Synthesis

International Scientific Conference of IT and Business-Related Research

APPLYING GIS IN GEOMODELING OF SOCIAL NETWORKS

PRIMENA GEOGRAFSKIH INFORMACIONIH SISTEMA U GEO-MODELOVANJU DRUŠTVENIH MREŽA

Verka Jovanović[1], Đorđe Vukelić[2]

[2]Singidunum University, Danijelova 32, Belgrade, Serbia
[2]RMS, Malta

Abstract:

The main characteristic of the new World Wide Web trend, also known as Web 2.0, is the ability to share information online and create different modes of collaboration. Web 2.0 has been replacing the static nature of the World Wide Web with the dynamic and real-time content. In such a transition process, the social nature of the World Wide Web has changed. Websites or other forms of Internet applications have started to become more community based, thus allowing interaction, collaboration and content-sharing among users. At the same time, it provides a different view of the complex social networking and cultural dynamics within a society. Accordingly, social media feeds are becoming increasingly geosocial in the sense that they often have a substantial geographic content. Such information contributes with the additional feature to social media (location) and provides additional context for the analysis of these data (topics and sentiment). As such, it represents an emerging alternate form of geographic information, which, with its volume and richness, opens new avenues and imposes research challenges for the understanding of dynamic events and situations. The use of Geographic Information System in GeoModeling of social networks is inherently complex, as it comprises the study of various types of content, connections and locations.

Key words:

GIS, geosocial, modeling, networks, Internet.

Apstrakt:

Osnovna karakteristika druge generacije Veb-a i internet usluga, poznatije pod terminom Veb 2.0, jeste sposobnost razmene informacija na Internetu i realizacije različitih vidova saradnje. Veb 2.o pruža dinamičko korisničko iskustvo u stvarnom vremenu, menjajući tako svoju prirodu i strukturu. Internet sajtovi ili drugi vidovi Internet aplikacija sve više se orjentišu ka zajednici i socijalizaciji, omogućavajući interakciju, saradnju i razmenu sadržaja između korisnika interneta i savremenih veb aplikacija kao i između samih korisnika. Naime, to pruža drugačiji pogled na složen proces društvenog umrežavanja i kulturnu dinamiku unutar samog društva. Shodno tome, sadržaji društvenih medija postaju sve više geo-socijalni u smislu da često nude značajan geografski sadržaj. To ukazuje na dodatno obeležje društvenih medija (lokacija) i pruža dodatni kontekst za analizu takvih podataka (teme i raspoloženje). Kao takva, ona predstavlja alternativni oblik geografskih informacija koji svojim obimom i bogatstvom otvara nove puteve i nameće istraživačke izazove neophodne za razumevanje dinamičkih događaja i situacija. Primena Geografskih informacionih sistema u geomodelingu društvenih mreža predstavlja proces koji je sam po sebi složen, jer obuhvata izučavanje različitih tipova sadržaja, veza i lokacija.

Ključne reči:

GIS, geosocijalni, modelovanje, društvene mreže, Internet.

1. INTRODUCTION

The term social networks refers to a different spectrum of digital interaction and information platforms. It includes blogs (*e.g.* Twitter, Blogger, WordPress), digital social services (*e.g.* Facebook, Google+) and multimedia content sharing services (YouTube). Although they are different, these social media services share the common goal of enabling the general public to make a contribution, disseminate and exchange information (Kaplan & Haenlein, 2010).

Along with the emergence of Web 2.0, ubiquitous computing and corresponding technological advancements, social media have become extremely popular over the last decade. At the same time, social media content is rapidly increasing. In 2012, Facebook announced that its system deals with petabytes scale data as it processes 2.5 billion content elements and over 500 TB of data daily (Borthakur, 2012).

The geographic content of social media content represents a new type of geographic information. It does not fall under the established geospatial community definitions of crowdsourcing (Fritz *et al.*, 2009) Instead, the type of geographic information that can be "collected" from social media feeds can be referred to as ambient geographic information (AGI) (Stefanidis *et al.*, 2013). Although it might be related to the originator of feed, it is most of the time embedded in the content of these feeds, often across the content of numerous entries rather than within a single one and has to be somehow extracted.

