PhD Dissertation:

Integration of SDI Services: an evaluation of a distributed semantic matching framework

Lorenzino Vaccari

April 28, 2009

1. Interoperability in Spatial Data Infrastructures (SDIs)
   - The SDI phenomenon
   - Information systems' interoperability
   - State of the art
2. A P2P semantic matching framework
   - Motivating scenario
   - Supporting the scenario: the OpenKnowledge (OK) system
   - Matching in OK
   - SDI services implementation
3. SPSM Evaluation
   - Final evaluation: two experiments
   - Evaluation of Structure Preserving Semantic Matching
   - Evolution experiment
   - Classification experiment
   - Performance evaluation
4. Conclusions and future work
   - Conclusions
   - Future work

Integration of geo-information

The Digital Earth initiative

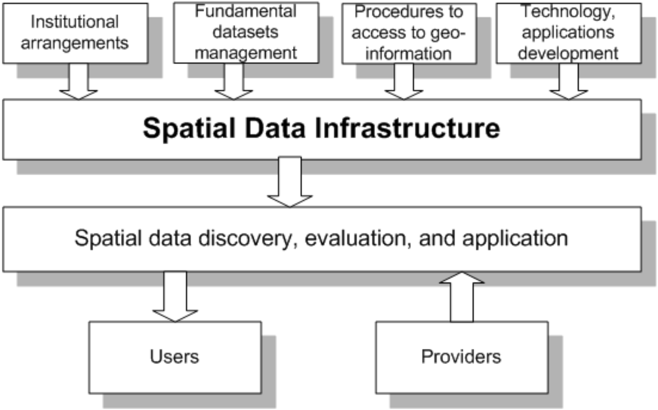
- First introduced by Al Gore US vice president in 1998
- Requirements:
  - Computational Science
  - Mass storage
  - Satellite images
  - Broadband networks
  - Metadata
  - **Interoperability**

Motivation

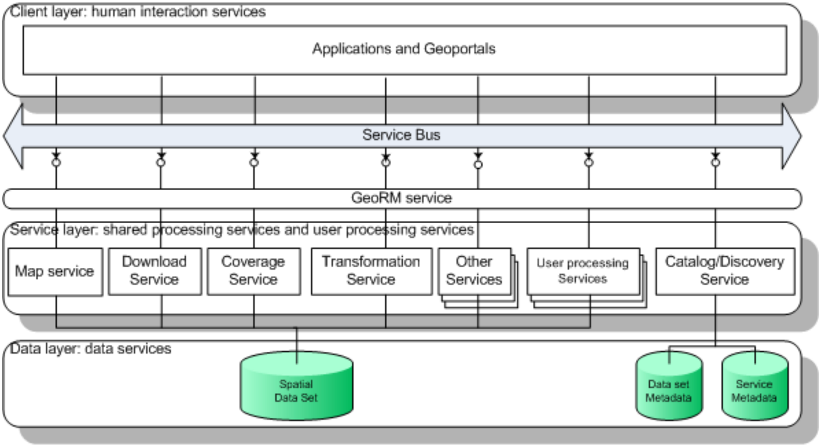
Initiatives for collection and dissemination of Geographical data

- Shared Environmental Information System (SEIS)
- Infrastructure for Spatial Information in Europe (INSPIRE)
- Global Earth Observation System of Systems (GEOSS)
- Global Monitoring for Environment and Security (GMES)

Spatial Data Infrastructure (SDI) components



SDI technological implementation



Heterogeneity of geo-data

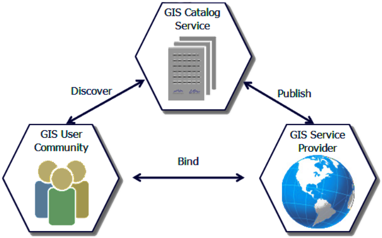
Geo-data heterogeneity

- Different syntax
- Different structure
- Different semantics
- Specifically for geo-data
  - Different precisions, lineage methods ⇒ Integration alignment issues
  - Different topological models of the same Earth's feature
  - Different representation formats (e.g. raster, vectorial)

Geo-service interoperability

Geo - Service Oriented Architecture

- Open Geospatial Consortium specifications
  - Geo-metadata: ISO 19115/ISO19139
  - Geo-Catalog: CAT
  - Geo-Services:
    - Web Map Service (WMS),
    - Web Feature Service (WFS),
    - Gazetteer (WFS-G),
    - ...



