Towards Context-Aware Search and Analysis on Social Media Data

Leon R.A. Derczynski
University of Sheffield, UK
leon@dcs.shef.ac.uk

Bin Yang
Aarhus University, Denmark
byang@cs.au.dk

Christian S. Jensen
Aarhus University, Denmark
csj@cs.au.dk

ABSTRACT

Social media has changed the way we communicate. Social media data capture our social interactions and utterances in machine readable format. Searching and analysing massive and frequently updated social media data brings significant and diverse rewards across many different application domains, from politics and business to social science and epidemiology.

A notable proportion of social media data comes with explicit or implicit spatial annotations, and almost all social media data has temporal metadata. We view social media data as a constant stream of data points, each containing text with spatial and temporal contexts. We identify challenges relevant to each context, which we intend to subject to context aware querying and analysis, specifically including longitudinal analyses on social media archives, spatial keyword search, local intent search, and spatio-temporal intent search. Finally, for each context, emerging applications and further avenues for investigation are discussed.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

1 Introduction

Social media data typically consists of non-curated, short messages that are shared among people, instead of being visited manually or crawled by an automatic agent. Messages may be distributed through an explicit network of friends and followers, openly visible or privately, according to the sender's preferences. There is no publishing delay and the barrier of entry is low, often only requiring an email address. This leads to substantial volumes of content constantly being created, and an expectation of data currency.

Online social networks are a source of "big data". Our social graphs are made explicit; our interactions are recorded; our utterances are saved in machine readable format; we can be heard across the world as easily as across the room. Twitter alone generates a million messages every five minutes; a four-day stream comprises around 10^9 messages. As a result of online social networking, massive volumes of diverse social media data that capture a sample of all human discourse are accessible online.

EDBT/ICDT '13 March 18 - 22 2013, Genoa, Italy

Further, a notable proportion of text and photos posted online are explicitly geotagged [13], and studies suggest it is possible to infer the locations of about half the remaining such content [9]. Every message has at least a creation time as temporal context. Thus, messages contain explicit and implicit *spatio-temporal* (ST) metadata. The stream of messages can be viewed as people creating an ever-growing set of ST points, each with a corresponding piece of text – similar to conventional spoken communication.

Analyses that consider these ST aspects indicate messages previously perceived as low impact may sometimes in fact have significant meaning. For example, analysing tweets for mentions of personal health (e.g., "`ugh, not feeling good today #sick`") has lead to accurate micro-level geo-social models for predicting whether and when someone will become ill. Such models have used social media message content and social graphs together with detected co-location with other (potentially) ill people to predicate an individual's future health [13].

There are notable big data challenges involved in querying and analysing social media data, with correspondingly significant and diverse rewards. The challenges can be broken into three contexts: temporal, spatial, and spatio-temporal. We focus on data-centric challenges in each context and cover example applications.

2 Temporal Challenges

Social media messages generally come with a creation time-stamp recorded as metadata. This provides an accurate temporal annotation of every message. Querying and analysis on social media data is called for in two temporal scenarios: one may be concerned with online, real-time reactions to social messages as they arrive in the stream but also with historical analyses, e.g., identifying sentiment changes about a past event such as a national election. Challenges related to the temporal context of social media data can be split into those concerned with *emerging aspects* of data and those concerned with *historical aspects* of data, as shown in Figure 1.

As high arrival rates are a key characteristic of social media streams, researchers have begun to answer questions on the emerging aspects, such as emerging-topic (news) detection from social media and prediction of future trends. We will shortly discuss challenges in time-sensitive querying over social media data.

In addition, challenges exist with historical aspects of social media archives, which are beyond the challenges of longitudinal analyses over document archives [16] and that are not yet well understood. Thus, we explore aspects of longitudinal analyses on social media archives; such analyses can offer, for example, insight into the evolution of different social actions.

