Reading Tweeting Minds: Real-time Analysis of Short Text for Computational Social Science

Zhe Wang
University of Cambridge
United Kingdom
zw267@cantab.net

Daniele Quercia
Yahoo! Research, Barcelona
Spain
dquercia@yahoo-inc.com

Diarmuid Ó Séaghdha
University of Cambridge
United Kingdom
do242@cam.ac.uk

ABSTRACT

Twitter status updates (tweets) have great potential for unobtrusive analysis of users' perceptions in real time, providing a way of investigating social patterns at scale. Here we present a tool that performs textual analysis of tweets mentioning a topic of interest and outputs words statistically associated with it in the form of word lists and word graphs. Such a tool could be of value for helping social scientists to navigate the overwhelming amounts of data that are produced on Twitter. To evaluate our tool, we select three concepts of interest to social scientists (i.e., privacy, serendipity, and Occupy Wall Street), build ground truths for each concept using the Grounded Theory approach, and perform a quantitative assessment based on two widely-used information retrieval metrics. To then offer qualitative assessments complementary to the quantitative ones, we run a user study involving 32 individuals. We find that simple information-theoretic association measures are more accurate than frequency-based measures. We also spell out under which conditions these metrics tend to work best.

Keywords

Grounded Theory approach, Association Measures, Concept Extraction, Linguistic Evaluation

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Human Factors

1. INTRODUCTION

"Computational social science" is a new discipline that aims at using large archives of naturalistically-created behavioral data (e.g., emails, tweets, Facebook contacts) to answer social science questions [5]. However, in using real-time web data, one faces a number of challenges, and here we investigate a specific one: how to use Twitter to understand people's beliefs about a variety of social issues in real-time (e.g., privacy and serendipity).

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France

One way of interpreting what people mention on Twitter is to analyze the use of language in their status updates. Analyzing the use of language in digital archives is an established area of interest in computer science [7, 6] and a growth area in social science [5]. However, in these communities, researchers either have worked with properly structured archived texts (e.g., academic papers, news articles) [4] or have not been concerned with real-time analysis. To address these limitations, we design and implement a tool for extracting associations that Twitter users make when mentioning specific keywords of interest. Given a keyword related to an abstract concept (e.g., serendipity), our tool identifies the statistically relevant word associated with it, returns the words as a list ranked by degree of relevance, and provides a word graph showing the different facets of the concept.

We apply our tool to three Twitter streams extracted for three representative topics that have been studied by social scientists: privacy[2], serendipity [8], and the Occupy Wall Street movement. We make three main contributions:

- We consider an existing technique called PMI (Point-wise Mutual Information), which is able to extract word associations in *static* data archives (Section 2). To investigate people's perspectives about a keyword over a period of time and to counter the limitations of data collection facilities of Twitter, we engineer PMI and make it work on *real-time* data - we call this adaptation RT-PMI (Real-time PMI). We then incorporate it into a concept extraction tool.
- Using the three Twitter streams, we evaluate the tool based on two information retrieval metrics (Section 3). To this end, we have generated gold standards in the form of sets of words that result from hand-coding relevant literature. We show that RT-PMI outperforms the competitive baseline of frequency analysis, which identifies relevant word associations simply based on the frequency with which the keyword is associated with other words.
- We finally prototype a visualization interface that shows word lists and word graphs for each of the three keywords (Section 4). We evaluated it by having 32 participants assess the interface's usefulness and helpfulness. As one expects, compared to frequency analysis, RT-PMI has been found to produce both lists of word associations that are more accurate and word graphs that more effectively reflect the different facets of each keyword.

The main contribution of this paper is not technological but methodological. This is a preliminary study that investigates whether applying simple information theory measures to very short pieces of text can give useful insights for computational social science. In so doing, we compare quantitative assessments with qualitative ones

and test the extent to which information retrieval metrics reflect actual user experience.

