Looking at CAREY: Controlling Climatological Regions in State of Emergency

Maribel Acosta, Marlene Goncalves, and Maria-Esther Vidal

Departamento de Computación
Universidad Simón Bolívar
Caracas 89000, Venezuela
{macosta,mgoncalves,mvidal}@ldc.usb.ve

Abstract. We present CAREY, an alert system notification for regions in state of emergency. CAREY implements the mediator-wrapper architecture on top of the Geospatial Web to visualize risky regions in terms of their weather conditions. We illustrate the formalization of the problem of detecting risky regions as a two-fold problem that relies on annotations of sensor data with the ontology of Observations and Measurements (O&M) to enhance state-of-the-art data mining and ranking techniques. First, sensor observations are clustered according to their Geospatial information then, proximate regions are clustered into micro-climate areas in terms of the similarity of their weather conditions. Finally, Top-k Skyline techniques are used to identify the top-k areas that best meet criteria of risk among the areas that are incomparable with respect to this condition. We demonstrate the capabilities of CAREY.

1 Introduction

Rapid changes in the world climate possibly produced by greenhouse gas emissions, are perturbing the livelihoods of large populations and firing extreme events that may cause high numbers of casualties. Event notification systems that provide a push-based discovery of areas that are more likely to have a disaster, may help government agencies to rapidly assist affected people. To achieve this goal, we developed CAREY, a ClimAtological contRol of regions in state of EmergencY tool, to identify regions that best meet a weather risk condition. CAREY is tailored to constantly receive sensor observations of weather conditions which are represented as RDF properties; the ontology O&M-OWL is used to describe the semantics of the observations. Sensor observations are grouped according to their Geospatial information; micro-climate areas are created from proximate regions by clustering them in terms of the similarity of their weather conditions. Clustering techniques [2, 3] are enhanced with the semantics encoded in the O&M-OWL ontology to exploit the meaning of the observations values during the clustering process. Then, micro-climate areas are ranked to identify the ones that are possibly in risk. A risk condition is modeled as a set of thresholds on the values of the sensor data; for example, a risk condition may establish that a region could be in risk when the values of humidity are at least 90% and

the temperature is at least $100°F$. Skyline and top-k ranking techniques are used to identify the top-k critical regions with respect to a risk condition.

We demonstrate the benefits of the approach and show the following key issues: the discovery capability of the two-fold approach by modeling a real-world domain, and the scalability by demonstrating how CAREY is able to identify the most risky regions in a large space of observations. The demo is published at `http://maribelacosta.com/carey/`.

2 CAREY Architecture

CAREY is built on top of the Geospatial Web which comprises sensor data annotated with the O&M-OWL ontology and is accessible through a federation of Semantic Sensor Services. CAREY is based on the architecture of wrappers and mediators [4], and integrates data exported by the services (Figure 1). CAREY receives a risk condition that establishes the risk thresholds for each of the sensor observation parameters and the number of risk regions that can be assisted; the answer of a request is the top-k critical regions that best meet the risk condition.

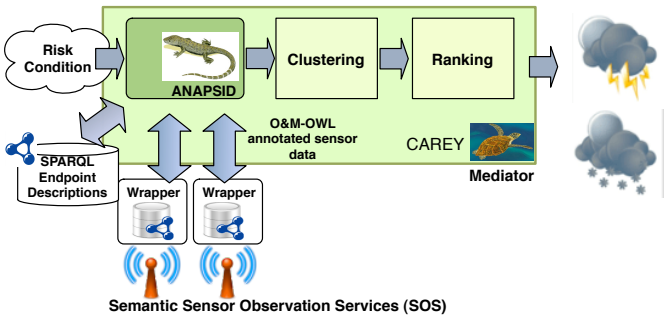


Fig. 1. The CAREY Architecture

ANAPSID [1] query engine is used to recover data from services and integrate all the observations in a unified format. *Wrappers* are created around Semantic Sensor Services to properly build the service URL and convert the results in appropriate formats. *Mediators* maintain information and statistics about the Web services. The following two components process the integrated sensor data:

Clustering: is an out-of-shell component implemented by WEKA, [1] that groups sensor data into clusters in two steps. First, geospatial information is considered to compute proximate regions; altitude, sea level, and geospatial coordinates are taken into account in this step. Then, sensor observations are used to cluster the regions based on the similarity of their weather conditions. These

