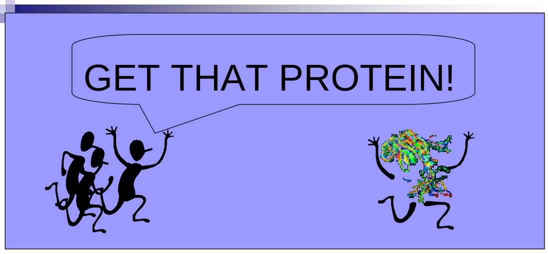
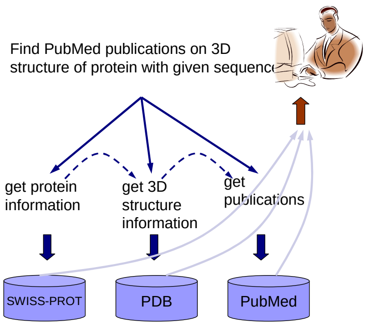
Towards a Semantic Web for Biomedical Informatics

Patrick Lambrix
Linköpings universitet



"Get that protein!"



- Find
- Divide & order
- Execute
- Combine

Two issues

- Creation of data and knowledge sources
- Searching for information in the data and knowledge sources

Data and knowledge sources

Types of sources

- Based on content
  - genes, proteins, interactions, …
  - different organisms
  - different properties
    (e.g. sequence, 3d-structure, …)
  - …

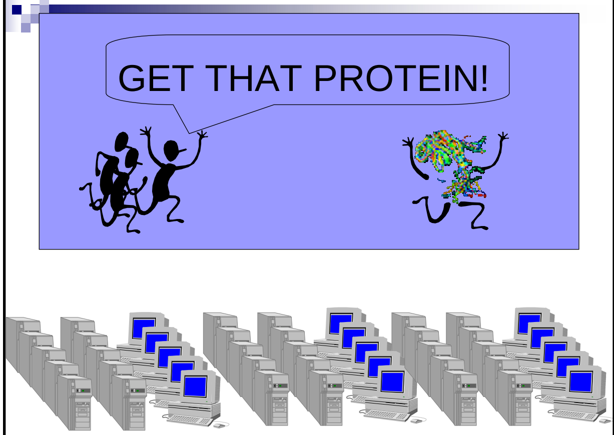
Types of sources

- Based on data quality
  - research results validated in lab
  - results manually extracted from research articles by curators
  - results automatically extracted from research articles by text mining systems
  - information extracted from research results and other sources by data mining systems

Types of sources

- Based on types of data
  - textual data
  - multimedia data (pictures, sound, video)
- Based on data models
  - text (document bases)
  - semi-structured data
  - structured data (databases)
  - logic-based formalisms (knowledge bases)

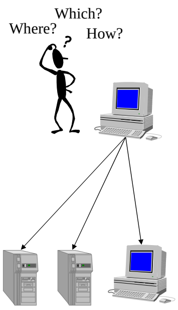
Searching for information in the data and knowledge sources



Locating relevant information

Vision: Web services

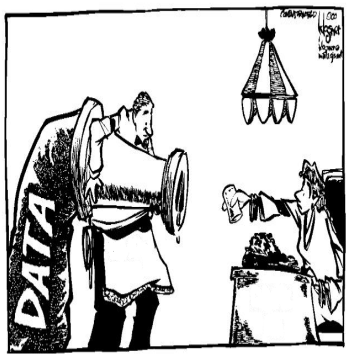
- Databases and tools (service providers) announce their service capabilities
- Users request services which may be based on task descriptions
- Service matchers find relevant services (composition) based on user needs and user preferences, negotiate service delivery, and deliver results to user



Retrieving relevant information

Vision:

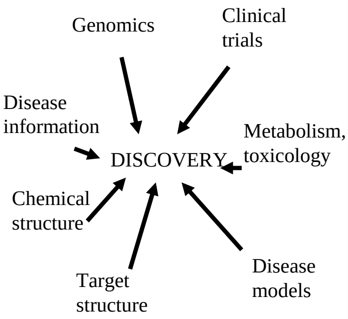
Based on the meaning of the query:
- only relevant information is retrieved
- all relevant information is retrieved



Integrating information

Vision:

Integrate data sources that are heterogeneous in content, data quality, data models, access methods, terminology



Today: syntactic Web

- A library of documents (web pages) interconnected by links
- A common portal to applications accessible through web pages, and presenting their results as web pages

A place where computers do the presentation (easy) and people do the linking and interpreting (hard).

Semantic Web

W3C: Facilities to put machine-understandable data on the Web are becoming a high priority for many communities. The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently. The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications.

What is the problem?

Example based on example on slides by P. Patel-Schneider

What information can we see…

Date: 13-15 June, 2005
Location: Linköping
Sponsors: IEEE, CERC, LiU
14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborating Enterprises (WETICE-2005)
Welcome to WETICE-2005

…

What information can a machine see…

7 - 1 1 m a y 2 0 0 2
S h e r a t o n w a i k i k i h o t e l
H o n o l u l u , h a w a i i , USA
WWW2 0 0 2
T h e e l e v e n t h i n t e r n a t i o n a l w o r l d w i d e w e b c o n f e r e n c e
R e g i s t e r e d p a r t i c i p a n t s c o m i n g f r o m
a u s t r a l i a , c a n a d a , c h i l e d e n m a r k , f r a n c e , g e r m a n y , g h a n a , h o n g k o n g , i n d i a , i r e l a n d , i t a l y , j a p a n , m a l t a , n e w z e a l a n d , t h e n e t h e r l a n d s , n o r w a y , s i n g a p o r e , s w i t z e r l a n d , t h e u n i t e d k i n g d o m , t h e u n i t e d s t a t e s , v i e t n a m , z a i r e
R e g i s t e r n o w
1 l o c a t i o n 5 d a y s l e a r n i n t e r a c t

Use XML markup with “meaningful” tags

<date> 13-15 June 2005 </date>

<location> Linköping </location>

<sponsors>IEEE, CERC, LiU </sponsors>

<name> 14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborating Enterprises (WETICE-2005) </name>

<welcome> Welcome to WETICE-2005 </welcome>

Machine sees …

<date>7 - 1 1 m a y 2 0 0 2</date>

<location>S h e r a t o n w a i k i k i h o t e l
H o n o l u l u , h a w a i i , USA</location>