Geo-service heterogeneity

Characteristics

- Discovering and integrating services is difficult task
- Usually invocation of a service: described in terms of its structure and data schema specifications
- Formal description of its functionality and the meaning of data are often missing
- Automatic composition: only the syntactical structure of the service can be verified
- Specifically for geo-services
  - Geography based information
  - Maps as implicit interfaces
  - Specific topological operations

Geo-service semantic heterogeneity

Characteristics

- At present: no standard notions are used for defining the semantics of a geographic web service
- *In today's GIS service architectures, the interfaces between agents, computational and human, are those of web services...* and...*the interface of a service is formally captured by its signature* (Kuhn, 2005)
- Signatures (name, inputs and outputs) of web services ⇒ tree-like structures/simple ontologies
- The terms of these tree-like structures implicity contain a classification of the background knowledge of the provider

State of the art

Geo-information integration

- Syntactic and structural aspects: Open Geospatial Consortium (OGC) standards
- Semantic aspects:
  - Various approaches use a central ontology to reduce the semantic heterogeneity problem
  - Semantic heterogeneity problem ⇒ problem of reasoning within the shared ontology

Ontology matching

- Techniques from different fields (e.g., statistics and data analysis, machine learning, linguistics)
- In our approach services are assumed to be annotated with the concepts taken from various ontologies

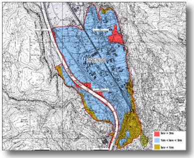
P2P model in GIS application

- P2P model applied to SDIs is very novel: focusses on the ways in which P2P paradigm can be used to support distribution and sharing of spatial information

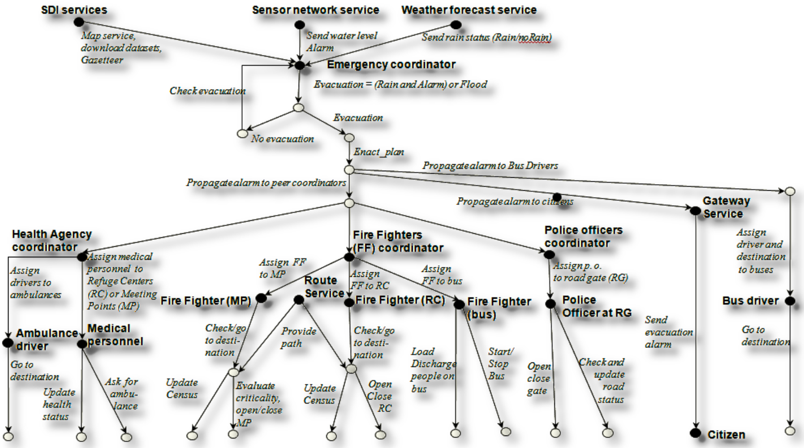
Emergency response (eResponse) scenario

Flooding event in Trento

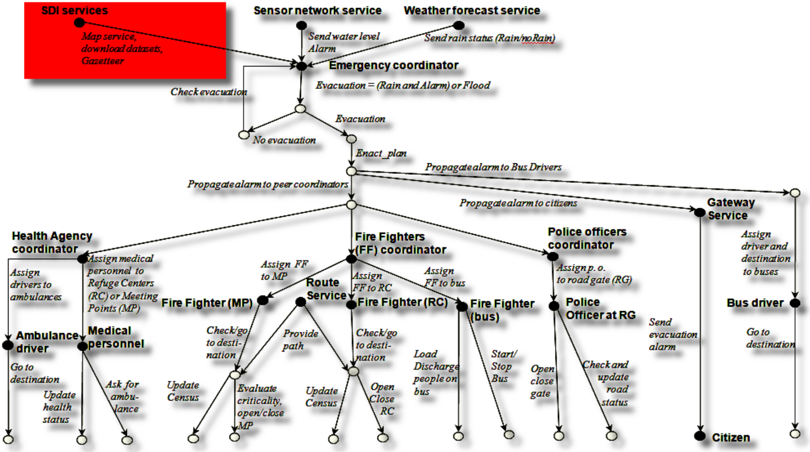
- eResponse scenario for the flooding in Trento (Italy)
- eResponse Coordination based on the Emergency plan of the municipality of Trento
- Main goal: people evacuation
  - We selected a subset of the operations from the plan: the ones related with the evacuation of the people from potential flooding



