Automatic Example Queries for Ad Hoc Databases

Bill Howe¹, Garret Cole², Nodira Khoussainova³, Leilani Battle⁴ {¹billhowe, ²gbc3, ³nodira, ⁴leibatt}@cs.washington.edu, University of Washington, Seattle, WA, USA



"Here is my data. Where do I start?"



- Unknown relationships, semantics, utility
- Temporary, time-sensitive applications

An ad hoc database is a collection of tables with unknown relationships gathered to serve a specific, often transient, often urgent, purpose.

Examples of Ad Hoc Databases

- A researcher assembles an ad hoc database of recent experimental results to prepare a paper or proposal.
- Emergency workers responding to a natural disaster assemble an ad hoc data-base from lists of addresses of nearby schools, locations of resources (e.g., ambulances), and contact information for emergency workers.
- A consulting business analyst assembles an ad hoc database from a set of spreadsheets provided by management for a short term engagement
- A security analyst assembles an ad hoc database from a set of application trace logs after an attack

Approach

SQLShare: Database-as-a-Service for Ad Hoc Data

- **1. Model** each operator independently using curated sts of example queries from the web
- **2. Compose** operators to generate a search space of example queries
- **3. Rank** each set using scores derived from configurable patterns called *idioms*
- Q: Are users willing and able to write SQL?
- A: Yes! But they need access to high-quality examples (c.f. Gray, Szalay et al. 2005; Howe 2010)

	RE	1 🔻 Natur	re Mapping Speci x	SDS joined with Seaf	fl ×			Lo	igged in: po	ckylan@wa	shington.ed
Your datasets		SDS joine	ed with Seaflow da	ta on time, 10-min	ute intervals <\vi	ewable by everyone	L	ast modified: Jan 12, 2011	2:30 AM 🔒	billhowe@wa	ashington.edu
All datasets		Data from R/V Thompson flowthrough system joined with seaflow data averaged over 10-minute									
Shared datasets		intervals.									
avorites		Click here to add a tag									
Recently viewed	»										
Jpload dataset New query Recent activity	0	SELECT FROM , WHERE ORDER	[billhowe].[sds [billhowe].[seaf sds.binid = seaf BY sds.binid	Thompson-lb.csv, low-Thompson_l-st low.binid	binned at 10-min ats, binned on 1	ute intervals] s O-minute interval	ds ls] seaflow				
POPULAR TAGS coffee salary Jw	3 2 2	DATASET PRE	EVIEW (Rows 1 - 100 of 74	7) □ next > last >>					Derive	dataset	Download
POPULAR TAGS coffee salary Jw gps	3 2 2 2	DATASET PRE	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5	7) <u>next > last >></u>					Derive (dataset	Download
POPULAR TAGS coffee salary Jw gps excel	3 2 2 2 2	DATASET PRE	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5 lat	7)] <u>next > last >></u> Ion	salinity	ocean_temp	fluorescence	transmission	Derive of binid	dataset I	Download flow
POPULAR TAGS coffee salary JW gps excel sample	3 2 2 2 2 2 2	DATASET PRE	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5 lat 50.1174993515015	7) <u>next> last>></u> lon -137.5	salinity 32.4525	ocean_temp 7.5375	fluorescence	transmission 3.5225	Derive of binid 5/5/2010 2:19:59	dataset I	Download flow 2.3300(
POPULAR TAGS coffee salary w gps sexcel sample swetha	3 2 2 2 2 2 2 2	DATASET PRE << first < pr binid 5/5/2010 2:19:59 AM	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5 lat 50.1174993515015	7) <u>next> last>></u> lon -137.5	salinity 32.4525	ocean_temp 7.5375	fluorescence 0.18	transmission 3.5225	Derive of binid 5/5/2010 2:19:59 AM	pop beads	Download flow 2.3300(
POPULAR TAGS coffee salary Jw gps axcel sample swetha poverage	3 2 2 2 2 2 2 2 1	DATASET PRE << first < pr binid 5/5/2010 2:19:59 AM 5/5/2010 2:10:50	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5 lat 50.1174993515015	7) <u>next > last >></u> lon -137.5	salinity 32.4525	ocean_temp 7.5375	fluorescence 0.18	transmission 3.5225	Derive of binid 5/5/2010 2:19:59 AM 5/5/2010 2:40:50	pop beads	flow 2.33000
POPULAR TAGS coffee salary uw gps excel sample sample swetha poverage cse	3 2 2 2 2 2 2 1 1	DATASET PRE << first < pr binid 5/5/2010 2:19:59 AM 5/5/2010 2:19:59 AM	EVIEW (Rows 1 - 100 of 74 rev 1 2 3 4 5 lat 50.1174993515015 50.1174993515015	7) <u>next > last >></u> lon -137.5 -137.5	salinity 32.4525 32.4525	ocean_temp 7.5375 7.5375	fluorescence 0.18 0.18	transmission 3.5225 3.5225	Derive of binid 5/5/2010 2:19:59 AM 5/5/2010 2:19:59 AM	pop beads diatoms	Download

- 1. Streamlined for a single workflow:
- 2. No DDL; schema inferred from data
- 3. Views as first-class citizens
- 4. Unfettered sharing; cloud-hosted
- 5. Full SQL; no restrictions

http://sqlshare.escience.washignton.edu





Query

Modeling Each Operator

Join

Finding: "Good" joins characterized by linear relationships among a handful of set properties

	-
Feature	Expression
/ • 1• 1•	

Project

Finding: Important attributes appear near the far left or far right of the table.

Weighted Positional Score

Select

Finding: Good queries return around 5-10% of the tuples in the table/join.

Histogram of Percentage of Tuples Returned by Example Queries for the SDSS Dataset

12.000

Group by

Grouping column: a column is selected if it has manageable distinct values.

Aggregates: In most cases, project count(*). If a separate numeric column is also discovered, demonstrate functions sum, min, max, and avg.

max/min cardinality	max/min(x , y)
cardinality difference	abs(x - y)
intersection cardinality	$ x \cap y $
union cardinality	$ x\cup y $
Jaccard similarity	$\frac{ x \cap y }{ x \cup y }$

We train a decision tree over these features using existing sets of example queries (with > 80% precision and recall)

Buid a graph (V,E) where V is the set of tables and E is the set of "good" joins.

Each quey is a minimum spanning tree of a connected component of this graph.



A histogram of the weighted positional scores of columns from the example queries of the SDSS database.



Union

Idea: two tables are good candidates for a union if they share sequence of columns with matching data types.

Finds more matches than just considering column name matches.

Evaluating "Good" Queries

What makes a good set of queries? It is application-dependent. e.g. for a DB class, the queries should demonstrate various SQL concepts vs. if user is familiar with SQL but not with the schema, better to have queries that refer to important tables or views.

What is an idiom? A function I : Q \rightarrow [0, 1]. Takes a query and outputs score between 0 and 1. Examples:

- outputs 1 if query includes GROUP BY clause, 0 otherwise
- rewards queries with more joins

How to use idioms to select the set of starter queries?

Represent queries as vectors of idiom

scores

	I ₁	l _m
q ₁	$I_1(q_1)$	$I_m(q_1)$
-		
-		
•		
-		

Goals:

- maximize idiom scores - select diverse set of queries

Score of a set of *k* queries:







