

Research Statement

Weili (Lily) Wu

My major research interest is databases, especially data mining and distributed database systems. Advances in technology (e.g. high performance networks, parallel architectures, etc.) and increasing demand for new applications (e.g. data mining, multimedia, etc.) have triggered research in distributed database systems. There are tremendous research issues centered around distributed database systems such as computing performance, data mining, data warehousing, and optimal I/O and minimized communication costs.

My research work spans several areas from spatial data mining to distributed database systems. The following sections discuss my current and future work in each area.

Spatial Data Mining

Knowledge discovery in databases (or data mining) has become an important research area. Data mining is a nontrivial process to extract implicit, previously unknown, and potentially useful information from data in database systems. Geo-spatial data mining [9], a subfield of data mining, is a process to discover interesting and potentially useful spatial patterns embedded in spatial databases. Efficient tools for extracting information from geo-spatial data sets are crucial to organizations which own, generate and manage large geo-spatial data sets. These organizations are spread across many domains including ecology and environmental management, public safety, transportation, public health, business logistics, travel, and tourism.

Challenges in spatial data mining arise from the following issues. First, classical data mining generally works with numbers and categories. In contrast, spatial data is more complex and includes extended objects such as points, lines, and polygons. Appropriate spatial modeling is required for analyzing and mining spatial datasets. Second, classical data mining usually deals with explicit inputs, whereas spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often implicit. Third, classical data mining treats each input as independent of other inputs, whereas spatial patterns often exhibit continuity and high spatial autocorrelation among nearby features. Fourth, modeling spatial context (e.g. autocorrelation) [10] is a key challenge in classification problems that arise in geospatial domains. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships and assuming independence between different data points. However, this approach violates the key property of spatial data, which is spatial autocorrelation. Like temporal data, spatial data values are influenced by values in their immediate vicinity. Ignoring spatial autocorrelation in the modeling process leads to results which are a poor fit and unreliable.

My current work on spatial data mining has focused on predicting location problems. We have proposed PLUMS (Predicting Locations Using Map Similarity) [14], a new

framework for supervised spatial data mining problems. PLUMS searches the space of solutions using a map-similarity measure, which is more appropriate in the context of spatial data. We have shown that compared to state-of-the-art spatial statistics approaches such as the SAR (Spatial Autoregression Model) model [15], PLUMS achieves comparable accuracy but at a fraction of the computational cost. Furthermore, PLUMS provides a general framework for specializing other data mining techniques for mining spatial data. We have also exploited different classification approaches for modeling spatial context (e.g. spatial autocorrelation) in the framework of spatial data mining. We compared and contrasted MRF (Markov Random Fields) and SAR using a common probabilistic framework. MRF [16, 19, 20] is a popular model to incorporate spatial context into image segmentation and land-use classification problems. The SAR model, which is an extension of the classical regression model for incorporating spatial dependence, is popular for prediction and classification of spatial data in regional economics, natural resources, and ecological studies. Our study shows that the SAR model makes more restrictive assumptions about the distribution of features and class shapes (or decision boundaries) than MRF. We also observed an interesting relationship between classical models that do not consider spatial dependence and modern approaches that explicitly model spatial context. The relationship between SAR and MRF is analogous to the relationship between Logistic Regression and Bayesian Classifiers [17]. We have provided theoretical results using a probabilistic framework as well as experimental results validating the comparison between SAR and MRF.

In future work, we will bring other data mining techniques, including clustering and association rules, within the PLUMS framework. We also plan to explore other search algorithms, new map-similarity measures, and non-uniform parameter spaces and determine their dominance zones. We would also like to study and compare other models that consider spatial context in the classification decision process. We would also like to extend the Graph cut [18] solution procedure for SAR.

Distributed Database Systems

A distributed database system contains a distributed database (DDB) and the software to manage the DDB. In a distributed database system, multiple logically related databases can be distributed over a computer network. Many machines can be used to solve one problem and data may be replicated at different sites, and hence improve performance, reliability, availability and extendibility. Applications of distributed database systems include multi-plant manufacturing, a chain of department stores, a bank with many branches, airlines, and military command and control.

Research issues in distributed database systems include distributed database design, query processing, concurrency control, fragment replication, and reliability. Our research work concentrates on the optimization problem of data replicas. Data replication is an important research topic in distributed database systems [8, 11, 12]. Such systems are built in a computer network with a certain topological structure. Multiple copies of data

replicas are placed at different locations to achieve high data availability in the presence of possible network link failures. The best placement of fragments for a given number of data replicas in a computer network to minimize query response time and maximize throughput is a very difficult problem and has been studied extensively in the literature for various protocols [5, 13]. In [12], we presented a sufficient and necessary condition for the optimality of a placement of an odd number of data replicas in a ring network with a majority voting protocol. As a corollary, we gave theoretical proof of a recent conjecture of Hu et. al in [6] that uniformly distributing placement is optimal. We also presented a simple and efficient algorithm to find optimal placements in a tree network with a majority voting protocol. However, when the number of replicas is even, it was pointed out by [6] with a counterexample that uniformly distributing placement may not be optimal and that the optimal placement may have a very complicated pattern. In [3], we explored the even more complicated case in which the number of replicas is $2t$ where t is an odd number. We presented and proved that in this case, an optimal placement can be obtained by first, dividing $2t$ replicas into t pairs and then uniformly distributing t pairs in the ring network while each pair is placed at two adjacent locations.