2. OUR PROPOSAL

Our goal is to extract conceptual associations that people make when discussing a specific topic, under the assumption that these associations can be mined from the contexts in the topic it is mentioned. For instance, the keyword "serendipity" might co-occur with "accident" and "fortunate" (reflecting its unexpectedness) and with"pleasant" and "happy" (reflecting its enjoyability).

Towards the goal of extracting meaningful word associations, we build a tool that consists of three layers: 1) *Real-time unstructured data manager*; 2) *Association extraction module*; and 3) *Application layer*, which will be introduced as below.

2.1 Real-time Unstructured Data Manager

The data manager collects unstructured data and cleans it for word association extraction. The pre-processing includes tokenization, filtering stop words and non-English tokens. Since the tool is tailored to the analysis of real-time and big streams of data, we have collected English tweets (Table 1) from the public Twitter API using the university's servers. The choice of tweets is motivated by two main reasons. First, tweets are *real-time* reports of what millions of people around the world are seeing, feeling and doing, and its real-time feature could dynamically captures the public's attitudes towards a social issue. Secondly, the *massive* amount of tweets (in last October, there were 250 million tweets per day) may reveal patterns of individual and group behavior with unprecedented breath and depth [5].

Although automated text analysis of tweets suffers from data sparsity [3], it is unclear to which extent the length of a tweet will impact the extraction of meaningful conceptual associations. Intuitively, the more usable words, the better the word associations extracted. To capture this intuition, we define the following metric: *number of usable words per tweet after pre-processing*, including removing stop words (e.g., "if", "the"), user names, hashtags(e.g., "#privacy") and URLs. In Section 3, we will test how the accuracy of word associations one can extract varies with the number of usable words: we can do so because the data streams associated with the three keywords under study (i.e., privacy, serendipity, and Occupy Wall Street) conveniently differ on *number of usable words per tweet* given similar sample sizes (Table 1).

2.2 Association Extraction Module

Given a collection of unstructured data as input, the association extraction module determines the relevance of each word to the target keyword. To quantify the abstract concept of relevance, we employ *statistical association measures* . Given a keyword (e.g., serendipity), association measures assign a score to each associated word: a higher score corresponds to better relevance. Two widely-used association measures are *frequency analysis* and *PMI analysis* (Point-wise Mutual Information).

Frequency analysis. This is the easiest approach to extract word associations. The association score is simply the observed frequency of two words' co-occurrence. Its use is motivated by the assumption that associated word pairs will in general occur more frequently than arbitrary combinations by chance.

The disadvantage of frequency analysis is that it is unreliable in the presence of irrelevant content or noise. For example, say we want to investigate the nature of "serendipity". To this end, we should collect tweets containing the keyword "serendipity". When doing so, besides the tweets about people's serendipitous experiences, we found a lot of tweets about the advertisements of serendipitous products like gift cards and about a recent movie titled "Serendipity"

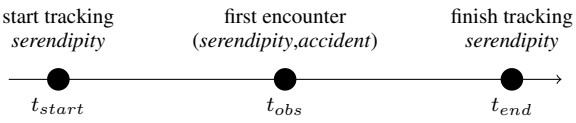


Figure 1: Example timeline: for rapidly-changing data streams, it is very difficult to go back in time.

(24.1% of tweets were about the movie). That is why frequency analysis gives the highest ranking to the word "movie" and does not rank among the top-10 any word about the nature of "serendipity", such as "chance" and "fortunate".

PMI analysis. To overcome this problem, we consider Point-wise Mutual Information (PMI). For a word pair (w_i, w_j), PMI returns an association score: $score(w_i, w_j) = \log(\frac{1}{p(w_i,w_j)}p(w_i)p(w_j))$, where $p(w_i, w_j)$ is the probability of w_i and w_j occurring together, and $p(w_i)$ and $p(w_j)$ are the respective probabilities of w_i and w_j co-occurring with any term. PMI analysis compares the observed co-occurrence probability of two terms to the chance probability of their co-occurrence given a unigram model of independent interactions. One considers that words relevant to a keyword will occur with it far more often than would be expected by chance. Take again the analysis of the keyword "serendipity". Although the word "movie" co-occurs with "serendipity" most frequently, it is also a word of high frequency (very common) in general. After performing PMI analysis (which discounts for general frequency), the highest ranked word becomes "fortunate" and the ranking of the word "movie" drops down.

