Demo: Using Personalized PageRank for Keyword Based Sensor Retrieval

Lorand Dali
Jožef Stefan Institute
Jamova 39, Ljubljana
Slovenia
+38614773933
lorand.dali@ijs.si

Alexandra Moraru
Jožef Stefan Institute
Jamova 39, Ljubljana
Slovenia
+38614773144
alexandra.moraru@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jamova 39, Ljubljana
Slovenia
+38614773377
dunja.mladenic@ijs.si

ABSTRACT
The paper proposes a system for advanced sensor retrieval. Sensors provide a large volume of data about the environment. The abundance and variety of the sensor data has given rise to increased interest in the development of applications involving sensor networks. The large volumes of data produced by such sensor networks require special techniques not only for management and processing but also for its discovery. We propose and implement a system for sensor search, based on matching user's given keywords against information extracted from standardized sensor descriptions. For ranking the results we use the Personalized PageRank algorithm and apply filtering based on geo-location. Illustrative examples are presented to show how a user can interact with our system and what the benefits of ranking the search results are.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, retrieval models.*

General Terms
Algorithms, Experimentation.

Keywords
Personalized PageRank, Keyword Search, Sensors.

1. INTRODUCTION
The integration of embedded devices, such as sensors, mobile phones or RFID tags, into the Web is becoming more and more popular. Standards designed to leverage this integration are getting widely used, enabling discovery and data access through standard protocols and APIs [1]. The data that such embedded devices can provide, fosters the development of new applications. These applications can vary from simple monitoring of environmental conditions that one can access for finding the less polluted track for jogging to complex processing of sensor data in advanced (stream) data mining applications for solving global problems such as water supply management, hazards detection or traffic fluidization.

Systems such as Pachube[1], SensorMap[2] or Sensorpedia[3] provide web services for accessing sensors descriptions and related data sets. These systems enable search based on tags, exact match of sensor names or use faceted search. However, regardless of the application a user is interested in, the first step required is the selection of the desired sensors. This step implies searching over thousands of sensor descriptions where ranking of the search results is imperative.

Searching for sensors that provide the observations for a desired application can be challenging when they originate from more vendors that are using different ways to describe and classify them. Often characteristics such as owner, location, observed phenomena are important for searching sensors, but they are not all the time described using the same structure forms so that they could be easily queried. Therefore extracting the relevant information from sensor description is often required.

We propose and implement a system for keyword based sensor search, which looks for matches in text descriptions of sensors that contain information regarding the observed phenomena, names and description given by the owners. Next, we apply the Personalized PageRank algorithm for ranking, considering observed phenomena and the platforms and networks on which the sensors are deployed. Finally we do filtering based on user's geographical preferences.

The rest of the paper is structured as follows. Section 2 briefly describes the related work in the areas of sensor search and object-level information retrieval. Section 3 describes the dataset used. Section 4 introduces the search and ranking models, Section 5 shows a couple of illustrative examples, and finally the conclusions are drawn.

2. RELATED WORK
The existing systems providing web services for accessing sensor data support simple keyword search based on tags describing the sensors and filtering of the results using predefined categories of sensors [2]. A different approach for sensor search is presented in [3], using semantic descriptions of sensors. The sensor metadata is structured using semantic MediaWiki, which creates a set of interlinked pages and also provides RDF graph representation.

WWW'11, March 28th– April 1st, 2011, Hyderabad, India.

