

УДК 004.724.4

Д. В. Ландэ^{1,2}, А. Н. Грайворонская², Б. А. Березин¹

¹Институт проблем регистрации информации НАН Украины,
ул. Н. Шпака, 2, 03113 Киев, Украина

² Национальный технический университет Украины «КПИ»
Проспект Победы, 37, 03056 Киев, Украина

Мультиагентная модель распространения информации в социальной сети

Приведены правила функционирования мультиагентной сети, определены статистические закономерности, связанные с «жизненным циклом» агентов — информационных сообщений. Выявлены статистические закономерности, относящиеся к количеству лайков и репостов, распределение которых, как видно по результатам моделирования, соответствуют распределению Вейбулла. Данные моделирования проверены путем исследования реальной сети микроблогов Twitter.

Ключевые слова: социальная сеть, моделирование, распределение Вейбулла, мультиагентная система, распространение информации.

Введение

Моделирование распространения информации в социальных сетях позволяет исследовать соответствующие информационные процессы, выявлять закономерности, которые могут использоваться как в изучении механизмов передачи информации в таких сетях, так и уровня ее воздействия на людей, что является актуальной задачей [1].

В данной работе рассматривается мультиагентная модель распространения информации, ключевым элементом в которой является сообщение [2]. Каждое сообщение может вызывать разные типы реакций, такие как: положительные/отрицательные комментарии; выражение поддержки или отторжения (like/dislike — лайк/дизлайк); текст сообщения может быть скопирован, размещен на другой странице (repost — репост); также в одном сообщении может быть размещена ссылка на другое (link — линк). Поэтому в качестве агента в модели будем рассматривать сообщение, а эволюция агента будет связана с событиями, которые с ним происходят. В качестве основной характеристики агента введем «энергию» (E), отображающую актуальность сообщения и степень интереса к нему. Естественно, что старение информации или отрицательная реакция уменьшают энергию

© Д. В. Ландэ, А. Н. Грайворонская, Б. А. Березин

сообщения, а положительная реакция или появление ссылки на сообщение увеличивают его энергию.

Правила эволюции агента в модели

Агент появляется с начальной энергией E_0 , и с каждым дискретным отсчетом времени его энергия уменьшается на 1. Рассмотрим события, типичные для социальных сетей: лайк, дизлайк, репост, линк (предоставления ссылки на одного агента другим). Эти события влияют на энергию агента следующим образом: лайк повышает энергию на 1, дизлайк уменьшает на 1, репост повышает на 2, линк повышает на 1. С другой стороны, вероятность того, что какое-то из этих событий произойдет, зависит от актуальности сообщения, интереса к информации в нем, что в терминах модели выражается энергией. В связи с этим определим вероятности того, что с сообщением с энергией E произошло определенное событие следующим образом:

$$p_{like}^{(E)} = p_{l_0} \varphi(E); \quad p_{dislike}^{(E)} = p_{d_0} \varphi(E); \quad p_{repost}^{(E)} = p_{r_0} \varphi(E),$$

где p_{l_0} , p_{d_0} , p_{r_0} — параметры модели, а φ — это некоторая монотонно неубывающая функция от текущей энергии агента со значениями в $[0, 1]$. При падении энергии до 0, агент «умирает» и больше не рассматривается.

Моделирование динамики информационного потока

Моделирование динамики всего информационного потока начинается с одного агента. Появление нового агента возможно двумя способами. Первый заключается в копировании существующего агента с помощью операции репост. Также возможно самозарождение агента, что отвечает публикации нового сообщения. Таким образом, в каждый момент времени с определенными вероятностями, с каждым из агентов может произойти любое из событий. Также в любой момент времени с вероятностью p_s может появиться новый агент в результате самозарождения.

Рассмотрим жизненный путь одного агента. Агент появляется с начальным значением энергии E_0 , и далее его энергия изменяется в зависимости от событий, которые с ним происходят. Будем считать, что возможны два события: лайк и репост. За единицу времени может произойти одно из этих событий, оба одновременно, или не произойти ни одного.

