Aissi *et al. Hum. Cent. Comput. Inf. Sci.* (2015) 5:30
DOI 10.1186/s13673-015-0045-y

Human-centric Computing and Information Sciences
a SpringerOpen Journal

RESEARCH Open Access

A spatial data warehouse recommendation approach: conceptual framework and experimental evaluation

Saida Aissi[1*], Mohamed Salah Gouider[1], Tarek Sboui[2] and Lamjed Ben Said[1]

*Correspondence:
saida.aissi@gmail.com
[1] High Institute of Management, SOIE Laboratory, Tunis, Tunisia
Full list of author information is available at the end of the article

Abstract

Spatial data warehouses store a large amount of historized and aggregated data. They are usually exploited by Spatial OLAP (SOLAP) systems to extract relevant information. Extracting such information may be complex and difficult. The user might ignore what part of the warehouse contains the relevant information and what the next query should be. On the other hand, recommendation systems aim to help users to retrieve relevant information according to their preferences and analytical objectives. Hence, developing a SOLAP recommendation system would enhance spatial data warehouses exploitation. This paper proposes a SOLAP recommendation approach that aims to help users better exploit spatial data warehouses and retrieve relevant information by recommending personalized spatial MDX (Multidimensional Expressions) queries. The approach detects implicitly the preferences and needs of SOLAP users using a spatio-semantic similarity measure. The approach is described theoretically and validated by experiments.

Keywords: Recommendation, Personalization, Spatial data warehouse, Spatial OLAP system, Semantic similarity, Spatial similarity

Background

Data warehouses (DW) are being considered as efficient components of decision support systems [7]. They are usually structured according to the multidimensional structure (also called a cube), which facilitates a rapid navigation within different levels of data granularity (from coarser to finer level and vice versa). Spatial Data warehouses (SDW) store a large amount of historized spatial data which have specific characteristics such as topology and direction. SDW can be explored by SOLAP systems (Spatial On-Line Analytical Processing) to enable spatial online analysis. SOLAP systems combine both GIS and OLAP (On-Line Analytical Processing) technologies. They offer, to users, opportunities for spatial analysis of geo-localized data by allowing them to visualize and navigate through aggregated spatial data according to a set of dimensions with different levels of granularity. SOLAP users can exploit spatial data warehouses by launching a sequence of MDX (Multidimensional Expressions) queries.

Still, extracting interesting information from SDW could be complex and difficult; Users might ignore what part of the warehouse contains the relevant information and

what the next query should be. The user facing a large amount of complex spatial data does not know where to find relevant information, and how to use them [15, 29].

On the other hand, recommendation systems aim to help users to navigate large amounts of data, and to discover relevant information according to their preferences and analytical objectives. Hence, developing a Spatial OLAP (SOLAP) recommendation system would facilitate information retrieval in spatial data warehouses.

In this paper, we propose to enhance spatial data warehouse exploitation by recommending personalized MDX queries to the user taking into account his preferences and analysis needs. The user's needs and preferences regarding the data stored in the SDW are detected implicitly during the recommendation process using a spatio-semantic similarity measure. To the best of our knowledge, there is no developed recommendation approaches in the field of SOLAP systems taking into account the specific characteristics of spatial data.

More precisely, our contribution consists in proposing (1) a framework for using spatio-semantic similarity measures to search the log of an SOLAP server to find the set of candidate relevant queries matching the current query. (2) Generating recommendations and classifying the recommended queries in order to present first the most relevant ones. (3) Implementing the approach and evaluating its efficiency.

This paper is organized as follows: "Recommendation approach: state of the art" presents a state of the art on recommendation approaches in DWs. "A spatio-semantic similarity measure between MDX queries" presents the proposed spatio-semantic similarity measure between developed MDX queries. "Motivating example" presents a motivating example explaining the usefulness of our proposal. "Personalized recommendations of SOLAP queries: conceptual framework" presents the theoretical framework of the proposed approach. "Experimental evaluation" presents the set of experiments conducted to test the effectiveness of the proposal. Finally, the conclusion and future works are presented in "Conclusion".

Recommendation approach: state of the art

Recommendation is a research topic which aims to help user finding relevant information according to his preferences and analytical objectives. Recommendation has been the subject of several studies in the fields of Information Retrieval [1] and in the Web Usage Mining [28]. In recent years, several academic studies have been conducted for personalizing OLAP systems [5, 9, 11, 13, 22, 24, 37] and databases [6, 34, 43, 44].

In the context of databases [41], propose a classification of databases recommendation techniques into three categories: (1) 'Current-state' approaches exploiting the content and the schema of the current query result as well as the database instance, (2) 'History-based' approaches using the query logs for recommendation and (3) 'External sources' approaches exploiting resources external to the database. Current-state approaches could be based on (1) the local analysis of the properties of the result of user's query or (2) the global analysis of the properties of the database. However, the classification proposed by [41] does not include a category for hybrid techniques mixing current-state, history based or external source techniques.

In the context of OLAP, [16–21] define recommendation as a process that exploits previously stated information requirements as well as user's previous queries on the DW

and what they did during the previous session in order to recommend the next query to the current user.

Aissi et al. [2] distinguish two main research orientations in the domain of OLAP recommendation: (1) Collaborative recommendation approaches and (2) Individual recommendation approaches. In the collaborative recommendation approaches [17, 20, 30], the system recommends alternatives queries to the current user using his query and the query log containing previous user's queries. In the individual recommendation approaches [25, 26, 40], the system provides alternatives and anticipated recommended query using a user profile.

Finally, Negre et al. [33] propose a classification of OLAP recommendation approaches into four types: (1) Methods exploiting a user profile, (2) methods based on expectations, (3) methods exploiting query logs and (4) hybrid methods.

While recommendation has been widely explored in the context of traditional database and OLAP systems, to the best of our knowledge, there is no developed recommendation approaches in the field of SOLAP systems that takes into account the specific characteristics of spatial data as well as the specific analytical needs of spatial data warehouse users [4].

In the next sections, we detail our proposal of a Spatial DW recommendation approach.

A spatio-semantic similarity measure between MDX queries

The basic idea of the approach is to recommend personalized MDX queries to the current user of SOLAP system. As part of our approach, we propose to detect implicitly the user's preferences by comparing the preferences of the current user with the preferences of the previous users of the data warehouse. The idea of exploiting the similarity between user's preferences to provide recommendations is a popular technique in collaborative filtering recommendation approaches in several domains such as the classification of opinions [35], the transactional databases [40] and traditional non-spatial data warehouses [17, 33].

