Fuzziness in Information Extracted from Social Media Keywords

Shahnaz N. Shahbazova[1(✉)] and Sabina Shahbazzade[2]

[1] Azerbaijan Technical University, Baku 1073, Azerbaijan
shahbazova@gmail.com, shahbazova@berkeley.edu
[2] The George Washington University, Washington DC 20052, USA
shahbazzade@gwu.edu

Abstract. Social media becomes a part of our lives. People use different form of it to express their opinions on variety of ideas, events and facts. Twitter, as an example of such media, is commonly used to post short messages – tweets – related to variety of subjects.

The paper proposes an application of fuzzy-based methodologies to process tweets, and to interpret information extracted from those tweets. We state that the obtained knowledge is fully explored and better comprehend when fuzziness is used. In particular, we analyze hashtags and keywords to extract useful knowledge. We look at the popularity of hashtags and changes of their popularity over time. Further, we process hashtags and keywords to build fuzzy signatures representing concepts associated with tweets.

1 Introduction

In this paper, we describe a simple methodology of analyzing a set of Twitter hashtags. The main focus of the method is investigation of temporal aspects of this data. We are interested in analysis of hashtags from the point of view of their dynamics. We identify groups of hashtags that exhibit similar temporal patterns, look at their linguistic descriptions, and recognize hashtags that are the most representative of these groups, as well as hashtags that do not fit the groups very well. The presented and used method is based on a fuzzy clustering process. Once the clusters are created we examine obtained clusters in detail and draw multiple conclusions regarding variations of hashtags over time. Further, we construct fuzzy signatures of political parties based on analysis of hashtags and noun-phrases extracted from a set of tweets associated with US elections of 2012. We use obtained signatures to analyze similarities between issues and opinions important for each party.

The paper is divided into the following sections. We start with a brief introduction to the concepts of tweets, fuzzy sets, and fuzzy clustering, Sect. 2. Section 3 provides a brief description of used data: hashtags collected from *Hashtagify.me*; and tweets associated with US elections 2012. Further, we focus on analysis of hashtags – we provide some examples of hashtag popularity; describe a data pre-processing leading to representation of popularity changes. Section 4 contains discussion and conclusion.

S. O. Kuznetsov et al. (Eds.): RCAI 2018, CCIS 934, pp. 138–144, 2018.
https://doi.org/10.1007/978-3-030-00617-4_13

2 Hashtags and Clustering

2.1 Tweet and Hashtags

Twitter – one of the most popular online message systems – allows its users to post short messages called tweets. According to *dictionary.com* [14], the definition of a tweet is:

> *"... 2. (Digital Technology) a very short message posted on the Twitter website: the message may include text, keywords, mentions of specific users, links to websites, and links to images or videos on a website."*

The users posting these messages include special words in the text. These words – hashtags – are easily recognizable and play the role of "connectors" between messages. An informal definition of hashtags – obtained from Wikipedia [15] – is as follows:

> *"A **hashtag** is a type of label or metadata tag used on social network and microblogging services which makes it easier for users to find messages with a specific theme or content. Users create and use hashtags by placing the hash character (or number sign) # in front of a word or unspaced phrase, either in the main text of a message or at the end. Searching for that hashtag will then present each message that has been tagged with it."*

As it can be induced, hashtags carry quite a weight regarding marking and identifying topics the users wants to talk about or draw attention to. The spontaneous way hashtags are created – there are no restrictions regarding what a hashtag can be – is their crucial feature. This allows for forming a true image of the users' interests, things important for them, and things that draw their attention. As the result, any type of analysis of hashtag data could lead to a better understanding of the users' attitudes, as well as detection of events, incidents, and calamities.

2.2 Fuzzy Sets

Fuzzy set theory [13] aims at handling imprecise and uncertain information in various domains. Let D represents a universe of discourse. A fuzzy set F with respect to D is defined by a membership function μ_F: $D \rightarrow [0,1]$, assigning a membership degree $\mu(d)$ to each $d \in D$. This membership degree represents the level of belonging of d to F. A fuzzy set can be represented as pairs:

$$\mathrm{F} = \left\{ \frac{\mu(d_1)}{d_1}, \frac{\mu(d_2)}{d_2}, \ldots \right\}$$

For more information on fuzzy sets and systems, please consult [17, 18].

2.3 Fuzzy Clustering

One of the most popular methods of analysis of data focuses on identifying clusters of data-points that exhibit substantial levels of similarity. There are multiple methods of clustering data that differ in their ability to find data clusters, and their complexity [4, 6, 7, 12].

Among many clustering algorithms there are ones that utilize fuzzy methodology [1, 2, 5]. In such a case, clusters of data-points do not have sharp boarders. In general, data-points belong to clusters to a degree. In the fuzzy terminology, we talk about a degree of belonging (membership) of a data-point to a given cluster. As the result, there are points that fully belong to a given cluster – membership value of 1, as well as points that belong to a cluster to a degree – membership values between 0 and 1. Such an approach provides more realistic segregation of data – very rarely we deal with a situation that everything is clear, and data can be divided into sets of data-points that are "clean", i.e., contain points that simply belong or do not belong to clusters.

