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

Dongwoo Kim and **Yohan Jo** and **Il-Chul Moon** and **Alice Oh**
dw.kim@kaist.ac.kr, yohan.jo@kaist.ac.kr, icmoon@smslab.kaist.ac.kr, alice.oh@kaist.edu
KAIST Computer Science Department
Daejeon, Republic of Korea

Abstract

This paper presents the results of a study using Twitter lists to infer the characteristics of users, especially about their interests. Gathering and structuring users' interest has been challenging because it often requires expensive data such as users' history logs or user surveys. Our approach overcomes this limitation and acquires informative words representing user interests by using Twitter data which is open to the public and analyzing it with the χ^2 feature selection algorithm. Furthermore, by using the tweets of all the users in a Twitter list, we can discover latent user characteristics that are not present in the tweets of individual users. We crawled Twitter list data that includes about ten percent of the Twitter user population, the lists they belong to, and the tweets of all the members of the lists. We used a standard feature selection algorithm and a supervised classification algorithm to verify the semantic coherence of the lists. Then we conducted a user survey which confirmed that lists serve as good groupings of Twitter users with respect to the perceived characteristics of the users. The user survey also confirmed that the words extracted from each set are representative of all the members in the list whether or not the words are used explicitly by those members.

1. Introduction

Twitter[1], a popular microblogging service, has recently added a new capability for users to create *lists*[2] 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". Predicting the set of related words for a user serves a goal similar to that of previous studies that have looked at ways to gather more information about users to develop and improve applications such as personalized web search (Teevan, Dumais, and Horvitz 2005), recommender systems (Shepitsen et al. 2008), and social search (Carmel et al. 2009). These previous studies use histories of user activities on the web or the desktop to model the user. While that approach may be fine in some settings (e.g., intranet within a company), in many cases it may not be preferred, or it may be downright impossible, to collect such data. 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.

Although Twitter has been in service for a few years now, the list functionality is very new. It was announced to the Twitter public in November of 2009, just one month before this study, so our data, analyses, and results reflect the premature state of the problem. Nevertheless, we were able to collect enough data, carry out sound analyses, and complete a user study to generate insightful and meaningful results. Using our Twitter crawled data of just three million Twitter users (about 10% of total users), we counted almost one million lists created, and almost four hundred thousand users who are in at least one list. See Table 1 for exact numbers. Our analyses of the list data are based on traditional machine learning approaches, looking at classification and feature selection methods. Using just simple methods and basic assumptions, the results of our analyses show that, at least for a significant subset of the Twitter lists, simple classification methods work well, and basic feature selection algorithm is effective in discovering words that are representative of the various lists. The most challenging aspect of this research was in evaluating our claim that the words found by the lists actually represent accurate information about the users in those lists. We designed and carried out a user study to test the relationship between the words and the users, and the results of this preliminary study validate our claim.

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

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

[2] http://help.twitter.com/forums/10711/entries/76460

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.

This paper is organized in the following structure. Section 2 discusses related research on Twitter, user profiling, and personalization and groupization. Section 3 provides descriptions about the current usages of twitter lists, i.e. basic statistics and qualitative analysis on the characteristics of the twitter lists. Section 4 explains the relations between twitter lists and tweets by extracting features explaining the characteristics of twitter lists. Section 5 demonstrates the effectiveness of the extracted features through machine learning experiments. Section 6 verifies the validity of the extracted features through user surveys. Finally, Section 7 discusses gained insights from the machine learning and the user studies as well as the future applications of the features.

2. Related Work

We will situate our work in four related research areas: research on microblogging, research on user profiling using web search and desktop behavior, research on using social media for user information, and research on applying user profiles for personalization and groupization of social search.

