Query Recommendation Systems Based on the Exploration of OLAP and SOLAP Data Cubes

Olfa Layouni[1(✉)], Assawer Zekri[1], Marwa Massaâbi[1], and Jalel Akaichi[2]

[1] BESTMOD Laboratory, Institut Supérieur de Gestion de Tunis,
Université de Tunis, Tunis, Tunisia
layouni.olfa89@gmail.com, assawer.zekri@gmail.com, massaabi.marwa@gmail.com
[2] College of Computer Science, King Khalid University, Abha, Saudi Arabia
jalel.akaichi@kku.edu.sa

Abstract. Business Intelligence systems refer to technologies and tools responsible for collecting, storing and analyzing data to improve decision-making. In BI systems, users interact with data warehouse by formulating and launching sequences of queries aimed at exploring multidimensional data cubes. However, the volumes of data stored in a data warehouse can be very large and diversified. So, a big amount of irrelevant information returned as results to the user could make the data exploration process inefficient. That's why, it's necessary to help the user by guiding him in his exploration. In fact, query recommendation systems play a major role in reducing the effort of decision-makers to find the most interesting information. Several works dealing with query recommendation systems were presented in the last few years. This paper aims at providing a comprehensive review of literature on a query recommendation based on the exploration of data cubes. A benchmarking study of query recommendation methods is proposed. Several evaluation criteria are used to identify the existence of new investigations and future researches.

Keywords: Query recommendation systems · Business intelligence systems · Data · Analysis · Cube · Data warehouse

1 Introduction

Business Intelligence (BI) system represents the tools that are used to collect, store and analyze data in order to make the best decision [4–6]. The BI system is realized by applying two different steps. The first step is the Extract, Transform and Load data. The ETL tools are responsible for extracting data from different heterogeneous sources, providing the integration and data cleansing according to a target schema or data structure, loading and storing data in a data warehouse. The second step is to analyze data by using an analysis server such as: OLAP or Spatial OLAP server. It is a rapid and flexible way for analysts to navigate, explore and analyze the large amount of data stored in

G. De Pietro et al. (eds.), *Intelligent Interactive Multimedia Systems and Services 2017*,
Smart Innovation, Systems and Technologies 76, DOI 10.1007/978-3-319-59480-4_33

the data warehouse. Indeed, the user can make analysis reports by using: some reporting tools, dashboards, navigation and statistical tools. These tools offer capabilities to explore data and support the analysis process. To analyze data, users interactively navigate a data cube by launching sequences of OLAP or SOLAP queries over a traditional or spatial data warehouse, respectively. The problem appeared when the user may have no idea of what the forthcoming query should be. As a solution and to help the user in his navigation, we need a recommendation system.

The remainder of this paper is organized as follows: Sect. 2 introduces the concepts of recommendation in data warehouse systems. Section 3 presents an overview of several different approaches presented in the field of query recommendation based on the exploration of data cube. Section 4 presents a comparative study that provides a general, comparative view of the different approaches that have been presented. Section 5 concludes the paper.

2 Recommendation System

Recommendation system defined as a system that gives the possibility to generate recommendations of items like books, movies, music, queries, etc.; and products that might interest users [9,16]; those recommendations give the possibility to help the user by guiding him to find relevant information. The current generation of recommendation system is usually categorized into a content-based method, a collaborative method and a hybrid method [1,13,17].

In various studies [17,21–23], we find that the authors described the characteristics of the general algorithm of a recommender system for the exploration of data. These characteristics are the inputs, the outputs and the recommendation steps.

The inputs of the algorithm can be a log of sessions of queries, a schema or an instance of the relational or multidimensional database, a current session that contains the queries launched by the current user and a profile or the behavior of the user.

The outputs of the algorithm can be a query, a set of ordered queries or a set of tuples that can be similar or interested in the current user.

