Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance

Maged N. Kamel Boulos a,∗, Antonio P. Sanfilippo b, Courtney D. Corley b, Steve Wheeler c

aFaculty of Health, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK
bTechno-Social Predictive Analytics Initiative, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA
cFaculty of Education, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK

ABSTRACT

This paper explores Technosocial Predictive Analytics (TPA) and related methods for Web “data mining” where users’ posts and queries are garnered from Social Web (“Web 2.0”) tools such as blogs, micro-blogging and social networking sites to form coherent representations of real-time health events. The paper includes a brief introduction to commonly used Social Web tools such as mashups and aggregators, and maps their exponential growth as an open architecture of participation for the masses and an emerging way to gain insight about people’s collective health status of whole populations. Several health related tool examples are described and demonstrated as practical means through which health professionals might create clear location specific pictures of epidemiological data such as flu outbreaks.

© 2010 Elsevier Ireland Ltd. All rights reserved.

doi:10.1016/j.cmpb.2010.02.007

1. Background

The rapid development and growing popularity of the Social Web, also commonly known as “Web 2.0”, and its applications, coupled with the increasing availability of mobile Internet-enabled devices (such as netbooks and Apple iPhone), has ensured that the Internet is now a truly integral part of everyday life. The Social Web is a particularly dynamic medium that captures the “pulse” of the society in real time. Blogs, micro-blogs such as Twitter (http://twitter.com/), and social networking sites such as Facebook (http://www.facebook.com/), enable people to publish their personal stories, opinions, product reviews and a great deal more in real time. They can also share geotagged alerts and reports about their current location and allow others to follow their whereabouts, all in real time, using location-aware social services such as Microsoft Vine (http://www.vine.net/). Such applications have greater power when coupled to Social Web ‘mashup’ services and aggregators. Such tools weave data from disparate sources into a new, compound data source or service [1,2]. A useful example of the latter is HealthMap, the global disease alert map (http://www.healthmap.org/en).
The dynamic nature and continuously updated features of the Social Web make it a fertile environment for intelligence gathering in a variety of disciplines, enabling users to tap into the ‘wisdom of the crowds’. Further, the Social Web is useful for gleaning the collective ‘impression of the masses’ regarding current matters, events and products [3]. In fact, it is not uncommon these days for example, to see shoppers (including those who do not make their purchases through the Internet) consulting online user reviews and ratings on sites such as Amazon.com about products they are planning to buy before making their purchase decisions.

This paper will briefly review a number of emerging technologies and tools, including Technosocial Predictive Analytics pioneered at Pacific Northwest National Laboratory, USA (http://predictiveanalytics.pnl.gov/), that can be used to exploit these inherent Social Web features in real or near-real time and harness them for public health, environmental and national security surveillance purposes.

2. A brief introduction to Technosocial Predictive Analytics and some related technologies and tool examples

2.1. Who Is Sick: the power of user-generated maps

‘Who Is Sick’ (http://whoissick.org/sickness/, Fig. 1) was launched in 2006 as a location-aware service for tracking and monitoring sickness, with the aim of providing current and local sickness information by the public to the public. People can anonymously post information about their own illness and use a familiar Google Maps interface to search and filter illness by symptoms, sex, age, and location. The maps act as a crude syndromic surveillance tool in real time. A person feeling unwell can instantly check what sicknesses are circulating within their own locale. Health-conscious individuals conduct similar scans and take any necessary preventive measures in preparation. People traveling to another area of the country or abroad can use the service to learn if there are any sicknesses going around that they need to be aware of. One should be warned however, that information quality is always dependent upon the accuracy of user-reported details and on sufficient numbers of people (who are actually sick) using the tool to report their symptoms in a timely manner.

2.2. We Feel Fine: differentiating between informative and affective text on the Social Web

Created in 2006, ‘We Feel Fine’ (http://www.wefeelfine.org/) is a Web-based applet and API (Application Programming Interface—http://wefeelfine.org/api.html) that can be used to assess the fluctuating effects of real world events and factors such as the Stock Index or the weather on people’s feelings and emotions, and ascertain a crude idea about the ‘general mood’ of populations at different dates and locations around the world [4].

