Constraint-based Information Integration

Craig Knoblock
University of Southern California

Thanks to Jose Luis for some of these slides
Constraint Satisfaction and Propagation for Integration

- Integrating data from multiple sources often involves reasoning about the information
- Constraints provide a approach to expressing relationships and filtering data
Outline

Part I
- Constraint satisfaction in SmartClients
- Constraint propagation in Heracles

Part II (Dr. Jose Luis Ambite)
- Building your own applications in Heracles
SmartClients [Torrens et al, 2002]

- Cast an integration problem as a Constraint Satisfaction Problem (CSP)
- Given a request, the server retrieves the required data and sends the data and the CSP to the client
- Client solves the CSP locally
  - Large complex problem transmitted in small amount of space
  - Provides fine-grained user interaction with the data
Example Problem

I live in Bern, Switzerland, and would like to visit colleagues in Princeton (New Jersey), and London. I would like to spend at least two days in each place, and will need to travel in the first two weeks of February.

1. 1st leg from Bern to Princeton: flights from ZRH/BSL/GVA to JFK/EWR/PHL on the dates from 1st to 10th February,

2. 2nd leg from Princeton to London: flights from JFK/EWR/PHL to LGW/LHR/LCY on the dates from 4th to 12th February, and

3. 3rd leg from London to Bern: flights from LGW/LHR/LCY to ZRH/BSL/GVA on the dates from 6th to 14th February.
Constraint Satisfaction Problem

- CSP
  - Variables
  - Domains for each variable
  - Constraints on variables
    - Supports both hard and soft constraints

- Algorithms and techniques for solving CSPs have been widely studied
Example CSP Graph

Variables Arr T₁, Dep T₁, Start-Time and End-Time contain information about Dates and Times.
Formalization of Example: Variables

\[ X = \{ DT_0, \ldots, DT_{n-1}, AT_0, \ldots, AT_{n-1}, \text{Airports}_0, \ldots, \text{Airports}_n, \text{Flights}_0, \ldots, \text{Flights}_{n-1}, \text{AirCrafts}, \text{Fares}, \text{Airlines}, \ldots \} \]

is a set of variables. There are several kinds of variables:

- \( DT_i \) and \( AT_i \) represent the dates and times on which the traveler could depart and arrive respectively.
- \( \text{Airports}_i \) represents the possible airports near the departure for leg \( i \) of the itinerary.
- \( \text{Flights}_i \) stands for the possible flights in between the airports of \( \text{Airports}_i \) and \( \text{Airports}_{i+1} \).
Formalization of Example: Domains

\[ D = \{ D_1, \ldots, D_n \} \] is the set of domains. There are several kinds of domains depending on the type of the associated variable:

- For variables \( DT_i \) or \( AT_i \): the domain contains all possible departure and arrival times for the \( leg_i \).
- For variables \( Airports_i \): the domain is a set of airports for the departure of the \( leg_i \).
- For variables \( Flights_i \): the domain is the set of possible flights from \( Airports_i \) to \( Airports_{i+1} \).
- For variables \( AirCrafts, Fares \) and \( Airlines \): the domain is the set of different aircrafts, the set of available fares or the set of airline companies respectively.
Formalization of Example: Constraints

$C = \{C_1, \ldots, C_k\}$ is the set of constraints. Basically, there are two kinds of constraints: those imposed by the user’s preferences and those imposed by flight schedules. There are constraints on the variables $Flights_i$, $Airports_i$, $DT_i$ and $AT_i$ that guarantee that the flight is compatible with the airports, departures times and arrival times. A binary constraint in between $AT_i$ and $DT_{i+1}$ takes into consideration that the flight for $leg_{i+1}$ departs after the flight for $leg_i$ arrives. Then most of the user’s preferences are expressed by means of constraints between $Flight_i$ variables and $Aircrafts, Airlines, Fares$ and other variables.
Constraint Graph for Example
Architecture for SmartClients

