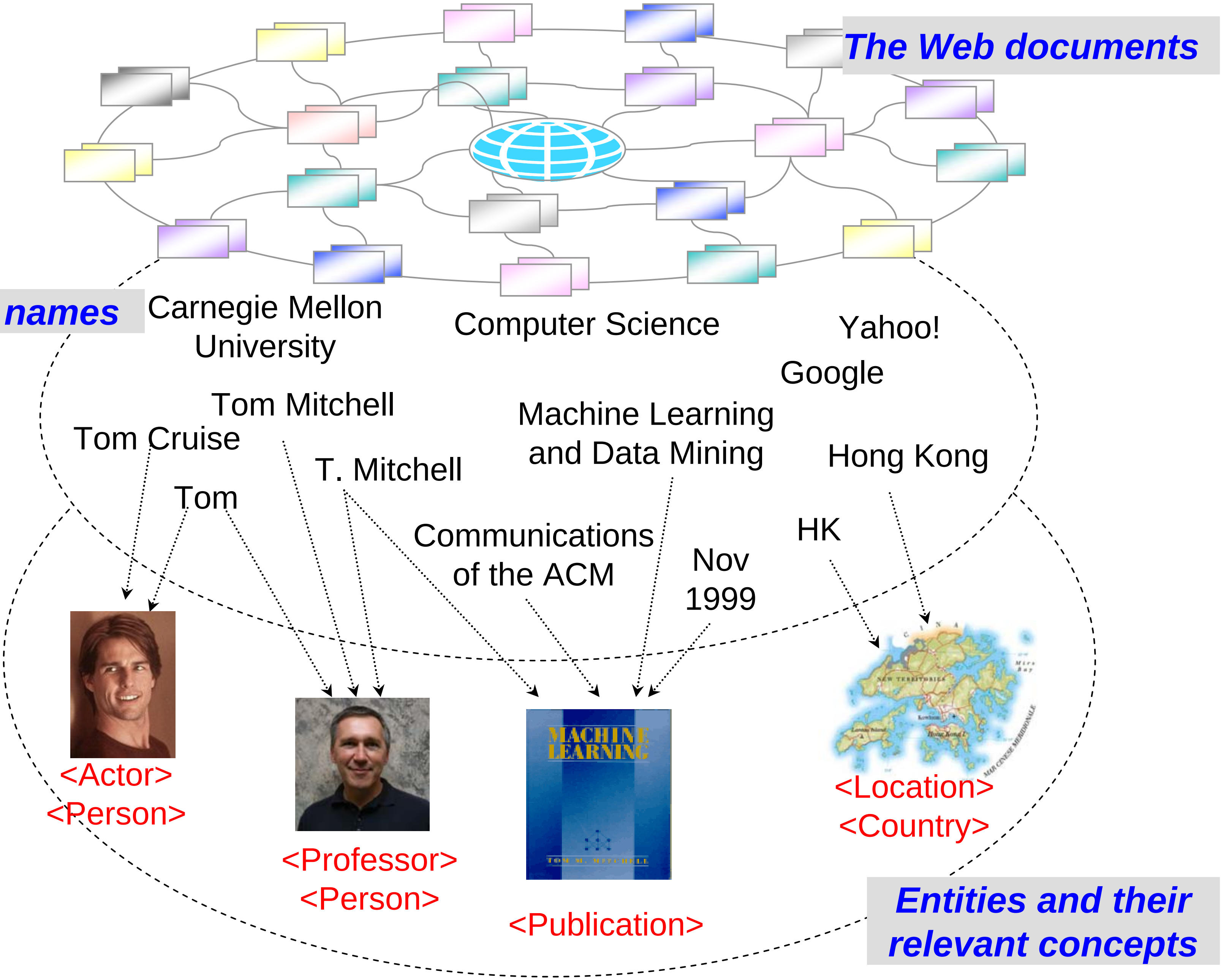
Extracting Information from Web Documents based on Conceptual Entity Tree Correspondence

Introduction

- WWW is the largest and richest information repository available today
- The distributed and decentralized nature cause the web to grow enormously
- However, the nature of the web create problems for user – difficult to find the right information or answer
- The aim of this work – extract and represent conceptual entities from the web, to enhance the retrieval of more specific and precise information

Named Entities

- Named-entity (NE) is a word or word sequences that denotes a particular individual or instance in the real world (e.g. Tom Mitchell, Google)
- NE signal prominent piece of information in web documents
- NE usually appear in many alias forms and a couple of NE may reflect a single instance



The Web documents

names

Carnegie Mellon University

Computer Science

Yahoo!

