Towards Keyword-Based Search over Environmental Data Sources

David Álvarez-Castro, José R. R. Viqueira(✉), and Alberto Bugarín

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela, Santiago de Compostela, Spain
david.alvarez.castro@rai.usc.es,
{jrr.viqueira,alberto.bugarin.diz}@usc.es

Abstract. This paper describes the problem of keyword-based search over environmental data sources. Based on a number of assumptions that simplify this general problem, a prototype of a search engine for environmental data was designed, implemented and evaluated. This first solution serves as a proof of concept that illustrates its applicability in different domains, for both expert and non-expert users. The requirements analysis undertaken and the subsequent design and implementation helped in the identification of a number of new research challenges.

Keywords: Keyword-based search · Environmental data
Geo-spatial data · Fuzzy query language · Information retrieval
Scientific data

1 Introduction

The monitoring and modelling of our environment (Earth Surface) is a task to which much effort is daily devoted, especially by public administrations. Most of the produced data is used by scientist of different areas to study and try to predict many disparate types of phenomena. A well-known and probably most representative example of the above is weather forecasting. As a consequence, very large amounts of geo-spatial data are produced nowadays by different means, including sensing devices and environmental modelling processes. Most of these data is only used by experts (mainly scientists). For example, environmental, socio-demographic and geographic datasets, which include properties such as rain-fall, winds, sea and air temperature, humidity, demography and economic development may be used in the field of public health to model the behaviour of diseases such as Cholera [1] and Influenza [10]. However, such data might also be of great importance in other applications areas. An obvious example of the above is the clear impact that meteorological data has in tourism and other applications related to open air activities.

This article is based upon work from COST Action KEYSTONE IC1302, supported by COST (European Cooperation in Science and Technology).

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 99–110, 2018.
https://doi.org/10.1007/978-3-319-74497-1_10

Keyword-based search technologies have already proved their effectiveness for the development of search engines in the area of Information Retrieval [7]. An example that we will use as a relevant metaphor or reference for the present work is the searching of books, which enables not only finding a book of interest for the user, but additionally to determine which parts of the book are of greater relevance for the specified set of keywords[1]. Recently, keyword-based search technologies have also been adapted to work over structured data sources, like relational databases, XML and linked open data sources [14]. The advantage over structured query languages is that the user does not need to know neither a specific syntax nor the data source schema. If we focus on geo-spatial data, keyword-based search technologies have been developed in the area of Geographic Information Retrieval [9] to search non structured datasets based on both textual and spatial criteria. Keyword-based search is also enabled to explore catalogues that contain metadata of structured geo-spatial[2] and environmental[3] datasets. Notice however that these catalogues enable the user to find a given dataset if it has been annotated with appropriate keywords, but they do not provide information about the parts of the dataset which are of greater relevance for the query, as Google Books does. Despite of all the above achievements, and to the best of these authors knowledge, none of the currently available technologies enable a scientist expert to search for datasets and specific areas and time periods within those datasets relevant for him/her. For instance, searching for areas of the planet involving "high water temperatures and rainfall close to densely populated areas" is of interest for an epidemiologist interested in cholera, since these are the conditions which help the Vibrio Cholerae bacteria to proliferate [1].

Based on the above, in this paper the problem of Keyword-Based Search over Environmental Data Sources is described. Besides, based on some assumptions that simplify the above problem, a first prototype of a search engine for environmental data was designed, constructed and evaluated. The prototype has the typical architecture composed of a Data Crawler, an Index Structure and a Search Engine. Besides, as a demonstrator of the search technology, a simple web map based graphical user interface was also developed, which enables the specification of search expressions and the navigation through spatial and temporal dimensions of the retrieved data. The construction of this prototype helped in the identification of research challenges that define interesting directions for future work.

The remainder of this paper is organized as follows. Section 2 provides a deeper description of the problem of searching environmental data sources with keywords. Related work is discussed in Sect. 3. Section 4 provides a brief description of the developed prototype. Performance evaluation is addressed in Sect. 5 and Sect. 6 concludes the paper and outlines potential future work directions.

