

Научная статья
УДК 004.9
doi:10.37614/2949-1215.2023.14.7.004

ПОРЯДКОВАЯ ОЦЕНКА ПОПУЛЯРНОСТИ ПУБЛИКАЦИЙ ДЛЯ ЗАДАЧ МОНИТОРИНГА ОНЛАЙНОВЫХ СОЦИАЛЬНЫХ МЕДИА

Максим Геннадьевич Шишаев^{1✉}, Владимир Витальевич Диковицкий²

^{1, 2}*Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия*

¹*shishaev@iimm.ru, <https://orcid.org/0000-0001-7070-7878>*

²*dikovitsky@iimm.ru, <https://orcid.org/0000-0003-0329-9979>*

Аннотация

В работе изучена возможность использования различных способов формирования порядковых оценок популярности сообщений в онлайн-социальных медиа. На примере экспериментальных данных, собранных из сети «ВКонтакте», показано, что использование в качестве основы порядковой оценки популярности количественных индикаторов лайков, репостов и комментариев дает эквивалентный результат. Отмечается необходимость использования метрик популярности, нормированных по количеству просмотров или размеру аудитории сообщества, при анализе нескольких онлайн-сообществ одновременно.

Ключевые слова:

онлайн-социальные медиа, оценка популярности

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования Российской Федерации, тема научно-исследовательской работы — «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (регистрационный номер 122022800551-0).

Для цитирования:

Шишаев М. Г., Диковицкий В. В. Порядковая оценка популярности публикаций для задач мониторинга онлайн-социальных медиа // Труды Кольского научного центра РАН. Серия: Технические науки. 2023. Т. 14, № 7. С. 35–42. doi:10.37614/2949-1215.2023.14.7.004.

Original article

RANKING ASSESSMENT OF THE POPULARITY OF PUBLICATIONS FOR ONLINE SOCIAL MEDIA MONITORING

Maxim G. Shishaev^{1✉}, Vladimir V. Dikovitsky²

^{1, 2}*Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia*

¹*shishaev@iimm.ru, <https://orcid.org/0000-0001-7070-7878>*

²*dikovitsky@iimm.ru, <https://orcid.org/0000-0003-0329-9979>*

Abstract

The paper examines the possibility of using various methods for generating ranking estimates of the popularity of messages in online social media. Using experimental data collected from the VKontakte network as an example, it is shown that using quantitative indicators of likes, reposts and comments as the basis for an ranking assessment of popularity gives an equivalent result. The need to use popularity metrics, normalized by the number of views or the size of the community audience, is noted when analyzing several online communities at the same time.

Keywords:

online social media, popularity assessment

Acknowledgments:

the study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic "Methodology for creating information and analytical systems to support the management of regional development based on formative artificial intelligence and big data" (registration number of the research topic 122022800551-0).

For citation:

Shishaev M. G., Dikovitsky V. V. Ranking assessment of the popularity of publications for online social media monitoring // Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2023. Vol. 14, No. 7. P. 35–42. doi:10.37614/2949-1215.2023.14.7.004.

Введение

Количественная оценка степени популярности публикаций является одной из типичных вспомогательных задач, возникающих при мониторинге онлайн-медиа с той или иной прикладной целью [1]. Несмотря на то, что практически все онлайн-медиа используют в качестве атрибутов сообщений те или иные счетчики пользовательской реакции, в прикладных задачах понятие популярности, как правило, трактуется более широко и используется для более глубокого анализа поведения аудитории, чем простые отметки «нравится/не нравится» или иные индикаторы пользовательской реакции, т. е. мы исходим из того, что реакция аудитории не тождественна популярности сообщения. Это заставляет искать более изощренные способы оценки популярности публикаций в онлайн-социальных медиа, использующие как непосредственно индикаторы пользовательской реакции, так и различные метаданные сообщений (время публикации, авторство, размер аудитории и т. п.).

В настоящей работе на экспериментальных данных, полученных в результате мониторинга нескольких региональных онлайн-сообществ в сети «ВКонтакте», проведен сравнительный анализ возможностей использования различных индикаторов реакции аудитории в качестве основы для расчета порядкового показателя популярности контента, примененного для ранжирования сообщений по степени их популярности.

