Enabling Social Media Research Through Citizen Social Science

Rob Procter[1], William Housley[2], Matthew Williams[2], Adam Edwards[2], Pete Burnap[3], Jeffrey Morgan[3], Omer Rana[3], Ewan Klein[4], Miranda Taylor[4], Alex Voss[5], Chris Choi[5], Panos Mavros[6], Andy Hudson Smith[6], Mike Thelwall[7], Tristan Ferne[8], Anita Greenhill[9]

[1]Department of Computer Science, University of Warwick
[2]School of Social Sciences, Cardiff University
[3]School of Computer Science & Informatics, Cardiff University
[4]School of Informatics, University of Edinburgh
[5]Department of Computer Science, St Andrews University
[6]Centre for Advanced Spatial Analysis, UCL
[7]University of Wolverhampton
[8]Internet Research & Future Services, BBC R&D
[9]Manchester Business School
[1]*rob.procter@warwick.ac.uk*

Abstract. This paper explores how 'citizen social science', may help professional social scientists deal with the challenges of exploiting the growing range and volume of 'born digital' social data. We outline a social media analytics platform that we have developed and describe how we plan to use crowdsourcing to improve the performance of our tools.

Keywords. Social media, Twitter, crowdsourcing, citizen social science, curation, annotation

Introduction

This paper explores how 'citizen social science', as a new example of the wider citizen science arena[1], may help social scientists deal with the challenges of exploiting the growing range and volume of 'born digital' social media data. We report on work in progress to establish and exploit the potential of crowdsourcing for large-scale social media data curation and analysis. Our aim in this research is to explore the benefits and limitations, and develop ways of maximising the former

[1] http://scienceforcitizens.net/

while minimising the latter. Specifically, our objective is to devise approaches to crowdsourcing in this context that are scalable but do not sacrifice the quality of contributions and investigate how these can be used to improve the performance of computationally-generated annotations.

The rapid growth of the Web as a publishing tool, and the recent explosion of social media such as blogs (and micro-blogs such as Twitter) and social networking sites (such as Facebook) presents both an opportunity and a challenge to social researchers. Data that can shed light on people's habits, opinions and behaviour is available now on a scale never seen before, but this also means it is impossible to analyse using conventional methodologies and tools.

We are building COSMOS[2], a platform providing an integrated suite of tools for harvesting, archiving, analysing and visualising social media data streams for use by social researchers (Burnap et al., 2013; Edwards et al., 2013), with the capability to link with other kinds of data, e.g. from ONS via open APIs. A critical task in the COSMOS research workflow is annotation of incoming social media streams. We have developed a range of computational tools (language detection, gender assignment, location, sentiment, tension, topic discovery). However, despite the growing sophistication of computational tools for social media analysis, they are not sufficiently reliable to substitute for human expertise. Hence, what is needed is a way to combine computational tools with human expertise in ways that make the best of their respective strengths (Procter et al., 2013a; 2013b). This human expertise is essential for benchmarking and improving the performance of computationally generated annotations and analyses, and curating datasets. If this is to be feasible, then human expertise needs to be readily available and in numbers sufficient to deal with the quantities of data.

One way for providing this expertise is through volunteer efforts in the manner of crowdsourcing (Doan et al., 2011), as is now widely exploited under the rubric of citizen science[3] and which projects such as Galaxy Zoo[4] have already demonstrated the potential for in the physical sciences.

To test the feasibility of 'citizen social science' for social media analytics we are building a web-based tool, which volunteers will be able to use to access social research collected by COSMOS and perform simple annotation tasks. These volunteered annotations will then be used to check and improve the quality of the COSMOS computationally-generated annotations.

Our approach is modelled on a crowdsourcing facility now being piloted by the BBC to put massive, searchable media archives online using a combination of algorithms and crowdsourcing (Raimond and Lowis, 2012). BBC Research & Development has built a browsable and searchable online archive, which uses crowdsourcing to validate and improve the quality of computationally-generated annotations. Registered users can listen to programmes in the archive, add new annotations and vote on the quality of existing annotations.

We begin by outlining the ways in which we generate social media annotations computationally. We then outline the BBC pilot and how we plan to build on that to improve the quality of computer-generated annotations for social media dataset curation and analytics.

Computer-generated annotations of social media

COSMOS harvests and annotates content from a number of social media sources. In this paper, we will focus on Twitter for the purposes of illustrating its capabilities and the challenges of improving the quality and reliability of our analysis tools.

