International Workshop on Visual Analytics (2012)
K. Matkovic and G. Santucci (Editors)

Visual Analytics of Microblog data for pandemic and crisis analysis

I. C. Pritchard, R. Walker and J. C. Roberts

School of Computer Science, Bangor University, UK

Abstract

Microblogging is a rich and plentiful source of data that contains potentially valuable information amidst noise. This work presents software that extracts useful metadata from a microblog dataset to explore and analyze the data for the detection and exploration of crisis events. The developed software (Vambutu) was successfully used to examine an artificially-generated dataset for the onset and source of an illness outbreak. A part of speech tagger was used to divide microblog posts into their component parts for the purposes of identifying posts pertaining to first, second, and third-hand experiences. A successful demonstration of this ability revealed clearly identifiable patterns for first-hand experiences. For example, for the word pneumonia we found patterns that were not apparent when all posts pertaining to pneumonia were examined at once. This promising result demonstrates the potential as a tool for filtering out irrelevant noise during the occurrence of a crisis event.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—GUI H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Search process

1. Introduction

Making sense of the large quantities of data that are stored and collected on a daily basis is a challenging task, especially in pressure-filled situations [TC06]. Microblogging is one such area where a large amount of data is generated on a continuous basis, some of it contradictory, and much of it irrelevant to any particular area of study. Visual analytics is one method that can help make sense of such data by allowing the user to make informed decisions and draw reasonable conclusions following an interactive visual exploration of the data when presented in an appropriate way [KMSZ06]. Representing the geospatial location of microblog posts (where such information is available) can enhance the semantic content of posts.

The small size of the microblogs leads to a high frequency of posts, and a tendency to post during or immediately after an event rather than several days later [JSFT07]. Disaster response agencies wish to take advantage of this short latent period between event and post to help inform and direct their efforts in the event of a crisis [KPA09], and although rumors and misinformation is present in microblogging it is possible to estimate their reliability [CMP11].

In the event of a crisis, posts pertaining to that occurrence would likely see key words describing the crisis (e.g. “earthquake”, “illness”, “fire”) rise in frequency within the data. In order to be able to detect these sudden word frequency increases, a means of categorising words must be developed so that words of potential interest may be monitored. Plotting the featured posts on a map would also help determine the location of the incident. Whilst those people who are having first-hand experience of an event may post from the location of that event, many other posts may refer to second or third-hand knowledge of an event, and may be posted some distance away from the event itself. Therefore one of our aims is to determine which posts may relate to first-hand experience, and which posts refer to second or third-hand experience.

We wish to analyze the microblog data by means of a deictic analysis (especially person deixis on the words) to investigate what types of words are used in the microblogs and where they are used in the texts, and to create a visual analytic tool that allows the user to explore and visualize the posts. Through this approach it would be possible to discern if the content of the microblog is about the user, direct first person experience, or is about other people. Potentially this

DOI: 10.2312/PE/EuroVAST/EuroVA12/055-059

could be used to detect retweets by verb tense or pronoun identification. For this paper we use the IEEE VAST Challenge 2011 dataset [GWLN]. This data represents a spread of a crisis over 21 days. The advantage of using the VAST dataset in the first instance, is because of the known ground truth in this dataset, consequently the software can be compared against a set of testable hypotheses.

In this work we present Vambuta: our software that performs Visual Analytics of Micro Blogs Using Text Analysis, which is developed in Java using the Processing.org library. We present how we analyze each message by their part of speech (POS), such as: verb, participle, pronoun, adverb and conjunction, and introduce our visual analytic interface. Parts of speech occur in nearly every natural language [Vou03] and the parts are largely independent of semantics, and are instead a function of grammar. POS taggers often employ a multi-stage approach, with statistical models being used to train POS taggers on manually annotated corpora in order to enable them to tag previously unseen text with a high degree of accuracy [Rat96]. Although POS taggers have been used for Twitter analysis, our work extends this by providing a visual analytic interface where users can select and choose different parts of speech. The advantage of using POS is that logical conjunctions can be made over the microblogs, where the blogs that contain (say) specific nouns or verbs can be selected, or the usage (frequency) of different parts of speech can be visualized. Furthermore, we evaluate the use of a POS tagger for crisis analysis, and suggest an alternative methodology for future work.