2.1 Longitudinal Analyses on Social Media Archives

As suggested by Figure 1, storing social media streams produces massive social media archives that may record changes in entities

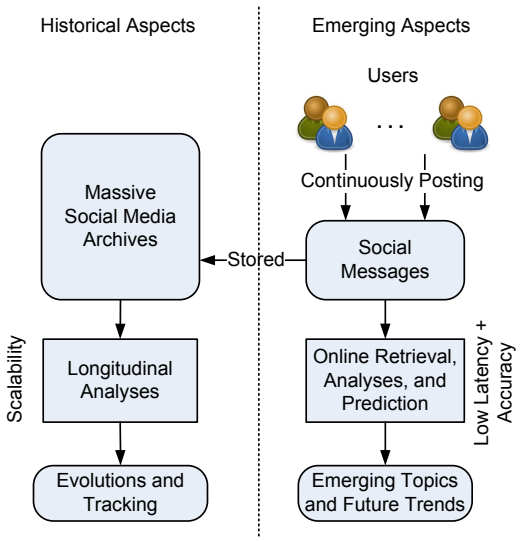

Figure 1: Emerging and historical aspects of social media data.

and relations between them. Longitudinal analyses on people's social graphs can reveal the evolution of social relationships, which helps better understand social behaviours and predict social relationships. Focused on specific events, longitudinal analyses on social media data reveal discourses around events, e.g., opinion swings during election campaigns or riots. Not all messages are of interest at the time they arrive. However, after a high-impact event such as a revolution or market crash, demand emerges for retrospective analyses of sentiment-significant social data [11] leading up to the event.

Further, correlating the results of longitudinal analyses on social media with other data may yield new insight. Micro-blog reports can act as a sampling of real-world activity. Time series analyses on tweets provide advance information on market behaviour [20]. As a concrete example, it happens that the West Nile Virus affects certain types of birds. Thus increased deaths in these birds often precede outbreaks among humans. In this case, analysing historical reports of dead crows from social media can yield a model for predicting virus outbreaks [15] (see Figure 2).

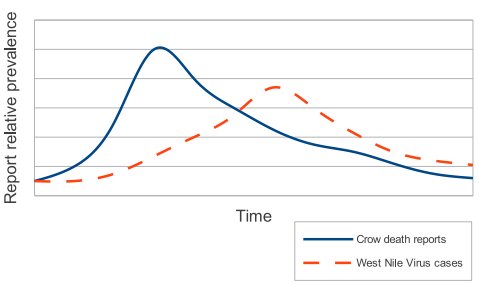

Figure 2: Crow death reports in social media precede incidences of West Nile Virus in humans.

Businesses can also benefit from longitudinal analyses. For example, a shop may retrieve and analyse historical reviews from social media in order to judge responses to advertising campaigns or to improve services.

Such analyses typically operate on large social media archives. The massive data volumes call for scalable analysis techniques. First, scalable temporal indexing is required. Tree-based indexing methods may not be the first choice, as tree structures are non-trivial to distribute among computing nodes [3]. Adaptive and multi-resolution grids may be more suitable, as they can be more easily distributed among nodes. Next, secondary indexing is typically useful, as after obtaining data covering a pertinent temporal extent, other data attributes need to be processed in an efficient way. For example, social graphs in a particular temporal range need to be formulated based on, e.g., reply-relationships of tweets using secondary indexing. The benefits of coupling or decoupling temporal indexing and secondary indexing remain to be understood.

2.2 Time-Sensitive Indexing

Longitudinal analyses rely on static information (e.g., indexing and searching a static social media archive) as opposed to dealing with streaming data. The information required to rank documents within given temporal bounds depends on the size and "position" of the temporal window between the bounds. This means that for text-based retrieval, indices must support time-sensitive TF and IDF measurement over temporal cross-sections of varying size [16]. On top of this, systems also need to support time-sensitive querying, and real-time analysis demands they support real-time message indexing. To manage this, indices must support high-frequency document creation, instead of offline indexing of snapshots. The demands for high-velocity document indexing are considerable: for example, some 2012 olympic events were commented on at a rate of 38 000 tweets per minute.[1] This gives the temporal dimension of social media message collections a high density. Unlike web page archives, social media messages are not intrinsically grouped into per-day clusters, but have per-second timestamps, and do not come in small discrete bursts followed by gaps, but rather are generated constantly and at a high rate. How to effectively manage and retrieve data with different temporal context criteria and such high density is a considerable challenge.

