A PERSONALIZED GEOWEB SEARCH ENGINE BASED ON USER INTENT RECOGNITION

Regita Pramesti Nur Cahyani (1), Chih-Yuan Huang (1)
[1] National Central University, 300 Zhongda Rd., Zhongli District, Taoyuan City, 320, Taiwan
Email: regitapramesti96@gmail.com; cyhuang@csrsr.ncu.edu.tw

KEY WORDS: Geospatial Search Engine, Personalized Search, Semantic, Ontology, User Intent

ABSTRACT: Geospatial Web (GeoWeb) represents the collection of web resource that contain geospatial components, such as maps, geocoded images, web services hosting geospatial data, etc. Similar to general web resources, GeoWeb resources are scattered on the widely-distributed Internet, where the identification and integration of GeoWeb resources become a challenging task. In order to facilitate geospatial data discovery, we argue that a GeoWeb search engine is necessary. While GeoWeb Crawlers can proactively discover GeoWeb resources, establishing semantic relationships between GeoWeb resource helps discover relevant resources. However, in order to further facilitate GeoWeb resource discovery, we believe that a GeoWeb search engine needs to provide personalized search results. In this paper, we present the GeoWeb resource ontology, which contains necessary classes of a GeoWeb resource that integrate many domains ontologies. This ontology will help map and discover semantic relationships between concepts. To provide a personalized search, user models are constructed to illustrate the importance of concepts and their relations that represent user intents and interests. Therefore, in this research, we aim at analyzing user intents via user background information, search history, selected resources, etc., which are then applies to personalized search results. With the proposed solution, a GeoWeb search engine can provide personalized results and consequently help users find targeted GeoWeb resources more efficiently.

1. INTRODUCTION

Geospatial data is data about events or phenomena with all the combination of its spatial and temporal information. Geospatial data attracts attentions from many application domains, e.g., Geography Information System (GIS), Remote Sensing, Internet of Things (IoT), atmospheric science, etc. Geospatial datasets are diverse in nature, including digital imagery, thematic layers, and diverse sensor feeds (Georgos M. and Anthony S., 2011). To support comprehensive applications, applications need to obtain related geospatial data easily, efficiently, and effectively. These datasets are collected, stored, distributed by a variety of organization or agencies, and widely-distributed on the Internet, which causes issues in data discovery. Currently, applications and users search for targeted data by using data portals or Spatial Data Infrastructures (SDI). However, a user may not know where to find the data portals or SDIs that provide the data or services the user needs (Huang Chih-Y. and Chang H. 2016).

In order to facilitate geospatial data discovery, many Geospatial Web (GeoWeb) search engines and web portals ware built. However, these work still cannot achieve what a user really wants and return irrelevant result. That can happen largely due to the many variations in the context and background of the user, as well as the ambiguities of the text (Malthankar S. V. and Kolte S., 2016). While GeoWeb Crawlers can proactively discover GeoWeb resources, establishing semantic relationships between GeoWeb resource helps discover relevant resources, we believe that a GeoWeb search engine needs to provide personalized search results. Personalized Web Search (PWS) is one of the general search techniques that is intended to provide improved quality of search results, and is tailored to the individual needs of users (Malthankar S. V. and Kolte S., 2016). During the search process, service providers will try to improve search quality by utilizing personalization of user profile.

Therefore, the discussion regarding the user intent of the GeoWeb search engine is carried out. To overcome this problem, we present the GeoWeb resource ontology, which contains necessary classes of the GeoWeb resource that integrate many domains ontologies. This ontology will help discover semantic features of concepts in the classes. In term of personalized search, user models are constructed to represent user intents and interests. We aim at analyzing user intents via user background information, search history, selected resources, etc., which are then applies to personalized search results. The datasets used in this research are from GeoWeb Crawler of the previous study (Huang Chih-Y. and Chang H. 2016).

