Web Document Clustering: A Feasibility Demonstration

Zamir & Etzioni

Web Clustering Requirements

- Consider
  - Are the following requirements necessary?
  - Sufficient?
  - Do they apply if you are at the search engine?
- Relevance
  - Groups should separate relevant from irrelevant responses
- Browsable
  - Clustering should produce summaries that can be browsed directly

Web Clustering Requirements (2)

- Overlap
  - Clusters should allow a single doc to be in multiple clusters
- Snippet Tolerant
  - Clustering should be good if have only snippet as people will not wait for entire doc download
- Speed
  - Clustering should work at 1000+ / second
- Incremental
  - Clustering should take in new info without having to start over

Clustering Algorithm (1)

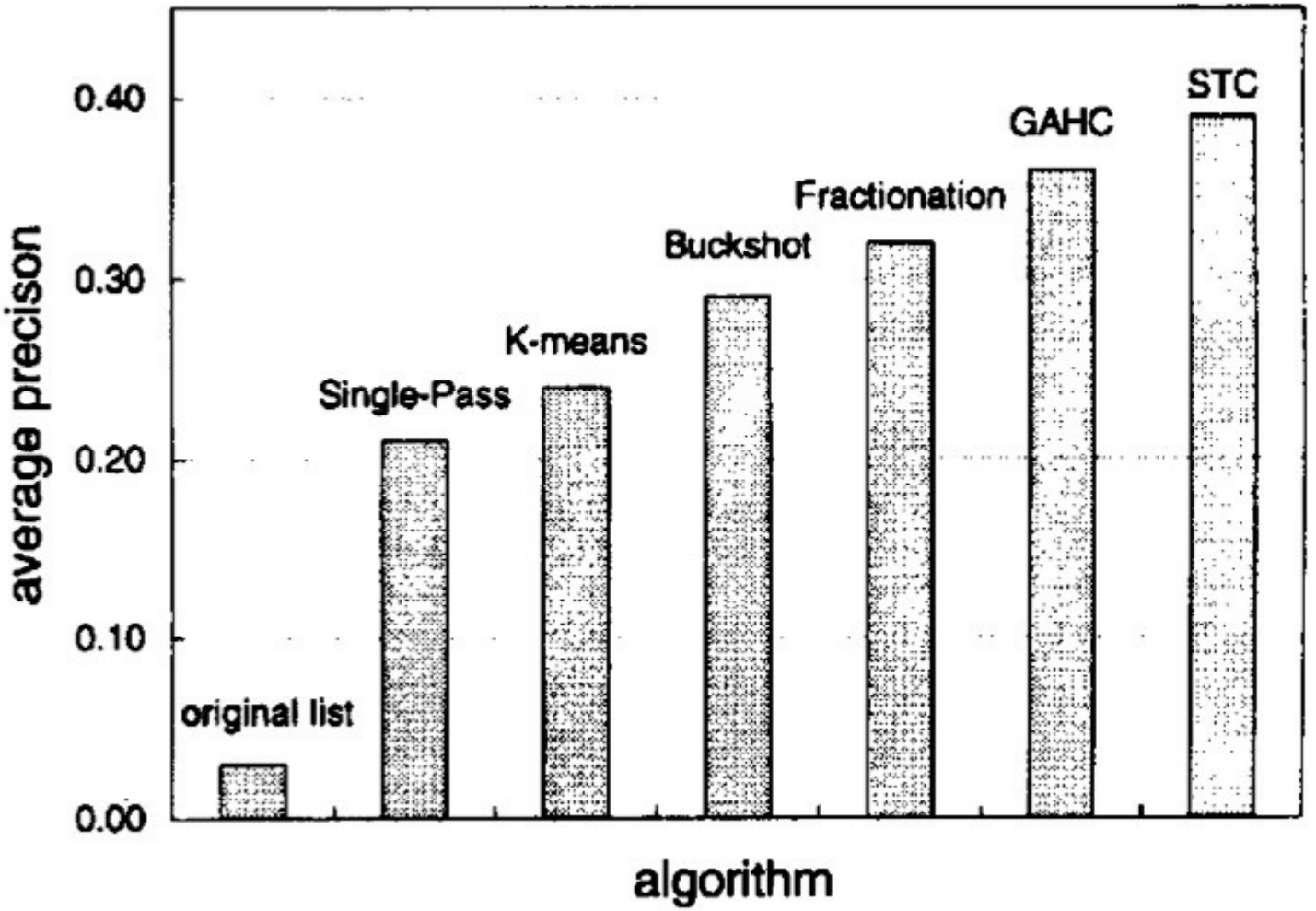
- Clean Documents
  - Lightly stem, mark sentence boundaries
  - Remove html tags and punctuation
- Identify Base Clusters
  - Any one or more word phrase shared by 2 or more documents
  - Score each phrase using s = B*f(P)
    - B is the number of documents
    - P is the number of words that are:
      - Not stop words
      - Appear in at least 3 documents
      - Appear in fewer than 40% of documents
    - f() is function such that f(p)=small if p=1, linear for 2<=p<=6 and constant thereafter

