

АВТОМАТИЗИРОВАННАЯ ОЦЕНКА ЭМОЦИЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

УДК 519.87: 004.02

Елена Евгеньевна Лунева,
к.т.н., доцент кафедры Автоматики и компьютерных систем
Томского Политехнического университета (ТПУ)
Тел.: (3822) 41-89-07
Эл. почта lee@tpu.ru

Александр Александрович Ефремов,
ассистент кафедры Автоматики и компьютерных систем
Томского Политехнического университета (ТПУ)
Тел.: (3822) 41-89-07
Эл. почта alexeyefremov@tpu.ru

Павел Иванович Банокин,
ассистент кафедры Автоматики и компьютерных систем
Томского Политехнического университета (ТПУ)
Тел.: (3822) 41-89-07
Эл. почта banokin@tpu.ru

В статье излагается способ автоматизированной оценки эмоций пользователей социальных сетей относительно некоторой темы, отличающийся возможностью учета в результирующей оценке как «влиятельности» пользователей, так и наличия повторных сообщений. Входными данными для анализа эмоций пользователя относительно определенной темы являются его релевантные сообщения, для автоматизированного разбора которых предлагается использовать алгоритмы нечеткой логики, в частности при обработке модификаторов, встречающихся в сообщениях. В статье приведены экспериментальные данные, демонстрирующие расчет результирующей эмоциональной оценки сообщений по определенной тематике.

Ключевые слова: оценка эмоций пользователей, социальная сеть, нечеткая логика, характеристика, терм, модификатор, граф, автоматизация, способ.

Elena E. Luneva,
PhD, Associate Professor, Department of Automatics and
Computer Systems, Tomsk Polytechnic University
Tel.: (3822) 41-89-07
E-mail: lee@tpu.ru

Alexandr A. Yefremov,
Teaching assistant, Department of Automatics and Computer
Systems, Tomsk Polytechnic University
Tel.: (3822) 41-89-07
E-mail: alexeyefremov@tpu.ru

Pavel I. Banokin,
Teaching assistant, Department of Automatics and Computer
Systems, Tomsk Polytechnic University
Tel.: (3822) 41-89-07
E-mail: banokin@tpu.ru

AUTOMATED SENTIMENT ANALYSIS EVALUATION OF SOCIAL NETWORK USERS BASED ON FUZZY LOGIC

In the article the method of automated sentiment analysis evaluation of social network users is represented. The advantage of suggested method consists in ability taking into account user's authority as well as the fact of several messages from one user. As input data for the sentiment analysis with respect to a certain topic is its relevant messages. Automated analysis of these messages is obtained with application of fuzzy logic algorithms. Particularly when in messages contains fuzzy hedges. The paper presents experimental data showing the steps of calculation of the resulting sentiment evaluation of messages on a certain topic.

Keywords: sentiment analysis evaluation, social network, fuzzy logic, characteristic, term, hedge, graph, automation, method.

1. Введение

В настоящее время социальные сети, насчитывают сотни миллионов пользователей, и их количество постоянно растет [1]. При этом, обычной практикой является использование социальных сетей для выражения своего мнения по интересующим темам или событиям, а также для отзывов об используемых продуктах, услугах. Эти данные могут быть использованы для решения таких актуальных задач как выявление потенциальных клиентов определенного продукта или услуги, прогнозирование потребности в продукте или услуге, а также оценка общественного мнения, настроения граждан относительно политической ситуации, в том числе, для прогнозирования итогов выборов [2–4].

Сложность автоматизированного решения перечисленных выше задач заключается в том, что мнение пользователей выражено не в количественном, а в качественном виде, в словесной форме с использованием нечетких характеристик предмета сообщения, например, как «интересный», «изящный» и т.д. В основе анализа сообщений пользователей социальных сетей лежит оценка эмоциональной составляющей этих сообщений.

