Designation of Situation Model in Twitter using Maximal Frequent Sequences

Anna Atyagina[1,2], Yulia Ledeneva[1], and René Arnulfo García-Hernández[1]

[1] Universidad Autónoma del Estado de México,
Unidad Académica Profesional Tianguistenco,
Toluca, Estado de México, Mexico

[2] Omsk F.M. Dostoevsky State University, Omsk,
Russia

atyagina@gmail.com, yledeneva@yahoo.com, renearnulfo@hotmail.com

Abstract. Hashtag is definitely one of the most significant features of Twitter which now is spread all over the social networking services. It can serve different functions, and one of the most important is the designation of situation models. Using the method of Maximal Frequent Sequences we proved that the main idea of all data of one hashtag can be described in two or three phrases as a summary processed using the given method. We demonstrate how the recognition of situation models can be done automatically and fast. Also this method can be used for analysis of hashtag combinations and reconstruction of concepts based on the results of 1-grams and 2-grams, as we presented in detailed example of analysis of the following hashtags: #GalaxyFamily, #RussianMeteor, #Grammys and #10Dec hashtags.

Keywords: Maximal Frequent Sequences, hashtag, Twitter, social media, situation model.

1 Introduction

More than 500 million messages appear on Twitter daily [1]. Although one of Twitter's main features is brevity of messages that are transmitted with it (no more than 140 characters), there is a real problem of organization and systematization of information on the service.

Hashtag is a feature that helps systematize process of communication. It has changed people's interaction and ways to find information within and outside of Twitter. Hashtags are key words or phrases that begin with a # symbol followed by any combination of Twitter permitted nonblank characters. They can occur in any part of tweets. Users simply may add # in front of any word. Hashtags can be used for searching messages, following a certain thread or topic, and therefore can mark a set of tweets focusing on a certain topic described by the hashtag.

At the same time there is still a problem of an immediate understanding of different hashtags, popular or not. Sometimes to get the main meaning of hashtag, user of

Anna Atyagina, Yulia Ledeneva, and René Arnulfo García-Hernández

Twitter has to look through dozens of tweets. It can be almost impossible if someone wants to understand, for example, overall chronic of the day with its main trending hashtags or keywords.

The purposes of our work are to confirm the function of hashtags as indicators of models of situations using Maximal Frequent Sequences (MFSs); also demonstrate how the designation of situation models can be done automatically and fast; and present examples of analysis of hashtags data, based on the results of 1-grams and 2-grams Frequent Sequences (FSs).

2 Related Works

In the past years, methods of data mining were actively applied to the analysis of the information provided on Twitter, including use of hashtags. The researches are devoted to different topics, from automatic identification of diverse sentiment types using hashtags as long as some other Twitter features [2] to the most popular terms or topics of discussion on Twitter for better insight into the collective viewpoint and subjects of interest of the typical Twitter user [3]. The structure of gatekeeping in Twitter by means of statistical analysis of the political hashtags #FreeIran, #FreeVenezuela and #Jan25, each of which reached the top position in Twitter Trending Topics (list of the most tweeted topics ranked by Tweeter proprietary algorithm), is explored in [4].

Some of the researches are mostly practically oriented and are objected to recommend how to use hashtags more effective. Zangerle et al. [5] recommend an approach that aims at creating a more homogeneous set of hashtags for a particular message. At the same time, Ma et al. [6] propose methods to predict the popularity of new hashtags on Twitter by formulating the problem as a classification task. Page [7] presents the investigation on the role of hashtags as means of self-branding.

Although hashtags are a subject of big interest for researches, there are not many researches devoted to the functions of hashtags. There are some new terms as gametags, rhetorical hashtag, bashtag, hijacking or grasstag that serve different purposes [8]. The phenomenon of "hijacking on Twitter" is described in [9], where some certain jumps in polarity are shown, caused by "hijackers" engaged in a particular type of hashtag war. Another new definition of blacktag is described in [10] and connects Twitter studies with cultural and racial studies.

Data mining on Twitter hashtags provides interesting results when it comes to forecasting some situations, human decisions and actions. Big data gives an opportunity to automatize an analysis of users' interaction with each other and referring to the important events. For example, [11] introduced the method of political elections forecasting in different countries based on Twitter messages. Another, more specific research forecasts the results of 2013 elections in Pakistan and 2014 elections in India [12].

