DOI: 10.15593/2499-9873/2021.4.05
УДК 39.04.03

**Т.А. Шестаков, Ю.А. Леонов,
А.А. Кузьменко, А.С. Сазонова, Р.А. Филиппов**

Брянский государственный технический университет, Брянск, Россия

# ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ИНФОРМАЦИИ О ПОЛЬЗОВАТЕЛЯХ СОЦИАЛЬНЫХ СЕТЕЙ

Большую роль в информатизации общества стали играть социальные сети. Специалисты со всего мира исследуют данные социальных сетей для решения различных задач, таких как создание востребованного контента, проведение рекламных кампаний, удовлетворение информационных потребностей социума, обеспечение государственной безопасности и др. Под анализом социальных сетей понимается решение таких задач, как определение тональности текста, определение целевого портрета аудитории, поиск ассоциативных правил, расчет показателей эффективности деятельности сообщества и визуализация данных.

Рассмотрена актуальность решения задачи, проведен анализ результатов предшествующих работ. Изучена реакция аудитории на контент, построен целевой портрет подписчиков различных сообществ, исследована зависимость между интересами пользователей.

Исходными данными исследования являются социальные сети, а точнее, информационные сообщения, мнения, подсети и сообщества, отдельные пользователи, внешние узлы. Рассмотрена классификация систем анализа социальных сетей (таких как Brand Analytics, IQBuzz, Agorapulse, Semantic Force,Talkwalker) по следующим признакам: пользователи, методы анализа, объекты анализа, источники данных, особенности. Для определения реакции аудитории на контент был применен метод определения тональности текста посредством анализа комментариев к контенту. Метод кластерного анализа был применен для определения целевого портрета пользователей в конкретном сообществе. Для поиска закономерностей между интересами пользователя в работе был рассмотрен частотный анализ наборов элементов. Поиск ассоциативных правил проводился с помощью алгоритма Apriori. Результаты работы представлены в виде графиков и диаграмм.

В ходе работы был использован комплексный подход к решению задач, что позволило создать автоматизированную информационно-аналитическую систему, которая может использоваться как аналитический инструментарий в данной сфере.

**Ключевые слова:** данные, методы анализа данных, анализ соцсетей, целевой портрет пользователя соцсети, зависимости между интересами пользователей соцсетей, реакция аудитории на контент, системы аналитики социальный сетей, анализ наборов элементов, ключевых показателей эффективности, ассоциативные правила.

**T.A. Shestakov, Ju.A. Leonov,
A.A. Kuzmenko, A.S. Sazonova, R.A. Filippov**

Bryansk State Technical University, Bryansk, Russian Federation

# INTELLECTUAL ANALYSIS OF INFORMATION ABOUT USERS OF SOCIAL NETWORKS

Social networks began to play an important role in the informatization of society. Experts from all over the world are researching social network data to solve various tasks, such as creating popular content, conducting advertising campaigns, meeting the information needs of society, ensuring state

security, etc. The analysis of social networks is understood as the solution of such tasks as determining the tonality of the text, determining the target portrait of the audience, searching for associative rules, calculating community performance indicators and data visualization.

The article considers the relevance of solving the problem, analyzes the results of previous work, examines the audience's reaction to content, builds a target portrait of subscribers of various communities, examines the relationship between user interests.

The initial data of the study are social networks, or rather informational messages, opinions, subnets and communities, individual users, external nodes.The paper considers the classification of social network analysis systems (such as Brand Analytics, IQBuzz, Agorapulse, Semantic Force, Talkwalker) according to the following criteria: users, analysis methods, objects of analysis, data sources, features.To determine the audience's reaction to the content, the method of determining the tonality of the text was applied by analyzing comments to the content. The cluster analysis method was used to determine the target profile of users in a particular community. To find patterns between the user's interests in the work, the frequency analysis of sets of elements was considered. The search for associative rules was carried out using the Apriori algorithm. As a result, the works are presented in the form of graphs and diagrams.

In the course of the work, an integrated approach to solving problems was used, which made it possible to create an automated information and analytical system that can be used as analytical tools in this area.

**Keywords:** data, data analysis methods, social network analysis, target portrait of a social network user, dependencies between the interests of social network users, audience reaction to content, social network analytics systems, analysis of sets of elements, key performance indicators, associative rules.

## Введение

Благодаря бурному развитию социальных сетей стали публично доступны большие объемы персональных данных пользователей, такие как комментарии, фото, видео, аудиоинформация, геотеги и др. Это открывает большие возможности для решения исследовательских и бизнес-задач, которые сложно эффективно решать без большого объема данных.

Специалисты со всего мира используют данные социальных сетей для создания и моделирования социальных, экономических и других процессов, направленных на решение государственных задач с целью создания инструментов воздействия на данные системы, а также создания аналитических систем и бизнес-приложений.

Создание таких систем имеет ряд особенностей и проблем, которые необходимо решать. Первая сложность создания таких систем – это большие объемы данных, что является как достоинством, так и недостатком. Большие объемы позволяют получить более точные результаты исследований, но требуют построения сложной распределенной архитектуры системы, позволяющей увеличивать эффективность системы пропорционально добавляемой вычислительной мощности. Вторая проблема – обработка и хранение данных социальных сетей требует разработки специальных алгоритмов, позволяющих учитывать спе-

цифику предметной области, а также инфраструктурные решения. Имеются и другие проблемы, связанные с приватностью данных, ограничением доступа к данным, слабой структурированностью данных.

Под анализом социальных сетей понимается решение таких задач, как определение тональности текста, целевого портрета аудитории, поиск ассоциативных правил, расчет показателей эффективности деятельности сообщества и визуализация данных.

