Geolocalized Filtering of Open Data Datasets for Mobile Devices

Riccardo Paolillo and Michele Orru'
Able Srl
Rome, Italy
firstname.lastname@able-srl.it

***Abstract*—Open data are has an high economic potential, allowing a transparent and democratized access to data. At the same time, geographic data are playing a fundamental role since it can easily filter pertinent data among the wide multitude of available data, such as finding the nearest wifi networks or the nearest museums. In this work we present a solution to exploit geolocalization on mobile devices in order to find, among the available Open Data, the most relevant data according to the position of the user. Starting from some public datasets, an initial data augmentation has been introduced to include location information whereas it was possible and useful. Then an optimized filtering system was developed for geographic contents through the design of a specific back-end.**

***Index Terms*—open-data, geographic information, LBS**

I. Introduction

In the last decade, the Open Government doctrine has been spreading more and more founding its cultural model on the principle according to which all the activities of governments and state administrations must be open and available in order to promote effective actions and guarantee widespread control over the management of public affairs. This movement pushes a new interaction model between citizen and Public Administration: from the "classic" user to a citizen that is directly involved in the government choices [1]. Particular impetus to this movement was given by the Obama administration, which in 2009 promulgated the so-called "Open Government Delegate" Memorandum, [2], a provision that codifies the principles of the "open" philosophy within institutions and administrations, prescribes tasks, processes and organizational models that public bodies are called to follow in compliance with the Directive, defining three essential keywords: Transparency, Participation and Collaboration.

Transparency: institutions are required to provide citizens with data and information on decisions taken and on their actions, in order to create a system of trust within the local community towards the work and choices made. *Participation*: citizen participation in the Public Administration choices increases the effectiveness of administrative actions and improves the quality of decisions. *Collaboration*: the collaboration envisage a direct involvement of citizens in the activities of the Public Administration and tends to include the institutions within a collaborative and participatory network composed of public bodies, non-profit organizations and a community of citizens.

To implement these principles it is therefore necessary that the P.A. make the widest possible amount of data available to the public available to it: such data, if released in specific ways, are called open data ("Open Data") [3]. So considerable amounts of data of all kinds are made available to the citizen, ready to be used in the most disparate ways.

A. Open Data

Open Data are data collections (datasets), publicly accessible, without patents or proprietary licenses that limit their diffusion or re-use. Open Data is a specific type of *Open content* that can be considered its father focused on the spread of creative works [3]. According to supporters of the Open Data movement, the data should be treated as common goods because:

- the data belongs to the human race
- the data produced by the public administration, as paid with public money, must return to the tax payers in the form of open and universally available data
- any restrictions on data and their use represent a limitation of the community's development potential
- the data is necessary to facilitate the execution of common human activities
- better is access to data, greater is the rate of discovery in the scientific field,

B. Open Data Quality

To distinguish the different formats that can be used in the coding of datasets, W3C has proposed a cataloging model that classifies them based on their characteristics on a scale of values from 1 (one star) to 5 (five stars) [4]:

* It is the basic level, consisting of unstructured files: for example a document in Microsoft Word format, a file in Adobe PDF format. A single star indicates the simple availability of information and data online, in any form, as long as it is distributed with an open license. The data distributed in this format is readable and printable by users, can be stored locally on a PC and is easy to publish. However they are not in open format and no processing is possible on them.
** This level indicates structured data but encoded with a proprietary format, for example a document in Microsoft Excel format. The data characterized by the two stars are not in open format as a proprietary software is needed to process them, however they can normally be converted - being structured data - into open data;
*** This level indicates structured data encoded in a non-proprietary format, for example the format .csv (Comma Separated Values), data that can be manipulated without having to use proprietary software;
**** This level indicates structured data encoded in a non-proprietary format, which are equipped with a URI2 that makes them addressable on the network and therefore usable directly online, through inclusion in a structure based on the RDF3 model. Four stars therefore indicate the fact that the single data of a dataset, available online in an open format (typically XML / RDF) can be invoked through a specific URL. This allows you to point to the data or a set of data from an application or access it from within a program that can then process it in various ways.
***** This level indicates those that are referred to as Linked Open Data (LOD). Those open data, that is, that in addition to responding to the characteristics indicated in the previous point also present, in the structure of the dataset, links to other datasets. The Linked Open Data therefore allows to combine the contents of different datasets thanks to formal constructs formulated according to the RDF model. This exponentially increases the value of mutually correlated datasets, allowing the transition from the data level to the information level and therefore to the knowledge level and thus providing a structured context framework starting from the correlation of information from different sources.