In the standard form of PMI, the three probabilities, $p(w_i)$, $p(w_j)$, and $p(w_i, w_j)$ are calculated as: $p(w_i, w_j) = freq(w_i, w_j)/N$, $p(w_i) = freq(w_i)/N$ and $p(w_j) = freq(w_j)/N$, where $freq(w_i, w_j)$ is the number of tweets containing the word pair (w_i, w_j), $freq(w_i)$ and $freq(w_j)$ are the number of tweets containing either w_i or w_j, and N is the total number of co-occurrences and is constant for a given corpus.

However, standard PMI scores are difficult to compute when applied to dynamically changing big data streams that cannot be processed in batch, which is our scenario. Let us consider the example case that we are tracking co-occurrences for the term "serendipity" between a start time t_{start} and an end time t_{end} (Figure 1). At an intermediate time t_{obs} we observe the word "accident" co-occurring with "serendipity". In order to calculate the PMI association between "accident" and "serendipity" we need to know the marginal co-occurrence probability $p(accident)$ between t_{start} and t_{end}. Nevertheless, the "search back" for real-time big data is difficult, sometimes impossible, not only due to data being massive, but also due to throttled data collection facilities. For example, although there are around 250 million tweets posted every day, one can use Search API to query only for tweets posted in the recent past (for up to $1,500$ tweets within the last 7 days). This leads to a problem: if an associated term appears for the first time a week or more after the start of data collection, it would be impossible to estimate the marginal frequency of that term over the period (t_{start}, t_{end}).

Instead of collecting the complete tweets containing "accident", we can alternatively crawl a general sample of tweets and count *freq*(accident) in that sample. However, to accommodate this solution, we reformulate the standard definition of PMI as what we call RT-PMI (*Real-time-PMI*). The derivation starts with rewriting the standard PMI formula as: $score(w_i, w_j) = \log(\frac{1}{p(w_i)}p(w_i|w_j))$. Then, the conditional probability $p(w_i|w_j)$ can be estimated as: $p(w_i|w_j) = freq(w_i, w_j)/\sum_k freq(w_k, w_j)$, where $freq(w_i, w_j)$ is the number of tweets containing w_i and w_j posted in a given

	Keyword-specific corpora			General Sample
	Privacy	Serendipity	Occupy Wall Street	
Time Span	Jan-Mar, 2012	Jan-Mar, 2012	Jan-Mar, 2012	Feb-Mar, 2012
#tweets	199,627	192,108	251,673	3,232,350
#distinct words	7,539	7,338	1,833	18,510
#usable words per tweet (after pre-processing)	6.7	5.5	4.7	-
#non-English tokens per tweet	1.5	1.3	3.8	-

Table 1: Statistics of the Twitter datasets.

time span, say, $[t_1, t_N]$ in Figure 1, and $\sum_k freq(w_k, w_j)$ is the frequency of all the words co-occurring with the keyword w_j, e.g., "serendipity", during $[t_1, t_N]$. Practically, $freq(w_k, w_j)$ is counted in a topic-specific corpus, which could be collected by querying "serendipity" with the Twitter APIs, and $p(w_i)$ is estimated in a random sample of all the tweets that were posted during the same time period in which the topic-specific corpus was gathered (e.g., $[t_1, t_N]$ in Figure 1). We call this random sample of tweets the "general sample". That is because the sample was collected with the so-called *Sample method* available on Twitter *Streaming API*, which guarantees that the sample is random. The statistics of the general sample of our Twitter datasets are reported in the last column of Table 1. In our evaluation, we will show the extent to which the general sample works as a good approximation of what has been mentioned on the Twitter platform in general. It is important to point out that the two forms of PMI, RT-PMI and standard PMI, are very similar: RT-PMI is simply a convenient transformation of the standard PMI, which relaxes the estimation of $p(w_i)$ in a general sample. This relaxation allows us to compute PMI scores in real-time data streams, extending the applications of a standard PMI approach when the data collection facilities are throttled.

