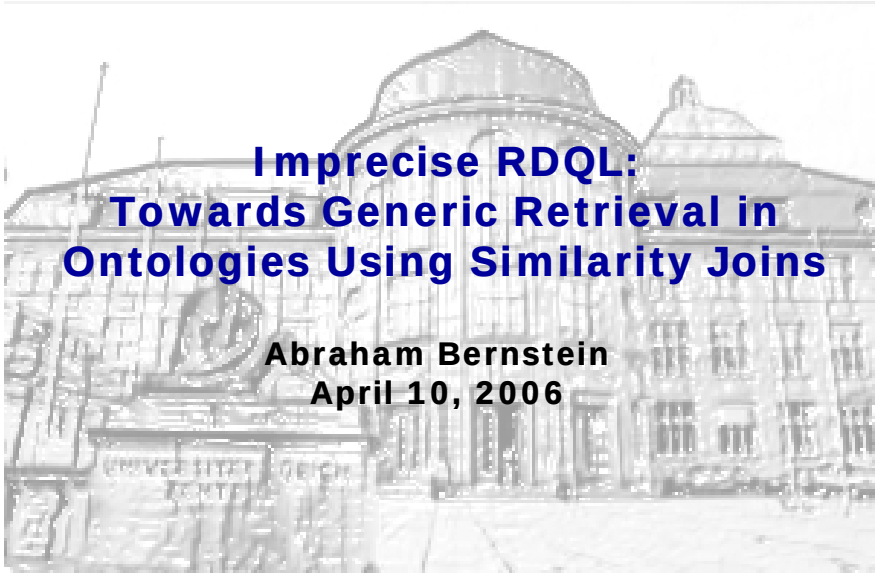
Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins



Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Abraham Bernstein
April 10, 2006

University of Zurich DIS Dynamic and Distributed Information Systems

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Agenda

- ***The Problem*:**
  Realistic Querying and the Semantic Web
- ***A Possible Solution*:**
  Similarity Joins
- ***So what is Similarity*:**
  A brief Excursion into Similarity
- ***The Evaluation*:**
  Retrieval Performance in OWL-S
- ***Limitations and Conclusions***

University of Zurich DIS Dynamic and Distributed Information Systems

2

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Surfing the Beach

- Questions:
  - Where to rent a surf board?
  - Which beach service offers best prices?
  - Where to find the biggest waves?

University of Zurich DIS Dynamic and Distributed Information Systems

3

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

The Web is a …

- Distributed, interlinked repository of Documents
- Everybody can add and/or remove their documents and links anywhere
- Problems:
  - It is in (many different) natural languages → difficult to process by machines
  - Difficult to query
  - There is no trust, authentication, authorization built into the infrastructure

University of Zurich DIS Dynamic and Distributed Information Systems

4

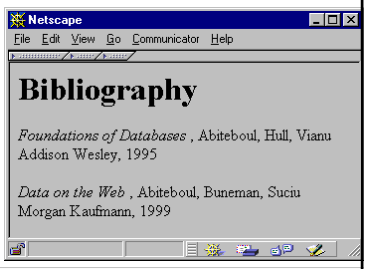
The Semantic Web is a …

- Distributed, interlinked repository of *Knowledge Bases (KBs)*
- Everybody can add and/or remove their KBs and interlink them
- Advantages:
  - It uses standardized concepts from Knowledge Representation (KR)
    → easy to process by machines
  - "Trivial" to query
  - Trust, authentication, authorization built into the infrastructure are achieved by clever use of logic
- Problems:
  - There is no guarantee of consistency between KBs

Example Web vs. Sem. Web

Web

```
<h1> Bibliography </h1>
<p> <i> Foundations of Databases </i>
      Abiteboul, Hull, Vianu
      <br> Addison Wesley, 1995
<p> <i> Data on the Web </i>
      Abiteboul, Buneman, Suciu
      <br> Morgan Kaufmann, 1999
...
```



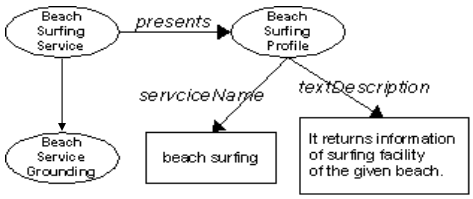
Sem. Web

- KB 1:
  - penguin(Pingu)
  - ∀ x: penguin(x) ⇒ bird(x)
- KB2:
  - ∀ y: bird(y) ⇒ can_fly(y)
- can_fly(Pingu)?