[1] http://www.pachube.com/

[2] http://atom.research.microsoft.com/sensewebv3/sensormap/

[3] http://www.sensorpedia.com

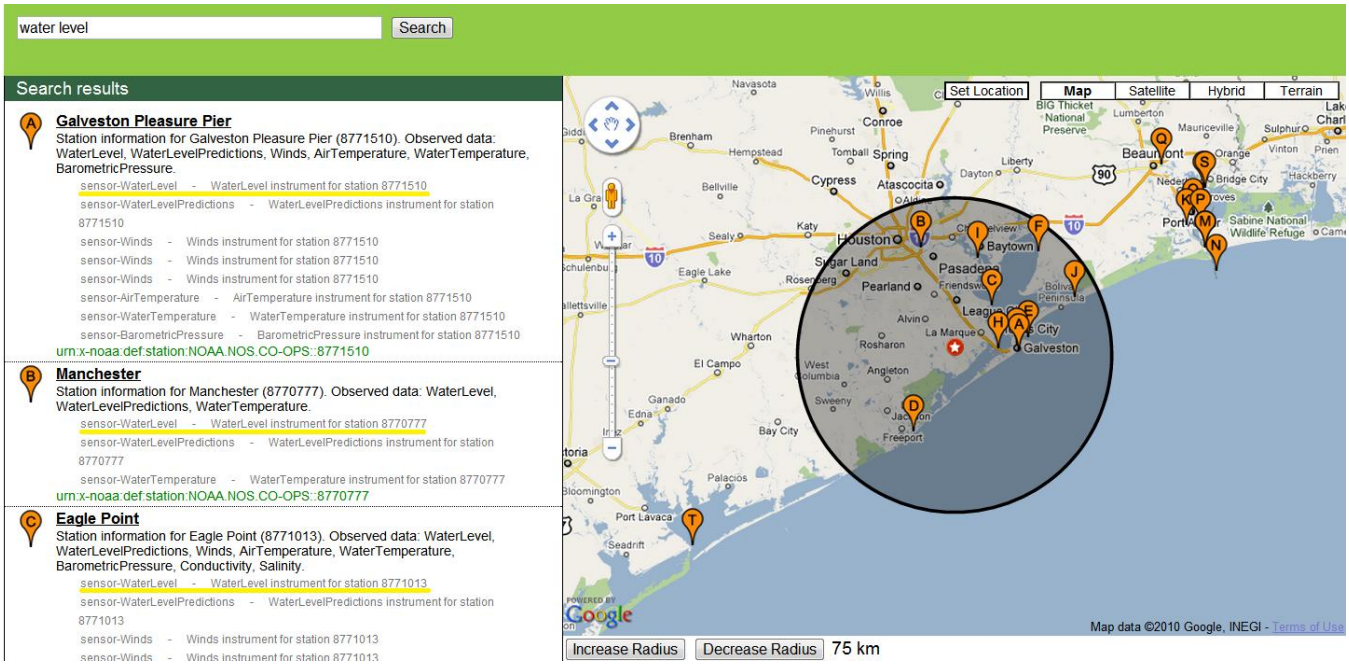

Figure 1 User Interface for Sensor Search

A SPARQL wrapper is then used to query the metadata repository and to obtain the relevant metadata pages. Further, an extended PageRank algorithm is applied on the result set, taking into consideration two types of links resulted from the metadata representation, i.e. links between pages and links – edges – between graph nodes. The system proposed in this article differs from that work in the structures used for metadata representation, the latter requiring a semantic representation and in applying Personalized PageRank algorithm instead of PageRank.

Researchers have used several graph-based algorithms in ranking search results. The most popular of them is PageRank [4] which regards web pages as nodes and hyperlinks as edges in a graph. Various adaptations of PageRank look at objects from the open data cloud or records from databases as nodes, and to the relations between them as edges. In this case the edges get weights according to the importance of the relation which they represent. PageRank is a query independent ranking model; this limitation was addressed by research on Personalized PageRank [5] and ObjectRank **Error! Reference source not found.** where the ranking is made query specific by allowing jumps to nodes which were matched by keyword search only.

3. DATA DESCRIPTION

The datasets chosen for experimentation contains description of sensors in the area of ocean tides and currents, available online[4]. The motivation for choosing this dataset is given by the large number of standardized sensor descriptions provided. The representation format is SensorML[5], facilitating parsing and extraction of relevant metadata. The sensor descriptions contain the following concepts: networks, platforms, sensors and observed property. The relations between these concepts are: each sensor can observe one property (i.e. air temperature, water salinity, etc.),and is attached to one platform; each platform is deployed in one network and can have one or more sensors attached. In addition, each platform is given the latitude and longitude for its location. For the concepts defined, the following text descriptions have been taken into consideration for keyword search:

- platform, sensor and property names, given by system owners;
- standard name and definition of the property observed; These properties are described in the Marine Metadata Interoperability (MMI) ontology, under the Climate and Forecast (CF) standard names parameter vocabulary[6].
- sensor description given by owner.