Queries launched by the user over the SOLAP system are key elements for analyzing user's behavior and preferences. The idea is to identify the similarity between the preferences of SOLAP users thought their MDX queries triggered on the system and to use this similarity to recommend, to the current user, personalized MDX queries. Developing a similarity measure between MDX queries is then a fundamental step in the recommendation process.

As SOLAP users handle spatial complex data (having specific characteristics such as topology, direction and distance) [14, 23, 31], when comparing MDX queries two aspects of similarity are considered, namely (1) semantic similarity and (2) spatial similarity.

Basic formal definitions

In order to introduce and itemize the similarity measure, we propose, in this section the formal definitions of the basic concepts used in our proposal for measuring spatio-semantic similarity between MDX queries.

Cube (multidimensional spatial data warehouse), schema and dimensions:

An N-dimensional Cube C is denoted C = (D1, D2,…,Dn, F) where:

For each i ∈ [1, n], Di is a dimension table of schema Sch (Di) = {Ai0,…,Aij}. For each dimension i ∈ [1, n], each attribute Aij describes a level of a hierarchy, j represents the depth of this level. Ai0 represents the lowest level which equals the primary key of Di.

We present in Fig. 1a multidimensional schema of a DW that was intended to support strategic analysis of the crop (production). The constellation schema diagram is presented using the formalism of Malinowski et al. [32]. It allows the analysis of the weight and the amount of the production according to the dimensions zone, time and product. It allows answering queries such as: "what is the total production of biological products in 2014 produced in North regions?", "what is the total production of high quality products in the suds regions in 2014?"

Query references:

Given a cube *C* and an MDX query *qc* over *C*. We define the set of the references corresponding to an MDX query as follows: $Rqc = \{R_1,...,R_n, M_1,...,M_N\}$ where: $\{R_1,...,R_n\}$: is the set of dimension members *Di* adduced in the MDX query. It represents the set of members of the dimension *Di* that is adduced from the SELECT and WHERE clause. $\{M_1,...,M_N\}$: is the set of measures used in the MDX query.

Example: Given the following MDX queries:

$$q1 : SELECT \{[Product].[All\ Products].[cereal]\}\ ON\ COLUMNS,$$
$$FROM\ [Production]\ WHERE\ \{[Measures].[Weight]\}\ AND$$
$$\{[Region].[All\ Rgion].[Region1].[zone1]$$

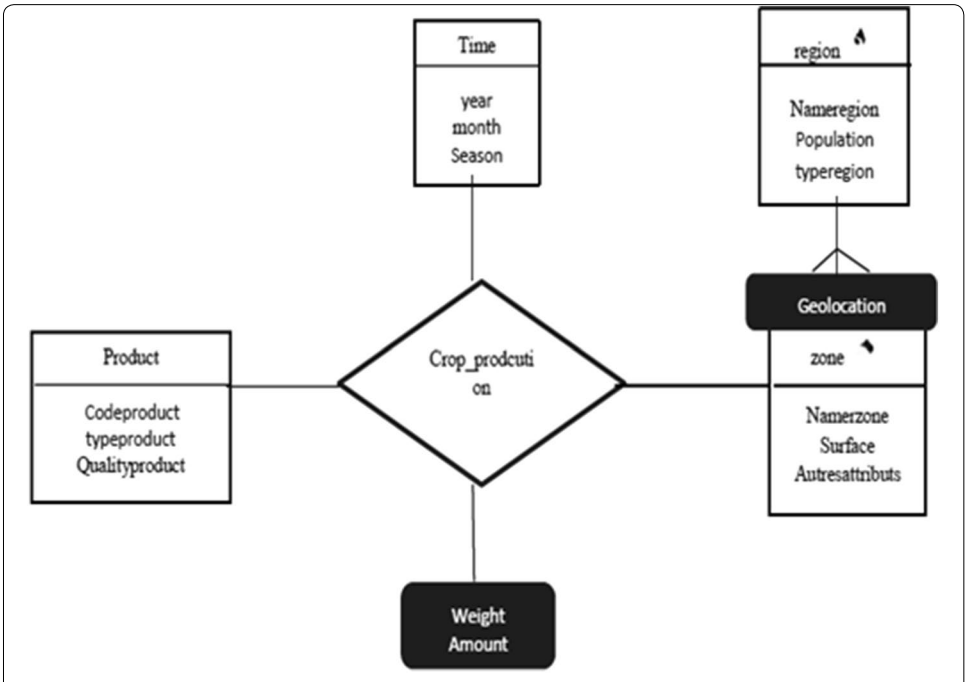

Fig. 1 A constellation schema for the analysis of the crop production

The references of the query *q1* are *Rq1 = {cereal, zone1, weight}*

$$q2 : SELECT \{[Product].[All\ Products].[vegetables]\}\ ON\ COLUMNS,$$
$$FROM\ [Production]\ WHERE\ \{[Measures].[amount]\}\ AND$$
$$\{[Region].[All\ Region].[Region1].[zone2].[zone3]\}$$

The references of the query *q2* are *Rq2 = {cereal, zone2, zone3, weight}*

Similarity measure for comparing MDX queries

Similarity measure between spatial MDX queries includes semantic similarity measure, topological similarity measure, metric similarity measure and direction similarity measure. To compute the similarity measure between two spatial MDX queries, we propose to compute the distance between these queries. That is, we compute the semantic distance, the topological distance, metric distance and direction distance.

Several similarity measures between concepts are proposed in the literature. The similarity measures are based on knowledge representation model offered by ontologies and semantic networks [39]. The concepts in our proposal are represented by the query references. To compute the semantic distance between references of each query we use an edge counting method by applying the Rada distance [36] using an application ontology representing the different concepts of the data warehouse model (dimensions, measures and attributes). The Rada distance computes the minimum number of edges which separate the query references in the ontology. We opted for the Rada distance because it is simple, accurate and efficient [36, 39].

Our approach for measuring spatio-semantic similarity is detailed in previous papers [2, 3]. In this paper, we illustrate how to compute spatial distance between MDX queries (topological distance, metric distance and direction distance). The illustration is based on the DW presented in Fig. 1.