Database with information for the CSP

SERVER

Generation of CSP

Initial specification

smart agent

CLIENT

Resolution of CSP
Pros and Cons

- **Pros**
  - Elegant approach that exploits past work on CSPs
  - Minimizes the data retrieval and supports complex reasoning and integration of the data

- **Cons**
  - Assumes that all data can be retrieved before any reasoning about the data
  - In the travel planning, assumes that prices are the same on any date and there are no issues with flight availability
Heracles Constraint Integration Framework

- Framework for building integrated applications
- Interleaves planning and information gathering
- Uses a constraint reasoner to decide what sources to query and to integrate the results
The Travel Assistant

**Meeting With:** Jim Hendler
**Company Name:** DARPA

**Meeting:** CoABS PI Meeting
**Location:** Washington, DC

**Starting At:**
- **Month:** Feb
- **Day:** 16
- **Year:** 2001
- **Time:** 01:00 PM

**Ending At:**
- **Month:** Feb
- **Day:** 18
- **Year:** 2001
- **Time:** 03:00 PM

**Leaving From:**
- **Street:** 2700 University Park
- **City:** Los Angeles
- **State:** CA

**Traveling To:**
- **Street:** 1120 19th ST NW
- **City:** Washington
- **State:** DC

**Destination weather:** Partly Cloudy
**Forecast:**
- **Hi:** 57
- **Low:** 46

**Distance (miles):** 2294

**Mode to Destination:** Fly

**Click to Expand**
Dynamically Updates Slots as Information Becomes Available
Supports Informed Choices
Changes Propagate Throughout
User Can Specify High-Level Preferences
Constraint Networks for Integrating Information

Components:
- Representation of the variables
- Representation of constraints
- Hierarchical template representation
- Constraint propagation and cycle detection
Constraint Networks for Integrating Information

Components:

- Representation of the variables
- Representation of constraints
- Hierarchical template representation
- Constraint propagation and cycle detection
Constraint Variables

- Constraint network consists of a set of variables such as:
  - MeetingStartTime
  - MeetingLocation

- Each variable depends on a set of ancestors.

- Variables are related by constraints that determine the possible values of a solution
Constraint Networks for Integrating Information

Components:
- Representation of the variables
- Representation of constraints
- Hierarchical template representation
- Constraint propagation and cycle detection
Constraint Representation

- Constraints are computable components:
  - Local calculations based on Xquery
    - MeetingStartTime + MeetingDuration --> MeetingEndTime
  - Web and Database Wrappers
    - ITN: DepartureAirport, ArrivalAirport, Date --> Flights
    - Yahoo Weather: City, Date --> Weather predication
  - External Programs (Outlook, Planners, etc)
    - Outlook Calendar: Date --> Meetings
Constraint Structure

Constraint

- Arguments: input and output variables
- Call:
  - Construct table with inputs and corresponding calls (http requests, SQL queries, etc) to sources (wrappers, DBs, etc) [using XML Query]
  - Calls are executed and results stored in a table
- Output
  - Restructure source results into desired output [using XML Query]
Drive or Take a Taxi?

- Departure Date: Sep 30, 2000
- Return Date: Oct 2, 2000
- Departure Airport: LAX
- Duration: 3 days
- Parking Rate: $7.00/day
- Parking Total: $21.00
- Driver Distance: 15.1 miles
- Get Taxi Fare: $23.00

Select Mode To Airport: Drive
Constraint Networks for Integrating Information

Components:
- Representation of the variables
- Representation of constraints
- Hierarchical template representation
- Constraint propagation and cycle detection
Hierarchically-Partitioned Constraint Networks

- Template:
  - Groups related variables and constraints
  - Organizes information for computation and presentation to user

- Templates organized hierarchically
  - Template decomposed into subtemplates
  - Choose among alternative subtemplates
Template Structure

Template

- Arguments: input and output variables
- Variables: name, type, default values
- Constraints
- Expansions: alternative subtemplate calls
- GUI specification
Partitioned Constraint Network

Who

Company

Subject

Dest Weather

Starting Time

Origin Weather

Ending Time

Origin Addr.