Overall use case



Overall use case



SDI services

Selection & clustering

- Gazetteer service
- Map request
- Download request

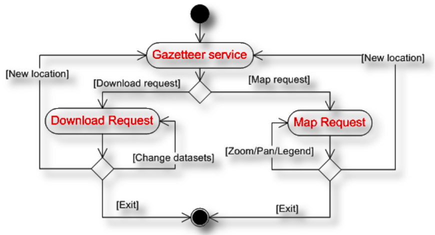


Figure: Clustering SDI services

OpenKnowledge (OK) EU project

Open, distributed, P2P system

- *Interaction-centric* approach: peers share Interaction Models (*IMs*)
- *Semantic P2P* approach:
  - Distributed storage
  - Decentralized address register
  - Symmetric roles of each peer
  - Semantic matching:
    - Discover and compose peer services
    - Locate shared IMs
- Service choreography mechanism: Lightweight Coordination Calculus (LCC) (Robertson, 2004) language

LCC language

LCC characteristics

- Tasks/processes are formalized by Interaction Models (IMs), written in LCC
- IMs written in LCC protocols: workflows
- Uses roles for agents and constraints on message sending to enforce social norms and behaviors

LCC Example

$a(r1, A1)$ ::
  $ask(X) \Rightarrow a(r2, A2) \leftarrow need(X)$ *then*
  $update(X) \leftarrow return(X) \Leftarrow a(r2, A2)$

$a(r2, A2)$ ::
  $ask(X) \Leftarrow a(r1, A1)$ *then*
  $return(X) \Rightarrow a(r1, A1) \leftarrow get(X)$

Figure: Double arrows ($\Rightarrow, \Leftarrow$) indicate message passing between roles, single arrow ($\leftarrow$) indicates constraint satisfaction.

How do we use matching in OK?

Different purposes

- To allow peers (service providers) to determine how similar their own service descriptions are to those required by IM constraints (service invocations)
- To allow peers to understand how they may satisfy the requirements of IM constraints. This is done through building up a map between each element of their service descriptions to each element of IM constraints.
- To discover model of interactions (IMs)

Matching in OK: LCC Example

LCC Example: Map Provider role

$a(ga\_sp, P) ::$
$askMap(Version, Layers, Width,$
$Height, Format, XMin\_BB$
$YMin\_BB, XMax\_BB, YMax\_BB)$
$\Leftarrow a(ga\_sr, R)$ *then*
$returnMap(Map) \Rightarrow a(ga\_sr, R)$
$\leftarrow requestMap(Version,$
$Layers,$
$Width,$
$Height,$
$Format,$
$XMin\_BB,$
$YMin\_BB,$
$XMax\_BB,$
$YMax\_BB, Map)$ *then*
$a(ga\_sp, P)$

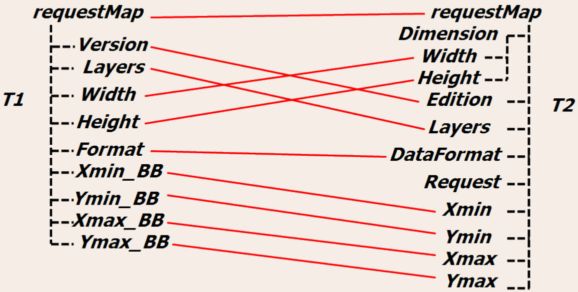
Web service signature

```
public class MapProvider
     extends OKCFacadeImpl{
....
public boolean requestMap{
     Argument
          Dimension(Height,
          Width),
     Argument Edition,
     Argument Layers,
     Argument DataFormat,
     Argument Request,
     Argument Xmin, Ymin,
     Argument Xmax, Ymax,
     Argument Map{
          ...
     }
}
```

Which kind of matching solution ?