As social media has advantages in reporting current and recent data, indexing must be capable of supporting high-quality results from newly received data; there is little opportunity to sacrifice accuracy for efficiency, especially in systems that support the detection of emerging events. Ensuring result accuracy under such adverse conditions is a tough challenge.

3 Spatial Challenges

We are witnessing a proliferation of mobile web users and geo-positioning capabilities, which significantly changes the nature of web search. It is reported that 20% of all Google searches and 53% of mobile searches on Bing have local intent.[2] Although temporal metadata is explicit (timestamps), spatial metadata can be more difficult to acquire [9]. The explicit and implicit spatial contexts in social media messages provide a unique setting for spatial search; importantly, they also bring opportunities to improve the quality of search results. Here, we discuss how social media can influence and benefit two different types of spatial search, *spatial keyword search* and *local intent search*, and how systems can be developed to support these. An outline is given in Figure 3.

This section discusses the application of existing spatial search techniques within a novel environment. We touch upon topics covered in earlier work by the authors and others. Reference [2] offers an overview of the field and contains references to additional work.

3.1 Spatial Keyword Search

Spatial web objects are prevalent on the web. For example, the web page of a restaurant is a spatial web object where a street address

[1] http://blog.twitter.com/2012/08/onlyontwitter-road-to-gold.html
[2] http://searchengineland.com/microsoft-53-percent-of-mobile-searches-have-local-intent-55556

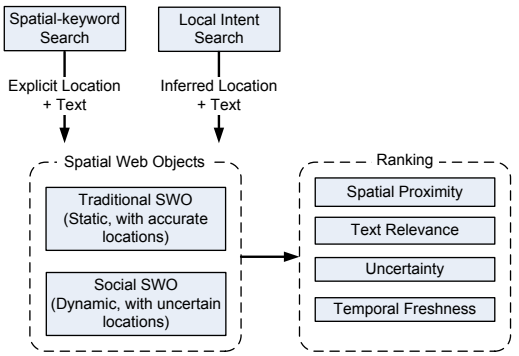

Figure 3: Processing spatial aspects of social media data.

encodes the object's location and the menus provided by the restaurant are the object's text. Using a ranking function, result objects are ordered according to spatial proximity to a query location and textual similarity to query keywords. This way, objects matching both spatial and textual arguments in a query can be identified.

More specifically, a spatial keyword query $Q_{SK} = (\lambda, \psi)$ takes an explicit, typically precise, location λ along with keywords ψ, and returns a ranked list of spatial web objects with high spatial proximity and textual relevance to the query. The content that is queried is a set of *spatial web objects* $\mathcal{O}$, where each object $o \in \mathcal{O}$ has a pair of attributes $o = (\lambda, \psi)$, where λ encodes an accurate location and ψ is a text value [2].

Social media messages are spatial web objects. Considerable amounts of social media content has a qualitative (e.g. textual) or quantitative (e.g. GPS) reporting location. It is reported that some 25% of tweets from mobile devices are accompanied by GPS coordinates,[3] and typically, the most active authors are also those who report their locations [13]. The text contained in messages may also contain spatial references (e.g., "`Amazing rainbow at the ARoS museum.`"). Further, alternative information sources obtained from social media can approximate the reporting location, e.g., author profiles [7] and social graph-based inference from friends [9]. Thus, social media provides large corpora for spatial keyword search, with a new form of spatial web objects. Meanwhile, social media also brings several challenges to existing techniques for spatial keyword searches, detailed below.

