GIScience 2021 Short Paper Proceedings

Title

Geo-Event Question Answering Systems: A Preliminary Research Study

Permalink

https://escholarship.org/uc/item/9cs309kd

Authors

Kazemi Beydokhti, Mohammad
Duckham, Matt
Griffin, Amy
et al.

Publication Date

2021-09-01

DOI

10.25436/E2KW2T

Peer reviewed

Geo-Event Question Answering Systems: A Preliminary Research Study

Mohammad Kazemi Beydokhti
Department of Geospatial Science, RMIT University, Melbourne, Australia
s3763411@student.rmit.edu.au

Matt Duckham
Department of Geospatial Science, RMIT University, Melbourne, Australia
matt.duckham@rmit.edu.au

Amy Griffin
Department of Geospatial Science, RMIT University, Melbourne, Australia
amy.griffin@rmit.edu.au

Vedran Kasalica
Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
v.kasalica@uu.nl

Abstract

Designing a Geospatial Question Answering (GeoQA) system that takes a user's GIS-related domain question, understands how to gather the required data, how to analyse it, and how to present the results in a suitable format is arguably among the most important "moonshots" in the GeoAI field. In this study, we focus specifically on answering geo-event questions. This work begins by presenting a prototype process for generating workflows to answer geo-event questions by providing annotations of the domain, comprising a tool taxonomy we created from descriptions of geo-operations, a data type ontology obtained from the Core Concept Data types (CCD) ontology, and the annotations of the mentioned geo-operations with respect to the input/output pairs. Finally, the generated workflows are post-processed to restrict the solution space and provide more structured solutions. The results of this research provide a step towards the implementation of a geo-event QA system capable of answering diverse geo-event questions defined by users.

1 Introduction

Current Question Answering (QA) systems rely mainly on Information Retrieval (IR) and Knowledge-Based (KB) methods to automatically answer questions from the respective [3]. However, various studies have addressed the inefficiency of current generic QA systems for answering geospatial question types, which lead to more specialised research focuses [12]. A good example is GeoQA, which addresses spatial questions and their corresponding answers in depth in different aspects, including geospatial semantics [5], GIS workflow composition [4], spatial language processing [3] and answering geo-analytical questions [8].

Studies within GeoQA research address its different research challenges. In terms of answering questions, some studies focus on more general spatial questions that do not require an elaborate set of geo-operations to answer them [10, 9]. These approaches focus on automated translations of natural language questions into query languages over knowledge bases. For example, the question 'Which cities are within 200 km of Berlin?' can be answered by retrieving the geometry of Berlin from a knowledge base and then computing the spatial buffer of the selected geometry. The recent studies on this area are mainly working on extending existing knowledge graphs with geographic semantics [10], capturing geospatial semantics and syntactics in geospatial questions [13], and translating questions into executable queries [9].

On the other hand, there are a few studies that have created a system for addressing geo-analytical questions, which consist of transformations that involve spatial concepts more commonly generated by professionals in geography and the spatial sciences. As stated in [12], answering geo-analytical questions is a challenging problem for two main reasons. First, the answers to geo-analytical questions are not known a-priori, therefore, it is quite unlikely their answers will be accessible through information retrieval. Second, the system needs to capture the right potential tools and data to answer a question. Analytical workflows can be considered a suitable solution to address these two issues. Generating analytical workflows as answers to these types of questions has been proposed in different works [6, 8].

In this work, we focus on a specific type of geo-analytical question that has not been a focus of previous studies: geo-event questions. Geo-events are most succinctly defined as something that happens [2]. We address the problem of answering geo-event questions in two steps. First, we utilize the process of automated composition of workflows for a specific geo-event question. Second, the candidate solutions from the previous step are post-processed in order to narrow down the search space and get us closer to the actual answers.

2 Methodology

This section is divided into two subsections which discuss the corresponding conceptual basis of our approach. In Section 2.1, we demonstrate the process of automatically composing workflows for a sample geo-event question using the Automatic Pipeline Explorer (APE) framework. In Section 2.2, we propose two approaches for post-processing the generated solutions: *intensional* and *extensional*.

