Geospatial Knowledge Hypercube (Demo Paper)

Zhaonan Wang, Bowen Jin, Wei Hu, Minhao Jiang, Seungyeon Kang, Zhiyuan Li, Sizhe Zhou, Jiawei Han, Shaowen Wang
University of Illinois Urbana-Champaign
{znwang,bowenj4,weih9,minhaoj2,sk129,zhiyuan5,sizhez,hanj,shaowen}@illinois.edu

ABSTRACT

Today a tremendous amount of geospatial knowledge is hidden in massive volumes of text data. To facilitate flexible and powerful geospatial analysis and applications, we introduce a new architecture: *geospatial knowledge hypercube*, a multi-scale, multidimensional knowledge structure that integrates information from geospatial dimensions, thematic themes and diverse application semantics, extracted and computed from spatial-related text data. To construct such a knowledge hypercube, weakly supervised language models are leveraged for automatic, dynamic and incremental extraction of heterogeneous geospatial data, thematic themes, latent connections and relationships, and application semantics, through combining a variety of information from unstructured text, structured tables, and maps. The hypercube lays a foundation for many knowledge discovery and in-depth spatial analysis, and other advanced applications. We have deployed a prototype web application of proposed geospatial knowledge hypercube for public access at: https://hcwebapp.cigi.illinois.edu/.

CCS CONCEPTS

• **Information systems** → *Language models*; *Information extraction*; *Clustering and classification.*

KEYWORDS

Knowledge Hypercube, Geographic Information Retrieval, Weakly-Supervised Text Classification

ACM Reference Format:
Zhaonan Wang, Bowen Jin, Wei Hu, Minhao Jiang, Seungyeon Kang, Zhiyuan Li, Sizhe Zhou,, Jiawei Han, Shaowen Wang. 2023. Geospatial Knowledge Hypercube (Demo Paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625629

1 INTRODUCTION

Today a tremendous amount of geospatial knowledge is hidden in massive volumes of text data (e.g., news reports, research papers, and social media). One primary example task is geoparsing [12], a sub-task of Named Entity Recognition (NER), aiming to extract named geographic entities, or toponyms, from unstructured text corpus. Geoparsing task is closely related to other geographic information retrieval (GeoIR) tasks, such as toponym disambiguation [6], location description recognition [3], and geocoding [2]. While there have been a series of studies working on improving the performance on these GeoIR tasks [8, 13, 18], limited attention has been paid to the underlying knowledge structure between geospatial and other dimensions (e.g., temporal, thematic). The extracted high dimensional structured knowledge lays a foundation for many knowledge discovery tasks, such as ontology construction, event extraction, and other geospatial applications. Compared to the typically incomplete and imperfect results of metadata, latent connections across different knowledge dimensions can be derived to lower barriers for even other cross-disciplinary applications.

In this demo paper, we introduce a new architecture: *geospatial knowledge hypercube*, a multi-scale and multidimensional knowledge structure that integrates information from geospatial dimensions, thematic themes and diverse application semantics, computed and extracted from massive spatial-related text data. Specifically, to construct such a knowledge hypercube, weakly supervised machine learning approaches [10, 17] need to be developed for automatic, dynamic and incremental extraction of heterogeneous geospatial data, thematic themes, latent connections and relationships, and application semantics, through combining a variety of information from unstructured text, structured tables, maps, and image data [4, 5]. The cube construction will also need to develop new methods for recognizing geospatial entities and inferring geospatial relationships. We have deployed a prototype web application of proposed geospatial knowledge hypercube, and keep it evolving by adding more new features.

2 GEOSPATIAL KNOWLEDGE HYPERCUBE

Given a geospatial text corpus $\mathcal{D}$, the knowledge hypercube for $\mathcal{D}$ is a multidimensional data structure. Each document $d \in \mathcal{D}$ lies in one or several multidimensional cube cells to characterize the textual content of the document from multiple aspects. The formal definition of the geospatial knowledge hypercube is shown as follows:

Definition 2.1. Geospatial Knowledge Hypercube. A geospatial knowledge hypercube for a text corpus $\mathcal{D}$ is a n-dimensional structure $C = (\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_n)$, where $\mathcal{L}_i$ is the i-th hypercube dimension. Each document $d \in \mathcal{D}$ resides in one or several hypercube cells $\{(l^k_{t_1}, ..., l^k_{t_n})\}_k$ in C, where $(l^k_{t_1}, ..., l^k_{t_n})$ is one cell and $l^k_{t_i}$ is the label of d in dimension L_i for the k-th related cell.