Going beyond Twitter, RT-PMI could also be used to extract word associations in other online data sources. For example, if we want to investigate the word associated with "privacy" on web blogs, then the keyword-specific corpus can be collected by searching "privacy", and the corresponding general sample can be collected by randomly sampling all web blogs.

2.3 Application Layer

The third and last component of our tool is the application layer. Here we build an application that allows users to visualize a word list and a word graph associated with each of the three keywords. For example, for our tweets mentioning the keyword "serendipity", the application shows the top-10 words are ranked by their *association scores* (either frequencies or PMI scores) with the keyword (Figure 2(a)): the word "fortunate" has the highest ranking, which indicates that it is the most relevant word in Twitter users' mentions of "serendipity". The application also visualizes a word graph (Algorithm 1). Furthermore, we find that if the two words occur together more than would be expected by chance (the PMI score exceeds a threshold t), they are likely to be grouped together to shape one aspect of a keyword. For instance, in Figure 2(b), we can see that the graph clusters the words "fate-coincidence-destiny" together - these words seem to reflect the out-of-control aspect of "serendipity".

3. QUANTITATIVE EVALUATION

3.1 Gold Standards

To evaluate the tool in a quantitative way, one needs to compare the tool's output to a human-created gold standard reflecting subjects' prior understanding of a topic. For example, words like "fortunate" and "accident" co-occurring with "serendipity" may reveal the unexpectedness aspect of serendipity in people's minds. To build such a gold standard, for a given keyword we use the *Grounded Theory* approach [1], which is a systematic framework in the social sciences involving theory-driven content analysis and provides us with a systematic way to minimise the bias due to differences in comprehension among annotators. More specifically, we use *line-by-line* coding, which generates a set of words conceptually associated with the keywords, and the procedure unfolds in three steps:

1. Collecting documents. The gold standard should cover the topic of interest as comprehensively as possible. For each of the three topics we studied, i.e. privacy, serendipity and Occupy Wall Street, we collected a set of non-microblog documents. For example, for privacy we coded documents from a variety of areas, including technology, healthcare, and legal regulation. More specifically, the documents fall into three categories: 1) recent news articles from online media; 2) academic papers; and 3) recent reports from public organizations or governments. This collection includes: 15 news

Algorithm 1 Word List Visualization

Input:
k : Keyword (e.g., serendipity);
w_i : Top-N word list mentioning k;
t : Threshold for filtering edges in word graphs;
C : Keyword-specific corpus (i.e., tweets with keyword k);
$E = \{\phi\}$: Edge list of the resulting word graph.

Procedure:
1: **for** each word $(w_i) \in W$ **do**
2: put edge (k, w_i) in E;
3: **end for**
4: **for** each word pair (w_i, w_j) **do**
5: compute PMI(w_i, w_j) in C ;
6: **if** PMI$(w_i, w_j) \geq t$ **then**
7: put edge (w_i, w_j) in E;
8: **end if**
9: **end for**

Output:
Edge list E of word graph.

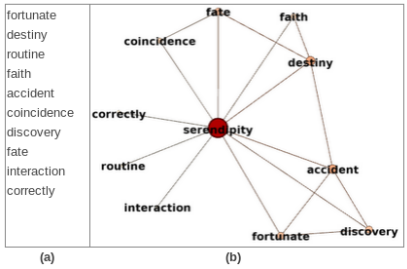


Figure 2: (a) Word list produced by RT-PMI; (b) Corresponding word graph.

