An Algebraic Language for Automatic Semantic Data Integration on the Hidden Web

Presented By
Md. Shazzad Hosain
Outline

- Hidden Web vs. Shallow Web
- Motivation
- Problems and Focus of Our Research
- BioFlow
- Integration Algebra, called Integra
- Related Works
- Future Works
- Summary
Welcome from the Chair

We welcome your interest in the Department of Computer Science. As a Department within the College of Liberal Arts and Sciences we are dedicated to educational excellence as well as scholarly achievement. The Department offers core academic strength in many areas including artificial intelligence, networking, databases, distributed and parallel computing, security, graphics and visualizations, and software engineering. A high-quality, innovative, education, focusing on the fundamentals of computer science while emphasizing technology, prepares students for employment and advanced studies.

We currently offer undergraduate (Bachelor of Science and Bachelor of Arts) and graduate degrees (Master of Science and PhD) in Computer Science, as well as an undergraduate degree in Information Systems Technology and a certificate in Scientific Computing. In addition, graduate students now have the opportunity to complete their Ph.D. work in Computer Science with a concentration in Bioinformatics and Computational Biology.

We have recently revised our curriculum to better serve our student population. Hallmarks of the new undergraduate curriculum include a high degree of hands-on experience with real-world systems, increased personal attention; undergraduate participation in research projects; and a carefully crafted progression of course content as students advance through the program. The increase in lab courses is supported by six state-of-the-art computer teaching labs. Undergraduate students now have the opportunity to enhance their knowledge by choosing among a variety of concentrations including Software Engineering, Databases, Web Technology and Computer Gaming, to name only a few. Study abroad opportunities are available with European schools including the Polytechnic School of Nantes, France.

In recent years, the Department has hired several new faculty members, leading to a substantial increase in interdisciplinary scholarship. In fact, the Department of Computer Science is a major focus for interdisciplinary activities within the university and has the potential for gaining much greater prominence. Our annual R&D expenditures are between $2 and $3 million. The Department’s commitment to community service includes a new joint degree program with Focus: HOPE, as well as...
**Used 2008 Toyota Sequoia 4x4 Limited**

- Black
- 8 Cylinder

**Price:** $34,999

Mileage: 3,000

**Private Seller**

Email this seller
Get this seller's phone number


**New 2009 Toyota Avalon**

- 2 Exact Matches

Starting at: $34,939

1 at this price

**Toyota of Waterford**

- Where your trade-in is worth MORE! OPEN EVERY SATURDAY FROM 10AM - 5PM!!!!
- Where your trade is worth more
- On M-59 just West of Telegraph

Contact this dealer

**Advertised by:**

**2009 Sonata**

Lease it today for just $100
Shallow Web vs Hidden Web

- Shallow Web mostly unstructured or semi-structured and/or natural language based text
- Hidden Web has some regular structure since it is generated from the database behind dynamically
- Study shows hidden web is about 91,000 tera bytes while shallow web 167 tera bytes
Best used vehicles under $20000
Best used cars under $20000. Our lists of reliable cars from our latest Annual Car Reliability Survey. ... Also, see our complete list of reliable used vehicles and those used cars to avoid (available ... Price ranges are what you'd pay for a typically equipped car with average mileage. Less than $4000, $4000-$6000 ... www.consumerreports.org/.../cars/used-cars/...used/.../overview/ - Cached - Similar - 📧 🔗

10 Best Used Cars for Less Than $19999 -- Inside Line
16 Aug 2009 ... And awesome is available for less than $20000. ... Because this price point is pretty lofty, we left out any vehicle more than a decade old or ... This is undeniably a list of enthusiast cars, real performance machines, ... All prices come straight from Edmunds.com and reflect True Market Value ...
www.edmunds.com/insideline/do/Features/articleId=155146 - Cached - Similar - 📧 🔗
Motivating Example

- List all used cars price less than $25,000 with mileage no more than 50,000, having dealer location within 50 miles of zip code 48202.
Find Your Car

Use any of the criteria below to start your search, or find by AT Car ID.