Анализ существующих научных работ, посвященных тематике автоматизированной оценки эмоций пользователей по их сообщениям [2–6] показал, что в данных работах не учитываются взаимоотношения между пользователями, а также имеются ограничения, связанные с обработкой нечетких характеристик с модификаторами, используемых в сообщении. Данные ограничения влияют на результирующую оценку эмоций группы пользователей относительно некоторой темы. Например, в случае, когда один пользователь формирует несколько сообщений, в результате в работах [2–5] учитывается суммарное количество всех сообщений вне зависимости от того, какое количество пользователей их оставило. Кроме того, в современных социальных сетях пользователи не только высказывают мнения; позиция некоторых пользователей может оказывать влияние на других пользователей. Соответственно, при анализе эмоций пользователя можно учитывать «авторитет» (влиятельность) пользователей, либо брать в расчет только пользователей с высокой степенью влияния, как задающих тренд общественного мнения относительно некоторого продукта, темы или события.

Целью данной работы является разработка автоматизированного способа оценки эмоций пользователей социальных сетей с учетом перечисленных выше ограничений.

2. Теоретический анализ

Входными данными для анализа эмоций пользователя относительно определенной темы являются его релевантные сообщения. Единицей сообщения является предложение. В каждом предложении могут высказываться положительные или отрицательные эмоции относительно некоторых характеристик рассматриваемой темы (продукта или предмета). Т.о. существует множество характеристик K_i , $i = [1..n]$, где

n – количество характеристик. Например, если в сообщении дается отзыв на электронную книгу, то примером таких характеристик могут быть {«Качество батареи», «Удобство интерфейса», «Качество экрана», «Устройство в целом», ...}. Для оценки продукта относительно каждой характеристики K_i пользователи могут использовать множество термов $T_{ij}, j = [1...m_i]$, где m_i – количество таких термов. Например, если $K_i =$ «Качество экрана», то используемыми термами могут быть $T_i =$ {«Хрупкий», «Удобный», «Плохой», ...}. Для оценки эмоций пользователя необходимо знать эмоциональную оценку каждого из используемых термов T_{ij} , которые можно представить в виде функций принадлежности лингвистической переменной. Однако для этого необходимо предварительно задать такие функции принадлежности для всех термов, что представляет собой достаточно трудоемкую задачу. Кроме того, даже если предварительно определить характеристики и термы, относительно которых будет оцениваться сообщение, то возможна ситуация, когда пользователи социальных сетей будут использовать в том числе и другие термы и характеристики, не попавшие в множество K и T . Однако эти характеристики и термы тоже следует учитывать для получения результирующей эмоциональной оценки сообщения. Таким образом, первым этапом предлагаемого способа является ввод исходных данных в виде характеристик, термов и модификаторов, как указано выше.

Для того чтобы сформировать эмоциональную оценку каждого терма без необходимости вручную задавать функции принадлежности, был проведен поиск готовых программных решений (компонентов), предоставляющих данные о значениях функции принадлежности нечеткого терма T_{ij} . Среди найденных программных решений можно выделить такие как General Inquirer, Subjectivity Clues Lexicon, SentiWordNet [6]. В данной работе в качестве готового решения, содержащего эмоциональную окраску термов, был выбран компонент

SentiWordNet 3.0, по причине того, что этот продукт поддерживается производителем и находится в состоянии постоянного развития, а также не является коммерческой разработкой. Компонент SentiWordNet 3.0 предоставляет данные о позитивности, негативности и объективности слов английского языка. При этом, в зависимости от смысла одного и того же слова, оценка может быть различной. Объективность слова означает констатацию некоторого факта без эмоциональной окраски. Например, терм «зеленый» в предложении «трава зеленая» только констатирует факт, что трава зеленая.

Для формирования функций принадлежности термов введем следующее множество-носитель $P = \{P$ («Позитивный»), O («Объективный»), N («Негативный»)». Исходя из рассуждения, что термин не может быть одновременно позитивным, негативным и объективным с максимальной степенью принадлежности, введем правило: $\mu_P + \mu_O + \mu_N = 1$, где μ_P, μ_O, μ_N – значение функции принадлежности для элементов P, O, N носителя нечеткого множества («Эмоциональная окраска»). Тогда каждый терм можно представить дискретной функцией принадлежности, где в качестве носителя используется нечеткое множество «Эмоциональная окраска». Таким образом, и для заранее заданных термов T , относительно которых проводится анализ, и для всех прочих, при помощи компонента SentiWordNet 3.0 может быть построена функция принадлежности для проведения анализа.

