Set Containment Joins: The Good, The Bad and The Ugly

Karthikeyan Ramasamy[*]
UW-Madison
karthik@cs.wisc.edu

Jignesh M. Patel[†]
UM-Ann Arbor
jignesh@eecs.umich.edu

Jeffrey F. Naughton
UW-Madison
naughton@cs.wisc.edu

Raghav Kaushik
UW-Madison
raghav@cs.wisc.edu

Abstract

Efficient support for set-valued attributes is likely to grow in importance as object-relational database systems, which either support set-valued attributes or propose to do so soon, begin to replace their purely relational predecessors. One of the most interesting and challenging operations on set-valued attributes is the set containment join, because it provides a concise and elegant way to express otherwise complex queries. Unfortunately, evaluating these joins is difficult, and naive approaches lead to algorithms that are very expensive. In this paper, we develop a new partition based algorithm for set containment joins: the Partitioning Set Join Algorithm (PSJ), which uses a replicating multi-level partitioning scheme based on a combination of set elements and signatures. We present a detailed performance study with a complete implementation in the Paradise object-relational database system. Our results show that PSJ outperforms previously proposed set join algorithms over a wide range of data sets.

1 Introduction

The data modeling community has long realized that set valued attributes provide a concise and natural way of modeling complex data [RKS98]. Recently, there has been a resurgence of interest in set-valued attributes from two different perspectives. First, commercial O/R DBMS [Sto96] are beginning to support set-valued attributes, which is likely to lead to their use in "real" applications. Second, the rise of XML as an important data standard increases the need for set-valued attributes, since it appears that set-valued attributes are key for the natural representation of XML data in relational systems [SHT+99]. Unfortunately, although sets have been fairly well studied from a data-modeling viewpoint [Zan83], very little has been published about the efficient implementation of operations on set-valued attributes. In this paper, we consider the implementation of a particularly challenging operation over set-valued attributes, the set-containment join.

Many real world queries can be easily expressed using set containment joins. Consider a simple relation that describes a document and set of hyper-links that point to it.

DOCUMENT(did, {hyper-links-in}, actual-document)

Suppose document d_1 is more important than d_2 if d_1 is linked-to by a superset of the documents that link to d_2. We can find pairs of documents d_1 and d_2 where d_1 is more important than d_2 with the following query:

```
SELECT d1.did, d2.did
FROM DOCUMENT d1, DOCUMENT d2
WHERE d2.hyper-links-in ⊂ d1.hyper-links-in
```

The algorithms available for implementing set-containment joins depend upon how set-valued attributes are stored in the database. As described in [KJD00], sets can be stored in the *nested internal representation* (set elements are stored together along with the rest of the attributes) or the *unnested external representation* (set elements are scattered and stored in a separate relation). To the best of our knowledge, current commercial O/R DBMS use the unnested external representation. Since the unnested external representation reduces to standard SQL2 relations under

[*]Work supported in part by a grant from the Microsoft Corporation

[†]Portions of this research were done while the authors were at NCR Corporation, Madison, WI

Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.

the covers, set containment joins on the unnested external representation can be evaluated by rewriting the queries into SQL2 (with no sets) and evaluating these rewritten queries. On the other hand, with the nested internal representation, the most obvious algorithm for evaluating set-containment joins is nested loops. Two questions immediately arise: (1) Are there better algorithms than nested loops? (2) How do these algorithms compare in efficiency with the rewrite in SQL2 approach that is most logical for the unnested external representation?

This paper attempts to answer these questions by proposing a new partition-based join algorithm for set containment joins, which we call PSJ. Partition-based algorithms certainly dominate join algorithms in scalar and spatial domains, so it is natural to suspect that a partition-based algorithm will be the algorithm of choice for set-containment joins.

This paper makes two main contributions. First, it presents the new algorithm PSJ for set containment joins. Second, it includes an extensive performance study of three set containment algorithms: the traditional SQL approach on the unnested external representation, signature nested loops and PSJ on the nested internal representation. Our experience with an implementation in the Paradise object-relational database system [PYK+97] shows that PSJ yields significant speedup over both the SQL-based approach and signature nested loops. An added benefit of this algorithm is that, like all partition-based algorithms, it is trivially parallelizable. Finally, our results present a strong case for storing sets in the nested internal form, since PSJ and even signature nested loops outperform the rewritten queries over the unnested external representation.

1.1 Related Work

Joins have been studied extensively in relational [MK76], [Bra84], [DKO+84], [DNS91] and spatial domains [LR96], [PD96]. Pointer joins for efficiently traversing path expressions in object-oriented databases has also been studied extensively [DLM93], [SC90]. However, there is very little previous work on set containment joins. The only reported work of which we are aware is the work by Helmer and Moerkotte [HM96], [HM97]. These papers investigate nested loops algorithms for computing a set containment join and propose a new signature based hash join. We discuss these algorithms in Sections 3 and 4.2

1.2 Paper Organization

The rest of the paper is organized as follows. Section 2 defines the problem of set containment and the notation used in the paper. Various storage representations for sets, the SQL approach and signature nested loops joins are explained in detail in Section 3. The partition based set join algorithm is outlined in Section 4. Section 5 presents a detailed performance study of all the algorithms. The conclusions and future work are presented in Section 6.

2 Problem Definition and Notations

For the rest of the paper, we consider the two relations $R(a, \{b\})$ and $S(c, \{d\})$ containing the set valued attributes $\{b\}$ and $\{d\}$ respectively. Since set is a type constructor, attributes b and d can be of any arbitrary type and we assume that these types provide an equality predicate that compares the equivalence of two set elements. Also we do not assume any order among the set elements. The set containment join, $R \bowtie_{\{b\} \subseteq \{d\}} S$, pairs tuples in relation R and S such that $\{b\}$ is subset of $\{d\}$. Table 1 describes the notation used in the rest of the paper.

3 Previously Proposed Algorithms

Options for algorithms for set containment joins heavily depend on how the set valued attributes are stored in the database. In order to make this paper self-contained, we briefly discuss the options for storing set-valued attributes.

3.1 Storage Representations for Sets

Various representations for sets are possible depending on the following two characteristics: **nesting** (set elements are clustered or scattered) and **location** (set elements are either stored with the rest of the attributes internally or vertically partitioned and stored externally). As outlined in [KJD00], the two main representations for sets are:

- **Nested Internal**: Here the set elements are grouped together and stored with the rest of the attributes in the tuple.
- **Unnested External**: In this representation, the set-valued attribute is stored in a separate relation. For each set-valued attribute in a relation, two relations are created: (1) A base relation that stores the other non set-valued attributes and an identifier, and (2) An auxiliary relation that stores each element of the set-valued attribute as a tuple with the (corresponding) identifier.

