Report on novel spatial-keyword search techniques

Akrivi Vlachou, Postdoctoral Researcher

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

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

Abstract

Nowadays an increasing amount of web-accessible information on spatial objects becomes available to the public every day. Apart from the spatial location of an object (e.g., a point of interest), additional descriptive information typically includes textual description as well as various ratings, often user generated. Modern applications employ spatio-textual queries, which take into account both the spatial location of an object and its textual similarity to retrieve the most relevant objects. However, existing applications provide a limited functionality to the users. For example, several meaningful queries cannot be expressed by existing approaches and motivate our novel prototype system. In the first chapter, we address this limitation by supporting ranked retrieval of objects of interest by taking into account the quality of facilities in their vicinity, but also their textual similarity to user defined keywords. In the second chapter, we analyze the properties of geotagged photos of Flickr, and propose novel location-aware tag recommendation methods. Both of the above techniques are novel spatial-keyword search methods.

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

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

Chapter 1

Preference Queries

Nowadays an increasing amount of web-accessible information on spatial objects becomes available to the public every day. Apart from the spatial location of an object (e.g., a point of interest), additional descriptive information typically includes textual description as well as various ratings, often user-generated. Modern applications employ spatio-textual queries, which take into account both the spatial location of an object and its textual similarity to retrieve the most relevant objects. Arguably, existing applications do not support effective spatio-textual retrieval of objects based on the quality of other facilities in their neighborhood. In this report, we address this limitation by supporting ranked retrieval of objects of interest by taking into account the quality of facilities in their vicinity, but also their textual similarity to user defined keywords. To this end, we propose a novel query type, termed *top-k spatio-textual preference query*, which is not currently supported by existing approaches. Moreover, we present a unified framework for query processing and we study many variations of the problem, namely for (i) range queries, (ii) influence queries, and (iii) nearest neighbor queries, and we de-

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

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

sign I/O efficient query processing algorithms. Among the benefits arising is the low programming cost at which the framework can be easily extended to cover other complex query types. We also suggest an alternative indexing approach that empowers search methods for the proposed query types. Last but not least, we evaluate all methods and their performance by means of experimental evaluation.

1.1 Introduction

An increasing number of applications support location-based queries, which retrieve the most interesting spatial objects based on their geographic location. Recently, spatio-textual queries have attracted much attention, as such queries combine location-based retrieval with textual information that describes the spatial objects. Most of the existing queries only focus on retrieving objects that satisfy a spatial constraint ranked by their spatial-textual similarity to the query point. However, users are quite often interested in spatial objects (*data objects*) based on the quality of other facilities (*feature objects*) that are located in their vicinity. Such features objects are typically described by non-spatial numerical attributes such as quality or ratings, in addition to the textual information that describes their characteristics. In this report, we propose a novel and more expressive query type, called *spatio-textual preference query*, for ranked retrieval of data objects based the textual relevance and the non-spatial score of feature objects in their neighborhood.

Consider for example, a tourist that looks for *“hotels that have nearby a good **Italian** restaurant that serves **pizza**”*. Fig. 1.1 depicts a spatial area containing hotels (data objects) and restaurants (feature objects). The quality of the restau-

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

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

rants based on existing reviews is depicted next to the restaurant. Each restaurant also has textual information, such as pizza or steak, which describes additional characteristics of the restaurant. The tourist specifies also a spatial constraint (in the figure depicted as a range around each hotel) to restrict the distance of the restaurant to the hotel. Obviously, the hotel h_2 is the best option for a tourist that poses the aforementioned query. In the general case, more than one type of feature objects may exist in order to support queries such as *"hotels that have nearby a good* ***Italian*** *restaurant that serves* ***pizza*** *and a cheap coffeehouse that serves* ***muffins****"*. Even though spatial preference queries have been studied before [23, 24, 18], their definition ignores the available textual information. In our example, the spatial preference query would correspond to a tourist that searches for *"hotels that are nearby a good restaurant"* and the hotel h_1 would always be retrieved, irrespective of the textual information.

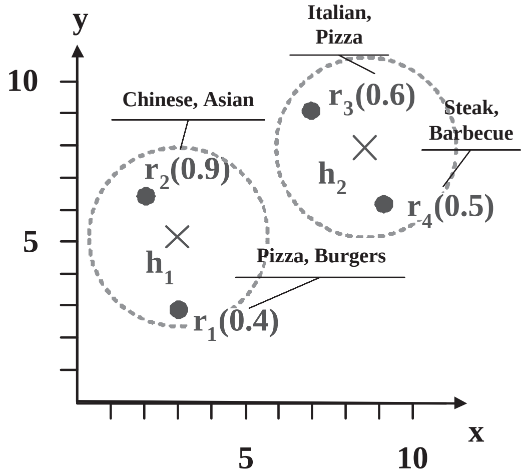


Figure 1.1: Associating spatial data objects with aggregate data.

In this report, we define top-k spatio-textual preference queries and provide a paradigm for processing this novel query type. A main difference compared to

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

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

traditional spatial preference queries [23, 24, 18] is that the rank of a data object changes depending on the query keywords, which renders techniques [18] that rely on materialization inappropriate. Most importantly, processing spatial preference queries is costly in terms of both I/O and execution time [23, 24], because it may require searching the spatial neighborhood of all data objects before reporting the top-k. Thus, extending spatial preference queries for supporting also textual information is challenging, since the new query type is more expensive due to the overhead imposed by the similarity of the query keywords to the facilities' textual descriptions.

We develop and evaluate two different paradigms for spatio-textual preference queries that rely on spatio-textual indexing techniques. The first method, called *Spatio-Textual Data Scan* (*STDS*), computes the spatio-textual score $\tau(p)$ of *each* data object $p \in O$ and then reports the k data objects with the highest score. The main focus of this algorithm is to reduce the cost required for computing the spatio-textual score of the data objects. A different strategy followed by our second method, called *Spatio-Textual Preference Search* (*STPS*), is to retrieve first highly ranked and relevant feature objects and then search for data objects nearby those feature objects. The main challenge we tackle with this approach is determining efficiently the best feature objects from all different feature sets that do not violate the spatial constraint.

The remainder of this report is organized as follows: Section 2.5 overviews the related work. In Section 1.3, we define the spatio-textual preference query. We present the experimental evaluation in Section 1.8 and we conclude in Section 1.9.

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

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

1.2 Related Work

Recently several approaches have been proposed for spatial-keyword search. In a previous seminal work [10], the problem of distance-first top-k spatial keyword search is studied. To this end, the authors propose an indexing structure (IR^2-Tree) that is a combination of an R-Tree and signature files. The IR-Tree was proposed in another conspicuous work [8, 15], which is an spatio-textual indexing approach that employs a hybrid index that augments the nodes of an R-Tree with inverted indices. The inverted index at each node refers to a pseudo-document that represents all the objects under the node. During query processing, the index is exploited to retrieve the top-k data objects, defined as the k objects that have the highest spatio-textual similarity to a given data location and a set of keywords. Moreover, in [17] the *Spatial Inverted Index (S2I)* was proposed for processing top-k spatial keyword queries. The S2I index maps each keyword to a distinct aggregated R-Tree or to a block file that stores the objects with the given term. All these approaches focus on ranking the data objects based on their spatio-textual similarity to a query point and some keywords. This is different from our work, which ranks the data objects based on the quality and relevance of the facilities in their spatial neighborhood.

Prestige-based spatio-textual retrieval was studied in [6]. The proposed query takes into account both location proximity and prestige-based text relevance. The m-closest keywords query [25] aims to find the spatially closest data objects that match with the query keywords. The authors in [7] study the spatial group keyword query that retrieves a group of data objects such that all query keywords appear in at least one data object textual description and such that objects are

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

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

nearest to the query location and have the lowest inter-object distances. These approaches focus on finding a set of data objects that are close to each other and relevant to a given query, whereas in this report we rank the data objects based on the facilities in their spatial neighborhood.

Ranking of data objects based on their spatial neighborhood without supporting keywords has been studied in [22, 9, 23, 24, 18]. Xia *et al.* studied the problem of retrieving the top-k most influential spatial objects [22], where the score of a data object p is defined as the sum of the scores of all feature objects that have p as their nearest neighbor. Yang *et al.* studied the problem of finding an optimal location [9], which does not use candidate data objects but instead searches the space. Yiu *et al.* first considered computing the score of a data object p based on feature objects in its spatial neighborhood from multiple feature sets [23, 24] and defined top-k spatial preference queries. In another line of work, a materialization technique for top-k spatial preference queries was proposed in [18] which leads to significant savings in both computational and I/O cost during query processing. The main difference is that our novel query is defined in addition by a set of keywords that express desirable characteristics of the feature objects (like “pizza” for a feature object that represents a restaurant).

Finally, spatio-textual similarity joins were recently studied in [3]. Given two data sets, the query retrieves all pairs of objects that have spatial distance smaller than a given value and at the same time a textual similarity that is larger than a given value. This differs from the top-k spatio-textual preferences query, because the spatio-textual similarity join does not rank the data objects and some data objects may appear more than once in the result set.

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

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

1.3 Problem Statement

Given an *object* dataset O and a set of c *feature* datasets $\{F_i \mid i \in [1, c]\}$, in this report, we address the problem of finding k data objects that have in their spatial proximity highly ranked feature objects that are relevant to the given query keywords. Each data object $p \in O$ has a spatial location. Similarly, each feature object $t \in F_i$ is associated with a spatial location but also with a *non-spatial score* $t.s$ that indicates the goodness (quality) of t and its domain of values is the range $[0, 1]$. Moreover, t is described by set of keywords $t.\mathcal{W}$ that capture the textual description of the feature object t. Figure 1.2 depicts an example of a set of feature objects that represent restaurants and shows the non-spatial score and the textual information. Table 1.1 provides an overview of the symbols used in this report.

Symbol	**Description**
O	Set of data objects
p	Data object, $p \in O$
c	Number of feature sets
F_i	Feature sets, $i \in [1, c]$
t	Feature object, $t \in F_i$
$t.s$	Non-spatial score of t
$t.\mathcal{W}$	Set of keywords of t
$dist(p, t)$	Distance between p and t
$sim(t, \mathcal{W})$	Textual similarity between t and $\mathcal{W}$
$s(t)$	Preference score of t
$\tau_i(p)$	Preference score of p based on F_i
$\tau(p)$	Spatio-textual preference score of p

Table 1.1: Overview of symbols.

The goal is to find data objects located nearby feature objects that (i) are of high quality and (ii) have a high similarity to the user specified keywords. Thus, the score of the feature object t captures not only the non-spatial score of the

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

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

	name	rating	x	y	textual description
r_1	Beijing Restaurant	0.6	1	2	Chinese, Asian
r_2	Daphne's Restaurant	0.5	4	1	Greek, Mediterranean
r_3	Espanol Restaurant	0.8	5	8	Italian, Spanish, European
r_4	Golden Wok	0.8	2	3	Chinese, Buffet
r_5	John's Pizza Plaza	0.9	8	4	Pizza, Sandwiches, Subs
r_6	Ontario's Pizza	0.8	7	6	Pizza, Italian
r_7	Oyster House	0.8	6	10	Seafood, Mediterranean
r_8	Small Bistro	1.0	3	7	American, Coffee, Tea, Bistro

Figure 1.2: Feature Set (Restaurants)

	name	rating	x	y	textual description
c_1	Bakery & Cafe	0.6	4	1	Cake, Bread, Pastries
c_2	Coffee House	0.5	4	7	Cappuccino,Toast, Decaf
c_3	Coffe Time	0.8	3	10	Cake, Toast, Donuts
c_4	Cafe Ole	0.6	6	2	Cappuccino, Iced Coffee, Tea
c_5	Royal Coffe Shop	0.9	5	5	Muffins, Croissants,Espresso
c_6	Mocha Coffe House	1.0	10	3	Macchiato, Espresso, Decaf
c_7	The Terrace	0.7	6	9	Muffins, Pastries, Espresso
c_8	Espresso Bar	0.4	7	6	Croissants, Decaf, Tea

Figure 1.3: Feature Set (Coffeehouses)

feature, but its textual similarity to a user specified set of query keywords.

Definition 1 *The **preference score** $s(t)$ **of feature object** t based on a user-specified set of keywords $\mathcal{W}$ is defined as $s(t) = (1-\lambda)\cdot t.s+\lambda\cdot sim(t, \mathcal{W})$, where $\lambda \in [0, 1]$ and $sim()$ is a textual similarity function.*

The textual similarity between the keywords of the feature and the set $\mathcal{W}$ is measured by $sim(t, \mathcal{W})$ and its domain of values is the range $[0, 1]$. The parameter λ is the smoothing parameter that determines how much the score of the feature should be influenced by the textual information. For the rest of the report, we assume that the textual similarity is equal to the Jaccard similarity between the keywords of the feature objects and the user-specified keywords: $sim(t, \mathcal{W}) = \frac{|t.\mathcal{W} \bigcap \mathcal{W}|}{|t.\mathcal{W} \bigcup \mathcal{W}|}$.

For example, consider the restaurants depicted in Figure 1.2. Given a set of keywords $\mathcal{W} = \{italian,\ pizza\}$ and $\lambda = 0.5$ the restaurant with the highest preference score is *Ontario's Pizza* with a preference score $s(r_6) = 0.9$, while the

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

Co- financed by Greece and the European Union

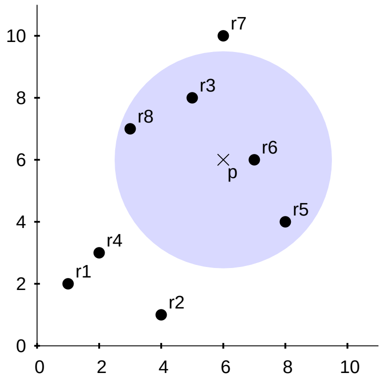


Figure 1.4: An accute example of a $STPQ$ query.

score of *Beijing Restaurant* is $s(r_1) = 0.3$, since none of the given keywords are included in the description of *Beijing Restaurant*.

Given a spatio-textual preference query Q defined by an integer k, a range r and c-sets of keywords $\mathcal{W}_i$, the preference score of a data object $p \in O$ based on a feature set F_i is defined by the scores of feature objects $t \in F_i$ in its spatial neighborhood, whereas the overall spatio-textual score of p is defined by taking into account all feature sets F_i, $1 \leq i \leq c$.