Для того чтобы получить четкое значение t_{ij} рассматриваемого терма T_{ij} следует выполнить процедуру дефаззификации [7–9]. Т.к. функция принадлежности является дискретной, предлагается выполнять дефаззификацию по первому максимальному значению (FOM) [10]. В случае если максимальных значений более одного, лицо, проводящее анализ, выбирает одно из значений по своему усмотрению. Если после дефаззификации четкое значение рассматриваемого терма T_{ij} равно « O » («Объективный»), то данный терм при оценке эмоций

пользователя не следует учитывать. Таким образом, учитываются только позитивные или негативные термы.

Для случая, когда терм используется с модификатором, к функции принадлежности следует применить операторы Заде (увеличения и уменьшения нечеткости) [7, 8]. Например, для модификатора «very» («очень») значения функции принадлежности должны быть возведены во вторую степень. После применения оператора Заде необходимо выполнить нормализацию полученных значений по формуле (1), для того, чтобы сохранился смысл термина с модификатором.

$$\mu_k = \frac{\mu_k^2}{\sum_{j=1}^3 \mu_j^2}, \quad (1)$$

где μ_k – значение функции принадлежности нечеткого термина $T_{ij}, k = [1...3]$.

Например, если сравнивать термы «good» (хорошо) и «very good» (очень хорошо), то очевидно, что последнее имеет более высокую позитивную эмоциональную окраску по сравнению с первым термом и степень принадлежности к значению P («Позитивный») должна увеличиться. Применение оператора Заде, без нормализации не позволяет получить данное соответствие. На рис. 1 показаны функции принадлежности для термов «good», «very good» без нормализации и «very good» с нормализацией.

В каждом сообщении может быть найдено l – термов, позитивных или негативных. Суммарная эмоциональная оценка одного сообщения –

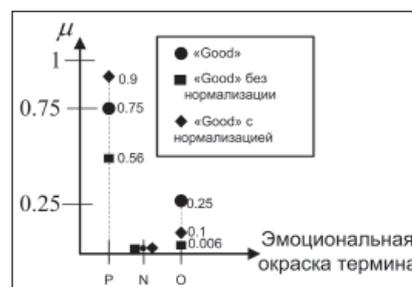


Рис. 1. Функции принадлежности для термов «good», «very good» без нормализации и «very good» с нормализацией

S может быть рассчитана по следующей формуле:

$$S = \frac{\sum_{i=1}^l t_i \cdot (-1)^k}{l}, \quad (2)$$

где $k=1$, если терм является позитивным, $k=-1$, если терм является негативным.

Формула (2) также может быть использована для расчета эмоциональной оценки по одной из характеристик. Значение, рассчитанное по формуле (2), лежит в диапазоне от $[-1; 1]$. Для того чтобы разделить сообщения по степени позитивности введем лингвистическую переменную «Эмоциональная характеристика» с термами: «Очень позитивный» (ОП), «Позитивный» (П), «Нейтральный» (Н), «Негативный» (НЕ) и «Очень негативный» (ОНЕ), при этом носителем является множество действительных чисел в интервале $[-1; 1]$. Учитывая, что для данной лингвистической переменной выделено всего пять термов, при разработке функций принадлежности была взята пятибалльная оценочная система [10]. Исходя из рассуждения, что значение эмоциональной характеристики не может относиться с максимальной степенью принадлежности нескольким термам, введем правило, что сумма функций принадлежности всех термов лингвистической переменной «Эмоциональная характеристика» должна составлять 1, т.е. $\forall x \in [-1...1] \sum_{i=1}^5 \mu(x)_i = 1$. На рис. 2 показаны заданные функции принадлежности:

Суммарная оценка по всему объему анализируемых сообщений будет не точной, если не учитывать, что один пользователь может оста-

вить более одного сообщения. При этом эмоциональная оценка каждого из сообщений от одного пользователя, может быть либо разной, либо попадать в одну категорию. Учитывая, что существует множество причин, по которым оценка одного пользователя может отличаться от сообщения к сообщению, например, пользователь мог поменять мнение, пользователь мог находиться в плохом или хорошем настроении, когда писал отзыв, и т.д., предлагается использовать среднее арифметическое значение эмоциональной оценки пользователя по нескольким сообщениям. Таким образом, если существуют пользователи, которые оставили более одного сообщения, то суммарная оценка формируется так, как описано ниже. Анализируется p сообщений, и количество пользователей, оставивших данные сообщения равно u , где $u \leq p$. При этом количество сообщений, которые оставил j -ый пользователь ($j = 1...u$) равно c_j , и по каждому сообщению получена оценка S_i , $i = 1...p$, то на основе этих данных следует вычислить эмоциональное отношение пользователя к анализируемой теме SS_j , $j = 1...u$, по следующей формуле:

$$SS_j = \frac{\sum_{i=1}^{c_j} S_i}{c_j}$$

Еще одним аспектом, который следует учитывать при формировании суммарной оценки, является показатель «влиятельности» пользователя. Для вычисления влияния пользователя сегодня используются различные показатели, предоставляемые такими сервисами как «Klout», «PeerIndex», «Kred»

и др. [11]. В большинстве случаев показатель влияния или авторитета пользователя зависит от количества подписчиков пользователя, количества репостов [2–5], частоты публикаций. Наиболее популярным среди приведенных сервисов является «Klout» [11], предоставляющий данные по показателю «Klout score», который измеряется в диапазоне от 0 до 100. Данный сервис также предоставляет API-интерфейс, позволяющий получить данные показателя для каждого пользователя таких социальных сетей как Twitter, Facebook, LinkedIn и некоторых других. Показатель влияния часто учитывается при формировании суммарной оценки для некоторых методов экспертной оценки. Рассматривая пользователей как экспертов, процент отзывов β , относящихся к некоторой эмоциональной характеристике (рис. 3), может быть рассчитан по формуле:

$$\beta = \frac{\sum_{i=1}^q (R_i \cdot \mu_i)}{\sum_{i=1}^u (R_i)} \cdot 100\%, \quad (3)$$

где, R_i – показатель «Klout score» i -ого пользователя, u – количество пользователей, q – количество отзывов, относящихся к рассматриваемой эмоциональной характеристике.

Однако, показатель «Klout score» вычисляется только для пользователей, подключенных к данному сервису. Согласно исследованию [11] сервис «Klout» на сегодняшний день охватывает более 600 млн. пользователей различных социальных сетей. Поэтому, кроме данного показателя влияния, дополнительно предлагается также формировать независимый от какого-либо сервиса альтернативный показатель. В качестве такого предлагается использовать один из показателей центральности пользователя, например, центральность по близости, отражающая, насколько быстро распространяется информация в сети от одного участника к остальным [12]. Расчет данного показателя выполняется на основе социального графа G [12], построенного исходя из следующих рассуждений.

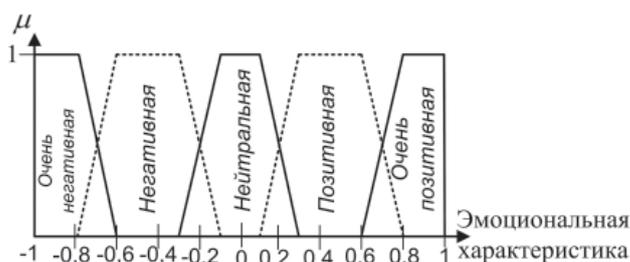


Рис. 2. Функции принадлежности термов лингвистической переменной «Эмоциональная характеристика»

В современных социальных сетях распространение информации происходит через механизм подписки [12]. Если пользователь *A* является подписчиком пользователя *B*, то пусть существует дуга, направленная от вершины *B* к вершине *A*. Анализируемое множество отзывов пользователей, посвященных избранной тематике (например, модель электронной книги) выбирается из всего множества пользователей по хэш тегам [3, 4], а также заголовкам отзывов.

