Methods of Knowledge Discovery in Tweets

Sunmoo Yoon, RN, PhD[1], Suzanne Bakken, RN, DNSc[1,2]
[1] School of Nursing and [2]Department of Biomedical Informatics,
Columbia University, New York, NY, United States

Abstract

The purposes of this methodological paper are: 1) to describe web mining methods for knowledge discovery in Tweets, and 2) to illustrate application of the methods using the topic of physical activity. Methods described include: 1) structure mining to discover structures (macro-, meso-, and micro-level) of Tweet networks using social network analysis, and 2) content mining to discover Tweet contents using n-gram based text analysis and sentiment analysis. Specific web mining tools for each step of the web mining process (e.g., NodeXL, ORA, Pajek, Weka) are detailed. Our novel application of web mining methods was useful in understanding multiple dimensions of physical activity. The methods that we applied may be useful to others wishing to mine social media for health-related purposes.

Keywords, web mining, knowledge discovery, structure mining, content mining, web 2.0, Twitter

Introduction

Twitter is a short message microblogging service system that has exponentially grown during the past year. 15 million people around the world use Twitter to share feelings, observations, and activities of their daily lives. Messages shared using Twitter, commonly referred to as Tweets, are stored in log files and can be analyzed for content. Because of its pervasive influence, Twitter is used as a vehicle in measuring the pulse of public opinion (e.g., movie ratings, emotional responses to events). Thus, Twitter has the potential to be used to gain an understanding of health-related behaviors and to promote healthy behaviors and lifestyles. The purposes of this methodological paper are: 1) to describe web mining methods for knowledge discovery in Tweets, and 2) to illustrate application of the methods using the topic of physical activity.

Web Mining

Web mining methods are applied to discover meaningful knowledge from Web-based data such as blogs, news, and social media, and can be categorized into three types: Web structure mining, Web content mining and Web usage mining (Figure 1). Web structure mining aims to discover useful knowledge regarding communication structures. Web content mining extracts useful knowledge (e.g., discussion topics, opinions, positive or negative sentiments expressed) from Web content. The purpose of Web usage mining is to discover user access patterns.

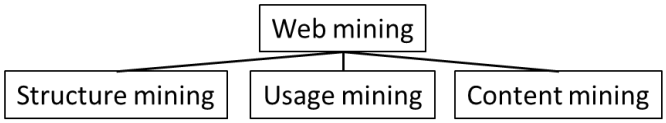

Figure 1. Types of web mining

1) Tools Used in Web Mining

Since web mining of social media is in its infancy, there is no single web mining tool available to comprehensively explore Tweet structures and contents. An overview of major web mining steps with applicable tools is shown in Table1.

Table 1. Overview of major web mining steps with the suggested tools

	Import data	Prepare data	Analyze data				
			Structure Mining			Content Mining	
			Macro	Meso	Micro	n-gram	Sentiment
NodeXL	√	√					
Notepad++		√					
ORA			√	√			
Pajek					√		
Weka						√	
Twitter Sentiment							√

2) Steps in Web Mining of Tweets

Web mining steps are illustrated in Figure 2 and described in the following paragraphs.

