Building Multi-Resolution Event-Enriched Maps From Social Data

Faizan Ur Rehman[*]
Science and Technology Unit,
Umm Al-Qura University, KSA
fsrehman@uqu.edu.sa

Imad Afyouni
Technology Innovation Center
Wadi Makkah, KSA
iafyouni@gistic.org

Ahmed Lbath
LIG, University of Grenoble
Alpes, France
Ahmed.Lbath@imag.fr

Sohaib Ahmad Khan[†]
Science and Technology Unit
Umm Al-Qura University, KSA
skhan@gistic.org

Saleh Basalamah
College of Computer and
Information Systems, Umm
Al-Qura University, KSA
smbasalamah@uqu.edu.sa

Mohamed Mokbel
Department of Computer
Science and Engineering,
University of Minnesota, USA
mokbel@cs.umn.edu

ABSTRACT

This paper discusses the next generation of digital maps, by positing that maps in future will intelligently self-update themselves based on distinctive events extracted dynamically from social media streams or other crowd-sourced data. To realize this concept, the challenges include developing a scalable and efficient system to deal with a variety of unstructured data streams, applying NLP and clustering techniques to extract relevant information from these streams, and inferring the spatio-temporal scope of detected events. This paper demonstrates *Hadath*, a system that extracts live events from social data by encapsulating incoming unstructured data into generic *data packets*. The system implements a hierarchical in-memory indexing scheme to support efficient access to data packets, as well as for memory flushing purposes. Data packets are then processed to extract *Events of Interest* (EoI), based on a multi-dimensional clustering technique. Next, we establish the spatial scope and the level of abstraction of each event. This allows us to show live events in correspondence to the scale of the view – when viewing at a city scale, we see events of higher significance, while zooming in to a neighborhood highlights events of a more local interest. The final output creates a unique and dynamic map browsing experience.

CCS Concepts

•Information systems → Geographic information systems; Wrappers (data mining); Data streaming;

Keywords

Event-Enriched Maps; Crowdsourced Data; Spatio-Temporal Scope

1. INTRODUCTION

While it is the norm nowadays to use digital mapping applications to use live data to find directions, traffic congestion states or places of interest, we posit that the next generation of maps will contain the additional functionality of showing live events at different spatio-temporal resolutions, and which are extracted dynamically from a variety of sources, starting from online social media and crowd-sourced data, to open governmental data and other online news sources. Within this context, there is a real opportunity to enrich current maps with knowledge extraction tools that take advantage of information retrieval, data management, and sentiment analysis techniques. Analyzing crowdsourced data can provide deep insights about surrounding events of interest (EoI). For instance, with the explosive growth in size of microblog data (e.g., Twitter, Flickr, and Yelp), fruitful insights can be extracted and displayed (examples of discovered findings are illustrated in Figure 1). However, designing an efficient and scalable system that extracts live events and infers their spatial and temporal scopes, so that they can be displayed in a clear, non-cluttered manner, remains a challenging task.

The challenge of displaying live events on a map is threefold. Firstly, these events need to be extracted from unstructured data streams efficiently, while preserving accuracy and conciseness. Secondly, to display such events on a map, their spatial scope must be established, so that as a user changes the zoom level, only events of appropriate scope are displayed. For example, a soccer match may be displayed at the city scale, the opening of a new restaurant at sub-urban scale, and a house-warming party at the neighborhood scale. Thus, not only is it necessary to extract the events themselves, but also to establish their spatial scope, so that they can be displayed appropriately in a clutter-free manner. Finally, all of this has to be done in real-time so that live streams can be handled, and up-to-date events at multiple resolutions can be detected. To address these challenges, we propose *Hadath*, a system that han-

[*]Also affiliated with LIG, University of Grenoble Alpes, France
[†]Also affiliated with Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan

Series ISSN: 2367-2005

10.5441/002/edbt.2017.78

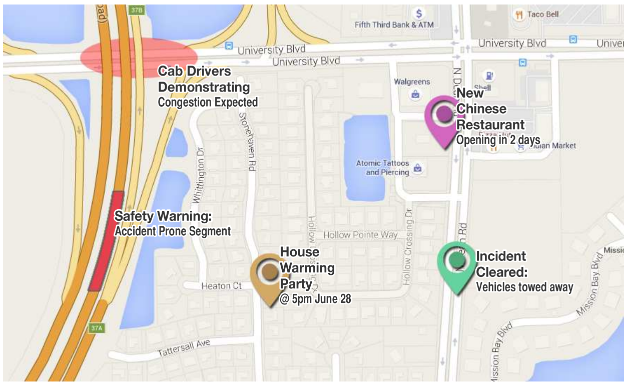


Figure 1: Conceptual illustration of Event-Enriched Maps: Findings automatically discovered from live streams, such as a restaurant opening, a neighborhood party, an accident prone road segment, warnings on demonstrations and emergency cases.

dles unstructured social streams, particularly Twitter data, and implements different algorithms for the efficient extraction, clustering, and mapping of live crowdsourced events. *Hadath* consists of several components as follows. Data collection involves gathering social data with different forms: data chunks and streams. The data wrapping and cleaning component digests streaming data, and prepossesses data to generate structured data packets from unstructured streams. The data manager stores data by implementing a indexing scheme to allow efficient and scalable access to raw data, as well as to extracted events. The events of interest detection module classifies and extracts events based on a multidimensional and hierarchical clustering technique, which defines the spatial scope and the level of abstraction of detected events. The query engine creates best query plan based on map zoom level, spatial and temporal characteristics, and executes the query plan in order to retrieve EOIs efficiently. The visualizer provides a new dimension to existing maps by illustrating extracted knowledge from live collected data as live events at different levels of abstraction. The remainder of this paper is as follows. Section 2 highlights related work and challenges from different perspectives. Section 3 introduces our proposed architecture with results; while Section 4 draws conclusions and discusses future work.

2. RELATED WORK

This section highlights different challenges and state-of-the-art techniques related to: 1) digital mapping, 2) events of interest detection, and 3) performance and scalability perspectives.

1. State-of-the-art mapping technologies: Today's maps are often crowd-sourced, and make use of *'Volunteered Geographic Information (VGI)'*, where users can seed maps with their own content. Researchers, authorities, and industries generate thousands of map-based analytics every year to meet their social and economic needs [6]. In addition, 'Live Maps' now contain real-time updates of bus schedules, traffic conditions, restaurant opening hours, and road accidents, among others. With the wide spread of social networks, people start to post their own social contributions on live maps, such as Foursquare check-ins, Flickr images, tweets [7], and Yelp reviews. Moreover, NLP techniques were embedded to extract spatially-referenced news from online newspapers and tweets [10]. However, current maps still lack intelligence in extracting knowledge about new events occurring at different spatio-temporal resolutions.

2. Discovering Events of Interest along with Spatio-Temporal Scope: Detection of irregular happenings and trends from social data, mainly Twitter data, is already a topic of many scientific articles [2]. This mainly includes: 1) earthquake detection along with their centers or other natural disasters; 2) extracting and localizing breaking news from tweets as presented in TwitterStand [10]; 3) discovering incidences related to traffic conditions [8] from twitter and user generated data; and 4) detection of unspecified hot topics based on text similarity [1], density or with wavelet spatial analysis [5]. The work presented in [4] is very close to our work with respect to detecting spatial/temporal extents of events, but was only distinguishing between local and global scales, without putting focus on mapping those events to the different spatio-temporal resolutions in digital maps.

3. Performance and Scalability Perspectives: Several works have presented systems that visualize geo-tagged social streams on maps, such as Flickr images[1], tweets [7], Yelp reviews, and spatially-referenced news [10]. Particularly, NewsStand [10] is a scalable system that extracts news from RSS feeds and visualize them on a world wide map. Furthermore, the system can apply spatio-temporal and keyword-based filtering of news. However, this system displays news at different spatial scales by only ranking them based on the number of views, without detecting and clustering events of interest along with their spatial scope. With the large volume of incoming streams, data indexing and the distributed processing of data represent an essential part of any system that implements *'event-enriched maps'*.

