Grouper: A Dynamic Clustering Interface to Web Search Results

Oren Zamir and Oren Etzioni
Department of Computer Science and Engineering
University of Washington

Doğan Altunbay and M. Burak Şenol

Introduction

- long ordered list of document «snippets»
- an example: Google

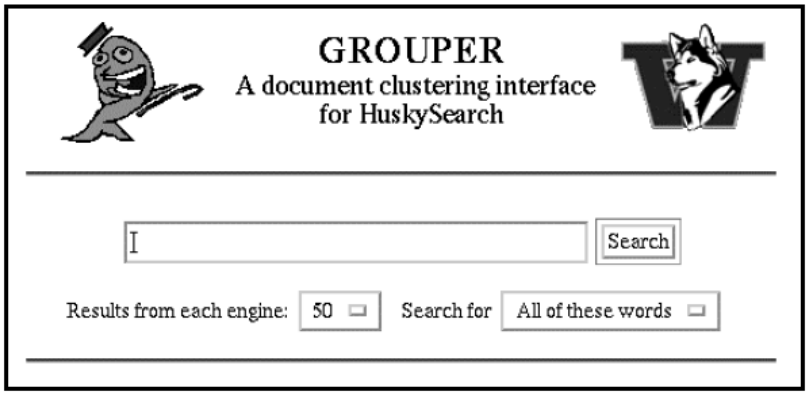
- main goal of the paper is to make search engine results easy to browse by clustering them

- post retrieval clustering is used

- an interface to the HuskySearch meta-search service

STC Algorithm

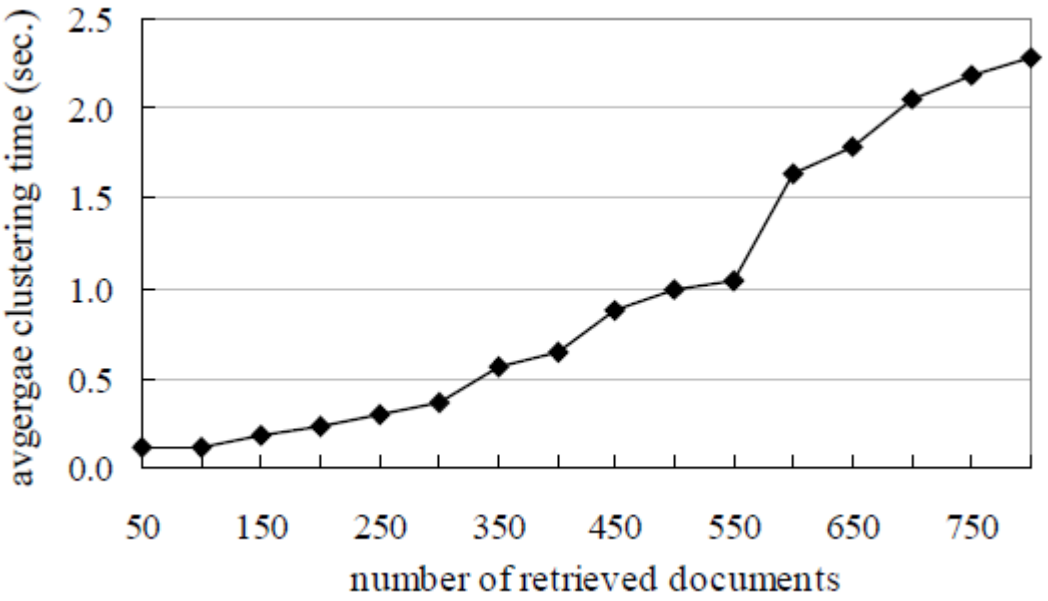
- Suffix Tree Clustering
  - works in linear time
  - based on identifying phrases
  - can create *overlapping clusters*
  - works incrementally
  - robust
  - does not need the number of clusters as an input



Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) ● Ahavat Israel – The Amazing Jewish Website! ● Israel and Judaism ● Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) ● Publications and Data of the BANK OF ISRAEL ● Consulate General of Israel to the Mid-Atlantic Region ● The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) ● Interactive Israel tourism guide - Jerusalem ● Ambassade d'Israel ● Travel to Israel Opportunites
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) ● Israel at Fifty: Our Introduction to The Six Day War ● Machal - Volunteers in the Israel's War of Independence ● HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) ● Israel Hotel Association ● Israel Association of Electronics Industries ● Focus Capital Group - Israel

