УДК 004.021

ОЦЕНКА ПОПУЛЯРНОСТИ АВТОРОВ СОЦИАЛЬНОЙ СЕТИ С ПОМОЩЬЮ ПОИСКА ЭКСПЕРТОВ НА ПРИМЕРЕ СЕРВИСА TWITTER

Р.М. Миниахметов, Е.О. Цацина

В данной работе рассмотрена смешанная языковая модель, применяемая для поиска экспертов в таких областях как анализ социальных сетей и информационный поиск, предложена адаптация этой модели для социальной сети Twitter. Рассмотрены метрики популярности в социальной сети Twitter. Предложена формула оценки популярности пользователей социальной сети Twitter с учетом оценки релевантности их сообщений заданной теме, а также описан прототип системы для сбора данных и оценки популярности по предложенной формуле.

Ключевые слова: анализ социальных сетей, информационный поиск, интеллектуальный анализ данных, поиск экспертов, анализ популярности.

Введение

В настоящее время одной из актуальных задач в области анализа социальных сетей и информационного поиска является задача поиска экспертов. Задача поиска экспертов заключается в нахождении пользователей, сообщения которых более релевантны заданной теме [1]. Большинство подходов, применяемых в задаче поиска экспертов при анализе социальных сетей, не учитывают историю сообщений пользователя и основываются только на анализе профиля пользователя или его связей в социальном графе [1, 2].

На основе поиска экспертов в данной работе предложена формула оценки популярности пользователей социальной сети Twitter. Предложенная формула, в отличие от существующих сервисов для оценки популярности, учитывает не только базовые метрики популярности социальной сети Twitter, но и содержание (контекст) сообщений пользователей.

Рассмотренный в работе подход — использование экспертного балла автора, будет более интересен и полезен для пользователя, который захочет подписаться на сообщения автора интересующих его тем, а не на автора, который просто имеет много подписчиков.

Статья организована следующим образом. В разделе 1 рассмотрены смешанная языковая модель для поиска экспертов и метрики популярности для социальной сети Twitter. В разделе 2 предложены адаптация смешанной языковой модели, применяемой в задаче поиска экспертов, для социальной сети Twitter и формула для оценки популярности в этой сети, описан прототип системы для сбора необходимых данных и вычисления оценки популярности по предложенной формуле. В разделе 3 представлены результаты вычислительных экспериментов. В заключении суммируются основные результаты работы и описываются направления дальнейших исследований.

1. Анализ предметной области

1.1. Смешанная модель для поиска экспертов

В информационном поиске выделяют языковые модели (query likelihood language model) [5], которые описывают зависимость между текстом документа и поисковым за-

просом как вероятность появления в документе d термов t из запроса q:

$$P(d \mid q) = P(q \mid d)P(d) \tag{1}$$

Предполагается, что появление термов t_i в запросе q — независимое событие:

$$P(q \mid d) = \prod_{t \in q} P(t_i \mid d) \tag{2}$$

В данной работе для нахождения экспертов была выбрана смешанная языковая модель [3], которая определена следующим образом.

Пусть между запросом q и документом d_j существует семантический слой в виде набора тем $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$. Каждая тема θ_m семантически связана как с запросами, так и с документами. Также, каждый документ и запрос связаны с множеством тем. В этом случае связь между запросом и документами моделируется не напрямую, а через связь с темами. Таким образом, получим:

$$P(q \mid d_j) = \sum_{m=1}^{k} P(q \mid \theta_m)P(\theta_m \mid d_j), \tag{3}$$

где $P(q \mid \theta_m)$ — вероятность того, что запрос q относится к теме θ_m, а $P(\theta_m \mid d_j)$ — вероятность того, что тема θ_m относится к документу d_j.

Подсчитать с какой вероятностью запрос q относится к автору e можно при помощи следующей формулы:

$$P(q \mid e) = \sum_{d_j \in D_e} P(q \mid d_j)P(d_j \mid e) = \sum_{d_j = D_e} \sum_{m=1}^{k} \prod_{t_i \in q} P(t_i \mid \theta_m)P(\theta_m \mid d_j)P(d_j \mid e) \tag{4}$$

Смешанная языковая модель для определения экспертов используется дополнительный параметр — тему. Данная модель не связывает термы из запроса и сообщения автора напрямую, тем самым не предполагает появление всех термов из запроса в сообщениях авторов. Это является ключевым фактором выбора данной модели для адаптации и применения на данных социальной сети Twitter, так как максимальный размер сообщения в социальной сети Twitter составляет всего 140 символов и вероятность появления всех термов из запроса в таком коротком сообщении невелика.

Данная модель позволяет выделить темы из множества сообщений авторов и связать запрос с более релевантной темой, что помогает пользователю выбирать наиболее интересных авторов социальной сети.