[1] https://books.google.com/
https://www.hathitrust.org/.

[2] http://www.opengeospatial.org/standards/cat.

[3] http://www.unidata.ucar.edu/software/thredds/current/tds/catalog/InvCatalogSpec.html.

2 Problem Statement

2.1 Environmental Data Sources

Broadly speaking, environmental data is generated by some either observation (Observation Data) or modelling (Modelling Data) process. In any case, as it is defined by the Observations and Measurements (O&M) standard specification of the Open Geospatial Consortium (OGC)[4], values of some property (*Observed Property*) of some entity of interest (*Feature of Interest - FOI*) are generated by some process (*Observation Process*) for different time instants (Phenomenon Time). Depending on the nature of the *Process* and *FOI* involved, the generated data may fit either an entity-based conventional data model or a scientific array-based data model. Thus, for example, Fig. 1(a) illustrates an entity-based dataset, which records Current-Temperature-Depth (CTD) profiles generated at various sampling stations of the coast of Galicia (NW Spain) by the regional oceanographic agency INTECMAR[5]. At each sampling campaign (weekly approximately), a ship travels through all the sampling stations (*FOIs*), and uses a CTD device (*Process*) to generate values for different properties (water temperature is illustrated in the figure) at different depths. On the other hand, Fig. 1(b) illustrates an array-based dataset, which records a 3D (2 spatial and 1 temporal dimensions) array of Sea Surface Temperature (SST) data generated by the National Oceanographic and Atmospheric Agency (NOAA) of the US government. Daily observation data obtained from sensors on board of different types of platforms (satellites, ships, buoys) is interpolated using the Optimum Interpolation method to generate a daily grid of 0.25° of spatial resolution.

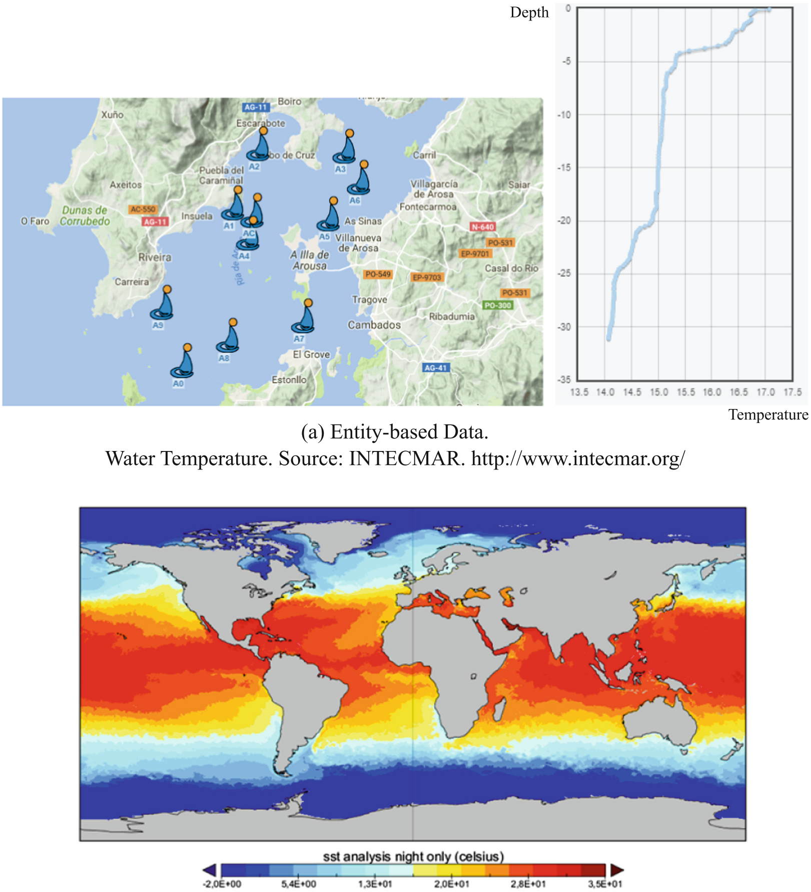
It has to be remarked that, although environmental data is mainly numeric data, keywords are also present in both data and metadata. In particular, the names of entity types ("Sampling station", "Municipaliy", "Road", etc.), properties ("Sea Surface Temperature", "Elevation", "Rainfall", "Population density", etc.) and processes ("CTD", "MODIS", "Optimum Interpolation", etc.) may be used by a keyword-based search engine to locate relevant data. Besides, text properties of entities (most commonly names and descriptions) also contain keywords that may be used in queries.

