МЕТОДЫ АНАЛИЗА И ОБРАБОТКИ ДАННЫХ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

Т. В. Батура, Ф. А. Мурзин, А. В. Проскуряков, Д. О. Сперанский*

Институт систем информатики им. А. П. Ершова СО РАН,
630090, Новосибирск, Россия
*Новосибирский национальный исследовательский государственный университет,
630090, Новосибирск, Россия

УДК 519.68; 681.513.7;
316.472.45; 007.51/52

Статья посвящена проблемам анализа и обработки данных, получаемых из социальных сетей. Были изучены некоторые формальные характеристики социальных сетей, введены соответствующие понятия, модели и методы, которые могут быть полезны для анализа информации, получаемой из социальных сетей. Для анализа межличностных отношений предлагается использовать так называемый анализ предпочтений. Предложены различные модификации динамической теории Латане социального влияния применительно к рассматриваемым задачам. В работе также предложено обобщение алгоритма реферирования новостных и обзорных статей с использованием Link Grammar Parser. Рассмотрена возможность применения этого алгоритма для оценки релевантности сообщений, оставляемых в социальных сетях, статьям, публикуемым в Интернете. Данный подход является полезным при решении задачи определения источника распространения информации. В статье кратко описан разработанный программный комплекс, позволяющий извлекать информацию из социальных сетей, проводить обработку, анализ и визуализацию данных.

Ключевые слова: анализ социальных сетей, обработка данных на естественном языке, теория социального влияния Латане, источник распространения информации, межличностный анализ.

This work focuses on the data processing and analysis of online social networking services. We examine several formal definitions of various characteristics (numerical and structural) and introduce appropriate concepts, models and methods that could be useful for the analysis of information obtained from social networks. Preference analysis is proposed for the study of interpersonal relations. Various modifications of Latané's dynamic theory of social impact are proposed.
The generalization of the summarization algorithm of news articles and reviews using the Link Grammar Parser proposed in article. We have considered the possibility of applying this algorithm to assess the relevance of posts to the articles published on the Internet. This approach is useful in solving the problem of determining the source of information dissemination. The paper briefly describes a software system developed by us allowing to carry out extracting, processing, analyzing and visualizing data from online social networking services.

Key words: social networks analysis, data mining, Latané's social impact theory, interpersonal relationships, source of information dissemination.

Введение. Современный мир теперь уже сложно представить без социальных сетей. Они являются удобным средством общения между людьми, кроме того, открывают новые

возможности для анализа потоков информации и поведения людей в процессе общения. Совокупный анализ структуры социальных графов и текстовых данных, получаемых из социальных сетей, по мнению многих специалистов является наиболее эффективным методом исследования взаимодействий между участниками сети [1, 2].

В статье предложены различные количественные характеристики, графы и множества, которые могут быть вычислены или построены на основе информации, полученной из социальных сетей. Для анализа межличностных отношений предлагается использовать так называемый анализ предпочтений [3, 4]. Авторами была предпринята попытка адаптировать динамическую теорию Латане социального влияния [5, 6] к социальным сетям, что позволяет вычислять уровень влияния окружающих людей на мнение конкретного человека.

Интуитивно ясно, что отношения между пользователями и лексика текстовых сообщений, создаваемых этими пользователями, связаны между собой. В четвертом разделе статьи предложено математическое описание некоторых множеств и функций, естественным образом при этом возникающих. Приведены наиболее важные характеристики рассмотренных множеств.

Пятый раздел посвящен анализу непосредственно текстовых сообщений. В Интернете существует довольно большое количество новостных и обзорных статей. Естественно предположить, что в некоторых случаях они являются источниками информации, обсуждаемой на форумах и в социальных сетях. Поэтому представляет интерес задача оценки релевантности текста статей сообщениям, оставляемым пользователями. За основу взят алгоритм реферирования статей, описанный в работе [7]. Предложено обобщение этого алгоритма, и рассмотрена возможность его применения для определения релевантности.

Для проведения экспериментов в рамках данного исследования был создан программный комплекс для извлечения, обработки и анализа пользовательских данных. В частности, в системе имеется модуль, позволяющий извлекать данные из крупнейших социальных сетей (Twitter, Facebook, „ВКонтакте“). Кроме того, этот модуль имеет возможность функционального расширения практически на любую социальную сеть в зависимости от предоставляемого API.

1. Количественные характеристики, отношения и множества, вычисляемые на основе данных, получаемых из социальных сетей. При анализе социальных сетей целесообразно рассматривать ряд числовых и нечисловых характеристик, отношений и множеств, естественным образом связанных с пользователями сети и сообщениями, циркулирующими в ней. Важным является условие, чтобы эти характеристики могли быть вычислены или построены при помощи соответствующих алгоритмов.

Обозначим T — сообщение („твит“) социальной сети, u — пользователя сети, который может создавать и пересылать сообщения.

Одноместные характеристики:

$Followers\_Count(u)$ — количество людей, которые читают сообщения данного пользователя (т. е. подписчиков этого пользователя);

$Friends\_Count\,(u)$ — количество друзей у данного пользователя (пользователь сам заносит некоторых людей в список друзей);

$Retweets\,(T)$ — количество пересылок данного сообщения.

Множества:

$Followers\,(u)$ — подписчики данного пользователя;

$Friends\,(u)$ — друзья данного пользователя;

$Mentions\,(u)$ — имена пользователей, упоминаемые в сообщениях данного пользователя;

$Hashtags\,(u)$ — хэштеги, которые встречаются в сообщениях данного пользователя;

$Urls\,(u)$ — внешние ссылки, которые встречаются в сообщениях данного пользователя.