Обозначим через ε_t значение энергии агента в момент времени t . Тогда значение энергии в следующий момент времени можно записать следующим образом:

$$\varepsilon_{t+1} = \varepsilon_t + \delta_t,$$

где δ_t является случайной величиной со значениями в $\{-1, 0, 1, 2\}$. В соответствии с правилами изменения энергии, введенными выше, увеличение энергии на

2 соответствует тому, что произошли одновременно лайк и репост; увеличение на 1 — произошел только репост; энергия не меняется, если был лайк; уменьшается на 1, если не произошло ни одно из событий. Следовательно, можно указать условное распределение δ_t при известной энергии ε_t :

$$\begin{aligned} P(\delta_t = 2 | \varepsilon_t = E) &= p_{like}^{(E)} p_{repost}^{(E)}, \\ P(\delta_t = 1 | \varepsilon_t = E) &= (1 - p_{like}^{(E)}) p_{repost}^{(E)}, \\ P(\delta_t = 0 | \varepsilon_t = E) &= p_{like}^{(E)} (1 - p_{repost}^{(E)}), \\ P(\delta_t = -1 | \varepsilon_t = E) &= (1 - p_{like}^{(E)}) (1 - p_{repost}^{(E)}). \end{aligned}$$

Данные формулы справедливы при $E > 0$. Далее используем обозначение $P_{\Delta}^{(E)} = P(\delta = \Delta | \varepsilon = E)$.

Процесс изменения энергии агента можно рассматривать как целочисленное случайное блуждание с переходными вероятностями:

$$p_{ij} = \begin{cases} p_{j-i}^{(i)}, & (j-i) \in \{-1, 0, 1, 2\}, \quad i > 0, \\ 1, & i = j = 0, \\ 0, & \text{в остальных случаях.} \end{cases}$$

Так как значение энергии в следующий момент времени зависит только от значения энергии в предыдущий момент времени, то стохастическая последовательность $(\varepsilon_0, \varepsilon_1, \dots, \varepsilon_t, \dots)$ является марковской цепью с переходными вероятностями p_{ij} . Граф, соответствующий такой марковской цепи, показан на рис. 1.

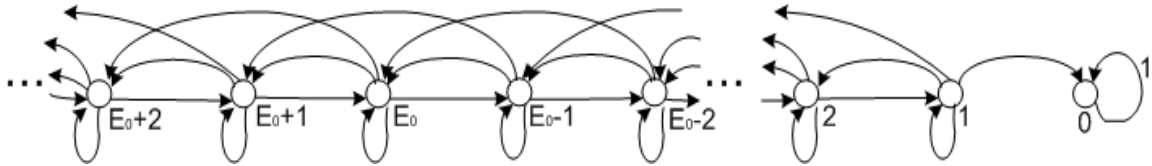


Рис. 1. Граф, соответствующий случайному блужданию энергии агента

Можно явно указать вероятностное распределение последовательности $(\varepsilon_0, \varepsilon_1, \dots, \varepsilon_t, \dots)$ через переходные вероятности:

$$P((\varepsilon_0, \varepsilon_1, \dots, \varepsilon_t, \dots) = (E_0, E_1, \dots, E_t, \dots)) = \prod_{i=1}^{\infty} p_{E_{i-1}E_i}.$$

Или аналогичное распределение для $(\delta_0, \delta_1, \dots, \delta_t, \dots)$:

$$P((\delta_1, \delta_2, \dots, \delta_t, \dots) = (\Delta_1, \Delta_2, \dots, \Delta_t, \dots)) = \prod_{i=1}^{\infty} P_{\Delta_j}^{(E_0 + \sum_{j=1}^{i-1} \Delta_j)}$$

при условии, что $\Delta_i \in \{-1, 0, 1, 2\}$, и если для некоторого $k: \sum_{i=1}^k \Delta_i = -E_0$ (энергия агента упала до 0 на k -м шаге), то $\Delta_i = 0, \forall i > k$. Для последовательностей $(\Delta_1, \Delta_2, \dots, \Delta_t, \dots)$, которые не удовлетворяют этим условиям, вероятность равна 0. Далее рассмотрим только те последовательности, которые удовлетворяют указанным условиям.