2.2 Description of the Search Problem

Roughly speaking, the problem consists in the evaluation of a set of search conditions expressed with keywords to determine which are the relevant data sources that fulfill the condition and which are the regions of space and the periods of time which are relevant inside each data source. Contrary to conventional catalogues, the result is not just a list of data sources, but also a ranking of relevant regions and time periods, which should be used to download relevant data from each data source. In general, the following types of search restrictions, expressed with keywords, should be supported by a search engine.

[4] http://www.opengeospatial.org/standards/om.

[5] http://www.intecmar.gal/.



(a) Entity-based Data.
Water Temperature. Source: INTECMAR. http://www.intecmar.org/

(b) Array-based Data.
Sea Surface Temperature. Source: NOAA OISST. http://www.ncdc.noaa.gov/oisst

Fig. 1. Illustration of environmental data sources.

Spatial restrictions. They determine the areas of geographic space of interest for the user. They combine keywords of names of entity types and/or entities with keywords that represent spatial relationships, either topological, distance or directional. Examples of these restrictions are the following: “in Madrid”, “in a Hotel”, “close to Hilton”, “far from sampling station”, “near the coast” and “north of Santiago”.

Temporal restrictions. They determine the periods of time of interest for the user. They combine keywords of names of entity types and/or entities with keywords that represent temporal relationships. Examples of these restrictions are the following: “during a storm”, “just before Katrina’ and “during monsoon”.

Data restrictions. They determine the values of data properties which are of interest for the user. They combine keywords of names of data properties with linguistic values, defined as terms which express value constraints for the properties. Examples of these restrictions are the following: "high sea surface temperature", "low rainfall" and "normal wind speed".

Additionally, a search engine might also support restrictions on other dimensions of the data, such as data provenance (data generated by a specific modelling software or by a concrete type of sensor) and data quality.

3 Related Work

Keyword-based search technologies are at the basis of modern search engines for digital libraries, document management and the web. Related to them, much work has been undertaken during the last decades in the area of Information Retrieval (IR) [7]. In this area, faceted search may combine text with available document metadata to improve the effectiveness. Regarding structured data sources, Full Text Search (FTS) functionality is included in query languages like SQL[6] and XQuery[7]. Limitations of these technologies include [5,14]: (i) only simple text search with limited scoring is supported, (ii) the keyword-based search functionality, supported by abstract data types is not really integrated into the DBMS query optimizer and (iii) the user must have structured query language skills and must know the database schema. To solve the first limitation more advanced ranking approaches have been proposed, including probabilistic [4] and authority-based [2] rankings. For the second limitation [5] proposes an architecture where keyword-based search functionality is supported in a layer on top of a relational storage engine. To cope with the last limitation, the user should submit a set of keywords and the system should response with a ranked list of interconnected tuples. To achieve this two main approaches have been identified in [14], namely Schema Based and Graph Based approaches. In the former a ranked list of relational expressions is generated from the keywords, which are next evaluated to obtain the expected result. In the latter, the database is materialized into a graph and graph algorithms are used to obtain the result.

Spatial Data Management (SDM) is a topic to which much research efforts have been devoted during the last 30 years. Many solutions have been developed for spatial data modelling, efficient spatial data storage and access and spatial query processing and optimization. In spite of this, research challenges still exist specific sessions are still present in the main Data Management conferences like VLDB and SIGMOD. Main application areas of SDM are Geographic Information Systems (GIS) and Environmental Data Management. Related to the present work are the advances in the so called Geo-spatial Semantic Web [8,11] and in Geographic Information Retrieval (GIR) [9]. More specifically related is

[6] https://www.iso.org/standard/31368.html.