Хэштег — это слово или набор слов, записанных без пробелов, начинающихся с символа „#“. Является одной из форм метаданных. Короткие сообщения в блогах или социальных сетях типа Facebook, Twitter или Instagram могут быть помечены символом „#“ перед ключевыми словами или фразами, не содержащими пробелов. Они встречаются в любых предложениях, например: „Ах, как жалко, что закончилась #Олимпиада2014!“ Хэштеги позволяют группировать похожие сообщения. По заданному хэштегу можно найти набор сообщений, содержащих его.

Числовые характеристики, ассоциированные с множествами:

$Count\_Mentions_u\,(v)$ — количество упоминаний пользователя v пользователем u;

$Count\_Hashtags_u\,(v)$ — количество употреблений хэштега v пользователем u;

$Count\_Urls_u\,(v)$ — количество упоминаний внешней ссылки v пользователем u;

$Count\_Retweets_u\,(u_1)$ — количество сообщений, пересланных пользователем u, полученных от пользователя u_1.

При анализе процессов, происходящих в коллективах, в исследованиях по социологии, психологии, экономике часто рассматривают трехместные отношения предпочтения [3]. Согласно [4], запись $i \underset{k}{|\!-} j$ означает, что i предпочтительнее j по мнению k. Критерии предпочтительности могут быть самыми различными: профессионализм, который можно разбить на разные виды деятельности, что породит спектр новых критериев; умение руководить людьми; коммуникабельность; восприимчивость к инновациям; психологическая устойчивость и т. д. На базе этих отношений складывается неформальная структура коллектива, важность выявления которой в ряде случаев не требует комментариев.

Трехместные отношения:

$Mentions_u\,(u_1, u_2)$ — пользователь u упоминает пользователя u_1 не реже, чем u_2;

$Hashtags_u\,(h_1, h_2)$ — пользователь u употребляет хэштег h_1 не реже, чем хэштег h_2;

$Urls_u\,(url_1, url_2)$ — пользователь u упоминает ссылку url_1 не реже, чем ссылку url_2;

$Retweets_u\,(u_1, u_2)$ — пользователь u пересылает сообщения, полученные от пользователя u_1, не меньшее число раз, чем полученные от пользователя u_2.

Числовые характеристики, ассоциированные с трехместными отношениями:

$$N\_Mentions_u\,(u_1, u_2) = Count\_Mentions_u\,(u_1) - Count\_Mentions_u\,(u_2)\,;$$

$$N\_Hashtags_u\,(h_1, h_2) = Count\_Hashtags_u\,(h_1) - Count\_Hashtags_u\,(h_2)\,;$$

$$N\_Urls_u\,(url_1, url_2) = Count\_Urls_u\,(url_1) - Count\_Urls_u\,(url_2)\,;$$

$$N\_Retweets_u\,(u_1, u_2) = Count\_Retweets_u\,(u_1) - Count\_Retweets_u\,(u_2)\,.$$

Можно считать, что приведенные выше функции также позволяют определять силы влияния различных факторов. Например, функция $N\_Mentions_u\,(u_1, u_2)$ позволяет вычислить силу влияния пользователя u_1 по сравнению с u_2 на пользователя u, т. е. силу влияния с учетом предпочтений пользователя u. Например, если пользователь u= '@navalny' упоминает 7 раз пользователя u_1='@KSHN' и 1 раз пользователя u_2='@kudriavtsev', то $N\_Mentions_u\,(u_1, u_2) = 7 - 1 = 6$. Насколько информативна функция $Retweets_u\,(u_1, u_2)$, пока не ясно.

2. Анализ отношения предпочтения. В данном разделе приведены краткие сведения из работы [4] с целью показать, каким образом проводится так называемый анализ предпочтений. Решение такого рода задачи осуществляется в два этапа.

Первый этап включает сбор информации, в результате чего формируются так называемые индивидуальные анкеты. В анкете содержится информация о парном ранжировании членов коллектива по заданному критерию. Точнее, двух членов коллектива сравнивает третий. При этом анкета ассоциируется с третьим членом коллектива. Рассматривая совокупность всех анкет, в итоге получаем трехместное отношение. С математической точки зрения индивидуальной анкете соответствует булева матрица.

На втором этапе осуществляется обработка информации. Используется некоторый алгоритм, преобразующий семейство булевых матриц (анкет) во взвешенный граф. Структура такого графа отражает в некоторой мере структуру коллектива и может анализироваться послойно, в зависимости от весов связей. Предполагается, что максимальный положительный отклик гарантируется отправителю импульса при подключении к так называемому „пути наибольшей симпатии“ или, по физической аналогии, к „пути пробоя“. Такого рода пути можно выделить в упомянутом выше графе. Важным также является выделение первого адресата импульса, т. е. через кого „входить“ в коллектив.

В процессе анализа социальных сетей возникают естественные отношения предпочтения, а именно трехместные отношения, упомянутые во втором разделе. Интересно, что, в отличие от социологов, мы сравниваем между собой не только людей, но также хэштеги и ссылки.

2.1. *Описание данных и алгоритма.* Введем следующие обозначения: n — количество участников группы (коллектива); A_k– индивидуальная анкета $(1 \leqslant k \leqslant n)$. Индивидуальная анкета представляет собой булеву (содержащую только нули и единицы) антисимметричную матрицу с нулями на главной диагонали, т. е. имеем

$$A_k = (a_{ij}^k), (1 \leqslant i, j \leqslant n) a_{ij}^k = 0, \left(i \neq j \to a_{ij}^k = \bar{a}_{ji}^k\right),$$

где черта обозначает отрицание $\bar{0} = 1$, $\bar{1} = 0$. Значению элемента $a_{ij}^k = 1$ матрицы A_k соответствует отношение $i \overline{\vdash}_k j$. Совокупность таких анкет A_k, $(1 \leqslant k \leqslant n)$ образует входное множество данных.

