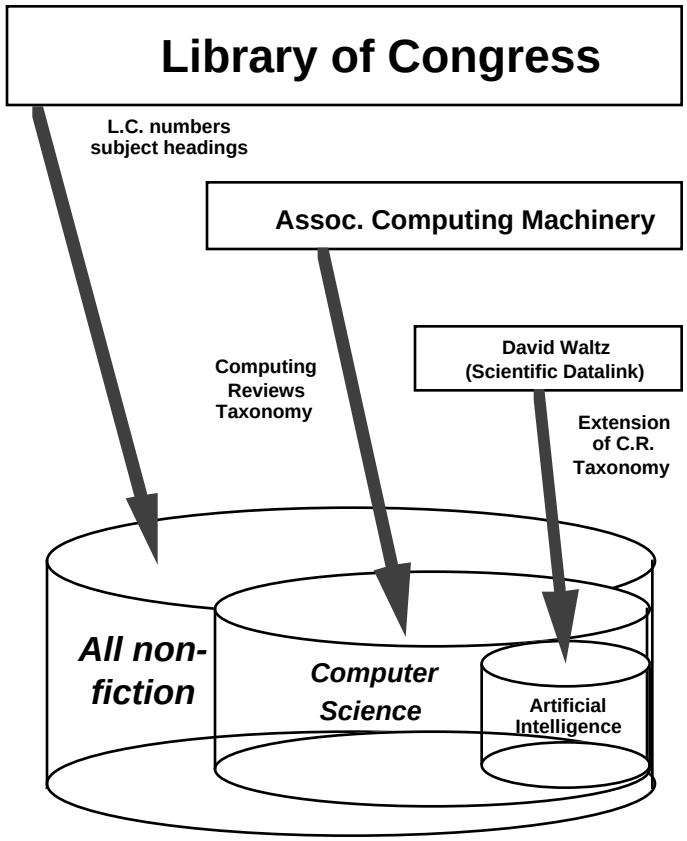
Intro

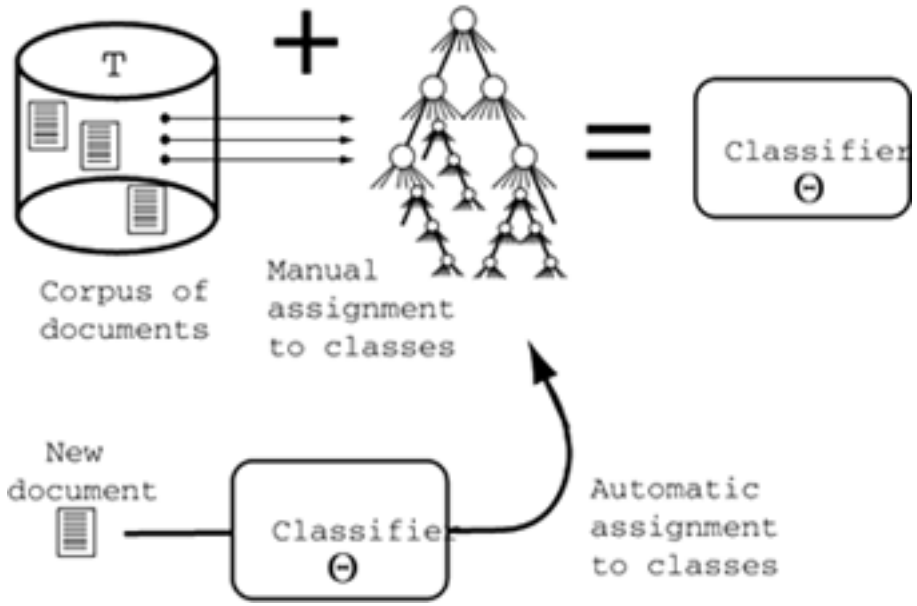
- *Encyclopedia Brittanica*
- *La Jolla Research Group*

Nested taxonomies

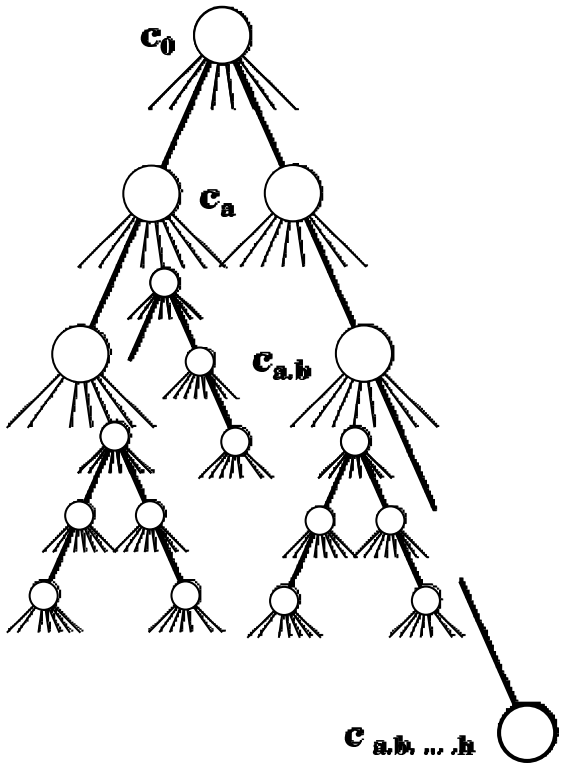
- Increasing topical precision
- Fewer people affected