2.1 Automated composition of workflows

The APE framework [7] was recently proposed as an intuitive system that automatically composes executable workflows based on the problem specification. Based on our input datasets, our final goal, and a large set of available operations, APE will generate all possible workflows which take the input datasets and generate the desired output. APE relies on two main components: domain knowledge and workflow specification. Domain knowledge (provided by the domain experts) includes all the information about the tools and data types and how to use them, while workflow specification (provided by the end user) requires the description of the input data and the final output data based on a data type ontology.

Recently, the APE framework functionalities were demonstrated in a geospatial case study [8]. The study defined the data type taxonomy using different core concepts of spatial information, known as the CCD ontology. In addition, a tool taxonomy was defined based on the CCD ontology to specify all input types and the output type of the collected geo-operations. In their study, tool annotations were all based on different data type properties and APE generated solutions for their five geo-analytical questions quite effectively. However, in many cases, describing operations based on data types alone do not provide sufficient constraints to generate efficient solutions using APE. Let us take the example of SQL operations described in database query language, where operations are mainly based on tables inputs and all operations return tables as outputs. In this case, APE will give us an explosion of solutions and we will end up with an enormous number of possible workflows.

Another example where this approach might be problematic in the geospatial domain is with the use of map algebra, which input and output primarily rasters for all operations. Accordingly, it seems that we need more detailed descriptions of geo-operations than just their data type to provide a higher level of abstraction for specifying tools.

Brauner in his PhD thesis [1] presented six different descriptions of geooperators in a framework known as geooperator categories. This universal view about geooperator categories derived from different perspectives on geoprocessing operations as documented in the literature. The list of categories along with their corresponding definitions and examples are provided in Table 1.

Table 1 Geooperator categories with their definitions and examples

Geooperator Categories	Definition	Example
Legacy	GIS software the geooperator is implemented in.	ArcGIS, GRASS
Geodata	Refers to the data model and data properties.	Vector, Raster
Formal	Mathematical characteristics of geooperators.	Arity, Symmetry
Geoinformatics	Relating a GIScience concept to a geooperator.	Overlay, Map Algebra
Technical	Refers to implementation or technical details.	Linux, Windows
Pragmatic	Application for which a geooperator can be used.	Hydrology

In order, in order to automate the process of generating workflows in this study by using APE, we created our taxonomy of tools based on the Brauner's geooperator categories to include more information about the tools than just data type. Also, we utilized the CCD ontology for creating the data type taxonomy and for describing data types.

2.2 Postprocessing generated workflows

APE ranks the candidate workflows by their length, assuming that the shorter workflows are better than longer ones. However, to date, a very few studies worked on postprocessing the generated solutions to narrow down the solution space as well as to provide more structured solutions. For this purpose, in the current study we present two different post-processing approaches for grouping equivalent workflows: *intensional* and *extensional*.

The intensional approach groups equivalent workflows whose tool steps are semantically equivalent (i.e., equivalent in query intensions). Let us say we have the following workflows generated by APE with a length of three:

Workflow 1: Intersect → Buffer → v.select

Workflow 2: Intersect → v.buffer → v.select

Here, *Intersect* and *Buffer* are the ArcGIS tools of those names and *v.buffer* and *v.select* are the corresponding GRASS GIS tools. The only difference between these workflows relates to the second tool listed in each workflow. Although the *Buffer* and *v.buffer* geoprocessing tools are from two different software environments, they are semantically equivalent based on their output results, which each create a buffer zone for each geometry layer. By knowing this equivalency, workflows 1 and 2 can be grouped together using the intensional approach.

The extensional approach refers to grouping equivalent workflows that return the same outputs (i.e., query extensions) by running the input data through the workflows and comparing their output results. The main difference between the extensional and intensional approaches is that we might have workflows with different tools that are not semantically equivalent, but that return the same outputs. In the next section, we will define a similarity measure to check the equivalency of workflow outputs.

3 Results and discussions

3.1 Automated composition of workflows results

Currently, our repository has only 40 geoprocessing tools annotated based on geooperator categories[1]. Therefore, it is not possible at this stage to answer to all geo-event questions as it needs rich tool annotations. In this section, to illustrate our approach, we instead take one sample geo-event question and then explain the process of automated composition of workflows for it.