[7] https://www.w3.org/TR/xpath-full-text-10/.

the semantic integration of data sources [12,13] and spatio-textual indexing ([6] provides a nice tutorial).

Finally, regarding the discovery of geo-spatial and environmental data sources, standards have been proposed that define metadata schemes[8] and catalogue service interfaces[9]. Despite of all the above efforts, it does not exist an effective technology for keyword-based search over geo-spatial environmental data sources, which enables both the discovery of relevant data sources and the effective and efficient search of their contents to determine relevant spatial zones and time periods for subsequent data download.

4 Prototype Description

In this section we describe the most relevant components of our prototype KEYWORDTERM [3], for keyword-based search over environmental data sources. Two types of data sources are considered in the current implementation of KEYWORDTERM: (i) Entity-based data sources whose data is accessible through the Open Geospatial Consortium (OGC) standard Web Feature Service (WFS)[10] interface and (ii) Raster Array data sources whose data is accessible through the NetCDF Subset[11] Web Service interface defined by Unidata. Future implementation of the system should support other well-known data access web service interfaces. Beyond data access, both data sources are required to implement also the OGC Web Map Service (WMS) standard interface, to ease data visualization tasks. In the current implementation entity-based data sources were simulated using Geoserver[12], whereas Unidata's Thematic Real-time Environmental Distributed Data Service (THREDDS)[13] was used to simulate raster data sources. It is assumed that the coordinates of all the data sources use the same projected coordinate reference system.

A *Crawler* periodically scans the data sources and updates a Spatio-temporal-textual index structure, which is next used during the search process. To ease this process, the existence of a catalog with an already harmonized vocabulary is assumed. Therefore, both data source discovery and semantic data integration and fusion are tasks, which should be undertaken in a real scenario, lie out of the scope of the current prototype.

The *Index Structure* has two parts. A first part provides access methods to evaluate data restrictions, i.e., restriction over the values of the scanned properties. Relevant structures are illustrated in Fig. 2(a). The property name vocabulary is shown at the left side of the figure. Each property name has a set of fuzzy

[8] https://www.iso.org/standard/53798.html.

[9] http://www.opengeospatial.org/standards/cat. http://www.unidata.ucar.edu/software/thredds/current/tds/catalog/InvCatalogSpec.html.

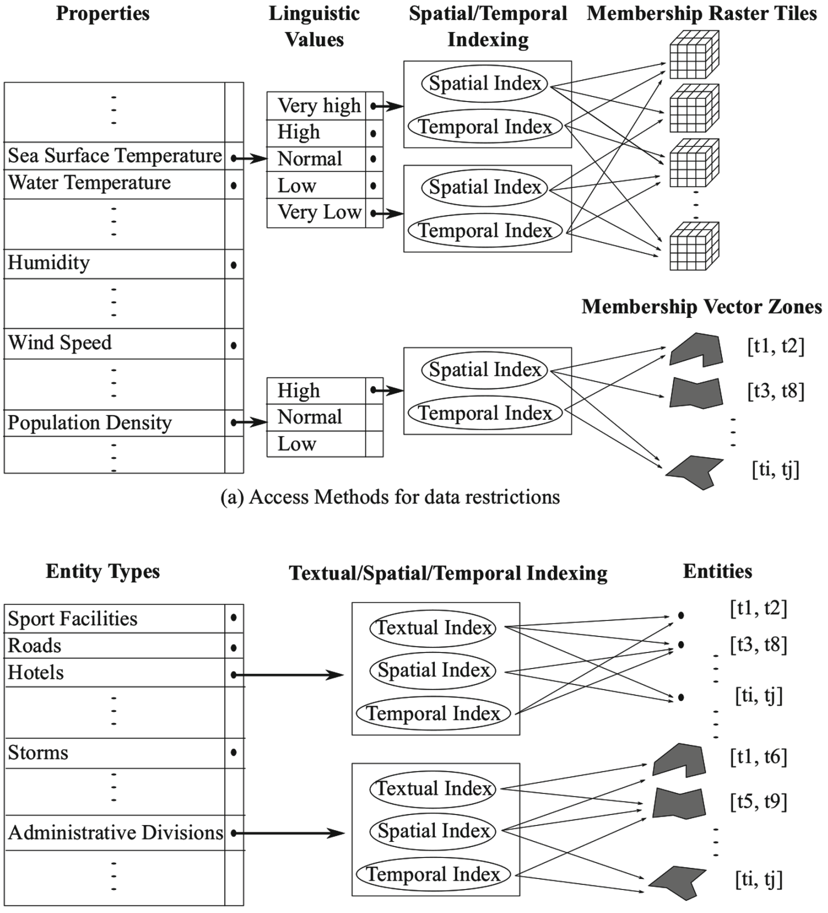
[10] http://www.opengeospatial.org/standards/wfs.

