OVERVIEW OF DATA EXPLORATION TECHNIQUES

Stratos Idreos, Olga Papaemmanouil, Surajit Chaudhuri
SIGMOD 2015, Melbourne
USER INTERACTION
express interests

query/results recommendations

User Interface Layer

DB

collaborate

annotate

visualize results

assisted query formulation
User Interface Layer

DB
User Interface Layer

Data Visualization

DB
User Interface Layer

Data Visualization

Exploration Interface

DB
data visualization

visualization tools

visual optimizations

automatic visualization

User Interface Layer

Data Visualization

Exploration Interface

DB
data visualization

visualization tools

visual optimizations

automatic visualization

User Interface Layer

Data Visualization

Exploration Interface
Back in 1982...

window-based “sophisticated” browser for relational DBs

browser for multiple relations/tuples

rich query language for icon-oriented DBs

visual editor of text objects

browser for geographical data

TIMBER, VLDB’82
user-driven visualizations
user-driven visualizations

back-end queries: data selection, partition into panes

Polaris, INFOVIS 2002
user-driven visualizations

visual specifications

Polaris

back-end queries

data cubes

transformations (group by, sort)

back-end queries: data transformations (group, sort, aggregate within each pane)

Polaris, INFOVIS 2002
visual specifications

mappings (shape, size, color)

back-end queries: graphical transformations (render and visualize)

user-driven visualizations

DB user-driven visualizations

Polaris, INFOVIS 2002
collaborative exploration

live annotations

exploration for sky objects/patterns

AstroShelf, SIGMOD ’12
Live Annotations

subscriptions to interesting objects

collaborative exploration

exploration for sky objects/patterns

AstroShelf, SIGMOD ‘12
Live Annotations

collaborative exploration

stream based notifications

exploration for sky objects/patterns

AstroShelf, SIGMOD ‘12
data visualization

visualization tools

visual optimizations

automatic visualization

User Interface Layer

Data Visualization

Exploration Interface

DB
automatic visualization

User Interface Layer

Data Visualization

request views

interesting? insightful?

review views

manual, repetitive exploration for best visualization(s)
auto-ranked visualizations

VizDeck

model “good” charts

saved decks/replay logs

search, select, promote, discard, save, share

filter across charts, recommend, rank

VizDeck, SIGMOD ‘12
automatic visualizations

user query

Q₁, Q₂, ..., Qₙ

aggregations/
single-attribute
group-by

informative queries

high deviation
from overall dataset

utility

DB

visualization engine

% sales/ region

sales over time

SeeDB, PVLDB‘13
resolution reduction

user query

SciDB

query results

Visualization

expensive, ineffective on big data sets

Scalar, Big Data Vis '13
resolution reduction

user query → SciDB → query results → Visualization

user query → Data Reduction → reduced results → Visualization

modified query plans filter/aggregate/sample at given resolution

Scalar, Big Data Vis ‘13
approximate visualizations

SELECT X, AVG(Y)
FROM R(X, Y)
GROUP BY X

Blais et al, PVLDB '15
approximate visualizations

user query

SELECT X, AVG(Y)
FROM R(X,Y)
GROUP BY X

Sampling

reduced results

Visualization

approximate chart

clear ordering
less samples

Blais et al, PVLDB ‘15
approximate visualizations

user query

SELECT X, AVG(Y)
FROM R(X,Y)
GROUP BY X

Sampling

reduced results

Visualization

approximate chart

correct order?
sample more

min # samples for correct order?

Blais et al, PVLDB ‘15
approximate visualizations

SELECT X, AVG(Y)  
FROM R(X,Y)  
GROUP BY X

<table>
<thead>
<tr>
<th>#samples</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[60,90]</td>
<td>[20,50]</td>
<td>[10,40]</td>
<td>[40,70]</td>
</tr>
<tr>
<td>20</td>
<td>[64,84]</td>
<td>[30,48]</td>
<td>[15,35]</td>
<td>[45,65]</td>
</tr>
<tr>
<td>21</td>
<td>[66,84], l</td>
<td>[30,48]</td>
<td>[17,35]</td>
<td>[46,64]</td>
</tr>
<tr>
<td>70</td>
<td>[66,84], l</td>
<td>[40,47]</td>
<td>[17,32], l</td>
<td>[46,53]</td>
</tr>
</tbody>
</table>

