



Efficient Acquisition of Web Data through Restricted Query Interfaces

Simon Byers Claudio Silva
AT&T Labs



Juliana Freire
Bell Labs

Lucent Technologies
Bell Labs Innovations



The Hidden Web

QUICK SEARCH FOR PROPERTIES

Choose Property: Enter Price Range From: ,000
 Type: To: ,000

Special Markets: All Luxury Resort Military University

Enter ZIP Code: Enter a City and select a State: Outside USA:

SEARCH FOR PROPERTIES VIA MULTIPLE LISTING NUMBERS

Enter a City: Select a State:
 And a ML Number:
 (All three fields are required)

Find the nearest Jiffy Lube Service Centers

Address:
 City:
 State:
 or
 ZIP Code:
 Find stores within: miles

 If using Address, include City/State or Zipcode.

Travelocity.com
 A Sabre Company

Home Dream Plan Go Flights Lodging Cars Vacations Cruises Deals & Rewards My Stuff

Round Trip / One Way Multiple Destinations Flight Arrival / Departure Info Dream Maps Deals

Search for flights and fares

1 Where would you like to go?
 From: Search for the closest airport
 To: Add more destinations
 Round Trip One Way

2 When do you prefer to travel?
 I need to choose specific dates and/or times
 Show me the best-priced trips on my dates [within 6+ hrs] of the times selected below
 I want to choose specific flights on my dates

Depart: Apr 27
 Return: May 6

My dates are flexible
 Show me the best economy fares for travel during the period:
 Starting: Through the end of:

3 How many travelers are there? Anyone under age 18 traveling alone?
 Total number of travelers:
 How many are aged 2 thru 11?

4 How would you like us to price this trip? Tip to find low fares
 Economy class with restrictions
 Economy class without restrictions
 Business class
 First class

5 What is your preferred maximum number of connections?

Dealer Locator

ADDRESS OR CROSS STREETS

 (e.g., 1000 North First Street or Main & Pine St.)

CITY / STATE OR ZIP CODE

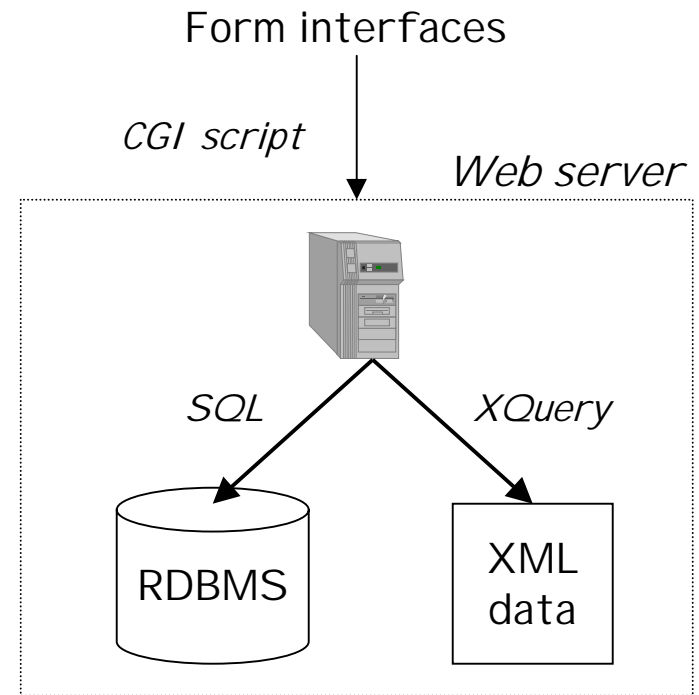
 (e.g., Torrance, CA or 90501)

OR DEALER

 (e.g., your Honda dealer)

Web Databases

- ◆ 80% of Web data is accessible through *limited* query interfaces
- ◆ Form interfaces provide a **simple** way to query and filter data, but...
- ◆ they can be **too restrictive**
- ◆ Too many queries may be needed to retrieve the desired information, e.g., find me all the stores in NJ, and
- ◆ it may take too long



Materializing Web Databases

- ◆ Enable data exploration - richer queries
- ◆ Improve performance
- ◆ Useful in a number of applications:
 - Comparison shopping services, job search sites, etc
- ◆ Back-door approach: content provider may create a *special*, more powerful interface
- ◆ What if the content provider does not cooperate?
 - E.g., vendor A wants to track the expansion strategy of vendor B

How to *scan* the database through the limited interface?

