SAT: Spatial Awareness from Textual input *

Dmitri V. Kalashnikov Yiming Ma Sharad Mehrotra
Ramaswamy Hariharan Nalini Venkatasubramanian Naveen Ashish

Information and Computer Science
University of California, Irvine

1 Motivation

Recent events (WTC attacks, Southeast Asia Tsunamis, Hurricane Katrina, London bombings) have illustrated the need for accurate and timely situational awareness tools in emergency response. Developing effective situational awareness (SA) systems has the potential to radically improve decision support in crises by improving the accuracy and reliability of the information available to the decision-makers. In an evolving crisis, raw situational information comes from a variety of sources in the form of situational reports, live radio transcripts, sensor data, video streams. Much of the data resides (or can be converted) in the form of free text, from which events of interest are extracted. Spatial or location information is one of the fundamental attributes of the events, and is useful for a variety of situational awareness (SA) tasks.

This demonstration will illustrate our approach, techniques and solutions for obtaining spatial awareness from raw input text. There are several challenges that arise in obtaining spatial awareness from raw text input - modeling/representation, event extraction and disambiguation, querying, reasoning and visualization. We specifically focus on illustrating solutions for (a)modeling and representation that captures spatial uncertainty in text and (b) efficient indexing and processing of various types of spatial queries to support reasoning of spatial information. Our solutions are implemented in the context of a prototype system called SAT (spatial awareness from text)that models and represents (potentially uncertain) event locations described in free text and incorporates several types of spatial queries of interest in SA applications. We demonstrate SAT in the context of 2 real-world applications that derive spatial information from text at different phases of the disaster response process.

• Offline spatial analysis of data from the Sept 11, 2001 WTC attacks to retrieve relevant events and the response as it occurred.

• Online, real-time assistance to field personnel using real time communication transcripts between dispatchers and first responders from 911 call centers in Los Angeles area.

Such tools enable social scientists and disaster researchers to accurately analyze transcribed communication logs and situational reports filed by the first responders after major disasters. These techniques can also be used to support real-time triaging and filtering of relevant communications and reports among first responders (and the public) during a crisis. Our primary objective is to

* This work was supported by NSF grants 0331707, 0331690

design database solutions to support applications where the real world is being monitored (potentially using a variety of sensing technologies) to support tasks such as situation assessment and decision-making.

An illustrative example. Consider a scenario during the response to the September 11, 2001 WTC attacks that demonstrates the need for spatial awareness. The following are excerpts from two real reports[1] filed by *Port Authority Police Department* (PAPD) Officers:

1. "*... the PAPD Mobile Command Post was located on West St. north of WTC and there was equipment being staged there ...*"
2. "*... a PAPD Command Truck parked on the west side of Broadway St. and north of Vesey St. ...*"

These two reports refer to the same location, i.e. the same command post – a point-location in the New York, Manhattan area. However, neither of the reports specify the exact location of the events; they do not even mention the same street names. Our objective is to represent and index such reports in a manner that enables efficient evaluation of spatial queries and subsequent analysis using the spatial data. Our system should have efficient supports to commonly used spatial queries, such as range, NN, spatial join, and so on. For instance, the representation must enable us to retrieve events in a given geographical region (e.g., around World Trade Center). Likewise, it should enable us to determine similarity between reports based on their spatial properties; e.g., we should be able to determine that the above events might refer to the same location(assuming a temporal correlation of the events).

To support spatial analyses on free text reports, merely storing location in the database as free text is not sufficient either to answer spatial queries or to disambiguate reports based on spatial locations. For example, spatial query such as 'retrieve events near WTC', based on keywords alone, can only retrieve the first report mentioned earlier. So instead, we need to project the spatial properties of the event described in the report onto the 2-dimensional domain Ω and answer queries within this domain. In this paper, we model uncertain event locations as random variables that have certain probability density functions (*pdfs*) associated with them. Assisted by GIS and probabilistic modeling tools, we map uncertain textual locations into the corresponding pdfs defined in Ω. Given that a large number of spatially uncertain events can potentially arise during crisis situations[2], the focus of our project is on developing scalable solutions for effective and efficient processing of such spatially-uncertain events.

2 Research Challenges and Solutions to be Demonstrated

Development of an end-to-end approach for spatial awareness from textual input must address four practical challenges – (1) modelling uncertain spatial events, (2) representation, (3) indexing, and (4) query processing. In Figure 1, we show

[1] original audio data available in converted text form

[2] For instance, more than 1000 such events can be extracted from just 164 reports filed by Police Officers who participated in the disaster of September 11th, 2001.

the major components of SAT. In the remaining of this section, we briefly describe the functionalities of SAT components and demonstrate the potentials of SAT in handling these challenges.