Google

Tom Mitchell

Tom Cruise

T. Mitchell

Machine Learning and Data Mining

Hong Kong

Tom

Communications of the ACM

Nov 1999

HK

<Actor>
<Person>

<Professor>
<Person>

<Publication>

<Location>
<Country>

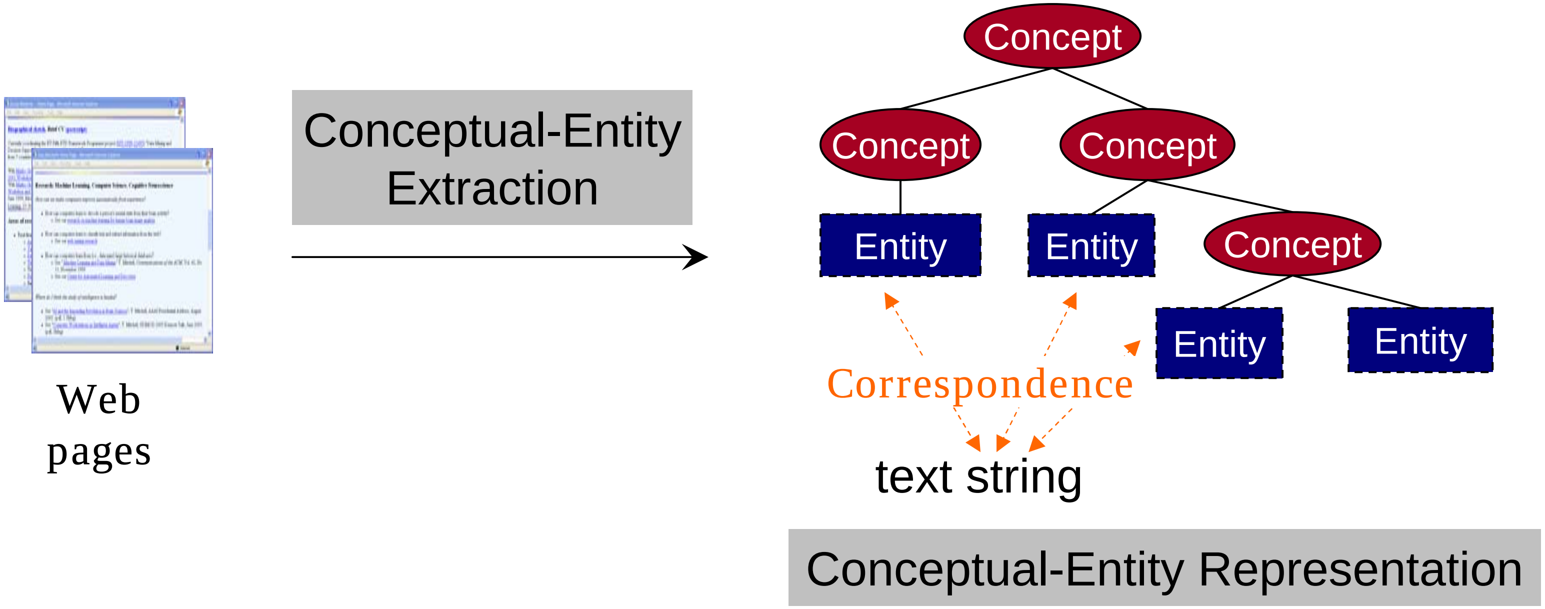
Entities and their relevant concepts

Concept-based Entities

- Traditionally, the recognition of NE is limited to a small set of broader, predefined categories (e.g. PERSON, LOCATION, ORGANIZATION, DATE, etc)
- This become a limitation in information seeking context – especially when user request for a more concise piece of information
- The categories of interest should be more diverse, refined and concept-based