Structure Preserving Semantic Matching (SPSM) (Giunchiglia et al., 2008)

$$Similarity(T1, T2) = 0.64$$



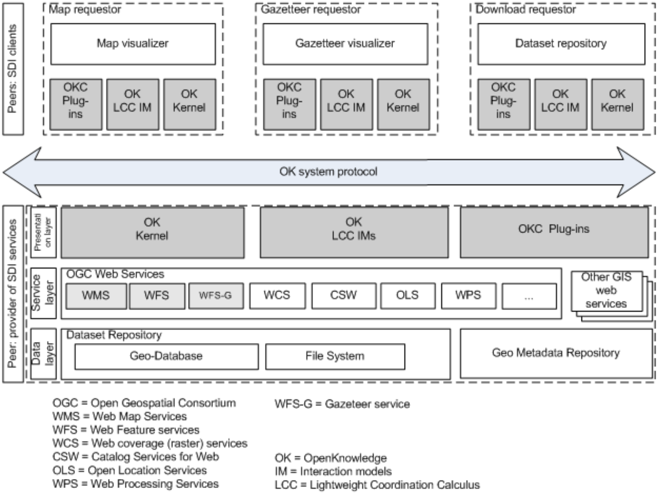
SPSM

Based on

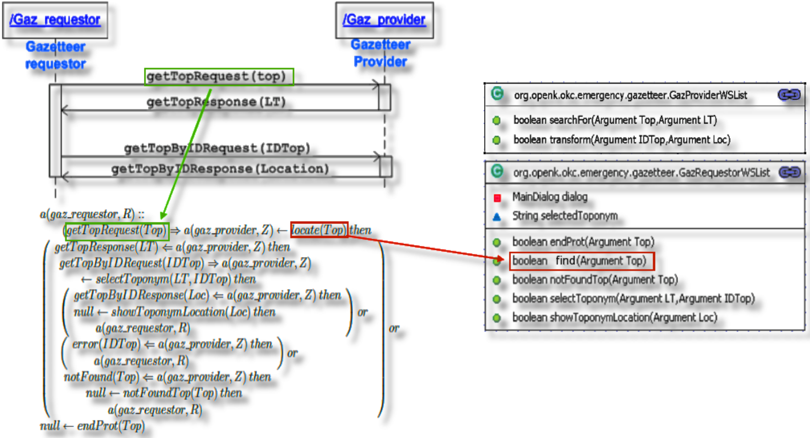
- The S-match algorithm
- A formal theory of abstraction (Giunchiglia & Walsh, 1992). The semantic matching preserve some structural properties (e.g., functions are matched to functions and variables are matched to variables)
- A tree edit-distance algorithm

$$TreeSim(T1, T2) = 1 - \frac{min \sum_{i \in S} n_i \cdot Cost_i}{max(|T1|, |T2|)} \quad (1)$$

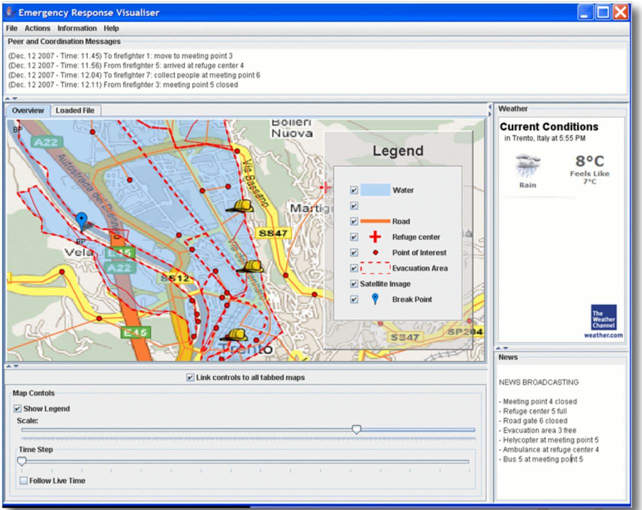
SDI services implementation architecture



Gazetteer service



The emergency GUI



Experiments

Evolution experiment

- How robust is SPSM when ontologies evolve ?
- Syntactic and semantic alteration operations on real world GIS Web service operation signatures
- The probability, assigned to each alteration operation, has been changed from the lower value (0.1) to the maximum value (0.9)

Classification experiment

- Does SPSM retrieve similar web services ?
- Comparison between a manual classification and the one computed by SPSM

Evolution experiment: syntactic and semantic alterations

Evaluation setup: dataset

- 80 trees were built out of the ESRI Geographic web services
- 4 alteration operations + 1 combination: Meaning and syntactic alterations
- 20 alterations for each tree, for each alteration operation and for each probability
    - total matching tasks (including 10 statistical repetitions): ca. 700.000