As can be easily understood, the latest levels indicate rather high Open Data quality, as these data can be easily integrated.

C. Open Data Position

A problem for the effective explotation of open data is that there are often too many and the downloading and processing in particular on mobile terminals is really expensive. Thus in this work we propose a system architecture to associate positions with the linked open data in order to limit the transfer of information to mobile devices, and increase the effectiveness of the information published by the various administrations

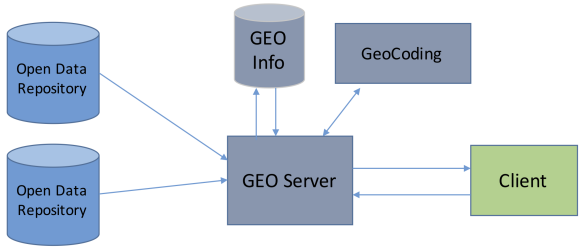

Fig. 1. Geolocalized OD system model

This paper we present a solution to ”exploit” this data in the most useful way for the citizen, introducing the use of geolocalization on mobile devices in order to find, among the available Open Data, those more interesting or those located nearby, such as a museum, a library or a Wifi access point. Starting from some public datasets, a filtering system was then developed for these contents through a specific back-end. Particular attention was paid to the way in which the queries on these datasets are made, in order to improve their efficiency and execution time.

II. System Model

The proposed system architecture is depicted in figure 1. The system cose is the GEO Server that associte the location data elaborating the dataset information. The

It includes a ”client side” that periodically takes care of operating the data augmentation downloadning the files from the open data providers and insertig it in the local database; The”server side” expose the new augmented infromation layer to the clients that can make geolocalized queries on the open database. Each dataset will have its own table - dynamically created based on predefined xml layouts present and each row of these tables contains the latitude and longitude field.

A. Google Geocoding

Severl Open Dataset provide the location information in an implict form. Thus to inlcude this data in the augemnted dataset, a geocoding service is used offered. A popular ”free” service is offered by Google: it converts a specific address into geographical coordinates (and vice versa, called reverse Geocoding). Request and response are made on the HTTP protocol and the output is produced in JSON.

A requested example is:

```
https://maps.googleapis.com/maps/api/
geocode/json?address=1600+Amphitheatre
```

```
+Parkway,+Mountain+View,+CA
```

This request produces a complex JSON output that can be parsed to acquire the goegraphical coordinates of the data.

B. GEO Sever

Since the GEO Server stores a copy of open datasets augmented with the geographical coordinates, it is possible to query the server to receive the open data filtered on their location (and that of mobile client).

Each request is serverd by an independent thread which operates according to the following life cycle:

- It process the HTTP request and extract: selected dataset, latitude, longitude, range, limit and offset.
- It perform the geolocated query on the database;
- It convert the query results into an JSON file;
- It set up an appropriate Response Header HTTP;
- It send the generated JSON file to the client;

III. Geolocated Query

In this section we describe the server query methodology used to filter in a efficient way all the reconds that are nearby the requesting client. The query result is the result-set of geolocated records sorted in ascending order based on the distance from a geographic point. We assume that the user provide in the query its goegraphical coordinates (latitude and longitude) and that each dataset is augmented with (lat,lon) information.