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

RDQL – RDF Data Query Language

- **Jena – Implementing the Semantic Web Recommendations** [Carroll *et al.* '03]




```
SELECT ?s1,?p1
WHERE (?s1 presents ?p1)
      (?p1 serviceName "beach surfing")
```

?s1	?p1
Beach Surfing Service	Beach Surfing Profile

University of Zurich DIS Dynamic and Distributed Information Systems

7

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

iRDQL – Extending RDQL With Similarity Joins

- **3 additional la**
  - **IMPRECISE,**

So how do we define similarity?

```
SELECT ?s1,?p1,
WHERE (?s1 pres
      (?p2 servi
IMPRECISE ?p1,?p2
SIMMEASURE Levenshtein
OPTIONS IGNORECASE false THRESHOLD 0.7
```

Similarity Join

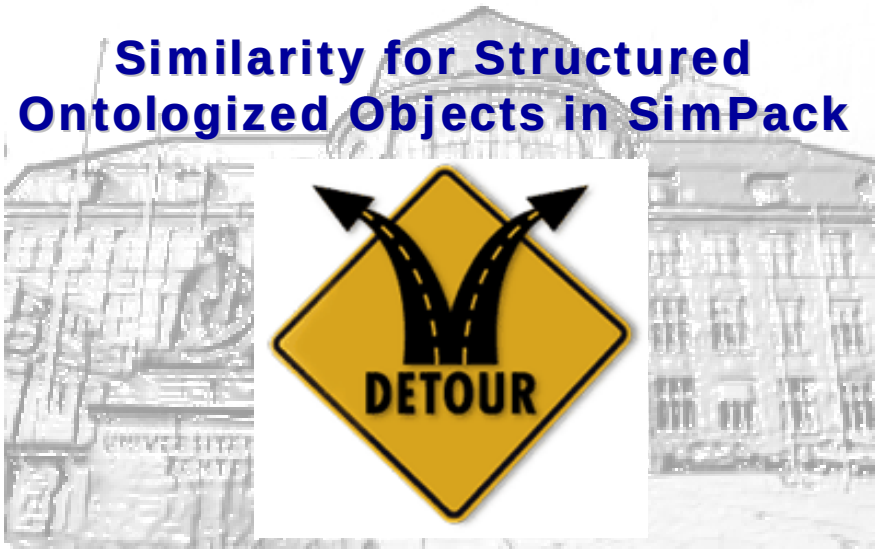
?s1	?p1	?p2	sim
Beach Surfing Service	Beach Surfing Profile	Beach Surfing Profile	1.0
Beach Broker Service	Beach Broker Profile	Beach Surfing Profile	0.85
Abstract Broker Service	Abstract Broker Profile	Beach Surfing Profile	0.7

University of Zurich

8

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Similarity for Structured Ontologized Objects in SimPack



University of Zurich DIS Dynamic and Distributed Information Systems

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Current Status

- Generic Java library of similarity measures for the use in ontologies
- Measures from the literature, adapted for ontologies
  - Computer Science, AI, Psychology, Linguistics, ...
- 41 measures in total of all similarity measure categories
- Similarity measure categories
  - feature vectors (8)
    - properties of objects (OWL ObjectProperty, DatatypeProperty, ...)
  - strings or sequences of strings (24)
    - textual description of objects
  - trees and graphs (6)
    - tree/graph comparison
  - information theory (3)
    - amount of information contained in objects

University of Zurich DIS Dynamic and Distributed Information Systems

10

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Measures

- **Featur vector-based (8)**
  Cosine, Dice, Jaccard, Overlap, …

Mapping (e.g.)

