Инженерный вестник Дона, №3 (2014)
ivdon.ru/ru/magazine/archive/n3y2014/2528

Применение теории графов в интеллектуальной методике анализа социальных медиа для мониторинга популярности кандидатов в период предвыборной кампании

В.И. Носко

Южный федеральный университет, Ростов-на-Дону

Аннотация: В настоящее время все большую значимость приобретают вопросы анализа блогосферы в период событийных ситуаций, связанных в первую очередь с проведением предвыборных компаний. В статье проводится анализ популярности украинских кандидатов в президенты в украинском сегменте социальных сетей и блогосфере за весь официальный период предвыборной кампании. Для анализа структуры предлагаются разнообразные методики исследования, моделирования и оценки активности блогосферы: теория графов, анализ социальных сетей, обработка естественного языка. Исследование показывает, кто и каких кандидатов преимущественно обсуждает, структуру высказываний и пересечение интересов аудиторий.

Ключевые слова: граф социальной сети, обработка данных, большие данные, граф упоминаний, кандидаты в президенты, теория графов, анализ социальных взаимодействий, social network graph, data mining, big data, opinions graph, presidential candidates, graph theory, social network analysis

В настоящее время все большую значимость приобретают вопросы анализа блогосферы в период событийных ситуаций, связанных в первую очередь с проведением предвыборных кампаний [1]. Исследование популярности кандидатов в блогосфере за весь официальный период предвыборной кампании имеет большое значение, как для общества, так и для самих кандидатов [2]. Предлагаются разнообразные методики исследования, моделирования и оценки активности блогосферы и социальных медиа, в которых выражаются мнения [3]. В задаче разработки методики мониторинга популярности кандидатов предвыборных кампаний применяются наработки из следующих смежных областей знаний [4, 5]: теория графов, анализ социальных сетей, обработка естественного языка.

В настоящей работе рассмотрен ряд аспектов применения теории графов в интеллектуальной методике анализа социальных медиа и блогосферы на примере мониторинга популярности кандидатов предвыборной кампании в современной Украине в апреле-мае 2014 года. Мнения собирались с помощью поиска по ключевым словам, которыми являются фамилии и прозвища кандидатов в президенты [6]. Сбор велся с ограничением по географии: только в украинских блогах – в самих блогах и в комментариях из блогов. Период сбора данных: с 1 апреля 2014 по 24 мая 2014 (официальный период предвыборной кампании кандидатов в президенты). За все время было собрано 37650 постов с упоминанием кандидатов, которые написали 2268 авторов.

Список источников также довольно широк: в него входят как платформы блогов, так и социальные сети: Livejournal.com, Vkontakte.ru, Liveinternet и, в меньшей степени, Facebook и Twitter.

На первом этапе формируются данные по ключевым словам. Перед исследованием мы сформулировали два ключевых вопроса, на которые было интересно получить ответы:

1. Можно ли хотя бы приблизительно предсказывать результаты выборов, собирая и анализируя результаты упоминаний кандидатов в социальных сетях?

2. Можно ли выявить лидирующих блогеров, которые влияют на общество, и, если да, то каких кандидатов они обсуждают преимущественно?

Ниже представлен фрагмент программного кода автоматизированного сбора данных на языке Python (см. листинг 1). Программа структурирует полученную информацию и сохраняет следующие поля:

- Заголовок комментария
- Ссылка на комментарий
- Текст комментария
- Автор комментария
- Блог, где размещен комментарий
- Дата публикации комментария
- Дата сохранения комментария, поста в системе (базе данных)

```
for site in set(sites):
    item = UkrPresidentItem()
    item['comment_text'] = unicode(" ".join(site.xpath('div/div/div').extract()))
    if u'Добкин' or u'Добкін' or u'Допу' or u'Допа' in item['comment_text']:
        item['Target'] = unicode('Dobkin')
        item['Label'] = unicode(" ".join(site.xpath('h3/a').extract()))  #comment_title
        item['Source'] = (" ".join(site.xpath('ul/li[2]/a/@href').extract())).split(" ", 1)[0]
#comment_author
        item['Type'] = "Directed"
        item['Weight'] = "1"
        item['comment_blog'] = " ".join(site.xpath('ul/li[3]/text()').extract())
        item['comment_url'] = " ".join(site.xpath('h3/a/@href').extract()) #h3/a/@href
        item['comment_dateposted'] = " ".join(site.xpath('ul/li/text()').extract())
```

```
item['comment_datesaved'] = datetime.now(tz=None)
items.append(item)
```

