Статистическая модель поиска целевых объектов в социальной сети

Д. И. Сафиканов, А. А. Артамонов, Ю. Е. Фомина, А. И. Черкасский

***Аннотация*—С учетом растущей роли социальных сетей их анализ является актуальным направлением в сфере информационных технологий, развитие которого позволяет решать широкий спектр прикладных задач. Однако при анализе социальных сетей встает ряд сложностей, связанных с объемом данных, их неоднородностью и неструктурированностью. Одним из путей решения данных проблем является использование агентных технологий. В статье описана методология использования статистических моделей для агентного поиска целевых объектов, предполагающая следующую последовательность действий. На первом этапе аналитики формируют обучающую выборку целевых объектов посредством интерактивного поиска в социальной сети. В дальнейшем проводится анализ обучающей выборки с целью определения критериев оценки, их ранжирования по важности и присвоения им весовых коэффициентов. В результате формируется статистическая модель целевого объекта. На основе обучающей выборки определяются граничные значения (маркеры) для отнесения новых объектов к целевым. На следующем шаге производится настройка агентов на автоматическое выполнение целевого поиска.**

В ходе практической реализации предложенная методология показала себя эффективным инструментом для идентификации целевых объектов в социальной сети «ВКонтакте». Данная методология может быть потенциально масштабирована для использования в различных социальных сетях.

***Ключевые слова*—социальная сеть, целевой объект, статистическая модель, многокритериальный анализ, агентный поиск, анализ больших данных.**

I. Введение

Одной из отличительных черт развития информационных технологий в XXI веке является стремительно увеличивающаяся роль социальных сетей в различных сферах жизни общества – социальной, экономической, политической и культурной. Социальные сети можно определить как интерактивные многопользовательские веб-сайты, которые позволяют зарегистрированным на них пользователям размещать информацию о себе и поддерживать коммуникацию с другими пользователями, при этом содержание (контент) веб-сайта создается преимущественно самими пользователями (так называемый user-generated content) [1]. Из приведенного определения следует, что социальные сети – это, прежде всего, инструмент коммуникации между пользователями и/или группами пользователей, обладающий рядом преимуществ, среди которых можно выделить следующие:

- возможность мгновенного обмена информацией;
- возможность поддержания трансграничной коммуникации;
- открытый (преимущественно бесплатный) доступ для пользователей;
- относительная свобода от цензуры;
- возможность реализации анонимного взаимодействия с другими пользователями [2];
- наличие единого канала для передачи и хранения данных разных типов (текстовых данных, изображений, аудио- и видеофайлов и т. д.) [3].

Социальные сети представляют собой источник больших объемов динамично обновляющихся данных, анализ которых позволяет решать широкий спектр задач, к которым можно отнести:

1. Поиск и выявление целевых объектов.
2. Выявление мнений и точек зрения субъектов и социальных групп по различным социальным вопросам, исследование социальной напряженности [4].
3. Обеспечение целенаправленного распространения информации [5].
4. Формирование групп для реализации определенных коллективных действий, осуществление дистанционного управления и координации таких действий [6].

Однако при анализе социальных сетей возникают следующие сложности. Так, перед исследователями встает необходимость решения проблемы обработки и анализа больших объемов данных. Количество пользователей социальных сетей постоянно увеличивается. Кроме того, как было указано ранее, информация в социальных сетях представлена различными типами данных – текст, аудио- и видеоформат – и при этом зачастую неформализована, что также усложняет ее обработку.

Задачу анализа социальных сетей можно решать с использованием агентных технологий. Под агентной технологией будем понимать программу, автономно функционирующую в информационной среде (в данном случае – в социальной сети) и решающую конкретную задачу (например, поиск, сбор данных). Нужно отметить, что для достижения целей, поставленных пользователем, агенты могут взаимодействовать друг с другом, образуя

Статья получена 1 мая 2024 года.

Д. И. Сафиканов, Национальный исследовательский ядерный университет «МИФИ» (e-mail: d.i.safikanov@mail.ru).

А. А. Артамонов, Национальный исследовательский ядерный университет «МИФИ» (e-mail: aartamonov@kaf65.ru).

Ю. Е. Фомина, Национальный исследовательский ядерный университет «МИФИ» (e-mail: yulya-fomina-1994@mail.ru).