Exploiting manual training



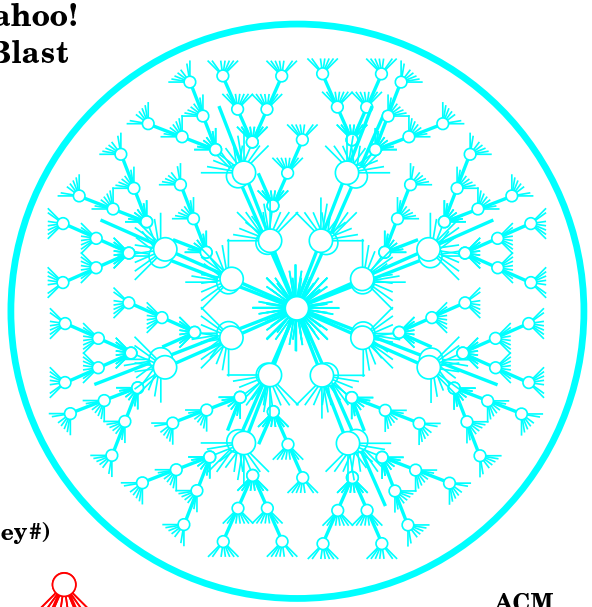
Hierarchic classification



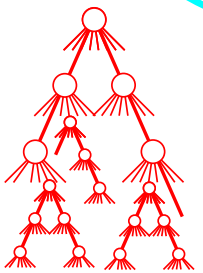
	Pr(belief)
UseNet	Pr(belief | UseNet)
alt	Pr(belief | alt)
alt.religion	Pr(belief | alt.religion)
alt.religion.scientology	Pr(belief | alt.religion.scientology)

Docking

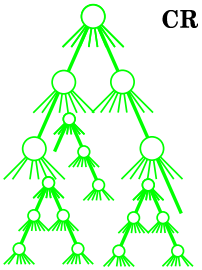
Library of Congress (Subject Headings)
Ency. Britannica (Propaedia)
Yahoo!
eBlast



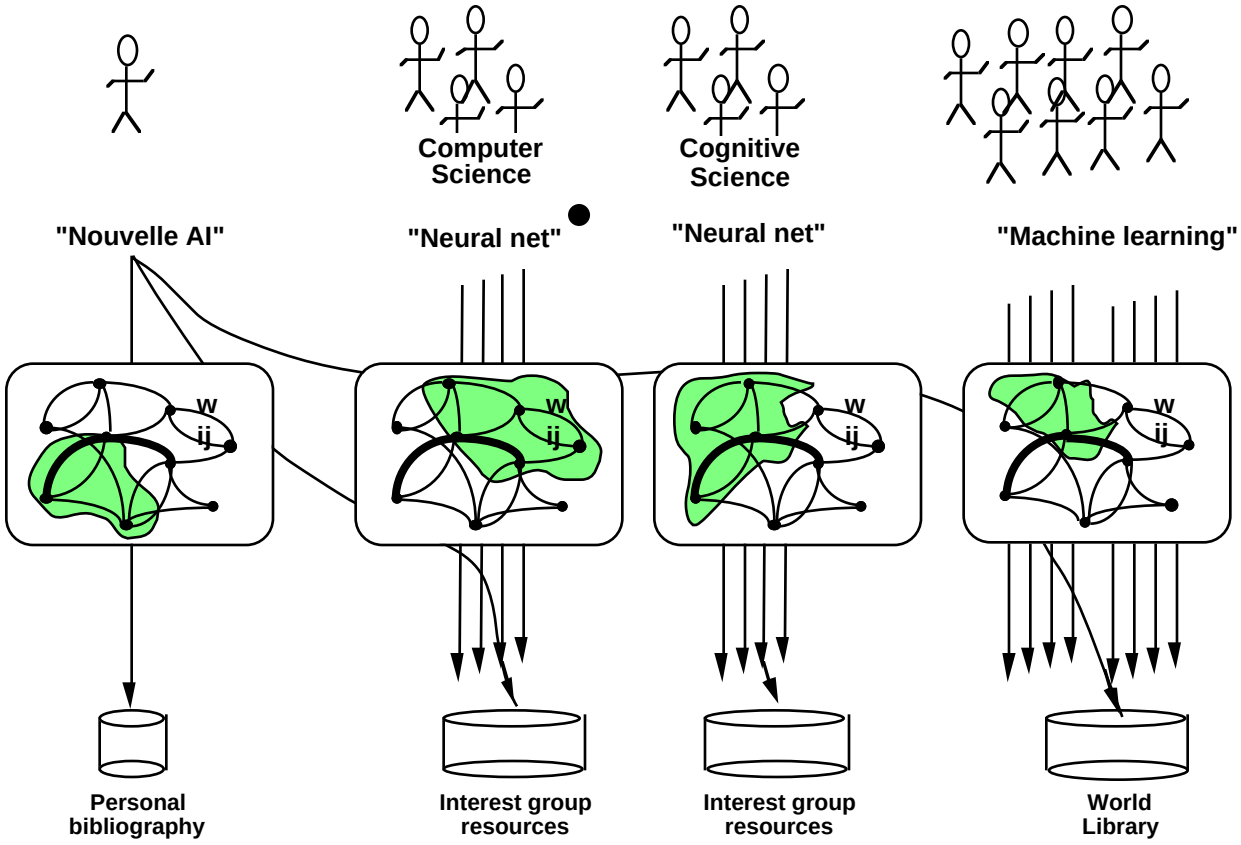
WestLaw (Key#)



