УДК 519.8

В.А. Громов, А.С. Конев
Днепропетровский национальный университет им. Олеся Гончара

ПРИМЕНЕНИЕ МЕТОДОЛОГИИ ПРОГНОЗИРОВАНИЯ С ПОМОЩЬЮ КЛАСТЕРИЗАЦИИ ДЛЯ РАННЕЙ ИДЕНТИФИКАЦИИ ПОПУЛЯРНЫХ ТЕМ В СОЦИАЛЬНОЙ СЕТИ «TWITTER»

Рассмотрен метод определения популярности темы в социальной сети «Twitter», нацеленный на раннюю идентификацию тем, которые с некоторого момента демонстрируют взрывной рост популярности; прогнозирование здесь осуществляется в рамках парадигмы прогнозирования на основе кластеризации; в качестве алгоритмов кластеризации применяются алгоритмы, основанные на теории моделирующего поля. Метод предполагает определение настоящей (а не указанной в заголовке) темы сообщения с последующим прогнозированием популярности той или иной темы. В ходе широкомасштабного вычислительного эксперимента были выявлены характерные варианты «предвзрывной» динамики популярности тем; некоторые из них оказались эквивалентными эмпирическим приёмам предсказания роста популярности темы, они известны специалистам по размещению рекламы в указанной социальной сети («краб», «бабочка Песавенто» и др.).

Розглянуто метод визначення популярності теми в соціальній мережі «Twitter», націлений на ранню ідентифікацію тем, що з певного моменту демонструють вибухове зростання популярності; прогнозування тут здійснюють у межах парадигми прогнозування на основі кластеризації; як алгоритми кластеризації застосовують алгоритми теорії моделюючого поля. Метод передбачає визначення дійсної (а не зазначеної у заголовку) теми повідомлення із подальшим прогнозуванням популярності тієї чи іншої теми. У процесі широкомасштабного обчислювального експерименту виявлено характерні варіанти «передвибухової» динаміки зростання популярності теми; деякі з них еквівалентні емпіричним прийомам прогнозування зростання популярності теми, які відомі спеціалістам із розміщення реклами в зазначеній соціальній мережі («краб», «метелик Песавенто» та ін.).

The paper outlines a method to estimate possible popularity for a certain topic for social network «Twitter»: the method designed to precociously identify the topics able to demonstrate «explosive» growth in popularity rate; predictive clustering is employed to predict topics popularity; one makes use modeling field theory algorithms as clustering techniques. First of all, the method ascertains the real (not written in hash-tag) topics of twits, and then predict popularity rates for the topics. In the course of wide-ranging simulation, typical variants of «pre-explosive dynamics» were revealed; some of them were turned out to be equal to heuristic techniques to predict topics popularity well-known for PR community working in the network («crab», «Pesavento's butterfly» etc).

Ключевые слова: прогнозирование на основе кластеризации, теория моделирующего поля, «Twitter».

Введение. Всевозрастающая роль социальных сетей в формировании и анализе общественного мнения обусловливает пристальное внимание, которое уделяют политтехнологи и специалисты по проведению рекламных кампаний работе именно с этим сегментом медиа-пространства. Правильная оценка популярности темы на этапе её появления и начального развития делает возможным проведение эффективной рекламной кампании и оптимизацию расходов по соответствующей статье. Особый интерес представляет раннее выявление (идентификация) так называемых взрывных тем (всплесков), то есть тем, которые, будучи некоторое время сравнительно малопопулярными, внезапно приобретают значительную популярность.

Одной из наиболее популярных и динамично развивающихся социальных сетей является «Twitter». Здесь следует отметить, что постоянный интерес к идентификации взрывных тем на начальных стадиях существования темы со стороны специалистов по размещению рекламы в социальных медиа обусловил, с одной стороны, появление большого количества эмпирических приёмов предсказания роста популярности той или иной темы в указанной социальной сети, подобных фигурам технического анализа («краб», «бабочка Песавенто» и т.д.), а с другой – значительного числа работ, связанных с определением популярности темы с применением того или иного механизма извлечения знаний.

Так, в работе [14] представлена система PoliTwi, предназначенная для выявления политических тем в «Twitter», популярность которых начинает расти; авторы отмечают, что информация в «Twitter» распространяется быстрее, чем в других информационных средах. Рост популярности оценивается здесь с помощью отношения числа сообщений, связанных с определённой темой, в предыдущий период к числу сообщений в настоящий момент времени с учётом величины стандартного отклонения (распределение числа считается гауссовым). Работа [3] посвящена определению характеристик, описывающих эмоциональную окрашенность того или иного твита, а также сравнению методов оценки указанных характеристик.