Even the behavior of stock exchange can be predicted with Twitter data. Using behaviorist approach to the economy along with data mining methods, [13] demonstrate how collective mood influences on the decision making at the stock

exchange market. Authors discovered that the connection between Twitter users' mood and changes of Dow Jones Index exists in 86.7% of situations processed.

Big events are of particular interest for hashtags' researchers. Twitter's reaction to important sport and cultural events is analyzed in [14]. One of the examples is an analysis of American soccer fans tweets during the World Cup 2014 [15].

Atiagina [16] presents the classification of hashtag functions that we rely on. Hashtags can be used for different purposes, and carry different information. Five functions are included in the classification: designation of situation models in order of compression, inclusion in the overall context/trends, actualization and expression, self-presentation, promotion [16]. In this paper, we consider to the first and the most significant function which is designation of situation models in order of compression.

In the analysis of Twitter messages, it is appropriate to rely on the general principles of text compression mechanism. It allows to compress text in transmission and expand it in perception, without losing the most important and significant information. This is largely possible due to the fact that an adequate understanding between the author and the reader is based on a code system common to the sender and recipient and includes mental scenarios, concepts, categories, rules and strategies. On the basis of cognitive theory of language use, T.A. Van Dijk [17, 18] proposes the following idea: we understand the text only when we understand the situation referred to. Therefore the model of situation is necessary for us as the basis of interpretation of the text. The use of models is explained why listener understands implicit and unclear sections of code: in this case, they activate the corresponding fragments of situational model.

Hashtag serves as a model of situation that compresses the context of tweet but, generally, can be expanded by the reader in perception [16]. If hashtag exists and is widely used, at the same time it would describe the same situation for the most of the people that are using this hashtag. It means that they are likely to use similar words (or other hashtags) in messages with the same hashtag.

In this paper, we propose the method that is based on extraction of MFSs which helps us to show that among the large amount of messages with the same hashtag users are going to use the words with the same semantics and, in general, the main idea of all the messages can be described in one or two phrases as a summary.

3 Proposed Method

An ngram is a sequence of n words. We say that an ngram occurs in a text if these words appear in the text in the same order immediately one after another. We call an ngram frequent (more accurately, β-frequent) if it occurs more than β times in the text, where β is a predefined threshold. Frequent ngrams—we will also call them frequent sequences (FSs)—often bear important semantic meaning: they can be multiword expressions, idioms or otherwise refer to some idea important for the text [19, 20]. An ngram can be a part of another, longer ngram. FSs that are not parts of any other FS are called Maximal Frequent Sequences (MFSs) [21, 22].

Only MFSs should be considered as bearing important meaning, while non-maximal FSs (those that are parts of another FS) should not be considered. Other

Anna Atyagina, Yulia Ledeneva, and René Arnulfo García-Hernández

additional motivation may be cost vs. benefit considerations: there are too many non-maximal FSs while their probability to bear important meaning is lower. In any case, MFSs represent all FSs in a compact way: all FSs can be obtained from all MFSs by bursting each MFS into a set of all its subsequences. FSs can express ideas both important and specific for the document [19].

Using data mining techniques and extraction of MFSs [21, 22] in accordance to the task involves the following steps:

1. Data collection using Twitter API.
2. Data preprocessing. Each Tweet contains a lot of information that needs to be removed or altered for adequate treatment and allocation of MFSs. Thus, functional words and phrases that are present in the interface of Twitter in each message (such as "expand", "answer" etc.) are being automatically removed by the program Format Text. Links to the other resources are substituted by the word "@ liga" and hashtags are substituted by the word @hash. Also links to user names containing the @ symbol are replaced with the word "name".
3. Extraction of MFSs. This is done as developed in [21, 22]. At the first stage, the program converts all data to binary files, which are then used to analyze the data. All documents (which are, in this case, Twitter messages) are loaded into the program so that each message is assigned its own line. Words are automatically highlighted in capital letters, separated by commas. The program automatically scans the text and extracts the words that are repeated often. Depending on what is specified maximum number of words within the sequence of the frequency (the average maximum frequency commonly used sequences containing from two to six words) can produce different results for later analysis.