3.2 Join Algorithms for Unnested External

If sets are stored in the unnested external representation, set-containment joins can be expressed and evaluated using standard SQL2 constructs. This approach is important to study, because (a) it is the simplest to add to any RDBMS, and (b) perhaps because of (a), to our knowledge the commercial O/R DBMSs all use this approach. As discussed in Section 3.1, in this representation, a relation with a set-valued attribute is decomposed into two relations. A set containment operation can then be expressed using SQL over these

$\mid R \mid$	Relation cardinality of R (# of tuples)	$\mid S \mid$	Relation cardinality of S (# of tuples)
r_R	Average set cardinality of R	r_S	Average set cardinality of S
σ	Selectivity of $R \bowtie_{\{b\} \subseteq \{d\}} S$	f	False drops as a percent of $\sigma \mid R \mid\mid S \mid$
IO_{seq}	Cost of a sequential I/O	IO_{rand}	Cost of a random I/O

Table 1: Notations

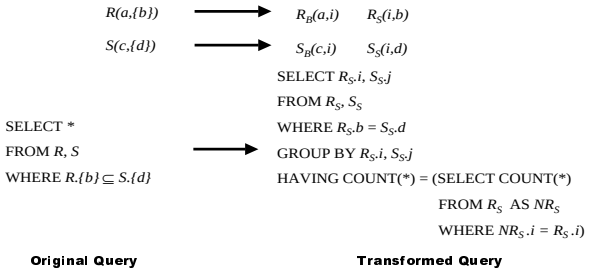


Figure 1: Original and Transformed SQL Queries (excluding final joins for a and c)

decomposed relations. If R and S are the two relations being joined, and R_S and S_S are the corresponding decomposed auxiliary set relations, then the original and transformed queries are shown in Figure 1.

The rewritten query involves a correlated nested sub-query and hence it is expensive to evaluate. A possible optimization is to use magic-sets rewriting [SPL96] and transform the original query into the set of queries shown in Figure 2, thus evaluating the inner query only once (as opposed to once for every tuple produced by the outer block). Our experiments show empirically that even this approach performs very poorly unless the set sizes and relation sizes are small; in fact, in many cases, it is so bad that the algorithm can arguably be called "ugly".

3.3 Signature Nested Loops Algorithm for Nested Internal

The signature nested loops algorithm proposed by [HM97] attempts to reduce the cost of evaluating the containment predicate by approximating sets using signatures and evaluating the join predicate by comparing these signatures. A signature is a fixed length bit vector that is computed by applying a function M iteratively to every element e in the set and setting the bit determined by $M(e)$. If the containment predicate $s \subseteq t$ is to be satisfied for two signatures s and t, then the following condition is necessary: *For all bit positions that are set to 1 in signature s, the corresponding bits in signature t should be set to 1.* However, this condition is not sufficient since signatures are only an approximate representation for the set (unless the signature length is equal to the size of the domain of the set). Hence using signatures to evaluate a predicate will yield false drops. The actual sets must be examined to eliminate these false drops.

The signature nested loops algorithm operates in three phases: the *signature construction phase*, the *probing phase*, and the *verification phase*. During the signature construction phase, the entire relation R is scanned, and for every tuple $t_i \in R$, a signature s_i is constructed. A triplet (c_i, s_i, OID_i) is computed and stored in an intermediate relation R_{sig}; here c_i is the set cardinality and OID_i is the physical record identifier (rid) of the tuple. The same process is repeated for the relation S and an intermediate relation S_{sig} is created. Next, the algorithm proceeds to the probing phase, where the tuples of R_{sig} and S_{sig} are joined. For every pair $(c_i, s_i, OID_i) \in R_{sig}$ and $(c_j, s_j, OID_j) \in S_{sig}$, two conditions must be verified (i) $c_i \leq c_j$ and (ii) $s_i \wedge s_j = s_i$, where the wedge represents the bit-wise and of the two signatures. If both the conditions are satisfied, then the pair (OID_i, OID_j) is a possible candidate for the result. During the final verification phase, the tuples referred to in the candidate (OID_i, OID_j) pairs are fetched and the subset predicate is evaluated on the actual set instances, producing the final result.

The main issue in the signature nested loop join algorithm is reducing the number of false drops to minimize the cost of the verification phase. The false drop probability depends on the number of bits used in constructing the signature. The greater the signature length, the smaller will be the false drop probability. However, larger signatures lead to more bit comparisons per signature, thereby increasing the execution time of the probing phase. Hence, it is necessary that the chosen signature size be such that further increases in the number of bits do not significantly reduce the false drop probability. Based on the definition of false drop probability, we derive an equation for the optimal signature length (F) as

$$F = \frac{-r_S}{ln\left(1 - \left(\frac{f\sigma}{1-\sigma(1+f)}\right)^{1/r_R}\right)} \quad (1)$$

The detailed derivation is presented in [KJJK00].

Note that even with the signatures of an ideal length, this algorithm compares signatures for every pair of tuples in the cross product of R and S. If R and S each has one million tuples, there are one trillion comparisons. This is discouraging enough to be considered "bad."

4 Partitioned Set Join (PSJ)

In this section, we propose a new algorithm for the nested internal representation that is based upon partitioning. In general, partition based algorithms for joins (scalar and spatial) attempt to optimize join execution by partitioning the problem into multiple

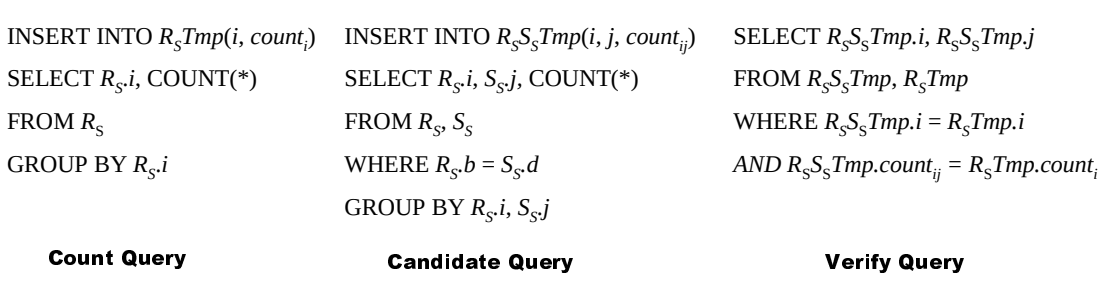


Figure 2: Magic Sets Rewriting

smaller subproblems using a partitioning function. First, the relation R is partitioned into k partitions, $R_1, R_2, \ldots, R_k$. Similarly, the relation S is partitioned into $S_1, S_2, \ldots, S_k$ using the same function. Note that we are using a generalization of the classical definition of partitioning in that one tuple may be mapped to multiple partitions.

The algorithm proposed in this section, called the Partitioned Set Join Algorithm (PSJ), uses a two level partitioning scheme. It operates in three phases:

- **Partitioning Phase**: Each tuple of R is sent to exactly one partition based on the first level partitioning function h. Each tuple of S, in general, is replicated across multiple partitions using (the same) h.
- **Joining Phase**: Each partition of R is joined with its counterpart in S using a second level partitioning function that operates on signatures. Hence false drops are possible.
- **Verification Phase**: The tuple pairs that the join phase indicates could join, are compared to remove any false drops.

The subsequent sections describe each of the phases in detail.