[2] www.cosmosproject.net
[3] http://en.wikipedia.org/wiki/Citizen_science
[4] http://www.galaxyzoo.org

Gender, location and language

To identify the gender of the tweeter, the name the user added to their profile is extracted from the tweet meta-data. The first name is mapped on to the 40k Namen database – a database of over 44,000 names from 54 countries around the world – with each name classified as male, female, or unisex (Michael, 2007; Morgan et al., 2013). One limitation to this approach is that there are clearly more than 44,000 names in use around the world, so crowdsourcing could assist in classifying previously unclassified names.

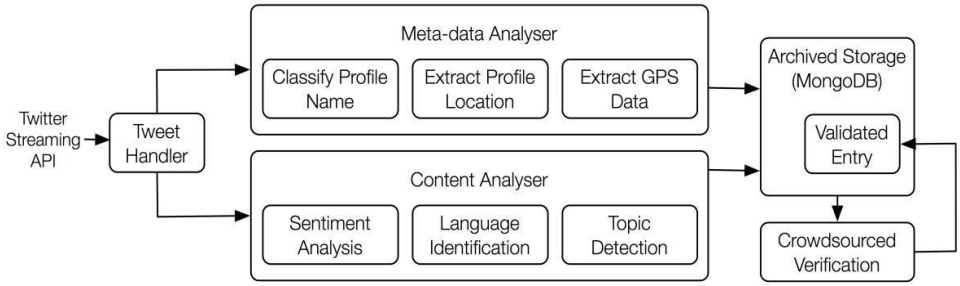

Figure 1: COSMOS annotation workflow.

Twitter enables users to provide their 'location' as an attribute in their profile. This can provide information about where the user lives. There are granularity issues with this approach. Some users simply state 'UK', others state 'Cardiff, UK', and some provide a locality down to area or street level. Some tweets contain GPS metadata. However, our investigations suggest that very few users enable GPS on their tweets (0.85%). To enhance accuracy, we plan to use crowdsourcing to analyse the text of the user's last *n* tweets to determine if there are clues in the text to suggest their location. Language is identified using the Language Detection Library for Java, which can identify 53 different written languages from a text sample. As with names, there will be languages, such as Welsh, that are not among the known languages. Crowdsourcing could be used to extend the number of languages COSMOS can detect by classifying 'new' languages.

Topics

In order to help researchers gain an overview of topics that are prominent in a corpus, COSMOS provides clustering algorithms. COSMOS clusters tweets incrementally in chronological order, using a sliding window of adjustable size. This makes it possible to investigate how topics change over time or to investigate topics within a specific time range.

COSMOS provides two algorithms: incremental online clustering, using cosine distance, and latent Dirichlet allocation (LDA). LDA is a probabilistic algorithm, which requires the number of clusters to be specified *a priori* and assigns each tweet to the cluster with the highest probability. These probabilities make it easy to identify tweets that are most representative of a cluster as well as outliers. The incremental algorithm compares each tweet to clusters that have already been formed and either assigns it to the nearest cluster, based on cosine distance, or creates a new one. The incremental algorithm is better suited for real-time clustering because it is faster and doesn't require the number of clusters to be specified. However, it is more sensitive to differences in datasets and requires more parameter tuning to obtain good results. In both, each tweet is represented as a list of word counts or 'features' and any term not considered a feature is ignored. Consequently, it is the number and quality of features that determines the quality of the resulting

clusters.

Clustering performance can be adjusted by tuning parameters for specific corpora and research questions, such as selecting appropriate features, including keyword inclusion and exclusion, top term exclusion, feature weighting and feature number specification. Crowdsourcing could be used to improve clustering performance by ranking cluster quality and harvesting candidate cluster labels.

Sentiment

Sentiment is an important aspect of online communication. Emotional exchanges can have different dynamics to more emotion-free communications and it is impossible to fully understand exchanges if their affective component is ignored. SentiStrength is a sentiment analysis program that has been purpose-built for analysis of social web texts, such as tweets, Facebook wall posts and short blog posts. It estimates sentiment content in two dimensions: the strength of positive sentiment on a scale of 1 (no positive sentiment) to 5 (very strong positive sentiment) and the strength of negative sentiment on a scale of -1 (no negative sentiment) to -5 (very strong negative sentiment).