For example, imagine someone is planning a night out. They search for "*good bar in north copenhagen*", and their phone provides GPS coordinates. The search's origin is shown as a black dot in Figure 4. Social messages with GPS positions are shown in Table 1, from which a ranking can be produced. In the table, spatial similarity is given by function $loca$ and textual similarity by $text$.

o	**Microblog text**	$loca$	$text$
A	So drunk last night at @BarSyv	0.7	0.6
B	Out shoe shopping!!! #louboutintime	0.9	0.0
C	Who pays $9 for a beer?!	0.6	0.5
D	wow found cph's greatest cocktail bar lol	0.1	1.0
E	Traffic. Traffic everywhere. Need a drink.	0.4	0.2

Table 1: Sample results for five ST-located messages given the query "*good bar in north copenhagen*".

3.2 Rankings

Ranking functions in spatial keyword search combine spatial proximity and textual relevancy between queries and spatial web objects [2]. For example, a ranking function may integrate location proximity and textual similarity in a weighted sum manner as shown in Equation 1.

$$rank = \alpha \cdot loca + (1 - \alpha) \cdot text, \quad (1)$$

where a parameter α balances the relative importance of spatial proximity and text relevance.

As diverse locations exist in social media, rankings, especially the spatial proximity part, on social media is different from existing ones. Uncertain locations render rankings according to spatial proximity uncertain. Thus, probabilistic top-k and skyline functionality may be called for.

[3]From observations of a 24-hour sample of the Twitter "garden hose" feed; see also an animated visualisation of this sample.

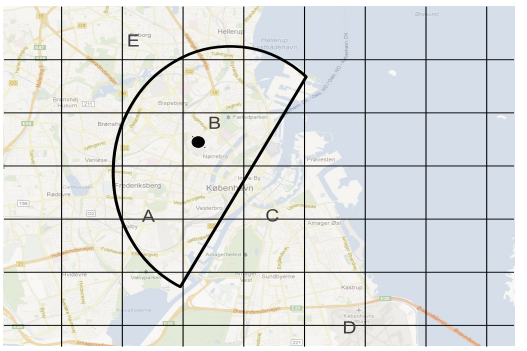

Figure 4: Sample spatial search results for "*good bar in north copenhagen*". The dot is the query location for spatial keyword search, and the semi-ellipse region is the inferred relevant region for local intent search. A-E are messages (with GPS coordinates) considered for ranking.

3.3 Huge Size and High Update Rates

Existing spatial keyword queries normally work on a static or infrequently updated set of spatial web objects, e.g., a collection of crawled web pages containing various *points-of-interest* (POIs – e.g., hotels, restaurants) [2]. In contrast, social media is streaming data, and new messages appear rapidly, which results in massive volumes of dynamic spatial web objects.

The streaming characteristic of social media introduces two novel types of continuous spatial keyword queries that we propose here: *Static Continuous Spatial-Keyword Queries (SCSKQ)* and *Moving Continuous Spatial-Keyword Queries (MCSKQ)*. Such queries are registered once and then logically run continuously over the set of spatial web objects, in contrast to traditional one-time queries.

In an SCSKQ, the location argument is fixed, e.g., a query issued by a user who is sitting in an office. The continuous aspect of SCSKQ is from the streaming spatial web objects (e.g., the social messages). As new messages arrive, Q_{SCSKQ} maintains an up-to-date result containing spatial web objects that are relevant w.r.t. the location and text arguments.

In contrast, the location argument $Q_{MCSKQ}.\lambda$ in an MCSKQ is moving, e.g., a query issued by a user who is traveling on a bus. Thus, the continuous aspect of MCSKQ is due to both the streaming spatial web objects and the moving query locations. A Q_{MCSKQ} keeps updating its result as the query location changes and as new messages arrive. Both SCSKQ and MCSKQ differ from the continuous queries discussed in the literature [18], where the continuous aspect is only due to the moving query locations, while the spatial web objects remain static.

The huge data volumes and the high update frequencies of social messages require that the technology supporting continuous spatial keyword queries must be scalable and capable of handling rapid changes. This motivates a rethinking of current techniques, such

as IR-trees [17]. Multi-level indexing should be considered, where main-memory indexing, which is able to exploit the parallel computation capabilities provided by modern computing hardware [14], is used on content with high update rates, while disk based indexing is exploited for static data.