На основе информации об авторах отзывов и их подписчиках может быть построен социальный граф. Если сообщения по анализируемой теме оставили *u* пользователей, то социальный граф может включать количество вершин больше *u*, т.к. содержит, в том числе, и тех подписчиков, которые не оставляли сообщения по анализируемой теме. Далее, только для *u* пользователей следует вычислить показатель центральности (R_i , где $i = 1 \dots u$, u – количество пользователей). Центральность по близости рассчитывается по следующей формуле из [12]:

Дальнейшую оценку эмоций пользователей по анализируемому множеству сообщений можно осуществлять либо с учетом полученных показателей центральности R_i , либо предоставить лицу, проводящему анализ, возможность учитывать только сообщения «влиятельных» пользователей. При этом процент отзывов, относящихся к некоторой эмоциональной характеристике (рис. 3) рассчитывается по формуле (3).

3. Экспериментальный анализ способа оценки эмоций пользователей приложений, функционирующих по принципу социальных сетей

Исходя из рассуждений и соотношений, приведенных в разделе «Теоретический анализ», в данном разделе приводятся экспериментальные данные, полученные в соответствии с разработанным способом. Для проведения экспериментов было разработано программное обеспечение, позволяющее выбрать сообщения из социальной сети

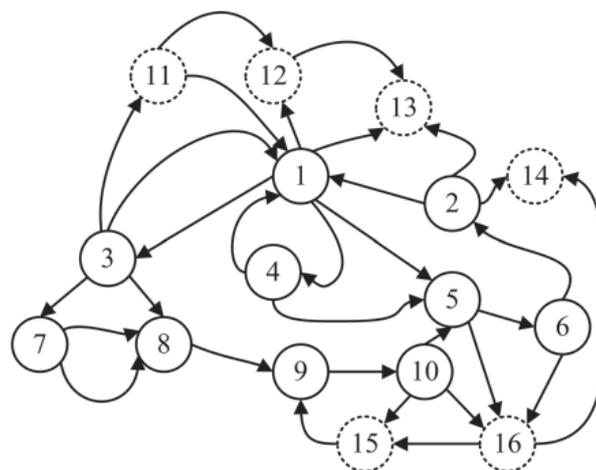


Рис. 3. Пример графа, построенный при анализе сообщений из социальной сети Twitter

Пунктиром обозначены пользователи, являющиеся читателями, но не оставлявшие сообщения по анализируемой тематике

«Twitter» по нескольким случайным тематикам. Далее, программно была проанализирована информация о пользователях и построены графы, один из которых приведен в качестве примера на рис. 3.

Расчет эмоциональной оценки *S* для каждого сообщения был заменен случайно сгенерированным числом в диапазоне от -1 до 1. Если каким-либо пользователем был выполнен «репост» сообщения, то использовалась эмоциональная оценка исходного сообщения. В качестве исходных данных в таблице 1 представлен список сообщений, полученных от пользователей 1–10, и их эмоциональная оценка. По причине наличия нескольких сообщений от некоторых пользователей, был рассчитан пара-

метр *SS* для каждого пользователя, а также центральность по близости для каждого пользователя. В качестве примера в табл. 1 также приведены рассчитанные параметры для графа, представленного на рис. 3.

Используя рассчитанные значения параметра *SS*, оценивается степень принадлежности μ к нечетким термам лингвистической переменной «Эмоциональная характеристика» (табл. 1). В таблице используются следующие обозначения: m_{ij} – *j*-ое сообщение *i*-ого пользователя; rm_{ij} – репост сообщения m_{ij} .

На основе данных, представленных в табл. 1 можно рассчитать процентное распределение эмоций пользователей по различным эмоциональным характеристикам (2).