To give an example, consider we are constructing a geospatial knowledge hypercube about "aging dams" from a text corpus $\mathcal{D}$. The dimensions $\mathcal{L}_i$ can include dam names, dam locations, and document topics, which are related to the aging dam problem. Each $d \in \mathcal{D}$ can be assigned to one or several cells, *e.g.,* (Englewood Dam, Colorado, Human), (Carters Dam, Georgia, Ecosystem), *etc.*

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3625629

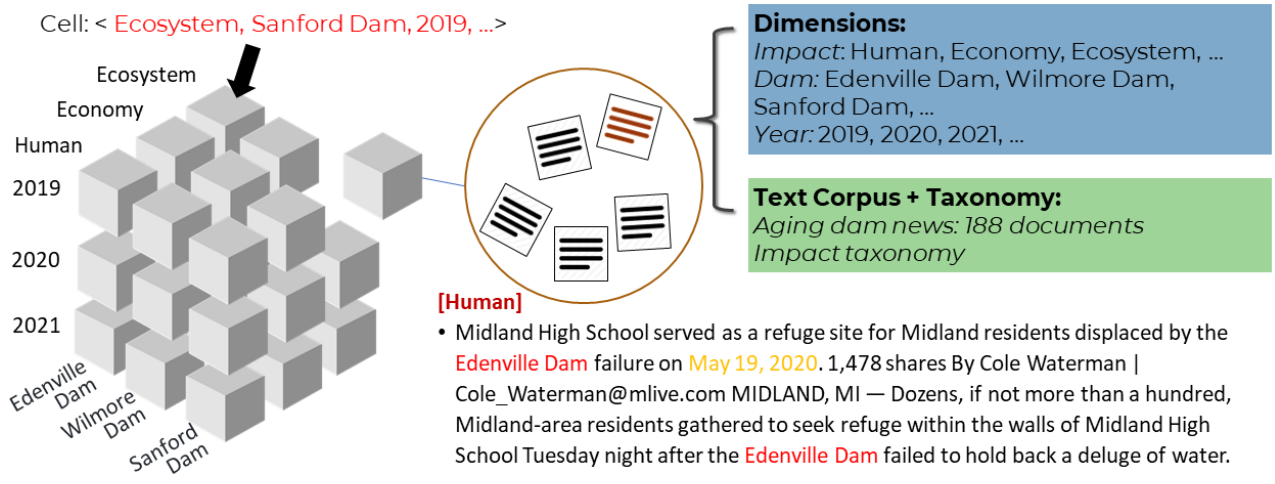

Figure 1: Construction of Geospatial Knowledge Hypercube from Unstructured Text Corpus

3 HYPERCUBE CONSTRUCTION

In this section, we discuss how to construct the geospatial knowledge hypercube. The hypercube construction can be formulated into a multidimensional text analysis problem where each dimension can be constructed separately. The dimensions of interest are categorized into two groups: knowledge span detection (for dimensions such as dam names and dam locations) and topic discovery (for dimensions such as document topic). For knowledge span detection, we adopt the language model-based named entity recognition method; while for topic discovery, we conduct document categorization with label names only.

3.1 Knowledge Span Detection

We propose to adopt knowledge span detection with pretrained language models (PLMs) [1, 7] for some dimensions such as dam names and dam locations. The PLMs are pretrained on large text corpora, which have the basic language understanding ability. The encoding process of PLM can be formulated as follows:

$$\boldsymbol{h}_{w_1}, ..., \boldsymbol{h}_{w_n} = \text{PLM}(d) = \text{PLM}(w_1, ..., w_n), \tag{1}$$

where d is the document, w_i is the i-th token in d and $\boldsymbol{h}_i$ is the output representation vector by PLM for w_i. The PLMs are further fine-tuned on token classification tasks (*e.g.,* named entity recognition [11]) to enable their ability for knowledge span detection. Specifically, the token hidden states are fed into a classifier $f(\cdot)$ to make token label prediction, *i.e.,* $y_{w_i} = f(\boldsymbol{h}_{w_i})$. After fine-tuning, the PLM and classifier $f(\cdot)$ are utilized for zero-shot knowledge span detection on our given corpus $\mathcal{D}$ for some specific dimensions $\mathcal{L}_i$ (*e.g.,* dam names and dam locations).

