Web Mining – The Ontology Approach

Ee-Peng Lim
Nanyang Technological University
Singapore

Work with: Aixin SUN, Dion GOH, Ming YIN, Myo-Myo NAING, Zhen SUN, Maria MARISSA and other members of G-PortalDL group.

The International Advanced Digital Library Conference (IADLC), Nagoya, 2005

Why This Talk?

- Hector Garcia-Molina at JCDL2005
  “Digital Libraries Initiatives: What I learned (and didn't) in 10 years”
- World Wide Web Tsunami:
  - Enormous volume and coverage of content
    (Everything is free? Heterogeneity?)
  - Large number of users
    (No difference between producers and consumers)
  - Vast number of computers and devices
    (Many different applications are possible)

Some interesting statistics

- Google indexes
  - > 8 billion web pages
  - > 2 billion images
  - > 1 billion Usenet messages
- Nielsen's May 2004 survey
  - An average surfer went online 30 times for > 24 hours in total during a month
    - 1 time per day
    - 45 mins per day

What are the implications to Digital Libraries?

- Search: OPACS vs Google
- Browse: Books vs Web Pages/E-Articles
- Classification System: Dewey Decimal Classification vs Yahoo! & DMOZ
- Definition of Terms: Encyclopedia vs Wikipedia (www.wikipedia.org/)
- Users: Library cards vs User blogs

Wikipedia


Digital library - Wikipedia - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://en.wikipedia.org/wiki/Digital_library
Go
Google wikipedia
Search Web 15 blocked AutoFill Options wikipedia
article
discussion
edit this page
history
Wikimedia needs your help in its US$200,000 fund drive. See our fundraising page for details.
Digital library
From Wikipedia, the free encyclopedia.
A digital library comprises digital collections, services and infrastructure to support lifelong learning, research, scholarly communication and preservation.
A digital library is, like a traditional library, a collection of books and reference materials. Unlike a traditional library, however, the collection of a digital library is, as you would expect, digital, and is usually served over the World Wide Web. Some of the largest and most successful digital libraries are Project Gutenberg, ibiblio and the Internet Archive.
Some people have criticized that digital libraries are hampered by copyright law, because works cannot be shared over different periods of time in the manner of a traditional library. The content is, in many cases, public domain or self-generated content only. Some digital libraries, such as Project Gutenberg, work to digitize out-of-copyright works and make them freely available to the public.
Genre publisher Baen Books has made many of its titles available electronically through the Baen Free Library.
WIKIPEDIA
The Free Encyclopedia
navigation
Main Page
Community portal
Current events
Recent changes
Random article
Help
Contact us
Donations
search
Go
Search
toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Internet

User Blog


Why are you worshipping the ground I blog on? - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://xiaxue.blogspot.com/
Go
Blogger
SEARCH
BlogThis!
GET YOUR OWN BLOG
FLAG?
NEXT BLOG»
The Essential Links
The Story thus far... & FAQs!
Read this before you email me asking me questions everyone else has asked before, thus resulting in my ignoring you and you getting pissed.
CHARACTERS In My Life
Don't know who I'm talking about? Fear not! Click here to find out! It's outdated though, I can't be bothered to update it. Cheerios!
Photos For your viewing pleasure
Teleports you to immediately my Multiply page! And add me as an online buddy while you are there, will ya?
MEDIA CENTER
TV interviews? Press interviews? Radio interviews? Wanna see/read/hear them?
Internet

What are the implications to CS researchers?

- Large amount of Web information waiting to be processed
- Semantic Web
- But there are technical challenges!
  - Unstructured and semi-structured content
  - Links, links, links….
  - Large of discipline
  - Dynamic Web

Use of Web Mining

- Types of Web mining:
  - Web content mining
  - Web usage mining
  - Web link mining
  - Web information extraction
- Web mining for addressing the challenges
- Ontology-based web (content) mining

Outline

- What is an Ontology?
- Ontology-based Web Mining
- Homepage Mining
- Homepage Relationship Mining
- Conclusion


Ontology - Wikipedia, the free encyclopedia - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://en.wikipedia.org/wiki/Ontology
Go
Create account / log in
article discussion edit this page history
WIKIPEDIA
The Free Encyclopedia
navigation
Main Page
Community portal
Current events
Recent changes
Random article
Help
Contact us
Donations
search
Go Search
Ontology
From Wikipedia, the free encyclopedia.
This article discusses ontology in philosophy. For the term in computer science, see ontology (computer science).
In philosophy, ontology (from the Greek όντος = part. of ειναι = being and λόγος = word/speech) is the most fundamental branch of metaphysics. It studies being or existence as well as the basic categories thereof—trying to find out what entities and what types of entities exist. Ontology has strong implications for the conceptions of reality.
Some philosophers, notably of the Platonic school, contend that all nouns refer to entities. Other philosophers contend that some nouns do not name entities but provide a kind of shorthand way of referring to a collection (of either objects or events). In this latter view, mind, instead of referring to an entity, refers to a
Internet

Ontology

- Genersereth and Nilsson:
  - Ontology is an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them.
- Ontology is to be shared and reusable
- Usually refer to abstract concepts and relationships (or properties)
- Rarely used for concept and relationship instances