3.4 Location Diversity

Locations appearing in social media are diverse, ranging from quantitative GPS coordinates to qualitative text mentions (e.g., "`ARoS Museum`"), from points to regions (e.g., the campus of Aarhus University), and from precise locations to uncertain locations (e.g., those obtained by some inference). Moreover, a social media message may be associated with more than one location, each of which may be in various formats and which may even be conflicting.

Location mentions in short social messages are of varying reliability [7]. Aside from the linguistic challenges in location extraction and disambiguation in a novel domain, there is a requirement to manage the resultant location uncertainty. While some message locations are given with GPS coordinates, others may be presented with only region-level descriptions (e.g., "Swiss canton of Vaud").

Spatial search is non-trivial in the imprecise social media setting. One cannot always confidently disambiguate terse place names; a USA-based message claiming to be from the city of Reading can be from at least four locations, for example. This creates ambiguity. It may only be possible to say the location refers to Reading, OH with 90% certainty and Reading, PA with 10% certainty.

For these reasons, when retrieving messages based on a given query location $Q_{SK}.\lambda$ and the location $o.\lambda$ of the message considered for result inclusion, location uncertainty needs to be handled. Different confidence values may be assigned to the same objects by different linguistic processing methods, which can be used as weights in result ranking. Locations may be given as large regions with high confidence, as well as sub-regions with reduced confidence. Human assistance, possibly gathered through implicit feedback, may be used to reduce uncertainty in some cases.

Location diversity from social media poses challenges to spatial keyword search and introduces novel types of functionality. For example, uncertain locations and multiple locations yield uncertain spatial web objects and thus may call for probabilistic spatial keyword search.

For example, given the query text "*wine bars*" from a mobile device that has a GPS location in Columbus, OH, the following results could all be considered.

1. A check-in at "*Bob's Wine Bar*", from a user with country "*USA*" and location "*Reading*";
2. A picture that has the caption "*Great wines here*", with EXIF data containing geo-coordinates close to the query location;
3. A message with no explicit location information, but containing the phrase "*Ohio's best wine bars*";
4. A link to a page titled "*Wine bars of the midwest*" embedded in a message originating in a neighbouring state.

The ranking function needs to accommodate certainty and proximity, as well as point- and region-based locations.

Existing approaches for answering spatial keyword queries typically assume that each spatial web object has an accurate and single point location, making them unable to contend with queries on spatial web objects with diverse (especially uncertain) locations. A general indexing and query processing framework that can contend with diverse locations and multiple search types is desirable.

3.5 Inferring Query Regions

A spatial keyword query may lack an explicit location parameter. Such queries rely instead on a region of interest being inferred from the query text.

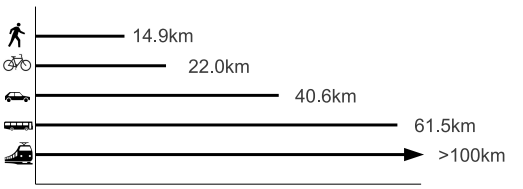

Figure 5: Maximum commute distance according to transport mode for a major Nordic city [6].

Re-using the previous example, consider a query for "*good bar in north copenhagen*" without GPS coordinates. To provide results for the query, a region of interest is first inferred from query text, in this case from "*north copenhagen*". Spatial usage of the preposition "*in*" distinguishes the location from the remaining text, so "good bar" is the non-spatial part of the query text. The region of interest is modelled by a semi-ellipse as shown in Figure 4. This region can then be used for selecting candidate content and ranking results [2].

Information about the movements of people and the locations they visit, obtained from analysing spatial contexts of social media data, informs solutions determining regions of interest. Specifically, travel patterns, willingness to travel, and people's movement radii are very useful for inferring regions of interest.

