Implementing and Monitoring Quality Preferences in Data by using Spatial Ranking Method

[1]Sajja Kiran Kumar, [2]Kunapareddy Rajani Devi
[1,2]Dept. of IT, Nalanda Institute of Engineering & Technology, Siddharth Nagar, Sattenapalli, Guntur, Affiliated to JNTUK, Kakinada, AP, India

Abstract
In order to handle spatial data efficiently, as required in computer aided design and geo-data applications, a database system needs an index mechanism that it help to retrieve data items quickly according to their spatial locations However, traditional indexing methods are not well suited to data objects of non-zero size located in multidimensional spaces In this paper we describe a dynamic index structure called an R-tree which meets this need, and give algorithms for searching and updating it. We present the results of a series of tests which indicate that the structure performs well, and conclude that it is useful for current database systems spatial applications

Keywords
Query Processing, Spatial Databases

I. Introduction

With the popularization of geo tagging information, there has been an increasing number of Web information systems specialized in providing interesting results through location-based queries. However, most of the existing systems are limited to plain spatial queries that return the objects present in a given region of the space. In this paper, we study a more sophisticated query that returns the best spatial objects based on the features (facilities) in their spatial neighborhood. Given a set of data objects of interest, a top-k spatial preference query returns a ranked set of the k best data objects.

The score of a data object is defined based on the non-spatial score (quality) of feature objects in its spatial neighborhood. On the other hand, the score of a feature object does not depend on its spatial location, but on the quality of the feature object. Such quality values can be obtained by a rating

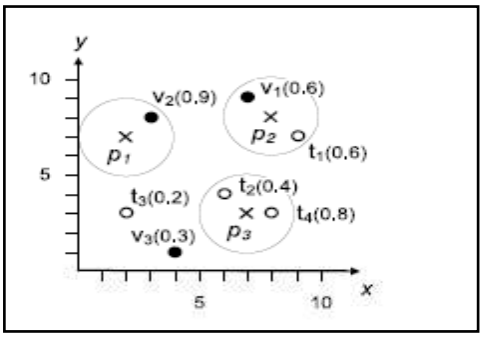

Fig. 1: Spatial Area Containing Data and Feature Objects

For example, fig. 1, presents a spatial area containing data objects p (hotels) together with feature objects t (restaurants) and v (caf´es) with their respective scores (e.g. rating). Consider a tourist interested in hotels with good restaurants and caf´es in their spatial neighborhood. The tourist specifies a spatial constraint (in the figure depicted as a range around each hotel) to restrict the distance of the eligible feature objects for each hotel. Thus, if the tourist wants to rank the hotels based on the score of restaurants, the top-1 hotel is p3(0.8) whose score 0.8 is determined by t4. However, if the tourist wants to rank the hotels based on caf´es, the top-1 hotel is p1(0.9) determined by v2. Finally, if the tourist is interested in both restaurants and caf´es (e.g. summing the scores), the top-1 hotel is p2(1.2). Top-k spatial preference queries are intuitive and comprise a useful tool for novel location-based applications. Unfortunately, processing top-k spatial preference queries is complex, because it may require searching the spatial neighborhood of all data objects before reporting the top-k. Due to this complexity, existing solutions are costly in terms of both I/Os and execution time. In this paper, we propose a novel approach for processing spatial preference queries efficiently. The main difference compared to traditional top-k queries is that the score of each data object is defined by the feature objects that satisfy a spatial constraint (for example range constraint). Therefore, pairs of data and feature objects need to be examined to determine the score of an object. Our approach relies on mapping of pairs of data and feature objects to a distance-score space, which in turn allows us to identify the minimal subset of pairs that is sufficient to answer all spatial preference queries. Capitalizing on the materialization of this subset of pairs, we present an efficient algorithm that improves query processing performance by avoiding examining the spatial neighborhood of data objects during query execution.