Цель научного исследования – получение аналитической информации для проведения эффективной рекламной компании в социальных сетях. Задачи системы интеллектуального анализа данных – сбор, мониторинг и актуализация данных социальных сетей, а также проведение оперативного и интеллектуального анализа.

## 1. Теория

Существуют различные зарубежные и отечественные системы мониторинга и анализа данных в социальных сетях.

Системы анализа социальных сетей можно классифицировать по следующим признакам:

1. Методы анализа данных. В основном выделяют два класса методов, используемых в анализе социальных сетей: метод статистического анализа (СА) и метод анализа графов. Для проведения семантического анализа текста и анализа тональности текста (АТТ) используются методы классификации, для определения целевой аудитории – статистические методы и методы кластеризации. Визуальный анализ (ВА) используется для демонстрации полученных данных и зависимостей. Также часто имеется возможность поиска по ключевым словам (ППКС) для последующего анализа связанного контента. Наличие ретроспективного анализа (РА) позволяет рассматривать динамику изменения объектов.

2. Объекты анализа. Система может анализировать различные объекты социальной сети: информационные сообщения, мнения, подсети и сообщества, отдельных пользователей, внешние узлы.

3. Набор источников данных. Чем больше источников данных имеет система, тем более точными могут быть результаты исследований с помощью таких технологий, как Big Data и глубинные нейронные сети.

4. Пользователи системы. В зависимости от целевой аудитории системы могут отличаться методы и объекты анализа. Также для ком-

мерческих организаций (КО) важным является наличие API-системы, возможность выгрузки отчетов. Для использования систем в государственных организациях (ГО) необходимо соответствовать определенным стандартам и быть включенным в единый реестр российского программного обеспечения. Для научных и образовательных учреждений (ОУ) важным фактором является возможность использования систем в научных целях, наличие хорошей документации и ценообразовательная политика компании.

5. Особенности. Каждая система имеет дополнительные характеристики, которые отличают их от конкурентов [1, 2]. Из наиболее популярных систем можно выделить: Brand Analytics, IQBuzz, Agorapulse, SemanticForce, Talkwalker. Каждая система имеет свои особенности и работает в определенной области аналитики и сбора данных (таблица).

Сравнительная таблица систем аналитики социальных сетей

| Признак | Brand Analytics | IQBuzz | Agorapulse | Semantic Force | Talkwalker |
|---|---|---|---|---|---|
| Пользователи | КО, ГО | КО | КО | КО, ГО, ОУ | КО |
| Методы анализа | АТТ, СА, ППКС | АТТ, СА, ППКС, РА, ВА | СА, ВА | АТТ, СА, ВА | АТТ, СА, ВА, РА |
| Объекты анализа | Информационные сообщения | Информационные сообщения, мнения, сообщества, пользователи, внешние узлы | Информационные сообщения, мнения, сообщества | Информационные сообщения, мнения, сообщества | Информационные сообщения, мнения, сообщества, изображения |
| Источники данных | VK, Facebook, OK, Instagram, YouTube, Telegram, СМИ | LiveJournal, VK, YouTube, Instagram, Twitter | Facebook, Twitter, LinkedIn, Google+, Instagram | Facebook, Twitter, VK, OK, YouTube | Facebook, Twitter, LinkedIn, Google+, Instagram |
| Особенности | Поддержка 67 языков, выгрузка отчетов | API, ретроспектива до 10 лет, выгрузка отчетов | CRM для сегментации аудитории, отложенный постинг | API, интеграция с Google Analytics, рубрикации текста | Поддержка 187 языков, выгрузка отчетов |

Как видно из таблицы, в основном системы разрабатываются для работы с коммерческими организациями. Системы, состоящие в Едином реестре российского ПО, работают также с государственными организациями и научными учреждениями.

В рассмотренных системах часто применяются следующие методы анализа: анализ тональности текста, статистический и визуальный анализ. Инновационные системы внедряют методы анализа с помощью поиска по ключевым словам, ретроспективного анализа, анализа изображения. Наиболее популярными объектами анализа являются информационные сообщения, мнения и сообщества [3].

Выбор источников данных зависит от регионального расположения целевой аудитории компании и сферы аналитической деятельности. Так, системы, предназначенные для Америки и Европы, анализируют данные из Facebook, Twitter, Instagram, Google+, в то время как отечественные системы уделяют особое внимание VK, OK, YouTube, Instagram и СМИ.

## 2. Данные и методы

Материалом исследования являются социальные сети, а точнее информационные сообщения, мнения, подсети и сообщества, отдельные пользователи, внешние узлы.

Данные о сообществах (Community):

$$\begin{gathered}\text{Community} = \{\ \text{CCM},\ \text{CC},\ \text{CD},\ \text{CP}, \text{CS}\},\ \text{CC} \in \text{Categories}, \\ CP \in \text{Posts},\end{gathered} \tag{1}$$

где CCM (count of community members) – количество участников в сообществе, CC (community category) – категория сообщества, CD (community description) – описание группы, CP (community posts) – список записей, CS (community subscribers) – список подписчиков.

Пользователей (User) описывает следующий набор данных:

$$\begin{gathered}\text{User} = \{\text{UA},\ \text{US},\ \text{UCo},\ \text{UCi},\ \text{UCF},\ \text{UI},\ \text{UE}\}, \text{UCo} \in \text{Coutries}, \\ \text{UCi} \in \text{Cities}, \text{US} = \{\text{Male},\ \text{Female}\}, \text{UI} \subset \text{Categories},\end{gathered} \tag{2}$$

где UA (user age) – возраст, US (user sex) – пол, UCo (user country) – страна, UCi (user city) – город, UCF (count of friends) – количество дру-

зей, UI (user interests) – интересы пользователя, UE (user education) – образование.