2. RELATED WORK

The volume of publicly available geospatial data on the internet is exponentially growth. To realize a Geospatial search engine, one of the first steps is to proactively discover Geospatial resources on the World Wide Web (WWW) or internet. A previous study (Christopher B. et.al., 2014) tried to utilize a web crawler built on top of the Google search engine called Geospatial Search Engine (GSE). This study runs daily the procedure to update map server layers and metadata, and to eliminate servers that go offline. This work represents an important step toward realizing the potential of a publically accessible tool for discovering the global availability of geospatial data. However, this study can't fully handle websites that contain shapefile data, because that metadata is not directly accessible from a website client interface. The other study (Huang Chih-Y. and Chang H. 2016) proposed a Geospatial Web (GeoWeb) Crawler, an extensible Web crawling framework that can find various types of GeoWeb resources, such as Open Geospatial Consortium (OGC) web services, Keyhole Markup Language (KML) and Environmental Systems Research Institute, Inc (ESRI) Shapefiles. This research applied the Bloom filter (Bloom B. H., 1970) to filter out the redundant URLs. The proposed crawling framework then used as base of a GeoWeb search engine of this research.

In addition of the crawler system, supporting the diverse needs of various users that may be accessing the same dataset for different applications remains a challenging issue. In order to overcome this challenge, (Georgos M. and Anthony S., 2011) offers a demonstration on how intelligent systems can assist geospatial queries to improve retrieval accuracy by customizing results based on preference patterns. However, this study lacks of aggregation of individual components of multidimensional user profiles. In the other hand, due to an expert usage of Information Retrieval (IR) and Semantic Web search in the computer science fields, the problem of user preference, and intent have been tried to solve by many researchers. An ontology based user model (Jiang X. and Tan Ah-H., 2009), called user ontology, for providing personalized information service in the Semantic Web was proposed. The proposed user ontology model utilizes concepts, taxonomic relations, and non-taxonomic relations in a given domain ontology to capture the users' interests. Specifically, we present a set of statistical methods to learn a user ontology from a given domain ontology and a spreading activation procedure for inferencing in the user ontology. However, this proposed model need to expand to collect user information during run time due to learning system that applied in the ontology learning.

In summary, while many systems have been proposed to build geospatial search engine that can give the most relevant results corresponding user's intents, most existing system can't fully capture a user's interest. To build more precise geospatial search engine, it is critical to explore effective ways of combining semantic relations with concepts for representing a user's interests on the geospatial search engine.

3. PROPOSED METHOD

This study aims at building a Geospatial Web (GeoWeb) search engine by utilizing geospatial ontology to gain semantic feature and integrate it with personalized search. The whole design of the system can be seen in the (Figure 1). A GeoWeb crawler provides geospatial services from internet as a base of GeoWeb search engine datasets. Meanwhile, these datasets normalized corresponding with the system requirement, GeoWeb Ontology is created. These normalize data and the GeoWeb Ontology will use to create linked data for semantic feature. This linked data then used to provide semantic data for search engine. When a user searches for data, the user-specified terms are combined in order to retrieve a potential geospatial resources. Before the search process begins, the user data will be collected to provide personalized. Looking at the block diagram of the research, the process will be divided into several steps as follows:

a. User Insert a Specific Term

   In this step, users will input the keyword that present their intent or interest. After receiving the user's keyword, system will process that term with removing stop words and doing stemming process. The result of this process will be checked again to make the most general term, so it will more easily to determine the user intents by using tokenization.

b. Data Retrieval from Linked Data

   In this step, the output term of tokenization process will be filtering with system's dictionary, so that the system can know that term is belong to what concept (Phenomena, Spatial, and Temporal). After the system know which concept it is, then the system will use it to query data from linked-data and get data that user intents by leveraging the semantic feature.

c. Give The Result Back to the User

In this step, main process is just doing the ranking evaluation and giving the results back to the user. The personalization will play its role in this step in order to score which resource should be display on top.

