REPORT ON PROTOTYPE IMPLEMENTATION OF PAUSANIAS

AKRIVI VLACHOU

The prototype was implementation with the help of:
Vasilis Soumakis (Master Student)
George Tsatsanifos (PhD Student)

Our demonstration is available at:
http://www.idi.ntnu.no/~vlachou/pausanias.html

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION, LIFELONG LEARNING AND RELIGIOUS AFFAIRS
MANAGING AUTHORITY

NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co- financed by Greece and the European Union

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

1 CONTENTS

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION, LIFELONG LEARNING AND RELIGIOUS AFFAIRS
MANAGING AUTHORITY

NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co- financed by Greece and the European Union

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

2 Introduction

Classic web search engines, such as Google and Yahoo, provide efficient algorithms for retrieval of ranked search results given a set of keywords. However, there are also applications that provide the user the opportunity to search using geotagged criteria. These applications are Google Maps, Bing Maps etc. Such applications depict points of interest on the map and combine their location with the keywords provided by the associated document(s). These queries are so called spatial queries.

Although in many cases spatial constraints seem to suffice, sometimes there is a need to combine spatial and textual information to find points of interests. These queries are called Spatio-Textual Queries. There is an active research interest in spatio-textual queries and more specifically in the efficient retrieval of topK Spatio-Textual Queries.

This project employs 5 different queries ranging from simple Boolean Spatial Range Queries to TopK Spatio-Textual Range Queries. In this way the user will be able to choose among 5 queries to satisfy his/her constraints.

In the following, there will be an explanation of the data structure and the algorithms that Pausanias employed.

3 IR-Tree

In order to provide queries that employ both spatial and textual information, Pausanias utilized a data structure called IR-Tree. IR-Tree stands for Inverted R Tree. It basically means that there is an R tree,that will be analyzed soon in the following paragraphs, together with an inverted index to store the textual information, in our case keywords.

An R tree is used for indexing multi-dimensional information and is especially important for spatial queries. The key idea of this data structure is to store nearby objects and represent them as MBRs (Minimum Bounding Rectangles). Like in B-Trees there is a balance so all leaf nodes are on the same level. Unlike B-Trees which can only represent one-dimensional information efficiently, R-Trees also have a root node which employs a big MBR.

Each MBR also has overlapping MBRs inside of it. At the leaf level, each rectangle describes a single object. As with most trees, the searching algorithms (e.g., intersection, containment, nearest neighbor search) are rather simple. The key idea is to use the bounding boxes to decide whether or not to search inside a subtree. In this way, most of the nodes in the tree are never read during a search. Like B-trees, this makes R-trees suitable for large data sets and databases, where nodes can be paged to memory when needed, and the whole tree cannot be kept in main memory.

The key difficulty of R-trees is to build an efficient tree that on one hand is balanced (so the leaf nodes are at the same height) on the other hand the rectangles do not cover too much empty space and do not overlap too much (so that during search, fewer subtrees need to be processed). For example, the original idea for inserting elements to obtain an efficient tree is to always insert into the subtree that requires least enlargement of its bounding box. Once that page is full, the data is split into two sets that should cover the minimal area each. Most of the research and improvements for R-trees aims at improving the way the tree is built and can be grouped into two objectives: building an efficient tree from scratch (known as bulk-loading) and performing changes on an existing tree (insertion and deletion).

The search algorithm is similar to that of a B+ Tree. The input is a search MBR and the process starts from the root node while traversing down the tree. Each node contains some rectangles and pointers to the child node. The searching continues down the tree and in each visited node it has to be decided if the search box overlaps with the corresponding MBR of the node. If yes, then the corresponding child node has to be searched also. The searching is completed until all overlapping nodes have been traversed. When a leaf node is reached, its bounding boxes are tested against the query search box and if its objects lie within the search box then these objects are put into the result set.

The insertion of an object consists of traversing down the tree and choosing a node where one of its rectangles needs the least enlargement. However, there are many algorithms and lots of variations of R-Trees from the original R-Tree. It is unwise to describe them in this document.

