MiTexCube: MicroTextCluster Cube for Online Analysis of Text Cells

Duo Zhang, ChengXiang Zhai, Jiawei Han
University of Illinois at Urbana-Champaign

Multidimensional Text Databases

Product Reviews

Epinions.com
a Shopping.com company

cnet reviews

Brand	Product	…	Reviews
Sony	VAIO		…Good LCD…

Aviation Reports


DC-8 Airborne Laboratory / NASA Ames Research Center - Moffet Field

Time	Airport	…	Narratives
9901	AUS		… bad weather…

Research Papers

dblp.uni-trier.de
Computer Science Bibliography

The DBLP Computer

acm PORTAL

Year	Conference	…	Abstract
2007	KDD	…	…Time Series…

Unstructured data

Structured data

Motivation

- A fundamental challenge in Online Analytical Processing (OLAP) of multidimensional text database is to summarize the content in its text cells.
  1. Neutral Summarization

     Give the most representative documents within a text cell
  2. Query-Specific Summarization

     Give the most relevant documents based on a query within a text cell, which also cover the content of the text cell well

Table 1: An example of text database in ASRS

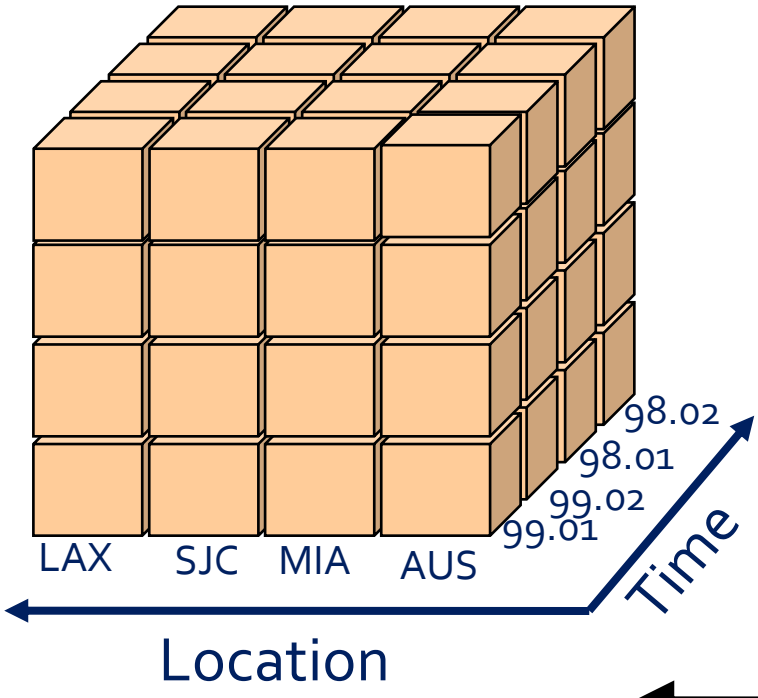
ACN	Time	Airport	···	Light	Narrative
101285	199901	MSP	···	Daylight	Document 1
101286	199901	CKB	···	Night	Document 2
101291	199902	LAX	···	Dawn	Document 3

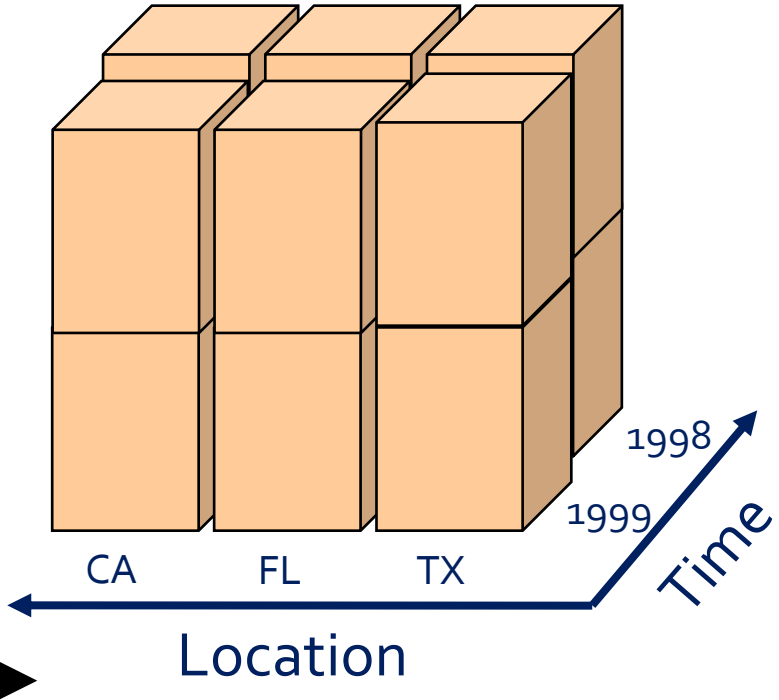
- For example:
  1. What are those narratives talking about during night in Jan. 1999? Or in the whole year 1999?
  2. What have the pilots mentioned about the landing at LAX in 1999?