3.2 Text Classification

In order to classify each document $d \in \mathcal{D}$ into the corresponding document topics cell in the hypercube, we propose to leverage extremely weakly-supervised text classification methods [9, 10, 14–16] which only use class names, where each class name corresponds to a document topic pre-defined in the structure. In this case, we do not require human efforts with domain-specific knowledge to manually label each document, which is expensive and time-consuming to access in real applications. Specifically, during the experiments, we used LOTClass [10], which finds category-indicative words in the corpus for each label name and generalizes the prediction model with self-training on the documents with the unlabeled data.

4 SYSTEM DEPLOYMENT

We have deployed a prototype web application of geospatial knowledge hypercube. The front and back-end components of the system are illustrated in Figure 2.

4.1 System Components

Geospatial knowledge hypercube offers geospatial data analysis and visualization through geoparsing, utilizing user input data. Users can upload news articles, which are then processed to extract geospatial information. The extracted location data is highlighted in the text and visualized to the user, allowing for the intuitive identification of locations on the map using the Google Maps API. Moreover, the framework enables users to directly map specific text in the article to corresponding locations on the map for immediate visual reference. The entities of the spaCy library obtained from the geoparsing are visually distinguished with different colors to enhance the visualization experience.

To ensure scalability and reliability, the framework follows a micro-services architecture. The front-end leverages the React framework, enabling the flexible design of interactive interfaces and integration with various visualization libraries. The back-end, implemented using Flask, handles geoparsing efficiently using the spaCy model. During system deployment, the backend language model is pre-loaded to reduce response time for user requests. Containerization using Docker is employed for system deployment, offering scalability benefits and minimizing deployment time by providing independence from the deployment environment. This approach lays a solid foundation for seamless system expansion in response to increasing user demands in the future.

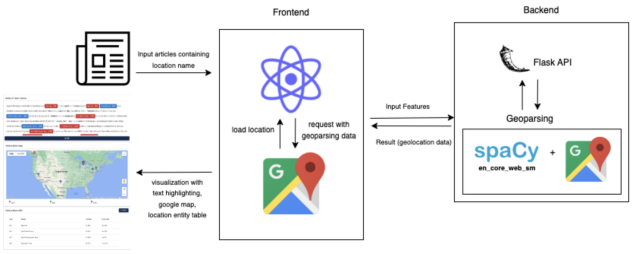

Figure 2: Web Application Components of Geospatial Knowledge Hypercube

4.2 Use Case: Aging Dam

We have collected a corpus consisting of 188 Google News on dam failures. U.S. dams are threatened by age-induced fragility and increased hydrologic stresses due to climate change. In many cases, communities and infrastructure below the dams have also increased dramatically over time, increasing the exposure to dam failure. Our geospatial knowledge hypercube system offers a unique capability to combine text and geospatial data analytics. To be specific, we focus on seven aspects for constructing aging dam hypercube, including *document type, date, location, dam name, watershed, government and organization, impact.*

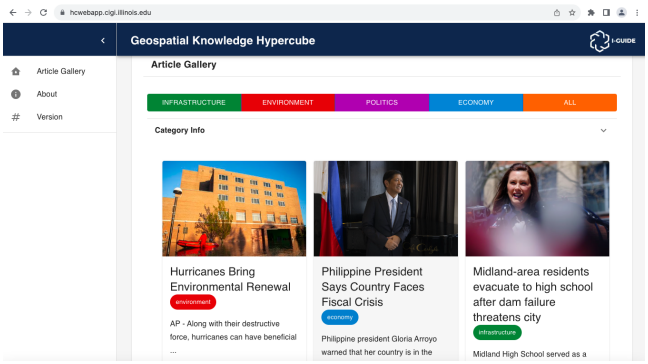

Figure 3: Hypercube Web App Input Page for Users to Select News based on Label-only Document Classification Results

Illustrated in Figure 3, users can select a target news based on the pre-trained label-only document classification results. Leveraging the pre-trained language model on the back end, we extracted different types of geo-entities as demonstrated in Figure 4: LOC for location in red, ORG for organization in blue, and FAC for facility in green, and further pinpointed them onto a map. This combination between text and geospatial contexts enables the direct linkage between this two types of data modalities. As illustrated in Figure 5, when users clicking on this marker on the map, the text document will be filtered correspondingly to only the sentences where this location is mentioned. This text-map linkage facilitates user's interaction to quickly gain spatially explicit knowledge. When we have bunches of articles at scale, it will save a great amount of time and efforts for the users. In the current version, our application also supports basic spatial analysis tools (e.g., buffer analysis) and export