О существующих подходах к оценке популярности контента

Единое понимание популярности в существующих работах по данной тематике отсутствует, что делает формальное ее определение нетривиальной задачей [2]. В целом определение популярности публикации опирается на имеющиеся в распоряжении индикаторы, характеризующие реакцию аудитории. К таким индикаторам в большинстве онлайн-социальных сетей относятся количество просмотров, лайков (отметок «нравится»), комментариев и репостов (пересылок сообщения целиком в другие онлайн-сообщества). Исследования в этой области показывают, что популярность публикации может быть выражена разными способами, кроме того, она может меняться со временем в зависимости от различных неформализованных факторов [3]. Задачу также усложняет и то, что популярность идентичного контента может быть различной в разных социальных сетях или даже в разных группах одной сети.

Кроме количественных индикаторов реакции аудитории на публикацию, при расчете оценки популярности могут использоваться различные модификаторы, например, астрономическое время публикации сообщения или объем аудитории онлайн-сообщества, в котором опубликовано сообщение. Важным обстоятельством является также то, осуществляется ли оценка популярности сообщений в пределах одного сообщества или же нескольких. В последнем случае, очевидно, оценки в абсолютных величинах не подходят, поскольку количественные значения индикаторов популярности, при прочих равных, прямо зависят от объема аудитории онлайн-сообщества.

Наряду с указанными выше «базовыми» индикаторами, для оценки популярности прибегают к использованию и других параметров публикаций. В работе [4], например, эта оценка осуществляется с учетом гомофильности аудитории и популярности автора. В исследовании [5] предлагается использовать различия в поведении и интересах аудитории. Еще один подход к определению популярности контента основан на анализе его распространения в социальных сетях.

Существующие способы оценки популярности можно разделить на три категории, каждая из которых применяется в соответствующих видах прикладных задач анализа контента онлайн-медиа [6]: 1) номинальные оценки; 2) количественные оценки; 3) порядковые оценки. В случае номинальных оценок требуется отнести элемент контента (сообщение) к одной из заданных категорий. Как правило, рассматривается двухклассовая идентификация, в результате которой сообщение относится к релевантным или не релевантным. При количественных оценках тем или иным способом оценивается величина популярности. Наконец, при порядковых оценках требуется ранжировать сообщения по степени их популярности. В данной работе рассматривается последний случай.

Характеристика экспериментальных данных

В работе использовались ранее собранные данные о публикациях («постах») в нескольких открытых региональных онлайн-сообществах социальной сети «ВКонтакте». В качестве образца данных рассматривался отдельный пост (сообщение), а в качестве его атрибутов — количества лайков, комментариев и репостов за период с января 2021 г. по декабрь 2022 г. В подборку были включены сообщества без специфической тематической привязки, но локализованные по составу подписчиков в небольшом регионе (в нашем случае — Кировско-Апатитском регионе Мурманской области).

Таким образом, использовавшийся для экспериментов набор данных (датасет) может быть охарактеризован разнообразием тематики публикаций и относительно стабильным составом аудитории в основном из числа жителей рассматриваемого региона. Для содержательного анализа было произведено его тематическое моделирование, в результате которого выделены наиболее обсуждаемые темы. Размеры полученных тем (количество сообщений, попавшее в соответствующий теме кластер) представлены в виде графика на рис. 1. Всего в состав датасета вошли данные из 20 онлайн-сообществ, имеющие 296 тыс. активных подписчиков, что обеспечивает их репрезентативность в контексте предполагаемого практического использования полученных результатов.