4 Type of Queries

4.1 Boolean Spatial Range Query

In this subunit I will describe the first of the five queries that Pausanias supports. **Boolean Spatial Range Query** is a type of query that returns all possible points of interests that are within the provided range.

The input of this query are Range in meters, a dataset to search for (Hotels, Restaurants, Bars all located in the USA) and a user location in the form of latitude, longitude. If the input is valid, then all that remains is to search the R-Tree starting with the root node.

Initially, the user's marker is tested against the root MBR to discover if the minimum distance is less than or equal to the given range. If it is, the search algorithm traverses down the tree in each of the root's node children. In each newly visited node, again we have a test between the user's marker and the node's MBR. If there is a match then the corresponding node is added to a List data structure. This data structure contains the next nodes that will have to be traversed. The process of finding a point of interest continues and every time the iteration reaches a leaf node there is an actual distance metric between the user's marker and the objects of the leaf node. Each object that lies within the given range is added to the result set.

4.2 Boolean Spatio-Textual Range Query

Boolean Spatio-Textual Range Queries employ the same logic as Boolean spatial queries do. The only exception is that now we have a combination of spatial and textual information. Therefore, both B-Trees and R-Trees have to be utilized as well to provide correct search results. This combination consists of the IR-Tree.

The searching algorithm is the same as with the one that Boolean Spatial Range Query employs, but now in its newly visited node there is a procedure to test the node's set of keywords with the existing query keywords. The caveat here is that there has to be a **total** match between the number of query keywords and the number of each node's keywords. Otherwise, this node and the subsequent subnodes that belong to this node are pruned. This way, this algorithm achieves high pruning rates that consequently results in reduced query execution time.

4.3 TopK Spatial Range Query

This query consists of a given dataset, a range in meters, a user location expressed in the form of longitude and latitude and a topK variable that denotes the number of best results to be searched and retrieved.

The logic behind this query is quite simple. To begin with, we only need an R-Tree as we have only spatial constraints. So, the search algorithm is somewhat same as with the ones used from Boolean Range Queries but with the only exception that now there is a data structure that holds the topK results. This data structure persists its contents in decreasing order and using a scoring function: $F = \mathrm{alpha} * \mathrm{minDistance} + (1 - \mathrm{alpha})/\mathrm{Jaccard}$, where alpha is a parameter set with a value of 0.5 to show equal value between textual and spatial information. MinDistance is the minimum distance between the user's location and the corresponding MBR or actual point and Jaccard is the Jaccard between the query keywords and the node's keywords.

This data structure is being traversed and each time the object with the lower score is extracted. If this object is a point of interest then it is handed to the result set. Otherwise, the contents of this object are extracted and each one is tested if it satisfies the spatial requirements. Then if each object satisfies the spatial constraints, it is calculated with the scoring function and if its score is lower that the minScore of the lower in score item in the data structure, then this object becomes the new lower item in the data structure. Otherwise, it is discarded and its subtree is also pruned.

4.4 TopK Spatio-Textual Range Query

This type of query utilizes an IR-Tree but the searching algorithm is the same as the one used from the query in 2.4.

4.5 TopK Nearest Neighbor Query

This type of query is the classic nearest neighbor query but enhanced with topK capabilities. It uses an IR-Tree to complete the query.

The logic behind this query is that there is the same data structure with the same scoring function combined with the procedure that the classic nearest neighbor query uses.

5 Architecture

5.1 Presentation Layer

The front-end or else web user interface consists of the main page of the website where relevant information about Pausanias are displayed. Furthermore there are five other website pages, each one dedicated for the purpose of each query.

The web client consists of HTML5 pages along with CSS3 and JavaScript/JQuery. Pausanias utilizes the Bootstrap 3 framework to provide responsive capabilities and the JQuery library to ease the programmatic complexity related to using plain JavaScript to make AJAX requests. Bootstrap 3 is a web application framework that provides responsive elements such as buttons, forms etc. and a 12 column grid to arrange elements. The programmer has the advantage that he/she can use multiple ''rows'' of up to 12 columns in each ''row'' to arrange the html elements. In many cases just making the interface using grid is enough to both display nicely in large and small screens. In our case, it needed a little more to be sure that the user interface will scale as much nicely as it can get on every device.