Twitter is a popular microblogging service, and although it is fairly new, it has been studied by quite a few researchers because of its unique characteristics of short messages, unidirectional relationships among users, and frequent realtime updates via mobile web. Work by Java, et al. (Java et al. 2007) was one of the first, and it presents their early findings about general usage of tweets and the social network on Twitter. A recent paper by Naaman, et al. (Naaman, Boase, and Lai 2010) presents an analysis of tweets according to different categories, and based on those categories, classification of the Twitter users into *meformers* and *informers*. Cheong and Lee (Cheong and Lee 2009) have conducted interesting research on the Twitter trends to retrieve *collective intelligence* using tweets and the demographics of the authors of those tweets. Besides these, there is a growing literature on Twitter studies, and a well-updated list is at danah boyd's website[3].

There is good prior research on inferring about users by tracking their web and desktop behaviors. Early work by Claypool, et al.(Claypool et al. 2001) looked at a user's clickthrough behavior to infer what he is interested in. A much more comprehensive user profile, was made by Teevan (Teevan, Dumais, and Horvitz 2005), utilizing desktop activities as well as web search activities. These and many others were able to construct good user models, but only for users with enough personal data, and this poses a serious problem for use with general web search because it is not possible to collect such detailed data for many web users. To overcome that problem, a recent study (Xue et al. 2009) proposed building user language models by individual users, user groups, and global users, and using those language models for collaborative personal search. It still has the problem that this type of user clickthrough data is not easily available and may cause issues with users who are concerned about privacy.

An alternative to using clickthrough or desktop behavior is using social media such as blogs and social annotation sites. Blogs can be used in interesting ways for research on user characteristics by using the contents of the blog entries as well as social network data gathered from permalinks and comments. Besides finding out which topics a blogger tends to write about, there are interesting twists such as predicting a blogger's personality (Nowson and Oberlander 2007) based on his blog contents. Another interesting paper using blogs is Nitin et al. (Agarwal et al. 2009), which finds *familiar strangers*, bloggers who are in close distance and are interested in similar topics. Social annotations sites, such as *del.icio.us*[4] and *Flickr*[5], also offer a large amount of useful data for finding out more about users. On those sites, users bookmark websites and images, and tag them with words and phrases that describe those resources. Michlmayr et al. (Michlmayr and Cayzer 2007) have looked at learning user profiles by looking at the bookmarks and the corresponding tags, and then using those profiles to match the webpages that are more relevant than general web search. Xu and colleagues (Xu et al. 2008) have looked at the the interrelationships among users and webpages, which can be inferred from their bookmarks and tags, for personalized search. All of these are good motivations for this work, in which we infer user characteristics from the tweets of the users in the same list.

In the applications domain, we are motivated by personalized social search, as defined by (Evans and Chi 2009) and (Carmel et al. 2009) and inspired by insights about collaborative search (Golovchinsky, Qvarfordt, and Pickens 2009) and social information seeking (Chi 2009). A good example of personalized social search within an Intranet setting is illustrated in (Teevan, Morris, and Bush 2009) which uses detailed information such as users' work task groups and desktop activities. Our work suggests a solution to an integral part of such a system–inferring detailed information about users–for the general web by using a very simple dataset, easily gathered from Twitter.

3. Twitter Lists

In Twitter, if user *A follows* user *B*, then *B* is *A*'s *friend*, and whenever *B* writes a short message, called *tweet*, on Twitter,

[3] http://www.danah.org/TwitterResearch.html

[4] http://www.del.icio.us

[5] http://www.flickr.com

Total Number of Users Checked	3,332,509
Number of Users in At Least One List	398,455
Number of Unique Lists	909,597
Number of Unique List Names	322,873

Table 1: Basic statistics about our Twitter lists data.

it appears on *A*'s timeline. Twitter users can be individual persons, celebrities, organizations, news media, and other miscellaneous categories. Many Twitter users follow hundreds or even thousands of users, and thus the number of tweets in such user's timeline can be overwhelmingly large. Thus, grouping one's friends into lists can alleviate a potentially serious problem of information overload. A user can add to his lists any other Twitter users including her/himself, even the ones he is not following. A user can define lists such as "food", "golf", or "mathematicians", and he can click on one of those lists to view the tweets by a subset of his friends (and non-friends) designated by that list. Once a list is created, the list can be shared with other users, so that other users do not need to create the same list of their own. In this section, we explain how we crawled Twitter users' data and how they use the list function.