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

clusters correspond to micro-climate areas or zones that characterize a particular climate that is considerable different from the climate of the surrounding areas. Micro-climate areas are common in regions with high elevations as the ones in the Andes Mountains, where each region can have itself a particular climate, which may vary several times in a day. Clusters represent micro-climate areas and their centroids correspond to a vector of the mean values of the sensor observations. The observations in a micro-climate area are similar, i.e., their sum of squares to the cluster centroid is minimal. Different cluster algorithms are used to create the micro-climate areas, e.g., X-means [3]. O&M-OWL annotations are used to determine distances to the centroids, e.g., the condition of closeness of two temperatures depends on the unit of the measurements: Celsius or Fahrenheit. Micro-climate regions are computed independently of the risk condition.

Ranking: this component implements ranking techniques to identify the top-k regions that best meet a risk condition among the non-dominated regions. Non-dominated regions are micro-climate areas with at least one observation value in the centroid, that is better than the same observation parameter of the centroids of the other areas. These areas have also at least one parameter in the centroid whose value is worse or equal than the value of this parameter in the centroids of the other non-dominated areas. Furthermore, a region is top-k, if it is among the k regions with the smallest distance to the risk condition. Information encoded in the O&M-OWL ontology is also used to determine the distance of an observation to the risk condition.

3 Demonstration of Use Cases

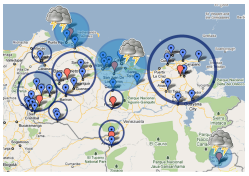
Consider a network of weather stations located all around Venezuela; each station measures variables such as temperature, rainfall, etc. Sensor data is used to determine regions with micro-climates, i.e., zones where their weather conditions considerably differ from nearby areas. Suppose some government disaster control agencies are interested in determing regions in state of emergency in terms of how close their measurements are to a risk condition. Suppose only temperature and rainfall are considered, which are both equally important. Further, assume that a region is in risk, if its temperature is less than $10°F$ and its rainfall is equal or greater than 50mm in a day. Thus, a region will be more risky, if there is no other region with a lower temperature and a higher rainfall among the regions with temperature less than $10°F$ and rainfall equal or greater than 50mm. Note that to compare two observations, the type of the parameters and the units of measurements encoded in the O&M ontology need to be considered.

Formally, a set of more risky micro-climate regions is comprised of regions that are non-dominated by any other region in terms of the former criteria; this set of regions is called Skyline. A region A dominates a region B, if B has greater value in temperature and less in rainfall than A. Based on this, regions 1, 2, 3, and 4 of Table in Figure 2(a) are the non-dominated ones. Finally, to attend the top-3 most critical regions, distances between the temperature and rainfall of the regions to the risk conditions are considered. Table in Figure 2(a) illustrates the

Euclidean distance values. These values are used as scores to identify that regions 4, 2 and 3 are the ones that have rainfall and temperature measurements that best meet the risk condition. Figure 2(b) illustrates the top-3 critical regions.

Region	Sensor Observations		Euclidean Distance
	Rainfall	Temperature	
1	99	8.4	49.02
2	90.2	6	40.67
3	95	6.6	45.12
4	87	1.6	37.94

(a) Euclidean **Distance** to the Risk Condition



(b) The Top-3 Critical Regions of Weather Stations

Fig. 2. Risky Micro-climate Areas

We will demonstrate four scenarios with sensor data from Venezuela stations; data was provided by WeatherUnderground [2] circa March 2010, Dec 2010, Feb 2011, and June 2011. Geospatial information and sensor observations such as temperature, dew point, sea level pressure, and relative humidity are used to identify micro-climate areas. Four risk conditions will be illustrated as well as the benefits of using semantics encoded in the O&M-OWL ontology.

4 Conclusions

We present a two-fold approach to identify the top-k critical regions that best meet a risk condition. Regions correspond to cluster of weather stations with similar sensor observations and geospatial information. Critical regions are modeled as Skyline points while the top-k most critical correspond to the Top-k Skyline regions with respect to a risk condition. We will demonstrate the capabilities of the clustering and ranking techniques implemented in CAREY.

References

1. M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. ANAPSID: AN Adaptive query ProcesSing engIne for sparql enDpoints. In *Accepted at ISWC*, 2011.
2. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. The analysis of a simple -means clustering algorithm. In *Symposium on Computational Geometry*, pages 100–109, 2000.
3. D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.
4. G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

[2] `http://www.wunderground.com/`