Definition 2 *The **preference score** $\tau_i(p)$ **of data object** p based on the feature set F_i is defined as:* $\tau_i(p) = max\{s(t) \mid t \in F_i : dist(p,t) \leq r \text{ and } sim(t, \mathcal{W}_i) > 0\}$.

The $dist(p,t)$ denotes the spatial distance between data object p and feature object t and, in this report, we employ the Euclidean distance function. Continuing the previous example, Figure 1.4 shows the spatial location of the restaurants in Figure 1.2 and a data point p that represents a hotel. The preference score of p based on the restaurants in its neighborhood (assuming $r = 3.5$ and $\mathcal{W} =$

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

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

$\{italian,\ pizza\}$) is equal to the score of r_6 ($\tau_i(p) = s(r_6) = 0.9$), which is the best restaurant in the neighborhood of p.

Definition 3 *The overall* ***spatio-textual score*** $\tau(p)$ ***of data object*** p *is defined as:* $\tau(p) = \sum_{i \in [1,c]} \tau_i(p)$.

Figure 1.3 shows a second feature set that represents coffeehouses. For a tourist that looks for a good hotel that has nearby a good Italian restaurant that serves pizza and a good coffeehouse that serves espresso and muffins, the score of p would be $\tau(p) = s(r6) + s(c5) = 0.9 + 0.78233 = 1.6833$.

Problem 1 Top-k Spatio-Textual Preference Queries(STPQ)*: Given a query* Q*, defined by an integer* k*, a radius* r *and* c*-sets of keywords* $\mathcal{W}_i$*, find the* k *data objects* $p \in O$ *with the highest spatio-textual score* $\tau(p)$*.*

1.4 Indexing Principles

The main difference of top-k spatio-textual preference queries to traditional spatio-textual search is that the ranking of a data object does not depend only on spatial location and textual information, but also on the non-spatial score of the feature object. In particular, the preference score $s(t)$ of feature object t is defined by its textual description and its non-spatial score, while the spatial location is used as a filter for computing the preference score $\tau_i(p)$ of data object p. Thus, efficient indexing of the textual description and the non-spatial score of feature objects is a significant factor for designing efficient algorithms for the STPQ query.

Indexing principles: In this report, we assume that the data objects O are indexed by an R-Tree, denoted as *rtree*. However, for the feature objects, it is impor-

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

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

tant that the non-spatial score and the textual description are indexed additionally. Each dataset F_i can be indexed by any spatio-textual index that relies on a spatial hierarchical index (such as the R-Tree). However, each entry e of the index must in addition maintain: (i) the maximum value of $t.s$ of any feature t in the sub-tree, denoted as $e.s$, and (ii) a summary ($e.\mathcal{W}$) of all keywords of any feature t in the sub-tree. To ensure correctness of our algorithms, the main property that needs to hold for any t stored in the sub-tree rooted by the entry e is

$$s(e) = (1 - \lambda) \cdot e.s + \lambda \cdot sim(e, \mathcal{W}) \geq s(t)$$

The above property guarantees that the preference score $s(t)$ of a feature object t is bounded by the score $s(e)$ of its ancestor node e. The efficiency of the algorithms directly depends on the tightness of this bound. In turn, this depends on the similarity between the textual description and the non-spatial score of the features objects that are indexed in the same node.

The remaining question is whether existing spatio-textual indexes (or adaptations thereof) can be employed to support the STPQ query. For example, if the IR^2-tree is augmented to also store in all nodes the maximum non-spatial score of the subsumed feature objects, then it can be used for supporting top-k spatio-textual preference queries. In this case, the summary of the keywords of an entry can be a signature of the keywords in the sub-tree and the above property holds as the signature of a non-leaf entry is the superimposition of all signatures of its child entries. Unfortunately, in this case, the index is not build by grouping together in the same node feature objects with similar textual description and non-spatial score, thus it leads to loose bounds. Consequently, similar or relevant objects are

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

Co- financed by Greece and the European Union

stored throughout the index, instead of being clustered in the same node, which hinders efficient pruning.

On the other hand, even though there exist indexes that take the textual description into account during the index construction, such the IR-Tree [8], these indexes assume that each keyword of an object is associated with a real value. Differently, in this report, we follow a Boolean model for the keywords and using an index such as the IR-Tree would lead to store redundant information. In contrast, we design a novel indexing method for STPQ queries that captures their salient characteristics and exploits all aspects, namely the spatial, the non-spatial score and the textual information of the feature objects.

Indexing based on Hilbert Mapping: Regarding the textual description of feature objects, let w denote the number of distinct keywords in the vocabulary. Then, for each feature t the keyword $t.\mathcal{W}$ can be represented as a binary vector of length w. For instance, assuming a vocabulary $\{pizza, burger, spaghetti\}$, we can use an active bit to declare the existence of the *"pizza"* keyword at the first place, *"burger"* at the second, and *"spaghetti"* at the last. Moreover, we suggest a mapping of the binary vector to a Hilbert value, denoted as $\mathcal{H}(t.\mathcal{W})$. For the above w=3 keywords, the defined order is $000, 010, 011, 001, 101, 111, 110$ and 100. The benefit of this order is that it ensures us that vectors with distance 1 have only one different keyword, while if the distance is w', then the maximum number of different keywords is bound by w'. This means that consecutive vectors in the afore-described order have only few different keywords, which means that objects with sequential Hilbert values are highly similar also based on the Jaccard similarity function.

Using the Hilbert mapping of the textual information, each feature object t can

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

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

be represented as a point in the 4-dimensional space $\{t.x$, $t.y$,$t.s$,$\mathcal{H}(t.\mathcal{W})\}$. Any spatial index, such as a traditional R-Tree, built on the mapped 4-dimensional space fulfills the above property and can be used for answering STPQ queries efficiently. The reason is that this indexing mechanism can identify effectively the promising branches of the hierarchical structure at a low cost, since during the index construction the similarity of the spatial location, the non-spatial score as well as the textual description is taken into account. We call this indexing technique *SRT-index*. In terms of structure, the SRT-index resembles a traditional R-Tree that it is built on the Hilbert value of the keywords, the spatial location and the non-spatial score of the feature objects altogether. Notably, the exact spatial index used for indexing the mapped space does not affect the correctness of our algorithms, but only their performance. In our experimental evaluation, we use bulk insertion [12] on our novel indexing technique.

We should highlight that an important benefit of the SRT-index is that it also takes into account the spatial location, which combined with the textual information and the non-spatial score, achieves a beneficial grouping of feature objects for query processing. Even though the dominant factors for computing the score of a feature objects are its non-spatial score as well as its textual relevance to the given query, the spatial location is also important for discarding early feature objects that do not satisfy the spatial constraint of the STPQ query. Thus, if the spatial location would have been ignored by the index, this would cause an I/O overhead which is associated with its filtering properties and query selectivity.

To summarize, the SRT-index overcomes the difficulty that other indexing approaches face, being unable to identify in advance what are the branches of the index that store highly ranked and relevant feature objects to the query. More im-

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

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

portantly, the search methods proposed in the following sections capitalize on this specialized indexing scheme to boost the performance of query processing.

1.5 Spatio-Textual Data Scan (STDS)

Our baseline approach, called spatio-textual data scan (*STDS*), computes the spatio-textual score $\tau(p)$ of *each* data object $p \in O$ and then reports the k data objects with the highest score. Algorithm 1 shows the pseudocode of *STDS*. The R-Tree that indexes the data objects is traversed once and for each object the score $\tau(p)$ is computed. In more detail, for a data object p, its score $\tau_i(p)$ for every feature set F_i is computed (lines 3-4). The details on this computation for range queries are described in Algorithm 2 that will be presented in the sequel. Interestingly, for some data points p we can avoid computing $\tau_i(p)$ for all feature sets. This is feasible because we can determine early that some data objects cannot be in the result set P. To achieve this goal, we define a threshold τ which is the k-th highest score of any data object processed so far. In addition, we define an upper bound $\widehat{\tau}(p)$ for the spatio-textual preference score $\tau(p)$ of p, which does not require knowledge of the preference scores $\tau_i(p)$ for all feature sets F_i: $\widehat{\tau}(p) = \sum\limits_{i\in[1,c]} \{ \begin{array}{ll} \tau_i(p), & if\ \tau_i(p)\ is\ known \\ 1, & otherwise \end{array}$. The algorithm tests the upper bound $\widehat{\tau}$ based on the already computed $\tau_i(p)$ against the current threshold (line 4). If $\widehat{\tau}$ is smaller than the current threshold, the remaining score computations are avoided. After computing the score of p, we test whether it belongs to P (line 5). If this is case, the result set P is updated (line 6), by adding p to it and removing the data object with the lowest score (in case that $|P| > k$). Finally, if at least k data

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

objects have already been added to P, we update the threshold based on the k-th highest score (line 8).

Algorithm 1: *Spatio-Textual Data Scan* (STDS)

```
   Input: Query $Q = (k, r, \{\mathcal{W}_i\})$
   Output: Result set $P = \{P[1] \dots P[k]\}$ sorted based on $\tau(p)$
1  $P = \emptyset$; $\tau = -1$;
2  foreach $p \in O$ do
3      for $i = 1 \dots c$ do
4          if $\widehat{\tau}(p) > \tau$ then  $\tau_i(p) = F_i.computeScore(Q, p)$ ;
5          ;
6      if $\tau(p) > \tau$ then
7          update($P$) ;
8          if $|P| \geq k$ then
9              $\tau = \tau(P[k])$ ;
10 return $P$ ;
```

The remaining challenge is to compute efficiently the score based on the spatio-textual information of the feature objects. The goal is to reduce the number of disk accesses for feature objects that are necessary for computing the score of each element $p \in O$. Algorithm 2 shows the computation of preference score $\tau_i(p)$ for range queries for feature set F_i. First, the root entry is retrieved and inserted in a heap (line 1). The heap maintains the entries sorted based on $s(e)$. In each iteration (lines 2-11), the entry e with the highest score is processed, following a best-first approach. If e is a data point and within distance r from p (line 5), then the score $\tau_i(p)$ of p has been found and is returned (line 7). If e is not a data point, then we expand it only if it satisfies the query constraints (line 9). More detailed, if the minimum distance of e to p is smaller or equal to r and its textual similarity is larger than 0, e is expanded and its child entries are added to the heap (line 11). Otherwise, the entire sub-tree rooted at e can be safely pruned.

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

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

Algorithm 2: *Spatio-Textual Score Computation on F_i ($computeScore(Q, p)$)*

```
Input: Query $Q$, data object $p$
Output: Score $\tau_i(p)$
$heap$.push($F_i$.root);
while (not $heap$.isEmpty()) do
    $e \leftarrow heap$.pop() ;
    if $e$ is a data object then
        if ($dist(p, e) \leq r$) then
            $\tau_i(p) = s(e)$ ;
            return $\tau_i(p)$
    else
        if ($mindist(p, e) \leq r$) and ($sim(e, \mathcal{W}_i) \geq 0$) then
            for childEntry in $e$.childNodes do
                $heap$.push(childEntry) ;
```

Correctness and Efficiency: Algorithm 2 always reports the correct score $\tau_i(p)$. The sorted access of the entries, combined with the property that the score of the entry is an upper bound, ensures the correctness of Algorithm 2. Moreover, it can be shown that Algorithm 2 expands the minimum number of entries, in the sense that if an entry that is expanded by Algorithm 2 was not expanded, it could lead to computing a wrong score. This is because only entries with score higher than any processed feature object are expanded, and such entries may contain in their sub-tree a feature object with score equal to the score of the entry.

Performance improvements: The performance of *STDS* can be improved by processing the score computations in a batch. Instead of a single data object p, a set of data objects $\mathcal{P}$ can be given as an input to the Algorithm 2. Then, an entry is expanded if the distance for *at least* one p in $\mathcal{P}$ is smaller than r. When a feature object is retrieved, for any p for which the distance is smaller than r the score is

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

Co- financed by Greece and the European Union

computed and those data objects p are removed from $\mathcal{P}$. The same procedure is followed until either the heap or $\mathcal{P}$ is empty. Algorithm 1 can be easily modified to invoke Algorithm 2 for all data objects in the same leaf entry of *rtree*. For sake of simplicity, we omit the implementation details, even though we use this improved modification in our experimental evaluation.

1.6 Spatio-Textual Preference Search (STPS)

The second approach, called Spatio-Textual Preference Search (*STPS*), follows a different strategy than the *STDS* algorithm from Algorithm 1. It first computes highly ranked and relevant feature objects, and then, retrieves the data objects in their neighborhood. In a nutshell, the goal here is to find sets of feature objects $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$ such that $t_i \in F_i$, where $1 \leq i \leq c$, and the score of each t_i is as high as possible. Intuitively, if we find a neighborhood in which highly ranked feature objects exist, then the neighboring data objects are naturally highly ranked, as well.

In the general case, a data object may be highly ranked even in the case where some kind of feature object does not exists in its neighborhood. For example, consider the extreme case where all data objects have only one type of feature object in their spatial neighborhood. For ease of presentation, we denote as $\emptyset$ a virtual feature object for which it holds that $dist(p, \emptyset) = 0$, $dist(t_i, \emptyset) = 0$ and $s(\emptyset) = 0\ \forall t_i, p$. This virtual feature object is used for presenting unified definitions for the case where the spatio-textual score of the top-k data objects are defined based on less than c feature objects.

Definition 4 *A valid combination of feature objects is a set* $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$

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

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

such that (i) $\forall i\ t_i \in F_i$ *or* $t_i = \emptyset$ *and (ii)* $dist(t_i, t_j) \leq 2r \forall i, j$*. The score of the valid combination* $\mathcal{C}$ *is defined as* $s(\mathcal{C}) = \sum_{1 \leq i \leq c} s(t_i)$*.*

The following lemma proves that it is sufficient to examine only the valid combinations of feature objects $\mathcal{C}$ in order to retrieve the result set of a top-k spatio-textual preference query.