At the heart of ‘We Feel Fine’ is a data crawler that automatically scours the Web every 10 min, harvesting data in the form of human feelings from a large number of blogs and social pages such as those hosted by MySpace (http://www.myspace.com/) and Blogger (http://www.blogger.com/). ‘We Feel Fine scans’ blog posts for occurrences of the phrases “I feel” and “I am feeling”, and once any of these is found, the system goes backward to the beginning of the sentence and forward to the end of the sentence, and then saves the full sentence in a database. Saved sentences are next scanned against a manually compiled list of about 5000 pre-identified “feelings” (adjectives and some adverbs). If a valid feeling is found, the corresponding sentence is considered to represent one person who feels that way. ‘We Feel Fine’ also attempts to extract the username of the post’s author from the URL of his/her blog post (e.g., ‘steve-wheeler’ in http://steve-wheeler.blogspot.com/) and uses it to automatically traverse the blog hosting site to locate that user profile page. From the profile page, it is often possible to extract the age, gender, country, state, and city of residence of the blog owner. ‘We Feel Fine’ can also retrieve the local weather conditions for the blog owner’s city at the time the post was written. This process results in about 15,000 to 20,000 saved feelings (with associated data) per day. For display purposes, the top 200 feelings were manually assigned colours that loosely correspond to the tone of the feeling, e.g., happy
positive feelings are graphically represented using bright yellow and sad negative feelings are depicted in dark blue. The applet also offers a number of filtering options for querying the underlying database (Fig. 2).

While ‘We Feel Fine’ functions as intended with its pre-coordinated search strategy and list of feelings, it is rather limited and cannot fully capture all possible affective expressions and differentiate between them and informative sentences in Social Web text. One can easily appreciate how complex this task is by searching for some keywords: such as ‘flu’ at BingTweets (http://bingtweets.com/) and observing the nonstop ‘tweets’ in the resulting real-time Twitter feed pane: there will almost always be people ‘tweeting’ about their own flu sickness (e.g., “back to bed to try and sleep off this flu. My sinuses feel numb!”—counts as one flu case) and others just sending informative ‘tweets’ about flu (sharing a news item, some related research or information, or asking a question, e.g., “How business travelers can avoid swine flu: http://bit.ly/zBOuY”—does not count as a flu case). The challenge here is to devise an algorithm that can reliably differentiate between the two types of ‘tweets’ automatically and in real time. Ni et al. [5] describe one solution using a machine learning method for classifying informative and affective blog posts.

‘We Feel Fine’ also attempts to perform some geographical classification of the feelings it continually gathers from the blogosphere, but again it does so in a rather limited fashion. The limitations extend beyond looking at users’ profiles for location details (which are not always available), because there may also be issues to individual posts. A post by, for example a user from the USA, might relate to that user’s experience at a totally different location to the one they are normally residing in, e.g., the Blogger’s holidays in Italy. Fortunately, more sophisticated technologies exist today, such as MetaCarta (http://www.metacarta.com/), which can be applied to sift through Social Web text and extract much more comprehensive geographical references, while intelligently handling any geographical name ambiguities or inconsistencies in the scanned text [6].

2.3. Google Flu Trends: an example of infodemiology and infoveillance in action

Eysenbach [7,8] defines ‘infodemiology’ as “the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy”. He further goes on to define a related concept, ‘infoveillance’, as a similar science whose “primary aim is surveillance”. Infodemiology and infoveillance rely on automated and continuous analysis of unstructured, free text information available on the Internet. This includes analysis of search engine queries (or what Eysenbach calls the “demand” side), as well as of what is being published on Web sites, blogs,
Twitter, etc. (or the “supply” side, to use again Eysenbach’s terminology).

Building on the same ‘infodemiology’ and ‘infoveillance’ concepts described by Eysenbach, researchers at Google found that certain search terms are good indicators of flu activity [9]. Google Flu Trends (http://www.google.org/flutrends/) uses aggregated Google search data to estimate flu activity up to two weeks quicker than traditional systems. Such an early detection of disease activity (in close to real time), when followed by an appropriate rapid response, has the potential of reducing the impact of both seasonal and pandemic influenza. Trends (continuously updated) are currently available from Google for Australia, New Zealand, the United States (Fig. 3), Mexico and sixteen more countries. A related project, Infovigil (http://www.infovigil.com/), has recently been launched by Eysenbach and colleagues at the Centre for Global eHealth Innovation, Toronto, Canada, and is introduced in [8]. (See also the Technosocial Predictive Analytics approach to monitoring of seasonal influenza later in this paper.)