Данные о записях (Post):

$$\text{Post} = \{\text{PD},\ \text{PV},\ \text{PL}, \text{PC}, \text{PR}\}, \tag{3}$$

где PD (post description) – подпись, PV (post views) – количество просмотров, PL (post likes) – количество лайков, PC (post comments) – количество комментариев, PR (post reposts) – количество репостов.

Данные о комментариях (Comment):

$$\text{Comment} = \{\text{CT},\ \text{CL}, \text{CV}\}, \tag{4}$$

где CT (comment text) – текст комментария, CL (comment likes) – количество лайков комментария, CV (comment views) – количество просмотров комментария.

Анализ данных выполняется с помощью следующих методов: Data Mining, статистический анализ, визуальный анализ и ретроспективный анализ.

Data mining – это процесс обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [4]. Для решения задач определения реакции аудитории на контент, целевого портрета подписчиков и зависимостей между интересами пользователей использовались методы Data Mining.

***Определение реакции аудитории на контент***

Одной из задач интеллектуального анализа данных в социальных сетях является определение реакции аудитории на контент. Данную задачу можно решить с помощью методов определения тональности текста посредством анализа комментариев к контенту. Решение данной задачи поможет определить настроение аудитории и будет полезно при выборе группы для рекламы товаров или услуг [5].

Данная задача может быть решена с помощью двух классов методов: методы, основанные на определении тональности в тексте по заранее составленным тональным словарям, и методы машинного обучения, такие как байесовский классификатор, метод *k*-ближайших соседей, метод опорных векторов.

1. *Тональные словари*. В методе с использованием тональных словарей по совокупности найденной эмотивной лексики текст может быть оценен по шкале, содержащей количество негативной и позитивной лексики.

2. *Байесовский классификатор (Naive Bayes)* – широкий класс алгоритмов классификации, основанных на принципе максимума апостериорной вероятности.

3. *Метод k-ближайших соседей (KNN)* – метрический алгоритм для автоматической классификации объектов или регрессии.

4. *Метод опорных векторов (SVM)* – набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа [6].

В ходе исследования выбран наиболее точный метод для определения тональности текста.

## 3. Модель

***Определение целевого портрета подписчиков сообщества***

Для определения целевого портрета пользователей в конкретном сообществе был применен метод кластерного анализа.

Для кластеризации был использован метод *k*-средних. Это метод кластерного анализа, целью которого является разделение *n* наблюдений (из пространства) на *k* кластеров.

Алгоритм *k*-means разбивает набор *x* на *k* наборов *S*1, *S*2, ..., *Sk* таким образом, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра (центр кластера). Введем обозначение – $S = \{S1, S2, ..., Sk\}$. Тогда действие алгоритма *k*-means равносильно минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров:

$$V = \min \sum_{i=1}^{k} \sum_{X \in S_i} \rho(X, \mu_i)^2, \tag{5}$$

где $\mu_i$ – центр *i* кластера, $i = 1, \ldots, k$, $\rho(X, \mu_i)$ – функция расстояния между *x* и $\mu_i$.

В качестве функции расстояния может выступать евклидово расстояние:

$$\rho(X,\mu)=\sqrt{\sum_{i=1}^{d}(x_i-\mu_i)^2}, \quad (6)$$

где $X=\{x_1,x_2,...,x_k\}$, $x\in R^d$, $d$ – размерность вектора $X$.

Преимуществами алгоритма являются: сравнительно высокая эффективность при простоте реализации, высокое качество кластеризации, возможность распараллеливания алгоритма [7].

Аналитик вводит количество кластеров, на которые необходимо разбить аудиторию сообщества, далее в каждом кластере определяются наиболее встречаемые характеристики пользователей, такие как возраст, пол, город, интересы и др.

***Определение зависимостей между интересами пользователей***

Частотный анализ наборов элементов и изучение ассоциативных правил могут быть использованы для поиска закономерностей между интересами пользователей [8]. Одним из наиболее популярных алгоритмов поиска ассоциативных правил является Apriori.

Apriori – это алгоритм для частотного анализа наборов элементов и изучения правил ассоциации в реляционных базах данных.

Пусть дано $I$ – множество характеристик пользователя, называемых элементами, $I=\{i_1,i_1,...,i_n\}$, $D$ – множество транзакций, где каждая транзакция имеет уникальный идентификатор и $D\subseteq I$, $D=\{t_1,t,...,t_n\}$.

Ассоциативным правилом является импликация вида

$$X\to Y, \text{ где } X,Y\subseteq I. \quad (7)$$

Чтобы выбрать правило из множества всех возможных правил, используются ограничения на различные меры значимости. Наиболее известными ограничениями являются минимальный порог поддержки и минимальный порог достоверности.

Поддержка правила показывает частоту, с которой набор $X\to Y$ встречается в множестве транзакций [9]. Поддержка набора $X$ по отношению к $D$ определяется как отношение числа транзакций $t$ в базе данных, содержащих набор $X$, к общему числу транзакций:

$$S(X\to Y)=\frac{\left|\{t\in D;X\subseteq t\}\right|}{|D|}. \quad (8)$$

Достоверность правила показывает частоту, с которой в совокупности данных соблюдается $X \to Y$ [10]. Значение достоверности правила по отношению к набору транзакций $D$ является отношением числа транзакций, которые содержат как набор $X$, так и набор $Y$, к числу транзакций, содержащих набор $X$:

$$C(X \to Y) = \frac{S(X \cup Y)}{S(X)}. \tag{9}$$

Эффективность маркетинга в социальных сетях зависит от значений ключевых показателей эффективности (*KPI*) в группах рекламодателей.

