Research Accomplishments

Shashi Shekhar
Department of Computer Science, University of Minnesota
200 Union Street S.E. # 4-192, Minneapolis, MN - 55455
TEL: (612) 624-8307  FAX: (612) 625-0572

October 13, 2000
Abstract

My research over the past decade at the University of Minnesota has primarily dealt with the
general research area of Data and Knowledge Engineering in the subareas of storage methods,
query processing and data mining, focusing on the unique needs of the application domains of
spatial databases, including transportation, ecology, climatology and army. Within the subarea
of storage methods, we have worked on both clustering and declustering techniques for spatial
databases. Within the subarea of query processing, we have worked on parallel strategies for
range queries on spatial data, sequential strategies for direction based selection, and semantic
query optimization. Within the emerging subarea of data mining, we have worked on neural
networks and now started work on the modelling of spatial dependence for mining spatial
datasets.

For the past 5 years, our work has focused primarily on spatial databases. We have also
worked on preparing a textbook and a survey paper with the goal of developing an undergraduate
course for students studying Computer Science, Geographic Information Systems and related
disciplines. This survey work has helped us identify the following subareas needing further
research:

Mining Spatial Datasets: Spatial datasets often exhibit high auto-correlation, which violates
the independent identical distribution assumption made by data mining techniques based
on classical statistics. We are exploring new data mining methods which can model auto-
correlation.

Processing Direction Predicates: Spatial data models (e.g. the Open Geo-data Interchange
Standard) are maturing in the areas of topological (e.g. overlap) and metric (e.g. distance)
predicates. Range query processing strategies are quite mature for processing selection
queries based on topological or metric predicates. However, range query strategy is quite
ineffective for processing selection queries based on directional relationships (e.g., left,
front). We are developing new query processing strategies for selection queries based on
direction predicates.

Parallel processing of Spatial Predicates: Main memory based parallel processing strate-
gies for spatial queries are important for high performance applications like flight simula-
tion. Traditional declustering techniques (e.g. range partitioning, hashing) perform poorly
on spatial datasets since the workload imposed by extended spatial objects can vary by
orders of magnitude depending on the query, the location of the object, and the spatial
properties of the object. We have developed systematic techniques for declustering col-
lections of extended spatial objects using the ideas of local load-balancing, similarity and
data-distribution functions.

Storage Methods for Spatial Graphs: Spatial databases traditionally use geometry based
storage and indexing techniques (e.g. R*-trees) for all kinds of spatial datatypes. We
showed the graph connectivity properties are more important than the geometric proximity
properties for spatial graphs (e.g. roadmaps). We developed a connectivity clustered
access method (CCAM) which almost always outperforms traditional spatial indices based
on geometric properties.

A set of transparencies are enclosed at the end of this writeup to illustrate the impact of our
research.
1 Recent Accomplishments

1.1 Survey of Spatial Databases

In recent years, we have devoted considerable time to consolidate the research to date in spatial databases by organizing numerous workshops and preparing survey papers[1, 2]. We are completing a textbook towards this purpose as well. Copies of one of the survey papers[1] and book[3] description (e.g., preface, table of content) are enclosed in the collection of ten representative publications. This document does not describe either the survey paper or the book.

1.2 Data Mining : Mining Spatial Databases for Location Prediction

We are currently working on a project related to spatial data mining. Spatial data mining is a process of discovering interesting and useful but implicit spatial patterns. People have been using maps, the common way of representing spatial data, to extract useful spatial information for thousands of years. One famous example is related to the London Asiatic Cholera which ravaged the city in the mid-nineteenth century. A scientist plotted all the known cases of cholera on a map and noticed that the centroid of the points on the map was a water pump. When the pump was turned-off, the cholera began to subside. The goal of data mining is to ‘automate’ this search for “nuggets” of information embedded in very large quantities of data.

The current approach is to use classical data mining tools after ”materializing” spatial relationships and assuming identical and independent distributions (i.i.ds) of the data sets. This assumption is not true for spatial attributes with spatial auto-correlation, a key property distinguishing spatial data from non-spatial data. Positive spatial autocorrelation implies that attribute values of neighboring data samples tend to systematically affect each other. The general implication of this property is that the conventional assumption of independence which underlies much of classical statistical theory may be inapplicable in the case of spatial data. Specifically, the standard Gauss-Markov assumptions in regression modeling do not apply to spatial data.