2. Related Work

Microblogs are a relatively recent technological development, and as such, the academic body of work covering their usage and analysis does not extend back very far [CL10]. However, analysis of microblog data has been performed by some researchers. For instance, Java et al. [JSFT07] discovered 21% of posts "mentioned" another user (i.e., prefixed a username with an @ symbol in order to involve that user in a dialog). URLs were found to be present in 13% of posts. While, work by Huberman et al. [HRW09] indicated that a disproportionately large number of posts were being made by some posters, and that users with a high number of followers tended to produce a higher number of posts, as did those users who followed a large number of other users.

Researchers have developed software to visualize microblogs, display information about the people who are tweeting and their relationships, and the GPS location of the tweets. E.g., TweetTracker continually monitors keywords of a crisis; Twitinfo monitors Twitter for keywords and displays the result on a timeline; and Singh et al. have used pixels to plot trends [SGJ10]. While White and Roth [WR10] use geospatial location as a means of visually exploring information contained within microblog posts. White and Roth state that this tool could be used for crime trend analysis within a particular area of interest. Indeed, MacEachren et al. [MRJ*11] discuss applications of twitter for crisis management, while other researchers have focused to identify outbreaks of illness. For instance, Achrekar et al. [AGL*11] conducted an experiment on twitter data, their results support the idea that microblog data can be used as a means of accurately identifying the location of outbreaks of influenza (and potentially other illnesses). Culotta [Cul10] conducted a similar experiment whereby approximately 6.5 million posts were examined for influenza-related key words. For the VAST Challenge dataset, Bertini et al. [BBF*11] used the Stanford Named Entity Recognizer (NER), while Cenydd et al. [aCWP*11] used the first three days as the corpus with a relative entropy metric, and Bosh et al. [BTW*11] displayed high frequency words using tag clouds.

3. Data Analysis & algorithmic decisions

The IEEE VAST 2011 Challenge dataset [GWLN] contains 1,023,077 microblog posts over a period of 21 days. Each blogger has a unique ID that identified which posts had been made by them, the number of unique IDs present in the dataset is 73,928. The dataset consists of real-life posts, along with created tweets, which have been cleaned and anonymized. The blogs all occur in Vastopolis, an imaginary city, and contain 'ground truth'. The posting frequency for the first 18 days of the dataset remains relatively stable after which there is an increase in posting frequency for the remaining three days. Through statistical analysis using SPSS and a histogram, the challenge dataset represents a Normal distribution ($M = 13.84, SD = 0.78$). This contradicts work by Huberman et al. [HRW09] who suggested that a logarithmic distribution is found from real data. While there are limitations to this dataset, it provides a useful test dataset for our purposes, and because our focus is on the content of the blogs, the data is suitable for our situation. After initial analysis, we cleaned the dataset by removing foreign tweets, 48 posts with no message, and six posts that contained an unusual high level of unusual characters. After cleaning the data was tagged.

3.1. Tagging the data

We use the Stanford log-linear Part-Of-Speech Tagger [TKMS03] for this work, which has a high level of tagging accuracy in English [GE09]. It uses a MaxEnt (Maximum Entropy) model in conjunction with the Penn Treebank tagset to obtain its high degree of accuracy. The software is written in Java and is open source. Some analysis on this POS tagger shows that accuracy is reduced (93.47%) when used on sources that do not conform to accepted rules of punctuation and grammar [EHH10] such as microblog messages. Consequently, recent work has focused on developing Twitter POS tagger, corpus, and tagset in order to address the problems associated with tagging non-standard text using traditional POS taggers [GSC*11]. This corpus (1827

manually annotated messages) however does not compare favorably with the Penn Treebank corpus (4.5 million annotated words), and the accuracy (89.37%) of the new tagger is not yet high enough to justify its use in place of the Stanford POS for our work. Therefore, although the Stanford POS was not designed for tagging microblog data, and tagging errors may occur, for the purposes of this research the Stanford Log-linear Part-Of-Speech Tagger was be used.