How to do that efficiently?

Some simple useful strategies

- ◆ Pose the most general queries allowed by the interface to minimize the number of queries:
 - leave optional attributes unspecified
 - choose most general values for obligatory attributes (e.g., all property types)
 - given a choice of obligatory attributes, select the one with smallest domain (e.g., state instead of zipcode)

Not always enough to guarantee full coverage...

Generating covers of hidden databases accessible through NN-query interfaces

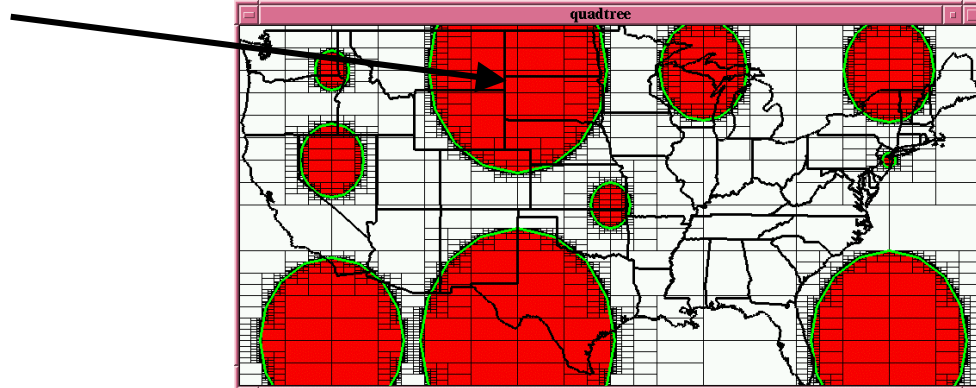
- ◆ NN-query: *find the 10 closest stores to zipcode 07974*
- ◆ Naïve strategy: pose one query per zipcode
 - *over 10,000 queries*
 - *it takes from hours to days to retrieve all the data*

Is it possible to find a smaller set of zipcodes that return the same answers?