<sponsors>R e g i s t e r e d p a r t i c i p a n t s c o m i n g
f r o m
a u s t r a l i a , c a n a d a , c h i l e d e n m a r k ,
f r a n c e , g e r m a n y , g h a n a , h o n g
k o n g , i n d i a , i r e l a n d , i t a l y ,
j a p a n , m a l t a , n e w z e a l a n d , t h e
n e t h e r l a n d s , n o r w a y , s i n g a p o r e ,
s w i t z e r l a n d , t h e u n i t e d k i n g d o m ,
t h e u n i t e d s t a t e s , v i e t n a m ,
z a i r e</sponsors>

<name>WWW2 0 0 2 T h e e l e v e n t h i n t e r n a t i o n a l
w o r l d w i d e w e b c o n</name>

<welcome>1 l o c a t i o n 5 d a y s l e a r n
i n t e r a c t</welcome>

But what about …

<date> 13-15 June 2005 </date>

<place> Linköping </place>

<sponsors>IEEE, CERC, LiU </sponsors>

<conf> 14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborating Enterprises (WETICE-2005) </conf>

<introduction> Welcome to WETICE-2005 </introduction>

Machine sees …

<date>7 - 1 1 m a y 2 0 0 2</date>

<n o l u>S h e r a t o n w a i k i k i h o t e l
H o n o l u l u , h a w a i i , USA</n o l u>

<sponsors>R e g i s t e r e d p a r t i c i p a n t s c o m i n g
f r o m
a u s t r a l i a , c a n a d a , c h i l e d e n m a r k ,
f r a n c e , g e r m a n y , g h a n a , h o n g
k o n g , i n d i a , i r e l a n d , i t a l y ,
j a p a n , m a l t a , n e w z e a l a n d , t h e
n e t h e r l a n d s , n o r w a y , s i n g a p o r e ,
s w i t z e r l a n d , t h e u n i t e d k i n g d o m ,
t h e u n i t e d s t a t e s , v i e t n a m ,
z a i r e</sponsors>

<o n a>WWW2 0 0 2
T h e e l e v e n t h i n t e r n a t i o n a l w o r l d w i d e
w e b c o n</o n a>

<l o c a t i o>1 l o c a t i o n 5 d a y s l e a r n
i n t e r a c t </l o c a t i o>

Adding “Semantics” – first approach

External agreement on meaning of annotations

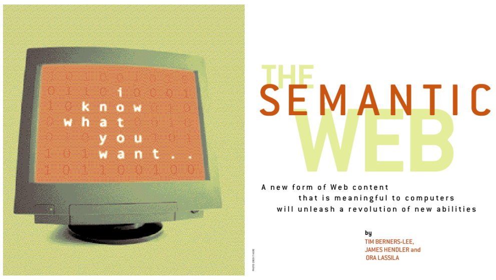
- Agree on the meaning of a set of annotation tags
- Problems with this approach:
  - Inflexible
  - Limited number of things can be expressed

Adding “Semantics” – second approach

Use on-line ontologies to specify meaning of annotations

- Ontologies provide a vocabulary of terms
- New terms can be formed by combining existing ones
- Meaning (semantics) of such terms is formally specified

Scientific American, May 2001:



- First step towards the vision:
  adding semantic annotation to web resources

Semantic annotations based on ontologies

- Locating information
  - Web service descriptions use ontologies
  - Users use ontologies when formulating requests
  - Service matchers find services based on meaning
- Retrieving relevant information
  - Reduce non-relevant information (precision)
  - Find more relevant information (recall)
- Integrating information
  - Relating similar entities in different databases

KETM for biomedical informatics

Schedule

- Monday
  - Introduction
  - Knowledge representation and search
    - Biomedical ontologies and alignment of biomedical ontologies
    - Standards for molecular interaction data
  - Creation of data and knowledge sources
    - Probabilistic Graph Models for Gene Regulatory Networks
- Tuesday
  - Creation of data and knowledge sources
    - Text mining for Biomedicine
    - Applied Biomedical Textmining at AstraZeneca
  - Demo/exercise session

Biomedical Ontologies

Biomedical ontologies

- Definition
- Use
- Components
- Knowledge representation

Ontologies

"Ontologies define the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary."

(Neches, Fikes, Finin, Gruber, Senator, Swartout, 1991)

Definitions

- Ontology as specification of a conceptualization
- Ontology as philosophical discipline
- Ontology as informal conceptual system
- Ontology as formal semantic account
- Ontology as representation of conceptual system via a logical theory
- Ontology as the vocabulary used by a logical theory
- Ontology as a meta-level specification of a logical theory

(Guarino, Giaretta)

Definitions

- An ontology is an explicit specification of a conceptualization (Gruber)
- An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. (Swartout, Patil, Knight, Russ)
- An ontology provides the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base. (Bernaras, Lasergoiti, Correra)
- An ontology is a formal, explicit specification of a shared conceptualization (Studer, Benjamins, Fensel)

Example

GENE ONTOLOGY (GO)

```
immune response
  i- acute-phase response
  i- anaphylaxis
  i- antigen presentation
  i- antigen processing
  i- cellular defense response
  i- cytokine metabolism
    i- cytokine biosynthesis synonym cytokine production
      …
    p- regulation of cytokine biosynthesis
     …
  …
  i- B-cell activation
    i- B-cell differentiation
    i- B-cell proliferation
  i- cellular defense response
  …
  i-  T-cell activation
    i- activation of natural killer cell activity
     …
```

Example Ontologies

- Knowledge representation ontology: frame ontology
- Top level ontologies: TLO, Cyc
- Linguistic ontologies: GUM, WordNet
- Engineering ontologies: EngMath, PhysSys
- Domain ontologies: CHEMICALS, Gene Ontology, Open Biomedical Ontologies

Ontologies used …

- for communication between people and organizations
- for enabling knowledge reuse and sharing
- as basis for interoperability between systems
- as repository of information
- as query model for information sources

Key technology for the Semantic Web

Biomedical Ontologies - efforts

OBO – Open Biomedical Ontologies
http://www.obofoundry.org/
(over 50 ontologies)

" The mission of OBO is to support community members who are developing and publishing ontologies in the biomedical domain. It is our vision that a core of these ontologies will be fully interoperable, by virtue of a common design philosophy and implementation, thereby enabling scientists and their instruments to communicate with minimum ambiguity. In this way the data generated in the course of biomedical research will form a single, consistent, cumulatively expanding, and algorithmically tractable whole. This core will be known as the "OBO Foundry". ."

OBO Foundry

1. open and available
2. common shared syntax
3. unique identifier space
4. procedures for identifying distinct successive versions
5. clearly specified and clearly delineated content
6. textual definitions for all terms
7. use relations from OBO Relation Ontology
8. well documented
9. plurality of independent users
10. developed collaboratively with other OBO Foundry members

Biomedical Ontologies - efforts

National Center for Biomedical Ontology
http://bioontology.org/index.html
Funded by National Institutes of Health

”The goal of the Center is to support biomedical researchers in their knowledge-intensive work, by providing online tools and a Web portal enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. A major focus of our work involves the use of biomedical ontologies to aid in the management and analysis of data derived from complex experiments.”