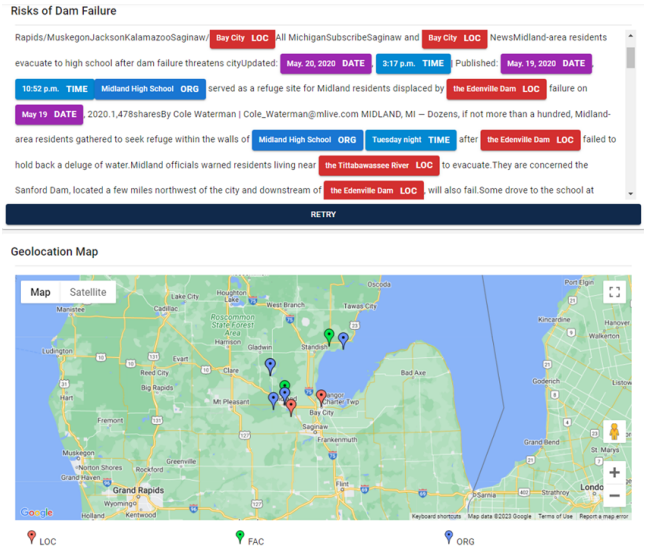

Figure 4: Hypercube Visualization of Multi-dimensional Named Entity Recognition (NER) and Geocoding Results

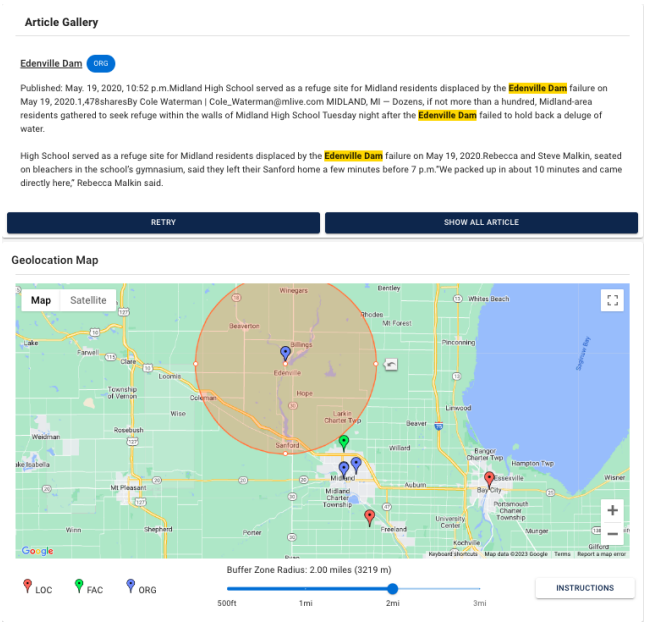

Figure 5: Text-Map Linkage for Fast Geographic Information-based Filtering and Buffer Analysis

of the extracted location entities with geographic coordinates into a json file, which opens more possibilities for cross-platform usage.

In addition to the extracted geospatial information, the web application can also visualize the recognized temporal information, such as time and date (tagged in purple in Figure 4), and classify sub-documents based on vocabulary expansion results of LOTClass [10]. Four labels, namely *infrastructure*, *human*, *ecosystem*, *economy*, are used as additional inputs in the current version. As demonstrated