Keyword	Word List	Average Precision %	
		Inter	Merge
Privacy	RT-PMI	15.2	16.2
	Frequency Analysis	4.7	5.1
	Random List	3.4	2.8
Serendipity	RT-PMI	11.4	12.2
	Frequency Analysis	5.7	6.1
	Random List	2.4	2.1
Occupy Wall Street	RT-PMI	4.5	4.5
	Frequency Analysis	5.1	5.0
	Random List	1.7	1.8

Table 2: Average Precision and Top-10 Precision of the three techniques for the three keywords of interest.

articles, 4 academic papers and 4 reports for privacy; 22 academic papers for serendipity; and 15 news reports for Occupy Wall Street.

2. Annotating the documents. Three annotators – postgraduates at top-ranked universities in the USA and UK – coded the gold standards . For each keyword, the annotators separately read each document line-by-line and highlighted any word they felt to be related to the keyword. We then combined their annotations for each keyword in two ways and generated two distinct annotation versions: one is the *merged version* of the three sets of annotations, and the other is the *intersection* among them.

3. Validating annotations. To quantitatively validate the gold standards, we measure *agreement* among annotators defined as the ratio of the size of the merged word sets over the size of the intersected sets. The agreement for each keyword is 92% (Privacy), 77% (Serendipity), and 83% (Occupy Wall Street). High agreements show good similarity between the two versions of the gold standards. In a qualitative survey, we have additionally asked respondents to provide any word they would find relevant to any of the three keywords. Based on their answers, we have found that all the words respondents have specified happen to be present in the three gold standards, further confirming their validity.

3.2 Metrics

Using the three gold standards, we evaluate RT-PMI against randomized word list (baseline) and frequency analysis, and we do so in terms of two widely-used information retrieval metrics.

Average Precision. This metric evaluates the average performance of our tool, giving us the overall comparison among RT-PMI, frequency analysis and baseline.

Precision/Recall Curves. We plot precision (fraction of the words returned by the tool that are part of the gold standard) at different recall levels (fraction of the ground truth's words returned by the tool). Precision/Recall curves complement the average precision metric, in that, they show how precision changes with recall.

Having the three evaluation metrics at hand, we are now able to compare the performance of RT-PMI against those of randomized word list (baseline) and frequency analysis.

3.3 Results

From the *Average Precision* (AP) in Table 2 ("inter" is short for the intersected version among the three annotators, and "merge" stands for the merged version), we can see that, for each keyword, both RT-PMI analysis and frequency analysis unsurprisingly outperform random lists, suggesting that the association measures we employ can effectively retrieve people's conceptual associations mentioning a keyword from unstructured tweets. Since the results for the two versions of gold standard (intersection and merged) are almost the same, to ease explanation, we will only discuss those for the merged version hereafter.

For that version, RT-PMI performs better than frequency analysis for all the three metrics. Indeed, RT-PMI's average precision is 10% higher for privacy and 6% for serendipity. More importantly, its precision/recall curves strike the best balance for both privacy (Figure 3(b)) and serendipity (Figure 3(a)) and do not fall down as quickly as the frequency analysis' curves. In Figure 3(a), one sees that RT-PMI outperforms frequency analysis with a difference of 40.0% at zero recall. That is because the top-3 words for frequency analysis are generic ones (i.e., "love", "sweet" and "movie"), while those for RT-PMI are more relevant (i.e.,"fortunate", "accident", and "coincidence").

By contrast, the results for Occupy Wall Street (OWS) are poor for all metrics and techniques. That is because, compared to the twitter streams of privacy and serendipity, that of OWS:

1. Contains far fewer distinct words: 1,833 compared to 7,539 for privacy and 7,338 for serendipity.
2. Only covers one third of the gold standard vocabulary: 31.6% compared to 91.2% for privacy and 84.6% for serendipity.
3. Consists of tweets with far fewer usable words: each tweet contains, on average, 4.7 words that can be potentially used by RT-PMI or frequency analysis; this number is, instead, 6.7 for privacy and 5.5 for serendipity.

