

The Cornell University logo, featuring the word "CORNELL" in white, serif, all-caps font on a red square background.

CORNELL



Finding the Boundaries of Information Resources on the Web

Pavel Dmitriev

Cornell University

Joint work with Carl Lagoze, Boris Suchkov

Problem

- Given a Web site, cluster the pages on it so that every cluster represents a **coherent document**

Compound Documents

- Authors do not think in terms of physical Web pages
- We want to recover their original intention
- Examples:
 - A news article consisting of several html pages
 - An entry in online encyclopedia

Compound Documents

But...

- Users have different goals, not necessarily similar to the author's

Problem

- No single “right” way to define what a resource is

Problem

- No single “right” way to define what a resource is

But

- A particular user always knows what they want!

Solution

- Use Machine Learning to define a compound document according to the needs of a particular user or class of users!

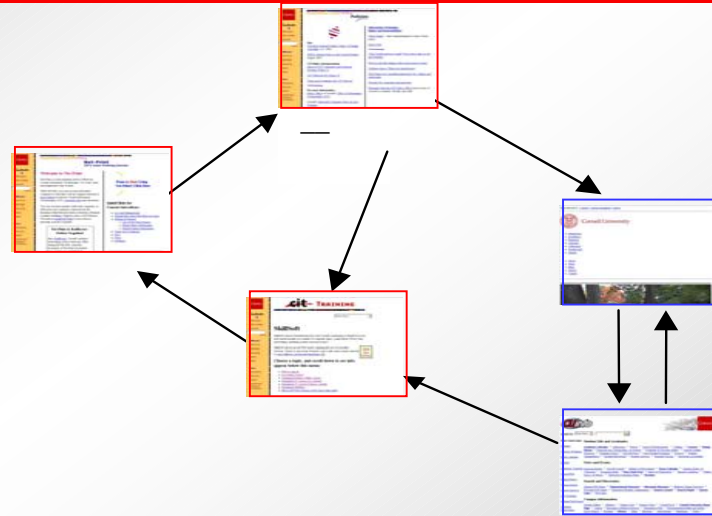
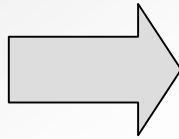
Applications

- Information Retrieval
 - Search
 - PageRank, etc.
- Collection generation & automatic metadata extraction
- Information Extraction
- Navigation

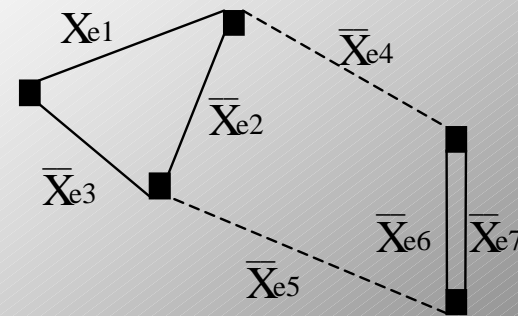
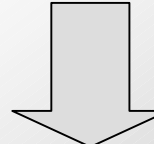
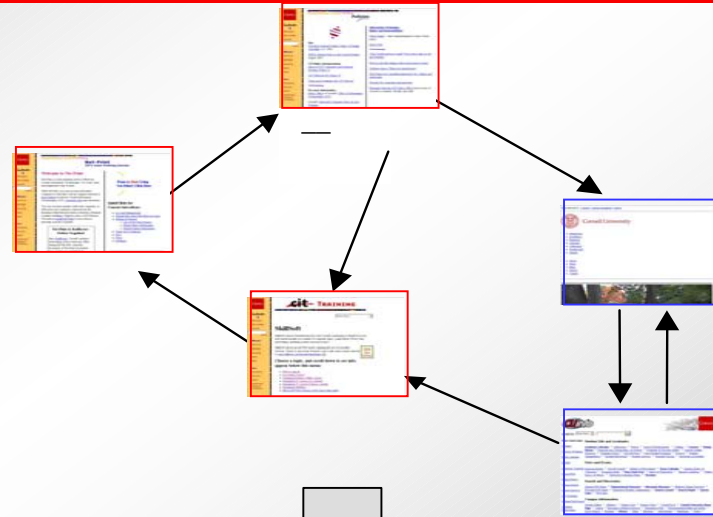
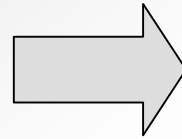
General Framework