Label	Similar Words Used in Aging Dam Corpus	Example Documents
Infrastructure	'infrastructure', 'heritage', 'development', 'network', 'structure', 'equipment', 'technology', 'system', 'property', 'structures', 'architecture', 'construction', 'dam', 'education', 'embankment', 'grid', 'facilities', 'security', 'powerhouse', 'sanitation' ...	'Brazilian mine tragedy will not be the last tailings dam disaster.' 'Brumadinho Dam Collapse: A Tidal Wave of Mud.' 'Guajataca Dam's Failure Highlights Puerto Rico's Infrastructure Issues.'
Human	'human', 'man', 'humans', 'individual', 'people', 'person', 'mankind', 'animal', 'normal', 'woman', 'civilian', 'humanity', 'civil', 'serious', 'player', 'moral', 'domestic', 'single', 'professional', 'mortal', 'live', 'american', 'male', 'internal', 'us' ...	'Russia dam collapse: At least 15 dead, others missing at gold mine.' 'Quebec officials warn of possible dam failure, forcing evacuation of 250 people.' 'Dam failure caused 11 deaths.'
Ecosystem	'ecosystem', 'ecosystems', 'wildlife', 'vegetation', 'forest', 'global', 'nature', 'park', 'climate', 'canopy', 'native', 'ecology', 'biodiversity', 'habitats', 'eco', 'fauna', 'rainforest', 'plant', 'indigenous', 'fragile', 'conservation', 'tree', 'society' ...	'Holtwood dam whitewater park on the Susquehanna river Raystown Lake in Huntingdon County is the largest lake wholly in Pennsylvania.' 'He had reportedly refused to follow orders to increase dam spillway capacity to handle severe floods.'
Economy	'economic', 'economically', 'financial', 'economical', 'economy', 'economics', 'commercial', 'agricultural', 'industrial', 'economies', 'business', 'uneven', 'international', 'entertainment', 'employment', 'impact', 'annual', 'income' ...	'Michigan Legislature seeks $500M to fund dam repairs.' 'BHP prepares for fresh battle against $6.6 bln Brazil dam lawsuit.' 'Michigan pegs Midland flooding damage to be $175 million.'

Table 1: Vocabulary Expansion Results based on Label-only Document Classification

in Table 1, LOTClass is trained to find semantically close words to the given labels and utilize these expanded word sets to further classify document in an unsupervised manner. For instance, given label word "environment", the model would find related seed words, such as "ecosystem", "wildlife", "vegetation", to guide the language model to make predictions. The labels can be utilized by users to quickly filter a large pool of documents and concentrate on one specific news topic, which is especially beneficial when users face a large volume of data.

5 CONCLUSION AND FUTURE WORK

In this demo paper, we introduce a new knowledge structure, namely *geospatial knowledge hypercube*, for weakly-supervised multidimensional knowledge retrieval. To demonstrate the functionality of hypercube, we have prototyped and deployed a web application with a use case on aging dam news corpus. The current version has shown promising potential to support cross-interdisciplinary knowledge discovery and applications. In the next step, we shall further improve the system interface, at the same time implement more use cases and additional functions, such as geospatial relationship inference, spatio-temporal event extraction and visualization.

ACKNOWLEDGMENTS

This work is supported by the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) funded by the National Science Foundation (NSF) under award No. 2118329.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.
[2] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which melbourne? augmenting geocoding with maps. ACL.
[3] Yingjie Hu and Jimin Wang. 2020. How do people describe locations during a natural disaster: an analysis of tweets from Hurricane Harvey. *arXiv preprint arXiv:2009.12914* (2020).
[4] Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023. Patton: Language Model Pretraining on Text-Rich Networks. *ACL*.
[5] Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. 2023. Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks. In *The Eleventh International Conference on Learning Representations*.
[6] Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. 2016. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *Knowledge Engineering and Knowledge Management (EKAW)*. Springer, 353–367.
[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692* (2019).
[8] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). In *ACM SIGSPATIAL*. 1–4.
[9] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-Supervised Neural Text Classification. In *CIKM*.
[10] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *EMNLP*. 9006–9017.
[11] Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *EACL*. 1–8.
[12] Jimin Wang and Yingjie Hu. 2019. Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In *3rd Workshop ACM SIGSPATIAL on GeoHumanities*. 1–6.
[13] Jimin Wang, Yingjie Hu, and Kenneth Joseph. 2020. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS* 24, 3 (2020), 719–735.
[14] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised Text Classification Based on Keyword Graph. In *EMNLP*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2803–2813. https://doi.org/10.18653/v1/2021.emnlp-main.222
[15] Yunyi Zhang, Minhao Jiang, Yu Meng, Yu Zhang, and Jiawei Han. 2023. PromptClass: Weakly-Supervised Text Classification with Prompting Enhanced Noise-Robust Self-Training. *arXiv preprint arXiv:2305.13723* (2023).
[16] Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, and Jiawei Han. 2023. The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study. In *Proceedings of the ACM Web Conference 2023*. 1626–1637.
[17] Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. In *WSDM*. 429–437.
[18] Bing Zhou, Lei Zou, Yingjie Hu, and Yi Qiang. 2023. TopoBERT: Plug and Play Toponym Recognition Module Harnessing Fine-tuned BERT. *arXiv preprint arXiv:2301.13631* (2023).