Biomedical Ontologies - efforts

- Gene Ontology Consortium (GO): molecular function, biological process, cellular component
- Standards and Ontologies for Functional Genomics (SOFG): meeting and website
- Proteomics Standards Initiative
- Plant Ontology consortium

Biomedical Ontologies - efforts

- International Health Terminology Standards Development Organisation

  http://www.ihtsdo.org

SNOMED CT (**S**ystematized **No**menclature of **Med**icine-**C**linical **T**erms)

Ontologies in biomedical research

- many biomedical ontologies

  e.g. GO, OBO, SNOMED-CT

- practical use of biomedical ontologies

  e.g. databases annotated with GO

GENE ONTOLOGY (GO)

```
immune response
  i- acute-phase response
  i- anaphylaxis
  i- antigen presentation
  i- antigen processing
  i- cellular defense response
  i- cytokine metabolism
     i- cytokine biosynthesis
        synonym cytokine production
        …
     p- regulation of cytokine
        biosynthesis
     …
  …
  i- B-cell activation
    i- B-cell differentiation
    i- B-cell proliferation
  i- cellular defense response
  …
  i- T-cell activation
     i- activation of natural killer
        cell activity
     …
```

Components

- concepts
  - represent a set or class of entities in a domain

    immune response
  - organized in taxonomies
    (hierarchies based on e.g. *is-a* or *is-part-of*)

    immune response is-a *defense response*

- instances
  - often not represented in an ontology
  (instantiated ontology)

Components

- relations

  R: C1 x C2 x … x Cn

  Protein hasName ProteinName

  Chromosone hasSubcellularLocation Nucleus

Components

- axioms

  'facts that are always true'

  The origin of a protein is always of the type 'gene coding origin type'

  Each protein has at least one source.

  A helix can never be a sheet and vice versa.

Different kinds of ontologies

- Controlled vocabularies
  - Concepts
- Taxonomies
  - Concepts, is-a
- Thesauri
  - Concepts, predefined relations
- Data models (e.g. EER, UML)
  - Concepts, relations, axioms
- Logics
  - Concepts, relations, axioms

Taxonomy - GeneOntology

id: GO:0003674 name: molecular_function
def: "Elemental activities, such as catalysis or binding, describing the actions of a gene product at the molecular level. A given gene product may exhibit one or more molecular functions."

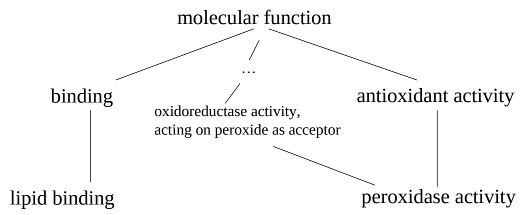
id: GO:0015643 name: binding
def: "The selective, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule."
is-a: GO:0003674 ! molecular_function

id: GO:0008289 name: lipid binding
is_a: GO:0015643 ! binding

id: GO:0016209 name: antioxidant activity
def: "Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can itself be more easily oxidized than the substance protected."
is_a: GO:0003674 ! molecular_function

id: GO:0004601 name: peroxidase activity
def: "Catalysis of the reaction: donor + H2O2 = oxidized donor + 2 H2O."
is_a: GO:0016209 ! antioxidant activity
is_a: GO:0016684 ! oxidoreductase activity, acting on peroxide as acceptor

Taxonomy - GeneOntology



Thesaurus

- graph
- fixed set of relations

  (synonym, narrower term, broader term, similar)

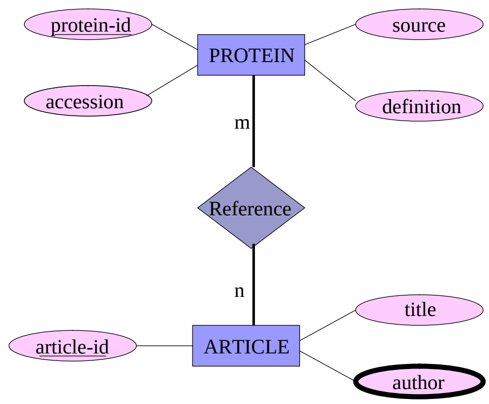
Thesaurus - WordNet

thesaurus, synonym finder
=> wordbook
=> reference book, reference, reference work, book of facts
=> book
=> publication
=> print media
=> medium
=> means
=> instrumentality, instrumentation
=> artifact, artefact
=> object, inanimate object, physical object
=> entity
=> work, piece of work
=> product, production
=> creation
=> artifact, artefact
=> object, inanimate object, physical object
=> entity

OO Data models

- EER
  entity types, attributes, relationships,
  cardinality constraints, taxonomy
- UML
  classes, attributes, associations,
  cardinality constraints, taxonomy, operations
- Taxonomy/inheritance – semantics?
- Intuitive, lots of tools, widely used.

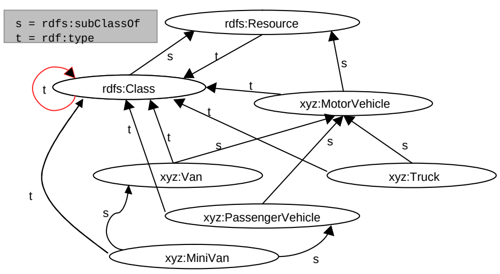
Entity-relationship



RDF + RDF Schema

- Basic construct: sentence: *Subject Predicate Object*
  - Encoded in XML
  - Can be seen as ground atomic formula
  - Represented as graph
- RDF Schema
- Editors, query tools exist

RDF Schema - example



Logics

- Formal languages
- Syntax, semantics, inference mechanisms

Logics

Reasoning services used in

- Ontology design
  - Check concept satisfiability, ontology satisfiability and (unexpected) implied relationships
- Ontology aligning and merging
  - Assert inter-ontology relationships.
  - Reasoner computes integrated concept hierarchy/consistency.
- Ontology deployment
  - Determine if a set of facts are consistent w. r. t. ontology.
  - Determine if individuals are instances of ontology concepts.
  - Query inclusion.
  - Classification-based querying.