[11] http://www.unidata.ucar.edu/software/thredds/current/tds/reference/NetcdfSubsetServiceReference.html.

[12] http://geoserver.org/.

[13] http://www.unidata.ucar.edu/software/thredds/current/tds/.

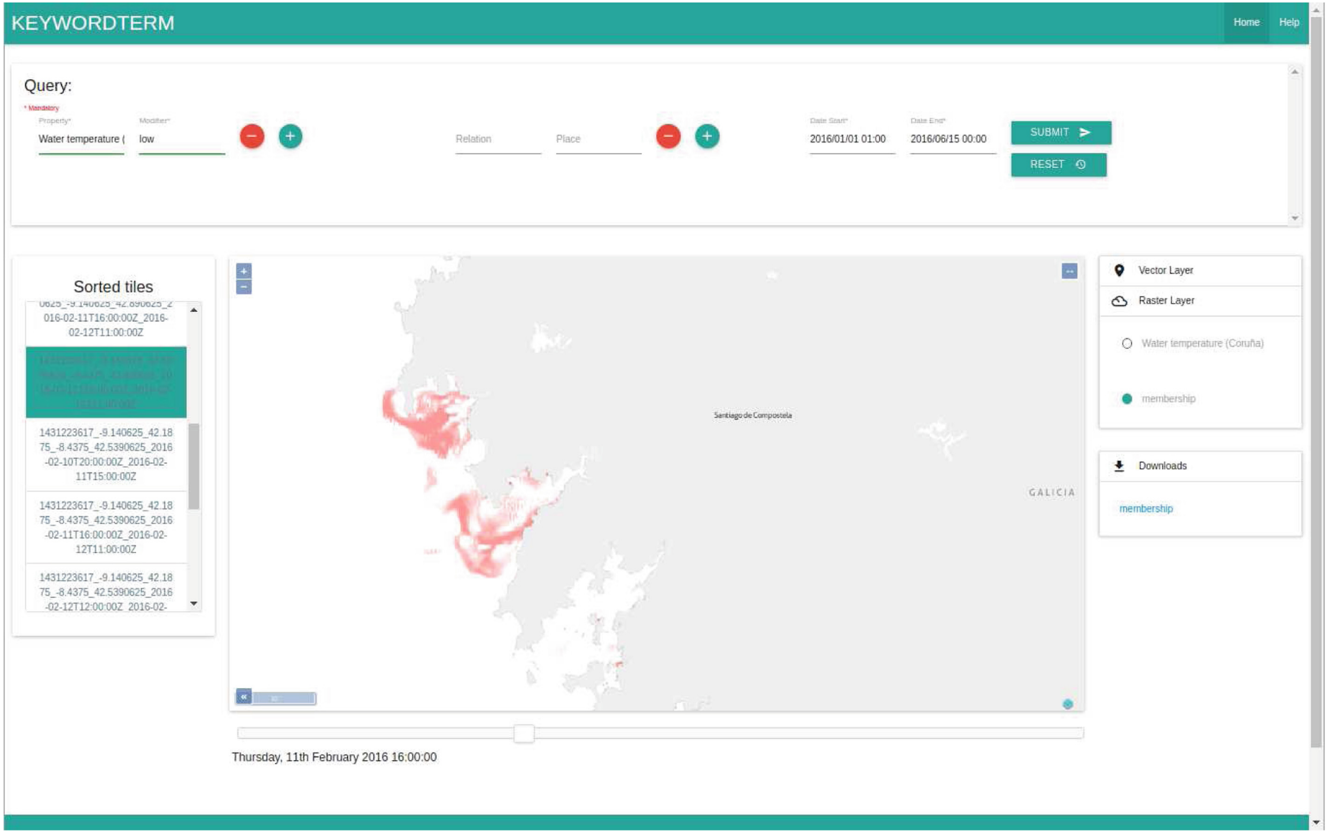
linguistic values, whose semantics are endowed by means of fuzzy sets [15]. At the current prototype it is assumed that those fuzzy linguistic values are defined by some expert, however, in a real scenario such expert knowledge should be combined with knowledge extracted from the data sources to generate those values. For each linguistic value, spatial and temporal indexes provide access to either membership raster tiles (for raster array properties) or membership spatio-temporal vector zones (for entity-based properties). The decomposition of the spatial dimension into tiles follows a regular Quadtree-based subdivision of space, whereas temporal semantics were considered for the temporal dimension. The second part of the structure provides methods to evaluate spatial and temporal restrictions. The implemented structure is illustrated in Fig. 2(b). The root of the structure is the Entity Type vocabulary. The entities of each type are