3.1 Dataset

In order to collect list information, we started crawling from the followers of Ashton Kutcher, one of the most popular people on Twitter who has more than 4 million followers. And then we crawled users' profiles and lists information from his follower list. Because many users make their account and never use again, we defined a user who has at least one friend as an active user. Finally, 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[6]. 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. A list name reflects the meaning of the list for the user who created it, The majority of the list names are created by a simple word or dash split concatenation of two or three words. The analysis shows that a large number of lists share the same names, like friends, news and music. Table 2 shows the top 20 list names and their frequencies. Besides these simple words, there were a lot of unique list names. To get the total number of names used for lists, we removed duplicate names. The number of unique list names are approximately one third of the total number of the crawled lists. Details are described in Table 1.

Table 3 compares the average number of followers and followees of all users versus the users who belong to at least one list. The users who are in at least one list have significantly larger numbers of followers/followees compared to all users. The number of followers shows the largest difference, we conjecture that this is because the list is a new feature, and so users use lists for grouping celebrities rather than ordinary people. The maximum capacity of a list is set to 500, and in our dataset, the average size of the lists is 36.7.

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

list name	freq	list name	freq
friends	31,267	politics	3,078
news	15,216	design	1,866
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 2: Frequencies of top 20 list names in Twitter.

	All User	In Lists
Avg. # of following lists per user	0.83	6.98
Avg. # of followers per user	98.97	1595.31
Avg. # of followees per user	107.59	551.11

Table 3: Basic statistics about Twitter network

3.2 Usage of List

Twitter, unlike many other social network services (e.g., Facebook[7]), does not require mutual consensus to make a connection between two people. Users just click the following button to make a connection with others, and this makes Twitter act as a news media service, in addition to a social network service. Therefore users who want to find more valuable information follow many other people, and thus their timelines may overflow with a flood of tweets. These situations lead users to group other users according to different categories. In our collected dataset, users use lists in various ways, and we found several different categories of lists.

- Celebrities(e.g. celebrities, famous, ...) - One of the reasons that make Twitter popular is the existence of celebrities and their enthusiastic activities. The lists in this category take a large portion of the overall lists.
- Friendly Relationships(e.g. friends, amigos, ...) - As we mentioned before, users might miss tweets of their friends and family due to the flood of tweets. The list functionality prevents this from happening.
- Organizations(e.g. microsoft, google-employees, KAIST, MIT, ...) - This group of lists consists of companies, organizations, education facilities, etc. They use Twitter for public relations.
- Interests(e.g. music, game, ...) - This category is used to organize people who have similar interests or expertise in specific areas.
- Geographic Information(e.g. Korea, Hollywood ...) - Geographic locations also can be one way of organizing lists.

[7]http://www.facebook.com

(a) U.S. Cities

List Name	Frequency
hollywood	495
chicago	322
seattle	173
austin	167
boston	167
atlanta	164
vancouver	161
los-angeles	137

(b) World Countries

List Name	Frequency
canada	124
iran	108
australia	105
china	103
japan	80
mexico	77
india	72
venezuela	62

Table 4: The number of lists that represent U.S. cities and world countries.

Here we used *gazetteer*[8] to extract country names and city names. Table 4 shows the most frequently used city and country names.

Besides the above categories, users use lists in many different ways. Some of the lists, like "conversationlist", are made by a 3rd-party program automatically[9]. This list consists of people that the list owner mentioned or replied to recently. We can also infer information about lists that was not intended by the list owner. For example, a user creates a list 'MIT' as an organization, but we can infer geographical information because we know that 'MIT' is in Cambridge.

In this paper, we will focus only on organizations, interests and geographic information groups, since there are few topics shared by the people in the first two lists, celebrities and friendly relationships.

There is not a clear rule for classifying lists into these categories, and, we cannot assert their purposes from their names. However, it is possible to find lists that are similar. In Section 5, we will explore a method of finding similar lists using classification techniques.