ACM
CR Taxonomy



Adaptive lenses



Semantics of hierarchy

- 'Classical' BT/NT hyernymy
- AI's IS_A inheritance
- Single authorship (EB,ACM) vs. consentual (DMOZ)
- General survey prose (EB)
- Pedagogical ordering (ACM, CSE Curricula)
- Special categories
  - General, Misc (EB, ACM)
  - People, Conferences, Publications (DMOZ)

Algorithmic details

- Hierarchic structure vs. pairwise match
- Bipartite matching:
  - Accomplishes pairwise matches only
- Form "Tree association graph"
  - Integrates holistic hierarchical constraints
- Transform to MaxClique
- Relaxing to MaxClique
- Convergence properties of replicator equations, QPOPT

Algorithmic details

- vis a vis other optimization procedures
- [Mark Chaisson, UCSD undergrad]

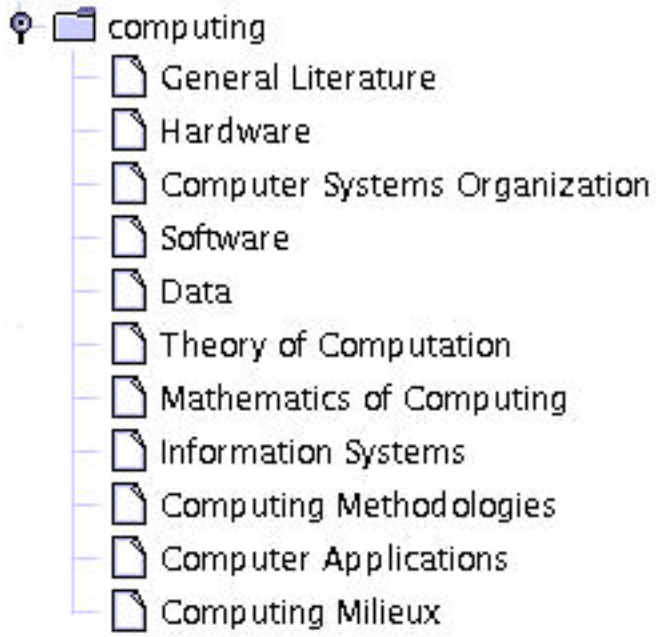
Methodological preliminaries

- HierML
  - Rubrics
  - example texts
- SMOOSH operator
  - Coalescing children's rubrics to generate larger textual samples

