Clustering data with heterogeneous spatiotemporal reference: towards web-mining of event-related knowledge

B. De Longueville[1], M. Hardy[1]

[1]Institute for Environment and Sustainability, Joint Research Centre, via E. Fermi,1, 21020 Ispra, Italy
Email: bertrand.de-longueville@jrc.ec.europa.eu, matthew.hardy@jrc.ec.europa.eu

1 Introduction

Since Web 2.0 provided Internet with colloquial read-and-write functionality, the quantity of digital information accessible online is growing at an even more overwhelming rate than previously. As a consequence, scientists are faced with a ‘data tsunami’ from which it is increasingly arduous to extract valuable information (Shah et al. 2010). Knowledge discovery from large amounts of online contents is known as web mining (Etzioni 1996), and is a challenging field of research.

When the information created online by users has a spatial reference, it is known as Volunteered Geographic Information (VGI , Goodchild 2007). The potential of web mining of VGI has been demonstrated in numerous use cases related to e.g., natural hazards (De Longueville et al. 2010), environmental monitoring (Gouveia et al. 2004), socio-economic studies (Vaccari et al. 2009) or outdoor activities (Pultar et al. 2008), thus highlighting the relevance of web mining research to Geographic Information Science.

2 Background and problem statement

Data clustering, a prominent group of data mining techniques, is the unsupervised classification of patterns into groups (clusters)(Jain et al. 1999). A wide variety of spatiotemporal clustering techniques and algorithms have been applied to detect events in fields like epidemics (Rogerson 2001), crime (Johnson 2010) or meteorology (Hsu & Li 2010).

Data clustering involves similarity measurements between features in order to group those that are the most similar (Fischer et al. 1996). These features can be represented n-dimensional vectors noted:

$$x = (x_1, x_2, x_3, \ldots, x_n) \qquad (1)$$

where each x_x correspond to a specific measurement that characterizes the feature.

The Euclidean distance is commonly used to evaluate the similarity between features. It is defined as follows (Jain et al. 1999):

$$e(x_i,x_j) = (\Sigma_{k = 1 \text{ to } d} (x_{i,k} - x_{j,\,k})^2)^{1/2} \qquad (2)$$

where $x_{i,k}$ and $x_{j,\,k}$ are the kth values of feature vectors x_i and x_j respectively, and d is the dimension of these vectors.

In spatiotemporal clustering, some of the measurements x_x describe the position of the feature in space and time (e.g.: latitude, longitude, date, time) (Gong et al. 2006). On the basis of the similarity measurement between spatiotemporal features, various clustering algorithms (hierarchical, partitional, density-based, *etc*.) can be applied, depending on the nature of the events that are investigated (Getis & Ord 2010).

But whereas spatiotemporal clustering techniques are usually designed to deal with discrete, comparable objects such as sensor observations or tabular data records (Miller & Han 2001), VGI obtained through web-mining can be heterogeneous in terms of quality and accuracy (Flanagin & Metzger 2008). In particular, De Longueville et al. 2009 emphasized that VGI have often place names as spatial reference (e.g., town, region, country, etc.), resulting in different levels of spatial accuracy when looked-up in a gazetteer. Oppositely, the temporal reference of VGI is usually accurate because of the creation of one ‘time stamp’ when VGI is encoded on a mobile device, and of a second time stamp when VGI posted online.

The calculation of spatial similarity between VGI features is, thus, an issue as the significance of the difference between $x_{i,k}$ and $x_{j,k}$ may be questionable when these variables are expressed at different accuracy levels (e.g., when features are referenced at country level while others are accurately located using GPS signal).

In consequence, current spatiotemporal clustering techniques cannot be successfully applied to data with heterogeneous spatial reference such as VGI. Our research question is to advance Geographic Information Science with spatiotemporal clustering methods that are suitable to extract event-related knowledge from such data.

In the remainder of this paper, possible methods are described, using a real-life dataset as an illustration. This work can serve as a theoretical base for future works aiming at performing web-based event detection.

3 A VGI dataset for illustration and testing

A VGI dataset related to forest fires harvested on the web through the Flickr[1] photo sharing website exemplifies our research question. 12883 pictures have been retrieved, which were taken between June 01 2009 and October 01 2009, and their title, description or tags contained the words ‘forest’ and ‘fire’ – or translations and their synonyms in French, Italian, Spanish, Portuguese, Greek and Catalan.