Our Definition

- A set of concepts (C) and relationships (R) between the concepts

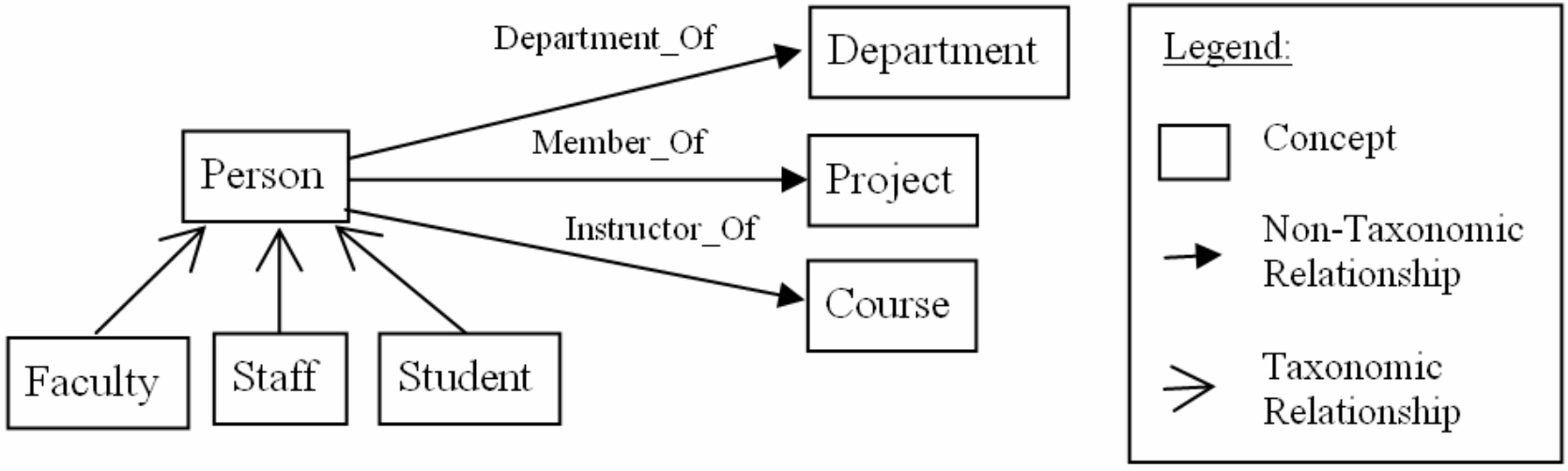


Fig. 1 An University Ontology Example

Ontology Research

- Ontology construction
  - Manual approach: OntoEdit
  - Automatic approach: OntoLearn
- Ontology representation languages
  - Traditional: CycL, Ontolingua, etc..
  - Web standards: XML, RDF
  - Web-based ontology specification languages: OIL, DAML+OIL, XOL, SHOE

Ontology-based Web (Content) Mining

- Types of web content mining
  - Web page classification
  - Web clustering
  - Web extraction

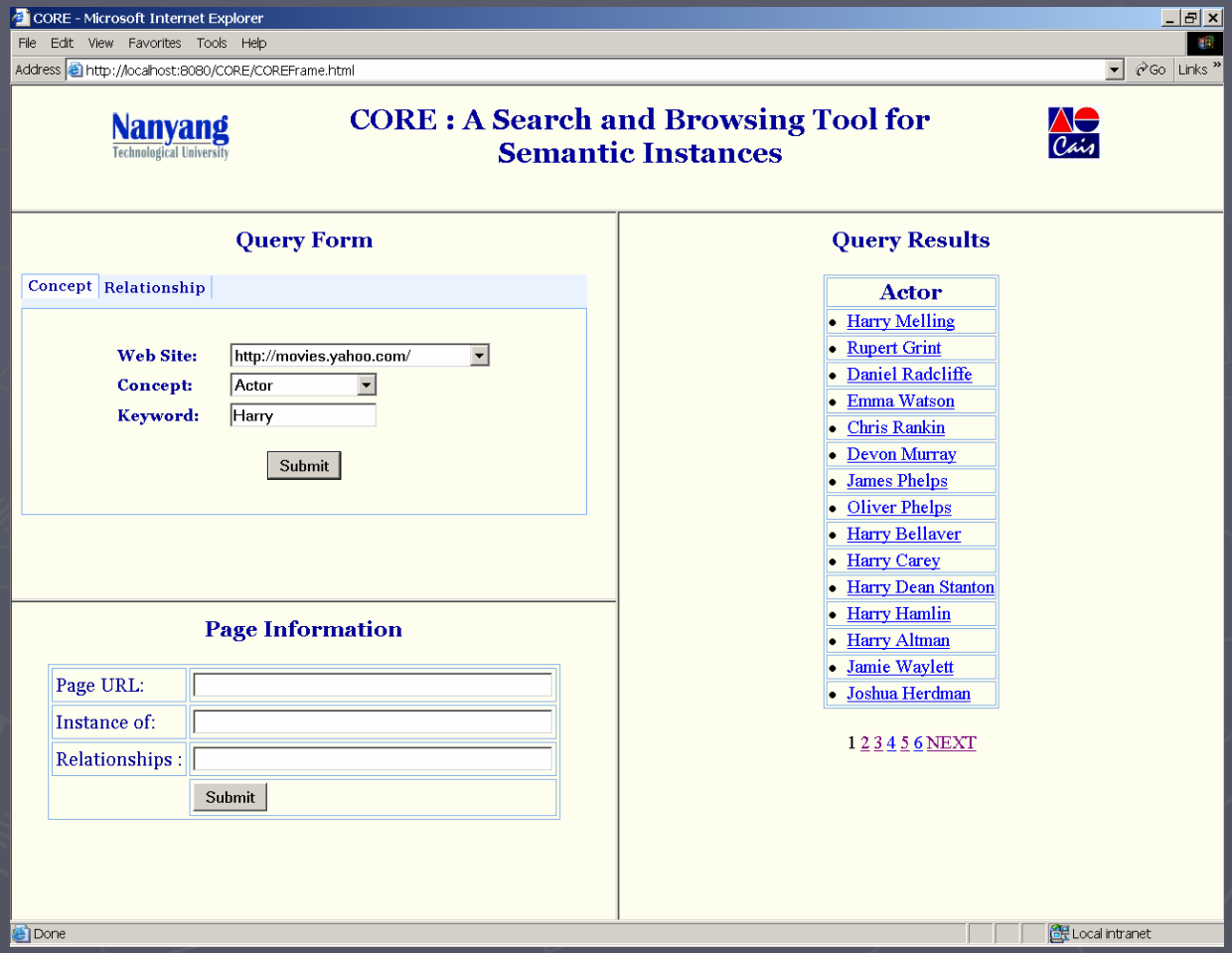
Web content mining + Ontology

- Known instances of ontology entities as additional features
  - Example: Ontology-based Web site structure mining
- Ontology provides background semantic structures for mining
  - Example: Ontology-based Web classification - classifying Web pages as concept instances and Web page pairs as relationship instances