As Table 1 shows, the OWS stream contains a high proportion of non-English tokens (URLs, hashtags(e.g., "#privacy"), usernames(e.g., "@chi")): 3.8 tokens in a tweet are non-English compared to 1.5 for privacy and 1.3 for serendipity. This alone can explain the poor results and suggests that, as one expects, the most important performance factor when analyzing short pieces of text is the number of usable words in them, which, in the dataset under study, should safely be above the range $[4.7, 5.5]$.

4. USER STUDY

Our quantitative evaluation has established that RT-PMI outperforms the two other techniques, yet it might not necessarily improve the user experience. To fix that, we conduct a user study to evaluate our methodology in an user-oriented style.Furthermore, we emailed individual researchers who co-authored reputable papers in the three topics and two departmental mailing lists. By doing so, we obtained 32 respondents: 10 for privacy, 12 for serendipity, and 10 for Occupy Wall Street.

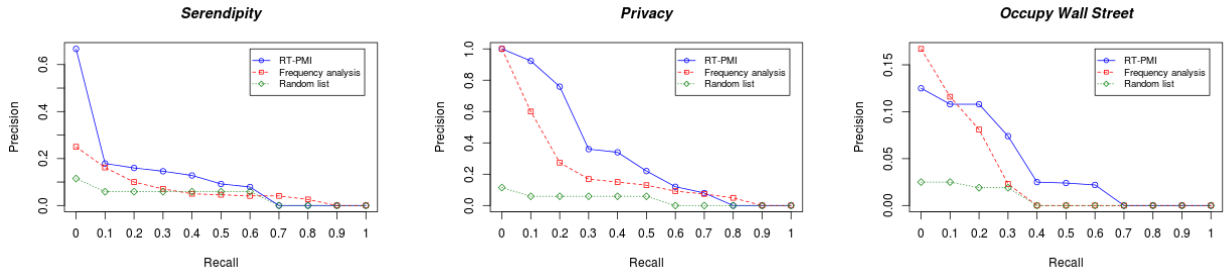
4.1 User Ratings

For each keyword, every participant is asked to rate: *a)* two word lists (one generated by RT-PMI, and the other by frequency analysis); and *b)* two word graphs (one with 10 nodes and the other with 20). Ratings are on a Likert scale ranging from 1 to 5 (where 1 corresponds to "no relevance/help at all", and 5 corresponds to "extremely relevant/helpful") and are expressed along two dimensions:

Effectiveness: *"To which extent does the [word lists or graphs] reflect your prior knowledge about the concept of [keyword]"*

Helpfulness: *"To which extent does the [word lists or graphs] suggest aspects of [keyword] or making "connections" that you had not thought before"*

The average user ratings are shown in Table 3 and confirm the results of the quantitative evaluation, suggesting that the two information retrieval metrics successfully capture an important part of the user experience. For both privacy and serendipity, RT-PMI outperforms the frequency analysis in terms of both effectiveness (serendipity: $t(11) = 12.6, p < 0.001$; privacy: $t(9) = 5.55, p < 0.001$) and



(a) Serendipity (b) Privacy (c) Occupy Wall Street

Figure 3: Precision/Recall curves for the three keywords of interest.

Topic	Metrics	Word Lists		Word Graphs	
		RT-PMI	Frequency Analysis	10 nodes	20 nodes
Privacy	Average Effectiveness	4.11 ± 0.80	3.00 ± 0.93	4.13 ± 0.67	3.13 ± 0.93
	Average Helpfulness	4.33 ± 0.75	3.11 ± 1.04	4.38 ± 0.83	3.13 ± 0.31
Occupy Wall Street	Average Effectiveness	1.50 ± 0.81	1.80 ± 0.08	1.40 ± 0.05	1.40 ± 0.05
	Average Helpfulness	1.50 ± 0.13	1.80 ± 0.03	1.60 ± 0.05	1.70 ± 0.04
Serendipity	Average Effectiveness	4.33 ± 0.67	1.67 ± 0.49	3.67 ± 0.65	2.75 ± 1.42
	Average Helpfulness	3.33 ± 0.65	1.33 ± 0.49	3.33 ± 0.77	2.33 ± 1.30