In order to build the Pausanias Website, I decided that Pausanias would benefit from less complex architecture. Therefore the web client is a ''thin'' client, which means that all the processing is done on the application tier. The client just sends requests to the server and receives responses from the server. These responses have to be parsed and extracted in order to display them to the user.

All the communication between the client and the server is done using REST web services. REST (Representational State Transfer) is a software architecture style consisting of guidelines and best practices for creating scalable web services. The data format that REST uses is mainly JSON but this is not mandatory as it can use both XML and plain objects. However, this implementation uses JSON as a data transfer method because it is really easy to parse JSON objects with JavaScript.

Lastly, the website utilizes Google Maps in order to display the points of interest to the user and also to provide the user the capability to mark his/her location or the location he/she wishes to search for.

5.2 Service Layer

The website Pausanias is based upon an embedded Http Server called Jetty. Jetty is a modern application container that can support both standalone mode of web applications and embedded mode. Pausanias is based upon an embedded server to provide the capability of easy transferring among different hosting solutions.

Pausanias is distributed within a runnable JAR file and a folder containing critical content in order to run the website. This JAR file can be started with double clicking on it, or with the command line. If Pausanias is started with the command line there is an extra capability to stop it, in contrast with double clicking the JAR file where stopping the web server requires killing the process.

Upon starting, Pausanias receives requests in port 8081 and displays logging information in the command line.

The back-end of Pausanias consists of Servlets who act as REST web services. These servlets are five in total, each one for a particular kind of query. The input required varies according to whether there is need for topK results and also whether or not we need the use of IR-Tree.

The communication between the Servlets and the data layer is done using a specialized framework specific to these types of queries. This framework supports communication as well as creation of R-Trees, B-Trees and of course IR-Trees. Although the framework is quite old, fortunately it is open source so with a little tweaking I made it to work a little more efficiently.

