A Data Mining Framework for Analysing Geospatial-Temporal Data

Raunak Sarbajna   
*Department of Computer Science   
Lamar University*Beaumont, Texas, USA  
rsarbajna@lamar.edu

Sujing Wang  
*Department of Computer Science   
Lamar University*Beaumont, Texas, USA  
rsarbajna@lamar.edu

*Abstract*—Change analysis and automatic storytelling are essential techniques in understanding patterns and trends in multifaceted, time-series geospatial-temporal data. In this paper, we introduce a new data mining framework for analyzing the change patterns of spatio-temporal data. It includes change analysis techniques and automatic storytelling methodology for spatio-temporal data. We evaluate the effectiveness of our framework through case studies involving Twitter emotion data and North American Drought data. The experimental results show that our framework can discover interesting change patterns and useful information from spatial-temporal data.

Keywords—change analysis, storytelling, sentiment analysis, polygon, spatio-temporal data.

# Introduction

Analyzing change in spatial-temporal data is critical for many applications including developing early warning systems that monitor environmental conditions, detecting political unrest and crime monitoring. Change analysis models are essential in understanding larger patterns and trends in multifaceted, time-series geographic data. The purpose of this study is to detect spatio-temporal changes within sequential (time-series) geospatial-temporal data.

Modern technology digitizes wide sources of information constantly, with hour after hour of data being stored and most of it being unprocessed raw data. This is especially true with remote sensing networks and resultant geo-spatial imagery. Much of this data has meta-information associated with it which can reveal patterns or trend if properly classified and analysed. Performing basic change analysis on a subset of this data gives us a lot of insights that would otherwise go unrecognised.

A common understanding is that most big data available today is either archival, media or web scrapes [25]. However, a large source of that data is actually from Geographic Information Systems (GIS), and the tools available to interpret them easily are lacking [26]. The multiplicity of APIs have standardized access and structuring, but they limit much of the meta-data associated with them. Most publicly available (i.e. non-governmental) 'big data' sources with spatial components revolve around data scraped from mobile software platforms, including twitter, Instagram, Snapchat and reviews on mapping apps [26]. The recent API changes by Google Maps, for instance, aptly show that data content, and thus its meaning, is subject to regulation that is outside the control of researchers. The goal of this research project is to detect and analyze how the patterns of features change over time and space in spatio-temporal data and automatically story-telling based on the time-serials of spatio-temporal data.

Our approach provides a change monitoring framework which creates a change graph that captures the changes in spatial land uses clusters and a change summarization framework that creates specific change summaries based on the change graph based on the change story types.

Our research contributions are summarized as follows:

1. A novel change analysis framework that works on vectored polygonal datasets
2. New change predicates that are data agnostic and can work on a large spectrum of data
3. A new measure of interestingness to aid in generating automated storytelling based the change analysis results

The rest if the paper is structured as follows. Section 2 reviews previous literature on the subject and discusses related work. Section 3 introduces our data mining framework and lays out the methodology in detail. Section 4 evaluates the framework with case studies on drought datasets pre and post Harvey, pre and post California wildfires and twitter emotion polygons. Section 5 provides a conclusion and discusses potential future expansions to the framework.

# Related Work

A survey of the classical change detection algorithms can be found in the Lu et al. [3] paper and tells us that the integrated GIS and remote sensing approaches yield the best results. However, they are very sensitive to registration accuracies between images. Thus, images must be properly orthorectified and georeferenced, especially because the changes in the emotion polygons are so subtle. This assumes the emotions are to be treated as just another feature in the map, like any other category.

Since our data is primarily in an urban environment, with all the grid like rigidity that entails, it is a good idea to look at change detection algorithms optimized for urban environments. One of the hardest aspects to measure is to distinguish between change and no-change, as well as different kinds of change. Comparing image differencing, image regression, tasselled-cap transformation and chi square transformation, Ridd and Liu [3] find image differencing to be the most consistent, with a sustained overall accuracy of >80%.

It is useful to have a programming-oriented study comparing several of the change detection algorithms using MATLAB, rather than pure application-oriented comparison, in order to have a benchmark. Minu and Shetty [5] analyzed image differencing, image ratioing, change vector analysis, tasseled cap transformation and principal component analysis for efficiency and effectiveness. Although their area of study was not urban but a variety of land use/ land cover, change vector analysis gave the best overall accuracy.