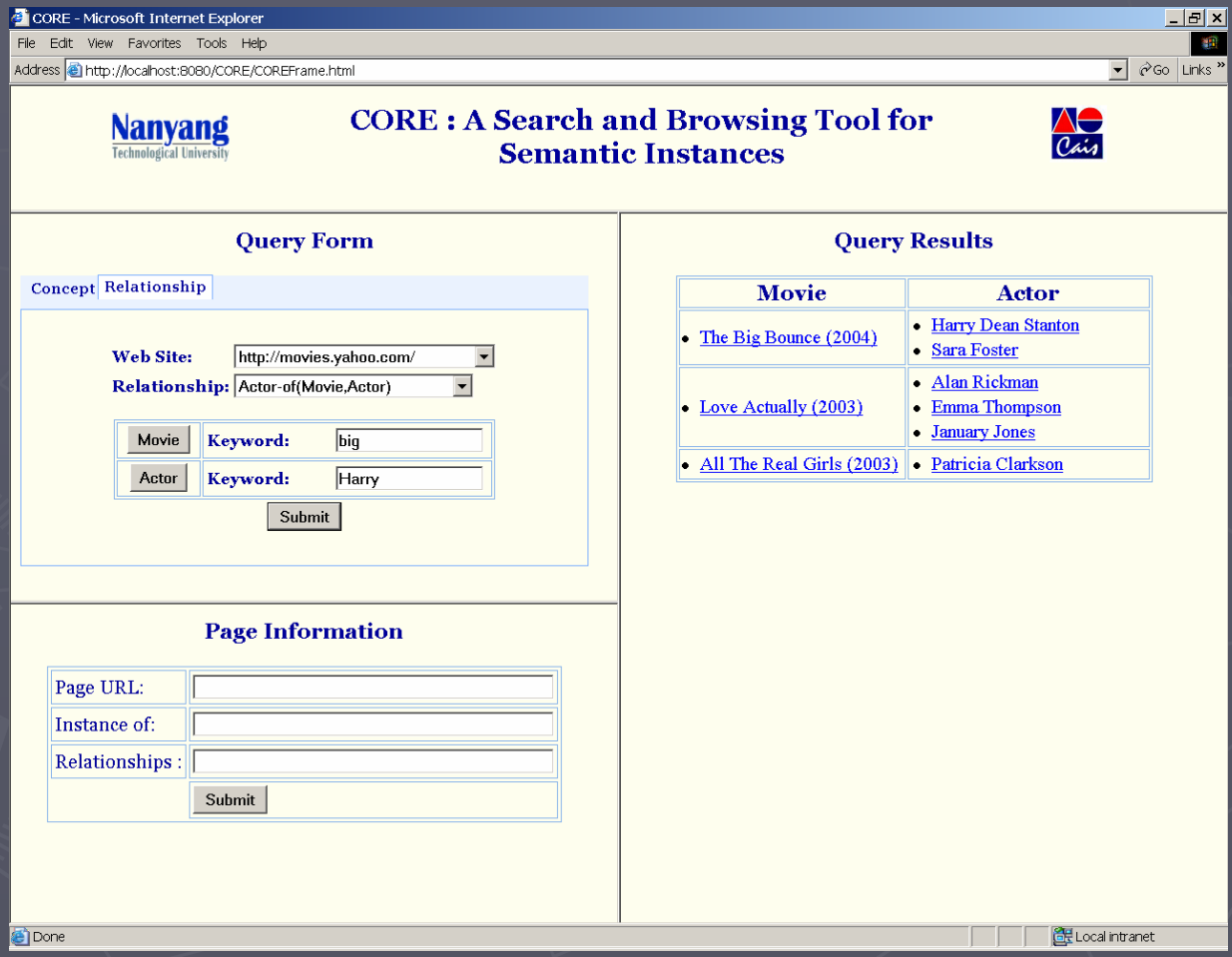
DL Applications of Ontology-based Web Mining

- Improved search to Web data
- Better browsing capabilities
- Personalization of Web data access