4 Corpus

For this paper, four popular hashtags were processed in English, Russian or mixed languages. We've postulated models of situations to each of these hashtags that we're going to prove using method of MFS's lately.

The processed hashtags are:

#GalaxyFamily (English, 11133 tweets) — promo hashtag, community of Galaxy phones' users as a family.
#Grammys (English, 9890 tweets) — highlighting the famous musical ceremony, the most relevant artists' names and actions of event in 2014.
#RussianMeteorit (Russian and English, 3952 tweets) — meteorite fall in Chelyabinsk in February 2013.
#10dec (Russian and English, 2456 tweets) — protests against the results of State Duma elections on December, 10th, 2011.

We provide all results for all processed hashtags but more detailed description presented later will be devoted to #GalaxyFamily hashtag. It is a promoted hashtag that, at the same time, serves at least two functions: designation of model of situation and marketing. In this case, the situation itself is created by the company (Samsung) but still is a relevant objective as long as thousands of Twitter users tweeted with this

hashtag on their own, voluntarily. Our #GalaxyFamily corpus consists of 11133 tweets published during 2013 and contained hashtag #GalaxyFamily, mostly in English. Working specifically with this hashtag, we've processed 135448 words and got 30989 frequent sequences.

5 Experiments and Analysis

5.1 Indicating Meaning of Hashtags

Previously described method serves to identify the most popular MFS in the corpus which can help to make a quick summary of the chosen hashtag. In the Table 1, three types of results for each hashtag are provided: the MFSs, the longest FSs and the most frequent 1-gram (or in other words the most frequent words that are met in the corpus).

Generally, these results can demonstrate the main meanings and ideas of each hashtag and help to identify it to those Twitter users who are not aware of the model of situation. Both MFSs and 1-grams demonstrate connection with the model of situation that can be used then in different ways: quick and automatized recognition of models of situations, analysis and linguistic interpretation of users' behavior and opinions, etc. In the next part of the paper, we provide the detailed example of this analysis.

The proposed method is language independent: we see that results in different languages are valid and for the same hashtag they have the same meaning (#RussianMeteor, #10dec).

During our experiment we met a problem of a spam or commercial tweets that use popular hashtags to promote other information. Sometimes, for example, with hashtag #RussianMeteor, the most frequent MFSs can be considered as a spam. In other situations (#10dec, #Grammys) they are not a spam. Comparing these results to the longest MFSs found, we conclude that the most effective way of automatic summarization of hashtags data would be combination of the most frequent MFSs with the longest ones. It is also important because some of the results can be in the language that user doesn't understand. For example, the most frequent result for the hashtag #GalaxyFamily is in Turkish although the main language of the hashtag is English but other results serve to understand the model of situations.

Although MFSs are considered as the most significant, shorter n-grams also can help to explain or to analyze hashtags. For example, each hashtag has its own group of the most frequent 1-grams that can help to reconstruct the overall model of situation and confirm our hypothesis. Frequent 2-grams can be used both to reconstruct the model situation and to find the frequent hashtag combinations.

5.2 Example of Detailed Analysis: #GalaxyFamily Hashtag

5.2.1 Maximal Frequent Sequences

Usage of Maximal Frequent Sequences helps us not only to understand the main idea of the hashtag but also to provide different kind of analysis that can be useful for marketing specialist as well as for linguist.

Anna Atyagina, Yulia Ledeneva, and René Arnulfo García-Hernández

Table 1. The results are the MFSs, the longest FSs and the most frequent 1-gram for hashtags #GalaxyFamily, #RussianMeteor, #Grammys and #10Dec.