The main method SentiStrength uses is a lexicon of 2,310 words and word stems with a predefined sentiment polarity and strength. For example, *angry* is a negative term with strength -4. If fed a sentence, SentiStrength will match all the words with its lexicon and assign the sentence the highest positive score of any matching term and the highest negative score of any matching term (Thelwall and Buckley, 2012). This method is supplemented by a set of linguistic rules to cover things like negations, questions and booster words (e.g. very). In addition, there are rules for identifying expressions of sentiment in ways that are in non-standard English. These include emoticons and emphatic spellings through repeated letters. For instance, the word angggggrrrrry would score -5 rather than -4 (the default for angry) due to emphatic spelling. Combining the word list and the linguistic rules gives approximately human level accuracy in the sense that (carefully selected, accurate) humans agree with each other about the same amount as they agree with SentiStrength (Thelwall and Buckley, 2012).

SentiStrength sometimes does not perform well on collections of topic-specific texts due to extensive exhibiting unusual sentiment language. For example, tweets about the UK riots used negative terms that are relatively rare in general social web texts, such as ‘baton’, ‘fire’, and ‘arrest’. In response, a method has been developed to customise SentiStrength for specific topics. It works by identifying the appropriate mood for the collection of texts and then identifying new potential sentiment-bearing terms that are candidates to be added to the lexicon for the topic, as well as suggestions for changing the sentiment weights of existing terms (Thelwall and Buckley, in press). One application of crowdsourcing would be to assist in customising SentiStrength by selecting candidate sentiment-bearing terms and adjusting their weight.

BBC crowdsourcing pilot

BBC Research & Development is running an experiment with the BBC’s World Service radio archive to demonstrate a way to put massive media archives online using a combination of algorithms and crowdsourcing.[5] We think we can automatically generate metadata for the archive that is good enough to kick-start crowdsourced metadata improvement.

The archive has around 50,000 digitised programmes from the World Service English-language radio service (Raimond and Lewis, 2012) from over 50 years. It has high-quality audio, but limited

[5] http://worldservice.prototyping.bbc.co.uk

metadata. We bootstrap the online archive by generating metadata automatically. We run the audio through a speech-to-text process using CMU Sphinx with the HUB4 acoustic model. This generates quite noisy transcripts, which are not normally readable, but from which we can still extract topics. For the extracted topics we use linked data entities from DBPedia[6], so that everything in the system is a 'thing' with a unique URI. Using this data, we built a browsable and searchable online archive[7], which uses crowdsourcing to validate and improve the machine-generated annotations. Registered users can listen to programmes in the archive, add new topics and vote existing topics up or down.
We identified a number of potential user groups, including BBC production staff, academic researchers and fans of radio, the World Service, particular programmes and topics. So far, it has been used mainly up by fan communities and some BBC staff. The number of registered users is fairly small (1300 by March 2013), but there has been a significant amount of activity.

About half of the registered users are active (i.e. they've carried out some action in the prototype) and so far they've listened to 8,533 distinct programmes (17% of the entire collection), taken action on 4429 distinct programmes (9%). On these programmes where activity has happened, users have added 7085 new tags (mean of 1.6 per programme) and voted on tags 34,000 times (mean of 8 votes per programme). From our initial work we appear to have a long tail distribution of how many times a programme has been listened to and tagged, and this corresponds to programmes we have promoted on the prototype or that have been linked to by the active user groups. Along with these 'defined' activities, users have also contacted us with corrections for existing metadata. We have seen two primary kinds of user; one is people who want primarily want to listen to programmes in the archive and might tag things whilst they are there, the other is people who either want to help or see tagging as an enjoyable task in itself. This latter group have done a lot of tagging, either around topics or around particular programmes. This is consistent with studies that have found it is often a small number of participants who do a large amount of the work (Dunn and Hedges, 2012).

The plan is to feed back the crowdsourcing into the topic extraction algorithms to improve them. For example, it has been noticed that people often down-vote particular tags. One way to feed this back into the algorithms is to reduce the confidence score wherever this is the computationally-generated tag.

Crowdsourcing for social media curation and analysis

The BBC crowdsourcing pilot provides a useful template for citizen social science and for how crowdsourcing may be used to improve the quality of computationally-generated annotations. However, there are some important features of the latter that may dictate that we have to employ different solutions. This enables us to define a series of specific objectives for this project.

First, given the potential size of social media datasets, to identify ways to select a representative sample for annotation by crowdsourced effort. This sample must be chosen so as to maximise the value of the crowdsourcing for improving the quality of computationally-generated annotations, while keeping the effort required within feasible bounds.

