Semantically Enriching VGI in Support of Implicit Feedback Analysis

Andrea Ballatore and Michela Bertolotto

School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
{andrea.ballatore,michela.bertolotto}@ucd.ie

Abstract. In recent years, the proliferation of Volunteered Geographic Information (VGI) has enabled many Internet users to contribute to the construction of rich and increasingly complex spatial datasets. This growth of geo-referenced information and the often loose semantic structure of such data have resulted in spatial information overload. For this reason, a semantic gap has emerged between unstructured geo-spatial datasets and high-level ontological concepts. Filling this semantic gap can help reduce spatial information overload, therefore facilitating both user interactions and the analysis of such interaction. Implicit Feedback analysis is the focus of our work. In this paper we address this problem by proposing a system that executes spatial discovery queries. Our system combines a semantically-rich and spatially-poor ontology (DBpedia) with a spatially-rich and semantically-poor VGI dataset (OpenStreetMap). This technique differs from existing ones, such as the aggregated dataset LinkedGeoData, as it is focused on user interest analysis and takes map scale into account. System architecture, functionality and preliminary results gathered about the system performance are discussed.

Keywords: Geographic Information Systems, Geo-Ontologies, DBpedia, LinkedGeoData, OpenStreetMap, Volunteered Geographic Information, Implicit Feedback Analysis

1 Introduction

In recent years several spatial datasets have been published and have become available through open-source licenses. Crowd sourcing has played a major role in this process, enabling many Internet users to collaborate in the construction of rich and increasingly complex units of reusable geographical knowledge. As Volunteered Geographic Information (VGI) proliferates, the need for reduction of information overload and for semantic structure has become prominent. Since the late 90s, the lack of semantic structure on the Internet has become problematic for several applications and has promoted several initiatives in the context of the so-called *Semantic Web* [3]. This problem is very evident in the spatial domain and even more in VGI repositories where finding a meaningful structure of associations between geographic entities is far from straightforward.

We believe that the semantic gap between rather unstructured geographical data and higher level spatial and non-spatial concepts is one of the key challenges for reducing information overload in modern Geographical Information Systems. Our previous work has tried to address spatial information overload by generating personalised maps that match user interests and facilitate their tasks [20]. In this context, filling the semantic gap would help identify relevant information and generating meaningful user profiles.

Geo-ontologies have proven useful in the attempt to semantically enrich spatial datasets. Although existing VGI projects offer rich and complex datasets, because of the semantic gap, it is not easy to combine them with such geo-ontologies in the reduction of information overload. For this reason, we propose an approach for exploiting ontologies to perform spatial exploratory queries in the context of spatial knowledge and data discovery. Within this context, we have developed a system that integrates a spatially rich but semantically poor vector dataset (OpenStreetMap) with a spatially poor but semantically rich ontology (DBpedia), partially exploiting the mapping offered by an aggregated dataset (LinkedGeoData). When the user clicks on a geo-location on the map, the system processes a spatial query extracting spatial features from the vector dataset. These features are then mapped to ontological nodes, showing semantic contents to the user.

This system was developed in support of our implicit user profiling efforts that rely on the analysis of mouse movements on an interactive map to gather information about the user spatial interests. Compared to existing projects such as LinkedGeoData, our system is specifically designed to enhance semantic content in the field of implicit feedback analysis. In order to emphasise the user perspective in the interaction with spatial data, the map scale is taken into account during the extraction of ontological concepts, selecting features that are visible on the map at a given scale.

This is a new contribution as to the best of our knowledge no other similar project takes this aspect into account. Therefore we aim at overcoming this limitation for our purposes.

Indeed, when a user interacts with a Web map, the extended semantic knowledge extracted by such a system can be used to further refine the insight into the user's spatial interests. In this paper we show that our approach to extracting explicit semantic relationships between vector geographic data and ontological entities can benefit spatial user profiling through unobtrusive implicit feedback indicators, such as mouse movements.

The remainder of this paper is organised as follows: Section 2 discusses related work in the area of geo-spatial ontologies, volunteered geographic information and geographical crowdsourced information. Section 3 outlines the system architecture including a detailed description of the technologies and web services used. Section 4 describes the system functionality, while Section 5 details the system Web graphical user interface and describes a preliminary evaluation. Finally, Section 6 draws a conclusion, identifying limitations and directions for future work.

2 Related work

Finding meaning in unstructured - or loosely structured - online information has been the main challenged addressed by the set of initiatives under the name *Semantic Web* at the beginning of the millennium [3]. After a decade, ontologies are still considered one of the key elements of the Semantic Web and are technologically supported by languages such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Given that there is no universally accepted definition, in this work we indicate as an *ontology* a document that formally defines the relations among terms. Ding et al. provide a detailed survey of the field of ontologies for the Semantic Web [7].

Personalisation is one of the fields in which ontologies have been utilised to build user profiles in location-based services [27, 12], to develop context-aware mobile systems [17] and to refine Web searches through ontological user profiles [23]. Although ontologies have a long and well-established tradition, little work has been done in the area of spatial implicit feedback analysis.

