Latent Semantic Analysis

Eliminating extraneous items, thus producing **coherent segments** is an important problem in Web Accessibility

Our Approach

- Latent Semantic Analysis
- Computes linguistic similarity of text documents
- Primarily used for analyzing text documents based on linguistic features in the context of NLP and IR applications
- Uses Singular Value Decomposition technique to decompose a term-document matrix
- Computes a low rank approximation of the original matrix which represents semantic relatedness among the documents better than the original matrix
- **Our Approach**
  - Extend LSA with visual and structural features of web elements
  - Expanded feature set includes both visual features, as well as words as features
  - Each web element is considered as a document and each term represents either a visual, or linguistic attribute of a web element

Coherent Segmentation

- **Feature Extraction**: traverse the DOM tree of the web page to collect attribute value of web elements
- **attribute data of a web page element includes text-attributes, (x,y)-positions, width, height, font, parent-id, div-id, table-id, form-id, p-tag-id, UL-id, LI-id, etc.**

- **Singular Value Decomposition**: decompose the original matrix **W** into three components
- **get the low rank approximation** **W_k**

- **Construct Similarity Matrix**: use **W_k** to compute the element-element similarity matrix **M**

- **Clustering Web Elements**: cluster the web elements bottom-up into larger coherent segments based on similarity matrix **M**

Experimental Evaluation

- **Data set**: web pages from different domains (e.g., news, shopping, education, email, encyclopedia, weather forecast, etc)
- A total of 5,478 segments were identified and labeled manually in those web pages

- **Accuracy of Segmentation Algorithm**

Interface and User Study

- **User Interface**: HearSay
  - A non-visual web browser: provides basic screen reading functionalities
  - **Comparison of task completion time.**
  - **Tasks**: finding specific information in news articles, Wiki, weather reports, shopping sites, emails, etc.
  - **Avg. time and shortcuts pressed to complete each task**

Future Work

- **Incorporate the idea in mobile devices, and many other web mining and IR applications**
  - Acknowledgements: This work was supported in part by
    NSF grants CNS-0751083 and IIS-0808678.