А. И. Черкасский, Национальный исследовательский ядерный университет «МИФИ» (e-mail: aicherkasskij@mephi.ru).

мультиагентные системы [7], [8].

В рамках данной статьи описано использование статистической модели для агентного поиска целевых объектов в социальной сети «ВКонтакте». Под целевым объектом понимается сущность в социальной сети (например, профиль пользователя), отвечающая набору определенных характеристик, зависящему от конкретной задачи. Так, в качестве целевых объектов могут рассматриваться следующие категории пользователей:

- «фейковые» пользователи или боты [9];
- лидеры общественного мнения (так называемые инфлюенсеры) [10];
- сторонники или противники различных политических движений, идеологических направлений и т. д. [11];
- пользователи, потенциально опасные для окружающих (например, сторонники движения «колумбайнеров») [12], и другие.

II. Литературный обзор

Важность задачи выявления определенных целевых объектов нашла широкое отражение в научной литературе.

Значительная часть существующих исследований сосредоточена на выявлении спамеров и «фейковых» профилей, которые могут поставить под угрозу безопасность и конфиденциальность данных других пользователей. Исследовательская группа под руководством М. Файр [13] предложила метод идентификации «фейковых» пользователей в Facebook с помощью специальной информационной системы с использованием алгоритмов машинного обучения, которые учитывают силу связей между пользователями. Система сканирует список друзей пользователя и оценивает степень «доверия» к каждому из них. Однако точность метода относительно невелика. Были разработаны более точные методы идентификации ботов и спамеров. Например, Ф. Морстаттер и др. [14] предлагают алгоритм под названием BoostOR для обнаружения ботов в Twitter. Набор характеристик (а именно, соотношение ретвитов, средняя длина твита и время между твитами и др.) используется для выявления ботов среди реальных пользователей. Алгоритм обнаружения был протестирован на реальных наборах данных из Twitter, и результаты показали более высокую точность, чем другие существовавшие на тот момент подходы. Точный метод идентификации спамеров также был предложен А. Дж. Бану и др. [15], которые подготовили набор данных из 30 000 профилей китайской социальной сети Renren, вручную разделили эти профили на спамеров и обычных пользователей, а затем применили алгоритм машинного обучения для обнаружения спамеров в других наборах данных. По заявлениям авторов, предложенный метод характеризуется высокой точностью обнаружения спамеров (99%).

Помимо обнаружения спамеров и ботов, также существует значительный корпус литературы по выявлению пользователей, которые могут влиять на мнения других пользователей (так называемые инфлюенсеры). Пользователей этого типа можно обнаружить, проанализировав их активность в социальной сети (М. Трусов и др. [16]). Авторы предположили, что, если пользователь увеличивает время, проведенное им в социальной сети, и связанные с ним пользователи также становятся более активными, то этого человека можно отнести к инфлюенсерам. Напротив, если активность пользователя растет, но нет никаких изменений в активности связанных людей, этот пользователь, вероятно, не оказывает влияния на других. Однако применение данного метода ограничено, поскольку далеко не во всех социальных сетях возможно вычислить время, проведенное пользователями в них. Другой метод поиска инфлюенсоров основан на анализе моделей их поведения в социальных сетях. П. Харриган [17] на основе интерактивного анализа Twitter выявил закономерность, что у влиятельных людей больше подписчиков, они часто используют хэштеги и заглавные буквы, они чаще публикуют сообщения, но их твиты просматриваются меньшее количество раз. Эти метрики могут быть использованы для разработки специальной модели для машинной идентификации инфлюенсеров. Похожая модель, которая объединяет интеллектуальный анализ данных с социальными вычислениями, была предложена К. Люнгом и др. [18].

Некоторые авторы также предлагают комплексные подходы к выявлению разных категорий пользователей. Фазин и др. [19] разработали два метода классификации пользователей Twitter в качестве инфлюенсеров, спамеров или обычных пользователей. Первый метод учитывает силу связей между пользователями и использует простой линейный классификатор для отнесения пользователей к той или иной категории. Другой метод основан на сравнении паттернов твитов пользователей с паттернами прототипных инфлюенсеров или спамеров. Преимуществом второго метода является его более высокая эффективность в условиях малого количества доступных данных.