The last decade has witnessed another major Internet phenomenon, sometimes called *neogeography* in the literature [25]. During this transition, the traditional GIS field has progressively become Web-based and universally accessible through Web technologies, a process described by Haklay et al. [13]. Once collaborative tools reached a certain maturity, online efforts have started, resulting in several high-impact GIS-related projects, following the crowdsourcing phenomenon. The mainly non-spatial project Wikipedia still represents the most visible product of these processes. This transformation of users from passive end points of GIS geographical services to active contributors has been defined *Volunteered Geographic Information* [11].

As the importance of crowdsourcing and Web GIS grows steadily, the semantic gap has become increasingly problematic and started attracting further initiatives. Spatial ontologies have traditionally helped geographical knowledge to be efficiently stored, processed and shared among institutions. Fonseca et al. have conducted influential work in this area [9, 10]. Several projects have sprung in the area of crowdsourcing and specifically in VGI. In this work we chose OpenStreetMap[1] as the data repository for bottom-up user-generated geographical vector data. Being one of the foremost VGI projects, OpenStreetMap has attracted a lot of contributors and researchers [14]. The quality of the OpenStreetMap spatial data and the VGI production mode are object of controversy, similarly to Wiki debate [8]. The project constitutes a typical example of a vast and fast-growing dataset built by users on an limited underlying semantic structure [4, p.383]. OpenStreetMap does not impose a formal semantic structure but rather suggests a 'core recommended feature set and corresponding tags'[2].

In the non-spatial domain, Wikipedia[3] is indeed the most impressive user-generated dataset available. As Völkel et al. pointed out, Wikipedia pages are

[1] http://www.openstreetmap.org

[2] http://wiki.openstreetmap.org/wiki/Map_Features

[3] http://www.wikipedia.org

very useful but they are not easy to be machine-processed [26]. Diverse efforts have been made to bridge this semantic gap over the past few years and progressively converged into DBpedia[4], a unified dataset offering a semantically structured version of Wikipedia [1, 6]. In order to structure and classify the pages, DBpedia relies on manually created cross-domain ontology, based on the most commonly used infoboxes within Wikipedia. From a spatial viewpoint DBpedia contains geo-coordinates for 390,000+ geographic locations but not the complex polygons and polylines that are available in OpenStreetMap.

Within this context, LinkedGeoData[5] aims to the challenging task of integrating such OpenStreetMap geometries with DBpedia and other semantically rich datasets [2]. Similarly to DBpedia, LinkedGeoData contains an ontology extracted from OpenStreetMap. The main dataset of the project contains about 3 billion RDF entities, creating a ontologically enhanced version of OpenStreetMap. For distributing its datasets, LinkedGeoData follows the 4 rules of the Linked Data initiative[6], whose increasing impact is analysed by Bizer et al. [5]. One limitation of the LinkedGeoData project is the fact that, out of a large amount of data available, only about 53,000 entities contain a direct link to a relevant DBpedia page [2, p.743]. While this matching produces good results when the entities considered are cities, it is less so for other semantic categories (e.g. countries and universities). However, in our work for map personalisation, many more categories of entities need to be taken into account.

In order to gain a better understanding of user spatial interests, this linkage between vector data and ontologies can be beneficial. Implicit feedback indicators, such as mouse movements and map navigational behaviour, can be used to infer user interests [18]. The explicit semantic associations provided by LinkedGeoData between the map presented to the user and its underlying ontological entities represent a step in this direction. Once these ontological relationships are established, it is possible to combine them with implicit indicators to enrich spatial and refine user profiles. But, in order to infer valid user interests from implicit feedback indicators such as mouse movements, a more detailed and comprehensive mapping between spatial data and ontological data is needed.

Another major topic that has not been not addressed in other projects is the impact of the map scale. When a user uses a typical interactive Web map, the semantic content of the displayed information changes depending on the scale. Our system takes this important parameter into account and allows a further exploration of the role of map scale in the user interest determination. This system, whose architecture is described in Section 3, aims at addressing these issues.

[4] http://dbpedia.org

[5] http://linkedgeodata.org

[6] http://www.w3.org/DesignIssues/LinkedData.html

3 System architecture

We have developed a system that contributes to fill the semantic gap in VGI data, bridging ontological concepts with geographical entities. The objective of the system is to retrieve semantic content from a geographical location, taking scale into account and finding ontological terms that can be used for user interests extraction. When the user clicks on the Web map, the system processes a spatial query mapping spatial features to semantic entities. This system has been developed as a module of our Web platform for map personalisation and visualisation outlined by McArdle et al. [20]. The module is a Web application that processes a spatial query and retrieves ontological results interacting with several Web services.

The core service of the system, which we called *Semantic Service*, is based on the Web development framework Grails[7], which provides an intuitive and effective environment for developing and deploying Web applications. Grails is also a suitable environment for creating and interacting with Web services. The company CloudMade[8] offers geographic-related Web services. Among others, a geocoding and reverse geocoding Web service that retrieve objects from the OpenStreetMap vector dataset data[9] are available. This service is particularly suitable to retrieve vector data without dealing with low level details of the underlying spatial DBMS, therefore we decided to use it as OpenStreetMap data provider.