In reality, an algorithm of recommendation is decomposed into three steps. The first step consists in choosing an approach for evaluating the used scores. In fact, in this step we can choose one of the categories of recommendations: a content-based, a collaborative and a hybrid method. The second step is the filter; this step consists in selecting the candidates' recommendations. The last step is the guide; this step consists in ordering the candidates' recommendations.

3 OLAP Query Recommendation Approaches

This section presents a thorough survey on the proposed approaches in the domain of query recommendation for helping users to explore data. Those approaches can be classified into two categories, the first category exploits the

OLAP data cube and so does the second category which exploits the Spatial OLAP data cube.

3.1 Methods Exploiting OLAP Data Cube

Data warehouse stores large volumes of consolidation and historized multidimensional data, to be explored and analyzed by various users. In fact, the user interacts with the data warehouse by launching sequences of OLAP queries aimed at exploring the multidimensional data cube. Since the volume of information to explore can be very huge and diversified, it is necessary to help the user to face this problem by guiding him: by proposing an OLAP query recommendation system in his data cube. In the literature, we can distinguish two different ways to explore OLAP data cube. The first way exploits the profile and so does the second way with the log of queries.

Using the Profile. A lot of researches recommend OLAP queries in the exploration of a data warehouse by exploiting the profile.

The works of Sarawagi et al. in [21–23] were proposed to help the user in his exploration of the OLAP data cube based on the atomization. For this reason, the authors proposed four different operators: DIFF, EXCEP, RELAX and INFORM. Those operators allow the return as results of all sets of tuples that can explain the anomalies detected through the different operators, in fact those operators give the possibility to recommend one or more queries. We find that the proposed algorithms are a recommendation method based on the content. Adding to that, we discover that some proposed operators execute the results obtained after launching the current query and other operators execute only the current query.

The recommendation method proposed by Bellatreche et al. in [7,8] treats the problem of OLAP query personalization. In this method, the authors took into account the particularities of OLAP queries. The proposed method gives the possibility to secure two principal objectives; the first is to compute the utility of query and the second is to display the best query to a user by taken into account not only his preferences but also his visualization constraints. The proposed method is composed in three different steps. The first step consists in using the user profile. In this step, authors proposed two functions Perso and MaxSubset which were used to compute the best subsets of references from the current query and to satisfy the proposed constraint. The second step consists in searching elements firstly by comparing between the stored preferences and a query; secondly by selecting an order set of references for a specific query. The third step, the authors proposed to build a personalization query by using the best references, sorted and recommended them in an ascending order. We deduce that this approach is based on the content method. In fact, this method doesn't take into consideration the previous queries launched in the cube and the sequencing of queries launched by the current user.

The recommendation method proposed by Jerbi in [11] treats the problem of OLAP analysis personalization within data warehouses. In fact, the user must launch several queries in order to obtain a result that can be similar or close to his preferences. The proposed method gives the possibility to improve the current query by using the preferences of the user, and recommends the best query for him. The first step in this method is to analyze OLAP data. An OLAP analysis is modeled through a graph where nodes represent the analysis contexts and edges represent the user operations. The second step is to build a model for user preferences on the multidimensional schema and values. The last step is to propose a framework including two personalization processes. The first process denoting OLAP query personalization depends on contextual preferences stored in a user profile. In this process, two phases must be performed: the selection and the integration of preferences. The second process is recommendation queries. In this process, two types of personalization have to be performed: a personalization of an explicit or a dynamic type. Moreover, the proposed recommendation framework supports recommendation scenarios: assisting the user in a query composition and suggesting the forthcoming and alternative queries. The system recommends a set of queries by comparing the user preferences and alternative queries. Consequently, this framework is based on the content method. Also, this method doesn't take into consideration the sequencing of queries launched by the current user; it takes only the last launched query and the current session.

Using the Log of Queries. A lot of studies recommend queries in the exploration of OLAP data cube by exploiting the log of queries.