Proposed Approach

- extract named entities and their concepts from web documents, and
- represent in a simple and flexible annotation structure → Conceptual Entity Tree Correspondence



Web pages

Conceptual-Entity Extraction

Concept

Concept

Concept

Entity

Entity

Concept

Entity

Entity

Correspondence

text string

Conceptual-Entity Representation

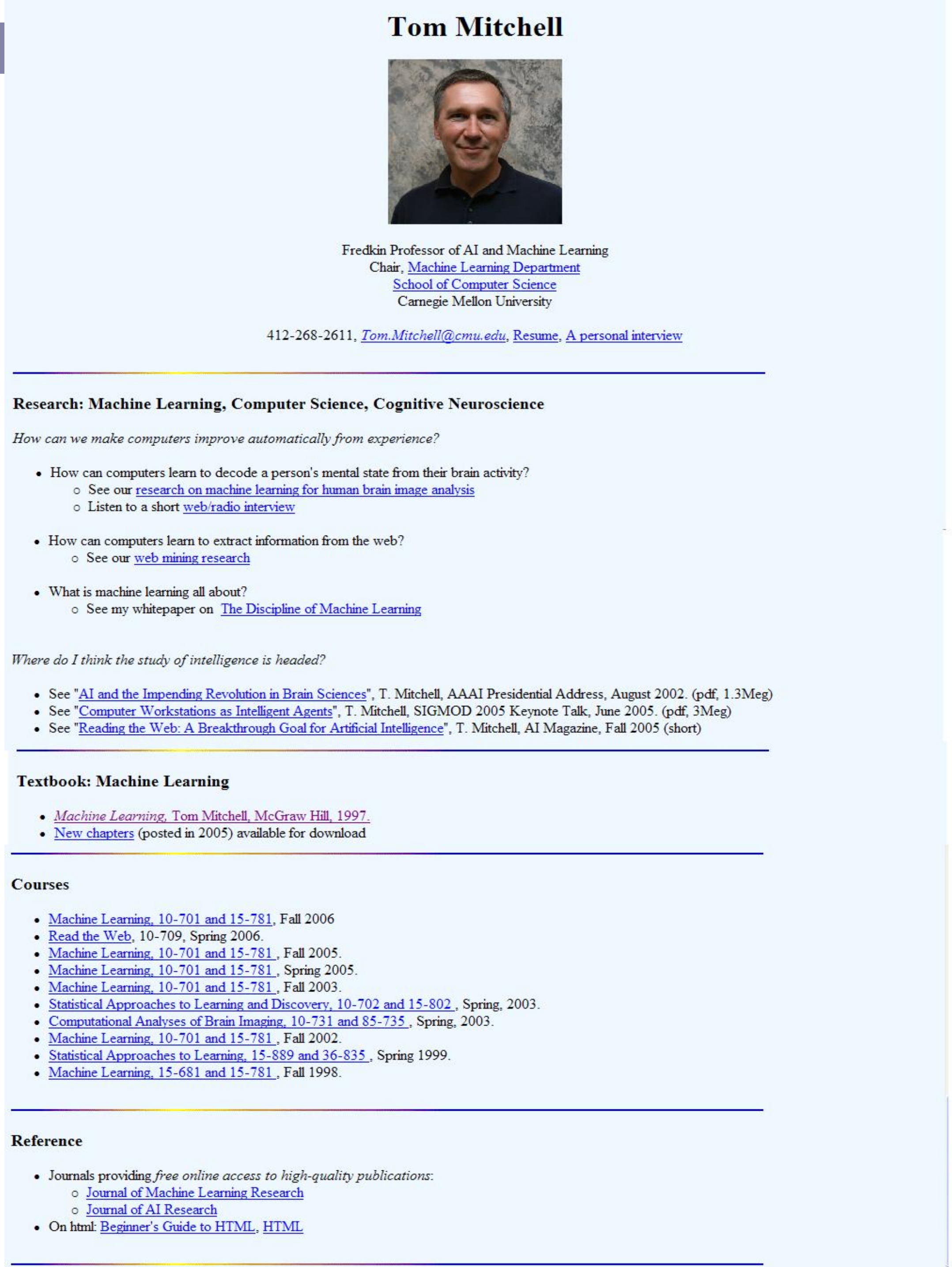
Conceptual Entities Extraction – Parsing Html Structure