Каждой индивидуальной анкете A_k соответствует ориентированный граф

$$G_k = \langle G_k, I_k \rangle, G_k = \{1, \ldots, n\}, \langle i, j \rangle \in I_k \leftrightarrow i \overline{\vdash}_k j.$$

Матрица A_k является для G_k матрицей смежности.

Результирующая матрица $Q = (q_{ki})$ вводится по правилу $q_{ki} = \sum_{j=1}^{n} a_{ij}^k$, т. е. в k-й строке отражено „суммарное мнение“ k-го участника об i-м участнике группы.

Далее, мнение k-го участника о коллективе в целом можно вычислить по формуле $Opinion = \sum_{i=1}^{n} q_{ki}$.

Величину $Rating = \sum_{k=1}^{n} q_{ki}$ назовем *рейтингом* i-го участника группы. Она отражает суммарное мнение всего коллектива о данном участнике.

Результирующей матрице Q соответствует граф

$$G = \langle G, I, w \rangle, G = \{1, ..., n\}, \langle i, j \rangle \in I \leftrightarrow i \neq j.$$

Вес ребра определяется по формуле $w(i, j) = q_{ij}$. Заметим, что данный граф является полным. Каждые две вершины соединены парой ребер противоположной ориентации,

которые могут иметь различный вес. Таким образом, алгоритм построения графа, являющегося нашей целью, полностью описан.

2.2. *Анализ построенного графа.* Для послойного анализа построенного графа рассмотрим подграфы с вершинами, веса которых больше наперед заданной величины. Пусть для каждого натурального числа t

$$G^t = \left\langle G, I^t \right\rangle,\ I^t = \{\langle i,j\rangle \in I | w(i,j) = t\},\bar{G}^t = \bigcup_{s\geqslant t} G^s = \left\langle G, \bigcup_{s\geqslant t} I^s \right\rangle.$$

Очевидно, что $t_1 \leqslant t_2 \to G^{t_1} \supseteq G^{t_2}$, $\bigcup_{s\geqslant 0} G^s = G$. Граф G^t будем называть *срезом* уровня t. Обычно для формирования более весомых связей требуется большее время. Поэтому, рассматривая G^t для различных t, можно видеть динамику развития связей во времени. Хотя иногда весомые связи могут образоваться очень быстро, например, как результат приглашения высококвалифицированного специалиста со стороны.

Для того чтобы выделить „пути наибольшей симпатии“, необходимо решить некоторый вариант задачи о коммивояжере. Это не очень удобно, если граф содержит большое количество вершин. В таком случае сначала можно рассматривать некоторый срез при подходящем t. Часть вершин в нем оказывается изолированной и отбрасывается. Полученный граф дополняется ребрами до полного, а веса наследуются, т. е. рассматриваются только значимые участники коллектива, но связи между ними учитываются все.

Переходим к графу $H^t = \{i \in G | \exists j \quad (w(i,j) \geqslant t)\}$. Для того чтобы не усложнять обозначения, будем считать, что мы работаем с исходным графом.

Пусть $l = \langle i_1, \ldots i_k\rangle$ — некоторый маршрут в графе G. Весом маршрута l называется величина

$$w(l) = \sum_{j=1}^{k-1} w(i_j, i_{j+1}).$$

Напомним, что если $i_1 = i_k$, то маршрут называется замкнутым. Будем использовать также следующие обозначения: $k \in l$ — вершина k содержится в маршруте l; $l_1 \subseteq l_2$ — маршрут l_1 является частью маршрута l_2; $Ent(k,l) = \{i | \langle k,l\rangle \subseteq l\}$ — вход из вершины k в маршрут l.

Один из вариантов задачи о коммивояжере может быть сформулирован как задача поиска максимального замкнутого маршрута без самопересечений, такого, что $w(l)$ достигает максимума, т. е. данный маршрут должен проходить через все вершины. При этом через каждую вершину он должен проходить только один раз. В силу полноты графа G такой маршрут существует, но, вообще говоря, их может быть несколько.

Обозначим множество всех таких маршрутов $L(G)$. По определению
$Ent(k,G) = \{Ent(k,l) | l \in L(G)\}$ — множество входов из вершины k в граф G.

Элементы $L(G)$ называются „путями наибольшей симпатии“; $Ent(k,G)$ показывает, на кого может направлять импульсы k-й участник коллектива, чтобы подключиться к этим путям.

Отметим еще одно обстоятельство. Реально могут выявляться парадоксальные цепочки, в которых нарушается транзитивность предпочтений. Например: „a лучше b“, „b лучше c“, но „c лучше a“. Можно ввести абсолютную и относительную меры транзитивности.

Абсолютная мера определяется по формуле

$$f(k) = Card\{\langle i_1, i_2, i_3\rangle \,|i_1 \underset{k}{\mapsto} i_2, i_2 \underset{k}{\mapsto} i_3, i_1 \underset{k}{\mapsto} i_3\},$$

где $Card$ — количество элементов в множестве.

Относительная мера может быть определена, например, так: $F_k = f(k)/n^3$.

Крупномасштабное нарушение транзитивности обычно свидетельствует о неустойчивости коллектива. Мелкомасштабное же нарушение обычно всегда присутствует. Более того, в большом коллективе постоянно образуются и распадаются малозначимые связи. Переходя на соответствующий уровень, можно элиминировать их и далее провести достаточно качественный анализ.

3. Теория Латане социального влияния и ее модификация. Далее рассмотрим, каким образом можно адаптировать теорию социального влияния, предложенную Латане [5, 6], к социальным сетям. Латане подчеркивал важность трех атрибутов отношений между получателем информации и источником: сила — это статус вовлекаемых агентов; расстояние между агентами — физическое или психологическое; количество источников, влияющих на получателя.