The recommendation method proposed by Sapia in [18–20] handles the problem that a user has no idea about the forthcoming query to request. Therefore, the author proposed a method to predict the next OLAP query in order to help him during the rest of the current session. The author proposed a method based on a probabilistic model: the Markov model. This model is used in order to return the similarity between two consecutive queries and predict the probability of occurrence for each prototype. In this method, the author proposed to build for each query in the log of sessions and in the current session a prototype corresponding to it. Then, he suggests to use a distance method to compare between those prototypes. Finally, he proposes to recommend a query to the current user which contains the highest probability of appearance between the prototype and the launched query. We note that this method is a collaborative recommendation method, which uses a statistical model the Markov model. In addition, we remark that this method takes into account the sequencing of queries and the previous queries.

The method proposed by Giacometti et al. in [10] gives the possibility to recommend for the current user the discoveries detected in the previous sessions saved in the log with the same unexpected data as the current session. This approach consists in two different steps. The first step is based on analyzing the query log to discover pairs of cells at various levels of detail for which the measure values differ significantly. The second step is based on analyzing a current query to detect if a particular pair of cells for which the measure values differ significantly can be related to what is discovered in the log. This approach is composed of two parts: the processing of the log and the computation of the recommendations. We find that this method is based on the results and the queries for each session in the log of recommending queries. Besides, it is a collaborative method.

The method proposed by Marcel et al. in [14] gives the possibility to recommend query to the user. The authors proposed a method for computing an intensional answer to an OLAP query by using the previous queries in the current session launched by the current user. The proposed method takes into account the intensional query answers. For this reason, it satisfied the three criteria proposed for the intensional answer: purity, completeness and dependency. So, we find that this method can be classified as mixed, partial and dependent. We describe the proposed method as following. The first step is to compute the extensional answer after the current user launching a query over the cube OLAP. In this method, authors proposed to use an expected cube for storing a model of the user expected values according to the data saved in the extensional answer. In the second step, for each query belonging to the current session, four steps must be done: execute the query over the cube, predict the expected extensional answer, improve the quality of the estimated values in the expected cube and build an intensional answer. We find that this method was suggested for generating recommendations OLAP queries in the context of the collaborative exploration of data cubes and it is based on the extensional and intensional answers. Furthermore, this method doesn't take into consideration the spatial queries and it is based only on the current session.

The method proposed by Aufaure et al. in [3] treats the problem of finding big numbers of possibilities of aggregations and selections that can be operated, all those possibilities may make the user experience disorientating and frustrating. In fact, authors proposed an approach for predicting and recommending the most likely next query to the current user. The proposed approach used a probabilistic user behavior model, which can be built by analyzing previous OLAP sessions and exploiting a query similarity metric. To this end, the authors proposed to analyze the query logs of a user, then, to cluster queries by using both a similarity metric and a Markov-based model based on the user behavior. Finally, by using this model, it is possible to recommend the next query for the current user. We find that this method doesn't only take into consideration sessions in the log and the sequencing of queries but also it takes into consideration queries launched by the current user in the past and the current query. This method is content-based method.

3.2 Methods Exploiting Spatial OLAP Data Cube

By using the new technologies such as: PDA, GPS, RFID, etc., we store different types of data. 80% of the obtained data represent spatial or location components. Spatial data warehouse have been used for storing and manipulating spatial data components [15]. Several spatial data types such as: point, surface, multipolygon, etc.; can be used to represent the spatial extent of real-world objects, which are collected for using the new technologies, in order to store the location or the movement of a real-world object like: car, train, ambulance, etc.

Spatial data types have a set of operations, which can be realized in order to represent spatial characteristics such as: topological, direction and metric distance [15]. A spatial data warehouse is made of multidimensional spatial model by integrating spatial measures and dimensions in order to take into account spatial components. In order to analyze and explore a spatial data warehouse, users need a SOLAP system. The SOLAP system obtained after the combination of Geographic Information Systems (GIS) with OLAP tools and operations. To navigate in the spatial data cube the current user launches a sequence of SOLAP queries over a spatial data warehouse. The problem appeared when the current user may have no idea of what the forthcoming SOLAP queries should be for this purpose we need a SOLAP queries recommendation system. In the literature, we find two methods for recommending SOLAP queries, the first one is proposed by [12,13] and the second one is proposed by [2].