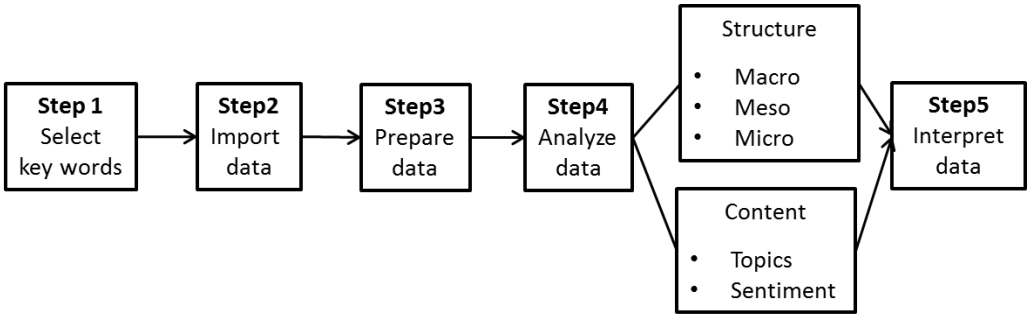

Figure 2. Steps of web mining

First, key terms and phrases are determined. Various terms can be used to cover a concept being searched (Step1). Second, Tweets are imported using the key term via NodeXL, which is a free, an open-source template for Microsoft Excel (http://nodexl.codeplex.com/) (Step2). Third, during the preparation step, Tweets are cleaned to facilitate readability in Weka data mining software. For example, symbols and unnecessary letters (“, @, ^, comma, space, ;, :, line feed, carriage return, www, http://*) can be removed using a free source-code editor, notepad++, followed by saving the file in CSV (comma separated value) format (Step3). Fourth, the imported Tweets are analyzed to discover the structure and content. For investigating macro level structure, the imported Tweet network files are opened in ORA (http://www.casos.cs.cmu.edu/projects/ora/software.html). The graphical visualization feature in ORA is used to examine overall structure of Tweet networks. The generating-measures feature in ORA supports calculation of various measures of Tweet networks such as density, reciprocity, centralization, and so on. For investigating meso-level structure, communities can be detected using Newman algorithm in ORA[1]. To examine micro-level structure, triad census can be used to see the dynamics of the communication among three people using open source software, Pajek 2.04 (http://pajek.imfm.si/doku.php?id=download#download).

In order to investigate the contents of Tweets, the cleaned Tweets are opened in Weka, and tokenized using n-gram parameters to generate a Term-Tweet frequency matrix. Conway and the colleagues[2] suggest that the combination of unigrams, bigrams, and trigrams is superior to unigrams alone so the recommended parameter settings are 1 to 3 for the n-gram based content analysis. Sentiment analysis determines the attitude towards a particular topic[3]. Attitudes can be explored using a sentiment analysis tool developed at Stanford University (http://twittersentiment.appspot.com/) which categorizes Tweets as positive or negative with classifiers built from machine learning algorithms such as Naïve Bayes, Maximum Entropy, and Support Vector Machine, and the tool has accuracy above 80%[4] (Step4). In interpretation phase, the investigator digests and makes sense of the information [5] (Step5).

Application of Web Mining for the Topic of Physical Activity

This section illustrates our application of web mining methods to discover new knowledge from Tweets using the topic of physical activity. We first took "physical activity" as a key search phrase (Step1). We imported 1000 Tweets mentioning physical activity via NodeXL (Step2) and then cleaned the data to remove symbols and unnecessary letters (Step3). Our analysis phase consisted of two tasks; structure mining and content mining (Step4). Detailed information regarding how to interpret the analysis results is described below (Step5).

1) Structure Mining

Our strategy to examine structure of Tweet networks was to approach the network from macro-, meso-, and micro-levels. This three-level approach offers more comprehensive views of the network than a single-level approach. Data are graphically visualized in Figure 3 and network characteristics are numerically summarized in Table 2. Additional details are provided in the following paragraphs.

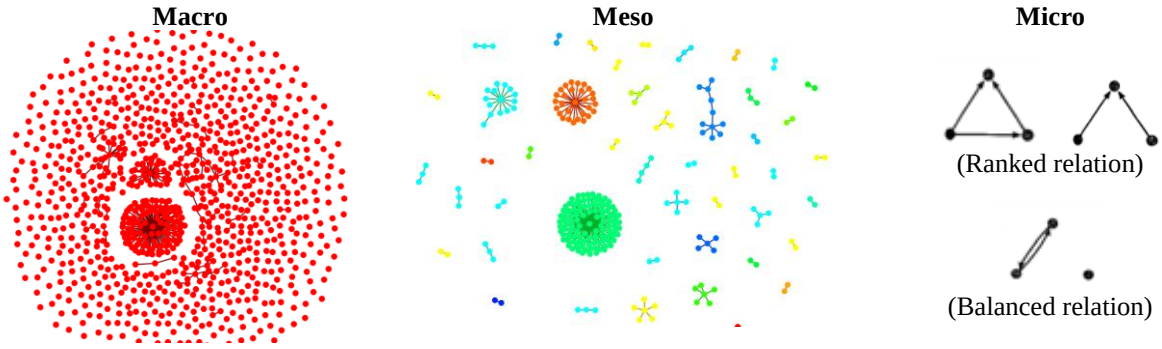

Figure 3 Graphical visualization of macro, meso, and micro-level structures of Tweet network mentioning physical activity; Macro level-structure showing many isolates, meso-level structure showing communities, micro-level structure showing the frequently occurring communication types among three people.