In addition, we propose an efficient algorithm for materialization and describe useful properties that reduce the cost of maintaining the materialized information. In summary, the main contributions of this paper are:

- We define a mapping of pairs of data and feature objects to the distance-score space that enables pruning of feature objects that do not affect the score of any data object.
- We prove that there exists a minimal subset of pairs that is sufficient to answer all top-k spatial preference queries.
- We propose an efficient algorithm for processing top-k spatial preference queries that exploits the materialized subset of points.
- In addition, we propose an effective algorithm for materialization, and we identify useful properties for cost-efficient maintenance of the materialized information.
- We show through an extensive experimental evaluation that our algorithm outperforms the state-of-the-art algorithms in terms of both I/Os and execution time.

II. Related Work

There are several representations and manipulation of vague regions, one of the most representatives is the Fuzzy Minimum Boundary Rectangle (FMBR)[4], that includes use of Fuzzy Logic to define degrees of memberships according to a membership function. In many geographical applications there is a need to model spatial phenomena not simply by sharp objects but rather through indeterminate or vague concepts. FMBR have been used to model such geographical data; it is considered an adequate tool to represent problems related with vague regions. FMBR is composed by two regions, the first region, called kernel, describes which part of the vague region belongs to it. The second region called boundary describes the fuzzy area of a vague region.

The Fuzzy Data Cube (FDC) is a fuzzy multidimensional structure to query data in a Data warehouse using OLAP tools. It was initially defined for the sales problem, but in a Different context, it is built by evaluating a membership function for each attributes stored in the data warehouse. The result is a degree of membership that is stored in the FDC. Part of this paper consists on to extend the definition of FDC considering a Spatial Database. Working with the semantic of the attributes in a data cube is another approach that has not been widely considered. The representation of this model helps to draw conclusions with a higher degree of uncertainty [6], based on the classification performance in order to assist the decision support tasks, which has been an application of data cubes.

A spatial database system (SDBS) is a database system offering spatial data types in its data model and query language and offering an efficient implementation of these data-types with their operations and queries. Typical operations on these data-types are the calculation of the distance or the intersection. Important query types are similarity queries, e.g.:

- Region queries, obtaining all objects within a specified query region and
- K-nearest neighbor (kNN) queries, obtaining the k objects closest to a specified query object.

Similarity queries are important building blocks for many spatial data mining algorithms - especially for our approach to 'generalized clustering'. Therefore, the underlying SDBS technology, i.e. spatial index structures, to support similarity queries efficiently, is sketched briefly in the following. Spatial index structures can be roughly classified as organizing the data space (Hashing) or organizing the data itself (search trees). In the following, we will introduce well-known representatives which are typical for a certain class of index Structures.

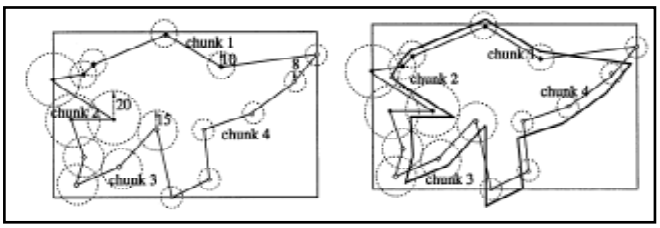

Fig. 2(a): A Polygon with four Chunks, (b). A Possible Realization of the Uncertain Polygon

III. Measuring the Quality Preferences

The treatment of spatial objects with indeterminate boundaries is especially problematic for the computer scientist who is confronted with the difficulties how to model such objects in a database system, so that they correspond to the user's intuition, how to finitely represent them in a computer format, how to develop spatial index structures for them, and how to draw them. Computer Scientists are accustomed to the abstraction process of simplifying spatial phenomena of the real world through the concepts of conventional binary logic, reduction of dimension, and cartographic generalization to precisely defined, simply structured, and sharply bounded objects of Euclidean geometry like points, lines, and regions. To define the Fuzzy Spatial Data Warehouse we begin with the dimensionality of the warehouse, based on the study case known, as risk areas nearby volcano. To handle the information we use a Geographical Information System that contains geo referenced data from maps of the State of Pueblo. In Figure , it is shown the Snow Flake Schema [7] to be used in the Fuzzy Spatial Data Warehouse, as an extension of the multidimensional model shown in Figure 2. Notice how each of the dimensions are laid out. The main reason behind to select the Snow Flake Schema, is that the fact table is a table that uses the main table to relate with other fact sub tables. In such way, the data warehouse has a tree like representation, where the root represents the principal fact table, and every node at the first level represents dimension and the remainder nodes with a level greater than 1 are called sub dimensional nodes. Consider that a geo referenced data follows a recursive definition [11] of spatial concept, thus, it can be used to define a geographic concept. That is a geographic concept is either (i) a geographic data element, or it is (ii) a set of geographic concepts. Analogously, the physical manifestation of a geographic concept, namely a geographic object, is also expressible recursively as either a geographic element, or as a set of geographic objects, this definition is very important because that gives us the opportunity to work with star schema without ambiguity [7].

