Spatial data retrieval using both quality and distance preference using top-k algorithm

S.PARTHIBAN
M.E (CSE) Second year
JJCET, Trichy
Trichirapalli, India
Parthi.parthi88@gmail.com

MR. LAKSHMI NARASIMAN M.E.
Associate Professor
JJCET, Trichy
Trichirapalli, India
lakshminar@gmail.com

ABSTRACT

Elaborates on the way quality should be taken into account in the development of Spatial Data Infrastructures (SDIs). A variety of quality concepts are described using four quality management viewpoints, i.e. a production-centered, planning-centered, customer-centered and a system-centered perspective. New ideas are introduced and discussed on how quality should be managed in the national SDI programmes. Most of the efforts concentrate on solving interoperability issues at data or system level, ignoring organizational issues. Spatial preference query ranks objects based on the qualities of features in their spatial neighborhood. Formally define spatial preference queries and propose appropriate indexing techniques and search algorithms for the spatial Dataset. In this paper, we study an interesting type of preference queries, which select the best spatial location with respect to the quality of facilities in its spatial neighborhood. The analysis and quantification of species/environment relationships is a key stone in predictive geographical modeling in ecology.

Keywords

Spatial data.

1. INTRODUCTION

Branch bound algorithm is proposed to handle the spatial data corresponding to the user's input query request. Feature join is Participate to deal the top-k spatial preference query efficiently. Influence sore can achieve with the proposed functions of MIN and MAX. The score of an object is defined by the quality of features (e.g., facilities or services) in its spatial neighborhood. To rank the contents of this database with respect to the quality of their locations, quantified by aggregating non-spatial characteristics of other features (e.g., restaurants, cafes, hospital, market, etc.) in the spatial neighborhood of the flat (defined by a spatial range around it).

Quality may be subjective and query-parametric. The locations of an object data set D (hotels) we used for our processing. Feature points are labeled by quality values that can be obtained from rating providers. The semantics of the aggregate function is relevant to the user's query. The SUM function attempts to balance the overall qualities of all features. It ensures that the top result has reasonably high qualities in all features.

For the MAX function, the top result obtained. It is used to optimize the quality in a particular feature, but not necessarily all of them. Our Ranking objects are spatial Ranking, Non spatial Ranking. Spatial ranking, which orders the objects according to their distance from a reference point, and Non-spatial ranking, which orders the objects by an aggregate function on their nonspatial values. Our top-k spatial preference query integrates these two types of ranking in an intuitive way.

A Brute Force approach (to be elaborated in Section 3.2) for evaluating it is to compute the scores of all objects in D and select the top-k ones. We propose alternative techniques that aim at minimizing the I/O accesses to the object and feature data sets, while being also computationally efficient. Our techniques apply on spatial-partitioning access methods and compute upper score bounds for the objects indexed by them, which are used to effectively prune the search space. The branch-and-bound (BB) algorithm and the feature join (FJ) algorithm for efficiently processing the top-k spatial preference query.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons.Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate. Data mining involves six common classes of asks: http://en.wikipedia.org/wiki/Data_miningcite_note-

Fayyad. Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors and require further investigation.

1.1 Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing data-driven inductive approaches to geographical analysis and modeling. Data mining offers great potential benefits for GIS-based applied decision-making. Recently, the task of integrating these two technologies has become of critical importance, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information contained therein. Among those organizations are: Offices requiring analysis or dissemination of geo-referenced statistical data. Public health services searching for explanations of disease clustering. Environmental agencies assessing the impact of changing land-use patterns on climate change. Geo-marketing companies doing customer segmentation based on spatial location.

1.2 Non-spatial data

It consists of attributes that are complementary to the spatial data, and describes what is at a point along a line or in a polygon. The attribute usually represents the properties or characteristics of the spatial data which may include socio-economic characteristics from census or from other sources.
For example, the attributes of a soil category could be the depth of soil, texture, erosion, drainage, etc, and for a geological category, they could be the rock type, its age, major composition, etc. the socio-economic characteristics could be the demographic and occupation data for a village or traffic volume data for roads in a city.