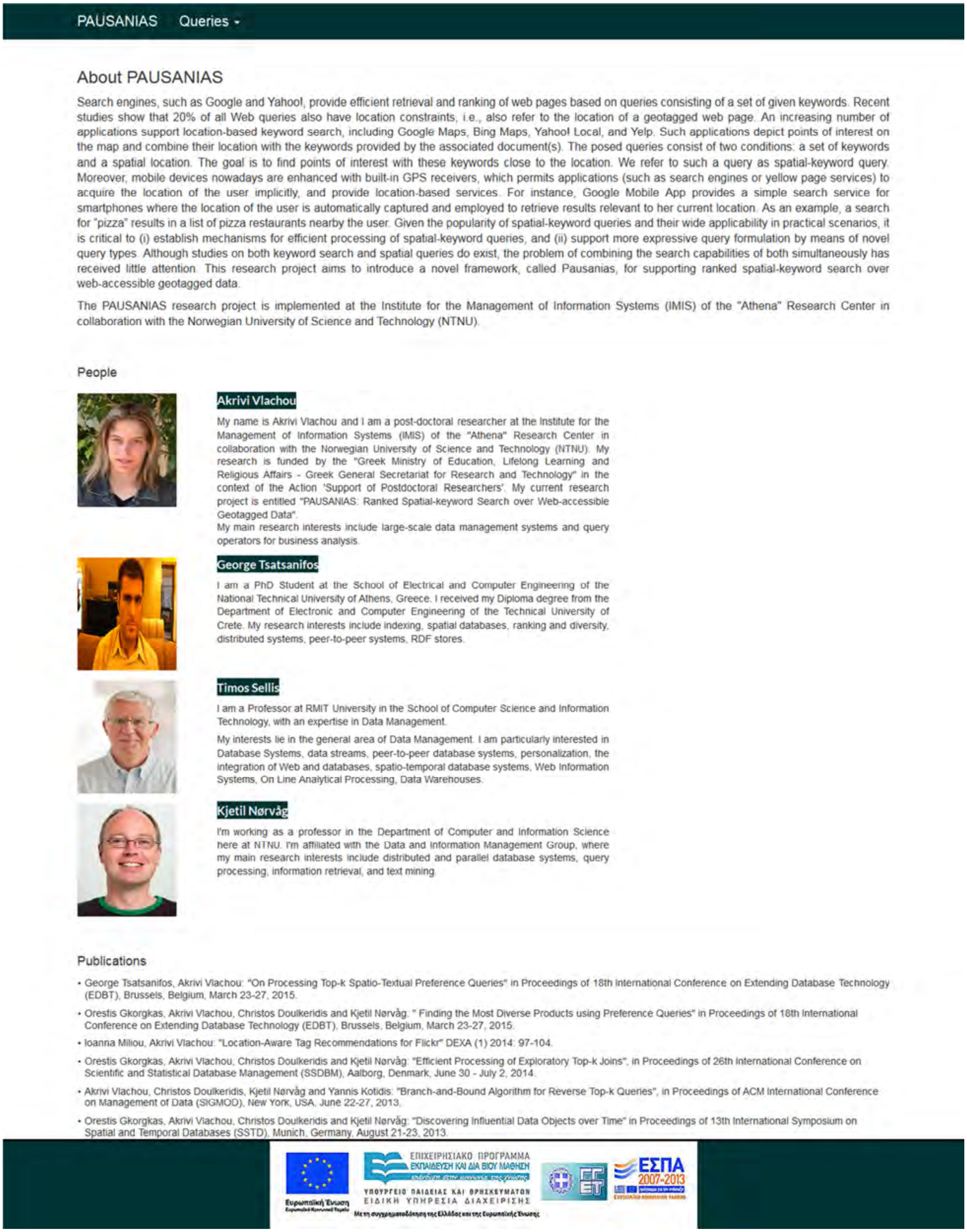
6 Website Manual

In this chapter I will present the website pages with screenshots as well as description of the process of finding POIs for the different queries that Pausanias supports.

In the following screenshot there is the starting page of Pausanias.

European Union
European Social Fund
OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION, LIFELONG LEARNING AND RELIGIOUS AFFAIRS
MANAGING AUTHORITY
Co- financed by Greece and the European Union
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.



About PAUSANIAS

Search engines, such as Google and Yahoo!, provide efficient retrieval and ranking of web pages based on queries consisting of a set of given keywords. Recent studies show that 20% of all Web queries also have location constraints, i.e., also refer to the location of a geotagged web page. An increasing number of applications support location-based keyword search, including Google Maps, Bing Maps, Yahoo! Local, and Yelp. Such applications depict points of interest on the map and combine their location with the keywords provided by the associated document(s). The posed queries consist of two conditions: a set of keywords and a spatial location. The goal is to find points of interest with these keywords close to the location. We refer to such a query as spatial-keyword query. Moreover, mobile devices nowadays are enhanced with built-in GPS receivers, which permits applications (such as search engines or yellow page services) to acquire the location of the user implicitly, and provide location-based services. For instance, Google Mobile App provides a simple search service for smartphones where the location of the user is automatically