Теоретически существуют последовательности $(\Delta_1, \Delta_2, \dots, \Delta_t, \dots)$, которые реализуются с положительной вероятностью, и при этом $k: \sum_{i=1}^k \Delta_i = -E_0$ может быть как угодно большим. Заметим, что k соответствует длине жизни агента и имеет смысл времени, на протяжении которого информационное сообщение остается актуальным. Следовательно, длина жизни агента должна быть конечной с большой вероятностью.

Обозначим длину жизни агента через τ_{E_0} с начальным значением энергии E_0 или, что то же самое, время, за которое из E_0 попали в 0. В реалистичной модели хотелось бы иметь оценку $P(\tau_{E_0} > T_{\max}) < \varepsilon$ для малого ε не очень большого значения T_{\max} для того, чтобы можно было вместо бесконечных последовательностей $(\Delta_1, \Delta_2, \dots, \Delta_t, \dots)$ рассматривать конечные.

Рассмотрим функцию $\rho_T(E) = P(\tau_E > T)$. Справедливо рекуррентное соотношение:

$$\rho_T(E) = P_2^{(E)} \rho_{T-1}(E+2) + P_1^{(E)} \rho_{T-1}(E+1) + P_0^{(E)} \rho_{T-1}(E) + P_{-1}^{(E)} \rho_{T-1}(E-1).$$

Систему таких рекуррентных соотношений можно решить, используя начальные условия:

$$\rho_0(E) = \begin{cases} 0, & E = 0, \\ 1, & E \neq 0. \end{cases}$$

При начальных параметрах $p_{i_0} = 0,4$, $p_{r_0} = 0,1$ из решения рекуррентного уравнения можно получить оценку $P(\tau_{E_0} > 1,5E_0) < 10^{-3}$. То есть время жизни агента ограничено $1,5E_0$ с большой вероятностью и, следовательно, для получения достаточно точных оценок распределения для $(\delta_0, \delta_1, \dots, \delta_t, \dots)$ можно рассматривать векторы конечной длины $T_{\max} = 1,5E_0$.

Представляет интерес распределение количества событий, которые происходят с агентами. Рассмотрим распределение количества лайков. Заметим, что если в момент времени t произошел лайк, то $\Delta'_t \in \{0, 2\}$, если же нет, то $\delta_t \in \{-1, 1\}$.

Обозначим $(\Delta'_1, \dots, \Delta'_{T_{\max}})$ — вектор, который удовлетворяет условию, что $\Delta'_t \in \{0, 2\}$, при $t = t_1, \dots, t_n$ и $\Delta'_t \in \{-1, 1\}$ иначе, где $0 < t_1 < \dots < t_n < T_{\max}$. Тогда для агента, получившего лайк, справедлива формула:

$$P\{like\} = \sum_{t_1 < \dots < t_n} \sum_{(\Delta_1, \dots, \Delta_{T_{\max}})} \prod_{i=1}^{T_{\max}} P_{\Delta_i}^{(E_0 + \sum_{j=1}^{i-1} \Delta_j)}.$$

Численные результаты

На рис. 2 показана плотность полученного распределения при начальных параметрах $p_{l_0} = 0,4$, $p_{r_0} = 0,1$. Точками, соединенными отрезками прямых линий, обозначены полученные значения P {агент получил n лайков}.