The key question then is to quantify the neighborhood influence of spatial data and incorporate it into classical data mining techniques. Spatial statistics has explored new models, e.g. spatial auto-correlation based regression (SAR) and Neighbor expectation maximization to account for spatial auto-correlation by modifying the model or the objective function. However, these methods are computationally extremely expensive, and do not scale up to medium or large data sets.

Our goal is to define and explore scalable techniques for spatial data mining. We have proposed a new technique for spatial data mining driven by ”map similarity” measures, which is a linear combination of spatial accuracy(e.g. average distance between predicted and actual location) and classification accuracy(e.g., fraction of sites classified correctly). Our results show that PLUMS yields spatial accuracy comparable to those achieved by SAR using order of magnitude fewer CPU cycles and less memory. Thus PLUMS scales for large data sets. We plan to explore the behavior of PLUMS in greater detail by studying the effect of alternative choices related to design issues on the solution quality. We are now developing formalism to characterize the dominance zone of PLUMS using algebraic models.

This work is supported by the National Science Foundation through infra-structure grants as well as by the Army Research Laboratory. Preliminary results[4] are published in the 1999 NSF Workshop on Data Mining in GIS, the SIGMOD 2000 Workshop on Data Mining as well as the
KDD 2000 Workshop on multimedia data mining. One of the reviews at the SIGMOD workshop on data mining is quoted below:

“When the area of spatial data mining was first introduced, there seemed to be a lot of promise. However, recent studies do not excite me, and I think the spatial data mining area has been stagnant for a while. This paper proposes using spatial autocorrelation, which appears to be natural. This may represent a breakthrough for the area of spatial data mining. While the success of this proposed technique is too early to judge, I encourage the authors to continue on this direction.” – A reviewer

1.3 Query Processing for Spatial Predicates

1.3.1 Strategies for processing direction predicate based selections

This work was primarily sponsored by the Army Research Laboratory (ARL) for querying battlefield information using direction predicates. Example queries included: list the enemy units behind a given ridge.

A major focus of this project was on the modeling and processing of directional relationships (e.g. north, left, front) for spatial databases. Previous work modeled directions as binary boolean relationships and performed qualitative reasoning by enumerating a large number of inference rules without an independent interpretation model. As a departure from traditional methods, we proposed a novel vector based framework to model direction as a spatial object. This object view of direction enables the definition of new spatial data types such as open shapes and oriented objects at the abstract object level. Open shapes include open lines and open rectangles and are characterized by partially defined boundaries and open infinite extents. The two-dimensional half-planes used widely in geometry are special cases of open rectangle. Secondly, a direction object can have its own attributes and operation sets. We further proposed equivalence classes of direction objects to unify spatial reasoning with different direction predicate sets. By defining an algebra on equivalence classes, we constructed a framework to model the semantics of direction predicates. The equivalence classes together with the algebra provide an independent interpretation model for qualitative direction reasoning. Many inference rules, related to direction modeling, that have appeared in the literature can be easily derived using this framework, significantly simplifying spatial reasoning.

This project also contributed a new, efficient, and scalable algorithm, called the open shape based strategy (OSS), to process direction predicate-based selection queries in spatial databases. Previous work on direction predicate based spatial query processing has focused on processing absolute directions using range query strategies and R-tree indexing. However, many direction queries depend on the orientation of reference objects (or the viewer) which may change due to motion. Classical methods are inefficient when orientation of the reference object is different from that of the global reference system. We proposed OSS to address this problem. OSS converts the processing of the directional queries to the processing of topological operations between open shapes and objects. It eliminates false hits at the earliest opportunity while recursively searching hierarchical indices like R-tree. Since OSS models the direction region as an open shape, it also eliminates the computation related to the embedding world boundary. The algebraic analysis and experimental results demonstrate that OSS consistently outperforms classical strategies (often by
an order of magnitude) in terms of both I/O and CPU cost. This work will have important and lasting impact on the design and implementation of spatial database search engines and spatial data modeling.

