PHEME: Veracity in Digital Social Networks

Leon Derczynski and Kalina Bontcheva

University of Sheffield, UK
{leon, kalina}@dcs.shef.ac.uk

Abstract. PHEME attempts to identify four kinds of false claim in social media and on the web, in real time: rumours, disinformation, misinformation and speculation. This brings challenges in modelling the behaviour of individual users, networks of users and information diffusion. This presentation proposal discusses the issues addressed by the project and the challenges it faces, in this emerging and rapidly-developing domain.

Keywords: social media, rumours, veracity, social network analysis

1 Introduction

Social networks are rife with lies and deception, half-truths and facts. The rapid spread of such information through social network sites and other online media can have immediate and serious consequences. For example, social sensors already inform live decision-making about incoming epidemics (such as West Nile Virus), fire engine dispatch (to handle fires in the Australian bush), and earthquakes in both Japan and the USA; in these cases, false information must be filtered quickly. To detect these rumours, large amounts of user-generated content need to be analysed quickly, yet it is not currently possible to carry out such complex analyses in real time. The PHEME project (http://www.pheme.eu) aims to model, identify, and verify claims annotated for truthfulness or deception as they spread across media, languages, and social networks.

Modelling user behaviour presents major challenges in this project. Individuals and groups send messages with different degrees of veracity, giving each author a different level of reliability. Further, people choose whether or not to propagate messages within their social network or to a new web venue. Capturing these behaviours, at both macro and micro levels, is a core part of PHEME.

2 Influence

Information diffusion plays a crucial role in a range of phenomena, including the spread of rumours over and between social networks.

Tracking information flow in implicit networks is a more challenging task because it involves: identifying identical information units; determining when this information was published; tracking the flow within this network; and inferring the implicit diffusion network.

Implicit networks created by dialogue can also be found over explicit social networks such as Twitter. For example, those contributing to a hashtag conversation are

not bound by explicit links such as follower or friend relations, but instead cause information to diffuse among those interested in that topic [1].

As a second stage, PHEME extends models to consider individual users and how they react to incoming rumours, critically seeking to characterise what types of incoming message are likely to be propagated. We model rumour spread using contact process models of epidemiology, whereby a disease is said to spread over a graph according to a Markov process. At each moment in time an ill individual (i.e. a user who currently believes a rumour) may infect a neighbour (i.e. share the rumour with them) or recover from the illness (i.e. realise they were mistaken due to new information), both with given probability.

In the context of the project, we will present models for influence, and models for the spread of changes over networks.

3 Influence and trust in social networks

Since PHEME deals with a wide variety of content, it is important to model trust as a mechanism for deciding which of a set of sources is the reliable one. Network models for trust are already used in information retrieval and have been adapted for Twitter. These typically take a PageRank-style score and propagate trust between nodes in the network.

With respect to influence, PHEME delivers methods for corroborating or refuting claims, leading to an estimate of a source's trust and reputation. This gives a means of capturing linguistic and network behavioural datasets [2] of users who trust / distrust sources.

The trustworthiness of a user/web site depends on the veracity of past content. The opposite is also true. It follows that veracity of a given message depends, amongst other things, on the trustworthiness of its author.

PHEME uses historical Twitter data dating back to 2009, the SWI, METER and APA news corpora, and the linked blogs, forums, and web content. These are searched for known false rumours, acquired from fact-checking websites, in order to create automatically large amounts of training data on past rumours and their spread. This in turn informs longitudinal user models.

We will present motivating examples of challenges in this scenario, and a summary of our existing research on longitudinal models over social networks.

4 Dynamic and transient information

PHEME addresses the spatio-temporal validity of claims, to find contradictions.

The temporal validity of facts needs to be taken into account when detecting contradictions and identifying rumours. It is possible to extract two truths that seem to contradict (e.g. "The president of the USA is George W Bush" and "The president of the USA is Barack Obama") but are in fact both accurate when the appropriate temporal information is added.