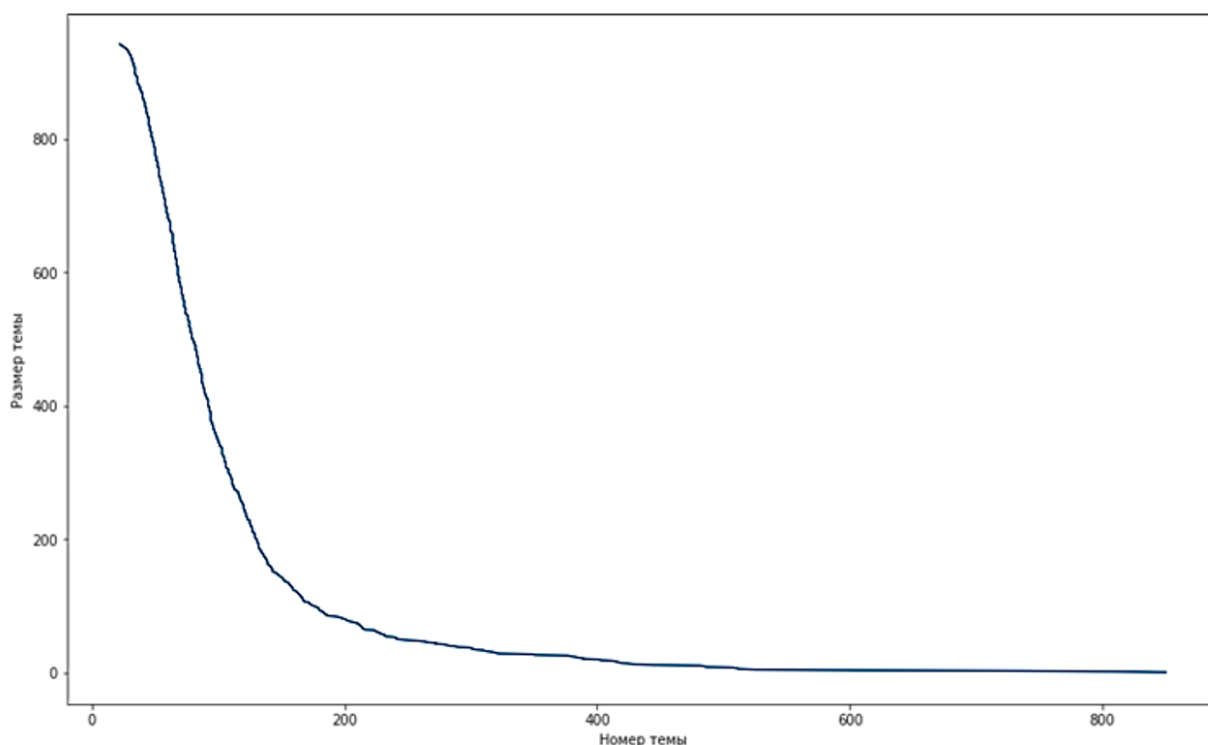


Рис. 1. Распределение сообщений датасета по темам

Специфичной особенностью данных онлайн-социальных сетей является доминирование сообщений небольшой длины (распределение постов по длине представлено на рис. 2).

Исходный датасет был подвергнут предварительной очистке, которая включала следующие шаги:

- удаление коротких постов, которые не содержат текста или содержат только ссылки или хештеги;
- удаление длинных постов, которые содержат более 100 слов, т. к. они являются аномальными для типичной онлайн-коммуникации пользователей, что дает основания предполагать их рекламный характер;
- удаление постов, которые содержат ненормативную лексику, оскорбления, спам, т. к. они могут исказить результаты анализа.