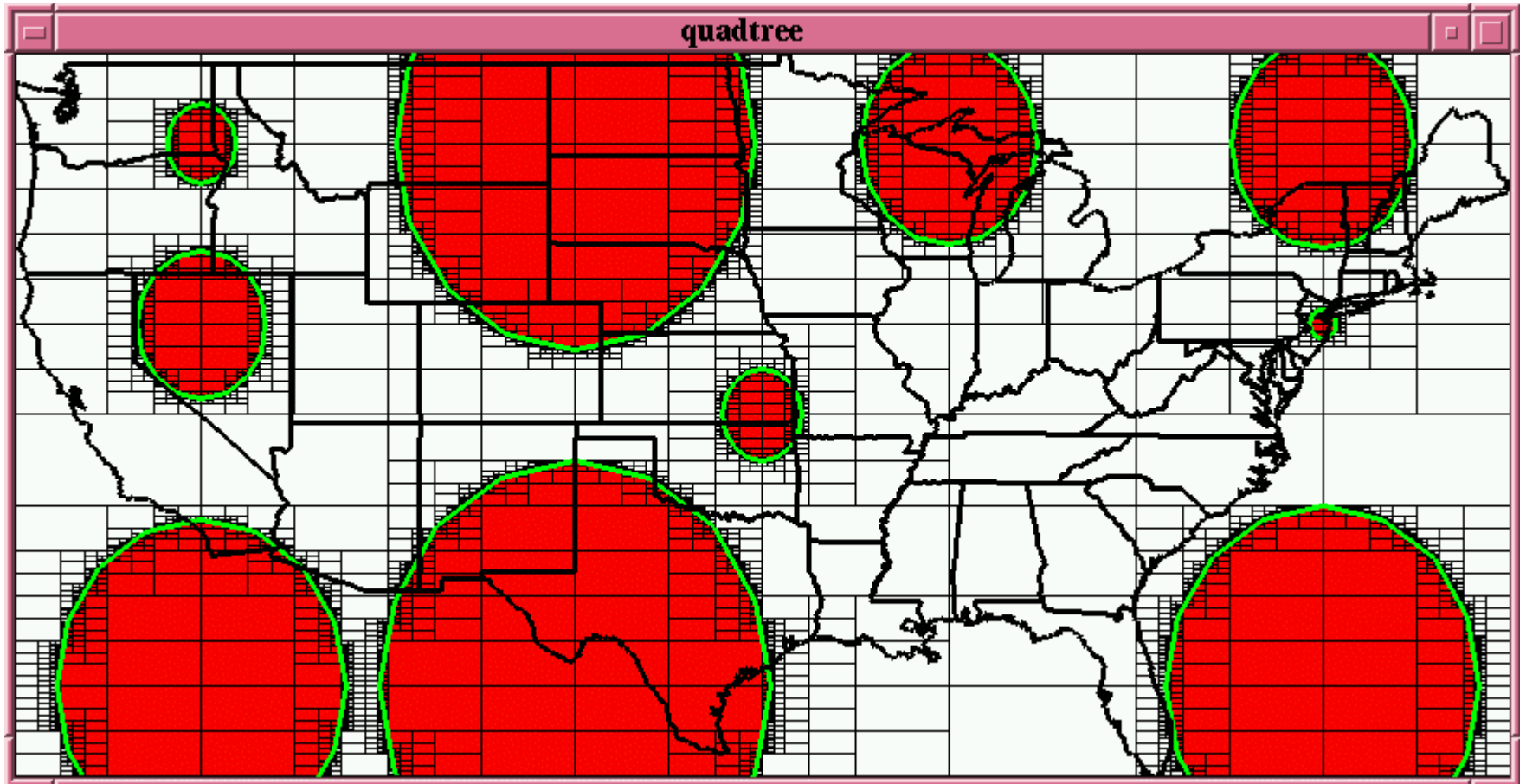
Algorithm

- ◆ Our algorithm is quite simple and is composed of two parts:
 - We use a spatial data structure to keep track of which parts of a region R that have already been covered by previous queries.
 - At any given point in time, we use the coverage information obtained thus far to determine where to perform the next query as to minimize the overlap of queries.

