Efficient Data Dissemination Using Locale Covers

Sandeep Gupta, Jinfeng Ni, and Chinya V. Ravishankar
Dept. of Computer Science and Engineering, University of California, Riverside
Riverside, CA - 92521, USA
{sandeep, jni, ravi}@cs.ucr.edu

Categories and Subject Descriptiors: H.2.8[Database Applications]:Spatial databases and GIS;

General Terms: Algorithms, Managment, Performance

Keywords: Location-dependant Data, Locale Covers, Dissemination, Accuracy, Latency

ABSTRACT

Location-dependent data are central to many emerging applications, ranging from traffic information services to sensor networks. The standard pull- and push-based data dissemination models become unworkable since the data volumes and number of clients are high.

We address this problem using *locale covers*, a subset of the original set of locations of interest, chosen to include at least one location in a suitably defined neighborhood of any client. Since location-dependent values are highly correlated with location, a query can be answered using a location close to the query point.

We show that location-dependent queries may be answered satisfactorily using locale covers, with small loss of accuracy. Our approach is independent of locations and speeds of clients, and is applicable to mobile clients.

1. INTRODUCTION

The growth of small and portable wireless computing devices has led to growing interest in location-dependent information services ranging from traffic information services to sensor networks. Mobile clients need data relevant to their current location. Unfortunately, a pull-based dissemination model is unworkable, since the typically huge number of clients would overload the server with pull requests. Push-based models such as broadcasting [6] fail because data volumes tend to be very large, so a client may have to wait a long time before data of interest appears in the broadcast.

Fortunately, values of location-dependent data are highly correlated with location. For example, temperature readings from adjacent sensors are similar. A client is therefore likely to be satisfied with data for a location sufficiently close to the query point. We develop the notion of a *locale cover*, a subset of the original set of locations, which includes at least one location in a suitably defined neighborhood of any client. We show that location-dependent queries may be answered satisfactorily from only this fraction of data, sacrificing accuracy by only a small amount. Our algorithms are independent of client locations and speeds, so our approach is very applicable to mobile clients.

Copyright is held by the author/owner.
CIKM'05, October 31 – November 5, 2005, Bremen, Germany
ACM 1-59593-140-6/05/0010.

1.1 Challenges

- The volume of data is to be broadcasted is large and therefore results in excessive bandwidth consumption and wastage of battery power at client terminal.
- Tracking each client's location may require high communication overhead, or even be impossible in applications such the satellite-based service Sirius [8], where bidirectional communication is not be available. Consequently, scheduling-based approaches such as [9, 1, 5, 2] are infeasible in this environment, since they assume knowledge of the access patterns or requests of each mobile client.
- Broadcasting data in arbitrary order would cause the access latency to be as long as the full broadcast cycle.
- Pull allows servers to remain stateless, but this model is inappropriate for our application scenarios. The push model is better, but requires servers to be stateful, and to keep track of what data is to be sent to which client, and at what time and therefore does not scale well.

1.2 Our Approach: Locale Covers

Central to our approach is the notion of locale cover, an idea interesting in its own right.

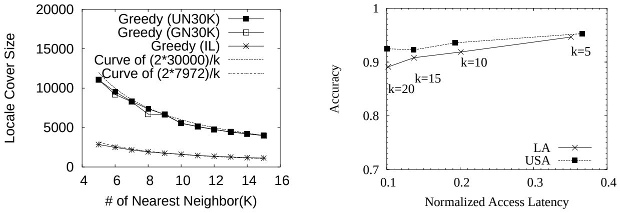
If P is some predicate defining closeness, the P*-locale* of a point $p \in \mathbb{R}^2$ is the set of all points q so that $P(p,q)$ holds. Given a region $R \subset \mathbb{R}^2$, a set of "sites" $X = \{s_1, s_2, \cdots, s_n\} \subseteq R$, and a proximity predicate P, the P*-locale cover* of R with respect to X is a subset $L \subseteq X$, so that, for any $p \in R$, $P(p,s)$ holds for at least one site $s \in L$. Thus, L covers all possible locales in R.

Definition The *compass* $\widehat{R}$ of a region R is the set of sites whose locations are in R or on its boundary.

Definition A k*-domain* D is a disk with $|\widehat{D}| = k$. That is, its compass has cardinality k.

Definition k*-domain locale cover:* $L \subseteq X$ is a k-domain locale cover if $\widehat{D} \cap L \neq \emptyset$ for all k-domains D.

Our algorithm for finding k-domain locale cover runs in two phases. In the first phase, it generates $\mathcal{F}$, where $\mathcal{F} = \{F \mid F \subseteq X, |F| = k$, and there is a disk $D \subseteq \mathbb{R}^2$ with $\widehat{D} = F\}$. The second phase computes the hitting set for $\mathcal{F}$. By theorem 1.1 the hitting set is the locale cover for $\mathcal{F}$.



(a) Sizes vs. k.

(b) Accuracy vs. latency.

Figure 1: k-domain locale covers.

THEOREM 1.1. *Let H be the hitting set for $\mathcal{F}$. For any disk D in $\mathbb{R}^2$ such that $\widehat{D} \geq k$, $H \cap \widehat{D} \neq \emptyset$.*

Further, we prove an elegant relationship (see theorem 1.2) between $\mathcal{F}$ and order-k Voronoi diagrams, so that that $\mathcal{F}$ can be computed in expected cost $O(nk^2 + nk \log^2 n)$.