This work[5] was published in journals (IEEE Transactions on Knowledge and Data Engineering, GeoInformatica) and conference proceedings (ACM International Symposium on Advances in Geographic Information Systems (ACMGIS), IEEE Knowledge and Data Engineering Exchange Workshop). One of our papers presented at ACMGIS’98 was voted among the best papers.

1.3.2 Parallel strategies for processing range queries

Visualization applications like flight simulators and virtual reality environments use spatial databases to represent actual terrain. Applications like these impose stringent restrictions on acceptable performance and response time. Sequential spatial databases do not meet these requirements, but parallel databases can. Our work on parallel spatial databases addresses two important issues: static declustering methods and dynamic load-balancing methods for the range-query problems. Declustering of spatial objects is used to reduce the static load imbalances in parallel formulations for range-query problem. In addition, dynamic load-balancing methods reduce the remaining load imbalances during the run-time for a parallel formulation.

Declustering and dynamic load-balancing are important issues in designing a high performance spatial database system. Declustering objects like polygons and chains of line-segments is difficult because the workload imposed by individual objects can vary by orders of magnitude depending on the query, the location of the object, and the spatial properties of the object. Dynamic load-balancing is difficult because transferring spatial objects between processors at run-time is often more expensive than processing the objects locally. We developed systematic techniques for declustering collections of extended spatial objects using the ideas of local load-balancing, similarity, and data-distribution functions. We also developed selective-replication based techniques for dynamic load balancing in spatial databases. We experimentally evaluated the proposed methods on a distributed memory MIMD machine (Cray T3D) and on a shared-memory machine (SGI Challenge). Our results showed that systematic declustering without dynamic load-balancing (DLB) is often better than the conventional technique of random declustering with DLB. The results also showed that systematic declustering is needed within the DLB to determine the units of work transfer. These are major new insights for parallelizing spatial databases and have been recognized by the community. In the words of one reviewer at the International Symposium on Large Spatial Databases (SSD95) ¹:

"this paper is going to be one of the most referenced paper in spatial databases". – a reviewer

This work[6, 7] was published in IEEE Computer's special issue on Shared Memory Multiprocessing, December 1996 (selection ratio 1 in 6). Complete results were published in IEEE Transactions on Knowledge and Data Engineering, one of the most influential journals in Spatial Databases. In addition, parts of the results were also published in international conferences including the International Symposium on Spatial Databases: 1995 (SSD 95), First Workshop on

¹the leading international conference on spatial databases

This work was sponsored by the Army Research Laboratory (ARL) and the implementation of declustering and load-balancing algorithms in context of range query processing strategies were used by the Battlefield Visualization Group at the ARL, Adelphi. The National Science Foundation supported this work through infra-structure grants funding the SGI Challenge computing facilities.

1.4 Storage Methods: Clustering Methods for Spatial Graphs

The goal of this work, sponsored by the National Science Foundation as well as the Federal Highway Administration, was development of technology for in-car navigation.

Current Spatial Database Management Systems (SDBMS) provide efficient access methods and operators for point and range queries over collections of spatial points, line segments, and polygons. However, it is not clear if existing spatial access methods can efficiently support graph computations which traverse line-segments in a spatial graph based on connectivity rather than geographic proximity. The expected I/O cost for many graph operations can be reduced by maximizing the Weighted Connectivity Residue Ratio (WCRR), i.e., the chance that a pair of connected nodes that are more likely to be accessed together are allocated to a common page of the file. Connectivity-Clustered Access Method (CCAM) is an access method for general graph that uses connectivity clustering. CCAM supports the operations of insert(), delete(), create(), and find() as well as the new operations, get-A-successor() and get-successors(), which retrieve one or all successors of a node to facilitate aggregate computations on graph. The nodes of the graph are assigned to disk pages via a graph partitioning approach to maximize the WCRR. CCAM includes methods for static clustering, as well as dynamic incremental reclustering, to maintain high WCRR in the face of updates, without incurring high overheads. We have also described possible modifications to improve the WCRR that can be achieved by existing spatial access methods. Experiments with graph computations on the Minneapolis road map show that CCAM outperforms existing access methods, even though the proposed modifications also substantially improve the performance of existing spatial access methods.