The method proposed by Layouni et al. in [12,13] gives the possibility to recommend SOLAP queries to the current user. The proposed approach consists of the three following steps. The first step consists in computing all the generalized sessions of SOLAP queries of the log. In this step, they proposed a new similarity measure in order to compare between SOLAP queries by taking into account spatial relationships: topological, direction and metric. Also, they propose to use the method of TF-IDF for extracting the spatial measures and spatial dimensions in a launched query. Besides, to do the last classification of SOLAP queries, they decide to choose the Hierarchical Ascendant Classification. The second step is the filter which consists in predicting the candidates SOLAP queries by computing the most similar sessions to the generalized current sessions and searching the set of candidates SOLAP queries. The last step is the guide that consists in ordering the candidates SOLAP queries. We find that this method was suggested for generating recommendations SOLAP queries in the context of the collaborative exploration of spatial data cubes. And it is based on the text query and not the results obtained.

The method proposed by Aissi et al. in [2] recommend a set of SOLAP queries for the current user, this set contains only five queries. This approach takes into account the specific characteristics of spatial data. But the proposed approach for recommending SOLAP queries have some disadvantages. The proposed algorithm eliminates all the old queries in the log. It takes into account only recent queries in the log. In order to recommend a set of queries, the proposed approach detects preferences of the current user and compares between queries by applying a spatio-semantic similarity measure.

4 Comparing Query Recommendation Approaches and Discussion

The following section presents a comparative study that provides a general comparative view of the different approaches that have been presented and discussed in the field of methods proposed for recommending queries. The different models are compared according to these criteria.

Objectives of the works: The works proposed by [7,8,11] have two objectives: the personalization and the recommendation queries in the exploration of the data warehouse but the works proposed by [3,10,12–14,18–23] have as objective the recommendation queries.

Recommendation SOLAP vs non-SOLAP queries: The works proposed by [3,7,8,10,11,14,18–23] recommend queries in the exploration of the data warehouse so they exploit an OLAP data cube. The work proposed by [12,13] recommend queries in the exploration of the spatial data warehouse so the authors in this method explore a SOLAP data cube.

Inputs of the algorithm of recommendation: The works proposed by [7, 8,11,21–23] exploit the profile, but the works proposed by [2,3,10,12–14,18–20] exploit the log. We find that the methods proposed by [7,8,11,21–23] take as inputs of the algorithm the profile of the current user. Also, the methods proposed by [3,10,12–14,18–20] take as inputs of the algorithm the log of sessions of queries. Indeed, we find that the inputs of the algorithm can be: a schema, an instance, the current query, the current session, the previous sessions and visualization constraints. In fact, we remark that a schema was used as input in the algorithms proposed by [10–14,18–23]; an instance was used as input in the algorithms proposed by [7,8,10–14,18–23]; the current query was used as input in the algorithms proposed by [2,3,7,8,10–14,18–23]; the current session was used as input in the algorithms proposed by [2,3,10–14,18–20]; the previous sessions were used as input in the algorithms proposed by [2,3,10,12,13,18–20] and visualization constraints were used as input only in the algorithm proposed by [7,8].

Output of the algorithm of recommendation: Comparing the different proposed methods, we find that the output of the methods proposed by [3,11–14,18–20] is a query, those proposed by [2,7,8,10,12–14] the output is a set of queries and only the method proposed by [21–23] the output is a set of tuples.

Category of recommendation proposed system: We find that the methods proposed by [3,7,8,11,21–23] are content-based methods and the methods proposed by [2,10,12–14,18–20] are collaborative methods.