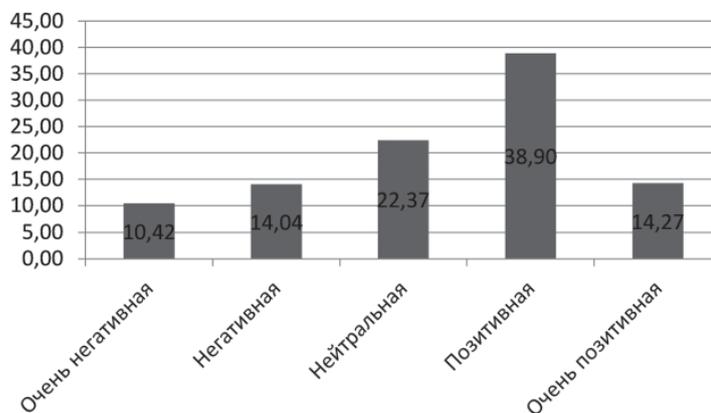


Рис. 4. Процентное распределение эмоций пользователей по различным эмоциональным характеристикам

Таблица 1

Рассчитанные параметры

Пользователь	Обозначение сообщения	S_j	SS_i	R_i	Эмоциональная характеристика / степень принадлежности m
1	$m_{1,1}$	0,67	0,76	0,48	ОП / 0,45; П / 0,55
	$m_{1,2}$	0,85			
2	$m_{2,1}$	-0,45	-0,45	0,39	HE / 1
3	$rm_{1,1}$	0,67	0,67	0,43	П / 0,59; ОП / 0,41
4	$rm_{1,1}$	0,67	0,18	0,39	Н / 0,73; П / 0,27
	$rm_{2,1}$	-0,45			
	$m_{4,1}$	0,34			
5	$m_{5,1}$	0,24	0,24	0,31	П / 0,74; Н / 0,26
6	$m_{6,1}$	-0,7	-0,7	0,36	ОНЕ / 0,5; Н 0,5
7	$rm_{1,1}$	0,67	0,67	0,18	П / 0,59; ОП / 0,41
8	$m_{8,1}$	0,93	0,19	0,22	Н / 0,72; П / 0,28
	$m_{8,2}$	-0,56			
9	$m_{9,1}$	-0,76	-0,76	0,23	ОНЕ / 0,7; HE / 0,3
10	$m_{10,1}$	0,3	0,3	0,28	П / 0,9; Н / 0,1

В качестве примера, на рис. 4 приведены полученные результаты для данных из табл. 1.

Представленные на рис. 4 результаты позволяют сделать вывод, что анализируемая тематика вызывает различные эмоции пользователей: от очень негативных до очень позитивных. Однако можно заключить, что около половины пользователей позитивно относятся к анализируемому вопросу.

Дальнейшее исследование в рамках представленного способа связано с разработкой точных и быстрых алгоритмов семантического разбора сообщений пользователя, а также с использованием быстрых алгоритмов [13] для вычисления «влиятельности» пользователей, и проведением анализа данных, полученных из нескольких социальных сетей.

4. Выводы

В данной работе представлен разработанный способ для автоматизированной оценки эмоций пользователей социальных сетей, предоставляющий возможность учитывать в результирующей оценке как «влиятельность» пользователей, так и наличие повторных сообщений.

Проведена серия экспериментов при помощи специально разработанного приложения, что показало работоспособность проведения процедуры оценки эмоций пользователей

социальных сетей, в соответствии с предложенным способом.

Предлагаемый способ может быть применен для решения таких задач как выявление потенциальных клиентов продукта или услуги, прогнозирование потребности в продукте или услуге, а также при оценке общественного мнения относительно некоторого события.