A. Distance calculation

Thus, to calculate the distance between two points on the earth's surface, we resort to the so-called haversine formula:

$$\mathcal{H}\left(\frac{d}{R}\right) = \mathcal{H}\left(\Delta\phi\right) + K\mathcal{H}\left(\Delta\lambda\right) \tag{1}$$

where d is the distance between two points, R is the terrestrial radius, $\Delta\phi$ is the difference between the latitudes of the two points (ϕ_1 and ϕ_2 respectively), $\Delta\lambda$ and the difference between the two longitudes (λ_1 and λ_2 respectively) and $K = cos(\phi_1)cos(\phi_2)$. The haversine corresponds to the half of the versine of θ angle, defined as:

$$\mathcal{H}\left(\theta\right) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2} \tag{2}$$

from which we can calculate the distance between two points on the earth's surface:

$$d = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + K\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{3}$$

B. SQL implementation

Starting from equation 3, we define the basic listing that operates the filterign orperation:

```
SELECT *, 2 * 6371 * ASIN(SQRT(POWER(
     SIN(((lat2-lat1) * pi()/180)/2),2)
   + COS(psi1 * pi()/180) * COS(psi2
   * pi()/180) * POWER(SIN(((long2-long1)
   * pi()/180)/2),2) ))
   as distance
   FROM dataset
   WHERE distance < range
   ORDER BY distance ASC
```

The problem with this query is that we need to compte the distance on every single record in the table to know whether it falls within the desired range or not, and this is too computational expensive. To significantly lower the timewe limit the distance calculation only to records that fall within a rectangular area on the earth shere. For this reason, the four vertices are defined according to the following listing:

```
SET @lat1 = lat - (range / 111.044736);
SET @lat2 = lat + (range / 111.044736);
SET @long1 = lng - (range /
  abs(cos(radians(lat) * 111.044736)));
SET @long2 = lng + (range /
  abs(cos(radians(lat) * 111.044736)));
SELECT *, 2 * 6371 * ASIN(SQRT(POWER(
    SIN(((lat-db_lat) * pi()/180)/2),2)
  + COS(lat * pi()/180) * COS(latitudine
  * pi()/180) * POWER(SIN(((lng-db_lon)
  * pi()/180)/2),2) ))
    as distance FROM tableName
    WHERE db_lat BETWEEN @lat1 and @lat2
    AND db_lon BETWEEN @long1 and @long2
    HAVING distance < range
    ORDER BY distance asc
```

Listing 1. Simple Query

C. Stored Procedure

To further optimize the queries to reduce their execution time we used a Stored Procedures (SP), i.e. a programs stored within the database, written in languages different (often derived from SQL) depending on the DBMS in use, which allow users to perform complex functions defined by the database administrator. SPs must not return values but can accept input and output parameters as well as generate Result Sets.

The defines SP for the geolocated query is:

```
CREATE PROCEDURE geoquery(IN lat DOUBLE,
      IN lng DOUBLE IN tableName VARCHAR(100),
      IN v_range INT, IN v_limit INT,
      IN v_offset INT)
BEGIN
  DECLARE lat1 FLOAT;
  DECLARE long1 FLOAT;
```

