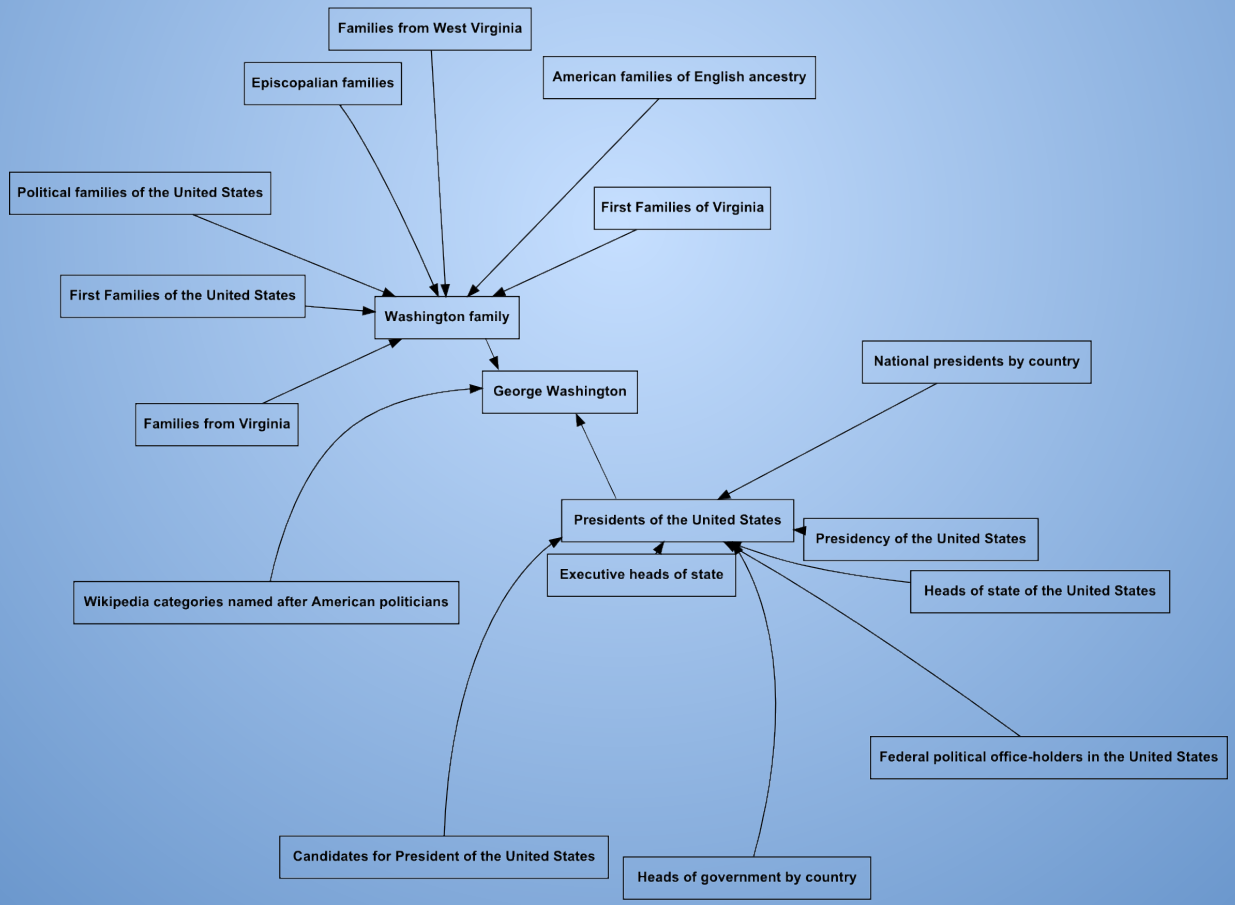
Semi-Supervised Wikipedia Category Suggestion