$$\mathbf{R_x} = \{type, name\} \Rightarrow \mathbf{x'} = \begin{pmatrix} 0 \\ name \\ type \end{pmatrix} \Rightarrow \mathbf{x} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathbf{R_y} = \{type, age\} \Rightarrow \mathbf{y'} = \begin{pmatrix} age \\ 0 \\ type \end{pmatrix} \Rightarrow \mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Cosine

$$sim_{cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_2 \cdot ||\mathbf{y}||_2} = \frac{0*1+1*0+1*1}{\sqrt{0^2+1^2+1^2} * \sqrt{1^2+0^2+1^2}} = \frac{1}{2}$$

University of Zurich DIS Dynamic and Distributed Information Systems

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Measures

- **String-based (24)**
  Levensthein, …

Mapping (extract terms from service file, e.g. from surfing_beach_service.owls)

$\mathbf{x} = [BeachSurfingProfile, hasInput, Beach, label, BEACH, hasOutput, Surfing]$

$\mathbf{y} = [BeachBrokerProfile, hasInput, Beach, label, BEACH, hasOutput, Broker]$

$$sim_{lev}(Rx, Ry) = \frac{xform(\mathbf{x}, \mathbf{y})}{xform\_wc(\mathbf{x}, \mathbf{y})} = \frac{3}{9}$$

Conversion from distance to similarity

$$\frac{1}{1+\frac{3}{9}} = 0.75$$

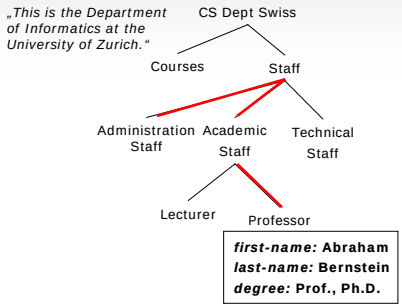
University of Zurich DIS Dynamic and Distributed Information Systems

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Measures

- **Graph-/Tree-based (6)**
  Shortest Path, Tree Edit Distance, …

Shortest path between 'Administration Staff' and 'Professor'



➡ 3

Conversion from distance to similarity

➡ $-1.0 * \log\left(\frac{3}{2.0 * 4}\right)$

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Measures

- **Information Theory (3)**
  Resnik, Lin, …

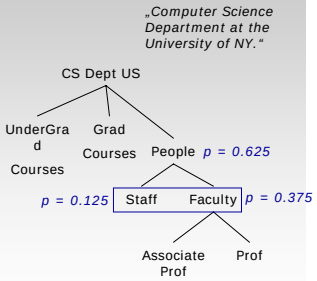
Most Recent Common Ancestor (MRCA) of 'Staff' and 'Faculty' is 'People'

$$sim_{resnik}(R_x, R_y) = -\log_2 P(MRCA(R_x, R_y))$$

➡ $\log_2\left(P(MRCA(Staff, Faculty))\right) = \log_2(0.625)$

$$sim_{lin}(R_x, R_y) = \frac{2\log_2 P(MRCA(R_x, R_y))}{\log_2 P(R_x) + \log_2 P(R_y)}$$

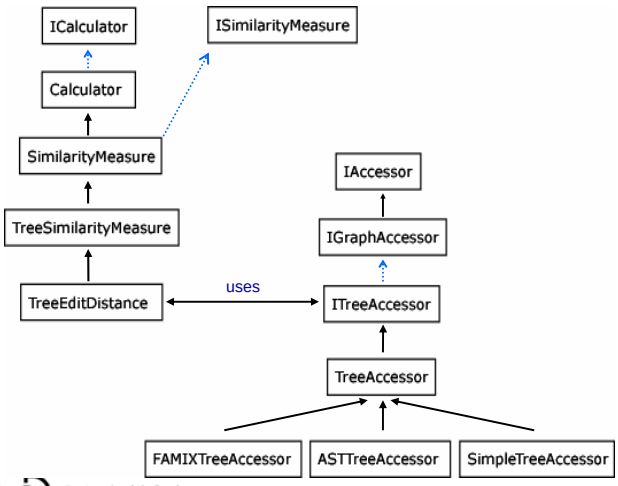
➡ $\frac{2\log_2 P(MRCA(Staff, Faculty))}{\log_2 P(Staff) + \log_2 P(Faculty)} = \frac{2\log_2(0.625)}{\log_2(0.125) + \log_2(0.375)}$



Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Architecture

- **Measures and data accessor**

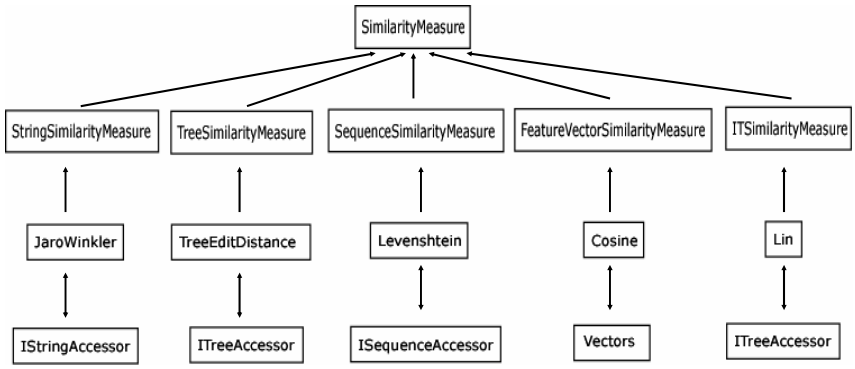


University of Zurich DIS Dynamic and Distributed Information Systems

15

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

SimPack—Architecture



- **Every measure category uses its own data accessor**
- **To use SimPack, write an appropriate data accessor for your data format**

University of Zurich DIS Dynamic and Distributed Information Systems

16

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

iRDQL –Evaluation Approach

- **Quantitative** evaluation using an OWL-S service retrieval test collection **[Klusch '05]**
- OWL-S Service-based **Precision, Recall** and **F-Measure** as performance measures

Precision / Recall / F-Measure:

$$prc = \frac{retrieved\,relevant\,services}{retrieved\,services}$$

$$rec = \frac{retrieved\,relevant\,services}{relevant\,services}$$

$$f = \frac{2 \cdot prc \cdot rec}{prc + rec}$$

University of Zurich DIS Dynamic and Distributed Information Systems

17

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

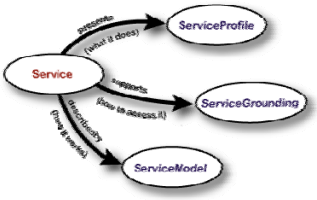
OWL-S Service Retrieval Test Collection

- 406 OWL-S services of 6 different domains
- 9 queries together with its correct answers

Query

```
<?xml version="1.0" encoding="UTF-8"?>
<service:Service rdf:ID="CITY_BROKER_SERVICE">
   <service:presents rdf:resource="#CITY_BROKER_PROFILE"/>
   <service:describedBy rdf:resource="#CITY_BROKER_PROCESS_MODEL"/>
   <service:supports rdf:resource="#CITY_BROKER_GROUNDING"/>
</service:Service>

<profile:Profile rdf:ID="CITY_BROKER_PROFILE">
   <service:isPresentedBy rdf:resource="#CITY_BROKER_SERVICE"/>
   <profile:serviceName xml:lang="en">
      hotel reservation booking service
   </profile:serviceName>
   <profile:textDescription xml:lang="en">
      Provide the best hotel reservation system in a given city.
   </profile:textDescription>
   <profile:hasInput  rdf:resource="#_CITY"/>
   <profile:hasOutput rdf:resource="#_BROKER"/>
   <profile:has_process rdf:resource="CITY_BROKER_PROCESS" />
</profile:Profile>
```



http://www.w3.org/Submission/OWL-S/

Query Relevance Set

```
city_broker_service.owls
city_broker_service2.owls
city_financial_agent_service.owls
city_financial_agent_service1.owls
urbanarea_financial_agent_service.owls
city_organization_service.owls
...
```