Q: What are the number of bushfires that occurred in the suburbs close to where the Canning River meets the Swan River in Perth?

To answer this question, the required inputs for the two main components of APE are prepared as follows:

Domain Modeling

The domain model is composed of a tool and type taxonomy and the operation annotations for capturing controlled geo-analytical concepts in the geospatial domain. For simplicity and conciseness, we have selected seven geoprocessing tools that are relevant to the sample question from the ArcGIS and GRASS GIS environments. Accordingly, we created the tool taxonomy for the selected tools based on the Brauner's geo-operation categories[2]. The seven tools have been parameterized based on the input types, output type, and the measurement scale level of attributes such as nominal, ordinal, ratio, etc. This results in 51 possible operations (Table 2). We used the formalized CCD ontology proposed in [11] for the data type taxonomy.

Table 2 Excerpt of the 51 parameterized geo-operations and their corresponding equivalent tool(s)

Geooperations	Parameterized tools	Equivalent tools
Intersect	Intersect_region_region_point_ordinal	
	Intersect_region_region_region_nominal	v.overlay_region_region_region_nominal
v.overlay	v.overlay_region_region_region_nominal	Intersect_region_region_region_nominal
	v.overlay_line_region_line_nominal	Intersect_region_line_line_nominal
Buffer	Buffer_point_region_nominal	MultipleRingBuffer_point_region_nominal v.buffer_point_region_nominal
	Buffer_point_region_ordinal	MultipleRingBuffer_point_region_ordinal v.buffer_point_region_ordinal

Workflow specification

The input datasets for the sample question consist of two river layers (Input type1 and Input type2) and the layer of Perth suburbs, which has the number of bushfires that occurred in each suburb (Input type3). All the input data sources are manually collected and provided as workflow inputs. The desired output is a map of nearest suburbs including the location of

---

[1] `https://github.com/GeoinformationSystems/GeooperatorBrowser`

[2] `https://github.com/MohammadUT/Geo_event-QA/blob/main/GeooperatorTaxanomy.jpg`

bushfires inside them (Output type). The input data type as well as the desired output type are annotated based on the CCD ontology as shown in Table 3.

Table 3 Inputs and output specifications in the CCD ontology

	Input specification			**Output specification**
CCD ontology dimensions	Input type1	Input type2	Input type3	Output type
CoreConceptQ	ObjectQ	ObjectQ	FieldQ	FieldQ
LayerA	LineA	LineA	VectorTesselationA	VectorTesselationA
NominalA	NominalA	NominalA	NominalA	NominalA

We set the required number of solutions to 50 and individually interpreted the generated workflows. Accordingly, APE could generate 35 workflows (70%) that return the correct answer, while six answers (12%) are invalid, and nine solutions (18%) are close to the actual answer, but do not completely match it (e.g., they provide a subset of the correct answer as shown here[3]). This diversity in the quality of the solutions is caused by the aforementioned similarity of the operation signatures, i.e., similarities between the input and output types. We could improve the quality of these results by expressing the user intents about the workflows by means of appropriate high-level constraints. However, automation of such constraints is not trivial and is left for future consideration.

Finally, we present the results of the two proposed post-processing approaches, intensional and extensional. We implemented a Python script that automatically retrieves the generated solutions from APE and equivalent tools (Table 2) as inputs, and groups the equivalent workflows based on the *intensional* approach. For the 50 generated solutions obtained from APE, this approach restricted the number of solutions to 24 groups, which means that 41% of the workflows were joined in the corresponding equivalence groups.

In order to compare the workflow outputs for the *extensional* approach, we take all the output geometries and measure how close these geometries are. For the sample question in this study in which the outputs are regions, we define a similarity measure by dividing the area of intersections by the total area. For this scenario, we considered two workflows to be equivalent when the similarity measure of their outputs is greater than 0.85. For identical outputs, the similarity measure will be equal to 1. The results showed the extensional approach restricted the 50 APE-generated solutions to only five groups. This approach grouped all those workflows that returned similar outputs, even if they have been parameterized differently, and this leads to grouping of a larger number of workflows compared to the intensional approach. Both approaches allowed us to have a better overview of the possible solutions. In addition, these classifications allow us to present more diverse solutions and explore different ways of solving the given problem.

