

# Experimental Evaluation of Filter Effectiveness \*

[Extended Abstract]

Lixian Han      Hongjun Zhu      Jianwen Su  
University of California at Santa Barbara

## ABSTRACT

Evaluation of a spatial join typically has two steps: a filter step that applies the join on approximations of spatial objects and a refinement step that determines the actual intersection of two spatial objects. In this paper, we study the effectiveness of filters with different approximations including minimum bounding rectangles, rectilinear approximations, minimum 5-corner bounding convex polygons, and convex hulls. We compare the effectiveness of filters by comparing the number of “false hits” in the filter result using these approximations. We analyze the impact of parameters such as approximation quality and aspect ratios of input polygons on the filter effectiveness.

## 1. INTRODUCTION

The conventional approach to evaluate a spatial join uses two steps: (1) a *filter step* that checks intersections of approximations of spatial objects; and (2) a *refinement step* that checks intersection of objects whose approximations intersect. The most commonly used approximations are *minimum bounding rectangles* (mbr’s). Although a majority of spatial join algorithms use mbr approximations, algorithms that use rectilinear polygons, trapezoids, and convex approximations were recently developed [2]. In this paper, we present an initial experimental study on filter effectiveness for four approximation methods, mbr, rectilinear approximations (rlin), minimum 5-corner bounding convex polygons (5ch), and convex hulls (ch). We investigate several factors (*approximation quality* and *aspect ratios*) that affect filter effectiveness using a specific approximation method. We studied the impact of approximation method and approximation quality on filter effectiveness, the relationships between aspect ratios, approximation quality and filter effectiveness, and filter improvements over mbr w.r.t. aspect ratios.

Our results provide some useful insight into query processing for optimizers. With statistical information about the app-

\*Supported in part by NSF grants IRI-9700370 and IIS-9817432. Authors’ addresses: Dept. of Computer Science, Univ. of California, Santa Barbara, CA 93106. Email: {lxhan, hongjunz, su}@cs.ucsb.edu. A full version of this paper will be available at <http://www.cs.ucsb.edu/~su/pub/cdb/>.

roximation quality of input data for different approximation methods and the aspect ratios of input data, an optimizer may be able to make better decisions about which approximation to use in the filter step.

## 2. EXPERIMENT SETUP

In order to study effectiveness of filters with different approximations and correlations between filter effectiveness and properties of input data, we perform experiments based on sets of simple polygons generated by a pseudo random algorithm. Each polygon is generated based on randomly generated number of sides, center position, positions of boundary points, and distances of boundary points to the center point. Only simple polygons are generated.

We consider two types of parameters:

- *approximation quality* **aq**: the ratio of the area of the actual spatial object to the area of the approximation.
- *aspect ratios*: There exist different versions of aspect ratio. We consider the following three definitions:
  - **ar1** of a spatial object  $o$  is the ratio of the area of  $o$  to the area of a circle that has the same perimeter as  $o$ .
  - **ar2** of a spatial object  $o$  is the ratio of the area of the largest rectangle inside  $o$  to the area of the mbr of  $o$ .
  - **ar3** of a spatial object  $o$  is the ratio of the lengths of the shorter edge over the longer one of the smallest rotated bounding rectangle of  $o$ .

For each parameter, we generate datasets each of which contains 100 randomly generated polygons such that the values of their corresponding parameters are in the same range.

To study the dependency among **aq**, **ar1**, **ar2**, **ar3**, and the effectiveness of the filter step, we perform a nested loop join on every pair of datasets. For each pairs of objects from the two input datasets in the join, we examine the intersections of their mbr, rlin, 5ch, and ch approximations (resp.).

## 3. EXPERIMENTAL RESULTS

*Approximation methods vs. filter effectiveness.* We compared the numbers of false hits (non-intersecting objects whose approximations intersect) in the filter result with different approximations on 1 million polygon pairs. The results (Fig.1) show that mbr is the least effective, the effectiveness increases for rlin and 5ch, and ch is the best. This is consistent with the work in [1].

*Approximation quality vs. filter effectiveness.* We examined the dependency between the approximation quality and the filter effectiveness. For each approximation method, we compute the ratio of false hits over the total number of approximation intersections. Fig.2 shows a representative