THEOREM 1.2. *There is a bijection between elements $F \in \mathcal{F}$ and cells in the order-k Voronoi diagram for X.*

We can infer that the size of $\mathcal{F}$ is reasonable, and contains $O(nk)$ subsets, since the order-k Voronoi diagram has $O(nk)$ Voronoi cells [3]. Several algorithms have been proposed to compute order-k Voronoi diagrams. (See the survey by Boissonnat [4] and references therein.) Once $\mathcal{F}$ is obtained, the greedy approach for finding hitting set is used to obtain k-domain Locale cover.

2. PERFORMANCE EVALUATION

This section presents the experimental results for the proposed technique of locale cover, using both synthetic and real datasets. Synthetic datasets, UN30K and GN30K, have 30,000 sites distributed in a unit square in random and Gaussian respectively. Real dataset IL represents 7972 road intersections in the state of Illinois. Dataset LA, obtained from [7], contains 627 traffic stations monitoring traffic speed on the freeways across Los Angeles County. All the site in dataset LA has average speed $\mu = 56.94$, standard deviation $\sigma = 16.66$. Dataset USA, obtained from http://weather.noaa.gov, contains 1,500 weather stations monitoring the temperature in USA. All the weather stations has average temperature $\mu = 28.24$, standard deviation $\sigma = 4.36$.

2.1 Size of k-Domain Locale Cover

Our first set of experiments evaluated the size of locale covers varying the value of k in the k-domain locale coverage problem. Figure 1(a) shows the sizes of locale covers for different k, for datasets UN30K, GN30k and IL. In these figures, the dashed line represents the curve of $\frac{2n}{k}$, for purposes of comparison. Clearly, we are able to obtain locale covers of approximately $\frac{2n}{k}$ for all the datasets, regardless of distribution.

2.2 Accuracy vs. Access Latency

Clearly, there is a trade-off between access latency and accuracy in using k-domain locale covers. As k increases, locale cover size decreases, lowering access latency. On the other hand, a smaller locale cover also lowers accuracy. The following experiment studies the trade-off between access latency and accuracy using k-domain locale cover. The experiment is conducted over traffic speed dataset LA, and the temperature dataset USA.

Given a query point q, let v be the reading (speed or temperature) at q's nearest neighbor from the original set of sites, and $\hat{v}$ be the reading at q's nearest neighbor from the locale cover. Accuracy is defined as $1 - \frac{|v - \hat{v}|}{v}$. We used 10000 query points randomly distributed in the region, and computed the average relative error.

We compare the access latency of broadcasting with k-domain locale covers with that of broadcasting the original dataset X. We measure the average access latency of k-domain locale cover L as $|L|/2$, and further normalize it to the optimal access latency $|X|/2$ of the original dataset X [6, 10].

Figure 1(b) plots the trade-off between accuracy and latency. As expected, as k increases from 5 to 20, both the access latency and accuracy decrease. Our approach using locale cover reduces access latency by an impressive 65%–90%, at a modest loss of accuracy (about 10%). Spatial correlation allows our technique to choose a small locale cover, which suffices for high accurate answer, with a significantly reduced access latency.

3. CONCLUSIONS

We introduce the notion of locale cover, and present several novel formulations and variants of the data dissemination problem for location-dependent data in broadcasting environments. Our schemes choose a small subset of sites that include a site in the neighborhood of all clients, regardless of their number or distribution. This method significantly reduce broadcast bandwidth and access latencies for clients, and scales well with the number of users and sites. Our experiments confirm the applicability and efficiency of our schemes.

Our notion of locale cover is very general and is likely to be applicable beyond the domain of data dissemination, for example, in spatio-temporal data mining, and approximate indices. We intend to investigate these possibilities in future work.

4. ACKNOWLEDGEMENTS

This work was supported in part by grants from the Tata Consultancy Services, Inc., the Digital Media Innovations Lab of the University of California and also by the DARPA award FTN F30602-01-2-0535.

5. REFERENCES

[1] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik. Broadcast disks: Data management for asymmetric communication environments. In *Proceedings of SIGMOD*, pages 199–210, 1995.

[2] D. Aksoy and M. Franklin. Scheduling for large-scale on-demand data broadcasting. In *Proceedings of IEEE INFOCOM*, pages 651–659, 1998.

[3] F. Aurenhammer and O. Schwarzkopf. A simple on-line randomized incremental algorithms for computing higher order voronoi diagrams. In *Proceedings of ACM Symp. on Computational Geometry*, pages 142–151, North Conway, New Hamsphire, US, June 1991.

[4] J.-D. Boissonnat, O. Devillers, and M. Teillaud. A semi-dynamic construction of higher order voronoi diagrams and its randomized analysis. *Algorithmica*, 9(4):329–356, July 1993.

[5] Q. Hu, D.-L. Lee, and W.-C. Lee. Dynamic data delivery in wireless communications environments. In *Proceedings of Workshop on Mobile Data Access*, pages 213–224, 1998.

[6] T. Imielinski, S. Viswanathan, and B. Badrinath. Data on air: Organization and access. *IEEE TKDE*, 9(3):353–372, May 1997.

[7] PeMS. Freeway performance measurement system. http://pems.eecs.berkeley.edu/Public/.

[8] SIRIUS. Sirius radio. http://www.siriusradio.com/.

[9] N. H. Vaidya and S. Hameed. Scheduling data broadcast in asymmetric communication environments. *Wireless Networks*, 5:171–182, 1999.

[10] B. Zheng, W.-C. Lee, and D. L. Lee. Search k nearest neighbors on air. In *4th International Conference On Mobile Data Management*, pages 181–195, Melbourne, Australia, January 2003.