2. BIG GEOSOCIAL DATA

Currently, the definition of big data (TechAmerica, 2012) is moving beyond the data volume, with two additional properties: velocity and variety.

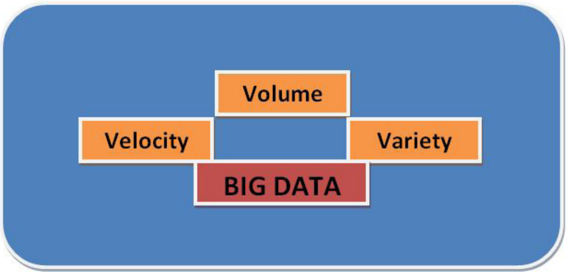

Figure 1. The three dimensions of big data

E-mail: dvukelic68@gmail.com

DOI: 10.15308/Synthesis-2015-620-623

The velocity refers to the rate at which the data is produced and the currency of its content, as well as the need for timely analysis. Such need to process data and extract information from them at the streaming rates is imposing substantially higher computational demands than the periodic (*e.g.* daily or weekly) processing of comparable information (*e.g.* the case of remote sensing data).

On the other hand, the variety is the diversity of the data sources and types that are processed and it depends on the degree to which the information to be extracted is distributed among such diverse sources. It is a very common scenario for an event to be communicated by the general public in fragments, as the individuals may only have a partial view of the event they are reporting. At the same line, the event may be communicated across numerous social networking channels by multiple users by means of various modalities (*e.g.* text in Twitter, images in Instagram, and videos in YouTube).

The example of the above mentioned situation are the Mumbai terrorist attacks in 2008, where Flickr imagery, Twitter streams, Google maps mashups and Wikipedia articles were set up immediately, to provide real-time coverage of the unfolding events (Arthur, 2008). Each piece of information separately is important for understanding the event, only the aggregate view of all these pieces would offer a far better understanding and the full complexity of the actual event. It is analogous to information aggregation in a Geosensor network (Stefanidis & Nittel, 2004), where each sensor contributes with a piece of information, but it is through aggregation across that the observed event is revealed in all its complexity. As regards social media, people also act as sensors, by reporting their observations in the form of multimedia feeds, and the challenge is to compose these fragmented contributions into a bigger picture, by overcoming the limitations of individual perception.

Since the first days of Geographic Information System (GIS), geospatial datasets have always been large volume datasets always near the edge of the computational capabilities of each era. This was true at the time of early seminal computer mapping software environments in the late 1960s and early 1970s, such as SYMAP and its use in Waldo Tobler's movie of the urban expansion of Detroit, and it is certainly true today. The NASA Earth Observing System Data and Information System (EOSDIS) is estimated to have over 7.5 PB (one PB is equal to 1,024 terabytes) of archived imagery and is currently generating approximately 5 TB of data daily. Furthermore, the proliferation of Google Earth has led to Google maintaining an archive of over 20 PB of imagery, from satellite to aerial and ground-level street view imagery (McKenna, 2014).

Technological advances will move the geospatial community further into big data territory, by broadening the variety of geospatial datasets that are collected and analyzed, and by improving the rates at which such data are generated.

It is very clear that there is a need to formulate the right model to process Geosocial data in order to extract knowledge from diverse social networks feeds.

3. GEOSOCIAL COMPLEXITY

Geosocial data are also differentiated from "traditional" geospatial information due to their complexity. Specifically, they are predominantly linked information; links are created among users to establish a social network, and among words to define the "topic' that is communicated through pieces of information.

As regards user connections, they can be established through specific actions. For instance, the user A by following, replying, referring to, or retweeting user B can establish a connection between them. However, it could be the same "topic" where different users are sending the information. Collection all of these connections could provide a view of the users as a networked community that can be represented as a graph.

Very similar to users, words are also connected, as they appear in the same messages, to form word networks that define the discussion topic. Defining the Geolocation tags or words related to them (name of the city, river *etc.*) in that particular discussion is another challenge.