Two trained tagger models are supplied with the Stanford Log-linear Tagger (SLT), one that uses only left context dependencies to tag words, whilst another uses bidirectional context dependencies to tag words. On the Wall Street Journal corpus, the bidirectional model was found to be more accurate (97.28% accuracy) when compared to the left word model (97.07% accuracy). Therefore to ascertain which model would be best for our purpose we analyzed both models on a shortened (random sampled) version of the tweet data containing 1668 messages. The time taken to complete the analysis was considerably longer for the bidirectional model (22 minutes and 44 seconds) than for the left words model (11 seconds). The SLT Part-Of-Speech Tagger performs less effectively on conversational speech when compared to formal speech [EHH10]. This is in part due to the lack of punctuation found within casual speech. The messages contained within the dataset were casual in nature. In our tests, inconsistencies were generated by both the left words and bidirectional models when dealing with superfluous or incorrect punctuation usage (such as excessive use of exclamation marks). However, the bidirectional model is consistent in identifying the extraneous punctuation as a noun, whilst the classification of the left words model depended on the number of exclamation marks. An inspection of the raw data showed near identical values for all categories excepting nouns (left words: 4033, bidirectional: 3742), and prepositions or subordinating conjunctions (left words: 1700, bi-directional: 1636). Both models returned a total of 23,762 tags with no errors recorded. In order to determine whether or not there was a statistical difference between the results for each model, the frequency values obtained were saved into a file format readable by statistical package SPSS 14. A paired samples t-test was carried out, the results of which found no statistical difference between the left words and bidirectional models ($t(36) = -0.10, p > 0.05$). The left word model was then used to repeat the tagging process for the entire dataset, which takes 74 minutes and 38 seconds to complete.

The parts of speech available for selection were: adjectives, proper nouns, adverbs, foreign words, nouns (singular or mass), personal pronouns, verbs (base form, including: imperatives, infinitives, and subjunctives), verbs (non-third-person singular present), and verbs (past tense). The decision to present these particular parts of speech to the user were informed by the results obtained during the analysis phase. These parts of speech represent those found to have the highest frequencies within the dataset (excepting those parts of speech that would not to add value to the software such as possessive endings and cardinal numbers).

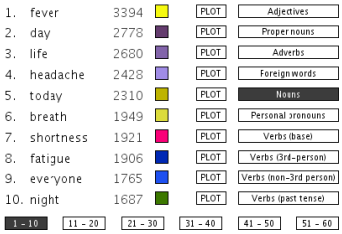

Figure 1: *The picture shows the tabular interface of Vanbuta, where users can select and plot different parts of speech. This diagram shows 'nouns' on the 18th May.*

3.2. Frequencies

In Vambuta we calculate the frequencies of each POS. A Frequencies class was made to implement the Comparable interface, which enables the Frequency objects to be sorted using the modified merge sort provided as part of the Java Collections Framework's Collections class. We have developed a basic tabular interface where the user can see the words , their ranks and can plot them on the screen (see Figure 1).

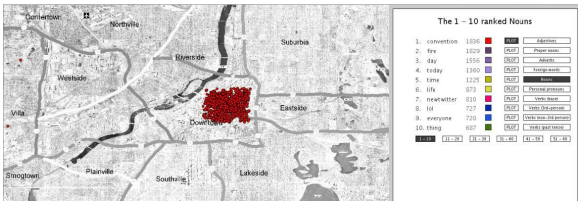

Figure 2: *Tweets which contain the selected words and are within the part-of-speech criteria are plotted by points on the map. Here the noun "convention" is displayed.*

The user can select different words to be visualized. Once the word to be plotted had been identified, each message was compared for string equality with both the target word and the correct part of speech. We need to confirm the relevant part of speech such to overcome possible ambiguities, for example, plotting posts containing the verb "fire" when the user wished to examine the noun "fire" to find a burning building. For example, the word "convention" is plotted in Figure 2; 1838 posts contain the word "convention" on the selected day, only two posts are not located within close proximity to the Vastopolis Convention Center. The user can to identify words of interest and then plot the location of posts containing that word on map.