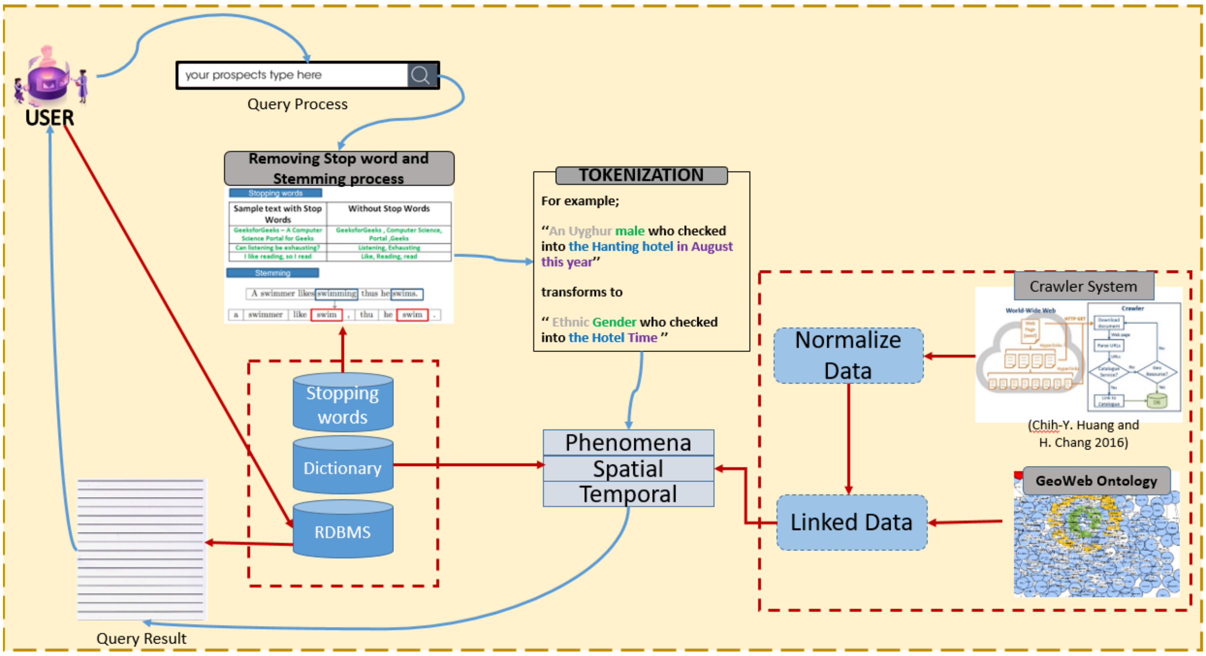

Figure 1. Block Diagram of How System Works

4. DESIGN EXPERIMENT

4.1Testing Limitation

a. The resource data used in this study is from GeoWeb crawler system.
b. There are 25 data that already linked with GeoWeb ontology, but in this data, there are just 3 service URL from Sensor Observation Service (SOS) available. The result will show in the array format.
c. Due to the linked-data limitation, the system just can find related user's term that the resource already linked with GeoWeb ontology. This user use keyword from Table 1. for testing the system.

Table 1. Keywords used for testing

User	*Phenomena*	*Spatial*	*Temporal*
	Sea water density	'POLYGON((20.608612082289973 -158.09158552099544,21.26046353009514 -158.09158552099544,21.26046353009514 -157.7086874598961,20.608612082289973 -157.7086874598961,20.608612082289973 -158.09158552099544))',0 Geocode into PA, united states	Begin: 2011-04-05T18:50:37.010Z End: 2011-05-24T21:18:35.784Z
	Sea water pressure	'POLYGON((20.96175 -158.0789462005226,21.265174916422367 -158.0789462005226,21.265174916422367 -157.70118333333332,20.96175 -	Begin: 2011-06-13T18:29:39.697Z End: 2011-07-30T00:50:57.310Z