Согласно теории социального влияния, уровень влияния, испытываемого агентом, может быть выражен следующей формулой

$$I_i = -S_i\beta - \sum_{j=1, j\neq i}^{N} \frac{S_j O_j O_i}{d_{i,j}^{\alpha}},$$

где I_i — количество социального давления, направленного на агента i; O_i — мнение i-го агента (± 1) по отношению к данному вопросу, значение $+1$ соответствует поддержке, и -1 соответствует сопротивлению предложению; S_i — сила социального влияния ($S_i \geqslant 0$); β — сопротивляемость агента к изменениям ($\beta > 0$); d_{ij} — расстояние между агентами i и j ($d_{ij} \geqslant 1$); α — степень ослабления расстояния ($\alpha \geqslant 2$); N — общее число взаимодействующих агентов.

Значение постоянной β обычно принимается равным 2 в соответствии с величиной, использованной в исследованиях Латане. Большее значение этой постоянной означает, что для изменения мнения требуется большее давление, меньшее значение соответствует меньшему усилию. Значение постоянной α также обычно принимается равным 2. Бóльшие значения α означают, что с ростом расстояния между источником и получателем требуется много бóльшая величина давления.

Величина d_{ij} определяется свойствами пары агентов, она может рассматриваться как показатель легкости общения (передачи информации). При задании данной величины могут учитываться возрастные, национальные, конфессиональные и другие различия. Формула для вычисления d_{ij} может включать в себя физическое расстояние, например, между населенными пунктами, в которых находятся агенты. Обычно учитывается факт, что легкость коммуникации подчиняется закону об обратной квадратичной зависимости от физического расстояния [6]. В случае с социальными сетями возможны различные подходы, в том числе такие, когда физическое расстояние не принимается во внимание.

Для анализа социальных сетей мы предлагаем модификацию формулы Латане в следующем виде:

$$I_u = -\beta \cdot \sum_{i=1}^{N} Count\_Mentions_u(u_i) - \sum_{i=1}^{N} \sum_{\substack{j=2 \\ i>j}}^{N} \frac{N\_Mentions_u(u_i, u_j)}{\rho^{\alpha}(u_i, u_j)}.$$

В этой формуле учитываются все пользователи, упоминаемые u. Можно считать, что все они на него влияют. В следующей формуле учитывается влияние только наиболее упоминаемого и наименее упоминаемого пользователей:

$$I_u^1 = -\beta \cdot \max_{i=1}^{N} \{Count\_Mentions_u(u_i)\} - \max_{i=1}^{N} \max_{\substack{j=2 \\ i>j}}^{N} \left\{ \frac{N\_Mentions_u(u_i, u_j)}{\rho^\alpha(u_i, u_j)} \right\}.$$

Также можно учитывать влияние только наиболее упоминаемого и следующего за ним по частоте упоминания:

$$I_u^2 = -\beta \cdot \max_{i=1}^{N} \{Count\_Mentions_u(u_i)\} - \min_{i=1}^{N} \min_{\substack{j=2 \\ i>j}}^{N} \left\{ \frac{N\_Mentions_u(u_i, u_j)}{\rho^\alpha(u_i, u_j)} \right\}.$$

Здесь $\rho(u_i, u_j)$ — расстояние между пользователями u_i и u_j. В этой формуле учитываются все пользователи, упоминаемые u, т. е. считается, что все они на него влияют. Расстояние можно, например, задать отношением „подписчик — подписчик подписчика — подписчик подписчика подписчика и т. д.“ Так как эти выкладки производятся относительно предпочтений пользователя u, то уместно будет воспользоваться французской железнодорожной метрикой:

$$\rho(u_i, u_j) = \begin{cases} \|u_i - u_j\|, & u_i - u = \lambda(u_j - u) \\ \|u_i - u\| + \|u_j - u\|, & u_i - u \neq \lambda(u_j - u) \end{cases},$$

где λ — заданный коэффициент, u — фиксированная выбранная точка, через которую обязательно должен проходить путь между u_i и u_j. Самое простое — это подсчитать количество ребер. Можно приписывать вес каждому ребру, но, на наш взгляд, здесь вес ребра не столь важен, т. к. в текущий момент времени пользователь u может не быть подписчиком, например, пользователя u_j, и поэтому находиться от него на далеком расстоянии, а в следующий момент времени уже стать подписчиком.

Внешнее влияние, например, влияние СМИ, также может быть учтено, если в основную формулу Латане добавить дополнительное слагаемое „$-O_i O_M S_{Mi}$“, где S_{Mi} — сила влияния внешних источников на агента i ($S_{Mi} > 0$); O_M — мнение внешнего источника (± 1). Для учета влияния масс-медиа Латане получил [5, 6] итоговую формулу

$$I_i = -S_i\beta - O_i O_M S_{Mi} - \sum_{j=1, j\neq i}^{N} \frac{S_j O_j O_i}{d_{i,j}^\alpha}.$$

Обычно внешний источник также моделируется как агент, но „вне окружающей среды“ и с расстоянием 1 до каждого агента ввиду своей „вездесущей“ природы. Величина S_{Mi} меняется в зависимости от агента, так как каждый человек испытывает различное давление СМИ. Эта величина аналогична иногда рассматриваемой „величине доверия“ агента к сообщениям, получаемым из внешних источников. Для социальных сетей аналогом СМИ можем считать хэштеги и внешние ссылки. Соответственно получаем формулу:

$$I_u = -\beta \cdot \sum_{i=1}^{N} Count\_Mentions_u(u_i) - \sum_{i=1}^{|Hashtags(u)|} \sum_{\substack{i=2 \\ i>j}}^{|Hashtags(u)|} Hashtags(u)(h_i, i_j) -$$

$$-\sum_{i=1}^{N}\sum_{\substack{i=2\\ i>j}}^{N}\frac{N\_Mentions_u(u_i,u_j)}{\rho^{\alpha}(u_i,u_j)},$$

в которой учитываются все пользователи, упоминаемые u, и все хэштеги.