captured and employed to retrieve results relevant to her current location. As an example, a search for "pizza" results in a list of pizza restaurants nearby the user. Given the popularity of spatial-keyword queries and their wide applicability in practical scenarios, it is critical to (i) establish mechanisms for efficient processing of spatial-keyword queries, and (ii) support more expressive query formulation by means of novel query types. Although studies on both keyword search and spatial queries do exist, the problem of combining the search capabilities of both simultaneously has received little attention. This research project aims to introduce a novel framework, called Pausanias, for supporting ranked spatial-keyword search over web-accessible geotagged data.

The PAUSANIAS research project is implemented at the Institute for the Management of Information Systems (IMIS) of the "Athena" Research Center in collaboration with the Norwegian University of Science and Technology (NTNU).

People

Akrivi Vlachou

My name is Akrivi Vlachou and I am a post-doctoral researcher at the Institute for the Management of Information Systems (IMIS) of the "Athena" Research Center in collaboration with the Norwegian University of Science and Technology (NTNU). My research is funded by the "Greek Ministry of Education, Lifelong Learning and Religious Affairs - Greek General Secretariat for Research and Technology" in the context of the Action 'Support of Postdoctoral Researchers'. My current research project is entitled "PAUSANIAS: Ranked Spatial-keyword Search over Web-accessible Geotagged Data".
My main research interests include large-scale data management systems and query operators for business analysis.