Ontology-based Concept Search

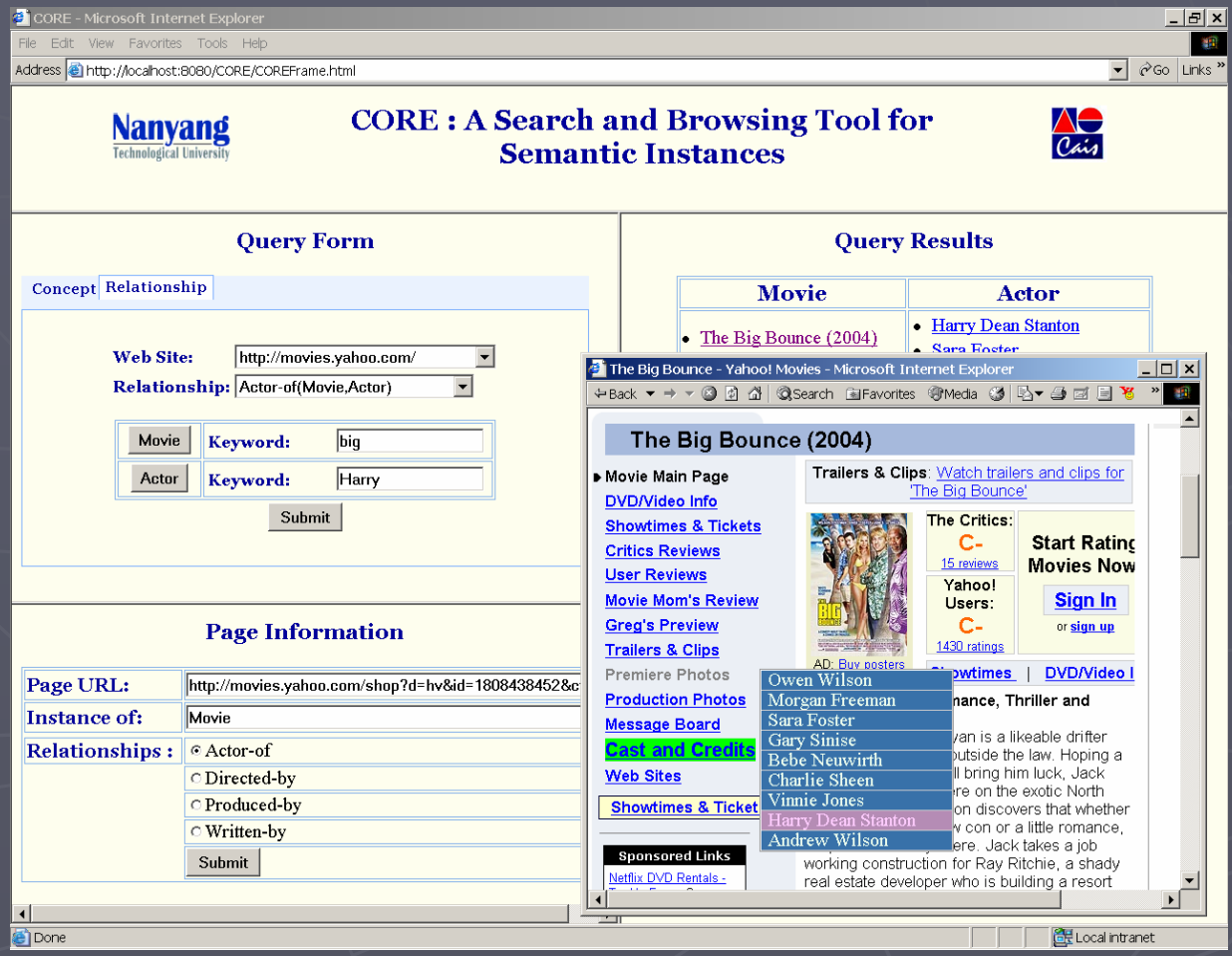


Ontology-based Relationship Search



Ontology-based Browsing

- Browsing Movie homepage

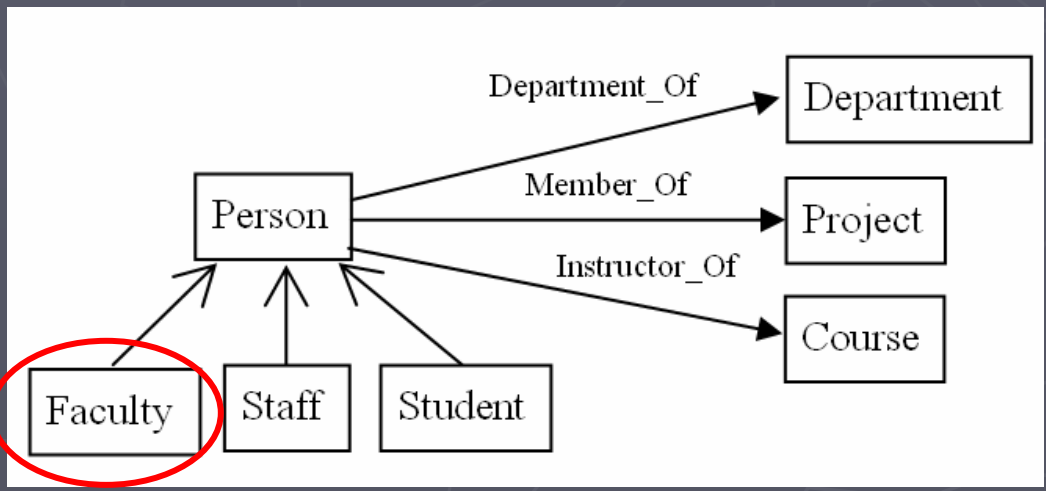


Our Ontology-based Web Content Mining Research

- Web page classification
- Homepage mining
- Homepage relationship mining
- Focus on web content from a given website

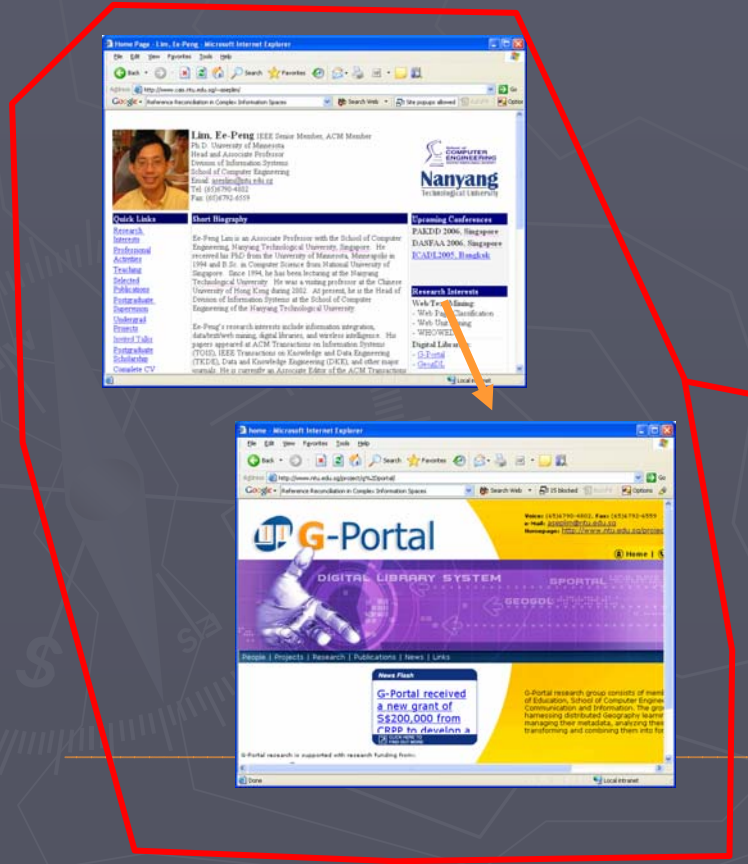
Homepage Mining

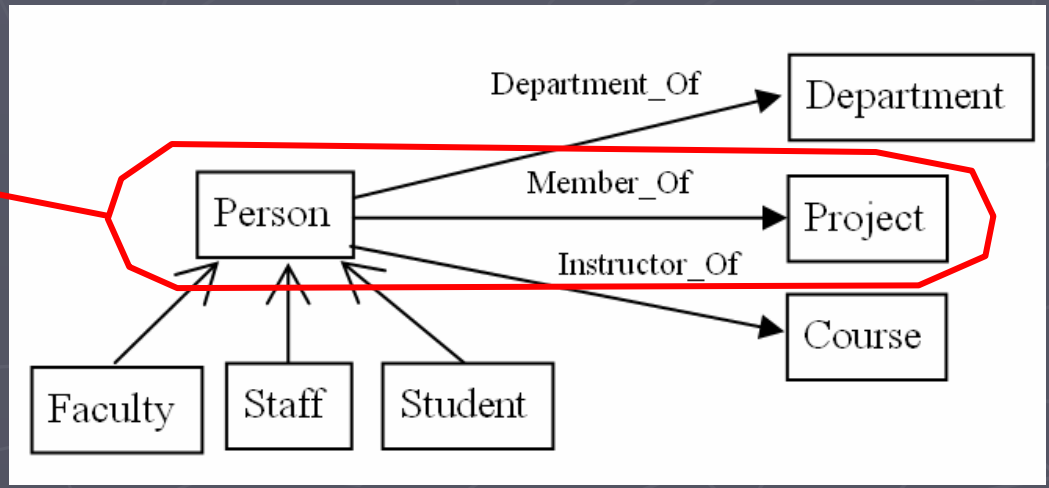
- Given an ontology consisting of concepts and a web site, find the homepages of concept instances



Homepage Relationship Mining

- Discovery of homepage pairs as related concept instances, or relationship instances





What are the technical challenges?