3.3. Word tracking and display

A full screen-capture of Vambuta is shown in Figure 3. Because the frequencies of the words change over time, we track the changes over time, such that users can select a word and track it's location and frequency count over time. The

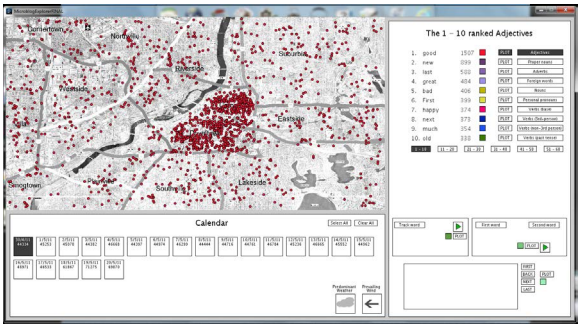

Figure 3: *The visual analytic interface of Vambuta, including: the map, tabular query interface for the parts of speech; the calendar interface, and the multi-word tracker interface.*

Figure 4: *(Top) A picture of microblogs that include "pneumonia". (Bottom) Picture showing the microblogs that include "pneumonia" and "me".*

frequency of the tracked word is displayed at each time step. A user may be able to draw informed conclusions from the change of frequency counts over different days. We use animation to display how the blogs change over time, and 'decaying plots' that fade the display of the older plots; a three-day window was used where the data of the previous day was 50% transparent and the day-before 75% of its opacity.

Multiple words can be selected and tracked. Multi-word tracking plots the data for posts that contain a conjunction of two or more user-selected words, a frequency of all the occurrences is generated, and the data is plotted on the map. This is an important use of the analytics. For example, plotting "pneumonia" alone is different to plotting "pneumonia" and "me" (Figure 4). Whereas in the first plot the data is noisy, in the second there is clear clustering around the hospitals. We do note, however, that some personal pronouns do not demonstrate clusters. E.g., in this Challenge dataset the use of "pneumonia" with second-person personal pronouns (e.g., "you") do not form interesting results, neither do the third-personal personal pronouns (e.g., "he", "she", "they").

4. Discussion & Conclusion

We have developed a prototype visual analytics tool (Vambuta) to explore POS tagged microblogging data. We have successfully used it to display the VAST 2011 Challenge dataset and are further developing the ideas for dynamic microblogging data. Through building up a POS model, we have demonstrated that it is possible to analyze the microblogs to investigate a crisis. The word tracker and multi-word tracker were found to be versatile features that enabled a range of useful analyzes. They allowed a word of interest to be tracked across time, regardless of where that word falls within the word frequencies of any given day. Word tracker was instrumental to discover that stomach pains were a common complaint (1,758 posts) on the 19th of May, but that diarrhea had become the more common ailment (2,176) by the 20th with instances of stomach pains decreasing to 887 posts. This discovery suggests a progression of symptoms rather than simultaneous onset of all symptoms. The multi-word tracker allowed users to understand the relationship between two words of interest (e.g. "shortness" and "breath"), and through this method users can distinguish between first, second, and third-hand experiences by combining a word of interest with a personal pronoun.

The decaying plots clearly showed the temporal and spatial proximity of different events. For instance, the word "terrible" was misspelled as "terible" was the 8th most frequent adjective (864 posts) on the 20th of May, and yet a time track revealed that the misspelling did not appear any other day. Whilst the decaying plots were found useful in situations where clear patterns were observed, these plots could add further clutter to plots that may already be dense.

POS does have drawbacks: it is slow to calculate and evaluates the complete dataset. It also relies on a comprehensive corpus (e.g., Penn Treebank). Other researchers have developed keyword monitoring algorithms in order to detect the onset of crisis events [AGL*11, KBAL11]. However, the challenge with these tools is that they monitor specific predefined keywords for certain anticipated events. We believe that the use of POS, frequency analysis and logical conjunctions over these parts creates a powerful and useful foundation to perform crisis analysis of unknown crisis. However, further work is required. One direction is to investigate comparison of tweets [GAW*11], another is to create a lightweight POS method that quickly tags the data and allows the corpus to dynamically change with the dynamic nature of microblogs, and be integrated with a Visual Analytics interface.

