The Role of XML Databases in Intelligent Search and Case-Based Reasoning

Larry Kerschberg
E-Center for E-Business; http://eceb.gmu.edu/
Department of Computer Science and Information and Software Engineering
Volgenau School of Information Technology and Engineering
George Mason University, Fairfax, Virginia, USA

Copyright 2007 by Larry Kerschberg, All Rights Reserved
Databases - An Historical Perspective
CBR at 30,000 feet
CBR Life-Cycle Models
XML/RDF/Semantic Web
Role of XML/RDF/Semantic Web in CBR
XML Databases to Support CBR
CBR Framework for Collaborative Semantic Search - Knowledge Sifter
CBR Meets 2.0 Challenge
Conclusions
## Databases - An Historical Perspective

<table>
<thead>
<tr>
<th>Years</th>
<th>Models</th>
<th>Systems</th>
<th>Players</th>
<th>Conferences Journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>Network (CODASYL) Hierarchical</td>
<td>IDMS - GE/Honeywell IMS from IBM System 2000</td>
<td><strong>Charles Bachman</strong>&lt;sup&gt;*&lt;/sup&gt; Edgar Sibley Mike Senko</td>
<td>SIGFIDET</td>
</tr>
<tr>
<td>1970s</td>
<td>Relational Model Entity/Relationship Functional Model</td>
<td>Ingres Oracle (Prototypes)</td>
<td><strong>E.F (Ted) Codd</strong>&lt;sup&gt;*&lt;/sup&gt; Stonebraker Larry Ellison, Peter Chen Kerschberg, Shipman</td>
<td>SIGMOD ACM TODS IEEE TKDE</td>
</tr>
<tr>
<td>1990s</td>
<td>Semi-structured Data Mining Scientific DB</td>
<td>Oracle, DB2, Informix Tamino (XML)</td>
<td>Many authors</td>
<td>JIIS SIGKDD Data Mining &amp; KD</td>
</tr>
<tr>
<td>2000s</td>
<td>Bioinformatics Stream Processing Event-Driven DBs RDF Databases</td>
<td>Aurora (Brown Univ.) Streambase</td>
<td>Zdonik Stonbraker</td>
<td>Journal of Data Semantics</td>
</tr>
</tbody>
</table>

*ACM Turing Award Winners
Hall of Luminaries

Charles Bachman
Ted Codd
Michael Stonebraker
Larry Ellison
Peter Chen
Stefano Ceri
Jennifer Widom
Steve Jobs
CBR at 30,000 Feet
CBR Life-Cycle Models
CBR Life Cycle Model

Retrieve

Problem

New Case

Learned Case

Previous Cases

Retrieved Case

New Case

Retain

General Knowledge

Reuse

Tested/Repaired Case

Confirmed Solution

Solved Case

Suggested Solution

Revise

From Aamodt & Plaza, 1994
CBR Model with Maintenance

From Roth-Berghofer and Iglezakis, GWCBR 2001
XML, RDF, and Semantic Web
Role of Metadata

- Metadata is data about data - describes the data and how it should be interpreted
  - Text, Numbers;
  - Class, Property, Task, etc.

- Metadata may be embedded within a document (e.g., tags) or external to the document (e.g., Relational DB Schema or a shared ontology).

- Embedded metadata provides the context and meaning for the data.

- Data DNA - Data knows everything that will possibly happen to it.
Metadata Standards Initiatives

- Dublin Core for library and Intellectual Property - hosted by OCLC in Dublin, Ohio
- **XML** - Extensible Markup Language
  - Provides the syntax for tagging document
  - XML Schema, XSLT, XML Protocol (SOAP)
- **RDF** - Resource Description Framework
  - Markup of Web resources, binary relations.
- Web Services and the Semantic Web
  - View the Web as a distributed information space
  - Allow computers, programs and agents to communicate in peer-to-peer using standard protocols.
## Dublin Core Metadata Types

<table>
<thead>
<tr>
<th>Content</th>
<th>Intellectual Property</th>
<th>Instantiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Creator</td>
<td>Date</td>
</tr>
<tr>
<td>Subject</td>
<td>Publisher</td>
<td>Format</td>
</tr>
<tr>
<td>Description</td>
<td>Contributor</td>
<td>Identifier</td>
</tr>
<tr>
<td>Type</td>
<td>Rights</td>
<td>Language</td>
</tr>
<tr>
<td>Source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The layered Semantic Web will have successive layers of knowledge, reasoning, learning, and trust.
XML, RDF and Relatives

XML (eXtensible Markup Language) is a markup language which indicates the meaning of the marked-up text.