- Tasks
  - Find the homepages
  - Assign it with the appropriate concept label
  - Identify the relationships among the homepages
- Challenges
  - Definition of concept instance is subjective
  - Web sites organize Web pages in different ways
  - Features for identifying relationship instances are limited

Homepage Mining using Web Units

- Idea:
  A more complete concept instance = homepage + support pages
- Web unit:
  - Exactly one homepage
  - Zero or more support pages
- Web unit-based homepage mining:
  - Finding Web units representing concept instances

Web Unit

Web Unit of a CS100 course

http://..path/course/CS100/CS100.html
http://..path/course/CS100/lecture-programs.html
http://..path/course/CS100/officehours.html
http://..path/course/CS100/instructor.html
http://..path/course/CS100/exams/final.html
http://..path/course/CS100/exams/prelim.html

Web Unit of a Professor

http://..path/user/johnson/index.html
http://..path/user/johnson/research.html
http://..path/user/johnson/publications.html
http://..path/user/johnson/activities.html
http://..path/user/johnson/students.html
http://..path/user/johnson/teaching.html
http://..path/user/johnson/contact.html

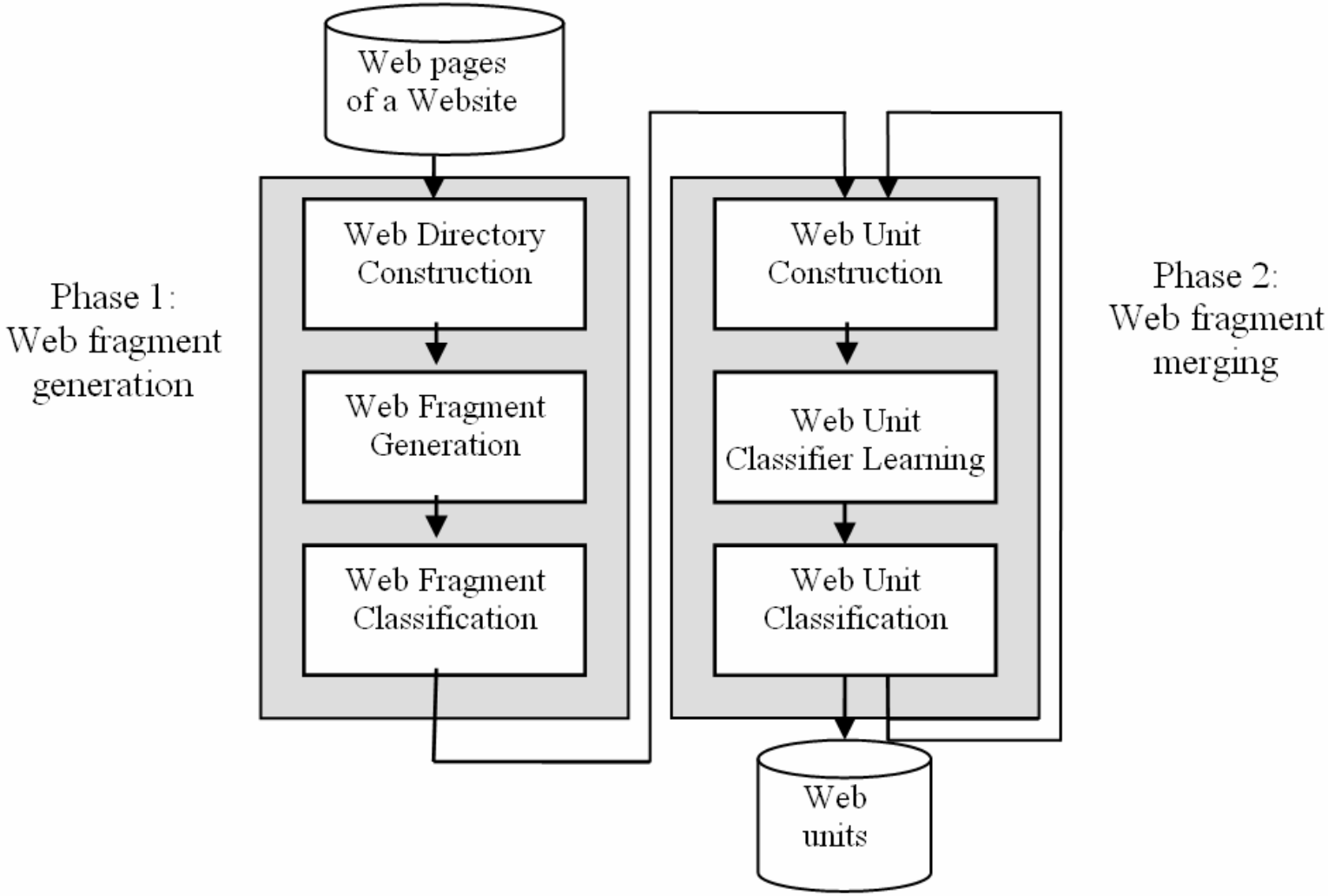
Web Unit-based Homepage Mining

- Two main tasks:
  - Find Web pages that form a Web unit and determine the role of each page
  - Assign concept labels to Web units
- Differences between web unit-based homepage mining and web page classification
  - Concept-relationship graph vs flat categories
  - Web units are not known beforehand

Iterative Web Unit Mining Method (iWUM)

1. Find homepages
2. Find some support pages for each homepage and construct initial set of web units (may be incomplete) – web fragments
3. Assign concept labels to web units
4. Construct larger web units
5. Reclassify web units
6. Repeat 4-5 until no or little changes to labels assigned

iWUM


Web pages of a Website
Phase 1: Web fragment generation
Web Directory Construction
Web Fragment Generation
Web Fragment Classification
Web Unit Construction
Web Unit Classifier Learning
Web Unit Classification
Phase 2: Web fragment merging
Web units

Observations on Web Units

- Observation 1
  - Web pages from the same Web folder are more semantically related
- Observation 2
  - Support pages are normally reachable from key page
- Observation 3
  - Key page is usually at the highest level of the Web folder containing the Web unit

Observations on Web Units

- Observation 4
  - Web units of same concept seldom have links between them
- Observation 5
  - Multi-page Web units of the same concept often reside in a set of folders (one for each) under a common parent folder
  - One-page Web units of the same concept often appear in the same folder
- Observation 6
  - Key page of the Web units of the same concept are often the link targets of a hub page

Web Fragment Generation

- Associate closely-related Web pages together
- Reduce the objects to be classified
- Reduce noise in training
- Steps:
  - Build a directory tree of folders and Web pages
  - Compute the *connectivity index* of each Web folder to measure the extent to which the Web pages and folders under the former are connected
  - Determine the candidate homepages in Web folder with small connectivity index values
    - Web page naming convention: common names for key pages are “index.html”, “index.htm”, etc..