Example of direction distance: *Given the following queries q1 and q2*

$$q1 : SELECT \{[Product].[All\ Products].[nonbiological]\}\ ON\ COLUMNS,$$
$$FROM\ [Production]\ WHERE\ \{[Measures].[Weight]\}\ AND\ \{[Region].$$
$$[All\ Region].[Region1].[Zone1]$$

The spatial references of the query q1 = {Zone 1}

$$q2 : SELECT \{[Product].[All\ Products].[biological]\}\ ON\ COLUMNS,$$
$$FROM\ [Production]\ WHERE\ \{[Measures].[Weight]\}\ AND$$
$$\{[Region].[All\ Region].[Region1].[Zone\ 2].[Zone\ 3]\}$$

The spatial references of the query q2 : {Zone 2, Zone 3}.

If the zone 1 exists in the north, the zone 2 in the northwest and the zone 3 in the south. Computing the directional distance between the query q1 and the query q2 refers to computing the directional distance between the couple of spatial references of each query.

$$direction - distance\,(q1,\ q2) = direction - distance\,(zone1, zone2) + direction - distance\,(zone1, zone3) = direction - distance\,(north,\ northwest) + direction - distance\,(north,\ south) = 2 + 6 = 8.$$

Example of metric distance: Given the previous queries q1 and q2. The spatial references of the query q1 = {Zone 1}. The spatial references of the query q2 are: {Zone 2, Zone 3}. If we have the distance between zone 1 and zone 2 is far and the distance between zone 1 and zone 3 is near. Computing the metric distance between the query q1 and the query q2 refers to computing the metric distance between the couple of spatial references of each query.

$$metric - distance\,(q1,\ q2) = metric - distance\,(zone1, zone2) + metric - distance\,(zone1,\ zone3) = 3 + 1 = 4.$$

Example of topological distance: Given two queries q1 and q2. The spatial references of the query q1 = {Zone 1}. The spatial references of the query q2 are: {Zone 2, Zone 3}.For example, we have zone 1 and zone 2 are disjoin and zone 1 contains Zone 3. The topological distance between q1 and q2is computed as follows:

$$topological - distance(q1,\ q2) = topological - distance\big(disjoin,\ contain\big) = 7.$$

Motivating example

In this section, in order to illustrate the usefulness of our approach for recommending personalized queries to SOLAP users, we continue with the example of the spatial data warehouse presented in Fig. 1.

Suppose two users of this spatial data warehouse; a User A and a User B. They are both responsible for agricultural animal production (e.g., red meat, white meat and milk) in the northern part of South Korea. The two users have the same preferences regarding the data stored in the data warehouse (agricultural animal production in the northern region).

The user A launch the following query q_A: "What is the total production of red meat in the area of Seoul?". We suppose that the user B has used the system and triggered the following query q_B "what is the total production of milk in the region of Suwon in 2014?." By analyzing the semantic and spatial similarity between the user's queries, we observe a semantic link (red meat and milk are two animal-based products) and a spatial link (Seoul and Suwon are two close regions located in the north of South Korea) between the two user's queries. Once the user A launches the query q_A, the system we developed recommends the query q_B to this user to assist him in the exploration of the spatial warehouse and to accelerate the process of research of relevant information. Thus, by analyzing the semantic and spatial similarity between queries, it is possible to implicitly identify the preferences and the analysis's objectives of the users in order to make useful recommendations.

Personalized recommendations of SOLAP queries: conceptual framework

The idea of the approach is to recommend personalized MDX queries to the user. The queries are adapted to user's preferences and objectives of analysis. Preferences of the

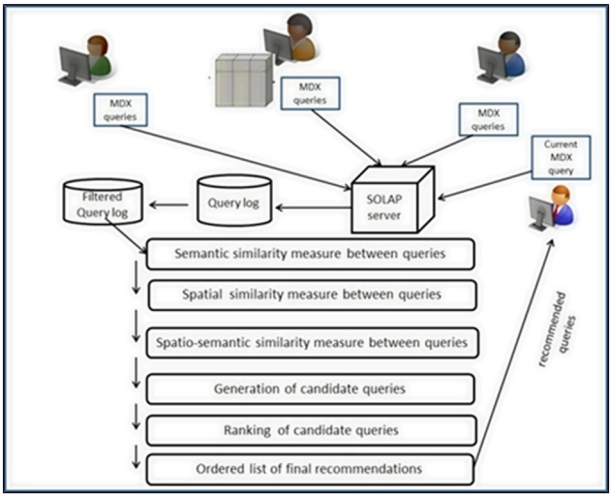

Fig. 2 Theoretical framework of SOLAP recommendation approach

user are detected implicitly using collaborative filtering technique by comparing the current user query with previous queries triggered by former users and recorded in a log file. The comparison between the current user query and previous queries is performed using the spatio-semantic similarity measure.

The recommendation approach is based on four main phases: (1) File log filtering, (2) Generation of candidate queries from the filtered log file, (3) Ranking of final recommendations and finally (4) the proposal of the relevant queries. Figure 2 explains the theoretical framework of the proposed approach.

In the following, we detail the different phases of the recommendation approach we proposed.

Log file filtering

The log file containing previous queries already launched on the DW can be very large because of the high number of queries and users. The time of recommendation can significantly increase. To address this problem, we propose to preprocess the log file to remove non-relevant queries in the recommendation process. The filtering criterion of the log file is the execution date (the age) of a query defined as a parameter of this phase to be settled by the user or the administrator of SOLAP system. Only relatively recent queries are considered in the recommendation process.

Generation of candidate queries

This phase allows to generate all candidate queries for recommendation from the initial log file after preprocessing. Generating candidate queries is based on measuring the spatio-semantic similarity between MDX queries. The approach we propose is based on a collaborative filtering technique that implicitly detects the current user preferences and needs by comparing his queries triggered on the system with the queries launched by previous users and logged in the log file. The most spatially and semantically similar queries to the current user query are presented in the list of candidate queries.

At this level, two methods for generating candidate queries are proposed. The first method is based on the selection of candidate queries having a similarity value, relative to the current query, equal or exceeding a predetermined threshold of spatio-semantic similarity. The spatio-semantic similarity threshold is a parameter defined by the user. The second method is based on the selection of the k most similar queries to the current query. The value of k is also a parameter specified by the user according to his preferences.

Each method has its advantages and disadvantages. The first method ensures a good quality of recommendations because queries that do not respect a defined threshold of similarity will be directly eliminated. However, this method may give an empty set of recommendations if the defined value of the threshold similarity is high.