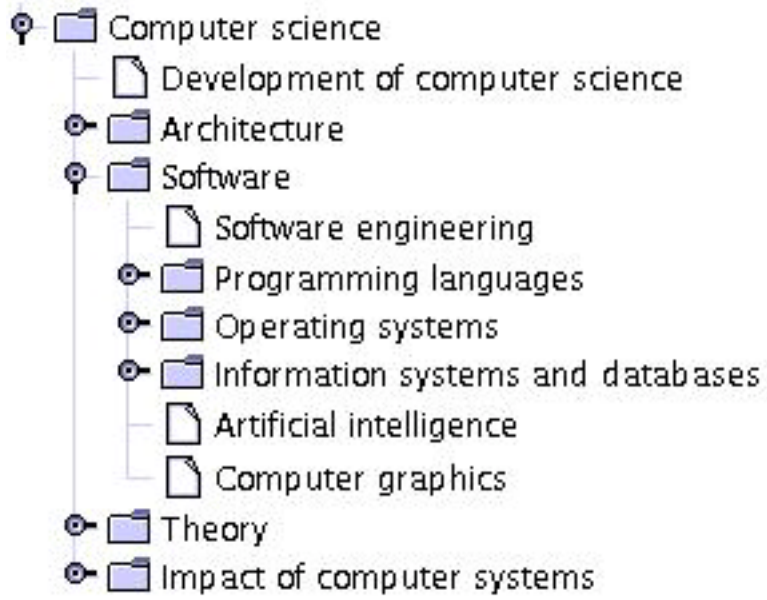
Data sets

- ACM
- Encyclopedia Britannica
- DMOZ: Open Directory
- Summary of data sets

ACM

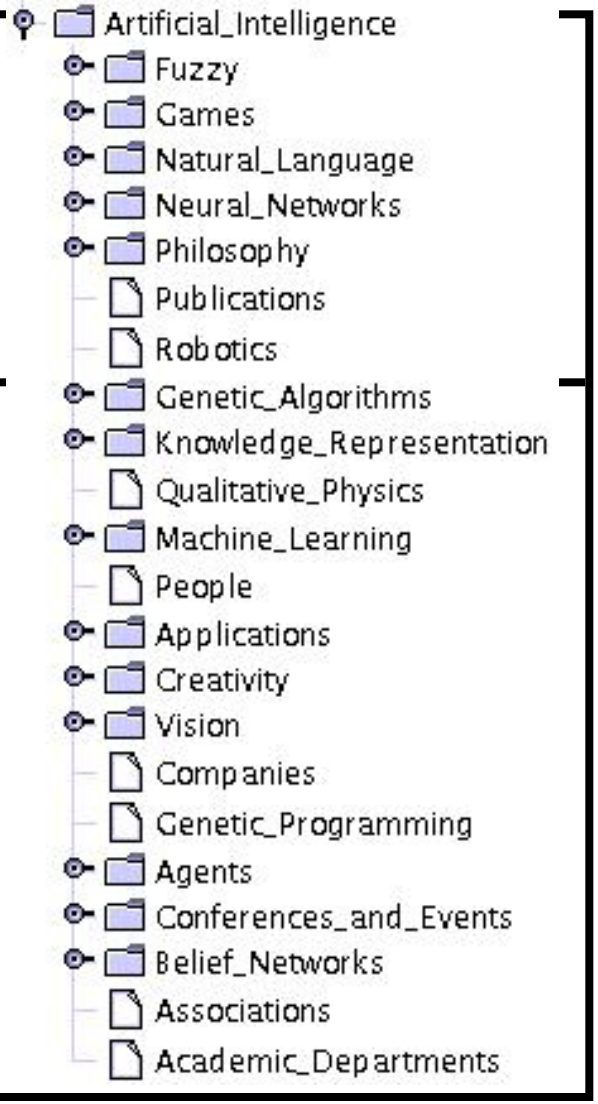


Encyclopedia Britannica



DMOZ: Open Directory

- Paritioning training data
  - DMOZ1 vs. DMOZ2
- Testing robustness across editors' selections



Summary of data sets

- ACM
  - 2,3 layers 'smooshed'
  - Full: examples of classified abstracts
- DMOZ
  - CS, AI, NatLang
  - First v. second partition
- INSPEC
- EB

Name	Nodes	Depth	Size
acm2	12	2	975
acm3	93	3	975
acm2full	12	2	9286
acm3full	93	3	9286
dmoz-nl1	8	3	4465
dmoz-nl2	8	3	4173
dmoz-ai1	74	5	17286
dmoz-ai2	75	5	18570
dmoz-cs	137	5	24544
inspec	202	4	20311
eb	50	5	1678

Experiments

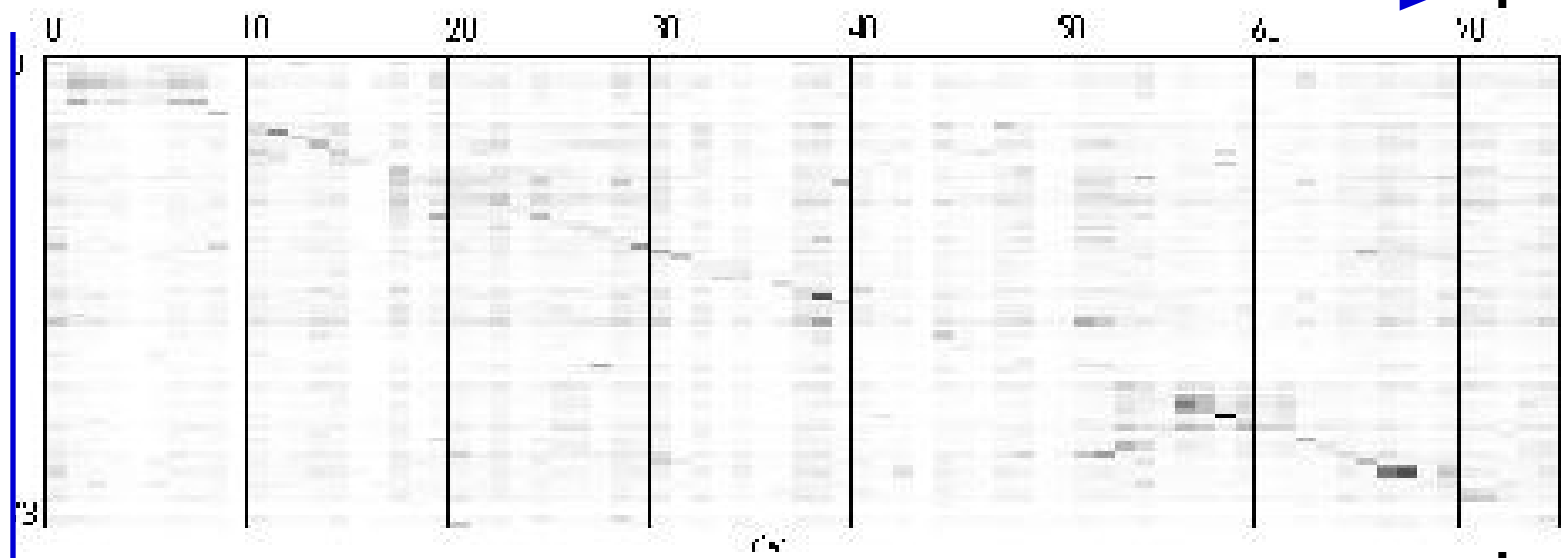
- DMOZ: AI1 vs. AI2
- "Docking" DMOZ:NL2 within DMOZ:AI1
- EB v. ACM
- INSPEC v. ACM

DMOZ: AI1 vs. AI2

- Similarity scores
- Bipartite match
- TAG/MaxClique
- Bipartite match
- TAG/MaxClique match