3.6 Travel Behaviour

Capturing travel behaviour permits generalisations about how people move and can help establish the region of relevance for a particular location. Spatial data in social media gives easy access to travel behaviour information, via explicit or inferred user locations. This can show what kinds of locations people visit and can help build models of personal travel behaviour. Such models inform three search modes. First, region-based retrieval is possible, to answer queries about behaviour in a specific region. Second, personalised query results can be improved, by e.g. emphasising results for locations near the user's daily commute. Finally, queries for POIs can be improved. For example, a query for "*cheap pizza*" may have a range of at most 2km; whereas "*recent festivals*" may have a wider range, perhaps within the same state or country.

3.7 Willingness to Travel

Knowing where and why people move and the regions that they tend to exist within can inform local intent search. Location information from social media messages, as well as message content, can reveal how likely an individual is to make certain journeys. Knowing an individual's maximum travel distance can bias POI inclusion. The willingness of a person to travel may be task-dependent; a user may not care about a coffee shop 5 kilometers away, even if rated higher than one near their office, but they might happily travel 20 kilometers to a furniture shop. Further, the maximum distance can depend on the user's chosen mode of transport, as indicated in Figure 5 (see also Section 4.3).

However, this kind of information has traditionally been difficult to extract, requiring extensive experiments and stratification of the people whose movements are monitored. Only sparse spatial information is required to model these preferences – perhaps just a "home" location in text format and a single report at a POI. Often, location information is explicitly present in messages, precisely revealing spatial behaviour. Collecting samples of messages for any given individual who reports their locations can provide a model for their willingness to travel.

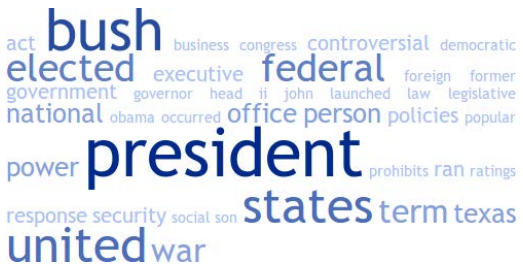

Figure 6: Terms used to describe the topic "*US president*" in 2007. Larger words represent greater term weight and higher expected frequency in documents on that topic.

Figure 7: Terms used to describe the topic "*US president*" in 2011.

4 Spatio-Temporal Challenges

In information retrieval, temporal querying generally aims to find either recent results or the provenances of entities [1]; in contrast, spatial search typically considers spatial proximity and local intent, but ignores the temporal aspect. Spatio-temporal search adds sophistication by giving biases according to both spatial and temporal proximity. We discuss how social media query benefits from spatio-temporal intent search, and we cover several spatial-temporal analytical applications, specifically topic-based retrieval, sentiment monitoring, spatio-temporal tracking, and locality-aware computational journalism.

Spatial behaviour changes over time. One may be interested in a business district, a residential region, or a restaurant district on different occasions and at different times of year. Apart from local intent queries, it is also helpful to take temporal context changes into account, e.g., *spatio-temporal intent search*. For example, time of day impacts search: daytime and nighttime activities center around different locations. We introduce three novel and demanding spatio-temporal query types and an example emerging application that draws upon all of them.

4.1 Topic-Based Retrieval

Retrieving results that match a topic enables querying without specific search terms, seeking instead documents that pertain to a certain subject. Conventionally, with topics modelled as clusters of words (e.g., a set of unique n-grams [10]) and with documents retrievable via inverted file indexing, finding documents relative to a topic can be done quickly, perhaps by employing an initial retrieval pass followed by ranking according to a similarity heuristic between the topic cluster and each candidate document. This becomes difficult in a spatio-temporal context; as well as the indexing challenges mentioned above, given that one can return temporally- and spatially-relevant documents, similarity measures are likely to under-perform given short texts as in tweets. Further, the clusters used to describe a topic will differ between regions reflecting e.g. dialect and attitudes, as well as over time. For example, see the tag clouds in Figures 6 and 7, showing how terms used in Wikipedia articles to describe the US president change from 2007 to 2011[4]. Improving topic-based search to account for spatial and temporal variations in topic models is a challenging task.