Filter step: We find that, in the filter step of the proposed algorithm, the method proposed by [2,3,7,8,11–14,21–23] gives the possibility to select the candidate's recommendation queries. Besides, for computing the candidate's recommendations the methods proposed by [14,21–23] apply the maximum entropy theory; also the methods proposed by [7,8,10,11] use a graphic model; as well the methods proposed by [3,18–20] apply the Markov model and the methods proposed by [2,12–14] apply distance model.

Guiding step: The guiding step was applied in the methods proposed by [11–13,18–20].

Intension and extension approach: We also find that the methods proposed by [2,3,7,8,11–14,18–20] applied to queries; only the method proposed by [10, 14,21–23] applied to the results obtained after launching a query.

Table 1. Comparative studies between the different approaches.

Proposed method		Proposed by								
		[21–23]	[7,8]	[11]	[18–20]	[10]	[14]	[3]	[12,13]	[2]
Objective	Recommending queries	*	*	*	*	*	*	*	*	*
	Personalization query		*	*						
Data warehouse		*	*	*	*	*	*	*	*	
Spatial data warehouse									*	*
Cube	OLAP	*	*	*	*	*	*	*		
	SOLAP								*	*
Inputs of the algorithm	Log				*	*	*	*	*	*
	Profile	*	*	*						
	Schema	*		*	*	*	*		*	
	Instance	*	*	*	*	*	*		*	
	Current query	*	*	*	*	*	*	*	*	*
	Current session			*	*	*	*	*	*	*
	Visualization constraints		*							
	Previous sessions				*	*		*	*	*
Output of the algorithm	Query			*	*		*	*	*	
	Set of queries		*			*			*	*
	Set of tuples	*								
Intension approach			*	*	*		*	*	*	*
Extension approach		*				*	*			
Approach	Content-based method	*	*	*				*	*	
	Collaborative method				*	*	*		*	*
Filter	Select candidates recommendations	*	*	*			*	*	*	*
	Compute candidates recommendations	The maximum entropy theory	A graphic model	A graphic model	A Markov model	A graphic model	The maximum entropy theory and a distance model	A Markov model	A distance model	A distance model
Guide				*	*				*	
Manipulation language	SQL	*		*	*	*	*	*		
	MDX		*			*	*	*	*	*
	Spatial MDX								*	*

Manipulation language: We find that the methods proposed by [3,10,11,14, 18–23] used the SQL language, the methods proposed by [3,7,8,10–14] used the MDX language and the methods proposed by [2,12,13] used the MDX language with Spatial functions. Table 1 reports a comparison of the above approaches according to the presented criteria.

5 Conclusion

In this paper, we have done an overview of the developed and suggested query recommendation approaches. Each approach is presented and discussed, then, a comparative study between the different proposed works is presented in order to compare and evaluate them in terms of some criteria.The relative novelty of the domain leaves many challenges, opportunities and extended studies open for future work, which we addressed most of them in our deduced research gaps. The proposed work allows us to have a overall vision on the different proposals and takes advantage of the studied contributions in an optimized way in order to introduce our future work which is the proposal of a new approach on a trajectory query recommendation to describe the object movement in space over the time.