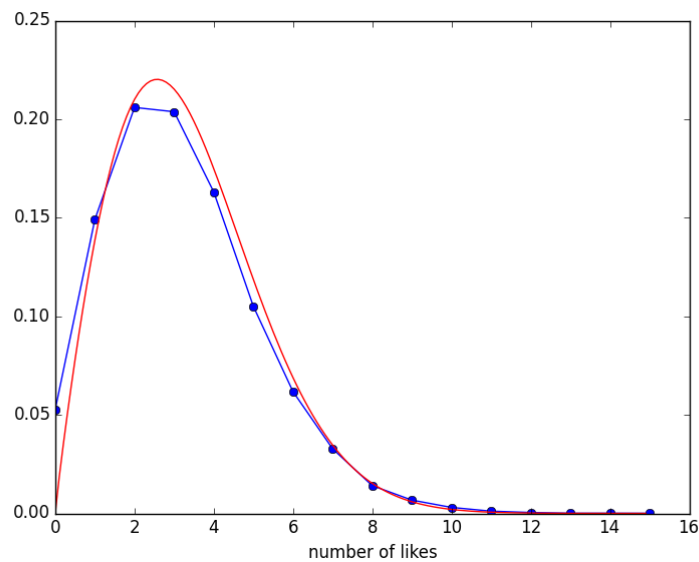


Рис. 2. Плотность распределения количества лайков, полученных агентом при начальных параметрах $p_{l_0} = 0,4$, $p_{r_0} = 0,1$

Плавная кривая на рис. 2 соответствует плотности распределения Вейбулла [3]:

$$f(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Параметры распределения Вейбулла k и λ были получены методом максимального правдоподобия. При указанных начальных параметрах, полученные значения $k = 1,9$, $\lambda = 3,8$ соответственно.

На рис. 3 представлен аналогичный результат при начальных параметрах $p_{l_0} = 0,3$, $p_{r_0} = 0,1$. В данном случае полученные параметры распределения $k = 1,9$, $\lambda = 3,8$ соответственно.

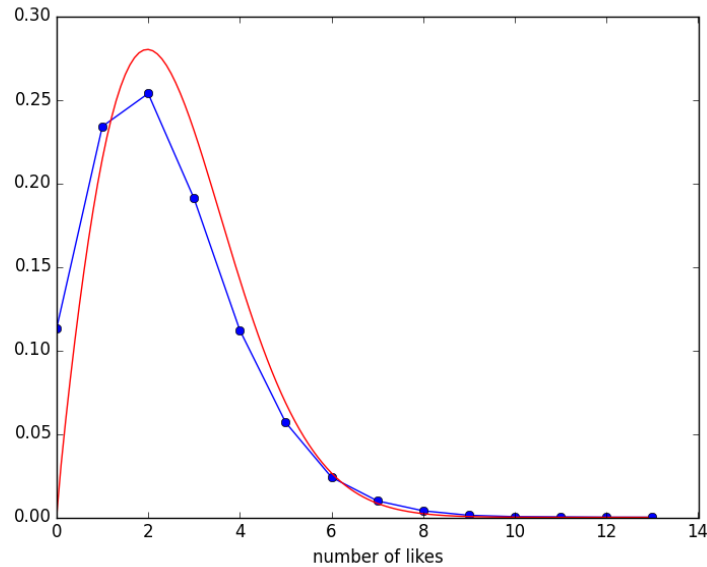


Рис. 3. Плотность распределения количества лайков, полученных агентом при начальных параметрах $p_{l_0} = 0,3$, $p_{r_0} = 0,1$

Исследование реальных информационных потоков

Полученные результаты моделирования сравнивались с проведенными авторами результатами исследования жизненного цикла новостных сообщений в сети микроблогов Twitter, где, в частности, анализировались характеристики роста количества специальных репостов (ретвиттов) выбранных сообщений [4]. Распределение лайков и ретвиттов в этом случае, как и в модели, соответствовало стандартному распределению Вейбулла, причем параметр k с высокой точностью совпал с модельным (рис. 4).