We evaluated various spatial access methods for spatial graphs (e.g. roadmaps) and various algorithms for storing the attributes of spatial graphs [8, 9, 10, 11, 12]. We have also designed a new access method, CCAM, to efficiently support graph operations and aggregate route evaluation queries over general spatial graphs such as road maps. CCAM adapts the heuristic graph-partitioning approach to cluster the nodes of a given graph into file pages using the connectivity relationship. A detailed description of this method appears in [13, 14]. We summarize the main ideas in this subsection.

In our design, a graph is modeled as a list of nodes, and each node has attributes named successor-list and predecessor-list, which represent the outgoing and incoming edges. The predecessor-list facilitates updating the successor-lists during the insertion and deletion of nodes. We focus on developing access methods which can provide efficient support of Get-A-successor() and Get-successors(), along with the traditional operations of Find(), Insert() and Delete(), etc.

The goal of an access method is to minimize the total cost of frequent operations and aggregate queries for accessing the nodes and edges of a given spatial graph. In [14], we prove the following result:
Theorem 1  The expected cost of accessing an edge and the expected cost of spatial graph operations are minimized by maximizing the Weighted Connectivity Residue Ratio (WCRR), where

$$WCRR = \frac{\text{Sum of weights over edges }(u,v) \text{ such that } \text{Page}(u) = \text{Page}(v)}{\text{Total sum of weights over all edges}}.$$  

The weight associated with an edge \(e(u,v)\) represents the relative frequency of a query accessing nodes \(u\) and \(v\) together. This result can be understood in the context of the Get-successors() operation as well. The high value of WCRR implies that a high fraction of the neighbors of the given node are in the same page as the given node. Thus, fewer pages need to be retrieved to main memory to process the Get-successors().

We propose dynamic reclustering strategies to handle dynamic updating effects. The principle of our dynamic reclustering strategy is to reorganize a small set of pages which are connected in the page access graph [15]. Alternate heuristic methods to maintain a high WCRR without incurring a high cost of reorganization during insertion and deletion are identified and evaluated. Analytical models and experiments [14] with spatial graphs like the Minneapolis road map show that CCAM achieves a higher WCRR and lower I/O costs for spatial graph operations relative to other spatial access methods such as Grid File.
2 Past Accomplishments

2.1 Storage Methods: Declustering Methods for Spatial Datasets

We have worked with declustering problems for spatial databases in the area of characterizing the problem and its difficulty, along with designing declustering methods for different data distributions. The detailed results can be found in [16, 17, 18].

A perfect declustering method should divide the load of processing a large query uniformly over all processes and disks. Perfect declustering methods can be designed for some query sets, including linear queries on rows and columns of matrices. However, in [17], we show that no disk allocation method can be perfect for the set of all orthogonal range queries over spatial data.

**Theorem 2** There is no perfect allocation method for orthogonal range queries over more than five disks.

Thus, we should focus on identifying useful subsets of orthogonal range queries and should devise methods which can parallelize those subsets. We evaluated mapping-function-based disk-allocation methods such as the Linear allocation method and the Lattice allocation method in [17]. The former was shown to be perfect for path queries (rows, columns, diagonals, anti-diagonals), on simple data distributions. The latter was shown to be perfect for small range queries and for some path queries on simple data distributions.

In [18], we proposed a new similarity-based declustering technique which can take advantage of the available information about query distribution, data distribution, data-item size and constraints on partition size. The declustering technique is based on the max-cut partitioning of a similarity graph with data-items as nodes. To handle the insertion of new data items during dynamic updates, we proposed the incremental max-cut technique, using a local window strategy and the max-cut similarity criterion, to best suit dynamic allocation. We showed that the proposed method can provide optimal speed-up for a query-set if there exists any declustering method which can do so.

**Theorem 3** If there exists a perfect allocation/partition method for a query set $Q_s$, then max-cut declustering is also perfect with respect to the query set $Q_s$.

This implies that the proposed max-cut declustering approach provides optimal speed-up for several applications, including row/column queries over uniformly distributed data and partial match queries with only one unspecified attribute over cartesian product files. Our experiments in the context of spatial range queries and grid files showed that the similarity-based declustering methods outperform conventional methods for non-uniform data distributions [18]. They outperform conventional methods even for declustering grid files that store uniform data distributions, if the grid file is created without a statically defined grid directory.