4. Lists and Terms

In the previous section, we looked at the basic statistics of lists and their usages. 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 χ^2 feature selection(Manning, Raghavan, and Schütze 2008), 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. Note that, we only crawled the recent 3,200 tweets instead of the entire history of tweets due to the limitation of the Twitter API. After crawling the data, we selected the lists of interest, because the entire set of lists are too many to compute. We set the following guidelines to choose the candidate lists.

- The selected lists should not contain users who are celebrities or related with advertising.
- The selected lists represent common interests.

[8]http://world-gazetteer.com/

[9]http://conversationlist.com/

(a) Author

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

(b) Coffee

Word	Value
sumatra	439
roaster	413
peaberri	343
sidamo	294
guatemala	293
barista	282
kopi	270
luwak	270
k-cup	252
arabica	242

(c) Cycle

cancellara	404
boonen	385
hincapi	383
peloton	383
cadel	364
velonew	356
interbik	355
wiggin	345
schleck	335
mtb	330

(d) Fitness

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

(e) Food

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

(f) Game

virtua	176
x360	176
destructoid	176
splosion	175
gaiden	171
nxe	162
hmv	159
harmonix	158
metroid	157
cutscen	156

(g) Mom

prayersforanissa	287
blissdom	201
anissa	199
mommyblogg	182
stroller	166
wahm	165
blogher09	165
gno	165
blogworld	157
breastfeed	155

(h) Photograph

e-sess	243
partnercon	225
tradeshow	223
ppa	222
twitterless	215
35mm	197
kubota	194
5dmkii	190
shootsac	188
whcc	188

(i) Swim

swimmer	491
fina	399
locht	375
peirsol	375
breaststrok	343
grever	282
lzr	249
scm	220
3413	218
kukor	218

(j) Tech

wpf	658
asp.net	626
telstra	503
sharepoint	492
azur	471
vs2010	470
mvc	456
autech	454
pdc	453
ie8	442

Table 5: Top 10 words and χ^2 values

We collected 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. After selecting the lists of interest, we applied the *Porter stemmer*[10] to consolidate the word counts of variations from a single word into one count. We also removed mention tags, mentioned user-ids and hyperlinks.

As shown in previous research(Java et al. 2007), 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 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. A word's χ^2 value is defined as follows.

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

$N_{e_t e_c}$ is count of documents that have the values of e_t and e_c. e_t is equal to 1 when the document contains term t, and e_c is equal to 1 when the document is in class c. N is the observed frequency in the corpus D, and E is the expected frequency. E_{11} is the expected frequency that t and c occur together in a document.

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 χ^2 value of each word in the selected groups of lists. Table 5 shows the top 10 χ^2 values and words in each group. Intuitively, most of the words explain their list well. *k-cup*, *wahm* (abbrev. for Work-At-Home-Mom), and *35mm* are example terms that make sense to users in the lists that each word comes from, but may not be known by other users. *Tech* and *game* groups contain many such words, which are mostly names of mostly proper nouns, such as names of games, developers, and software. Several community names appear in the table, since many people in the lists are connected through those communities, for example, *nanowrimo*, *whcc*, etc. There are many other interesting words that are not listed in the table. For example, *h1n1* obtained a high χ^2 value in the *mom* group, mainly because the side effects of the H1N1 vaccine to kids were a hot issue recently.

We also calculated the χ^2 values of celebrity groups and lists that are not likely to contain common interests, such as *conservationlist* which is an automatically created list for each user, reflecting who he has had conversations with recently. Compared to the 10 groups we have specified as interest groups, these lists do not show an interesting result. Top words from *conversationlist* are *xcode, noch, mail*, etc., and these words do not belong to a common interest. Also, the average χ^2 value of top 10 words in *conservationlist* group was 115.72 which is considerably lower than the average χ^2 value of the 10 groups, 310.58. The χ^2 value is a good indicator of the discriminating power of a word with respect to a document in a corpus

[10]http://tartarus.org/ martin/PorterStemmer

Keyword	Precision	Recall	F-Measure
Mom	0.552	0.687	0.772
Game	0.985	0.708	0.824
Photo	0.984	0.719	0.831
Food	0.952	0.588	0.727
Author	0.861	0.626	0.725
Tech	0.980	0.671	0.797
Cycle	0.952	0.675	0.790

Table 6: Machine learning model performance for classifying Twitter users into a keyword category.