George Tsatsanifos

I am a PhD Student at the School of Electrical and Computer Engineering of the National Technical University of Athens, Greece. I received my Diploma degree from the Department of Electronic and Computer Engineering of the Technical University of Crete. My research interests include indexing, spatial databases, ranking and diversity, distributed systems, peer-to-peer systems, RDF stores.

Timos Sellis

I am a Professor at RMIT University in the School of Computer Science and Information Technology, with an expertise in Data Management.

My interests lie in the general area of Data Management. I am particularly interested in Database Systems, data streams, peer-to-peer database systems, personalization, the integration of Web and databases, spatio-temporal database systems, Web Information Systems, On Line Analytical Processing, Data Warehouses.

Kjetil Nørvåg

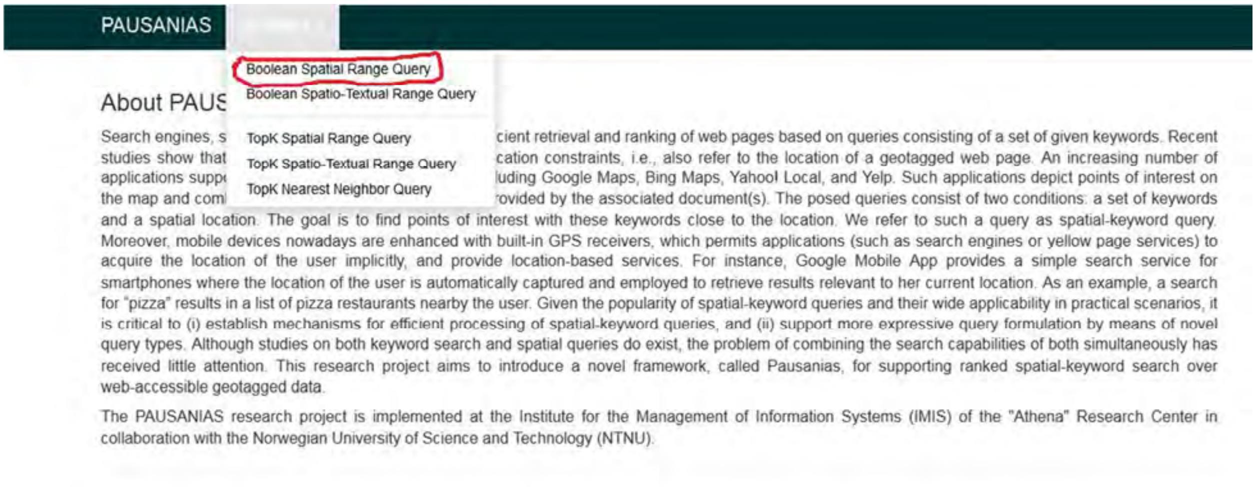
I'm working as a professor in the Department of Computer and Information Science here at NTNU. I'm affiliated with the Data and Information Management Group, where my main research interests include distributed and parallel database systems, query processing, information retrieval, and text mining.

Publications

- George Tsatsanifos, Akrivi Vlachou: "On Processing Top-k Spatio-Textual Preference Queries" in Proceedings of 18th International Conference on Extending Database Technology (EDBT), Brussels, Belgium, March 23-27, 2015.
- Orestis Gkorgkas, Akrivi Vlachou, Christos Doulkeridis and Kjetil Nørvåg: " Finding the Most Diverse Products using Preference Queries" in Proceedings of 18th International Conference on Extending Database Technology (EDBT), Brussels, Belgium, March 23-27, 2015.
- Ioanna Miliou, Akrivi Vlachou: "Location-Aware Tag Recommendations for Flickr" DEXA (1) 2014: 97-104.
- Orestis Gkorgkas, Akrivi Vlachou, Christos Doulkeridis and Kjetil Nørvåg: "Efficient Processing of Exploratory Top-k Joins", in Proceedings of 26th International Conference on Scientific and Statistical Database Management (SSDBM), Aalborg, Denmark, June 30 - July 2, 2014.
- Akrivi Vlachou, Christos Doulkeridis, Kjetil Nørvåg and Yannis Kotidis: "Branch-and-Bound Algorithm for Reverse Top-k Queries", in Proceedings of ACM International Conference on Management of Data (SIGMOD), New York, USA, June 22-27, 2013.
- Orestis Gkorgkas, Akrivi Vlachou, Christos Doulkeridis and Kjetil Nørvåg: "Discovering Influential Data Objects over Time" in Proceedings of 13th International Symposium on Spatial and Temporal Databases (SSTD), Munich, Germany, August 21-23, 2013.