		157.70118333333332,20.96175 -158.0789462005226))',0 Geocode into PA, united states	
	Sea water salinity	'POLYGON((42.82809829711914 -73.99040222167969,42.828198297119144 -73.99040222167969,42.828198297119144 -73.99030222167968,42.82809829711914 -73.99030222167968,42.82809829711914 -73.99040222167969))',0 Geocode into Auvergne-Rhone-Alpes, Savoie, france	Begin: 2013-12-09T04:00:00Z End: 2017-07-22T15:00:00Z
	Sea water temperature		
	Sea water turbidity		

5. RESULT AND ANALYSIS

5.1 GeoWeb Ontology

This ontology built by integration of many domain ontologies. The structure and the built-classes are created in this research, including the knowledge based. This ontology considers about 3 mains needed concept of GeoWeb search engine, which are phenomena, spatial and temporal. Figure 2. shows the classes in the ontology of *owl: Thing*. Figure 3. shows the classes in the phenomena object. Figure 4. shows the classes in the spatial object. Figure 5. shows the classes in the temporal object. Figure 6. shows the graph visualization of all the classes.

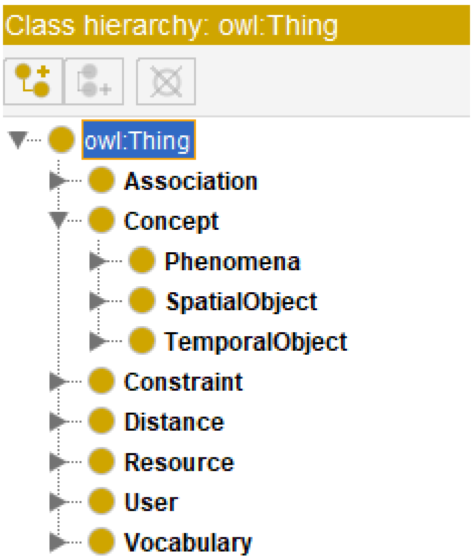

Figure 2. The classes in the ontology of *owl: Thing.*

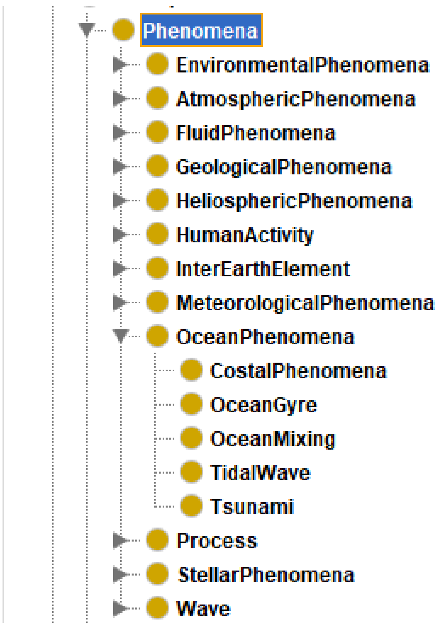

Figure 3. The classes in the phenomena object

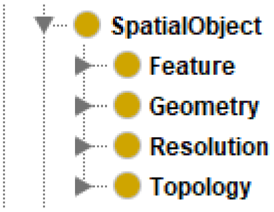

Figure 4. The classes in the spatial object.

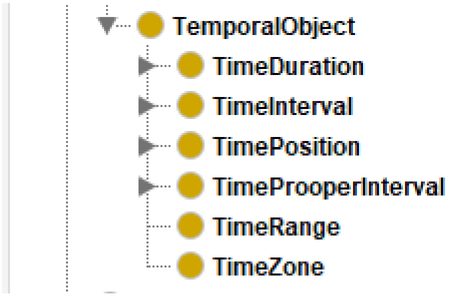

Figure 5. The classes in the temporal object.