Листинг 1. – Фрагмент программы: формирование графа при сборе данных о кандидатах

Для формирования графа в программном коде используются служебные поля [4]:

- Label – подпись к узлу
- Source – «источник», автор комментария или поста
- Target – «цель», кандидат, которого упомянул в тексте автор комментария
- Weight – вес ребра
- Type – тип графа, значение Directed делает граф ориентированным

Данные сохраняются в no-SQL базу данных MongoDB [7]. Результат можно просматривать, например, в специальной программе Robomongo [8].

На втором этапе производится структурный анализ получившегося графа.

Анализируя популярность кандидатов в блогосфере стоит рассмотреть четыре визуализации одного и того же графа, отражающих структурные особенности сетевых взаимодействий пользователей:

1. Взвешенный граф популярности кандидатов.
2. Взвешенный граф активности комментаторов и блогеров.
3. Распределенный граф популярности кандидатов.
4. Распределенный граф активности комментаторов и блогеров.

Рассмотрим сначала сводный взвешенный граф популярности кандидатов (см. рис. 1). В результате обработки постов и комментариев видно, что представленность кандидатов в блогосфере распределена следующим образом: абсолютным лидером по числу упоминаний является Тимошенко. Следующими топовыми кандидатами являются Порошенко и Ярош.

Больший узел (кружок) означает, что данный кандидат чаще других упоминался в комментариях кандидатов, стрелки от узла указывают, кто именно его упоминал [3,4]. Узлы ранжированы по цвету и размеру: от голубого до красного, голубой цвет узла указывает на то, что это рядовой комментатор, желтые, оранжевые и красные узлы выделяют кандидатов в президенты.

При этом Порошенко, которому социологические опросы предсказывали победу на президентских выборах был не очень популярен в блогосфере в первые две недели предвыборной кампании. Возможная причина заключается в том, что Порошенко не делал

громких заявлений в масс-медиа и не приходил на топовые политические ток-шоу, такие как Шустер-LIVE и «Свобода слова». Однако он смог набрать вес и по итогам двух месяцев вышел на второе место.

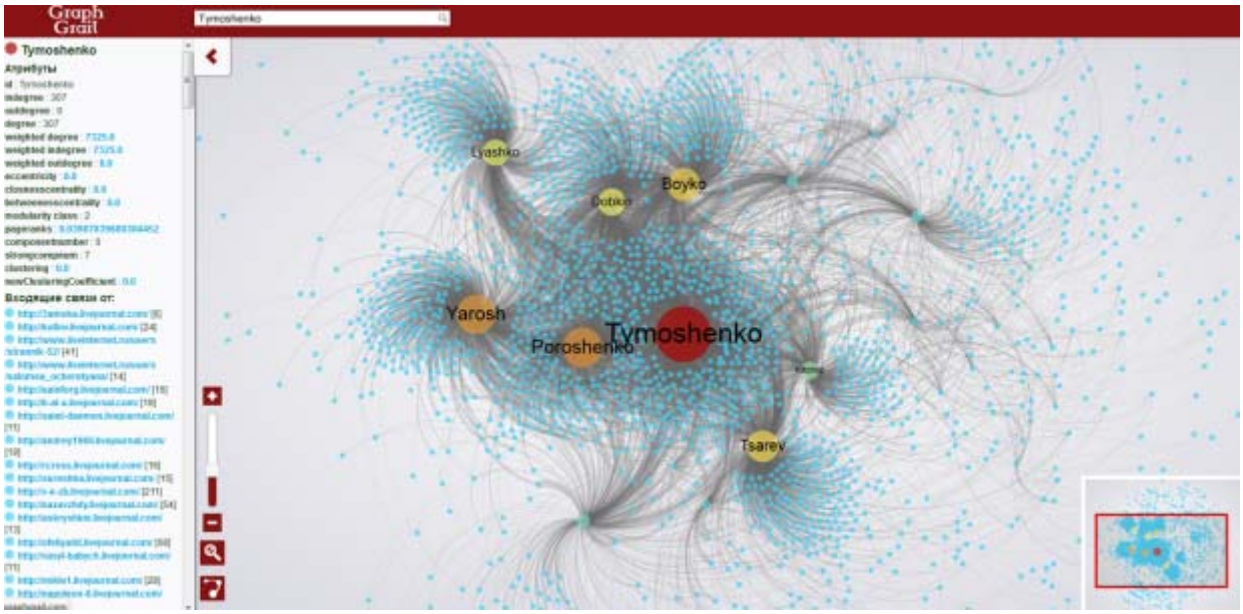

Рис. 1. – Граф популярности кандидатов в президенты

Полная картина лидеров упоминаний выглядит следующим образом (см. рис.2):