Литература

1. URL: <http://www.statisticbrain.com/> (дата обращения: 16.02.2015)
2. Dalal M.K., Zaveri M.A. Opinion Mining from online user reviews using Fuzzy Linguistic Hedges // Applied Computational Intelligence and So. Computing Volume. – 2014. Vol. 2014, № 735942. – P. 1–9.
3. Bollen J., Mao H., Zeng X.-J. Twitter mood predicts the stock market // Journal of Computational Science. – 2011. Vol. 2, № 1. – P. 1–8
4. Amit P., Tejashree W., Swati Rao M. Review of Online Product using Rule Based and Fuzzy Logic // Smiley’s International Journal of Computing and Technology. – 2014. Vol. 1, P. 39–44
5. Haque Md., Rahman T. Sentiment analysis by using fuzzy logic // International Journal of Computer Science, Engineering and Information Technology (IJCEIT). – 2014. Vol. 4, № 1. – P. 33–48
6. Denecke K. Using SentiWordNet for multilingual sentiment analysis // Proceedings International Confer-

ence on Data Engineering. 2008. – № 4498370. – P. 507–512

7. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. – Москва: Мир, 1976. – 165 с.

8. Leekwijck W.V., Kerre E.E. De-fuzzification: criteria and classification // Fuzzy Sets and Systems. – 1999. Vol. 108, № 2. – P. 159–178

9. Ефремов А.А. Новые операции над нечеткими числами и интервалами // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2013 – №. 1(27). – С. 95–99

10. Ходашинский И.А. Нечетко-логическое оценивание величин // Известия Томского политехнического университета. – 2003. –Т. 306, № 3. – С. 10–15.

11. Инструменты мониторинга в социальных сетях. URL: <http://habrahabr.ru/post/173341/> (дата обращения: 16.02.2015)

12. Freeman L.C. Centrality in social networks, conceptual clarifications // Soc. Networks. – 1979. Vol. 1. – P. 215–236.

13. Song Q., Wang X.F. Survey of speedup techniques for shortest path algorithms // Journal of the University of Electronic Science and Technology of China. – 2012. Vol. 41, № 2. – P. 176–184.

References

1. URL: <http://www.statisticbrain.com/> (date of handling: 16.02.2015)
2. Dalal M. K., Zaveri M.A. Opinion Mining from online user reviews using Fuzzy Linguistic Hedges // Applied Computational Intelligence and So. Computing Volume. – 2014. Vol. 2014, № 735942. – P. 1–9.
3. Bollen J., Mao H., Zeng X.-J. Twitter mood predicts the stock market // Journal of Computational Science. – 2011. Vol. 2, № 1. – P. 1–8
4. Amit P., Tejashree W., Swati Rao M. Review of Online Product using Rule Based and Fuzzy Logic // Smiley’s International Journal of Computing and Technology. – 2014. Vol. 1, P. 39–44
5. Haque Md., Rahman T. Sentiment analysis by using fuzzy logic // International Journal of Computer Science, Engineering and Information Technology (IJCEIT). – 2014. Vol. 4, № 1. – P. 33–48

6. Denecke K. Using SentiWordNet for multilingual sentiment analysis // Proceedings International Conference on Data Engineering. 2008. № 4498370. – P. 507–512
7. Zadhe L. The Concept of a Linguistic Variable and its Applications to Approximate Reasoning. – M.: Mir, 1976. – 165 P.
8. Leekwijck W.V., Kerre E. E. De-fuzzification: criteria and classification // Fuzzy Sets and Systems. – 1999. Vol. 108, № 2. – P. 159–178
9. Yefremov A.A. New operations on fuzzy numbers and intervals // Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki. – 2013 – №. 1(27). – С. 95–99
10. Khodashinsky I.A. Nechetkologicheskoe evaluation values // Izvestiya Tomskogo politehnicheskogo universiteta. – 2003. –Т. 306, № 3. – S. 10–15.
11. Monitoring tools in social networks. URL: <http://habrahabr.ru/post/173341/> (date of handling: 16.02.2015)
12. Freeman, L.C. Centrality in social networks, conceptual clarifications // Soc. Networks. – 1979. V. 1. – pp. 215–236.
13. Song, Q., Wang, X.F. Survey of speedup techniques for shortest path algorithms // Journal of the University of Electronic Science and Technology of China . – 2012. – V. 41, № 2. – pp. 176–184.