Search for
- All cars
- New cars
- Certified cars
- Used cars

Search Area
within
50 miles

of ZIP Code
48202 [Required]
** Post Your Car For Sale Here $19.95! **

Audi

BMW

BMW 330 ci convertible, asking $29K OBO

Cadillac

1971 El Dorado Cadillac 80% restored
Sedan DeVille, V8, 4.5L, 155 HP, 51,000 Miles, $4,280

Chevrolet

2003 Chevy Silverado 1500

Chrysler

Beautiful 1978 New Yorker

For-Sale

1977 Datsun 280 Z $10,000
1987 Sterling 825s $2500
2000 2DR Oldsmobile Alero - $4000.00

Cars for Sale

Sell Your Car for $19.95
Tips for Selling Your Car
Should I Buy A Used Car?
How to Buy a Used Car
Bargain Hunting For A Used Car
How to Buy a New Car
Help Buying a New Car
Should You Buy A Certified Pre-Owned Car?
Dealer Directory
Vehicle History Report
Auto Insurance
Extended Auto Warranty
Extended Warranty FAQ
Lemon Detector
2001 Acura CL Type-S, 45K miles

2001 Acura
2001 Acura CL Type-S 2 door luxury coupe in excellent condition. The car has a CARFAX clean title guarantee. Black exterior with tan leather interior, looks excellent with no wear at all. Equipped with a 3.2L V6 SOHC 24V FI Engine, 5-Speed Automatic Transmission, Bucket Seats, Center Console, Garage Door Opener, Heated Front Seats, Power Heated Mirrors, Keyless Entry System, Lighted Entry System, Power Brakes, Power Moonroof, Rear Window Defroster, Wood Interior Trim, 17 Inch Alloy Wheels, Power Driver Seat...
Can we automate the process? Or

Can a system automatically crawl the hidden Web and find answer for us?
Outline

- Hidden Web vs. Shallow Web
- Motivation
- Problems and Focus of Our Research
- BioFlow
- Integration Algebra, called Integra
- Related Works
- Future Works
- Summary
Problems of Integration

- Automatic resource discovery
- Automatic form submission
- Extraction of relations
- Column name identification
- Object consolidation
- Query for the answer
Automatic Resource Discovery

- List all used cars priced less than $25,000 with mileage no more than 50,000, having dealer location within 50 miles of zip code 48202.

- Which site to go?
- What to submit?
- How to order link and combine?
Automatic Form Submission

- Submit with name value pairs
- List cars within radius 50 miles of zip-code 48202
- Schema matching problem


- COMA, Cupid, Ontobuilder etc.
Extraction of Relations


DeLa, DEPTA, RoadRunner etc.
Column Name Assignment


- DeLa etc.
# Object Consolidation

- Record Linkage or key identification
- TANE, GORDIAN etc.

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
<th>startAddress</th>
<th>endAddress</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
<td>48202</td>
<td>Estchester, NY</td>
<td>630</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
</tbody>
</table>
Steps of Automatic Integration

- Automatic Resource Discovery
- Automatic Form Submission
- Automatic Information Extraction as Wrapper generation
- Column Name Identification
- Semantic Reconciliation
- Query over integrated view

Schema Matching
Key Identification
The Idea

- Exploit the power of declarative language
- BioFlow

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

AutoTrader

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

CarSearch

<table>
<thead>
<tr>
<th>startAddress</th>
<th>endAddress</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>48202</td>
<td>Estchester, NY</td>
<td>630</td>
</tr>
<tr>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
</tbody>
</table>
The Idea

- Link CarSearch, GoogleMap as R1
- Combine AutoTrader, R1 as R2
- Select * from R2 where car = “bmw”, mileage < 50000, price < 20000, distance < 50

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
<th>startAddress</th>
<th>endAddress</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
<td>48202</td>
<td>Estchester NY</td>
<td>630</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
</tbody>
</table>
The Advantage

- If resource discovery and pipeline definition can be automated
- Build visual interface

Focus of Our Research

- Develop an autonomous integration system
- Develop a declarative language, BioFlow
- Develop an algebraic language for BioFlow, called Integra
Outline

- Hidden Web vs. Shallow Web
- Motivation
- Problems and Focus of Our Research
- BioFlow
- Integration Algebra, called Integra
- Related Works
- Future Works
- Summary
BioFlow


BioFlow

- BioFlow is a semantic Extension of SQL
- Like SQL it has two types of statements
  - Resource Definition
    - Storage Definition
    - Web Function Definition
  - Data Manipulation or Query statements
Storage Definition

create datatable AutoTrader
{
    car varchar(60),
mileage integer,
price integer,
dealer varchar(60),
address varchar (100),
distance float
} [format table [at URL]] ;
Web Function Definition

define function AutoTrader
    extract car varchar(80), mileage integer, price integer,
        dealer varchar(80), address varchar(80), distance float
    using wrapper FastWrap, matcher OntoMatch
from URL www.autotrader.com
submit (distance integer, zip-code integer,
    searchType varchar (40));
Data Manipulation or Query Language

- We have extended relational algebra operators to Integra operators.
- Also introduced one unary operator called transform operator that converts Web pages to relations.
Relational Algebra Approach

\[
\text{CarSearch} \bowtie_{\text{location} = \text{endAddress}} \text{GoogleMap}
\]

\[
\pi_{\text{car}, \text{mileage}, \text{price}, \text{distance}}(\text{AutoTrader})
\]

\[
\pi_{\text{carType}, \text{mileage}, \text{price}, \text{distance}}(\text{CarSearchGoogleMap})
\]
Relational Algebra not Enough for Automatic Integration

