Content-based Ontology Ranking

Mathew Jones & Harith Alani

Ontology Ranking

- Is crucial for ontology search and reuse!
  - Especially when there is a large number of them available online
- Just like most things, there are many ways to evaluate and rank ontologies
- Some suggested approaches are based on assessing:
  - Philosophical soundness (e.g. OntoClean)
  - General properties such as metadata, documentation (e.g. Ontometric)
  - User ratings
  - Authority of source
  - Popularity (e.g. Swoogle)
  - Coverage
  - Consistency
  - Accuracy
  - Fit for purpose
  - …

Ontology Ranking by Swoogle

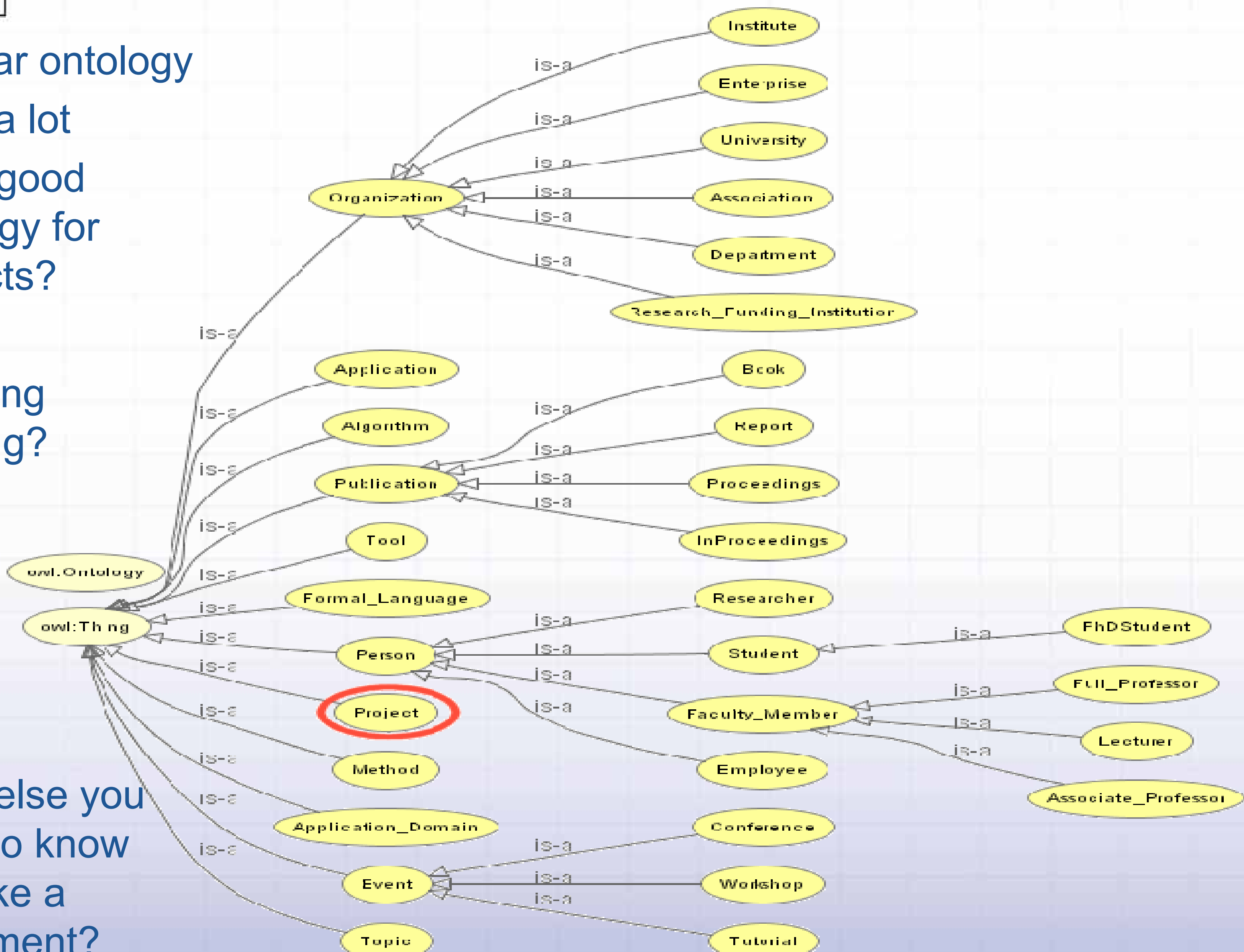
- Swoogle ranks ontologies using a variation of PageRank
  - The more links an ontology receives from other ontologies the higher its rank
- Page Rank of ontologies is sometimes insufficient
  - Many ontologies are not connected to others
  - Ontology popularity gives no guarantees on quality of specific concepts' representation
  - There is a need to extend this ranking to take into account other ontology characteristics
- Searching is based on concept names
  - Searching for Education will find ontos containing this concept

What to look for in an ontology?!

- Popular ontology
- Used a lot
- Is it a good ontology for Projects?

- Anything missing?

- What else you need to know to make a judgement?