Dest. Addr.

Origin Addr.

Distance

Travel Mode

Depart Time

Depart Airport

Dist. toAirport

Depart Time

Depart Airport

Dist. toAirport

Arrival Time

Parking Lot

Taxi Fare

Flight Num

Parking Rate

Mode toAirport

Arrival Airport

Parking Lot

Parking Rate
Template Hierarchy for the Travel Assistant

Diagram:
- Trip
  - ModeToDestination
    - Drive
    - Fly
    - Taxi
  - ModeHotel
    - Hotel
    - NoOvernight
  - ModeNext
    - Trip (Return Home)
    - Trip (Return Office)
    - Trip (New Leg)
  - End Trip

Diagram structure:
1. ModeToDestination
   - Drive
   - Fly
   - Taxi
2. ModeHotel
   - Hotel
   - NoOvernight
3. ModeNext
   - Trip (Return Home)
   - Trip (Return Office)
   - Trip (New Leg)
4. End Trip
Dynamic Constraint Networks

Generalization of Constraint Networks

- Variables can be active or inactive
- Normal Constraints
  \[ x_1 = k_1 \land \ldots \land x_m = k_m \rightarrow x_n = k_n \]
- Activity constraints:
  \[ x_1 = k_1 \land \ldots \land x_m = k_m \rightarrow \text{active}(x_n) \]
- Inactive variables do not participate in the network, i.e., do not propagate constraints
Heracles: Template Selection

- Core network
  - Computes values of template selection vars
  - Always active

- Template selection variables
  - Inputs to activity constraints: determine the choice of subtemplates, i.e., which additional variables are active
Constraint Networks for Integrating Information

Components:
- Representation of the variables
- Representation of constraints
- Hierarchical template representation
- Constraint propagation and cycle detection
Constraint Propagation

- **Approach**
  - When a variable is assigned a value, re-compute the value sets and assigned values of all dependent variables
  - Proceeds recursively until no values are changed or a cycle is detected

- **Core network**
  - Propagates all variables through the core network
  - Remaining variables are computing when a template is opened

- **Does not perform full CSP**
  - Less costly
  - Does not require all information in advance
  - Makes choices locally, so may fail to find optimal assignment
Cycle Detection

- Address cyclic *interactive* networks
  - Multiple input paths:
    - region/country/city vs. lat/long
  - Conversion rounding errors:
    - lat/long, temperature, …

=> Cycle detection in constraint propagation
Interactive Cyclic Network

1. Find City Center (C1)
2. Copy City Center (C2)
3. Closest City (C3)
Interactive Cyclic Network

<table>
<thead>
<tr>
<th>Sequence of events</th>
<th>User time</th>
<th>Visited Variables</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>1</td>
<td>Ø</td>
<td>LA</td>
</tr>
</tbody>
</table>

- **City (v1)**
  - **C1: Find City Center**
    - City center (v2)
    - **C2: Copy**
      - Lat/Long (v3)
    - **C3: Closest City**
Interactive Cyclic Network

<table>
<thead>
<tr>
<th>Sequence of events</th>
<th>User time</th>
<th>Visited Variables</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>1</td>
<td>Ø</td>
<td>LA</td>
</tr>
<tr>
<td>2)</td>
<td>1</td>
<td>v1</td>
<td>34N118W</td>
</tr>
<tr>
<td>3)</td>
<td>1</td>
<td>v1 v2</td>
<td>34N118W</td>
</tr>
</tbody>
</table>
### Interactive Cyclic Network