- Look at one web site at a time
- Two stages:
 - Learn a profile of a compound document (cDoc) from examples
 - Use the profile to identify cDocs on new web sites

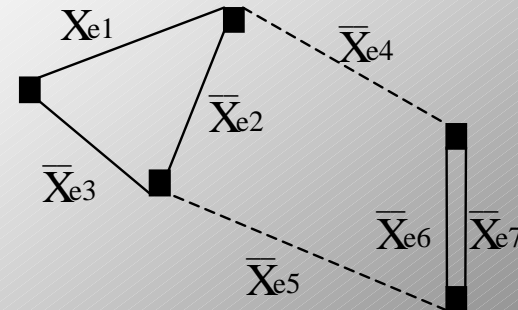
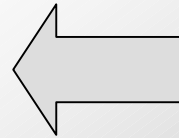
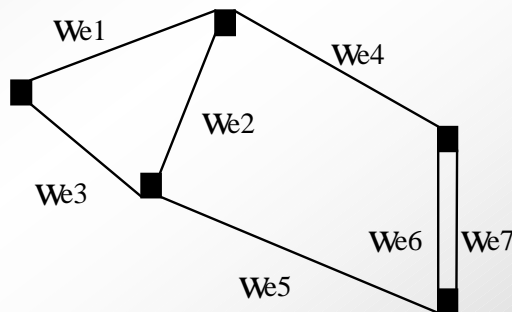
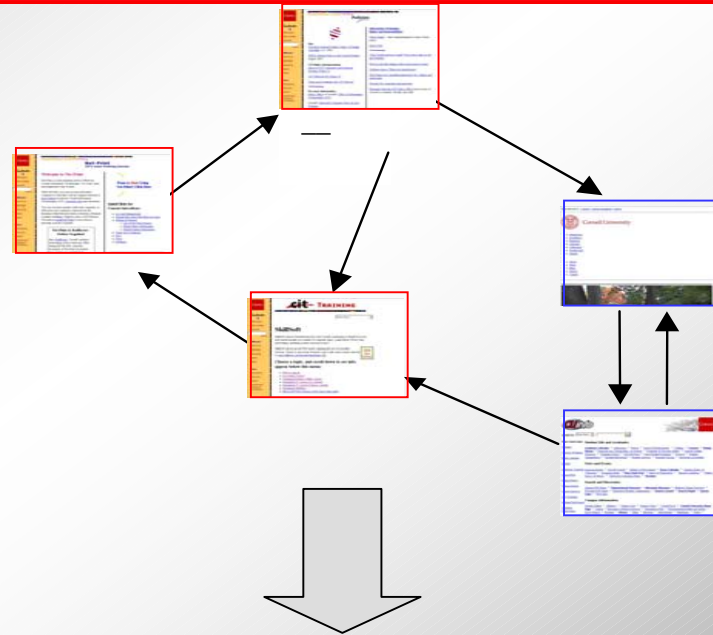
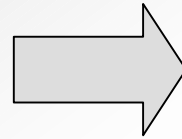
Learning cDoc Profile (I)