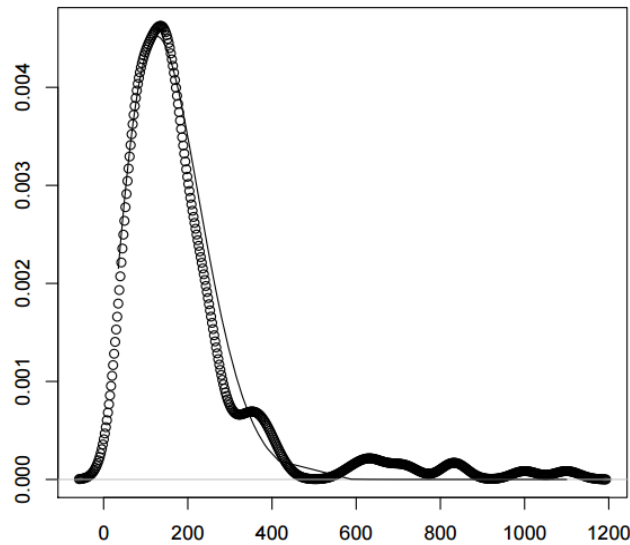


Рис. 4. Плотность распределения количества ретвитов, полученных из реальной сети (аппроксимация распределением Вейбулла при $k = 1,9$, $\lambda = 180$)

Программные средства, разработанные с помощью языка программирования *R* для статистических расчетов, состояли из трех частей:

- средств сканирования и накопления данных о росте количества ретвитов для сообщений отдельного пользователя сети в режиме он-лайн (для сообщений от газеты Нью-Йорк Таймс был выбран цикл сканирования в 15 с);

- обработки накопленной информации путем аппроксимации данных о росте количества ретвитов функцией Вейбулла, получения соответствующих коэффициентов масштаба и формы (нелинейная аппроксимация с помощью метода наименьших квадратов), а также вычисления производной для оценки скорости роста количества ретвитов сообщений и построения необходимых графиков;

- накопления полученных результатов для дальнейшего анализа. С этой целью данные, полученные с помощью программных средств на языке *R*, импортируются во внешние базы данных.

Таким образом, накапливались записи, включающие текст каждого сообщения, время его публикации в сети, значения коэффициентов масштаба и формы, графики роста количества ретвитов и лайков, графики аппроксимации данных о росте ретвитов функцией Вейбулла, график скорости роста количества ретвитов и т.д.

Выводы

В результате описанных исследований построена мультиагентная модель «жизненного цикла» новостей в информационных сетях.

В результате моделирования выявлены статистические закономерности, относящиеся к количеству лайков и репостов отдельных сообщений, распределение которых, как видно по результатам моделирования, соответствуют распределению Вейбулла.

Данные моделирования проверены путем исследования реальной сети микроблогов Twitter. Совпадение результатов моделирования и параметров распреде-

ления реальной сети позволяют говорить о закономерности, присущей реальным сетям, а также об адекватности модели.

Найденные закономерности могут использоваться как при изучении механизмов распространения информации в социальных сетях, так и при определении аномалий в жизненном цикле отдельных сообщений, информационных операций, искусственно поддерживаемых информационных кампаний [5].

1. *Конкурентная разведка в компьютерных сетях* / [Додонов А.Г., Ландэ Д.В., Прищепа В.В., Пуятин В.Г.]. — К.: ИПРИ НАН Украины, 2013. — 248 с.

2. *Грайворонська А.М., Ланде Д.В. Дослідження інформаційних потоків як динамічних мультиагентних систем* // Системний аналіз і інформаційні технології: матеріали 17-й Міжнародної науково-технічної конференції SAIT 2015, Київ, 22–25 червня 2015 р. / УНК «ІПСА» НТУУ «КПІ». — К.: УНК «ІПСА» НТУУ «КПІ», 2015. — С. 62–63.

3. *Нешиной В.В. Математико-статистические методы анализа в библиотечно-информационной деятельности* / В.В. Нешиной. — Минск: Белорус. гос. ун-т культуры и искусств, 2009. — 203 с.

4. *CTEDAS: A Twitter-based Event Detection and Analysis System* / Li R., Lei K.H., Khadiwala R., Chang K.C. // Data Engineering (ICDE), 2012 IEEE 28th International Conference, 2012. — P. 1273–1276.

5. *Додонов А.Г. Распознавание информационных операций: мультиагентный подход* / А.Г. Додонов, Д.В. Ландэ, В.А. Додонов // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2016): материалы VI междунар. науч.-техн. конф. (Минск 18–20 февраля 2016 года). — Минск: БГУИР, 2016. — С. 253–256.

Поступила в редакцию 19.03.2016