4.1 Partitioning Phase

This phase uses a partitioning function h that operates on the set elements. The partitioning phase begins by reading the relation R. For each tuple r of R, the following steps are executed

1. A 3-tuple (c_i, s_i, OID_i) is computed, where c_i is the set cardinality, s_i is the signature of the set instance, and OID_i is the OID of the tuple.
2. A random element e_R is picked from $r.\{b\}$.
3. The 3-tuple is sent to the partition determined by $h(e_R)$.

Observe that the 3-tuple for each tuple of R is sent only to one partition. Now the relation S is read. For each tuple s of S, the following steps are executed

1. A 3-tuple (c_i, s_i, OID_i) is computed.
2. *For each* element $e_S \in s.\{d\}$, the 3-tuple is sent to the partition determined by $h(e_S)$.

Note that if $r.\{b\} \subseteq s.\{d\}$ then the partition determined by $h(e_R)$ will contain the 3-tuples corresponding to r and s. Hence the algorithm computes containment correctly.

4.2 Joining Phase

During the joining phase, each partition of R is joined with its counterpart in S. There are various algorithms that could be used in this phase. However, at this point, the tuples in each partition do not carry the actual set instances since they are approximated by signatures. Hence the join algorithm in this phase has to operate directly on signatures. In this phase, we use a partition based in-memory algorithm using signatures.

The joining algorithm works in two steps: the *build step* and the *probe step*. In the build step, an array A of size equal to the number of bits in the signature is constructed. Now the partition R_i is scanned and each 3-tuple (c_i, s_i, OID_i) is read. A bit position m that is set to 1 is chosen randomly from the signature. The 3-tuple is inserted into $A[m]$. At the end of first step, the signatures from partition R_i have been partitioned.

During the probe step, partition S_i is scanned. For each 3-tuple (c_j, s_j, OID_j) the chain of signatures in $A[n]$ is examined whenever bit n is set to 1 in s_j. The containment predicate is evaluated (as in Section 3.3) for each signature encountered in the chain and the candidate pairs (OID_i, OID_j) are inserted into a temporary relation. These candidate pairs potentially satisfy the containment relationship.

This phase of the algorithm is similar to signature hash join (SHJ) proposed in [HM97]. We use a single bit in the signature to determine the array index for R. SHJ in general uses more bits (a partial signature) to determine the array index. For S, SHJ requires all possible subset signatures to be enumerated for a given partial signature to determine the chains to be probed. This enumeration is exponential.

4.3 Verification Phase

In the verification phase, we examine the actual R and S tuples to determine whether they satisfy the join

condition. The main issues involved in this phase are speeding up set containment verification and avoiding random seeks while fetching the tuples. Refer to [KJJK00] for a full discussion of the techniques used to accomplish these goals.

4.4 Estimation of Number of Partitions and Signature Size

The performance of PSJ depends two factors: the number of partitions (P_{PSJ}) and the signature size (F_{PSJ}). The desired number of partitions further depends on two parameters: the average set cardinality and the relation cardinality. Even though the speedup is expected to increase as the number of partitions is increased, in practice, the overhead associated with each partition prevents such unbounded speedup.

In order to estimate the desired number of partitions, we employ a detailed analytical model which accounts for the overheads. Based on this model, we estimate the ideal number of partitions as

$$P_{PSJ} = \left(\frac{|R||S|(1-(1-\frac{1}{F})^{r_S})}{Z} \right)^{1/3} \quad (2)$$

where $Z = 2IO_{rand} + 2IO_{seq} + H$

The derivation of this equation is presented in [KJJK00]. The fudge factor H accounts for various system dependent factors. The fudge factor is likely to vary across systems. For a given system, H can be determined by choosing a sample set of data and running the algorithm for various partitions.

Since partitioning avoids many redundant comparisons, one can expect the signature size to be lower (when compared to Sig-NL). Also, as the number of partitions is increased the signature size is expected to get lower. We derive an equation for signature size.

$$(1-e^{-r_S/F_{PSJ}})^{r_R} - \frac{f\sigma P_{PSJ}}{\left(1-\left(1-\frac{1}{F_{PSJ}}\right)^{r_S}\right) - \sigma P_{PSJ} - f\sigma P_{PSJ}} = 0 \quad (3)$$

We use bisection method to solve this equation. There is a cyclic dependency between equations (2) and (3). Hence both the equations have to be solved simultaneously. We use these equations to determine the appropriate combination of partitions and signature size in our experiments for PSJ. As we shall see in Section 5.8 and Section 5.9, fortunately the performance curves as a function of the number of partitions and signature size are rather flat. So these equations do not have to be exact for reasonable performance.

5 Performance Evaluation

In this section, we evaluate the performance of the three set containment algorithms: the SQL approach for the unnested external representation (**SQL**), and the signature nested-loops (**Sig-NL**) and **PSJ** algorithms for nested internal. As a special case, we also ran PSJ with one partition which we call **PSJ-1**. The special case of one partition is important when applicable, because it has no partitioning overhead. We first describe our implementation of these algorithms and then present results from various experiments designed to investigate the performance of these algorithms under various conditions.

5.1 Implementation

Paradise is a shared nothing parallel object-relational system developed at the University of Wisconsin-Madison [PYK+97]. We implemented sets using the ADT mechanism in Paradise. The set ADT implements a number of set-oriented methods, including: create-iterator, which returns an iterator over the elements of the set; and set operators which are implemented by type specific methods invoked by the query engine when comparison and assignment are performed on sets. For more details on the implementation, refer to [KJD00], [RPN].

We implemented signature-nested loops (Sig-NL) and PSJ as join algorithms in the system, and extended the optimizer to recognize set containment join operations in queries. For the SQL approach, magic set optimization was used to rewrite the correlated nested query as shown in Section 3.2. In order to ensure that the optimizer did not choose bad plans, optimal physical plans for each query were fed into the system rather than the queries themselves.

5.2 Experimental Setup and Data Generation

In our experiments, the total size of the non set-valued attributes in a tuple was 68 bytes. The average size of each set element was 30 bytes. We ran the experiments on an Intel 333 MHZ Pentium processor with 128MB of main memory running Solaris 2.6. We used a 4GB disk for storing the database volume. The disk was mounted as a raw device. It provided an I/O bandwidth of 6 MB/sec. Paradise was configured with a 32MB buffer pool. Though this buffer pool size may seem small compared to current trends in memory, we used this value since we wanted to test data sets that were much larger than the buffer pool. As will be seen in the following sections, with this buffer pool size, some experiments take many days to run. Each experiment was run against a cold buffer pool to eliminate the effect of file caching. The data generator for the BUCKY benchmark [CDN+97] was modified to generate data synthetically. The data generator takes as input the cardinality of the relations R and S, the average cardinality of the set valued attributes in the two relations, the size of the domain from which the set elements are drawn, and a correlation value. For each tuple, the set-valued attribute is generated as follows. First, the data generator divides the entire domain into 50 smaller sub-domains. The set elements are drawn from these sub-domains. Set elements are correlated if they are drawn from the same sub-domain. Correla-

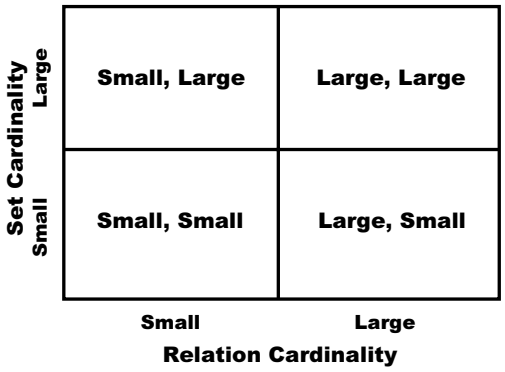


Figure 3: Taxonomy of Set Distributions

tion of a set instance is defined as the percentage of the set elements that are drawn from a single sub-domain. For example, if the set cardinality is 10, a correlation of 90% implies that 9 set elements are picked from one sub-domain and 1 element is randomly chosen from one of the remaining 49 sub-domains. All the experiments used a correlation of 10% unless otherwise specified. Joining tuples were generated such that every R tuple joins with exactly one S tuple.