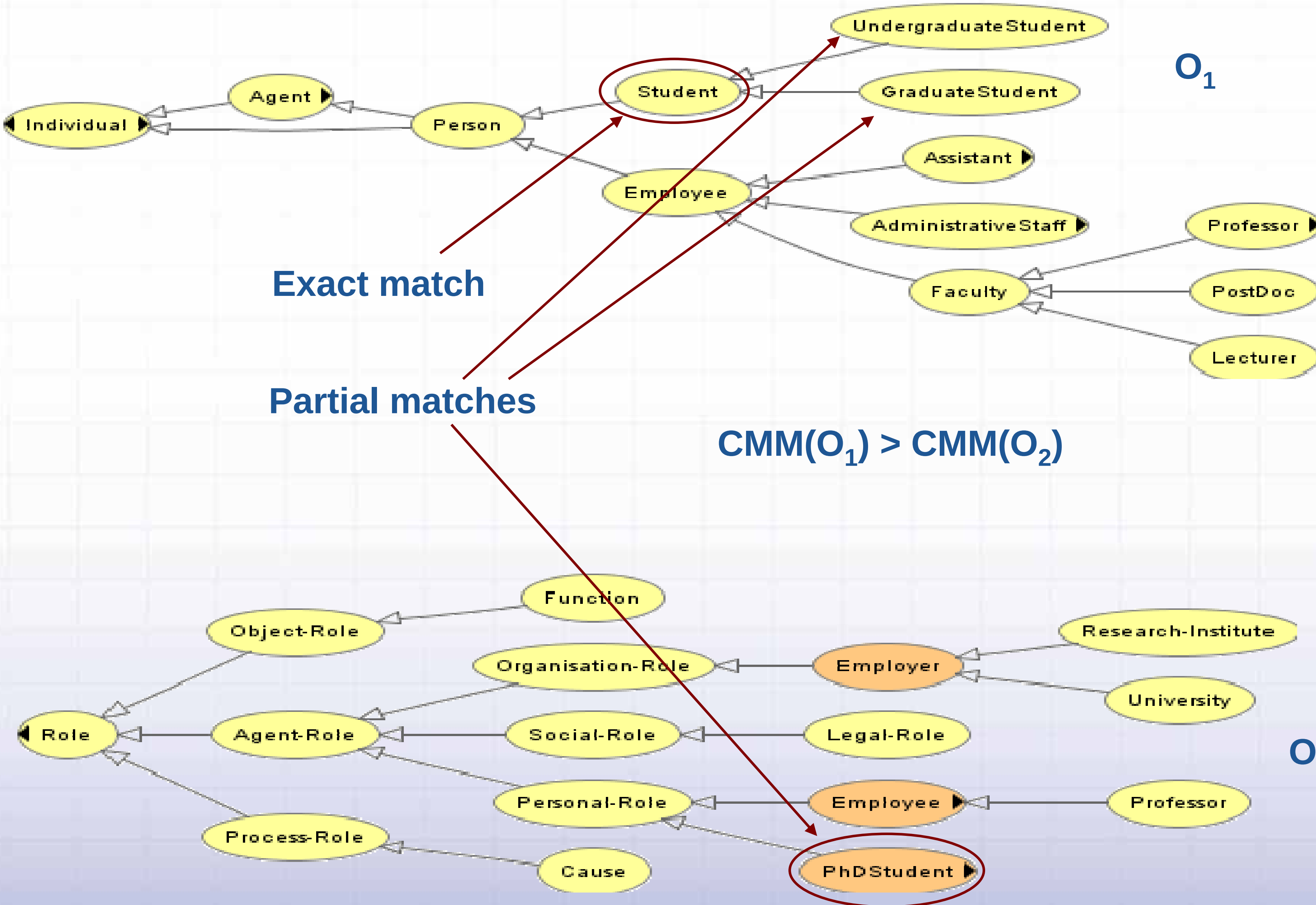
Ontology Ranking

- Our approaches:
  - Ranking based on structure analysis of concepts
    - Prototype system named AKTiveRank
    - Tries to measure how “rich” and “close” are the concepts of interest
    - Check KCap 2005 and EON 2006 for more info about AKTiveRank
  - Ranking based on content coverage
    - Measures how well the ontology terminology covers a given domain

Ranking base on Structure Analysis

- AKTiveRank: Uses as input the search terms provided by a knowledge engineer
  - Same as when searching with Swoogle
- Retrieves a list of ontology URIs from an ontology search engine
  - Not hard wired into any specific ontology search tool
- Applies a number of measures to each ontology to establish its rank with respect to specific characteristics
  - Class Match Measure
    - Evaluates the coverage of an ontology for the given search terms
  - Density Measure
    - Estimates the “semantic richness” of the concepts of interest
  - Semantic Similarity Measure
    - Measures the “closeness” of the concepts within an ontology graph
  - Betweenness Measure
    - Measures how “graphically central” the concepts are within an ontology
- Total score is calculated by aggregating all the *normalised* measure values, taking into account their *weight factors*

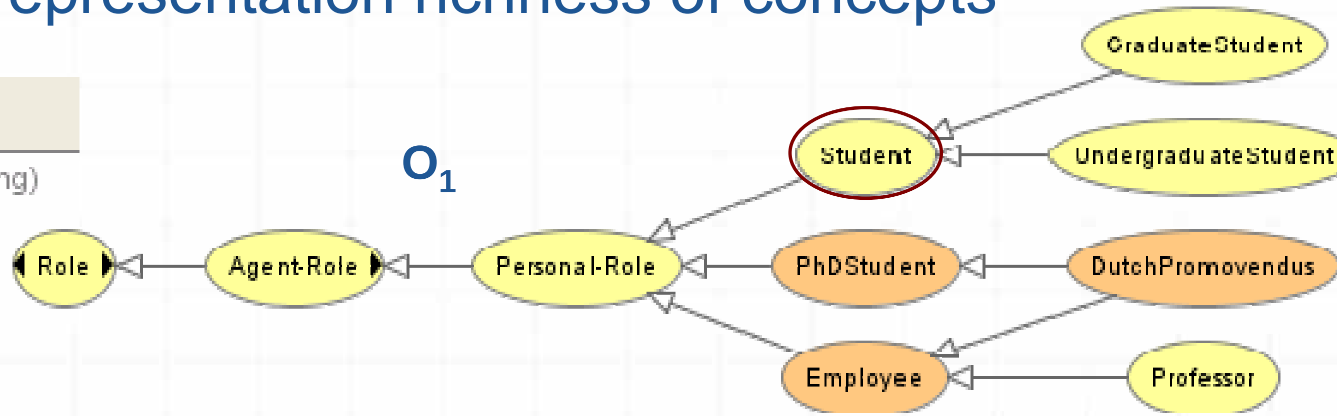
Class Match Measure (CMM)



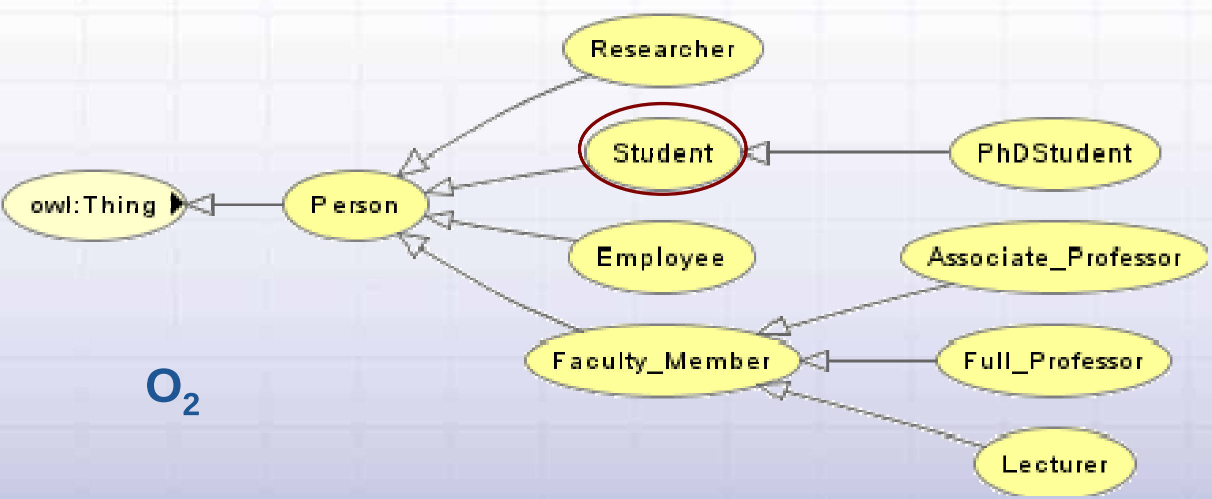
Density Measure (DEM)

