# Определение близости групп в социальных сетях на основе анализа текста с использованием больших данных

**А.С. Мухин[1], И.А. Рыцарев[1,2]**

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

**Аннотация.** Статья посвящена определению близости групп в социальных сетях. Объектом исследования были выбраны данные социальной сети Вконтакте. В рамках работы были собраны, обработаны и проанализированы текстовые данные сообществ социальной сети Вконтакте. Для преодоления проблем связанных с превышением лимитов, установленных социальной сетью были проведены исследования в области оптимизации сбора данных социальной сети. Был разработан программный инструмент, который обеспечивает сбор и последующую обработку необходимых данных из указанных ресурсов. Были исследованы и применены существующие алгоритмы текстового анализа большого объема данных.

## 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. Таким образом, рассматриваемая в рамках данной работы задача определения близости групп 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 из социальной сети

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 комплексна 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.

В рамках данного исследования при помощи разработанного программного комплекса было собрано более 8000 постов и более 280000 комментариев к ним из двух наиболее популярных сообществ города Самара (Подслушано Самара, Услышано Самара) и сообщества студентов Самарского Университета (Подслушано Самарский университет).

Потоковые данные, полученные из социальных сетей, содержат в себе множество служебной информации. Для дальнейшего анализа важны лишь те данные, которые представляют интерес, поэтому необходимо отделить служебную информацию от нужной. Модуль предобработки программного комплекса производит структурирование полученных данных, а также фильтрует «полезные» и служебные поля.

## 3. Определение близости групп с применением технологии BigData

Для определения близости групп были рассмотрены несколько метрик для сравнения словарей: евклидово расстояние, расстояние городских кварталов и расстояние Махаланобиса. Выбор пал на евклидово расстояние, так как оно является наиболее подходящим для проведения данного эксперимента по следующим критериям:

1) Наиболее применяемая и универсальная метрика;
2) Евклидово расстояние вычисляется по исходным, а не по стандартизованным данным.

Для вычисления данной метрики между группами были сформированы вектора признаков, при помощи объединения двух и более словарей в один общий. Каждому слову в словаре был присвоен свой вес, тем самым каждая группа принимала вид вектора признаков(слов) с собственными весами. В данной работе в качестве веса было принято решение использовать частоту употребления слова в тексте.

Подобный подход для подсчета весов слов в словарях с использованием традиционных методов и технологий при увеличении объемов и количества анализируемых словарей требует огромных вычислительных ресурсов и занимает длительное время, поэтому было принято решение использования технологии BigData и вычислительных кластеров для выполнения данной работы. На этом этапе был разработан алгоритм, который с применением технологии MapReduce отбрасывал неинформативные части словаря (слова менее трех и более пятнадцати символов) а также подсчитывал частоту употребления слов в тексте. В итоге, были получены три словаря, элементы которого обладали собственными весами, один из которых представлен на рисунке 1.

На следующем шаге было принято решение использовать два словаря (для групп Подслушано Самара и Услышано Самара) для получения общего словаря, а оставшийся (для группы Подслушано Самарский Университет) использовать для тестового подсчета. Общий словарь состоял из пересекающихся слов с пересчитанными весами по формуле 2.

$$g(g_1, g_2) = \left(\frac{g_1 + g_2}{n}\right) \quad (1)$$

где: $g(g_1, g_2)$ - вес в искомом словаре; $g_1$- вес слова в первом словаре; $g_2$- вес слова во втором словаре.

Последним действием были подсчитаны расстояния между получившимся словарем сначала до групп, на основе которого он был создан, а после расстояние до тестовой группы. Для замера данной величины использовалась формула евклидового расстояния между двумя группами.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

```
('что', 1780)
('пожалуйст', 1078)
('анон', 881)
('карт', 625)
('помог', 584)
('ребят', 570)
('сказал', 568)
('можн', 489)
('есл', 401)
('был', 368)
('только', 303)
('андрей', 294)
('скажит', 291)
('город', 234)
('наход', 231)
('самар', 228)
('потер', 217)
('спасиб', 213)
('марат', 207)
('человек', 207)
('подскаж', 199)
('когда', 197)
('сказал', 195)
('говорил', 191)
('теперь', 184)
('очень', 183)
```

**Рисунок 1**.Часть полученного словаря для группы Подслушано Самара.

Результаты подсчета представлены в таблице 1.

**Таблица1.** Результаты подсчета расстояний между группами.

| Название | Евклидово расстояние |
|---|---|
| Подслушано Самара | 188,32 |
| Услышано Самара | 173,11 |
| Подслушано Самарский Университет | 365,98 |

Взглянув на результаты, можно сделать вывод, что расстояния между двумя первыми группами очень приближены. Это говорит о том, что общий словарь составлен достаточно точно и хорошо отражает контекст сообщений в представленных группах. Проанализировав расстояние до тестовой группы можно заметить близость всех трех значений, но также видно, что оно увеличилось относительно двух других в два раза. Можно предположить, что данная группа немного отлична от двух других.

## 4. Заключение

В ходе работы был разработан комплекс программных модулей, позволяющих найти расстояние между сообществами социальной сети Вконтакте. В результате работы был получен общий словарь слов, на основе которого были определены степени близости между 3 сообществами. В дальнейшем результаты работы могут быть использованы в разработке алгоритмов определения близости групп и сообществ большего объема с использованием технологии BigData.

**5. Литература**

[1] Khotilin, M.I. Visualization and Cluster Analysis of Social Networks / M.I. Khotilin, A.V. Blagov // CEUR Workshop Proceedings. – 2016. – Vol. 1638. – P. 843-850.

[2] Xu, X. Scan: a structural clustering algorithm for networks // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2007. – P. 824-833.

[3] Blagov, A. Big Data Instruments for Social Media Analysis / A. Blagov, I. Rytcarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015. – P. 179-184.

# Determining the proximity of groups in social networks based on text analysis using big data

**A.S.Mukhin[1], I.A.Ritsarev[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.** The article is devoted to the definition of such groups in social networks. The object of the study was selected data social network Vk. Text data was collected, processed and analyzed. To solve the problem of obtaining the necessary information, research was conducted in the field of optimization of data collection of the social network Vk. A software tool that provides the collection and subsequent processing of the necessary data from the specified resources has been developed. The existing algorithms of text analysis, mainly of large volume, were investigated and applied.