Аналогично получаем формулу, в которой учитываются все пользователи и внешние ссылки в Интернете, упоминаемые пользователем u:

$$I_u=-\beta\cdot\sum_{i=1}^{N}Count\_Mentions_u(u_i)-\sum_{i=1}^{|Urls(u)|}\sum_{\substack{j=2\\ i>j}}^{|Urls(u)|}Urls_u(url_i,url_j)-$$

$$-\sum_{i=1}^{N}\sum_{\substack{j=2\\ i>j}}^{N}\frac{N\_Mentions_u(u_i,u_j)}{\rho^{\alpha}(u_i,u_j)}.$$

4. Некоторые интересные множества пользователей социальной сети и их характеристики. При анализе социальных сетей интерес представляет следующая ситуация. Предположим, что две рассмотренные группы людей употребляют лексику двух разных типов. Ребра в графах соответствуют взаимным цитированиям или другим критериям, связанным с использованием лексики. В качестве весов вершин берутся частоты использования слов. Веса ребер также могут быть определены посредством различных частот слов, используемых ими одновременно. Иначе говоря, интересно пытаться учесть семантическую составляющую сообщений пользователей, т. е. случай, когда разными словами описано одно и то же явление, событие и пр. В этом случае важно рассматривать не только пересечение данных графов или их симметрическую разность, но также некоторые другие подграфы, точнее, некоторые подмножества вершин. Например, множество вершин, „достаточно близких“ к пересечению. Формальные определения даны ниже.

Предположим, имеются два графа $G_i=(V_i,E_i), i=1,2$, где V_i — множество вершин графа G_i; E_i — множество ребер графа G_i. Естественным образом определяются объединение $V_1\cup V_2$, пересечение $V_1\cap V_2$ и симметрическая разность $V_1\Delta V_2=(V_1\cup V_2)\setminus(V_1\cap V_2)=$ $=(V_1\backslash V_2)\cup(V_2\backslash V_1)$.

Обозначим для краткости $H=V_1\cap V_2$, $\mathrm{adj}(x,y)\leftrightarrow E(x,y)\vee E(y,x)$. Введем еще несколько обозначений: $Adj(x)=\{y:adj(x,y)\}$ — множество вершин, смежных с вершиной x; $AdjH(x)=Adj(x)\cap H$ — множество вершин, смежных с вершиной x и лежащих в H; $CAdj(x)=Adj(x)\backslash AdjH(x)$ — множество вершин, смежных с вершиной x и лежащих вне H.

Далее считаем, что заданы две функции: $\omega_i:V_i\to N$ — функция, задающая веса вершин; $r_i:E_i\to N$ — функция, задающая веса ребер.

Можно определить весовую функцию $\omega:V_1\cup V_2\to N$ для вершин, заданную на объединении графов

$$\omega(x)=\begin{cases}\omega_1(x),\ \text{если } x\in V_1\backslash V_2\\ \omega_2(x),\ \text{если } x\in V_2\backslash V_1\\ \dfrac{\omega_1(x)+\omega_2(x)}{2},\ \text{если } x\in V_1\cap V_2\end{cases}.$$

Аналогично может быть определена функция $r:E_1\cup E_2\to N$.

Теперь можем записать числовые характеристики некоторых подграфов:

$$\alpha_i = \sum_{x \in V_i} \omega_i(x), \lambda_i = \sum_{x \in V_1 \bigcap V_2} \omega_i(x), \lambda = \sum_{x \in V_1 \bigcap V_2} \omega(x), \mu = \sum_{x \in V_1 \Delta V_2} \omega(x),$$

$$\beta_i = \sum_{e \in E_i} r_i(e), \beta(x) = \sum_{y \in Adj(x)} r(x,y), \gamma(x) = \sum_{y \in CAdj(x)} r(x,y).$$

Рассмотрим множества $V_i' = \{x' \in V_i : \exists x \in H \ (adj(x,x'))\}$. Наибольший интерес представляет множество тех вершин, которые сопряжены с вершинами из пересечения, а сами лежат вне него $L = (V_1' \cup V_2') \backslash H$.

Наиболее интересными числовыми характеристиками являются следующие.

1. Толерантность вершины

$$T(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \sum_{y \in AdjH(x)} \omega(y).$$

2. Степень защиты вершины

$$D(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \sum_{y \in CAdj(x)} \omega(y).$$

3. Совместность множеств V_1 и V_2

$$Q = \sum_{x \in H} \omega(x) + \sum_{x \in V_1 \Delta V_2} T(x) - \sum_{x \in V_1 \Delta V_2} D(x).$$

Приведенные выше формулы можно обобщить таким образом, чтобы учесть веса ребер.

1. Толерантность вершины с учетом весов ребер

$$T'(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \frac{1}{\beta(x)} \cdot \sum_{y \in AdjH(x)} \omega(y) \cdot r(x,y).$$

2. Степень защиты вершины с учетом весов ребер

$$D'(x) = \frac{\omega(x)}{|\alpha_1 - \alpha_2|} \cdot \frac{1}{\gamma(x)} \cdot \sum_{y \in CAdj(x)} \omega(y) \cdot r(x,y).$$

3. Совместность множеств V_1 и V_2 с учетом весов ребер:

$$Q' = \sum_{x \in H} \omega(x) + \sum_{x \in V_1 \Delta V_2} T'(x) - \sum_{x \in V_1 \Delta V_2} D'(x).$$

5. Анализ сообщений пользователей. В настоящее время в Интернете существует огромное количество новостных и обзорных статей на фильмы, игры, цифровую технику и т. д. После прочтения этих статей пользователи социальных сетей обмениваются впечатлениями и мнениями об их содержании. Чтобы определить источник распространения информации, важно установить соответствие между Интернет-статьей и сообщениями,