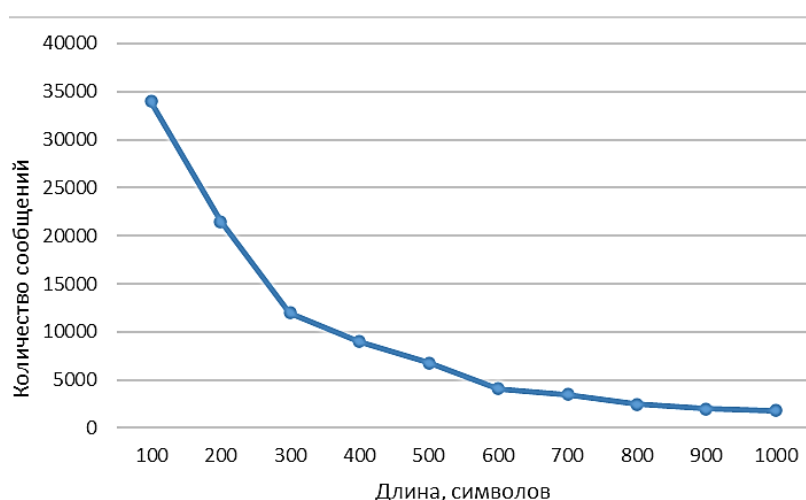


Рис. 2. Количество образцов в выборке в зависимости от длины сообщения

Формальное представление и сравнительный анализ метрик популярности

Для формального представления популярности сообщения необходимо прежде всего определиться с концептуальным пониманием, прикладной интерпретацией данного свойства. Интерпретация популярности, в свою очередь, опирается на концептуальное представление задачи. Фрагмент подобного представления, включающий наиболее существенные компоненты предметной области, представлен в виде концептуального графа на рис. 3. В данной статье мы исходим из концептуального трактования популярности, как меры интенсивности реакции аудитории на сообщение (публикацию).

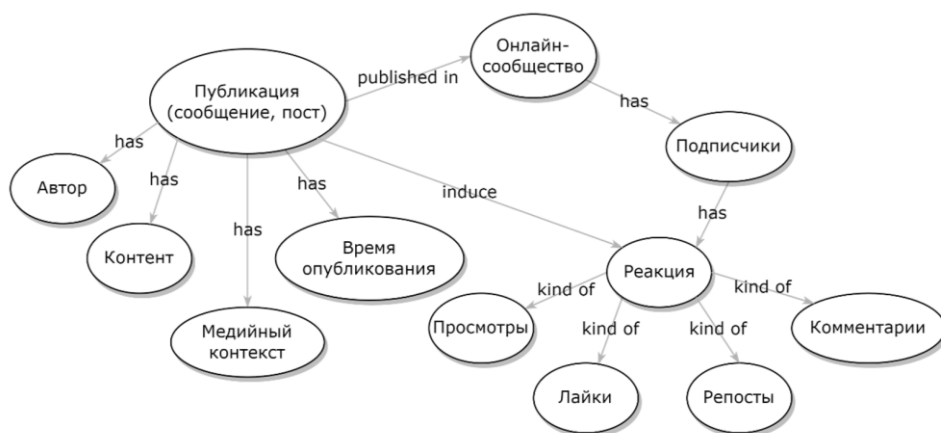


Рис. 3. Концептуальный граф задачи

Формальное представление популярности основывается на некоторых наблюдаемых и измеримых признаках публикации. В качестве таковых, как правило, выступают количество просмотров, репостов, комментариев и лайков. Кроме того, могут учитываться дополнительные параметры сообщения (метаданные), т. е.:

$$Popularity = f(s, r, c, l, M),$$

где s , r , c , l — количество просмотров, репостов, комментариев и лайков соответственно, а M — метаданные сообщения, включающие авторство, временные характеристики, медийный контекст появления и другие параметры, характеризующие ситуацию, в которой опубликовано сообщение.

Вид функции f зависит от используемой интерпретации понятия популярности публикации, которая, в свою очередь, определяется прикладной задачей, в контексте которой нам необходимо оценить популярность. Областью значений функции, в зависимости от вида оценки — количественная, номинальная или порядковая, является либо числовое значение, либо одно из заданных номинальных значений, либо отношение порядка на множестве сообщений.

В данной работе рассматривается задача порядковой оценки популярности, т. е. решающее значение имеет не величина признака как таковая, а порядок объектов, отсортированных по этому признаку. Исходя из чего, способы расчета значений признаков объектов, дающие в результате идентичный порядок сортировки объектов, являются эквивалентными. Для проверки этого утверждения в отношении используемых нами данных были проведены отдельные расчеты показателя популярности, основанные на количестве лайков, репостов и комментариев, после чего объекты (сообщения) были отсортированы и оценена степень сходства получившихся трех последовательностей.

Схожесть последовательностей оценивалась как среднее относительное смещение позиции объекта в двух последовательностях:

$$s = \left(\sum_{i=1}^N \frac{p_1(x_i) - p_2(x_i)}{N} \right) / N,$$

где N — количество объектов; $p_1(x_i)$, $p_2(x_i)$ — позиция объекта x_i в 1-й и 2-й последовательностях соответственно.

Размер исходного датасета составил 100 тыс. образцов. При сравнении рассматривались только образцы с ненулевыми значениями соответствующих меток.

