Available online at www.sciencedirect.com
SCIENCE DIRECT
Information and Software Technology 47 (2005) 785–795
INFORMATION AND SOFTWARE TECHNOLOGY
www.elsevier.com/locate/infsof

Multi-way spatial join selectivity for the ring join graph

Jun-Ki Min[a], Ho-Hyun Park[b,*], Chin-Wan Chung[c]

[a]School of Internet-Media Engineering, Korea University of Technology and Education, Byeongcheon-myeon, Cheonan, Chungnam 330-708, South Korea
[b]School of Electrical and Electronics Engineering, Chung-Ang University, 221, Huksuk-Dong, Dongjak-Gu, Seoul 156-756, South Korea
[c]Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, 373-1, Kusong-Dong, Yusong-Gu, Daejeon 305-701, South Korea

Received 1 April 2004; revised 8 January 2005; accepted 10 January 2005
Available online 7 April 2005

Abstract

Efficient spatial query processing is very important since the applications of the spatial DBMS (e.g. GIS, CAD/CAM, LBS) handle massive amount of data and consume much time. Many spatial queries contain the multi-way spatial join due to the fact that they compute the relationships (e.g. intersect) among the spatial data. Thus, accurate estimation of the spatial join selectivity is essential to generate an efficient spatial query execution plan that takes advantages of spatial access methods efficiently. For the multi-way spatial joins, the selectivity estimation formulae only for the two kinds of query types, tree and clique, have been developed. However, the selectivity estimation for the general query graph which contains cycles has not been developed yet. To fill this gap, we devise a formula for the multi-way spatial ring join selectivity. This is an indispensable step to compute the selectivity of the general multi-way spatial join whose join graph contains cycles. Our experiment shows that the estimated sizes of query results using our formula are close to the sizes of actual query results.

Keywords: Spatial data; Spatial join selectivity; Multi-way join; Databases

1. Introduction

In the past few decades, the research on spatial database management systems (SDBMSs) has actively progressed since the applications using the spatial information such as geographic information systems (GIS), computer aided design (CAD), multimedia systems and satellite image database, and location based service (LBS), have increased

The spatial join is a common spatial query type which requires high processing cost due to the high complexity and large volume of spatial data. Thus, to reduce the overall processing cost, the spatial join is processed in two steps (the *filter step* and the *refinement step*) [5,11]. As shown in Fig. 1(a), the filter step evaluates tuples whether they satisfy the constraints of a given spatial query, using the MBR (Minimum Bounding Rectangle) approximation. The refinement step (Fig. 1(b)) checks the candiate tuples (i.e. the outputs of the filter step) using computational geometric algorithms whether the output tuples really satisfy the constraints of the given spatial query. This paper, like most related spatial database literature, focuses the query processing on the filter step [2,18].

Many spatial queries include the multi-way spatial join because the spatial queries mainly compute relationships (e.g. intersect) among spatial data such as '*Find all buildings which are adjacent to roads that intersect with boundaries of districts*'. The multi-way spatial join combines m ($m>2$) spatial relations using $m-1$ or more spatial predicates.

Since the multi-way spatial join combines tuples from m spatial relations into a single m-tuple whenever the combination satisfies the join conditions (e.g. intersect), the estimated number of the multi-way spatial join result is:

$$\#\text{ of all possible } m\text{-tuples} \cdot \mathrm{Pr}ob(\text{an } m\text{-tuple is a solution}) \qquad (1)$$

The front part of the above formula equals the cardinality of the Cartesian product of m relations, and the latter part is

* Corresponding author. Tel.: +82 2 820 5345; fax: +82 2 825 1584.
E-mail addresses: jkmin@islab.kaist.ac.kr (J.-K. Min), hohyun@cau.ac.kr (H.-H. Park), chungcw@islab.kaist.ac.kr (C.-W. Chung).

0950-5849/$ - see front matter
doi:10.1016/j.infsof.2005.01.002

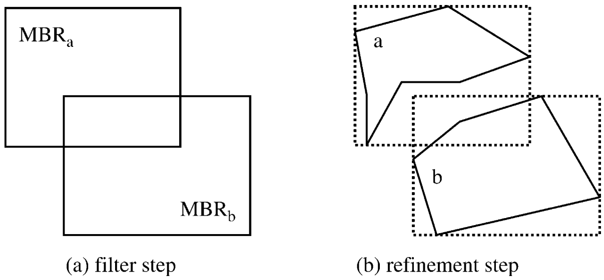

Fig. 1. Spatial join processing steps.

the multi-way spatial join selectivity which means the probability that an m-tuple satisfies the join predicates.