5. Classifying Lists

We perform a machine learning experiment to classify new lists into a keyword category with tweet contents from the training lists. Specifically, we hypothesize that the set of words we found using the χ^2 analysis represents the distinct characteristics of respective Twitter list. To test this hypothesis, we use a machine learning model identifying Twitter users who seem to be related to a certain interest groups, e.g, *computer gaming*, and the model is trained with the bag-of-words for each list. With the trained model, we classify Twitter users belonging to the various lists, and we can infer similar lists about a keyword when the majority of the users in the list are classified positive. This experiment is done by using a SVM classifier using 500 words with the highest χ^2 values.

First, we evaluated the performance of a learning model trained on the bag-of-words using the χ^2 feature selection with 10-fold cross validation. Table 6 shows the precision, the recall and the F-measure statistics from the training dataset. Overall, the learners have high precision and mediocre recall rates. In all of the keyword categories, the users were classified with above 85% of precision. The recall rate is relatively low because the number of users in a certain list is very small compared to the entire user set. To increase recall, the classifiers have to sacrifice precision, but our fundamental goal here is accurately identifying a Twitter list with user classification, so the presented high precision is more satisfactory than higher recall. Also, it should be noted that the bag-of-words alone can produce a satisfactory precision for this problem.

After examining the classifier performance, we applied the classifier to the Twitter users in the test lists that are not used for training. Table 7 displays the sizes of the test lists and the percentages of users classified positive by the classifiers for keyword. For instance, we trained a classifier to identify Twitter users who are knowledgeable about "photo", and we applied the classifier to the users belong to 11 test Twitter lists. Among the 11 test Twitter lists, one Twitter list is identified by the list about "photo", and the remaining ten lists are not. When we apply the classifier

(a) Test Twitter Lists about the Keyword

Keyword	Affiliated Users	Users classified Positive	Avg. %
Mom	33	16	48.48%
Game	51	16	31.37%
Photo	13	9	69.23%
Food	25	10	40.00%
Author	91	54	59.34%
Tech	212	101	47.64%
Cycle	81	55	67.90%

(b) Test Twitter Lists not about the Keyword

Keyword	Affiliated Users	Users classified Positive	Avg. %
Mom	473	8	1.69%
Game	455	5	1.10%
Photo	493	7	1.42%
Food	481	1	0.21%
Author	415	1	0.24%
Tech	294	1	0.34%
Cycle	425	0	0.00%

Table 7: Classification results of users from the test Twitter lists.

to the list about "photo", the classifier resulted that 9 users, 69% of the number of members in the list, are positive in showing interests in "photo". Then, we can classify that the theme of the Twitter list is photo. Unlike the classification of the positive Twitter list, the other 10 lists show that only 1.42% of the list members are classified positive, and this will result that none of the Twitter lists will be identified as photo-related Twitter lists. This list classification demonstrates the potential of using bag-of-words from tweets of existing Twitter lists to understand the characteristics of new Twitter lists with high precision.

6. 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 (high χ^2 words) are informative characteristics of users.

6.1 Method

The survey was conducted through a website[11] with 37 subjects whose ages range from 20 to 32. 25 subjects are graduate students or undergraduate students in the CS department at our university, and the rest are graduates.

A subject is given ten words and hyperlinks to three Twitter users, and for each word, the subject is asked to choose

[11]http://uilab.kaist.ac.kr/research

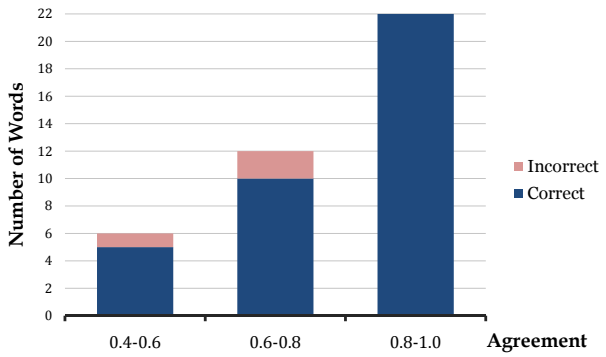

Figure 1: User survey result. Red color represents incorrect words and blue color correct words.