В настоящем исследовании в качестве механизма извлечения знаний применяется метод прогнозирования, принадлежащий парадигме прогнозирования с помощью кластеризации (predictive

clustering), причём в качестве элемента временного ряда выступает число ретвитов, которые получила та или иная тема в определённый промежуток времени.

Классическая постановка задачи прогнозирования предполагает анализ единственного временного ряда и построение на его основе прогнозной модели для получения оценок последующих наблюдений рассматриваемого ряда. Здесь возможно как построение одной модели для всех имеющихся в распоряжении исследователя наблюдений (как в случае применения классических методов прогнозирования – ARIMA, GARCH и др.), так и декомпозиция ряда на отдельные участки с помощью алгоритмов кластеризации с последующим построением значительного числа подмоделей, каждая из которых связана со своим кластером, как это предполагается в парадигме прогнозирования на основе кластеризации [2]; методы, построенные в рамках указанной парадигмы, позволяют эффективно прогнозировать хаотические временные ряды в силу того, что участки ряда, связанные с различными областями странного аттрактора, «оказываются» в различных кластерах с каждым из которых связана своя прогнозная модель [5;6].

В то же время задачи прогнозирования, характерные для медицинской статистики, при построении моделей потребления электроэнергии [9;11], динамики земной коры (предсказание землетрясений) и атмосферных фронтов (прогноз погоды) имеют значительное количество «родственных» временных рядов, которые допускают совместный анализ и построение единых прогнозных моделей.

Следует подчеркнуть, что кластеризация данных, осуществляемая в рамках прогнозирования на основе кластеризации, и построение тем или иным способом прогнозных моделей, отвечающих полученным кластерам, может рассматриваться как способ автоматической экстракции знаний, что является сильной стороной методов, принадлежащих указанной парадигме.

В работе [4] рассматривается возможность применения алгоритмов построения деревьев решений для построения алгоритмов прогнозирования на основе кластеризации. В исследовании [11] предложен алгоритм прогнозирования на основе характерных последовательностей (*pattern sequence-based forecasting* (*PSF*)), основывающийся на методе *k*-средних; представлены результаты применения указанного алгоритма в прогнозировании суточных продаж на испанском рынке электроэнергии. Поскольку метод *k*-средних предполагает знание точного числа реально существующих кластеров, авторы провели широкомасштабный вычислительный эксперимент для определения оптимального числа кластеров, то есть числа кластеров, позволяющего построить наиболее эффективную прогнозную модель. Работа [6] посвящена применению метода «муравьиных колоний» для получения кластеров и построения прогноза для хаотических временных рядов; в публикации [5] с этой целью применялся алгоритм Уишарта – данный алгоритм кластеризации не требует предварительного знания числа кластеров, их оптимальное число определяется в процессе функционирования алгоритма.

В работе [15] кластеризируются не сами отрезки временных рядов, а коэффициенты регрессионных многочленов; алгоритм был применён для прогнозирования индонезийских биржевых индексов. В [16] аналогичный подход применялся для прогнозирования множественных («родственных») временных рядов. Концепция «лейтмотива» (motif) была введена в работе [13] как инструмент извлечения наиболее значимой информации с целью получения «хорошей» кластеризации. В исследовании [10] предлагается применять алгоритм нечётких *c*-средних для выделения из временного ряда нечётких информационных «гранул» с последующим построением прогноза с помощью нечёткой самоорганизующейся карты высокого порядка. Следует отметить также работу [8], в которой представлен обзор нейросетевых моделей, относящихся к парадигме эволюционирующих коннекционистских открытых систем (ECOS); парадигма предполагает выделение и обучение локальных моделей путём кластеризации с последующей их структуризацией для построения единой модели.