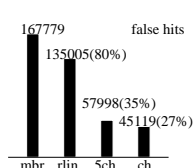


Fig. 1. Effectiveness

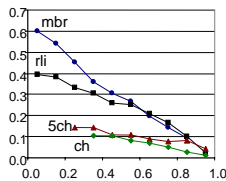


Fig. 2. aq vs. Effectiveness

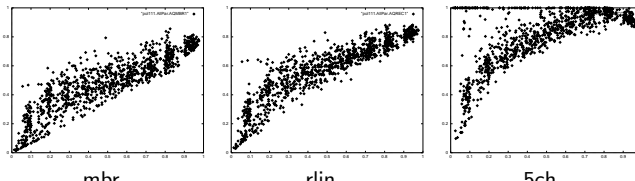


Fig. 3. ar1 vs. aq

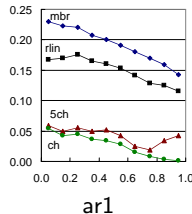


Fig. 4. ar1, ar3 vs. Effectiveness

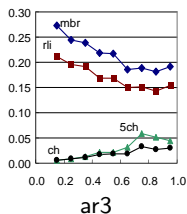


Fig. 4. ar1, ar3 vs. Effectiveness

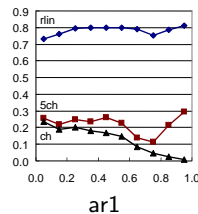


Fig. 5. ar1, ar3 vs. relative false hits

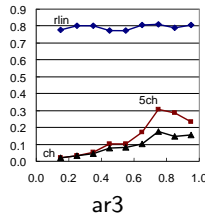


Fig. 5. ar1, ar3 vs. relative false hits

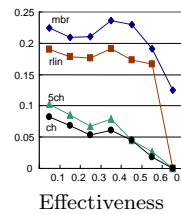


Fig. 6. ar2 with skewed datasets

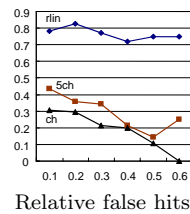


Fig. 6. ar2 with skewed datasets

result of several joins we performed. It reveals that for each approximation method, when the aqs of objects increase, the ratio of false hits over the total number of approximation intersections decreases. Clearly higher aq means smaller number of false hits, thus more effective filter.

**Approximation quality vs. aspect ratios.** We also studied the relationship between the approximation quality and aspect ratios. Fig.3 shows ar1 and aq of mbr, rin, and 5ch for 1000 polygons whose ar1s are evenly distributed from 0 to 1 ( $x$ -axis for ar1 and  $y$ -axis for aq). Other datasets show similar results. The result of ar1 and aq for ch are similar to that for 5ch. The figure also shows that for each approximation method, when ar1 increases, aq also increases. Moreover, the aq change for mbr and rin is almost at the same speed, slower than that for 5ch. However, when ar1 is close to 1, the aq of 5ch decreases. ar1 measures roundedness of a polygon. When a polygon has ar1 close to 1, it is close to a circle and its 5ch approximation will have more extra area outside the polygon. Thus aq of 5ch decreases when ar1 is close to 1.

For ar3, we have similar results (in full paper) for mbr and rin. However, aq for 5ch and ch is more evenly distributed and even decreases slightly when ar3 increases. This is due to our random polygon generator. When ar3 is small, the generator tends to generate polygons with smaller number of boundary points that are likely to be convex. But when ar3 is large, the generator produces polygons with more points and are less likely to be convex. Therefore, when ar3 is small, 5ch and ch of generated polygons are closer to the polygons themselves, and their aqs are closer to 1. When ar3 is large, this is not the case.

**Aspect ratios vs. filter effectiveness.** We also conducted experiments on the aspect ratios and filter effectiveness. Similar to the experiments we did for aq, we compute the ratio of false hits over the total number of approximation intersections. We performed same tests for each approximation method. Fig.4 shows the results for ar1 and ar3 from joins of datasets with same range of the aspect ratios. It shows that in general for each approximation method, when ar1 increases, the number of false hits decreases (filter becomes more effective). Note that when ar1 is close to 1, the false hits ratio for 5ch increases. This is consistent with the 5ch results from Fig.3. For mbr and rin, it is also true that when ar3 increases, the number of false hits decreases. However, the curves for 5ch and ch move up. This is due to the decreasing of aq observed previously.

**Relative false hits.** In another set of experiments, instead of comparing the ratio of false hits to total intersections, we compared the ratio of false hits for rin, 5ch, and ch over that for mbr, since mbr is least effective. Relative false hits ratio shows the improvements over mbr. Fig.5 exhibits the relative false hits ratios for ar1 and ar3. The results show that although the number of false hits for rin is smaller than mbr, the relative false hits ratio stays almost constant. When ar1 increases, the relative ratio for 5ch and ch decreases. These results further reveal the relationships among aq, aspect ratios, and filter effectiveness. For example, we know from Fig.3 that when the aspect ratios of input data increase, aq for mbr and rin increase. The aq of rin changes in the same speed as the aq of mbr. From Fig.2, aq and false hits seem to have a linear relationship. It follows that the relative false hits for rin keeps almost unchanged when ar1 of the input data change. Other curves can be explained similarly.

**Aspect ratio ar2.** Fig.6 shows results for ar2 based on our earlier tests in which two datasets in a join contain 1000 random polygons each, and we only count intersections of polygons whose aspect ratios are within the same range. (The datasets are skewed because the values of parameters of polygons are not evenly distributed. Thus the results are less conclusive. We are conducting more experiments and will report all results in the full paper.) Even with the skewed data, results for ar2 shows that aq and ar2 are related in a similar way to aq and ar1 though not as obvious. In Fig.6 (left) we can still see a general decrease of false hits when ar2 increases, though there are some ups due to data skew. The right figure is similar to the one for ar1 in Fig.5.

## 4. CONCLUSIONS

Our experiments show that among the three aspect ratios, ar1 is the best metric for filter effectiveness. In particular, the higher the aspect ratio, the higher the approximation quality is and the more effective the filter becomes. It is very interesting to compare the overall performance of spatial join evaluations using rin or 5ch filters vs. using mbr filters.

## 5. REFERENCES

- [1] T. Brinkhoff, H-P. Kriegel, R. Schneider, and B. Seeger. Multi-step processing of spatial joins. In *Proc. ACM SIGMOD*, 1994.
- [2] H. Zhu, J. Su, and O. H. Ibarra. Towards spatial joins for polygons. In *Proc. Int. Conf. on SSDBM*, 2000.