H. Reiterer & O. Deussen (Hrsg.): Workshopband Mensch & Computer 2012
München: Oldenbourg Verlag, 2012, S. 83-90

Mining Twitter for Cultural Patterns

Elena Ilina, Fabian Abel, Geert-Jan Houben

Web Information Systems, Delft University of Technology

Abstract

Adaptive applications rely on the knowledge of their users, their needs and differences. For instance, in the scope of the ImReal[1] project, a training process is adapted to users' origins using information on user cultural backgrounds. For inferring culture-specific information from available microblogging content, we monitor the usage of Twitter elements such as hashtags, web links and user mentions. We analyze how users from different cultural groups employ these elements when they tweet. This allows us to get insights on microblogging patterns for different cultural groups of Twitter users and an outlook into user preferences and traits towards sharing content with others, time preferences, and social networking attitudes. Potentially, such information can be used for adapting software applications in accord with user culture-specific behavioral traits.

1 Introduction

Adaptive applications such as e-learning environments benefit from knowledge of the cultural backgrounds of users. For instance, e-learning applications aiming to work with students from different cultural backgrounds benefit from a representation of culture-related aspects of the users. One of the case-studies of the ImReal project involves learning how to effectively communicate with people from other cultural backgrounds. In this case, culture-oriented user modeling could take place by considering cultural aspects of users and using them in adapting the application behavior accordingly to the user needs. The cultural-awareness is the research scope of the ImReal project investigating the impact of the augmentation of user experiences in the adaptation process. The user modeling realized in the U-Sem component is envisaged to provide needed culture-related information for the adaptive simulators.

As result of User Modeling (UM), user profiles representing user characteristics are created and used for adapting applications to user needs. When user-related information cannot be retrieved directly from the user, or is not available, adaptive applications might exploit user

[1] http://www.imreal-project.eu/

data derived from external sources like social networks. Twitter can also provide information on a user's geographic locations and use of languages. However, can we ascertain culture-oriented behavioral patterns of user behavior on microblogs? This question motivated us to investigate mining cultural patterns of user behavior on Twitter. In this work, we analyze microblogging behavioral patterns. We adopt the well-known Lewis model (Lewis 2000) of cultural dimensions, which is used for describing differences in communication of people belonging to different cultural groups. We base our investigation on the assumption that personality traits as defined by Lewis are also reflected in the way how users blog on Twitter. This allows us to identify differences between user groups. Our main contributions include an analysis of user behavior on Twitter for user groups of different cultural origin and culture-oriented user modeling insights based on user behavior in microblogs.

2 Related Work

Previous research on personalization and adaptive systems exploit information published in social network platforms in order to collect information on user traits and interests. For instance, (Abel et al. 2011a) uses Twitter for creating content-based user profiles, which are further aligned with news articles in their news recommendation experiments. For improving recommendation quality, (Abel et al. 2011b) exploit information from several social networks, including Twitter and Facebook. A generic adaptive system based on Twitter data was proposed in (Hannon et al. 2011). Nevertheless, there are not many existing approaches for collecting culture-related user traits. The recent work by (Gao et al. 2012) compares user behavior on Twitter and Weibo, linking identified behavioral patterns with the culture model by Hofstede. Hofstede studies social interactions with the help of cultural dimensions, relating people from different origins to criteria such as power distance, individualism or collectivism, uncertainty avoidance (Hofstede 2007).

The Lewis model of Cultures explores cultural differences in communication, which is based on three cultural dimensions including multi-active, linear-active and reactive types (Lewis 2000). The multi-active dimension is associated with people from Brazil and Spain, who are generally warm and loquacious. The linear-active dimension relates with cultures developed in countries such as Germany and the USA, who are good planners and like to operate with factual information. Japan and Vietnam are related with the reactive dimension associated with politeness and good listening skills. People from other countries are described having a mix of the three dimensions described. This research by Lewis indicates that people belonging to the same or similar cultural dimensions are also similar in their interpersonal communication. Since the main goal of this work is to model cultural-oriented user behavior based on micro blogging activities, we consider the Lewis model due to its focus on communication activities.