Важной особенностью парадигмы прогнозирования на основе кластеризации является наличие так называемых непрогонозируемых точек [5;6], то есть точек, для которых с помощью метода нельзя указать прогнозные значения – не существует кластера (определённого тем или иным способом, в зависимости от метода), центр которого был бы близок к участку ряда, по которому необходимо построить прогноз. Следует подчеркнуть, что наличие непрогонозируемых точек является существенным достоинством указанного класса методов: здесь есть возможность явно указать на невозможность прогноза в некоторой точке, в отличие от методов классической парадигмы, которые в любом случае будут давать прогноз, даже если он заведомо ошибочный. В [5] указывается на эффективность применения «непоследовательных» наблюдений (наблюдений, номера которых соответствуют некоторому шаблону) в алгоритмах данного класса.

В настоящей статье предложен новый метод прогнозирования популярности тем социальной сети «Twitter», в том числе идентификации взрывных тем; метод основан на применении алгоритмов теории моделирующего поля [12] для кластеризации участков временных рядов.

Дальнейшее изложение структурировано следующим образом: во втором пункте настоящей работы представлены описания алгоритмов кластеризации, позволяющих выделить настоящую (а не формально указанную в заголовке) тему сообщения и осуществить прогнозирование популярности реальной темы. В третьем пункте работы представлены результаты применения указанного алгоритма в прогнозировании модельного временного ряда – ряда Лоренца (который имеет хаотический характер) – и для выявления взрывных тем в англоязычном «Twitter». Наконец, в заключительном, четвёртом, пункте сформулированы выводы.

Применение методов прогнозирования, построенных в рамках классической парадигмы, требует наличия отрезков ряда весьма значительной длины для получения адекватных оценок параметров модели, что явным образом вступает в противоречие с рассматриваемой задачей, где для отдельной темы число наблюдений измеряется сотнями, максимум тысячами, и тем более с задачей ранней идентификации взрывных тем, в рамках которой чем меньшее число наблюдений применяется при идентификации взрывной темы, тем лучше.

В классической прогнозной парадигме указанное противоречие, по-видимому, неразрешимо, в то время как в рамках парадигмы прогнозирования на основе кластеризации можно предложить использование наблюдений не только рассматриваемого ряда, но и всей совокупности подобных («родственных») ему рядов. С этой целью в качестве исходных данных в настоящей работе рассматривалась вся совокупность временных рядов, порождённых направлением «финансы», в англоязычном сегменте «Twitter». Это позволило, с одной стороны, обеспечить весьма значительный объём данных (порядка 10000 точек), а с другой – гарантировать их известную однородность.

Другой существенно важной для построения эффективного прогноза особенностью рассматриваемой предметной области является отсутствие надёжных оснований для отнесения конкретного сообщения (твита) к той или иной теме: в некоторых сообщениях хеш-теги вообще отсутствуют, в значительном числе других случаев проставленные авторами хеш-теги не позволяют однозначно определить, к какой именно теме относится сообщение; более того, в некоторых случаях авторские хеш-теги неадекватны реальной теме сообщения. Это обусловливает необходимость определения реальной темы сообщения путём анализа текста сообщения методами интеллектуального анализа текстов (text mining).

Постановка задачи. Рассматривается множество кратких текстовых сообщений (твитов социальной сети «Twitter»). Необходимо, во-первых, основываясь на текстах сообщений, отнести каждое сообщение к той или иной «настоящей» теме – темы выделяются в процессе классификации сообщений; во-вторых, выбрав в качестве меры популярности темы число ретвитов сообщений, относящихся к теме, построить прогноз роста популярности темы с течением времени. Наконец, выделить среди всех возможных характерных вариантов динамики роста/падения популярности темы те, которые позволяют с заданным ограничением на величину ошибки предсказать появление взрывной динамики по результатам наблюдения ряда до наступления динамики указанного типа или на ранних ее этапах.

Математически первая задача формулируется как задача классической кластеризации, сформулированная для многомерных векторов, указывающих на принадлежность того или иного слова из стандартного словаря наиболее популярных слов («word bag»).

Вторая задача может быть сформулирована как задача прогнозирования на основе кластеризации. Здесь рассматриваются отрезки «родственных» временных рядов $y_0, y_1, \ldots, y_{t_s}$, $s = \overline{1, S}$, где S – общее число временных рядов; y_{t_s} – число наблюдений в s-м ряде. Необходимо построить прогнозные значения $\tilde{y}_{t_s+1}, \ldots, \tilde{y}_{t_s+K}$ для K следующих позиций во временном ряде с тем, чтобы ошибка прогнозирования не превышала заданной величины ε: $\left|\tilde{y}_{t_s+i} - y_{t_s+i}\right| < \varepsilon, i = \overline{1, K}$, $s = \overline{1, S}$, либо указать на невозможность прогноза для той или иной позиции $t_s + i, i = \overline{1, K}$, $s = \overline{1, S}$; доля такого рода позиций от общего числа позиций, для которых необходимо построить прогноз (KS), не должна превышать заданной величины α.