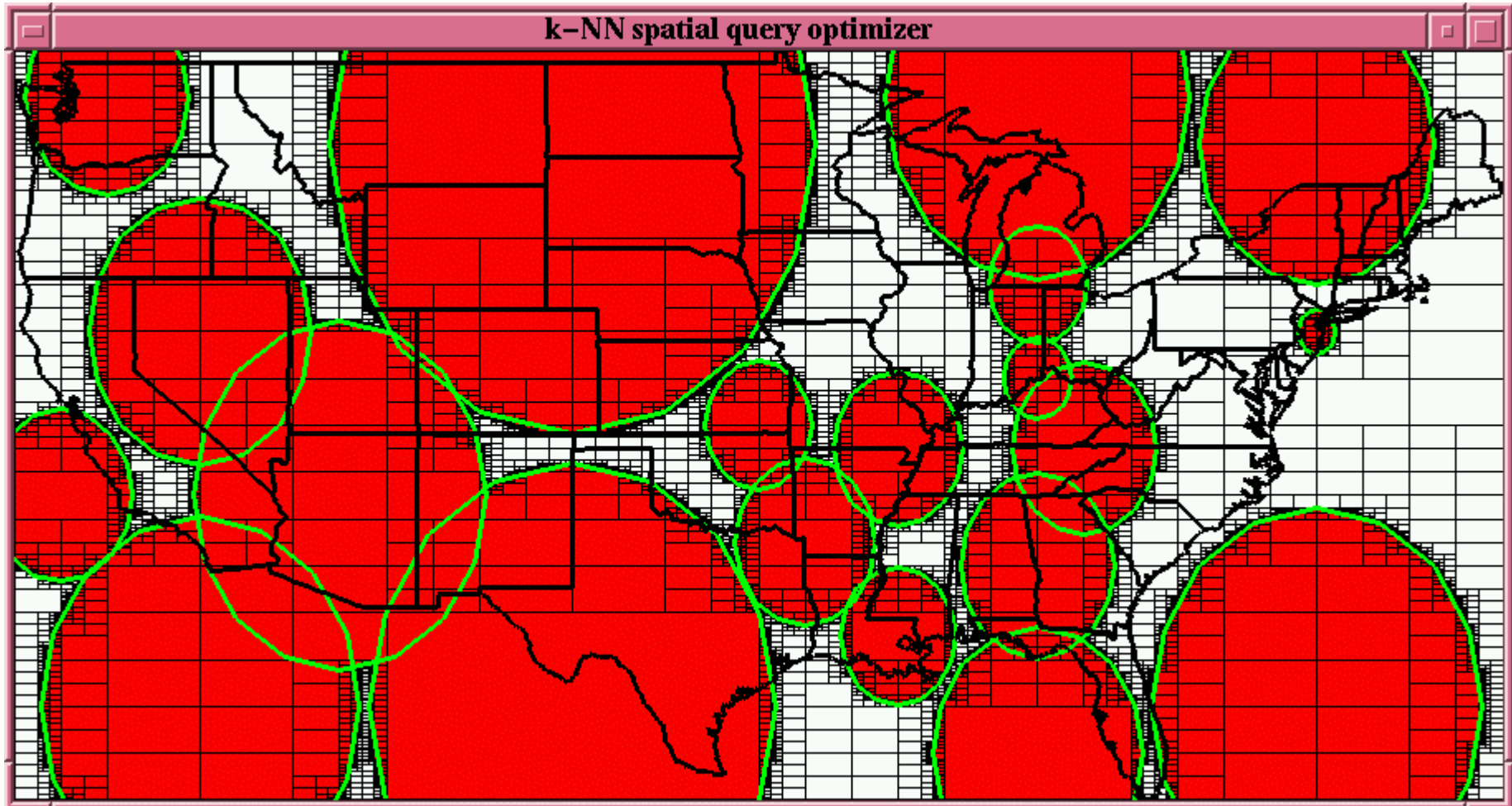
Query



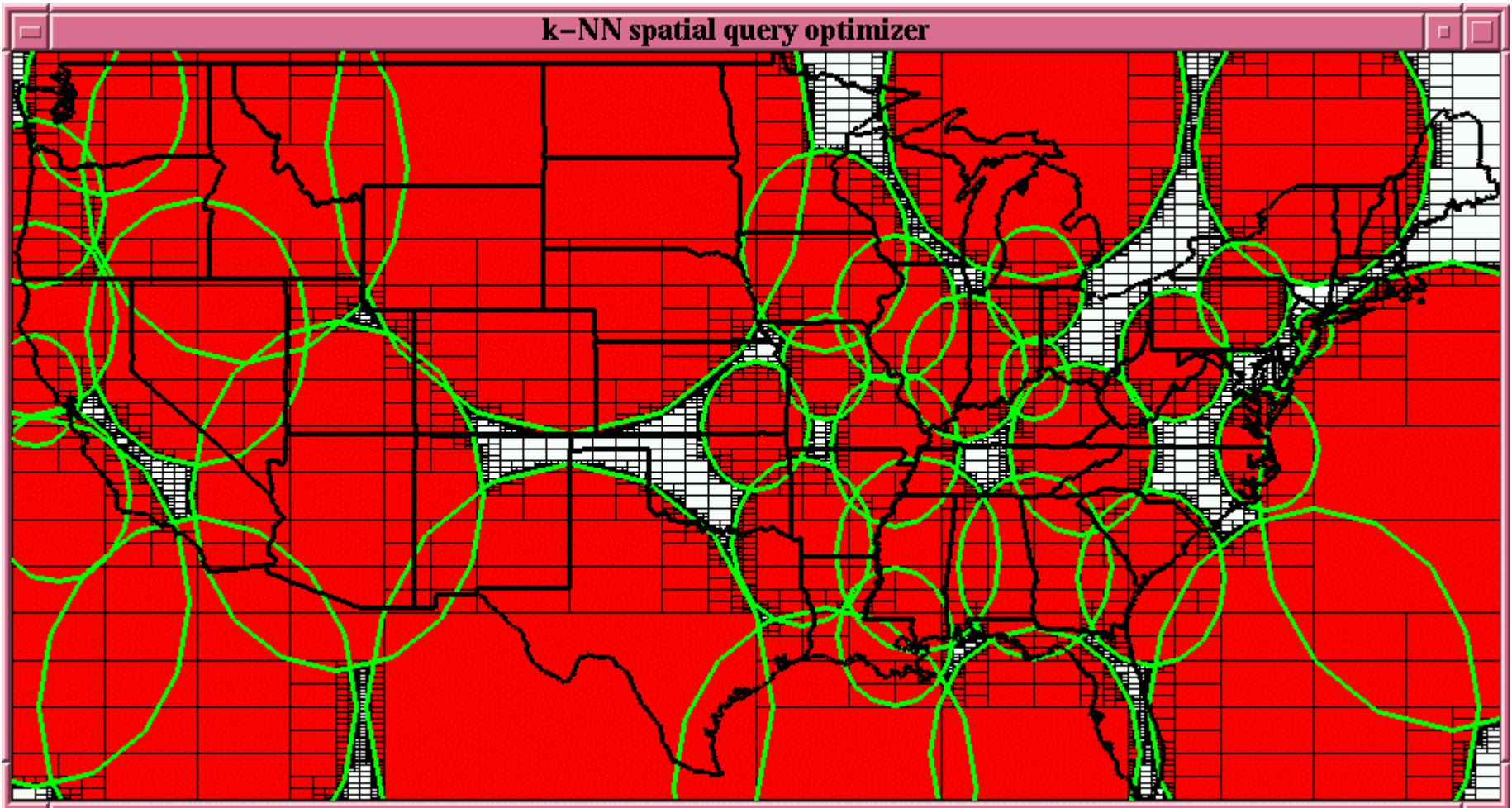
Coverage (snapshot I)