Second, this raises the question of how to recruit crowdsourcing contributions to match the volume of data (Willett et al., 2012). One challenge is to identify 'communities of interest' whose efforts may be leveraged. We also need to explore how to incentivise volunteer contributions (e.g. entertainment, games, prizes, peer esteem, recognition for participating in a research project, getting feedback on results) while maintaining the quality and to understand what appears to explain the

[6] http://dbpedia.org/About
[7] http://www.bbc.co.uk/blogs/researchanddevelopment/2012/11/the-world-service-archive-prot.shtml

interest in citizen social science, both in terms of scale of volunteered effort and the quality assurance of contributions. Examples from successful (and unsuccessful) citizen science projects will be instructive here.

Third, and linked with the above is the need to provide a range of options for contributing (e.g. voting on annotations, adding new annotations, etc.). To minimise the effort involved, we also need to investigate ways of linking annotation tasks as seamlessly as possible with volunteers' everyday uses of social media, so that rather than being experienced as additional work, it becomes a simple extension of their normal activities. One possibility for tweets would be to integrate annotation within an adapted Twitter client and to select content for annotation for presenting to individual volunteers that matches their social media usage and interests. In this way, we aim to increase both the scale and quality of the annotations crowdsourced.

As yet, we only have limited experience (e.g. Procter et al., 2013a; 2013b) on which to base estimates of the scale of crowdsourcing effort required for social research. The annotation effort required was quite modest (up to 15 volunteers annotating a few hundred tweets each). Determining a sampling strategy that balances effort required against quality improvement will be important for determining whether citizen social science can scale to add value to much larger corpora. Our ongoing work is aimed at exploring and resolving these issues, using the BBC pilot to identify lessons for crowdsourcing annotations and investigating how to translate these lessons to the context of social media research.

References

Burnap, P., Rana, O. and Avis, N. (2013): 'Making Sense of Self-Reported, Socially Significant Data Using Computational Methods'. International Journal of Social Research Methodology, Computational Social Science: Research Strategies, Design and Methods. Volume 16:2.

Doan, A., Ramakrishnan, R. and Halevy, A.Y. (2011): 'Crowdsourcing systems on the world-wide web'. Communications of the ACM, 54(4), 86-96.

Dunn, S. and Hedges, M. (2012): 'Crowd-Sourcing Scoping Study - Engaging the Crowd with Humanities Research'. Centre for e-Research, King's College London. http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/12/Crowdsourcing-connected-communities.pdf

Edwards, A., Housley, W., Sloan, L., Williams, M.L. and Williams, M. (2013): 'Digital Social Research and the Sociological Imagination: Surrogacy, Augmentation and Re-orientation'. International Journal of Social Research Methodology, Computational Social Science: Research Strategies, Design and Methods.

Michael, J. (2007): '40000 Namen, Anredebestimmung anhand des Vornamens'. *http://www.heise.de/ct/ftp/07/17/182/*

Morgan, J., Sloan, L., Housley, W., Williams, M.L., Edwards, A., Burnap, P. and Rana, O. (2013): 'Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter'. *Sociological Research Online*.

Norman, G., Norris, C., Gollan, J., Ito, T., Hawkley, L., Larsen, J., Cacioppo, J. and Berntson, G.G. (2011): 'Current emotion research in psychophysiology: The neurobiology of evaluative bivalence'. Emotion Review, 3, 3349-59.

Procter, R., Vis, F. and Voss, A. (2013a): 'Reading the riots on Twitter: methodological innovation for the analysis of big data'. International Journal of Social Research Methodology, Special Issue on Computational Social Science: Research Strategies, Design & Methods.

Procter, R., Crump, J., Karstedt, S, Voss, A. and Cantijoch, M. (2013b): 'Reading the riots: What were the Police doing on Twitter?' Policing and Society, Special issue on policing and cybercrime, April.

Raimond, Y. and Lowis, C. (2012): 'Automated interlinking of speech radio archives'. WWW 2012 Workshop on Linked Data on the Web, Lyon, France. http://ceur-ws.org/Vol-937/ldow2012-paper-11.pdf

Thelwall, M., Buckley, K. and Paltoglou, G. (2012): 'Sentiment strength detection for the social web'. Journal of the American Society for Information Science and Technology, 63(1).

Thelwall, M. and Buckley, K. (in press): 'Topic-based sentiment analysis for the social web: The role of mood and issue-related words'. Journal of the American Society for Information Science and Technology.

Willett, W., Heer, J. and Agrawala, M. (2012): 'Strategies for crowdsourcing social data analysis'. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, pp. 227-236. ACM.