GIS Open-Source software is one of the technologies that compose neogeography [13, p.2025]. Software packages released under licenses derived from the GNU public license have provided users and developers with increasing number of tools [24]. Similarly, one of the distinguishing aspects of VGI is the distribution of data under the Creative Commons license[10], which enables its free circulation on the Internet. As opposed to traditional ‘closed’ spatial datasets maintaining and licensed by a strongly vertical organisation, such open technologies offer the opportunity to increase the understanding of their properties, quality and possible usages, with an advantage for the whole community of users and developers. Therefore, we decided to adopt open technologies and data sources for this work.

Most of the open data produced by DBpedia and LinkedGeoData is stored and distributed in the Resource Description Framework (RDF). In order to manipulate those structures, our system relies on Jena, a semantic web framework for Java[11], which provides a set of functionality specific for RDF and other widely-used ontological formats. These projects offer Web services to query their datasets through the SPARQL language[12], designed specifically for RDF data, considered one of the key technologies of the Semantic Web [22]. This way com-

[7] http://www.grails.org

[8] http://cloudmade.com

[9] http://developers.cloudmade.com/projects/show/geocoding-http-api

[10] http://creativecommons.org

[11] http://jena.sourceforge.net

[12] http://www.w3.org/TR/rdf-sparql-query

plex queries can be executed remotely without maintaining a local copy of the data.

The system architecture is depicted in Figure 1. The *Semantic Service* hosts the main functionality of the system, discussed in detailed in Section 4.

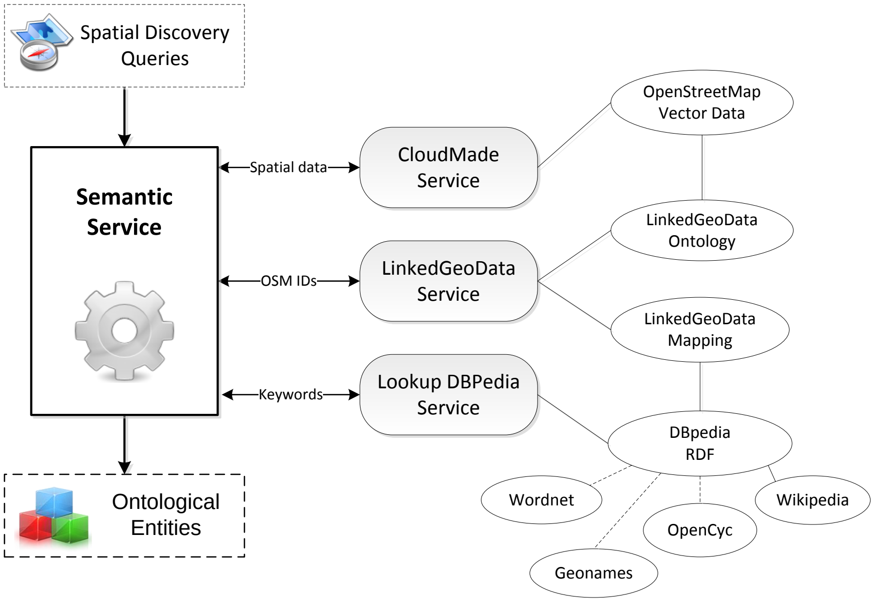


Fig. 1. System architecture

4 System functionality

Our system aims at integrating the OpenStreetMap vector dataset (spatially rich and semantically poor) with DBpedia (spatially poor and semantically rich) in order to discover ontological concepts and entities related to a given geo-location. This is done by issuing spatial discovery queries, whose main parameters are:

- **geo-location**: latitude and longitude (e.g. *53.3071, -6.2218*);
- **radius**: radius expressed either in screen pixels (dependent on the current map scale) or in meters (e.g. *20p* or *500m*);
- **scale**: map scale (e.g. *1/14,000*);
- **max results**: maximum number of OpenStreetMap objects retrieved from CloudMade in the given area (e.g. *10*).

For example, a valid exploratory query could be:

{ *geo-location = (53.3071, -6.2218), radius = (500m), scale = (1/14,000), max results = (10)* }

This query retrieves a maximum of 10 ontological entities located within a radius of 500 metres from the specified point (expressed in the lat/long coordinates 53.3071 and -6.2218) at a low scale (1:14,000, corresponding to the street level). The indicated target geographical area roughly covers the University College Dublin campus in Dublin, Ireland. The Semantic Service is then expected to return entities that are semantically related to the target geographical area, such as *education* as well as individual institutions located in the campus.

In this process, a critical issue is the role of the *map scale*. Commonly used Web maps (such as Google Maps and Visual Earth) take scale into account for two reasons, firstly to decide which objects need to be displayed, and secondly to choose a suitable visual style aiming to combine clarity and aesthetic coherence. In the context of implicit feedback analysis, scale plays an important role which, as Mac Aoidh et al. pointed out, has not been fully investigated [19]. Given the different visual content of a Web map at different scales, including this parameter in the semantic extraction is beneficial to further reduce the semantic gap.