**Sequence of events**

<table>
<thead>
<tr>
<th>Event</th>
<th>User time</th>
<th>Visited Variables</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>1</td>
<td>Ø</td>
<td>LA</td>
</tr>
<tr>
<td>2)</td>
<td>1</td>
<td>v1</td>
<td>34N118W</td>
</tr>
<tr>
<td>3)</td>
<td>1</td>
<td>v1 v2</td>
<td>34N118W</td>
</tr>
<tr>
<td>4)</td>
<td>Blocked!</td>
<td>(t(v3) = t(v1) \land v1 \in \text{vis}(v3))</td>
<td></td>
</tr>
</tbody>
</table>

**Diagram:**
- **C1: Find City Center**
- **C2: Copy**
- **C3: Closest City**
- **City (v1)**
- **City center (v2)**
- **Lat/Long (v3)**

**Interaction:**
- C1: Find City Center
  - City (v1)
  - City center (v2)
  - Lat/Long (v3)

- C2: Copy
  - Lat/Long (v3) → City center (v2)

- C3: Closest City
  - City center (v2) → City (v1)

Note: The interaction of C1 and C3 is shown with a cycle, indicating the interactive nature of the network.
Interactive Cyclic Network

C3: Closest City
  ↓
C1: Find City Center
  ↓
City center (v2)
  ↓
C2: Copy
  ↓
Lat/Long (v3)
  ↓
City (v1)

Sequence of events | User time | Visited Variables | Value
--- | --- | --- | ---
1) | 1 | Ø | LA
2) | 1 | v1 | 34N118W
3) | 1 | v1 v2 | 34N118W
4) Blocked! | | v1 ∈ vis(v3) |
5) | 2 | Ø | 40N70W
Interactive Cyclic Network

<table>
<thead>
<tr>
<th>Sequence of events</th>
<th>User time</th>
<th>Visited Variables</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>1</td>
<td>Ø</td>
<td>LA</td>
</tr>
<tr>
<td>4)</td>
<td></td>
<td>Block! t(v3) = t(v1) ^ v1 ∈ vis(v3)</td>
<td></td>
</tr>
<tr>
<td>6)</td>
<td>2</td>
<td>v3</td>
<td>NY</td>
</tr>
<tr>
<td>2)</td>
<td>1</td>
<td>v1</td>
<td>34N118W</td>
</tr>
<tr>
<td>3)</td>
<td>1</td>
<td>v1 v2</td>
<td>34N118W</td>
</tr>
<tr>
<td>5)</td>
<td>2</td>
<td>Ø</td>
<td>40N70W</td>
</tr>
</tbody>
</table>

- C1: Find City Center
- C2: Copy
- C3: Closest City
Interactive Cyclic Network

C3: Closest City

C1: Find City Center

C2: Copy

City (v1)

City center (v2)

Lat/Long (v3)

Sequence of events | User time | Visited Variables | Value
--- | --- | --- | ---
1) | 1 | Ø | LA
4) | | Blocked! $t(v3) = t(v1) \land v1 \in \text{vis}(v3)$
6) | 2 | v3 | NY
2) | 1 | v1 | 34N118W
7) | 2 | v3 v1 | 40N73W
3) | 1 | v1 v2 | 34N118W
5) | 2 | Ø | 40N70W
Interactive Cyclic Network

Sequence of events | User time | Visited Variables | Value
---|---|---|---
1) | 1 | Ø | LA
4) | Blocked! | $t(v_3) = t(v_1) ^ v_1 \in vis(v_3)$ |
6) | 2 | v_3 | NY
2) | 1 | v_1 | 34N118W
7) | 2 | v_3 v_1 | 40N73W
3) | 1 | v_1 v_2 | 34N118W
5) | 2 | Ø | 40N70W
8) | Blocked! | $t(v_3) = t(v_2) ^ v_3 \in vis(v_2)$ |
Discussion

- General framework for interleaving planning and information gathering
  - Retrieves information as needed
  - Gathers and integrates data in a uniform framework
  - Evaluates tradeoffs and selects among alternatives
  - Allows the users to explore alternatives
  - Supports a wide variety of information types: databases, web pages, images, video, etc.