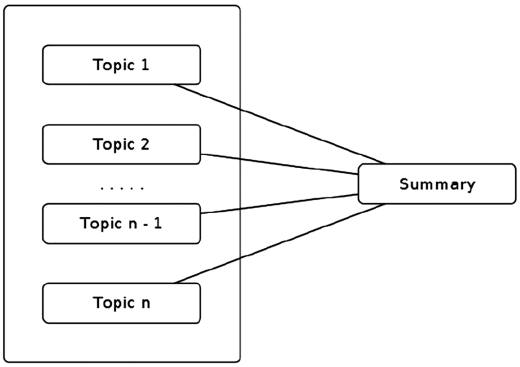

Рис. 1. Общая схема алгоритма

оставляемыми пользователями в сети. Тогда набор твитов одного или нескольких пользователей (к примеру, за какой-то период времени) в некотором смысле можно считать „рефератом“ обзора или новости. Для этого необходимо оценить релевантность сообщения статье. Естественно, что в таком „реферате“ может содержаться информация как из исходной статьи, так и из других источников.

Ниже приведено обобщение алгоритма реферирования [7] новостных и обзорных статей, базирующееся на основе использования синтаксического анализатора Link Grammar Parser. Рассмотрена возможность применения этого алгоритма к задаче оценки релевантности сообщений статьям.

5.1. *Базовый алгоритм определения релевантности сообщений новостным и обзорным статьям.* Базовый алгоритм определения релевантности описан в вышеупомянутой работе [7] и состоит из нескольких этапов.

1. Проводится предварительная обработка статьи, могут удаляться отдельные элементы, специальные обозначения, неподдерживаемые символы.

2. Вычисляются веса слов.

3. Выполняется разбиение статьи на топики. Вычисляются веса топиков. Под топиком будем понимать набор предложений, идущих не обязательно последовательно и охватывающих некоторую самостоятельную подтему в исходной статье. Отметим, что для достаточно большого объема статьи таких подтем всегда будет несколько.

4. Вычисляется оценка релевантности топиков и сообщения.

5. Вычисляется окончательная оценка.

Общая схема алгоритма изображена на рис. 1. Исходная статья разбивается на топики, каждый из которых сравнивается с сообщением, а затем на основе полученных оценок выводится окончательная оценка.

Считаем, что текст статьи представляет собой последовательность предложений, каждое из которых состоит из слов. Тогда статью можем представить в виде графов, вершины которых — слова, а направленные ребра показывают очередность следования слов. Пример построения графа приведен на рис. 2.

Для вычисления весов слов будем рассматривать такие графы. Веса слов можно вычислять различными способами. В частности, в работе [7] предлагается вычислять их по формуле:

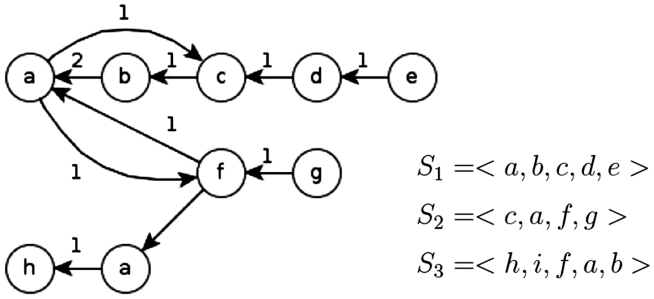

Рис. 2. Граф слов, ассоциированный с набором предложений S_1, S_2, S_3

$$W(v_i) = \frac{1-\lambda}{N} + \lambda \sum_{v_j \in In(v_i)} \frac{W(v_j)}{|Out(v_j)|},$$

где N — количество вершин в графе; $In(v_k)$ — множество вершин, соединенных входящими в v_k связями; $Out(v_k)$ — множество вершин, соединенных исходящими из v_k связями; λ — коэффициент затухания.

Чтобы разбить статью на топики, воспользуемся иерархическим методом кластеризации, например, Group-Average Agglomerative Clustering. Схожесть предложений определяется количеством совместно встречающихся слов. Схожесть топиков вычисляется по следующей формуле:

$$\mathrm{sim}(T_i, T_j) = \frac{1}{|T_i \cup T_j|(|T_i \cup T_j| - 1)} \sum_{S_n \in T_i \cup T_j} \sum_{\substack{lS_m \in T_i \bigcup T_j; \\ S_n \neq S_m}} \mathrm{sim}(S_n, S_m).$$

На каждом шаге мы объединяем наиболее схожие топики. Вес топика определяется весами входящих в него слов. Для работы алгоритма кластеризации также используется формула Ланса-Вильямса.

5.2. *Модель определения релевантности с использованием Link Grammar Parser.* Обобщим теперь данный алгоритм и попытаемся учитывать синтаксическую структуру предложений. Для этого на третьем этапе перед тем, как вычислять веса топиков, воспользуемся результатом работы системы Link Grammar Parser. Link Grammar Parser — это синтаксический анализатор английского языка, базирующийся на оригинальной теории синтаксиса английского языка. Программная система приписывает заданному предложению синтаксическую структуру, состоящую из множества помеченных связей, соединяющих пары слов. Пример синтаксического разбора предложения анализатором приведен на рис. 3. Подробное описание этого Link Grammar Parser можно найти в [8].