A complex research question that is worth asking is: what can map scale reveal about the user's spatial interests? For example, when a user chooses a scale of 1:30,000,000, a typical Web map only renders country borders, major lakes and capital cities. If the user operates for a long time over a geographic area at a very high scale, it is reasonable to expect that they are likely to be interested into high regional features such as rivers, urban areas and major infrastructures. Considering all the entities located in the target area, such as individual cinemas and theatres, would be semantically misleading. On the contrary, when the scale is as low as 1:10,000, the user can see objects at the street level such as individual buildings, restaurants and bus stops and therefore the Web map includes all the fine-grained details available in the dataset. A project such as LinkedGeoData does not take this problem into consideration, as it only provides links for cities and few other entities.

Our system follows this idea by retrieving different objects categories at different scales, following an approach sometimes called *what you see is what you get*. The definition of these 'semantic layers' associated with different scales mirrors closely the structure of the CloudMade map, whose internal structure and style are fully customisable. We have chosen to start with the default map style, called 'the original' in the CloudMade style editor[13], because it represents the general-purpose Web maps that have become ubiquitous in neogeography. For instance, the top semantic layer in the scale range *(1:28,000,000-1:15,000,000)* only includes *countries* and *capital cities*. On the contrary the lowest semantic layer, whose scale range is *(1:2,000-1:1)*, includes categories such as *restaurants*, *ATMs* and *shops*, exposing the smallest entities contained in the dataset.

The *radius* is also an important parameter that can have a high impact on the result of the spatial query. The CloudMade service accepts the query radius

[13] http://maps.cloudmade.com/editor

in meters. In order to match the user perception more closely, our Semantic Service also accepts the radius in screen pixels, converting them into meters on the map currently being displayed. Another relevant parameter is the *maximum number* of OpenStreetMap objects to be retrieved. This parameter can be used to tune the query scope, trying to strike a balance between a more inclusive result set potentially including irrelevant entities and a smaller result set, which might exclude relevant ones.

The query, including the aforementioned parameters, is processed by the Semantic Service as follows (see Figure 1):

1. *Retrieve OpenStreetMap objects from CloudMade service.*
2. *Retrieve DBpedia mapping from LinkedGeoData.*
3. *Extract key words from OpenStreetMap objects.*
4. *DBpedia lookup with extracted key words.*
5. *Heuristic to determine whether the DBpedia nodes are valid or not.*
6. *Extract ontological terms and categories.*
7. *Merge results and store them in a XML file.*

In **step 1** the system retrieves the OpenStreetMap objects located within a certain radius from the CloudMade service, taking scale into account. The OpenStreetMap nodes contain metadata and tags. For example, Figure 2 displays the OpenStreetMap node that represents University College Dublin.

```
<node id="83211617" lat="53.3071709" lon="-6.2218882"
  user="rorym" uid="23770" visible="true" version="2"
  changeset="432796" timestamp="2008-03-31T00:24:37Z">
    <tag k="amenity" v="university"/>
    <tag k="created_by" v="Potlatch 0.8a"/>
    <tag k="name:en" v="University College Dublin"/>
    <tag k="name:ga" v="An Coliste Ollscoile, Baile tha Cliath"/>
</node>
```

Fig. 2. University College Dublin in OpenStreetMap XML

It is possible to note the nature of semantic information contained in OpenStreetMap. The entity names and the *amenity* tag do not allow further semantic navigation, for example towards the concepts *education*, *school* and *college*, which are highly similar. Such data is unsuitable for implicit feedback analysis, because it does not show connections between those ontological concepts, which are useful to determine user's interests (e.g. in education). Therefore it is necessary to proceed to step 2 to get richer semantics.

In **step 2** the node ids are then extracted and matched on DBpedia nodes through the LinkedGeoData mapping dataset, described in Section 2. In the case of the University College Dublin node, the mapped entity on LinkedGeoData[14] does not contain any more information than the original OpenStreetMap node.

[14] http://linkedgeodata.org/page/node83211617

Afterwards, in **step 3**, key words are extracted from the node, trying to obtain useful semantic content. The extraction of key words from OpenStreetMap objects is executed by defining a subset of the tags as semantically relevant, ignoring the others. OpenStreetMap metadata, such as contributor's information and data sources (*created_by*, *user* and *source*) are discarded, as well as tags that do not seem to form points of interest for a general user (*abutters*, *smoothness*, *incline*, *voltage*). Semantically relevant tags are given a high priority in the key words list: the English name tag (*name:en*), when available, has the highest priority, followed by *amenity*, *shop*, *tourism*, *landuse*, *natural*. In the case of the node representing University College Dublin, the extracted keywords are 'university', 'college' and 'dublin', which are utilised in the following step.

In order to allow users to retrieve nodes by key words, DBpedia provides a Web service called *DBpedia Lookup*[15], designed and successfully used by [16] in collaboration with the BBC. The service takes key words and returns the URI of matching DBpedia nodes, if any. In **step 4**, the Semantic Service invokes DBpedia lookup and analyses the return URIs. In the example, the service returns the University Dublin College page as a first result[16].