Lemma 1 *The score of any data object* $p \in O$ *is defined by a valid combination of feature objects* $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$*, i.e.,* $\forall p : \exists \mathcal{C} = \{t_1, t_2, \ldots, t_c\}$ *such that* $\tau(p) = s(\mathcal{C})$

Proof. Let us assume that there exists p such that: $\tau(p) = \sum_{i \in [1,c]} \tau_i(p)$ with $\tau_i(p) = \{s(t_i) \mid t_i \in F_i : dist(p, t_i) \leq r \text{ } and \text{ } sim(t_i, \mathcal{W}_i) > 0\}$ and $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$ is not a valid combination of feature objects. Since $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$ is not a valid combination of feature objects, there exists $1 \leq i \neq j \leq c$ such that $dist(t_i, t_j) > 2r$ but also $dist(p, t_i) \leq r$ and $dist(p, t_j) \leq r$. Based on the triangular inequality it holds: $dist(t_i, t_j) \leq dist(p, t_i) + dist(p, t_j) \leq r + r \leq 2r$, which is a contradiction.

1.6.1 $STPS$ Overview

Algorithm 3: *Spatio-Textual Preference Search (STPS)*

```
Input: Query $Q$
Output: Result set $P$ sorted based on $\tau(p)$
while ($|P| \leq k$) do
    $\mathcal{C} = nextCombination(Q)$ ;
    $P = P \cup$ getObjects($\mathcal{C}$) ;
return $P$ ;
```

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

Co- financed by Greece and the European Union

Algorithm 3 describes the *STPS* algorithm. We assume that there exists an iterator that returns the valid combinations of feature objects sorted based on their score (we discuss the details on the implementation of the iterator in the following subsection). Line 2 retrieves the next combination, i.e., the valid combination that has the highest score of all valid combinations that have not been processed yet. Thereafter, in line 3, we retrieve all data points in the spatial neighborhood of these features. Data objects that have already been previously retrieved are discarded, while the remaining data objects p have a score $\tau(p) = s(\mathcal{C})$ and can be returned to the user incrementally. If at least k data objects have been returned to the user, the algorithm terminates without retrieving the remaining combinations of feature objects. Differently to the *STDS* algorithm, *STPS* retrieves only the data objects that certainly belong to the result set.

In line 3 *getObjects($\mathcal{C}$)* is called to retrieve from the R-Tree $rtree$ all data objects in the neighborhood of the feature objects in $\mathcal{C}$. This method starts from the root of the $rtree$ and processes its entries recursively. Entries e for which $\exists i$ such that $t_i \in \mathcal{C}$ with $dist(e, t_i) > r$ are discarded. The remaining entries are expanded until all objects p for which it holds $\forall i \; dist(p, t_i) \leq r$ are retrieved.

Consider for example the feature sets depicted in Figure 1.2 and in Figure 1.3. Given a query with $r = 3.5$, $\mathcal{W}_1 = \{italian,\ pizza\}$ and $\mathcal{W}_2 = \{espresso,\ muffins\}$, the restaurant and the coffeehouse with the highest scores are $r6$ and $c5$ respectively. Since it holds that $dist(r6, c5) \leq 2r$, the set $\mathcal{C} = \{r6, c5\}$ is a valid combination of feature objects. Assume that the set of data objects is $O = \{p1, p2, \ldots, p10\}$ as depicted in Figure 1.5. For the data objects $p6$, $p9$ and $p10$ it holds that $dist(pi, c5) \leq r$ and $dist(pi, r6) \leq r$, and their spatial-textual score is $\tau(p6) = \tau(p9) = \tau(p10) = 1.6833$. These data objects are guaranteed to be the

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

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

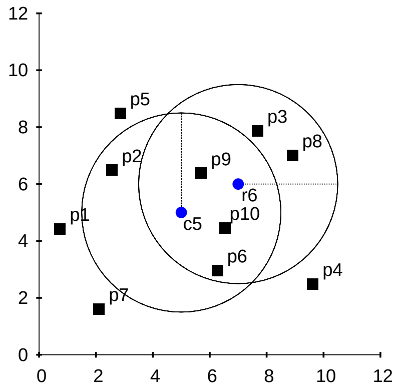


Figure 1.5: Finding the data objects within qualifying distance from $\mathcal{C} = \{r6, c5\}$.

highest ranked data objects and can be immediately returned to the user. If $k \leq 3$, our algorithm terminates without examining other feature combinations.

The remaining challenge is how to retrieve efficiently the valid combinations of feature objects $\mathcal{C}$ sorted based on their score $s(\mathcal{C})$, which is described in the following subsection.

Spatio-Textual Feature Objects Retrieval

Algorithm 4 shows our algorithm for retrieving the valid combinations $\mathcal{C}$ of feature objects sorted based on their spatio-textual score $s(\mathcal{C})$. The different spatio-textual indices that store feature objects of the feature sets F_i are accessed and the feature objects t_i are retrieved based on their score $s(t_i)$ that aggregates their non-spatial score, but also their textual similarity to the query keywords. The retrieved feature objects are maintained in a list $\mathcal{D}_i$ and are used to produce valid combinations $\mathcal{C}$ of feature objects. The main property of the spatio-textual index namely that the score $s(e)$ of an entry e is an upper bound of the score of any feature object t in the sub-tree pointed by e, enables efficient retrieval of the feature objects t_i sorted by

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

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

Algorithm 4: *Spatio-Textual Feature Objects Retrieval* $(nextCombination(Q))$

```
   Input: Query $Q$
   $heap_i$: heap maintaining entries of $F_i$
   $heap$: heap maintaining valid combinations of feature objects
   $\mathcal{D}_i$: set of feature objects of $F_i$
   Output: $\mathcal{C}$: valid combination with highest score
1  while ($\exists i$ : not $heap_i$.isEmpty()) do
2      $i \leftarrow$ nextFeatureSet() ;
3      $e_i \leftarrow heap_i$.pop() ;
4      while (not $e_i$ is a data object) do
5          for childEntry in $e_i$.childNodes do
6              $heap_i$.push(childEntry) ;
7          $e_i \leftarrow heap_i$.pop() ;
8      $\mathcal{D}_i = \mathcal{D}_i \cup e_i$ ;
9      $heap$.push(validCombinations($\mathcal{D}_1, \cdots, e_i, \cdots, \mathcal{D}_c$)) ;
10     $min_i = s(e_i)$ ;
11     $\tau = max_{1 \leq j \leq c}(max_1 + \cdots + min_j + \cdots + max_c)$ ;
12     $\mathcal{C} \leftarrow heap$.top() ;
13     if (score($\mathcal{C}$) $\geq \tau$) then
14         $heap$.pop() ;
15         return $\mathcal{C}$;
```

$s(t_i)$, since only the entry with the highest score needs to be expanded. The first feature object that is retrieved is guaranteed to be the next feature object with the highest score (lines 3-7). The remaining challenge is to find efficiently the valid combinations $\mathcal{C}$ of feature objects with the highest score.

We denote as max_i the maximum score of $\mathcal{D}_i$ and min_i the minimum score of $\mathcal{D}_i$. Thus, min_i represents the best potential score of any feature object of F_i that has not been processed yet. Moreover, in Alg. 4 the variables $heap_i$, $\mathcal{D}_i$, max_i, min_i, and $heap$ are global variables. They are initialized as following $heap_i$: the root of F_i, $\mathcal{D}_i = \emptyset$ and $heap = \emptyset$, $min_i = \infty$. Variable max_i is the score of the

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

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

highest ranked feature object of F_i and is set the first time the F_i index is accessed.

In each iteration Alg. 4 retrieves a feature object e_i that belongs to the feature set F_i (lines 3-7). The entries of the spatio-textual index responsible for the feature objects of F_i are inserted in $heap_i$, which keeps the entries e sorted based on $s(e)$. Moreover, for sake of simplicity, we assume that $heap_i.pop()$ will return the virtual feature object $t_i = \emptyset$ before $heap_i$ gets empty. When an entry is retrieved that corresponds to a feature object, e_i is inserted in the list $\mathcal{D}_i$ (line 8). Then, new valid combinations $\mathcal{C}$ are created by combining e_i with the previously retrieved feature objects t_j maintained in the lists $\mathcal{D}_j$ (line 9). For this, the method $validCombinations$ is called, which returns all combinations of the objects in $\mathcal{D}_j$ and e_i, by discarding combinations for which the condition $dist(t_i, t_j) \leq 2r\ \forall i, j$ does not hold. The new valid combinations are inserted in the $heap$ (line 9) that maintains the valid combinations sorted based on their score $s(\mathcal{C})$.

Alg. 4 employs a thresholding scheme to determine if the current best valid combination can be returned as the valid combination with the highest score. The threshold τ represents the best score of any valid combination of feature objects that has not been examined yet. The best score derives by assuming that the next feature object t_j retrieved from F_j has the same score $s(t_j)$ with the previously retrieved feature object of F_j that is equal to min_j, Since the feature objects are accessed sorted based on $s(t_j)$ this value is an upper bound. Obviously, for the remaining feature sets we assume that the new feature object t_j is combined with the feature objects that have the highest score. Thus, $\tau = max_{1 \leq j \leq c}(max_1 + \cdots + min_j + \cdots + max_c)$ (line 11) is an upper bound of the score for any valid combination that has not been examined yet. In line 13, we test whether the best combination of feature objects in the $heap$ has a score higher or equal to the

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

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

threshold τ. If so, the best combination in the heap is the next valid combination with the best score.

The order in which the feature objects of different feature sets are retrieved is defined by a pulling strategy, i.e., $nextFeatureSet()$ returns an integer between 1 and c and defines the pulling strategy. In addition, $nextFeatureSet()$ never returns i if $heap_i$ is empty.

Pulling Strategy: In the following, we proposed an advanced pulling strategy that prioritize retrieving from feature sets that have higher potential to retrieve the next valid combination $\mathcal{C}$.

Definition 5 *Given c sets of feature objects $\mathcal{D}_i$, the prioritized pulling strategy returns m as the next feature set such that $\tau = max_1 + \cdots + min_m + \cdots + max_c$.*

The main idea of the prioritized pulling strategy is that in each iteration the feature set F_m that is responsible for the threshold value τ is accessed. It is obvious that the only way to reduce τ is to reduce the min_m, since retrieving from the remaining feature sets cannot reduce τ. In addition, retrieving from the remaining feature sets cannot produce a new valid combination $\mathcal{C}$ of feature objects that has a higher score than τ. Thus, retrieving the next tuple from the feature set F_m can reduce the threshold τ and may produce new valid combinations that have a score equal to the current threshold.

1.7 Variants of Top-k Spatio-Textual Preference Queries

In this section, we extend our algorithms for processing spatio-textual preference queries based on alternative score definitions under a unified framework. We pro-

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

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

vide formal definitions for the alternative score definitions, namely *influence preference score* and *nearest neighbor preference score*. Moreover, we discuss for all query types the necessary modifications of algorithms for query processing and certain optimizations. Above all, we address these query types under a generalized framework that can be further expanded accordingly at a low programming cost.

For all variants the *STDS* algorithm, as defined in Algorithm 1 can be easily adapted. Only the function $computeScore(Q, p)$ must implemented according to the definition of each score variant. Thus, in Algorithm 2 each entry in line 11 will be prioritized according to score variant. In addition, the range restriction is upheld for the minimum distance of the index nodes. No further modifications are needed, thus in the following we focus on the modifications and optimizations needed for *STPS* algorithm.

1.7.1 Influence-Based $STPQ$ Queries

In contrast to the preference score defined in Definition 1 (in the following referred as range score), in this section we define an alternative score that does not pose a hard constrain on the distance, but reduces the score based on the distance instead. We call this score influence score.

Definition 6 *The* ***influence preference score*** $\tau_i(p)$ ***of data object*** p *based on the feature set* F_i *is defined as:* $\tau_i(p) = max\{s(t) \cdot 2^{\frac{-dist(p,t)}{r}} \mid t \in F_i : \; sim(t, \mathcal{W}_i) > 0\}$.

The overall spatio-textual score $\tau(p)$ of data object p is defined as for the case of the range score, and the query returns the k objects with the highest score.

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

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

Algorithm 5: *STPS for influence score*

```
Input: Query $Q$
Output: Result set $P$ sorted based on $\tau(p)$
$\tau = 0$ ;
$score = -1$ ;
while ($|P| \leq k$) or (score ¡ $\tau$) do
    $\mathcal{C} = nextCombination(Q)$ ;
    $\tau = s(\mathcal{C})$ ;
    $P = P \cup$ getObjects($\mathcal{C}$) ;
    score = score of $k$-th element of $P$ ;
return $P$ ;
```

$STPQ$ queries based on the influence preference score can be efficiently supported by the *STPS* algorithm with few modifications. Algorithm 5 shows the modified *STPS* for influence preference score. The algorithm continues until at least k data object have been retrieved and until we are sure that none of the remaining data objects can have a better score. In more details, $\mathcal{C} = nextCombination(Q)$ is the same with Algorithm 4 and returns the best combination based on score $s(\mathcal{C})$, but without discarding combinations that have a $distance > 2r$. Thus, in each iteration the combination of feature $\mathcal{C}$ with the highest $\tau(p) = \sum_{i \in [1,c]} \tau_i(p)$ is retrieved. Recall that for the case of the range preference score, all data objects that were located in distance smaller than r from all feature objects of $\mathcal{C}$ had a score equal to $s(\mathcal{C})$. Differently in the case of the influence prefrence score, the $s(\mathcal{C})$ is an upper bound for the score of all data objects based on $\mathcal{C}$. Therefore, *getObjects*($\mathcal{C}$) must be modified accordingly.

The best score of any unseen combination is $\tau = s(\mathcal{C})$, because this is the score for distance 0. Hence, if the k-th score of the P is smaller than τ, we have to retrieve additional objects.