IV. Conclusion

This work represents a part of the Intelligent Geographical Project that will model a geographic area in the same way as in the real world appears, taking advantage of the 66 Decision Support Systems, Advances in Information Technology. We have integrated ArcGIS technology with Fuzzy Theory, given the fact that linguistic variables get closer to coloquial language and they describe a geographic situation in a natural way. The main contribution of this work is the integration of Fuzzy Logic with Spatial Databases in order to help during the decision support and OLAPs querying processes. ArcGIS allows to obtain spatial features by the queries execution on maps. These spatial features are integrated into a multidimensional database allowing aggregation and disaggregating OLAP operations on it, avoiding the use of classic spatial access methods. In addition, spatial semantic is added to spatial and not spatial dimensions of the multidimensional model improving the decision making process. Finally, the Fuzzy Spatial Data Warehouse's design methodology proposed, simplify the use of the existing analysis tools for exploiting the potential of Data Warehouses.

References

[1] M. L. Yiu, X. Dai, N. Mamoulis, M. Vaitis,"Top-k Spatial Preference Queries", in ICDE, 2007.
[2] N. Bruno, L. Gravano, A. Marian,"Evaluating Top-k Queries over Web-accessible Databases", in ICDE, 2002.
[3] A. Guttman,"R-Trees: A Dynamic Index Structure for Spatial Searching", in SIGMOD, 1984.
[4] G. R. Hjaltason, H. Samet,"Distance Browsing in Spatial Databases", TODS, Vol. 24(2), pp. 265–318, 1999.
[5] R. Weber, H.-J. Schek, S. Blott,"A quantitative analysis and performance study for similarity-search methods in highdimensional spaces", in VLDB, 1998.
[6] K. S. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, "When is "nearest neighbor" meaningful?" in ICDT, 1999.
[7] R. Fagin, A. Lotem, M. Naor,"Optimal Aggregation Algorithms for Middleware", in PODS, 2001.
[8] I. F. Ilyas, W. G. Aref, A. Elmagarmid,"Supporting Top-k Join Queries in Relational Databases", in VLDB, 2003.
[9] N. Mamoulis, M. L. Yiu, K. H. Cheng, D. W. Cheung, "Efficient Top-k Aggregation of Ranked Inputs", ACM TODS, Vol. 32, No. 3, p. 19, 2007.

[10] D. Papadias, P. Kalnis, J. Zhang, Y. Tao,"Efficient OLAP Operations in Spatial Data Warehouses", in SSTD, 2001.
[11] S. Hong, B. Moon, S. Lee,"Efficient Execution of Range Topk Queries in Aggregate R-Trees", IEICE Transactions, Vol. 88-D, No. 11, pp. 2544–2554, 2005.
[12] T. Xia, D. Zhang, E. Kanoulas, Y. Du,"On Computing Top-t Most Influential Spatial Sites", in VLDB, 2005.
[13] Y. Du, D. Zhang, T. Xia,"The Optimal-Location Query", in SSTD, 2005.

Sajja Kiran Kumar, Pursuing M. Tech(IT), Nalanda Institute of Engineering & Technology, Siddharth Nagar, Sattenapalli, Guntur,Affiliated to JNTUK, Kakinada, A.P., India

Kunapareddy Rajani Devi, M.Tech working as Associate Professor in the dept of Information Technology, Nalanda Institute of Engineering & Technology, Siddharth Nagar, Sattenapalli, Guntur,Affiliated to JNTUK, Kakinada, A.P., India