- Schema Specific
- Exact value has to be specified
- Or, semantic matching not possible
Outline

- Hidden Web vs. Shallow Web
- Motivation
- Problems and Focus of Our Research
- BioFlow
- Integration Algebra, called Integra
- Related Works
- Future Works
- Summary
Integra Operators

- Extended select
- Extended project
- Extended minus
- Extended intersection
- New binary operators
  - Link
  - Combine
- Transform
Preliminary Concepts

- We consider Web pages as relations
- Terms are basic element in Integra
  - *a-term* is an attribute name
  - *v-term* is the attribute value
Semantic Equivalence

Term Similarity

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

AutoTrader

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

CarSearch

car ~ carType
Warren, Michigan ~ Warren, MI
Term Equivalence

- Two a-terms are equivalent if the terms are similar and the types are equal
  \[\text{car} \sim \text{carType}\]

- Two v-terms are equivalent if the terms are similar and their associated a-terms are equivalent
  \[\text{Warren, Michigan} \sim \text{Warren, MI}\]
Object Equivalence

- Each relation has a key
- The keys are equivalent
- And the key values are similar
Database Functions

- Semantic Reconciliation Function (μ)
- Key Discovery Function (κ)
- Extraction Function (ω)
Semantic Reconciliation Function ($\mu$)

- $U = \{\text{car, mileage, price, dealer, address, distance}\}$
- $V = \{\text{carType, mileage, price, phone, location}\}$
- $\mu(U, V) = \{\{\text{car, carType}\}, \{\text{mileage, mileage}\}, \{\text{price, price}\}, \{\text{address, location}\}\}$
Semantic Reconciliation Function ($\mu$)

- $U = \{\text{“Warren, Michigan”}\}$
- $V = \{\text{“Warren, MI”}\}$
- $\mu(U, V) = \{\{\text{“Warren, Michigan”}, \text{“Warren, MI”}\}\}$ or $\{\}$

- OntoMatch, Cupid, COMA

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>
Key Discovery Function ($\kappa$)

$\kappa$(CarSearch) = \{ \{carType\}, \{mileage\}, \{price\}\}

- TANE
- GORDIAN
Extraction Function ($\omega$)

- Given a Web page the function automatically identifies tabular data from the page
- $\omega$ (page) = R
- FastWrap, DeLa, DEPTA
### Extended Select Operation

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

**AutoTrader**

\[
\sigma_{\text{location}='Warren, MI'}(\text{AutoTrader})
\]

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>location</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
</tbody>
</table>

(a) selection on AutoTrader
Extended Project Operation

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

CarSearch

\[ \hat{\pi}^{\mu}_{car, price, address}(CarSearch) \]

<table>
<thead>
<tr>
<th>car</th>
<th>price</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>13000</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>17500</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>20500</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

(b) projection on CarSearch
Combine Operation

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

### AutoTrader

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

### CarSearch

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
<td>null</td>
</tr>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>null</td>
<td>Eastchester, NY</td>
<td>null</td>
<td>917-692-4108</td>
</tr>
</tbody>
</table>

(c) combine on AutoTrader and CarSearch
### Link Operation

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

CarSearch: $\chi^\mu_k$ GoogleMap

<table>
<thead>
<tr>
<th>startAddress</th>
<th>endAddress</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>48202</td>
<td>Estchester, NY</td>
<td>630</td>
</tr>
<tr>
<td>48202</td>
<td>Warren, MI</td>
<td>14.4</td>
</tr>
</tbody>
</table>

(c) link on CarSearch and GoogleMap
## Extended Minus Operation

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

### AutoTrader

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

### CarSearch

\[ \text{AutoTrader} \xrightarrow{\mu} \text{CarSearch} \]

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

(c) difference on AutoTrader and CarSearch
## Extended Intersection Operation

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Saturn Vue 2.2L</td>
<td>32758</td>
<td>11890</td>
<td>Autonet</td>
<td>Plymouth, Michigan</td>
<td>18</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
<tr>
<td>2004 BMW-coupe</td>
<td>36000</td>
<td>20500</td>
<td>Paradise Autos</td>
<td>Warren, Michigan</td>
<td>14.4</td>
</tr>
</tbody>
</table>