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

Co- financed by Greece and the European Union

In particular, $getObjects()$ retrieves the k points that have the highest influence score, by starting a top-k query on the R-Tree of the data objects. The root is inserted in a heap sorted by the influence score ($\tau(p) = \sum_{i\in[1,c]} \tau_i(p)\dot{2}^{\frac{-dist(p,t_i)}{r}}$). For non-leaf entries the influence score is computed based on the mindist. Then the influence score of an entry is an upper bound of any object in the subtree. After retrieving k data objects, we have retrieved the k data object with the highest influence score for this combination of feature objects. Further improvement can be done if $getObjects()$ stops retrieving data objects based on τ, which reduces the I/Os on the R-Tree. If τ is given to $getObjects()$ then it will return at most k data objects that have a score less than τ. $P = P\cup$ getObjects($\mathcal{C}$) merges the results while it removes objects that have been retrieved before. Thus, if an object that is already in the heap is retrieved again the score with the highest value is kept. For those feature objects the k data objects with the best $\tau(p)$ are retrieved. The score of the k-th retrieved data object is a threshold. If the $s(t)$ of the next combination is smaller then the threshold we stop retrieving other combinations.

Towards this end, we suggest a rank-aware approach for processing influence queries. To elaborate, once the first combination of highly ranked feature objects is composed, we can easily trace back to the structures that comprise the data objects the records which achieve distances from the feature objects that are qualified to be added to the result. However, there is a chance for the next combination of feature objects, even though it is less relevant, to have within closer proximity data objects in such a way that some of the previously retrieved records are outranked, and therefore, should be replaced. Hence, given that we already have an answer-set of k items that achieve certain scores and a new combination of feature objects, we can pre-determine the maximum distances to search for better ranked

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

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

data objects. This consideration will help us reduce disk accesses to the most essential I/Os. In addition, we need to know when to stop, in other words to recognize the situation where all remaining combinations of feature objects cannot be associated with any data object so as to outrank any element of the result-set. Therefore, it is of the utmost importance for processing influence queries to establish a rigorous method for accessing only the disk-pages that contain highly relevant and ranked records. Otherwise, by underestimating this intricate aspect of the problem we might end up accessing the whole R-Tree several times until all top-k objects are found.

In the following example we assume that $c = 1$ and $K = 1$ for simplicity, but we will soon generalize for any positive c and K. Let the feature object t_1 returned from the *nextCombination()* method. Then, the score achieved by any data object within d_1 distance from t_1 equals to $\frac{\sigma_1}{2^{d_1}}$, where σ_1 equals to the affine combination of t_1's ranking and textual similarity with the query for some smooting factor λ. Also, assume that the data object $o_1 = \mathrm{argmin}_o d(t_1, o)$ achieves the best score w_1 when compared to all data objects. Now, for best data object to be combined with the next best feature object t'_1 returned from *nextCombination()* with $\sigma'_1 \leq \sigma_1$, it holds to be considered at least as good as the previous combination that $\frac{\sigma'_1}{2^{d'_1}} \geq \frac{\sigma_1}{2^{d_1}}$. Otherwise, if there is none such object, we should proceed with the next feature object. Hence, it follows that,

$$d'_1 \leq d_1 + \underbrace{\log \frac{\sigma'_1}{\sigma_1}}_{\leq 0} \tag{1.1}$$

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

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

Algorithm 6: *Data Objects Retrieval within Feature Objects' Influence* $(getObjects(\mathcal{C}, \tau, \mu, k))$

```
Output: Result set $P$ sorted based on $\tau(p)$
$heap$.push($rtree$.rootNode) ;
$j \leftarrow 0$;
while not $heap$.isEmpty() or $j < k$ do
    $e \leftarrow heap$.top() ;
    $heap$.pop();
    if $e$.isLeaf() then
        if $j < \mu$ or $(s(e) > \tau$ and $j < k)$ then
            $P \leftarrow P \cup e$;
            $j \leftarrow j + 1$;
        else
            break;
    else
        for childEntry in $e$.childNodes do
            overlaps $\leftarrow$ true ;
            forall the $1 \leq x \leq |\mathcal{C}|$ do
                $\phi \leftarrow 0$;
                forall the $1 \leq y \leq |\mathcal{C}|$ do
                    if $x \neq y$ then
                        $w \leftarrow \mathcal{C}_y$.rank + jaccard($\mathcal{C}_y$.text,Q);
                        $\phi \leftarrow \phi + \frac{w}{2^{dist(\mathcal{C}_y, childEntry)}}$;
                if $d_x \geq \log_2 \frac{\mathcal{C}_x.rank + jaccard(\mathcal{C}_x.text, Q)}{\tau - \phi}$ then
                    overlaps$\leftarrow$false;
                    break;
            if overlaps then
                $heap$.push (childEntry);
return $P$;
```

Thereby, we manage to transform an influence query to a series of range queries where the radius is dynamically adapted according to the score of best answer found so far. In essence, the threshold distance d'_1 that is set by the last found item is further tightened in Eq. 1.1 analogously to the logarithm of the score ratio

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

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

of the feature objects, since the latter score is at most equal to the former. As a result, only the R-Tree nodes, either internal or leaves, that overlap with the area designated by a circle with t'_1 as its center and radius less than d'_1 are accessed during search. In effect, all branches of the tree that contain data objects that cannot outrank o_1 are pruned.

Moreover, we can follow the same convention for $c > 1$ to minimize the disk accesses performed during search. Specifically, let $\sum_{i=1}^{c} \frac{\sigma_i}{2^{d_i}}$ the score of the best ranked data object and $t_1, t_2, \cdots, t_c$ the most influential combination of feature objects. Again, we can determine an upper bound for the influence radius, given the previous combination of feature and data objects. Hence, it holds that, $\sum_{i=1}^{c} \frac{\sigma'_i}{2^{d'_i}} \geq \sum_{i=1}^{c} \frac{\sigma_i}{2^{d_i}}$. Therefore, when examining a data object with respect to feature object f_x, it follows that, $\frac{\sigma'_x}{2^{d'_x}} \geq \sum_{i=1}^{c} \frac{\sigma_i}{2^{d_i}} - \sum_{\substack{j=1\\ j\neq x}}^{c} \frac{\sigma'_i}{2^{d'_i}}$, which eventually leads to,

$$d'_x \leq \log_2 \frac{\sigma'_x}{\sum_{i=1}^{c} \frac{\sigma_i}{2^{d_i}} - \sum_{\substack{j=1\\ j\neq x}}^{c} \frac{\sigma'_i}{2^{d'_i}}}, 1 \leq x \leq c \tag{1.2}$$

In other words, the next retrieved object in order to be as good as the previous, it should be positioned in the area that corresponds to the intersection of c circles around the feature objects of the combination and radius given by the formula above. Hence, when examining the child nodes of an accessed R-Tree node we can drop it and proceed with the next if even one of its distances from the associated feature objects does not satisfy Eq. 1.2

Last but not least, Alg. 6 implements *getObjects()* from Alg. 5 and it addresses the general case where $c \geq 1$ and $k \geq 1$. We follow the same procedure, though, the sum corresponding to the previous influence score in Eq. 1.2 now corresponds to the score τ of the k-th previously retrieved item. In lines 13–24 we compute for

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

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

each feature object of the combination $\mathcal{C}$ the maximum distance from this specific object that the qualified data objects should keep. In particular, for each new influential combination we retrieve the top-μ, with $\mu \leq k$, items until a full set of k items is formed for all examined combinations. Of course, if there are more than μ items that outrank the previously retrieved items, then these objects are returned as well so as to update the answer-set accordingly (Alg. 6, line 7). We can stop early when the remaining influential combinations have no better preference score than the influence score of the k-th item in the result.

1.7.2 Implementation of Nearest Neighbor Score for Parameterized Query Processing

In the next variant of the range score (Definition 1), each data object takes as a score the goodness of the feature objects that are its nearest neighbors.

Definition 7 *The **influnce preference score** $\tau_i(p)$ **of data object** p based on the feature set F_i is defined as:* $\tau_i(p) = max\{s(t) \cdot 2^{\frac{-dist(p,t)}{r}} \mid t \in F_i : \ sim(t, \mathcal{W}_i) > 0\}$.

The overall spatio-textual score $\tau(p)$ of data object p is defined as for the case of the range score, and the query returns the k objects with the highest score. Again, *STDS* treats nearest neighbor queries similarly as in Alg. 2 with subtle changes. The range predicate is upheld in line 10, though the child entries in line 11 are prioritized according to their minimum distance from all data objects.

Regarding *STPS*, Alg. 3 is directly applicable for the nearest neighbor score by modifying the $\mathcal{C} = nextCombination(Q)$ of Alg. 4 and returns the best

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

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

combination based on score $s()$, but without discarding combinations that have a $distance > 2r$ as also in the case of the influence score. Generally, it is more difficult to retrieve the data objects that have the retrieved combination of feature objects as their nearest neighbor.

In order to retrieve efficiently the data objects for a combination $\mathcal{C}$, we have to first determine the area where the data points are located. Then, by enacting the appropriate query we retrieve them all. For each feature object t_i of $\mathcal{C}$, there exists a region in which all data points that fall into that region t_i is their nearest neighbor. Only the data objects in the intersection of all regions need to be retrieved. In fact, we compute incrementally the region for each feature object t_i of $\mathcal{C}$, which allows us to discard early combinations for which the intersection becomes empty. In order to compute this region we have to compare the location of t_i with the other feature objects of F_i. To elaborate, with the following steps we compute the convex space that is associated with feature object t_i with a process that resembles solutions for finding bichromatic reverse k nearest neighbors. Initially, the whole keyspace constitutes the influence area of t_i. This area is gradually refined and reduced accordingly.

1. We start by initializing a heap with the root node of the aggregated R-Tree that corresponds to F_i. The key of each heap element corresponds to the minimum distance of the associated MBR from t_i.

2. At each iteration, we pop the next node from the aggregated R-Tree with the minimum distance from t_i. Now, if the popped node is a leaf containing feature object p_k, then it corresponds to the center of another cell.

3. We compute the parameters of the bisector of the segment between t_i and

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

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

p_k, where $y = \alpha_k x + \beta_k$ with $\alpha_k = -\frac{x_{p_k} - x_{t_i}}{y_{p_k} - y_{t_i}}$ and $\beta_k = \frac{y_{t_i} + y_{p_k}}{2} - \alpha_k \frac{x_{t_i} + x_{p_k}}{2}$.

4. We compute all $b_{i,j}$ points where the bisectors intersect for all p_i, p_j pairs with $i < j$.

5. We insert into set $\mathcal{V}_{t_i}$ all $b_{i,j}$s such that dist$(b_{i,j}, t_i) \leq$ dist$(b_{i,j}, p_k)$, $\forall k$, where $p_k \neq p_i$ and $p_k \neq p_j$. The points in $\mathcal{V}_{t_i}$ bound the area where all comprised data objects have t_i as their nearest neighbor.

6. We keep retrieving t_i's nearest neighbors (which still correspond to the centers of neighboring Voronoi cells) until the next node's (either internal or leaf) minimum distance from t_i becomes greater than $2 \max_{i,j}$ dist$(t_i, b_{i,j})$ (computed in Alg. 7). Hence, beyond this point it can be easily shown by contradiction that there is no chance for the bisector of the segment between t_i and p_k to intersect the already formed influence area of t_i, regardless the angle. All other branches are effectively pruned.

7. We can further optimize this scheme by using the bounded area we computed for the previous feature class and carrying it to the next. If the intersection of $\mathcal{V}_{t_i}^{F_{i+1}}$ with $\mathcal{V}_{t_i}^{F_i}$ yields the empty-set $\emptyset$ at any point, then we can stop working on this combination $\mathcal{C}$ of feature objects, and proceed with the next. This way we can further reduce any unnecessary IOs.

In Alg. 7, we compute all intersections of the bisectors that formed between t_i and each of the neighboring cell center, say p_i and p_j. Now, if for any other cell center, say p_k, the intersection of the bisectors, say $b_{i,j}$, is closer to p_k that t_i, then this means that this particular vertex is not part of the Voronoi cell surrounding t_i. In other words, $b_{i,j}$ is obscured by p_k. Thereby, when each intersection is

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

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

```
Algorithm 7: getMaxVertexDistance ($t_i, \{p_1, \cdots, p_\nu\}$)
forall the $p_i \in \{p_1, \cdots, p_\nu\}$ do
    forall the $p_j \in \{p_1, \cdots, p_{i-1}\}$ do
        $\alpha_i \leftarrow -\frac{x_{p_i} - x_{t_i}}{y_{p_i} - y_{t_i}}$;
        $\alpha_j \leftarrow -\frac{x_{p_j} - x_{t_i}}{y_{p_j} - y_{t_i}}$;
        $\beta_i \leftarrow \frac{y_{p_i} + y_{t_i}}{2} - \alpha_i \frac{x_{p_i} + x_{t_i}}{2}$;
        $\beta_j \leftarrow \frac{y_{p_j} + y_{t_i}}{2} - \alpha_j \frac{x_{p_j} + x_{t_i}}{2}$;
        $x_{b_{i,j}} \leftarrow \frac{\beta_j - \beta_i}{\alpha_i - \alpha_j}$;
        $y_{b_{i,j}} \leftarrow \alpha_i x_{b_{i,j}} + \beta_i$;
        tmpDist $\leftarrow$ dist $(t_i, b_{i,j})$;
        forall the $p_k \in \{p_1, \cdots, p_\nu\} \setminus \{p_i, p_j\}$ do
            if tmpDist $>$ dist$(p_k, b_{i,j})$ then
                obscuredVertexFlg $\leftarrow$ true;
                break;
        if not obscuredVertexFlg and mxDist $<$ tmpDist then
            mxDist $\leftarrow$ tmpDist;
return mxDist;
```

obscured by any other, and hence, none non obscured vertex exists then $\mathcal{V}_{t_i}$ simply corresponds to the empty-set. Therefore, we know in advance that no data objects can ever exist to qualify for the examined combination $\mathcal{C}$ of feature objects, whose cells' interesection we try to compute with $\bigcap_i \mathcal{V}_{t_i}^{F_i}$, and thus, no unnecessary effort should be paid into finding the cells associated with the remaining feature objects, or reading any disk-page from the R-Tree with the data objects to find out that no record eventually overlaps with an empty space after all. We also note that we treat a little differently when p_i or p_j is on the same axis, either horizontal or vertical.

Notably, the task of exposing the influence area of each feature object is not a very difficult task for two dimensions. As a matter of fact, even for real datasets,

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

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

such as the distribution of postal codes in the US, each cell is formed by up to ten edges most of the times. Next, once we have determined the influence areas of feature objects $t_1, t_2, \cdots, t_c$, we start traversing recursively the tree hierarchy where the data objects are stored. More importantly, only the branches of the tree that overlap will *all* c influence regions are accessed.