Кроме того, еще одним перспективным направлением исследований является выявление пользователей, склонных к совершению самоубийства. Д. Рамирес-Сифуэнтес и др. [20] предполагают, что признаки суицидальных мыслей могут быть идентифицированы посредством анализа мультимодальных данных, таких как паттерны сообщений, связи между пользователями, аудиовизуальный контент. Для анализа этих данных используются статистические методы и методы машинного обучения.

III. Методология

В данном исследовании предлагаются следующие этапы реализации выявления целевого объекта в социальной сети:

- формирование аналитиком обучающей выборки, которая содержит целевые объекты, найденные в результате интерактивного поиска в социальной сети;
- составление статистической модели целевого

объекта;

• определение граничных значений характеристик для отнесения объекта к целевым.

Целью первого этапа является выявление характеристик объекта – явных и неявных, которые свидетельствуют о его принадлежности к целевым. Для этого формируется группа экспертов. Экспертами составляется список аккаунтов, принадлежащих целевым объектам, другими словами – формируется обучающая выборка. Для включения аккаунта в обучающую выборку экспертом проводится его верификация по нескольким открытым источникам информации (например, новостным ресурсам).

На следующем шаге осуществляется анализ аккаунтов: для этого производятся сбор и систематизация данных, открытых настройками приватности, с личных страниц пользователей. Критерии целевых объектов определяются набором полей профиля, которые различаются в зависимости от социальной сети. Например, в социальной сети «ВКонтакте» используются более 30 критериев, которые можно условно разделить на следующие категории: основная информация, контакты, деятельность, интересы, жизненная позиция. Критерии объектов в социальных сетях могут описываться различными типами данных: числовыми данными, текстовыми данными, символами, изображениями, аудио- и видеофайлами [21]. Разнообразие типов используемых данных представляет сложность с точки зрения автоматизации их сбора и обработки. Кроме того, некоторые поля личных анкет пользователей могут оставаться незаполненными или быть скрыты настройками приватности. Нужно отметить, что не все имеющиеся критерии могут соответствовать задачам целевого поиска, в связи с чем некоторые из них могут быть исключены экспертами из рассмотрения. Задачей экспертов является выделение ряда критериев, дающих полное описание целевых объектов.

Степень значимости того или иного критерия для идентификации целевых объектов различается, что может быть отражено в количественной форме. Для этого критерии ранжируются экспертами в зависимости от их относительной значимости. Оценки экспертов агрегируются, и каждому критерию присваивается определенный весовой коэффициент *Wi*, значение которого находится в диапазоне от 0 до 1, при этом все весовые коэффициенты нормированы к единице.

Итогом первого этапа является перечень критериев целевого объекта и их весовые коэффициенты (иначе – максимальные значения), формализованные в табличном виде. Иными словами, формируется статистическая модель целевого объекта. Термин «статистическая» применяется в отношении модели в том смысле, что при ее составлении используются только эмпирические данные.

ТАБЛИЦА I. ГРУППОВОЕ ЭКСПЕРТНОЕ РАНЖИРОВАНИЕ И ОПРЕДЕЛЕНИЕ ВЕСОВ КРИТЕРИЕВ

Эксперты / **Критерии**	**1**	**2**	**…**	**i**	**…**	**N**
1	R_{11}	R_{21}	…	R_{i1}	…	R_{N1}
2	R_{12}	R_{22}	…	R_{i2}	…	R_{N2}
…	…	…	…	…	…	…
J	R_{1j}	R_{2j}	…	R_{ij}	…	R_{Nj}
…	…	…	…	…	…	…
K	R_{1K}	R_{2K}	…	R_{iK}	…	R_{NK}
$W = \frac{\sum_{j=1}^{K} R_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{K} R_{ij}}$	W_1	W_2	…	W_i	…	W_N

Целью следующего этапа является получение диапазона интегральных значений характеристик объекта, при попадании в который объект поиска будет считаться целевым.

Как было указано ранее, критерии могут описываться различными типами данных. Каждому критерию присваивается детерминированная оценка: численная, балльная или по тезаурусу с указанием диапазона возможных значений. Так, нечеткие данные оцениваются экспертами по баллам, а для текстов формируется тематический тезаурус. Упорядочивание объектов по значениям характеристики «текст» осуществляется по частоте встречающихся в нем терминов из тематического тезауруса. Для анализа фотографии профиля пользователя используются нейросетевые технологии, которые определяют наличие или отсутствие определенных элементов на фотографии. Аудиозаписи пользователя анализируются по принципу нахождения в поле маркированных аудиозаписей из «аудио» тезауруса.