Наконец, третья задача сводится к выявлению среди характерных вариантов динамики тех временных рядов (выявленных в процессе решения второй задачи), которые позволяют осуществить раннюю идентификацию взрывных тем с заданными ограничениями на ошибки 1-го и 2-го рода.

Алгоритм построения прогноза. Для определения «настоящей» темы сообщения и для осуществления кластеризации, необходимой для построения прогноза, применялся один и тот же алгоритм кластеризации – адаптивная нейросетевая система, применяющая принцип максимального правдоподобия (*maximum likelihood adaptive neural system* (*MLANS*)); алгоритм основывается на принципах теории моделирующего поля (*modeling field theory* (*MFT*)), предложенной Л. И. Перловским [12].

Алгоритм обучения модели состоит из двух подсистем – моделирующей и ассоциативной:

1. Моделирующая подсистема производит оценку параметров распределений, составляющих гауссову смесь, а именно оцениваются компоненты векторов математических ожиданий многомерных гауссовых распределений, их дисперсионно-ковариационных матриц, а также математическое ожидание попадания вектора в тот или иной кластер.

2. Ассоциативная подсистема позволяет оценить для каждого вектора значения нечётких функций принадлежности кластерам.

Взаимная обусловленность параметров, оценка которых осуществляется в рамках ассоциативной и моделирующей подсистем, приводит к реализации алгоритма MLANS в виде итерационного процесса; критерием окончания данного итерационного процесса выступает малость изменения оценок параметров модели на двух соседних итерациях: в качестве метрики применяется расстояние Бхаттачария (Bhattacharyya).

С применением обозначений $x_n, n = \overline{1,N}$ для множества векторов, подлежащих кластеризации, $N_k, M_k, C_k, k = \overline{1,K}$ для соответственно ожидаемого числа векторов, принадлежащих k-му кластеру, математического ожидания и дисперсионно-ковариационной матрицы многомерного нормального распределения $pdf(x_n \mid k)$, описывающего плотность вероятности принадлежности данному кластеру, f_{kn} для значения нечёткой функции принадлежности n-го вектора k-му кластеру соотношения моделирующей подсистемы могут быть записаны в следующем виде:

$$N_k = \sum_{n=1}^{N} f_{kn},\ M_k = \frac{1}{N_k}\sum_{n=1}^{N} f_{kn} x_n,\ C_k = \frac{1}{N_k}\sum_{n=1}^{N} f_{kn}\left(x_n - M_k\right)^T\left(x_n - M_k\right);$$

а соотношения ассоциативной подсистемы –

$$f_{kn} = \frac{pdf(x_n \mid k)}{\sum_{i=1}^{K} pdf(x_n \mid i)}.$$

Расстояние Бхаттачария между плотностями двух нормальных распределений с математическими ожиданиями и дисперсионно-ковариационными матрицами M_1, M_2 и C_1, C_2 даётся выражением

$$\rho = \frac{1}{8}\Delta M \overline{C}^{-1} \Delta M + \frac{1}{2}\ln\left[\frac{\det \overline{C}}{\sqrt{\det\left(C_1 C_2\right)}}\right],$$

где $\Delta M = M_1 - M_2,\ \overline{C} = \frac{1}{2}\left(C_1 + C_2\right).$

К недостаткам метода следует отнести необходимость априорного знания числа кластеров, что обусловливает необходимость проведения дополнительного исследования зависимости качества прогноза от числа кластеров.

Для формирования выборки для задачи выделения «настоящих» тем был сформирован стандартный словарь наиболее популярных слов (список тем), в который были включены имена существительные и глаголы, наиболее часто встречающиеся в используемом корпусе текстов. Бинарные вектора, образующие выборку, формировались по следующему правилу: i-й компонент вектора был равен 1, если в сообщение входило слово (или однокоренное ему), стоящее на i-й позиции в словаре, и 0 в противном случае.