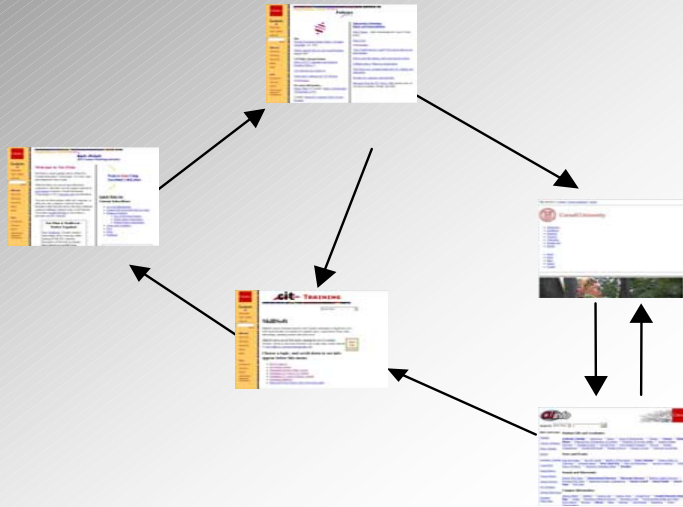
Learning cDoc Profile (II)



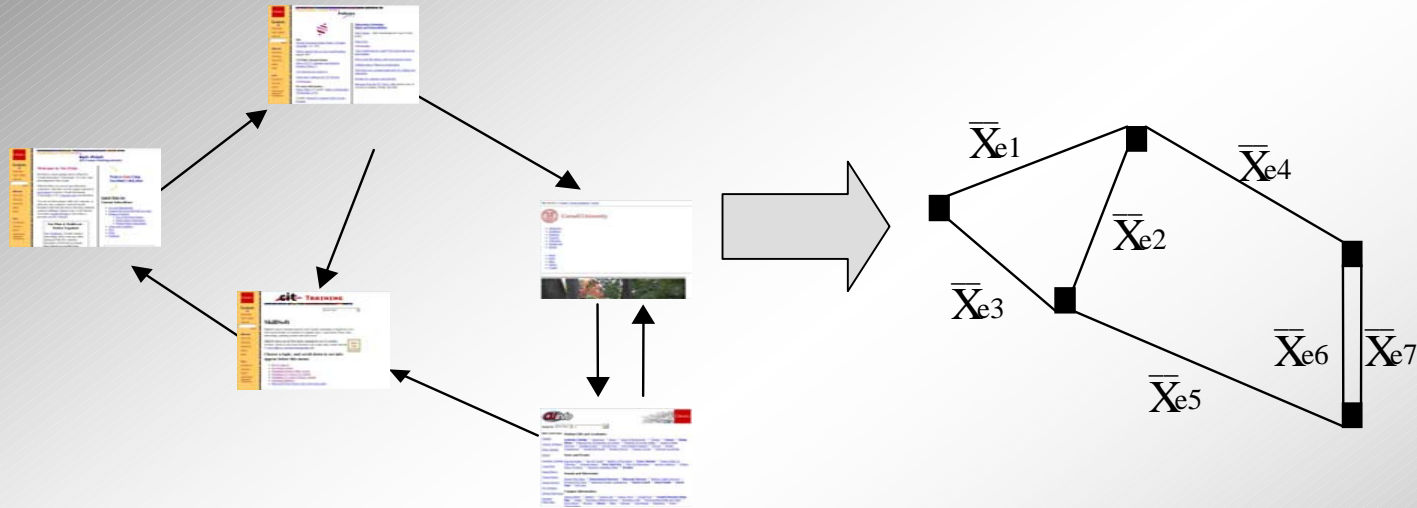
Learning cDoc Profile (III)