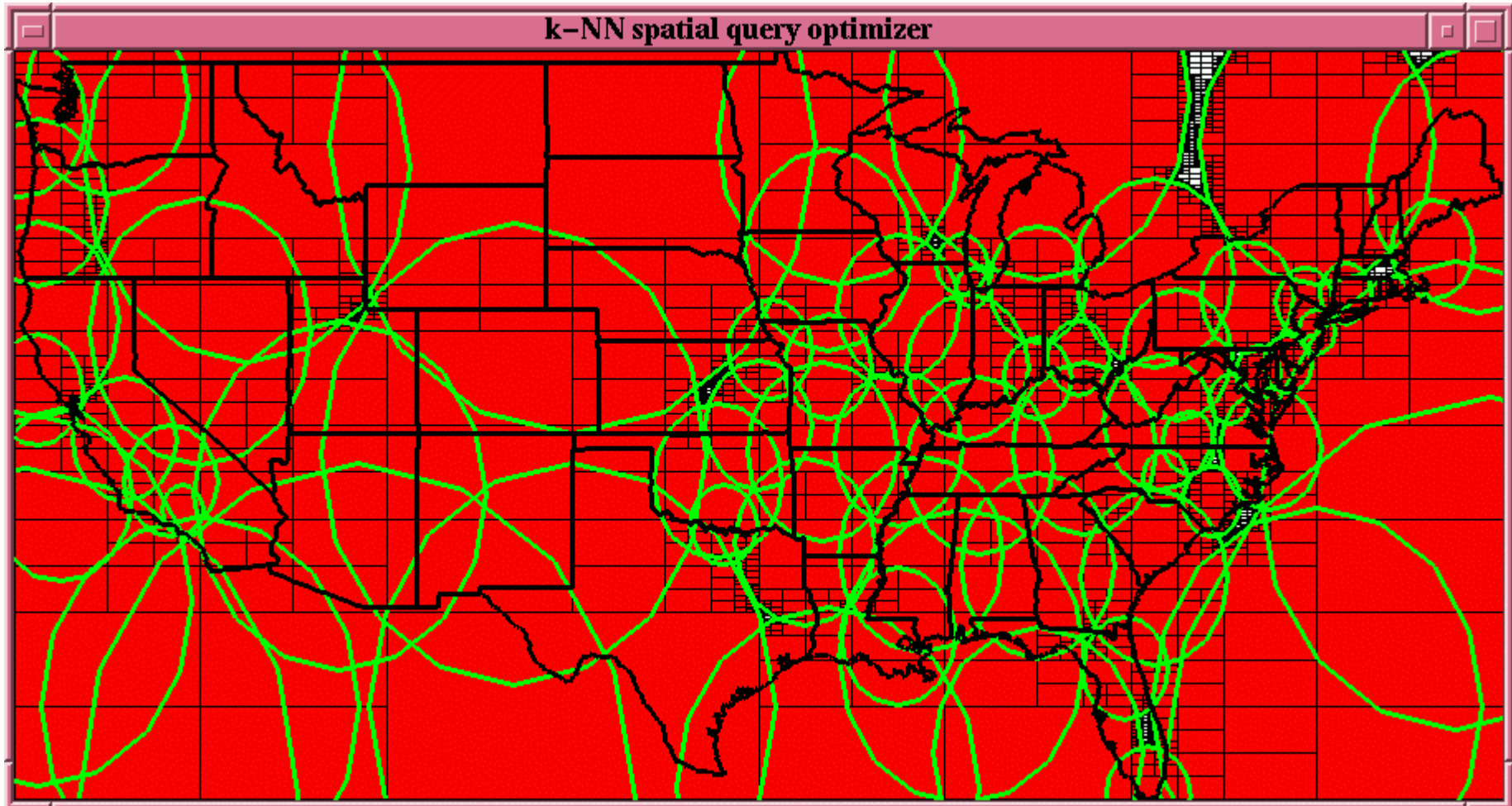
Coverage (snapshot II)