- consists of 3 main steps:
- Parsing html structure
  - In web pages, the ***structure*** and ***visual clues*** are important features to facilitate the extraction of information
  - web pages are designed for human to read, they will follow some widely accepted rules to enable us to easily read and understand the content
    - hierarchical structure of headings and labels
    - contents are in short information segments represented in list and table
  - parse web pages into html structure tree

Conceptual Entities Extraction – Recognizing Entities & Concepts

- Recognizing entities
  - In English, capitalization gives good evidence of named entities
  - identify entities by finding continuous capitalized words including lower-case functional words
- Deriving concepts
  - every level in html structure tree corresponds to a different granularity of information
  - derive concepts that describe named entities by analyzing the tree

An Example:



Tom Mitchell

Fredkin Professor of AI and Machine Learning
Chair, Machine Learning Department
School of Computer Science
Carnegie Mellon University

412-268-2611, *Tom.Mitchell@cmu.edu*, Resume, A personal interview

Research: Machine Learning, Computer Science, Cognitive Neuroscience

How can we make computers improve automatically from experience?

- How can computers learn to decode a person's mental state from their brain activity?
  - See our research on machine learning for human brain image analysis
  - Listen to a short web/radio interview
- How can computers learn to extract information from the web?
  - See our web mining research
- What is machine learning all about?
  - See my whitepaper on The Discipline of Machine Learning

Where do I think the study of intelligence is headed?

- See "AI and the Impending Revolution in Brain Sciences", T. Mitchell, AAAI Presidential Address, August 2002. (pdf, 1.3Meg)
- See "Computer Workstations as Intelligent Agents", T. Mitchell, SIGMOD 2005 Keynote Talk, June 2005. (pdf, 3Meg)
- See "Reading the Web: A Breakthrough Goal for Artificial Intelligence", T. Mitchell, AI Magazine, Fall 2005 (short)

Textbook: Machine Learning

- *Machine Learning*, Tom Mitchell, McGraw Hill, 1997.
- New chapters (posted in 2005) available for download