OLAP in a Data Cube

Table 1: An example of text database in ASRS

ACN	Time	Airport	···	Light	Narrative
101285	199901	MSP	···	Daylight	Document 1
101286	199901	CKB	···	Night	Document 2
101291	199902	LAX	···	Dawn	Document 3



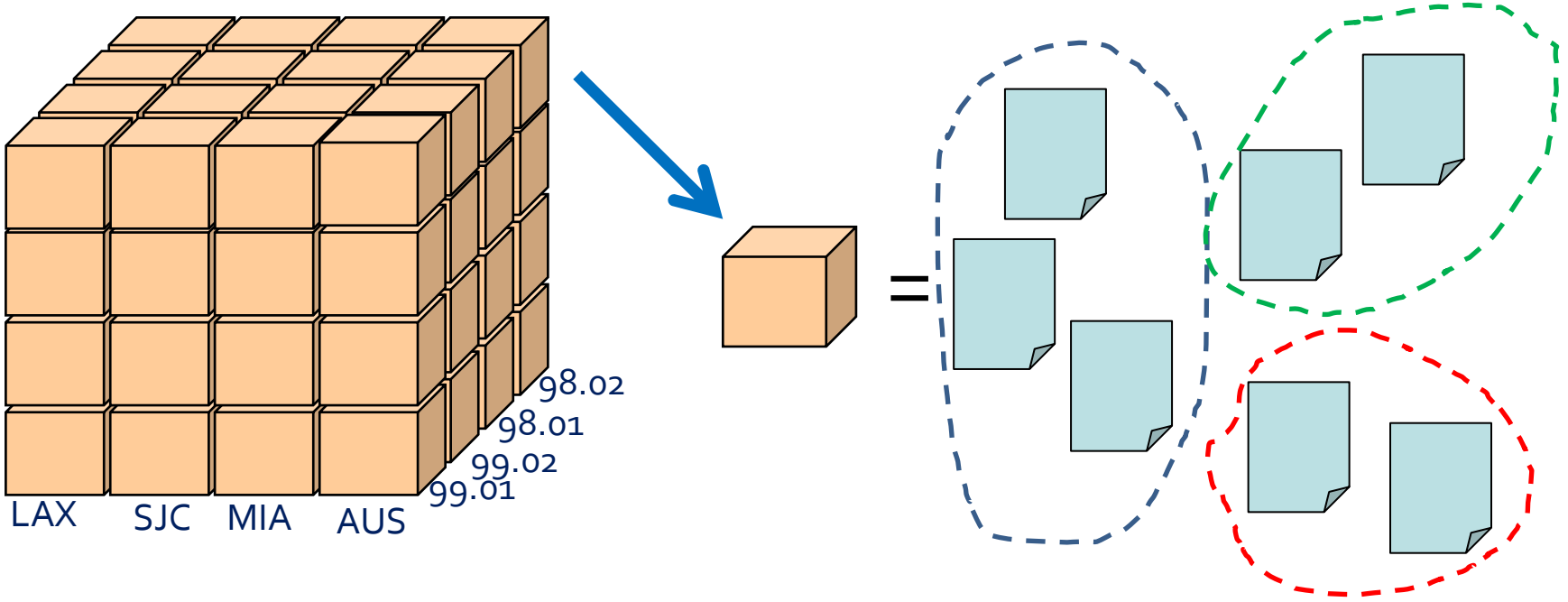


drill-down **roll-up**

Materialization with Micro-Clusters

Table 1: An example of text database in ASRS

ACN	Time	Airport	···	Light	Narrative
101285	199901	MSP	···	Daylight	Document 1
101286	199901	CKB	···	Night	Document 2
101291	199902	LAX	···	Dawn	Document 3