Полученные детерминированные значения оценки имеют разную размерность, то есть являются неаддитивными величинами. С целью уйти от разных физических размерностей значений критериев вводится понятие относительного значения критерия.

Интегральное значение характеристик объекта определяется по следующей формуле:

$$E_m = \sum_{i=1}^{N} W_i \cdot \nu_{mi} \text{ , где}$$

W_i – весовой коэффициент i-го критерия;

ν_{mi} – относительное значение критерия;

N – количество критериев;

m – общее количество целевых объектов в обучающей выборке.

Отметим, что интегральное значение характеристик объекта должно находиться в диапазоне от 0 до 1.

Все объекты обучающей выборки являются целевыми, для них интегральное значение характеристик лежит в диапазоне $[E_{min}, E_{max}]$. Если интегральное значение характеристик нового объекта, не входящего в обучающую выборку, больше E_{min}, вычисленного по обучающей выборке, то объект является целевым и вносится в массив целевых объектов. E_{min}, который разделяет целевые и нецелевые объекты, будет называться маркером целевого объекта. Вышеописанная методика позволяет агенту выявлять и маркировать целевые объекты с целью их дальнейшего анализа и решения прикладных задач.

IV. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В ходе эксперимента группа из пяти экспертов сформировала выборку, состоящую из 100 аккаунтов

целевых объектов в социальной сети «ВКонтакте» (обучающая выборка). Далее на основе структуры профилей пользователей «ВКонтакте» эксперты выделили 21 критерий для определения целевых объектов. Критерии были отранжированы каждым экспертом по степени важности, где 1 – наименее важный, 21 – наиболее важный. Результаты ранжирования были агрегированы (таблица II), после чего был подсчитан вес каждого критерия. Полученное значение коэффициента конкордации (K = 0,517) отразило высокую степень согласованности мнений экспертов.

ТАБЛИЦА II. РЕЗУЛЬТАТЫ РАНЖИРОВАНИЯ

Критерий	**Эксперт (m)**					**Сумма рангов, полученная каждым критерием (ΣX_i)**
	1	**2**	**3**	**4**	**5**	
1	20,5	20,5	19	20,5	16,5	97
2	20,5	20,5	19	20,5	16,5	97
3	17	18	8	18,5	2	63,5
4	7,5	1	8	18,5	7,5	42,5
5	18	19	21	17	16,5	91,5
6	15	16	10	16	16,5	73,5
7	12	3,5	11,5	13,5	7,5	48
8	5,5	3,5	4	13,5	7,5	34
9	5,5	3,5	4	13,5	7,5	34
10	7,5	7,5	11,5	13,5	7,5	47,5
11	3	7,5	4	10	7,5	32
12	3	7,5	4	10	7,5	32
13	1	7,5	4	10	7,5	30
14	3	3,5	8	4,5	2	21
15	16	17	1	4,5	2	40,5
16	19	15	19	4,5	16,5	74
17	12	12,5	16,5	4,5	16,5	62
18	12	12,5	16,5	4,5	16,5	62
19	12	12,5	15	4,5	16,5	60,5
20	12	12,5	13	4,5	16,5	58,5
21	9	10	14	4,5	16,5	54

Практическим итогом проведенных работ стало создание информационно-аналитической системы (далее – система), позволяющей автоматически собирать информацию из социальной сети и идентифицировать целевые объекты. С этой целью был разработан агент, осуществляющий сканирование социальной сети и сбор информации из нее. В систему были внедрены описанные выше модели и технологии маркирования объектов. В ходе эксперимента в систему была загружена обучающая выборка из 100 аккаунтов, и был определен целевой интервал интегральных значений характеристик целевого объекта $[E_{min}, E_{max}]$.

Для проведения тестовых испытаний системы была сформирована контрольная выборка объектов, состоящая как из заведомо целевых объектов, так и заведомо нецелевых объектов (заведомо ложные объекты). Ключевым условием включения заведомо целевых объектов в контрольную выборку является то, что они не должны входить в обучающую выборку, на основе которой разрабатывалась методика маркирования объектов. Контрольная выборка объектов формировалась экспертами в ручном режиме, всего в нее вошло более 50 аккаунтов. Объекты контрольной выборки были загружены в систему и промаркированы. По результатам тестовых испытаний заведомо целевые объекты контрольной выборки должны иметь значения, входящие в заданный интервал $[E_{min}, E_{max}]$, а заведомо нецелевые объекты должны быть вне данного интервала.