In **step 5** the system utilises a heuristic to determine whether the returned DBpedia node is valid or not, based on two criteria: (a) *geographic proximity* and (b) *tag matching*. To assess criterion (a), the system calculates the distance between the OpenStreetMap and the DBpedia node centroids. If the distance is lower than a threshold ϵ, the match is considered valid. A value for ϵ that seems to give reasonably good results is 50km, for example preventing frequent mismatches between European and North-American cities with the same name. In the case of University College Dublin, criterion (a) is fulfilled, with a distance smaller than 1 kilometre. When the geo-location is not available in the DBpedia node, criterion (b) is considered. All the tags present in the OpenStreetMap node are matched against the DBpedia node. If the matching tags ratio is higher than threshold σ, the node is considered to be valid. The default value of σ, based on a preliminary evaluation, is set to 0.5. The optimal values of ϵ and σ can be determined by further experimental evaluation.

In **step 6**, the retrieved valid nodes are then processed to extract ontological terms and categories, which enhance the semantic relevance of the results. For example, the DBpedia node representing University College Dublin contains, among others, the ontological term *university*[17]. By visiting this ontological term, it is possible to navigate to the parent term (*educational institution*) and, from there, to reach the terms *school* and *college*. Semantic similarities starting from the OpenStreetMap node can now be explored.

In **step 7**, all the results are merged and stored into an XML structure and can be either stored for further analysis for formatted in human-readable HTML code and displayed to a human user.

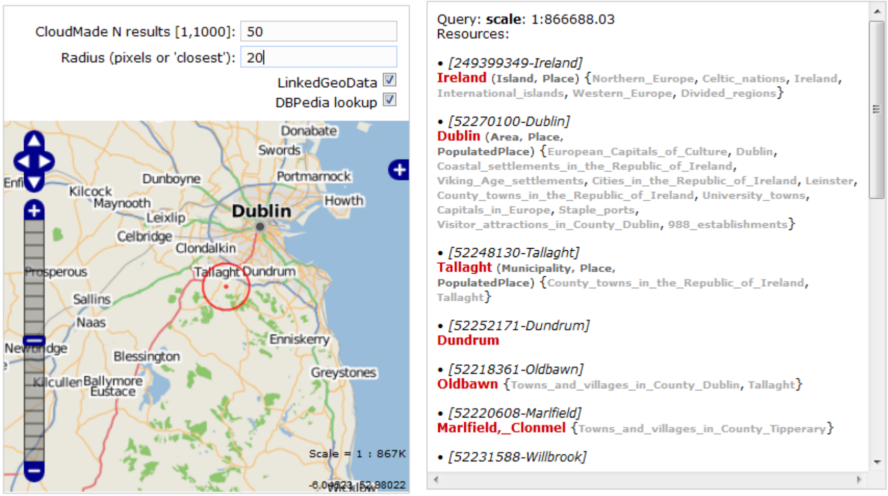
[15] http://lookup.dbpedia.org

[16] http://dbpedia.org/page/University_College_Dublin

[17] http://dbpedia.org/ontology/University

Section 5 outlines a user case which shows an interaction with the system and describes its Web GUI.

5 The Web graphical user interface



CloudMade N results [1,1000]: 50

Radius (pixels or 'closest'): 20

LinkedGeoData ☑

DBPedia lookup ☑

Query: **scale**: 1:866688.03
Resources:

- *[249399349-Ireland]*
Ireland (Island, Place) {Northern_Europe, Celtic_nations, Ireland, International_islands, Western_Europe, Divided_regions}

- *[52270100-Dublin]*
Dublin (Area, Place, PopulatedPlace) {European_Capitals_of_Culture, Dublin, Coastal_settlements_in_the_Republic_of_Ireland, Viking_Age_settlements, Cities_in_the_Republic_of_Ireland, Leinster, County_towns_in_the_Republic_of_Ireland, University_towns, Capitals_in_Europe, Staple_ports, Visitor_attractions_in_County_Dublin, 988_establishments}

- *[52248130-Tallaght]*
Tallaght (Municipality, Place, PopulatedPlace) {County_towns_in_the_Republic_of_Ireland, Tallaght}

- *[52252171-Dundrum]*
Dundrum

- *[52218361-Oldbawn]*
Oldbawn {Towns_and_villages_in_County_Dublin, Tallaght}

- *[52220608-Marlfield]*
Marlfield,_Clonmel {Towns_and_villages_in_County_Tipperary}

- *[52231588-Willbrook]*

Fig. 3. Web User Interface for Spatial Queries

In this section we present the Web user interface of the system, showing an example of its functionality, and describes a preliminary evaluation. The Web GUI allows the user to execute spatial queries in a user friendly way.