Поставим в соответствие предложению S_i граф $G_i(V_i, E_i)$, получаемый в результате работы Link Grammar Parser. В этом графе V_i — множество слов, а E_i — множество троек $< v_1, v_2, t >$, где $v_1, v_2 \in V_i$ — вершины, а t — тип связи. Таким образом, мы имеем $G_1, ..., G_k$ — графы предложений. Следующим этапом построим граф $G(V, E)$, полученный объединением графов предложений $G_1, ..., G_k$. Здесь $V = \bigcup\limits_{1 \leqslant i \leqslant k} V_i$ — множество всех слов, встреченных в предложениях; E — множество четверок $< v_1, v_2, t, n >$, где, как и прежде, $v_1, v_2 \in V_i$ — вершины, t — тип связи, а дополнительный параметр $n = |\{i : < v_1, v_2, t > \in E_i\}|$ — количество раз, когда встретилась тройка $< v_1, v_2, t >$.

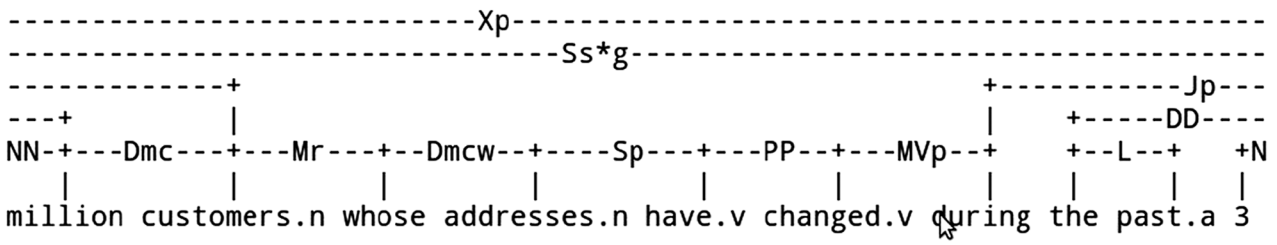

Рис. 3. Пример синтаксического разбора предложения в системе Link Grammar Parser

Определим $LinkLength_{ij}$ как взвешенную сумму чисел, которые соответствуют силе связей между v_i и v_j, где весовыми коэффициентами являются соответствующие значения функции W, определяющей вес для типов грамматических связей:

$$LinkLength_{ij} = \sum_{<v_i,v_j,t,n>\in E} nW\left(t\right).$$

Определим $PathLength_{ij}$, которая будет выступать в качестве весов ребер в окончательном графе. Этот показатель характеризует легкость перехода от одного слова к другому:

$$PathLength_{ij} = \frac{1}{LinkLength_{ij}}.$$

Для каждой вершины вычислим Clotheness Centrality — величину, обратную к среднему геодезическому расстоянию до остальных вершин

$$C_c\left(v_i\right) = \frac{N-1}{\sum\limits_{v_j\in V, v_j\neq v_i} d_G(v_i, v_j)}.$$

Для каждой вершины вычислим относительное изменение показателя Clotheness Centrality

$$Diff\left(v_i\right) = \frac{|C_c^1\left(v_i\right) - C_c^2\left(v_i\right)|}{C_c^1\left(v_i\right)}.$$

Наконец, определим предикат, который будет сигнализировать, входят ли вершины схожим образом в графы для топика и сообщения:

$$Sim\left(v_i\right) = (Diff_{avg} < 0,5 \wedge Diff\left(v_i\right) < 0,5) \vee (Diff_{avg} > 0,5 \wedge Diff\left(v_i\right) < Diff_{avg}).$$

Оценкой релевантности топика к сообщению будет служить относительное количество совместно встречающихся слов, для которых предикат верен

$$Score\left(T\right) = \frac{|\left\{v \in V:\ Sim\left(v\right)\right\}|}{|Set\left(T\right)|}.$$

Окончательная оценка — это взвешенная сумма оценок для топиков — вычисляется по формуле:

$$Score = \sum_{T} W(T)\ Score(T).$$

Таким образом, становится возможным установить, относится ли действительно сообщение или набор нескольких сообщений к выбранной статье. Можно ввести дополнительный коэффициент, помогающий определить отношение пользователя к фильму, игре, новости, описываемым в статье. И если вычисленная оценка релевантности удовлетворяет заранее заданному критерию, то можно будет получить более полную информацию об отношении пользователя к прочитанному материалу. Полезнее, конечно, рассматривать не отдельно взятого человека, а группы людей. Заметим, что этот процесс удобно распараллеливается.

6. Программная реализация. В процессе исследований был разработан программный комплекс, содержащий модули извлечения информации из социальных сетей, обработки, анализа и визуализации данных. Все модули реализованы на языке Python для широкого круга операционных систем, на которых может работать комплекс. Структура программного комплекса показана на рис. 1.

Модуль извлечения данных имеет возможность извлекать данные, в первую очередь, из крупнейших социальных сетей: Twitter, Facebook, „ВКонтакте“. Для доступа к каждой из них используется интерфейс прикладного программирования (API), авторизация производится при помощи протокола OAuth. На данный момент этот модуль имеет возможность функционального расширения практически на любую социальную сеть, в зависимости от предоставляемого API.

Основными трудностями на этапе извлечения данных были ограничения на количество запросов к серверам социальных сетей от определенного IP адреса. На начальном этапе реализации ограничение на доступ к серверам Twitter составляло не более 350 запросов в час, на данный момент от 60 до 720 в зависимости от типа запроса. Сервера „ВКонтакте“ имеют ограничение в 3 запроса в секунду. Серверы Facebook не имеют каких-либо ограничений на количество запросов, однако каждый запрос обрабатывается ими в течение 1–2 секунд. Для преодоления этих трудностей было решено использовать прокси серверы (на данный момент в модуль можно загружать их список). Также ввиду возможности модуля работать практически на любой операционной системе довольно легко увеличить количество машин, на которых работает модуль, т. е. можно использовать распределенную обработку.