University of Zurich DIS Dynamic and Distributed Information Systems

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Complete iRDQL Query

```
SELECT ?s, ?p, ?p1, ?m, ?m1

WHERE (?s rdf:type sv:Service)
      (?s sv:supports ?g)
      (?g rdf:type gr:WsdlGrounding)
      (?g sv:supportedBy ?s)
      (?s sv:presents ?p1)
      (?p1 rdf:type pr:Profile)
      (?p1 sv:isPresentedBy ?s)
      (?s sv:describedBy ?m1)
      (?m1 rdf:type px:ProcessModel)
      (?m1 sv:describes ?s)
      (?p rdf:type pr:Profile)
      (?p sv:isPresentedBy ?s1)
      (?s1 rdf:type sv:Service)
      (?p pr:serviceName ?sn)
      (?p pr:textDescription ?sd)
      (?p pr:hasInput ?in1)
      (?p pr:hasOutput ?out1)
      (?in1 px:parameterType ?in1PT)
      (?in1 rdfs:label ?in1L)
      (?out1 px:parameterType ?out1PT)
      (?out1 rdfs:label ?out1L)
      (?m rdf:type px:ProcessModel)
      (?m sv:describes ?s2)
      (?s2 rdf:type sv:Service)
      (?m pr:hasProcess ?x)
      (?x rdf:type px:AtomicProcess)
      (?x px:hasInput ?in2)
      (?x px:hasOutput ?out2)
      (?in2 px:parameterType ?in2PT)
      (?in2 rdfs:label ?in2L)
      (?out2 px:parameterType ?out2PT)
      (?out2 rdfs:label ?out2L)
AND ?sn =~ /beach surfing/i
AND ?sd =~ /It returns information.../i
AND ?in1 =~ /_BEACH/
AND ?out1 =~ /_SURFING/
AND ?in1 eq ?in2
AND ?out1 eq ?out2

USING sv for <Service.owl#>
      pr for <Profile.owl#>
      px for <Process.owl#>
      gr for <Grounding.owl#>

IMPRECISE ?p, ?p1
IMPRECISE ?m, ?m1

SIMMEASURE Levenshtein
OPTIONS IGNORECASE false THRESHOLD 0.7;
```

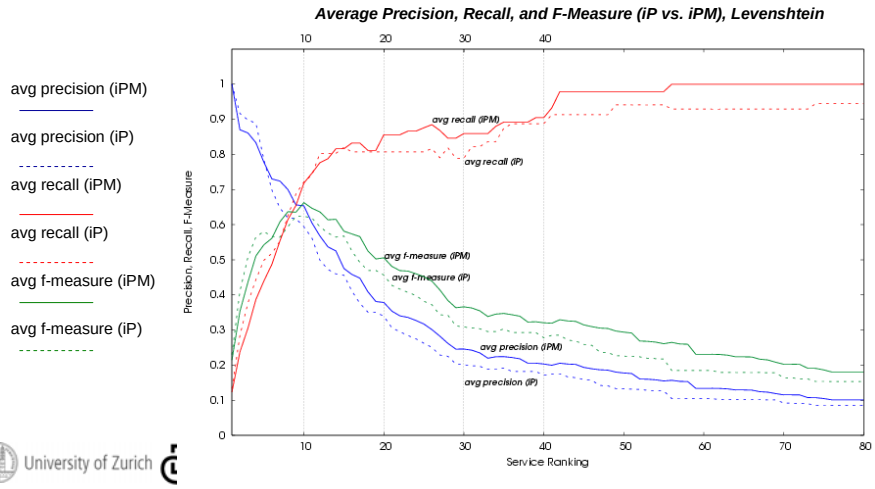
University of Zurich DIS Dynamic and Distributed Information Systems

19

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

iRDQL – Performance (*iP* vs. *iPM*)

- ***iP***: **IMPRECISE ?p1, ?p2**
- ***iPM***: **IMPRECISE ?p1, ?p2 & IMPRECISE ?m1, ?m2**