1.8 Experimental Evaluation

In this section, we scrutinize meticulously the performance of *STDS* and *STPS* for processing spatio-textual preference queries over large disk-resident data.

1.8.1 Methodology

The efficiency of all schemes is evaluated according to two distinct metrics: (i) the required I/Os, which is measured by the average number of disk-pages accessed (disk accesses) per query, and (ii) the average execution time required by a query (time consumed in the CPU and to read disk-pages).

Furthermore, our experimental evaluation examines six important parameters. More specifically, we investigate how the aforementioned metrics scale under six different scenarios: (i) as the query range r scales between 0.02 and 0.16, (ii) as the smoothing parameter λ between textual similarity and ranking in search ranges in $(0, 1)$, (iii) as the number of expected top-k items in a query is varied from 20 to 640, (iv) as the cardinality of objects' set $|O|$ varies from $50K$ tuples to $1M$, (v) as the features' set cardinality $|F_i|$ also varies from $50K$ tuples to $1M$, and (vi) as the number of feature sets c increases. Tested ranges for all parameters are shown in Table 1.2. The default values are denoted as bold.

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

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

Parameter	**Range**
Query radius (norm. in $[0, 1]$)	$.005$, **.01**, $.02, .04, .08$
Page-size (in bytes)	2048, **4096**, $8192, 16384$
Result-size	5, **10**, $20, 40, 80$
Smoothing parameter	$.1, .3$, **.5**, $.7, .9$
Queried keywords/Feature class	1, **3**, $5, 7, 9$
Objects set cardinality	$50K$, **100K**, $500K, 1M$
Features set cardinality	$50K$, **100K**, $500K, 1M$
Features classes queried	**2**, $3, 4, 5$
Indexed keywords	64, **128**, $192, 256$

Table 1.2: System parameters.

For evaluating our algorithm, we created real and synthetic datasets, as well. The real dataset, which was obtained from `factual.com`, describes hotels and restaurants. In more details we collected restaurant and hotels that are annotated by their location. Moreover, for the collected restaurants we extracted their rating and their textual description of the served food mentioned as cuisine. The possible values of keywords for the cuisine is around 130 and each restaurant description may contain one or more keywords. Our datasets contain collected hotels and restaurants for 13 US states that are the states for which `factual.com` lists sufficient enough data. We created synthetic clustered datasets of varying size, keywords and classes of feature objects. Approximately $10,000$ clusters constitute each synthetic dataset. The number of distinct keywords is set to 256 and each feature object is characterized by one or more keywords. When we vary one parameter, all others are set at their default values. The spatial constituent of all datasets has been normalized in $[0, 1] \times [0, 1]$. Every reported value is the average of executing $1,000$ randomly generated queries. The queries are generated in a similar way as the synthetic data and follow the same data distribution.

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

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

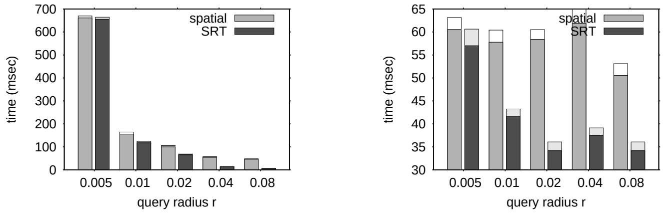
NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co- financed by Greece and the European Union

1.8.2 Results

This section presents the results of the experimental evaluation illustrated in Figures 1.6–1.18 for real and synthetic workloads, where we explore the impact of several parameters on IO and processing time. Overall, we reckon that there are profound implications from using alternative approaches for building aggregate R-Trees. First and foremost, the composite index outperforms the conventional index that relies on spatial information only.

In Figures 1.6(a) and 1.6(b), for real and synthetic workloads respectively, where only the query range is varied, we see that the smaller the radius becomes, the more similarly the two indices behave; for query processing is focused on finding qualified combinations of feature objects, which are quite a few for very small values of r, and then, select the most relevant ones. Therefore, the index which is built on the records' spatial information performs as well as the composite one for very low and restrictive r-values for being the factor that defines search performance. However, difference in performance becomes obvious when the query radius restriction is relaxed, greater r-values, and hence, finding relevant tuples in terms of textual description and good rank becomes most important. As shown in Fig. 1.6, the advantage of the composite index ameliorates performance a great deal.



(a) real workload (b) synthetic

Figure 1.6: Varied query radius for range queries.

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

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

In Figures 1.7 and 1.11, we get faster response with larger pages evidently, for real and synthetic workloads respectively. In particular, the time required for IO diminishes with page-size, whereas the time spent at the CPU increases. More specifically, larger disk-pages congregate more records, and therefore, more time is needed to process each disk-page, while the total number of disk-pages that constitute the R-Tree decreases significantly. Particularly in Fig. 1.7(c) we illustrate with a striped pattern separately the IO and the CPU-time required to compute the respective Voronoi cells for the nearest neighbor queries. Remarkably, the cost of finding the ifluence area of a specific combination of feature objects, in other words computing the intersection of the areas for each retrieved feature object relevant to the query with all comprised data objects having as their nearest neighbors these specific feature objects, is higher than the cost of finding highly ranked combination of feature objects from the aggregate R-Trees and retrieving k relevant data objects altogether. Nonetheless, this cost is slightly less singificant for the conventional index which is built based on the spatial information only, for records in close proximity are clustered together in consecutive disk-pages, if not in the same. On the other hand, for the composite index we have similar records in terms of textual description, rank *and* location clustered together. Hence, since computing the influence area of a combination of feature objects takes exclusively into account spatial information, the composite index exhibits a small overhead, even though the conventional approach is easily outperformed when total time is considered. Also, we note that for static data the cells can be pre-computed in a special structure, and therefore, significantly reduce the total cost.

Overall, execution time increases with result-size k as expected in Figures 1.8 and 1.12, for real and synthetic workloads respectively. Specifically, with

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

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

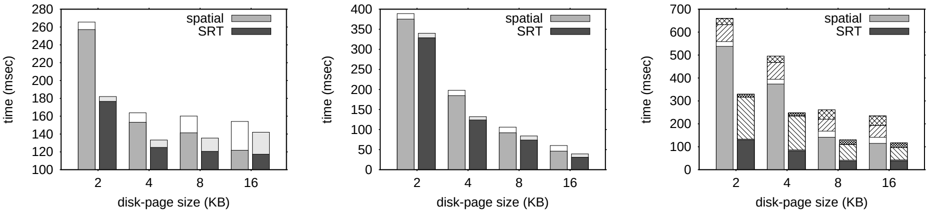
greater k-values, more combinations of feature objects are constructed to compose a larger answer-set of data objects. In practice, this is translated into multiple searches for each feature category, until the objects that constitute valid combinations are retrieved.

Regarding the trade-off parameter λ, we observe in Figures 1.9 and 1.13 that the composite index is in position of taking the most out of either of its constituents, namely rank-based, text-based, or spatial-based. When a query with $\lambda \to 1$ is issued, then our composite index takes advantage of the fact that it is partly build based on records' textual information. Likewise, for $\lambda \to 0$, the information which corresponds to the records' rank is used. On the contrary, the index that is built conventionally relying exclusively on the records' spatial information has no way of knowing a priory which branches of the already accessed R-Tree nodes are ranked higher. Thereby, all children nodes within the predefined range are accessed in tandem to be inserted into a priority heap from which the best combination will be retrieved. Again, score thresholds are used in a branch-and-bound fashion, even though they are not as effective as with the composite index. We note for the conventional index that objects with similar textual descriptions are stored throughout the index, regardless their rank; unlike the composite index where they are clustered together in consecutive disk-pages. As a result, a significant overhead is evident when searching for relevant objects all over the spatial-based index, ranking them, and combining them appropriately. In Figures 1.10 and 1.14 any reasonable number of queried keywords between 3 and 9 has little impact on performance, if any. However, we note that this would not be the case if an approach based on inverted files had been approached.

Furthermore, processing time also increases with either the number of indexed

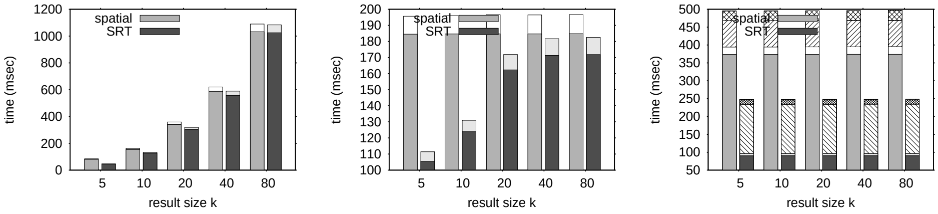
The research project is implemented within the framework of the Action Supporting Postdoctoral Researchers of the Operational Program "Education and Lifelong Learning" (Actions Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the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

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



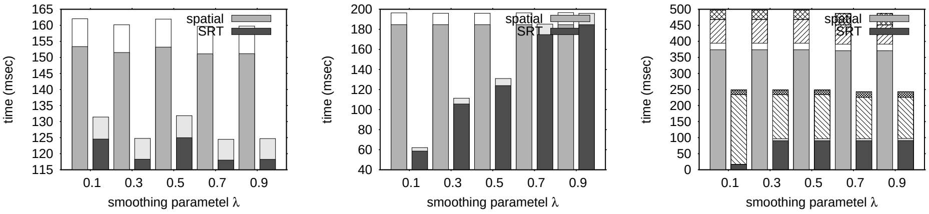
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.7: Varied page-size for real workload.



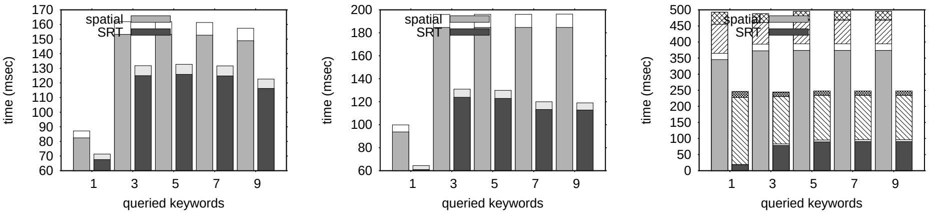
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.8: Varied result-size for real workload.



(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.9: Varied trade-off for real workload.



(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.10: Varied number of queried keywords.

feature objects in Fig. 1.15, or the number of indexed data objects in Fig. 1.16, although the former has a stronger impact on performance than the latter. This

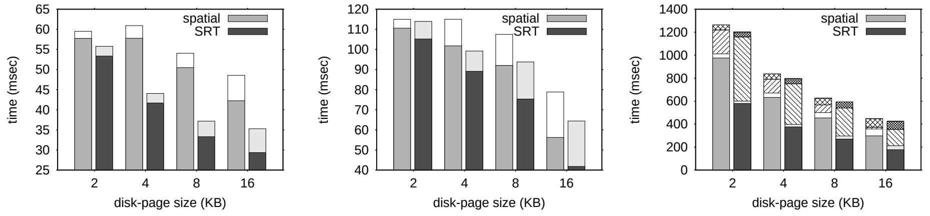
The research project is implemented within the framework of the Action Supporting Postdoctoral Researchers of the Operational Program "Education and Lifelong Learning" (Actions Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the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

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

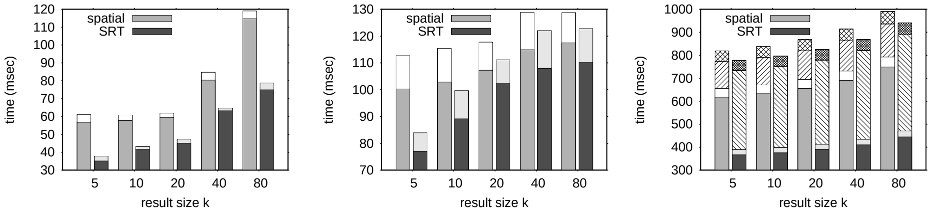
NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co- financed by Greece and the European Union



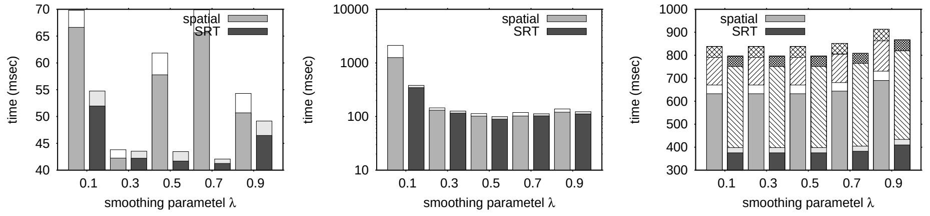
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.11: Varied page-size for synthetic workload.



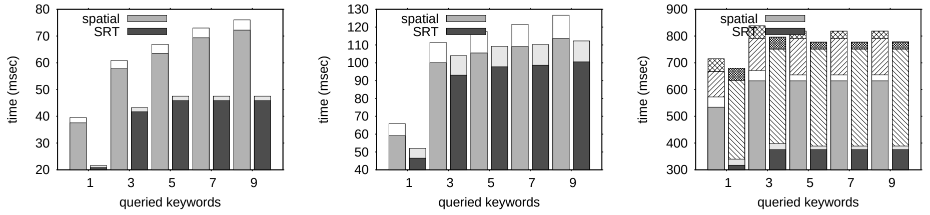
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.12: Varied result-size for synthetic workload.



(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.13: Varied trade-off for synthetic workload.



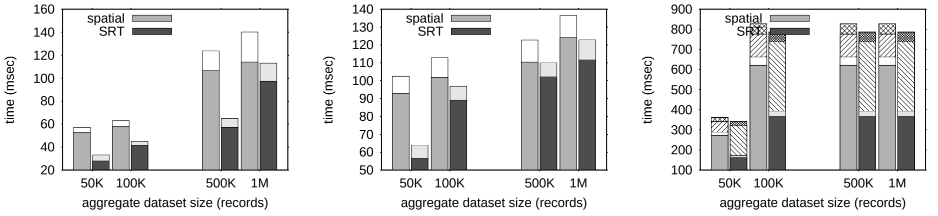
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.14: Varied number of queried keywords.

behavior can be easily explained: as the data structures grow bigger, more effort is required to find the best ranked items and their respective qualified data objects.

