Spatio-Temporal Narrative AI Agent for North America Event Analysis via MCP Architecture

Xiangxin Tang
Computer Science
Virginia Tech
906809608
xiangxin@vt.edu

Xing Gao
Computer Science
Virginia Tech
906747539
xingg@vt.edu

Yuxin Miao
Computer Science
Virginia Tech
906337428
yuxinm@vt.edu

Ziliang Chen
Computer Science
Virginia Tech
906831181
chenziliang@vt.edu

***Abstract*—In the era of big data, understanding the root causes of major socio-political hotspots requires tracing complex historical and geographical footprints. While traditional data management systems excel at isolated data retrieval, they lack the autonomous reasoning needed to uncover historical trajectories dynamically. To address this gap, we propose a Spatio-Temporal Narrative AI Agent designed to automatically discover and synthesize antecedent events leading up to a user-defined target event. Leveraging the GDELT 2.0 dataset focused on 2024 North American events hosted in a MySQL database, our system employs the Model Context Protocol (MCP) to bridge the gap between relational data and Large Language Models (LLMs). Through MCP tool integration, the LLM agent dynamically translates its reasoning steps into executable SQL queries (Text2SQL), autonomously traversing the MySQL database to retrieve preceding interactions among related actors and locations. By combining these dynamically queried structured records with unstructured news texts via a Retrieval-Augmented Generation (RAG) pipeline, the agent translates fragmented data into a coherent, chronological narrative of causality. Ultimately, this project demonstrates a novel approach to exploratory data analysis, showcasing how MCP-empowered agents can transform static relational databases into engines for dynamic spatio-temporal storytelling.**

I. Introduction

Large Language Models (LLMs) are revolutionizing database systems by shifting capabilities from mere textual retrieval to advanced semantic reasoning, planning, and tool use. When a major socio-political or geopolitical event breaks, understanding its context requires identifying its preceding triggers—often buried in massive, heterogeneous data lakes. The Global Database of Events, Language, and Tone (GDELT) 2.0 dataset captures these interactions at a global scale. However, exploring its raw tabular format requires analysts to manually write complex, iterative SQL queries to piece together historical timelines, creating a critical need for an intelligent system capable of automated "spatio-temporal storytelling". To solve this challenge, we propose a Spatio-Temporal Narrative AI Agent that acts as an autonomous tracing engine. Our project focuses on 2024 event data within North America (United States, Canada, and Mexico), centralized within a highly optimized MySQL relational database. The core innovation of our system is its retroactive event discovery mechanism, powered by the Model Context Protocol (MCP). Given a user-specified hotspot event, the LLM agent utilizes MCP tools to act as a data analysis agent. It autonomously generates and executes SQL queries directly against the MySQL database, enabling an iterative backward trace. By dynamically parsing involved entities (ActorName, ActorType1Code), locations (ActionGeo), and timeframes (SQLDATE), the agent fetches a logical chain of antecedent events. Furthermore, the system extracts the SOURCEURL attribute from these historical precursors to fetch unstructured news texts, feeding them into a RAG pipeline to comprehend contextual nuances. Following the CS5614 project guidelines, our workflow encompasses four main tasks: (1) Data Crawling and Preprocessing of the GDELT dataset into MySQL; (2) Database Design and Implementation optimized for temporal lookbacks; (3) LLM-based Data Analysis utilizing MCP tools for dynamic SQL generation and causal storyline construction; and (4) Data Visualization to render interactive event timelines. By seamlessly integrating MySQL database operations with MCP-driven LLM reasoning, this project delivers a powerful tool for discovering the hidden historical trajectories behind modern news events.

II. Related Work

The analysis of geopolitical events and risks through structured data has become a cornerstone of modern political geography and computational social science. A significant body of work leverages large-scale event datasets, such as the Global Database of Events, Language, and Tone (GDELT), to quantify and analyze the spatio-temporal dynamics of conflict and cooperation. This study builds upon and contributes to two primary strands of this literature: the measurement of geopolitical risk and the forecasting of violent events using narrative or "storytelling" approaches. The methodological lineage of our forecasting approach can be traced to work on spatio-temporal storytelling and inference. Dos Santos et al. (2014) proposed a framework that combines storytelling with Spatio-logical Inference (SLI) to forecast location-based events. Their core contribution lies in transforming sequences of entity interactions (storylines) into probabilistic rules. By applying relaxed logical operators to calculate a "distance to satisfaction" for these rules, their method provides a quantitative measure of how well a sequence of trigger events predicts a final violent event. This focus on the narrative