2.2 Query Processing

**Semantic Query Optimization**

Our work on query optimization has focused on semantic query optimization [19, 20, 21], and real-time search algorithms [22, 23, 24, 25, 26]. For brevity, the contributions of the work on
the implementation of rule based systems [?, ?], and query processing in heterogeneous federated databases [27] are not discussed here.

A semantic query optimizer transforms the original query into syntactically different but semantically equivalent queries, which may yield a lower-cost execution plan. Our work addressed the two major difficulties in realizing the promise. First, the cost of optimization can become extremely large for a semantic optimizer and wipe out the savings achieved in the query execution time. Second, the availability of effective query transformation rules may be severely limited.

A major contribution of our work is the design and evaluation of an effective stopping criterion to find the near-optimal trade-off between the additional optimization time and the improvement in query execution time. The stopping criterion minimizes the response time, i.e. the weighted sum of the optimization time and execution time of a query.

**Theorem 4** Search termination criterion \((\lambda \times \tau(i) \leq t(i))\) guarantees that the total response time \(RT(i)\) achieved is near-optimal, i.e., \(\frac{RT(i)}{RT_{opt}} \leq \max(1 + \lambda, 1 + \frac{1}{\lambda})\), where \(t(i)\) represents the cost of the best query execution plan so far, and \(\tau(i)\) represents the total optimization cost so far from start up to the current node \(Q_i\) in the search space of semantically equivalent queries.

We formally and experimentally demonstrated that this criterion can achieve a near-optimal response time. The detailed results appeared in the Journal of Data and Knowledge Engineering [20], and the Proc. of the VLDB Conf. [19].

To address the second problem, we contributed a new learning algorithm to extract query transformation rules from the data in the database for the semantic query optimizer. This work also addressed the issue of efficiently maintaining the set of learned transformations in the face of database updates and changes in usage patterns. This work was reported in IEEE Transactions on Knowledge and Data Engineering (special issue on discovery in databases) [21].

**Real-time query processing**

Minimization of total response-time is a fundamental problem in any on-line real-time system, including task scheduling for real-time operating systems and real-time heuristic search algorithms. It is believed to be an extremely difficult problem; however, little work had been done to characterize the difficulty. One of our major contributions [22], showed that its difficulty cannot even be characterized by the NP-completeness framework, since the optimization cost is algorithm-dependent. A simplified version of the problem can be analyzed. Another major contribution was to show the need to account for and control scheduling/search costs in order to find and execute the best solution within a given time-constraint. Ours was among the first attempts towards formalizing the difficulty of the response-time minimization problem. We designed two new algorithms, DYNORA and SARTS, which show much better deadline compliance and response-times than previously proposed algorithms. This work resulted in papers in IEEE Transactions on Software Engineering [22], the International Journal of AI Tools (special issue on real-time systems) [23], IJCAI [26], the AAAI National AI Conference [25], and the IEEE Conference on Tools for AI [28, 29], as well as a paper under revision for IEEE Transactions on Knowledge and Data Engineering [24].
Cooperative query processing in Federated Knowledgebases

A primary focus of my work has been on cooperating knowledge-based systems to facilitate interaction among a collection of knowledge-based systems to solve multi-disciplinary problems [?, ?, ?]. A major contribution of the work includes the self-assessment based cooperation model and supporting tools. This was the first work which provided a computational method based on evidential reasoning to decide if an expert system has enough knowledge to solve a given problem or if it needs to consult other expert systems. Another major contribution was the development of the Annotated Prolog language and its interpreter to support evidential reasoning and support logic programming.

Recent contributions include the use of the Dempster-Shafer theory and evidential reasoning for resolving attribute incompatibility in database integration [27], and new models for managing conflict [?, ?].

2.3 Data Mining : Query Transformation Rules, Neural Networks

Our work on learning query-transformation rules has made unique contributions to the area of learning and discovery in databases as well, since the existing learning algorithms from the domain of artificial intelligence are not capable of discovering query transformation rules. We contributed a new learning algorithm to detect interesting and useful data distribution patterns whose closure can provide effective query transformation rules. The algorithm was validated via formal analysis and application to real databases. This work was reported in IEEE Transactions on Knowledge and Data Engineering (special issue on discovery in databases) [21].