5.3 Set Distributions

There are many distributions involving set valued attributes because there are many degrees of freedom:

- Average set cardinality of relation R and S
- Relation cardinality of R and S
- Size of domain from which the set elements are drawn
- Degree of correlation among the elements.

Each parameter can influence the performance of the containment algorithm. In an effort to reduce the problem space, we restricted ourselves to varying the relation and set cardinalities. Based on these two parameters we have four possible quadrants as shown in Figure 3 and the experiments explore each of the quadrant in detail. We chose the response time as our performance metric.

5.4 Varying Relation Cardinality

In this set of experiments, we investigated the effect of varying the relational cardinality. The domain size was fixed at 10000. Since the join was not symmetric, we further refined the experiments based on different cardinalities of R and S :

- The relation cardinalities of R and S are varied together and the values are kept the same.
- The relation cardinality of S is kept constant at a large value and that of R is varied.
- The relation cardinality of R is kept constant at a large value and that of S is varied.

5.4.1 Vary Relation Cardinalites of R and S

In this experiment, the relation cardinality was varied for two values of set cardinality: 20 and 120. The results of these experiments are plotted in Figure 4. The numbers for the SQL approach for relation cardinalities greater than 20000 are not included in the figure since these runs took more than 24 hours. The main observation is that PSJ outperforms (or performs as well as) other algorithms consistently over the entire space of relation cardinality. On the other hand, the SQL approach starts getting worse from 10000 onwards. Section 5.5 discusses why the SQL approach performs poorly. Sig-NL and PSJ are analyzed in Section 5.6.

5.5 Performance of the SQL Approach

As seen from Figure 4, the SQL approach performs reasonably well at very small relation and set cardinalities. However, as the relation sizes increase (note the peak at 10000), the response time increases rapidly. The cost breakdown of the SQL approach shows that most of the time is dominated by candidate generation query.

- The input to the joins are two large set relations R_S and S_S.
- The number of intermediate tuples generated as a result of the join is also large.
- The number of groups generated from the aggregate operator is also large.

For a detailed cost breakdown of SQL approach, refer the expanded version of the paper [KJJK00]. Because of the aforementioned problems and consequent performance degradation, the SQL approach is not considered in the remaining sections.

5.6 Sig-NL Vs PSJ

The individual cost breakdown of these algorithms is shown in Figure 5, Figure 6 and Figure 7.

In general, the cost of these algorithms consists of three components: partitioning cost, comparison cost and verification cost. The cost of Sig-NL and PSJ-1 do not have any partitioning cost. The cost of Sig-NL can be broken down into signature creation cost (labeled as Rsig-creat and Ssig-creat in the graphs), join cost (labeled as Sig-join) and sort and verify costs (labeled as Sort and Verify). The cost of PSJ-1 is broken down into build cost (labeled as R-build), probe cost (labeled as S-probe), and sort and verify costs (labeled as Sort and Verify). The cost of PSJ is broken into partition creation and deletion cost (labeled as Part-creat and Part-delete), partition cost (labeled as Spart-time and Rpart-time), join cost (labeled as Part-join) and sort and verify costs (labeled as Sort and Verify). The comparison cost is high in Sig-NL. It decreases in PSJ-1 and is least in PSJ.

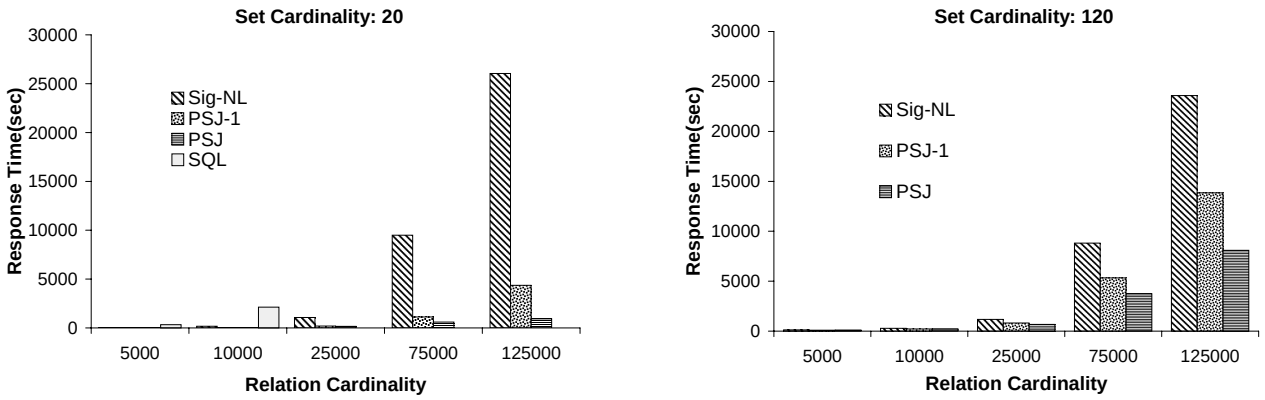


Figure 4: Varying Relation Cardinalities

The first observation is that PSJ outperforms PSJ-1 and Sig-NL consistently as seen from Figure 4. Sig-NL spends most of its execution time in comparing the signatures (see Figure 5), whereas the execution time of PSJ-1 is dominated by the signature probing cost (see Figure 6). Looking at Figure 7, we observe that the cost in PSJ is distributed across the partitioning, signature joining and the verification costs. Partitioning reduces the signature comparisons, but requires a partitioning phase. For PSJ to perform well the reduction in the number of comparisons from partitioning should be significant, and the partitioning cost should not be too high. The reduction in number of comparisons is dominant at higher relation cardinalities as seen in Figure 6 and Figure 7 (compare Part-Join in PSJ with S-probe in PSJ-1 and Sig-Join in Sig-NL). Hence PSJ consistently performs better at higher relation cardinalities. For lower relation cardinalities, the cost gained by avoiding unnecessary comparisons is not high.