3. SYSTEM OVERVIEW

This section presents *Hadath*, a system that retrieves data streams from social data (here we focus on Twitter data), efficiently manages and processes those streams in order to find Events of Interest (EoI), and visualizes those events in correspondence to their spatial and temporal scopes, thus creating *'multi-resolution event-enriched maps'*. Figure 2 illustrates the main components of our system architecture, which are described as follows:

• *Data collection* involves gathering data with different format. This includes digesting data streams and data chunks (i.e., historical tweets) from Twitter. In data chunks mode, we download the files that contain partial or full datasets. Digesting data streams is performed by running crawlers that collects bulks of streams based on windows of a specified temporal extent w (e.g., 30 minutes window).

• The *data wrapper* provides an efficient and generic mechanism with the aim of allowing new data sources (e.g., Flickr) to be easily plugged, by supporting new crawlers at the data collection level without affecting the other processing components. Major tasks for the data wrapper are: 1) to clean irrelevant fields and digest incoming streams into a unique data packet format; 2) to use specified string matching technique that detect and match candidate packets with our event classifier corpus in order to identify potential event

[1] https://www.flickr.com/map

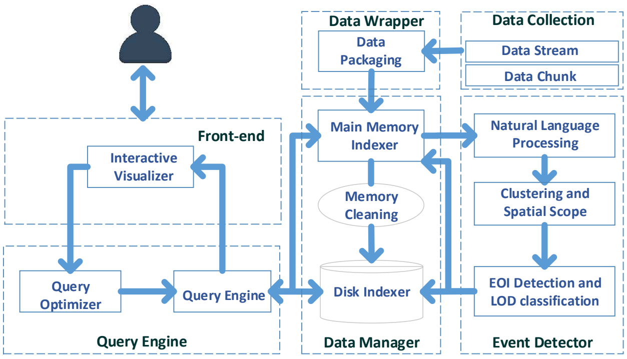


Figure 2: Hadath Architecture

classes and properties; and 3) to apply unspecified topic detection method that extract spatio-temporal peaks and unusual happenings based on the top frequent words. Figure 3 shows an example of data packets generated from twitter data with potential event flag, event class name, and event properties. Packets that show no relevancy with respect to the above steps are discarded at this phase.

- The *data manager* implements an in-memory spatial indexing scheme to allow efficient and scalable access to data packets. The spatial index is a multi-resolution data structure (similar to a partial quad tree [9]). Leaves in this data structure correspond to cells that represent the minimum bounding rectangles comprising data packets. Figure 4 displays a snapshot of indexed data packets at a fine level of the hierarchal tree, and with a single day specified as a time threshold. Cells are colored lighter to darker based on data packet counts; darker-colored cells are further expanded at deeper levels in the tree as compared to lighter-colored cells. *Hadath* employs a big data mechanism that continuously process data packets within the different cells on several execution nodes. The manager also indexes detected EoIs in order to fetch them efficiently based on the map zoom level and scope. Using this multi-resolution indexing scheme, hierarchical clustering of events can be applied for efficient determination of their content and spatial scopes. For temporal aspects and cleaning of EoIs, we took three parameters: a) *'birth time'* that indicates the existence of a new event in our system whenever we calculate the first cluster of data packets related to that event; b) *'time of occurrence'* that marks the actual happening time of the event (e.g., next Monday); and c) *'time to live'* (TTL) is the survival time of an event in our system. Whenever we receive new data packets related to an existing event, we increase its TTL by T number of hours. Moreover, processed data packets are moved to disk based on temporal and memory thresholds. The main task of disk indexer is to index outdated data packets and events on disk using an R*-tree spatial index to allow efficient retrieval for historical queries.

- The *event detector* module starts from leaf cells within the multi-resolution data structure to detect events at a local spatio-temporal scope. Within each leaf cell, our system adopts the graph analogy where each potential event data packet is considered as a *node* and the value of 'text similarity (TF-IDF)' between data packets as a weight of the bidirectional *edge*. Data packets with a high text similarity value are clustered using the graph-specific Louvain clustering algorithm [3]. The Louvain algorithm is suitable in our approach as, unlike most of the other clustering methods, it does not require a prior knowledge of the minimum number of clusters. For unspecified events that are not matching our training corpus, the NLP module detects frequent tags and keywords within local cells, in order to identify peaks at local and global scales.