Description Logics

- A family of KR formalisms tailored for expressing knowledge about concepts and concept hierarchies
- Based on FOPL, supported by automatic reasoning systems
- Basic building blocks: concepts (concepts), roles (binary relations), individuals (instances)
- Language constructs can be used to define new concepts and roles (axioms).
  - Intersection, union, negation, quantification, …
- Knowledge base is Tbox + Abox
  - Tbox: concept level - axioms: equality and subsumption (is-a)
  - Abox: instance level - axioms: membership, relations
- Reasoning services
  - Satisfiability of concept, Subsumption/Equivalence/Disjointness between concepts, Classification, Instantiation, Retrieval

Description Logics

Intersection

Signal-transducer-activity ∩ binding

Negation

¬ Helix

Quantifiers

∃ hasOrigin.Mitochondrion

∀ hasOrigin.Gene-coding-origin-type

DAML+OIL / OWL

- DAML+ OIL almost equivalent to SHIQ
- DAML+ OIL supports the full range of XML Schema data types
- OWL updated DAML+OIL

DAML+OIL Class Constructors

Constructor	DL Syntax	Example
intersectionOf	$C_1 \sqcap \ldots \sqcap C_n$	Human ⊓ Male
unionOf	$C_1 \sqcup \ldots \sqcup C_n$	Doctor ⊔ Lawyer
complementOf	$\neg C$	¬Male
oneOf	$\{x_1 \ldots x_n\}$	{john, mary}
toClass	$\forall P.C$	∀hasChild.Doctor
hasClass	$\exists P.C$	∃hasChild.Lawyer
hasValue	$\exists P.\{x\}$	∃citizenOf.{USA}
minCardinalityQ	$\geqslant n P.C$	⩾2hasChild.Lawyer
maxCardinalityQ	$\leqslant n P.C$	⩽1hasChild.Male
cardinalityQ	$= n\ P.C$	=1 hasParent.Female

- XMLS **datatypes** as well as classes
- Arbitrarily complex **nesting** of constructors
  - E.g., Person ⊓ ∀hasChild.(Doctor ⊔ ∃hasChild.Doctor)

EDBT 2002: DAML+OIL – p.13/32

DAML+OIL Axioms

Axiom	DL Syntax	Example
subClassOf	$C_1 \sqsubseteq C_2$	Human ⊑ Animal ⊓ Biped
sameClassAs	$C_1 \equiv C_2$	Man ≡ Human ⊓ Male
subPropertyOf	$P_1 \sqsubseteq P_2$	hasDaughter ⊑ hasChild
samePropertyAs	$P_1 \equiv P_2$	cost ≡ price
sameIndividualAs	$\{x_1\} \equiv \{x_2\}$	{President_Bush} ≡ {G_W_Bush}
disjointWith	$C_1 \sqsubseteq \neg C_2$	Male ⊑ ¬Female
differentIndividualFrom	$\{x_1\} \sqsubseteq \neg\{x_2\}$	{john} ⊑ ¬{peter}
inverseOf	$P_1 \equiv P_2^-$	hasChild ≡ hasParent^-
transitiveProperty	$P^+ \sqsubseteq P$	ancestor^+ ⊑ ancestor
uniqueProperty	$\top \sqsubseteq \leqslant 1 P$	⊤ ⊑ ⩽1hasMother
unambiguousProperty	$\top \sqsubseteq \leqslant 1 P^-$	⊤ ⊑ ⩽1isMotherOf^-

- Axioms (mostly) **reducible to subClass/PropertyOf**

EDBT 2002: DAML+OIL – p.14/32

OWL

- OWL-Lite, OWL-DL, OWL-Full: increasing expressivity
- A legal OWL-Lite ontology is a legal OWL-DL ontology is a legal OWL-Full ontology
- OWL-DL: expressive description logic, decidable
- XML-based
- RDF-based (OWL-Full is extension of RDF, OWL-Lite and OWL-DL are extensions of a restriction of RDF)

OWL-Lite

- **Class**, subClassOf, equivalentClass
- intersectionOf (only named classes and restrictions)
- **Property**, subPropertyOf, equivalentProperty
- domain, range (global restrictions)
- inverseOf, TransitiveProperty (*), SymmetricProperty, FunctionalProperty, InverseFunctionalProperty
- allValuesFrom, someValuesFrom (local restrictions)
- minCardinality, maxCardinality (only 0/1)
- **Individual**, sameAs, differentFrom, AllDifferent

(*) restricted

OWL-DL

- **Type separation (**class cannot also be individual or property, property cannot be also class or individual), Separation between DatatypeProperties and ObjectProperties
- **Class –complex classes**, subClassOf, equivalentClass, *disjointWith*
- intersectionOf, *unionOf, complementOf*
- **Property**, subPropertyOf, equivalentProperty
- domain, range (global restrictions)
- inverseOf, TransitiveProperty (*), SymmetricProperty, FunctionalProperty, InverseFunctionalProperty
- allValuesFrom, someValuesFrom (local restrictions), *oneOf, hasValue*
- minCardinality, maxCardinality
- **Individual**, sameAs, differentFrom, AllDifferent

(*) restricted

Defining ontologies is not so easy ...

The Celestial Emporium of Benevolent Knowledge, Borges

"On those remote pages it is written that animals are divided into:
a. those that belong to the Emperor
b. embalmed ones
c. those that are trained
d. suckling pigs
e. mermaids
f. fabulous ones
g. stray dogs
h. those that are included in this classification
i. those that tremble as if they were mad
j. innumerable ones
k. those drawn with a very fine camel's hair brush
l. others
m. those that have just broken a flower vase
n. those that resemble flies from a distance"

Slide from talk by C. Goble

Defining ontologies is not so easy ...

Dyirbal classification of objects in the universe

- Bayi: men, kangaroos, possums, bats, most snakes, most fishes, some birds, most insects, the moon, storms, rainbows, boomerangs, some spears, etc.
- Balan: women, anything connected with water or fire, bandicoots, dogs, platypus, echidna, some snakes, some fishes, most birds, fireflies, scorpions, crickets, the stars, shields, some spears, some trees, etc.
- Balam: all edible fruit and the plants that bear them, tubers, ferns, honey, cigarettes, wine, cake.
- Bala: parts of the body, meat, bees, wind, yamsticks, some spears, most trees, grass, mud, stones, noises, language, etc.