The second observation is that the gap between PSJ and the rest is smaller for set cardinality of 120. This is because the partitioning cost is higher for larger set cardinalities. In addition, the comparison cost also increases because of replication. Another contributing factor is the requirement of large signature sizes for lower set cardinalities of R. This unexpected phenomenon occurs because the probability that a given set instance in R joins with some set instance in S increases as its cardinality decreases. Hence in order to keep the false drops minimum, an increase in the signature size is required. For example, in Sig-NL when the relation cardinality of R (and S) was 25000, the required signature size was 181 bits for a set cardinality of 20 while it was 104 bits for a set cardinality of 120. This larger signature size has a much greater impact on Sig-NL and PSJ-1. Note however that as the average set cardinality of S increases, the signature size increases as expected.

The third observation is that PSJ-1 outperforms Sig-NL consistently. This is expected since several unnecessary comparisons are eliminated. Quantitatively, for a set cardinality of 20 and relation cardinality of 25000, Sig-NL requires 625 million comparisons whereas PSJ-1 requires only 80 million comparisons. When the set cardinality is 120, the number of comparisons increases since the expected number of bits set to 1 in the signature increases thereby causing more chains to be examined for a given set of S. Hence the performance gap between the two decreases.

We also conducted experiments where the cardinality of one relation was fixed and the other was varied. The trends observed were the same.

5.7 Varying Set Cardinality

In this experiment, we varied the set cardinality for two different relation cardinalities: 20000 and 100000 to explore the quadrants of small and large relation cardinalities. The signature size for Sig-NL and PSJ-1 and the number of partitions for PSJ were chosen using equations (2) and (3). The domain size was set at 10000. The results are plotted in Figure 8 and the cost breakdown of PSJ-1 and PSJ are shown in Figure 9 and Figure 10.

For a given relation cardinality, as the set cardinality increases, the gap between PSJ and the rest diminishes. In fact for a relation cardinality of 20000 when the set cardinality is 160, PSJ-1 marginally outperforms PSJ. This is because the partitioning cost increases rapidly with increasing set cardinality as seen in Figure 10. This happens because more partitions are required and replication is higher. At the larger relation cardinality of 100000, the set cardinality threshold beyond which PSJ-1 outperforms PSJ increases as expected.

5.8 Effect of Signature Size

In this experiment, we study the effect of signature size on the performance of Sig-NL and PSJ. Both algorithms use signatures for producing an intermediate