Coverage (snapshot III)



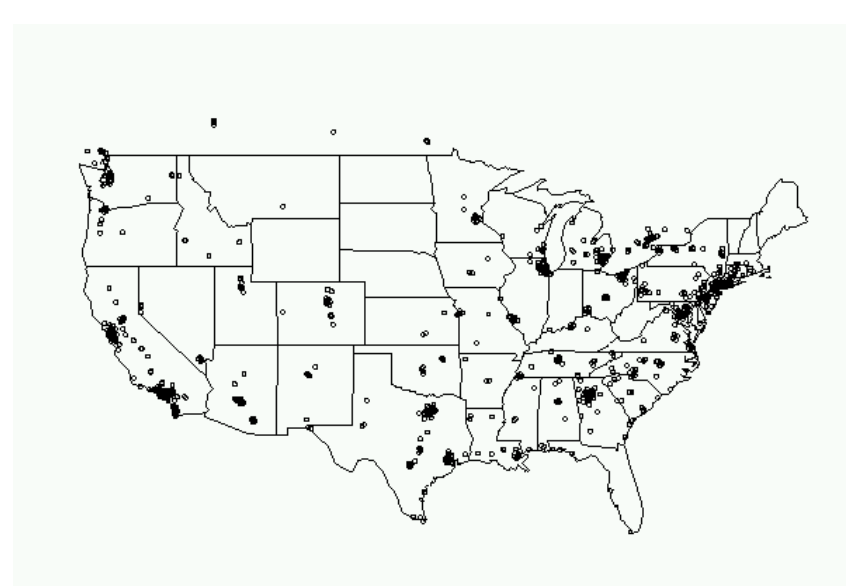
Coverage (snapshot I V)



Datasets



Dataset 1



Dataset 2

Acquisition by querying zip codes takes over 10,000 queries!

Experimental results

Let $\text{QUAD}(\mathcal{D})$ be the number of queries performed by our technique, where in general $\text{OPT}(\mathcal{D}) \leq \text{QUAD}(\mathcal{D})$. We can define an approximation factor, $\rho(\mathcal{D})$, to be the ratio between our algorithm and the optimum, that is,

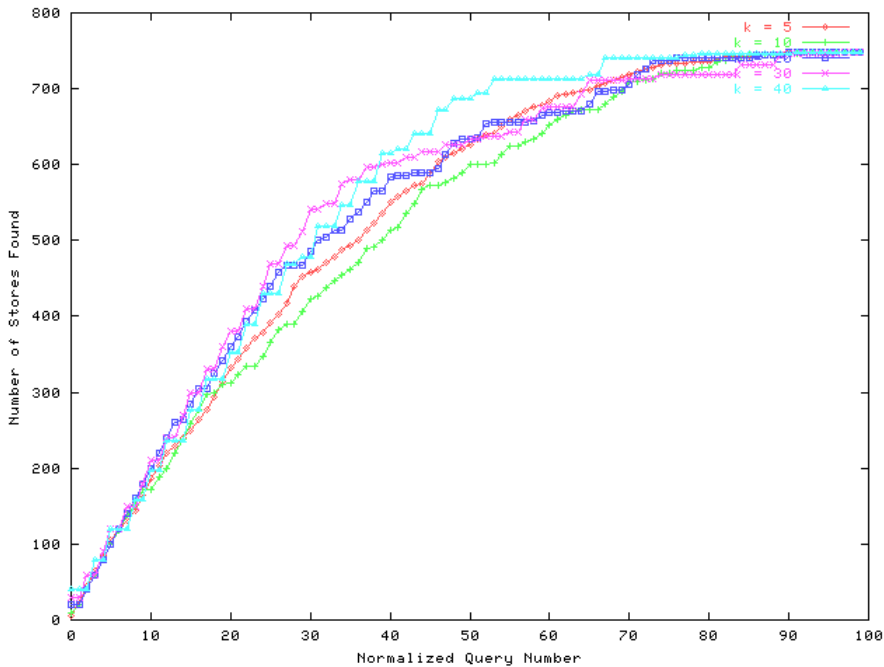
$$\rho(\mathcal{D}) = \frac{\text{QUAD}(\mathcal{D})}{\text{OPT}(\mathcal{D})}. \quad (1)$$