structure of events—their cascading nature, propagation, and sequencing—offers a powerful lens for understanding how discrete incidents coalesce into larger-scale outcomes. The authors demonstrate the efficacy of this approach in forecasting events in Afghanistan, showing that SLI can outperform traditional probabilistic methods in both precision and recall. While distinct in their primary objectives, these two bodies of work are deeply complementary. The risk measurement framework of Du et al. (2024) identifies where and when geopolitical risk is high, providing a valuable macro-level context. In contrast, the spatio-temporal storytelling method of Dos Santos et al. (2014) offers a mechanism to understand how that risk might materialize by modeling the progression from precursor events to a final outcome. Our work synthesizes these perspectives. We adopt the rigorous, GDELT-based empirical grounding exemplified by Du et al. to define our study area and validate the broader significance of the events we aim to forecast. Simultaneously, we operationalize the core tenets of the storytelling approach—entity interaction, spatial proximity, and temporal sequencing—to build a predictive model. By integrating the macro-level diagnostic power of risk assessment with the micro-level prognostic capability of spatio-temporal storytelling, this study aims to provide a more holistic and actionable understanding of geopolitical dynamics. This integrated approach not only helps in identifying potential flashpoints but also in understanding the narrative pathways that lead to them, thereby offering a more robust tool for analysis and decision-making.

III. Proposed Approach

The core methodology is divided into database construction and the design of the MCP-driven analytical framework.

A. Custom Database Creation (Data + Schema)

To satisfy the project requirements, we will ingest raw event data from the `GDELT 2.0` dataset [**?**], filtering specifically for North American geographic bounding boxes. A custom relational schema will be designed in **MySQL**, optimized with indices specifically tailored for:

- **Time-series range scans:** Ensuring rapid retrieval across specific temporal windows.
- **Geospatial queries:** Utilizing R-tree indexing for efficient spatial filtering.

B. MCP-Driven Analytical Pattern

Instead of forcing the LLM to write raw SQL, we will develop a custom **MCP (Model Context Protocol) Server** that exposes high-level, semantic database operations as callable tools [**?**].

The AI Agent will invoke these tools to fulfill three primary Spatio-Temporal Narrative tasks:

1) **Retrospective Event Lineage:** The agent will use MCP tools to dynamically query the MySQL database backward in time, identifying the chain of political or economic shifts that led to a specific North American event.
2) **Precursor Pattern Recognition:** The agent will synthesize historical data to summarize early warning signs of specific event types, identifying commonalities in past occurrences.
3) **Cross-Regional Perspective Analysis:** The agent will invoke tools to simultaneously query and compare localized reporting differences. For example, the system will be able to analyze and contrast narrative framing emerging from the Philadelphia and southern New Jersey areas against perspectives from Washington D.C. or Atlanta.

IV. System Design

A. System Overview

The proposed system is an AI-driven analytical platform designed to transform raw, multi-dimensional event data from the GDELT Project into coherent, context-rich narratives. By integrating an **Autonomous Agent** with **Retrieval-Augmented Generation (RAG)** and **Spatio-Temporal indexing**, the system allows users to explore the "how" and "why" behind event clusters across North America through natural language interaction.

B. System Architecture

The architecture follows a modular four-tier structure, specifically optimized for MySQL's Spatial and Vector features to meet the CS5614 project requirements for advanced data management and LLM integration.

1) I. Data Layer (MySQL DBMS):

- **Relational Storage:** MySQL (v8.4+ LTS) serves as the core engine to store structured event data from GDELT.
- **Spatial Data Management:** Utilizing MySQL's built-in Spatial Data Types (`POINT`, `GEOMETRY`) and Spatial Indexes (R-tree) to store and query the `ActionGeo_Lat/Long` of North American events. This allows for efficient proximity searches (e.g., `ST_Distance_Sphere`).
- **Integrated Vector Storage:** Leveraging the new `VECTOR` data type in MySQL 8.4+ to store semantic embeddings of event summaries and themes. This enables "Hybrid Retrieval"—combining traditional SQL filtering with semantic similarity search within a single database instance.