Все метрики можно разделить на несколько категорий.

***Метрики для оценки динамики подписчиков***

1. Количество подписок за период (Follows).

2. Количество отписок за период (Unfollows).

3. Количество просмотров (Views) – как правило, используется суммарный показатель по всем записям сообщества за период:

$$\text{Views} = \sum_{i=1}^{n} \text{PV}_i, \text{PV}_i \in \text{CPP}, \text{CPP} \subset \text{CP}, \tag{10}$$

где $n$ – количество записей, $\text{PV}_i$ – количество просмотров $i$-й записи, CPP – множество записей за определенный период.

4. Охват (Reach) показывает количество пользователей, которые хотя бы раз контактировали с записями сообщества:

$$\text{Reach} = \sum_{i=1}^{n} \left(\text{PL}_i + \text{PR}_i + \text{PC}_i\right), \text{PL}_i, \text{PR}_i, \text{PC}_i \in \text{CP}, \tag{11}$$

где $n$ – количество записей, $\text{PL}_i$ – количество лайков в $i$-й записи, $\text{PR}_i$ – количество репостов $i$-й записи, $\text{PC}_i$ – количество комментариев к $i$-й записи.

***Метрики для оценки обратной связи от аудитории***

Метрики, отражающие реакцию пользователей на контент. Наиболее известными метриками являются лайки, комментарии и репосты.

1. Уровень привлекательности (Love Rate, LR) – среднее количество лайков в пересчете на размер аудитории:

$$\text{LR} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{PL}}{\text{CCM}} \cdot 100\,\%, \text{PL}_i \in \text{CP}, \tag{12}$$

где $n$ – количество записей, $\text{PL}_i$ – количество лайков $i$-й записи.

2. Уровень общительности (Talk Rate, TR) – среднее количество комментариев в пересчете на размер аудитории:

$$\mathrm{TR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{PC}_i}{\mathrm{CCM}}\cdot 100\,\%, \mathrm{PC}_i \in \mathrm{CP}. \qquad (13)$$

3. Коэффициент распространения (Amplification Rate, AR) – показатель, определяющий заинтересованность пользователей в теме конкретной публикации:

$$\mathrm{AR} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{PR}_i \cdot 100\,\%, \mathrm{PR}_i \in \mathrm{CP}. \qquad (14)$$

4. Коэффициент вовлеченности аудитории (Engagement Rate, ER) – высокий уровень вовлеченности пользователей говорит о качестве и востребованности ресурса:

$$\mathrm{ER} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{PL}_i + \mathrm{PR}_i + \mathrm{PC}_i\right)\cdot 100\,\%\,, \mathrm{PL}_i, \mathrm{PR}_i, \mathrm{PC}_i \in \mathrm{CP}. \qquad (15)$$

5. Коэффициент вовлеченности по охвату (Engagement Rate by Reach, ERR) – метрика показывает отношение пользователей, которые хоть раз взаимодействовали с публикациями, к просмотрам:

$$\mathrm{ERR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\mathrm{PL}_i + \mathrm{PR}_i + \mathrm{PC}_i\right)}{\mathrm{PV}}\cdot 100\,\%\,, \mathrm{PL}_i, \mathrm{PR}_i, \mathrm{PC}_i \in \mathrm{CP}. \qquad (16)$$

6. Уровень вовлеченности в пересчете на пост (Engagement Rate of Post, ER Post) – показатель позволяет оценивать привлекательность конкретной публикации:

$$\mathrm{ER\,Post} = \frac{\left(\mathrm{PL} + \mathrm{PR} + \mathrm{PC}\right)}{\mathrm{CCM}}\cdot 100\,\%\,, \mathrm{PL}, \mathrm{PR}, \mathrm{PC} \in \mathrm{CP}. \qquad (17)$$

7. Пользовательский контент (User Generated Content, UGC) – метрика позволяет оценить количество записей, созданных пользователями сообщества [10]:

$$\mathrm{UGC} = \frac{x}{n}\cdot 100\,\%, \qquad (18)$$

где $n$ – количество всех записей, $x$ – количество записей, созданных участниками сообщества.

***Метрики для оценки коммуникации со стороны SMM-специалистов***

1. Частота генерации постов (Post Rate) – количество постов, размещенных в сообществе за отчетный период:

$$\text{Post rate} = \frac{N}{x} \cdot 100\ \%, \tag{19}$$

где $n$ – количество записей за $x$ дней.

2. Среднее время отклика (Response Time) – метрика, показывающая среднее время отклика администратора на сообщения пользователей. Это важный показатель качества обслуживания и уважения к клиентам [11]:

$$\text{Response time} = \frac{1}{n} \sum_{i=1}^{n} x_i - y_i \cdot 100\ \%, \tag{20}$$

где $n$ – количество сообщений, $x_i$ – дата отправки $i$-го сообщения клиента, $y_i$ – дата ответа администрации сообщества на $i$-е сообщение.

Существуют различные методы поиска скрытых закономерностей с помощью алгоритмов и машинного обучения, но не стоит упускать возможность анализа и интерпретации данных с помощью человека. Визуальный анализ данных позволяет представить большие объемы данных в таких графических представлениях, как двумерные и трехмерные графики, таблицы и деревья решений.

Данный вид анализа имеет следующие преимущества:

- позволяет анализировать зашумленные данные, в отличие от автоматических методов, которые могут плохо работать с такими данными;
- не требует реализации сложных алгоритмов;
- интуитивно понятен.