As against, the method of the k most similar queries allows to guarantee a minimum number of recommendations, however, the quality of a recommendation is not sufficiently controlled.

Candidate queries generation using the threshold similarity

The algorithm *CQGS* (Candidate Queries Generation using the threshold Similarity) extract from the filtered log file the set of candidate queries that are similar to the current query taking into account a predetermined similarity threshold s. The algorithm takes as input the number of queries in the filtered log file, the current user query, the SIM function (which computes the spatio-semantic similarity between two MDX queries) and the similarity threshold *s*. SIM function is used to compute the spatio-semantic similarity values between the current query and the queries presented in the filtered log file.

Algorithm CQGS (qc, n,SIM, s)

Input
qc : current user query
n : number of queries in the filtered log file
SIM :a function that computes the spatio-semantic similarity between two MDX queries
s :a similarity threshold
Output
Cquery :The set of candidate queries

Cquery ← Ø
For I in 1..n **do**
If *SIM(qi, qc)*≥ s
then Cquery ← Cquery U {qc}
end if
end for
return Cquery

Generation of top-k similar queries

Based on the method of ***k*** most similar queries, the number of candidate queries is fixed and the recommendation system generates, from the log file, the *k* most similar queries to the current user. Regarding this method, no similarity constraint is imposed. The algorithm *GTSQ* (Generation of Top-k Similar Queries) presents the principle of generation of the recommendations using the top-k similar queries method.

Algorithm GTSQ (qc, log F, n,SIM, k)

Input:
log F : Set of queries in the filtered log file
n :Number of queries in the filtered log file
k :Number of queries to be generated
SIM :a function that computes the spatio-semantic similarity between two MDX query
qc **:**The [2]Current query
Output:
Cquery :The set of candidate queries

Cquery ◄—— *q1*
for I in 2..k **do**
Cquery ◄—— *{q1} U {qi}*
end for
for I in k+1..n **do**
if *the similarity between qi and qc is greater than the similarity of an element of Cquery relative to qi*
then *remove from Cquery the least similar query relative to qc AND Add qi in Cquery*
end if
end for

Ranking of candidate queries

Once candidate queries are generated, we classify them to be recommended to the user by order of relevance. At this level, we define three ranking criteria (1) Candidate queries ranking according to their occurrence frequencies in the log file (according to their use frequency) (2) queries ranking according to their execution date (the most recent queries are the most favored) and (3), candidate queries ranking according to the spatial and semantic proximity relative to the current query (Queries having a high similarity are the first to be recommended).

In order to have an efficient recommendation process, we propose a combined ranking method that includes the three ranking criteria. The following ranking order of candidate queries will be applied in the recommendation system: (1) the spatio-semantic similarity relative to the current query, (2) the frequency of use of a query and (3) the execution date of a query. Our choice of this ranking order is motivated by the following reasons:

Spatial and semantic similarity is considered as the most important factor to be used in the recommendation process. Indeed, this criterion reflects the preferences and interests of the user detected through the spatio-semantic similarity measure. Providing a spatially and semantically relevant query to the current user's one refers to offering a query that responds more to the user needs and analysis objectives.

Also, we have classified the age factor as the last ranking criterion because the original log file have been filtered according to this criterion and only the queries that are relatively recent according to user preferences. Thus, at this level, the use of age factor just makes a final classification of the most relevant candidate queries. The algorithm *RANKQ (Ranking Queries)* explains the ranking process.

Algorithm RANKQ(Q, ORDOSIM, ORDOTEMPS,ORDOFREQ)

Inputs

Q :The set of candidate queries

ORDOSIM: A ranking function of candidate queries using the spatio-semantic similarity

*ORDOTEPMS:*A ranking function of candidate queries using the execution date of a query

ORDOFREQ: A ranking function of candidate queries by use frequency of a query

Outputs:

listerecommandation :ordered list of final recommendations

Q ← The set of candidate queries
list1 ← *ORDOSIM* (Q)
list 2 ← ORDOFREQ (list1)
listerecommandation ← *ORDOTEPMS* (list2)
return *listerecommandation*

Proposal of final recommendations

After generating and ordering the set of candidate queries, the relevant recommendations should be proposed to the user. At this level, we need to address two points: First, we must specify the maximum number of recommendations to be proposed. Second, we must specify the action to be taken in case the set of candidate queries is empty.

Concerning the first point, we propose to introduce a maximum number of five recommendations for the user. For the following reasons:

We believe that when the number of recommendations exceeds five queries, the user will be tempted to read, analyze and compare the proposed queries to make his choice. This will cause an increase in the reflection time and SDW exploitation. In some cases, the user may not take into consideration the recommendation system when the number of proposals is not relevant.

Regarding the second point, we have considered unnecessary to provide a default recommendation for the user when the set of candidate queries is empty taking into account on the following principle "It is better not to make a recommendation than to propose an irrelevant recommendation". This allows keeping a good perception of the recommendation process in the mind of the user and encouraging him/her to take seriously each proposal

Experimental evaluation

Similarity measure evaluation

In order to evaluate the efficiency of the proposed similarity measure, we used the technique of human evaluation based on the Spearman's correlation coefficient (Spearman 1904). We asked 15 human subjects to choose 30 pairs of spatial MDX queries, and assign the degrees of similarity between these queries which have different degrees of spatiosemantic relatedness as assigned by the proposed similarity. Then, we calculate the value of the Spearman's correlation coefficient that express the correlation between the similarity values given for the 30 pairs of queries using our proposal and the similarities values provided by human evaluation. The obtained value of the Spearman's coefficient is equal to 0.82. This value express that there is a high degree of positive correlation

(0.82) between the similarity values accorded to the evaluated queries, using the human evaluation technique and the similarity values accorded to the evaluated queries using our proposal. Hence, the obtained correlation coefficient proves the efficiency of the proposed measure.

We developed the CoSIM (ComputeSIMilarity) system using Java language. CoSIM system implements our proposal of the spatio-semantic similarity measure. It identifies the references of a given MDX query and computes the semantic meaure/distance, the spatial distance/measure and the spatio-semantic similarity measure/distance between two MDX queries according to the crop productionSDW presented in Fig. 1. An example of spatio-semantic similarity measures/distances computed between MDX queries using CoSIM is presented in Fig. 3.