3 Experimental Setup

We selected users from countries such as Germany and Brazil which are positioned in the apexes of the Lewis model of Cultures and representing linear-active and multi-active user groups. We also added users from USA and Spain as being close to the respective countries in the triangle of the Lewis model even though these countries are not located directly at the apexes of the triangle. This enabled to analyze the behavior of the aforementioned user groups. Japan is also not depicted in an apex of the Lewis model, however, since Japan is listed as one of the top countries[2] in terms of open accounts and the number of active users on Twitter, we selected Japan for representing a reactive user group. This is why we selected Twitter users from Germany and the USA for representing the linear-active group, users from Japan for representing the reactive group, and users from Brazil and Spain for representing the multi-active users group, as shown in Figure 1.

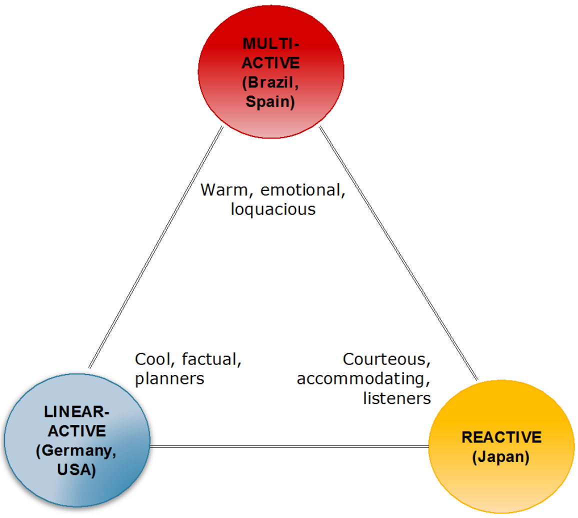

MULTI-ACTIVE (Brazil, Spain)
Warm, emotional, loquacious
Cool, factual, planners
Courteous, accommodating, listeners
LINEAR-ACTIVE (Germany, USA)
REACTIVE (Japan)

Figure 1: The Lewis Model of Cultures (Adapted)

Using the Twitter Public Streaming API[3], we selected a set of users tweeting from the respective geographic locations and having the locations indicated in their user profiles. When a user profile includes only a city name, we identify the country name using the Geonames API[4]. Next, we collected their tweets for a period of about two months. Afterwards, we

[2] http://semiocast.com/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan

[3] https://dev.twitter.com/docs/streaming-apis

[4] http://www.geonames.org/export/web-services.html

created user profiles by summarizing tweeting behavior based on the data extracted from the randomly selected 100 tweets per user. Our threshold of 100 tweets enabled us to aggregate user microblogging behavior for more than 1000 of users for each country, as shown in Table 1.

Country	Total Number Of Users	Users Posted 100 or More Tweets
Japan	4885	3479
Spain	4906	3448
Brazil	4910	3100
USA	1714	1325
Germany	2823	1770

Table 1: Dataset of Users (for whom we collected tweets from 2012-03-26 to 2012-06-01)

4 Features Analysis Results and Discussion

In order to mine cultural behavior patterns on Twitter, we analyzed usage of the following Twitter-specific features shown in Table 2: *Content-based* features including Uniform Resource Locators (URLs), hashtags and detected foreign languages[5], *Activity-based* features such as number of tweets from different geographic locations and the balance of tweeting during weekends versus weekdays, *Social Network-based* features such as user mentions, Twitter retweets, replies, number of friends and followers in the user network.

Feature	Description
URLs, Hashtags	Number of URLs and hashtags referred by a particular user in the content of the user tweets.
Languages	Number of languages automatically detected from the user tweets.
Locations	Number of tweeting locations for a user having a different location specified in the Twitter user profile.
Weekends	Number of tweets posted on weekends by the user.
Users	Number of user mentions detected in the user content
Friends, Followers	Number of friends and followers in the user's social network.
Retweets, Replies	Number of retweets and replies by the user.

Table 2: Twitter-specific User Features

Content-based Features. Table 3 shows statistics over content-based features. URLs are shared the most by user groups from the USA and Germany in average. URLs are shared least by the user group from Spain having about 31 URLs per 100 tweets published by a user in average. Users from Brazil and Japan use a similar number of URLs in average. The standard deviation in all user groups is however quite high, indicating that some of the users

[5] We employ java library available at: http://code.google.com/p/language-detection/

share more URLs, while others share less URLs. Linear-active users from Germany also lead the hashtag usage. However, the group from Spain shares more hashtags than the group from the USA per user in average, while users from Brazil and Japan share less hashtags compared to the aforementioned groups. The group from Japan, as reactive-users, share the lowest number of hashtags compared to others. Statistics on foreign languages detected in the user content shows that users from Japan use the most foreign languages, which is about 3 foreign languages detected per user in average, compared to other user groups having at most one foreign language detected in the user content. The language detection for users from Japan would appear to require further investigation. It seems that the USA user group tweet mostly in English, which explains the lowest mean value for the number of languages detected.

