Clustering Web Search engines

PAPER : WEB DOCUMENT CLUSTERING: A FEASIBILITY STUDY BY OREN ZAMIR AND OREN ETZIONI

PRESENTERS : NABEEL ABUBAKER AND HAMZEH AHANGARI

Introduction & Motivation

Normal search engines → Unordered top-ranked results

- Takes time to find the "appropriate" result/document

Clustering search engines → Top ranked results, but **clustered**

- Easier to find your "appropriate" document in the "appropriate" cluster

An Example: The famous “Jaguar”

Normal search engine might return:

- Jaguar, the car *in document 1.*
- Jaguar, the animal *in document 2*.
- Jaguar, the animal *in document 3*
- Jaguar, The car *in document 4*.
- Jaguar, the animal *in document 5*.
- Jaguar, The car *in document 6*.

Clustered search result:

- Cluster 1: **jaguar, the car.** 24 results *click to show*
- Cluster 1: **jaguar, the animal**. 18 results *click to show*

Contributions

STC : Suffix Tree Clustering, an incremental, lineartime algorithm.

Identified key requirements for web clustering methods:

- Relevance
- Browsable summaries
- Overlap
- Snippet-tolerance
- Speed
- Incrementality

They've created a prototype clustering web search engine called MetaCrawler-STC

- MetaCrawler : **Normal** search engine. (Actually a metasearch engine, 1994. Now called Zoo)
- MetaCrawler-STC : **Clustering** search engine.

Simple architecture

** The clustering should be performed on separate machines for better performance.*

Suffix Tree Clustering Algorithm

3 steps

- Cleaning
  - Light stemming
- Identifying Base Clusters
  - Making suffix tree
- Combining Base Clusters
  - Merging clusters based on similarity

Linear time

No need to define “number of clusters”

Not sensitive to similarity threshold

Suffix Tree Clustering Algorithm

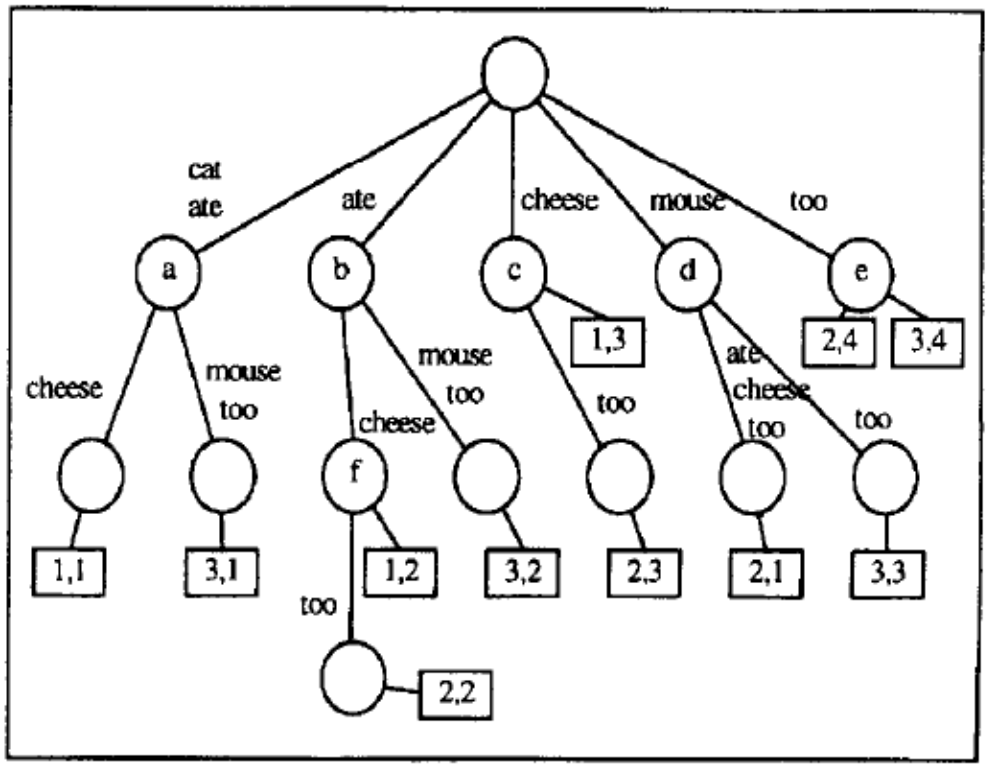
The suffix tree of the strings “cat ate cheese”, “mouse ate cheese too” and “cat ate mouse too”.