Slide from talk by C. Goble

Ontology tools

- Ontology development tools
- Ontology merge and alignment tools
- Ontology evaluation tools
- Ontology-based annotation tools
- Ontology storage and querying tools
- Ontology learning tools

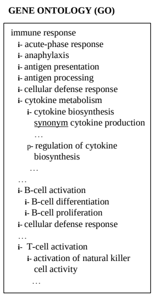
Ontology Alignment

Ontology Alignment

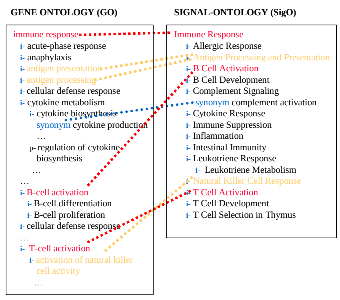
- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies
- Current issues

Ontologies in biomedical research

- many biomedical ontologies
  e.g. GO, OBO, SNOMED-CT
- practical use of biomedical ontologies
  e.g. databases annotated with GO



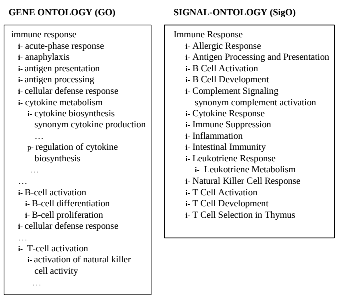
Ontologies with overlapping information



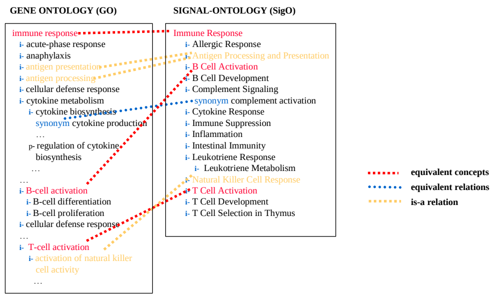
Ontologies with overlapping information

- Use of multiple ontologies
  e.g. custom-specific ontology + standard ontology
- Bottom-up creation of ontologies
  experts can focus on their domain of expertise

à important to know the inter-ontology relationships



Ontology Alignment

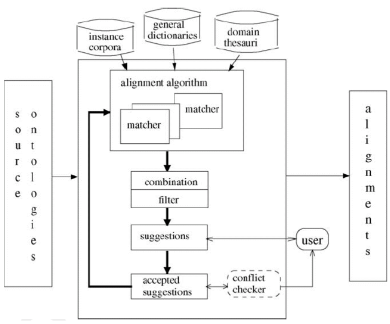


Defining the relations between the terms in different ontologies

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies
- Current issues

An Alignment Framework



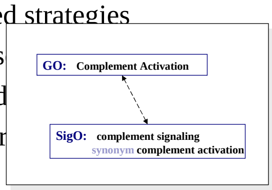
Classification

- According to input
  - KR: OWL, UML, EER, XML, RDF, …
  - components: concepts, relations, instance, axioms
- According to process
  - What information is used and how?
- According to output
  - 1-1, m-n
  - Similarity vs explicit relations (equivalence, is-a)
  - confidence

Matchers

Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-bas
- Instance-based
- Use of auxiliar

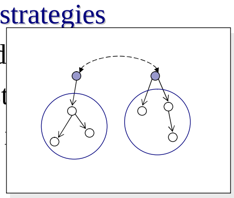


Example matchers

- Edit distance
  - Number of deletions, insertions, substitutions required to transform one string into another
  - aaaa → baab: edit distance 2
- N-gram
  - N-gram : N consecutive characters in a string
  - Similarity based on set comparison of n-grams
  - aaaa : {aa, aa, aa}; baab : {ba, aa, ab}

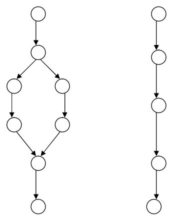
Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based
- Instance-based st
- Use of auxiliary



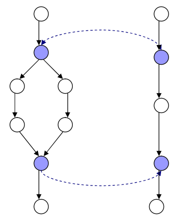
Example matchers

- Propagation of similarity values
- Anchored matching



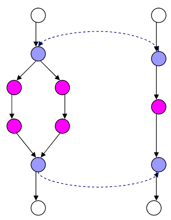
Example matchers

- Propagation of similarity values
- Anchored matching



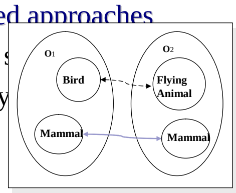
Example matchers

- Propagation of similarity values
- Anchored matching



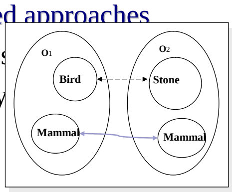
Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary



Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based s
- Use of auxiliary

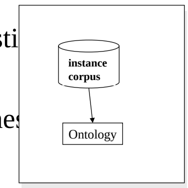


Example matchers

- Similarities between data types
- Similarities based on cardinalities

Matcher Strategies

- Strategies based on linguisti
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information



Example matchers

- Instance-based
- Use life science literature as instances

- Structure-based extensions

Learning matchers – instance-based strategies

- Basic intuition

  A similarity measure between concepts can be computed based on the probability that documents about one concept are also about the other concept and vice versa.

- Intuition for structure-based extensions

  Documents about a concept are also about their super-concepts.

  (No requirement for previous alignment results.)

Learning matchers - steps

- Generate corpora
  - Use concept as query term in PubMed
  - Retrieve most recent PubMed abstracts
- Generate text classifiers
  - One classifier per ontology / One classifier per concept
- Classification
  - Abstracts related to one ontology are classified by the other ontology's classifier(s) and vice versa
- Calculate similarities

Basic Naïve Bayes matcher

- Generate corpora
- Generate classifiers
  - Naive Bayes classifiers, one per ontology
- Classification
  - Abstracts related to one ontology are classified to the concept in the other ontology with highest posterior probability P(C|d)
- Calculate similarities

$$sim(C_1, C_2) = \frac{n_{NBC2}(C_1, C_2) + n_{NBC1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

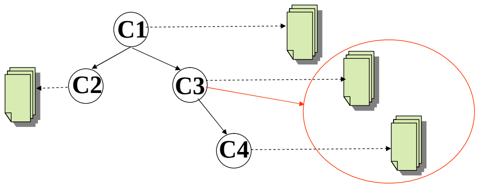
Basic Support Vector Machines matcher

- Generate corpora
- Generate classifiers
  - SVM-based classifiers, one per concept
- Classification
  - Single classification variant: Abstracts related to concepts in one ontology are classified to the concept in the other ontology for which its classifier gives the abstract the highest positive value.
  - Multiple classification variant: Abstracts related to concepts in one ontology are classified all the concepts in the other ontology whose classifiers give the abstract a positive value.
- Calculate similarities

$$\frac{n_{SVMC-C_2}(C_1, C_2) + n_{SVMC-C_1}(C_2, C_1)}{n_D(C_1) + n_D(C_2)}$$

Structural extension 'Cl'