1.2. Метрики популярности в социальной сети Twitter

Социальная сеть Twitter — это сервис для публичного обмена короткимим сообщения. Отношения между пользователями в Twitter несколько отличаются от других сетей, часто однонаправлены и базируются на совместных интересах. Пользователь может стать подписчиком определенного автора, и публичные сообщения этого автора будут отображаться в новостной ленте пользователя. Благодаря этой функции многие известные и популярные люди имеют аккаунт в этой социальной сети.

Одной из практически полезных задач в анализе социальных сетей является оценка популярности пользователей [7, 8]. Такая оценка помогает рекомендательным сервисам точнее

подбирать содержание сообщений и дает пользователям простой и быстрый способ для поиска нужной информации.

Для оценки популярности в сети Twitter используются различные метрики, основанные на следующих базовых параметрах самой сети Twitter:

1. Количество подписчиков пользователя (followers).
2. Количество повторных размещений сообщений (retweet) пользователя другими пользователями.
3. Количество отметок «понравилось» (favourites, likes), полученных от других пользователей у добавленного автором сообщения.

Проведенный анализ рейтингов популярности в сети Twitter показал, что большинство рейтингов, подобных Twitaholic [15], используют только базовые метрики сети Twitter или их комбинацию, но не учитывают информацию, содержащуюся в сообщениях пользователя.

Другой метрикой (не базовой), которую авторы работы [8] применили для анализа сети Twitter, является определение связей между авторами сети и их подписчиками. Связь рассматривается как наличие перекрестных подписок между авторами сети в рамках одной темы. Такая метрика используется при анализе многих социальных сетей, однако, не учитывает насколько автор связан с темами его сообщений. Степень связи автора и тем его сообщений может быть более интересна для тех пользователей сервиса, которые хотят получать больше информации по определенной теме.

В данной работе вводится такая метрика как экспертный балл, которая показывает, насколько автор является экспертом по интересующей пользователя теме.

2. Оценка популярности авторов в социальной сети Twitter на основе поиска экспертов

В данном разделе приводится описание смешанной модели, применяемой для поиска экспертов, на основе работы [17] и адаптация данной модели для социальной сети Twitter.

2.1. Определение тем сообщений автора

Пусть d_i — сообщение автора e в наборе сообщений D. Каждое сообщение d_i представляет набор термов (ключевых слов) $T_i = \{t_{i1}, t_{i2}, ..., t_{ik}\}$. Будем считать, что термы встречаются в сообщениях со схожим содержанием, если они часто появляются в наборах термов T_i. Определим тему θ для сети Twitter как набор независимых термов, которые часто встречаются в сообщениях одного автора.

Для нахождения часто встречающихся наборов термов использовался алгоритм FP-Growth [18].

Темы сообщений пользователя определяются на основе найденных шаблонов L_i. Для определения темы используется алгоритм кластеризации графов на основе клик:

1. Каждый шаблон $L_i = \{t'_{i1}, ..., t'_{ik}\}$ представляет собой вершину v_i графа G так, что $v_i = L_i$.
2. Если шаблоны имеют общие термы, то будем считать, что между вершинами графа, представляющего данные наборы термов, есть ребро.
3. Найдем все максимальные полные подграфы (клики) в графе G.
4. Для каждого подграфа g_i в графе G, необходимо объединить шаблоны L_i соответствующие каждой вершине подграфа.

Рассмотрим следующий пример. Пусть имеется подграф $g_i = \{v_1, v_3\}$, где $v_1 = L_1 = \{show, night, music\}$, а $v_3 = L_3 = \{new, dance, show\}$. Тогда темой будет объединение соответсвующих вершинам шаблонов L_1 и L_3, такое что $\theta_m = L_1 \cup L_3 = \{show, night, music\} \cup \{new, dance, show\} = \{show, night, music, new, dance\}$.

2.2. Определение авторов, сообщения которых более релевантны заданным темам

Пусть e — автор сообщений в Twitter и $q = \{t_{i1}, t_{i2}, ..., t_{ik}\}$ — запрос по теме, представленный набором термов и $d_i = \{t_{i1}, t_{i2}, ..., t_{ik}\}$ — сообщение автора. К первоначальному набору термов в запросе q будут добавлены термы из тех тем авторов, которые определяются как релевантные запросу [8]. Пусть $expert\_score(e, q)$ — экспертный балл автора e по теме q, который отражает насколько релевантны сообщения автора заданной теме. Тогда формула нахождения экспертов определена следующим образом:

$$expert\_score(e, q) = P(e \mid q) = P(e) \sum_{\theta_m \in \Theta} \prod_{t_i \in q} P(t_i \mid \theta_m) P(\theta_m \mid e), \tag{5}$$

где $P(e)$ определяет вероятность появления термов из запроса q в сообщениях пользователя e.