- Measures the representation richness of concepts

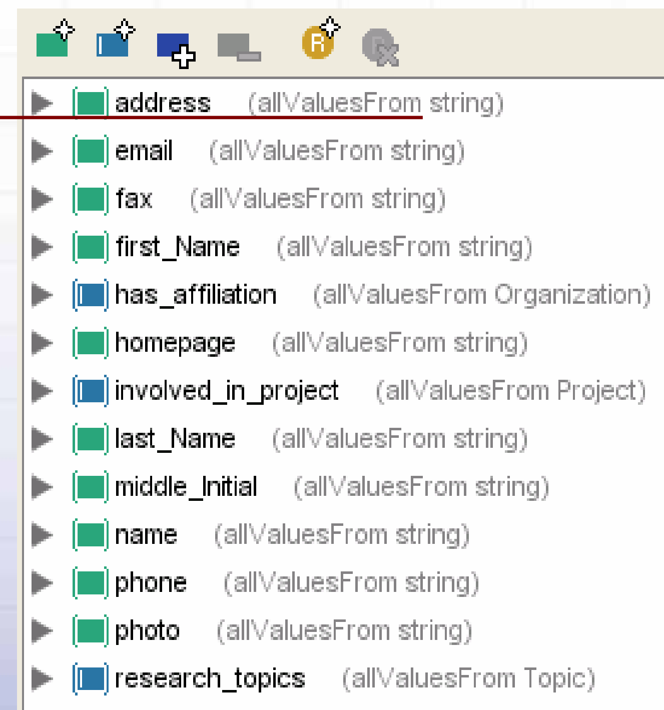
O_1



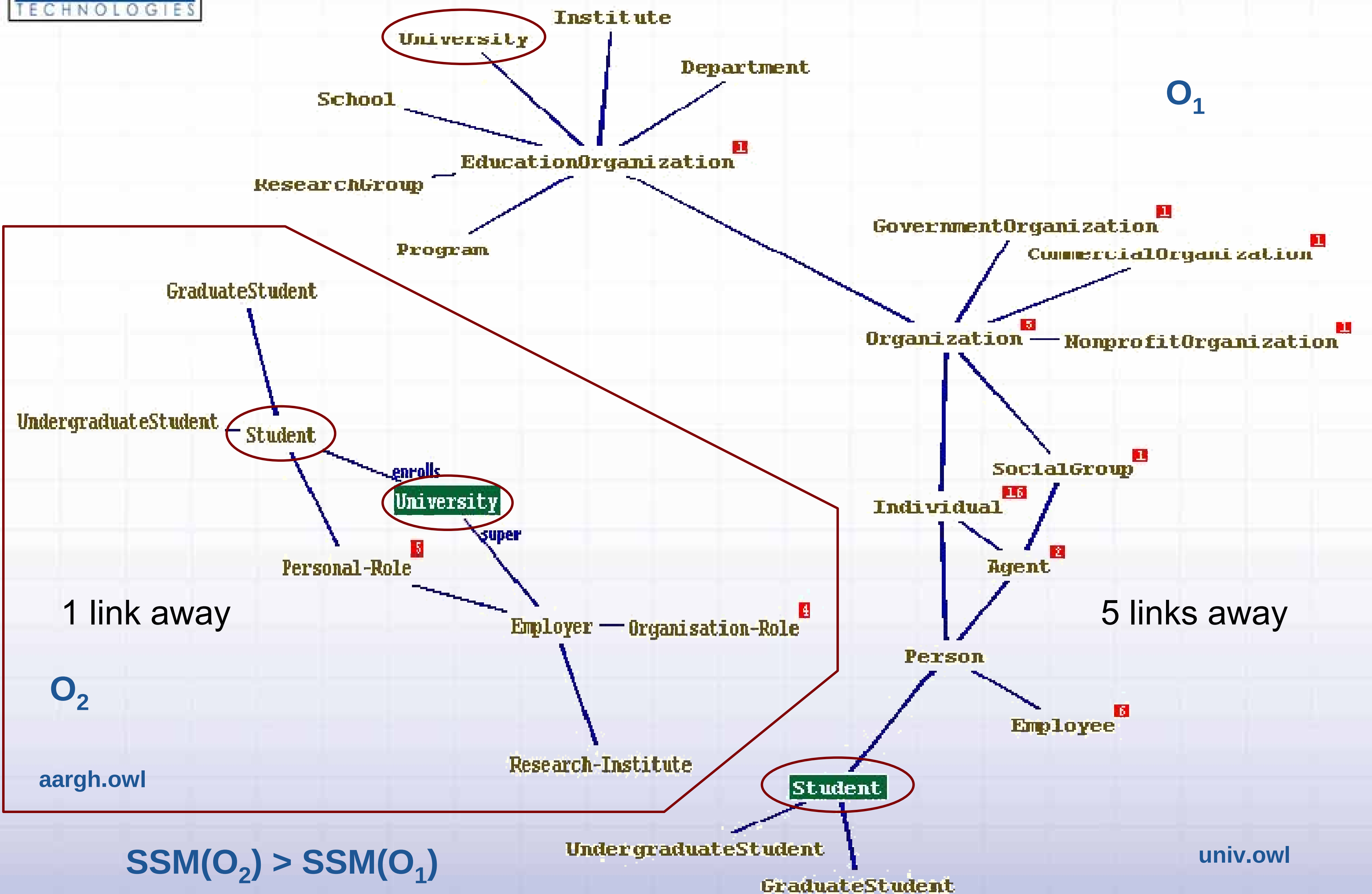
DEM(O_2) > DEM(O_1)



