Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users

Dongwoo Kim
Users&Information Lab, Computer Science Department
KAIST, 335 Gwahangno, Yuseong-gu, Deajeon, Korea
dw.kim@kaist.ac.kr

Yohan Jo
Users&Information Lab, Computer Science Department
KAIST, 335 Gwahangno, Yuseong-gu, Deajeon, Korea
yohan.jo@kaist.ac.kr

Il-Chul Moon
System and Modeling Simulation Lab, Electrical Engineering Department
KAIST, 335 Gwahangno, Yuseong-gu, Deajeon, Korea
icmoon@smslab.kaist.ac.kr

Alice Oh
Users&Information Lab, Computer Science Department
KAIST, 335 Gwahangno, Yuseong-gu, Deajeon, Korea
alice.oh@kaist.edu

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
We discuss our findings from a study using Twitter lists to infer the characteristics and interests of users. Gathering and structuring user interest has been challenging because it often requires expensive and/or proprietary data such as users' clickthrough logs or desktop histories. We show that by using the tweets of all the users in a Twitter list, we can discover characteristics and interests of the users in that list, even if the users as individuals do not tweet about those interests. We conducted an experiment in which we compared the user interests as found by our system using Twitter lists with those that are perceived by the human subjects in the user survey. The survey confirmed that Twitter lists reflect well the perceived characteristics and interests of the users in those lists. The user survey also confirmed that the words extracted from each set of lists are representative of all the members in the list even if the words are not used by those members.

Keywords
Microblogging

ACM Classification Keywords
H3.3. Information Search and Retrieval: Retrieval models.

Introduction

Twitter[1] has recently added a new capability for users to create lists of their Twitter friends. While this new functionality serves the original purpose of organizing friends such that users can quickly look at the activities of a designated subset of friends, Twitter lists can also be a valuable source for inferring meaningful characteristics of Twitter users. For example, if user A belongs to a list called "coffee", we can infer that A is judged to be a good resource for questions related to coffee. Furthermore, we can use the tweets of the users in that list to predict a set of words that are related to user A, such as "arabica", "k-cups", and "starbucks". We propose to analyze Twitter lists to discover more information about users, an approach with the following characteristics:

- It uses publicly available data.
- It uncovers user characteristics that are not explicit in the data.
- It models users as they are judged by other users.

Our analyses and results show three interesting insights about how users have begun to use Twitter lists. First, like other Web 2.0 tools such as wikis, blogs, and social annotations sites, Twitter with its list functionality has evolved to be complex and full of potential for interesting new research directions in social media analytics. Second, a user's related words we found through the Twitter lists include many words that the user did not use on Twitter. Those words were found because they appeared in the tweets of other users who are in the same list as that user. The reason for this was our hypothesis that the discriminating words for a list will apply to all the members of the list, and that hypothesis was confirmed positive by our user study. Finally, Twitter lists are unique in that when we look at a user and the names of the lists that he is in, those list names represent what other Twitter users think of that user. Thus, the related words found by our method are probably closest to a model of that user's reputation or expertise as judged by others, which is an inherently different model than that built from the user's own tagging, browsing, or desktop activities.

List Name	Freq	List Name	Freq
friends	31,267	politics	3,078
news	15,216	design	2,866
music	14,596	family	2,834
celeb	13,837	travel	2,724
sports	8,210	tv	2,618
celebrities	7,419	people	2,608
amigos	6,852	famosos	2,549
tech	5,735	fashion	2,523
media	4,233	famous	2,333
entertainment	3,640	social-media	2,275

Table 1 Frequencies of top 20 list names

Twitter Lists

We have crawled over 3.3 million users' profiles to check if they are in lists, and it is roughly 10% of the total users in Twitter, as reported by techcrunch[2]. After analyzing out data we found over nine hundred thousand lists, and about 12% of the users belong to at least one list. Next we analyzed the list names. The analysis shows that a large number of lists share the same names, like friends, news and music. Table 1 shows the top 20 list names and their frequencies.

Lists and Terms

If we assume a list is composed of people who share the same interests, then, can we extract their common characteristics from their tweets in the list? In this chapter, we tried the feature selection to find the representative words that differentiate one list from other lists. We first gathered the tweets of the users who are in the lists we crawled. After crawling the data,

[1] http://www.twitter.com

[2] http://www.techcrunch.com

(a) Author

Word	Value
nanowrimo	349
booksel	342
menuscript	328
novelist	274
wip	260
synopsi	259
amwrit	249
paperback	236
kirku	226
bestsel	221

(b) Food

Word	Value
foie	282
slaw	252
shallot	251
chowder	216
chard	209
gnocchi	208
heirloom	207
fennel	205
horseradish	204
leek	200

(c) Cycle

Word	Value
cancellara	404
boonen	385
hincapi	383
peloton	383
adel	364
velonew	356
interbik	355
wiggin	345
schlock	335
mtb	330

(d) Fitness

Word	Value
kettlebel	322
glute	316
metabol	310
tricep	263
whei	220
bodybuild	217
bicep	208
bosu	203
healthiest	198
squat	193

Table 2 Top 10 words and χ^2 values

we selected 10 groups of lists that contain the following *keywords, author, coffee, cycle, fitness, food, game, mom, photograph, swim*, and *tech*. Each group contains 2 or 3 lists, and each list consists of 67.5 users on average. The reason why we aggregated several lists into one group is to avoid over-fitting to a specific list.