The non-spatial data is mainly available in tabular records in analog from and needs to be converted into digital format for incorporation in GIS. The 1991 cense data of the country is now available in digital mode and thus its direct incorporation in a GIS database is possible. Maps or photographic data (spatial or non-spatial) can be fed to the GIS by converting it into a digital form, using any of the following devices.

1.3 Spatial database

Non spatial ranking can, which orders the objects by an aggregate function on their non spatial values. It consists of attributes that are complementary to the spatial data, and describes what is at a point of particular location. The attribute usually represents the properties or characteristics of the spatial data which may include socio-economic characteristics from census or from other sources. A spatial database is a database that is optimized to store and query data that is related to objects space, including points, lines and polygons.

Figure: 1.1 Spatial Database Management and Advanced Geographic Information Systems

While typical databases can understand various numeric and character types of data, additional functionality needs to be added for databases to process spatial data types. These are typically called geometry or feature. The Open Geospatial Consortium created the Simple Features specification and sets standards for adding spatial functionality to database systems.

1.4 Query processing

Query processing and optimization is a fundamental, if not critical, part of any DBMS. To be utilized effectively, the results of queries must be available in the timeframe needed by the submitting user—be it a person, robotic assembly machine or even another distinct and separate DBMS. The query processor turns user queries and data modification commands into a query plan - a sequence of operations (or algorithm) on the database from high level queries to low level commands.

1.5 Distance calculation

The distance between the reference point and the each object going to be calculated for every process of input query. Then the minimum distance between the reference point and the one object will be calculated. According to that distance the rank to the object will be provided, the object which holds the minimum distance to the reference point ranked as one.

1.6 Aggregation function

Spatial ranking orders the object according to the distance between the reference point. Non spatial ranking orders the object according to the features. The combination of both spatial and non spatial ranking gives the final rank for the object.

1.7 Quality sorting

From the large dataset we have to sort the data's based on the quality at a particular location.Defining the geographic entities in the spatial database systems. Describing the objects weather which based the quality sorted. Quality may be subjective or parametic. A customer may want to rank the contents of this database with respect to the quality of their locations, qualified by aggregating non spatial characteristics of other features.

1.8 Branch and bounding sector

Finding the distance between particular location and object locator. GP is still expensive as it examines all objects in D and computes their component scores. Now propose an algorithm that can significantly reduce the number of objects to be examined. The bounds are reasonably tight, in order to facilitate effective pruning.

1.9 Tree evaluation

Algorithm BB derives upper bound scores for non leaf entries in the object tree, and prunes those that cannot lead to better results. Based on the computation range defining the accurate quality measured from the real or synthetic dataset.

2. R-TREE

Spatial data objects often cover areas m multi-dimensional spaces and are not well represented by point locations For example, map objects like counties, census tracts etc., Occupy regions of non-zero size m two dimensions A common operation on spatial data search for all objects an area. For example, to find all counties that have land within 20 miles of a particular point. They replace the recursion stack of the regular top-down traversal with a priority queue. In addition to using the priority queue for nodes, objects are also put on the queue as leaf nodes are processed. The key used to order the elements on the queue is distance from the query object. In order to distinguish between two elements at equal distances from the query object, they adopt the convention that nodes are ordered before objects, while objects are ordered according to some arbitrary (but unique) rule.

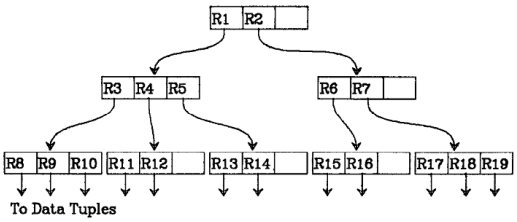

Figure 2.1 R-Tree General Structure

An R-tree 1s a height-balanced tree similar to a B-tree with index records in its leaf nodes containing pointers to data objects Nodes correspond to disk pages If the index is disk-resident, and the structure is designed so that a spatial search requires visiting only a small number of nodes The index is completely dynamic; inserts and deletes can be inter- mixed with searches and no periodic reorganization is required.