После автоматической маркировки объектов полученные данные были выгружены из системы и проанализированы, на основе чего были вычислены показатели точности (precision) и полноты (recall) системы.

Под точностью подразумевается доля объектов, действительно являющихся целевыми, относительно всех объектов, которые система отнесла к целевым. Точность может быть рассчитана по следующей формуле:

$$P = \frac{TP}{TP + FP}, \text{ где}$$

TP - истинно-положительное решение системы;

FP - ложно-положительное решение системы.

Под полнотой подразумевается доля правильно идентифицированных целевых объектов к общему числу целевых объектов в контрольной выборке. Полнота может быть рассчитана по следующей формуле:

$$R = \frac{TP}{TP + FN}, \text{ где}$$

TP - истинно-положительное решение системы;

FN - ложно-отрицательное решение системы [22].

По результатам тестового испытания точность и полнота системы превысили 0,8.

V. ЗАКЛЮЧЕНИЕ

Увеличение роли социальных сетей в жизни человечества является одной из значимых тенденций XXI века. Анализ социальных сетей позволяет решать широкий круг различных задач – политических, экономических и других.

В представленной статье описано использование статистической модели для агентного поиска целевых объектов в социальной сети. Под целевыми объектами понимаются пользователи социальной сети, отвечающие заданному набору характеристик и представляющие интерес для исследователей в рамках решения конкретной задачи. К таким задачам могут относиться превентивное выявление потенциально опасных пользователей, отслеживание деятельности пользователей, имеющих широкую аудиторию и влияние на нее – так называемых инфлюенсеров, выявление мнений субъектов и социальных групп по различным вопросам и другие.

Предложенная методология поиска включает формирование аналитиком обучающей выборки целевых объектов, найденных в результате интерактивного поиска. На базе обучающей выборки строится статистическая модель целевого объекта: выделяются характеристики (критерии) целевых объектов, которые ранжируются экспертами по важности, на основе чего соответствующим критериям присваиваются весовые коэффициенты. После этого обучающая выборка

маркируется, и определяются граничные значения характеристик целевого объекта. Минимальное граничное значение является критерием отнесения объектов социальной сети к целевой группе (маркером). На следующем этапе производится настройка агентов на автоматическое выполнение целевого поиска.

Вышеописанная методология была использована при создании специализированной информационно-аналитической системы, которая продемонстрировала высокую точность и полноту в ходе тестовых испытаний.

Одним из возможных направлений дальнейших исследований является адаптация методологии и созданной системы к поиску целевых объектов в других социальных сетях. Однако важно принимать во внимание следующие факторы:

1. Различия в архитектуре социальных сетей, типах контента, разметке страниц и т.д.
2. Политика социальной сети в отношении автоматизированного сбора данных: владельцы некоторых социальных сетей могут выступать против такой деятельности на их ресурсе. Как в зарубежной, так и в отечественной практике имеются случаи судебных разбирательств между владельцами и лицами, осуществлявшими такой сбор информации из социальных сетей (например, судебное разбирательство между LinkedIn и hiQ Labs в 2017-2019 гг.).
3. Используемые механизмы противодействия автоматизированному сбору данных из социальной сети. Например, социальная сеть LinkedIn отслеживает поведение пользователя (в т.ч. количество запросов) и в случае подозрения на автоматизированный сбор данных временно ограничивает доступ к учетной записи, а при повторном нарушении – окончательно блокирует ее.

Вышеперечисленные факторы оказывают определяющие влияние при разработке подходов к поиску целевых объектов в той или иной социальной сети.

Библиография

[1] Винник Д. В. “Социальные сети как феномен организации общества: сущность и подходы к использованию и мониторинг,” *Философия науки*, 2012, № 4, с. 113.

[2] Chang B., Xu T., Liu Q., Chen E. H. “Study on information diffusion analysis in social networks and its applications,” *International Journal of Automation and Computing*, 2018, no. 15, pp. 1-26, doi:10.1007/s11633-018-1124-0.