Evaluation setup: alteration operations

Original signature

find_Address_By_Point(point, address_Finder_Options, part)

1. Replace a node name with an unrelated one (Brown corpus) :
   point → *cable*
2. Add or remove a label in a node name (Brown corpus):
   find_Address_By_Point → *find_By_Point*
3. Alter syntactically a label (add, delete and change characters):
   find_Address_By_Point → *finm_Address_By_Poioat*
4. Replace a label in a node name with a related one (synonyms, hyponyms, hypernyms from Moby and WordNet 3.0):
   address_Finder_Options → *location_Finder_Options*
5. Combination of 3. and 4.:
   address_Finder_Options → *lfctin_Finder_Options*

Evaluation methodology

Modify(AlterationOperation, AlterationProbability, Signature):

$ExpScore \leftarrow 1$
$AltSignature \leftarrow Change(AlterationOperation, AlterationProbability, Signature)$
$ExpScore \leftarrow Decrease(ExpScore, AlterationOperation, AlterationProbability)$

return ExpScore,AltSignature

Recall, precision and F-measure quality measures computation. Ingredients:

- Expected Score: *ExpScore*
- User threshod: *CorrThresh*
- SPSM similarity value: *TreeSim*
- Variable acceptance (cut-off) threshold: *CutoffThresh*
- Results: average on 10 repetitions

Evaluation methodology: quality measures

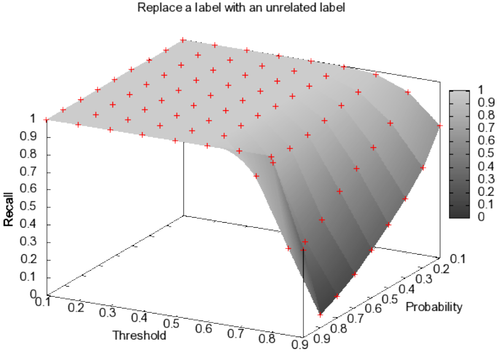
Quality measures

- $R = \{T2 \in AltSignatures \mid TreeSim(T1, T2) \geq CutoffThresh\}$
- $C = \{T2 \in AltSignatures \mid ExpScore(T1, T2) \geq CorrThresh\}$
- $TP = \{T2 \mid T2 \in R \wedge T2 \in C\}$
- $FP = \{T2 \mid T2 \in R \wedge T2 \notin C\}$

Table: Example ($CorrThresh = 0.6, AlterationProbability = 0.7$).

Cut-off threshold	$\lvert C\rvert$	$\lvert R\rvert$	$\lvert TP\rvert$	$\lvert FP\rvert$	$\lvert FN\rvert$	Recall	Precision	F-measure
0.1	593	1598	593	1005	0	1.000	0.371	0.541
0.2	593	1585	593	992	0	1.000	0.374	0.545
0.3	593	1568	593	975	0	1.000	0.378	0.549
0.4	593	1496	593	903	0	1.000	0.396	0.568
0.5	593	1391	593	798	0	1.000	0.426	0.598
0.6	593	758	588	170	5	0.992	0.776	0.871
0.7	593	642	513	129	80	0.865	0.799	0.831
0.8	593	397	315	82	278	0.531	0.794	0.636
0.9	593	143	112	31	481	0.189	0.783	0.304

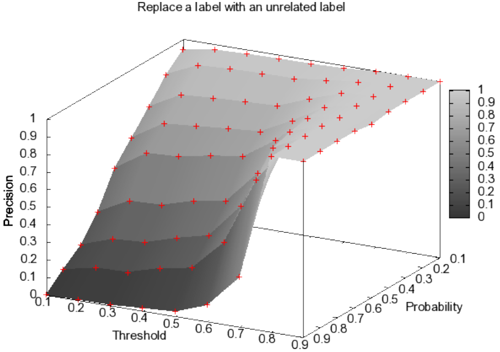
Evaluation results: recall



Replace a node name with an unrelated node name

The SPSM approach retrieves all the expected (relevant) correspondences until the empirically fixed threshold ($corrThresh = 0.6$), that mimics the user's tolerance to errors, is reached