- Generate classifiers
  - Take (is-a) structure of the ontologies into account when building the classifiers
  - Extend the set of abstracts associated to a concept by adding the abstracts related to the sub-concepts



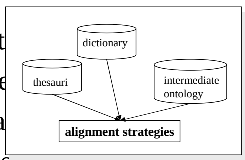
Structural extension 'Sim'

- Calculate similarities
  - Take structure of the ontologies into account when calculating similarities
  - Similarity is computed based on the classifiers applied to the concepts and their sub-concepts

$$sim_{struct}(C_1, C_2) = \frac{\sum_{C_i \subseteq C_1, C_j \subseteq C_2} n_{NBC2}(C_i, C_j) + \sum_{C_i \subseteq C_1, C_j \subseteq C_2} n_{NBC1}(C_j, C_i)}{\sum_{C_i \subseteq C_1} n_D(C_i) + \sum_{C_j \subseteq C_2} n_D(C_j)}$$

Matcher Strategies

- Strategies based linguist
- Structure-based strategie
- Constraint-based approa
- Instance-based strategies
- Use of auxiliary information



Example matchers

- Use of WordNet
  - Use WordNet to find synonyms
  - Use WordNet to find ancestors and descendants in the is-a hierarchy
- Use of Unified Medical Language System (UMLS)
  - Includes many ontologies
  - Includes many alignments (not complete)
  - Use UMLS alignments in the computation of the similarity values

Ontology Alignment and Merging Systems

	linguistic	structure	constraints	instances	auxiliary
ArtGen	name	parents, children		domain specific documents	WordNet
ASCO	name, label description	parents, children, siblings, path from root			WordNet
Chimaera	name	parents, children			
FCA-Merge	name			domain specific documents	
FOAM	name, label	parents, children	equivalence		
GLUE	name	neighborhood		instances	
HCONE	name	parents, children			WordNet
IF-Map				instances	a reference ontology
iMapper		leaf, non-leaf, children, related node	domain, range	instances	WordNet
OntoMapper		parents, children		documents	
(Anchor-) PROMPT	name	direct graphs			
SAMBO	name, synonym	is-a and part-of, descendants and ancestors		domain specific documents	WordNet, UMLS
S-Match	label	path from root	semantic relations codified in labels		WordNet

Combinations

Combination Strategies

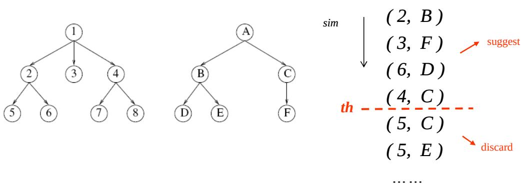
- Usually weighted sum of similarity values of different matchers
- Maximum of similarity values of different matchers

Filtering

Filtering techniques

- Threshold filtering

  Pairs of concepts with similarity higher or equal than threshold are alignment suggestions

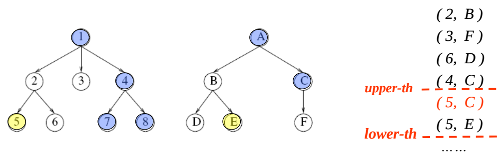


Filtering techniques

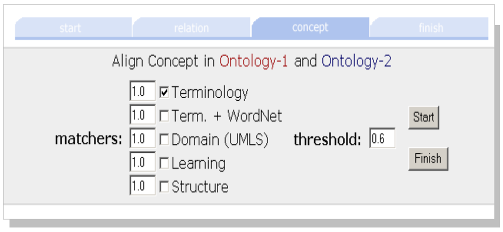
- Double threshold filtering

  (1) Pairs of concepts with similarity higher than or equal to **upper** threshold are alignment suggestions

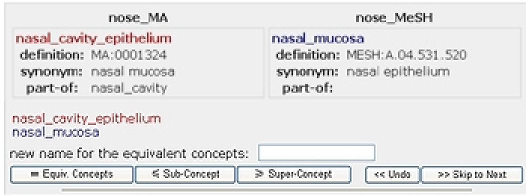
  (2) Pairs of concepts with similarity between **lower** and **upper** thresholds are alignment suggestions if they make sense with respect to the structure of the ontologies and the suggestions according to (1)



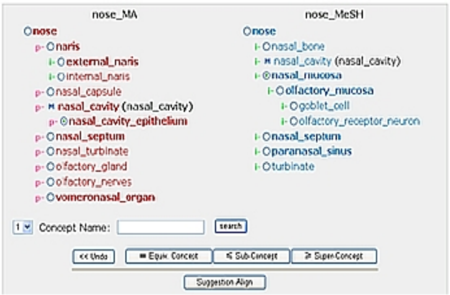
Example alignment system SAMBO – matchers, combination, filter



Example alignment system SAMBO – suggestion mode



Example alignment system SAMBO – manual mode



Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies
- Current issues

Evaluation measures

- Precision:

$$\frac{\text{\# correct suggested alignments}}{\text{\# suggested alignments}}$$

- Recall:

$$\frac{\text{\# correct suggested alignments}}{\text{\# correct alignments}}$$

- F-measure: combination of precision and recall

Ontology Alignment Evaluation Initiative

OAEI

- Since 2004
- Evaluation of systems
- Different tracks
  - comparison: benchmark (open)
  - expressive: anatomy (blind), fisheries (expert)
  - directories and thesauri: directory, library, crosslingual resources (blind)
  - consensus: conference

OAEI

- Evaluation measures
  - Precision/recall/f-measure
  - recall of non-trivial alignments
  - full / partial golden standard

OAEI 2008 – anatomy track

- Align
  - Mouse anatomy: 2744 terms
  - NCI-anatomy: 3304 terms
  - Alignments: 1544 (of which 934 'trivial')
- Tasks
  - 1. Align and optimize f
  - 2-3. Align and optimize p / r
  - 4. Align when partial reference alignment is given and optimize f

OAEI 2008 – anatomy track#1

- 9 systems participated
- SAMBO
  - p=0.869, r=0.836, r+=0.586, f=0.852
- SAMBOdtf
  - p=0.831, r=0.833, r+=0.579, f=0.832
- Use of TermWN and UMLS

OAEI 2008 – anatomy track#1

Is background knowledge (BK) needed?