Web Fragment Generation

- Find *Candidate Key Pages*
  - URL of the page ends with a "/"
  - The folder containing the page and the page share the same name, e.g., ...path/cs100/cs100.html
  - Page file name matches: *home, index, welcome, default, and homepage*

Web Fragment Generation and Classification

Example: Course CS100

1. http://..path/course/CS100/CS100.html [COURSE]
   http://..path/course/CS100/lecture-programs.html
   http://..path/course/CS100/officehours.html
   http://..path/course/CS100/instructor.html

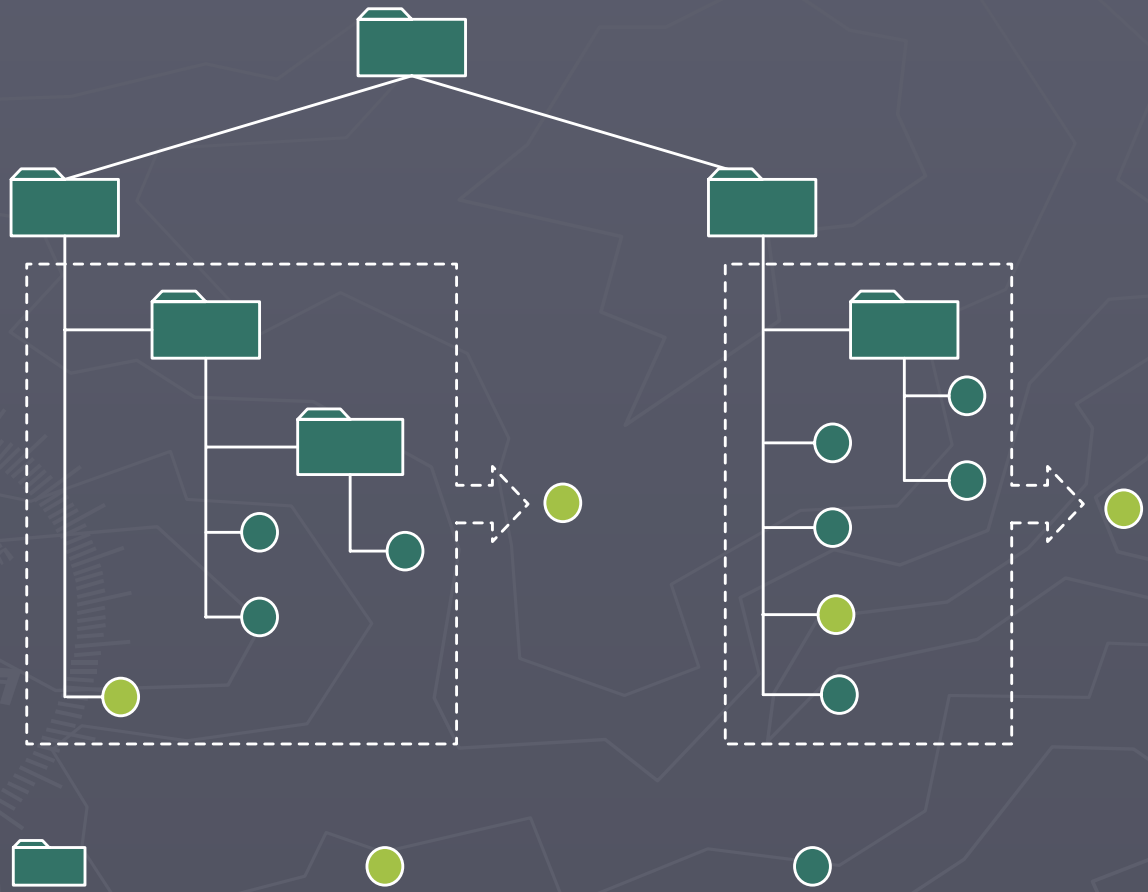
2. http://..path/course/CS100/exams/final.htm [NONE]

3. http://..path/course/CS100/exams/prelim.html [NONE]

Example: Prof Johnson

1. http://..path/user/johnson/index.html [PROF]
   http://..path/user/johnson/research.html
   http://..path/user/johnson/publications.html
   http://..path/user/johnson/activities.html
   http://..path/user/johnson/students.html
   http://..path/user/johnson/teaching.html
   http://..path/user/johnson/contact.html

Web Unit Construction



Web Unit Construction

1. http://..path/course/CS100/CS100.html [COURSE]
   http://..path/course/CS100/lecture-programs.html
   http://..path/course/CS100/officehours.html
   http://..path/course/CS100/instructor.html

2. http://..path/course/CS100/exams/final.htm [NONE]

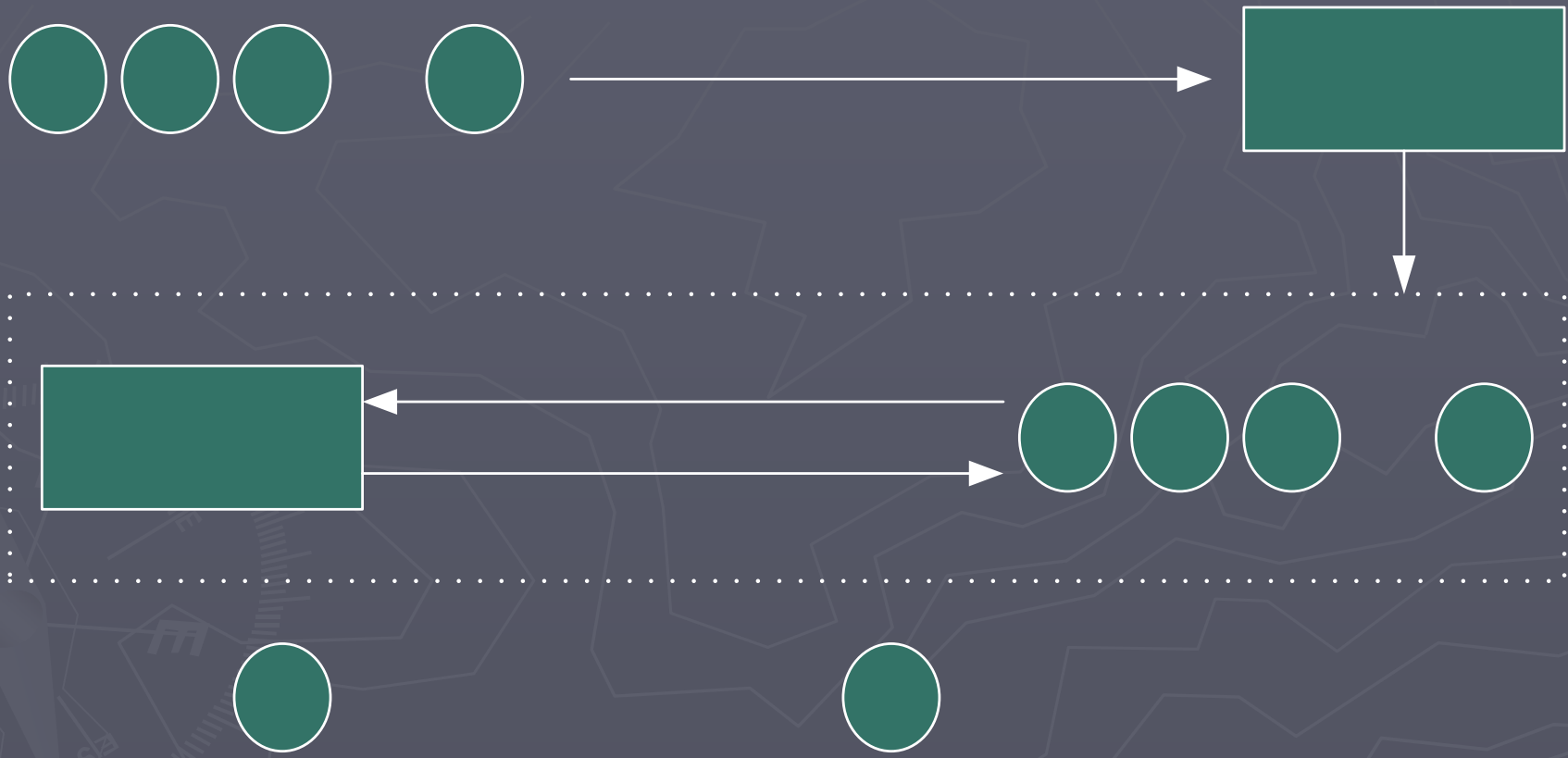
3. http://..path/course/CS100/exams/prelim.html [NONE]