2.4. Emergency and public health virtual situation rooms using 3D serious gaming technology: getting the ‘big picture’ in real time

Conventional situation rooms are routinely used to oversee public health emergencies and disaster management operations in real time. Nowadays, large amounts of emergency data are increasingly coming from a wide range of sources in real or near-real time and need to be cross-linked and visualized where they spatially belong on maps of the affected regions. Also, emergencies are usually managed by multi-professional teams who, not uncommonly, are distributed in multiple geographic locations. Because of these reasons, we have started to see physical situation rooms gradually being replaced (or combined) with virtual situation rooms that use online collaborative (and mostly 2D—two-dimensional) platforms such as Depiction (http://www.depiction.com/), in addition to conventional Web conferencing. These platforms, although usable and helpful, leave much to be desired, as they are lacking the ‘third dimension’, which is needed to create a proper perception of the emergency space, as well as a sense of co-presence of other virtual team members [1,10,11].

To overcome this limitation, Kamel Boulos [11,12] proposed the development of novel emergency and public health virtual situation rooms in a suitable 3D (three-dimensional) virtual world (sometimes also called ‘3D serious gaming platform’), where animated 3D graphical self-representations (avatars) of experts and professionals can collaborate together and discuss incident data in real time in a simulated 3D space representing the physical location where the emergency/public health incident of interest is unfolding and reflecting all changes taking place there (again in real time). The real-time link between the virtual world and the physical world incident would be two-way and multimodal involving geo-tagged physical/environmental sensor data feeds, citizen-contributed data (as found in Microsoft Vine and ‘Who Is Sick’), data gleaned via automatic analysis of Social Web content, textual and 3D spatialized audio/voice exchanges between virtual team members, video feeds, 3D simulations and animations, various Web mashups, and shared desktop applications, among other possibilities. Various ‘what-if’ scenarios can also be explored and tested in the virtual environment, making the proposed rooms much suited for real-time emergency and disaster management, e.g., for managing an influenza pandemic and planning and coordinating actions at global, regional and local levels.

To achieve the above vision, the proposed collaborative and interactive situation room platform would marry virtual globes or 3D mirror worlds (such as Google Earth™—http://earth.google.com/) and 3D virtual worlds (such as Second Life™—http://secondlife.com/ and OpenSim—http://opensimulator.org/) [1,11,12], and complement and tightly integrate them with other key technologies, e.g., real time, geo-tagged RSS—Really Simple Syndication data feeds and geo-mashups (using Web services such as Yahoo! Pipes—http://pipes.yahoo.com/). The platform would weave data and services in real time from different sources into a new rich ‘datascape’ that better reflects the current situation (the ‘big picture’) in novel visual ways that are easier to understand and manage. The platform has to be secure, enabling multiple distributed persons to “see” each other, visualize relevant data together in unique ways, conduct 3D simulation and training scenarios, and collaborate in real time, each according to their assigned role and access privileges. True stereoscopic vision can be added, if needed, using readily available technologies such as NVIDIA 3D Vision (http://www.nvidia.com/object/3D_Vision_Overview.html) for more realistic 3D visualization, with a better sense of 3D depth and relief. It is noteworthy that it is now possible to stream a 3D virtual world to a suitable mobile phone (see, for example, http://www.youtube.com/watch?v=XwRnjbkjnc) or other Internet-enabled, small form factor mobile devices, making this vision end-user device and platform-independent and thus suitable for those members of the emergency operations team who are on the move.

2.5. Technosocial Predictive Analytics: a systems science approach to proactive anticipatory reasoning

Events occur daily that challenge the health, security and sustainable growth of our society, and often find us unprepared for the catastrophic outcomes. These events involve the interaction of complex processes such as climate change, emerging infectious diseases, energy reliability, terrorism, nuclear proliferation, natural and man-made disasters, and geopolitical, social and economic vulnerabilities. If we are to prevent the adversities and leverage the opportunities that emerge from these events, integrated anticipatory reasoning has to become an everyday activity. Technosocial Predictive Analytics (TPA[13]) supports anticipatory reasoning through a multidisciplinary approach to strategic analysis and response that enables a concerted decision-making effort by analysts and policymakers.

There is increased awareness among subject-matter experts, analysts, and decision makers that a combined understanding of interacting physical and human factors is essential in addressing strategic decision making proactively. For example, multilevel studies that consider a broad range of biological, family, community, socio-cultural, environmental, policy, and macro-level economic factors are necessary to
prevent and mitigate the emergence of health threats such as childhood obesity [14] and drug addiction [15]. Integrated understanding of the infrastructural, social and ideational context in which contentious social movements operate is an essential analytical step in framing the emergence of violent behaviour [16]. Combined understanding of anthropogenic effects (e.g., chemical waste) and natural processes (e.g., solar variation) is needed to predict the impact of global warming [17]. Insights on human cognitive and emotional biases improve our understanding of economic decisions and their effect on market prices, returns, and the allocation of resources [18]. TPA provides theoretical and operational advancements of these insights.