Подсчет вероятностей $P(t_i \mid \theta_m)$, $P(\theta_m \mid e)$ производится с помощью модели скрытого распределения Дирихле (Latent Dirichlet Allocation) [19]:

$$P(e) = \frac{\sum_{t_i \in q, d_i \in D_e} f(t_i, d_i)}{\mid D_e \mid}, \; f(t_i, d_i) = \begin{cases} 0, & \text{если } t_i \cap d_i = 0 \\ 1, & \text{в противном случае,} \end{cases} \tag{6}$$

$$P(t_i \mid \theta_m) = \frac{count(t_i^{(\theta_m)}) + \beta}{count(t^{(\theta_m)}) + |T|\beta}, \quad P(\theta_m \mid e) = \frac{count(\theta_m^{(e)}) + \alpha}{count(\theta^{(e)}) + |\Theta|\alpha}, \tag{7}$$

где $count(t_i^{(\theta_m)})$ — частота появления терма $t_i \in \theta_m$ во всех сообщениях авторов, $count(t^{(\theta_m)})$ — частота появления всех термов из темы $\theta_m = \{t_1, ..., t_n\}$, $|T|$ — общее количество термов (слов) во всех сообщениях автора, $count(\theta_m^{(e)})$ — частота появления термов из темы θ_m во всем массиве документов автора D_e, $count(\theta^{(e)})$ — частота появления термов из всего набора тем $\Theta = \{\theta_1, ..., \theta_n\}$ во всем массиве документов автора D_e, $|\Theta|$ — общее количество тем авторов. Параметры α и β — априорные вероятности распределения тем относительно документов и термов относительно тем соответственно.

2.3. Формула популярности авторов

В результате анализа метрик была выявлена сильная корреляция между повторными размещениями сообщений автора (retweets) и метками «понравилось» у сообщений автора (favourites, likes), поэтому они объединены в одну метрику rt_likes.

Для получения формулы оценки популярности был проведен линейный регрессионный анализ базовых метрик сети Twitter и экспертного балла. Будем считать, что чем больше значения базовых метрик и экспертного балла, тем выше соответствующая оценка популярности. Тогда предполагаемое значение популярности для проведения множественной регрессии от трех независимых параметров определяется как модуль радиус-вектора в пространстве метрик $expert\_score$, rt_likes_score, $followers\_score$.

Таким образом, формула популярности с учетом экспертного балла имеет следующий вид:

$$popularity(e,q) = r_1 \cdot followers\_score(e) + r_2 \cdot rt\_likes\_score(e) + r_3 \cdot expert\_score'(e,q), \quad (8)$$

где r_1, r_2, r_3 — коэффициенты множественной регрессии.

Относительное количество подписчиков автора считается по формуле:

$$followers\_score(e) = \frac{followers\_num(e)}{\sum\limits_{e} followers\_num(e)} \quad (9)$$

Относительное количество повторных размещений сообщений и меток «понравилось» у сообщений автора, которые соответствуют теме запроса, определяется следующим образом:

$$rt\_likes\_score(e) = \frac{\sum_{d_i \in D_e} rt\_num(d_i) + likes\_num(d_i)}{\sum\limits_{e} \sum_{d_i \in D_e} rt\_num(d_i) + likes\_num(d_i)}, d_i \cup q \neq 0 \quad (10)$$

Нормированное значение экспертного балла вычисляется по следующей формуле:

$$expert\_score'(e,q) = \frac{expert\_score(e)}{\sum\limits_{e} expert\_score(e)} \quad (11)$$

2.4. Прототип системы расчета оценки популярности

Для вычисления оценки популярности авторов в социальной сети Twitter на основе поиска экспертов был разработан прототип программной системы. На рис. 1 изображен технологический цикл разработанного прототипа. Каждый этап работы алгоритма выполняется отдельной подсистемой, реализованных в виде утилит на языке Python.

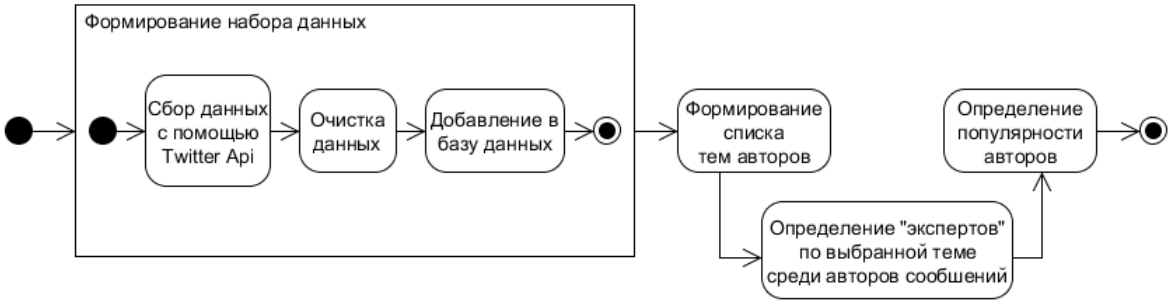

Рис. 1. Технологический цикл работы прототипа системы

Сбор данных реализован с помощью функций библиотеки tweepy [9] для языка Python, которая позволяет взаимодействовать с Twitter REST API [10]. С помощью REST API социальная сеть Twitter предоставляет доступ ко всем сущностям сети. Для сбора данных были использовныы следующие функции REST API:

1. $get\_user(id|screen\_name)$ возвращает всю доступную информацию (например, географическое положение, число подписчиков) о пользователе по указанному идентификатору или уникальному имени пользователя. Для получения информации об авторах был составлен список имен наиболее популярных авторов сети Twitter согласно рейтингу популярности Twitaholic [15].
2. $user\_timeline(id)$ возвращает список сообщений автора по его идентификатору.

Очистка полученных данных представляет собой следующие действия:

1. Отбрасывание сообщений, которые содержат символы не из таблицы символов ASCII, так как анализируются только англоязычные сообщения.
2. Отбрасывание авторов, которые имееют большой процент повторяющихся (одинаковых) сообщений.

Расчет коэффициентов в формуле популярности 8 производится на обучающей выборке, которая представляет собой информации об авторах, их сообщениях и темах. Для хранения обучающей выборки была использована свободная СУБД PostgreSQL [11]. Схема базы данных изображена на рис. 2.

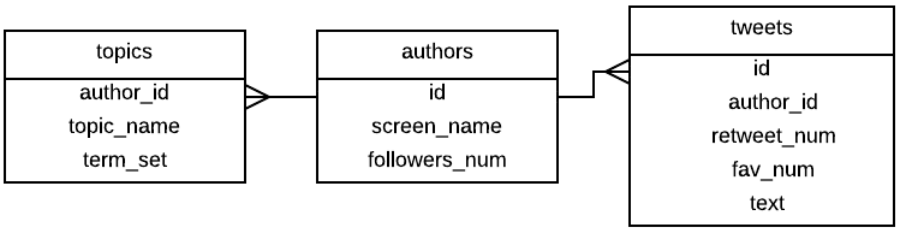

Рис. 2. Схема базы данных

Для формирования списка тем авторов используется алгоритм, описанный в подразделе 2.1. Определенные темы добавляются в базу данных и используются для расчета экспертного балла.

Определение авторов, сообщения которых релевантны заданной теме, выполняется по алгоритму из подраздела 2.2. Экспертный балл для каждого автора вычисляется по формуле 5. Результатом работы данной подсистемы является файл формата CSV, который содержит идентификатор автора и соответствующее ему значение экспертного балла.

Расчет оценки популярности авторов ведется по формуле 8. Для построения линейной регрессионной модели от трех переменных использовалась функция $LinearRegression()$, реализующая метод наименьших квадратов для построения регресии, из библиотеки scikit-learn [16] для языка Python. Результатом работы данной подсистемы является файл формата CSV, который содержит идентификатор автора и соответствующую ему оценку популярности.

Главным недостатком реализованного прототипа является достаточно продолжительное время работы на обучающей выборке, которая представляет собой большой набор данных. Возможны две основные причины этого недостатка:

1. Сбор и хранение больших объемов данных — запрос и передача данных по сети через REST API.
2. Расчет по смешанной языковой модели и модели линейной регрессии на больших объемах данных.

Среди возможных эффективных путей устранения указанного недостатка прототипа могут быть следующие:

1. Реализация онлайн-алгоритма [12] поиска экспертов и расчета оценки популярности, который не будет требовать хранения больших объемов данных.
2. Использование СУБД, которые способны эффективно обрабатывать и хранить большие объемы данных [14].
3. Оптимизация сбора обучающей выборки на основе репрезентативного сэмплинга [13].

3. Эксперименты

В этом разделе приведена информация о собранных данных для разработанной системы. Приведен результат работы примененной модели поиска экспертов и расчет популярности авторов социальной сети Twitter на основе предложенной формулы популярности. Проведен анализ полученных результатов.

3.1. Собранные данные

Для проверки работы системы было собрано около 12000 сообщений из социальной сети Twitter. Список авторов был составлен на основе рейтинга популярности в социальной сети Twitter [15], который, как и многие подобные рейтинги, строится по количеству подписчиков у автора. Для анализа работы системы были выбраны первые 100 авторов, которые имеют максимальное число подписчиков среди всех пользователей сети Twitter. Информация о собранных данных приведена в табл. 1.

Таблица 1

Информация о собранных данных

	Обучающая выборка	Тестовая выборка
Количество авторов	100	10
Количество сообщений автора	100	200
Всего сообщений	10000	2000

Обучающая выборка использовалась для проведения регрессионного анализа между значением популярности (*popularity*) и относительными значениями метрик сети Twitter ($followers\_score$, rt_likes_score), а также экспертным баллом ($expert\_score$).