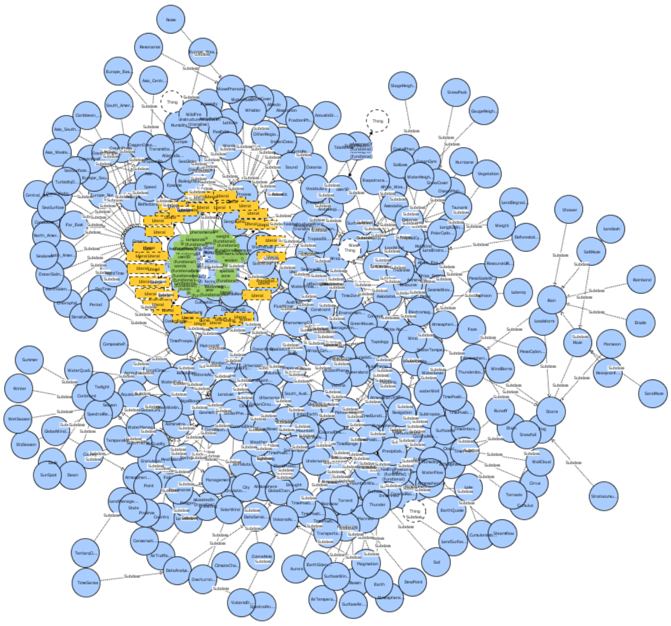

Figure 6. The graph visualization of all the classes.

5.2 System User's Interface

This interface is used to send query from user to get what the users intents are. The preliminary results show the process when user insert the keyword to the system and the system will send back the result. The results here a raw result in the array format. Figure 7. shows the user query field, place to insert user keyword. Figure 8. The user's keyword of "sea water density on sea water density on 2011-04-05T18:50:37.010Z until 2011-05-24T21:18:35.784Z". Figure 9. The preliminary results this system of the keyword of "sea water density on sea water density on 2011-04-05T18:50:37.010Z until 2011-05-24T21:18:35.784Z". Figure 10. The user's keyword of "sea water turbidity on france". Figure 11. The preliminary results this system of the keyword of "sea water turbidity on france".

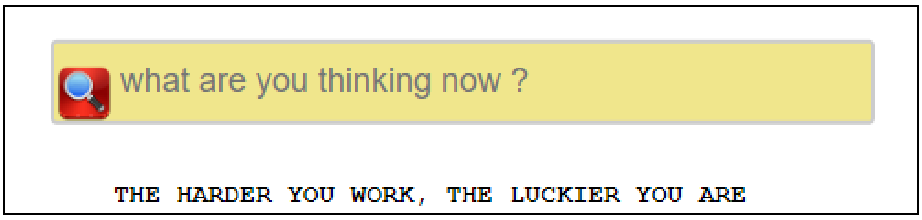

Figure 7. The user query field, place to insert user keyword.

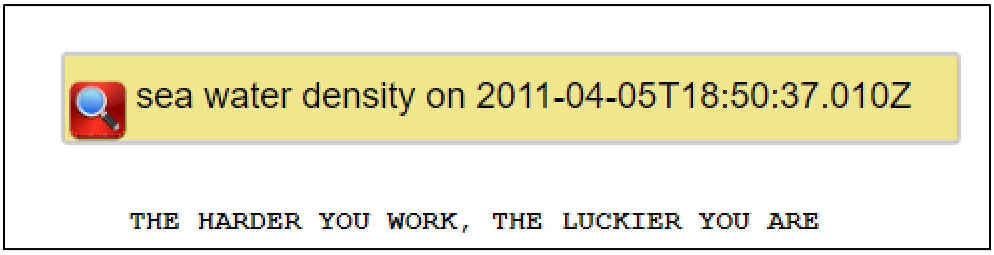

Figure 8. The user's keyword of "sea water density on sea water density on 2011-04-05T18:50:37.010Z until 2011-05-24T21:18:35.784Z".

array(0) { } array(3) { ["query_type"]=> string(6) "select" ["result"]=> array(2) { ["rows"]=> array(2) { [0]=> array(2) { ["ResourceURI type"]=> string(3) "uri" ["ResourceURI"]=> string(64) "http://oos.soest.hawaii.edu/thredds/sos/hioos/glider/sg139_9_agg" } [1]=> array(2) { ["ResourceURI type"]=> string(3) "uri" ["ResourceURI"]=> string(64) "http://oos.soest.hawaii.edu/thredds/sos/hioos/glider/sg523_1_agg" } } ["variables"]=> array(1) { [0]=> string(11) "ResourceURI" } } ["query_time"]=> float(0.017135143280029) }

Figure 9. The preliminary results of the keyword of "sea water density on sea water density on 2011-04-05T18:50:37.010Z until 2011-05-24T21:18:35.784Z".