RECQUERY system

In order to evaluate the efficiency of the proposed approach, we developed the prototype RECQUERY (RECommend QUERY). RECQUERY implements all phases of the approach to provide user with useful and relevant queries. RECQUERY system works as follows: From (1) a log file containing MDX queries previously triggered on the system and (2) the current MDX query triggered by the user and by applying a global recommendation algorithm with predefined parameters, the RECQUERY system recommend an ordered set of MDX queries that may interest the current user and help him to better progress in SDWSDW exploitation.

In a first phase, the current user runs a query on the spatial data warehouse stored in the Data Base Management System MySQL, through the GeoMondrian SOLAP server. During the recommendation process, RECQUERY accesses information in the SOLAP server and recommends to the current user an ordered set of queries. Figure 4 presents different interfaces of RECQUERY system.

Hereafter, we present a set of conducted experiments to evaluate the efficiency of the approach. First, the performance of RECQUERY system is tested regarding the

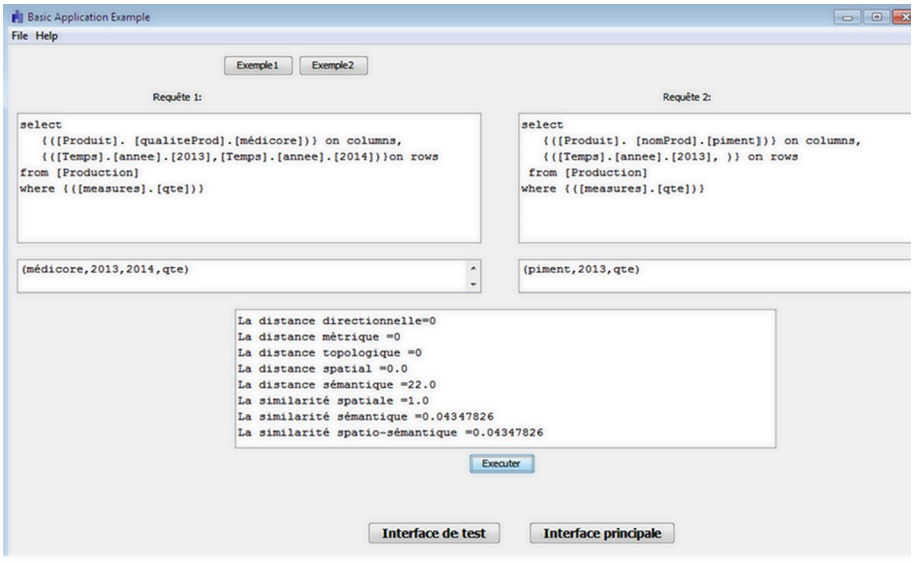

Fig. 3 CoSIM system: extracting MDX queries references

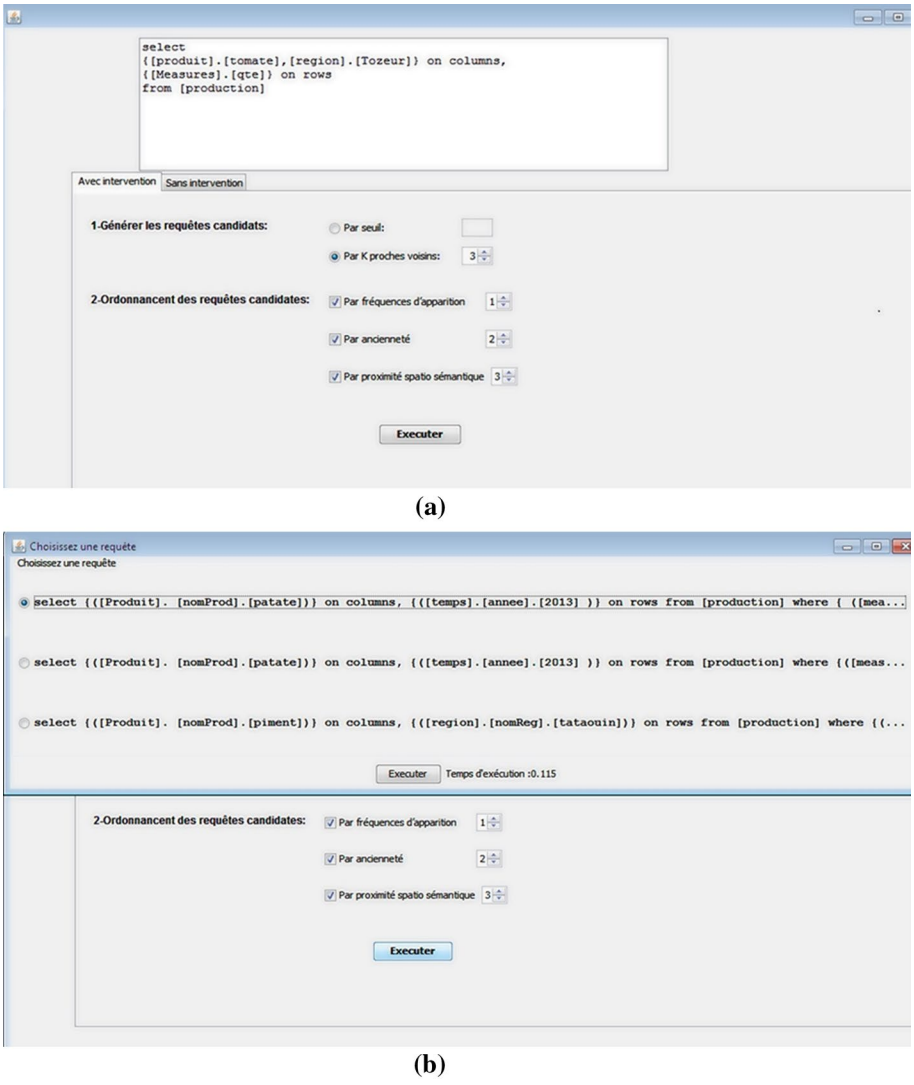

Fig. 4 RECQUERY; **a** generating and ranking of candidate queries, **b** proposal of personalized MDX queries to the user

time required to make recommendations for different size of the log file. Second, the approach is evaluated using the precision indicator.

Performance analysis

For Performance Analysis, We evaluate the time needed to present recommendations to the user for filtered log files with different sizes. The log size presents the total number of queries contained in the log file after preprocessing. We evaluate the time needed to present recommendations using the method of the k most similar queries to generate recommendations. k is fixed to five queries The results of this evaluation are presented in Fig. 5.