Было сформировано 6 различных последовательностей: P_{lc} , P_{cl} , P_{lr} , P_{rl} , P_{cr} , P_{rc} . Нижний индекс в данном случае означает следующее: P_{lc} , — последовательность, отсортированная по убыванию l , в которой образцы имеют ненулевое количество лайков и комментариев. Аналогичным образом интерпретируются обозначения остальных последовательностей, т. е. выполняются следующие наборы условий:

$$\begin{aligned} & \left\{ \begin{array}{l} o \in P_{lc}: \text{likes}(o) \neq 0 \wedge \text{comments}(o) \neq 0 \\ \forall o_i, o_j \in P_{lc}: \text{likes}(o_i) \geq \text{likes}(o_j), \forall i < j \end{array} \right\} \\ & \left\{ \begin{array}{l} o \in P_{cl}: \text{likes}(o) \neq 0 \wedge \text{comments}(o) \neq 0 \\ \forall o_i, o_j \in P_{cl}: \text{comments}(o_i) \geq \text{comments}(o_j), \forall i < j \end{array} \right\} \\ & \left\{ \begin{array}{l} o \in P_{lr}: \text{likes}(o) \neq 0 \wedge \text{reposts}(o) \neq 0 \\ \forall o_i, o_j \in P_{lr}: \text{likes}(o_i) \geq \text{likes}(o_j), \forall i < j \end{array} \right\} \\ & \left\{ \begin{array}{l} o \in P_{rl}: \text{likes}(o) \neq 0 \wedge \text{reposts}(o) \neq 0 \\ \forall o_i, o_j \in P_{rl}: \text{reposts}(o_i) \geq \text{reposts}(o_j), \forall i < j \end{array} \right\} \\ & \left\{ \begin{array}{l} o \in P_{cr}: \text{comments}(o) \neq 0 \wedge \text{reposts}(o) \neq 0 \\ \forall o_i, o_j \in P_{cr}: \text{comments}(o_i) \geq \text{comments}(o_j), \forall i < j \end{array} \right\} \\ & \left\{ \begin{array}{l} o \in P_{rc}: \text{reposts}(o) \neq 0 \wedge \text{comments}(o) \neq 0 \\ \forall o_i, o_j \in P_{rc}: \text{reposts}(o_i) \geq \text{reposts}(o_j), \forall i < j \end{array} \right\} \end{aligned}$$

где $\text{likes}(o)$, $\text{reposts}(o)$, $\text{comments}(o)$ — количество лайков, репостов и комментариев образца o соответственно; o_i — i -й элемент последовательности.

Последовательности сравнивались попарно. Итоги сравнения и размер соответствующих последовательностей приведены в табл. 1.

Таблица 1

Сравнение последовательностей, полученных путем отбора и сортировки элементов

Сравниваемые последовательности	P_{lc}, P_{cl}	P_{lr}, P_{rl}	P_{cr}, P_{rc}
Размер последовательности	66970	80310	59397
Среднее относительное смещение позиции объекта по всей последовательности	0,278	0,191	0,248

Можно отметить, что различие между последовательностями образцов, отсортированными по лайкам и репостам не превышает 20 % (в терминах относительного смещения элементов). В то же время последовательности, сформированные путем сортировки по количеству комментариев, значительно отличаются от прочих — на 25 % и более. Таким образом, использование в качестве основы для расчета показателя популярности сообщений количества лайков или количества репостов, потенциально, даст эквивалентный результат.

Можно также предположить, что наблюдаемые особенности последовательностей обусловлены значительной разницей в количественных значениях индикаторов реакции аудитории — лайков, комментариев и репостов. Для использованной в работе выборки средние значения этих индикаторов составили 53,7, 9,7 и 13,8 соответственно. Данное предположение подтверждается также видом зависимостей относительного смещения объектов в последовательностях от их значений индикаторов реакции. Данные зависимости для последовательностей P_{lc} , P_{cl} , P_{lr} , P_{rl} , P_{cr} , P_{rc} представлены на рис. 4. Для их построения в каждой из них выделялась серия подпоследовательностей, включающих K начальных элементов исходной последовательности — $\{o_1, \dots, o_K\}$, и для каждой подпоследовательности рассчитывалось среднее относительное смещение элементов. Поскольку последовательности отсортированы по убыванию соответствующих индикаторов реакции, то в начале оси абсцисс концентрируются последовательности, включающие элементы с наибольшими значениями лайков, репостов, комментариев. Можно заметить, что для последовательностей, включающих объекты с большими значениями индикаторов (расположены в начале оси абсцисс), среднее относительное смещение ниже. Таким образом, в отношении сообщений с высокими значениями индикаторов реакции использование любого из них в качестве основы для расчета показателя популярности даст схожий результат. Это подтверждается значениями коэффициентов корреляции последовательностей, отсортированных по лайкам и комментариям, комментариям и репостам, лайкам и репостам, которые составили 0,988, 0,990 и 0,994 соответственно.