Для проверки адекватности построенной модели были посчитаны средняя ошибка аппроксимации и коэффициент детерминации R^2. Средняя ошибка аппроксимации представляет среднее отклонение расчетных значений от фактических и составляет $0,32\%$, что меньше допустимых 15%. Данное значение свидетельствует об относительной точности построенной модели и о том, что ее можно применять для расчетов. Коэффициент детерминации R^2, характеризующий качество построенной модели, близок к 1 — составляет $0,9999855$ и означает, что модель достаточно точно объясняет поведение прогнозируемого параметра (популярности).

Построенная линейная регрессионная модель позволила оценить зависимость между перечисленными величинами и определить коэффициенты, которые используются при подсчете оценки популярности. На рис. 3–5 изображены графики полученных зависимостей.

График, изображенный на рис. 3 показывает, что зависимость популярности автора от количества подписчиков автора имеет линейный характер. График зависимости популярности автора от значения метрик понравившихся и пересланных сообщений сети Twitter изображен на рис. 4 и представляет собой логарифмическую функцию.

График зависимости между экспертным баллом и популярностью хоть и напоминает отдаленно логарифмическую функцию (рис. 5), имеет большое количество значений, которые плохо описываются этой функцией. Возможными причинами плохого объяснения зависимости популярности от экспертного балла могут являться как неподходящая модель линейной

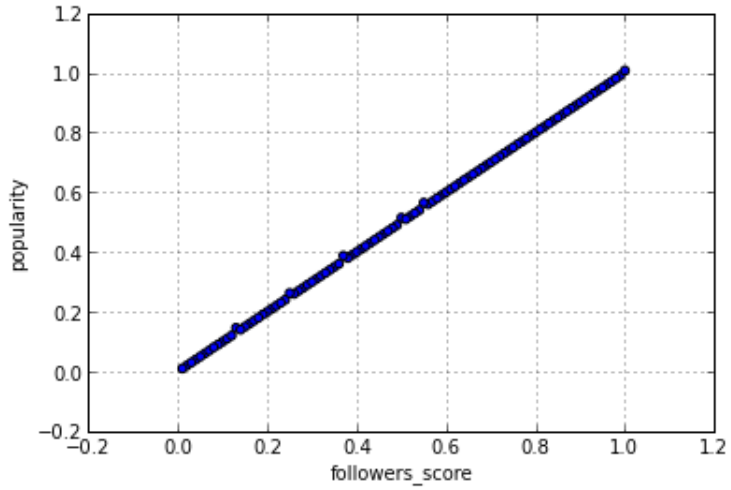

Рис. 3. Зависимость популярности автора от количества подписчиков автора

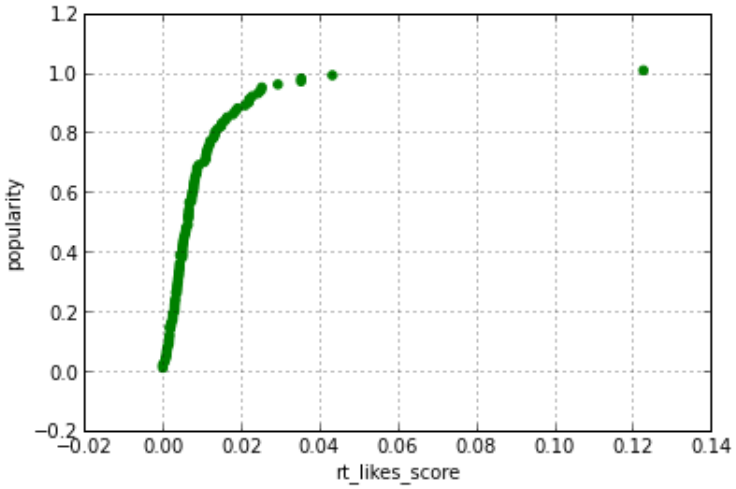

Рис. 4. Зависимость популярности автора от количества повторных размещений сообщений и меток «понравилось» у сообщений автора, соответствующих теме запроса

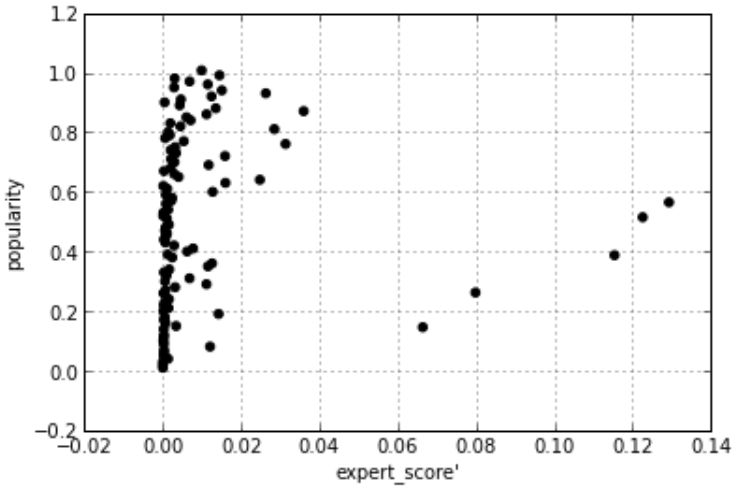

Рис. 5. Зависимость популярности автора от экспертного балла

регрессии, так и недостатки вычисления самой метрики экспертного балла. В то же вре-