The most frequent MFSs	The longest FSs	1-gram FSs
1. #GalaxyFamily [276] retweet uygulamas @liga @foto #galaxyfamily #milletineserikapatilamaz #benceeask #reklam #google [103] welcome to the #galaxyfamily you can learn all about your new phone in our guide to the galaxy at @liga [85] black friday deals on amazon now @liga #toysruskid #dwts thanksgiving #galaxyfamily	[2] i m a owner of the #s and i have to admit it s the best phone i ve had by far wouldn t change for the world thanks #galaxyfamily (30 words) [2] got my galaxy s saturday and i gotta say it was the best decision i made when i d comes to choosing a new cell phone #galaxyfamily(27) [2] to the #galaxyfamily i have had a galaxy phone for years and i have loved em all but the note might just be greatest phone ever (26)	[82] love [38] member [35] @samsungmobileus [32] shit [31] photo [27] fun [25] #apple [24] almost [24] lmao [24] her [24] his [23] apps [21] try [20] something [20] white [20] super [20] definitely [19] #follow [19] hi [19]#cybermondaymadness [19] nothing [19] wit [19] hello [18] @hashmtvstars [18] galaxys
2. #RussianMeteor 4 most frequent are spam not connected with the situation. after them there are the following: [15] rt @bbcbreaking #russianmeteor shower six of the most dramatic videos @liga #meteorit @liga [13] @liga ultimas imagenes del meteorito que cayo en rusia #meteorit @hashrussianmeteor [13] at least people injured in spectacular #russianmeteor shower interior ministry @liga @liga #meteorit	[5] coming up at pm @nasa experts discuss the #russianmeteor learn more about meteor listen to panel live at pm et @liga (21 words) [5] wow rt @wsj the explosion from the #russianmeteor is estimated to be as powerful as Hiroshima bombs @liga (18) [5] rt @nasa #russianmeteor is largest reported,meteor since Tunguska event impact was at utc still being @liga (17)	[11] kak (how) [10] jeto (this) [10] again [9] recorded [8] though [7] her [7] eshhe (again) [7] #meteorites [7] @neiltyson [7] window [7] loud [7] during [6] les [6] kilotons [6] answers

The most frequent MFSs	The longest FSs	1-gram FSs
3. #Grammys [19] see the best part of the #grammys also known as queen latifah marrying a bunch of happy couples via @upworthy @liga [18] fresh from the #grammys did you see what beyonce was wearing barely @liga [16] paul walker death tyrese gibson thanks fans during grammys #paulwalker #tyresegibson #grammys @liga	[2] the best part about watching the 7ashgrammys the day after is you re able to fast forward through the boring stuff which is most of it (26) [2] i m with trent on this one i tuned into the end because i was excited to see some of my fave artists collaborate left unhappy 7ashgrammys (27) [2] love how pharrell pointed to stevie to pick up his cue at the start when he missed it stevie s blind mate pointing won t do much 7ashgrammys (28)	[9] event [9] Ciara [9] JaredLeto [8] ridiculous [8] ni [8] loves [8] literally [8] exclusive [8] avec [8] apparently [7] noticed [7] lame [7] france [7] daftpunk [7] become [7] along [6] wit [6] water [6] themselves
4. #10dec [49] @liga #10dec #russia [40] rt @burmatoff voenkomaty segodnja perevypolnjat normu po prizyvu @liga soberut na bolotnoj vseh uklonistov #10dek #10dec (RUS to ENG: rt @burmatoff today military enlistment offices will exceed a plan @liga by finding all the objectors on Bolotnaya #10dek #10dec) [34] @liga #10dec #10dek #revolution	[3] nezavisimaja ocenka chislennosti akcij #10dec #10dec #feb v #omsk na osnove kart i fotografij @liga (15) (RUS to ENG: independent estimation of number of participans of the action based on maps and photos #10dec #10dec #feb in #omsk) [3] mitingi #10dec #10dec v #omsk ocenka chislennosti i jeffekta ubeditelnaja infografika na karte @liga (14) (RUS to ENG: protests #10dec #10dec in #omsk evaluation of number and effect, infographics on the map @liga (14)	[74] no [32] narod (people) [27] budet (will be) [24] net (no) [24] all [23] ru [21] ochen (very) [21] miting [20] with [20] policija (police) [20] policii (police) [20] est (is) [20] over [20] mitingujushhih (protesters) [19] rossija (Russia) [15] Navalnyj

Anna Atyagina, Yulia Ledeneva, and René Arnulfo García-Hernández

In this part of our paper, we provide detailed analyses of the example of #GalaxyFamily hashtag that was described previously. All the other results of processing are in English (the main language of the #GALAXYFAMILY hashtag). Firstly, we look through the MFSs that appear in our list and give a brief analysis of the results.

