ИЗВЛЕЧЕНИЕ МАКСИМАЛЬНЫХ ЧАСТОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДЛЯ ОБОЗНАЧЕНИЯ МОДЕЛИ СИТУАЦИИ В ТВИТТЕРЕ

Атягина А. (atiagina@gmail.com)[1, 2],
Леденева Ю. (yledeneva@yahoo.com)[1],
Гарсиа-Эрнандез Р. А. (renearnulfo@hotmail.com)[1],
Иссерс О. (isserso@mail.ru)[2],
Тапиа Фабела Х. Л. (joseluis.fabela@gmail.com)[1]

[1]Автономный Университет штата Мехико, Толука, Мексика
[2]Омский Государственный Университет
им. Ф. М. Достоевского, Омск, Россия

Ключевые слова: максимальные частотные последовательности, хэштег, Твиттер, модель ситуации, обработка информации

EXTRACTION OF MAXIMAL FREQUENT SEQUENCES FOR IDENTIFICATION OF SITUATION MODEL IN TWITTER

Anna Atiagina (atiagina@gmail.com)[1, 2],
Yulia Ledeneva (yledeneva@yahoo.com)[1],
Rene Arnulfo García-Hernández (renearnulfo@hotmail.com)[1],
Oksana Issers (isserso@mail.ru)[2],
Jose Luis Tapia Fabela (joseluis.fabela@gmail.com)[1]

[1]Autonomous University of the State of Mexico, Toluca, Mexico
[2]Omsk F. M. Dostoevsky State University, Omsk, Russia

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 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 #GalaxyFamily hashtag.

Key words: Maximal Frequent Sequences, hashtag, Twitter, social media, situation model

1. Introduction

One of main features of Twitter 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, because it is more than 500 million messages that appear on Twitter daily [Krikorian 2013].

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 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 key words.

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 [Atiagina 2014]. In this paper, we give consideration to the first and the most significant function which is designation of situation models in order of compression.

In this paper, we propose the method that is based on extraction of MFSs which helps 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. Automatic Text Summarization using MFSs are described in [Ledeneva 2008, Ledeneva 2014].

The purposes of our work are:

1) To confirm the function of hashtags as indicators of models of situations using Maximal Frequent Sequences (MFSs).
2) To demonstrate how the recognition of situation models can be done automatically and fast.
3) To demonstrate the examples of analysis of hashtags data based on the results of 1-grams and 2-grams Frequent Sequences (FSs).

2. Proposed method

An ngram is a sequence of n words. 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. An ngram can be a part of another, longer ngram. All ngrams contained in an FS are also FSs.

FSs that are not parts of any other FS are called Maximal Frequent Sequences (MFSs) [García-Hernández 2004; García-Hernández 2006]. For example, in the following text

(4) … *Mona Lisa is the most beautiful picture of Leonardo da Vinci* …

(5) … *Eiffel tower is the most beautiful tower* …

(6) … *St. Petersburg is the most beautiful city of Russia* …

(7) … *The most beautiful church is not located in Europe* …

the only MFS with $\beta = 3$ is *is the most beautiful*, while the only MFS $\beta = 4$ is *the most beautiful* (it is not an MFS with $\beta = 3$ since it is not maximal with this β). As this example shows, the set of MFSs with different thresholds do not have to contain one another.

The extraction of MFSs 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 extraction 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. 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 described in [García-Hernández 2006].

3. Experiments and analysis

3.1. Indicating the meaning of hashtag

Previously proposed 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 the hashtag #GalaxyFamily 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).

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.

During our experiment we have met a problem of a spam or commercial tweets that use popular hashtags to promote other information. Sometimes, for example, with other hashtags, the most frequent MFSs can be considered as a spam. But comparing these results with the longest MFSs found 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.

Table 1. Different types of results for the hashtag #GalaxyFamily

Hashtag	The most frequent MFSs	The longest FSs	1-gram FSs
#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

It is also important because some of the results can be in the language that the user doesn't understand. For example, the most frequent result for 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 also confirm our hypothesis. Frequent 2-grams can be used both to reconstruct the model situation and to find the frequent hashtag combinations.