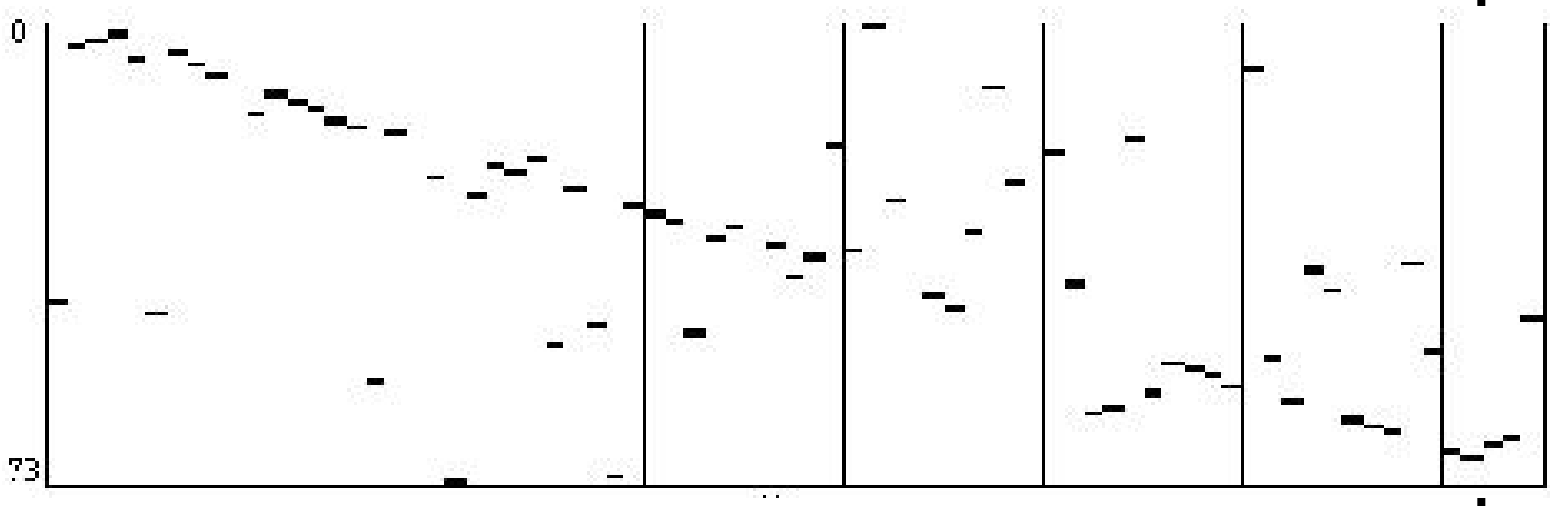
Similarity scores

DMOZ: AI1

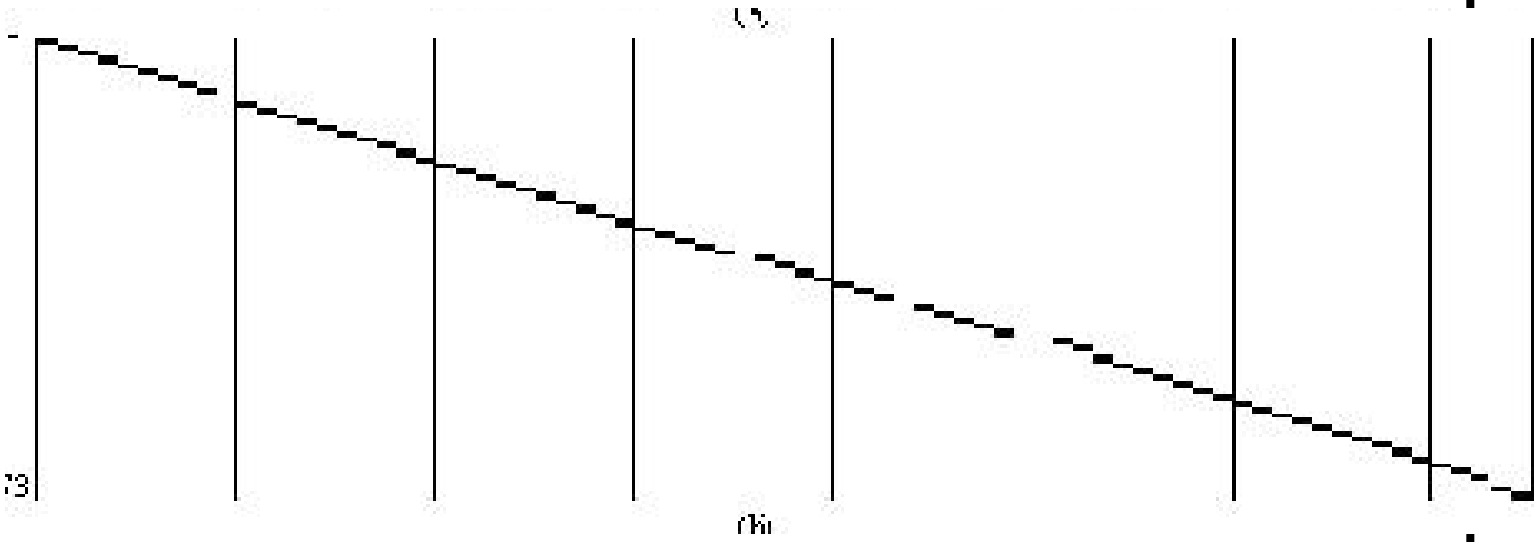


DMOZ: AI2

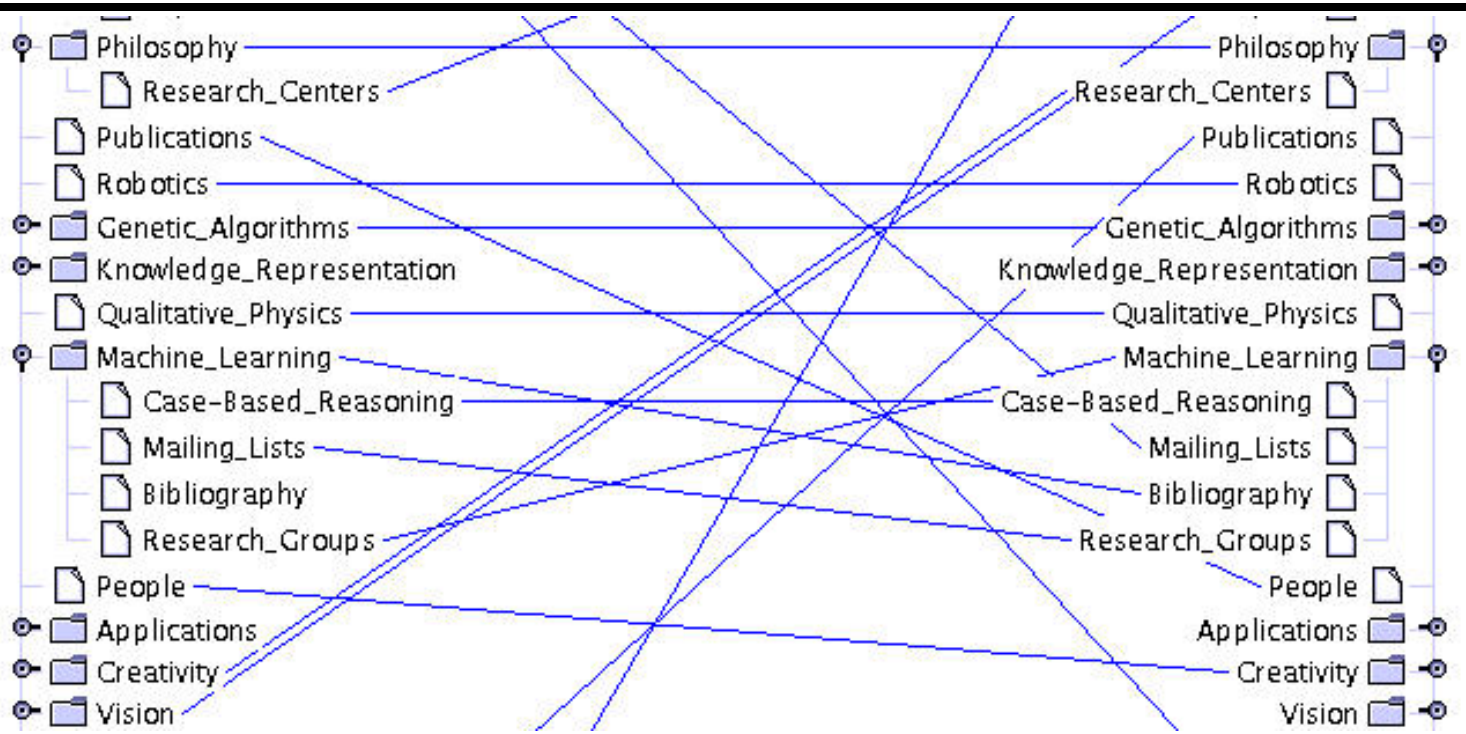
Bipartite match



TAG/MaxClique