Macro-level structure: In Figure 3, dots (nodes) represent Tweet users (e.g., individuals, companies, or organizations). The Tweet network has 1000 nodes. A link is a tie between two people representing 'reply to' or 'mention' of relationship. The link count was 253 meaning that 253 Tweets were replied to or mentioned. Most of Tweet users (694) were isolated users. Density (cohesion) of the network was close to zero (0.000253, normalized from 0 to1) indicating extremely few communication links among Tweet users, and suggesting that information rarely flows among Tweet users. Reciprocity 0.016 indicates that only 1.6% of Tweet users respond to each other. The total degree of centralization of 0.052 indicates that there is a distributed communication style among Tweet users.

Meso-level structure: Communities are the structures in-between the whole network and individuals and reflect groups of individuals with similarities[6]. Sixty-one communities were detected by Newman algorithm within the Tweet network. High values of modularity indicate good divisions with dense internal connections and sparse connections between different groups, and the modularity score of 0.81 in our physical activity network indicates that the detected groups are distinct. The Top 10 leaders of communities varied including government (nutrition_gov, gohealthypeople, voalearnenglish), advertisers (lovescopes, astrologyforyou), media (ahealthblog), experts (davidkittner), and individuals (thatkalyngirl, insektmute, megsauce).

Micro-level structure: The small elements of micro-level groups accumulate into network structures, and understanding those local dynamics is important to understanding the entire network structure. Studies report that similarities in small groups of people (triads) can demonstrate and predict overall structure[7]. Triad census counts the different subgraphs of three users in a network and includes 16 different triad types representing balanced, ranked, hierarchical, and forbidden communication styles [8]. The triad census analysis for our physical activity network found that the most frequent communication type was the ranked-type9 (relative frequency (RF) compared to random network=246.2), followed by the ranked-type5 (RF=183.0), and the balanced-type3 (RF=124.0). Relative

frequencies compare the analyzed network to a random network, e.g., the ranked-communication type9 appears 246 times more than in a random network.

Table 2. Summary of web mining results of Tweets mentioning physical activity

Structure Mining	
Macro	Density (0.000), Reciprocity (1.6%), Total centralization (5.2%)
Meso	61 Communities, Top 10 Leaders; government (3), individual leader (3) media (1), expert (1), advertiser (1)
Micro	Relative frequency of communication-types among 3 users comparing to a random network; ranked-type9 (246.2), ranked-type5 (183.0), balanced-type3 (124.0)
Content Mining	
Topics	Advantages of regular exercise, 7 benefits from Mayo clinic, disadvantages of lack of exercise
Sentiments	Positive attitudes (52%), negative attitudes (48%)

2) Content Mining

Analysis of Tweets text is a rapid and inexpensive way to glimpse public opinion, although tools for extracting meaningful information are in the infancy phase. We used n-gram based text analysis and sentiment analysis as content mining strategies in order to detect the main discussed topics and to assess overall mood associated with the topics.

Topic detection: N-gram based text mining technique was applied to handle the voluminous Tweets. Tweet texts can be viewed as a vector with one component corresponding to each terms in a dictionary, and for dictionary terms that do not occur in a document, the value is 0. Tweets were transformed and represented with n-gram (monogram, bigram, and trigram) terms. The Tweet-term frequency dictionary computed by the n-gram was generated to represent the Tweet corpus including 1000 Tweets in Table 3. Word vector represents the frequency of the terms. In the Table 3, the "good" term appears most frequently. One can compare the vector of "benefits" (0.0519) and "7 benefits regular" (0.0294). The main topics included: consequence of lack physical activity, seven benefits of physical activity from Mayo Clinic, and health benefits of physical activity including enhancing immunity from American Heart Association.