O_2

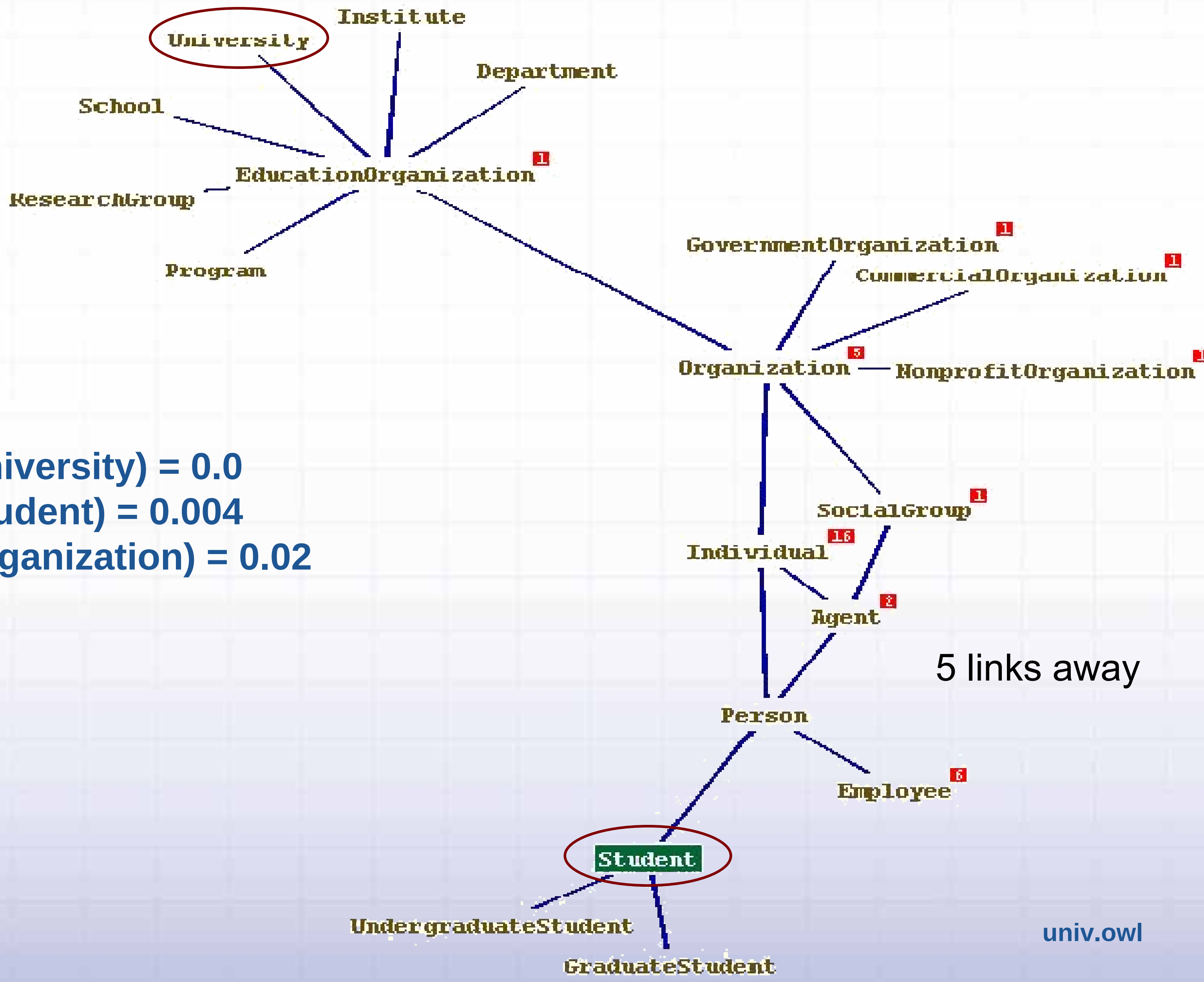


Semantic Similarity Measure (SSM)



SSM(O_2) > SSM(O_1)

Between Measure (BEM)



BEM(University) = 0.0
BEM(Student) = 0.004
BEM(Organization) = 0.02

5 links away

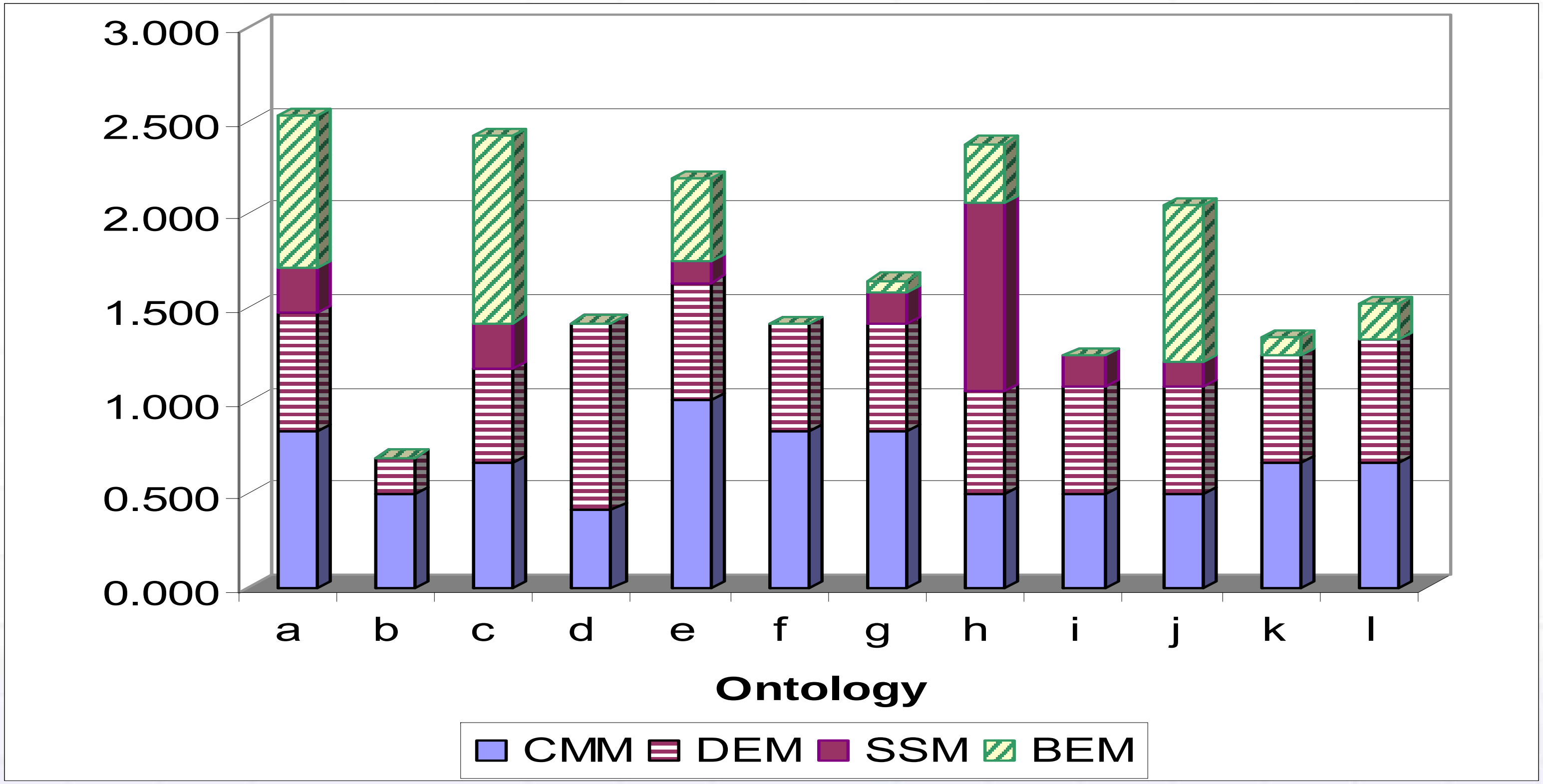
univ.owl

Example

- A query for “Student” and “University” in Swoogle returned the list below:

Pos.	Ontology URL
a	http://www.csd.abdn.ac.uk/~cmckenzi/playpen/rdf/akt_ontology_LITE.owl
b	http://protege.stanford.edu/plugins/owl/owl-library/koala.owl
c	http://protege.stanford.edu/plugins/owl/owl-library/ka.owl
d	http://reliant.teknowledge.com/DAML/Mid-level-ontology.owl
-	http://www.csee.umbc.edu/~shashi1/Ontologies/Student.owl
e	http://www.mindswap.org/2004/SSSW04/aktive-portal-ontology-latest.owl
f	http://www.mondeca.com/owl/moses/univ2.owl
g	http://www.mondeca.com/owl/moses/univ.owl
-	http://www.lehigh.edu/~yug2/Research/SemanticWeb/LUBM/University0_0.owl
h	http://www.lri.jur.uva.nl/~rinke/aargh.owl
-	http://www.srdc.metu.edu.tr/~yildiray/HW3.OWL
i	http://www.mondeca.com/owl/moses/ita.owl
j	http://triplestore.aktors.org/data/portal.owl
k	http://annotation.semanticweb.org/ontologies/iswc.owl
-	http://www.csd.abdn.ac.uk/~cmckenzi/playpen/rdf/abdn_ontology_LITE.owl
l	http://ontoware.org/frs/download.php/18/semiport.owl

AKTiveRank Results



- The figure shows the measure values as calculated by AKTiveRank for each ontology

Content based ranking ..

Revisiting how we search for ontologies

Content-based Ranking

- We observed how people search for ontologies on the Protégé mailing list
  - They tend to search for domains, rather than specific concepts

[protege-discussion] ontolgies and telecom services

[protege-discussion] Hi Help

protege-discussion-bounce@SMI.Stanford.EDU on behalf of Michelle Longsworth [mushe

To: protege-discussion@SMI.Stanford.EDU

Am new to domain ontology development, and really need to understand quickly as I need to develop a hospital domain ontology for my Master Thesis

Has any ontology been develope for hospital before???

Or can some suggestions and recommendations for getting started be given??

PLease am desperate

Mush

Content-based Ranking

- This approach tries to rank ontologies based on the coverage of their concept labels and comments, of the domain of interest

- Steps:
  - Get a query from the user (e.g. Cancer)
  - Expand query with WordNet
  - Retrieve a corpus from the Web that covers this domain
  - Analyse the corpus to get a set of terms that strongly relate to this domain
  - Get a list of potentially relevant ontologies from Google (or Swoogle)
  - Calculate frequency in which those terms appear in the ontology (in concept labels and comments)
  - First rank is awarded to the ontology with the best coverage of the "domain terms"

Getting a Query

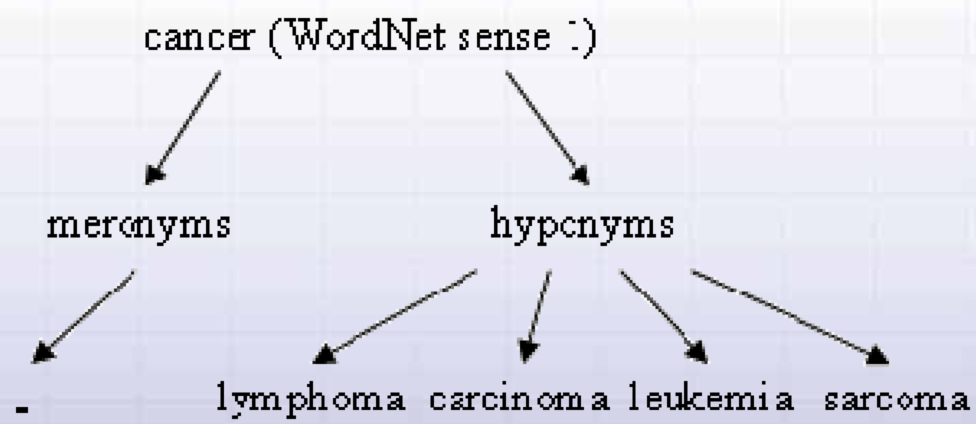
- The query is assumed to give a domain name
  - As in the ontology search queries on Protégé’s mailing list
  - Eg “Cancer” to search for an ontology *about* the domain of cancer
- An ontology that has the concept “Cancer” but nothing much else about the domain is no good!
  - The ontology needs to contain other concepts, related to the domain of Cancer

Expanding with WordNet

- Many documents found on the Web when searching for the given query (eg Cancer) were too general
  - Documents about charities, counselling, fund raisers, general home pages, etc.
  - Need to find documents that discuss the disease
    - Of course we first need to verify which meaning of the word Cancer is the user looking for (more on this later)

- Need to expand the query with more specific words
  - Which is what we usually do when searching online

- Expand query with meronyms and hypernyms of the given term



Finding & Analysing a Corpus

- Use the expanded query to search for documents on the Web
  - Those documents are downloaded and treated as a domain corpus
- Concepts associated with the chosen domain are expected to be frequent in a relevant corpus of documents
- Most discriminating words can be found using traditional text analysis
  - such as tf-idf (text frequency – inverse document frequency)
- The top 50 terms from the result of tf-idf analysis will be used to rank the ontologies
  - Ontologies that contain those terms are given higher ranks than others

Tf-idf with/without WordNet

1. cancer
2. cell
3. breast
4. research
5. treatment
6. tumor
7. information
8. color
9. patient
10. health
11. support
12. news
13. care
14. wealth
15. tomorrow
16. entering
17. writing
18. loss
19. dine
20. mine

1. cancer
2. cell
3. tumor
4. patient
5. document
6. carcinoma
7. lymphoma
8. disease
9. access
10. treatment
11. skin
12. liver
13. leukemia
14. risk
15. breast
16. genetic
17. tobacco
18. thymoma
19. malignant
20. gene

ic Google Search		Expanded Google Search	
	26. teddy		26. lesion
	27. bobby		27. blood
	28. betrayal		28. study
	29. portfolio		29. thyroid
	30. lincoln		30. smoking
	31. inn		31. polyp
	32. endtop		32. human
	33. menuitem		33. health
	34. globalnav		34. exposure
	35. cliphead		35. studies
	36. apologize		36. ovarian
	37. changed		37. information
	38. unavailable		38. research
	39. typed		39. drug
	40. bar		40. related
	41. spelled		41. associated
	42. correctly		42. neoplastic
	43. typing		43. oral
	44. narrow		44. bone
	45. entered		45. chemotherapy
	46. refine		46. body
	47. referenced		47. oncology
	48. recreated		48. growth
24. heard	49. delete	24. Tissue	49. medical
25. signposts	50. bugfixes	25. therapy	50. lung

Find Relevant Ontologies

- Now we need to find some ontologies about Cancer
- This is currently done by searching for owl files in Google given the word “Cancer”
  - Of course others sources can also be used, such as Swoogle
- The list of ontologies is then downloaded to a local database for analyses and ranking
  - Some ontologies will be unavailable or can not be parsed for any reason
  - Ontologies are stored in MySQL for future reuse

Scoring the Ontologies

- Map the set of terms found earlier to each ontology found in our search
  - Each ontology will be scored based on how well it covers the given terms

- The higher the term is in the tf-idf list, the higher its weight
  - So each word is given an importance value
  - This needs to be considered when assessing the ontologies
  - E.g. An ontology with concepts whose labels match the top ten tf-idf words would outrank an ontology with only the second ten words matching.