Figure 5 shows that the execution time varies linearly according the size of the original log file. However, it is still acceptable since it reaches a value equal to 2 s for a log file size equal to 3000 queries.

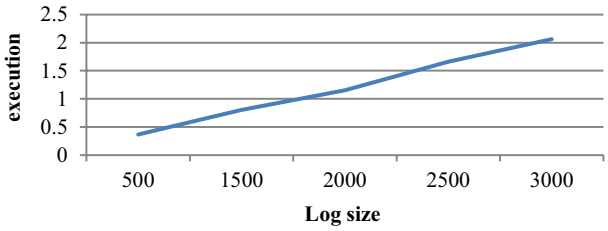

Fig. 5 Performance of RECQUERY

As there is no approach for SOLAP recommendation, we compare the performance analysis of our system with related work in (non-spatial) OLAP recommendation systems. We notice that OLAP recommendation may appear more performant than SOLAP (spatial OLAP) recommendation. For example, the work presented in [33] shows that the execution time does not exceed 0.6 s for a log file size equal to 3000 queries. This may appear more performant than our recommendation approach. But, we should notice here that for SOLAP recommendation, the execution time is expected to be longer than the one for non-spatial OLAP systems, since SOLAP systems handle spatial complex data [31] (having specific additional characteristics such as topology, direction and distance).

Precision evaluation

Evaluation technique

Precision is an indicator widely used to assess the quality and performance of recommendations in many areas [12, 35]. It reflects the proportion of recommendations that are in the user preferences [12]. In general, the precision is measured as follows:

$$Precision = |\{Relevant\ Recommandations\} \cap \{proposed\ Recommandations\}| / |\{Proposed\ Recommandations\}|$$

In our case, the proposed recommendations are presented by all the recommended queries (RC_{final}). The relevant recommendations are presented by all recommended queries accepted by the user (RC_{accp}). Thus, the precision is obtained as follows:

$$Precision = |\{RC_{final} \cap RC_{accp}\}| / \{RC_{final}\}$$

A human evaluation technique is used to calculate precision. 15 human subjects who already deal with Spatial OLAP systems and the MDX language are formed on the content of the agricultural spatial database presented in Fig. 1. We distinguished three groups of users. We asked each group to launch a set of queries to achieve some objectives of analysis. Analysis objectives are defined explicitly for each group of users. We asked the first group to focus on the analysis of organic agricultural production in the north, northeast and northwest of South Korea, the second group is interested in the analysis of animal agricultural production in the southern part of South Korea and the third group is interested in the analysis of vegetable agricultural production in all regions of South Korea. Thus we have defined explicitly for each group their analysis objectives and preferences in relation to the data in the SDWSDW. RECQUERY system will detect implicitly user's preferences and analysis objectives and propose personalized

recommendations to make information retrieval easier and faster by enhancing SDW exploitation. We asked each user to launch 10 queries to exploit the SDW and retrieve the needed information.

Precision evaluation for different similarity thresholds

The aim of this test is to study the efficiency of the recommendation system using the similarity threshold method for the generation of candidate queries. Figure 6 shows the results of this test.

Figure 6 shows that at least 25 % of the recommendations have been chosen by the user for relevant information retrieval. This precision rate increases proportionally according to the similarity threshold defined by the user. The precision rate reaches a value of 76 % for a threshold similarity equal to 0.8.

We also note that the precision varies proportionally to the defined similarity threshold. In fact, the higher the similarity threshold is, the greater the quality of recommendations and the precision rate are. However, for a similarity threshold which exceeds 0.8, the precision value starts to decrease. In fact, when the similarity threshold is relatively high, the number of candidate queries is reduced which implies a reduction of possible choices for the user. Note that the precision value depends also on the quality of the queries already registered in the log file. Indeed, the higher the number of queries having a high similarity with the current query is, the greater the precision value rate is.

Precision evaluation for different values of k

The purpose of this test is to evaluate the variation of the precision according to the number of candidate queries to be generated. We note that the number of the proposed queries is equal to five queries whatever is the value of k (the number of queries). Figure 7 shows the value of the precision obtained for different values of k.

Figure 7 shows that when the number of candidate queries is less than three queries, the precision rate is relatively low (at about 40 %) and increases linearly with k. The precision value remains almost unchanged when k is greater than 3, this is explained by the idea that the number of the proposed queries is equal to five queries whatever is the value of k fixed. We note that RECQUERY system gives good results by applying the method of the top-k most similar queries to generate recommendations. Indeed, at least 40 % of the proposed queries were triggered by users to progress in their analysis process, this rate is generally higher than 70 % (when k exceeds 2 queries) and reaches 76 % when the number of candidate queries is optimal.

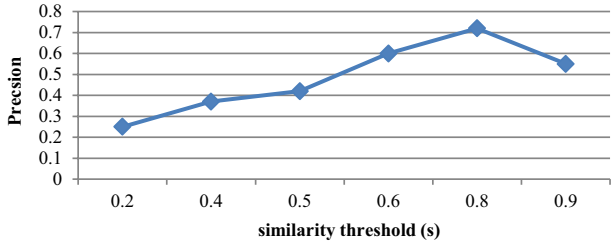

Fig. 6 Precision of RECQUERY for different similarity thresholds

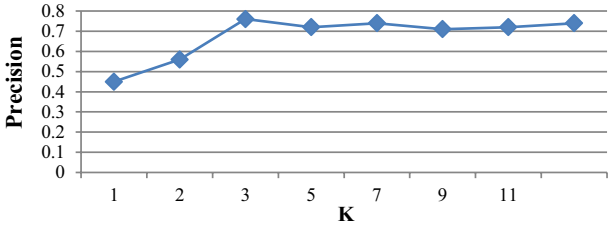

Fig. 7 Precision of RECQUERY for different values of k

As mentioned in the performance analysis, and since there is no approach for SOLAP recommendation, we compare the precision analysis of our system with related work in (non-spatial) OLAP recommendation systems. The precision in such systems are more or less precise than SOLAP (spatial OLAP) recommendation. For example, Giacometti et al. [17] developed an OLAP recommendation system that proposes to the user the next query based on the OLAP server query log. The execution of this system shows that the precision varies between 0.4 and 0.8 for a cluster quality above 0.7. A cluster may contain 50, 100 or 200 queries.

Such result may appear better than the result of our SOLAP recommendation system. However, we should take into account the fact that OLAP systems deal with non-spatial data, while SOLAP systems deal with complex spatial data [29, 38].