2.1 Distance browsing in spatial databases

The incremental and k-nearest neighbor approaches for browsing through a collection of spatial objects stored in an R-tree spatial data structure on the basis of their distances from an arbitrary spatial query object. Present a general incremental nearest neighbor algorithm that is applicable to a large class of hierarchical spatial data structures, and show how to adapt this algorithm to the R-tree.

The transformation process also reveals that the R-tree incremental nearest neighbor algorithm achieves more pruning than the R-tree k-nearest neighbor algorithm. Our R-tree adaptation leads to a considerably more efficient (and conceptually different) algorithm. This is because the presence of object bounding rectangles in the tree enables their use as pruning devices to reduce disk I/O for accessing the spatial descriptions of objects (stored external to the tree).

A node is not examined until it reaches the head of the queue. At this time, all nodes and objects closer to the query object have been examined. Initially, the node spanning the whole index space is the sole element in the priority queue. At subsequent steps, the element at the head of the queue (i.e., the closest element not yet examined) is retrieved, and this is repeated until the queue has been emptied. They start by locating the leaf node(s) containing q.

The distance between the reference point and the each object going to be calculated for every process of input query.Then the minimum distance between the reference point and the one object will be calculated.According to that distance the rank to the object will be provided, the object which holds the minimum distance to the reference point ranked as one. Spatial ranking orders the object according to the distance between the reference point.Non spatial ranking orders the object according to the features.The combination of both spatial and non spatial ranking gives the final rank for the object.

3. TOP K QUALITY

Our top-k spatial preference query integrates these two types of ranking in an intuitive way. Efficient solution for processing the top-k spatial preference query. A brute force approach for evaluating it is to compute the scores of all objects in D and select the top-k ones. Group evaluation technique that computes the scores of multiple points concurrently.

3.1 Evaluating top-k queries

In this section we present strategies for evaluating top-k queries, as defined in Section 2. Specifically, present a naive but expensive approach to evaluate top-k queries. Then, introduce our novel strategies. Adapt existing techniques for similar problems to our framework.

A number of simplifying assumptions in the remainder of this section. Specifically, we assume that the scoring function for all attributes return values between 0 and 1, with 1 denoting a perfect match. Also, assume that exactly one S-Source (denoted S and associated with attribute A0) and multiple R-Sources (denotedR1; : : : ; Rn and associated with attributes A1; : : :;An) are available. (The S-Source S could in fact be of type SR-Source. Random-access capabilities in our discussion.) In addition, Assume that only one source is

accessed at a time, so all probes are sequential during query processing. Thus a random access cannot zoom in on a previously unseen object, i.e., on an object that has not been previously retrieved under sorted access from a source. Therefore, an object will have to be retrieved from the S-Source before being probed on any R-Source. Exactly one S-Source S available objects in S are then the only candidates to appear in the answer to a top-k query.

This set of candidate objects as Objects(S). Besides, Assume that all R-Source R1; : : :;Rn "know about" all objects in Objects(S). In other words, given a query q and an object t 2 Objects(S), we can probe Ri and obtain the score ScoreAi (qi; t) corresponding to q and t for attribute Ai, for all i = 1;::: ; n. Of course, this is a simplifying assumption that is likely not to hold in practice, where each R-Source might be autonomous and not coordinated in any way with the other sources. For instance, in our running example the NYT-Review site might not have reviewed a specific restaurant.

4. REFERENCES

[1] Bruno.N., GravanoL., and Marian.A., (2002) "Evaluating Top-k Queries over Web-Accessible Databases," Proc. IEEE Int'l Conf. Data Eng.(ICDE).

[2] Chen.Y., and Patel.J.M., (2007) "Efficient Evaluation of All-Nearest- Neighbor Queries," Proc. IEEE Int'l Conf. Data Eng. (ICDE).

[3] Chen.Y.Y ,Suel.T., and Markowetz.A., (2006) "Efficient Query Processing in Geographic Web Search Engines," Proc. ACM SIGMOD. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Du.Y, D. Zhang, and Xia.T., (2005) "The Optimal-Location Query," Proc. Int'l Symp. Spatial and Temporal Databases (SSTD).

[5] Guttmann.A., (1984) "R-Trees: A Dynamic Index Structure for Spatial Searching," Proc. ACM SIGMOD.