(a) Access Methods for data restrictions

(b) Access Methods for spatial and temporal restrictions

Fig. 2. Illustration of textual spatial and temporal access methods.



KEYWORDTERM

Query:

* Mandatory

Property*: Water temperature (

Modifier*: low

Relation

Place

Date Start*: 2016/01/01 01:00

Date End*: 2016/06/15 00:00

Sorted tiles

0625_-9.140625_42.890625_2016-02-11T16:00:00Z_2016-02-12T11:00:00Z

1431223617_-9.140625_42.1875_-8.4375_42.5390625_2016-02-10T20:00:00Z_2016-02-11T15:00:00Z

1431223617_-9.140625_42.1875_-8.4375_42.5390625_2016-02-11T16:00:00Z_2016-02-12T11:00:00Z

1431223617_-9.140625_42.1875_-8.4375_42.5390625_2016-02-12T12:00:00Z_2016-02-

Thursday, 11th February 2016 16:00:00

○ Water temperature (Coruña)

● membership

Fig. 3. Screenshot of the web GUI, showing the output (in red) of the query “Low Water Temperature”. The darker the red colour the higher the fulfilment of the search condition for each point of the map. (Color figure online)

indexed by textual, spatial and temporal indexes. Both the spatial and temporal extent of each entity is recorded. Currently, the PostgreSQL+PostGIS spatial DBMS is used to implement both the index and the storage of entities. The raster array tiles are recorded with the scientific array DBMS SciDB.

The *Search Engine* is made up of three subcomponents. A concept discovery component is used by clients to discover the names of *Properties* and *Entity Types* that are available in the index. A search component enables the evaluation of combinations of data, spatial and temporal restrictions, based on the available *Properties*, *Entity Types* and *Entities*. The result set of spatio-temporal tiles, zones and entities may be visualized by clients through a WMS service. The query language has been limited to a conjunction of a three optional restrictions: (i) data restrictions expressed by conjunctions of expressions of the form "*LinguisticValue PropertyName*", such as "*High Temperature and Low Salinity*". (ii) Spatial restrictions expressed by conjunctions of expression of the form "*SpatialRelationshipName SpatialName*", where *SpatialName* is either the name of a *Spatial Entity Type* (for example "Near Hotel") or the name of a *Spatial Entity* (for example "Within Hilton"). These restriction are still not implemented in the current version. (iii) Temporal restrictions expressed by either a time instant or a time interval (notice that keywords are not enabled in the current implementation to express temporal restrictions, but they will be in future versions).