The process of collecting and representing this type of the data can be very complex. A collection of tweets captured during Hurricane Sandy (October 2012) resulted in a data corpus of nearly 10 million tweets, including over 4.7 million retweets (47.3% of the data corpus) and over 4.1 million (41.7% of the data corpus) links to external sites (*e.g.* news websites). The analysis of such highly complex datasets is a computational challenge, which is today typically addressed through graph analysis.

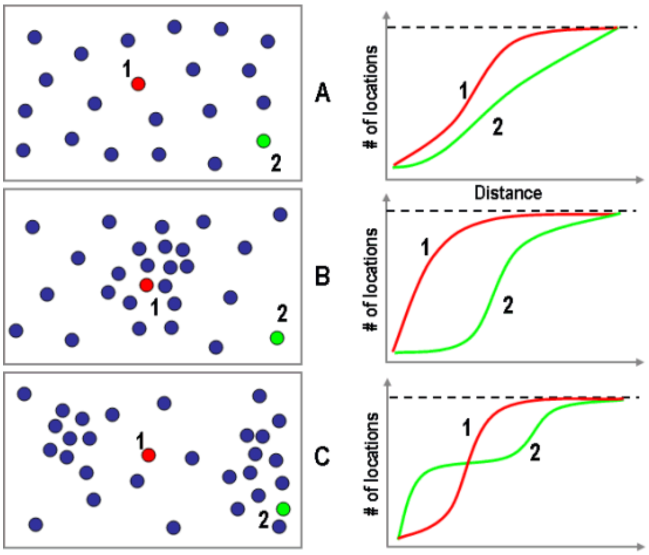

Figure 2. The relation between nodes in the networks and distance (Rodrigue *et al*, 2013)

Centrality, topological, and distance measures are fundamental metrics to support such analysis:

- *Centrality metrics*: Centrality is one of the most important structural attributes of a social network (Freeman, 1979). It is defined in terms of reachability of a given node in a network. In recent years, numerous measures have been proposed, including degree, closeness, betweenness, and flow betweenness (Scott, 2012). One of the betweenness measure reflects the influence of a node in the network, often measured as the number of times a node acts as a bridge along the shortest path between the two other nodes (Wasserman *et al.*, 1994). PageRank (Page *et al.*, 1999) and the Katz centrality (Katz, 1953; Borgatti, 2005) are two other measures that are closely related to the eigenvector centrality measure. Computing these features on large graphs is superlinearly expensive. The scalability of these algorithms in terms of three graph partitioning methods and GPU implementation of these measures is not an easy task.
- Topological features: Topological features are dependent on a structure of the graph. They mainly include degree distribution and clustering coefficient. The degree distribution is the probability distribution of the degree (of

DOI: 10.15308/Synthesis-2015-620-623

nodes) over the entire network. It is an important measure, as random graphs often have binomial or poison distributions, whereas real-world networks are highly skewed (Bollobas *et al*., 2001). The clustering coefficient determines the degree to which the nodes tend to cluster together (Huang, 2006).

- Similarity and distance measures: Searching for similar patterns in heterogeneous information networks is a complex task. Heterogeneous networks are directed graphs, which contain multiple types of objects or links. Recently, several new similarity measures have been defined for heterogeneous networks (Sun, 2012). The Figure 2 represents the relationship between the nodes in the networks and distance.

The graph analysis is a very important instrument for the analysis and understanding of social networks. However, the spatial component of the real physical place does not exist in any of these models.

4. MODELING GEOSOCIAL DATA

The Social media feeds allow for the first time to explore the physical presence of people together with their online activities, enabling us to link the cyber and physical spaces on a massive scale.

As such, it represents an emerging alternate form of geographic information, with the main particular characteristics being as follows:

a) Social media datasets are streaming big data that are best-suited for real-time analysis
b) Social media data are non-curated and their reliability varies substantially
c) The spatial distribution of social media contributions is non-uniform and heavily skewed

Currently, social media services offer a wide range of platforms using various technologies As a result, their content tends to be very diverse both in content—ranging from text to photos and videos—and in form, ranging from structured content to semi- or non-structured content. In addition, the form of raw social media data tends to be unstructured or ill-defined, thus making valuable knowledge hidden and limiting the capability to process it through automation (Sahito *et al*., 2011).