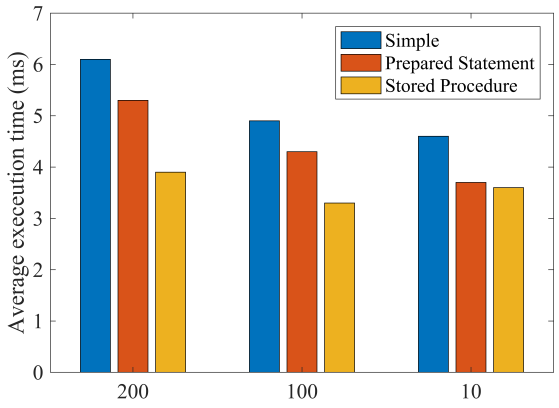

Fig. 2. Geolocalized Quesry Execution Delay

```
DECLARE lat2 FLOAT;
DECLARE long2 FLOAT;
SET lat1 = lat - (v_range / 111.044736);
SET lat2 = lat + (v_range / 111.044736);
SET long1 = lng - (v_range /
     abs(cos(radians(lat) * 111.044736)));
SET long2 = lng + (v_range /
     abs(cos(radians(lat) * 111.044736)));
SELECT *, 2 * 6371 * ASIN(SQRT(POWER(
     SIN(((lat-db_lat) * pi()/180)/2),2)
   + COS(lat * pi()/180) * COS(db_lat
   * pi()/180) * POWER(SIN(((lng-db_lon)
   * pi()/180)/2),2) )) as distance
        FROM tableName
        WHERE db_lat BETWEEN lat1 and lat2
        AND db_lon BETWEEN long1 and long2
     HAVING distance < v_range
     ORDER BY distance asc
     LIMIT v_limit
     OFFSET v_offset;
END;
```

Listing 2. Stored Procedure Query

IV. Performace Result

To evaluate the effectiveness of the proposed solution we performed 1000 requests to a database MySql with a variable number of stored Open DataSetsa and we report the average measured response time from the server. In the comparison we also include a stored procedure query that is executed by a prepared statement. It similar to the SP one but since it require a string concatenation its average response time si in the miggle of the other two queries.

Figure 2 reports the query execution time for the simple case, for the SP case and the SP case with prepared statement. It shows that the SP is faster query possible and that the execution time clearly increses with the number of records in the DB. Note that since each query operates on a table and that all the tables are independent, the proposed solution well scale in cloud architecture partitionag the dataset among a cluster of servers as in [5].

V. Related Work

Open data augmentation is an approach pursued by several works such as [6] and [7]. In [6] data are augment with information from crowd sourcing to assure transparency and service improvement. In [7] both time and location of the data are added with a scalable architecture.

Geographic data has been extensively used in research in the context of Location Based Services, Proximity services [8], IoT and monitoring devices [9] [10] [11].

Dedicated techniques [12] has been investigated to provide solution for creating RDF based open-data geographical resources and how this can be used for the semantic web.

Providing open-data with storage and mining solutions requires dedicated architecture and system design that involve the recent findings in the field of Big Data processing [13], sharding database solutions [14], NO-SQL database solutions, dedicated search engine [15].

High speed machine learning process [16], [17] [18] also supported by new dedicated hardware and methods [19] [20], allow to extract useful information from the stored data in business compliant times, also thanks to innovative solution to process geographic data [21]. With this regard, the use of Open Data represen is, in many cases, a valid training dataset to be used and elaborated through new emerging machine learning approach such as [22] [23] [24].

VI. Conclusion

We presented a system architecture able to provide the open data framewotk with an augmentation service that add the location information to the published data. This alows the mobile client to access only the useful dataset and thus to optmize the networking performance. We investigated the SQL solution for geolocated queires evaluating typical dealys that can be expected by the server. Future work can adapt the architecture to a NoSQL database.

References

[1] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Information systems management*, vol. 29, no. 4, pp. 258–268, 2012.

[2] T. M. Harrison, S. Guerrero, G. B. Burke, M. Cook, A. Cresswell, N. Helbig, J. Hrdinova, and T. Pardo, "Open government and e-government: Democratic challenges from a public value perspective," *Information Polity*, vol. 17, no. 2, pp. 83–97, 2012.

[3] F. A. Zeleti, A. Ojo, and E. Curry, "Exploring the economic value of open government data," *Government Information Quarterly*, vol. 33, no. 3, pp. 535–551, 2016.

[4] L. Van den Brink, P. Barnaghi, J. Tandy, G. Atemezing, R. Atkinson, B. Cochrane, Y. Fathy, R. García Castro, A. Haller, A. Harth *et al.*, "Best practices for publishing, retrieving, and using spatial data on the web," *Semantic Web*, no. Preprint, pp. 1–20, 2017.

[5] A. Detti, L. Bracciale, P. Loreti, G. Rossi, and N. B. Melazzi, "A cluster-based scalable router for information centric networks," *Computer networks*, vol. 142, pp. 24–32, 2018.