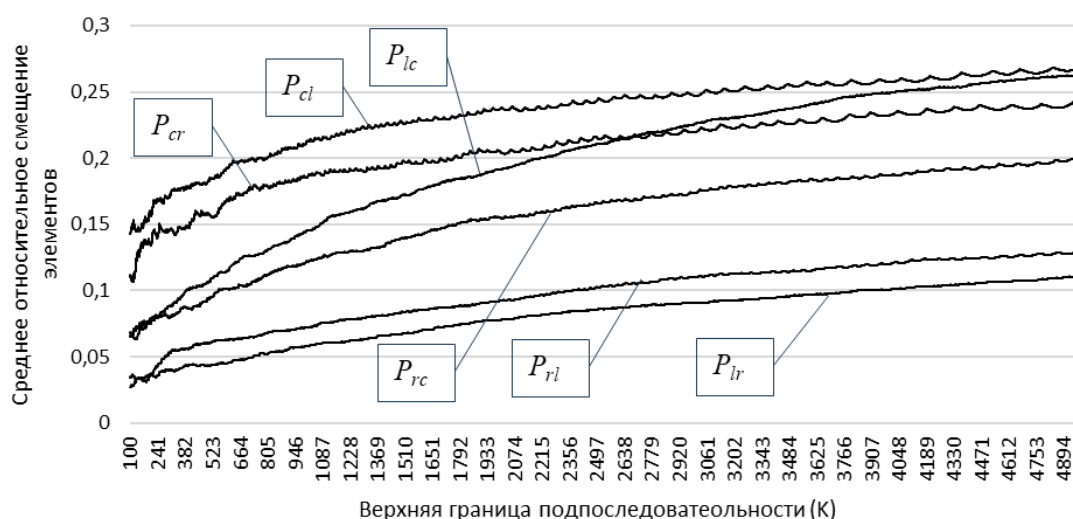


Рис. 4. Зависимость среднего относительного смещения объектов в последовательностях от значений индикаторов реакции (нарастающим итогом)

Мы видим, что для порядковой оценки популярности публикаций могут использоваться любые из трех основных индикаторов реакции. Однако следует иметь в виду, что в абсолютном измерении эти индикаторы, очевидно, зависят от размера аудитории. Это обстоятельство является важным, если мониторинг социальной сети подразумевает анализ данных сразу нескольких онлайн-сообществ (с разными количествами подписчиков). Таким образом, для выравнивания априорной разницы в «базовой» популярности и размере аудитории различных онлайн-сообществ, для оценки популярности данный показатель целесообразно дополнительно нормировать по количеству просмотров контента или количеству подписчиков сообществ.

Заключение

В данной работе рассматривается задача оценки популярности публикаций в онлайн-социальных медиа. Акцент сделан на порядковых оценках, когда имеет значение не абсолютное значение метрики, а их соотношение для разных сообщений из рассматриваемого множества. Анализ использованных в качестве примера экспериментальных данных, собранных из нескольких онлайн-сообществ социальной сети «ВКонтакте», показал, что для порядковых оценок использование таких индикаторов реакции аудитории, как количество лайков, репостов и комментариев, дает практически эквивалентный результат.

Интерес для дальнейших исследований представляет анализ возможных способов определения количественных оценок популярности, в частности, влияния различных дополнительных признаков сообщения. Многие подобные признаки существенно влияют на интенсивность реакции аудитории и, соответственно, должны учитываться при формировании оценки популярности сообщения. В этой связи можно сформулировать несколько предположений.