Multi-word phrase, not single word

Overlap

Incremental

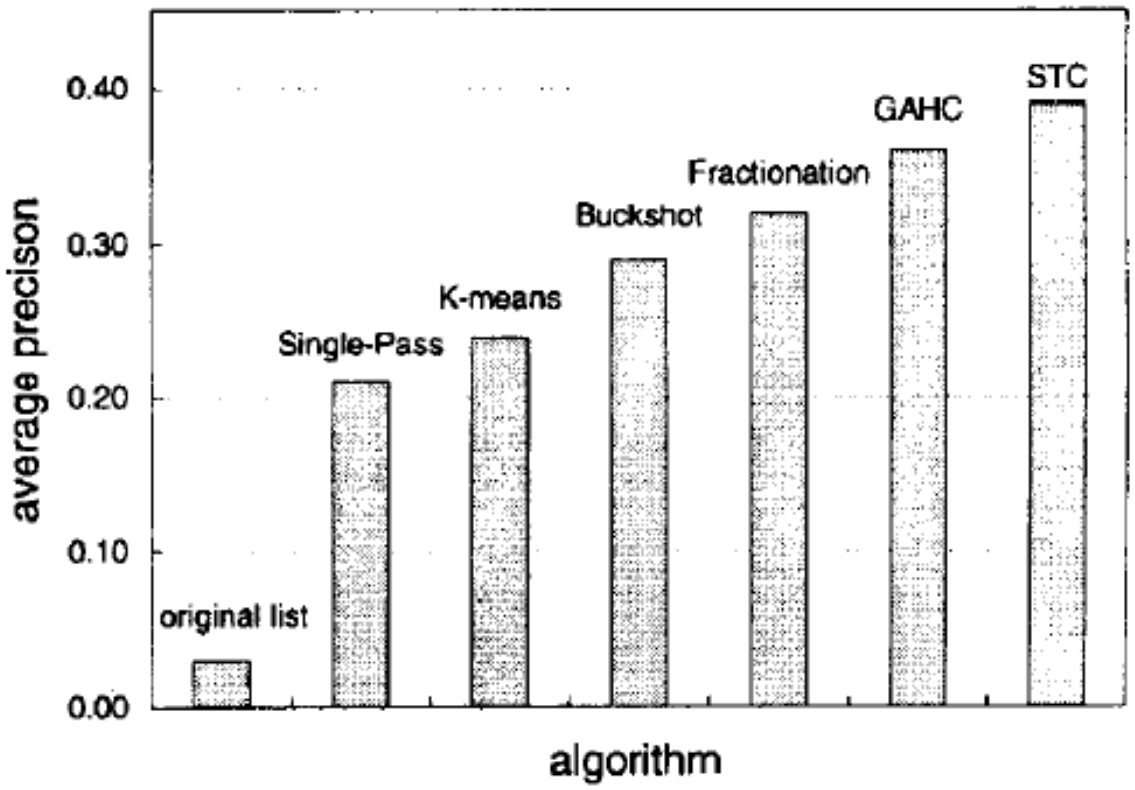
Order independent



Experiments

Critical reasons of STC success :