- Two scores are calculated using two formulas:
  - Class Match Score (CMS): to match with concepts labels
  - Literal Match Score (LMS): to match with comments and other text

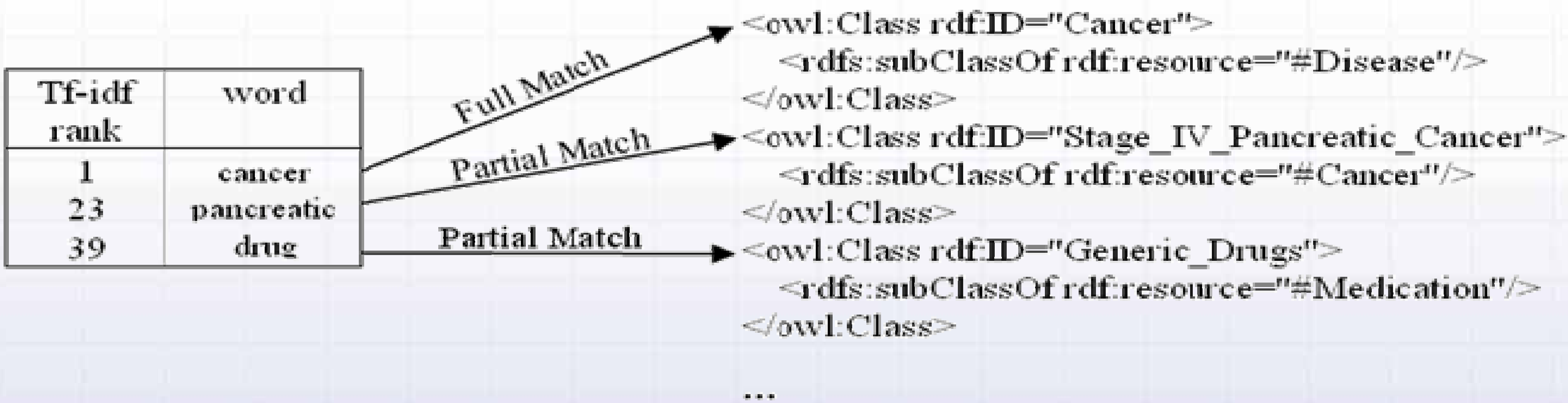
- Total score = α CMS + β LMS
  - α and β are weights to control the two scoring formulas

Class Match Score

$$CMS[o \in O] = \sum_{i=1}^{n} I(P_i, o) \times 5 \log (n + 2 - i)$$

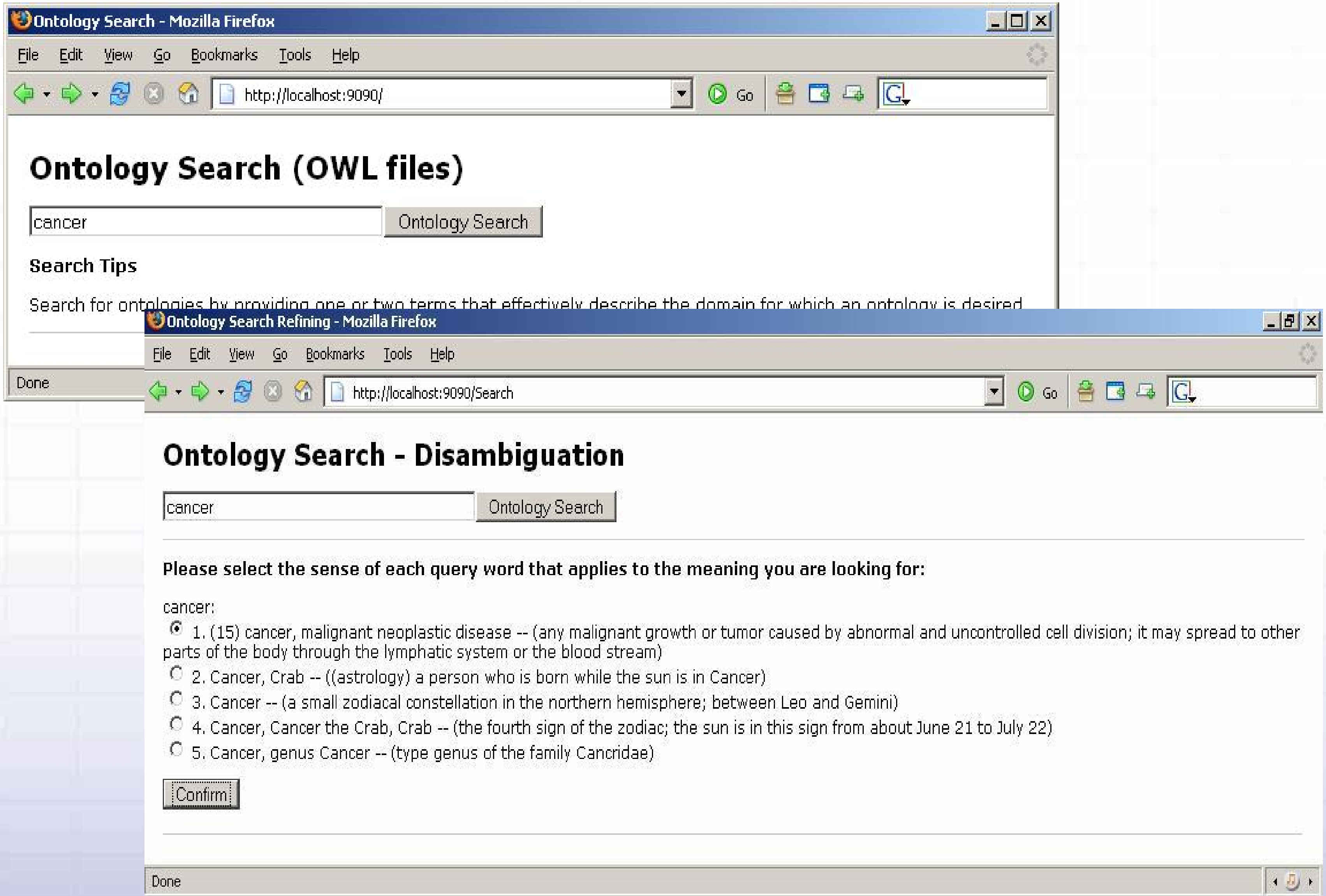
$$I(P_i, o) = \begin{cases} 1 & : \text{ if } o \text{ contains a class with label matching } P_i \\ 0.4 & : \text{ if } o \text{ contains a class with label which contains } P_i \\ 0 & : \text{ if } P_i \text{ does not appear in any of } o\text{'s class labels} \end{cases}$$