Извлеченные данные пользователей можно разделить на три категории: собственно пользовательские данные, такие как имя, ник, время регистрации; сообщения пользователей; связи между пользователями. После извлечения модуль обработки данных производит поиск маркеров хэштегов, упоминаний пользователей, ссылок и т. д. Далее производится нормализация текста сообщений в зависимости от настроек: либо с помощью стеммера Портера [9] (для большей скорости обработки), либо морфологическая нормализация на основе алгоритмов АОТ [10] с использованием библиотеки PyMorphy. Для хранения данных используется документоориентированная база данных MongoDb.

В модуле анализа данных используются различные алгоритмы кластеризации и классификации данных как самих пользователей и их связей, так и их сообщений. В модуле построения графовых структур имеется возможность для построения графов, отражающих связи пользователей. При этом могут использоваться данные, как исходные, так и полученные в результате анализа. В этом модуле также имеется возможность „выгрузки данных“ в программное средство для работы с графами Gephi [11], как в некотором

специальном формате, так и посредством http протокола. Модуль визуализации данных дает возможность на основе извлеченных данных строить графики зависимостей между различными показателями.

Заключение. При анализе социальных сетей решается довольно большой круг задач и применяются методы из различных областей знаний. Часто решение задач из одного класса связано с решением задач из другого класса. Поэтому к их решению приходится подходить комплексно. Для этого, безусловно, необходимы новые методы, алгоритмы, обнаружение новых характеристик, которые помогли бы в решении возникающих вопросов.

В данной статье предложены количественные и структурные характеристики, даны соответствующие понятия, модели и методы, которые могут быть полезны для анализа информации, полученной из социальных сетей.

Был разработан программный комплекс, позволяющий извлекать информацию из социальных сетей, проводить обработку, анализ и визуализацию данных. Модуль извлечения данных имеет возможность извлекать данные, в первую очередь, из крупнейших социальных сетей: Twitter, Facebook, „ВКонтакте“.

Программный комплекс при использовании одного компьютера позволяет в сутки выполнять от 8 до 250 тысяч запросов в зависимости от того, к какой социальной сети осуществляется запрос и в зависимости от быстродействия оборудования и пропускной способности каналов. Очевидно, что объем получаемой информации оказывается очень большим. При увеличении одновременно используемых компьютеров, т. е. при использовании распределенной системы получения и обработки данных, объем еще более возрастает. Поэтому сначала необходимо из всего объема информации выделить ту часть, которую можно было бы достаточно эффективно обработать и которая представляла бы интерес в соответствии с поставленными целями. Необходимо более детальное исследование этого нетривиального вопроса.

Для определения источника распространения информации предлагается использовать обобщенный алгоритм оценки релевантности сообщений, оставляемых в социальных сетях, статьям, публикуемым в Интернете. Чтобы учесть синтаксическую структуру текста, предлагается использовать результат работы синтаксического анализатора Link Grammar Parser.

Описанная модель представляется перспективной. Тем не менее, настоящая работа еще далека от завершения. Весьма вероятно, предложенный подход потребует доработки. Можно рассматривать множество вариаций обсуждаемого алгоритма: использовать различные методы кластеризации, меры схожести подтем, варьировать формулы для вычисления весов слов и пр. Чтобы найти наилучшую конфигурацию, при которой алгоритм обеспечит качественные результаты, следует протестировать различные комбинации вариаций этого алгоритма, а также привлечь экспертов для оценки качества работы алгоритма. Это довольно трудоемкий процесс. Также не исключено, что во время изучения полученных результатов тестирования станут видны способы улучшения описанного алгоритма.

Список литературы

1. Charu C. Aggarwal Social network data analytics. 2011.

2. Батура Т. В. Методы анализа компьютерных социальных сетей. // Вестник НГУ. Серия: Информационные технологии. Новосибирск. 2012. Т. 10. Вып. 4. С. 13–28.

3. Роджерс Э., Агарвала-Роджерс Р. Коммуникации в организациях. М.: Экономика, 1980.

4. Крючков В. Н., Мурзин Ф. А., Нартов Б. К. Исследование связей в коллективах и сетях ЭВМ на основе анализа предпочтений // Проблемы конструирования эффективных и надежных программ. Новосибирск. 1995. С. 136–141.

5. Nowak A., Szamrej J., Latané B. From private attitude to public opinion: a dynamic theory of social impact // Psychological Review. 1990. V. 97. P. 362–376.

6. Latané, B. The psychology of social impact // American Psychologist. 1981. V. 36. P. 343–356.

7. Kumar N., Srinathan K., Varma V. Using graph based mapping of co-occurring words and closeness centrality score for summarization evaluation // CICLing Proc. of the 13th Intern. Conf. on Computational Linguistics and Intelligent Text Proc. 2012. P. 353–365.

8. Sleator D., Temperley D. Parsing English with a Link Grammar. Pittsburgh: School of Computer Science Carnegie Mellon University, 1991.

9. Willett P. The Porter stemming algorithm: then and now // Program: Electronic Library and Information Systems. 2006. V. 40. N 3. P. 219–223.

10. Автоматическая Обработка Текста [Электрон. ресурс]. 2012. Режим доступа: http://aot.ru/.

11. Gephi, an open source graph visualization and manipulation software [Электрон. ресурс]. 2012. Режим доступа: https://gephi.org/.

Батура Татьяна Викторовна — канд. физ.-мат. наук, науч. сотр. Института систем информатики им. А. П. Ершова СО РАН; e-mail: tatiana.v.batura@gmail.com

Мурзин Федор Александрович — канд. физ.–мат. наук, зам. дир. по науч. работе Института систем информатики им. А. П. Ершова СО РАН; e-mail: murzin@iis.nsk.su

Проскуряков Алексей Викторович — асп. Института систем информатики им. А. П. Ершова СО РАН; e-mail: alexey.proskuryakov@gmail.com

Сперанский Данил Олегович — студент Новосибирского национальноого исследовательского государственного университета; e-mail: speranskydanil@gmail.com

Дата поступления: 15.05.2014