Another evaluation criterion that could make the comparison of OLAP and SOLAP systems difficult is the difference of the recommendation parameters and the difference of the recommendation processes used in both systems. For example, Giacometti et al. [20] used a group of similar queries (clustering) as a criterion to evaluate the recommendation precision. In our approach, we do not adopt such a clustering as we consider only similar queries (i.e., all candidate queries are similar).

Furthermore, in our approach, we have obtained a minimum of precision rate equal to 25 % regardless of the similarity threshold (s) or the number of candidate queries (k) (i.e., at least 25 % of the recommendations are chosen by users). The precision rate is above 40 % in most cases, and may reach 76 %. These values demonstrate the efficiency of our approach.

Optimal number of final recommendations

The purpose of this test is to define the optimal number (*f*) of final recommendations to be presented to the user. To do this, the value of the precision is evaluated according to different numbers of final recommendations. The results of this test are shown in Fig. 8.

The highest value of relevant recommendations is obtained for $f = 3$. When the number of final recommendations presented to the user exceeds three queries, the performance of the system decreases. In fact, when the user had a large number of recommendations (more than 3 queries), he will be disturbed and take more times to analyze and select the proposals. The user may also ignore the recommendation system and build his own query. Thus, the optimal number of final recommendations to submit to the user is three queries.

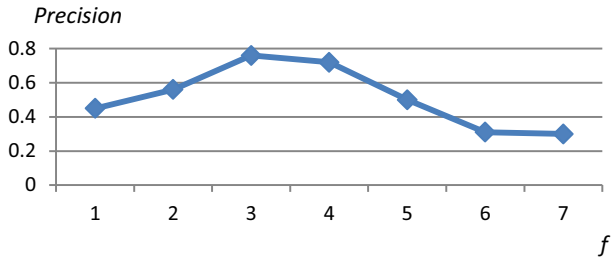

Fig. 8 Optimal number of candidate queries

Note that the number of final recommendations can slightly vary from one user to another according to his experience with MDX language as well as the structure and the content of the spatial data warehouse.

Conclusion

In this paper, we propose a personalization of SOLAP systems through a recommendation approach. The approach assists the user in spatial data warehouse exploitation through the recommendation of personalized MDX queries. The approach (1) detects the preferences and the analysis objectives of the user using a collaborative filtering technique, and (2) applies a spatio-semantic similarity measure between MDX queries to compare the analytical objectives of the users and their preferences. We presented a theoretical framework detailing the various phases of the approach namely (1) log file filtering, (2) generation of candidate queries (3) ranking recommendations, and (4) presentation of recommendations. Each step is explained by detailed algorithms presenting how these phases can be implemented.

We conducted also an experimental evaluation of the proposed system, we proposed the prototype RECQUERY that implements the different phases of the approach. We tested the quality of recommendations using human evaluation technique, and we presented the results of some experimental tests used to evaluate the efficiency of the approach. Experimental results show that recommendations can be computed efficiently and in most cases good and helpful recommendations are proposed. In fact, during various tests, at least 25 % of the recommendations have been triggered by the users to advance their information search process, this rate is in most cases above 40 % and reaches 76 %. The advantage of the approach is its flexibility since it allows to users and administrators to intervene in the system to choose the method of candidate queries recommendations and the ranking queries criteria. The importance of a flexible recommendation system is revealed in the work of Adomavicius and Tuzhilin [2].

As future works, we propose to develop the implicit extraction process of users preferences taking into account not just the triggered MDX queries on the system but also the SIG operations launched by the users on the system like pan, zoom and selection on spatial data.

At the recommendation process level, we propose to develop collaborative recommendation by performing firstly a clustering of SOLAP users into similar groups before applying the spatio-semantic similarity measure for the detection of user preferences.

Regarding the proposed spatio-semantic similarity measure, we propose to further develop this measure taking into considerations other similarity assessment criteria between spatial references such as geometric and non-spatial attributes of DW content.

Authors' contributions
SA and SG establish the general framework of the proposed approach, conduct and interpret the set of experiments and draft the article, TS proposes the plan of the paper and revised the paper after being writed. LBS proposes the global idea and directs the set of contributions presented in the paper and approved the final version of the article. All authors read and approved the final manuscript.

Author details
[1] High Institute of Management, SOIE Laboratory, Tunis, Tunisia. [2] Faculty of Sciences, University of Tunis El Manar, UR 11, ES 15, Tunis, Tunisia.

Acknowledgements
Iwould like to acknowledge le ISG students and teachers who participate in the evaluation of the recommendation approach and the in the human evaluation process of the proposed similarity measure. Thank you for your time. I would like also to acknowledge the ministry of agriculture for giving us the data to drive our experimental evaluation.

Compliance with ethical guidelines

Competing interests
The authors declare that they have no competing interests.

Received: 23 January 2015 Accepted: 21 August 2015
Published online: 08 October 2015