Nodes	Id	Weighted In-Degree
	Tymoshenko	7 325
	Poroshenko	4 988
	Yarosh	4 862
	Boyko	3 999
	Tsarev	3 816
	Dobkin	3 207
	Lyashko	3 019
	Klitchko	1 824
	Tyagnibok	1 160
	Gricenko	784
	Tigipko	763
	Simonenko	413
	Bogomolets	355
	Rabinovich	247
	Korolevskaya	156
	Kuzmin	112
	Klimenko	76
	Malomuzh	70
	Konovaluk	64
	Tsushko	56
	Grinenko	41
	Shkiryak	36
	Kuibida	36
	Saranov	30

Рис. 2. – Распределение количества упоминаний кандидатов

Итак, теперь у нас достаточно данных, чтобы подтвердить первую гипотезу, а именно: исследование в целом позволяет вести мониторинг предвыборной президентской кампании на Украине.

Два из трех кандидатов, получивших в результате наибольшее количество голосов, и в блогосфере оказались наиболее обсуждаемыми кандидатами, то есть существует положительная корреляция между анализом экспертов и политологов и мнением масс. В целом же задача предсказать исход выборов в исследовании не ставилась. Отметим, что в топ по упоминаниям также попали Царев и Ярош, их процент на выборах ничтожно мал, но своими действиями и высказываниями они оказывали значительное воздействие на мнения сторонников или противников того или иного сценария развития страны. И это воздействие выражается в активном обсуждении их позиций и концептов, которые они излагают.

Рассмотрим теперь взвешенный граф активности комментаторов (см. рис. 3): размер и цвет узлов указывает на количество постов и комментариев с упоминанием того

или иного кандидата, которые данный пользователь оставил. Чем больше и краснее узел (кружок), тем чаще данный автор упоминал кандидатов, стрелки от узла указывают, кого именно он упоминал.

Еще одно важное наблюдение: на графе визуально можно выделить кластеры авторов, которые обсуждают одних и тех же кандидатов. Первый большой кластер включает в себя Тимошенко и Порошенко. Голубые узлы авторов, которые находятся близко вокруг узлов кандидатов - это те блогеры, которые в своих публикациях упоминают преимущественно только этих двух кандидатов. Данный факт можно интерпретировать как то, что эти авторы схожи в некоторой степени в своих политических предпочтениях. А вот Ляшко и Царев находятся в противоположных «углах» графа - их электорат не пересекается совсем. Итак, кластеры помогают оценить схожесть, пересечение интересов аудитории, или наоборот, идентифицируют группы людей, которые имеют противоположные взгляды [9].

Рис. 3. – Граф активности комментаторов и блогеров

На графе активности комментаторов и блогеров можно выделить лишь нескольких лидеров, которые пишут много и часто:

- http://www.v-n-zb.livejournal.com/
- http://www.mikle1.livejournal.com/
- http://www.andriy-lopata.livejournal.com/

На этих активных пользователей стоит обратить особое внимание, как самим кандидатам, так и их пресс-службам, так эти пользователи во многом формируют

общественное мнение среди интернет-пользователей, интересующихся текущей политической ситуацией на Украине.

На примере Тимошенко рассмотрим параметры узлов (см. рис.4). Справа в информационной панели отображаются служебные данные по узлу, в том числе параметр Взвешенная входящая степень 7325 указывает, что фамилия «Тимошенко» встречалась в комментариях и постах 7325 раз (перепосты и цитирования учитываются).