Table 3: User ratings reflecting the extent to which word lists and word graphs are both effective and helpful.

helpfulness (serendipity: $t(11) = 8.66, p < 0.001$; privacy: $t(9) = 5.6, p < 0.001$). Again, for Occupy Wall Street, we register the lowest ratings, which range from 1.40 to 1.60. More interestingly, participants find word graphs to be effective (serendipity: $t(11) = 2.56, p = 0.026 < 0.05$; privacy: $t(9) = 9.00, p < 0.001$) and helpful (serendipity: $t(11) = 3.32, p = 0.007 < 0.01$; privacy: $t(9) = 7.64, p < 0.001$), especially the small ones with 10 nodes (as opposed to those with 20).

4.2 Qualitative Results

In a qualitative survey, most of the respondents found word lists to be beneficial for both serendipity (all 9 responses were positive) and privacy (4 out 6 were positive), but, again, not for Occupy Wall Street (*"most of the words in the lists are about privacy"*, and *"only a few words are relevant to OWS."*). From the responses, we gather that, in the case of Occupy Wall Street, respondents felt they were "overladed" by too many relevant words.

However, not all comments about the word lists for serendipity and privacy were positive. The most common issue raised by respondents was a list's lack of logical connections among words ("*[The word list] reflects the concept of serendipity, but the links among words are not very clear*").

By contrast, respondents felt that word graphs are able to show different facets of the same concept: all respondents, for example, have identified the cluster "accident-fortunate-discovery" in the graph of serendipity and "network-online" in that of privacy. ("*Word clusters effectively reveal different aspects of serendipity*", "*Yes, [with word graphs], connections between words become clear*", "*Groups of words are able to tell stories*")

Finally, respondents preferred smaller word graphs with 10 nodes as opposed to those with 20 nodes (Table 3): based on the respondents' comments, we learn that larger graphs are more likely to contain irrelevant words.

5. CONCLUSION

Based on a real-time adaptation of an existing word association measure, a textual analysis of *individual* tweets has produced interesting insights into the two concepts of serendipity and privacy, which have been widely studied in the past. Using the Grounded Theory approach, we have coded the most cited articles in those fields and obtained annotations that match (to a large extent) our tool's word lists and word graphs. This suggests that gathering and analyzing tweets in real time might well offer concise 'snapshot views' and understanding of research issues other than privacy and serendipity. However, analysis of user-generated datasets is not intended to replace other established methodologies but to complement them, providing yet one more way of investigating social patterns *at scale*.

Acknowledgments. We thank Stephann Makri and Anne Hsu for their feedbacks. This research was partially supported by the SocialSensor FP7 project (contract no. 287975).

6. REFERENCES

[1] J. M. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, 2008.

[2] A. J. Gill, A. Vasalou, C. Papoutsi, and A. N. Joinson. Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In *Proc. CHI*, 2011.

[3] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proc. CHI*, 2011.

[4] D. Hopkins and G. King. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1), 2010.

[5] Lazer, D. et al. Computational Social Science. *Science*, February 2009.

[6] G. Mark, M. Bagdouri, L. Palen, J. Martin, B. Al-Ani, and K. Anderson. Blogs as a collective war diary. In *Proc. CSCW*, 2012.

[7] D. Quercia, D. Ó Séaghdha, and J. Crowcroft. Talk of the city: Our tweets, our community happiness. In *Proc. ICWSM*, 2012.

[8] J. Rubin, V.L. Burkell and A. Quan-Haase. Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis. *Information Research*, 16(3), 2011.