Для данной задачи число кластеров равнялось числу тем в словаре (50). В качестве начальных условий для алгоритма выбирались «чёткие» значения $f_{kn}:\ f_{kn} \in \{0,1\},\ \sum_{k=1}^{K} f_{kn} = 1.$

После осуществления кластеризации сообщение считается отнесённым к той или иной «настоящей» теме, если оно принадлежит одному кластеру с бинарным вектором, точно соответствующим некоторой теме, то есть имеющим только одну ненулевую компоненту, а именно ту, номер которой совпадает с номером соответствующей темы в словаре.

Выделение «настоящих» тем позволило сформировать временные ряды, характеризующие изменение популярности темы (число ретвитов, относящихся к теме сообщений) с течением времени.

Для формирования выборки для задачи прогнозирования популярности темы из сформированных таким образом временных рядов конструировались вектора выборки из всевозможных наборов наблюдений временного ряда, номера которых удовлетворяют некоторому шаблону [5; 6]. Под шаблоном будем понимать фиксированную последовательность расстояний между номерами (в общем случае не следующих непосредственно друг за другом) наблюдений временного ряда. Тем самым обучающая выборка формировалась из непоследовательных (non-successive) наблюдений, что лучше отвечает природе рассматриваемой задачи. Для получения эффективного прогноза здесь применялись все возможные шаблоны длины 4, в которых разность между соседними элементами не превышает 10; при этом описанный выше алгоритм кластеризации применялся отдельно к выборке, полученной для каждого из шаблонов.

При прогнозировании усечённые на один элемент центры всех кластеров, построенные по всем шаблонам, сравниваются с вектором, получаемым наложением соответствующего шаблона на прогнозируемый временной ряд таким образом, чтобы последний элемент шаблона совпадал с позицией прогнозируемого элемента. В качестве прогнозного значения принимается последний элемент того центра, усечённый вариант которого оказался ближайшим (в смысле евклидовой метрики) к построенному таким образом вектору.

Для улучшения качества прогноза здесь применялся приём отсечения кластеров с низкой прогнозной ценностью: для этого вводилась дополнительная (проверочная) выборка, и значение этой характеристики (при заданном максимальном допустимом значении погрешности α) для k-го кластера вычислялось как $I_k(\alpha) = \sum_{i \in S_k} \frac{\bar{\varepsilon}_i}{\varepsilon_{ki}} \frac{1}{|V_i| - 1},\ \bar{\varepsilon}_i = \frac{1}{|V_i|}\sum_{j \in V_k} \varepsilon_{ji}$, где V_i – множество кластеров, которые могут быть применены для прогнозирования i-го наблюдения с ошибкой, меньшей заданного значения α; ε_{ji} – ошибка,

соответствующая прогнозу i-го наблюдения с применением центра j-го кластера, $j \in V_i$; $\bar{\varepsilon}_i = \frac{1}{|V_i|} \sum_{j \in V_k} \varepsilon_{ji}$ – средняя ошибка прогнозирования для i-го наблюдения по всем кластерам; S_k – множество таких наблюдений, что центр k-го кластера обеспечивает прогноз с ошибкой меньшей, чем заданное значение α.

Следует отметить, что применение приведенного выше метода прогнозирования с помощью тем, определяемых самими пользователями в хеш-тегах, приводило к существенно худшим прогнозным результатам.

Прогнозирование популярности тем социальной сети «Twitter». Для проверки эффективности работы алгоритма была применена совокупность временных рядов, порождённых системой Лоренца [7]; в качестве базовых применялись стандартные «хаотические» значения параметров $\sigma = 10, b = \frac{8}{3}, r = 28$; далее значение параметра r «зашумлялось» белым шумом с дисперсией, равной 1.0, и модифицированная таким образом система применялась для генерации временных рядов.