Clustering Algorithm (2)

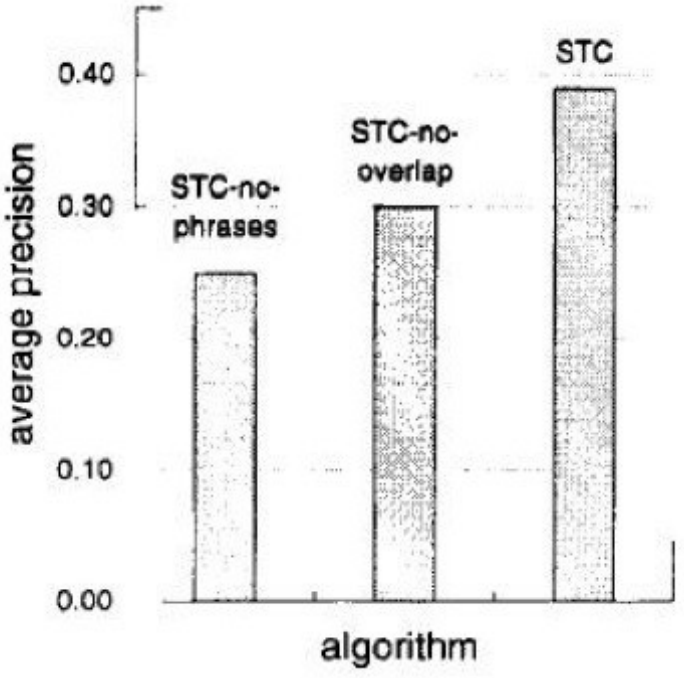
- Combine Clusters
  - Define sim=1 of two clusters if intersection/sizeis always greater than 0.5
  - In a single pass, merge any clusters with sim>0
    - Note that this may result in merging 2 clusters that have a sim=0 iff they share a clusters for which they have a sim > 0
  - Do this merge step on only the N highest scoring base clusters (a speed game, if you have a lot of documents you will have a lot of clusters)
- Display to user to N clusters

Evaluation

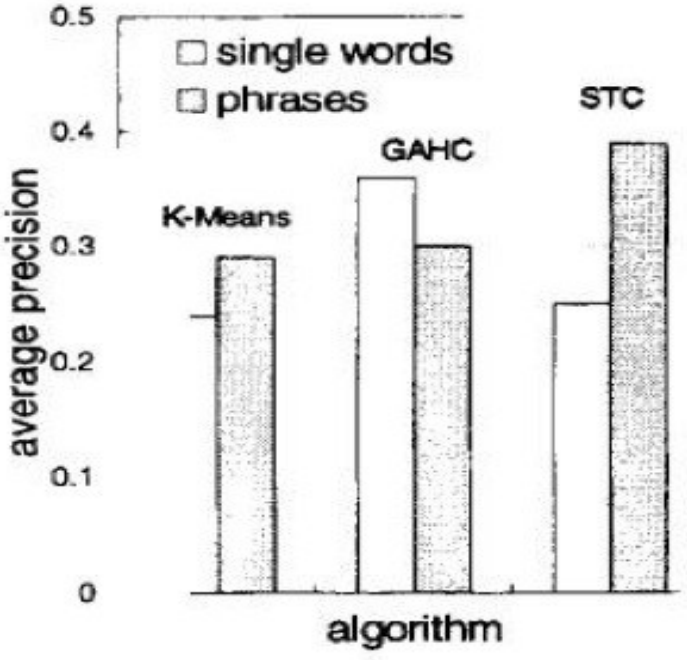
- Does it meet criteria they set out?
- Does it help?