## 4. Полученные результаты

***Определение тональности текста***

Во время исследования необходимо было выяснить наиболее точный метод определения тональности текста. Для обучения моделей был выбран набор данных RuTweetCorp, который включает в себя комментарии, распределенные на две группы: «заведомо положительные» (114,911 записи) и «заведомо отрицательные» (111,923 записи) [12, 13].

Результаты тестирования различных моделей тонального анализа текста представлены на рис. 1.

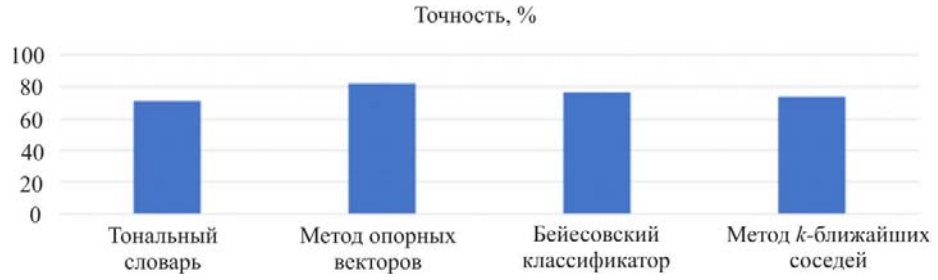


Рис. 1. Точность моделей определения тональности текста

Как видно на рис. 1, наиболее точным является метод опорных векторов с показателем 82 %, на последнем месте оказался метод, основанный на тональных словарях, с точностью 71 %. Следовательно, для определения тональности текста рекомендуется использовать метод опорных векторов.

***Определение целевого портрета пользователя***

На рис. 2 можно увидеть проекцию данных пользователей (пол, возраст, страна, город, интересы, количество подписчиков) в двумерное пространство и кластеры, на которые были разбиты данные с помощью алгоритма кластеризации *k*-средних.

Как видно на рис. 2, пользователи были разделены на два кластера, а данные о каждом пользователе были спроецированы в двумерное пространство. Как можно заметить, кластеры пересекаются – это нормальное явление при анализе пользователей одного сообщества.

На рис. 3 показано распределение кластеров между городом проживания и полом пользователя. Как видно на рис. 3, условием для максимизации расстояния до кластеров является разбиение аудитории по половому признаку.

Исходя из полученной информации, можно рассчитать средние значения по каждому кластеру и определить характеристики и размер аудитории для проведения целевой рекламной компании.

***Поиск ассоциативных правил***

Поиск ассоциативных правил проводился с помощью алгоритма Apriori. Входными данными алгоритма являлись интересы пользователей. Поиск ассоциативных правил осуществлялся на основе интересов пользователей одного из сообществ «ВКонтакте».