Cris Feo and Benjamin Perez

Introduction

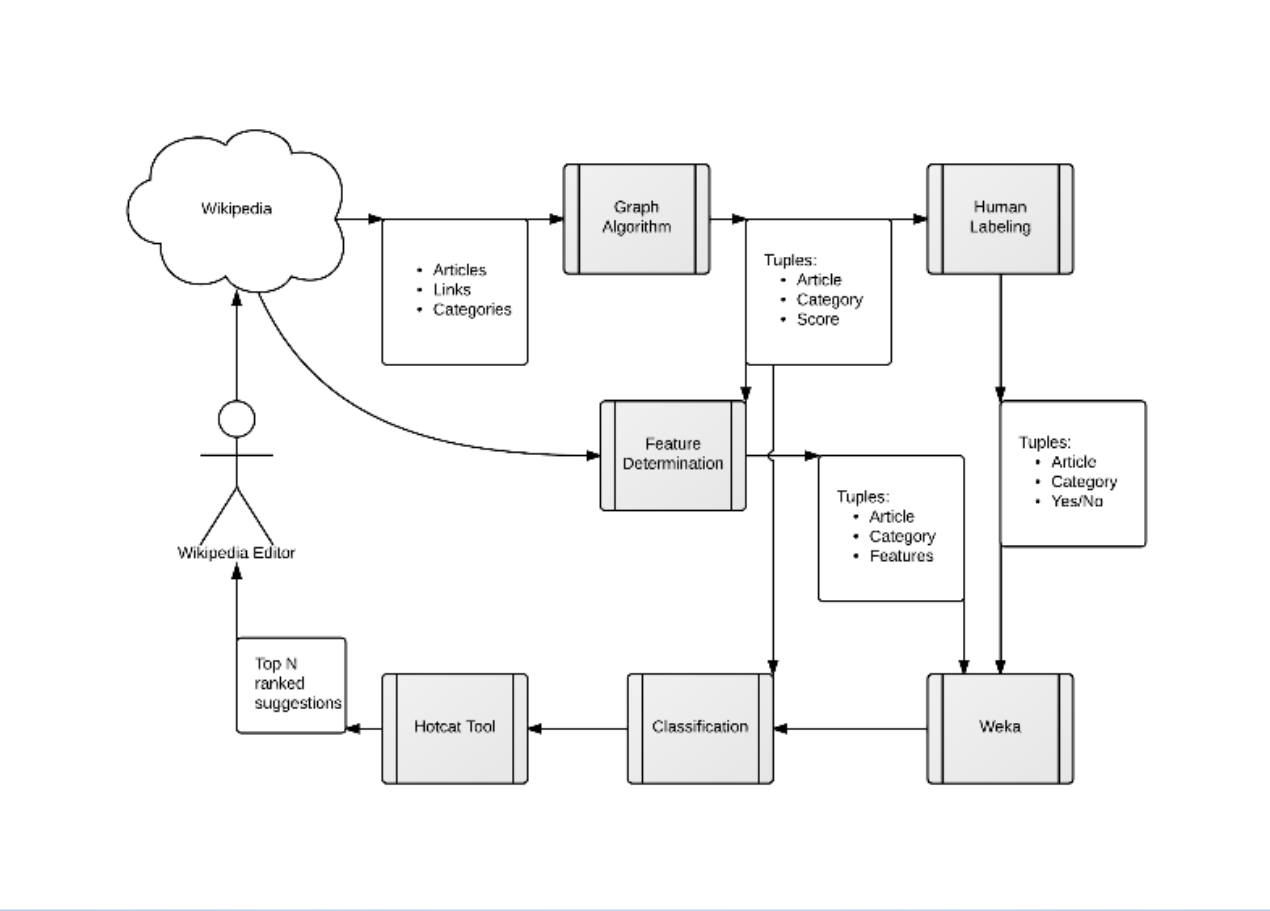
- Complexity of Wikipedia categories



Related Work

- Early work distinguishes between less than 20 categories. Accurate, but has problems with scale
  - Binary/Multi-Class Classifiers
  - Content Features
  - Network Features
- Recent work uses semi-supervised learning
  - Graph Algorithms, Label-Propagation
  - Good Preliminary Results
  - Use Link-Structure of Wikipedia

Anticipated Approach



Evaluation Criteria

- Good Suggestions
  - Remove existing category links from Wikipedia, check if algorithm suggests them
- Bad Suggestions
  - Label a subset of graph algorithm output using Wikipedia editors
  - Use this data for training and testing

Research Timeline

- Already Completed: Background work, infrastructure
- By Thanksgiving: Decide on a graph-based approach
- By Christmas: Have graph algorithms working, decide how to label data
- By Early February: Have labeled data, start working on filtering
- By Late March: Finish filtering suggestions, create a GUI tool
- Final Tasks: Finish report and integration, collect feedback from editors