Visual Analysis of Thematic, Social and Geospatial Patterns of Microblogging Content Using D3

Thomas Gründemann and Dirk Burghardt
Institute of Cartography, Technische Universität Dresden
Helmholtzstraße 10, 01069 Dresden, Germany
thomas.gruendemann@tu-dresden.de, dirk.burghardt@tu-dresden.de

Abstract

User generated content from microblogging platforms often have a direct or indirect thematic, social, geospatial and temporal reference. Visual analysis tools can give an insight into the structure of the data provided by such platforms. This paper describes a simple way to analyze and visualize the thematic, social and geospatial patterns by using microblogging content from the platform Twitter. For this study we used tweets collected during a soccer match. We mapped the network of the thematic relation by examining the co-occurence of hashtags in a single tweet and the network of the social relation by examining the relationships between the users. The Louvain unsupervised community detection algorithm was applied on the networks to discover different groups of topics and users. To map the geospatial distribution of the users, we extracted the location information from the user profile and from the geolocated tweets. The results of this work were presented by using the JavaScript library D3 in order to produce a dynamic and interactive data visualization.

Keywords: user generated content, volunteered geographic information, Twitter, visualization, visual analysis.

1 Introduction

In the recent years social network platforms such as Twitter are increasingly becoming popular as a media for the communication over the internet. The users produce, share and consume information over social links and this conversation creates patterns with identifiable structures. Visualization techniques can help us efficiently analyze and understand these structures.

Therefore the visualization of microblogging content is currently a popular topic in the scientific community. One example is TweetXplorer [8], which illustrates how a topic propagates in time and space according to the user retweets. Another example is ScatterBlogs2 [4] a tool for interactive filtering and visualization of real-time microblog messages.

Using Twitter as data source, we present a simple approach to understanding microblogging content and illustrate how one can start to find patterns in order to form better conclusions about the data.

2 Data Set and Tools

With about 320 million active users, which share about 500 million tweets per day [13], Twitter has gained tremendous popularity in the past few years. The sheer amount of user generated content, the variety of metadata and the public access through the API are advantages that makes Twitter a popular data source for researchers.

For our study we used a dataset resulting from a search query with the hashtag #fcbsvw through the Twitter Search API [14] on 20 April 2016. The hashtag #fcbsvw was the keyword for content related to the soccer match between the teams FC Bayern München and SV Werder Bremen on 19 April 2016 in Munich. It was the semifinal of the DFB Pokal 2016. The dataset contains 624 tweets covering the period 19 - 20 April 2016.

For the visualization of the thematic, social and geospatial patterns we used the JavaScript library D3. Embedded within an HTML webpage, D3 uses pre-built JavaScript functions to select elements, create SVG objects, style them using CSS, or add transitions, dynamic effects or tooltips to them for producing dynamic and interactive data visualizations [7]. There are two main advantages of D3 compared to other software tools or libraries. First, D3 use web browsers to represent the visualization. By avoiding proprietary software and plug-ins, the visualization is accessible on the widest possible range of devices, from desktop computers to tablets and even smartphones. Second, because of D3s concept of binding data to the DOM (Document Object Model) and the use of SVG, HTML5 and CSS standards it provides the ability to create own data visualization techniques with a number of interactive components.

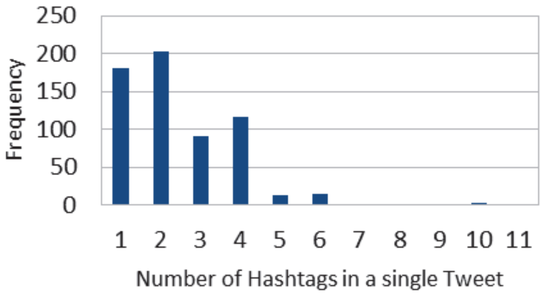
3 Analysis and Visualization

3.1 Analysis and Visualization of the Thematic Relation

Hashtags are keywords preceded by a hash (“#”) character, selected by users as being representative for a certain topic. By adding hashtags to the message, users build a semantic network linking their messages to each other. These networks can be represented as a graph, where nodes represent hashtags, and the edges represent the co-occurrence of hashtags within a single tweet. The more often two hashtags are used together, the stronger the edge is between them. To examine only the hashtags that occur together in a single tweet, can be a first step to discover topics and can help to get a better insight into the conversation.