[6] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall, and m.c. schraefel, "Linked open government data: lessons from data.gov.uk," *IEEE Intelligent Systems*, vol. 27, no. 3, pp. 16–24, May 2012. [Online]. Available: https://eprints.soton.ac.uk/340564/

[7] S. Neumaier and A. Polleres, "Enabling spatio-temporal search in open data," *Available at SSRN 3304721*, 2018.

[8] P. Loreti, L. Bracciale, and A. Caponi, "Push attack: Binding virtual and real identities using mobile push notifications," *Future Internet*, vol. 10, no. 2, p. 13, 2018.

[9] L. Bracciale, A. Catini, G. Gentile, and P. Loreti, "Delay tolerant wireless sensor network for animal monitoring: The pink iguana case," in *International Conference on Applications in Electronics Pervading Industry, Environment and Society*. Springer, 2016, pp. 18–26.

[10] P. Loreti, A. Catini, M. De Luca, L. Bracciale, G. Gentile, and C. Di Natale, "The design of an energy harvesting wireless sensor node for tracking pink iguanas," *Sensors*, vol. 19, no. 5, p. 985, 2019.

[11] L. Bracciale, P. Loreti, A. Detti, R. Paolillo, and N. B. Melazzi, "Lightweight named object: an icn-based abstraction for iot device programming and management," *IEEE Internet of Things Journal*, 2019.

[12] J. Goodwin, C. Dolbear, and G. Hart, "Geographical linked data: The administrative geography of great britain on the semantic web," *Transactions in GIS*, vol. 12, pp. 19–30, 2008.

[13] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[14] A. Detti, M. Orru, R. Paolillo, G. Rossi, P. Loreti, L. Bracciale, and N. B. Melazzi, "Application of information centric networking to nosql databases: the spatio-temporal use case," in *2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. IEEE, 2017, pp. 1–6.

[15] C. Gormley and Z. Tong, *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. " O'Reilly Media, Inc.", 2015.

[16] G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, M. Re, and S. Spanó, "Awsom, an algorithm for high-speed learning in hardware self-organizing maps," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2019.

[17] G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Matta, M. Re, F. Silvestri, and S. Spanò, "Efficient ensemble machine learning implementation on fpga using partial reconfiguration," in *International Conference on Applications in Electronics Pervading Industry, Environment and Society*. Springer, 2018, pp. 253–259.

[18] M. Matta, G. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, F. Silvestri, and S. Spanò, "Q-rts: a real-time swarm intelligence based on multi-agent q-learning," *Electronics Letters*, vol. 55, no. 10, pp. 589–591, 2019.

[19] M. Salerno, G. Susi, and A. Cristini, "Accurate latency characterization for very large asynchronous spiking neural networks," in *International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2011)*. SciTePress, 2011, pp. 116–124.

[20] G. L. Susi, L. F. Antón-Toro, L. Canuet, M. E. López, F. Maestú, C. Mirasso, and E. Pereda, "A neuro-inspired system for online learning and recognition of parallel spike trains, based on spike latency and heterosynaptic stdp," *Frontiers in neuroscience*, vol. 12, p. 780, 2018.

[21] G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, A. Nannarelli, M. Re, and S. Spanò, "N-dimensional approximation of euclidean distance," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2019.

[22] G. Susi, A. Cristini, and M. Salerno, "Path multimodality in a feedforward snn module, using lif with latency model," *Neural Network World*, vol. 26, no. 4, p. 363, 2016.

[23] C. Napoli, G. Pappalardo, G. M. Tina, and E. Tramontana, "Cooperative strategy for optimal management of smart grids by wavelet rnns and cloud computing," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 8, pp. 1672–1685, 2015.

[24] G. Capizzi, G. L. Sciuto, P. Monforte, and C. Napoli, "Cascade feed forward neural network-based model for air pollutants evaluation of single monitoring stations in urban areas," *International Journal of Electronics and Telecommunications*, vol. 61, no. 4, pp. 327–332, 2015.