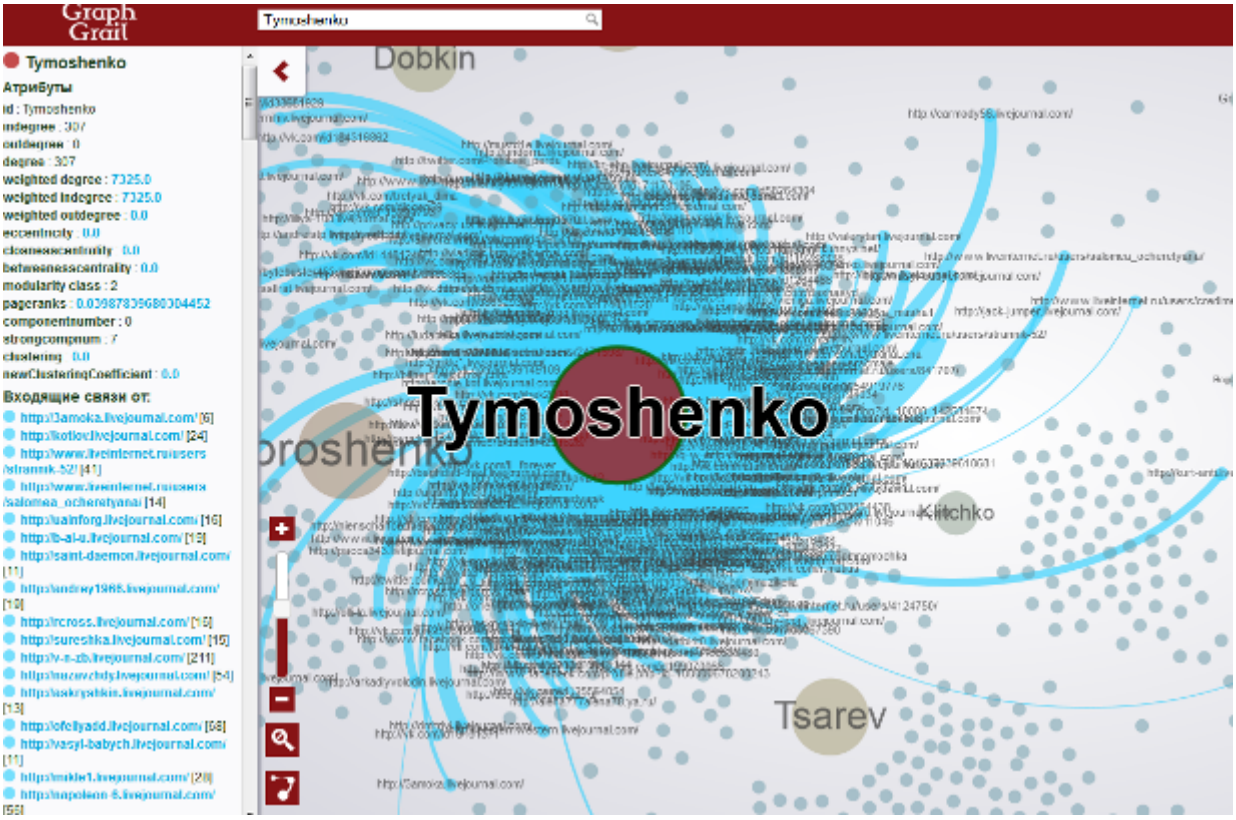

Рис. 4. – Блогеры, упоминавшие одного из кандидатов

Ниже расположен список авторов, которые упоминали Тимошенко, а также количество упоминаний в скобках для каждого пользователя. Например, пользователь v-n-zb.livejournal.com за все время сбора данных упомянул Тимошенко 211 раз (о чем выше уже сказано), а щелкнув по имени автора можно снова перейти к просмотру его параметров.

Несмотря на то, что Кличко свою кандидатуру снял, его продолжают упоминать, однако это может быть связано с его популярностью как боксера.

Еще одним интересным вопросом является распределение упоминаний различных кандидатов по блогерам: какие из блогеров упоминают всех кандидатов в целом равномерно и одинаковое количество раз, а какие сосредоточены в своих постах лишь на

одном-двух кандидатах? Ответом на данный вопрос может быть распределенный граф активности авторов, приведенный на рис.5.

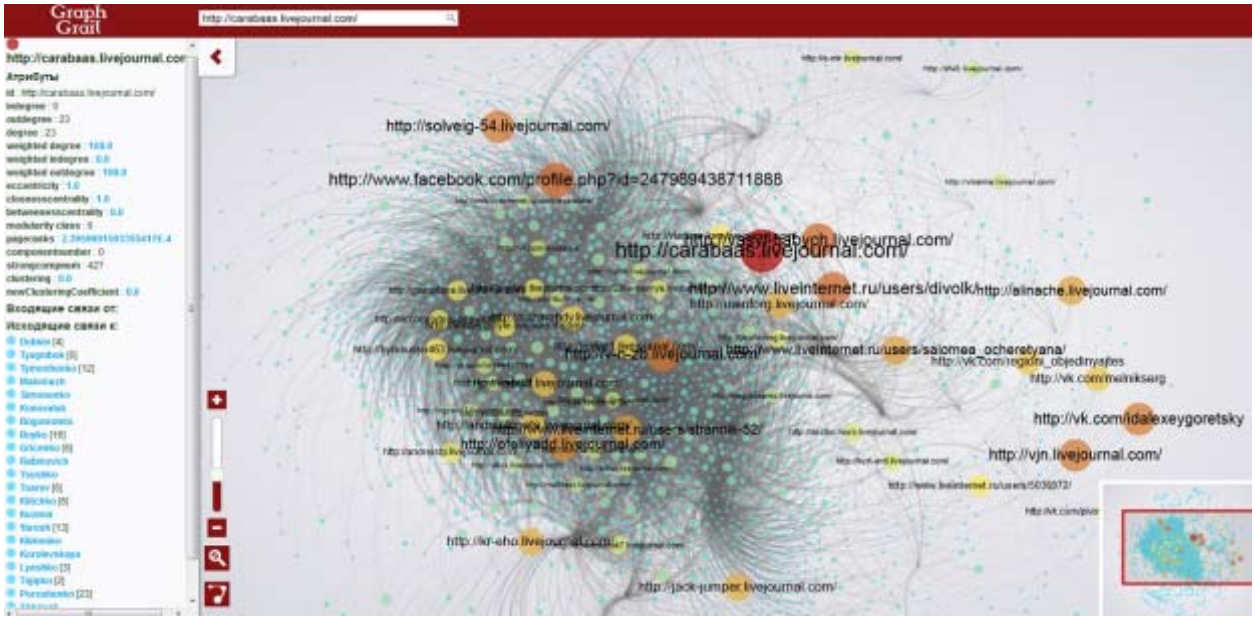

Рис. 5. – Распределённый граф активности авторов

Чем более равномерно и полно охватывает автор в своих постах кандидатов, тем больше его узел на данном графе. Стоит отметить, что авторы-лидеры по упоминаниям на этом графе совсем другие, чем на графе активности блогеров выше. На этом графе узел автора, который упомянул 10 кандидатов в своих публикациях будет по размеру в 2 раза больше, чем другой автор, который упомянул только 5 кандидатов. При этом по абсолютному количеству второй автор может быть впереди, и, таким образом, данная характеристика показывает кругозор того или иного автора, спектр его политических интересов [10]. Например, блогер carabaas.livejournal.com по абсолютному числу упоминаний находится на 23м месте с 108, но по охвату он в топе. Эту информацию можно использовать с тем, чтобы эффективно и точно выделять тех блогеров, на которых можно влиять с целью распространять нужные мнения по сети.

Итак, теперь мы можем подтвердить и вторую гипотезу: выявлять топовых блогеров, которые влияют на общество можно, и, более того, можно определять каких кандидатов они обсуждают преимущественно, каков спектр их охвата.

В результате проведенного исследования и мониторинга блогосферы удалось подтвердить обе гипотезы относительно статистических данных о кандидатах. Исследование показало, какие кандидаты имеют более высокий шанс победить на выборах, а также на каких площадках и кто именно преимущественно их обсуждает. Эта

информация потенциально может быть полезна в любых выборных кампаниях тем кандидатам, которые хотят получить электоральные преимущества и воздействовать на свою аудиторию.

Реализации технологии мониторинга агитационных действий с помощью разработанной методики и с использованием описанного алгоритма и теории графов будут полезны на разных этапах мониторинга социальных сетей и избирательного процесса – как во время избирательных кампаний, так и в периоды между ними. Также возможно применение системы сбора данных и формирования графа в любых сферах деятельности, где структура может быть представлена в виде графа с четко выраженными узлами и связями между ними.