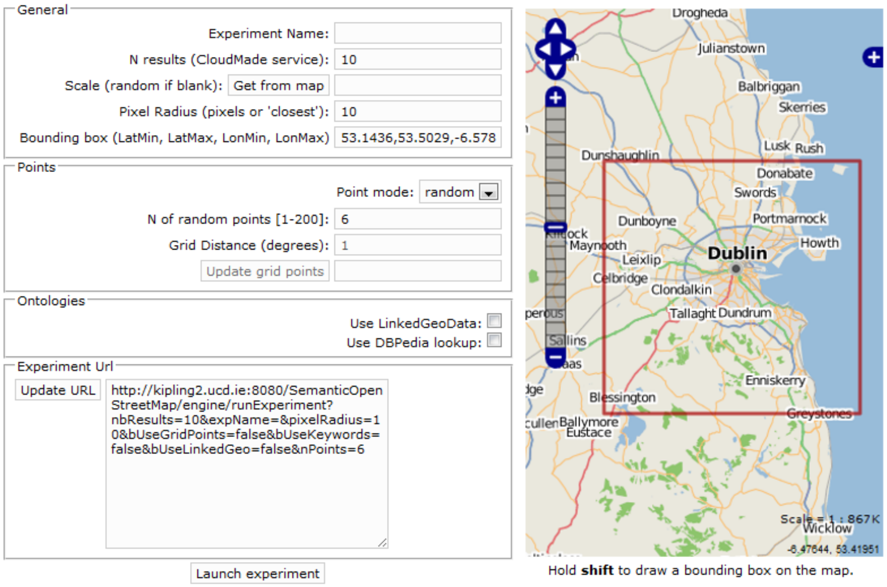
Discovery spatial queries can be performed on the system via an interactive Web page represented in Figure 3. This page can be used to execute spatial queries by clicking on an interactive Web map. The panel is split in two frames. The left frame contains the query parameters, which can be changed by the user, such as max number of results, radius and ontologies to be used. A CloudMade map is rendered to let the user choose a location in an intuitive way. The map (rendered with 'The Original' CloudMade style[18]). The user clicks on an area in the south of Dublin and a red circle with a red dot in its centre is displayed to represent the query area. In this query, the radius is set to 20 pixels. The map scale (about 1:860,000) is read automatically from the Web map and submitted to the server, which processes the request on-the-fly, as described in Section 4.

The query XML output is loaded in the panel in the frame on the right, converted into HTML. In this case the system has retrieved several DBpedia

[18] http://maps.cloudmade.com/editor

nodes (highlighted in bold) and has extracted ontological terms, including Ireland, Dublin and the suburbs Tallaght and Dundrum, located within the circle. Several DBpedia categories are also displayed, ranging from geographical concepts such as *Northern Europe* to political and historical ones (*Divided Regions*, *Celtic Nations*). The system has made explicit this implicit semantic content and has given to the user a list of abstract concepts contained within the selected spatial area.

Retrieving the results associated with a given location is useful to show the system capabilities, but it cannot give any insight on the properties of the datasets. In order to analyse the system behaviour on large geographical areas, we have built an interface to perform sampling experiments.



General

Experiment Name:

N results (CloudMade service): 10

Scale (random if blank): Get from map

Pixel Radius (pixels or 'closest'): 10

Bounding box (LatMin, LatMax, LonMin, LonMax) 53.1436,53.5029,-6.578

Points

Point mode: random

N of random points [1-200]: 6

Grid Distance (degrees): 1

Update grid points

Ontologies

Use LinkedGeoData: □

Use DBPedia lookup: □

Experiment Url

Update URL

```
http://kipling2.ucd.ie:8080/SemanticOpen
StreetMap/engine/runExperiment?
nbResults=10&expName=&pixelRadius=1
0&bUseGridPoints=false&bUseKeywords=
false&bUseLinkedGeo=false&nPoints=6
```

Launch experiment

Hold **shift** to draw a bounding box on the map.

Fig. 4. GUI for sampling a bounding box by running multiple queries

Through the page represented in Figure 4 the system can sample the datasets. The user can draw a bounding box on the map and choose several parameters of the sampling to be performed. The controls in the *ontologies* section of the page allows the user to select which ontologies have to be included in the experiment. In order to contrast the results, it is possible to enable or disable the LinkedGeoData mapping and the DBpedia lookup service.

The points to be analysed can be selected in 2 modes: *random* and *grid*. The former consists of selecting a number of random locations within the bounding box, while the latter distributes points on the bounding box at a regular distance

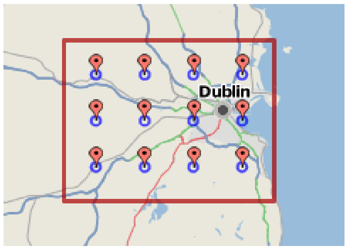


Fig. 5. Grid sampling a bounding box surrounding the Dublin area

(specified in the parameter *grid distance*, expressed in degrees). The map scale can either be set to a given value for all the points, or can be selected randomly for each point. This way it is also possible to observe what the impact of the scale is on the results.