2) II. Processing & Integration Layer (Hybrid RAG Pipeline):

- **ETL Pipeline:** A Python-based ingestion service filters GDELT records for North America (`CountryCode` in 'US', 'CA', 'MX').
- **Embedding Engine:** Converts non-structured GKG (Global Knowledge Graph) themes and event snippets into high-dimensional vectors using OpenAI's `text-embedding-3-small`.
- **Hybrid Retriever:** Coordinates queries by executing Structured Queries (e.g., filtering by `GoldsteinScale` and `Date`) and Semantic Queries (vector similarity) to provide the LLM with a comprehensive context.

3) III. Intelligence Layer (AI Narrative Agent):

- **Reasoning Engine:** Powered by a Large Language Model (GPT-4o).
- **Agentic Workflow:**
  1) **Query Planning:** Analyzes user's natural language to identify spatial, temporal, and thematic constraints.
  2) **Text-to-MySQL Tool:** Dynamically generates MySQL-compatible SQL code, including spatial functions (`ST_Within`) and vector distance functions (`VECTOR_DISTANCE`).
  3) **Self-Correction Loop:** Inspects MySQL error logs and regenerates corrected SQL if a query fails.
  4) **Narrative Synthesis:** Transforms raw query results (events, tones, locations) into a structured story.

4) IV. Presentation Layer (Interactive Interface):

- **Narrative Dashboard:** Built with Streamlit, providing a conversational interface.
- **Spatio-Temporal Visualization:** Uses Mapbox or Pydeck to render the "Storyline" on a map, dynamically highlighting geographic clusters and time intervals in North America.

C. Technical Stack

The following table summarizes the primary tools and languages utilized in the system implementation:

TABLE I
SYSTEM TECHNICAL STACK

Category	Tools/Language
Programming Languages	Python, SQL
DBMS	MySQL 8.4+ LTS
Spatial Engine	MySQL Spatial Extensions
Vector Indexing	MySQL Vector Store
Database Driver	`mysql-connector-python`
Orchestration	LangChain
LLM API	OpenAI GPT-4o
Frontend	Streamlit
Visualization	Pydeck / Leaflet

D. Data Sets

- **Source:** GDELT Project 2.0 (filtered for US, CA, MX in 2024).
- **Storage:** MySQL 8.4 Tables (`events_table` + `gkg_table`) with Spatial and Vector Indexes.
- **Key Fields:** `GLOBALEVENTID`, `SQLDATE/DATEADDED`, `ActionGeo_Lat/Long`, `EventCode` (CAMEO), `GoldsteinScale`, `V2Tone`, `V2Themes`, and `SOURCEURL`.

E. Narrative Workflow Example

User Input: "How did the wildfires in Canada last summer affect the public discourse in the Northern US?"

Agent Logic: 1. Generates SQL to find "Wildfire" themes in Canada (Summer 2023).
2. Generates SQL to find "Public Sentiment" events in Northern US for the same period.
3. Retrieves news snippets via Vector Search.

Output: A structured narrative explaining smoke migration, environmental policy mentions, and localized sentiment shifts.

V. APPENDIX

A. Task Assignment

- **Xiangxin Tang:** Project lead; ETL pipeline development; Data crawling and preprocessing of GDELT dataset.
- **Xing Gao:** Database design; Implementation of Spatial and Vector indices in MySQL; Query optimization.
- **Yuxin Miao:** MCP Server development; Agent reasoning logic; RAG pipeline integration for news text analysis.
- **Ziliang Chen:** Frontend development; Streamlit dashboard implementation; Pydeck/Mapbox spatio-temporal visualization.

B. Schedule

- **Feb 24 - Mar 10:** Data ingestion, MySQL schema setup, and baseline spatial query implementation.
- **Mar 11 - Mar 24 (Checkpoint 1):** Completion of the MCP Server and basic Text-to-SQL agent functionality.
- **Mar 25 - Apr 14 (Checkpoint 2):** RAG pipeline integration and refinement of the narrative synthesis engine.
- **Apr 15 - Apr 28 (Final):** Frontend dashboard completion, system testing, and final documentation.

REFERENCES

[1] Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global Data on Events, Language, and Tone, 1979–2012. ISA Annual Conference.

[2] Anthropic. (2024). Model Context Protocol Specification.