- Differs from HTML which deals with the presentation of information.

- XML is really a *meta-language*, a mechanism for representing other languages in a standardized way.

- The *interpretation*, *i.e.*, *the meaning*, of the tags is left to the community which uses that markup language.
RDF - Resource Description Framework

- RDF is a meta-model to describe “things” on the Web.

  - Things are resources in the RDF vocabulary.

- RDF model deals with:

  - **Resources** - a thing on the Web

  - **Properties** - a specific aspect, characteristic, attribute or relation the describes a resource.

  - **Statements** - consists of a specific resource, with a named property together with that property’s value.

  - The value can be either a resource or a literal (free text).
The RDF data model is defined as follows:

- There is a set of Resources.
- There is a set of Literals.
- There is a subset of Resources called Properties.
- There is a set of Statements, each element of which is a triple of the form:

  \[ \{\text{pred, sub, obj}\} \]

  Where pred is a property (member of Properties), sub is resource (member of Resources) and obj is either a resource or a literal (member of Literals).

RDF Schema - allows RDF resources to be typed.
Ontology

- **Ontology** is defined as the "science or study of being", *Oxford English Dictionary*

- Ontology building involves identifying the domain objects, their relationship to one another.

- **Semantic Web** researchers consider an ontology to consist of:
  - A set of knowledge terms, which includes the vocabulary,
  - the semantic interconnections, and
  - some rules of inference and logic for some particular domain of discourse.
Web Services Protocols
UDDI, WSDL and SOAP
Role of XML, RDF, and Semantic Web in CBR
XML Specification of a Case

From Coyle, Hayes, Cunningham, Representing Cases for CBR in XML, ICCBR, 1999

<case name="DUB-OSL #34">
  <features>
    <username>Coyle</username>
    <traveloffer>
      <origin>DUB</origin>
      <destination>OSL</destination>
      <departuretime>Mon, 2 Dec 2002 at 6:45 GMT</departuretime>
      <arrivaltime>Mon, 2 Dec 2002 at 12:00 CET</arrivaltime>
      <distance>1051</distance>
      <flighttime>255</flighttime>
      <hops>
        <numberofhops>2</numberofhops>
        <hop>
          <origin>DUB</origin>
          <destination>AMS</destination>
          <carrier>KLM</carrier>
          <departuretime>Mon, 2 Dec 2002 at 6:45 GMT</departuretime>
          <arrivaltime>Mon, 2 Dec 2002 at 9:20 CET</arrivaltime>
          <class>Coach</class>
        </hop>
        <hop>
          <origin>AMS</origin>
          <destination>OSL</destination>
          <carrier>KLM</carrier>
          <departuretime>Mon, 2 Dec 2002 at 10:10 CET</departuretime>
          <arrivaltime>Mon, 2 Dec 2002 at 12:00 CET</arrivaltime>
          <class>Coach</class>
        </hop>
      </hops>
    </traveloffer>
    <recommendation>5</recommendation>
  </features>
</case>
Definition of a Case in CaseML

- CaseML - a Case Markup Language (See Chen and Wu, ICCBR03, 2006)

- Classes in CaseML
  - **CaseBase** - class which acts as a container for cases;
  - **Case** - has one problem description and one solution description;
  - **Problem** - One problem has one or more features;
  - **Feature** - Feature contains *attribute-value pairs* or objects that are described by *domain ontologies*.
  - **Solution** - One solution has one or more features
  - **Similarity Assessment** - The class encapsulates the detail about how the case contained in this CaseBase would be assessed.
Definition of a Case in CaseML