1. Интенсивность реакции зависит от времени суток, в которое опубликовано сообщение. При этом для различных онлайн-сообществ и категорий пользователей «прайм-тайм», обеспечивающий максимально интенсивную реакцию в абсолютном измерении, будет разный — в одних случаях это будут вечерние часы, в других — дневные или утренние, или даже ночные.

2. Интенсивность реакции на сообщение зависит от инфомедийного контекста появления последнего, который характеризует спектр наиболее значимых тем общественного дискурса в момент появления сообщения. Например, в начале пандемии COVID-19 любое сообщение на эту тему имело намного большую популярность, нежели в настоящее время, когда острота проблемы коронавируса спала.

3. Интенсивность реакции зависит от характера темы сообщения — относится таковая к вирусным, характеризуемым резким всплеском и последующим падением интереса, или так называемым «гринлайн»-темам со стабильным во времени уровнем интереса аудитории.

Список источников

1. Shishaev M. Analysis of Online Social Networking When Studying the Identities of Local Communities / M. Shishaev, A. Fedorov, I. Datyev // *Digitalisation and Human Security: A Multi-Disciplinary Approach to Cybersecurity in the European High North: New Security Challenges* / eds. M. Salminen, G. Zojer, K. Hossain. Cham: Springer International Publishing, 2020. P. 267–293.
2. A survey on predicting the popularity of web content / A. Tatar [et al.] // *Journal of Internet Services and Applications*. 2014. Vol. 5(1). P. 8.
3. Szabo G. Predicting the popularity of online content / G. Szabo, B.A. Huberman // *Communications of the ACM*. 2010. Vol. 53(8). P. 80–88.
4. Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features / Y. Shang [et al.] // *Frontiers in Physics*. 2022. Vol. 10.
5. Ding K. Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks / K. Ding, R. Wang, S. Wang // *Proceedings of the 27th ACM International Conference on Multimedia: MM '19*. New York, NY, USA: Association for Computing Machinery, 2019. Social Media Popularity Prediction. P. 2682–2686.
6. Moniz N. A review on web content popularity prediction: Issues and open challenges / N. Moniz, L. Torgo // *Online Social Networks and Media*. 2019. Vol. 12. A review on web content popularity prediction. P. 1–20.

References

1. Shishaev M., Fedorov A., Datyev I. Analysis of Online Social Networking When Studying the Identities of Local Communities. Digitalisation and Human Security: A Multi-Disciplinary Approach to Cybersecurity in the European High North: New Security Challenges, eds. M. Salminen, G. Zojer, K. Hossain, Cham: Springer International Publishing, 2020, pp. 267–293.
2. Tatar A. et al. A survey on predicting the popularity of web content. Journal of Internet Services and Applications, 2014, vol. 5, no. 1, pp. 8.
3. Szabo G. Huberman B. A. Predicting the popularity of online content. Communications of the ACM, 2010, vol. 53, no. 8, pp. 80–88.
4. Shang Y. et al. Predicting the Popularity of Online Content by Modeling the Social Influence and Homophily Features. Frontiers in Physics, 2022, vol. 10.
5. Ding K., Wang R., Wang S. Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks. Proceedings of the 27th ACM International Conference on Multimedia: MM '19, New York, NY, USA: Association for Computing Machinery, 2019, Social Media Popularity Prediction, pp. 2682–2686.
6. Moniz N., Torgo L. A review on web content popularity prediction: Issues and open challenges. Online Social Networks and Media, 2019, vol. 12, A review on web content popularity prediction, pp. 1–20.

Информация об авторах

М. Г. Шишаев — доктор технических наук, главный научный сотрудник;

В. В. Диковицкий — кандидат технических наук, старший научный сотрудник.

Information about the authors

M. G. Shishaev — Doctor of Science (Tech.), Chief Research Fellow;

V. V. Dikovitsky — Candidate of Science (Tech.), Senior Research Fellow.

Статья поступила в редакцию 11.10.2023; одобрена после рецензирования 01.11.2023; принята к публикации 08.11.2023.
The article was submitted 11.10.2023; approved after reviewing 01.11.2023; accepted for publication 08.11.2023.