[3] Artamonov A. A., Ionkina K. V., Kirichenko A. V., Lopatina E. O., Tretyakov E. S., Cherkasskiy A. I. “Agent-based search in social networks,” *International journal of civil engineering and technology*, 2018, vol. 9, no. 13, pp. 28-35.

[4] Гребенюк А. А., Максимова А. С., Лэмер Л. Г. “Исследование социальной напряженности на основе больших данных электронных социальных сетей,” *Цифровая социология*, 2021, том 4, № 4, с. 4-12, doi:10.26425/2658-347X-2021-4-4-4-12.

[5] Jin D., Ma X., Zhang Y., Abbas H., Yu H. “Information diffusion model based on social big data,” *Mobile networks and applications*, 2018, vol. 23, pp. 717-722, doi:10.1007/s11036-018-1004-4.

[6] Steinert-Threlkeld, Z. “Spontaneous collective action: peripheral mobilization during the Arab Spring,” *American Political Science Review*, vol. 111, no. 2, pp. 379-403, doi:10.1017/S0003055416000769.

[7] Artamonov A. A., Leonov D. V., Nikolaev V. S., Onykiy B.N., Pronicheva L.V., Sokolina K.A., Ushmarov I.A. “Visualization of semantic relations in multi-agent systems,” *Scientific visualization*, 2014, vol. 6, no. 3, pp. 68-76.

[8] Antonov E. V., Artamonov A. A., Orlov A. V., Nikolaev V. S., Zakharov V. P., Khokhlova M. V., Kontsevaya Yu. S., Bonartsev A. P., Voinova V. V. “Processing of scientific and technical information in interdisciplinary research by methods of mathematical and linguistic directed search by the example of the study of biomaterials for tissue engineering,” *International Journal of Open Information Technologies*, 2022, vol. 10, no. 11, p. 137, doi:10.25559/INJOIT.2307-8162.10.202211.134-140.

[9] Shinde S., Mane S. B. “Malicious profile detection on social media: a survey paper,” presented at the 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, doi:10.1109/ICRITO51393.2021.9596322.

[10] Harrigan P., Daly T., Coussement K., Lee J. A. “Identifying influencers on social media,” *International Journal of Information Management*, 2021, vol. 56, pp. 1-11, doi:10.1016/j.ijinfomgt.2020.102246.

[11] Fraiwan M. “Identification of markers and artificial intelligence-based classification of radical Twitter data,” *Applied Computing and Informatics*, 2022, doi:10.1108/ACI-12-2021-0326.

[12] Veijalainen J., Semenov A., Kyppö J. “Tracing potential school shooters in the digital sphere”. In: Bandyopadhyay S. K., Adi W., Kim Th., Xiao Y. (eds) Information Security and Assurance. ISA 2010. Communications in Computer and Information Science, vol 76. Springer, Berlin, Heidelberg, doi:10.1007/978-3-642-13365-7_16.

[13] Fire M., Kagan D., Elyashar A., Elovici Y. “Friend or foe? Fake profile identification in online social networks,” *Social Networks Analysis and Mining*, 2014, vol. 4. pp. 1-23, doi:10.1007/s13278-014-0194-4.

[14] Morstatter F., Wu L., Nazer T., Carley M., Liu H. “A new approach to bot detection: Striking the balance between precision and recall,” presented at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 533-540, doi:10.1109/ASONAM.2016.7752287.

[15] Banu A. J., Ahamed N. N., Manivannan B., Vanitha K., Musthafa M. M. “Detecting Spammers on Social Networks,” *International Journal of Engineering and Computer Science*, 2017, vol 6, pp. 20240-20247, doi:10.18535/ijecs/v6i2.14.

[16] Trusov M., Bodapati A., Bucklin R. “Determining Influential Users in Internet Social Networks,” *Journal of Marketing Research*, 2010, vol. 47, pp. 643-658, doi:10.1509/jmkr.47.4.643.

[17] Harrigan P., Daly T., Coussement K., Lee J. A. “Identifying influencers on social media,” *International Journal of Information Management*, 2021, vol. 56, pp. 1-11, doi:10.1016/j.ijinfomgt.2020.102246.

[18] Leung C., Tanbeer S., Cameron J. “Interactive discovery of influential friends from social networks,” *Social Networks Analysis and Mining*, 2014, vol. 4, pp. 1–13, doi:10.1007/s13278-014-0154-z.