Many consequences and further investigations from my works remain to be exploited in future research. In ring networks, we determined the optimal placement for odd k and a special case of even k where $k = 4m + 2$. However, the structure of optimal placement in this special case seems unable to be generalized to $k = 2t$ for $t = 2m + 1$. Different applications have various requirements regarding data concurrency control. To meet these different requirements, there exist many database management protocols in the literature [1, 2]. This suggests many possibilities for future research. Many applications of distributed database systems do not require strict consistency of data replicas [7]. Recently, Olston and Widom [11] designed a new replication system, TRAPP, which provides each user a choice of tradeoff in replication precision and performance. This system has also generated many interesting research problems for my future study.

References:

- [1] P.A. Bernstein and N. Goodman, Concurrency control in distributed database system, in M. Stonebraker (ed.), *Reading in Database Systems* (Morgan Kaufmann Publishers, 1994) 548-566.
- [2] P.A. Bernstein, V. Hadzilacos, and N. Goodman, *Concurrency Control and Recovery in Database System*, (Addison-Wesley, 1987).
- [3] Weili Wu and Shashi Shekhar, Optimal locations in ring network for data replicas in distributed database with majority voting protocol, Submitted to *Journal of Parallel and Distributed Computing*
- [4] H.A. Eiselt, M. Gendreau, and G. Laporte, Location of facilities on a network subject to a single edge failure, *Networks* 22 (1992) 231-246.
- [5] A. El Abbadi, D. Skeen, and F. Cristian, An efficient, Fault-Tolerant protocol for replicated data management, in M. Stonebraker (ed.) *Reading in Database Systems*(Morgan Kaufmann Publishers, 1994) 577-591.

- [6] X. Hu, X. Jia, H. Huang, and D. Li, Placement of data replicas for optimal data availability in ring networks, to appear in *Journal of Parallel and Distributed Computing*.
- [7] J. Gray, P. Helland, P. O'Neil, and D. Shasha, The dangers of replication and a solution, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, 1996, pp.173-182.
- [8] D.B. Johnson and L. Taab, Complexity of network reliability and optimal resource placement problems, *SIAM J. Computing* 23 (1994) 510-519.
- [9] Sanjay Chawla, Shashi Shekhar, Weili Wu and Uygur Ozesmi, Extending Data Mining for Spatial Applications: A Case Study in Predicting Nest Locations, *Proceedings of 2000 ACM/SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000)*.
- [10] N.A. Cressie, *Statistics for Spatial Data (Revised Edition)*, Wiley, New York, 1993.
- [11] Olston, C., Widom, J., Offering a Precision-Performance Tradeoff for Aggregation Queries over Replicated Data *Twenty-Sixth International Conference on Very Large Data Bases (VLDB 2000)*, Cairo, Egypt, September 2000.
- [12] S. Shekhar and W. Wu, Optimal placement of data replicas in distributed databases with majority voting protocol, *Theoretical Computer Science*, 258 (2001) pp. 555-571.
- [13] A.B. Stephens, Y. Yesha, and K. Humenik, Optimal allocation for partially replicated database systems on tree networks, *Appl. Math. Lett.*, 8 (1995) pp. 71-76.
- [14] Sanjay Chawla, Shashi Shekhar, Weili Wu and Uygur Ozesmi, Predicting Locations Using Map Similarity (PLUMS): A new approach for supervised spatial data mining. *Proceedings of KDD-2000 Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*
- [15] J. LeSage, Regression Analysis of Spatial data, *The Journal of Regional Analysis and Policy* (Publisher: Mid-Continent Regional Science Association and UNL College of Business Administration), 27 (2): 83-94, 1997.
- [16] S. Li, Markov Random Field Modeling, *Computer Vision* (Publisher: pringer Verlag, 1995).
- [17] C. E. Warrender and M. F. Augusteijn, Fusion of image classifications using Bayesian techniques with Markov rand fields, *International Journal of Remote Sensing*, 20 (10), pp.1987-2002, 1999.
- [18]. Y. Boykov and O. Veksler and R. Zabih, Fast Approximate Energy Minimization via Graph Cuts, *International Conference on Computer Vision*, September 1999.
- [19] P.B. Chou and P.R. Cooper and M. J. Swain and C.M. Brown and L.E. Wixson, Probabilistic network inference for cooperative high and low levell vision. *In Markov Random Field, Theory and Applications* (Academic Press, New York, 1993).
- [20] S. Geman and D. Geman, Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), pp.721-741, 1984.