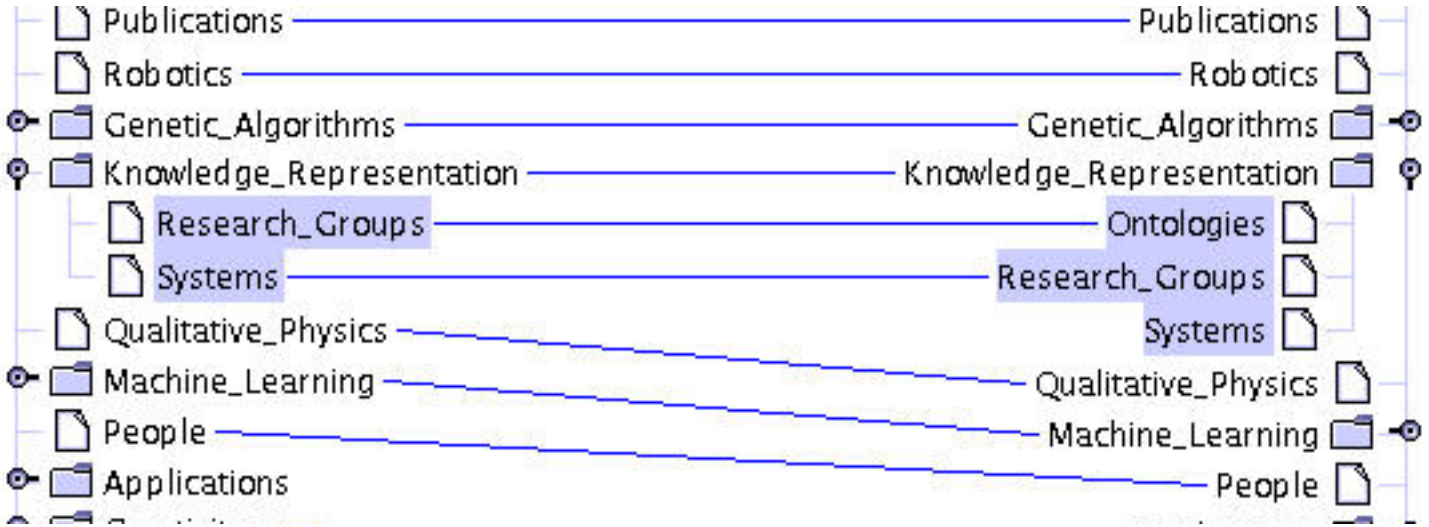


Bipartite match

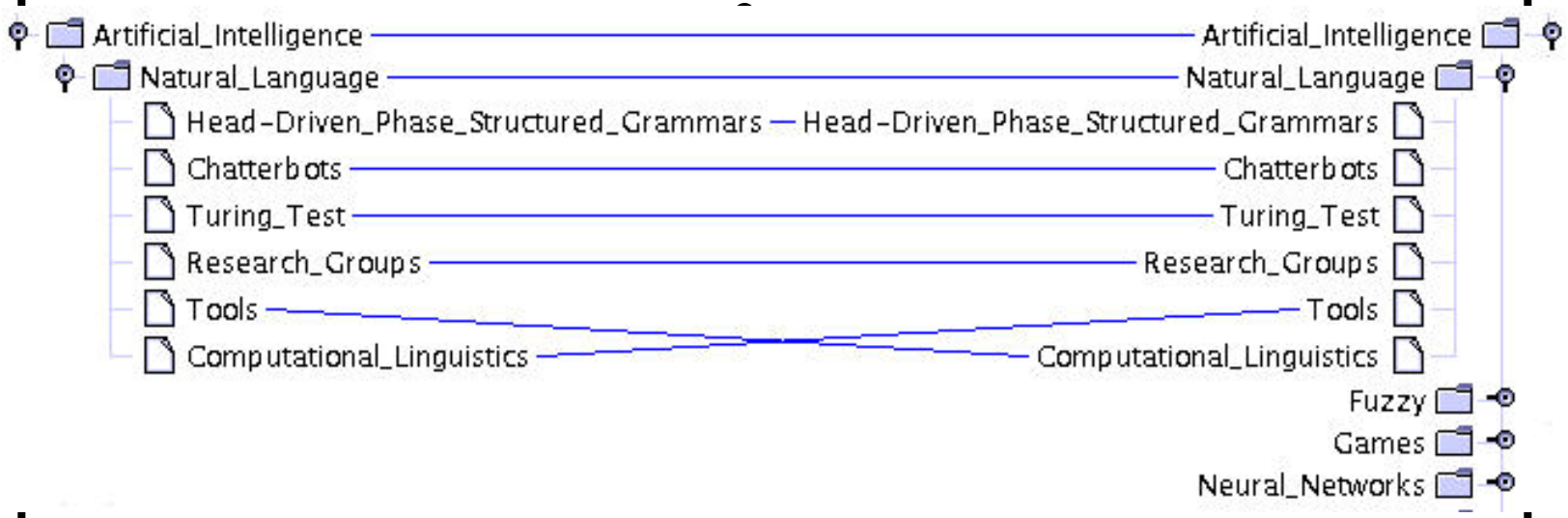


- DMOZ: AI1 vs. AI2

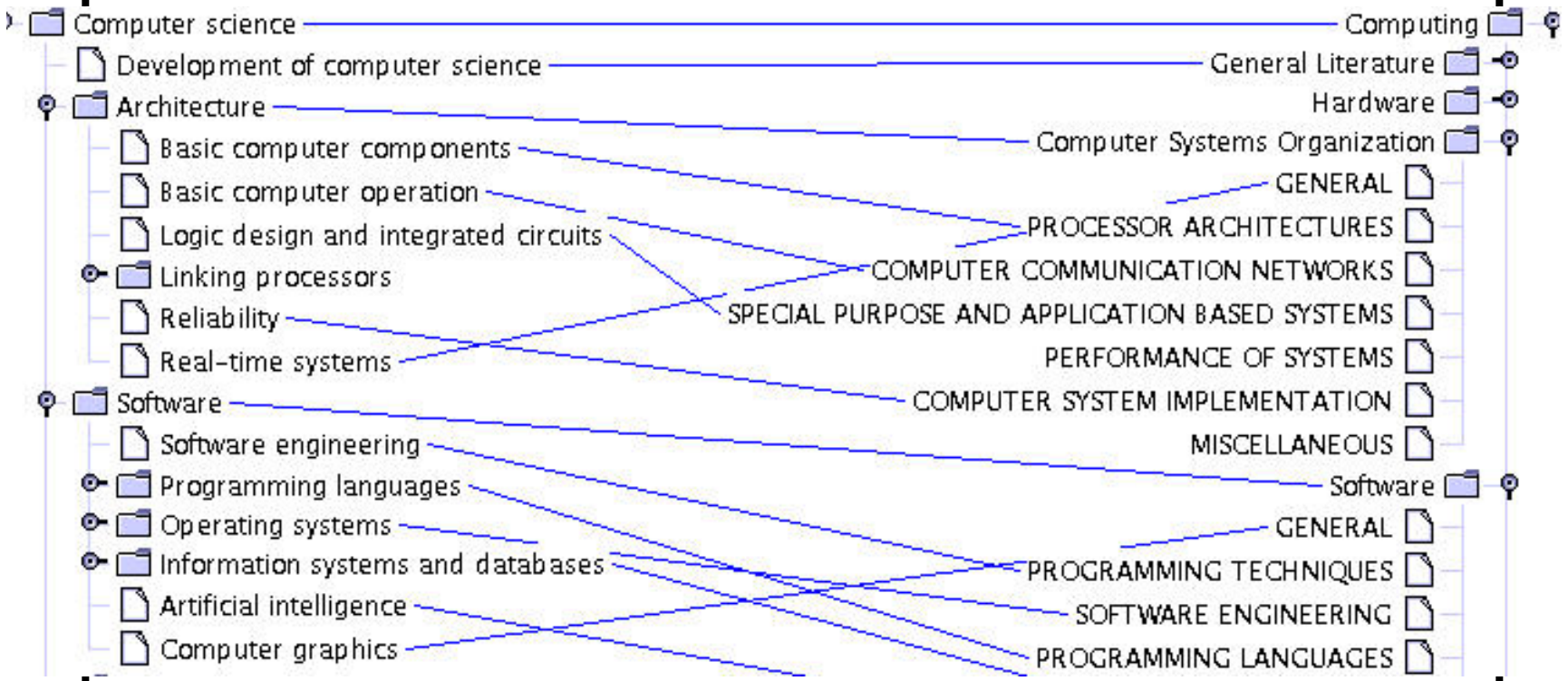
TAG/MaxClique match



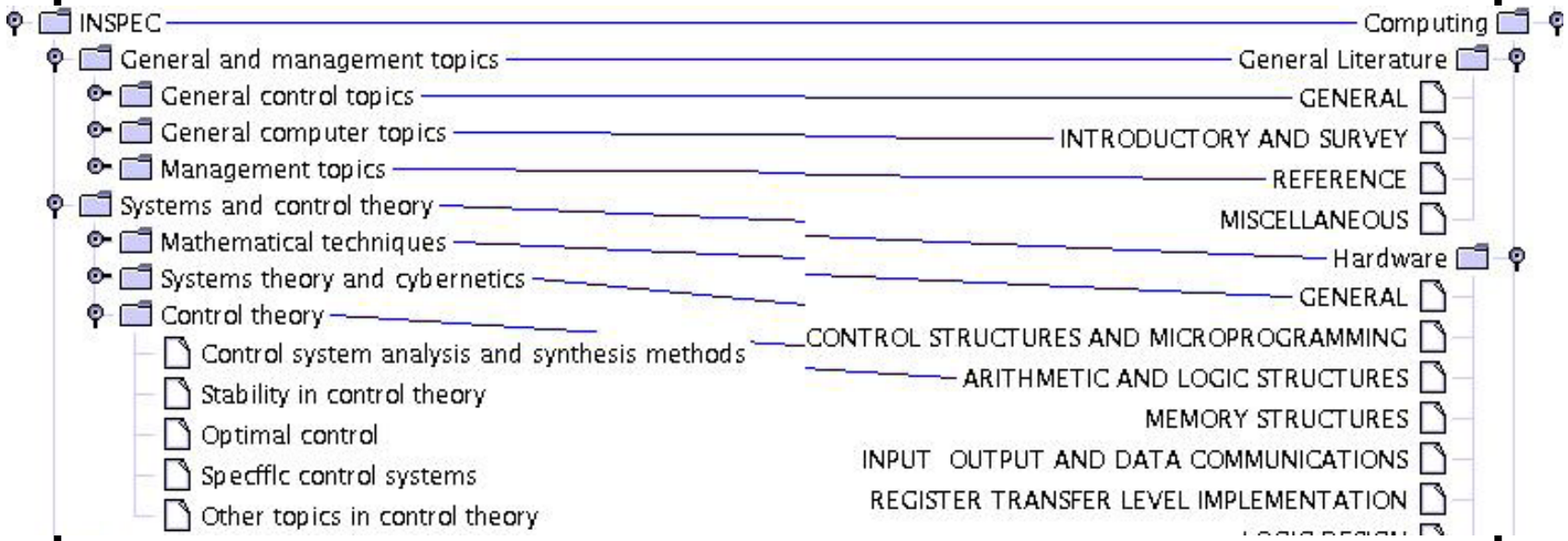
“Docking” DMOZ:NL2 within DMOZ:AI1



EB v. ACM



INSPEC v. ACM



2do: Iterative program

- sensitive to SCALE of match
- balance strengths/weaknesses of structure/bipartite match
- Edit distance of add/delete, merge operators

Goals of a match

- Integration of focused corpus (portal) into a broader context
- Differentiated language use
  - Exported phrases [Steier&Belew'97]
  - Inside/outside vocabulary
- Identifying survey texts