1. http://..path/course/CS100/CS100.html [COURSE]
   http://..path/course/CS100/lecture-programs.html
   http://..path/course/CS100/officehours.html
   http://..path/course/CS100/instructor.html
   http://..path/course/CS100/exams/final.htm
   http://..path/course/CS100/exams/prelim.html

Web Unit Classification

- Observations 5 and 6:
  - Multi-page Web units of the same concept often reside in a set of folders (one for each) under a common parent folder
  - Key pages of the Web units of the same concept are often the link targets of a hub page
- Improve Web unit mining accuracy
  - Web site structure features
  - Content features

Web Unit Classification



Web Site Structure Features

- Normalized classification score (each web unit) for each concept
- Organization of the web units within the web site
  - Closeness to the average depth for each concept
  - Highest in-link hub value for each concept
  - Precision support of the parent web folder for each concept
  - Recall support of the parent web folder for each concept
- Word features in the web page names and URLs
  - Each word (term) in page names and URL

Performance of iWUM

- Performance is measured by
  - Are the web units correctly constructed?
  - Are the web units correctly classified?
- Implication of homepage and support pages
- iWUM performs well on the WebKB dataset
  - 4 university websites: Cornell, Texas, Washington and Wisconsin
  - 4 concepts: Student, Course, Faculty, Project
- iWUM works better for more structured websites

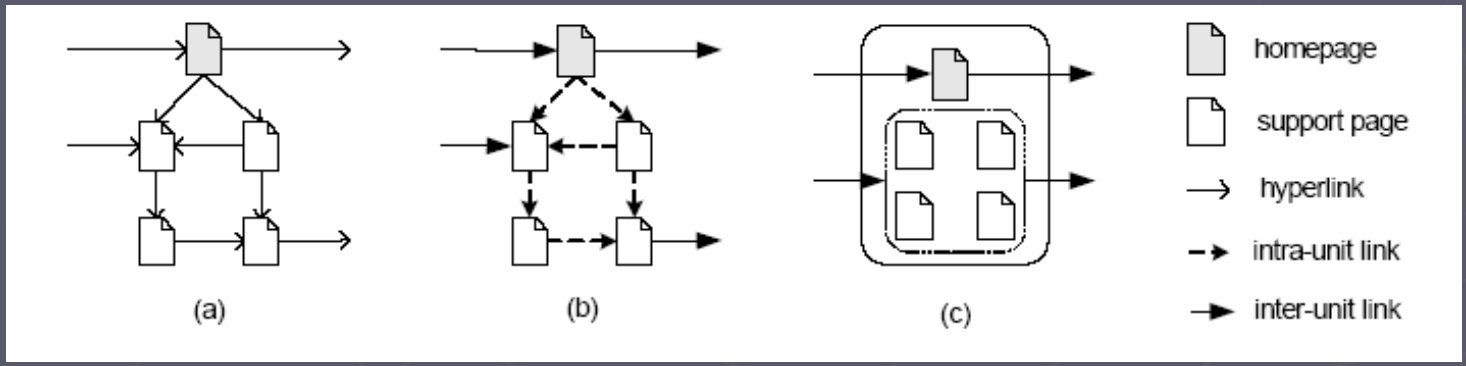
Web Unit-based Homepage Relationship Mining

- Assumptions
  - Web units are known by web unit-based homepage mining
  - Relationships can be determined based on background relation knowledge
  - Background relation are represented by inter-homepage features
- Our proposed method
  1. Candidate homepage pair generation
  2. Feature acquisition
  3. Classifier training
  4. Classification

Inter-Homepage Features

- Navigation Features (N)– links between web pages
  - Intra-unit links
  - Inter-unit links
- Relative Location Features (R) – location in web directory
  - Parent-child
  - Sibling
  - Ancestor-descendent
- Common-item Features (E) – shared by homepages
  - Email addresses
- Supplementary features (A) – additional features derived for some inter-homepage features

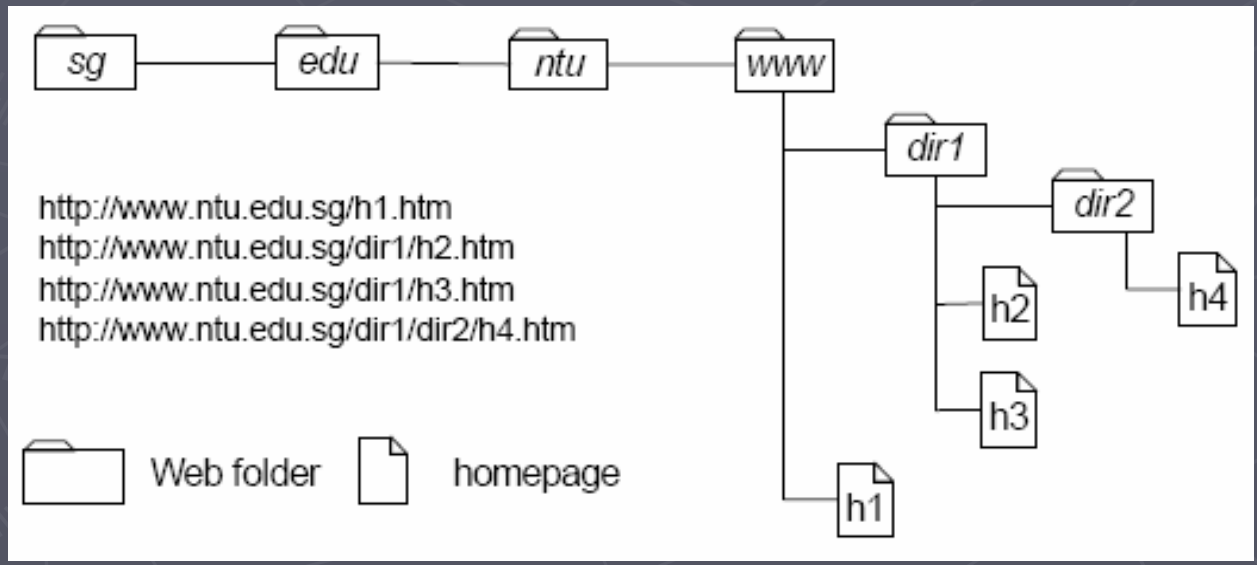
Navigation Features (N)