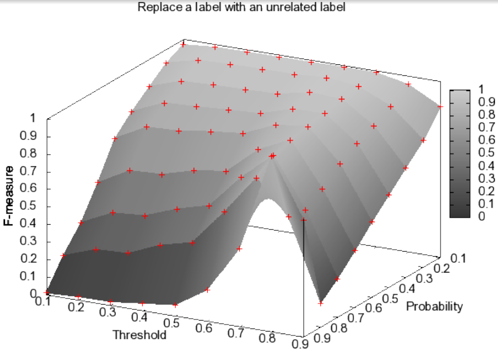
Evaluation results: precision



Replace a node name with an unrelated node name

Precision improves rapidly as the *TreeSim* cut-off threshold exceeds the empirically fixed threshold. Precision decreases steadily as a function of the alterations' probability while the *TreeSim* cut-off threshold is below the empirically fixed threshold

Evaluation results: F-measure



Replace a node name with an unrelated node name

Even when the probability of the alteration is very high the balance between correctness and completeness is good.

For instance, at the optimal *TreeSim* cut-off threshold (0.6), for an important alteration probability of 80%,

F-measure is higher than 74%. These data prove the robustness of the SPSM approach up to significant syntactic

modifications in the node names.

Evaluation results: SPSM vs. Baseline

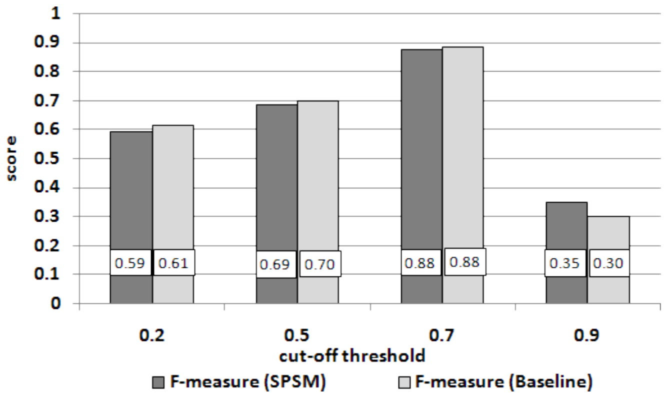


Figure: F-measure: Syntactic alteration

Evaluation results: SPSM vs. Baseline

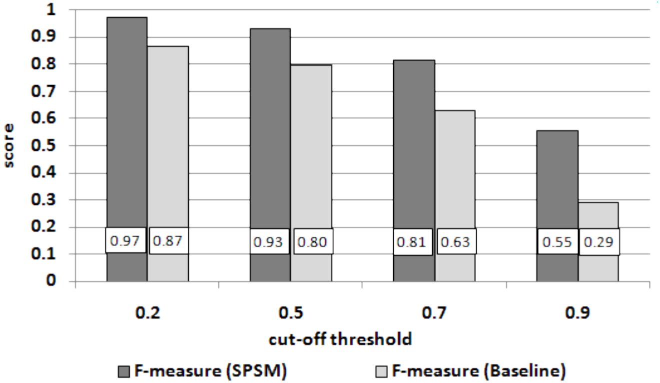


Figure: F-measure: Semantic alteration

Evaluation results

Alteration operations

- Robustness of the SPSM algorithm over significant ranges of parameters' variation (different alteration operations, alteration operations' probabilities, and cut-off threshold) was good and SPSM maintained a relatively high (over 60%) F-measure

SPSM vs. Baseline

- F-measure comparison
- Equivalent for syntactic alteration
- $>$ 20% for meaning alteration
- $\Rightarrow$ SPSM matcher: best of both worlds

Classification experiment

Evaluation setup: dataset

- Selected set (50) of GIS Web service operations from the previous dataset
  - Manual classification of the initial set of operations (WSDL files)
  - Deletion of some general (valid for all the groups) operations
  - Refinement of the classification by logically regrouping some operations

Evaluation methodology: example

Methodology

- $R = \{(Op_i, Op_j) \in OP^2 | TreeSim(Op_i, Op_j) \geq cutoffThresh\}$
- $C = \{(Op_i, Op_j) \in OP^2 | (Op_i, Op_j) \in RefAlign\}$
- $TP = \{(Op_i, Op_j) | (Op_i, Op_j) \in R \wedge (Op_i, Op_j) \in C\}$
- $FP = \{(Op_i, Op_j) | (Op_i, Op_j) \in R \wedge (Op_i, Op_j) \notin C\}$
- $FN = \{(Op_i, Op_j) | (Op_i, Op_j) \in C \wedge (Op_i, Op_j) \notin R\}$

In our example

- $cutoffThresh = 0.5$
- $|C| = |TP| \cup |FN| = 10$
- $|R| = |TP| \cup |FP| = 12$
- $|TP| = 8$
- $|FN| = 2$
- $|FP| = 4$
- $Recall = |TP|/|C| = 0.8$
- $Precision = |TP|/|R| = 0.67$
- $F - measure = 0.73$

Table: Manual classific.