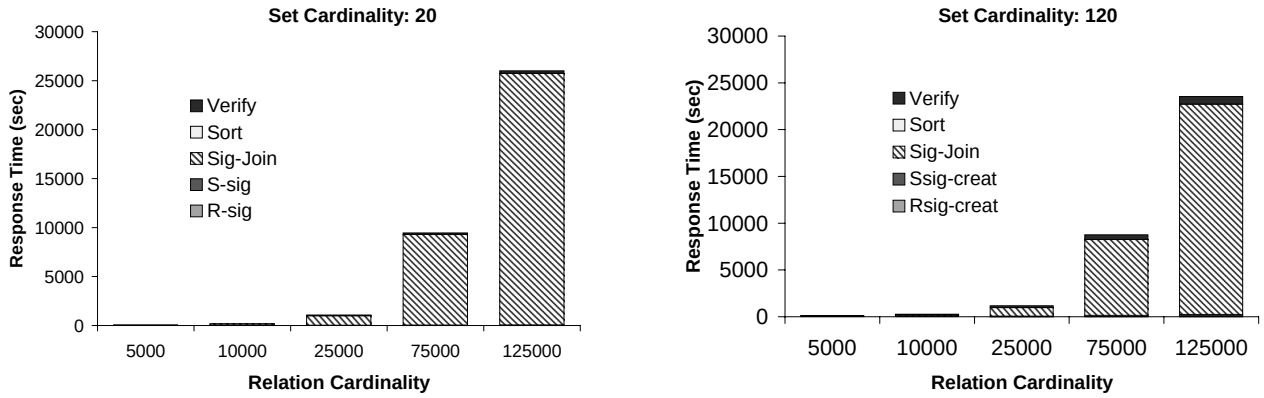


Figure 5: Cost Breakdown for Sig-NL

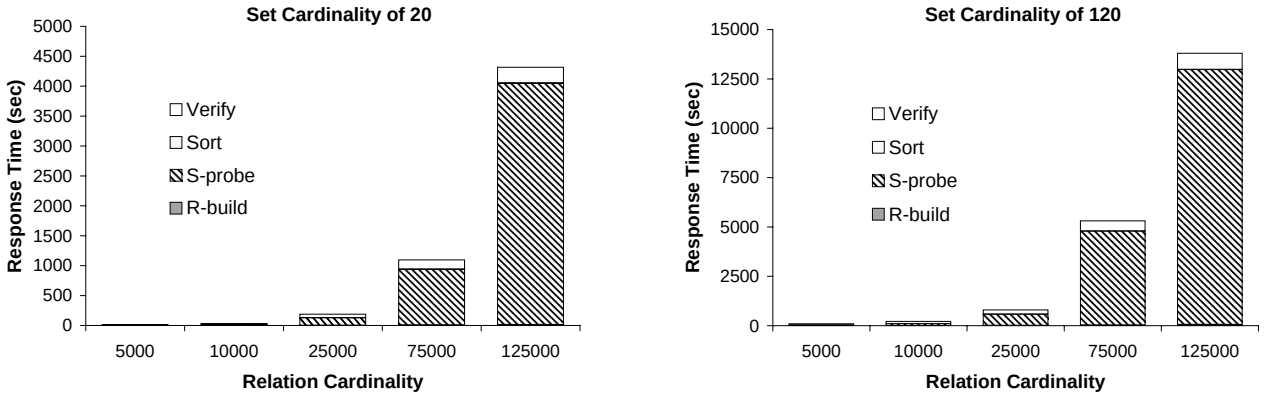


Figure 6: Cost Breakdown for PSJ-1

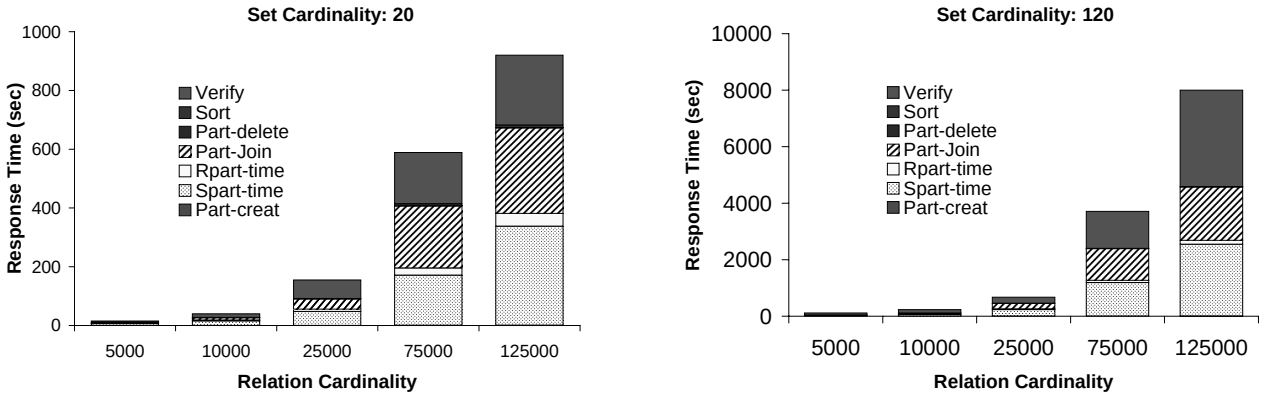


Figure 7: Cost Breakdown for PSJ

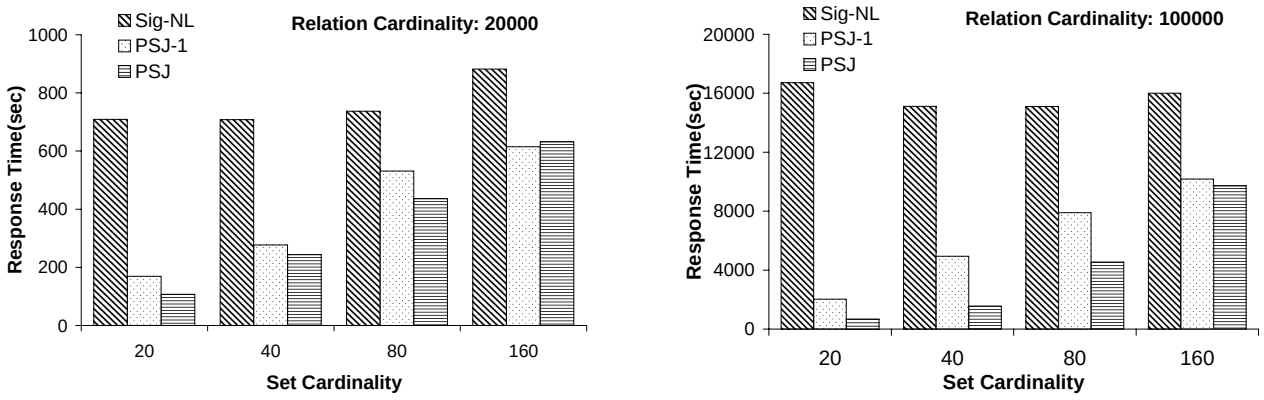


Figure 8: Varying Set Cardinality

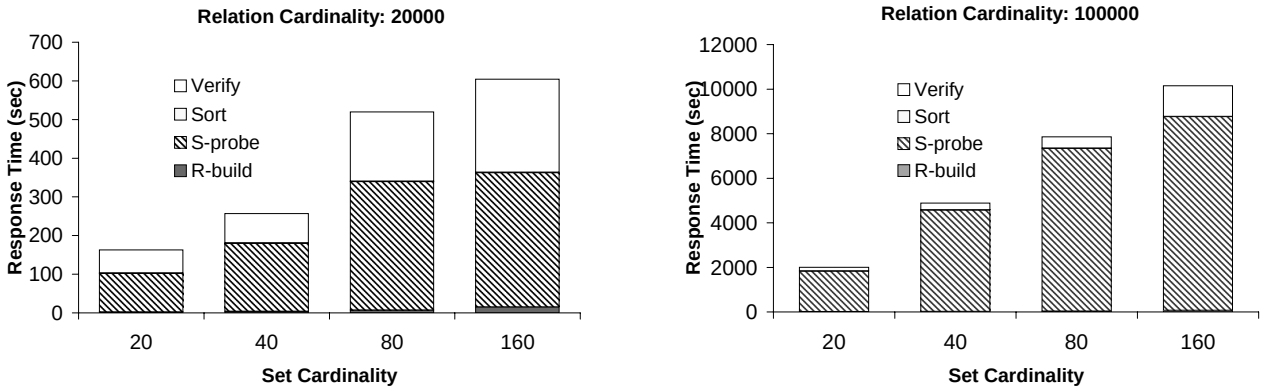


Figure 9: Cost Breakdown for PSJ-1

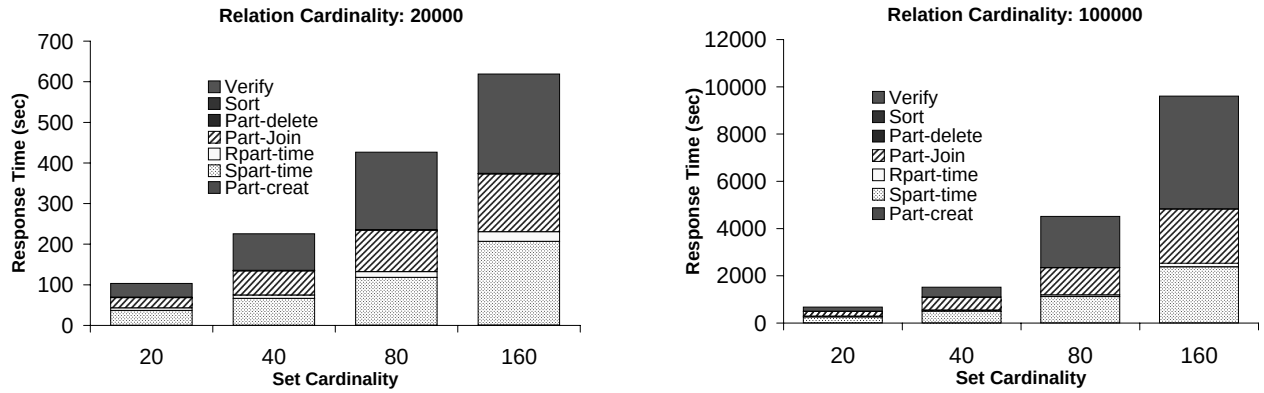


Figure 10: Cost Breakdown for PSJ

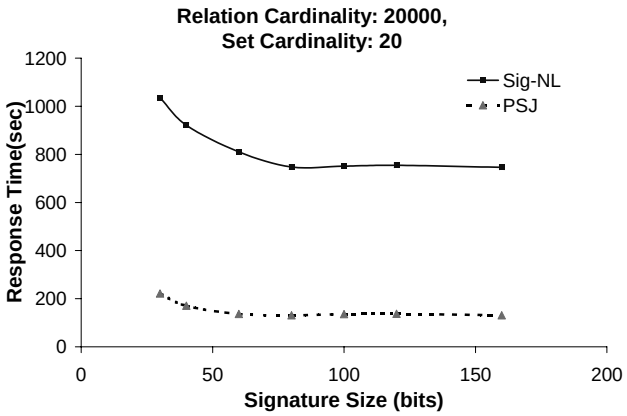


Figure 11: Effect of Signature Size

candidate set of result. As noted in Section 3.3, the number of false drops in the candidate set is influenced by the size of the signature. Hence the choice of signature size is important both in Sig-NL and PSJ. For this experiment, we used a relation cardinality of 20,000 for both R and S, an average set cardinality of 10 for R, and average set cardinality of 20 for S. The size of domain was fixed at 10,000. For PSJ, we used the optimal number of 42 partitions, as predicted by equation (2). The result of this experiment is plotted in Figure 11.