Similarly, claims have spatial constraints, especially when they pertain to elided contexts. For example, we may say "The president is Obama" and "The president is

Hollande"; without other knowledge, these are in conflict, but are in fact both true – just in distinct spatial regions.

PHEME seeks to develop tools for annotating and determining the temporal and spatial contexts of claims [3], and cross-referencing these to detect conflicting assertions, as a feature for detecting rumours, mis- and disinformation.

We will describe the anatomy of news articles, discuss existing approaches to temporal bounding, and the challenges in adapting these to social media texts.

5 Posts, networks and diffusion

PHEME involves the development of ontologies to model users, social posts and social network, as well contradictions, temporal bounding, and so on. The project involves building new and extended ontologies to model veracity, misinformation, social and information diffusion networks, rumours, disputed claims and temporal validity. It draws a distinction between content authors, receivers, and diffusers. This includes modelling the temporal validity of statements (e.g. Lenin was born in the Soviet Union vs. in Russia) and lexicalisations (e.g. Kaliningrad vs. Königsberg), based on work on adding temporal arguments to RDF triples.

We will give information on the ontological approach developed so far, accompanied by examples.

6 Representing dynamic flow in social graphs

The main challenge in browsing and visualisation of interlinked media and social network content is in providing a suitably aggregated, high-level overview. Timestamp-based list interfaces that show the entire, continuously updating stream (e.g. the Twitter timeline-based web interface) are often impractical, especially for analysing high-volume, bursty events.

The project incorporates visual analytics tools for collected veracity intelligence, including visualisations of geospatially and semantically referenced information, across news, media and social networks. Exploring the storytelling potential of big data visualization [4], the interactive components of PHEME are intended to increase the understanding of the complementary relationship between the explorative and communicative dimensions.

We will give examples illustrating the challenges in visualising this information and information regarding our planned approach.

7 Decision support

Modelling and reasoning with rumours is particularly challenging, due to the need to represent multiple possible truths (e.g. superfoods may cause vs. prevent cancer). The reasoning is parameterised further in accordance to the domains of the two use cases (healthcare and digital journalism).

PHEME addresses how the new veracity intelligence methods can be applied to a health-related use case, and how social media analysis can be integrated with public health monitoring and with analysis of the electronic patient record (EPR). Specific topics for demonstration are: Among other topics, PHEME will analyse the impact of public health concerns and health-related rumours, including issues of medications, trace elements and food additives (e.g. Alzheimer's, autism, Attention Deficit Hyperactivity Disorder).

The project also prototypes an open-source digital journalism tool, to support the cross-linking, verification, analysis, and visualisation of veracity, operating across media and languages. A real-time news platform, SwiftRiver, is to be used to test and develop rumour detection algorithms. Spatio-temporal knowledge plays also an important role. A key challenge is to identify the regionality of events (e.g., neighbourhood, city, or country level).

We will present these two motivating scenarios for rumour detection and discuss the potential impact and issues in each case.

8 Conclusion

PHEME addresses veracity in social networks, and attempts to identify rumour, misinformation, disinformation and speculation before it has a potentially harmful impact. The technologies required to do things come with big challenges, especially in the areas of user modelling, user interaction, and information diffusion. We describe a few of these specific challenges and present them in their context, as well as some of our proposed approaches.

9 Acknowledgments

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 611233, PHEME.

References

1. Magnani, M., Montesi, D., Rossi, L.: Conversation retrieval for microblogging sites. Information retrieval **15**(3-4) (2012) 354–372
2. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Machine learning and knowledge discovery in databases. Springer (2011) 18–33
3. Derczynski, L., Bontcheva, K.: Spatio-temporal grounding of claims made on the web, in PHEME. In: Proceedings of the 10th joint ACL-ISO workshop on Interoperable Semantic Annotation, ACL (2014)
4. Segel, E., Heer, J.: Narrative visualization: Telling stories with data. Visualization and Computer Graphics, IEEE Transactions on **16**(6) (2010) 1139–1148