The most popular MFSs as a result of processing #GalaxyFamily are a combination of hashtags, links and photos in Turkish that was used 276 times among our corpus:

(1) RETWEET UYGULAMAS @LIGA @FOTO #GALAXYFAMILY #MILLETINESERIKAPATILAMAZ #BENCEEASK #REKLAM #GOOGLE

It has its limits in automatically defining the model of the situation (researcher need to know Turkish). But our method provides more significant results. For example, the second frequent result can be widely used as a short description of this hashtag:

(2) WELCOME TO THE #GALAXYFAMILY YOU CAN LEARN ALL ABOUT YOUR NEW PHONE IN OUR GUIDE TO THE GALAXY AT @LIGA

We met this MFS 103 times in our corpus. It's the longest (with n = 20) and the most informative one. If we look further we can see some other MFSs that have almost the same meaning but consist of different words (examples 3-4):

(3) WELCOME TO THE #GALAXYFAMILY GET STARTED WITH GREAT TIPS, TRICKS, APP, RECOMMENDATIONS AT @LIGA (38 times)

(4) WELCOME TO THE #GALAXYFAMILY, WE'RE HERE TO HELP CHECK OUT OUR GUIDE TO THE GALAXY AT @LIGA (34 times)

Other popular messages in the list of MFS serve to connect product with different emotions (examples 5-9):

(5) WE'RE HAPPY YOU'RE PART OF OUR #GALAXYFAMILY (52 times)

(6) THANKS FOR THE LOVE #GALAXYFAMILY (42 times)

(7) WE LOVE HAVING YOU IN THE #GALAXYFAMILY (40 times)

(8) THANKS FOR BEING PART OF THE #GALAXYFAMILY (38 times)

(9) THANKS FOR BEING SUCH A LOYAL MEMBER OF THE #GALAXYFAMILY (38 times)

So we can conclude that emotions are an essential part of the concept. Hashtag #GalaxyFamily can be described as a marketing hashtag created to greet new clients and to create the image of Galaxy users and company itself as a kind family which is ready to welcome its new members.

Other important parts are some other brands than Galaxy products. Using method of MFSs we are able to identify the following: Amazon, Samsung, James Bond (examples 10-13).

(10) BLACK FRIDAY DEALS ON AMAZON NOW @LIGA #TOYSRUSKID #DWTS, THANKSGIVING #GALAXYFAMILY (85 times)

(11) CHECK OUT THE LATEST #GALAXYFAMILY PRODUCTS FROM @SAMSUNGMOBILEUS @LIGA (32 times)

(12) BOND, ALL JAMES BOND MOVIES ON BLU RAY FOR ONLY ON AMAZON TODAY ONLY @LIGA #DVD #BLURAY #GALAXYFAMILY (26 times)

(13) AMAZONS AFTER CHRISTMAS BLOWOUT SALE ON NOW @LIGA MGM GRAND #THEVOICE BOXING DAY UFC #UFC #THEGIFTER #GALAXYFAMILY (23 times)

Therefore we can conclude that using the proposed method, the main purpose of hashtag can be described in one or two MFSs. Although for better automatized results, as we've already mentioned, it is recommended to choose two or three MFSs out of issue just to elude the influence of spam information and one or two of the longest MFSs as well (see Table 1).

5.2.2 1-gram Frequent Sequences

1-grams can also be useful in understanding overall hashtag meaning. The frequency of the words can be useful as a basis of concept reconstruction both for marketing and scientific purposes. In this part we talk briefly about the analysis of results.

We divided the processed data into some groups with the exact significance. We didn't consider so called "stop words" which are pronounces, prepositions, articles etc. As a result, we have at least 6 significant groups – slots that are directly connected with the analyzed hashtag. The frequency of use of a particular word form is given in brackets.

Technology: photo (31), apps (23), smartphone (17), smartphones (15), #smartphone (12), software (14), data (12), pics (12), tech (11), videos (11), technology (9), droid (9), #photography (8), androids (7), android (5), #phone (6), content (4), etc.

Market: sold (11), business (11), money (10), service (9), credit (9), contract (8), purchase (8), prices (7), selling (7), etc.