The first observation is that for smaller signature sizes, Sig-NL is very expensive. This is because many elements in the domain hash to the same bit, thereby increasing the false drops. Such an increase in the false drops increases the time of the verification phase. As the signature size increases, the number of false drops reduces and hence the performance of Sig-NL improves. However, after a signature size of 80, increasing the signature length does not cause any significant improvement in the performance of Sig-NL. The second observation is that PSJ is relatively immune to the signature size. This is because partitioning reduces the number of false drops.

For this data set, the signature size for Sig-NL predicted by equation (1) was 173 bits. For PSJ with 42 partitions, the signature size predicted by equation (3) was 116 bits. Given the flatness of the PSJ curve, it is not important to get the signature size exactly right.

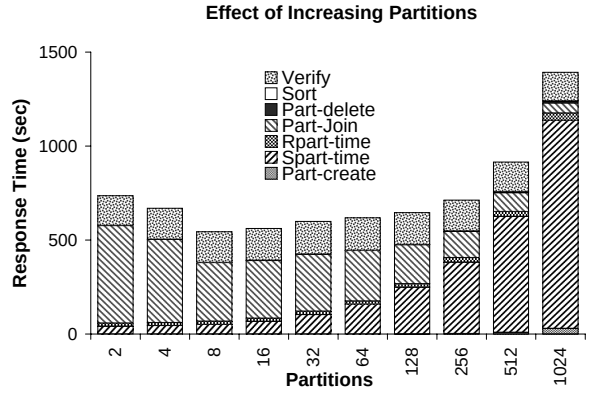
5.9 Effect of Increasing Partitions in PSJ

In this experiment, we study the effect of the number of partitions on the performance of PSJ. The relation cardinality of both relations was set at 20,000 and the set cardinality was set at 120. The set elements are drawn from a domain size of 10000. An appropriate combination of partitions and signature size was used as determined by equations (2) and (3). The results of this experiment is shown as the first graph in Figure 12. It shows the breakdown of total cost: partition creation and deletion times (the time to create and delete the partition files in SHORE, the storage manager used in Paradise), partition time (the time taken to insert tuples into the partition files), join time, sort time and verification time. From this figure, we observe that PSJ has three phases: the first phase, in which the total cost decreases gradually as the number of partitions is increased; the second phase in which the total cost is approximately constant; and the third phase in which the total cost starts increasing as the number of partitions becomes very large.

In order to further investigate the sharp increase in partitioning overhead, we plot both the total number of pages generated by the algorithm and the actual number of pages that the system uses (see the second graph in Figure 12. The actual number of pages generated by the system counts the number of disk pages that were created by the algorithm. This number is higher than the number of pages generated by the algorithm as it includes the per-tuple overhead, and the overhead due to fragmentation. The storage manager allocates pages in extents (a group of pages) and fragmentation occurs when there are unused pages in the extent. The graph shows that as the number of partitions increases, there is a corresponding increase in the size of the data generated (because of the increased replication of S tuples). However, the replication of each tuple is bounded by the set cardinality, and, consequently the increase in the amount of data generated slows down after 64 partitions. However, the actual number of pages required still continues increasing rapidly because of fragmentation. In addition, other costs like the the number of buffer pool pins and unpins, the cost of creating and deleting the partitions also increases with the number of partitions. Thus, the partitioning overhead increases sharply when the number of partitions is large. This experiment shows that the number of partitions has a critical impact on the performance of PSJ. The equation (2) can be used to estimate a reasonable number of partitions. For set cardinality of 120, the number of partitions chosen by the equation was 70.

5.10 Disk Space Requirements

Here we investigate the size of the intermediate space required for Sig-NL and PSJ. We do not consider PSJ-1 since it is an in-memory algorithm. We ran two experiments to examine the disk space requirements. In the first experiment, we set the relation cardinalities to 100,000 and varied the set cardinality, and in the second experiment, we set the set cardinality to 120 and varied the relation cardinalities. The results of these two experiments are plotted in Figure 13 and Figure 14 respectively. In these two figures, we plot the number of pages generated by each algorithm (labeled as Sig-NL-Gen and PSJ-Gen in the graphs) and the actual number of pages created on disk (labeled as Sig-NL-Actual and PSJ-Actual). The main observation



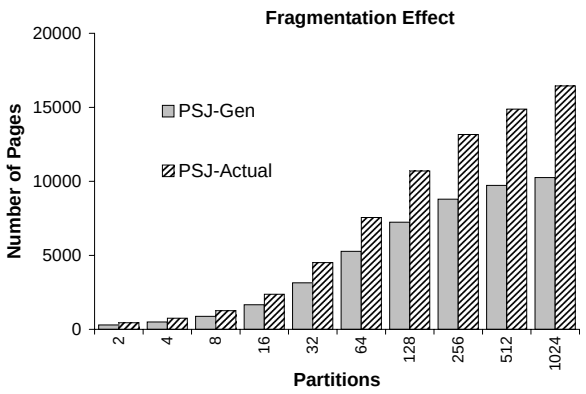


Figure 12: Effect of Increasing Partitions and Fragmentation for Set Cardinality of 120

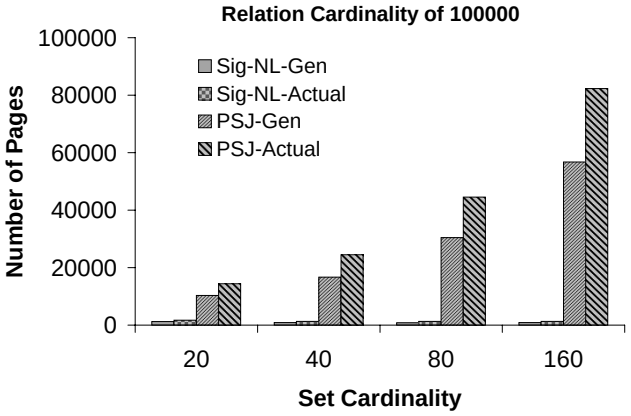


Figure 13: Disk Space, Varying Set Cardinality

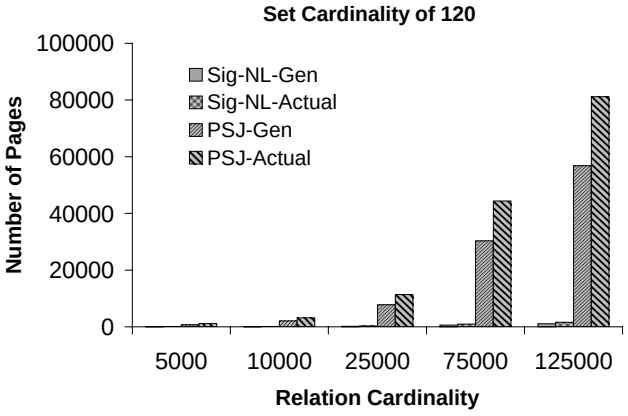


Figure 14: Disk Space, Varying Relation Cardinality

is that Sig-NL requires much less storage than PSJ as expected. The number of pages required by Sig-NL varies slightly because of the variation in signature size. Since the number of pages required by Sig-NL is so low, there is a high probability that these pages will remain in the buffer pool during the operation of the algorithm. PSJ on the other hand requires a large amount of intermediate storage that steadily increases as the cardinality increases. This behavior in PSJ is caused by the following two factors: a) the number of times the 3-tuple (as described in section 4.1) is replicated increases as set cardinality increases and b) the number of tuples per partition increases as the relation cardinality increases.