[19] Fazeen M., Dantu R., Guturu P. “Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches,” *Social Networks Analysis and Mining*, 2011, vol. 1, pp. 241-254, doi:10.1007/s13278-011-0017-9.

[20] Ramírez-Cifuentes D., Freire A., Baeza-Yates R., Puntí J., Medina-Bravo P., Velazquez D., Gonfaus J., Gonzàlez J. “Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis,” *Journal of Medical Internet Research*, 2020, vol. 22, no. 7, pp. 1-16, doi:10.2196/17758.

[21] Peng S., Cao L., Zhou Y., Ouyang Z., Yang A., Li X., Jia W., Yu S. “A survey on deep learning for textual emotion analysis in social networks,” *Digital Communications and Networks*, 2021, doi:10.1016/J.DCAN.2021.10.003.

[22] Hassan S. U., Ahamed J., Ahmad K. “Analytics of machine learning-based algorithms for text classification,” *Sustainable Operations and Computers*, 2022, no. 3, pp. 238–248, doi:10.1016/J.SUSOC.2022.03.001.

Statistical model for the identification of target objects in a social network

D. I. Safikanov, A. A. Artamonov, Yu. E. Fomina, A. I. Cherkasskiy

***Abstract*—Taking into account the growing role of social media, their analysis is a topical issue in information technology. Its development can allow solving a wide range of applied tasks. However, when analyzing social networks, a number of problems arise related to the volume of data, their heterogeneity and unstructured nature. One of the ways to solve these issues is the use of agent technologies. The article describes the methodology of the use of statistical models for agent-based search of target objects, which involves the following actions. At the first stage, analysts form a training set of target objects through an interactive search in a social network. Then, the analysis of the training set is carried out in order to determine the evaluation criteria, rank them by importance and assign weight coefficients to them. As a result, a statistical model of the target object is formed. Based on the training sample, boundary values (markers) are determined for marking new objects as target ones. In the next step, agents are configured to automatically perform the target search. In the course of practical implementation, agent search has proved to be an effective tool for identifying target objects in a social network VKontakte. The proposed methodology can potentially be scaled for use in other social networks.**

***Keywords*—social network, target object, statistical model, multi-criteria analysis, agent search, big data analysis.**

REFERENCES

[1] Vinnik D. V. “Social networks as a phenomenon of society organization: their nature and approaches to their use and monitoring,” *Philosophy of Science*, 2012, no 4, p. 113.

[2] Chang B., Xu T., Liu Q., Chen E. H. “Study on information diffusion analysis in social networks and its applications,” *International Journal of Automation and Computing*, 2018, no. 15, pp. 1-26, doi:10.1007/s11633-018-1124-0.

[3] Artamonov A. A., Ionkina K. V., Kirichenko A. V., Lopatina E. O., Tretyakov E. S., Cherkasskiy A. I. “Agent-based search in social networks,” *International journal of civil engineering and technology*, 2018, vol. 9, no. 13, pp. 28-35.

[4] Grebenyuk A. A., Maksimova A. S., Lemair L. G. “Study of social tension based on electronic social networks big data,” *Digital Sociology*, 2021, vol. 4, no. 4, pp. 4-12, doi:10.26425/2658-347X-2021-4-4-4-12

[5] Jin D., Ma X., Zhang Y., Abbas H., Yu H. “Information diffusion model based on social big data,” *Mobile networks and applications*, 2018, vol. 23, pp. 717-722, doi:10.1007/s11036-018-1004-4.

[6] Steinert-Threlkeld, Z. “Spontaneous collective action: peripheral mobilization during the Arab Spring,” *American Political Science Review*, vol. 111, no. 2, pp. 379-403, doi:10.1017/S0003055416000769.

[7] Artamonov A. A., Leonov D. V., Nikolaev V. S., Onykiy B.N., Pronicheva L.V., Sokolina K.A., Ushmarov I.A. “Visualization of semantic relations in multi-agent systems,” *Scientific visualization*, 2014, vol. 6, no. 3, pp. 68-76.