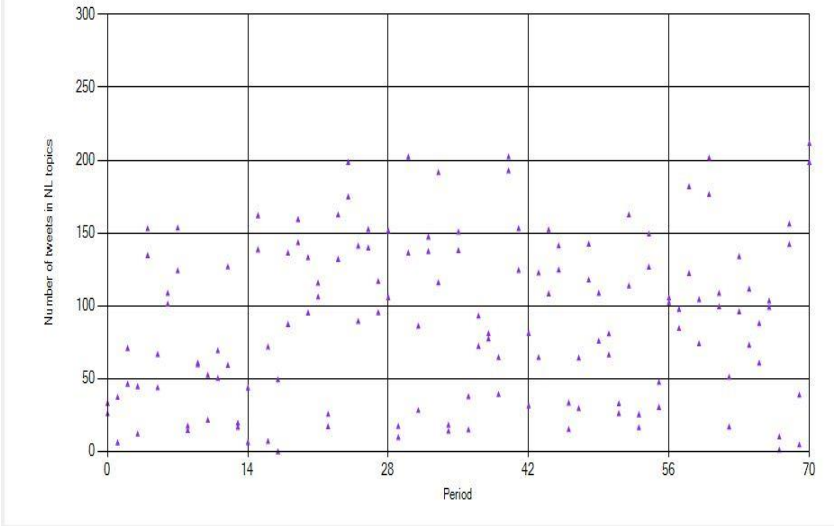
Для получения временных рядов здесь применялся метод Рунге-Кутта 4-го порядка точности: шаг интегрирования составлял 0.1, уровень шума соответственно может быть оценён как 10^{-5}. Для обеспечения выполнимости условий теоремы Такенса (движение наблюдаемой траектории в окрестности странного аттрактора системы), а следовательно, применимости прогнозных алгоритмов указанного класса [5] отбрасывались первые 800 значений временного ряда, полученного интегрированием системы. Тестирующее множество здесь состояло из 3000 наблюдений, размер обучающего множества составлял 10000 наблюдений.

Для ряда Лоренца и для тех временных рядов, характеризующих популярность тем в «Twitter», размеры которых позволяют это сделать, были получены (с помощью метода аналога [1]) оценки значений старших показателей Ляпунова. Во всех случаях значения показателей Ляпунова строго положительны, что является признаком хаотичности рассматриваемых рядов; в частности, для ряда Лоренца значение старшего показателя Ляпунова составило 0.91, что хорошо согласуется с представленными в литературе результатами [Там же].

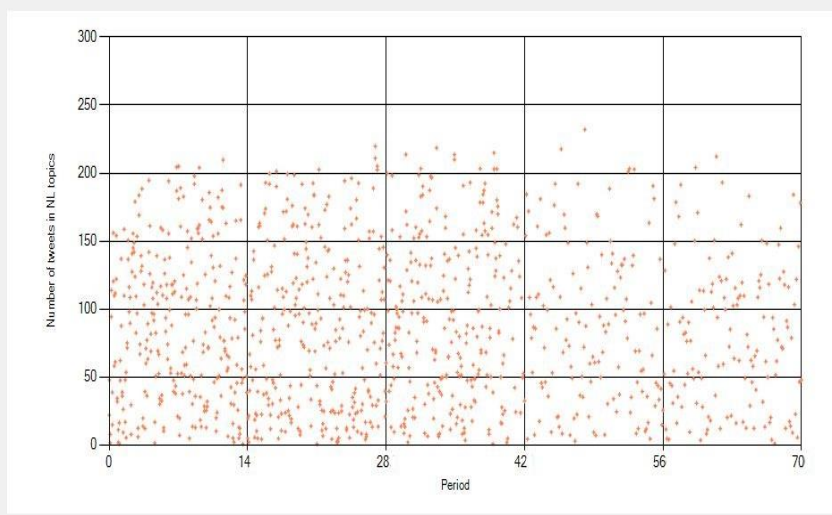
Среднеквадратичная ошибка (при прогнозировании на один шаг вперёд) для рядов, порождённых системой Лоренца, составила 2,2 %.

Для определения популярности той или иной тематики в социальной сети «Twitter» была сформирована обучающая выборка из 111000 сообщений, охватывающих период с 2006 по 2014 г.; выборка формировалась из сообщений англоязычных экономических и финансовых изданий, выкладываемых ими в «Twitter».

После необходимой предварительной обработки информации – очистки от ссылок, сокращений, хеш-тегов, имён пользователей, а также от вспомогательных частей речи – была проведена кластеризация с выделением сообщений, относящихся к той или иной «настоящей» теме; наиболее популярными оказались темы (в порядке убывания популярности): «нефть» («oil»), «биткойн» («bitcoin»), «инвестиции» («investments»), «кризис» («crisis»), «валюта» («currency»), «евро» («euro»), «долг» («debt»). 88 % тем в этом списке оказались именами существительными. В значительном числе случаев было установлено расхождение между «настоящей» темой и темой, которую ставит автор сообщения (хеш-тег).



а



б

Рис. 1. Твиты по «настоящим» темам «инвестиции» (*а*) и «биткойн» (*б*)

Примечание: по оси ординат отложено число ретвитов для того или иного твита.