The experiment can be executed with the button *launch experiment*. The system then executes a query on each point with the selected parameters, and all the results are merged. For the moment the complexity of the procedure is linear, so the time grows linearly with the amount of selected points. When completed, the experiment results are stored in the XML file containing all the spatial queries and the relevant ontological entities. The input and output of each query are included in the file, in order to be easily machine-processed and analysed. At the end of the procedure, the URI of the XML document is returned to the user, and the file can be used for any other analysis.

```
<resource index='0'>
  <osm id='52270446' name='Roundwood'>
      <location lon='-6.22481' lat='53.06347' />
  </osm>
  <dbpedia>
      <node>Roundwood</node>
      <location lon='-6.2333' lat='53.06' />
      <subjects>
        <category>Towns_and_villages_in_County_Wicklow</category>
      </subjects>
      <ontology>
         <term>Municipality</term>
         <term>Place</term>
         <term>PopulatedPlace</term>
      </ontology>
  </dbpedia>
</resource>
```

Fig. 6. A sample of the XML code returned by the system

Figure 5 shows an example of *grid* sampling of a large bounding on the Irish East coast with 12 points. When executing a small experiment with scale set to 14,000, radius equal to 10 pixels (equivalent to 50 meters at this scale), the results contain 13 DBpedia nodes, 43 categories and 27 ontological terms. One of the retrieved resources is the Irish village Roundwood, stored in the XML code in Figure 6 contains an OpenStreetMap structure (*osm*) and a DBpedia node (*dbpedia*), which contains a geo-location, one category and 3 ontological terms. This code can be easily processed and the relevant URI[19] can be accessed. Thus, links between the spatial vector data displayed to the user and the ontological and semantic richness of DBpedia have been enabled.

During the development of the system we carried out a preliminary evaluation with postgraduate students of the School of Computer Science and Informatics (University College Dublin). In order to engineer the heuristics, establish a suitable value for the thresholds and identify issues, we defined three alternative versions of the system, combining the phases in different ways:

- *Version 1*: OpenStreetMap + LinkedGeoData mapping
- *Version 2*: OpenStreetMap + Our heuristics
- *Version 3*: OpenStreetMap + LinkedGeoData mapping + Our heuristics

Version 1 and 2 are using only certain parts of the system, while Version 3 exploits its full functionality, similarly to the approach used by Mirizzi et al. [21]. The following parameters have been defined as constant: ϵ (maximum node distance) = 50km, σ (tag matching ratio) = 0.5, *maximum results* (CloudMade service) = 10, *radius* = 20 pixels, *scale* = random, *bounding box* = Dublin area.

The users have been presented with a set of 6 random spatial queries and the correspondent view on the CloudMade map scaled and centred to the query location. The queries had been performed with different versions of the algorithm (2 for each version), hiding this information from the user. The users have then been asked to rank the semantic relevance of each node (either DBpedia entity or ontological term) on a Likert scale from 1 to 5 (from *strongly uncorrelated* to *strongly correlated* with the visible map).

During this preliminary evaluation, several issues have been identified. Very little difference separates the performance of Version 2 and 3. The LinkedGeoData mapping, although very accurate when present, did not significantly improve the results, which were mostly return by our heuristics based on DBpedia lookup. In some cases, such as for the term *Smithfield*[20], which identifies numerous places in former British colonies, the heuristics succeeded and selected the correct neighborhood in Dublin, based on the geo-location. However, certain mismatches still occurred. When the OpenStreetMap feature contains a very frequent term in a certain geographical area and does not have other semantic content, a mismatch is likely to occur. For example, this typically happens with common surnames (e.g. *Smith*) that are highly frequent in the datasets. Overall, on 47 retrieved nodes, 4 were considered as irrelevant (8.5%).

[19] http://dbpedia.org/resource/Roundwood

[20] http://dbpedia.org/page/Smithfield

Although this preliminary evaluation has confirmed that the system works correctly in most cases, the sample is too small to draw general conclusions about it. The objective of this evaluation was mainly to identify design flaws that need to be addressed. In order to draw general conclusions about the system performance, an extensive evaluation is needed. In particular, more independent variables need to be defined, such as the thresholds, the radius and the maximum number of results, and a bigger sample of queries must be taken into account. This way, several aspects of the system behaviour can be better assessed and quantified.

Section 6 draws conclusions and outlines our plans for system evaluation and future work.

6 Conclusions and future work

In this paper we have described a system that aims to fill the semantic gap between spatial and non-spatial user-generated data. Starting from a semantically poor dataset (OpenStreetMap), the system performs spatial discovery queries and expands the semantic content of a given geographic location. This process relies on the LinkedGeoData dataset mapping and and extends it with the addition of key word extraction from OpenStreetMap nodes. Furthermore, through a novel heuristic technique, the system filters out uncorrelated entities. The system retrieves matching DBpedia nodes (semantically rich, spatially poor), collecting ontological knowledge related to the given location. This way semantic relationships between a geographical location and ontological concepts are established and can refine implicit feedback analysis and enable other applications. In this process the system takes into account the map scale at which the user is working for improved relevance during retrieval.

Although our preliminary evaluation seems promising, an extensive evaluation needs to be carried out in order to gain further understanding of the system strengths and weaknesses. The technique implemented in the system relies on several thresholds, described in Section 4. The optimal values of these thresholds should be established by experimental evaluation, based on performance metrics. A fundamental metric in this context is the relevance of related entities to human users, which was used in our preliminary evaluation described in Section 5. Defining query parameters (scale, geographical region, radius, etc) as independent variables, it is possible find out which conditions impact on the system performance and why.