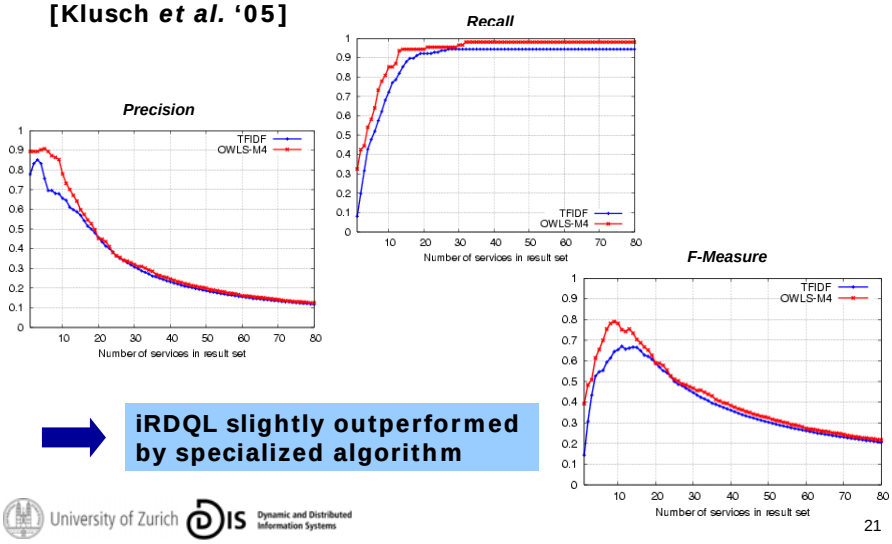


University of Zurich

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

iRDQL – Performance (iRDQL vs. OWLS-M4)

- **OWLS-M4: Matchmaking algorithm of OWLS-MX [Klusch *et al.* '05]**



iRDQL slightly outperformed by specialized algorithm

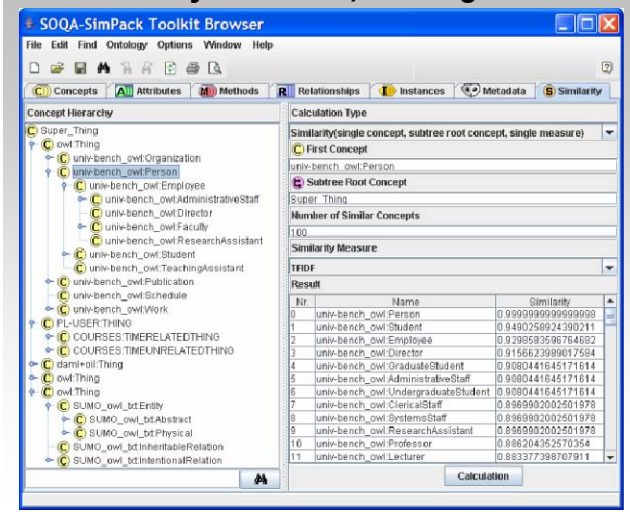
University of Zurich DIS Dynamic and Distributed Information Systems

21

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Other things to do with SimPack: SST—SOQA SimPack Toolkit

- **Similarity enabled, Protégé-like ontology browser**



1/ P. Ziegler, C. Kiefer, C. Sturm, A. Bernstein & K. Dittrich, **The SOQA-SimPack Toolkit**, 10th European Conference on Database Technology (EDBT), Munich, Germany. 2006.

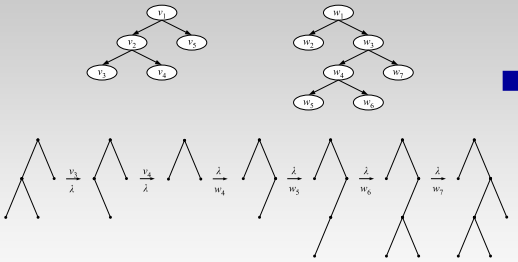
2/ P. Ziegler, C. Kiefer, C. Sturm, K. Dittrich, and A. Bernstein. **Generic Similarity Detection in Ontologies with the SOQA-SimPack Toolkit (Demo Paper).** To appear in 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD 2006). Chicago, USA, June 26-29, 2006.