Our work on learning query-transformation rules has extended to addressing the crucial issues of performance and generalization. In the performance area, we contributed one of the most scalable parallel formulations of the backpropagation learning algorithm for the neural networks [30, 31]. This implementation achieved one of the best recorded performances, in range of Giga CUPS ($10^9$ connections per second) on CM-5 and nCUBE machines. We also worked on the learning algorithms for Generalization Problems, which are traditionally modeled by statistical regression. We developed a new stochastic learning algorithm [32, 33, 34] for generalization problems which learns the globally optimal connection weights, in contrast to existing learning algorithms, which find sub-optimal weights.

A major contribution of this work was the application of neural network models to the Bond Rating problem [35]. This work pioneered the development of neural network models in the domain of finance and business, triggering panel discussions, workshops, conferences and industrial applications. This paper is among the most cited works in the area. It is used as a case study in the Neural Networks textbook by Khamu, and in articles by Maureen Caudill in AI magazine. It started a lively debate, as evidenced by its review in OR/MS today. It is listed in Scientific Software 1988. It was used as a model paper in the call for papers for the Neural Networks in Business track of the 24th Annual Hawaii Conf. on System Sciences, 1991. It received a best paper award from the Institute of Chartered Computer Professionals (India) in 1991.
2.4 Databases for Intelligent Transportation Systems

I have been working with the Center for Transportation Studies and the Intelligent Vehicle Highway Systems Institute for the last four years. I have also advised the Strategic Research Initiatives group at the Minnesota Department of Transportation on numerous data management issues. This has helped me understand the unique needs of the application domain which were reported in [36, 12].

One of my major contributions is the creation of a state-of-the-art driving simulator consisting of a spatial network management database, a Honda Accord vehicle, a SGI Onyx workstation, and a high resolution color projector. The spatial network management database contributed innovative access methods such as CCAM [14] for road networks and query optimization algorithms for network analysis, which outperform alternative methods in most situations. The simulator is being extensively used in safety research for transportation systems. We designed highly efficient access methods [11] and query processing algorithms [10] for processing collision detection between moving vehicle, and other objects in the driving world.

I recently worked with route evaluation [14] and path optimization [9], two of the most important problems in advanced traveler information systems. Surprisingly, database researchers had not examined the problem of route evaluation. We developed and evaluated efficient access methods and query processing algorithms for such route evaluations [14] We evaluated several algorithms for transitive closure (e.g., iterative), partial transitive closure (e.g., Dijkstra’s) and heuristic search (e.g., A* with the euclidean distance heuristic) for path optimization over the spatial network representing the road-map of Minneapolis [9]. We evaluated the effect of path length and edge cost distribution on the response time of different algorithms. A* outperformed transitive closure and partial transitive closure algorithms for most paths and edge cost distributions. Transitive closure algorithms won only when the path length was comparable to the diameter of the spatial network and edge costs were uniformly distributed.

Our work indicated the need for developing semantic heuristics for path optimization under criteria other than the shortest distance. It also indicated the need for more comprehensive work in evaluating transitive closure algorithms for path optimization. Finally, the work indicated that route evaluation and network overlay are different from transitive closure, and thus need further investigation.
3 Bibliography

References


A Slides Illustrating Recent Accomplishments
Contents

1 Recent Accomplishments ................................................................. 1
  1.1 Survey of Spatial Databases .................................................... 1
  1.2 Data Mining : Mining Spatial Databases for Location Prediction .............. 1
  1.3 Query Processing for spatial predicates ........................................ 2
    1.3.1 Strategies for processing direction predicate based selections ............. 2
    1.3.2 Parallel strategies for processing range queries ............................. 2
  1.4 Storage Methods: Clustering Methods for Spatial Graphs ....................... 3

2 Past Accomplishments .................................................................. 5
  2.1 Storage Methods: Declustering Methods for Spatial Datasets .................. 5
  2.2 Query Processing .................................................................. 5
  2.3 Data Mining : Query Transformation Rules, Neural Networks ................. 6
  2.4 Databases for Intelligent Transportation Systems .............................. 7

3 Bibliography ............................................................................ 8

A Slides Illustrating Recent Accomplishments .................................. 10