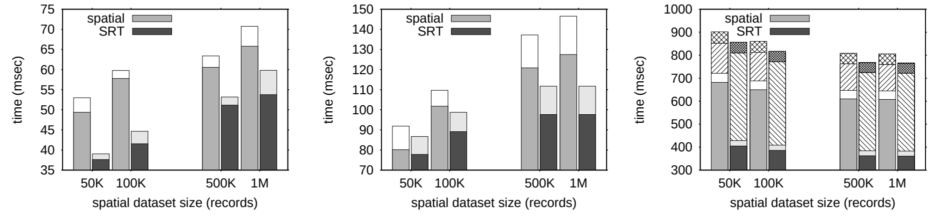
The research project is implemented within the framework of the Action Supporting Postdoctoral Researchers of the Operational Program "Education and Lifelong Learning" (Actions Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the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

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



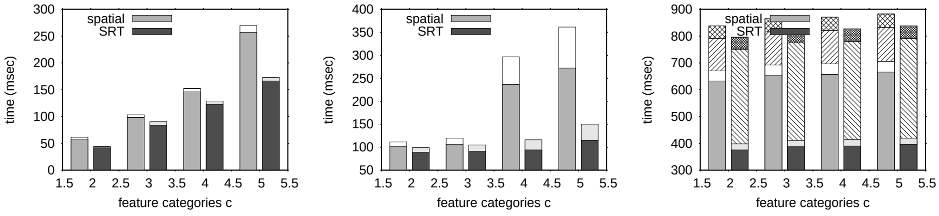
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.15: Varied number of features for synth.



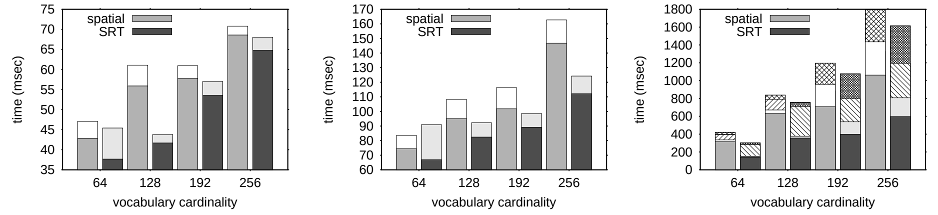
(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.16: Varied number of data objects for synth.



(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.17: Varied feature categories for synth.



(a) range queries (b) influence queries (c) nearest neighbor

Figure 1.18: Varied keywords for synthetic workload.

Surprisingly, performance slightly seems to improve for nearest neighbor queries. Keep in mind that under this specific setting the computation of the Voronoi cell

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

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

consumes exactly the same resources for it is computed in the same structure when only the number of data objects is varied. Presumably, with larger datasets more similar records are clustered together into the same disk-page. Thereby, less disk-pages are accessed overall since all records comprised by a specific Voronoi cell are congregated in just a few disk-pages, whereas, for a smaller R-Tree we access a larger proportion of its leaves. Of course, a few reads is sufficient for very small datasets. We also note that the size and the shape of the cell also affect performance. More importantly, the number of feature categories c affects performance more dramatically in Fig. 1.17, since the number of possible combinations of feature objects increases exponentially with c. Intuitively, the efforts and costs required to retrieve the best ranked combination increase with the number of possible combinations, which in turn, increases exponentially with c.

In Figure 1.18 we illustrate how performance is affected with respect to the number of keywords. Apparently, performance is impaired for increased number of keywords for all query types. However, CPU-time diminishes for range queries and influence queries since more records fit in each disk-page that have to be processed. On the other hand, we observe that both IO and CPU-time grow with the number of keywords for nearest neighbor queries. With more indexed keywords we have less records per disk-page and the cost of computing the Voronoi cell among more disk-pages naturally increases. Remarkably, the computation of the Voronoi cell is the decisive factor that defines the total time, as performance deteriorates for nearest neighbor queries with the number of keywords.

Last but not least, we noticed that range and influence queries are costlier for the real dataset. This is due to the data distribution: our real dataset which was extracted from `factual.com` consists of restaurants and hotels in the US form-

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

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

ing just a few clusters. On the other hand, our synthetic dataset is substantially larger and coined to form $10,000$ cluster approximately. Hence, the data from the latter dataset are more dispersed compared to the former. Conversely, nearest neighbor queries are more efficient for the real dataset for the same reason as the Voronoi cells are formed faster when processing data from just a few very condensed clusters.

1.9 Conclusions

Recently, the database research community has lavished attention on spatio-textual queries that retrieve the objects with the highest spatio-textual similarity to a given query. Differently, in this report, we addressed the problem of ranking data objects based on the quality of facilities in their spatial neighborhood and their textual similarity to user-specified keywords. Towards this end, we proposed a novel query type called *top-k spatio-textual preference queries*. We developed a framework for processing many forms of this novel query type. We make use of spatio-textual indices that are capable of processing efficiently spatial and textual information simultaneously. Our first approach, called *Spatio-Textual Data Scan* (*STDS*), first retrieves a data object and then computes its score, whereas the latter, called *Spatio-Textual Preference Search* (*STPS*), first retrieves highly ranked feature objects and then searches for data objects nearby those feature objects. More importantly, we develop algorithms for processing three forms of top-k spatio-textual preference queries, namely (i) range queries, (ii) influence queries, and (iii) nearest neighbor queries. Above all, our framework can be easily extended to cover complex query types at a low programming cost.

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

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

Furthermore, a salient characteristic of our approach is the alternative technique used for indexing aggregate data, which is suitable for processing top-k spatio-textual preference queries, as it ameliorates performance a great deal. Besides, there is a dearth of work on optimizations at the storage layer. Therewith, search is directed immediately towards the most promising records, in an effort to reduce I/Os and avoid accessing irrelevant disk-pages. We emphasize on the fact that hitherto approaches ignore the textual constituent of data, and thus, cause a significant I/O overhead. Finally, in our experimental evaluation, we put all methods under scrutiny to verify the efficiency and the scalability of our methods, partly by exposing the deficiencies of conventional approaches inept for processing top-k spatio-textual preference queries.

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

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

Chapter 2

Tag Recommendations

Flickr is one of the largest online image collections, where shared photos are typically annotated with tags. The tagging process bridges the gap between visual content and keyword search by providing a meaningful textual description of the tagged object. However, the task of tagging is cumbersome, therefore tag recommendation is commonly used to suggest relevant tags to the user and enrich the semantic description of the photo. Apart from textual tagging based on keywords, an increasing trend of geotagging has been recently observed, as witnessed by the increased number of geotagged photos in Flickr. Geotagging refers to attaching location-specific information to photos, namely about the location where a photo was captured. Even though there exist different methods for tag recommendation of photos, the gain of using spatial and textual information in order to recommend more meaningful tags to users has not been studied yet. In this report, we analyze the properties of geotagged photos of Flickr, and propose novel location-aware tag recommendation methods. For evaluation purposes, we have implemented a prototype system and exploit it to present examples that demonstrate the effectiveness

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

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

of our proposed methods.

2.1 Introduction

Flickr allows users to upload photos, annotate the photos with tags,view photos uploaded by other users, comment on photos, create special interest groups etc. Currently, Flickr stores one of the largest online image collections with more than 8 billion photos (March 2013[1]) from more than 87 million users and more than 3.5 million new images uploaded daily. The tags are important for users to retrieve relevant photos among the huge amount of existing photos. Since multimedia data provide no textual information about their content, tags bridge the gap between visual content and keyword search by providing a meaningful description of the object. Thus, to make their photos searchable, users are willing to annotate their uploaded images with tags [2]. Nevertheless, tags reflect the perspective of the user that annotates the photo and therefore different users may use different tags for the same photo. This can be verified by the fact that photos of Flickr that depict the same subject may be described by a variety of tags. Tag recommendation [20] is commonly used to provide to the user relevant tags and enrich the semantic description of the photo.

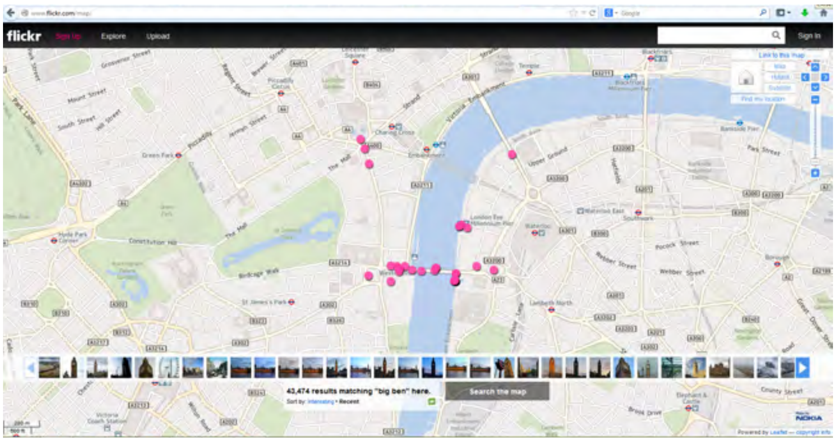
Flickr motivates its users to geotag their uploaded photos[2]. Geotagging means to attach to a photo the location where it was taken. Photos taken by GPS-enabled cameras and mobile phones are geotagged automatically and location metadata, such as latitude and longitude, are automatically associated with the photos. Flickr

[1] http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer

[2] http://www.flickr.com/groups/geotagging/

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

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



43,474 results matching "big ben" here.

Figure 2.1: Example of geotagged photos on a map in Flickr.

is able to read the spatial information (latitude and longitude) during the upload and place the photos on a map, as depicted in Figure 2.1. Furthermore, photos may be also geotagged manually by the user when the photo is uploaded. Currently, there is an increasing trend in the number of geotagged photos in Flickr.

Even though several recent studies [8, 5] examine how relevant web objects can be retrieved based on both the spatial and textual information, the gain of using spatial information in order to recommend more meaningful tags to users has not been studied yet. Nevertheless, it is expected that nearby photos may depict similar objects, thus sharing common tags with higher probability. In this report, we propose methods for tag recommendations based on both location and tag co-occurrence of the photos. In details, this report makes the following contributions:

- We create different data collections of geo-tagged photos of Flickr that are located in different cities and analyze their properties in terms of tag frequency, number of tags per photos and the type of tags commonly chosen by users. This study allows us to analyze the behavior of the users related to tagging and draw some important conclusions for our tag recommendation

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

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

methods.

- We introduce novel tag recommendation methods that take into account also the location of the given photo as well as the location of the existing photos. The key idea of our methods is that not only the similarity in terms of existing tags is important, but also the distance between the existing photos in which the tags appear.

- We implemented a prototype system for location-aware tag recommendations over photos of Flickr and evaluate experimentally our proposed method through examples that demonstrates the effectiveness of location-aware tag recommendation.

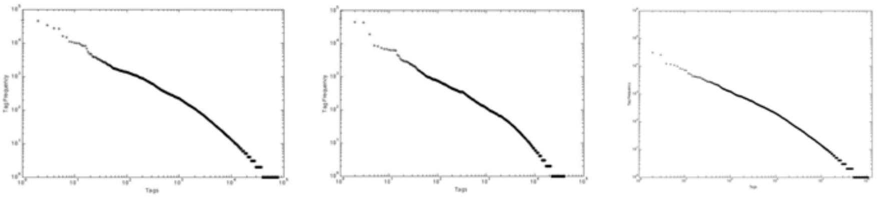
The remainder of the report is structured as follows. In Section 2.2 we describe our data collections and analyze their properties. Then, Section 2.3 presents an overview of the location-aware tag recommendations system and describes the proposed location-aware tag recommendation methods. In Section 2.4 we evaluate our proposed methods. Finally, in Section 2.5 we discuss related work and in Section 2.6 we provide some concluding remarks.

2.2 Data Collection

In this section we describe our data collections and provide statistics about the photo tags. In order to design our recommendation strategies it is important to first study the relevance and quality of the tags. What kind of tags are used for tagging is also important in order to understand which tags are useful for recommendations and how the tags relate to the location of the photo.

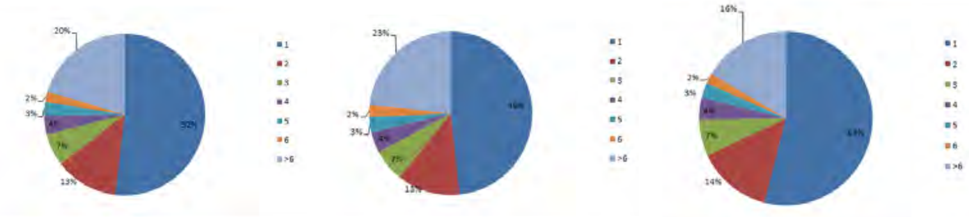
The research project is implemented within the framework of the Action Supporting Postdoctoral Researchers of the Operational Program "Education and Lifelong Learning" (Actions Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the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

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



(a) New York (b) Rome (c) London

Figure 2.2: Tag frequency distribution



(a) New York (b) Rome (c) London

Figure 2.3: Number of tags per photo

We have created three different data collections. Each of them contains 100.000 geotagged photos that are located in New York, Rome and London respectively. Table 2.1 summarizes the number of tags that appear in each collection and the number of unique tags per collection. The collected photos are a random snapshot of the geotagged photos located in the aforementioned cities. For each city the boundary is defined by the bounding box provided at `http://www.flickr.com/places/info/`. The photos were collected between December 2012 and February 2013 and each photo has at least one tag describing it.

Collection	**Tags**	**Unique tags**
New York	1.502.454	80.180
Rome	897.185	41.843
London	1.428.047	110.231

Table 2.1: General characteristics per collection.

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

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

2.2.1 Distribution of Tag Frequency

Our data collection of photos collected from Flickr located in New York contains 100.000 photos, with 1.502.454 tags in total, while the unique tags are 80.180. The photo collection of Rome has 897.185 tags in total and the unique tags are 41.843. Finally, the data collection of London has 1.428.047 tags in total and the unique tags are 110.231.

Figure 2.2 shows the distribution of the tag frequency on a log-log scale. The x-axis represent the set of unique tags order based on the frequency in descending order. The y-axis is the tag frequency. We observe that the tag frequency can be modeled by a power law for all data collections.