References

1. Adomavicius, G., Tuzhilin, A.: 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 (2005)
2. Aissi, S., Gouider, M.S., Sboui, T., Said, L.B.: Enhancing spatial data warehouse exploitation: a solap recommendation approach. In: Computer and Information Science, pp. 131–147. Springer (2016)
3. Aufaure, M., Kuchmann-Beauger, N., Marcel, P., Rizzi, S., Vanrompay, Y.: Predicting your next OLAP query based on recent analytical sessions. In: Proceedings Data Ware-housing and Knowledge Discovery - 15th International Conference, DaWaK 2013, Prague, Czech Republic, 26–29 August, pp. 134–145 (2013)
4. Badard, T.: L'open source au service du géospatial et de l'intelligence d'affaires. Geomatics Sciences Department (avril 2011)
5. Badard, T., Dubé, E.: Enabling geospatial business intelligence. Geomatics Sciences Department, Semptember 2009
6. Bédard, Y., Han, J.: Geographic Data Mining and Knowledge Discovery, 2e edn. Taylor & Francis, Boca Raton (2009)
7. Bellatreche, L., Giacometti, A., Marcel, P., Mouloudi, H., Laurent, D.: A personalization framework for OLAP queries. In: Proceedings DOLAP 2005, ACM 8th International Workshop on Data Warehousing and OLAP, Bremen, Germany, 4–5 November, pp. 9–18 (2005)
8. Bellatreche, L., Mouloudi, H., Giacometti, A., Marcel, P.: Personalization of MDX queries. In: 22èmes Journées Bases de Données Avancées, BDA 2006, Lille, 17–20 octobre 2006, Actes (Informal Proceedings) (2006)
9. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**(4), 331–370 (2002)
10. Giacometti, A., Marcel, P., Negre, E., Soulet, A.: Query recommendations for OLAP discovery-driven analysis. IJDWM **7**(2), 1–25 (2011)
11. Jerbi, H.: Personnalisation d'analyses décisionnelles sur des données multidimensionnelles. Ph.D. thesis, Institut de Recherche en Informatique de Toulouse - UMR 5505, France (2012)
12. Layouni, O., Akaichi, J.: A novel approach for a collaborative exploration of a spatial data cube. IJCCE Int. J. Comput. Commun. Eng. **3**(1), 63–68 (2014)

13. Layouni, O., Alahmari, F., Akaichi, J.: Recommending multidimensional spatial olap queries. In: Intelligent Interactive Multimedia Systems and Services 2016, pp. 405–415. Springer (2016)
14. Marcel, P., Missaoui, R., Rizzi, S.: Towards intensional answers to OLAP queries for analytical sessions. In: Proceedings DOLAP 2012, ACM 15th International Workshop on Data Warehousing and OLAP, Maui, HI, USA, November 2, pp. 49–56 (2012)
15. Marketos, G.: Data Warehousing & Mining Techniques for Moving Object Databases. Ph.D. thesis, Department of Informatics, University of Piraeus (2009)
16. Melville, P., Sindhwani, P.: Recommender systems. In: Encyclopedia of Machine Learning, pp. 829–838 (2010)
17. Negre, E.: Exploration collaborative de cubes de données. Ph.D. thesis, Université François Rabelais of Tours, France (2009)
18. Sapia, C.: On modeling and predicting query behavior in olap systems. In: Proceedings INT'L Workshop on Design and Management of Data Warehouses (DMDW 99), SWISS LIFE, pp. 1–10 (1999)
19. Sapia, C.: PROMISE: Predicting query behavior to enable predictive caching strategies for OLAP systems. In: Kambayashi, Y., Mohania, M., Tjoa, A (eds.) Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science, vol. 1874, pp. 224–233. Springer, Heidelberg (2000)
20. Sapia, C., Alexander, F., Erlangen-nürnberg, U.: Promise: modeling and predicting user behavior for online analytical processing applications. Ph.D. thesis submitted, Technische Universität München (2001)
21. Sarawagi, S.: Explaining differences in multidimensional aggregates. In: Proceedings of the 25th International Conference on Very Large Data Bases, VLDB 1999, pp. 42–53. Morgan Kaufmann Publishers Inc., San Francisco (1999)
22. Sarawagi, S.: User-adaptive exploration of multidimensional data. In: VLDB, pp. 307–316. Morgan Kaufmann (2000)
23. Sathe, G., Sarawagi, S.: Intelligent rollups in multidimensional olap data. In: Proceedings of the 27th International Conference on Very Large Data Bases, VLDB 2001, pp. 531–540. Morgan Kaufmann Publishers Inc., San Francisco (2001)