References

[aCWP*11] AP CENYDD L., WALKER R., POP S., MILES H., HUGHES C., TEAHAN W. J., ROBERTS J. C.: epspread - storyboarding for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology, Providence, Rhode Island, USA* (2011), pp. 311–312. 2

[AGL*11] ACHREKAR H., GANDHE A., LAZARUS R., YU S., LIU B.: Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (2011), IEEE, pp. 702–707. 2, 4

[BBF*11] BERTINI E., BUCHMULLER J., FISCHER F., HUBER S., LINDEMEIER T., MAASS F., MANSMANN F., RAMM T., REGENSCHEIT M., ROHRDANTZ C., SCHEIBLE C., SCHRECK T., SELLIEN S., STOFFEL F., TAUTZENBERGER M., ZIEKER M., KEIM D. A.: Visual analytics of terrorist activities related to epidemics. In *IEEE Conference on Visual Analytics Science and Technology, Providence, Rhode Island, USA* (2011), pp. 329–330. 2

[BTW*11] BOSCH H., THOM D., WORNER M., KOCH S., PUTTMANN E., JACKLE D., ERTL T.: Scatterblogs: Geo-spatial document analysis. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (oct. 2011), pp. 309 –310. 2

[CL10] CHEONG M., LEE V.: A study on detecting patterns in twitter intra-topic user and message clustering. In *Proceedings of the 20th International Conference on Pattern Recognition* (Washington, DC, USA, 2010), ICPR '10, IEEE Computer Society, pp. 3125–3128. 2

[CMP11] CASTILLO C., MENDOZA M., POBLETE B.: Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (New York, NY, USA, 2011), ACM, pp. 675–684. 1

[Cul10] CULOTTA A.: Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (2010), ACM, pp. 115–122. 2

[EHH10] EIDELMAN V., HUANG Z., HARPER M.: Lessons learned in part-of-speech tagging in conversational speech. In *Conference On Empirical Methods In Natural Language Processing* (2010), pp. 821–831. 2, 3

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (oct 2011), 289–309. 4

[GE09] GIESBRECHT E., EVERT S.: Is part-of-speech tagging a solved task? an evalutation of pos taggers for the german web as corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)* (2009), pp. 27–35. 2

[GSC*11] GIMPEL K., SCHNEIDER N., CONNOR B. O., DAS D., MILLS D., EISENSTEIN J., HEILMAN M., YOGATAMA D., FLANIGAN J., SMITH N. A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings ACL-HLT* (2011), Association for Computational Linguistics, pp. 42–47. 2

[GWLN] GRINSTEIN G., WHITING M., LIGGETT K., NEBESH D.: IEEE VAST Challenge 2011. URL: http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/. 2

[HRW09] HUBERMAN B., ROMERO D., WU F.: Social networks that matter: Twitter under the microscope. *First Monday 14*, 1 (2009), 8. 2

[JSFT07] JAVA A., SONG X., FININ T., TSENG B.: Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (New York, NY, USA, 2007), WebKDD/SNA-KDD '07, ACM, pp. 56–65. 1, 2

[KBAL11] KUMAR S., BARBIER G., ABBASI M., LIU H.: Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011), pp. 661–662. 4

[KMSZ06] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in visual data analysis. In *Proceedings of the conference on Information Visualization* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 9–16. 1

[KPA09] KIREYEV K., PALEN L., ANDERSON K.: Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond* (2009). 1

[MRJ*11] MACEACHREN A., ROBINSON A., JAISWAL A., PEZANOWSKI S., SAVELYEV A., BLANFORD J., MITRA P.: Geo-twitter analytics: Applications in crisis management. In *25th International Cartographic Conference, Paris, France* (2011). 2

[Rat96] RATNAPARKHI A.: A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing* (New Brunswick, New Jersey., 1996), vol. 1, Association for Computational Linguistics, pp. 133–142. 2

[SGJ10] SINGH V. K., GAO M., JAIN R.: Social pixels: genesis and evaluation. In *Proceedings of the international conference on Multimedia* (New York, NY, USA, 2010), MM '10, ACM, pp. 481–490. 2

[TC06] THOMAS J. J., COOK K. A.: A visual analytics agenda. *IEEE Comput. Graph. Appl. 26* (January 2006), 10–13. 1

[TKMS03] TOUTANOVA K., KLEIN D., MANNING C., SINGER Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (2003), pp. 252–259. 2

[Vou03] VOUTILAINEN A.: *The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press.* Oxford University Press, 2003, ch. Part 2, Chapter 11, Part-of-speech tagging. 2

[WR10] WHITE J., ROTH R.: Twitterhitter: Geovisual analytics for harvesting insight from volunteered geographic information. In *Proceedings of GIScience* (2010), vol. 2010. 2