4 Conclusion and Future Work

This paper focused on developing an automated mechanism for answering geo-event questions using APE framework to automatically compose workflows and two post-processing approaches to provide a more structured solutions. All the resources required for running the APE framework and the post-processing steps can be found on our GitHub repository[4].

---

[3] `https://github.com/MohammadUT/Geo_event-QA/blob/main/SolutionNo_2_length_2.png`

[4] `https://github.com/MohammadUT/Geo_event-QA`

The results of this study provide promising preliminary evidence for our future direction as about 88% of the generated solutions for the sample geo-event question were completely correct or close the correct answer and only 12% returned invalid workflows. Also, the results of post-processing revealed that it is possible to refine the solution space to a great extent in order to get closer to the correct solution, especially by applying the extensional method, which restricted the number of solutions to about 90%.

In this study, we attempted to highlight the importance of studying geo-event questions within the GeoQA field, as these have not been studied in detail in previous works. However, some of our future challenges in Geo-event QA are: 1) Capturing the semantic and syntactic structure of geo-event questions. 2) Ranking of the post-processed composed workflows. 3) Improving the precision of generated workflows in terms of obtaining higher number of completely correct answers by defining user intents using appropriate constraints. 4) Comparing our results in which tools were annotated based on geooperator categories with the results of recent study by [8] in which tool annotations were based on the CCD ontology alone to explore to what extent the use of different descriptions for the tools improves the precision of the generated workflows.

References

1 Johannes Brauner. Formalizations for geooperators-geoprocessing in spatial data infrastructures. 2015.

2 AP Galton. States, processes and events, and the ontology of causal relations. 2012.

3 Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. Place questions and human-generated answers: A data analysis approach. In *International Conference on Geographic Information Science*, pages 3–19. Springer, 2019.

4 Barbara Hofer, Stephan Mäs, Johannes Brauner, and Lars Bernard. Towards a knowledge base to support geoprocessing workflow development. *International Journal of Geographical Information Science*, 31(4):694–716, 2017.

5 Krzysztof Janowicz, Simon Scheider, Todd Pehle, and Glen Hart. Geospatial semantics and linked spatiotemporal data–past, present, and future. *Semantic Web*, 3(4):321–332, 2012.

6 Vedran Kasalica and Anna-Lena Lamprecht. Workflow discovery through semantic constraints: A geovisualization case study. In *International Conference on Computational Science and Its Applications*, pages 473–488. Springer, 2019.

7 Vedran Kasalica and Anna-Lena Lamprecht. Ape: A command-line tool and api for automated workflow composition. In *ICCS 2020*, pages 464–476. Springer, 2020.

8 Johannes F Kruiger, Vedran Kasalica, Rogier Meerlo, Anna-Lena Lamprecht, Enkhbold Nyamsuren, and Simon Scheider. Loose programming of gis workflows with geo-analytical concepts. *Transactions in GIS*, 25(1):424–449, 2021.

9 Haonan Li, Ehsan Hamzei, Ivan Majic, Hua Hua, Jochen Renz, Martin Tomko, Maria Vasardani, Stephan Winter, and Timothy Baldwin. Neural geospatial question answering. *Journal of Spatial Information Science*, 2009.

10 Dharmen Punjani, K Singh, Andreas Both, et al. Template-based question answering over linked geospatial data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, pages 1–10, 2018.

11 Simon Scheider, Rogier Meerlo, Vedran Kasalica, and Anna-Lena Lamprecht. Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, 2020(20):167–201, 2020.

12 Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. Geo-analytical question-answering with gis. *International Journal of Digital Earth*, pages 1–14, 2020.

13 Haiqi Xu, Ehsan Hamzei, Enkhbold Nyamsuren, et al. Extracting interrogative intents and concepts from geo-analytic questions. *AGILE: GIScience Series*, 1:1–21, 2020.