Of the non-trivial alignments:

- Ca 50% found by systems using BK and systems not using BK
- Ca 13% found only by systems using BK
- Ca 13% found only by systems not using BK
- Ca 25% not found

Processing time:

hours with BK, minutes without BK

OAEI 2008 – anatomy track#4

Can we use given alignments when computing suggestions?
→ partial reference alignment given with all trivial and 50 non-trivial alignments

- SAMBO
  - p=0.636 → 0.660, r=0.626 → 0.624, f=0.631 → 0.642
- SAMBOdtf
  - p=0.563 → 0.603, r=0.622 → 0.630, f=0.591 → 0.616

(measures computed on non-given part of the reference alignment)

OAEI 2007-2008

- Systems can use only one combination of strategies per task
  - à systems use similar strategies
    - text: string matching, tf-idf
    - structure: propagation of similarity to ancestors and/or descendants
    - thesaurus (WordNet)
    - domain knowledge important for anatomy task?

Evaluation of algorithms

Cases

- GO vs. SigO

GO: 70 terms GO-immune defense | SigO: 15 terms SigO-immune defense | GO: 60 terms GO-behavior | SigO: 10 terms SigO-behavior

- MA vs. MeSH

MA: 15 terms MA-nose | MeSH: 18 terms MeSH-nose | MA: 77 terms MA-ear | MeSH: 39 terms MeSH-ear

MA: 112terms MA-eye | MeSH: 45 terms MeSH-eye

Evaluation of matchers

- Matchers

  Term, TermWN, Dom, Learn (Learn+structure), Struc

- Parameters

  Quality of suggestions: precision/recall

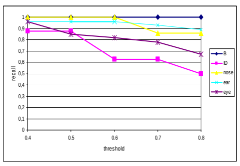
  Threshold filtering : 0.4, 0.5, 0.6, 0.7, 0.8

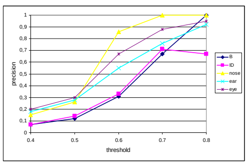
  Weights for combination: 1.0/1.2

KitAMO (http://www.ida.liu.se/labs/iislab/projects/KitAMO)

Results

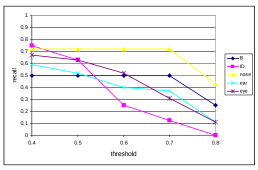
- Terminological matchers

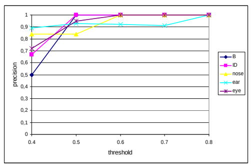




Results

- Basic learning matcher (Naïve Bayes)



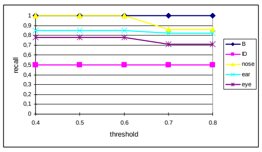


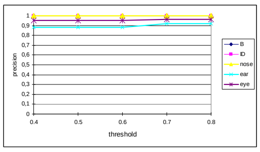
Naive Bayes slightly better recall, but slightly worse precision than SVM-single

SVM-multiple (much) better recall, but worse precision than SVM-single

Results

- Domain matcher (using UMLS)





Results

- Comparison of the matchers

  CS_TermWN ⊇ *CS_Dom* ⊇ *CS_Learn*

- Combinations of the different matchers
  - combinations give often better results
  - no significant difference on the quality of suggestions for different weight assignments in the combinations

  (but: did not check yet for large variations for the weights)

- Structural matcher did not find (many) new correct alignments

  (but: good results for systems biology schemas SBML – PSI MI)

Evaluation of filtering

- Matcher

  TermWN

- Parameters

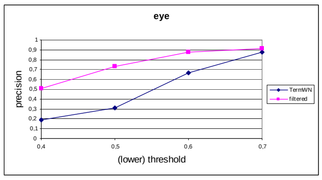
  Quality of suggestions: precision/recall

  Double threshold filtering using structure:

  Upper threshold: 0.8

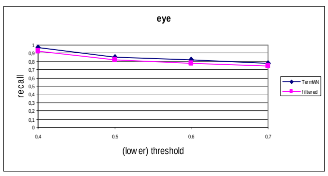
  Lower threshold: 0.4, 0.5, 0.6, 0.7, 0.8

Results



- The precision for double threshold filtering with upper threshold 0.8 and lower threshold T is higher than for threshold filtering with threshold T

Results



- The recall for double threshold filtering with upper threshold 0.8 and lower threshold T is about the same as for threshold filtering with threshold T

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies
- Current issues

Recommending strategies - 1

- Use knowledge about previous use of alignment strategies
  - gather knowledge about input, output, use, performance, cost via questionnaires
  - Not so much knowledge available
  - OAEI

(Mochol, Jentzsch, Euzenat 2006)

Recommending strategies - 2

- Optimize
  - Parameters for ontologies, similarity assessment, matchers, combinations and filters
  - Run general alignment algorithm
  - User validates the alignment result
  - Optimize parameters based on validation

(Ehrig, Staab, Sure 2005)

Recommending strategies - 2

- Tests
  - travel in russia
    QOM: r=0.618, p=0.596, f=0.607
    Decision tree 150: r=0.723, p=0.591, f=0.650
  - bibster
    QOM: r=0.279, p=0.397, f=0.328
    Decision tree 150: r=0.630, p=0.375, f=0.470

Decision trees better than Neural Nets and Support Vector Machines.

Recommending strategies - 3

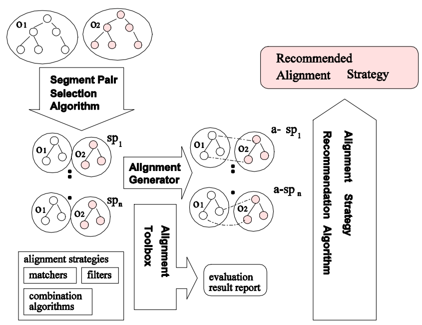
- Based on inherent knowledge
  - Use the actual ontologies to align to find good candidate alignment strategies
  - User/oracle with minimal alignment work
  - Complementary to the other approaches

(Tan, Lambrix 2007)

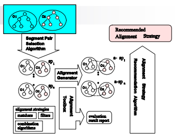
Idea

- Select small segments of the ontologies
- Generate alignments for the segments (expert/oracle)
- Use and evaluate available alignment algorithms on the segments
- Recommend alignment algorithm based on evaluation on the segments

Framework



Experiment case - Ontologies

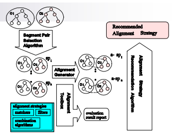


- NCI thesaurus
  - National Cancer Institute, Center for Bioinformatics
  - Anatomy: 3495 terms
- MeSH
  - National Library of Medicine
  - Anatomy: 1391 terms

Experiment case - Oracle

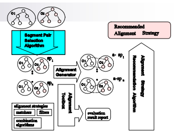
- UMLS
  - Library of Medicine
  - Metathesaurus contains > 100 vocabularies
  - NCI thesaurus and MeSH included in UMLS
  - Used as approximation for expert knowledge
  - 919 expected alignments according to UMLS