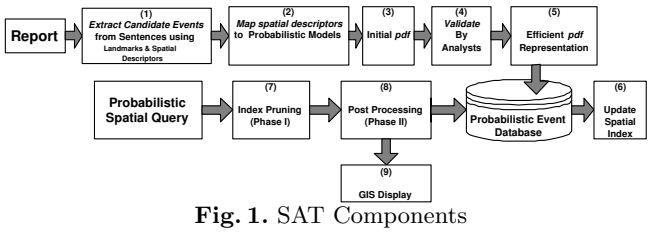

Fig. 1. SAT Components

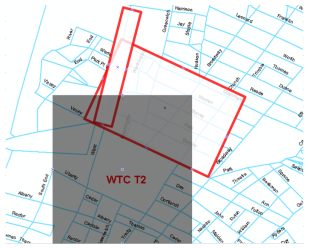

Fig. 2. WTC: Data and Query

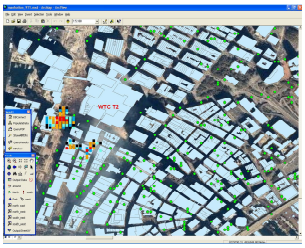

Fig. 3. GIS Interface

Modeling. Spatial uncertainty has been explored both in the GIS and in database literature. We extend the probabilistic model for spatial uncertainty developed in [1–3]. In the probabilistic model, an uncertain location ℓ is treated as a continuous random variable (r.v.) which takes values $(x, y) \in \Omega$ and has a certain probability density function (pdf) $f_\ell(x, y)$ associated with it. Interpreted this way, for any spatial region R, the probability that ℓ is inside R is computed as $\int_R f_\ell(x, y)dxdy$. However, to apply the probabilistic models in our context requires us to solve: (a) event extraction from text, (b) modelling spatial uncertainty in text. The first four components (1–4) of Figure 1 are meant for these tasks. First, automated tools are employed to extract events from text, including their spatial properties. The analyst oversees this process to correct errors arising from this process and also to resolve extraction ambiguities (cf. [4]). Next, we map the extracted textual location into the corresponding probabilistic representation in a semi-supervised fashion. Our modelling solution [5] takes a spatial expression (s-expression) as its input and outputs the desired pdf for the s-expression. It achieves that by analyzing landmarks and spatial descriptors (s-descriptors) mentioned in the s-expression. The analyst oversees these steps and adjusts the models if needed. We integrate this modelling process as a toolkit to the standard GIS system as shown in Figure 3. For example, our extraction and modelling tools can automatically determine the uncertainty regions of the two reports in Section 1, and display them as in Figure 2. An analyst can use the probabilistic modelling toolkit to further enhance the probabilistic models. Besides demonstrating the modelling process, we will also demonstrate the practical significance of using the probabilistic models. Showing together with query processing demonstration, we will show that simple bounding region models are not sufficient to answer analytical queries.

Representation. In our context, we need to be able to represent pdfs of complex shapes in the database. There are several known methods for such a representation, such as histograms and modeling pdf as a mixture of Gaussians or of other distributions. However, these solutions cannot scale well. In [6], we have proposed a novel compressed quad-tree representation. We have implemented this representation in the SAT component No. 5 in Figure 1. Coupled with our new indexing strategies, we will demonstrate significant performance boost in query response time. It is interesting to note that the existing solutions that also deal with probabilistic spatial queries [1–3] do not address the representation issues directly. The reason is that their empirical evaluation is carried out using only simple densities such as uniform and Gaussian.

Indexing and query processing. SAT efficiently supports several spatial query types – such as range, NN, and spatial join – commonly used in SA applications. For example, using a spatial region query, an analyst can express a query such as "find all the events, the location of which are around WTC". Figure 2 shows this query visually (shaded region). The system should compute the probability of the events inside this region, and filter away low probability events. In [6], we have proposed a novel grid base indexing approach. Compared to the state-of-arts techniques proposed in [1, 3], our new indexing scheme has 2–10 times speedup. The new index solution has been incorporated into SAT system as component 6,7 and 8 in Figure 1. In our demonstration, using both real and synthetic data, we will demonstrate the efficiency of the indexing solution on different types of spatial query.

3 Concluding Remarks

In this paper we presented a system – SAT – which builds spatial awareness and provides for reasoning with spatial locations from textual input. Such Situational Awareness (SA) applications abound in a variety of domains including homeland security, emergency response, command and control, process monitoring/automation, business activity monitoring, to name a few. We believe that techniques such as ours can benefit a very broad class of applications where free text is used to describe events.

References

1. Cheng, R., Kalashnikov, Prabhakar, S.: Querying imprecise data in moving object environments. TKDE **16** (2004)
2. Cheng, R., Kalashnikov, D., Prabhakar, S.: Evaluating probabilistic queries over imprecise data. In: SIGMOD. (2003)
3. Cheng, R., Xia, Y., Prabhakar, S., Shah, R., Vitter: Efficient indexing methods for probabilistic threshold queries over uncertain data. In: Proc. of VLDB. (2004)
4. Woodruff, A., Plaunt, C.: GIPSY: Georeferenced Information Processing SYstem. (1994)
5. Kalashnikov, D.V., Ma, Y., Hariharan, R., Mehrotra, S.: Spatial queries over (imprecise) event descriptions. Submitted for Publication (2005)
6. Kalashnikov, D., Ma, Y., Mehrotra, S., Hariharan, R.: Spatial indexing over imprecise event data. (In: Submitted to EDBT 2006)