Starting Page where the description of the website Pausanias is presented as well as a short bio of each researcher of the project. Finally there is a list of publications.

In order to search for Hotels that are at most 1000 meters from a position we have specified, we need to first select the webpage called Boolean Spatial Range Query.


PAUSANIAS
Boolean Spatial Range Query
Boolean Spatio-Textual Range Query
TopK Spatial Range Query
TopK Spatio-Textual Range Query
TopK Nearest Neighbor Query

About PAUS

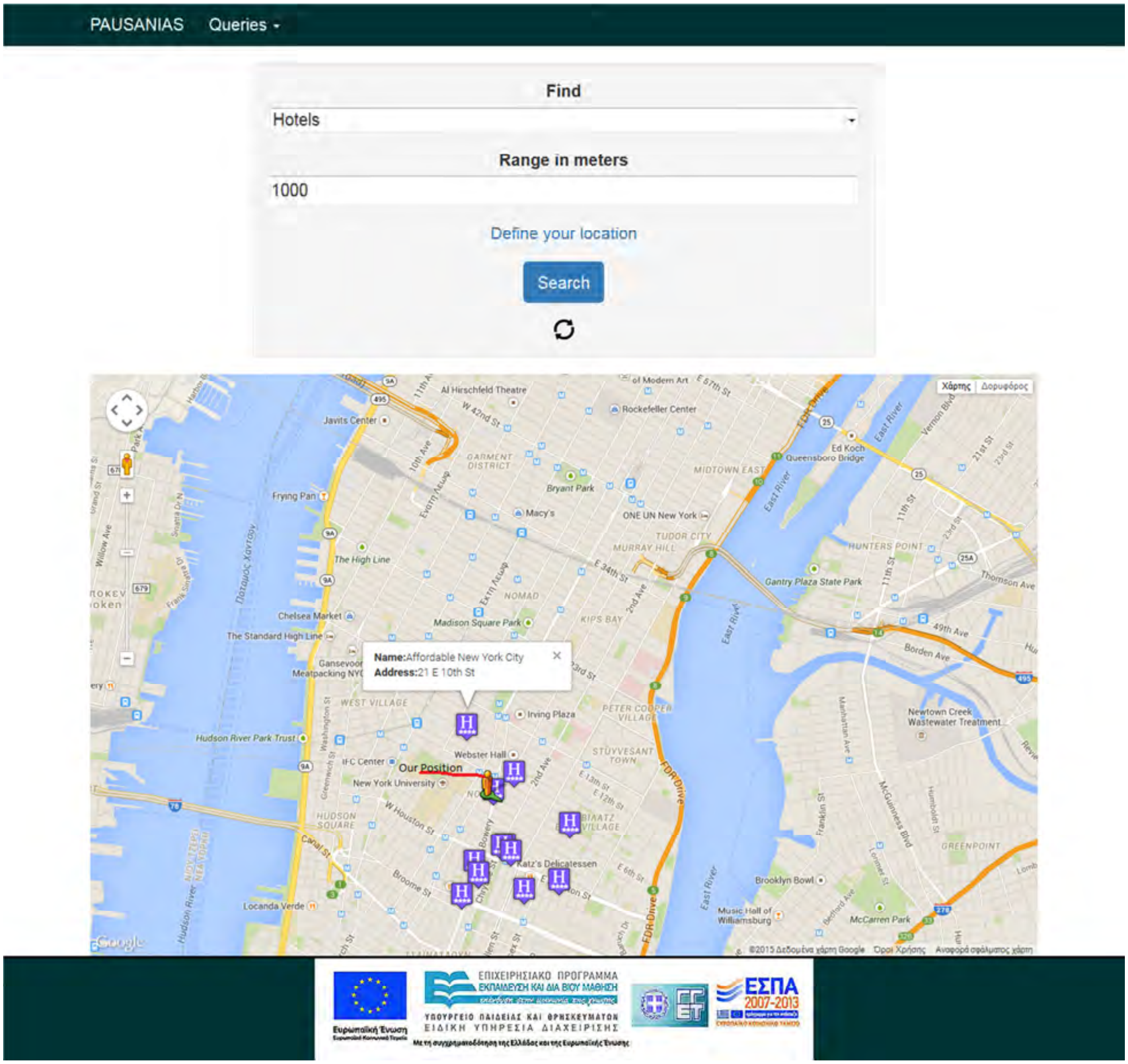
Search engines, s cient retrieval and ranking of web pages based on queries consisting of a set of given keywords. Recent
studies show that cation constraints, i.e., also refer to the location of a geotagged web page. An increasing number of
applications supp luding Google Maps, Bing Maps, Yahoo! Local, and Yelp. Such applications depict points of interest on
the map and com rovided by the associated document(s). The posed queries consist of two conditions: a set of keywords
and a spatial location. The goal is to find points of interest with these keywords close to the location. We refer to such a query as spatial-keyword query. Moreover, mobile devices nowadays are enhanced with built-in GPS receivers, which permits applications (such as search engines or yellow page services) to acquire the location of the user implicitly, and provide location-based services. For instance, Google Mobile App provides a simple search service for smartphones where the location of the user is automatically captured and employed to retrieve results relevant to her current location. As an example, a search for "pizza" results in a list of pizza restaurants nearby the user. Given the popularity of spatial-keyword queries and their wide applicability in practical scenarios, it is critical to (i) establish mechanisms for efficient processing of spatial-keyword queries, and (ii) support more expressive query formulation by means of novel query types. Although studies on both keyword search and spatial queries do exist, the problem of combining the search capabilities of both simultaneously has received little attention. This research project aims to introduce a novel framework, called Pausanias, for supporting ranked spatial-keyword search over web-accessible geotagged data.