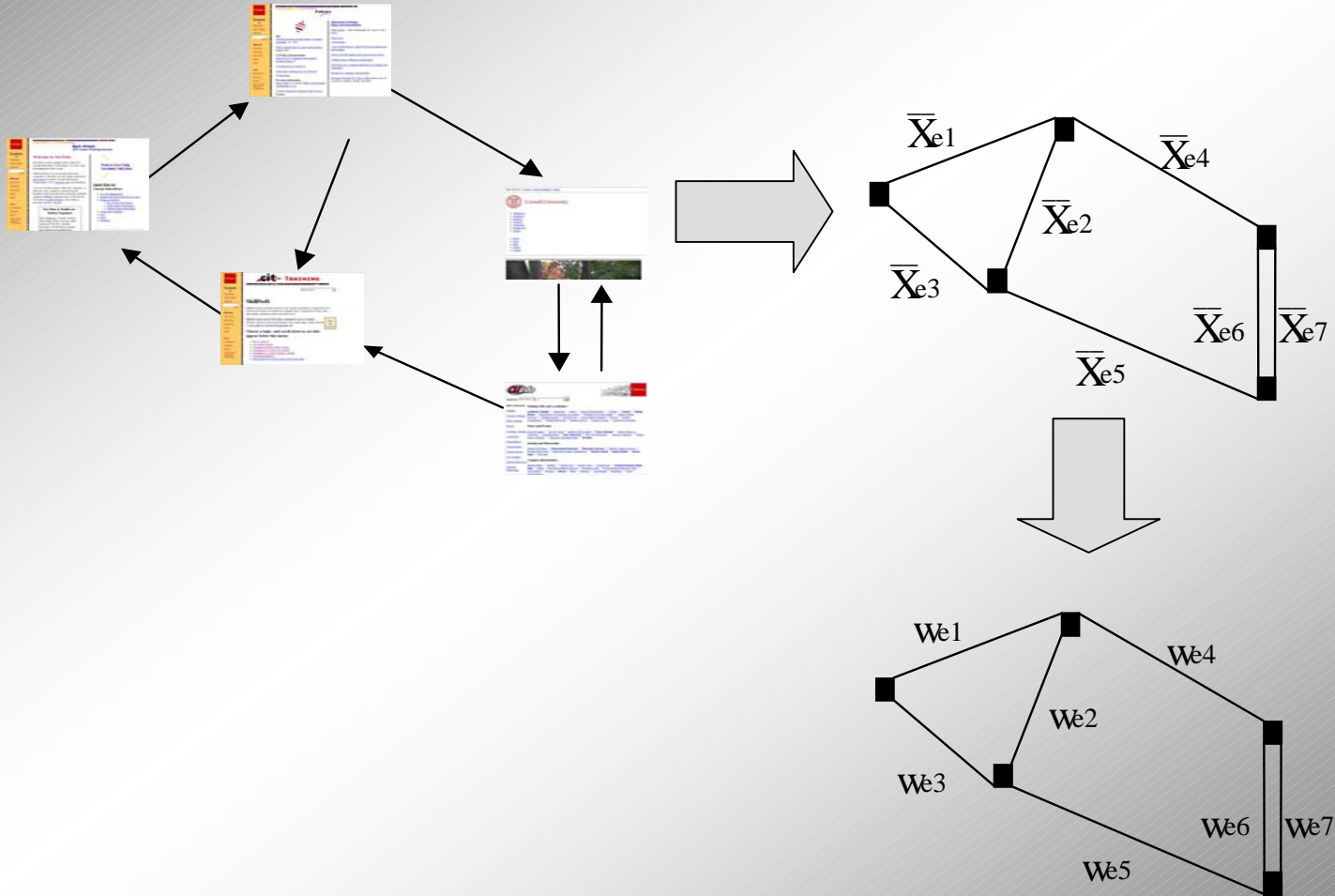
Processing New Site (I)



