Embedded cost model in mobile agents for large scale query optimization

M. Hussein, F. Morvan, A. Hameurlain
Paul Sabatier University, IRIT
July 2005
Outline

- Context & motivations
- Problem position
- Embedded cost model
- Performance evaluation
- Conclusion & perspectives
Data Integration: Mediators-Wrappers approach

- Optimizer
- Cost Model
- Mediator
- Wrapper
- DBMS
- Web
- Applications

- Optimizer
- Cost model
- Mediator
- Wrapper
- Program
- FMS
- Applications
Mediator cost model

• Several approaches:
  – Calibration (DU 95, GARDARIN 96, ZHU 03)
  – Historic (ADALI 96)
  – Generic (NAACKE 98)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Catalog</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Size of temporary relations</td>
<td>• Profiles of data sources</td>
</tr>
<tr>
<td>• Execution cost of operators</td>
<td>• Architecture</td>
</tr>
<tr>
<td></td>
<td>• Cost of wrappers</td>
</tr>
<tr>
<td></td>
<td>• Cost formulas</td>
</tr>
<tr>
<td></td>
<td>– Mediators</td>
</tr>
<tr>
<td></td>
<td>– Wrappers</td>
</tr>
</tbody>
</table>
Query optimization based on mobiles agents

Motivations

- Inaccuracy of estimations
  - Size of operands
  - Costs of operators
- Unavailability of resources (CPU, Memory, Network, Data)
- Centralization of optimization
  - Classical and dynamic optimization methods are centralized
  - Depended on the mediator which optimize the query

Execution model based on mobiles agents

- Each operator is executed by a mobile agent
- Agent can migrate, at run-time, from a mediator to another to execute its operator.
Need for an embedded cost model

• Necessary information to choose the migration site of the agent
  – The set of sites where the agent can migrate
  – The production cost of its operands
  – The cost of its migration
  – The execution cost of its operator

M : Mediator
W : Wrapper

These information are not certainly known by the mediator where the agent is located
Need for an embedded cost model

- Necessary information to choose the migration site of the agent
  - The set of sites where the agent can migrate
  - The production cost of its operands
  - The cost of its migration
  - The execution cost of its operator

These information are not certainly known by the mediator where the agent is located.

M : Mediator
W : Wrapper

MA: JOIN (R3,R5)
Need for an embedded cost model

• Necessary information to choose the migration site of the agent
  – The set of sites where the agent can migrate
  – The production cost of its operands
  – The cost of its migration
  – The execution cost of its operator

These information are not certainly known by the mediator where the agent is located
Extension of mediator cost model

Cost\_hashjoin (R1, R2) = Cost\_p (R1) + Cost\_hash + coût\_p (R2) + Cost\_probe + Cost\_migration

- The time-stamp makes it possible to compare the information embedded in the agent and those of catalogue in order to use the most recent
Components & characteristics

Components:
- The migration space: set of sites where the agent is susceptible to be executed.
- The profiles of its operands (Size, Cardinality, ...)
- The production cost of its operands (read a tuple, ...)

Characteristics:
- It is provided to the agent by the query optimizer
- Each agent has its own embedded cost model
- It is must be small to minimize the migration cost of the agent
Interactions between agent and mediator (1)

The agent interacts with the hosting mediator at two points:

- Installation phase
- Decision phase

Installation phase:
- The agent to the mediator
  - Its operands and their localizations
  - Its migration space and their characteristics

- The agent asks to the mediator to estimate
  - The availability of R2 and the available resources on the sites of its migration space
  - The bandwidth of the network between the agent site and the sites of the space of migration
Interactions between agent and mediator (2)

- **Decision phase**
  - The agent provides to the mediator
    - The revised profiles of its operands
    - The parameters of the production costs of its operands
  - The agent asks to the mediator to estimate
    - The size of the result of the operator
    - The execution cost on every site of the migration space

The mediator computes the metrics of the agent according to the most recent information
Objective is to validate if the agent chooses the most appropriate site

Behavior of mobile agent according to the estimation errors (R1, SF)

Query experimentation : join (R1, R2)

Experimentations
  – Local environment
  – Large scale environment
Local environment (1)

- **Error on \(|R1|\)**

![Graph showing variation on |R1| execution on S2 and proactive execution](image)

\[\text{JOIN (R1,R2)}\]

- R1
- S1
- R2
- S2
Local environment (2)

- Error on SF

JOIN (R1, R2)

R1
S1

R2
S2

![Graph showing response time vs. variation on SF]

- execution on S2
- execution on S1
- proactive execution

<table>
<thead>
<tr>
<th>Variation on SF</th>
<th>Response time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>20</td>
</tr>
<tr>
<td>-10%</td>
<td>19</td>
</tr>
<tr>
<td>-20%</td>
<td>18</td>
</tr>
<tr>
<td>-30%</td>
<td>17</td>
</tr>
<tr>
<td>-40%</td>
<td>16</td>
</tr>
<tr>
<td>-50%</td>
<td>15</td>
</tr>
<tr>
<td>-60%</td>
<td>14</td>
</tr>
<tr>
<td>-70%</td>
<td>13</td>
</tr>
<tr>
<td>-80%</td>
<td>12</td>
</tr>
<tr>
<td>-90%</td>
<td>11</td>
</tr>
</tbody>
</table>
Large scale environment (1)

- Error on ||R1||

JOIN (R1,R2)

R1

R2

Toulouse  Izmir

Paul Sabatier University
Toulouse III

0%  -10%  -20%  -30%  -40%  -50%  -60%  -70%  -80%  -90%

Variation on ||R1||

25  45  65  85  105  125  145  165  185

Response time (s)

execution on Toulouse
execution on Izmir
proactive execution
Large scale environment (2)

- Error on SF

JOIN (R1,R2)

Toulouse Izmir

Paul Sabatier University Toulouse III

JOIN (R1,R2)
Contributions

- Definition of an embedded cost model in mobile agent

- Definitions of the various interactions occurring between the agent and the hosting mediator.

- Storage of the information in the catalog with a timestamp in order to keep the catalog more up to date
Perspectives

• Define the methods which determine the migration space of agents

• Study the additional cost due to the transfer of the embedded cost model and its impact on the quality of the optimization

• Extension of the performance evaluation.
  – Increasing the number of site of experimentation environments
  – Increasing the complexity of query experimentations.
Thank you