Among these pictures, a first group of 1752 pictures (13.6% of the total) had a geographic reference expressed in latitude and longitude coordinates, while a second group of 7193 pictures (55.8%) where provided with one or several place names that could be looked-up in a Gazetteer[2]. An important proportion of these pictures had at least one town name as spatial reference (4550, 63.3 % of group 2) or a state (or county or equivalent, 2299 pictures 32%) or a country (928, 13%). Land features (e.g. lakes, mounts, national parks) were also frequently cited (for 840 pictures, 11.7%). For the third group, including the 3938 remaining pictures (30.6% of the total) no geographic reference at all was provided.

This emphasizes two fundamental points for our research. Firstly, users often use place names to locate a piece of knowledge they share with others. This creates spatial heterogeneity in the VGI dataset, as the resulting geographic objects are not comparable to the each other in terms of accuracy. Secondly, the VGI items from the second group should not be ignored only because of their poorer spatial accuracy, as their potential to provide knowledge about the events of interest seems to be rich. The challenge is thus to maximize the amount of event-related knowledge extracted from VGI, while locating such events with the highest possible accuracy.

[1] http://www.flickr.com

[2] for instance, http://developer.yahoo.com/geo/placemaker/

4 Possible methods

At this stage of our research, we identified three groups of possible methods to address our research question. They are further described in following sub-sections.

4.1 Adapting spatiotemporal clustering for datasets with heterogeneous spatial reference

A first possible group of methods would be to adapt well established spatiotemporal clustering algorithms in order to make them work efficiently on datasets with heterogeneous spatial reference.

For example, a multi-scale version of the Scan statistic can be further investigated. Scan statistic method is a powerful method to detect space-time clusters (Sikder & Woodside 2007). It consists in moving an n-dimensional window on the dataset, and to measure the number of features it contains at each position. In our case, such method could be run sequentially at various scales (i.e. with decreasing window size) ignoring at each iteration the features that are not accurately enough geo-referenced compared to the window size. The result would be that low-scale clusters including the whole dataset would be available to support the interpretation of high-scale clusters including only the most accurately geo-referenced features.

4.2 Rasterization and hot spot analysis

Another strategy could be to avoid having to calculate Euclidian distance between features by rasterizing spatial objects associated with VGI items, and to measure spatial correlation by map algebra operations.

This process is divided in four steps. Firstly, a temporal segmentation of the dataset is performed following arbitrary divisions (e.g. weeks, month) or using statistical methods such as Natural Breaks (Jenks & Coulson 1963). In a second step, the geographic object associated to each VGI feature is converted to a set of pixels with a numeric value equal to 1/n, where n is the number of pixels that overlap with the VGI features. This means that a VGI feature that has inaccurate spatial reference (e.g., ‘France’) will provide numerous pixels with low value, while a VGI feature accurately geo-referenced (e.g., GPS coordinates) will result in a limited number of pixels with high value. In a third step, pixel values derived from VGI features that are in the same ‘time slice’ are summed using a simple map algebra operation. Finally, a Hot Spot analysis (i.e. mapping of areas with extreme value; Haining 2003) is run on the final raster dataset to locate in space and time events described by heterogeneous VGI.

4.3 Knowledge discovery with ontology-based approach of the spatial dimension

A third promising approach aims to leverage spatiotemporal semantics contained in VGI dataset (i.e.: Oxfordshire is part of the United Kingdom). It seems particularly adapted to VGI, when spatial reference is primarily expressed as a place name. Sizov (2010) recently highlighted the latent spatial semantics of VGI and its potential for geospatial knowledge discovery.

We suggest to measure the spatial similarity between two VGI items (i.e. the spatial component of their similarity) using reasoners based on spatial ontologies provided by state-of-the-art gazetteers[3] instead of using spatial analysis techniques based on physical distance measurement.

On this basis, for example, an adapted Knox test can be designed. The Knox Test observes the spatial (‘close in space’) and temporal similarity (‘close in time’) for each pair of features to evaluate the spatiotemporal autocorrelation of a dataset (Rogerson 2001). An adapted Knox test should use relations between spatial concepts instead of distance between spatial objects to measure the spatial similarity.