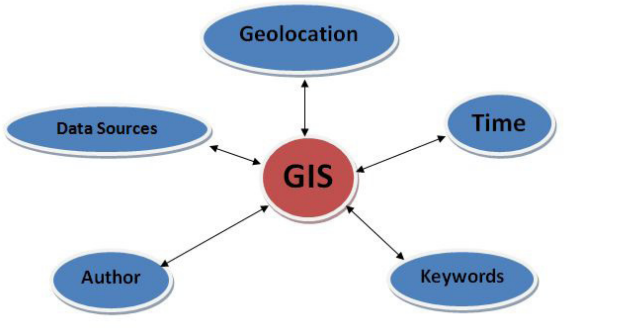

Figure 3. The components of a new model

GIS uses spatio-temporal (space-time) location as the key index variable. Modern GIS needs to move discussion away from broad theoretical positions and focus on a specific project area and its unique distribution of conditions and potential uses. In this direction, GIS might have a unique ability to be the right solution for Geosocial modeling of Social Networks.

The components of new model (see Figure 3) will include:

- ***Data sources***
  Data source serves as a directory of different source types that can provide data/information to the analysis framework. This information is uniquely identified and is linked to one author instance of the author component. In addition, each entry is associated with a time stamp indicating when the entry instance was published. Every Data source has the Application programming interface (API). The source API data are comprised of the data elements that are unique to each social media source. For instance, the Twitter API returns a tweet attribute that contains the content of a tweet, while the Flickr API returns a photo URL and has no equivalent to the tweet attribute. Such source-specific attributes, which are driven by the characteristics of each social media source, are considered source dependent and therefore, after the extraction of the data, the process of data normalization will be necessary.
- ***Author***
  The author's instances represent social media service users that contributed to the content. As user identification across sources is rather limited, each social media service creates its own author namespace, *i.e.*, a unique set of user names. As an author is identified by a tuple of a user name and the social media service identifier, different users can have the same identifier value in different services. It should be noted that the authors can be referenced to in the content of social media feeds, through which the underlying social network can be recovered.
- ***Geolocation***
  Geolocation information for social media feeds can be inferred indirectly from content analysis, or it can be extracted directly from the data itself. Also, the contributors themselves may directly provide Geolocation information, either in the form of exact coordinates or as a toponym (*e.g.* listing a city name) that in turn can be Geolocated using some service. It is important to highlight various forms of this Geolocation content, as already mentioned in the works of Croitoru *et al.* (2012) and Crooks *et al.* (2012).
- ***Time***
  Temporal information can be typically found in all social media platforms. In this model, time information is embedded with each entry instance along with a time stamp-type identifier.
- ***Keywords***
  As part of social media entry, users contribute with keywords or tags, such as hashtags (#) to emphasize views and ideas and engage other users (Romero *et al*., 2011). Hashtags also support the building of semantic networks by allowing individual tweets to be linked thematically based on their content. Unlike explicit tagging, implicit keyword may emerge from user conversations, when certain words become widely adopted as a result of a noteworthy event.
- ***GIS***
  Managing and integrating such diverse social media data requires the development of a unified conceptual data model that will support various data structures under a single scheme. By using GIS capability to operate with different type of the data (attribute and spatial), a new

DOI: 10.15308/Synthesis-2015-620-623

model would be able to "convert" social media data into structured Geosocial information, from which knowledge can be extracted through further analysis.

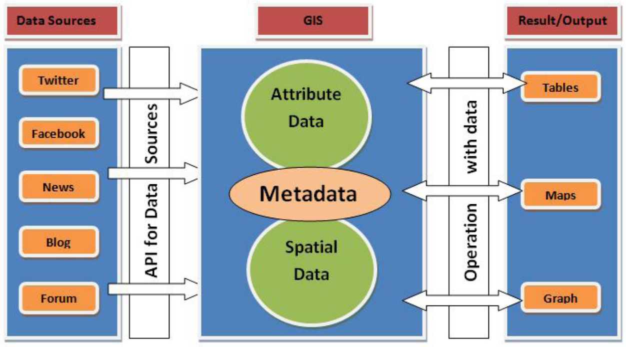

Figure 4. High level architecture of software solutions based on a new model

By means of the above-mentioned model, high level architecture of software solution is presented (see Figure 4). Input to the system from different data sources should be done via layer API for data sources, where the result will be "normalized" and "structured" information in a form of attribute or spatial data. GIS will process the data and include the specific metadata. The definition of metadata should be done for each system differently and must include predefinition (input from other system), collection of data (machine learning), and correction (user intervention). Based on the different operations and together with metadata, the layer operation with data will give the result/output. This result/output might assume different forms such as tables, maps, graph *etc.*