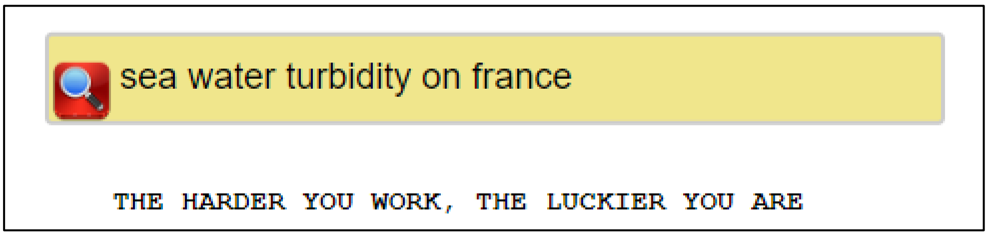

Figure 10. The user's keyword of "sea water turbidity on france".

array(0) { } array(3) { ["query_type"]=> string(6) "select" ["result"]=> array(2) { ["rows"]=> array(1) { [0]=> array(2) { ["ResourceURI type"]=> string(3) "uri" ["ResourceURI"]=> string(65) "http://sos.maracoos.org/stable/sos/hrecos/stationHRLCK8H-agg.ncml" } } ["variables"]=> array(1) { [0]=> string(11) "ResourceURI" } } ["query_time"]=> float(0.0084319114685059) }

Figure 11. The preliminary results of the keyword of "sea water turbidity on france".

6. CONCLUSION

GeoWeb search engine requires both semantic feature and personalized search to provide the high quality of information retrieval. That two functions can be expected to process what a search engine should do, especially geospatial search engine that need to consider about phenomena, spatial and temporal. Personalization is one of command knowledge to derive user intent recognition. Thus, the combination of semantic and personalize search is the best solution that can be achieved. This study uses many domain ontologies to create new ontology which can considerate with this kind of unique information of geospatial data, and also try to achieve user intent recognition to give a very relevant result to the user. The system works so good within the process from block diagram. When user insert the term or keywords, the results that given to the user is precise enough with user's intents. It shows in the preliminary result, even its still just a row data, but that results represent how GeoWeb works in a good direction. However, in order to give very high corresponding data with user intent, collecting user background and search history will help to enhance the relation within concept and the most related concept toward user's interests.

REFERENCES

Bloom B. H., 1970. Space/Time Trade-Offs in Hash Coding with Allowable Errors. Communication ACM, 13, pp. 422- 426.

Christopher B. et.al., 2014. A geospatial search engine for discovering multi-format geospatial data across the web. International Journal of Digital Earth, 9:1, pp. 47-62.

Georgos M. and Anthony S., 2011. Moving towards Personalized Geospatial Queries. Journal of Geographic Information System, 3, pp. 334-344.

Huang Chih-Y. and Chang H. 2016. GeoWeb Crawler: An Extensible and Scalable Web Crawling Framework for Discovering Geospatial Web Resources. ISPRS Int. J. Geo-Inf. 5(8), pp. 136.

Jiang X. and Tan Ah-H., 2009. Learning and inferencing in user ontology for personalized Semantic Web search. Elsevier. Volume 179, Issue 16, pp. 2794-2808.

Malthankar S. V. and Kolte S., 2016. Client Side Privacy Protection Using Personalized Web Search. Procedia Computer Science. 79, pp. 1029-1035.