На рис. 1, *а* представлены твиты, соответствующие «настоящим» темам «инвестиции» и «биткойн» (рис. 1, *б*) за 70 суток с 22 октября 2013 по 1 января 2014 г.; по оси ординат отложено число ретвитов, соответствующих тому или иному твиту. Сходимость алгоритма достигалась в среднем за 1000 итераций (рис. 2). В первом случае (рис. 3) средняя ошибка при прогнозировании на 1 шаг вперёд составила 5,5%, во втором – 25,4 %. Высокое значение ошибки прогноза во втором случае объясняется тем, что, несмотря на популярность темы «биткойн» (начиная с конца 2012 г.), алгоритм не смог отыскать достаточное число вариантов поведения, и тема «инвестиции» фигурирует в «Twitter» гораздо дольше и соответственно даёт лучшие результаты прогнозирования.

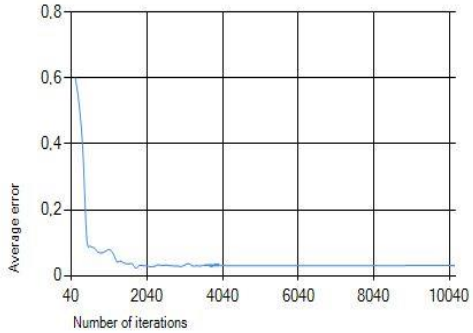
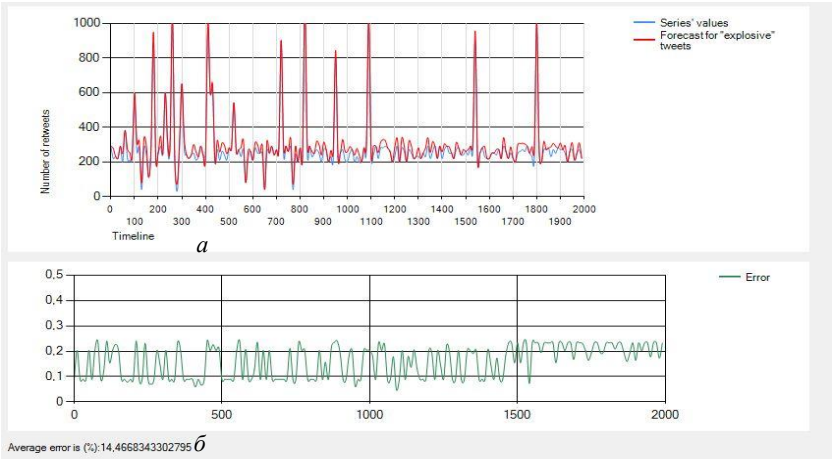


Рис. 2. Характер сходимости алгоритма: зависимость средней ошибки прогнозирования



а

б

Рис. 3. Прогнозирование взрывных тем:

а – реальное (обозначено тонкой линией) и прогнозное (обозначено жирной линией) число ретвитов; *б* – значение относительной погрешности

В таблице представлены значения ошибок 1-го и 2-го рода для статистической гипотезы «наблюдается предвзрывное поведение» для семи наиболее популярных тем (здесь ошибка 1-го рода, когда «предвзрывное» поведение игнорируется и алгоритм не прогнозирует «взрыв» популярности темы, ошибка 2-го рода, когда «невзрывное» поведение, которое не разовьётся во взрывное, идентифицируется алгоритмом как «предвзрывное»).

Ошибки прогнозирования взрывного роста популярности для наиболее популярных тем

Популярность темы	Тема	Ошибка 1-го рода	Ошибка 2-го рода
1	Нефть	14.46	19.23
2	Биткойн	25.36	42.12
3	Инвестиции	18.41	20.07
4	Кризис	17.59	31.24
5	Валюта	26.87	40.56
6	Евро	21.84	37.73
7	Долг	19.63	39.65

Выводы. Вышеприведенный анализ позволил сформулировать следующие выводы.

Необходимым элементом прогноза динамики популярности той или иной темы является классификации всего массива сообщений социальной сети с выделением «настоящих» (а не указанных в хеш-тегах) тем.

Применение методов парадигмы предсказания на основе кластеризации к совокупности временных рядов, дающих число ретвитов по той или иной теме, позволяет выявить характерные варианты динамики, предшествующие «взрыву» популярности темы.