Due to the high complexity of the spatial join operation and a large volume of spatial data, an accurate estimation for the spatial join selectivity has a great influence to the query optimizer and the spatial database management system.

Recently, the cost models for several kinds of spatial joins have been studied [10,12–15,18]. Especially, for the multi-way spatial join, the formulae only for the selectivity of the two kinds of query types, tree and clique, have been developed [14]. However, the formula for the selectivity of the general multi-way spatial join whose join graph contains cycles has not been derived yet. Therefore, the join selectivity which contains cycles could not be estimated accurately. Instead, only the tree typed spatial join and the clique typed spatial join have been used as approximated estimates. However, the selectivity of the tree typed spatial join overestimates the amount of the ring typed spatial join result since the tree typed spatial join does not have a cycle. Also, the selectivity of the clique typed spatial join underestimates the amount of the ring typed spatial join result since the clique typed spatial join considers only the mutually intersected spatial data.

Thus, we develop a formula for the selectivity of the ring typed spatial join. This is an indispensable step for computing the selectivity of the general multi-way spatial join whose join graph contains cycles.

Traditionally in the database area, the selectivity estimation problem, when a query graph contains cycles, was considered as very difficult [8]. Thus, this work should be considered as a theoretical break through to go forward the selectivity estimation problem of the general multi-way spatial join whose join graph contains cycles.

Our contributions are as follows:

- find properties of the result that satisfies the constraints of the multi-way spatial ring join.
- devise a formula for the selectivity for the multi-way spatial ring join.
- show the accuracy of the proposed formula through experiments.

The rest of this paper is organized as follows. Section 2 describes the multi-way tree spatial join selectivity and the multi-way clique spatial join selectivity through a survey of the previous work. Section 3 describes the multi-way ring spatial join among various multi-way spatial joins. Section 4 contains experimental results on uniformly distributed data in the two-dimensional space, which show the accuracy of the proposed formula. Section 5 concludes this paper.

2. Preliminaries

In this paper, like most related spatial database literature, we assume that all spatial data are uniformly distributed in the d-dimensional unit work space, $WS=[0,1)^d$ ('uniformity assumption of the placement distribution' [9]), and all spatial data are rectangles. The notations to be used in this paper are summarized in Table 1.

Formally, a multi-way spatial join can be expressed as follows [14]:

- Given m relations $R_1, R_2, \ldots, R_m$ and a query Q, where Q_{ij} is the binary spatial predicate between R_i and R_j, find all m-tuples $\{\{s_{R_1}, s_{R_2}, \ldots, s_{R_m}\} | \forall\, i,j : s_{R_i} \in R_i,\ s_{R_j} \in R_j$ and $Q_{ij} = TRUE\}$.

A multi-way spatial join can be modeled by a *query graph* G_Q whose nodes represent relations and edges represent spatial predicates.

Various spatial conditions (intersect, meet, include, etc [4,6]) can be applied to spatial join predicates. But, following the standard approach in the literature of spatial joins, the *intersect* (not disjoint) is considered as the default join predicate. Papadias et al. [16] showed how the spatial selectivity using intersect could be applied to the other spatial predicates.

2.1. Selectivity for the multi-way spatial tree join

Huang et al. [7] and Theodoridis et al. [18] provided the formula of the 2-way spatial join selectivity.

As shown in Fig. 2, in order for a spatial object ($=s_{R_i}$) of relation R_i and a spatial object ($=s_{R_j}$) of relation R_j to intersect, assuming the location of s_{R_i} is fixed, an end point of s_{R_j} (denoted by •) should exist within the area (dotted area in Fig. 2) that is an extension of s_{R_i} to the size of s_{R_j} in each dimension. This is the same as, assuming the location of s_{R_j} is fixed, an end point of s_{R_i} (denoted by ∘) should exist within the area (not seen in Fig. 2) that is an extension of s_{R_j} to the size of s_{R_i} in each dimension. Thus, the 2-way spatial

Table 1
Notations

Symbol	Description
s_{R_i}	A spatial object of relation R_i
$\lvert s_{R_i}\rvert$	The average length (on each dimension) of a spatial object in R_i
$\lvert\lvert S\rvert\rvert$	The number of elements of set or relation S