id	connectivity type	id	connectivity type	id	connectivity type
n_1	$u_s.h \rightarrow u_t.h$	n_9	$u_s.h \rightarrow p \rightarrow u_t.h$	n_{17}	$u_s.h \rightarrow p \leftarrow u_t.h$
n_2	$u_s.h \rightarrow u_t.s$	n_{10}	$u_s.h \rightarrow p \rightarrow u_t.s$	n_{18}	$u_s.h \rightarrow p \leftarrow u_t.s$
n_3	$u_s.s \rightarrow u_t.h$	n_{11}	$u_s.s \rightarrow p \rightarrow u_t.h$	n_{19}	$u_s.s \rightarrow p \leftarrow u_t.h$
n_4	$u_s.s \rightarrow u_t.s$	n_{12}	$u_s.s \rightarrow p \rightarrow u_t.s$	n_{20}	$u_s.s \rightarrow p \leftarrow u_t.s$
n_5	$u_s.h \leftarrow u_t.h$	n_{13}	$u_s.h \leftarrow p \leftarrow u_t.h$	n_{21}	$u_s.h \leftarrow p \rightarrow u_t.h$
n_6	$u_s.h \leftarrow u_t.s$	n_{14}	$u_s.h \leftarrow p \leftarrow u_t.s$	n_{22}	$u_s.h \leftarrow p \rightarrow u_t.s$
n_7	$u_s.s \leftarrow u_t.h$	n_{15}	$u_s.s \leftarrow p \leftarrow u_t.h$	n_{23}	$u_s.s \leftarrow p \rightarrow u_t.h$
n_8	$u_s.s \leftarrow u_t.s$	n_{16}	$u_s.s \leftarrow p \leftarrow u_t.s$	n_{24}	$u_s.s \leftarrow p \rightarrow u_t.s$

Relative Location Features (R)

- Parent-child: h2 and h4
- Sibling: h2 and h3
- Ancestor-descendent: h1 and h4



Experimental Dataset

- WebKB
  - Department-of (people, department)
  - Instructor-of (people, course)
  - Member-of (people, project)

University	Department-of		Instructor-of		Member-of	
	Pos	Neg	Pos	Neg	Pos	Neg
Cornell	183	0	32	7654	66	3594
Texas	197	0	42	7444	89	3851
Washington	161	6	65	12294	135	3372
Wisconsin	207	3	112	17108	102	5148

Experimental Results

- On the manually labelled web units

Features	Department-of			Instructor-of			Member-of		
	Pr	*Re*	*F1*	*Pr*	*Re*	*F1*	*Pr*	*Re*	*F1*
N	0.988	0.796	0.875	0.879	0.651	0.724	0.879	0.884	0.881
NR	0.987	1.000	0.994	0.877	0.673	0.737	0.879	0.884	0.881
NE	0.988	0.797	0.876	0.884	0.695	0.759	0.879	0.890	0.883
NRE	0.987	1.000	0.994	0.864	0.698	0.750	0.879	0.890	0.883

Experimental Results

- On the iWUM mined web units

	department-of			instructor-of			member-of		
University	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
Cornell	0.986	0.770	0.865	0.800	0.250	0.381	0.323	0.303	0.312
Texas	0.989	0.893	0.939	0.731	0.452	0.559	0.000	0.000	0.000
Washington	0.863	0.863	0.863	0.737	0.438	0.549	0.000	0.000	0.000
Wisconsin	0.968	0.884	0.924	0.812	0.500	0.619	0.477	0.618	0.538
MacroAve	0.952	0.853	0.898	0.770	0.410	0.527	0.200	0.230	0.213

Conclusion

- Ontology can be used to add semantics to web content
- We introduce two ontology-based web content mining problems
  - Homepage mining
  - Homepage relationship mining
  - Web Unit to model a concept instance

Future Research Opportunities

- Ontology can be incorporated in other web mining techniques
- Digital libraries can benefits from the additional semantics about web content
- Future research
  - Web unit-based searching
  - Link analysis among web units
  - Evolution of web units

Relevant Publications

- Yin Ming, Dion Hoe-Lian Goh, Ee-Peng Lim, "On Discovering Concept Entities from Web Sites," International Journal of Web Information Systems (IJWIS), accepted, 2005.
- Myo-Myo Naing, Ee-Peng Lim, Roger H.L. Chiang, "Extracting Link Chains of Relationship Instances from a Web Site," American Society for Information Science and Technology (JASIST), accepted, 2005.
- Aixin Sun and Ee-Peng Lim, "Web Unit Based Mining of Homepage Relationships," JASIST, accepted, 2005.
- A. Sun, E.-P. Lim, W.-K. Ng, J. Srivastava "Blocking Reduction Strategies in Hierarchical Text Classification," IEEE TKDE 16(10):1305-1308 , 2004.
- Myo Myo Naing, Ee-Peng Lim, Roger Chiang. "CORE: A Search and Browsing Tool for Semantic Instances of Web Sites," 7th Asia Pacific Web Conference (APWeb2005), Shanghai China, March 2005.
- A. Sun, E.-P. Lim, "Web Unit Mining: Finding and Classifying Subgraphs of Web Pages," ACM CIKM, 2003.
- A. Sun, E.-P. Lim, and W.-K. Ng, Performance Measurement Framework for Hierarchical Text Classification, JASIST 54(11):1014 – 1028, 2003.
- A. Sun, E.-P. Lim, "Web Classification Using Support Vector Machine," ACM WIDM 2002.

Thank You