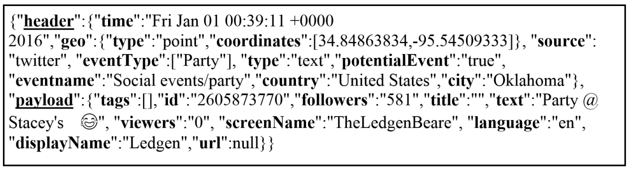

{"**header**":{"**time**":"Fri Jan 01 00:39:11 +0000
2016","**geo**":{"**type**":"point","**coordinates**":[34.84863834,-95.54509333]}, "**source**":
"twitter", "**eventType**":["Party"], "**type**":"text","**potentialEvent**":"true",
"**eventname**":"Social events/party","**country**":"United States","**city**":"Oklahoma"},
"**payload**":{"**tags**":[],"**id**":"2605873770","**followers**":"581","**title**":"","**text**":"Party @
Stacey's 😂", "**viewers**":"0", "**screenName**":"TheLedgenBeare", "**language**":"en",
"**displayName**":"Ledgen","**url**":null}}

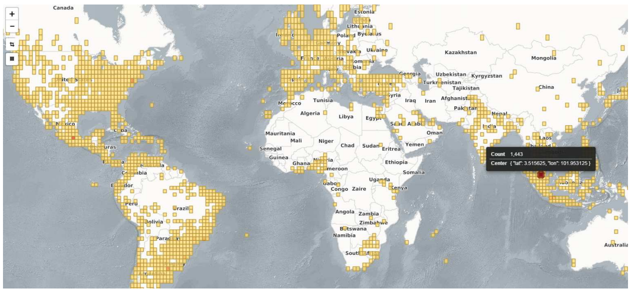
Figure 3: Sample Data Packet from Twitter Streams

As events can be discovered more efficiently on small-scale regions (starting from leaf nodes), a bottom-up approach for clustering close-by and similar events is developed, so that redundant events on different spatial resolutions can be aggregated, and their spatial scope can be upgraded. Visualization of EoIs with the same spatial resolution on maps does not make sense, since these events have different significance from spatio-temporal perspectives. For instance, events of someone's birthday cannot be displayed at a national level, except is this person is a celebrity, and that happening had spread throughout the country. Our hierarchical clustering technique for event aggregation works as follows. Starting from events at neighborhood/district level (i.e., which corresponds with leaf cells in our tree), the system clusters identical events at higher levels of abstraction, and incrementally increases their spatial scope. Local clusters are first compared with their siblings in the hierarchical tree, with the aim of aggregating and updating the scope of similar events. Merging two clusters (Cl_1, Cl_2) from two different cells ($C_{1,n}$, $C_{2,n}$) at a depth level n in the tree, will result in upgrading their spatial scope from zoom level k (e.g., corresponds to district level on map) to zoom level $k-1$ (e.g., corresponds to city level). An event cluster Cl_i is represented as follows:

$$Cl_i = \langle id, ptGeom, eventClass, eventProperties, packetIDs, imageURLs, iconId, zoomLevelStart, zoomLevelEnd\rangle$$

where 'id' is cluster identifier, 'pointGeom' is the centroid point location, 'eventClass' and 'eventProperties' depict the event class(es) and a list of top frequent meaningful words within the cluster, 'packetIDs' is the list of data packets identifiers forming that cluster, 'imageURLs' is the list top selected image URLs, 'iconId' is the icon identifier related to the event class, and 'zoomLevelStart, zoomLevelEnd' correspond to the multiple resolutions where this event is available to be displayed on map.

- *Hadath's query engine* supports efficient retrieval of in-memory and disk indexed events based on the main querying attributes, that are, the spatial, temporal dimensions, and the map levels of detail. The *visualizer* provides a new dimension to existing maps by illustrating extracted knowledge from live streams in the form of live events at different levels of abstraction. Figure-5 illustrates an example output of *Hadath* system by showing EoIs at different zoom levels including a) 'Grand Opening' at city-scale; b) 'Traffic Incident' at a locality-level and c) 'Birthday Party' at a neighborhood-level. The final output creates a unique and



Count 1,443
Center { "lat": 3.515625, "lon": 101.953125 }

Figure 4: A snapshot of indexed data packets at a fine level of the hierarchical tree

dynamic map browsing experience.