мя линейная регрессия хорошо подходит для объяснения зависимости между количеством подписчиков автора и его популярностью. Кроме того, на представленных графиках видно, что значения экспертного балла количественно существенно отличаются от значений числа подписчиков. Таким образом, построенная функция линейной регресии, предсказывающая значение популярности автора по этим трем параметрам, имеет сильную зависимость от количества подписчиков.

3.2. Результаты экспериментов

Далее приведем результаты работы системы. В качестве запроса был выбран один из найденных наборов термов, представляющий определенную тему $q^* =$ *'tonight play last show'*. Алгоритм нахождения тем авторов описан в подразделе 2.1. Результат применения предложенной адаптации смешанной модели поиска экспертов для авторов социальной сети Twitter приведен в табл. 2.

Таблица 2

Позиции авторов относительно релевантности их сообщений запросу q^*

№ п/п	Имя автора	Экспертный балл	Позиция в рейтинге Twitaholic
1	@taylorswift13	0,009396866984837462	5
2	@katyperry	0,007293893139221034	1
3	@TheEllenShow	0,003811523205488189	9
4	@JLo	0,0029136778336937533	10
5	@jtimberlake	0,0027675981458890678	8
6	@ladygaga	0,002531642429819456	4
7	@rihanna	0,0016395838934267649	7
8	@justinbieber	0,0015584543674876422	2
9	@britneyspears	0,0015263600436778333	6
10	@BarackObama	0,00013284391918131865	3

Результаты, приведенные в табл. 2, показывают, что авторы, количество сообщений которых в большей степени соответствовали заданному запросу по теме, занимают первые позиции, несмотря на то, что в выбранном рейтинге популярности занимают последние позиции. Например, автор @JLo переместился с 10 позиции на 4, а @TheEllenShow с 9 на 3, и наоборот, автор @BarakObama опустился с 3 позиции на 10, так как не является представителем массовой культуры и его сообщения в наименьшей степени соответствуют запросу.

В табл. 3 приведены результаты оценки популярности авторов с помощью предложенной в подразделе 8 формулы.

Из табл. 3 видно, что формула оценки популярности зависит как от экспертного балла, так и от метрик популярности, в частности, сильное влияние на распределение авторов оказывает количество подписчиков. Авторы с более высокой позицией в рейтинге популярности Twitaholic оказываются в верхней части полученного списка популярности, однако, автор @taylorswift13, сообщения которого являются более релевантными теме запроса q^* согласно экспертному баллу, также сохранил свою позицию вверху списка популярности.

Таблица 3

Оценка популярности авторов, с учетом экспертного балла по теме q^* и метрик популярности сети Twitter

№ п/п	Имя автора	popularity	Место в рейтинге Twitaholic
1	@katyperry	0,17349268242859633	1
2	@taylorswift13	0,16173712417285024	5
3	@justinbieber	0,16142120932868534	2
4	@ladygaga	0,12378441606619985	4
5	@BarackObama	0,11051037361616957	3
6	@britneyspears	0,10208204786496214	6
7	@rihanna	0,09785773438412347	7
8	@jtimberlake	0,09483032431326485	8
9	@TheEllenShow	0,09016066084936912	9
10	@JLo	0,08371571502915753	10

Таким образом, экспертный балл, даже с учетом большой зависимости популярности от количества подписчиков, оказывает влияние на оценку популярности.

Заключение

В работе представлена смешанная языковая модель для поиска экспертов, адаптированная для социальной сети Twitter, а также предложена формула оценки популярности авторов в социальной сети Twitter с учетом релевантности их сообщений заданной теме. Данная формула оценки популярности авторов учитывает как базовые метрики самой сети Twitter, так и контекст сообщений авторов. Реализован прототип системы, которая сначала ранжирует авторов согласно релевантности их сообщений заданной теме, и на основе полученной экспертной оценки (экспертный балл), производит оценку популярности.

Проведенные вычислительные эксперименты показывают, что предложенная формула оценки популярности имеет сильную зависимость от количества подписчиков автора, однако экспертный балл также влияет на значение популярности автора. Таким образом, авторы, чьи сообщения в большей степени соответствуют теме запроса, могут находится вверху рейтинга популярности, несмотря на то, что они, возможно, менее популярны в соответствии с рейтингом. Предложенная формула оценки популярности помогает найти авторов, которые пишут на определенные темы и также являются достаточно популярными пользователями сети. Формула оценки популярности может быть использована в рекомендательных сервисах для сети Twitter.