5. CONSLUSION

The emerging development of social media is imposing a new challenge to the Geoinformatics community. Such data ("big data") would have the particular characteristics that differentiate them from traditional geospatial datasets and that could enable us to monitor human observations at a massive scale and to cross-reference such data across a variety of sources and modalities (*e.g.* text, imagery, video, and audio). It presents a unique opportunity to validate information regarding the events as they unfold in space and time.

Managing and integrating such diverse social media data requires the development of a unified conceptual data model that will support the various data structures under a single scheme. Due to the particularities of the analysis that these data support, the model presented uses a hybrid mix of spatial and social analysis with the central role of GIS.

The recommendations for future research will be to create the prototypes of the software solutions based on the above-mentioned model.

REFERENCES

Arthur, C. (2008). *How Twitter and Flickr recorded the Mumbai terror attacks*. Retrieved March 01, 2015, from http://bit.ly/MIMz ().

Bollobas, B., Riordan, O., Spencer, J., & Tusnady, G. (2001). The degree sequence of a scalefree random graph process, *Random Structures & Algorithms*, 18(3), 279-290.

Borgatti, S.P. (2005). Centrality and network flow. *Social Networks*, 27, 55-71.

Borthakur, D. (2012). *Petabyte Scale Data at Facebook.* Retrieved March 03, 2015, from http://www.infoq.com/presentations/Data-Facebook

Croitoru, A., Stefanidis, A., Radzikowski, J., Crooks A.T., Stahl J., & Wayant, N. (2012). Towards a collaborative GeoSocial analysis workbench. Washington, DC: *COM-Geo.*

Crooks, A.T., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1), 124-147.

Freeman, L.C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215-239.

Fritz, S., MacCallum, I., Schill, C., Perge,r C., Grillmayer, R., Achard F., Kraxner F., & Obersteiner M. (2009). Geo-wiki.org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3), 345-354.

Huang, Z. (2006). Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. Workshop on Link Analysis: Dynamics and Static of Large Networks, Philadelphia, PA.

Katz, L. (1953). A new index derived from sociometric data analysis. *Psychometrika*, 18, 39-43.

McKenna, B. (2014). *What does a petabyte look like?* Retrieved March 02, 2015, from http://www.computerweekly.com/feature/What-does-a-petabyte-look-like

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking:Bringing order to the web. *Technical report.* Stanford, CA: Stanford InfoLab.

Romero, D., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags and complex contagion on twitter. Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, pp. 695-704.

Rodrigue, J.P., Comtois, C., & Slack, B. (2013). *The Geography of Transport Systems*. New York: Routledge.

Sahito, F., Latif, A., & Slany, W. (2011). Weaving twitter stream into linked data: A proof of concept framework. Proceedings of the 7th International Conference on Emerging Technologies, Islamabad, Pakistan, pp. 1-6.

Scott, J. (2012). *Social Network Analysis.* London, UK: SAGE Publications.

Stefanidis, A., Cotnoir, A., Croitoru, A., Crooks, A.T., Radzikowski, J., & Rice, M. (2013). Statehood 2.0: Mapping nation-states and their satellite communities through social media content. *Cartography and Geographic Information Science*, 40(2), 116-129.

Stefanidis, A., & Nittel, S. (2004). *GeoSensor Networks*. Boca Raton, FL: CRC Press.

Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2), 1-159.

TechAmerica. (2012). Demystifying big data: A practical guide to transforming the business of government. *TechAmerica Foundation*, *Federal Big Data Commission*, Washington, DC. Retrieved March 02, 2015, from http://bit.ly/11CpPDc.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. Cambridge, UK: Cambridge University Press.

DOI: 10.15308/Synthesis-2015-620-623