	Japan	**Germany**	**USA**	**Brazil**	**Spain**
			URLs Usage		
Maximum	235	200	107	116	190
Mean	32.0	37.5	42.5	32.1	30.7
St. deviation	26.0	27.4	26.1	29.7	24.9
			Hashtag Usage		
Maximum	343	326	356	434	431
Mean	7.6	34.4	28.7	14.7	29.5
St. deviation	17.8	36.1	31.5	23.7	28.7
			Number of Foreign Languages Detected		
Maximum	9	6	5	5	5
Mean	3.2	1.0	0.2	1.1	1.1
St. deviation	1.6	0.7	0.5	0.7	1.0

Table 3: Content-based Features Detected in 100 tweets per User in Average

Activity-based Features. Table 4 below shows that the linear-active users from Germany and the USA have the highest means of tweeting from different locations. The reactive group of users from Japan tweets the least from different geographic locations and tweet the most during weekends in average when compared to others. Users from the USA tweet the least on weekends in average.

Social Network-based Features. Table 5 below shows descriptive statistic over social network-based features including user retweets, replies, user mentions and the number of friends and followers in social networks of users. The statistics indicate that users from Spain retweet the most, while users from Japan retweet the least in average. Users from Germany reply the most in average and then followed by users from Spain, Japan, the USA and Brazil. The users from Japan mention other users the least in average. Users from Spain and the USA mention other users the most in average. Linear-active users from the USA and Germany have the largest number of friends and followers in average compared with other user groups. Brazilian users have the smallest number of friends, and Spanish users have a smaller number of followers in average.

	Japan	Germany	USA	Brazil	Spain
Tweeting from Different Locations					
Maximum	3	5	5	5	5
Mean	0.6	0.9	0.9	0.9	0.8
St. deviation	0.5	0.6	0.4	0.4	0.5
Tweeting on Weekends					
Maximum	76	74	62	73	88
Mean	28.6	25.3	23.5	24.3	24.0
St. deviation	9.7	9.7	8.9	9.8	9.7

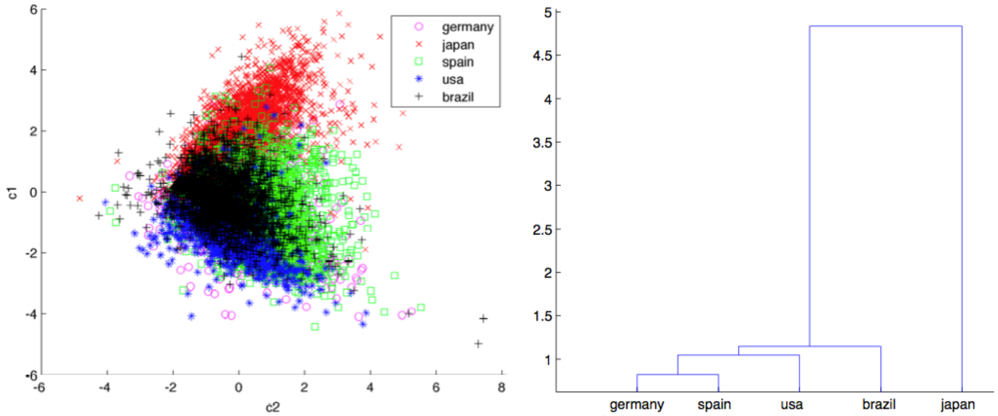
Table 4: Activity-based Features Detected in 100 tweets per User in Average

	Japan	Germany	USA	Brazil	Spain
User Retweets					
Maximum	100	92	96	86	99
Mean	8.2	14.9	15.0	14.3	23.2
St. deviation	11.6	14.2	13.4	13.0	16.1
User Replies					
Maximum	100	99	100	94	98
Mean	27.2	28.6	26.2	22.0	27.5
St. deviation	21.1	20.2	18.4	17.5	17.2
Other User Mentions					
Maximum	368	445	304	360	371
Mean	46.5	65.8	75.1	57.9	83.9
St. deviation	28.1	35.0	38.6	33.9	35.5
Number of Friends					
Maximum	4513	4658	4603	5060	5227
Mean	337.5	356.2	400.5	282,1	335.4
St. deviation	491.3	438.9	446.3	338.9	378.0
Number of Followers					
Maximum	4911	4941	4880	4886	4816
Mean	318.3	398.9	501.6	335.8	296.4
St. deviation	499.2	589.3	749.6	518.2	495.5