Quality: small (11), cute (11), hot (11), quality (9), quick (9), nuevo (6), exclusive (6), special (6), #new (4), etc.

Emotions: love (82), shit (32), super (20), fucking (12), epic (8), gorgeous (6), etc.

Personalities: member (38), her (24), his (24), girl (18), anyone (17), myself (16), #together (16), mine (13), wife (12), families (11), owners (10), everybody (9), kids (8), somebody (6), men (6), etc.

Brands: SamsungMobileUs (35), #Apple (25), Galaxys (18), HTC (13), Facebook (12), Windows (11), Blackberry (9), #Verizon (9), #Galaxycampus (8), #iPhones (6), #Amazon (8), Verizon (5), etc.

Overall view on the most frequent words in the corpus helps us to reconstruct the model of situation and the image of the product. Predictably there are a lot of words connected with market and technology, emotions or qualities. At the same time there are surprisingly many words connected with Galaxy competitors. As long as tweet producing is uncontrolled, products of other companies can be promoted using competitor's hashtag which is highly undesirable and can be identified using the method of frequent sequences.

5.2.3 Combinations of Hashtags

Another significant feature is combinations of hashtags when different hashtags go together in one tweet.

The way hashtags are used in the phrase can also be important and significant for a research. As long as hashtags are concepts or models of situation, they have more significance than the other words that are frequently met in corpus. The combinations of different hashtags can demonstrate the attitude to the situation, show similar situations, etc. One of the ways of extracting combinations of hashtags is analysis of 2-grams MFSs. Here we provide some of the brief results.

According to data processing based on 2-grams, hashtags mostly used with #GalaxyFamily are the following (the frequency of use of a particular word form is given in brackets):

#BlackFriday (54)
#Note (44)
#Android (36)
#Apple (25)
#CyberMondayMadness (19)
#Samsung (18)
#TeamGalaxy (17)
#TeamIPhone (17)
#Christmas (17)
#Smartphone (12), etc.

The information on the word order can be provided too, if needed.

Among these results we can see references to Galaxy itself (Samsung, TeamGalaxy), competitors (Apple, TeamIPhone, Android), a situation when Galaxy products can be useful, probably as a gift (Christmas), an appropriate time to buy this products (Cyber Monday, Black Friday) and just key words for the hashtag as Smartphone. These combinations can be used, for example, by marketing specialist as long as they indicate the most significant features of a product or the main competitors as considered by Twitter users.

At the same time there are some hashtags that are used by people to promote their messages with another already popular hashtag #GalaxyFamily:

#RT (22 times)
#Follow (19 times)
#FollowMe (14 times)
#NowPlaying (13 times).

All of these hashtags can be used just to promote the message or the user itself: probably, the theme of the messages is not associated with Galaxy products neither with music, etc.

6 Conclusions

In this paper, we described the method of Maximal Frequent Sequences as applied to Twitter data processing. Using it we were able to restore the models of the situation described with four different hashtags: #RussianMeteor, #Grammys, #GalaxyFamily, #10dec.

We confirm the function of hashtags as an indicator of models of situations using the described method. The results demonstrate us the main meanings and ideas of

each hashtag. Both MFSs and 1-grams show connection with the model of situation and can be used for quick and automatized recognition of these models. Our method is language independent: the results in different languages are valid, and for the same hashtag they have the same meaning (#RussianMeteor, #10dec).

During our experiment we met a problem of a spam or commercial tweets that use popular hashtags to promote other information. Our experiment provided a possible way to avoid its influence on final results. Comparing these results to the longest found MFSs we can conclude that the most effective way of automatic summarization of hashtags data would be combination of the most frequent MFSs with the longest ones.

Although MFSs are considered as the most significant, shorter n-grams also help to explain and to analyze hashtags. Using example of #GalaxyFamily hashtag we demonstrate that the most frequent 1-grams can help to reconstruct the overall model of situation and also confirm our assumption of hashtag as a model of situation. Frequent 2-grams can be used both to reconstruct the model situation and to find the frequent hashtag combinations. The results we got for other hashtags (such as the political one, #10dec) show that the same method can be applied to non-commercial hashtags as well.