An analysis of the data set shows that a total of 1,537 hashtags are included, of which 156 are unique. In about 90% of the tweets at least two hashtags occur together. Duplicates of hashtags within a single tweet were not included.

Figure 1: Hashtag occurrence in a single tweet in the data set.



Furthermore, it was examined how many hashtag pairs are contained in the data in order to find strong thematic relationships. A total of 1,778 hashtag pairs were found, and of these in turn 509 were unique.

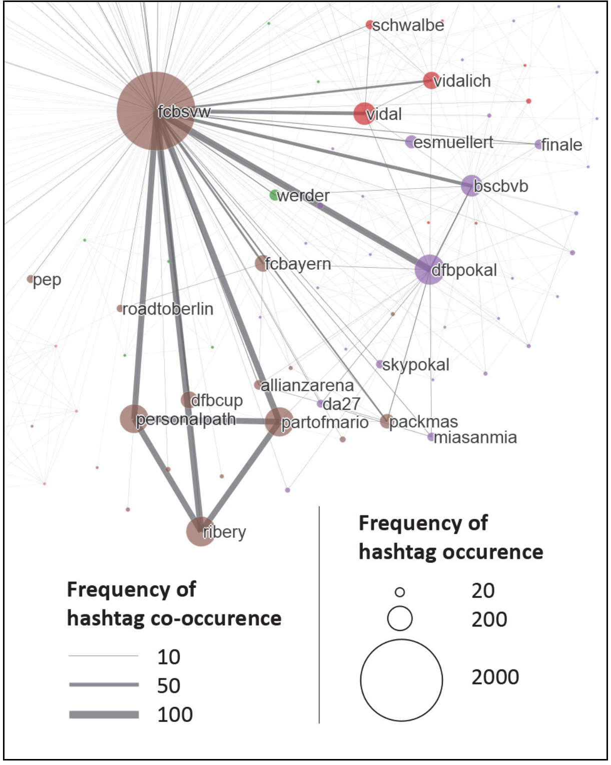
The D3 visualization of the hashtag network is shown in figure 2. The nodes represent the hashtags and the edges the occurrence of hashtag pairs in a tweet. The network is shown in a force-directed way, which means hashtags who are stronger connected are shown in closer proximity, while weaker connected hashtags are farther apart. The strength of their connectedness depends on their node degree. The size of the node indicates the number of the appropriate hashtag in the entire data set. With the interactive visualization the network can be moved, zoomed in and out. In addition, the nodes of the network can be selected and moved. By hovering the mouse cursor over a node a tooltip appears. It shows the tweets in which the respective hashtag is included. The tweets are ordered by the frequency of their occurrence in the data set.

Since tweets were searched by the given hashtag "fcbsvw", this node is the main theme and all other themes are connected to "fcbsvw". The number of hashtag pairs is represented by the width of the edge. Strong links exist between "partofmario", "ribery" and "personalpath" and results from a retweet with the image of players of FC Bayern München Mario Götze, Franck Ribéry and Thomas Müller in preparation for the match, which was shared 84 times. Strongly linked are also the hashtags "dfbpokal" and "bscbvb" (the second semi-final match between Hertha BSC and Borussia Dortmund a day later). And there is a strong link to "vidal" – another player of FC Bayern.

In order to find groups of topics a clustering method was applied. To detect these groups we used the Louvain Method [3]. The Louvian method is a greedy optimization method. First, the method identifies small groups by attempting to optimize local groups, then it aggregates the nodes belonging to the same group and performs the process again. A total of 7 groups were found with this method. Nodes that belong to the same thematic group, have the same color.

By examining the groups of the network and the tweets which can be seen in the node tooltips, the strong relationships between hashtags can be explained, such as the light red group: the player Vidal (hashtag vidal) feigns a foul (hashtag schwalbe – German for swallow, in English diving). The network shows many more topics but primarily additional issues of the day.

Figure 2: Thematic relationships of co-occurring hashtags of tweets with the hashtag #fcbsvw.