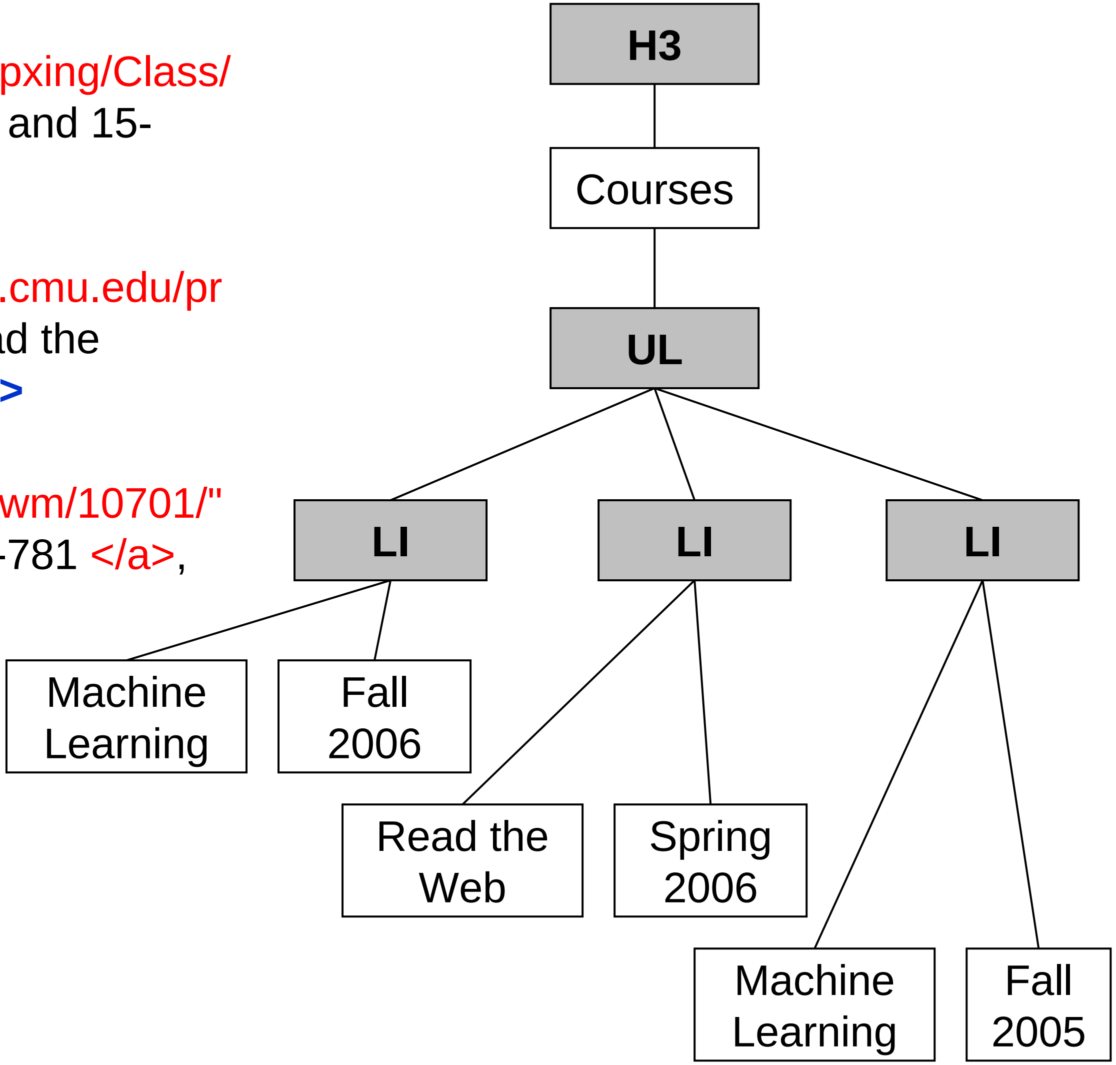
Courses

- Machine Learning, 10-701 and 15-781, Fall 2006
- Read the Web, 10-709, Spring 2006.
- Machine Learning, 10-701 and 15-781 , Fall 2005.
- Machine Learning, 10-701 and 15-781 , Spring 2005.
- Machine Learning, 10-701 and 15-781 , Fall 2003.
- Statistical Approaches to Learning and Discovery, 10-702 and 15-802 , Spring, 2003.
- Computational Analyses of Brain Imaging, 10-731 and 85-735 , Spring, 2003.
- Machine Learning, 10-701 and 15-781 , Fall 2002.
- Statistical Approaches to Learning, 15-889 and 36-835 , Spring 1999.
- Machine Learning, 15-681 and 15-781 , Fall 1998.

Reference

- Journals providing *free online access to high-quality publications*:
  - Journal of Machine Learning Research
  - Journal of AI Research
- On html: Beginner's Guide to HTML, HTML