For large data sets, the memory requirement for PSJ-1 is very high since the entire set of R signatures has to be accommodated. On the other hand, Sig-NL and PSJ adapt themselves to available amount of memory. Hence they are well suited to a multi-user environment.

6 Conclusions and Future Work

This paper investigates algorithms for computing a set containment join. These algorithms cover two possible implementations of set valued attributes: the unnested external representation and the nested internal representation. The unnested external representation is used by commercial O/R DBMSs for implementing set-valued attributes. In this case, set containment join is implemented using a standard SQL2 query. For the nested internal representation, this paper considers two algorithms. The first is a variation of nested loops (Sig-NL) that uses signatures to speed up the evaluation of the join predicate. The second algorithm is PSJ, a new partition based algorithm that is proposed in this paper. This algorithm is based on a two level partitioning scheme by using set elements to partition relation R and replicate relation S. Within each partition, it uses an in-memory algorithm based on partitioning of signatures.

This paper also presents a detailed performance study of the three algorithms. The performance space of these algorithms is summarized in Figure 15. For small data sets and small set cardinalities, PSJ works well. The SQL approach and Sig-NL performs reasonably well for extremely small data sets and small set cardinalities; however, as the relation or the set cardinality size increases the performance degrades very

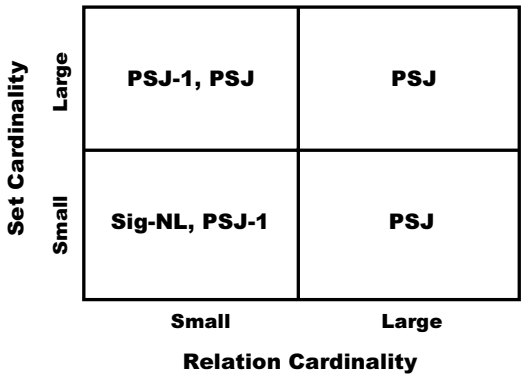


Figure 15: Performance Space of Set Containment Algorithms

rapidly. PSJ with one partition is usable at higher set cardinalities provided there is enough memory. Elsewhere, PSJ is the algorithm of choice.

Since the native SQL approach performed so poorly, we are investigating how the benefits of PSJ can be achieved even in systems that use the unnested external set representation. One obvious approach would be to execute the join by: (a) converting the inputs from the unnested external format to a temporary nested internal approach, (b) doing the join, (c) reconverting the output. In this way the nested internal approach is just an internal data structure of the join algorithm. Clearly this will be much faster than the native SQL over unnested external approach (which took days in some of our tests.) In future work we plan to investigate this and other alternatives.

References

[Bra84] K. Bratbergsengen. Hashing methods and relational algebra operations. In *Proceedings of International Conference on Very Large Databases (VLDB)*, pages 323–333, 1984.

[CDN+97] M. Carey, D. Dewitt, J. F. Naughton, M. Asgarian, P. Brown, J. E. Gerke, and D. N. Shah. The bucky object-relational benchmark. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1997.

[DKO+84] D. Dewitt, R. Katz, F. Ohlken, L.Shapiro, M.Stonebraker, and D. Wood. Implementation techniques for main memory database systems. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 1–8, 1984.

[DLM93] D. Dewitt, D. Lieuwen, and M. Mehta. Pointer-based join techniques for object-oriented databases. In *PDIS*, 1993.

[DNS91] D. Dewitt, J. Naughton, and D. Schneider. Parallel sorting on a shared-nothing architecture using probabilistic splitting. In *PDIS*, Miami Beach, 1991.

[HM96] S. Helmer and G. Moerkotte. Evaluation of main memory join algorithms for joins with set comparison join predicates. Technical report, University of Mannheim, 1996.

[HM97] S. Helmer and G. Moerkotte. Evaluation of main memory join algorithms for joins with set comparison join predicates. In *Proceedings of International Conference on Very Large Databases (VLDB)*, Athens, Greece, 1997.

[KJD00] K.Ramasamy, J.Naughton, and D.Maier. High performance implementation techniques for set valued attributes. Technical report, Computer Sciences Department, University of Wisconsin, Madison, 2000.

[KJJK00] K.Ramasamy, J.Patel, J.Naughton, and K.Raghav. Set containment joins: The good, the bad and the ugly. Technical report, Computer Sciences Department, University of Wisconsin, Madison, 2000.

[LR96] M. Lo and C. Ravishankar. Spatial hash-joins. In *Proceedings of ACM SIGMOD Conference on Management of Data*, Montreal,Quebec, May 1996.

[MK76] M.W.Blasgen and K.P.Eswaran. On the evaluation of queries in a relational database system. Technical report, IBM, 1976.

[PD96] J. Patel and D. DeWitt. Partition based spatial merge join. In *Proceedings of ACM SIGMOD Conference on Management of Data*, Montreal,Quebec, May 1996.

[PYK+97] J. Patel, J. Yu, N. Kabra, K. Tufte, B. Nag, J. Burger, N. Hall, K. Ramasamy, R. Lueder, C. Ellman, J. Kupsch, S. Guo, J. Larson, D. DeWitt, and J. Naughton. Building a scalable geo spatial dbms: Technology, implementation and evaluation. In *Proceedings of ACM SIGMOD Conference on Management of Data*, Tucson, Arizona, May 1997.

[RKS98] M. Roth, H. Korth, and A. Silberschatz. Extending relational algebra and calculus for nested relational databases. *ACM Transactions on Database Systems*, 13(4):389–417, December 1998.

[RPN] K. Ramasamy, J. Patel, and J. Naughton. Efficient pairwise operations for nested set valued attributes. Working paper.

[SC90] E. J. Shekita and M. J. Carey. A performance evaluation of pointer based joins. In *Proceedings of ACM SIGMOD Conference on Management of Data*, pages 300–311, 1990.

[SHT+99] J. Shanmugasundaram, G. He, K. Tufte, C. Zhang, D. DeWitt, and J. Naughton. Relational databases for querying xml documents: Limitations and opportunities. In *Proceedings of International Conference on Very Large Databases (VLDB)*, Scotland, 1999.

[SPL96] P. Seshadri, H. Pirahesh, and T.Y. C. Leung. Complex query decorrelation. In *Proceedings of IEEE Conference on Data Engineering(ICDE)*, pages 450–458, 1996.

[Sto96] M. Stonebraker. *Object-relational DBMS: The Next Great Wave*. Morgan Kaufmann, 1996.

[Zan83] Carlo Zaniolo. The database language gem. In *Proceedings of 1983 ACM SIGMOD Conference on Management of Data (SIGMOD)*, San Jose, California, May 1983.