Table 5: Social Network-based Features Detected in 100 tweets per User in Average

Overall, based on the descriptive statistics and comparing mean values of features for different cultural groups, we found behavioral differences between linear-active and reactive user groups, while multi-active users having similarities with linear-active and reactive groups are difficult to separate. Spanish refer the most to other users and they are quite similar in their behavior with the USA group, while Brazilians share fewer links, and, only mention more other users than Japanese. Reactive users share the least of hash tags and user mentions, and tweet the least from different geo-graphic locations. They tweet more on weekends compared with other user groups. Linear-active users have larger social networks and share more URLs in average compared with other user groups.

Furthermore, we investigate whether it is possible to distinguish between user groups based on the features analyzed. Based on the Multivariate Analysis of Variance, we draw scatter plots showing clusters of user groups. Two canonical variables help to distinguish between user groups. They are calculated from the means of the feature values. The first canonical *c1* variable helps to differentiate the red cluster for people from Japan from the blue cluster for people from the USA. The second canonical variable *c2* enables to distinguish between users from Spain (green cluster) and users from Brazil (black cluster). In Figure (b), clusters of user groups are plotted against the distance between clusters. The link between Germany and Spain shows that these clusters' have the smallest distance between their means, while the cluster of Japan has the greatest distance to other clusters. The link between the USA and Brazil clusters is shorter than the link between the USA and Japan clusters. This enables us to separate these user groups, however, user groups from Spain and Germany are difficult to separate. We explain it by possible cultural similarities between two user groups and how they behave on Twitter with the features analyzed.



(a) User Clusters (b) Distances between Clusters

Figure 2: Cultural Group Clusters

The feature set does not allow us to map our model to the Lewis' model explicitly. However, our findings revealed some interesting microblogging patterns. Linear-active users share the most of URLs, and reactive persons from Japan tweet the most on weekends while sharing the least of hashtags in average compared to others.

5 Conclusions and Further Research

We analyzed microblogging behavior on Twitter for user groups from Germany, USA, Spain, Brazil and Japan and created group profiles describing user behavior for different cultural backgrounds. We have found that Japanese users behave very differently. They tweet more on weekends, and share the least hashtags and user mentions when compared with other user groups. Users from the USA and Germany generally share more URLs and have more friends compared with others. Users from Spain and Brazil stay apart in a way that they have some similarities with the rest of groups. Potentially, the information about user behavior can be further exploited for designing adaptable applications fitting to user needs. In further work, we will perform in-depth analysis on a larger user dataset and add more Twitter-specific features to our model in order to find further insights on cultural differences. The next goal will be predicting cultural origins of users, which is paramount for adaptive e-learning applications requiring knowledge on user cultural backgrounds.

Acknowledgements

This work is partially sponsored by the ImREAL project (http://imreal-project.eu).

References

Abel, F., Gao, Q., Houben, G.J. & Tao, K. (2011a). Analyzing User Modeling on Twitter for Personalized News Recommendations. In *User Modeling, Adaptation and Personalization, UMAP 2011.* Girona, Spain: Springer LNCS 6787. 1-12.

Abel, F., Herder, E., Houben, G., Henze, N. & Krause, D. (2011b). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems 22*(3), 1–42.

Gao, Q., Abel, F., Houben, G.J. & Yu, Y. (2012). A Comparative Study of User's Microblogging Behavior on Sina Webo and Twitter. In *UMAP 2012, Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization*. Montreal, Canada: Springer LNCS.

Hannon, J., Knutov, E., De Bra, Pechenizkiy, P. M., Smyth, B. & McCarthy, K. (2001). Bridging Recommendation and Adaptation: Generic Adaptation Framework - Twittomender compliance study. *Proceedings of 2nd DAH'2011 Workshop on Dynamic and Adaptive Hypertext.*1-9.

Hofstede, G. (2007). A european in asia†. *Asian Journal of Social Psychology 10* (1). 16–21.

Lewis, R. (2000). *When cultures collide: Managing successfully across cultures*. London: Nicholas Brealey Publishing.

Contact Information

Elena Ilina
EEMCS, Web Information Systems
P.O. Box 5031, 2600 GA Delft
The Netherlands