The human brain provides a basic framework for memory and prediction in which decision making emerges as a process of pattern storage, recognition and projection rooted in our experience of the world and driven by perception and creativity [19]. Qualities such as the ability to focus on what is perceived to be most important “in the moment” and the capacity to make quick decisions by insight and intuition make human judgment uniquely effective [20,21]. However, the same qualities can also be responsible for fallacious reasoning when judgment lacks sufficient knowledge and expertise [22], and is biased by factors such as “group-think” [23,24], limitations in memory and focus attention [25,26], positive framing [27] and increased confidence in extreme judgments and highly correlated observables [28]. TPA addresses these gaps by using knowledge management and modeling to inform human judgment and analytical gaming to neutralize biased reasoning. More specifically, TPA helps analysts and policymakers provide better proactive analysis and response by enabling naturalistic decision making while countering adverse influences on human judgment through a combined set of capabilities that (a) provide multidisciplinary knowledge reach-back to inform analysis and response during decision making; (b) supplement the expertise of the analyst and policymaker with simulated scenarios generated by computational models that integrate human and physical factors, and (c) engage analysts and policymakers within a gaming environment that stimulates creative critical reasoning through visual analysis and collaborative/competitive work, as described in Fig. 4a.

TPA offers a collaborative visual analytic environment with user-friendly access to predictive models in the areas of security, energy and the environment that are informed by knowledge management processes, as shown in Fig. 4b.
Such an environment promotes model integration through innovative algorithms that enable the combination of diverse modeling paradigms such as Bayesian networks and system dynamics. Integrated modeling is informed by a Knowledge Encapsulation Framework that facilitates the distillation and vetting of relevant knowledge from heterogeneous data streams, including social media. The TPA collaborative working environment enables analysts and policymakers to stress-test the validity of their analysis and policy plans using techniques such as role-playing and serious gaming. Gaming data are stored over time and analyzed to make inferences about strategic decision making in context from social interaction during game-play.

TPA provides an ideal systems science approach for public health informatics with reference to challenges such as emerging infectious diseases. Current epidemic models characterize effects of transportation, demographic traits, schools, healthcare access, and host-to-host dynamics. TPA can extend...
these existing modeling paradigms through the integration of socio-cultural models to provide a more accurate characterization of how public health responses reflect factors such as social order, individual and community-based health knowledge, and behaviour change due to public health communications in mass media and emerging Web technologies.

One specific application of the TPA approach to public health informatics is the monitoring of seasonal influenza through Web and social media. As described earlier in this paper, many Internet users formulate search queries and post entries in blogs and discussion forums using terms related to influenza-like-illness (ILI) to identify sites and other users that offer medical diagnosis and advice [9,29]. There is a clear sense in which an increase/decrease in the number of ILI-based searches and posts in blogs and discussion forums reflects a higher/lower public concern for the spreading of influenza and can therefore be used to monitor seasonal influenza. Such an insight is corroborated by observed correlations of ILI-based blog posts and surveillance reports from sentinel healthcare providers. For example, Fig. 5 plots ILI-based blog posts (source: http://www.spinn3r.com) with outpatient ILI Surveillance Network (ILINet) by the US Centers for Disease Control and Prevention on a weekly basis for the period from October 05, 2008 to January 25, 2009. The evaluation of Pearson’s correlation coefficient for the two data series yields a strong ($r = 0.545$) and statistically significant (95% confidence) correlation value. Such results and analyses can greatly improve the quality of existing models of seasonal influenza.

3. Conclusions

Technosocial Predictive Analytics is still in its infancy, but nevertheless provides health professionals with a powerful method of exploiting vast, continuous streams of user-generated content on a rapidly proliferating worldwide social network. As the technology develops it will provide more sophisticated and reliable tools for public health use. It must be emphasized however, that such techniques are only applicable where large proportions of a society regularly use Social Web services. There are still many people and sometimes entire communities who do not use the Web on a regular basis. For such communities there may be no easy way to gain general health data, and it is inadvisable to extrapolate findings from other communities in an attempt to predict their health state. However, as Social Web use increases exponentially due to technology and connectivity becoming more widely available and affordable, this situation will rapidly change. There is therefore a need to monitor the continuously shifting demographic characteristics of a variety of Social Web services that populations are using. Each social network service and its prevailing user characteristics such as age, gender and prevalent user geographic locations should be considered as these data can help in better interpreting any intelligence or results derived from these services.

**References**