A web based Graphical User Interface (GUI) was developed to demonstrate the functionality of the search engine and to analyse issues related to the interaction with the user, including the specification of the search restrictions and the navigation through the result. A screenshot of this interface is provided in Fig. 3 in landscape orientation. At the top of the interface, three control sets enable the specification of data, spatial and temporal restrictions. The result ranking of spatio-temporal tiles (it is reminded that spatial restrictions are still not implemented) is listed below at the left side of the interface. Currently each tile is depicted with spatial and temporal coordinates, however a better graphical representation of the spatial and temporal context of each tile is currently being implemented. At the centre of the screen a map shows the membership values of the selected tile (low water temperatures are shown in the figure). The temporal dimension of the current tile may also be navigated with the time slider at the bottom of the interface. At the right of the interface the user may chose to view in the map either the result membership or the input data. A download link enables the user to get the data of the current tile from the data source.

5 Performance Evaluation

A preliminary evaluation of the performance of the prototype was done. Although the evaluation is not exhaustive it already enables to derive some important conclusions. The index structure uses tiles that record 20 time instants and 180 × 360 spatial cells (10.4 MB aprox). Two datasets of different sizes (441 and 1928 Tiles) were created by crawling environmental raster data from the region of Galicia (NW Spain). The prototype was installed in a virtual machine

Table 1. Index access and membership calculation times in seconds.

Search expression (DB size)	Accessed tiles	Result tiles	Index	Membership
Low elevation (1928 tiles)	33	33	0.036	11.087
Low elevation (441 tiles)	33	33	0.003	8.526
Low air temperature and Normal wind speed (441 tiles)	6/11	6	0.021	480.086

with 6 cores (3.60 GHz) and 6 GB of RAM, running in a machine with a total of 8 cores and 8 GB of RAM.

Table 1 shows for three different search expressions over the two databases: (i) the number of membership tiles that are accessed through the index, (ii) the number of membership tiles of the result, (iii) the time (seconds) needed to navigate the index and (iv) the time (seconds) required to access and process the membership values. A first important issue is that the index access time is very small compared to the membership calculation for theses database sizes. However, it is important to note that increasing the database size around 4.5 times caused an increase of the index time of around 12 times, whereas the membership calculation was increased less than 1.5 times. It is expected therefore the a good indexing technique is a key issue to use very large datasets. Another important issue to note is the high cost of the computation of membership intersections to evaluate the last search expression. A parallel implementation of theses calculations is the only feasible approach. To complete this section, Fig. 4 shows the searching time with respect to the number of tiles selected by the index for queries involving just one property. It is obvious that search selectivity has great impact in the response time, however it is important to remind that high selectivity will, in general, come from the intersection of various restrictions, and the cost of computing those intersections is very high.

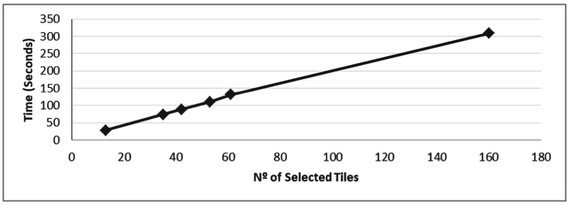


Fig. 4. Response time with respect to query selectivity.

6 Conclusions and Future Work

A first proof of concept prototype of a search engine for environmental data sources was designed and implemented. The functionality and performance of the tool is still very limited, however it has served as a starting point to get a more clear idea of the problem of keyword-based search over structured environmental dataset and to identify various research challenges that may guide future research efforts in this topic. Such issues of potential future work include the following.