4. SEARCH AND RANKING

The goal of the search is to retrieve and rank a list of sensors based on the user's request. The user provides a keyword query, a geographic location (given by latitude and longitude coordinates) and a distance (interpreted as a radius around the location). We first limit all the sensors to the given location and range. In order to rank the results of a keyword search, we have implemented a query dependent version of the PageRank algorithm.

4.1 PageRank

In this section we briefly describe the PageRank algorithm. PageRank was introduced for web search in order to provide a global, query independent ranking of web pages. The input for PageRank is a directed graph, and it gives a score to each of the nodes as a result. PageRank is based on the random walk model, i.e. a large number of users walk the graph choosing at each step to either walk to a neighbor of the current node or to jump to any random node in the graph. The number of users which are expected to be at a given node at a moment in time gives the score of that node.

[4]Center for Operational Oceanographic Products and Services, http://tidesandcurrents.noaa.gov/index.shtml

[5] http://www.opengeospatial.org/standards/sensorml

[6] http://mmisw.org/orr/#http://mmisw.org/ont/cf/parameter

The scores are computed recursively with the following equation:

$$\boldsymbol{p} = d \cdot \boldsymbol{M} \cdot \boldsymbol{p} + (1-d) \cdot \boldsymbol{u}, \qquad \boldsymbol{p}, \boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{M} \in \mathcal{M}(n)$$

Where n is the number of nodes in the graph, $\boldsymbol{p}$ is the PageRank vector containing the score for each node and is initialized with 0, $\boldsymbol{M}$ is the transition matrix constructed in the following way:

$$\boldsymbol{M}[i,j] = \begin{cases} 5, & \boldsymbol{i} \text{ and } \boldsymbol{j} \textit{ measure the same thing} \\ 4, & \boldsymbol{i} \text{ and } \boldsymbol{j} \textit{ are on the same platform} \\ 1, & \boldsymbol{i} \text{ and } \boldsymbol{j} \textit{ are on the same deployment} \\ 0, & \textit{otherwise} \end{cases}$$

Moreover, to eliminate nodes which do not link to any other node we consider a sink node k such that $\boldsymbol{M}[i,k] = 1, \forall i$ and $\boldsymbol{M}[k,i] = 0, \forall i$. Finally the columns of M are normalized to sum up to 1. $\boldsymbol{u}$ is the jump vector and its entries are $\boldsymbol{u}[i] = \frac{1}{n}, \forall i$. d is the damping parameter, and represents the probability of walking to a neighboring node versus jumping. In our experiments we have set d to 0.8, and the number of iterations to 5.

4.2 Personalized PageRank

Personalized PageRank is a version of PageRank adapted to be query dependent. The query dependency is achieved by assuming that the subset Q of nodes matched by the keyword search are important apriori. This is implemented by putting a constraint on the jumps in the random walk model. More exactly, the random walker is allowed to jump only to a node which is a member of Q. This has implications only on the jump vector $\boldsymbol{u}$ whose values are now $\boldsymbol{u}[i] = \frac{1}{|Q|}$ if $i \in Q$ and 0 otherwise.

4.3 Filtering of Search Results

As the final result we return platforms, not sensors. The platform's score is the sum of the scores of the sensors it contains. Finally, the score of the platform is adjusted by dividing it with the number of radiuses it is away from the location which the user has specified. This means that the user can give a small radius to be very strict about the location of the results, and a larger radius if he is more relaxed.