The PAUSANIAS research project is implemented at the Institute for the Management of Information Systems (IMIS) of the "Athena" Research Center in collaboration with the Norwegian University of Science and Technology (NTNU).

Then we will specify what we are looking for, the range in meters from our position and finally our position on the map. In this example we are looking for “Hotels that are at most 1000 meters from our position”, which is in NOHO, New York. In each Hotel we can click on it and display information such as its name and its address.

European Union
European Social Fund
OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION, LIFELONG LEARNING AND RELIGIOUS AFFAIRS
MANAGING AUTHORITY
Co- financed by Greece and the European Union
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.



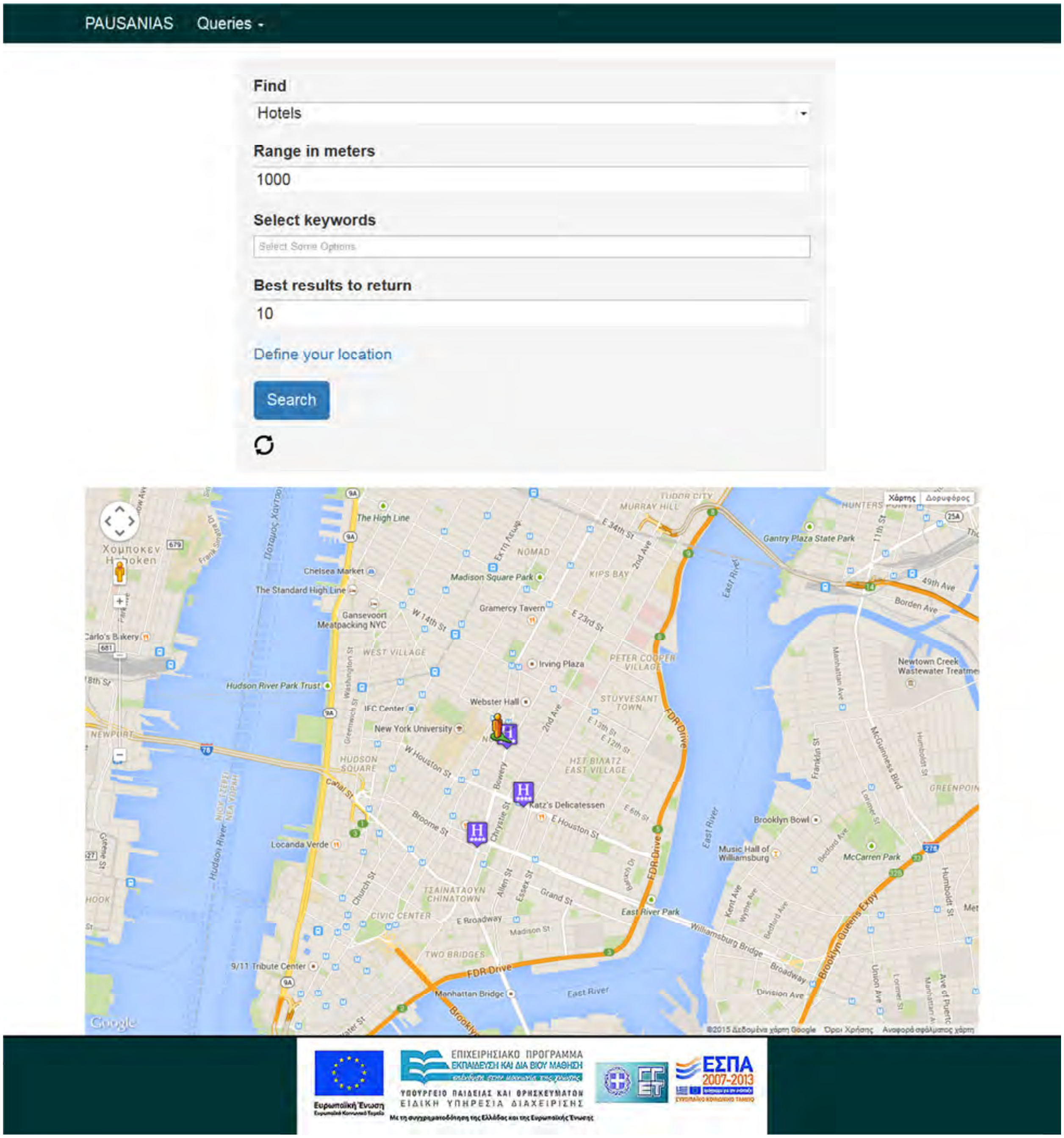
Find: Hotels

Range in meters: 1000

Name:Affordable New York City
Address:21 E 10th St

If we want to search for restaurants also in this area then all we have to do is to select a different dataset, in this case Restaurants, and click the “Search” button. When we want to clear out all the results we press the 2 arrows ⟳

Now in order to search for “top-10 Hotels that are within 1000 meters from our position and have parking”, we click the “Queries” option in the top of the page and go to the page “TopK Spatio-Textual Range Query”. In this page we select the dataset we want to search for, we denote the range in meters from our position, the keywords that we want to include in our search and finally our position.



PAUSANIAS Queries

Find: Hotels

Range in meters: 1000

Select keywords: Select Some Options

Best results to return: 10

Define your location

Search

All in all these are the effective web pages. By the word "effective" I mean that are in total 5 webpages devoted for each query. I described only two, namely "Boolean Spatial Range Query" and "TopK Spatio-Textual Range Query". The process of finding Points of Interest is the same in the remaining three webpages.

European Union European Social Fund
OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING investing in knowledge society
MINISTRY OF EDUCATION, LIFELONG LEARNING AND RELIGIOUS AFFAIRS MANAGING AUTHORITY
Co- financed by Greece and the European Union
NSRF 2007-2013 programme for development EUROPEAN SOCIAL FUND

The research project is implemented within the framework of the Action «Supporting Postdoctoral Researchers» of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.