4. DEMONSTRATION SCENARIO

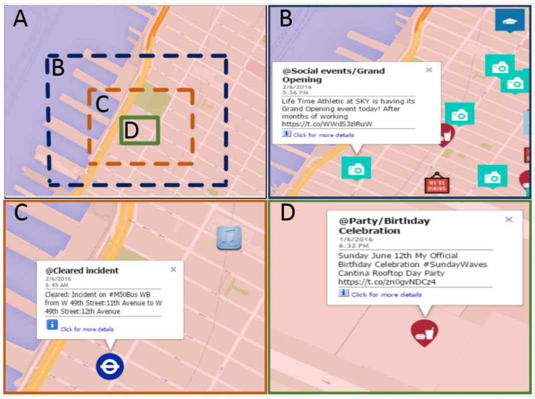
Attendees will be able to interactively use our *Hadath* system, and enjoy discovering events of different levels of abstraction on a world wide map with a smooth and fast panning and zooming capabilities. Either (near) real-time or historical events can be browsed on map with a calendar option specifying a certain time threshold. This demo is intended to show the usage and efficiency of our prototype. For this purpose, several visualizations are made possible including: i) interactive tag/word clouds of events that are dynamically adapted when changing the specified spatio-temporal scope (i.e., by zooming, panning or applying a rectangular range selection); ii) statistical plots and histograms that illustrates the number of raw data packets as well as clusters of events at different zoom levels; and finally 3) the multi-resolution event-enriched map visualization, where events of higher significance are displayed at higher abstraction levels.

5. CONCLUSION

This paper introduces a system, called *Hadath*, that builds multi-resolution event-enriched maps by handling social data streams, and by developing different algorithms for the efficient extraction, clustering, and mapping of live events. Hadath wraps incoming unstructured data streams into data packets, that is, a generic structured format of a potential event. These packets are then processed to extract EoIs based on a hierarchical clustering technique, which defines the spatio-temporal scope for each event. The system can provide valuable knowledge from crowd-sourced data to authorities, market firms, event organizers, and end-users to help in decision making. In future, we plan to merge more data sources (e.g., Flickr, online newspapers) to increase correctness and conciseness of detected events. Furthermore, an extensive performance evaluation of the different solutions need to be conducted with respect to closely-related systems.

6. REFERENCES

[1] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum. See what's enblogue: real-time emergent topic identification in social media. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 336–347. ACM, 2012.



@Social events/Grand Opening
2/6/2016
5:56 PM
Life Time Athletic at SKY is having its Grand Opening event today! After months of working https://t.co/WWdS3zlRuW
Click for more details

@Cleared incident
2/6/2016
6:45 AM
Cleared: Incident on #M50Bus WB from W 49th Street:11th Avenue to W 49th Street:12th Avenue
Click for more details

@Party/Birthday Celebration
1/6/2016
6:32 PM
Sunday June 12th My Official Birthday Celebration #SundayWaves Cantina Rooftop Day Party https://t.co/zn0gvNDCz4
Click for more details

Figure 5: An example output of *Hadath* system. A) Overview map of the area. B) EoIs at a city-scale that are of general interest to the residents. C) EoIs at a locality-level and D) EoIs at a neighborhood-level that are progressively more specific in their spatial scope.

[2] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Comput. Intell.*, 31(1):132–164, Feb. 2015.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[4] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard. Multiscale event detection in social media. *Data Min. Knowl. Discov.*, 29(5):1374–1405, Sept. 2015.

[5] S. B. Kaleel and A. Abhari. Cluster-discovery of twitter messages for event detection and trending. *Journal of Computational Science*, 6:47–57, 2015.

[6] J. Krygier and D. Wood. *Making maps: a visual guide to map design for GIS.* Guilford Press, 2011.

[7] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T. M. Ghanem, S. Ghani, and M. F. Mokbel. Taghreed: a system for querying, analyzing, and visualizing geotagged microblogs. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 163–172. ACM, 2014.

[8] F. U. Rehman, A. Lbath, M. A. Rahman, S. Basalamah, I. Afyouni, A. Ahmad, and S. O. Hussain. Toward dynamic path recommender system based on social network data. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, IWCTS '14, pages 64–69, New York, NY, USA, 2014. ACM.

[9] H. Samet. The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)*, 16(2):187–260, 1984.

[10] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *Commun. ACM*, 57(10):64–77, Sept. 2014.