Incorrect Word	osp	taper	kenya
Subjects' Answer	tech	fitness	author
Our algorithm	photograph	swim	coffee
Subjects' Agreement	0.44	0.60	0.67

Table 8: Three incorrect words and subjects' agreements

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. (Stemming was not used in this survey, since it often distorts words and recovering the original forms is not a trivial task.) For each list, one Twitter user that the list follows 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. They are also asked to specify where they found hints among *the user's ID*, *the user's tweets*, *other users' tweets*, or *other*. The subjects were allowed to look up a dictionary or to search the Web. The complete survey form is available at our research website[11]. Through this survey, we sought to evaluate how well a list's high χ^2 words can be applied to a randomly chosen user in the list.

6.2 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. Figure 1 illustrates the result of the survey. For each word, agreement is defined as the largest proportion of the subject answers. The graph shows that the forty high χ^2 words that were used in the survey have agreement values of at least 0.4, and thirty of them have agreement values of at least 0.67, which means two thirds of human decisions hold consensus. The most agreed

Word	Agreements	Meaning in Context
descenza	0.40	swimmer's name
osp	0.44	website name
schafer	0.44	game desiner's name
prayersforanissa	0.50	slogan
azure	0.53	computer system name
mythic	0.53	game developer's name
artisan	0.60	food community name
popcap	0.60	game company name
preschool	0.60	common word
taper	0.60	technical term in swimming

Table 9: 10 words with the lowest agreements. Agreement is defined as the largest proportion of the subjects' answers. The third column describes how each word is used in context.

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. In Figure 1, incorrect words are represented with red color(light). Table 8 lists the incorrect words and their user agreement values. The user agreement values are equal to or lower than 0.67, meaning that the words' right answers are not obvious to humans as well. Looking at how the incorrect words are used in context helps understand why they have low user agreements. In the original dataset, the word *osp* represents *opensourcephoto.net* website, which was mentioned by people in the *photograph* list. As *osp* is both a community name and an abbreviation, most subjects could not catch the correct meaning. A *taper* or *tapering* in swimming means the reduction of workload prior to a competition. In the dataset, as *taper* is used as a technical term, the subjects had trouble identifying the meaning of the word in the case they could not find this word in the tweets of the given Twitter users. For the word *kenya*, many chose the *author* category instead of *coffee*, and their reason was *the user's tweets*. This may be because the given Twitter user was writing a book regarding travels and the tweets included several country names and travel terms. It turns out that such very specific words as *osp* are difficult to generalize to all the people in the list, in which case the χ^2 feature selection is not appropriate; on the other hand, the χ^2 feature selection is good at finding such technical terms as *taper* that are applicable to all the people in the list, whereas it is not easy for ordinary people to infer those terms.

Ten words with the lowest agreement values are listed in Table 9. Seven words are names, and only one is a common word. Because our subjects showed low agreement for these words, we cannot judge how relevant these words are for the respective lists. To evaluate the performance of our algorithm on these words, we need another way to test how representative these words are of the users in the respective lists. The best way would be to ask the users in the lists directly whether they are interested in those words, so we conducted another survey that asked the members of the lists whether they have recently searched the web for the words. We sent out tweets containing the request "Please pick words that you have recently searched for" along with the words from the list that he is in, and also words from some of the other lists. We sent out the tweets to five hundred people in the lists, but we only got eight replies. Five out of the eight users responded by saying that they have searched for at least one of the words for the correct list, and three users responded by saying they have not searched for any of the words. Although this survey does not provide a decisive evidence due to the small number of responses, the result is in favor of our hypothesis. A new design of the survey is left for future work.