Note that by definition $\rho(\mathcal{D}) \geq 1$; one being the best possible query schedule for retrieving all sites with a k -NN query interface.

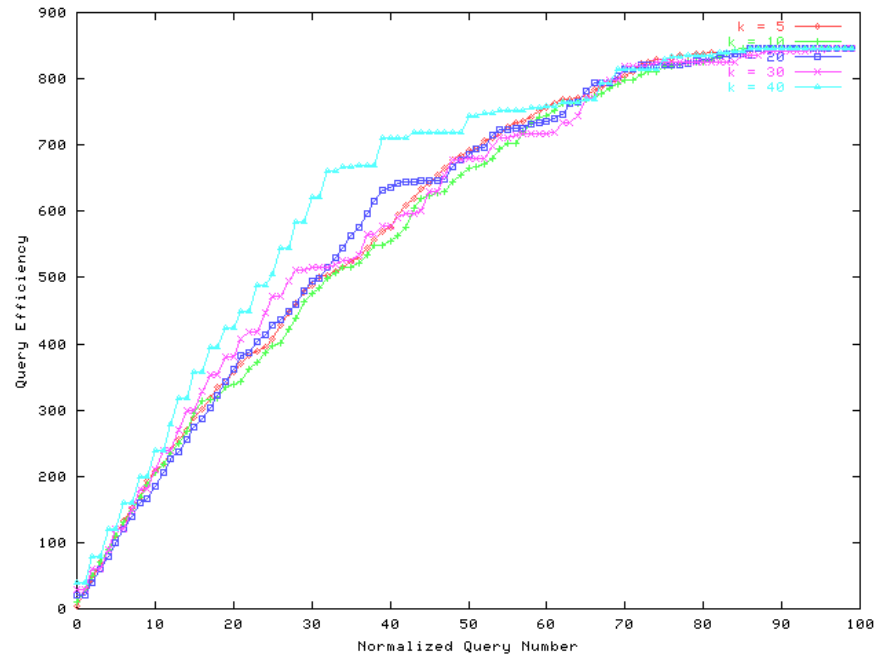
	$k = 5$		$k = 10$		$k = 20$		$k = 30$		$k = 40$	
	QUAD	ρ	QUAD	ρ	QUAD	ρ	QUAD	ρ	QUAD	ρ
Dataset 1	411	2.7	191	2.6	94	2.5	60	2.4	42	2.2
Dataset 2	435	2.6	207	2.4	99	2.3	65	2.3	54	2.6

Table 1: Query performance with varying k . In the table, we show: (1) **QUAD**, the number of queries necessary to find all the sites in a given dataset; (2) ρ , the approximation factor of the queries. See text for details.

Experimental results (cont.)



Dataset 1



Dataset 2

Acquiring Web Data

- ◆ Automate Web navigation and data retrieval
 - WebVCR, W4F, Perl scripts
- ◆ Optimize the execution of Web queries
 - many queries may be required for a full scan over the hidden database
 - optimize wrapper (access and extraction)
 - exploit parallelism, smart source selection, re-ordering, etc
- ◆ Disguise requests
 - anonymizing proxies, random time intervals,

Related Work

- ◆ Mediators and restrictive query interfaces (e.g., Information Manifold, TSI MMI S, Web Integrator)
 - previous works did not consider the *coverage problem*, or nearest-neighbor interfaces
- ◆ Xyleme project from INRIA: build a world-wide XML warehouse
 - considers only documents that can be reached through traditional crawling - blind to the hidden Web
- ◆ Optimization of Web queries
 - query scrambling, adaptive techniques
- ◆ Wrapper creation
 - WebVCR, W4F, NoDoSe, Ariadne

Full technical report available at

<http://www.research.att.com/~csilva/papers>