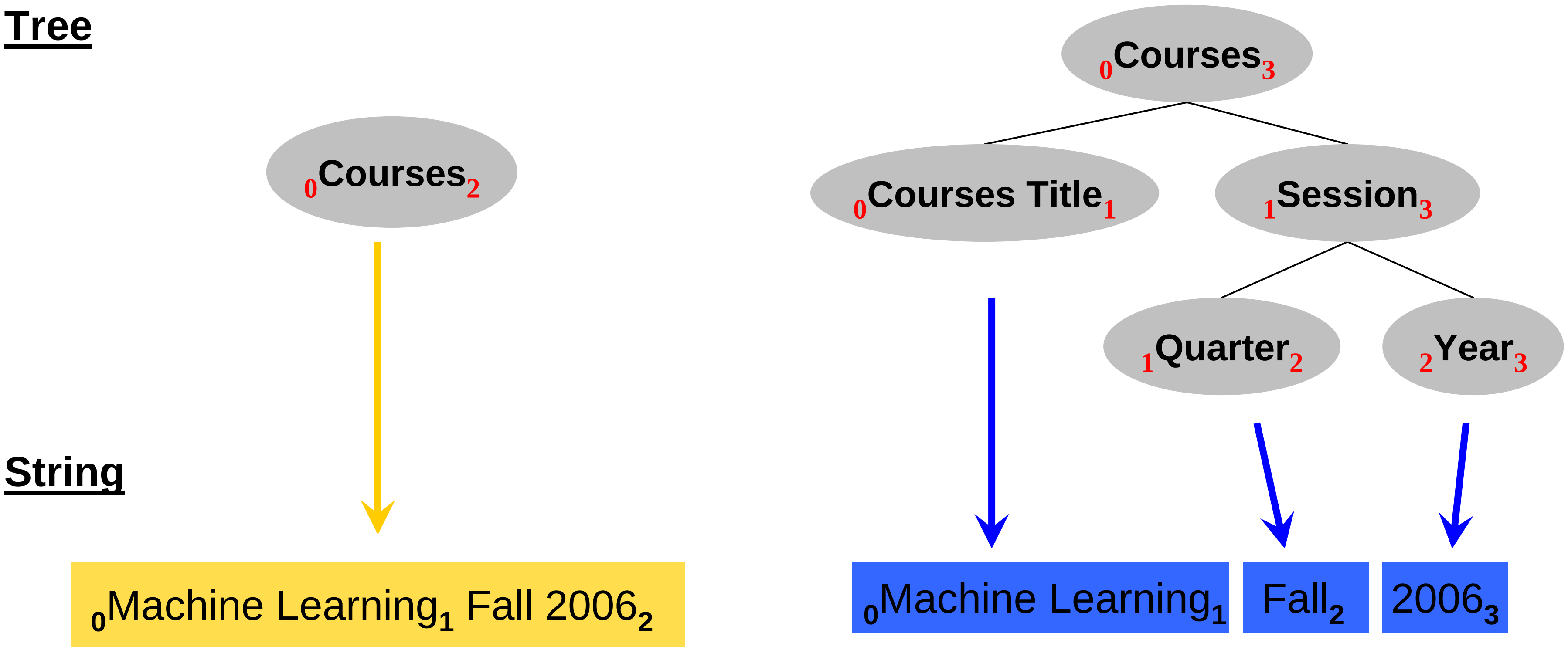
```
<h3><a name="courses">Courses</a></h3>
<ul>
   <li><a
     href="http://www.cs.cmu.edu/%7Eepxing/Class/
     10701/">Machine Learning, 10-701 and 15-
     781</a>, Fall 2006</li>
   <li><a
     href="http://www.cs.cmu.edu/afs/cs.cmu.edu/pr
     oject/theo-21/www/index.html">Read the
     Web</a>, 10-709, Spring 2006. </li>
   <li><a
     href="http://www.cs.cmu.edu/%7Eawm/10701/"
     >Machine Learning, 10-701 and 15-781 </a>,
     Fall 2005. </li>
   ….
</ul>
```



Conceptual Entities Representation

- use conceptual entity tree correspondence to capture the conceptual entity, its tree representation and the mapping (correspondence) between these two
- the correspondence is encoded on the representation tree by attaching to each concept node an interval of the entity in the string
- we do not define what are “primitive” concepts, thus the correspondence can be applied at any level of granularity