As shown in previous research[1], most tweets are daily chatters that are talking about user's life and what they are currently doing. To show that a list could represent interests of specific topics, we need to exclude these daily tweets. We used the χ^2 feature selection, a standard tool in text processing, because given a corpus, the χ^2 feature selection gives lower scores to commonly used words across the entire corpus and higher scores to words occurring within a few classes of documents. The length of one tweet is limited to 140 characters, so it is too short to consider one tweet as a single document. Instead we consider the aggregation of all tweets written by one user as one document. After that, we calculate the one-versus-rest word in the selected groups of lists. Table 2 shows the top 10 χ^2 values and words in each group. Intuitively, most of the words explain their list well.

User study

We ran a user study to test the hypothesis that for a Twitter list, the words with high χ^2 values are representative of the people in the list even if they do not use the words explicitly. That is, by using the χ^2 feature selection on tweets grouped by the lists we can find latent characteristics of users. Our user survey revealed that words with high χ^2 values are informative characteristics of users.

Method

The survey was conducted through a website[3] with 37 subjects. A subject is given ten words and hyperlinks to three Twitter users, and for each word, the subject is asked to choose one of the three Twitter users that they think knows about the word best. We generated four questionnaires. For each questionnaire, three Twitter lists are randomly chosen out of the nine pre-destined lists whose names contain: *author, coffee, fitness, food, game, mom, photograph, swim*, and *tech*. For each list, three or four words that have χ^2 values of higher than 120 were randomly selected. Also For each list, one of the Twitter users in the list was randomly chosen. Subjects were asked to explore the Twitter pages of the given Twitter users, and match each word to the most appropriate user.

Result

Although there is no ground truth, subjects' decisions can be considered as correct answers, because human can understand what each user is likely to say through the user's ID, bio, lists, friends as well as tweets. Hence, we compared our χ^2 result to the subjects' answers to verify that high χ^2 words in a list are effective in explaining the characteristics of the people in the list. For each word, agreement is defined as the largest proportion of the subject answers. The most agreed decision being considered as the correct answer, our high χ^2 words produced only three incorrect results out of forty, the accuracy being 0.925. Our user survey shows that the combination of the Twitter list functionality and the χ^2 feature selection is an efficient tool for inferring user characteristics. In about a third of

[3] http://uilab.kaist.ac.kr/ICWSM10

the subjects' answers, the subjects said they had found hints from user IDs, bios, and background images. A user's Twitter profile, including the screen ID, bio, and background image would require either manual effort or a more complex model to extract information computationally. Although our algorithm uses only tweets, it works as well as human judgments. Thus, the algorithm can be applied to data that have no summary or meta data available.

Discussions and Future Work

Our user study shows that our approach yielded good agreements between human decision and χ^2 words, even for the words that are not in the users' timeline. These results imply that the suggested high χ^2 words could be the latent characteristic of the users in the respective lists. The Twitter list can be a valuable information source in the various applications and research areas, when considered together with the features that we did not look at, such as hyperlinks, users' profile, network structure. We propose the followings list of potential research on Twitter lists.

- *Social Search* The rationale behind the social search is that users would trust the relevance judgments of their friends more than the overall popularity of Web pages. We can apply our work to prior work on social search[2] and groupization for personalized search[3] in this way: for a given query term, we can look for it within our χ^2 words to identify an appropriate Twitter list that would contain the most useful "friends" for that search.
- *Expert Recommend System* A Twitter list may consist of users with expertise on some topic. Although interest and expertise are different, we can extend our work here to discover experts on Twitter. This may involve categorizing Twitter lists into interest-related and expertise-related lists.
- *Information Source* Many users are sharing up-to-date news and events on Twitter. As most geographic lists are composed of people who live in or know well about the locations, tweets in these lists serve as local news.
- *Social Network Analysis* The following relationship on Twitter enables users to easily are unidirectional, whereas other popular social networking sites, such as Facebook, allow only bi-directional relationships. Thus, traditional social networking analyses do not fit Twitter in the same way as the other sites. However, some Twitter lists such as "friends" or "conversationlist", probably contain many bidirectional relationships within them. Looking at the network structure of those lists would be an interesting research direction.

Although Twitter list is a brand new feature, already a large number of people have started to use it. Hence, its potential as an information source is huge. We plan to study further in the directions outlined above, as well as continuing with our approach in this paper with more mature data and sophisticated models

References

[1] Java, A., Song, X., Finin, T., and Tseng, B., Why we twitter: understanding microblogging usage and communities. WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, Aug 2007.

[2] Chi, E.H. Information seeking can be social. IEEE Computer 42(3):42-46, 2009.

[3] Teevan, J., Dumais, S. and Horovitz, E. Discovering and Using groups to improve personalized search. Conference on Web Search and Data Mining, 2009.