	Op_1	Op_2	Op_3	Op_4
Op_1	1	1	1	0
Op_2	1	1	1	0
Op_3	1	1	1	0
Op_4	0	0	0	1

Table: SPSM classification

	Op_1	Op_2	Op_3	Op_4
Op_1	1	0.76	0.22	0.52
Op_2	0.76	1	0.57	0.54
Op_3	0.22	0.57	1	0.12
Op_4	0.52	0.54	0.12	1

Evaluation results

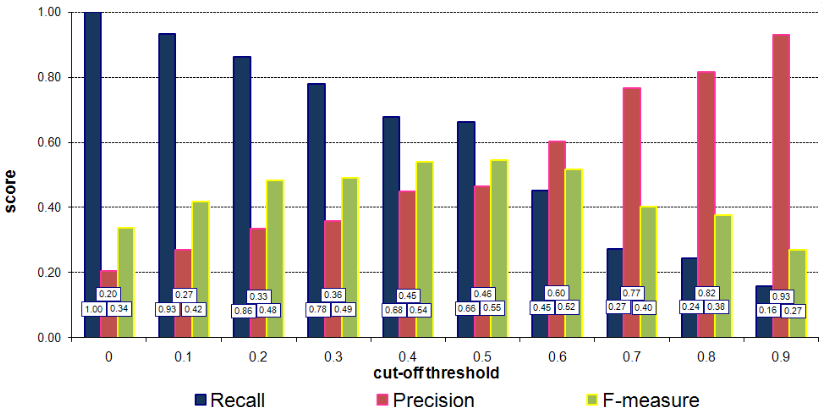


Figure: Classification results: best F-measure: 52%

Performance evaluation

More than 700.000 matching tasks

- Setup: standard laptop Intel Centrino Core Duo CPU-2Ghz, 2GB RAM, Windows Vista O.S., no applications running but a single matching system.
- Average numbers of the parameters of the WSDL operations: 4
- Efficiency: execution time per matching task: 43 ms
- Quantity of main memory during matching tasks: less than 2.3Mb (than the standby level)
- SPSM could be employed to find and integrate similar web service implementations at runtime

Conclusions

Summary

- State of the art of interoperability among distributed and heterogeneous SDIs
- OK system application to a distributed SDI scenario
- SPSM approach evaluation with important results:
  - Evolution experiment: $>$ 20% in comparison to the baseline
  - Classification experiment: best F-measure around 52%
  - Performance: SPSM is robust and can be used at run-time

Application scenarios: ontologies evolve !

- Geo Web service discovery
- Geo Web service composition
- Geo-sensor networks

Future work

Application and evaluation

- Geo-catalog of the Autonomous Province of Trento
- Geo-sensor networks in a real world emergency scenario
- Extensive evaluation on different kinds of geo-services (e.g., GRASS package)
- Geo-data similarity evaluation (e.g. INSPIRE themes)

Extending the SPSM solution

- Incorporating domain specific preferences
- Use domain specific (GIS) and/or multilingual thesauri, e.g. Gemet, Agrovoc and Eurovoc for semantic matching
- Extension of SPSM to perform spatial matching

Thank you for your attention !

QUESTIONS ?

This work has been supported by:

- The University of Trento (http://www.unitn.it)
- The EU project OpenKnowledge (http://www.openk.org)
- The Autonomous Province of Trento (http://www.provincia.tn.it)

Evaluation measures

Definitions

- TP: True positives
- FP: False positives
- FN: False negatives
- Relevant: $C = TP \cup FN$
- Retrieved: $R = TP \cup FP$

Quality measures

- $Precision = \frac{|TP|}{|R|}$
- $Recall = \frac{|TP|}{|C|}$
- $F - measure = \frac{2*Recall*Precision}{Recall+Precision}$