Evaluation – Why does it work?

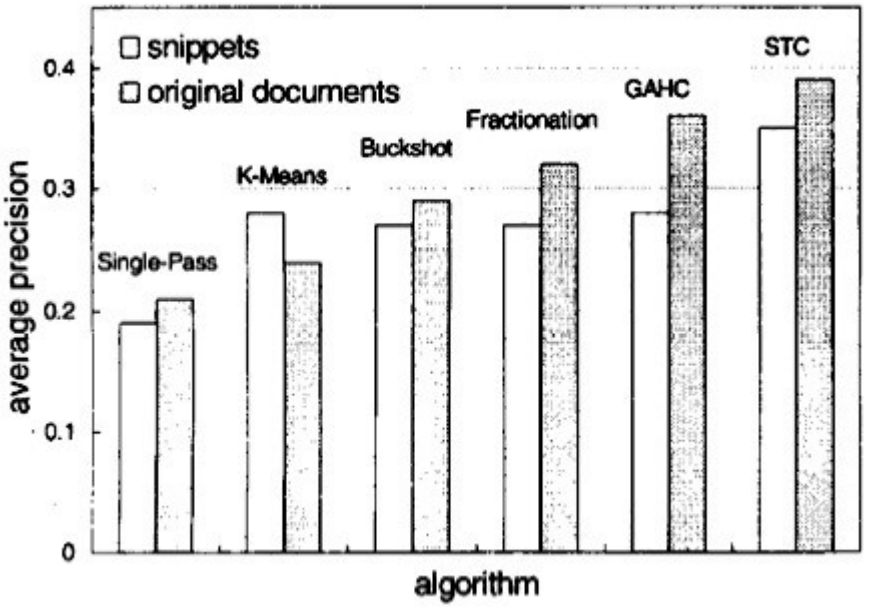


Parts of STC and their effect on precision

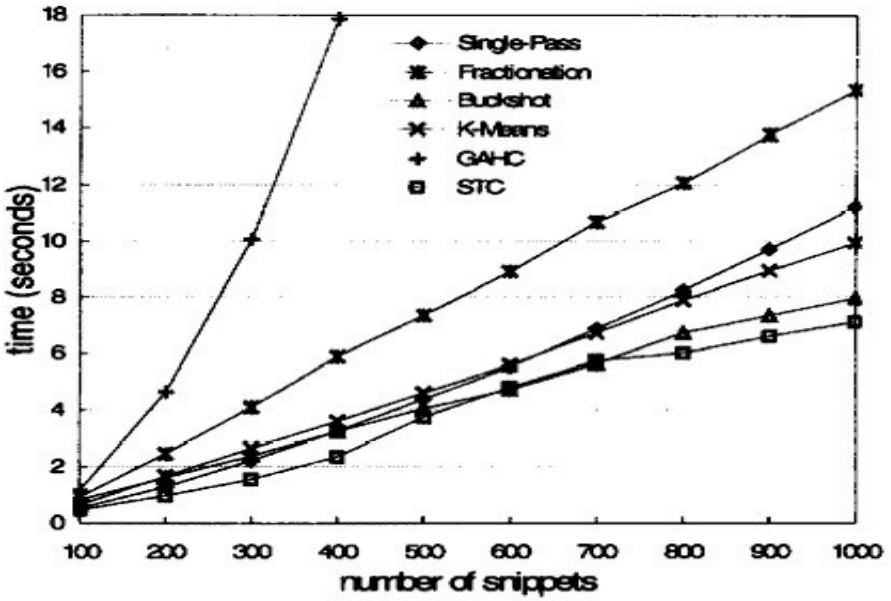


STC vs other algorithms with and without phrases

Comparison Stats



Accuracy given “snippets” vs whole document



Time required for clustering