University of Zurich DIS Dynamic and Distributed Information Systems

22

Other things to do with SimPack:
Detecting Similar Java Classes

- **Measures the tree edit distance between source code ASTs**



Edit distance is 6

(2x delete, 4x insert)

Conversion to similarity

$$\frac{(|T_1|+|T_2|)-\#ed}{|T_1|+|T_2|}=\frac{12-6}{12}=\frac{2}{3}$$

Tobias Sager, Abraham Bernstein, Martin Pinzger, Christoph Kiefer. **Detecting Similar Java Classes Using Tree Algorithms**. To appear in MSR '06: Proceedings of the 2006 International Workshop on Mining Software Repositories, China, Shanghai, May 22-23, 2006.

University of Zurich DIS Dynamic and Distributed Information Systems

23

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Other things to do with SimPack:
Detecting Similar Java Classes

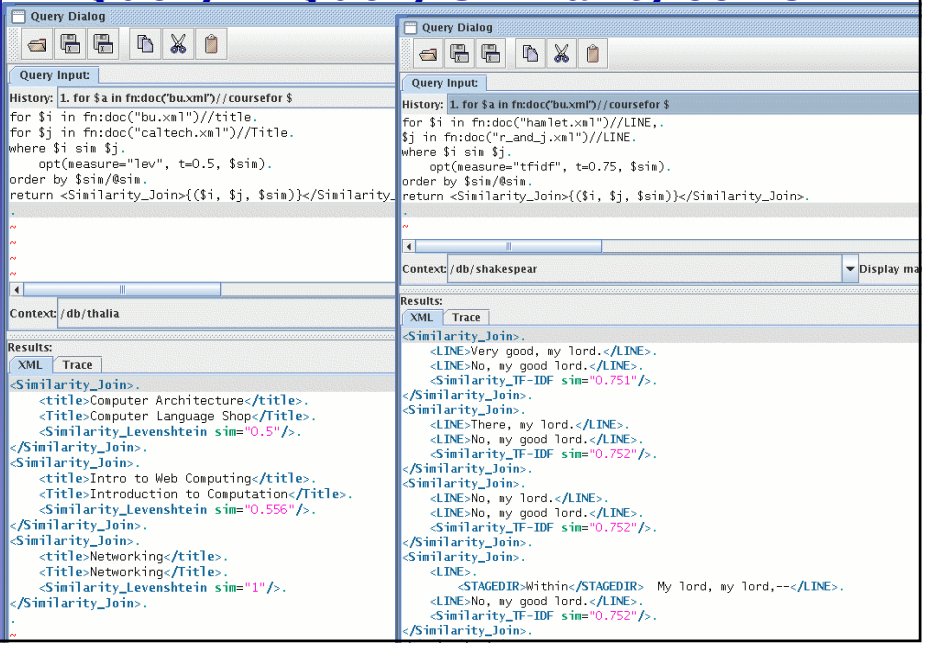
- **Visualize a software project's code evolution steps**
- **org.eclipse.compare version 3.0 vs. 3.1**
- Diagonal shows classes that have changed/kept them same between versions
- Horizontal/vertical show similarities between different classes

University of Zurich DIS Dynamic and Distributed Information Systems

24

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Other things to do with SimPack: iXQuery—XQuery Similarity Joins



Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Conclusions and Outlook

- **Similarities are (surprisingly) useful**
- **Combination of RDQL and similarity measures**
- **Generic IR-based approach only slightly outperformed**
- **Inspired by Cohen's work [Cohen '00]**
  - **no flat tables**
  - **aggregated (ontologized) objects**
- **Performance improvements**
- **Switch to SPARQL**

University of Zurich DIS Dynamic and Distributed Information Systems

26

Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins

Thank you…

- **iRDQL makes it possible…**

University of Zurich DIS Dynamic and Distributed Information Systems

27