На основе реализованной модели поиска экспертов исследования могут быть продолжены по следующим основным направлениям:

1. Поиск экспертов в режиме реальном времени (онлайн-обработка). Например, для рекомендательного сервиса.
2. Полученный экспертный балл может быть использован в качестве базовой оценки в алгоритме поиска экспертов, учитывающем связи авторов [1].

Также исследования могут быть продолжены в направлении улучшения предложенной регрессионной модели для оценки значения популярности: поиск новой модели, которая

будет точнее описывать зависимость экспертного балла автора и его популярности, что позволит увеличить корреляцию значения популярности и сообщений автора по заданной теме и уравновесить по значимости используемые метрики.

Полученные результаты могут быть использованы при поиске ответов на следующие вопросы:

1. Что представляет собой популярность в социальных сетях и как ее оценивать?
2. Какие метрики социальных сетей позволяют точнее и корректнее оценить популярность пользователей?
3. Какие агрегированные значения базовых метрик могут быть использованы при определении популярности?
4. Является ли полученная оценка популярности оценкой реальных пользователей сети или результатом различных способов искусственного увеличения показателей базовых метрик популярности сети?

Литература

1. Zhang, J. Expert Finding in A Social Network / J. Zhang, J. Tang, J. Li // Lecture Notes in Computer Science. — 2007. — Vol. 4443. — P. 1066–1069.

2. Bozzon, A. Choosing the Right Crowd: Expert Finding in Social Networks / A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, G. Vesci // Proceedings of the 16th International Conference on Extending Database Technology (EDBT '13). — 2013. — P. 637–648.

3. Zhang, J. A Mixture Model for Expert Finding / J. Zhang, J. Tang, L. Liu, J. Li // Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '08). — 2008. — P. 466–478.

4. Manning, C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schütze — Cambridge University Press, 2008. — 496 p.

5. Ponte, J. A Language Modeling Approach to Information Retrieval / J. Ponte, W. Croft // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98). —1998. — P. 275–281.

6. Hiemstra, D. A Linguistically Motivated Probabilistic Model of Information Retrieval / D. Heimstra // Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries.— 1998. — P. 569–584.

7. Steck, H. Item Popularity and Recommendation Accuracy / H. Steck // Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (RecSys '11). — 2011. — P. 125–132.

8. Cha, Y. Incorporating Popularity in Topic Models for Social Network Analysis / Y. Cha, B. Bin, C. Hsieh, J. Cho // Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13). — 2013. — P. 223–232.

9. Документация Python библиотеки для работы с Twitter API
URL: `http://pythonhosted.org//tweepy/` (дата обращения: 24.04.2014).

10. Документация Twitter REST API
URL: `https://dev.twitter.com/docs/api/1.1` (дата обращения: 24.04.2014).

11. Stonebraker, M. The POSTGRES Next-generation Database Management System / M. Stonebraker, G. Kemnitz // Communications of the ACM. — 1991. — Vol. 34, No. 10. — P. 78–92.

12. Borodin, A. Online Computation and Competitive Analysis / A. Borodin, R. El-Yaniv — Cambridge University Press, 1998. — 432 p.

13. Янцен, Д.Д. Алгоритм репрезентативного сэмплинга для параллельных систем баз данных / Д.Д. Янцен, М.Л. Цымблер // Параллельные вычислительные технологии (ПаВТ'2014): труды международной научной конференции (1–3 апреля 2014 г., г. Ростов-на-Дону). — Челябинск: Издательский центр ЮУрГУ, 2014. — С. 381.

14. Пан, К.С. Подход к разбиению сверхбольших графов с помощью параллельных СУБД / К.С. Пан // Вестник ЮУрГУ. Серия "Вычислительная математика и информатика". — 2012. — № 47(306). Вып. 2. — С. 127–132.

15. Рейтинг популярности пользователей в социальной сети Twitter «Twitaholic» URL: `http://twitaholic.com/` (дата обращения: 24.04.2014).

16. Документация Python библиотеки для интеллектуального анализа данных «scikit-learn» URL: `http://scikit-learn.org/stable/documentation.html` (дата обращения: 24.04.2014).

17. Zhu, H. Finding Experts in Tag Based Knowledge Sharing Communities / H. Zhu, E. Chen, H. Cao // Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management (KSEM '11). — 2011. — P. 183–195.

18. Han, J. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach / J. Han, J. Pei, Y. Yin, R. Mao // Data Mining and Knowledge Discovery. — 2005. — Vol. 8, No. 1. — P. 53–87.

19. Blei, D. Latent Dirichlet Allocation / D. Blei, A. Ng, Y. Yin, M. Jordan // The Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.