5. EXAMPLE

Figure 1 illustrates the user interface for sensor search and the results for the keywords "water level". The grey circle indicates the area of interest and the orange flags represent sensor platforms (letters are illustrating the ranking). It can be observed that the results are relevant to the keyword and also the platforms with higher scores are in the area of interest. The advantage of performing the proposed ranking is that of obtaining more results closer to the area of interest, as it can be observed in Figure 2 and Figure 3 (getting 10 results instead of 3). This is explained by the weights given as we consider relevant also sensors located on the same platform or those that are in the same deployment (e.g. the sensors situated on a platform are usually related to each other, from the point of view of the applications in which they are used).

6. CONCLUSIONS

We have presented a system[7] for sensor search that uses Personalized PageRank for ranking and takes into consideration geographical preferences of the user. The examples given show how the system obtains more relevant sensor platforms in the user's area of interest. In addition, the system provides a friendly user interface.

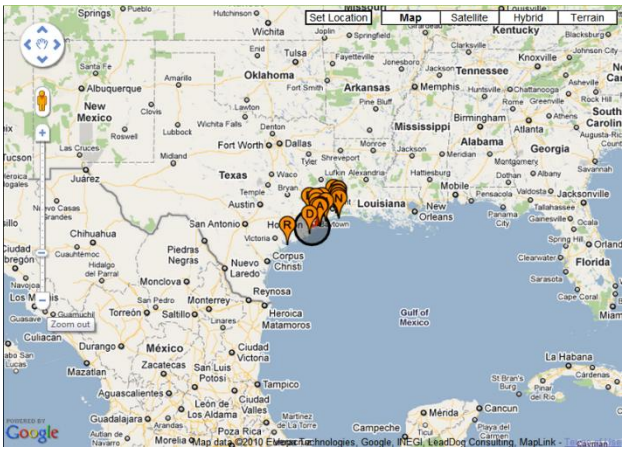

Figure 2 Search results with ranking

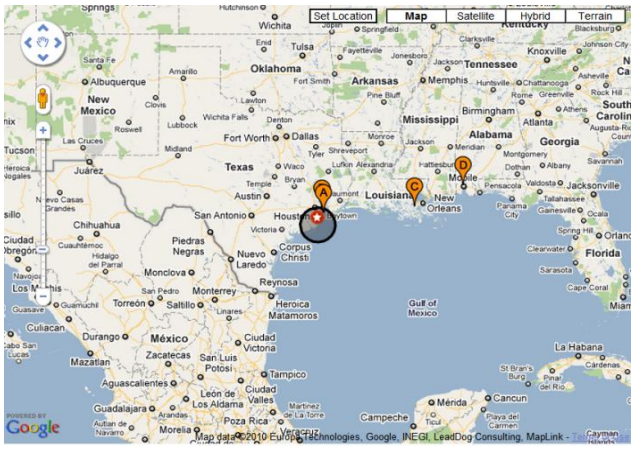

Figure 3 Search results without ranking

7. ACKNOWLEDGEMENTS

This work was supported by METANET (ICT- 249119 – NoE)

8. REFERENCES

[1] Botts, M., Percivall, G., Reed, C., Davidson, J. 2007. OGC White Paper OGC® Sensor Web Enablement: Overview And High Level Architecture. White Paper. OpenGIS.

[2] Gupta, V., Udupi, P., Poursohi, A. Early lessons from building Sensor.Network: an open data exchange for the web of things, *In Proc. of 8th IEEE International Conference* on *Pervasive Computing and Communications Workshops (PERCOM Workshops).* (March 29 2010-April 2 2010)

[3] Jeung, H. et al. 2010. Effective Metadata Management in Federated Sensor Networks. *In Proc. of The Third IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing.* (Newport Beach, California, USA, June 7-9, 2010).

[4] Bianchini, M., Gori, M., Scraselli, F. *Inside PageRank.* Transactions on Internet Technology, February 2005

[7] http://sensors.ijs.si/static/index.html

[5] Jeh, G., Widom, J., Scaling personalized web search. In WWW, Budapest, 2003

[6] Balmin, A., Hristidis, V., Papakonstantinou, Y. *Authority-based keyword queries in databases using ObjectRank*. In VLDB, Toronto, 2004