописуемый	0.10557280900008414
добрый	ласковый	0.11808289631180313
демографический	материнский	0.12294198069297069
весь	который	0.1323941314820628
бойцовский	псиный	0.14188366967896682
конструктивный	толковый	0.14719713457755823
немногочисленный	северокорейский	0.14719713457755823
жаркий	симпатичный	0.17497135267460984
глухой	слепой	0.18350341907227397
который	свой	0.18678856996069937
больной	тяжёлый	0.1970449314530338
который	этот	0.20425166740679124
импровизационный	спорный	0.20943058495790523
желанный	материнский	0.2254033307585166
оперативный	рекламный	0.2294482496288779
один	свой	0.24773882819618875
свой	такой	0.24790689577395286
исторический	художественный	0.2639354343699736
большой	весь	0.2826676724548497
вкусный	питьевой	0.2928932188134524
гастрономический	незабываемый	0.2928932188134524
грязный	потный	0.2928932188134524

Рисунок 1. Результат расчета дистанций между векторами дистанций.

Представленные на рисунке 1 рассчитанные дистанции отфильтрованы по значению дистанции (0 – близко, 1 – далеко). Предложенные пары слов можно (условно) разделить на три категории (предложенная интерпретация результатов и разделение не является точным, а является лишь точкой зрения автора статьи):

- Темно-серые – наиболее точные совпадения (40%);
- Белые – пару слов можно (условно) считать синонимами (33%);
- Серые – словосочетания-антонимы (27%).

Результаты работы предложенного подхода позволяют сделать предположение, что при анализе текстовых данных можно легко построить граф взаимозаменяемости слов и при дальнейшем анализе использовать его с целью извлечения контекстного смысла из набора данных.

4. Заключение

В данной работе было проведено исследование возможности применения конверсационного анализа для анализа текстовых данных социальных сетей. Исследование показало, что данный подход имеет место при анализе контекста с целью построения логических цепочек между текстами. Основной проблемой является интерпретация результатов так как закономерности могут быть неявными и варьироваться в зависимости от контекста употребления текстовых единиц. В дальнейшем автор планирует продолжить исследование данной области с применением подходов, основанных на машинном обучении и другими методами области NLP.

5. Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (№ 18-37-00418, № 19-29-01135, № 19-31-90160) и Министерства науки и высшего образования Российской Федерации в рамках выполнения государственного задания Самарского университета и ФНИЦ «Кристаллография и фотоника» РАН.

6. Литература

[1] Рыцарев, И.А. Исследование и анализ сообщений пользователей социальных сетей с использованием технологии BigData / И.А. Рыцарев, А.В. Куприянов, Д.В. Кирш // Сборник трудов ИТНТ-2019 – Самара: Новая техника, 2019. – С. 748-752.

[2] Мухин, А.В. Определение близости групп в социальных сетях на основе анализа текста с использованием больших данных / А.В. Мухин, И.А. Рыцарев // Сборник трудов ИТНТ-2019 – Самара: Новая техника, 2019. – Т. 4. – С. 757-760.

[3] Рыцарев, И.А. Кластеризация медиаконтента из социальных сетей с использованием технологии bigdata / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927.

[4] Исупова О.Г. Конверсационный анализ: представление метода //Социология: методология, методы, математическое моделирование (4М). – 2002. – № 15. – С. 33-52.

Text data analysis using conversion analysis

I.A. Rytsarev[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. In this paper, we propose an algorithm for analyzing text data based on conversion analysis. Currently, natural languages are developing dynamically. New semantic units constantly come into spoken language. Under these conditions, chains of dependency graphs of semantic units are constantly being rebuilt. In this paper, we propose a method for determining synonyms based on conversion analysis. The proposed method was tested on data that was collected from social networks.