3.2. Example of detailed analysis of hashtag: #GalaxyFamily

For this paper, we have use the hashtag in English:

#GalaxyFamily (English, 11,133 tweets) is used as promo hashtag in the community of Galaxy phones' users as a family.

This hashtag, 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 11,133 tweets published during 2013 and contained hashtag #GalaxyFamily, mostly in English. Working specifically with this hashtag, we have processed 135,448 words and got 30,989 frequent sequences.

3.2.1. Maximal Frequent Sequences

Usage of Maximal Frequent Sequences helps 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.

In this section, we present opportunities that provide from Maximal Frequent Sequences to the researcher of analyzing hashtag. We use 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*

For example, the second frequent result can be 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 is the longest (with n = 20) and the most informative. If we look further we can see some other MFSs that have almost the same meaning but consist of some other 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 different brands than Galaxy products. Using method of MFSs we are able to identify the following brands: 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 @SAMSUNG-MOBILEUS @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).

3.2.2. 1-gram Frequent Sequences

1-grams can also be useful in understanding of overall hashtag meaning. Also the frequency of the words can be useful as a basis of concept reconstruction both for marketing and scientific purposes. In this section, we present the analysis of results.

We divided the processed data into some groups with the exact significance. We haven’t considered 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.

1) 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.
2) Market: sold (11), business (11), money (10), service (9), credit (9), contract (8), purchase (8), prices (7), selling (7) etc.
3) Quality: small (11), cute (11), hot (11), quality (9), quick (9), nuevo (6), exclusive (6), special (6), #new (4) etc.
4) Emotions: love (82), shit (32), super (20), fucking (12), epic (8), gorgeous (6) etc.
5) 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.
6) 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 method of frequent sequences.

3.2.3. 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 researcher. As long as hashtags are concepts or models of situation, they have more significance than the other words that are frequently met in corpus. Combinations of different hashtags can demonstrate the attitude to the situation, show similar situations, etc. One of the ways of extracting the combinations hastags 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. If needed the information on the word order can be provided too.

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 the 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.

4. Conclusion

In this article, we proposed the method of Maximal Frequent Sequences to apply to Twitter data processing. Using it we were able to restore situation models described for the hashtag #GalaxyFamily.

We confirm that the function of hashtags is an indicator of situation models using the described method. The results demonstrate 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 models of situations.

During our experiment we have met a problem of a spam or commercial tweets that use popular hashtags to promote other information. But 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 can help to explain or to analyze hashtags. Using example of #GalaxyFamily hashtag we demonstrate that the most frequent 1-grams can help to reconstruct the overall situation model and also confirm the hypothesis 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 proposed method can be applied to a rapid designation of the situation model with any hashtag. Also this method is valid to reconstruct the overall concept of the situation and to determine the key words or opinions connected with it for marketing or other purposes. As a future work, we use sintactic ngrams to the designation of situation model [Sidorov 2013].

References

1. *Atiagina A.* Twitter as a new discoursive practice [Twitter kak novaja discoursivnaja practika]. Ph. D. Thesis (Автореферат на соискание степени канд. фил. наук). Omsk, 2014.
2. *García-Hernández R. A., Martínez-Trinidad J. F., Carrasco-Ochoa J. A.* (2004), A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text, CIARP'2004, LNCS, Vol. 3287, pp. 478–486.

3. *García-Hernández R. A., Martínez-Trinidad J. F., Carrasco-Ochoa J. A.* (2006) A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 514–523. Springer, Heidelberg
4. *Krikorian R.* (2013), New Tweets per second record, and how! Available at https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
5. *Ledeneva Y., Gelbukh A., García-Hernández R. A.* (2008). Terms Derived from Frequent Sequences for Extractive Text Summarization. LNCS 4919, pp. 593–604, Springer-Verlag, 2008.
6. *Yulia Ledeneva, René García-Hernández, Alexander Gelbukh.* Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization. Springer-Verlag LNCS vol. 8404, pp. 466–480, 2014. DOI 10.1007/978-3-642-54903-8_39.
7. *Sidorov G.* (2013). N-gramas sintácticos no-continuos. Polibits, 48, pp. 67–75.