Another area that needs further exploration is the navigation of ontological entities and concepts, starting from the ones returned by the system, which do not necessarily include all relevant concepts. A body of work exists on similarity metrics between DBpedia nodes. Kobilarov et al. utilises a metric based on distance in the DBpedia graph [16], while Hassanzadeh and Consens investigate several similarity measures based on an approximate string matching [15]. Starting from a set of nodes returned by the system, it is possible to explore similar concepts, showing correlation between geographically and semantically close entities and see how this relates to the user interests.

From a more quantitative viewpoint, it would be interesting to use the *experiment* interface to study and contrast the datasets spatial and semantic statistical properties. Comparing the system results in different geographical areas (e.g. Europe vs North-America) and semantic areas (e.g. human-built entities vs natural entities) can give insight not only on the system performances, but also on the properties of these open datasets. We plan to further investigate the possibilities of implicit feedback analysis and spatial personalisation offered by such open spatial datasets and ontologies, whose visibility and impact are bound to increase in the near future, in academia and industry alike.

Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support and thank Paul Huynh for his contribution to the system implementation.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The Semantic Web. 4825, 722–735 (2007)
2. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a spatial dimension to the web of data. In: International Semantic Web Conference. pp. 731–746 (2009)
3. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific American. 284(5), 28–37 (2001)
4. Bishr, M., Kuhn, W.: Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In: AGILE 2007. pp. 365–387. Springer, Aalborg, Denmark (2007)
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems. 5(3), 1–22 (2009)
6. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A crystallization point for the Web of Data. Journal of Web Semantics. 7(3), 154–165 (2009)
7. Ding, L., Kolari, P., Ding, Z., Avancha, S.: Using ontologies in the Semantic Web: A survey. Ontologies pp. 79–113 (2007)
8. Flanagin, A., Metzger, M.: The credibility of volunteered geographic information. GeoJournal. 72(3), 137–148 (2008)
9. Fonseca, F., Egenhofer, M.: Ontology-Driven Geographic Information Systems. In: Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems. pp. 14–19. ACM-GIS (1999)
10. Fonseca, F., Egenhofer, M., Agouris, P., Câmara, G.: Using ontologies for integrated geographic information systems. Transactions in GIS. 6(3), 231–257 (2002)
11. Goodchild, M.: Citizens as sensors: The world of volunteered geography. GeoJournal. 69(4), 211–221 (2007)
12. Haav, H., Kaljuvee, A., Luts, M., Vajakas, T.: Ontology-Based Retrieval of Spatially Related Objects for Location Based Services. In: On the Move to Meaningful Internet Systems: OTM 2009. pp. 1010–1024. Springer (2009)

13. Haklay, M., Singleton, A., Parker, C.: Web Mapping 2.0: The Neogeography of the GeoWeb. Geography Compass 2(6), 2011–2039 (2008)
14. Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Computing 7(4), 12–18 (2008)
15. Hassanzadeh, O., Consens, M.: Linked Movie Data Base. In: Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), Madrid, Spain (2009)
16. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. pp. 723–737. Springer (2009)
17. Korpipää, P., Häkkilä, J., Kela, J., Ronkainen, S., Känsälä, I.: Utilising context ontology in mobile device application personalisation. In: Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia. pp. 133–140. ACM (2004)
18. Mac Aoidh, E., Bertolotto, M., Wilson, D.: Analysis of implicit interest indicators for spatial data. In: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems. p. 47. ACM (2007)
19. Mac Aoidh, E., Bertolotto, M., Wilson, D.: Understanding geospatial interests by visualizing map interaction behavior. Information Visualization 7(3), 275–286 (2008)
20. McArdle, G., Ballatore, A., Tahir, A., Bertolotto, M.: An Open-Source Web Architecture for Adaptive Location Based Services. In: Proceedings of the 14th International Symposium on Spatial Data Handling (SDH), Hong Kong. vol. 38(2), pp. 296–301 (2010)
21. Mirizzi, R., Ragone, A., Di Noia, T., Di, E.: Ranking the Linked Data: The Case of DBpedia. In: Web Engineering. 10th International Conference, ICWE 2010. Vienna, Austria. pp. 337–354. Springer (2010)
22. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query (2006)
23. Sieg, A., Mobasher, B., Burke, R.: Web Search Personalization with Ontological User Profiles. In: Proceedings of the 16th ACM International Conference on Information and Knowledge. pp. 525–534. ACM (2007)
24. Steiniger, S., Bocher, E.: An Overview on Current Free and Open Source Desktop GIS Developments. International Journal of Geographical Information Science 23(10), 1345–1370 (2009)
25. Turner, A.: Introduction to Neogeography. O'Reilly Media, Inc. (2006)
26. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland. pp. 585–594. ACM (2006)
27. Yu, S., Spaccapietra, S., Cullot, N., Aufaure, M.: User Profiles in Location-Based Services: Make Humans More Nomadic and Personalized. In: Databases and Applications. Proceedings of the IASTED International Conference on Databases and Applications. Innsbruck, Austria. ACTA Press (2004)