References

1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749
2. Aissa S, Gouider MS (2012) A new similairty measure for spatial personalization. Int J Database Manag Syst 4(4):1–12
3. Aissi S, Gouider MS, Sboui T, Bensaid L (2014) Enhancing spatial data cube exploitation: a spatio- semantic similarity perspective. ICIST 2014. Commu Comp Inf Sci Springer 465:121–133
4. Aissi S, Gouider M S, Sboui T, Bensaid L (2015) Personnalisation OLAP et SIG: etude comparative et perspectives de personnalisation SOLAP. J Decis Syst (in press)
5. Aligon J, Golfarelli M, Marcel P, Rizzi S, Turricchia E (2011) Mining preferences from OLAP query logs for proactive personalization. In: Proceedings advances in databases and information systems, Vienna, Austria, pp 84–97
6. Ballatore A, McArdle G, Kelly C, Bertolotto M (2010) RecoMap: an interactive and adaptive map-based recommender. Proceedings of the 2010 ACM Symposium on Applied Computing. ACM, pp 887–891
7. Bédard Y (1997) Spatial OLAP. 2ème Forum annuel sur la R-D, Géomatique VI: Un monde accessible, Montreal
8. Bédard Y, Merrett T, Han J (2001) Fundaments of spatial data warehousing for geographic knowledge discovery. Geographic data mining and knowledge discovery. Taylor and Francis, London
9. Bellatreche L, Giacometti A, Marcel P, Mouloudi H, Laurent D (2005) A Personalization framework for OLAP queries. DOLAP 05:9–18
10. Bellatreche L, Giacometti A, Marcel P, Mouloudi H (2006) A personalization of MDX queries. Bases de Données Avancées, BDA'06 Proceedings
11. Bentayeb F (2008) K-means based approach for OLAP dimension updates. ICEIS 08 Proceedings, pp 531–534
12. Bellogín A, Castells P, Cantador I (2011) Precision-oriented evaluation of recommender systems: an algorithmic comparison. RecSys'11 Proceedings of the fifth ACM conference on Recommender, pp 333–336
13. Biondi P, Golfarelli M, Rizzi S (2011) Preference-based datacube analysis with MYOLAP. ICDE Proceedings, pp 1328–1331
14. Bruns HT, Egenhover MJ (1996) Similarity of spatial scenes. Proceedings of the International Symposium on Spatial Data Handling, pp 31–42
15. Favre C, Bentayeb F, Boussaid O, Darmont J, Gavin G, Harbi N, Kabachi N, Loudcher S (2013) Les entrepôts de données pour les nuls… ou pas!. 2ème Atelier aide à la Décision à tous les Etages (EGC/AIDE 13)
16. Garrigós I, Pardillo J, Mazón JN, Zubcoff J, Trujillo J, Romero R (2012) A Conceptual modeling personalization framework for OLAP. J Database Manag 23(4):1–16
17. Giacometti A, Marcel P, Negre E (2008) A Framework for Recommending OLAP Queries. International Workshop on Data Warehousing and OLAP, ACM, pp 73–80
18. Giacometti A, Marcel P., Negre E, Soulet A (2010) Query recommendations for OLAP discovery-driven analysis. IJDWM Proceedings, pp 1–25
19. Glorio O, Mazón J, Garrigós I, Trujillo J (2010) Using web-based personalization on spatial data warehouses. EDBT/ICDT Workshops
20. Giacometti A, Marcel P, Negre E (2011) Query recommendations for OLAP discovery-driven analysis. IJDWM 7(2):1–25
21. Glorio O, Mazón J, Garrigós I, Trujillo J (2012) A personalization process for spatial data warehouse development. Decis Support Syst 52:884–898
22. Golfarelli M, Rizzi S (2009) Expressing OLAP preferences. SSDBM Proceedings, pp 83–91

23. Goyal R, Egenhofer M (2001) Similarity of Cardinal Directions. In: Jensen C, Schneider M, Seeger B, Tsotras V (eds) Seventh International Symposium on Spatial and Temporal Databases
24. Jerbi H, Ravat F, Teste O, Zurfluh G (2009) Preference-based recommendations for OLAP analysis. DaWaK Proceedings, pp 467–478
25. Jerbi H, Pujolle G, Ravat F, Teste O (2010) Personnalisation de Systèmes OLAP annotés. In: Proceedings of the 28th congrésINFORSID, pp 327–341
26. Jerbi H, Pujolle G, Ravat F, Teste O (2011) Recommandation de requêtes dans les bases de données multidimensionnelles annotées. Ingénierie des Systèmes d'Information, vol 16, no 1, pp 113–138
27. Kalthoum R, Hédia M, Ghédira K (2013) Theoretical formulas of semantic measure: a survey. J Emerg Technol Web Intell 5(4):333–342
28. Kaur D, Kaur R (2014) Minimizing the repeated database scan using an efficient frequent pattern mining algorithm in web usage mining. Int J Res Advent Technol 2(6):2321–9637
29. Khalissa DA, Frihi I, Boukhalfa K, Alimazighi Z (2013) De la Conception d'un Entrepôt de Données Spatiales à un Outil Géo-Décisionnel pour une Meilleure Analyse du Risque Routier. Proceedings of the congré INFORSID, pp 181–196
30. Khemiri R, Bentayeb F (2010) Utilisation des vues matérialisées pour la personalization des entrepôts de données. In: Proceedings of Atelier et Systèmes Décisionnels, ASD10, Tunisie (Sfax) 2010, pp 121–129
31. Li B, Fonseca F (2006) Tdd—a comprehensive model for qualitative spatial similarity assessment. Spat Cognit Comput 6(1):31–62
32. Malinowski E, Zimányi E (2004) Representing spatiality in a conceptual multidimensional model. Proceedings of the 12th annual ACM International workshop on Geographic information systems. ACM Press, pp 12–22
33. Negre E (2011) Quand la recommandation rencontre la personnalisation. Ou comment générer des recommandations (requêtes MDX) en adéquation avec les préférences de l'utilisateur. Tech Sci Inform 30(8):933–952
34. Petit M, Claramunt C, Ray C, Calvary G (2008) A design process for the development of an interactive and adaptive GIS. W2GIS Proceedings, pp 96–106
35. Poirier D, Fessant F, Tellier I (2010) De la classification d'opinion à la recommandation: l'apport des textes communautaires. Revue Traitenent autionatique de langues (TAL) 51:19–46
36. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. Syst Man Cybern IEEE Trans 19(1):17–30
37. Ravat F, Teste O, Zurfluh, G (2007) Personnalisation de bases de données multidimensionnelles. INFORSID Proceedings, pp 121–136
38. Ravat F, Teste O (2008) Personalization and OLAP databases. Ann Inform Syst New Trends Data Warehous Data Anal Springer 3:71–92
39. Rezgui K, Mhiri H, Ghédira K (2013) Theoretical formulas of semantic measure: a survey. J Emerg Technol Web Intell 5(4):333–342
40. Sapia C (2000) PROMISE : predicting query behavior to enable predictive caching strategies for OLAP systems. DaWaKV Proceedings, pp 224–233
41. Stefanidis A, Wang C, Xu L, Curtin KM (2009) Multilayer scene similarity assessment. ICDM Workshops 2009, pp 622–629
42. Tahir A, McArdle G, Bertolotto, M (2014) A Geovisual Analytics Approach for Mouse Movement Analysis. Int J Data Min Model Manag 6(4):315–332
43. Thalhammer T, Schrefl M, Mohania M (2001) Active data warehouses: complementing OLAP with analysis rules. Data Knowl Eng 39(3):241–269
44. Wilson D, Lipford H, Carroll E, Karr P, Najjar N (2008) Charting new ground: modeling user behavior in interactive geovisualization. Proceedings of the GIS'08 16th International Symposium on Advances in Geographic Information Systems. ACM, New York, pp 1–4