### AutoTrader

<table>
<thead>
<tr>
<th>carType</th>
<th>mileage</th>
<th>price</th>
<th>phone</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 Acura CL Type-S</td>
<td>45000</td>
<td>13000</td>
<td>917-692-4108</td>
<td>Eastchester, NY</td>
</tr>
<tr>
<td>2003 Audi Allroad</td>
<td>55000</td>
<td>17500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
<tr>
<td>2004 BMW coupe</td>
<td>36000</td>
<td>20500</td>
<td>586-756-4240</td>
<td>Warren, MI</td>
</tr>
</tbody>
</table>

### CarSearch

\[ \text{AutoTrader} \hat{\cap}_{\kappa} \text{CarSearch} \]

<table>
<thead>
<tr>
<th>car</th>
<th>mileage</th>
<th>price</th>
<th>dealer</th>
<th>address</th>
<th>distance</th>
<th>phone</th>
</tr>
</thead>
</table>
define function AutoTrader
    extract car varchar(80), mileage integer, price integer, dealer varchar(80),
    address varchar(80), distance float
    using wrapper FastWrap, matcher OntoMatch
from URL www.autotrader.com
submit (distance integer, zip-code integer, searchType varchar (40));

Call AutoTrader (50, 48202, “all cars”) ;
Properties of Integra

- Compositionality and closure property
  - All operators take one or two input relations and output one relation

- Integra reduces to relational algebra
  - We can find equivalent relational algebra expressions for the Integra operators as the following commutative diagram.

```
\[
\begin{array}{c}
\mathcal{R} \\
\downarrow \vartheta \\
\mathcal{R}' \\
\end{array}
\xrightarrow{\vartheta(Q)}
\begin{array}{c}
\mathcal{Q} \\
\downarrow \vartheta \\
\mathcal{O} \\
\end{array}
\]
```

Commutative Diagram
Translating Integra to RA

We take an indirect approach

Integra to Relational Algebra Mapping process


Outline

- Hidden Web vs. Shallow Web
- Motivation
- Problems and Focus of Our Research
- BioFlow
- Integration Algebra, called Integra
- Related Works
- Future Works
- Summary
Related Works

Operators
- Match Join Operator
- Merge Operator
- Join Merge Operator
- Fusion Operator
Fusion Operator

- Fusion operator is n-ary operator

$$\phi_{F,C,R,S}(R)$$

$R = r_1, r_2, ..., r_n$ \hspace{1cm} $n$ input relations

$F = f_1, f_2, ..., f_m$ is a list of $m$ attributes

$CR = cr_1, cr_2, ..., cr_k$ is a list of $k$ conflict resolution functions

$S = s_1, s_2, ..., s_l$ is a list of $l$ intra-group sort key.
First it takes outer union of input relations
Then fuses data identified by $F$ according to $CR$

```
SELECT Name, RESOLVE(Age, max), VOTE(Student)
FUSE FROM EE_Students, CS_Students
FUSE BY Name
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>AGE</th>
<th>STUDENT</th>
<th>CAR</th>
<th>PHONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>22</td>
<td>no</td>
<td>Ford</td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>22</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>22</td>
<td>yes</td>
<td>VW</td>
<td></td>
</tr>
<tr>
<td>Charly</td>
<td>22</td>
<td>yes</td>
<td>Pontiac</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAME</th>
<th>AGE</th>
<th>STUDENT</th>
<th>CAR</th>
<th>PHONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul</td>
<td>26</td>
<td>yes</td>
<td>Chevy</td>
<td></td>
</tr>
<tr>
<td>Paul</td>
<td>26</td>
<td>yes</td>
<td>Chevy</td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>25</td>
<td>yes</td>
<td>Pontiac</td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>25</td>
<td>yes</td>
<td>VW</td>
<td></td>
</tr>
<tr>
<td>Charly</td>
<td>25</td>
<td>yes</td>
<td>Chevy</td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>22</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary</td>
<td>22</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary</td>
<td>22</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAME</th>
<th>AGE</th>
<th>STUDENT</th>
<th>CAR</th>
<th>PHONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>22</td>
<td>no</td>
<td>Ford</td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>22</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>22</td>
<td>yes</td>
<td>VW</td>
<td></td>
</tr>
<tr>
<td>Charly</td>
<td>22</td>
<td>yes</td>
<td>Pontiac</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: EE Students

Table 5.2: CS Students
Strengths of Integra

- The operators are schema independent
- Object identification or key discovery is automatic
- No user specific conflict resolution functions
- Most importantly, Integra has a set of algebraic operators that provide a declarative query language called BioFlow
Outcome of the Research


Future Works

- Extend the algebra for nested relations
- Query plan generation and query optimization
- Performance analysis of the algebra based on its component functions
Summary

- We have proposed an integration algebra, called Integra
- Integra extends relational algebra with a number of user defined functions
- We have shown Integra is relational complete
- Integra is closed and thus compositional complete
Thanks
Questions?