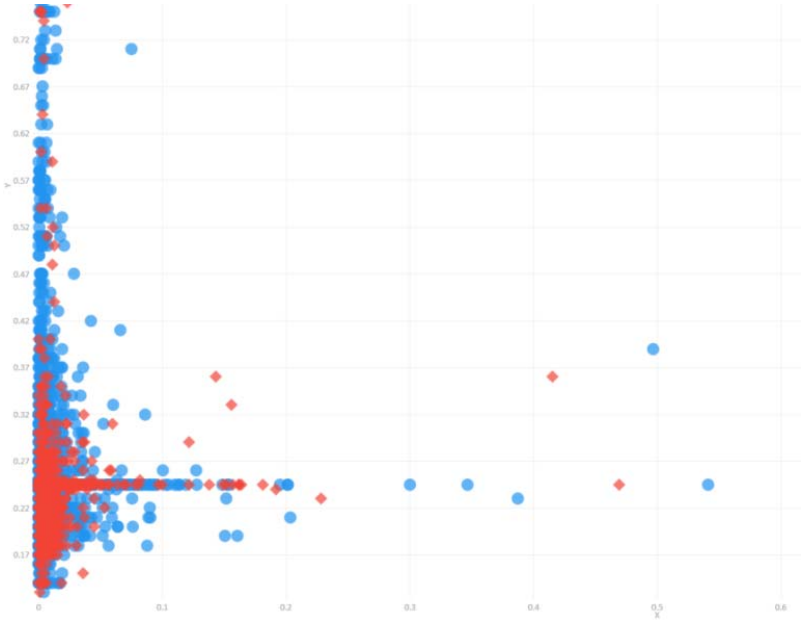

Рис. 2. Кластеризация пользователей сообщества в двумерном пространстве

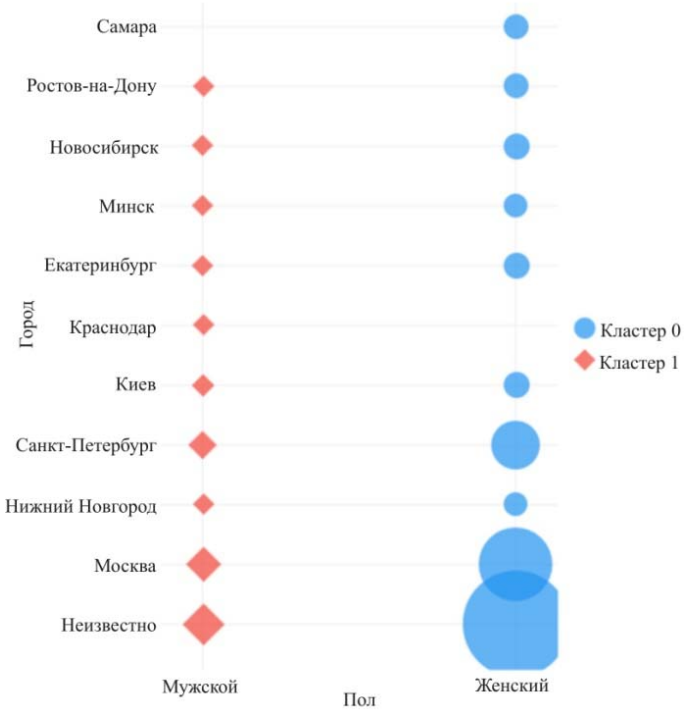


Рис. 3. Кластеризация пользователей сообщества
по полу и городу проживания

Большинство полученных правил с большой поддержкой ($C$) и достоверностью ($S$) являются тривиальными, например «Творчество → Юмор» ($C$ = 0,94 и $S$ = 0,65) или «Фотография → Юмор» ($C$ = 0,92 и $S$ = 0,27). Однако существование данных зависимостей можно было предположить и без поиска ассоциативных правил. Также были найдены более интересные зависимости, например «Интернет-СМИ, Образование → Юмор» ($C$ = 0,93 и $S$ = 0,05), «Юмор, Фотография, Литература → Творчество» ($C$ = 0,77 и $S$ = 0,04).

На основе полученных данных можно искать новую аудиторию в соответствующих, связанных по интересам сообществах.

***Визуальный анализ***

Визуальное представление позволяет быстрее анализировать большие объемы информации. Так, на примере одного сообщества социальной сети «ВКонтакте», являющегося СМИ, был проведен визуальный анализ (рис. 4).

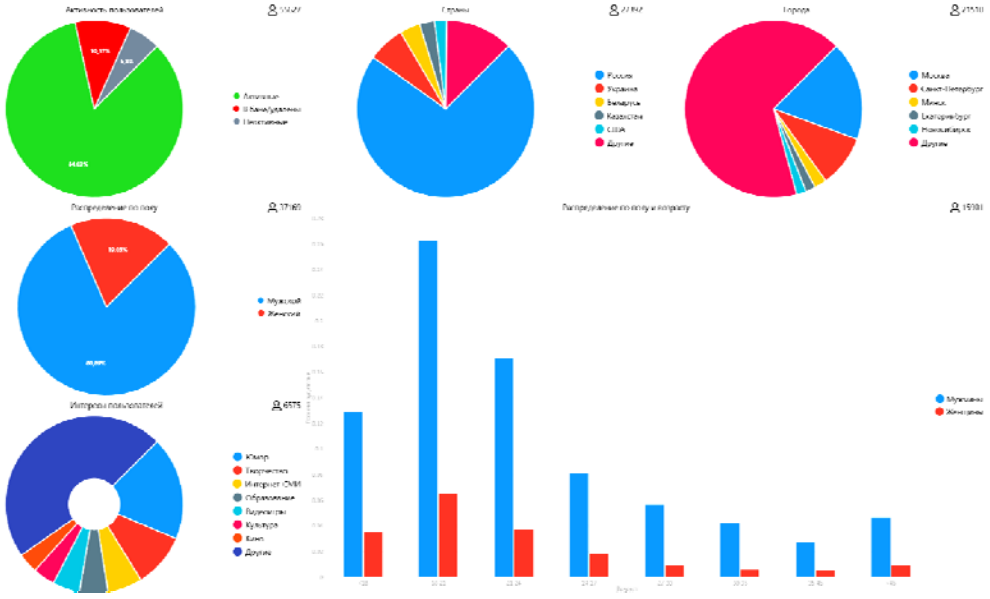

Рис. 4. Визуальный анализ сообщества социальной сети «ВКонтакте»

Данный пример визуального анализа состоит из шести графиков и диаграмм, которые включают в себя: активность, страны и города проживания, распределение по полу, распределение по полу и возрасту и диаграмму интересов пользователей сообщества.

По данным диаграммам можно сделать следующие выводы: большинство пользователей проживают на территории России, более 80 % аудитории сообщества – мужчины, основная возрастная аудитория – от 18 лет до 21 года, пользователи, подписанные на данное сообщество, интересуются юмором, творчеством, интернет-СМИ, образованием и видеоиграми, больше 84 % являются активными пользователями социальных сетей.

## Заключение

Результатом исследования стала информационно-аналитическая система по анализу социальных сетей, которая предназначена для автоматизации сбора, мониторинга и актуализации информации об объектах в социальных сетях.

Составляющими компонентами данной системы являются:

1. *Сбор данных.* Система может иметь множество источников данных, таких как социальные сети, блоги и сайты.

2. *Трансформация данных.* На данном этапе происходит объединение данных из разных источников в единую систему хранения. Также на этом этапе выполняется очистка, масштабирование и кодирование данных.

3. *Оперативный анализ.* На данном этапе происходит автоматическая агрегация данных с помощью технологии обработки OLAP [10]. Причина использования OLAP – высокая скорость обработки данных. Структура реляционных БД удобна для операционных баз данных (системы OLTP), но сложные многотабличные запросы в ней выполняются относительно медленно.

4. *Интеллектуальный анализ.* Сочетает в себе инновационные подходы к обработке данных, такие как Data Mining, Big Data и машинное обучение.

На данном этапе определяется тональность комментариев в сообществе и под записями. Это может быть полезно для исследований в области маркетинга в социальных сетях.

Кластеризация пользователей сообществ поможет определить целевой портрет аудитории. На основе полученных результатов может быть выстроена соответствующая целевая рекламная компания с использованием данных о географическом расположении, возрасте, интересах целевого пользователя.

С помощью поиска ассоциативных правил могут быть найдены закономерности в интересах пользователей.

На основе статистического анализа были вычислены KPI-показатели сообществ, которые являются числовыми показателями эффективности деятельности сообщества.

5. *Представление результатов.* На данном этапе найденные результаты представляются в виде графиков, диаграмм, таблиц и других визуальных объектов. Визуальный анализ позволяет преобразовать сложные данные в наглядные образы, которые дают возможность пользователю выявлять закономерности.

Одними из перспективных направлений аналитики в социальных сетях являются анализ изображений и семантический анализ текста. Развитие данных направлений поможет лучше понимать интересы и настроение пользователя и строить более точные аналитические модели.

Дальнейшее развитие системы может способствовать созданию новых моделей на основе машинного обучения, позволяющих строить достоверные прогнозы относительно разных показателей, например определение реакции пользователей на определенные записи.

*Исследование не имело спонсорской поддержки. Авторы заявляют об отсутствии конфликта интересов.*

## Список литературы

1. Губанов Д.А., Новиков Д.А. Чхартишвили А.Г. Социальные сети. Модели информационного влияния, управления и противоборства: учеб. пособие. – М.: Издательство физико-математической литературы, 2010. – 228 с.

2. Белов В.С. Информационно-аналитические системы. Основы проектирования и применения: учеб. пособие. – М.: Евразийский открытый институт, 2010. – 112 с.

3. Selection of rational schemes automation based on working synthesis instruments for technological processes / Yu.A. Leonov, E.A. Leonov, A.A. Kuzmenko, A.A. Martynenko, E.E. Averchenkova, R.A. Filippov. – Yelm, WA, USA: Science Book Publishing House LLC, 2019. – 192 p.

4. Чубукова И.А. Data Mining – М.: Интернет-университет информационных технологий (ИНТУИТ), 2016. – 470 с.

5. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – 2012. – Вып. 11 (18). – С. 7–10.

6. Осипова Ю.А., Лавров Д.Н. Применение кластерного анализа методом k-средних для классификации текстов научной направленности. – М.: Изд-во МСиМ, 2017. – С. 108–121.

7. Hipp J., Guntzer U., Nakaeizadeh G. Algorithms for association rule mining – a general survey and comparison. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 2000, Tübingen, Germany. – 2000. – P. 58–64

8. Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems / E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova / In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds) // Advances in Intelligent Systems and Computing [Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). IITI 2017. Varna, Bulgaria, 14-16 September 2017]. – 2017. – Vol. 679. – P. 95–103. – DOI:10.1007/978-3-319-68321-8_10

9. Ramezani R., Saraee M., Nematbakhsh M.A. MRAR: mining multi-relation association rules // Journal of Computing and Security. – 2014. – Т. 1, № 2. – P. 133–158.

10. Иваничев И. KPI в SMM. 30+ метрик эффективности маркетинга в социальных сетях // Интернет-агентство «Текстерра». – Электрон. дан. – 2018. – URL: https://texterra.ru/blog/kpi-v-smm-metriki-effektivnosti-marketinga-v-sotsialnykhsetyakh.html (дата обращения 25.10.2020).

11. Сенаторов А.А. Контент-маркетинг: стратегии продвижения в социальных сетях – Москва: Альпина Паблишер, 2020. – 160 с.

12. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109–116.

13. Intelligent system of classification and clusterization of environmental media for economic systems / A.A. Kuzmenko, L.B. Filippova, A.S. Sazonova, R.A. Filippov // Advances in Economics, Business and Management Research [Proceedings of the International Conference on Economics, Management and Technologies 2020 (ICEMT 2020) Jalta, Krym, Rossia, 19-21 may 2020.]. – 2020. – Vol. 139. – P. 583–586.

## References

1. Gubanov D.A., Novikov D.A. Chkhartishvili A.G. Sotsial'nye seti. Modeli informatsionnogo vliianiia, upravleniia i protivoborstva: uchebnoe posobie [Social networks. Models of information influence, management and confrontation: study guide]. Moscow, Izdatel'stvo fiziko-matematicheskoi literatury, 2010, 228 p.

2. Belov V.S. Informatsionno-analiticheskie sistemy. Osnovy proektirovaniia i primeneniia: uchebnoe posobie [Information and analytical systems. Fundamentals of design and application: tutorial]. Moscow, Evraziiskii otkrytyi institut, 2010, 112 p.

3. Leonov Yu.A., Leonov E.A., Kuzmenko A.A., Martynenko A.A., Averchenkova E.E., Filippov R.A. Selection of rational schemes automation based on working synthesis instruments for technological processes. Yelm, WA, USA, Science Book Publishing House LLC, 2019, 192 p.

4. Chubukova I.A. Data Mining. Moscow, Internet-Universitet Informatsionnykh Tekhnologii (INTUIT), 2016, 470 p.

5. Kotel'nikov E.V., Klekovkina M. V. Avtomaticheskii analiz tonal'nosti tekstov na osnove metodov mashinnogo obucheniia [Sentiment analysis of texts based on machine learning methods]. *Komp'juternaja lingvistika i intellektual'nye tehnologii*, 2012, iss. 11 (18), pp. 7–10.

6. Osipova Iu.A., Lavrov D.N. Primenenie klasternogo analiza metodom k-srednikh dlia klassifikatsii tekstov nauchnoi napravlennosti [Application of cluster analysis by the k-means method for classification of scientific texts]. MSiM Publ., 2017, pp. 108–121.

7. Hipp J., Guntzer U., Nakaeizadeh G. Algorithms for Association Rule Mining – A General Survey and Comparison. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Tübingen, Germany, 2000, pp. 58–64.

8. Leonov E.A., Leonov Yu.A., Leonov, Kazakov Yu.M., Filippova L.B. Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems. *Advances in Intelligent Systems and Computing*, 2017, vol 679, pp. 95-103 – DOI:10.1007/978-3-319-68321-8_10

9. Saraee M., Nematbakhsh M.A., Ramezani R. MRAR: Mining Multi-Relation Association Rules. *Journal of Computing and Security*, 2014, vol 1, no. 2, pp. 133–158.