4.2 Sentiment Monitoring

Sentiment analysis attempts to extract mood and object-related sentiment from subjective texts. It is a growing field [11], attracting great funding from commercial entities interested in tracking political, product, and brand attitudes. Monitoring regional attitudes over time and in reaction to events offers a formal base that may enable data-driven change. Sentiment analysis has strong spatial requirements (e.g., linking messages such as "`The Kwik-e-Mart manager is sleeping on the job again..`" to a specific Kwik-e-Mart location, or finding regions with mixed feelings about issues or people) as well as temporal requirements as mentioned in Section 2 (enabling, for example, the tracking of historical attitudes preceding large events such as riots and elections). Accumulating time series about mood changes in given regions can inform the prediction of future social and political events nearby.

4.3 Transport Modes

The modes of transport chosen by a city's inhabitants is a valuable metric for evidence-based city planning, for monitoring and controlling pollution levels, and for determining willingness to travel to a given point of interest. Previous studies have shown maximum commute distances based on mode of transport [6] (see Figure 5), but this data is time-consuming to extract, as it requires personal interviews and tracking, and the results often only apply to a single city or region. Huge spatio-temporal samples can be collected quickly and cheaply from social media data, leading to simple trajectory extraction. However, the frequency of spatio-temporal samples is typically lower than in GPS-based tracking systems, making the extracted trajectories less accurate. Analysing trajectories and text content can suggest a mode of transport with some confidence, which is useful for returning personalised results and for informing spatial search. Mode of transport information enables evidence-based infrastructure planning, such as the successful introduction of "bicycle highways" in Copenhagen after the discovery that some residential estates were just beyond the distance that citizens felt comfortable cycling along city roads.[5] Effective management of spatio-temporal social media data makes collection of such information cheap and fast.

4.4 Local Computational Journalism

The ability of social media to near-instantly report news is well-known (e.g., for providing notice of earthquakes [5]). Extracting significant real-world events from spatio-temporally-positioned live media text holds the potential to enable fully automatic detection and reporting of news – an advancement in computational journalism [4]. This method can be much faster than manually discovering, researching, and writing a story. Where traditionally, journalists had to go and find stories and content, e.g. by talking to people, keeping an eye on court records, or monitoring local emergency services, social media pushes breaking local stories directly to the journalist. While work exists on event detection in social media [10, 8, 12], the spatial aspect of stories is novel, especially in the context of effective data management and retrieval.

This spatially-aware event-reporting scenario places significant

[4]Generated with thanks to http://tagcrowd.com.

[5]http://news.xinhuanet.com/english/world/2012-04/17/c_131533236.htm

and novel demands upon a data management system under constant use, especially with regards to spatio-temporal search. Not only do systems have to be able to filter and cluster location information, dealing with uncertainty and irregular spatial regions (Section 3.1), they also need to be able to identify stories from the background noise of social media messages and track the development of these stories in both space and time. Current high-level challenges include separating emerging events from background noise based on time-series trends and, once an event is detected, identifying the regionality of events (e.g., street, city, or country level) [19].

Regionality is important because different events are relevant at varying scales. Without this, it is difficult to judge which audience the story is for, so the spatial aspect is critical. For example, a burst pipe may only affect a small neighbourhood and is often not relevant to the whole city. Also, a traffic jam affects road-users along specific routes, but not people who do not travel near it. However, a tainted water supply is often relevant to not only a city, but also to its surrounding area. Critically, hyper-local events (such as spontaneous art performances, burst pipes, and traffic jams) may be interesting, but are not discovered by conventional journalism as effectively as they can be by social media.

In addition to regionality, it is important to determine whether the events identified are of interest. Some types of event may occur relatively frequently and may be of low impact. This requires historical information about events in a region of interest, so that past occurrences can be found. Frequent, region-specific occurrences should have less emphasis than rare ones. For example, heavy downtown traffic may emerge as an event with specific spatial and temporal locality, but if it happens every day at about 4 p.m., each successive occurrence is less interesting as a piece of emerging news. However, downtown explosions are likely less frequent than heavy traffic, and so explosion-type events could be ranked higher. It is also important to consider the spatial aspect of these events: heavy rainfall may be common in England but rare in Ethiopia, and the events should be weighted accordingly.