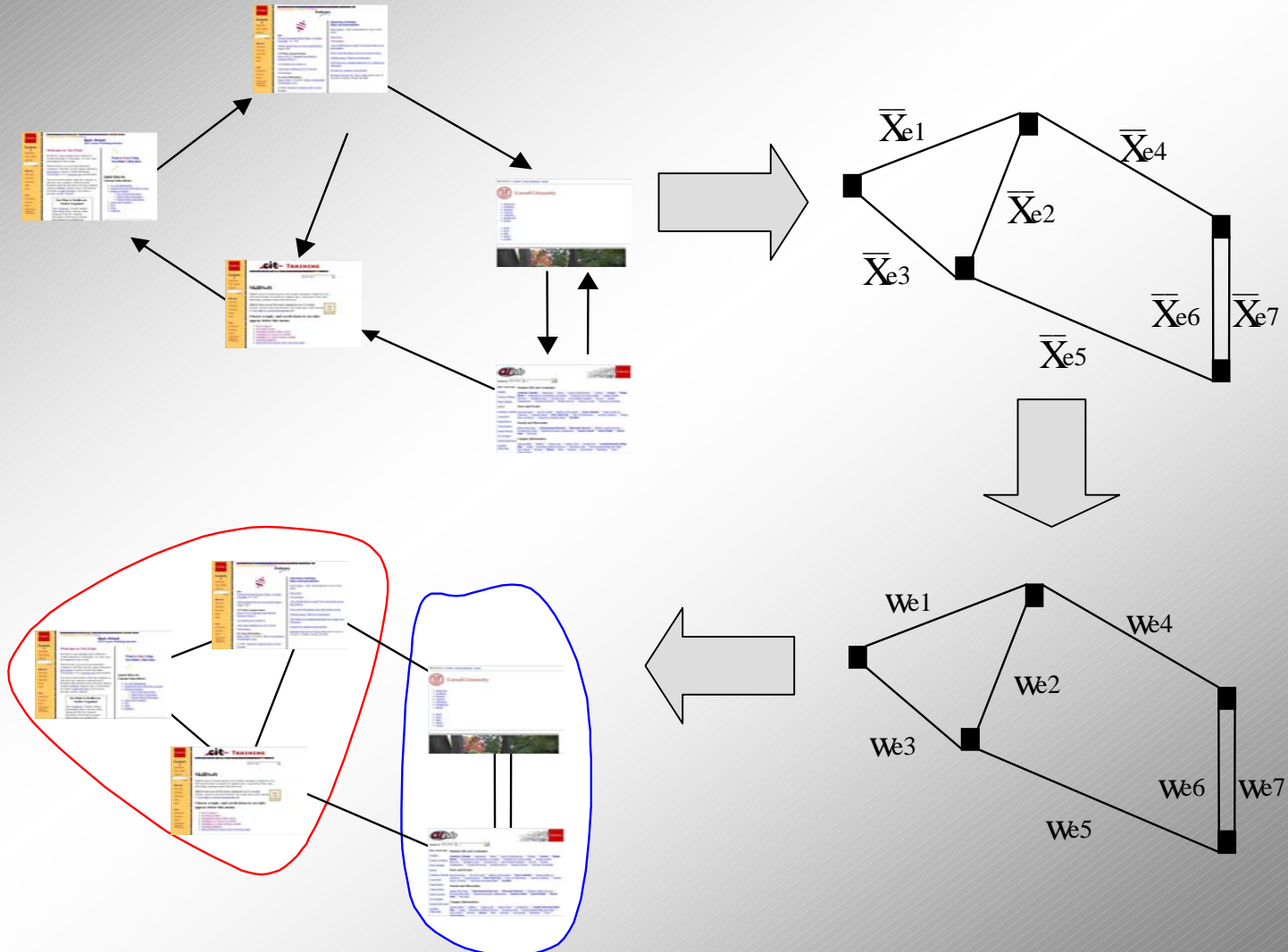
Processing New Site (II)



Processing New Site (III)



Processing New Site (IV)



Experimental Results

- Experimented on 50 web sites on educational topics
- Identified correctly most of the compound documents
- A small number of training web sites was enough to learn a good mapping
- The approach was relatively insensitive to the choice of training examples
- No single feature was vital for performance

Conclusion

- The ability to identify cDocs may lead to improvements of several fundamental web applications
- We present a unified framework to identify cDocs, which naturally combines structural and content features of web pages
- Experiments on a set of 50 educational web sites showed that our approach can identify well most of cDocs in the dataset

Future Work

- Improve learning and clustering algorithms to allow overlapping compound documents
- Introduce a post-processing step to refine the resulting clustering
- Investigate active learning approaches to finding compound documents
- Develop new information retrieval, link analysis, and metadata generation algorithms based on the extracted compound documents