Experiment case – alignment strategies

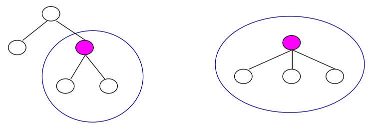


- Matchers and combinations
  - N-gram (NG)
  - Edit Distance (ED)
  - Word List + stemming (WL)
  - Word List + stemming + WordNet (WN)
  - NG+ED+WL, weights 1/3 (C1)
  - NG+ED+WN, weights 1/3 (C2)
- Threshold filter
  - thresholds 0.4, 0.5, 0.6, 0.7, 0.8

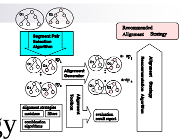
Segment pair selection algorithms



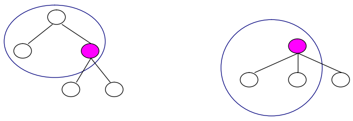
- SubG
  - Candidate segment pair = sub-graphs according to is-a/part-of with roots with same name; between 1 and 60 terms in segment
  - Segment pairs randomly chosen from candidate segment pairs such that segment pairs are disjoint



Segment pair selection algorithms



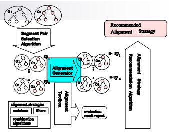
- Clust - Cluster terms in ontology
  - Candidate segment pair is pair of clusters containing terms with the same name; at least 5 terms in clusters
  - Segment pairs randomly chosen from candidate segment pairs



Segment pair selection algorithms

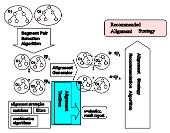
- For each trial, 3 segment pair sets with 5 segment pairs were generated
- SubG: A1, A2, A3
  - 2 to 34 terms in segment
  - level of is-a/part-of ranges from 2 to 6
  - max expected alignments in segment pair is 23
- Clust: B1, B2, B3
  - 5 to 14 terms in segment
  - level of is-a/part-of is 2 or 3
  - max expected alignments in segment pair is 4

Segment pair alignment generator



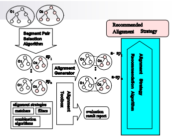
- Used UMLS as oracle

Alignment toolbox



- Used KitAMO as toolbox
- Generates reports on similarity values produced by different matchers, execution times, number of correct, wrong, redundant suggestions

Recommendation algorithm



- Recommendation scores: F, F+E, 10F+E

F: quality of the alignment suggestions

- average f-measure value for the segment pairs

E: average execution time over segment pairs, normalized with respect to number of term pairs

- Algorithm gives ranking of alignment strategies based on recommendation scores on segment pairs

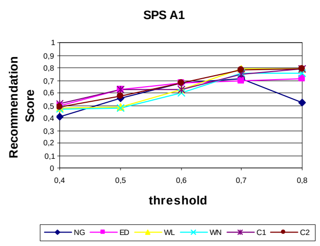
Expected recommendations for F

- Best strategies for the whole ontologies and measure F:

1. (WL,0.8)
2. (C1,0.8)
3. (C2,0.8)

Results

SubG, F, SPS A1



Results

- Top 3 strategies for SubG and measure F:

A1: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)
A2: 1. (WL,0.8) 2. (WL,0.7) 3. (WN,0.7)
A3: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)

- Best strategy always recommended first
- Top 3 strategies often recommended
- (WL,0.7) has rank 4 for whole ontologies

Results

- Top 3 strategies for Clust and measure F:

B1: 1. (C2,0.7) 2. (ED,0.6) 3. (C2,0.6)
B2: 1. (WL,0.8) (WL, 0.7) (C1,0.8) (C2,0.8)
B3: 1. (C1,0.8) (ED,0.7) 3. (C1,0.7) (C2,0.7) (WL,0.7) (WN,0.7)

- Top strategies often recommended, but not always
- (WL,0.7) (C1,0.7) (C2,0.7) ranked 4,5,6 for whole ontologies

Results

- SubG gives better results than Clust
- Results improve when number of segments is increased
- 10F+E similar results as F
- F+E
  - WordNet gives lower ranking
  - Runtime environment has influence

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Recommending ontology alignment strategies
- Current Issues

Current issues

- Systems and algorithms
  - Complex ontologies
  - Use of instance-based techniques
  - Alignment types (equivalence, is-a, …)
  - Complex alignments (1-n, m-n)
  - Connection ontology types – alignment strategies

Current issues

- Evaluations
  - Need for Golden standards
  - Systems available, but not always the alignment algorithms
  - Evaluation measures
- Recommending 'best' alignment strategies

Further reading

Starting points for further studies

Further reading ontologies

- KnowledgeWeb (http://knowledgeweb.semanticweb.org/) and its predecessor OntoWeb (http://ontoweb.aifb.uni-karlsruhe.de/)
- Lambrix, Tan, Jakoniene, Strömbäck, Biological Ontologies, chapter 4 in Baker, Cheung, (eds), *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, 85-99, Springer, 2007. ISBN: 978-0-387-48436-5.

(general about ontologies)

- Lambrix, Towards a Semantic Web for Bioinformatics using Ontology-based Annotation, *Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, 3-7, 2005. Invited talk.

(ontologies for semantic web)

- OWL, http://www.w3.org/TR/owl-features/ , http://www.w3.org/2004/OWL/

Further reading ontology alignment

- http://www.ontologymatching.org

(plenty of references to articles and systems)

- Ontology alignment evaluation initiative: http://oaei.ontologymatching.org

(home page of the initiative)

- Euzenat, Shvaiko, *Ontology Matching*, Springer, 2007.
- Lambrix, Tan, SAMBO – a system for aligning and merging biomedical ontologies, *Journal of Web Semantics*, 4(3):196-206, 2006.

(description of the SAMBO tool and overview of evaluations of different matchers)

- Lambrix, Tan, A tool for evaluating ontology alignment strategies, *Journal on Data Semantics*, VIII:182-202, 2007.

(description of the KitAMO tool for evaluating matchers)

Further reading ontology alignment

- Chen, Tan, Lambrix, Structure-based filtering for ontology alignment,*IEEE WETICE workshop on semantic technologies in collaborative applications*, 364-369, 2006.

(double threshold filtering technique)

- Tan H, Lambrix P, A method for recommending ontology alignment strategies, *International Semantic Web Conference*, 494-507, 2007.

  Ehrig M, Staab S, Sure Y, Bootstrapping ontology alignment methods with APFEL, *International Semantic Web Conference*, 186-200, 2005.

  Mochol M, Jentzsch A, Euzenat J, Applying an analytic method for matching approach selection, *International Workshop on Ontology Matching*, 2006.

(recommendation of alignment strategies)