3.2 Analysis and Visualization of the Social Relation

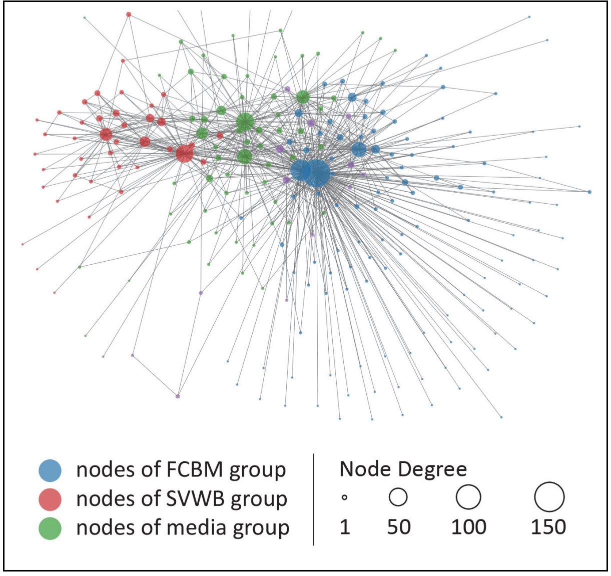
The users are connected by conversation. On Twitter, the connection is realized as a follower or as a friend. Followers are users who follow "me" and friends are users who "I" follow. The visualization of the relationships in a network gives an insight into the structure of the conversation.

The data set contains tweets from 358 different users. For the graphical representation of the network, all followers and friends of the users were fetched. Then, the connections between the users of the data set were extracted from the followers and friends lists.

Figure 3 shows the relationships among the users. The nodes represent the users and the edges the existence of a link between the users. Not shown are the 93 unconnected users and small networks of less than 10 users. The social network shown in figure 3 consists of 241 nodes with 720 edges. For further consideration it was not relevant whether the users are followers or friends. It was only considered if a connection exists. The network is shown again in a force-directed way, in which stronger connected users are shown in closer proximity, while weaker connected users are farther apart, depending on their node degree. The size of the node provides information about the number of nodes connected to the user.

By hovering the mouse cursor over a node, a tooltip appears showing information about the respective user such as the user description from his profile, the number of connections to other users in the network or the content of the tweet the user posted.

Figure 3: Relationships among the users who posted a tweet with the hashtag #fcbsvw.



The Louvain Method for detecting groups was also applied to the social network. Nodes that belong to the same social group, have the same color. Three big clusters can be detected. The group with the red nodes (SVWB group) consists of supporters of the soccer club SV Werder Bremen. Its largest node is the official Twitter account of the soccer club. The group with the blue nodes (FCBM group) consists of supporters of the soccer club FC Bayern München with its largest node the official Twitter account of the soccer club. The group with the green nodes (media group) consists mainly of news agencies and journalists who report on the topic.

One can clearly see the supporters of the soccer clubs by the radial distribution of the connections to the official accounts. There are also very few connections among the SVWB group and the FCBM group. Such polarized groups are often found in soccer. The supporters of the respective clubs are also linked among themselves very low. The dissemination of information happens mainly through the nodes of the media group and the nodes of the official accounts of the soccer clubs.

3.3 Analysis and Visualization of the Geospatial Distribution

Microblogging content from Twitter often contains geospatial information and is available from two different sources. Either from the profile of the user or from the tweet when the user optionally choose to provide location information using a smartphone with GPS capabilities or a predefined location from Twitter such as a city or country with its associated bounding box.

In the data set none of the tweets had an explicit coordinate of the location. Therefore the location which the user provides in his profile were used. In about two-thirds of the user profiles we found an entry in the location field. The location string obtained from the user profile must first be translated into geographic coordinates. Typically, a gazetteer is used to perform this task. A gazetteer takes a location string as input, and returns the coordinates of the location that best correspond to the string. The granularity of the location is generally coarse. Most of the profiles provided a city as location. There are several online gazetteers, including Bing, Google, and OpenStreetMap. In our study, we used the Nominatim service from OpenStreetMap. For more than 80% of the users with an entry in their location field coordinates could be assigned. The rest of the users made entries of a fictional place.

In order to verify the location information of the user profiles, the 1,000 of the most recent tweets from the timeline of the user were collected. In the next step all geolocated tweets were extracted. For 77 users at least one geolocated tweet was found in the timeline. 56 of the 77 users made an entry in the location field in their user profile. If the assigned coordinate of the entry in the location field of the user profile was located within the boundary box of the place in the geolocated tweet, then the location information in the user profile was confirmed. To locate the 21 users who did not made an entry in the location field, the coordinate of the geolocated tweets was used that occurred most frequently.

Figure 4: Spatial distribution of the users who posted a tweet with the hashtag #fcbsvw.