Tweets	Visible	# total words	# correct words
Yes	Yes	4	4
Yes	No	4	4
No	No	32	29

Table 10: The survey result categorized by whether or not words are explicitly used by the given Twitter users. The first column indicates if tweets contain words. The second column indicates if words are visible to the subjects, that is, words occur within the latest 200 tweets.

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. About a third of 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 requires either manual effort or a more complex model to extract information computationally. Although our algorithm uses only tweets, which are full of slangs, abbreviations, onomatopoeias, and shortened words, it works as well as human judgements. Thus, the algorithm can be applied to data that have no summary or meta data available. Table 10 categorizes the survey result according to whether words were used explicitly in tweets. In most cases, the words of questions were not used in the tweets of the given Twitter users. Regardless, we managed to find, from those users, informative words on which subjects made high agreements and our algorithm reached a high accuracy by hitting 29 words out of 32. For those words explicit in the tweets, our algorithm made the same decisions as the subjects.

7. Discussions and Future Work

We have discussed in 5 and 6 how we confirmed the representative power of the χ^2 words through machine learning techniques and the user study. First, we showed that the χ^2 words are good features to classify users and lists according to user interests. Second, 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. Though we do not cover in this paper, actually a user could belong to multiple lists. In this case, there would be multiple sets of characteristics for a given user.

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. The followings are possible applications of the results of our study, and other 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. Prior work in social search can benefit from the result of our research because for a given query term, we can match it with the χ^2 words to identify an appropriate list that would contain the most useful "friends" for that search.
- *Expert Recommend System* A list consists of users who have similar interests or expertise. Considering inter-list and intra-list structures and list itself together would be useful to find experts on Twitter.
- *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.

Acknowledgement

This paper is the longer version of CHI workshop paper with the same title. The original paper is presented at the workshop on microblogging at the ACM conference on human factors in computer systems (CHI2010).

References

Agarwal, N.; Liu, H.; Murthy, S.; Sen, A.; and Wang, X. 2009. A social identity approach to identify familiar strangers in a social network. *Proceedings of International Conference on Weblogs and Social Media (ICWSM 2009)* 1–8.

Carmel, D.; Zwerdling, N.; Guy, I.; Ofek-Koifman, S.; Har'el, N.; Ronen, I.; Uziel, E.; Yogev, S.; and Chernov, S. 2009. Personalized social search based on the user's social network. *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*.

Cheong, M., and Lee, V. 2009. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining*.

Chi, E. H. 2009. Information seeking can be social. *IEEE Computer* 42(3):42–46.

Claypool, M.; Brown, D.; Le, P.; and Waseda, M. 2001. Inferring user interest. *IEEE Internet Computing*.

Evans, B. M., and Chi, E. H. 2009. An elaborated model of social search. *Information Processing and Management* 1–23.

Golovchinsky, G.; Qvarfordt, P.; and Pickens, J. 2009. Collaborative information seeking. *IEEE Computer* 42(3):47–51.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. 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*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. Introduction to information retrieval. 482.

Michlmayr, E., and Cayzer, S. 2007. Learning user profiles from tagging data and leveraging them for personal(ized) information access. *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)* 1–7.

Naaman, M.; Boase, J.; and Lai, C.-H. 2010. Is it really about me? message content in social awareness streams. *CSCW2010*.

Nowson, S., and Oberlander, J. 2007. Identifying more bloggers. *Proceedings of International Conference on Weblogs and Social Media (ICWSM) 2007* 1–7.

Shepitsen, A.; Gemmell, J.; Mobasher, B.; and Burke, R. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*.

Teevan, J.; Dumais, S.; and Horvitz, E. 2005. Personalizing search via automated analysis of interests and activities. *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.

Teevan, J.; Morris, M.; and Bush, S. 2009. Discovering and using groups to improve personalized search. *Conference on Web Search and Data Mining*.

Xu, S.; Bao, S.; Fei, B.; Su, Z.; and Yu, Y. 2008. Exploring folksonomy for personalized search. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*.

Xue, G.-R.; Han, J.; Yu, Y.; and Yang, Q. 2009. User language model for collaborative personalized search. *Transactions on Information Systems (TOIS* 27(2).