- The semantic integration and fusion of different data sources for the same property or entity type. Such data integration must take into account data quality and provenance, but it might also consider other factors such as the overall reliability or authority of each data source.
- In the current prototype it is assumed that fuzzy linguistic values are provided by some expert. However, in a real scenario, the membership at each location and time should be calculated taking into account expert knowledge, knowledge extracted from the data source and also knowledge extracted from each user feedback and profile.
- The current search language is a compromise between the simplicity of list of keywords and the complexity of a full fledged structured query language. More investigation should be undertaken in this topic to find the best compromise between simplicity and expressiveness, considering the uncertainty in the definition of terms and operators involved in the queries.
- The ranking of the result tiles is currently done by aggregating membership values. In an scenario with more than one data provider for each Property and Entity Type, more sophisticated algorithms could be devised both to produce membership values for each location and time and to rank tiles and data sources based on those values.
- It was shown in Sect. 5 that the performance of the current implementation is far from being acceptable. To solve this, better spatio-temporal-textual indexing structures should be devised and distributed implementations of both crawling and searching should be undertaken, together with a rigorous evaluation of the efficiency of the implementations.

References

1. Baker-Austin, C., Trinanes, J.A., Taylor, N.G., Hartnell, R., Siitonen, A., Martinez-Urtaza, J.: Emerging vibrio risk at high latitudes in response to ocean warming. Nature Clim. Change **3**(1), 73–77 (2013)
2. Balmin, A., Hristidis, V., Papakonstantinou, Y.: ObjectRank: authority-based keyword search in databases. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, pp. 564–575. VLDB Endowment (2004)
3. Álvarez-Castro, D., Viqueira, J.R., Bugarín, A.: Aproximación a la búsqueda basada en términos sobre conjuntos de datos medioambientales. In: Molina, J.J.G. (ed.) XXI Jornadas de Ingeniería del Software y Bases de Datos, V Congreso Español de Informática, pp. 381–384. Ediciones Universidad de Salamanca (2016)
4. Chaudhuri, S., Das, G., Hristidis, V., Weikum, G.: Probabilistic information retrieval approach for ranking of database query results. ACM Trans. Database Syst. **31**(3), 1134–1168 (2006)
5. Chaudhuri, S., Ramakrishnan, R., Weikum, G.: Integrating DB and IR technologies: what is the sound of one hand clapping? In: Second Biennial Conference on Innovative Data Systems Research (CIDIR 2005), Asilomar, CA, USA, pp. 1–12, 4–7 January 2005

6. Chen, L., Cong, G., Jensen, C.S., Wu, D.: Spatial keyword query processing: an experimental evaluation. In: Proceedings of the 39th international conference on Very Large Data Bases, PVLDB 2013, pp. 217–228. VLDB Endowment (2013)
7. Croft, W., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Pearson, Upper Saddle River (2010)
8. Janowicz, K., Scheider, S., Pehle, T., Hart, G.: Geospatial semantics and linked spatiotemporal data - past, present, and future. Semant. Web **3**(4), 321–332 (2012)
9. Jones, C.B., Purves, R.S.: Geographical information retrieval. Int. J. Geogr. Inf. Sci. **22**(3), 219–228 (2008)
10. Lowen, A.C., Mubareka, S., Steel, J., Palese, P.: Influenza virus transmission is dependent on relative humidity and temperature. PLOS Pathog. **3**(10), 1–7 (2007). https://doi.org/10.1371/journal.ppat.0030151
11. Patroumpas, K., Giannopoulos, G., Athanasiou, S.: Towards geospatial semantic data management: strengths, weaknesses, and challenges ahead. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 301–310. ACM, New York (2014)
12. Regueiro, M.A., Viqueira, J.R., Stasch, C., Taboada, J.A.: Semantic mediation of observation datasets through sensor observation services. Future Gener. Comput. Syst. **67**, 47–56 (2017)
13. Vilches-Blázquez, L.M., Villazón-Terrazas, B., Corcho, O., Gómez-Pérez, A.: Integrating geographical information in the linked digital earth. Int. J. Digit. Earth **7**(7), 554–575 (2014)
14. Yu, J.X., Qin, L., Chang, L.: Keyword search in relational databases: a survey. IEEE Data Eng. Bull. **33**(1), 67–78 (2010)
15. Zadeh, L.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)