В ходе широкомасштабного вычислительного эксперимента были выявлены характерные варианты «предвзрывной» динамики популярности тем; часть из них оказались эквивалентными эмпирическим приёмам предсказания роста популярности темы, известных специалистам по размещению рекламы в указанной социальной сети («краб», «бабочка Песавенто» и др.).

Библиографические ссылки

1. **Малинецкий, Г.Г.** Современные проблемы нелинейной динамики [Текст] / Г.Г. Малинецкий, А.П. Потапов. – М.: Эдиториал УРСС, 2000. – 336 с.
2. **Blockeel, H.** Top-down induction of clustering trees [Text] / H. Blockeel, L. De Raedt, J. Ramon // 15th International Conf. on Machine Learning. – 1998. – P. 55 – 63.
3. **Bravo-Marquez, F.** Meta-level sentiment models for big social data analysis [Text] / F. Bravo-Marquez, M. Mendoza, B. Poblete // Knowledge-Based Systems. – 2014. – Vol. 69. – P. 86 – 99.
4. Analysis of Time Series Data with Predictive Clustering Trees [Text] / S. Dzeroski, V. Gjorgjioski, I. Slavkov, J. Struyf // Proceedings of the 5th international conference on Knowledge discovery in inductive databases. – 2006. – P. 63 – 80.
5. **Gromov, V.A.** Predictive clustering on non-successive observations for multi-step ahead chaotic time series prediction [Text] / V.A. Gromov, E.A. Borisenko // Neural Computing & Applications. – 2015. – Vol. 74. – P. 1214 – 1226.
6. **Gromov, V.A.** Chaotic time series prediction with employment of ant colony optimization [Text] / V.A. Gromov, A.N. Shulga // Expert Systems with Applications. – 2012. – Vol. 39. – P. 8474 – 8478.
7. **Jackson, E.A.** The Lorenz System: I. The Global Structure of its Stable Manifolds [Text] / E.A. Jackson // Physica Scripta. – 1985. – Vol. 32, №5. – P. 469 – 475.
8. **Kasabov, N.K.** Evolving connectionist systems for adaptive learning and knowledge discovery: Trends and directions [Text] / N.K. Kasabov // Knowledge-Based Systems. – 2015. – Vol. 74. – P. 831 – 840.
9. **Koprinska, I.** Correlation and instance based feature selection for electricity load forecasting [Text] / I. Koprinska, M. Rana, V.G. Agelidis // Knowledge-Based Systems. – 2015. – Vol. 74. – P. 841 – 852.
10. The modeling and prediction of time series based on synergy of high-order fuzzy cognitive map and fuzzy c-means clustering [Text] / W. Lu, J. Yang, X. Liu, W. Pedrycz // Knowledge-Based Systems. – 2014. – Vol. 70. – P. 242 – 255.
11. Energy time series forecasting based on pattern sequence similarity [Text] / F. Martınez-Alvarez, A. Troncoso, J.C. Riquelme, J.M. Riquelme // IEEE Transactions on Knowledge and Data Engineering. – 2011. – Vol. 23, № 8. – P. 1230 – 1243.
12. **Perlovsky, L.** Neural networks and intellect: using model-based concepts [Text] / L. Perlovsky. – N.-Y.: Oxford University Press, 2001. – 496 p.
13. **Phu, L.** Motif-Based Method for Initialization the K-Means Clustering for Time Series Data [Text] / L. Phu, D.T. Anh // AI 2011: Advances in Artificial Intelligence. – 2011. – Vol. 7106. – P. 11 – 20
14. PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis [Text] / S. Rill, D. Reinel, J. Scheidt, R.V. Zicari // Knowledge-Based Systems. – 2014. – Vol. 69. – P. 24 – 33.
15. A Novel Evolving Clustering Algorithm with Polynomial Regression for Chaotic Time-Series Prediction [Text] / H. Widiputra, H. Kho, R. Pears, N. Kasabov // Neural Information Processing. – 2009. – Vol. 5864. – P. 114 – 121.

16. **Widiputra, H.** Multiple Time-Series Prediction through Multiple Time-Series Relationships Profiling and Clustered Recurring Trends [Text] / H. Widiputra, R. Pears, N. Kasabov // Advances in Knowledge Discovery and Data Mining. – 2011. – Vol. 6635. – P. 161 – 172.

Надійшла до редколегії 01.05.2015