The overall goal is to prime locality-aware computational journalism [4]; instead of journalists seeking out stories, spatially-aware social media monitoring should push stories to them. The initial challenge is to accurately identify local emerging events and extract news stories not found elsewhere, such as at major news outlets. Secondary goals include determining which events are the most interesting, and ranking them accordingly.

5 Conclusion

Data-centric computing for social media that is aware of spatial and temporal contexts presents formidable research challenges. Solving these challenges promises better personal interactions with computers, supporting information-seeking behaviour, providing fresh and more relevant results. Spatio-temporal social media analyses at large scale help predict trends in a variety of domains, ranging from medical and sociological to business and political. Effective context-aware search and analysis over this genre of information enables a wide range of powerful technologies.

Acknowledgments This work was supported by EU funding under grants 287863 (TrendMiner[6]) and 264994 (Geocrowd[7]).

6 References

[1] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *Proc. TempWeb workshop*, pages 1–8, 2011.

[2] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu. Spatial keyword querying. In *Proc. International Conference on Conceptual Modeling*, pages 16–29, 2011.

[3] V. P. Chakka, A. Everspaugh, and J. M. Patel. Indexing large trajectory data sets with SETI. In *Proc. CIDR*, 2003.

[4] S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In *Proc. CIDR*, pages 148–151, 2011.

[5] P. Earle, D. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6):708–715, 2012.

[6] K. Halldórsdóttir, L. Christensen, T. Jensen, and C. Prato. Modelling mode choice in short trips – shifting from car to bicycle. In *Proc. European Transport Conference*, 2011.

[7] B. Hecht, L. Hong, B. Suh, and E. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proc. HFCS*, pages 237–246, 2011.

[8] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using twitter. In *Proc. ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32, 2011.

[9] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? Inferring home locations of Twitter users. In *Proc. ICWSM*, pages 511–514, 2012.

[10] M. Mathioudakis, N. Bansal, and N. Koudas. Identifying, attributing and describing spatial bursts. *PVLDB*, 3(1–2):1091–1102, 2010.

[11] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In *Proc. @NLP can u tag #usergeneratedcontent?! workshop*, 2012.

[12] A. Ritter, O. Etzioni, and S. Clark. Open domain event extraction from Twitter. In *Proc. SIGKDD*, pages 1104–1112, 2012.

[13] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Proc. ICWSM*, pages 322–329, 2012.

[14] D. Sidlauskas, S. Saltenis, and C. S. Jensen. Parallel main-memory indexing for moving-object query and update workloads. In *Proc. SIGMOD*, pages 37–48, 2012.

[15] R. Sugumaran and J. Voss. Real-time spatio-temporal analysis of west nile virus using twitter data. In *Proc. Int'l Conference on Computing for Geospatial Research and Applications*, pages 39–40, 2012.

[16] G. Weikum, N. Ntarmos, M. Spaniol, P. Triantafillou, A. Benczúr, S. Kirkpatrick, P. Rigaux, and M. Williamson. Longitudinal analytics on web archive data: It's about time! In *Proc. CIDR*, pages 9–12, 2011.

[17] D. Wu, G. Cong, and C. S. Jensen. A framework for efficient spatial web object retrieval. *The VLDB Journal*, 21(6):797–822, 2012.

[18] D. Wu, M. L. Yiu, C. S. Jensen, and G. Cong. Efficient continuously moving top-k spatial keyword query processing. In *Proc. ICDE*, pages 541–552, 2011.

[19] J. Xu, A. Bhargava, R. Nowak, and X. Zhu. Socioscope: Spatio-temporal signal recovery from social media. In *Proc. ECML-PKDD*, pages 644–659, 2012.

[20] X. Zhang, H. Fuehres, and P. Gloor. Predicting asset value through twitter buzz. *Advances in Collective Intelligence*, 113:23–34, 2011.

[6] http://www.trendminer-project.eu/
[7] http://www.geocrowd.eu/