- Multi-word phrase
- Overlap



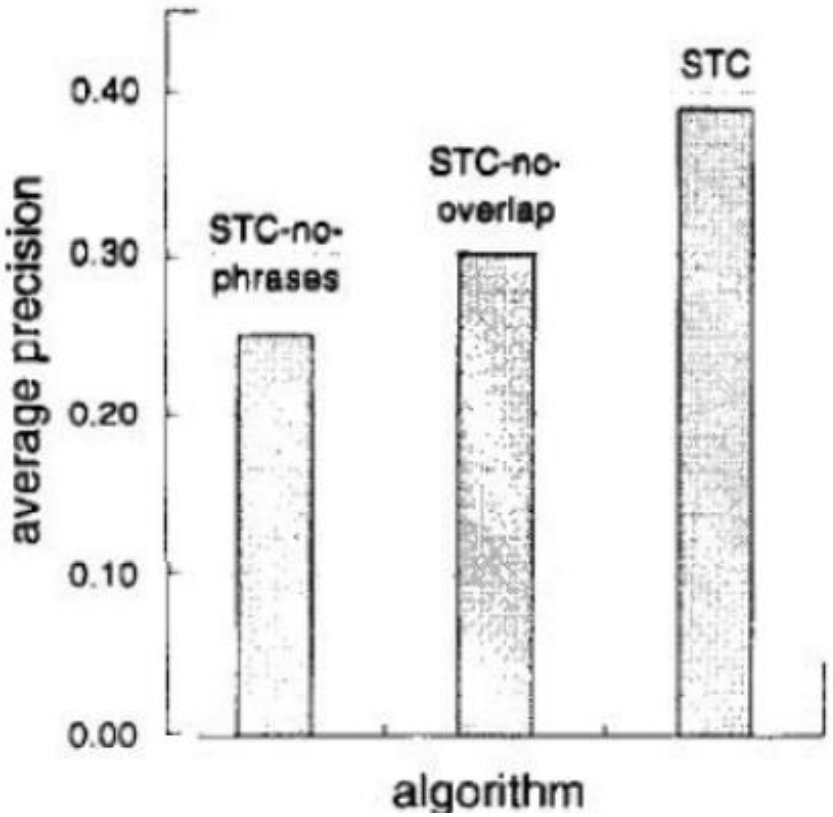
Experiments

Impact of multi-word phrase

- STC-no-phrases

Impact of overlap

- STC-no-overlap



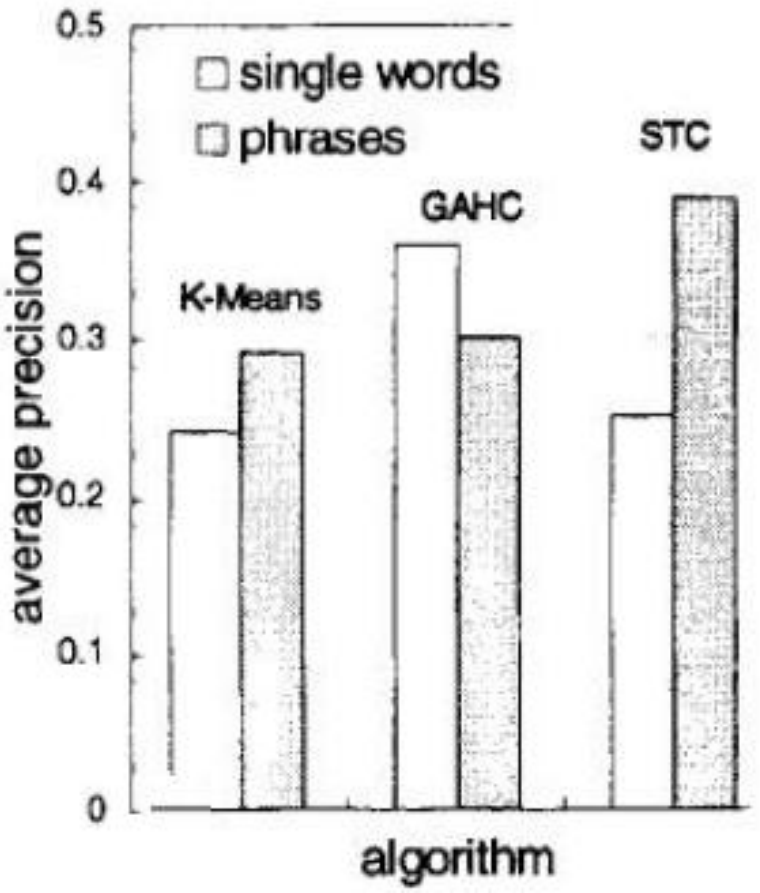
Experiments

Improve other algorithms by multi-word phrase

vs single-word

+/- impact

- Big impact on STC
- Not dramatic for others
- Can’t be plugged in!
- Inextricable STC’s part