sampling phases/ confidence intervals

Blais et al, PVLDB ‘15
visualization management

user query

overlapping user queries

DB

replicated db operations

query results

Visualization

memory operations on big data

Ermac, PVLDB ‘14
visualization management

user query → DB → query results → Visualization

visual specifications → DVMS

transformations to pixel space
visual optimizations

logical visual plans → physical query plans

reduced rendering time

Ermac, PVLDB ‘14
User Interface Layer

Data Visualization

Exploration Interface

DB

DB

exploration interfaces

automatic exploration

assisted query formulation

novel query interfaces
exploration interfaces

- automatic exploration
- assisted query formulation
- novel query interfaces

User Interface Layer

- Data Visualization
- Exploration Interface

DB
manual vs automatic
data exploration

long, imprecise, labor-intensive process

manual

SQL query formulation

query execution

result review

predicate adjustment
manual vs automatic data exploration

long, imprecise, labor-intensive process

manual

SQL query formulation

query execution

result review

predicate adjustment

auto

capture user interests

optimize query execution

reduce user effort

recommend data/queries
manual vs automatic data exploration

---

**Manual Process:**
- SQL query formulation
- Query execution
- Result review
- Predicate adjustment

**Long, Imprecise, Labor-intensive Process**

**Automatic Process:**
- Capture user interests
- Optimize query execution
- Reduce user effort
- Recommend data/queries
explore by example

effectiveness vs efficiency
sampling areas? sampling size?

AIDE, SIGMOD’14/ VLDB’15
explore by example

relevant areas to predict

Attribute B

Attribute A
explore by example

uniform sampling across domain

Attribute B

Attribute A

AIDE, SIGMOD’14/ VLDB’15
explore by example

discover relevant area

sampling around relevant objects

predicted relevant area

AIDE, SIGMOD’14/ VLDB’15
explore by example

refined predicted relevant areas

sampling around boundaries

AIDE, SIGMOD’14/ VLDB’15
result recommendations

YMALDB, VLDBJ ’13
result recommendations

query

YMALDB

interesting queries

DB

additional results

query

extract query faSets

expands attributes

rank faSets

top-k queries

selection predicates based on original query

add attributes from table schema

freq(result)/freq(DB)

YMALDB, VLDBJ ’13
result
recommendations

query

DB

YMALDB

DB

additional results

query

results

interesting queries

DB

extract query faSets

YMALDB

DB

expand attributes

rank faSets

top-k queries

selection predicates based on original query

add attributes from table schema

freq(result)/freq(DB)

"title, year, genre of Scorsese movies" + "title, year, genre, country of Scorsese movies" = many Scorsese movies are related to Italy

YMALDB, VLDBJ ’13
exploration interfaces

User Interface Layer

Data Visualization

Exploration Interface

DB

automatic exploration

assisted query formulation

novel query interfaces
keyword-based query suggestions

how we can discover relevant queries?
keyword-based query suggestions

keywords

Template Matcher

ranked templates

SQL Query Generator

suggested queries

Sample Results/Visualization

Template Repository

"database gray"

template on title/authors?
template on title?

SQLSUGG, ICDE’11
keyword-based query suggestions

Template Matcher

Template Repository

SQL Query Generator

Sample Results/Visualization

Keywords

Template generation

Paper

Template 1

Template 2

title_year

Paper — Author

id=p_id

SQLSUGG, ICDE’11
keyword-based
query suggestions

Template Matcher

Template Repository

SQL Query Generator

Sample Results/Visualization

template relevance = f (entity relevance & importance)

entity relevance ➔ keyword frequency in entity

entity importance ➔ importance of data nodes

SQLSUGG, ICDE’11
equi-join inference

table A

<table>
<thead>
<tr>
<th>A₁</th>
<th>A₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

table B

Cartesian product

inference algorithm

informative tuple

sample

goal join predicate

discover all positives
eliminate all negatives

minimize user effort

goal predicate:

Bonifati et al, EDBT’14
equi-join inference

sample

goal join predicate

inference algorithm

informative tuple

prune predicates

prune predicates with uninformative tuples

label tuple that prunes as many predicates as possible

Bonifati et al, EDBT’14
graphical query specification

result visualization

answers

non-answers

DataPlay, PVLDB ’13
graphical query specification

result visualization

answers

non-answers

pivot relation

add, remove query constraints

query /visualization recommendations

semantic query tuning by local syntactic modifications

DataPlay, PVLDB ’13
graphical query specification

result visualization

answers

non-answers

pivot relation → add, remove results

query corrections

search limited to local modifications

DataPlay, PVLDB ’13
query recommendations

query

Charles,
CIDR ’13
query recommendations

Different data partitions

query
results
Charles
queries
selected query

weight

weight, height

<5 | >5
<5 | >5
<20 | <30
<5 | >5
>20 | >30

Quality: simplicity, breadth, balance

Charles, CIDR ’13
query refinement

conditional query

select species from birds
where color= {red: 80%, blue: 20%}

ranked results by match probability
sensitivity of user predicates
query refinements w/ quality improvement

 Merlin, ICDE ’14
exploration interfaces

automatic exploration

assisted query formulation

novel query interfaces
no-keyboard interfaces

query context + gesture recognition = query intend

query space + search pattern = query template

GestureDB, CIDR ’13
no-keyboard interfaces

touch recognition → gesture recognition → map touch to operators

novel database kernel

touch input → quick response

dbTouch, CIDR ’13
Interactive Exploration through Data Prefetching & Query Approximation

MIDDLEWARE TECHNIQUES
interactive data exploration

SQL query formulation → query execution → result review → predicate adjustment

ad-hoc, non-optimized, labor-intensive process

interactive: small latency bounds on user wait time
middleware optimizations

query

results

middleware

DB

prefetching

query approximation

online processing

csample-based processing

speculative query execution

result reuse

structure-aware prefetching

query

approximation

DB
sample-based processing

- accuracy vs response times
- sample construction & selection
- error approximation
off-line data synopses

join synopses: sample distinguished joins
congressional samples: biased sampling for group-by queries
incremental maintenance: equi-depth & compressed histograms

Aqua, SIGMOD ’99
select avg(sessionTime) 
FROM table 
WHERE city="SF" 
WITHIN 1 SEC

Results 190+/- 5.89 (95% confidence)

BlinkDB, EuroSys ’13
data impressions

query & time/error bounds

approximate results

impressions during data loading

adaptive sampling to exploration focus

+ multi layer sampling and processing to meet user bounds

SciBORG, CIDR ’11
middleware optimizations

query approximation

- online processing
- sample-based processing

prefetching

- speculative query execution
- result reuse
- structure-aware prefetching
speculative query execution

1. predict follow-up queries
2. execute queries
3. cache results
speculative query execution

1. predict follow-up queries
2. execute queries
3. cache results
speculative query execution

1. predict follow-up queries
2. execute queries
3. cache results

exploration space reduction
query enumeration
query ranking
cube exploration

user query

SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone

DICE, ICDE ’14
User query:
```
SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone
```

Cube exploration operators:
- `WHERE month="m1"` (parent)
- `WHERE month="m1" AND week="w1"
AND hour="h1"` (child)
- `WHERE month="m1" AND week="w2"` (sibling)

Cube exploration

DICE, ICDE ’14
cube exploration

user query
SELECT AVG (iops) FROM events
WHERE month=“m1” AND week=“w1”
GROUP BY zone

speculative queries
Q(month=“m1”)
... 
Q(month = “m12”)
Q(hour =“h1”)
... 
Q(hour =“ h24”)
Q(week=“w2”)
... 
Q(week=“w3”)

DICE, ICDE ’14
cube exploration

user query
SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone

speculative queries
Q(month="m1")
...
Q(month = “m12”)
Q(hour =“h1”)
...
Q(hour =“ h24”)
Q(week="w2")
...
Q(week="w3")