Properties in CaseML

- **domainOntology** - one case base belongs to one domain which has a URL that points to its definition;
- **hasProblem** - this property establishes the relationship between the Case and Problem classes;
- **hasDescription** - relates Feature class to domain ontology;
- **hasSolution** - relationship between the Case and Solution classes.
- **hasSimilarityAssessment** - property points to multiple Similarity Assessment classes, indicating multiple assessment algorithms;
- **hasMeasureMethod** - relates the SimilarityAssessment class to specific measurement method identified as a Web resource.
- **hasAdaptationRule** - relates Feature class with RuleML-specified adaptation rule.
Structure of a Case

Namespaces:
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w4.org/2000/01/rdf-schema#"
xmlns:caseml="http://grid.zju.edu.cn/caseml#"
CBR Life Cycle and CBROnto Task Structure

Role of XML & RDF Databases in Support of CBR
Role of Databases in CBR

- Direct support for two main life-cycle tasks:
  - **Retrieve**
    - Use query languages such as XML-based XQuery, SQL/XML, or RDF-Based SPARQL, RDQL, RQL, etc.
  - **Retain**
    - Store large collection of “case” instances in formats such as: relational, native-XML, or RDF.
    - Create indexes to allow fast retrieval of cases based on features, context, etc.
XML Databases

- Native XML Databases
  - Defines a logical model for an XML document, versus defining just the data in the document.
  - Model must include elements, attributes, PCDATA, and document order. Examples include the XPath DM and Document Object Model (DOM).
  - Document-based storage - entire document can be stored and retrieved
  - Node-based storage - individual nodes of the document stored and retrieved.

- Vendors
  - Berkeley DB from Oracle,
  - Tamino by Software AG
RDF Databases

- Oracle Spatial 10g includes an open, scalable, secure and reliable RDF management platform. Based on a graph data model, RDF triples are persisted, indexed and queried, similar to other object-relational data types.

- IBM’s Web Ontology Manager is a lightweight, Web-based tool for managing ontologies expressed in Web Ontology Language (OWL).

- IBM’s IODT, IBM’s toolkit for ontology-driven development.

- IBM Semantic Layered Research Platform - IBM SLRP is a family of open-source Semantic Web software components including an enterprise RDF store, query engine, web application framework, RCP development libraries, etc.

- SemWeb for .NET supports persistent storage in MySQL, Postgre, and Sqlite; has been tested with 10-50 million triples; supports SPARQL.
CBR Meets Web 2.0 Challenge

Distributed Heterogeneous Collaborative Filtering for Case Discovery and Learning
Combine web-based authoritative (recommender, collaborative) sources:
- Amazon (Books); iTunes (Music); Netflix and IMDB for Movies; ... 

Access Web 2.0 Collaborative Markup Applications
- Wikipedia - Collaborative Encyclopedia
- Delicious for tagged URLs
- Flikr for pictures

The entire Internet and Web constitute the Case Base.
Search for emergent case patterns by querying the markup tags across heterogeneous domains.
Create a semantic web of concepts from a domain model
Conclusions

- Databases have not played major role in CBR, partly because the case bases have been small.

- However, XML, RDF, and the Semantic Web will change this and the CBR community should explore the use of DBMS to support the REs Live Cycle.

- Extend CBR to resource discovery in Web 2.0 - your new Case Base

- Invitation to the CBR community for a Special Issue of the Journal of Intelligent Information Systems.
The Functional Approach to Data Management: Modeling, Analyzing and Integrating Heterogeneous Data

Peter Gray
Larry Kerschberg
Peter King
Alex Poulovassilis

Springer
Unabashed Journal Plug

Journal of Intelligent Information Systems
Integrating Artificial Intelligence and Database Technologies

Editors-in-Chief
Larry Kerschberg
Zbigniew Ras
Maria Zemankova
Knowledge Sifter Meta-Model