Миниахметов Руслан Марсович, аспирант, кафедра системного программирования, Южно-Уральский государственный университет (г. Челябинск, Российская Федерация), miniakhmetovrm@susu.ac.ru.

Цацина Елизавета Олеговна, студент 4 курса, кафедра системного программирования, Южно-Уральский государственный университет (г. Челябинск, Российская Федерация), decemberliz92@gmail.com.

Поступила в редакцию 9 мая 2014 г.

Bulletin of the South Ural State University
Series "Computational Mathematics and Software Engineering"
2014, vol. 3, no. 2, pp. 63–76

TWITTER USERS POPULARITY ESTIMATION USING EXPERT FINDING

R.M. Miniakhmetov, South Ural State University (Chelyabinsk, Russian Federation)**,**
E.O. Tsatsina, South Ural State University (Chelyabinsk, Russian Federation)

In this paper we have considered mixed language model that is used for experts finding in areas such as social network analysis and information retrieval, and proposed an adaptation of this model for the social network Twitter. We also have reviewed Twitter popularity metrics and proposed Twitter users' popularity estimation approach based on expert finding, which allows to rank users according to the probability of user being an expert in given query, and have implemented a prototype for data collection and popularity estimation, based on our approach.

Keywords: social network analysis, information retrieval, data mining, expert finding, popularity analysis.

References

1. Zhang J., Tang J., Li J. Expert Finding in A Social Network. Lecture Notes in Computer Science. 2007. Vol. 4443, No. 3. P. 1066–1069.
2. Bozzon A., Brambilla M., Ceri S., Silvestri M., Vesci G. Choosing the Right Crowd: Expert Finding in Social Networks. Proceedings of the 16th International Conference on Extending Database Technology (EDBT '13). 2013. P. 637–648.
3. Zhang J., Tang J., Liu L., Li J. A Mixture Model for Expert Finding. Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '08). 2008. P. 466–478.
4. Manning C., Raghavan P., Schütze P. Introduction to Information Retrieval. Cambridge University Press, 2008. 496 p.
5. Ponte, J., Croft J. A Language Modeling Approach to Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98). 1998. P. 275-281.
6. Hiemstra, D. A Linguistically Motivated Probabilistic Model of Information Retrieval. Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries. 1998. P. 569–584.
7. Steck H. Item Popularity and Recommendation Accuracy. Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11). 2011. P. 125–132.
8. Cha Y., Bin B., Cho J. Incorporating Popularity in Topic Models for Social Network Analysis. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13). 2013. P. 223–232.
9. Twitter API library for Python documentation
URL: `http://pythonhosted.org//tweepy/` (accessed: 24.04.2014).

10. Twitter REST API Documentation
URL: `https://dev.twitter.com/docs/api/1.1` (accessed: 24.04.2014).

11. Stonebraker M., Kemnitz G. The POSTGRES Next-generation Database Management System // Communications of the ACM. 1991. Vol. 34, No. 10. P. 78–92.

12. Borodin A., El-Yaniv R. Online Computation and Competitive Analysis. Cambridge University Press, 1998. 432 p.

13. Yantsen D.D., Zymbler M.L. Algoritm reprezentativnogo sempling dlya parallel'nykh sistem baz dannykh [A Representative Sampling Method for Parallel DBMS] // Parallel'nyye vychislitel'nyye tekhnologii (PaVT'2014): trudy mezhdunarodnoy nauchnoy konferentsii (1-3 aprelya 2014 g., g. Rostov-na-Donu) [Parallel Computational Technologies (PCT'2014): Proceedings of the International Scientific Conference (April 1-3, 2014, Rostov-on-Don)]. Chelyabinsk, Publishing of the South Ural State University, 2014. P. 381.

14. Pan C.S. Podkhod k razbiyeniyu sverkhbol'shikh grafov s pomoshch'yu parallel'nykh SUBD [An Approach for Very Large Graph Partitioning by Means of Parallel DBMS] // Vestnik YuUrGu. Seriya "Vychislitel'naya matematika i informatika" [Bulletin of the SUSU. Series «Computational Mathematics and Software Engineering»]. 2012. No 47(306). Iss. 2. P. 127–132.

15. Tracking the most popular users of a certain microblogging/social network tool «Twitaholic».
URL: `http://twitaholic.com/` (accessed: 24.04.2014).

16. Machine learning library for Python documentation
URL: `http://scikit-learn.org/stable/documentation.html` (accessed: 24.04.2014).

17. Zhu H., Chen E., Cao H. Finding Experts in Tag Based Knowledge Sharing Communities. Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management (KSEM '11). 2011. P. 183–195.

18. Han J., Pei J., Yin Y., Mao R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery. 2005. Vol. 8, No. 1. P. 53–87.

19. Blei D., Ng A., Yin Y., Jordan M. Latent Dirichlet Allocation. The Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022.

Received 9 May 2014.