Speed


avgergae clustering time (sec.)
2.5
2.0
1.5
1.0
0.5
0.0
50
150
250
350
450
550
650
750
number of retrieved documents

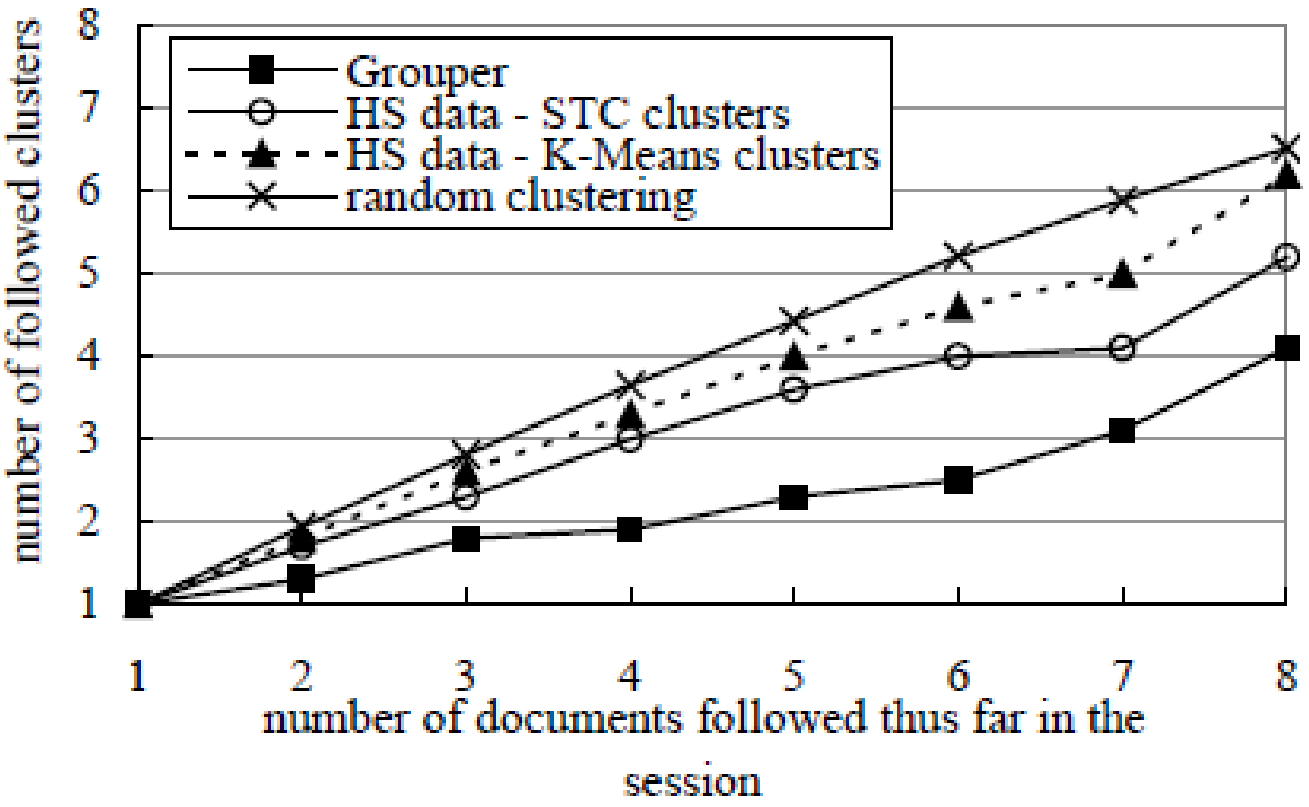
Evaluation

- The system is evaluated using session logs
  - Logs store the behavior of both search Uis, Grouper and HuskySearch
- Evaluation Perspectives
  - **Cluster Coherence**
    - How many clusters are visited after a particular number of documents are followed?
  - **Grouper UI vs. Ranked List**
    - How does the clustering affect efficiency of retrieval?