Table 3. Physical activity Tweet term frequency dictionary

n-gram	Frequency*	n-gram	Frequency
good	0.0607 (70)	condition every human	0.0284 (29)
benefits	0.0519 (53)	destroys good condition	0.0284 (29)
regular	0.0499 (51)	lack activity destroys	0.0284 (29)
weight	0.0450 (47)	physical exercise save	0.0284 (29)
brain	0.0401 (39)	heart	0.0215 (29)
lack	0.0401 (39)	heart #didyouknow #exercise	0.0196 (21)
7 benefits regular	0.0294 (30)	immunity	0.0157 (16)

*Frequency is represented as a vector form. A number in parentheses () represents the actual number of times the term appeared in 1000 Tweets mentioning physical activity

Sentiment analysis: Sentiment analysis detects the opinions or attitudes towards a particular topic. We explored the attitudes of Twitter users associated with content expressed in Tweets. Figure 4 shows the sentiment expressed in Tweets mentioning physical activity from July 1st to July 31st, 2011; 52% of Tweets mentioning physical activity (green) reflect positive attitudes and 48% of Tweets reflect negative attitudes.

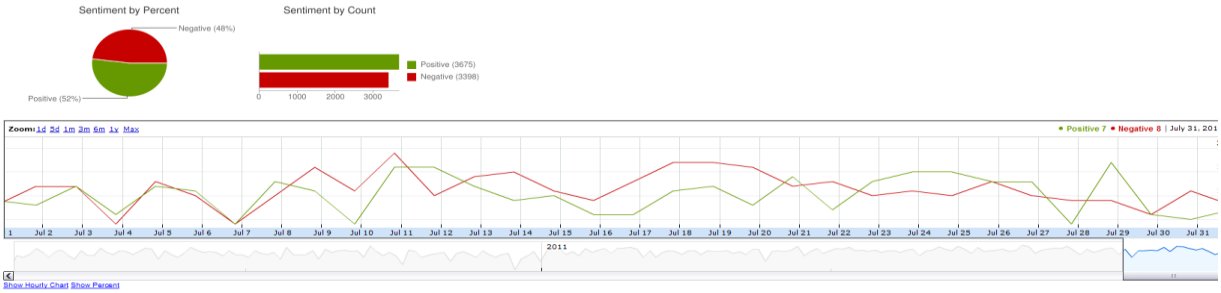

Figure 4. Sentiments expressed in Tweets mentioning physical activity from July 1st to July 31st, 2011

Discussion

The methods described in this paper support analysis of Tweet network structures and Tweet contents. Moreover, the tools applied are open source and downloadable, thus broadly available to other researchers. In contrast to other approaches, the methods and associated tools that we applied required minimal programming skills. Our novel application of web mining using social network analysis and n-gram based analysis with sentiment analysis efficiently produced new knowledge about comprehensive Tweet network structures and contents of Tweets mentioning physical activity. Our structure mining strategy uncovered multi-level views of the Tweet network. Our Tweet content mining strategy revealed the most frequently occurring topics and associated sentiments.

Conclusion

The methods that we applied may be useful to others wishing to mine social media for health-related purposes.

Acknowledgement: This study was funded by 3T32NR007969-09S1.

References

1. Newman MEJ, Girvan M: Finding and evaluating community structure in networks, Physical Rev E 2004; 6.
2. Conway M, Doan S, Kawazoe A, Collier N: Classifying disease outbreak reports using n-grams and semantic features, Int J Med Inform 2009; 78:e47-58.
3. Argamon S: Opinion Mining and Sentiment Analysis, Computational Linguistics 2009; 35:311-312.
4. Go A, Bhayani R, Huang L: Twitter Sentiment Classification using Distant Supervision, retrieved on Aug 14, 2010; available at http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf
5. Gershon N, Eick SG: Information visualization applications in the real world, Ieee Computer Graphics and Applications 1997; 17:66-66.
6. Reichardt J, Bornholdt S: When are networks truly modular?, Physica D-Nonlinear Phenomena 2006; 224:20-26
7. Newcomb T: The acquaintance Process. Edited by New York, Holt, Rinehart & Winston, 1961.
8. Wouter de Nooy, Andrej Mrvar, Batagelj V: Exploratory social network analysis with Pajek. Edited by New York, Cambridge University Press, 2005.