Conceptual Entity Tree Correspondence



Tree

${}_0$Courses${}_3$

${}_0$Courses${}_2$

${}_0$Courses Title${}_1$

${}_1$Session${}_3$

${}_1$Quarter${}_2$

${}_2$Year${}_3$

String

${}_0$Machine Learning${}_1$ Fall 2006${}_2$

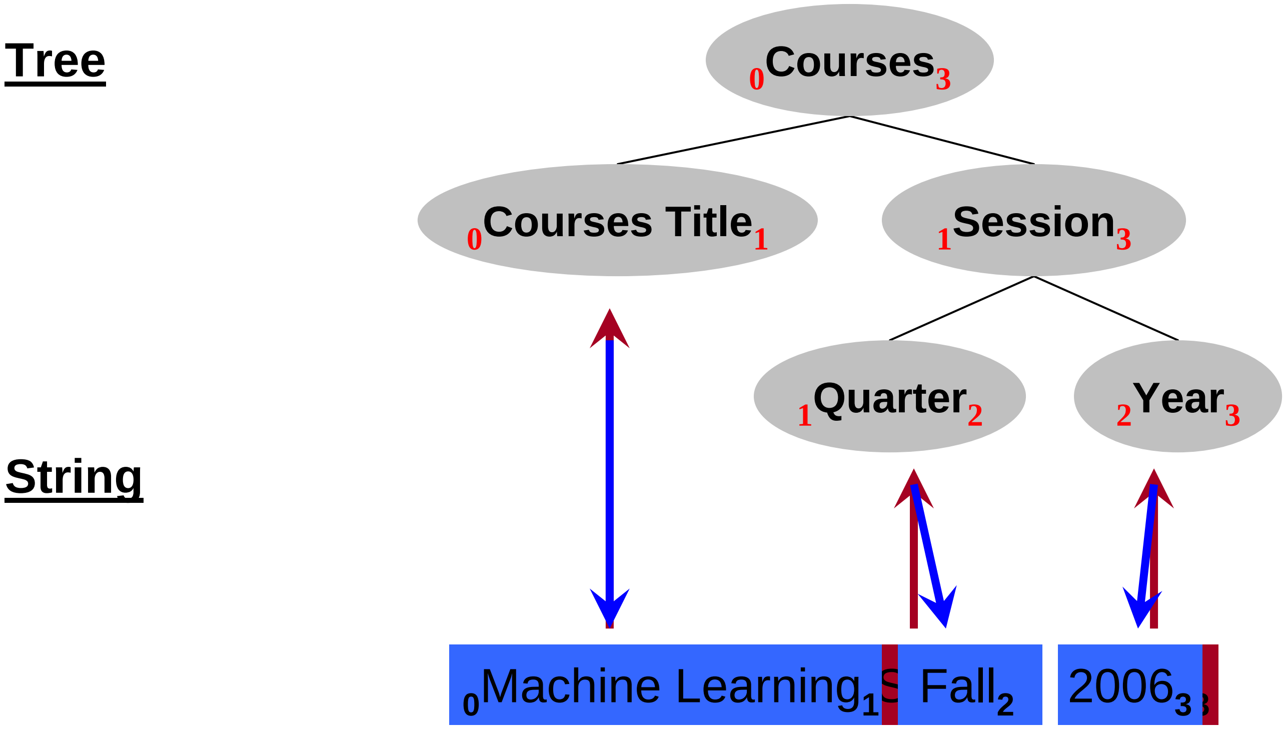
${}_0$Machine Learning${}_1$

Fall${}_2$

2006${}_3$

Application to Information Extraction – Example-based Learning

- Learn new conceptual entities based on the correspondence



Tree

${}_0$Courses$_3$

${}_0$Courses Title$_1$

${}_1$Session$_3$

${}_1$Quarter$_2$

${}_2$Year$_3$

String

${}_0$Machine Learning$_1$S Fall$_2$

2006$_3$3

Application to Information Extraction – Information Retrieval

- Conceptual entity tree correspondence model can be used to annotate a web document by enriching the texts with concepts
- enable information retrieval system to return more precise answer

```
<Courses>
   <Courses Title> Machine Learning
   <Session> Fall 2006

<Courses>
   <Courses Title> Read the Web
   <Session> Spring 2006
```

References

- C. Boitet, and Y. Zaharin, “Representation Trees and String-Tree Correspondences”, in *Proceedings of the 12th International Conference on Computational Linguistics* (*COLINGS 1988*), Budapest, Hungary, August 1988, pp. 59 – 64
- D. DiPasquo, “*Using HTML Formatting to Aid in Natural Language Processing on the World Wide Web*”, Senior Honors Thesis, 1998
- M. Pasca, “Acquisition of Categorized Named Entities for Web Search”, in *Proceedings of the 13th ACM International Conference on Information and Knowledge Management* (*CIKM 2004*), ACM Press, Washington, D.C., USA, 8 - 13 November 2004, pp 137 – 145.
- P.J. Cheng, H.C. Chiao, Y.C. Pan, and L.F. Chien, “Annotating Text Segments in Documents for Search”, in *Proceedings of the 2005 IEEE/WIC/ACM Conference on Web Intelligence* (*WI 2005*), IEEE Computer Society Press, Compiegne University of Technology, France, 19 – 22 September 2005, pp. 317 – 320.

Thank You