The method used here is based on a fuzzy clustering method called FANNY [9]. The optimization is performed via minimizing the following objective function

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mu_{iv}^{r} \mu_{jv}^{r} d(i,j)}{2 \sum_{j=1}^{n} \mu_{jv}^{r}}$$

where n is a number of data-points, k is a number of clusters, μ is a membership value of a data-point to a cluster, $d(i,j)$ is a distance or difference between points i and j.

The selection of that approach has been dictated by the fact that we do not want to create fictitious centers of clusters, as it happens in widely popular fuzzy clustering method FCM [3]. Additionally, there is its new implementation in R programming language [15] that is used here.

2.4 Cluster Quality and Visualization

Clusters contain multiple data-points that are distributed in the space embraced by the clusters' boundaries. Some of these points are quite inside – have high values of membership, while some are close to the boundaries – have small values of membership while at the same time they have comparable values of membership to other clusters. An interesting measure indicating quality of a cluster, i.e., demonstrating that data-points that belong to this cluster are well fitted into it, is called silhouette width [11]. This measure is represented by the following ratio for a given element i from a cluster k:

$$s(i,k) = \frac{OUT(i) - IN(i,k)}{\max(OUT(i), IN(i,k))}$$

with

$$OUT(i) = \min_{j \neq k} \left(\frac{\sum_{m=1}^{N_j} d(i,m)}{N_j} \right), \qquad IN(i,k) = \frac{\sum_{m=1}^{N_k} d(i,m)}{N_k}$$

where $d(i,m)$ is a distance (or a difference) between data-points i and m, N_k is a size of cluster k, N_j is a size of any other cluster. The value of $s(i,k)$ allows us to identify the

closest cluster to a point i outside the cluster k. Positive values of silhouette indicate good separation of clusters.

The process of visualization of multi-dimensional clusters is fairly difficult. A possible solution could be a projection of clusters into selected dimensions. But then, the issue is which dimensions to choose. In his paper, we use an approach introduced in [10]. The approach called CLUSPLOT is based on a reduction of the dimension of data by principal component analysis [8]. Clusters are plotted in coordinates representing the first two principal components, and are graphically represented as ellipses. To be precise, each cluster is drawn as a spanning ellipse, i.e., as a smallest ellipse that covers all its elements.

3 Collected Data

3.1 Hashtag Data

The process of data analysis is performed on real data representing popularity of hashtags. The data are obtained from the website *Hashtagify.me*, and contain information about 40 different hashtags. The popularity ratings have been obtained for the period of nine weeks. The sample of data for a few selected hashtags is shown in Table 1.

Table 1. Popularity of selected hashtags

Hashtag	Popularity (# weeks in the past)									
	-nine	-eight	-seven	-six	-five	-four	-three	-two	-one	zero
#KCA	100.0	100.0	100.0	100.0	100.0	93.0	84.3	85.2	81.1	75.9
#callmebaby	0.0	0.0	30.3	20.9	65.5	98.0	95.1	89.1	85.7	87.3
#SoFantastic	56.9	71.5	72.7	73.5	75.8	97.4	92.3	36.6	37.7	48.8
#Nepal	44.0	44.5	42.3	42.5	42.6	42.5	43.2	73.8	80.8	63.6
#NepalQuake	0.0	0.0	0.0	0.0	0.0	0.0	0.0	59.9	72.5	53.1
#iphone	80.4	80.8	81.9	81.2	79.8	81.9	83.2	84.6	84.1	82.8

The values presented in Table 1 show the popularity (as % relative to other hashtags) for every week. For example, *#KCA* was the most popular hashtag for the first five weeks. However after the fifth week, its popularity has started decreasing. The hashtag *#callmebaby* did not even exist for the first few weeks, than rapidly gained popularity, and after two weeks its popularity has been around 85%. Very similar behavior can be observed for the *#NepalQuake*. Its popularity in the last few weeks has been in the range from 53% to 72% [13, 16]. The hashtag *#iphone*, on the other hand, is characterized via a continuous – with some small fluctuations – level of popularity: 80 to 84%.

Table 2. Changes in popularity of selected hashtags

Hashtag	Popularity (# weeks in the past)				
	zero vs nine	zero vs seven	zero vs five	zero vs thee	zero
#KCA	−24.1	−24.1	−24.1	−8.4	75.9
#callmebaby	87.3	57.0	21.8	−7.8	87.3
#SoFantastic	−8.1	−23.9	−27.0	−43.5	48.8
#Nepal	19.6	21.3	21.0	20.4	63.6
#NepalQuake	53.1	53.1	53.1	53.1	53.1
#iphone	2.4	0.9	3.0	−0.4	82.8

In order to analyze behavioral patterns of hashtags a simple processing of data has also been performed. Here, we are interested in the percentage of changes of popularity of hashtags. The data are presented in Table 2. Here, the calculations have been done using a very simple formula [19, 20]:

$$change_{zero\ vs\ N} = popularity_{week:zero} - popularity_{week:N}$$

The calculated change represents a difference between the popularity value for the current week *week:zero* and the popularity value for a considered *week:N*. For example, the value $change_{zero\ vs\ nine} = -24.1$ means that the popularity of *#KCA* in week ***zero*** is *−24.1* lower than its popularity in week ***-nine***.