Tag	**Freq.**
NYC	47940
New York	45809
NY City	33941
manhattan	27282
NY	26717
USA	15957
City	14637
New	10952
Brooklyn	10741
2012	10126

Table 2.2: New York.

Tag	**Freq.**
rome	56660
italy	44842
roma	44225
italia	19281
Lazio	8883
2012	8374
Europe	7534
Rom	6917
square	6851
iphoneography	6464

Table 2.3: Rome.

Tag	**Freq.**
London	68250
UK	30839
England	25760
2012	12459
kenjonbro	11693
trafalgar square	11090
United Kingdom	10023
Westminster	8404
fuji hs10	7981
SW1	7282

Table 2.4: London.

Tables 2.2- 2.4 show the 10 most popular tags for New York, Rome and London respectively. For the New York collection there exist 41.230 tags with tag frequency 1, which are the less popular. To give an example of their relevance we report 10 random of them: walmart, resort, people mover, kristin, bougainvillea, pixie, aviso, World Heritage Site, Beggar, ox. Similarly, for Rome there exists 20.197 tags with frequency 1, while for London there are 59.559 tags with frequency 1.

By observing the distribution of the tags in the each collection, but also by

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

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

looking at the most popular tags, it is obvious that the most popular tags should be excluded by our recommendation method because these tags are too generic to be helpful for recommendation. Recall that the popular tags include tags such as: NYC, New York, Rome, Italy, London, UK. Similar, the less popular tags with very small frequency (i.e., equal to 1) should be also excluded by our recommendation method, since these tags include words that are misspelled, complex phrases and very specific tags. For example consider the tags: drwho, loo, boring, SF, #noon, dv. Due to their low frequency it is expected that those tags can be useful only in very specific cases and thus are not suitable for recommending to other photos.

2.2.2 Distribution of Number of Tags per Photo

In Figure 2.3 the number of tags per photo are depicted. More precisely, the percentage of photos that have 1, 2, 3, 4, 5, 6, >6 tags for each data collection are depicted. In addition, we consider (Figure 2.4) also the distribution of the number of tags per photo for New York. Figure 2.4 is in log-log scale and the x-axis represents the set of photos ordered based on the number of tags per photo (descending order), while the y-axis refers to the number of tags of each photo. We notice that a high percentage of photos, i.e, approximate 20%, has a high number of tags (more than 6 tags) and there even exist photos with more than 50 tags. Similar results have been also obtained for the other two data collections.

Thus, some photos have a very high number of tags, so that these tags cannot be considered to be representative for the photo. Therefore, our recommendation methods will not use such photos. Moreover, approximately 50% of the photos

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

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

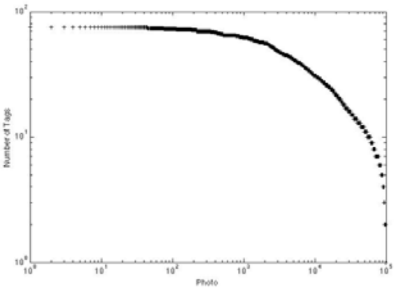


Figure 2.4: Number of tags per photo for New York.

have only one tag, and again these photos can not be used for tag recommendation that relies on co-occurrence of tags. On the other hand, the fact that a high percentage of photos have only one tag motivates the need for tag recommendation, since all these photos would benefit by a more detailed description.

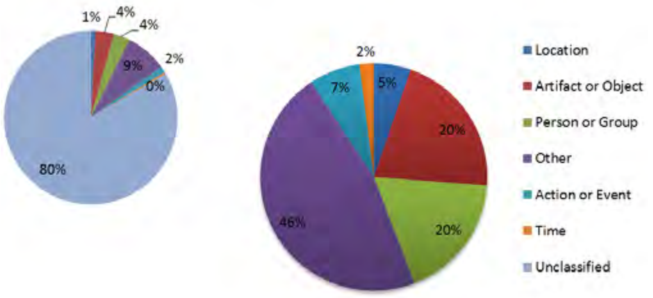


Figure 2.5: Tags per WordNet category for New York.

2.2.3 Analysis based on WordNet

Finally, we analyze what and how users tag by categorizing the tags based on WordNet. We use the broad categories of WordNet and if there exist multiple categories for one tag, this tag is associated with the category of the highest rank. Figure 2.5 presents the distribution of tags for New York based on the most popu-

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

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

lar categories of WordNet. Following this approach, approximate 20% of the tags can be categorized based on WordNet, leaving 80% of the tags without any category. We depict also in higher details the categorization of the 20% of the tags. By taking into account only the tags that can be categorized, the most frequent categories are ”person or groups” (appr. 20%) and ”artifact or object” (appr. 20%), followed by ”action or event” (appr. 8%), ”location” (appr. 5%), and ”time” (appr. 2%). The category ”Other” (appr. 45%) contains the tags that belong to some category of WordNet, but do not belong to any of the aforementioned categories. We can conclude that the users tag photos not only based on their features, but also based on the information the photo depicts, such as the time taken or the event and the location that is depicted. Similar results hold also for London and Rome data collection.

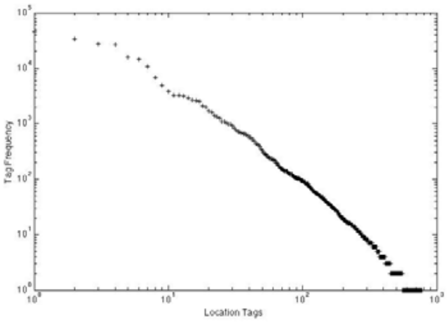


Figure 2.6: Location tag frequency distribution (New York)

Since in this report location tags are important, we analyze in more details the location based tags. For the New York data collection it holds that from all unique tags only 777 refer to a location based on WordNet. For the Rome data collection only 411 tags are tags referring to a location based on WordNet, while for London there exist 877 unique location tags. Figure 2.6 depicts the frequency

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

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

of the location based tags in log-log scale for New York data collection. The x-axis represents the set of unique location tags order based on the frequency in descending order. The y-axis is the tag frequency. We observe that the tag frequency can be modeled by a power law and this holds also for the other data collections.

Tag	Freq.
New York	45809
New York City	33941
manhattan	27282
NY	26717
USA	15957
City	14637
Brooklyn	10741
United States	6788
america	4853
park	3842

Table 2.5: New York.

Tag	Freq.
rome	56660
italy	44842
italia	19281
Lazio	8883
Vatican City	3067
city	2433
Latium	2002
Piazza	1781
town	604
Umbria	401

Table 2.6: Rome.

Tag	Freq.
London	68250
UK	30839
England	25760
trafalgar square	11090
United Kingdom	10023
Westminster	8404
City	5332
Great Britain	4303
Britain	3870
surrey	2196

Table 2.7: London.

Tables 2.5-2.7 show the 10 most popular location-based tags for New York, Rome and London respectively. There exist 227 tags with location-based tag frequency 1 for the New York collection. To give an example of their relevance we declare 10 random of them (for the New York collection): vienna, Nepal, Ohio, Bali, Calgary, praia, oslo, Cali, Rio de Janeiro, liverpool, St. Petersburg. Similar for Rome and London there exists 130 and 235 tags with frequency 1. Due to the small number of tags that can be categorized as location tags based on WordNet, but also due their relatively low frequency (i.e., Table 2.6) it is not possible to enhance our recommendation method using the WordNet categories.

2.3 Recommendation Methods

In this section we describe our recommendation methods. The input of our methods is a photo p that is described by a location given by the owner of the photo

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

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

and a set of tags $\{t_1, t_2, \dots\}$. The goal is to recommend to the use a set of relevant tags $\{t'_1, t'_2, \dots\}$ that could augment the description of p. Our methods rely on *tag co-occurrence*, i.e., the identification of tags frequently used together to annotate a photo. Furthermore, we enhance tag recommendation by taking explicitly into account the location of photos, in order to derive more meaningful co-occurring tags.

2.3.1 System Overview

Figure 2.7 gives a crisp overview of our location-aware tag recommendation system. Our system is built on an existing collection of photos that are geotagged, such as a subset of geotagged photos provided by Flickr. This information is necessary in order to identify frequently occurring tags, as well as to discover keywords that are used together as tags in many photos.

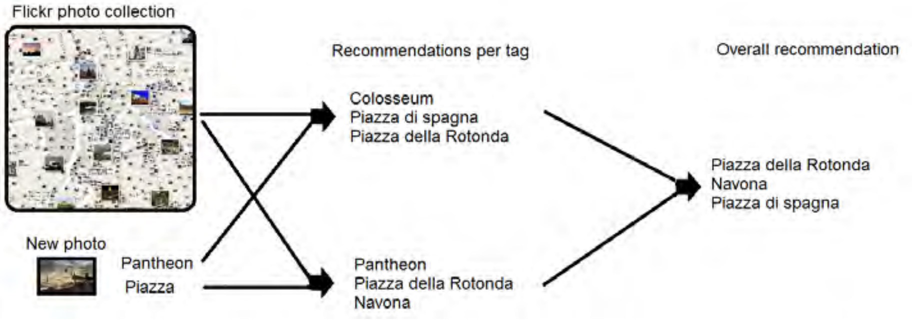


Figure 2.7: System overview.

We adopt a two-phase approach: in the first phase, a set of frequently co-occurring tags is discovered for each input tag $\{t_1, t_2, \dots\}$, while in the second phase, these sets of tags are combined to produce the final tag recommendation. In more details, for each given tag t_i a ranked list of n relevant tags to t_i is retrieved

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

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

based on the tag co-occurrence and the distance between the given photo and the photos in which the tags co-occur. Each tag is associated with a score that expresses its relevance to given tag t_i. Then, in the second phase, the different lists of relevant tags are combined, by aggregating their partial scores, so that the k most relevant tags are recommended to the user.

Even though different aggregation functions are applicable, we employ a plain strategy of summing the partial scores. Thus, for each tag t'_i, the overall score is defined as the sum of its scores in the ranked lists. Our goal is to produce more qualitative recommendations, by taking into account the location of the photo as well as the location of the existing tags.

2.3.2 Tag Recommendation Methods

We employ three different tag recommendation methods: (a) simple tag co-occurrence, (b) range tag co-occurrence, and (c) influence tag co-occurrence. The first method is location-independent and is used as a baseline, while the other two are novel, location-aware methods for tag recommendation.

Simple Tag Co-occurrence Method (Baseline)

The simplest way to measure the relevance of an existing tag to a given tag is tag co-occurrence. Assuming that t_i is the given tag and t_j an existing tag, then we denote $\mathcal{P}_i$ (or $\mathcal{P}_j$) the sets of photos in which tag t_i (or t_j) appear. To compute the co-occurrence of tags t_i and t_j, we need a metric for set similarity. One commonly used metric to express the similarity based on co-occurrence is the Jaccard coefficient, which is defined as the size of the intersection of the two sets divided

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

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

by the size of their union. Thus, for tags t_i and t_j, the Jaccard similarity is defined as:

$$Jaccard(t_i, t_j) = \frac{|\mathcal{P}_i \cap \mathcal{P}_j|}{|\mathcal{P}_i \cup \mathcal{P}_j|}.$$

Range Tag Co-occurrence Method

One major shortcoming of the simple tag co-occurrence method is that it does not take into account the location of the photo. Intuitively, it is expected that photos that are taken at nearby locations will share common tags, while photos taken far away from each other are less probable to be described by they same tags. This intuition guides the design of both location-aware methods that we propose. Given a radius r and a geo-tagged photo p, we define as $\mathcal{R}(p)$ the set of photos in our data collection that have a distance smaller than r to the location of the given photo p. In other words, photos in the set $\mathcal{R}(p)$ have been geo-tagged with a location that is within distance r from the location of the input photo p. Then, we define a novel measure that combines tag co-occurrence with location information:

$$Range(t_i, t_j) = \frac{|\mathcal{P}_i \cap \mathcal{P}_j \cap \mathcal{R}(p)|}{|\mathcal{P}_i \cup \mathcal{P}_j|}.$$

In this way, for tag co-occurrence, we take into account only the pairs of photos in which both tags appear and are geo-tagged withing a distance r. On the other hand, we divide with the total number of photos in which at least one of the tags appears, thus giving a penalty to tags that appear very often in photos that are distant to each other (i.e., outside the range r).

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

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

Influence Tag Co-occurrence Method

One drawback of range tag co-occurrence method is that a radius r needs to be defined as input, and it is not always straightforward how to set an appropriate value, without knowing the distribution of the locations of existing photos. Moreover, the defined range enforces a binary decision to whether a photo will be included or not in the tag co-occurrence computation, based on its distance being above or below the value r. For example, a very small value of radius may result in no photos with the given tag being located into the range, while on the other hand a large radius may result in most (or all) of the photos being located inside the range. Summarizing, the recommended tags are quite sensitive to the value of the radius, which is also hard to define appropriately.

To alleviate this drawback, we propose also a more robust and stable method than the plain range tag co-occurrence method. Given a radius r and a geo-tagged photo p, we define the *influence score* of two tags t_i and t_j as:

$$inflscore(t_i, t_j) = \sum_{p' \in \mathcal{P}_i \cap \mathcal{P}_j} 2^{\frac{-d(p',p)}{r}}$$

, where $d(p', p)$ is the distance between the locations of p and p'. Then the relevance of a given tag t_i and an existing tag t_j is computed as:

$$Influence(t_i, t_j) = \frac{inflscore(t_i, t_j)}{|\mathcal{P}_i \cup \mathcal{P}_j|}.$$

The key idea behind the influence score is that tags that co-occur in nearby photos have a higher influence than tags that co-occur in distant photos. This is nicely captured in the above definition by the exponent, which gradually decreases the

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

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

contribution of any photo p' the further it is located from p. Compared to the range tag co-occurrence method, this method does not enforce a binary decision on whether a photo will contribute or not to the score. Also, even though a radius r still needs to be defined, this practically has a smoothing effect on the influence score (rather than eliminating some photos), thus the score is not very sensitive to the value of r.