Литература

1. Розин М.Д., Свечкарев В.П., Конторович С.Д., Литвинов С.В., Носко В.И. Проблемы мониторинга социальных сетей как площадки социальной коммуникации рунета // Научная мысль Кавказа. Междисциплинарные и специальные исследования, 2011, №2. С.65-77.

2. Розин М.Д., Свечкарев В.П., Конторович С.Д., Литвинов С.В., Носко В.И. Исследование социальных сетей как площадки социальной коммуникации рунета, используемой в целях предвыборной агитации // Инженерный вестник Дона, 2011, №1. URL: ivdon.ru/ magazine/archive/n1y2011/397

3. Конторович С.Д., Литвинов С.В., Носко В.И. Методика мониторинга и моделирования структуры политически активного сегмента социальных сетей // Инженерный вестник Дона, 2011, №4 URL: ivdon.ru/ru/magazine/archive/n4y2011/642

4. Носко В.И. Система автоматизированного построения графа социальной сети // Инженерный вестник Дона, 2012, №4. URL: ivdon.ru/magazine/archive/n4p2y2012/1428

5. Newman, Mark E.J. "The structure and function of complex networks." SIAM review 45, no. 2 (2003): pp.167-256.

6. Bird Steven. Natural Language Processing with Python. – O'Reilly Media Inc, 2009. – 482 c.

7. MongoDB is an open-source document database, and the leading NoSQL database. Written in C++. URL: mongodb.org

8. Robomongo 0.8.4. Shell-centric cross-platform mongodb management tool. URL: robomongo.org

9. Губанов Д.А. Социальные сети: модели информационного влияния, управления и противоборства / Под ред. чл.-кор. РАН Д.А. Новикова / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили. – М.: Изд-во физ.-мат. лит., 2010. – 228 с.

10. Заруцкий, С.А. Система выбора и настройки метода агрегирования как элемент инструментария СППР в сфере региональной политики // Управление экономическими системами: электронный научный журнал, Издательство: Кисловодский институт экономики и права (Кисловодск), 2013, №11. URL: uecs.ru/marketing/item/2544-2013-11-20-05-54-16

References

1. Rozin M.D., Svechkarev V.P., Kontorovich S.D., Litvinov S.V., Nosko V.I. Problemy monitoringa sotsial'nykh setey kak ploshchadki sotsial'noy kommunikatsii runeta. Nauchnaya mysl' Kavkaza. Mezhdistsiplinarnye i spetsial'nye issledovaniya [Social network monitoring challenges for social media communications in Runet as a platform], 2011, №2. pp.65-77.

2. Rozin M.D., Svechkarev V.P., Kontorovich S.D., Litvinov S.V., Nosko V.I. Inženernyj vestnik Dona (Rus), 2011, №1. URL: ivdon.ru/magazine/archive/n1y2011/397

3. Kontorovich S.D., Litvinov S.V., Nosko V.I. Inženernyj vestnik Dona (Rus), 2011, №4 URL: ivdon.ru/ru/magazine/archive/n4y2011/642

4. Nosko V.I. Inženernyj vestnik Dona (Rus), 2012, №4. URL: ivdon.ru/magazine/archive/n4p2y2012/1428

5. Newman, Mark E.J. "The structure and function of complex networks." *SIAM review* 45, no. 2 (2003): pp.167-256.

6. Bird Steven. Natural Language Processing with Python. – O'Reilly Media Inc, 2009, p. 482

7. Mongodb. MongoDB is an open-source document database, and the leading NoSQL database. Written in C++. URL: mongodb.org

8. Robomongo 0.8.4. Shell-centric cross-platform mongodb management tool. URL: robomongo.org

9. Gubanov D.A. Sotsial'nye seti: modeli informatsionnogo vliyaniya, upravleniya i protivoborstva [Social networks: the models of information influence, control and confrontation]. Pod red. chl.-kor. RAN D.A. Novikova D.A. Gubanov, D.A. Novikov, A.G. Chkhartishvili. – M.: Izd-vo fiz.-mat. lit., 2010. – p. 228

10. Zarutskiy S.A. Upravlenie ekonomicheskimi sistemami: elektronnyy nauchnyy zhurnal, Izdatel'stvo: Kislovodskiy institut ekonomiki i prava (Kislovodsk), 2013, №11. URL: uecs.ru/marketing/item/2544-2013-11-20-05-54-16