Cluster Coherence

- **Hypothesis:** Users will tend to follow documents from relatively few clusters.
- **Metric:** Average number of followed clusters as a function of the number of documents followed in the session.
    - Grouper's clusters
    - Random clustering
    - STC clustering on HuskySearch results
    - K-means clustering on HuskySearch results

Cluster Coherence

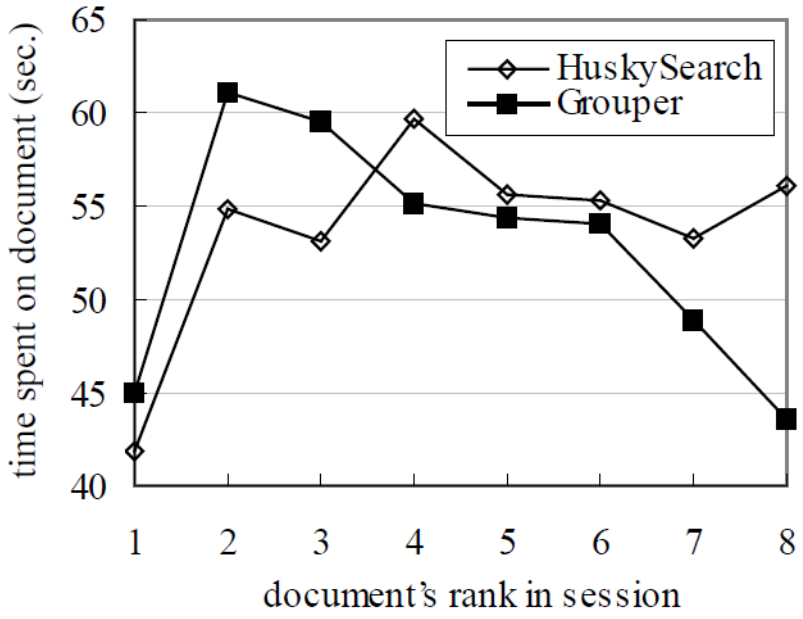


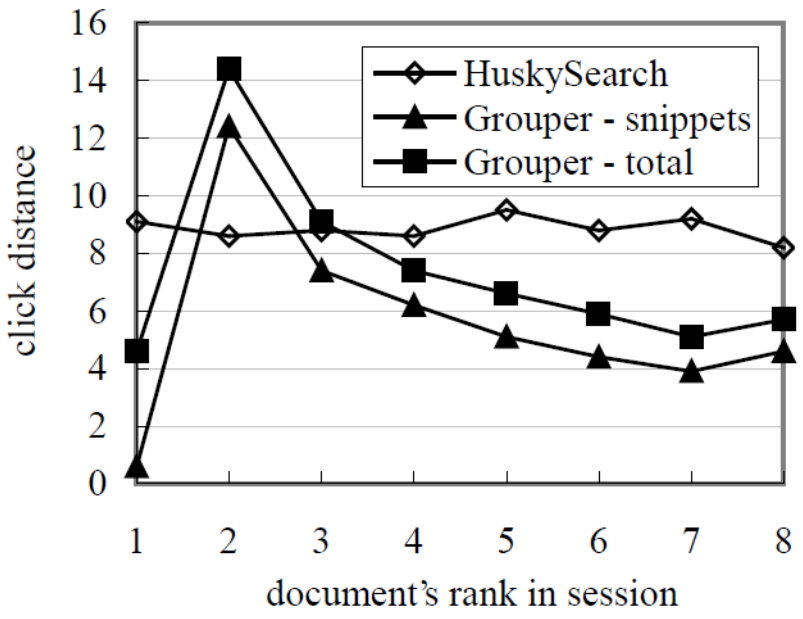
Grouper UI vs Ranked List

- **Number of documents followed**
  - A document is followed if the user clicked on it.
- **Time Spent**
  - Elapsed time between two successive document requests, including network latency, reading time, and traversal in results.
- **Click Distance**
  - **Ranked List:** Number of documents between two clicks.
  - **Grouper Interface:** Number of clusters and snippets between two clicks.

Grouper vs. Ranked List

Num. of Docs. Followed:	0	1	2	3	4	5	6	7	8+
% of HuskySearch sessions	53.0	26.9	8.4	4.2	2.3	1.6	1.1	0.7	1.9
% of Grouper sessions	46.0	25.2	10.2	6.0	3.9	2.4	1.8	1.4	3.2





Conclusion

- The paper introduces a clustering interface to HuskySearch meta search engine.
- Two issues are forwarded unresolved to Grouper II:
  - Grouper should provide a view of non-merged base clusters, which may be helpful for novice users.
  - For scaling considerations, clusters should be presented hierarchically so users can navigate the results more efficiently.