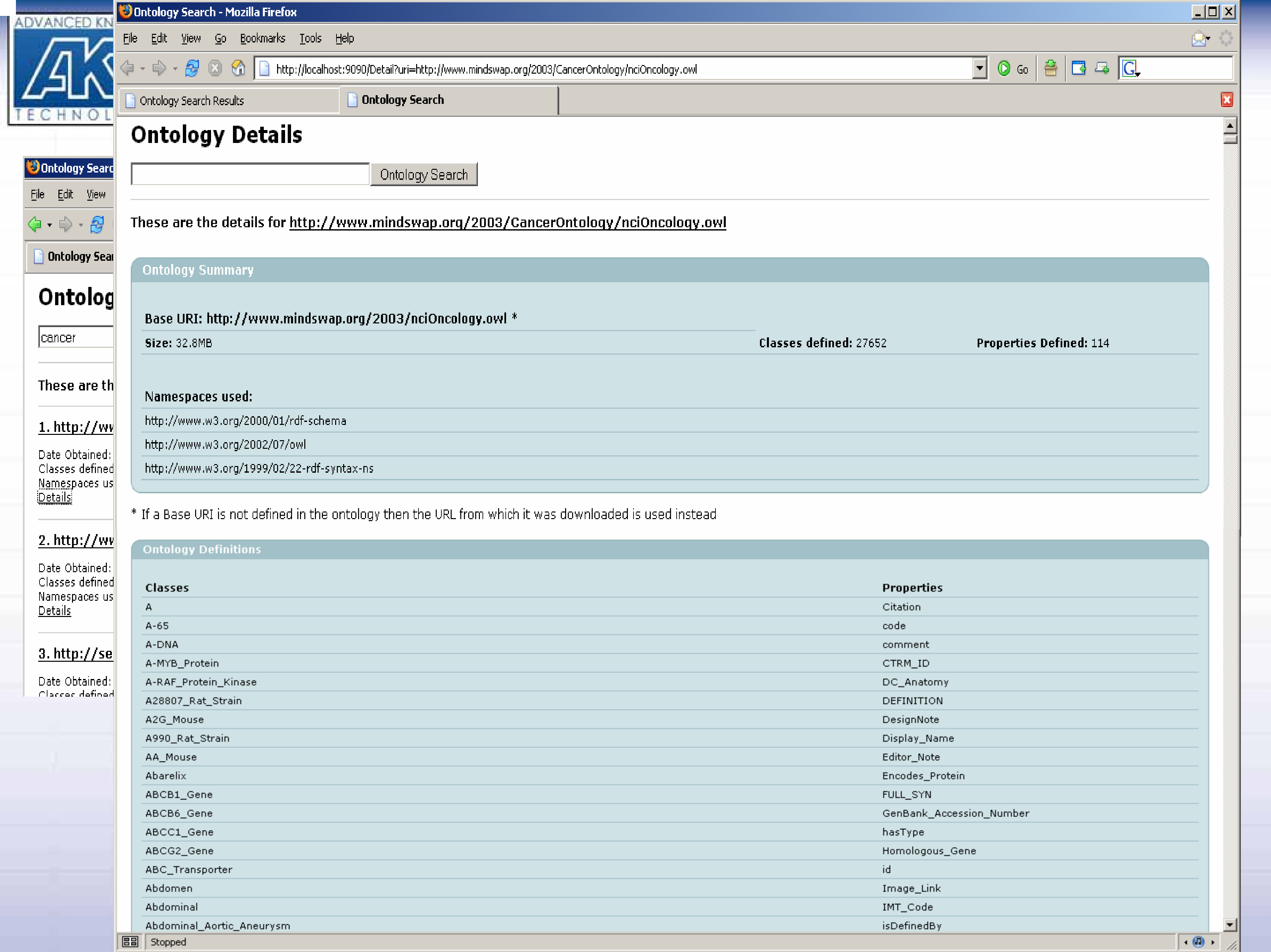
- Uses weights to control exact and partial matching
  - Eg 1 for a full match, 0.4 for a partial match, 0 for no match



$$\begin{aligned} CMS &= 1 \text{ x } 5 \text{ x } \log(52-1) + 0.4\text{x } 5 \text{ x } \log(52-23) + 1 \text{ x } 5 \text{ x } \log(52-39) \\ &= 8.538 + 2.925 + 2.229 \\ &= 13.691 \end{aligned}$$

Interface


Ontology Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:9090/
Go
Ontology Search (OWL files)
cancer
Ontology Search
Search Tips
Search for ontologies by providing one or two terms that effectively describe the domain for which an ontology is desired
Done
Ontology Search Refining - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:9090/Search
Go
Ontology Search - Disambiguation
cancer
Ontology Search
Please select the sense of each query word that applies to the meaning you are looking for:
cancer:
1. (15) cancer, malignant neoplastic disease -- (any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream)
2. Cancer, Crab -- ((astrology) a person who is born while the sun is in Cancer)
3. Cancer -- (a small zodiacal constellation in the northern hemisphere; between Leo and Gemini)
4. Cancer, Cancer the Crab, Crab -- (the fourth sign of the zodiac; the sun is in this sign from about June 21 to July 22)
5. Cancer, genus Cancer -- (type genus of the family Cancridae)
Confirm
Done



Ontology Details

Ontology Search

These are the details for http://www.mindswap.org/2003/CancerOntology/nciOncology.owl

Ontology Summary

Base URI: http://www.mindswap.org/2003/nciOncology.owl *

Size: 32.8MB **Classes defined:** 27652 **Properties Defined:** 114

Namespaces used:

http://www.w3.org/2000/01/rdf-schema

http://www.w3.org/2002/07/owl

http://www.w3.org/1999/02/22-rdf-syntax-ns

* If a Base URI is not defined in the ontology then the URL from which it was downloaded is used instead

Ontology Definitions

Classes	Properties
A	Citation
A-65	code
A-DNA	comment
A-MYB_Protein	CTRM_ID
A-RAF_Protein_Kinase	DC_Anatomy
A28807_Rat_Strain	DEFINITION
A2G_Mouse	DesignNote
A990_Rat_Strain	Display_Name
AA_Mouse	Editor_Note
Abarelix	Encodes_Protein
ABCB1_Gene	FULL_SYN
ABCB6_Gene	GenBank_Accession_Number
ABCC1_Gene	hasType
ABCG2_Gene	Homologous_Gene
ABC_Transporter	id
Abdomen	Image_Link
Abdominal	IMT_Code
Abdominal_Aortic_Aneurysm	isDefinedBy

Experiment

- Searching for ontologies about "Cancer"
- Use different sets of weights to calculate final ranks
- Compare results with ranks given by human experts
  - This helps to find out which settings produce the best results
- The list of ontologies used in the experiment are:

ID	URL
1	http://semweb.mcdonaldbradley.com/OWL/Cyc/FreeToGov/060704/FreeToGovCyc.owl
2	http://www.inf.fu-berlin.de/inst/agnbi/research/swpatho/owldata/swpatho1/swpatho1.owl
3	http://www.mindswap.org/2003/CancerOntology/nciOncology.owl
4	http://sweet.jpl.nasa.gov/ontology/data_center.owl
5	http://compbio.uchsc.edu/Hunter_lab/McGoldrick/DataFed_OWL.owl
6	http://www.cs.umbc.edu/~aks1/ontosem.owl
7	http://homepages.cs.ncl.ac.uk/phillip.lord/download/knowledge/ontologyontology.owl
8	http://www.daml.org/2004/05/unspsc/unspsc.owl
9	http://envgen.nox.ac.uk/miame/MGEDOntology_env_final.owl
10	http://www.fruitfly.org/~cjm/obo-download/obo-all/mesh/mesh.owl

Experiment 1

- Experimenting with exact and partial matching of class labels
  - To test the effect of partial matching on the overall result

- Three sets of weights are used:

Experiment	Exact Match	Partial Match
a	1	0.4
b	1	0
c	1	1

Results of Experiment 1

Ontology ID	Experiment 1(a) Ranks (1,0.4)	Experiment 1(b) Ranks (1,0)	Experiment 1(c) Ranks (1,1)
1	3	3	2
2	6	8	6
3	1	1	1
4	9	10	7
5	5	4	5
6	2	2	4
7	10	9	10
8	4	6	3
9	7	7	8
10	8	5	9

- Ranks for some ontologies remained relatively stable
  - Indicating having few class labels that partially match the words retrieved from the domain corpus
- Other ranks fluctuated, such as for ontologies 4,8,10
  - These ontologies contain more partially matching class labels than the other ontologies

Experiment 2

- Experimenting with matching class labels as well as comments
  - To test the effect of matching comments on the overall result
- Three sets of weights are used:

Experiment	Class Match	Text Match
a	1	0.25
b	1	0
c	1	1

Results of Experiment 2

Ontology ID	Experiment 2(a) Ranks (1,0.25)	Experiment 2(b) Ranks (1,0)	Experiment 2(c) Ranks (1,1)
1	3	3	3
2	6	7	7
3	1	1	1
4	9	6	9
5	5	5	6
6	2	2	2
7	10	9	10
8	4	4	4
9	7	8	8
10	8	10	5

- Some ranks fluctuated, such as for ontologies 4,10
  - Ontology 10 is well commented, while ontology 4 is not!
  - Matching with comments increased the total scores of commented ontologies
  - Note that these comments had Cancer related words

Evaluation

- To evaluate the ranks given by the system, we need humans to rank those ontologies
- Evaluation involved three “experts”
  - Two 3rd year medical students which enough knowledge about the chosen domain
  - One computer science lecturer with a lot of experience in medical ontologies
- The experts were given the freedom to browse and visualise the ontologies in Protégé
- Each expert was asked to provide a rank for each ontology, and a short comment

Example Result from an Expert

Ontology URL:	http://semweb.mcdonaldbradley.com/OWL/Cyc/FreeToGov/060704/FreeToGovCyc.owl	Rank:	6
Comments:	Too General, nothing specific		
Ontology URL:	http://www.inf.fu-berlin.de/inst/ag-nbi/research/swpatho/owldata/swpatho1/swpatho1.owl	Rank:	4
Comments:	Lots of anatomy- some could be cancer related, lots of cardiopulmonary and breathing tract concepts defined		
Ontology URL:	http://www.mindswap.org/2003/CancerOntology/nciOncology.owl	Rank:	1
Comments:	Contains genetic, anatomical & treatment related concepts for cancer		
Ontology URL:	http://sweet.jpl.nasa.gov/ontology/data_center.owl	Rank:	10
Comments:	Completely useless		
Ontology URL:	http://compbio.uchsc.edu/Hunter_lab/McGoldrick/DataFed_OWL.owl	Rank:	7
Comments:	Contained some very basic cancer concepts. Lots of other medical concepts, mostly unrelated		

Ranking Results by Experts

- These are the results provides by our three experts
- Note that the average Pearson Correlation Coefficient between these results is 0.8, indicating high agreement
  - PCC value of +1 is a perfect match, 0 is no correlation, -1 is an inverse relationship

Ontology ID	Participant 1	Participant 2	Participant 3	Average Rank
1	6	6	5	6
2	4	5	4	3
3	1	1	1	1
4	10	10	9	10
5	7	3	3	3
6	3	4	8	5
7	8	9	7	8
8	8	8	10	9
9	5	7	6	7
10	2	2	2	2

Comparison of Results

- Ranks are compared using Pearson Correlation Coefficient values
- Compare results of experiments 1 and 2 against ranks given by experts

α is weight for class labels

β is weight for comments

Class match weights / α & β values	Full: 1 Partial: 0	Full: 1 Partial: 0.4	Full: 1 Partial: 1
α: 1 β: 0	0.304331206	0.142133811	0.08291139
α: 1 β: 0.25	0.627757665	0.426401433	0.236889685
α: 1 β: 1	0.568535244	0.509312822	0.497468338

- Same as above, but using a corpus made up from Wikipedia pages only

Class match weights / α & β values	Full: 1 Partial: 0	Full: 1 Partial: 0.4	Full: 1 Partial: 1
α: 1 β: 0	0.341748977	0.260578653	0.165822779
α: 1 β: 0.25	0.651446633	0.426401433	0.272423138
α: 1 β: 1	0.521157307	0.509312822	0.438245917

Results

- Best result was when:
  - Partial matching was ignored (partial weight = 0)
  - Some emphases is given to literal text matching (β = 0.25), but not much more than that!
    - Results deteriorated with $\beta = 1$
  - Limiting the corpus to Wikipedia
    - This generated slightly better results, but nothing significant!
    - Wikipedia might not be a suitable corpus for some domains

Conclusions

- Some broad ontologies ranked high in our system, but disliked by the experts for being too general
  - They contained many of the terms found in the corpus, but with minimum detail
  - Overall focus of the ontologies was not on the chosen domain
  - Perhaps an ontology should be penalised if it had many terms that are definitely not related to the domain
  - Adding extra tests might also help to filter out such ontologies, such as density and betweenness
- Evaluation was based on only 3 people!
  - No statistical significance can be claimed
  - Difficult for people to assess an ontology
- Use of Wikipedia was good, but limiting the corpus to it is unwise
  - Some domains might not be well covered in Wikipedia
  - Of course finding a good corpus on the web can not be guaranteed either
- Use of WordNet is good for disambiguating query terms
  - But WordNet might not cover the given term
  - Cost of an additional layer of user interaction

Further Work

- Get someone to continue this work
- More test, using different settings
- Compare and perhaps merge with AKTiveRank
- Penalise ontologies with terminology that is *outside* the given domain of interest