5 Conclusions and future works

This extended abstract described and illustrated advances of Geographic Information Science that are required to improve web-mining techniques of event-related knowledge, and outlined possible approaches towards this achievement. Next steps will include testing candidate methods, by applying them to real-life VGI datasets and by comparing the outcome with reference spatiotemporal data from independent sources.

References

De Longueville, B., Luraschi, G., Smits, P., Peedell, S. & De Groeve, T., 2010. Citizens as sensors for natural hazards: a VGI integration workflow. *Geomatica*, 64(1) (in press).

De Longueville, B., Smith, R.S. & Luraschi, G., 2009. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatiotemporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*. Seattle, Washington: ACM, pp. 73-80. Available at: http://portal.acm.org/citation.cfm?id=1629907 [Accessed January 22, 2010].

Etzioni, O., 1996. The World-Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, 39(11), 65-68.

Fischer, M.M., Scholten, H.J. & Unwin, D.J., 1996. *Spatial analytical perspectives on GIS*, CRC.

Flanagin, A.J. & Metzger, M.J., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72, 137-148.

Getis, A. & Ord, J.K., 2010. The Analysis of Spatial Association by Use of Distance Statistics. In *Perspectives on Spatial Data Analysis*. pp. 127-145. Available at: http://dx.doi.org/10.1007/978-3-642-01976-0_10 [Accessed April 29, 2010].

Gong, B. et al., 2006. Event Discovery in Multimedia Reconnaissance Data Using Spatio-Temporal Clustering. In *Proc. of the AAAI Workshop on Event Extraction and Synthesis (EES'06)*.

Goodchild, M., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2, 24-32.

Gouveia, C. et al., 2004. Promoting the use of environmental data collected by concerned citizens through information and communication technologies. *Journal of Environmental Management*, 71(2), 135-154.

Haining, R.P., 2003. *Spatial data analysis: theory and practice*, Cambridge University Press.

Hsu, K. & Li, S., 2010. Clustering spatial-temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources*, 33(2), 190-200.

Jain, A., Murty, M. & Flynn, P., 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3), 316-323.

Jenks, G.F. & Coulson, M., 1963. Class intervals for statistical maps. *International Yearbook of Cartography4*, 3, 119-134.

[3] For example, the Yahoo Geoplanet gazetteer contains a rich set of relations between spatial concepts, such as: is ancestor (i.e. includes), is sibling, is neighbour, overlaps, belongs to, has common ancestor with, *etc*. See http://developer.yahoo.com/geo/geoplanet/guide/api-reference.htmlfor a detailed reference.

Johnson, S., 2010. A brief history of the analysis of crime concentration. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-77950652185&partnerID=40&md5=e42cdef8031130769e54718940bf4be9 [Accessed April 28, 2010].

Miller, H.J. & Han, J., 2001. Geographic data mining and knowledge discovery : an overview. In *Geographic data mining and knowledge discovery*. CRC Press, pp. 3-33. Available at: http://books.google.com/books?id=fJTD26iYxpkC.

Pultar, E., Raubal, M. & Goodchild, M.F., 2008. GEDMWA: Geospatial Exploratory Data Mining Web Agent. In W. Aref et al., eds. *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008)*. Irvine, CA, USA.

Rogerson, P.A., 2001. Monitoring Point Patterns for the Development of Space-Time Clusters. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 164(1), 87-96.

Shah, A.R. et al., 2010. Applications in Data-Intensive Computing. In Elsevier, pp. 1-70. Available at: http://www.sciencedirect.com/science/article/B7RNF-4YM596H-4/2/2d4b9333be3089c483107b7f407f0039 [Accessed April 28, 2010].

Sikder, I. & Woodside, J., 2007. Detection of Space-Time Cluster. In *Information and Communication Technology, 2007. ICICT '07. International Conference on*. pp. 139-143.

Sizov, S., 2010. GeoFolk: Latent spatial semantics in web 2.0 social media. In *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. pp. 281-290. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-77950885442&partnerID=40&md5=c2021257c7e70baf8cf55876b8a66717 [Accessed April 29, 2010].

Vaccari, A. et al., 2009. Towards the SocioScope: an information system for the study of social dynamics through digital traces. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle, Washington: ACM, pp. 52-61. Available at: http://portal.acm.org/citation.cfm?id=1653782 [Accessed April 28, 2010].