[8] Antonov E. V., Artamonov A. A., Orlov A. V., Nikolaev V. S., Zakharov V. P., Khokhlova M. V., Kontsevaya Yu. S., Bonartsev A. P., Voinova V. V. “Processing of scientific and technical information in interdisciplinary research by methods of mathematical and linguistic directed search by the example of the study of biomaterials for tissue engineering,” *International Journal of Open Information Technologies*, 2022, vol. 10, no. 11, p. 137, doi:10.25559/INJOIT.2307-8162.10.202211.134-140.

[9] Shinde S., Mane S.B. “Malicious profile detection on social media: a survey paper,” presented at the 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, doi:10.1109/ICRITO51393.2021.9596322.

[10] Harrigan P., Daly T., Coussement K., Lee J. A. “Identifying influencers on social media,” *International Journal of Information Management*, 2021, vol. 56, pp. 1-11, doi:10.1016/j.ijinfomgt.2020.102246.

[11] Fraiwan M. “Identification of markers and artificial intelligence-based classification of radical Twitter data,” *Applied Computing and Informatics*, 2022, doi:10.1108/ACI-12-2021-0326.

[12] Veijalainen J., Semenov A., Kyppö J. “Tracing potential school shooters in the digital sphere”. In: Bandyopadhyay S. K., Adi W., Kim Th., Xiao Y. (eds) Information Security and Assurance. ISA 2010. Communications in Computer and Information Science, vol 76. Springer, Berlin, Heidelberg, doi:10.1007/978-3-642-13365-7_16.

[13] Fire M., Kagan D., Elyashar A., Elovici Y. “Friend or foe? Fake profile identification in online social networks,” *Social Networks Analysis and Mining*, 2014, vol. 4. pp. 1-23, doi:10.1007/s13278-014-0194-4.

[14] Morstatter F., Wu L., Nazer T., Carley M., Liu H. “A new approach to bot detection: Striking the balance between precision and recall,” presented at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 533-540, doi:10.1109/ASONAM.2016.7752287.

[15] Banu A. J., Ahamed N. N., Manivannan B., Vanitha K., Musthafa M. M. “Detecting Spammers on Social Networks,” *International Journal of Engineering and Computer Science*, vol 6, 2017, pp. 20240-20247, doi:10.18535/ijecs/v6i2.14.

[16] Trusov M., Bodapati A., Bucklin R. “Determining Influential Users in Internet Social Networks,” *Journal of Marketing Research*, 2010, vol. 47, pp. 643-658, doi:10.1509/jmkr.47.4.643.

[17] Harrigan P., Daly T., Coussement K., Lee J. A. “Identifying influencers on social media,” *International Journal of Information Management*, 2021, vol. 56, pp. 1-11, doi:10.1016/j.ijinfomgt.2020.102246.

[18] Leung C., Tanbeer S., Cameron J. “Interactive discovery of influential friends from social networks,” *Social Networks Analysis and Mining*, 2014, vol. 4, pp. 1–13, doi:10.1007/s13278-014-0154-z.

[19] Fazeen M., Dantu R., Guturu P. “Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches,” *Social Networks Analysis and Mining*, 2011, vol. 1, pp. 241-254, doi:10.1007/s13278-011-0017-9.

[20] Ramírez-Cifuentes D., Freire A., Baeza-Yates R., Puntí J., Medina-Bravo P., Velazquez D., Gonfaus J., Gonzàlez J. “Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis,” *Journal of Medical Internet Research*, 2020, vol. 22, no. 7, pp. 1-16, doi:10.2196/17758.

[21] Peng S., Cao L., Zhou Y., Ouyang Z., Yang A., Li X., Jia W., Yu S. “A survey on deep learning for textual emotion analysis in social networks,” *Digital Communications and Networks*, 2021, doi:10.1016/J.DCAN.2021.10.003.

[22] Hassan S. U., Ahamed J., Ahmad K. “Analytics of machine learning-based algorithms for text classification,” *Sustainable Operations and Computers*, 2022, no. 3, pp. 238–248, doi:10.1016/J.SUSOC.2022.03.001.

Safikanov D. I., National Research Nuclear University MEPhI (e-mail: d.i.safikanov@mail.ru).
Artamonov A. A., National Research Nuclear University MEPhI (e-mail: aartamonov@kaf65.ru).
Fomina Yu. E., National Research Nuclear University MEPhI (e-mail: yulya-fomina-1994@mail.ru).
Cherkasskiy A. I., National Research Nuclear University MEPhI (e-mail: aicherkasskij@mephi.ru).