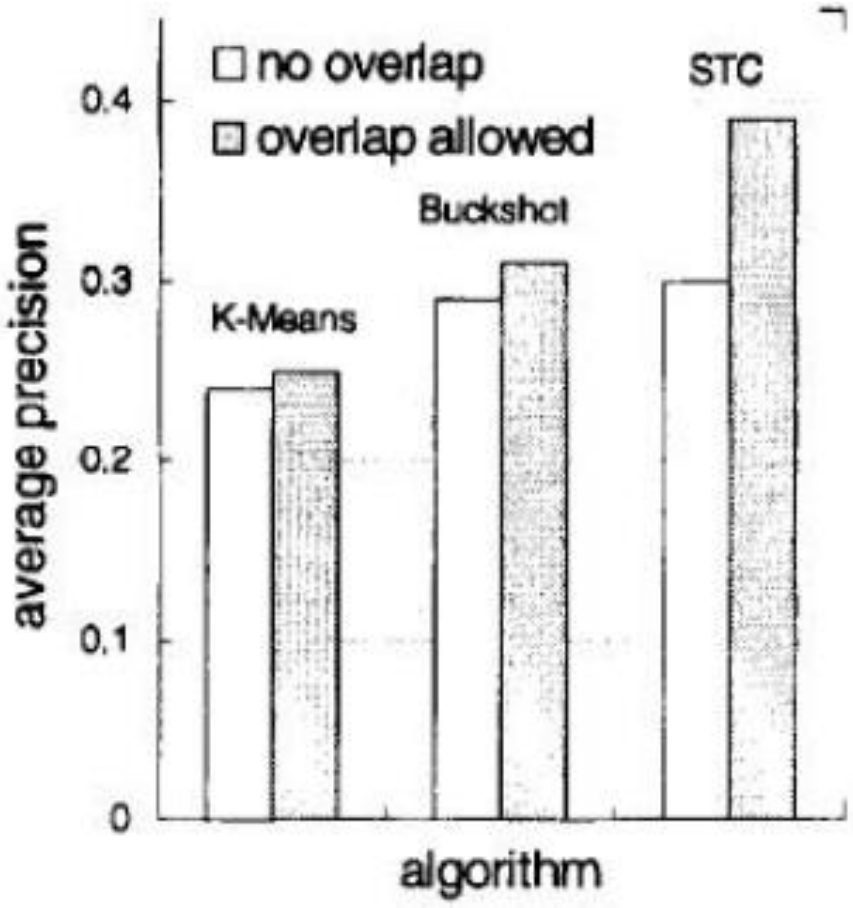
Experiments

Improve other clustering algorithms by overlapping

vs no-overlapping

+/- impact

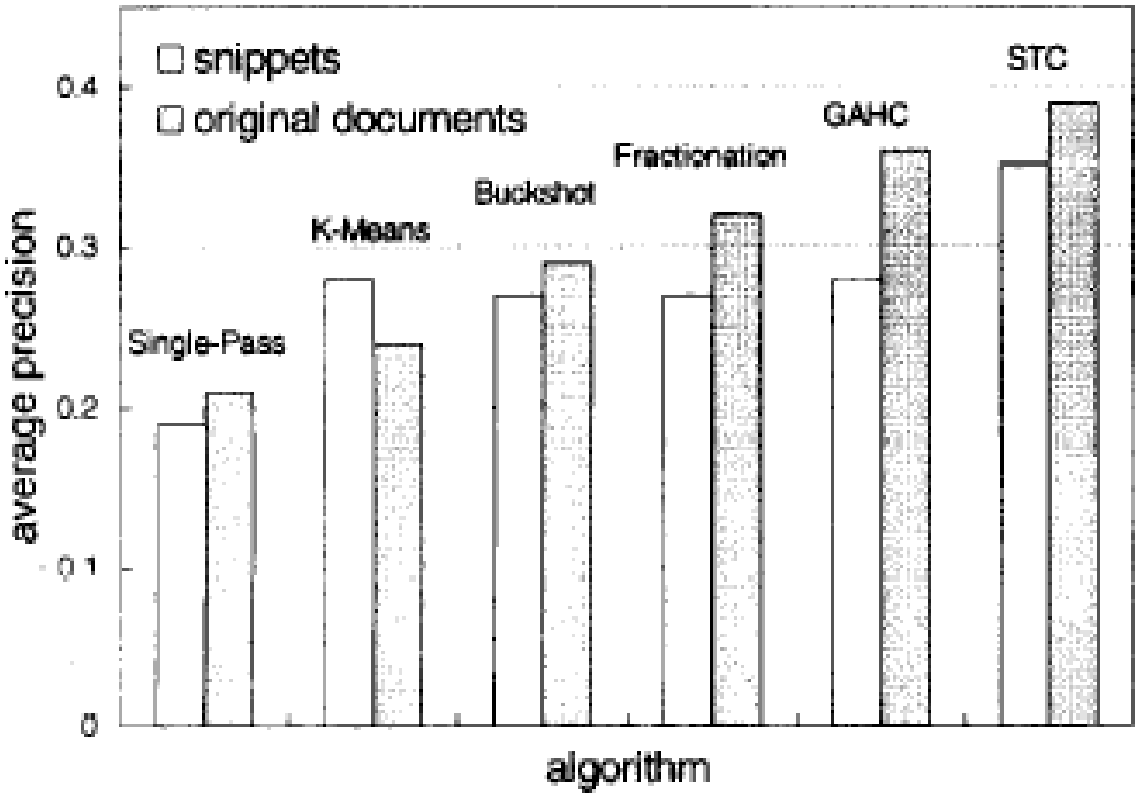
- Big impact on STC
- Not dramatic for others
- Can't be plugged in!
- Inextricable STC's part



Experiments

Snippets vs whole document

Decrease in quality is apparent but small



Reference(s)

Zamir, Oren, and Oren Etzioni. "Web document clustering: A feasibility demonstration." *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998.

Thank you

☺