2.4 Experimental Evaluation

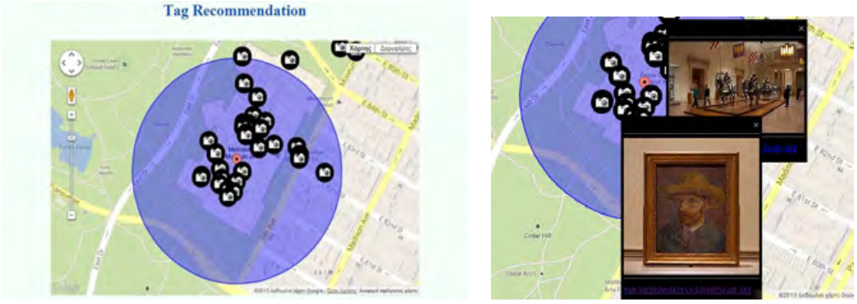
2.4.1 Prototype System

In order to evaluate experimentally our proposed recommendation methods we implemented a prototype system. Our prototype system displays to the user a map by using Google maps and the user can upload a new photo by providing its location (latitude and longitude). Then, in order to use the tag recommendation methods the user is asked to give the radius of interest as well as some initial tags. The recommendation query is posed and the systems displays on the map to the user the location of the new photo, the photos that participate in the recommendation query as well as the recommended tags(Figure 2.8(a)). The user can as depicted in Figure 2.8(b).

In our example, the new photo is uploaded at the location of the Metropolitan Museum of Art in New York (latitude:40.7789 and longitude:-73.9637) and one tag is given by the user namely 'The Metropolitan Museum of Art'. The user decides to use the Influence Recommendation Method and sets the radius to 200 and requests the 3 best matching tags. The recommendation tags are: 'The Met',

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

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



(a) Recommendation Example

(b) Photos used for Recommendation

Figure 2.8: Example of prototype system.

’Greek and Roman art’ and ’Manhattan’.

2.4.2 Experimental Evaluation

In this section, we provide examples of the proposed recommendation methods of Section 2.3. To this end, we take into account also the conclusions drawn in Section 2.2. Therefore, to avoid tags that are too generic to be helpful for recommendation, we exclude from the recommendation tags that appear in more than 10% of the photos. Also, we remove from our photo collection photos that have more than 30 tags, as these tags cannot be considered to be representative for the photo. Moreover, photos that have only one tag cannot be used for tag recommendation that rely on co-occurrence of tags, therefore such photos are also removed from the photo collections.

In order to measure the distance between two photos, we convert the longitude and latitude of each photo to the Universal Transverse Mercator (UTM) projected coordinate system. Then, we apply the Euclidean distance in this transformed space.

In our first example we use the New York data collection. Assuming a user

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

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

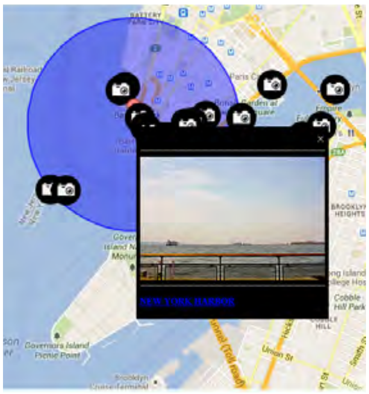


Figure 2.9: Example of recommendation.

that uploads to Flickr a photo taken at the Battery Park $(40.703294, -74.017411)$ in the Lower Manhattan of New York. The user gives one tag to the photo namely ”New York Harbor”. Figure 2.9 shows our prototype system for this query. The recommendation results are shown in Table 2.8. In this example we study how the radius influences our two approaches, while the Baseline fails to recommend relevant tags (”Newtown Creek”, ”Maspeth, New York” and ”DUGABO”). We notice that Range is more sensitive to the radius than Influence. Table 2.8 shows also the number of photos that fall into the region of radius r. This explains the behavior of Range, as for small radius values there exist too few photos to make meaningful recommendations.

Our next example uses again the New York data collection and this time a new photo is located nearby Time Square and the query point location is 40.756116, -73.986409. The given tag by the user is ”Broadway”. The results are depicted in Table 2.9. In this example, we notice that even for small radius the Influence method manage to retrieve relevant tags, while Range fails for small radius due to the low number of existing photos. On the other hand, both Range and Influence manage to retrieve relevant tags for higher radius values, while Baseline returns

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

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

more general tags like ”Madison Ave”.

In the following example we use the Rome data collection. We assume that the given photo is located in Vatican City (query location: 41.903491,12.453214) and it is annotated with the tag ”Museum” and the results are shown in Table 2.10. We notice that for small values of radius Range fails to return relevant tags due to the low number of existing photos. On the other hand Influence is influenced by very co-occurred tags like ”painting” even for higher radius values, because these two tags appear at many photos together and even if the distance is larger their score is aggregated and alters the final result.

In the next example (Table 2.11) we study the case of a photo that is annotated by 2 tags before the tag recommendation. We use the Rome data collection and we assume that the photo is taken at Piazza della Rotonda in front of Pantheon (41.899134, 12.47681). We set the radius equal to 100 since in the historical center of Rome there are many nearby photos. Location-aware tag recommendation manages to give relevant tags also for generic terms like ”Piazza”. For ”Piazza” and ”pantheon” query, the Baseline returns the same results as ”Piazza” because there is a higher co-occurrence between this tag and the others, while for the location-aware approaches the results are the same as ”pantheon” because there are more photos with this tag nearby the given location.

Finally, we examine another example in which 2 tags are given (”Buckingham Palace” and ”park”). This time we use the London data collection and the photo is located on the Birdcage Walk in front of the St. James’s Park (51.501011, -0.133268). The radius is set to 500 and the results are depicted in Table 2.12. This example tries to illustrate a hard case, as one of the tags (i.e, ”Buckingham Palace”) is not directly related to the location and the other tag (i.e., ”park”) is

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

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

quite generic. We notice that Range fails to return ”St. James’s Park” as a recommended tag, which is probably the most related term based on the location, but still both Range and Influence manage to recommend more relevant tags than the baseline.

2.5 Related Work

Automatic tag recommendation in social networks has emerged as an interesting research topic recently [21]. Especially in the case of Flickr, tag recommendation has been studied in [20, 11]. In more details, [20] presents different tag recommendation strategies relying on relationships between tags defined by the global co-occurrence metrics. On the other hand, in [11] tag recommendation methods are studied that are personalized and use knowledge about the particular user’s tagging behavior in the past. Nevertheless, none of the above methods takes into account the locations of photos. SpiritTagger [16] is a geo-aware tag suggestion tool for photos, but the proposed approach relies on the visual content (such as global color, texture, edge features) of the photo and on the global and local tag distribution. In contrast, our approach takes into account the tag co-occurrence and the distance between the given and the existing photos.

An overview of the field of recommender systems can be found in [1]. A framework that decouples the definition of a recommendation process from its execution and supports flexible recommendations over structured data has been proposed in [13, 14]. Neighborhood-based tag recommendation is studied in [4]. The neighborhood is defined based on a graph and tags are propagated through existing edges.

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

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

In [19] the authors also focus on geo-tagged photos and propose methods for placing photos uploaded to Flickr on the World map. These methods rely on the textual annotations provided by the users and predict the single location where the image was taken. This work is motivated by the fact that users spend considerable effort to describe photos [2, 20] with tags and these tags relate to locations where they were taken.

2.6 Conclusions

Tag recommendation is a very important and challenging task, since it helps users to annotate their photos with more meaningful tags, which in turn enables retrieving relevant photos from large photos collections such as Flickr. Nowadays, more and more photos are geotagged, and therefore in this report we investigate how to improve tag recommendation based on the spatial and textual information of the photos. To this end, we analyzed the tags of geotagged photos collected from Flickr and proposed two different location-aware tag recommendation methods. Our experiments show that location-aware tag recommendation is promising and the location of a photo improves the quality of the recommendation. In the future, we aim to investigate in depth how different existing recommendation methods can be improved by combining them with the photo locations.

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

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

Radius	Photos	Range	Influence
500	1098	Frederic Bartholdi, nite, lens adapters	One New York Plaza, Statue of Liberty, Harbor
1000	3828	One New York Plaza, Harbor Statue of Liberty	One New York Plaza, Statue of Liberty, Harbor
1500	6117	One New York Plaza, Harbor, Statue of Liberty	Liberty Island, Statue of Liberty, Harbor
2000	8816	Harbor, One New York Plaza, Statue of Liberty	Liberty Island, Staten Island Ferry, Statue of Liberty

Table 2.8: New York Harbor (Baseline recommends: ”Newtown Creek”, ”Maspeth, New York”, ”DUGABO”).

	Baseline	Range		Influence	
		100	1000	100	1000
1	peeps	Times Square	Times Square	Times Square	Times Square
2	Hood	nikkor 24-70mm f2.8	theatre	lights	theatre
3	Madison Ave	Silver Efex Pro2	Theater District	Theater District	Theater District
4	Lexington Ave	lights	Musical	neon	Musical

Table 2.9: Broadway.

Radius	Photos	Range	Influence
100	219	Musei Vaticani, heritage, DMC-GF1	painting, Musei Vaticani, Vatican Museum
500	11486	Musei Vaticani, Vaticano, Vatican	Musei Vaticani, painting, Vaticano
1000	14450	Musei Vaticani, Vaticano, Vatican	museo, painting, Musei Vaticani, Vaticano
1500	17914	Musei Vaticani, Vaticano, Vatican	museo, Musei Vaticani, Vaticano

Table 2.10: Museum (Baseline recommends: ”museo”, ”Musei Vaticani”, ”sculpture”).

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

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

Query	Baseline	Range	Influence
Piazza	Navona, spagna, popolo	pantheon, Rotonda, della	pantheon, Navona, Rotonda
pantheon	colosseum, piazza di spagna, Piazza della Rotonda	Piazza della Rotonda, temple, Dome	Piazza della Rotonda, temple, Dome
Piazza and pantheon	Navona, spagna, popolo	Piazza della Rotonda, temple, Dome	Piazza della Rotonda, temple, Dome

Table 2.11: Rome at Piazza della Rotonda (radius=100).

	Baseline	Range	Influence
1	hyde	roadrace	the mall
2	Green Park	Piccadilly London	Green Park
3	the mall	Road Race Cycling	st james park'
4	Constitution Hill	the mall	Piccadilly London

Table 2.12: ”Buckingham Palace” and ”park”.

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

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

Bibliography

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(6):734–749, 2005.

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*, pages 971–980, 2007.

[3] P. Bouros, S. Ge, and N. Mamoulis. Spatio-textual similarity joins. *PVLDB*, 6(1):1–12, 2012.

[4] A. Budura, S. Michel, P. Cudré-Mauroux, and K. Aberer. Neighborhood-based tag prediction. In *Proceedings of Extended Semantic Web Conference (ESWC)*, pages 608–622, 2009.

[5] X. Cao, G. Cong, B. Cui, C. S. Jensen, and Q. Yuan. Approaches to exploring category information for question retrieval in community question-answer archives. *ACM Transactions on Information Systems*, 30(2):7, 2012.

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

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

[6] X. Cao, G. Cong, and C. S. Jensen. Retrieving top-k prestige-based relevant spatial web objects. *PVLDB*, 3(1):373–384, 2010.

[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In *Proc. of the Int. Conf. on Management of Data (SIGMOD)*, pages 373–384, 2011.

[8] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. *PVLDB*, 2(1):337–348, 2009.

[9] Y. Du, D. Zhang, and T. Xia. The Optimal-Location query. In *Proc. of the Int. Symposium on Spatial and Temporal Databases (SSTD)*, pages 163–180, 2005.

[10] I. D. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In *Proc. of Int. Conf. on Data Engineering (ICDE)*, pages 656–665, 2008.

[11] N. Garg and I. Weber. Personalized, interactive tag recommendation for flickr. In *Proceedings of ACM Recommender System Conference (RecSys)*, pages 67–74, 2008.

[12] I. Kamel and C. Faloutsos. Hilbert r-tree: An improved r-tree using fractals. In *Proc. of the Int. Conf. on Very Large Data Bases (VLDB)*, pages 500–509, 1994.

[13] G. Koutrika, B. Bercovitz, and H. Garcia-Molina. Flexrecs: expressing and combining flexible recommendations. In *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, pages 745–758, 2009.

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

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

[14] G. Koutrika, B. Bercovitz, R. Ikeda, F. Kaliszan, H. Liou, and H. Garcia-Molina. Flexible recommendations for course planning. In *Proceedings of International Conference on Data Engineering (ICDE)*, pages 1467–1470, 2009.

[15] Z. Li, K. C. K. Lee, B. Zheng, W.-C. Lee, D. L. Lee, and X. Wang. Ir-tree: An efficient index for geographic document search. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 23(4):585–599, 2011.

[16] E. Moxley, J. Kleban, and B. S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *Proceedings of Multimedia Information Retrieval*, pages 24–30, 2008.

[17] J. B. Rocha-Junior, O. Gkorgkas, S. Jonassen, and K. Nørvåg. Efficient processing of top-k spatial keyword queries. In *Proc. of the Int. Symposium on Spatial and Temporal Databases (SSTD)*, pages 205–222, 2011.

[18] J. B. Rocha-Junior, A. Vlachou, C. Doulkeridis, and K. Nørvåg. Efficient processing of top-k spatial preference queries. *PVLDB*, 4(2):93–104, 2010.

[19] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 484–491, 2009.

[20] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of International World Wide Web Conference (WWW)*, pages 327–336, 2008.

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

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

[21] Y. Song, L. Zhang, and C. L. Giles. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, 5(1):4, 2011.

[22] T. Xia, D. Zhang, E. Kanoulas, and Y. Du. On computing top-t most influential spatial sites. In *Proc. of the Int. Conf. on Very Large Data Bases (VLDB)*, pages 946–957, 2005.

[23] M. L. Yiu, X. Dai, N. Mamoulis, and M. Vaitis. Top-k spatial preference queries. In *Proc. of Int. Conf. on Data Engineering (ICDE)*, pages 1076–1085, 2007.

[24] M. L. Yiu, H. Lu, N. Mamoulis, and M. Vaitis. Ranking spatial data by quality preferences. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 23(3):433–446, 2011.

[25] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. In *Proc. of Int. Conf. on Data Engineering (ICDE)*, pages 688–699, 2009.

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

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