10. Ivanichev I. *KPI v SMM. 30+ metrik jeffektivnosti marketinga v social'nyh setjah* [KPIs in SMM. 30+ metrics of social media marketing effectiveness]. *Internet-agentstvo «Teksterra»*. – Jelektron. dan. – 2018. – URL: https://texterra.ru/blog/kpi-v-smm-metriki-effektivnosti-marketinga-v-sotsialnykhsetyakh.html (Accessed 25 November 2020).

11. Senatorov A.A. Kontent-marketing: strategii prodvizheniia v sotsial'nykh setiakh [Content Marketing: Social media promotion strategies]. Moscow, Al'pina Pablisher, 2020, 160 p.

12. Rubtsova Iu. Avtomaticheskoe postroenie i analiz korpusa korotkikh tekstov (postov mikroblogov) dlia zadachi razrabotki i trenirovki tonovogo klassifikatora [Automatic construction and analysis of the corpus of short texts (microblogging posts) for the task of developing and training a tone classifier]. *Inzheneriia znanii i tekhnologii semanticheskogo veba*, 2012, vol. 1, pp. 109–116.

13. Kuzmenko A.A., Filippova L.B., Sazonova A.S., Filippov R.A. Intelligent System of Classification and Clusterization of Environmental Media for Economic Systems. *Advances in Economics, Business and Management Research*, 2020, vol. 139. pp. 583–586.

Статья получена: 10.11.2021

Статья принята: 25.11.2021

Опубликовано: 26.01.2022

## Сведения об авторах

**Шестаков Тимофей Андреевич** (Брянск, Россия) – студент, кафедра «Компьютерные технологии и системы», Брянский государственный технический университет (241035, Брянск, б-р 50-летия Октября, 7, e-mail: libv88@yandex.ru).

**Леонов Юрий Алексеевич** (Брянск, Россия) – кандидат технических наук, доцент, доцент, кафедра «Компьютерные технологии и системы», Брянский государственный технический университет (241035, Брянск, б-р 50-летия Октября, 7, e-mail: yorleon@yandex.ru).

**Кузьменко Александр Анатольевич** (Брянск, Россия) – кандидат биологических наук, доцент кафедры «Компьютерные технологии и системы», Брянский государственный технический университет (241035, Брянск, б-р 50-летия Октября, 7, e-mail: alex-rf-32@yandex.ru).

**Сазонова Анна Сергеевна** (Брянск, Россия) – кандидат технических наук, доцент, доцент, кафедра «Компьютерные технологии и системы», ФГБОУ ВО «Брянский государственный технический университет» (241035, Брянск, б-р 50-летия Октября, 7, e-mail: libv88@yandex.ru).

**Филиппов Родион Алексеевич** (Брянск, Россия) – кандидат технических наук, доцент, доцент, кафедра «Компьютерные технологии и системы», Брянский государственный технический университет (241035, Брянск, б-р 50-летия Октября, 7, e-mail: redfil@mail.ru).

## About the authors

**Timofei A. Shestakov** (Bryansk, Russian Federation) – Student, Department of Computer Technologies and Systems, Bryansk State Technical University (7, b-r 50th anniversary of October, Bryansk, 241035, e-mail: alex-rf-32@yandex.ru)

**Iurii A. Leonov** (Bryansk, Russian Federation) – Ph. D. in Engineering, associate professor, associate professor, Department of Computer Technologies and Systems, Bryansk State Technical University (7, b-r 50th anniversary of October, Bryansk, 241035, e-mail: yorleon@yandex.ru)

**Aleksandr A. Kuz'menko** (Bryansk, Russian Federation) – Ph. D. in Biology, associate professor in the Department of Computer Technologies and Systems, Bryansk State Technical University (7, b-r 50th anniversary of October, Bryansk, 241035, e-mail: alex-rf-32@yandex.ru)

**Anna S. Sazonova** (Bryansk, Russian Federation) – Ph. D. in Engineering, associate professor, associate professor, Department of Computer Technologies and Systems, Bryansk State Technical University (7, b-r 50th anniversary of October, Bryansk, 241035, e-mail: libv88@yandex.ru)

**Rodion A. Filippov** (Bryansk, Russian Federation) – Ph. D. in Engineering, associate professor, associate professor, Department of Computer Technologies and Systems, Bryansk State Technical University (7, b-r 50th anniversary of October, Bryansk, 241035, e-mail: redfill@mail.ru)

**Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:**

Интеллектуальный анализ информации о пользователях социальных сетей / Т. А. Шестаков, Ю. А. Леонов, А. А. Кузьменко, С. А. Сазонова, Р. А. Филиппов. – текст: непосредственный. – DOI: 10.15593/2499-9873/2021.4.05 // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2021. – № 4. – С. 72–91.

**Цитирование статьи в изданиях РИНЦ:**

Интеллектуальный анализ информации о пользователях социальных сетей / Т. А. Шестаков, Ю. А. Леонов, А. А. Кузьменко [и др.] // Прикладная математика и вопросы управления. – 2021. – № 4. – С. 72–91. – DOI: 10.15593/2499-9873/2021.4.05

**Цитирование статьи в references и международных изданиях**

**Cite this article as:**

Shestakov T.A., Leonov Ju.A., Kuzmenko A.A., Sazonova A.S., Filippov R.A. Intellectual analysis of information about users of social networks. *Applied Mathematics and Control Sciences*, 2021, no. 4, pp. 72–91. DOI: 10.15593/2499-9873/2021.4.05 *(in Russian)*