We also studied two novel methods that are recent developments and are showing promising results: Neighborhood Correlation Image and Comprehensive Change Detection Method, both of which are optimized for remote sensing imagery but can be adapted to vectorized maps without loss of generality.

The change detection model using Neighborhood Correlation Image (NCI) logic works because of the obvious fact that the same geographic area (e.g., a 3x3 pixel window) on two dates of imagery will tend to be highly correlated if little change has occurred, and uncorrelated when change occurs [1]. Computing the piecewise correlation between two data sets demonstrates that NCIs contain change information and that NCIs may be powerful tools for change detection.

A high-performance remote sensing method called Comprehensive Change Detection Method (CCDM) integrates spectral-based change detection algorithms and a novel change model called Zone, which extracts change information from two Landsat image pairs [2]. This can be easily modified to work on the Twitter-based emotional grading maps. This method is simple, easy to operate, widely applicable, and capable of capturing anthropogenic changes like our area of interest.

Storytelling techniques are an effective summarization method to succinctly organize extensive information. Traditional storytelling has been mostly successful on news articles, blogs, as well as structured databases. However, traditional storytelling techniques tend to perform poorly on social media content, such as Twitter, where text lacks proper form and function [11]. Moreover, the ability to support dynamic storylines as they evolve is critical to modelling fast moving social media streams such as Twitter. Dos Santos et al. [21] introduced a set of methods to automatically derive stories over linked entities in tweets. They model a story as a graph of entities propagating through spatial regions in a temporal sequence, and controls search space complexity by suggesting regions of exploration. They developed algorithms to conduct storytelling to model tweets over space and time, reasoning over spatio-temporal features, and devise spatio-temporal storylines based on connectivity strength.

Kumar et al. [14] proposed an efficient storytelling implementation that embeds the CARTwheels [15] redescription mining algorithm which utilizes induced classification trees to model redescriptions in an A\* search procedure, using the CARTwheels to supply next move operators on search branches to the A\* search procedure. Vocht et al. [15] proposed the implementation of an optimized algorithm controlling the pathfinding process to obtain more homogeneous search domain and retrieve more links between adjacent hops in each path to improve the semantic relatedness of concepts mentioned in a story by increasing the relevance of links between nodes through additional domain delineation and refinement steps. Chen et al. [20] proposed a multimodal imitation learning via generative adversarial networks (MIL-GAN) method to directly model users' interests as reflected by various data by imitating users' demonstrated storylines. MIL-GAN model is designed to learn the reward patterns given user-provided storylines and then applies the learned policy to unseen data. Santos et al. [21] combined storytelling and Spatio-logical Inference (SLI) to generate rules of interaction among entities and measure how well they forecast a real-world event.

Hossain et al [13] introduced Google Fusion Tables(GFT) that offers collaborative data management in the cloud for data scientists to enable the integration of increasingly complex geospatial data to support storytelling. The paper focused on introduction of overview of map processing in GFT, the architecture overview of GFT, and how to scale to large datasets, massive and complex polygon datasets. GFT provides a useful tool for storytelling through interactive maps.

Kumar et al. [14] formulated storytelling as a generalization of redescription mining. Stories are defined as chains of redescriptions. They proposed an efficient storytelling algorithm as A\* search around the outputs of a CARTwhells redescription mining algorithm. The efficiency and scalability of the proposed algorithm were evaluated by three application case studies: word overlaps in large English dictionaries, exploring connections between gene sets in a bioinformatics data set, and relating publications in the PubMed index of abstracts.

Hossain et al. [19] proposed an approach to automatically construct stories between entities in large document collections that can help from directed chains of relationships, with support for co-referencing, evidence marshaling, and imposing syntactic constrains on the story generation process. A new optimization techniques based on concept lattice mining is used to rapidly construct stories on massive datasets.

Chen et al. [20] introduced an approach, multimodal imitation learning via generative adversarial networks (MIL-GAN) for generating storyline on unseen data. It can directly model users’ interests as reflected by various data. This approach is used to learn the reward patterns given user-provided storylines and then applies the learned policy to unseen data.

Santos et al. [21] introduced three methods of association analysis, Distance-based Byesian Inference, Spatial Association Index, and Spatio-logical inference, to capture relatedness among real-world events in high data volumes, and to model similar events that are described disparately under high data variability. It takes as input a set of geotemporally-encoded text streams about violent events called “storylines”. This study demonstrated that spatio-temporal storytelling is able to capture important associations among violent events reported in social media and traditional datasets.