DICE, ICDE ’14
cube exploration

Query Formulation

user wait time, t

Result Review

Speculative Execution

Query Execution

time

DB

DICE, ICDE ’14
cube exploration

maximize query probability
total speculation time < t

<table>
<thead>
<tr>
<th>QUERY</th>
<th>Probability</th>
<th>Exec Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁</td>
<td>0.3</td>
<td>22</td>
</tr>
<tr>
<td>Q₂</td>
<td>0.25</td>
<td>20</td>
</tr>
<tr>
<td>Q₃</td>
<td>0.25</td>
<td>35</td>
</tr>
<tr>
<td>Q₄</td>
<td>0.15</td>
<td>70</td>
</tr>
<tr>
<td>Q₅</td>
<td>0.05</td>
<td>35</td>
</tr>
</tbody>
</table>

DICE, ICDE ’14
result reuse

- identify (likely) **overlapping** results
- **cache** them
- reduce query execution time (user wait time)
semantic windows

user-defined window properties

overlapping results/windows

window prefetching

which order?

2D exploration space

Kalinin et al, SIGMOD ’14
semantic windows

user-defined window properties

overlapping results/windows

utility-based result ranking & result prefetching

2D exploration space

Kalinin et al, SIGMOD ’14
semantic windows

extend & prefetch

online performance vs query completion time

adjust prefetching size to output progress

Kalinin et al, SIGMOD ’14
data diversification

k representative tuples with max total pairwise distance
data diversification

query

DB

diversified results

k representative tuples with max total pairwise distance

Query Output

Max Diversified Set Search

Diversified Output k= 3

<table>
<thead>
<tr>
<th>T_1</th>
<th>d(T_1, T_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_2</td>
<td>d(T_2, T_3)</td>
</tr>
<tr>
<td>T_3</td>
<td>random tuple</td>
</tr>
<tr>
<td>T_4</td>
<td>d(T_4, T_3)</td>
</tr>
<tr>
<td>T_5</td>
<td>d(T_5, T_3)</td>
</tr>
</tbody>
</table>
data
diversification

Query Output

Max Diversified Set Search

Diversified Output k= 3

<table>
<thead>
<tr>
<th>Query</th>
<th>Output</th>
<th>Max Diversified Set Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>d(T1, T3)</td>
<td>d(T2, T1) + d(T2, T3)</td>
</tr>
<tr>
<td>T2</td>
<td>d(T2, T3)</td>
<td>d(T4, T1) + d(T4, T3)</td>
</tr>
<tr>
<td>T3</td>
<td>d(T4, T3)</td>
<td>d(T5, T1) + d(T5, T3)</td>
</tr>
<tr>
<td>T4</td>
<td>d(T1, T3)</td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>d(T5, T3)</td>
<td></td>
</tr>
</tbody>
</table>

random
tuple
**Query Output**

<table>
<thead>
<tr>
<th>T₁</th>
<th>d(T₁, T₃)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₂</td>
<td>d(T₂, T₃)</td>
</tr>
<tr>
<td>T₃</td>
<td>d(T₄, T₃)</td>
</tr>
<tr>
<td>T₄</td>
<td>d(T₅, T₃)</td>
</tr>
</tbody>
</table>

**Max Diversified Set Search**

<table>
<thead>
<tr>
<th>T₁</th>
<th>T₂</th>
<th>T₃</th>
<th>T₄</th>
<th>T₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(T₂, T₃)</td>
<td>d(T₄, T₃)</td>
<td>d(T₅, T₃)</td>
<td>d(T₁, T₃)</td>
<td>d(T₂, T₃)</td>
</tr>
</tbody>
</table>

**Diversified Output k=3**

<table>
<thead>
<tr>
<th>T₁</th>
<th>T₂</th>
<th>T₃</th>
<th>T₄</th>
<th>T₅</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d(T₂, T₃)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>d(T₄, T₃)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>d(T₅, T₃)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data diversification**

k representative tuples with max total pairwise distance
interactive data diversification

overlapping diversified results

long Time-To-Insight

- cache diversified results and use most promising
- regression model predicts max diversification of a set

DivIDE, SSDBM’14
**interactive data diversification**

- search space **pruning** through regression model
- best/first fit search for max total diversification among cached and new results

**DivIDE, SSDBM’14**
structure-aware prefetching

- prefetching for interactive spatial query sequences
- model structures of past spatial queries in graph
- identify guiding structure in past two queries: iterative pruning
- cache the predicted next location

SCOUT, VLDB’12