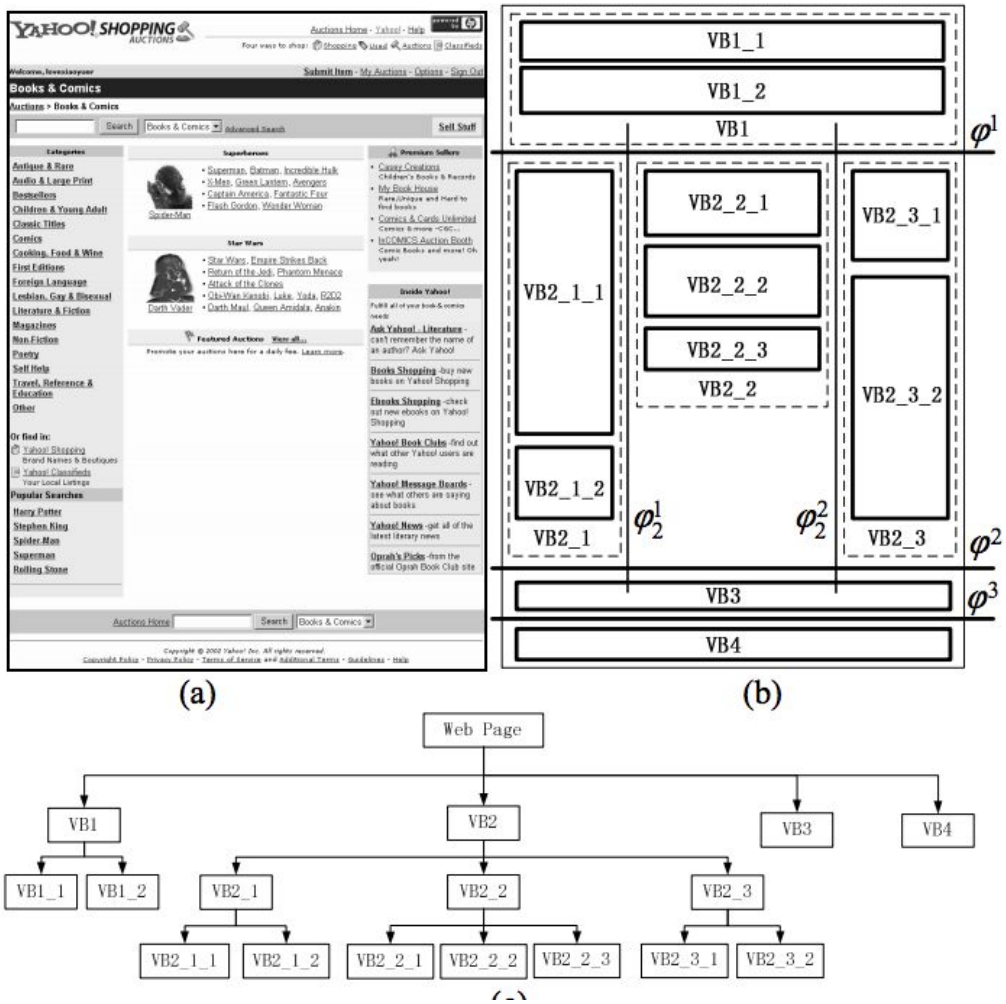
Joanna Wu

Computer Vision and Semantic Annotation Write Up

- Computer Vision web page analysis examples
    - “Extracting Content Structure for Web Pages based on Visual Representation”
        - Web pages these days are really complex, contain many elements and different topics
        - Researchers have been using many different techniques, from database techniques to semantic parsing to topic distillation to segment the pages
        - Proposes a VIPS (Vision-based Page Segmentation) algorithm to determine structure
        - Basic object is the leaf node in the DOM (document object model, they also have layout blocks which are groups of basic objects
        - Each level in the DOM tree is parsed one by one, to check if the current object is a single block. If not, then it’s children are parsed. When parsed, the blocks are put into a pool and checked for visual separators. Putting the content structure back together is the creation of the content structure



- an example using Yahoo
- Algorithm is efficient, uses top-down analysis
- When tested against human analysis, 97% of pages were corrected detected
- Next steps: adaptive content to consider mobile usage

- “Computer Vision-based Analysis of Web Page Structure for Assistive Interfaces”
  - Goal oriented around better web experiences for impaired users
  - Uses two algorithms- segmentation and classification algorithm
    - Segmentation: recursively divide image into tree
    - Classification: label regions of trees into roles on page
    - “Divide and conquer” method
    - Differs from VIPS method because of this method, uses image of the rendered page as evidence, uses Bayesian framework over hand-coded heuristics

                - Advantages: can be more accurate to the intention of a web page designer than the look of the internal code which is created for functionality
                - ARIA friendly
        - “Experimental web-scraping using the Google Cloud Platform”
            - They created an experiment flow by using images of websites, labelling the data on the websites, training a model on the data, and then evaluating the predictions made on that data and extracting information
            - Results: 96.88 precision
            - Advantages: use of an API makes building features in the future faster, is more accessible, and a vision based approach is easier to maintain
- Semantic Annotation
    - Benefits: Content reuse and implication of new knowledge
    - Gather text, use NLP on the text to split sentence and tag parts of speech, classify any unique entities, make connections and recognize relationships between known entities, represent knowledge in a framework, stored in a database. Can be combined with other data sets of knowledge
    - Advantages: reduced operational costs, smarter search and more complex queries, better representation of knowledge, more durable and long lasting storage of knowledge
    - Similarly requires split training and testing data
    - “Semantic Annotation of Web Pages Using Web Patterns”
        - Adding metadata knowledge to web content
        - Goals of simplifying querying and improve relevance of answers
        - Relies heavily on the intuition of work of web designers, to fulfill and mimic user goal
        - Classify web patterns as organization of user controls, other common UI components
        - This paper focuses on web pages selling products

Fig. 1. A web page with marked patterns. The patterns found are graphically marked on the page: *Sign on possibility*, *Price information*, *Purchase possibility*, and *Rating*.

-
- Elements are evaluated by proximity, similarity, continuity and closure
- Simplify querying by breaking queries down into key words
- Created a data set, and ran algorithm: taking the plain text, and data entity, then comparing to create pattern.
- Based on plain text, not HTML which is a big advantage
- Key characteristics of web patterns are independent of language environment