3.2 Presidential Election 2012 Data

The created data set focuses on elections in United States. We have selected elections of 2012. The main reason for such a selection is an importance and scope of the 2012 elections. The elections were a very large event in the US history. They consist of the following elections: (1) the 57th presidential election; (2) Senate elections; and (3) House of Representative elections.

The first step in collecting tweets of the members of parties has been creation of a list of Twitter accounts of members of a parliament and most important members of parties. Twitter has a feature called *twitter list* where you can create a collection of Twitter accounts for people to follow. Almost all parties share lists of party members, parliament members or party related accounts. Such lists enable to promote the party's ideas, and make it easy to follow news related to the party. Also, some websites offer such lists for individuals to follow. In order to create our own lists for each party, we merge all accounts that appear in those party lists in one list. Such created list will be used to collect tweets.

The collection process has been done using the Twitter Search API. We have constructed a program – twitter search collector – that periodically collects and stores tweets using eight different API keys. The program requests only tweets that have an ID higher than the last tweets we collected with the previous/last usage of the collector.

The details regarding number of tweets associated with each party are presented in Table 3.

Table 3. Collected tweets statistics

Party	Number of tweets	Number of accounts
Republican	95193	560
Democratic	95731	361
Libertarian	13202	63
Green	8625	175
Justice	62612	43
Socialism and liberation	2128	16

4 Conclusion

The presented here analysis of temporal aspects of hashtags – their popularities over time and the changes of these popularities – is an attempt to look at dynamic nature of the user-generated data. The application of fuzzy clustering shown here provides a number of interesting benefits related to fact that categorization of hashtags is not crisp. The further investigation of fuzzy-based measures leads to interesting conclusions.

The construction of fuzzy signatures based on frequency of occurrence of hashtags is an interesting approach to express importance of opinions and issues represented via tweets' hashtags and noun-phrases. A simple process of constructing such signatures is presented here. Once the signatures are obtained, they are used to compare the importance of opinions/issues articulated by groups of individuals represented by the signatures. These processes have been applied to tweets representing US elections 2012.

References

1. Wu, K.L., Yang, M.S.: Alternative c-means clustering algorithms. Pattern Recogn. **35**, 2267–2278 (2002)
2. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
3. http://dictionary.reference.com. Accessed 8 May 2015
4. http://www.r-project.org. Accessed 8 May 2015
5. https://twitter.com/. Accessed 8 May 2015
6. http://www.wikipedia.org. Accessed 8 May 2015
7. Klir, G., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, Upper Saddle River (1995)
8. Pedrycz, W., Gomide, F.: Fuzzy Systems Engineering: Toward Human-Centric Computing, Wiley-IEEE Press, New York (2007)
9. Pal, A., Mondal, B., Bhattacharyya, N., Raha, S.: Similarity in fuzzy systems. J. Uncertainty Anal. Appl. **2**(1), 18 (2014)
10. Pappis, C.P., Karacapilidis, N.I.: A comparative assessment of measures of similarity of fuzzy values. Fuzzy Sets Syst. **56**(2), 171–174 (1993)
11. Gerstenkorn, T., Manko, J.: Correlation of intuitionistic fuzzy sets. Fuzzy Sets Syst. **44**(1), 39–43 (1991)

12. Dumitrescu, D.: A definition of an informational energy in fuzzy sets theory. Stud. Univ. Babes-Bolyai Math. **22**(2), 57–59 (1977)
13. Dumitrescu, D.: Fuzzy correlation. Studia Univ. Babes-Bolyai Math. **23**, 41–44 (1978)
14. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**(1), 87–96 (1986)
15. Shahbazova, S.N.: Development of the knowledge base learning system for distance education. Int. J. Intell. Syst. **27**(4), 343–354 (2012). Wiley Periodicals, Inc., Wiley - Blackwell
16. Shahbazova, S.N.: Application of fuzzy sets for control of student knowledge. Appl. Comput. Math. Int. J. **10**(1), 195–208 (2011). Special Issue on Fuzzy Set Theory and Applications. ISSN 1683-3511
17. Koshelova, O., Shahbazova, S.N.: Fuzzy multiple-choice quizzes and how to grade them. J. Uncertain Syst. **8**(3), 216–221 (2014). www.jus.org.uk
18. Abbasov, A.M., Shahbazova, S.N.: Informational modeling of the behavior of a teacher in the learning process based on fuzzy logic. Int. J. Intell. Syst. **31**(1), 3–18 (2015). Wiley Periodicals, Inc., Wiley - Blackwell
19. Shahbazova, S.N.: Modeling of creation of the complex on intelligent information systems learning and knowledge control (IISLKC). Int. J. Intell. Syst. **29**(4), 307–319 (2014). Wiley Periodicals, Inc., Wiley - Blackwell
20. Zadeh, L.A., Abbasov, A.M., Shahbazova, S.N.: Fuzzy-based techniques in human-like processing of social network data. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **23**(1), 1–14 (2015). Special issue on 50 years of Fuzzy Sets