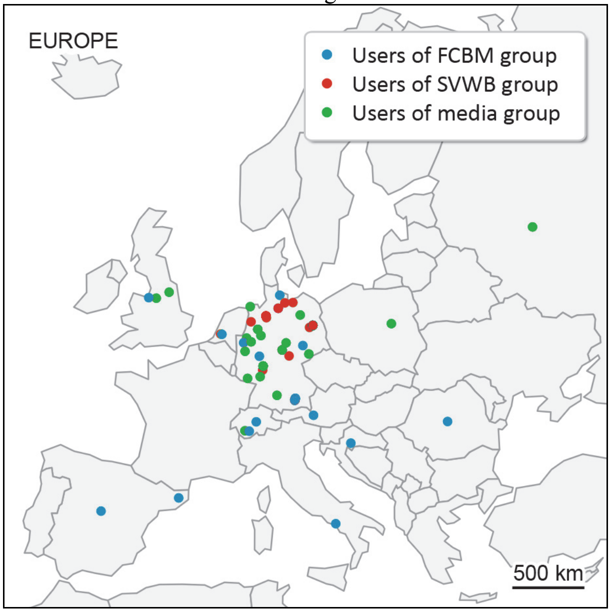


Figure 4 shows the spatial distribution of the user locations for the European part. The location of a user is identified by a dot on the map. The color corresponds to the user network. The spatial distribution is mainly focused on Germany, because it is an event in Germany. The user groups are very unequally distributed. Users of the SVWB group focus on the north and the middle of Germany. Whereas the spatial distribution of the FCB group also extends to other European countries, due to the higher level of awareness of the soccer club. The user of the media group focus on Germany, but are also found in other European countries.

4 Conclusions and Future Work

We have presented a simple way to visualize thematic, social and geospatial patterns of a microblogging conversation on Twitter. With a small data set of about 600 tweets we could show that distinctive patterns emerge from the conversation. The JavaScript library D3 provides the possibilities to visualize the pattern and to interact with the graphical representation. In this work we have demonstrated how the visualization tool D3 can help a user gradually obtain an insight into Twitter data.

To get a deeper insight into the conversation of the users and their relationships, future work should consider other information that relate to each other. For example images or hyperlinks in the tweet that are found together with the respective hashtags could be analyzed and visualized. Linking the views to show their mutual interaction and the consideration of temporal parameters will also lead to a better understanding of the thematic, social and geospatial relationships.

5 References

[1] Reda Alhajj, Jon Rokne. Encyclopedia of Social Network Analysis and Mining. Springer Publishing Company, Incorporated. 2014.

[2] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski and D. Thom. Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. In Computing in Science and Engineering, 15(3): 72–82, 2013.

[3] Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre. Fast unfolding of community hierarchies in large networks. In Journal of Statistical Mechanics: Theory and Experiment. P1008. 2008.

[4] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner and T. Ertl. ScatterBlogs2: Real-Time Monitoring of Microblog Messages through User-Guided Filtering. IEEE Transactions on Visualization and Computer Graphics, 19(12): 2022-2031, 2013.

[5] K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing Communication on Social Media: Making Big Data Accessible. In Proc. 15th ACM Conference on Computer Supported Cooperative Work, Workshop on Collective Intelligence as Community Discourse and Action, 2012.

[6] E. Meeks. D3.js in Action. Manning Publications, Shelter Island, NY, 2014.

[7] Scott Murray. Interactive Data Visualization for the Web: An Introduction to Designing with D3. O'Reilly Media, Inc., Sebastopol, CA, 2013.

[8] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski. Understanding Twitter Data with TweetXplorer. In Proceedings of the 2013 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1482-1485, 2013.

[9] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In International AAAI Conference on Weblogs and Social Media, 2013.

[10] N. Pervin, T. Q. Phan, A. Datta, H. Takeda and F. Toriumi, F. 2015. Hashtag popularity on twitter: Analyzing co-occurrence of multiple hashtags. In Social Computing and Social Media. Springer. 169–182. 2015.

[11] M. Russel. Mining the social web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more (2nd ed.). Sebastopol, CA: O'Reilly, 2013.

[12] Shamanth Kumar, Fred Morstatter and Huan Liu. Twitter Data Analytics. Springer Publishing Company, Incorporated, 2013.

[13] Twitter. About. https://about.twitter.com/company [Online; accessed 24-April-2016].

[14] Twitter. Using the Twitter Search API. https://dev.twitter.com/rest/public/search, 2016. [Online; accessed 24-April-2016].