We consider the proposed method as valid for a rapid designation of the model of situation with any hashtag. Also this method is valid to reconstruct the overall concept of the situation and to determine the keywords or opinions connected with it for marketing or other purposes. As a future work, we use syntactic ngrams for extraction stage to the designation of situation model [23].

References

1. Krikorian, R.: New Tweets per second record, and how! Available at https://blog.twitter.com/2013/new-tweets-per-second-record-and-how (2013)
2. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using Twitter hashtags and smileys. In: COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, pp. 241–249 (2010)
3. Cheong, M.: 'What are you tweeting about?': A survey of Trending Topics within Twitter. Clayton School of Information Technology, Australia.
4. Bastos, M.T., Raimundo, R. L. G., Travitzki, R.: Gatekeeping Twitter: message diffusion in political hashtags, Media, Culture & Society, vol. 35 (2), pp. 260–270 (2013)
5. Zangerle, E., Gassler, W., Specht, G.: On the impact of text similarity functions on hashtagrecommendations in microblogging environments. Springer-Verlag, available at: http://link.springer.com/article/10.1007%2Fs13278-013-0108-x (2013)
6. Ma, Z., Sun, A., Cong, G.: On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology, vol. 64 (7), pp. 1399–1410 (2013)
7. Page, R.: The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. Discourse and Communication, vol. 6 (2), pp. 181–201 (2012)

8. Zimmer, B.: Hastag, You're It. Spectrum.IEEE.org. North American, vol. 50 (4), p. 24 (2013)

9. Hadgu, A.T., Garimella, K., Weber, I.: Political hashtag hijacking in the U.S. In: WWW '13 Companion Proceedings of the 22nd international conference on World Wide Web companion, pp. 55–56 (2013)

10. Sharma, S.: Black Twitter? Racial Hashtags, Networks and Contagion. New formations: a journal of culture/theory/politics, vol. 78, pp. 46–64 (2013)

11. Tsakalidis, A., Papadopoulos, S., Cristea, A.: Predicting elections for multiple countries using Twitter and polls. IEEE Intelligent Systems, vol. 30, pp. 10–17 (2015)

12. Kagan, V., Stevens, A., Subrahmanian, V.S.: Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. IEEE Intelligent Systems, vol. 30, pp. 2–5 (2015)

13. Bollen, J., Mao, H., Zeng, X.: Twitter Mood Predicts the Stock Market. Journal of Computational Science, no. 2, pp. 1–8 (2011)

14. Popescu, A.-M., Pennacchiott, M.: Dancing with the Stars, NBA Games, Politics: An Exploration of Twitter Users' Response to Events. In: The Fifth International AAAI Conference on Weblogs and Social Media (2011)

15. Yu, Y., Wang, X.: World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. Computer in Human Behavior, vol. 48, pp. 392–400 (2015)

16. Atiagina, A.: Twitter as a new discoursive practice [Twitter kak novaja discoursivnaja practika (in Russian)]. Ph.D. Thesis, Omsk State University, Omsk (2014)

17. Dijk Van, T.A., Kinch, B.: Strategies and understanding of connected text [Strategii ponimaniia sviaznogo teksta (in Russian)]. In: New in foreign linguistics [Novoe v zarubezhnoi lingvistike (in Russian)], vol. 23, pp. 153–211 (1988)

18. Dijk Van, T.A.: Language, cognition, and communication [Iazyk, poznanie, kommunikatsiia (in Russian)], Moscow (1989)

19. Ledeneva, Y., Gelbukh, A., García-Hernández, R.A.: Terms Derived from Frequent Sequences for Extractive Text Summarization. LNCS 4919, pp. 593–604, Springer-Verlag (2008)

20. Ledeneva, Y., García-Hernández, R., Gelbukh, A.: Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization. LNCS 8404, pp. 466–480, DOI 10.1007/978-3-642-54903-8_39 (2014)

21. Garcia-Hernandez, R.A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J.A.: A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text. LNCS 3287, pp. 478–486 (2004)

22. García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. LNCS 3878, pp. 514–523 (2006)

23. Sidorov, G.: N-gramas sintácticos no-continuos. Polibits, vol. 48, pp. 67–75 (2013)