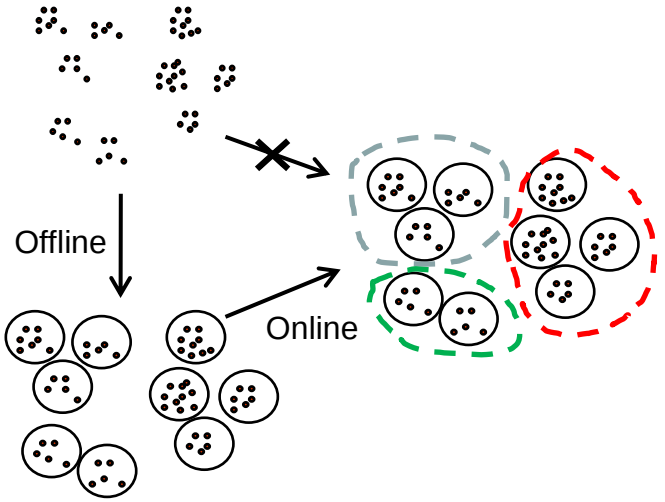


MicroTextCluster Cube

Table 2: An Example of a MiTexCluster Cube

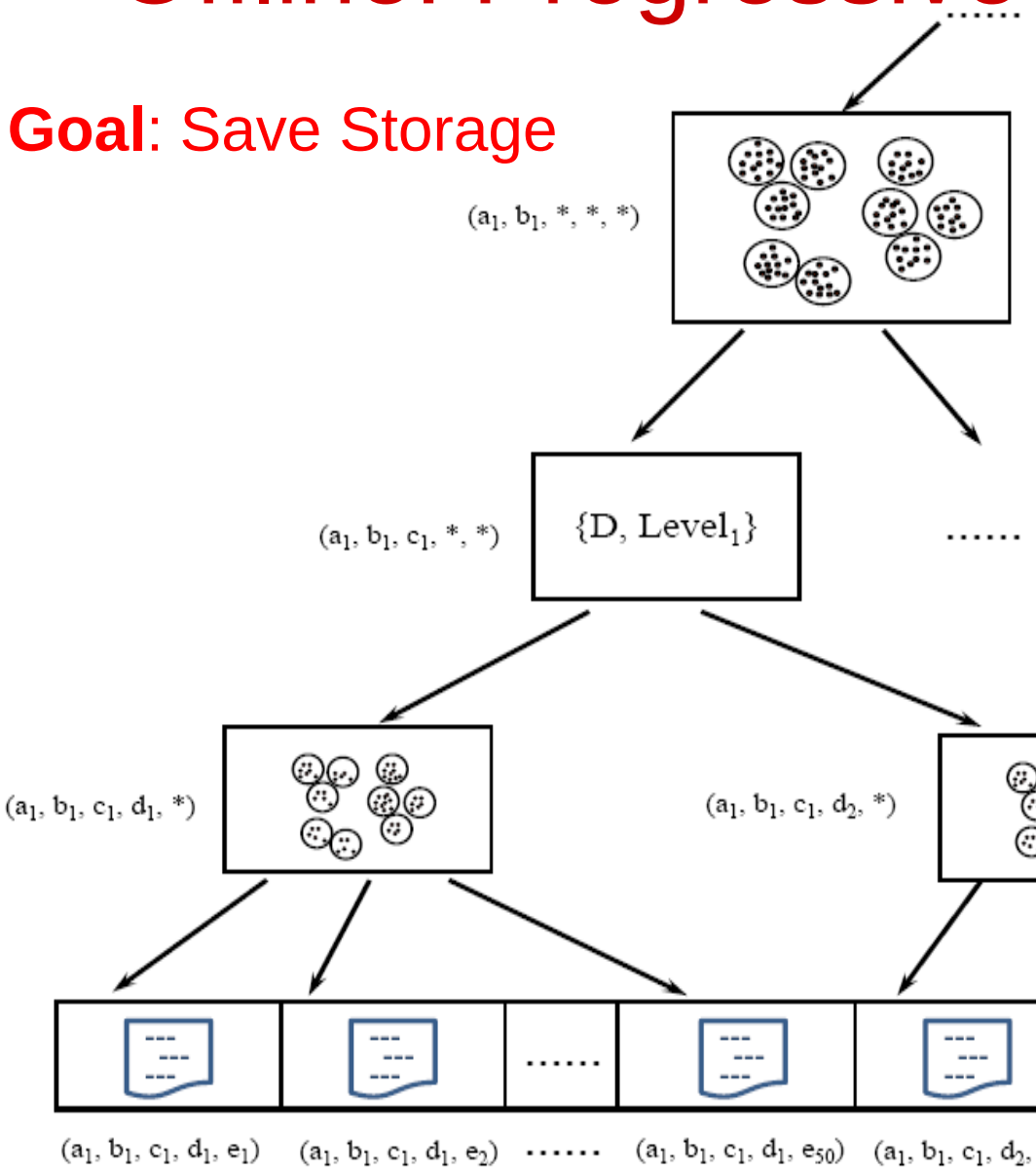
Cell	Doc ID	Content	Micro-Text-Clusters
(Time=1999, Location=TX)	d_1	... due to stronger than forecasted winds and weather going ...	(weather 2.5, wind 1.2, ...), 3
	d_2	... I think that the weather, headwinds, shrinking dew-point/temperature contributed to the fuel emergency ...	
	d_3	... After an hour, the weather had not much improved. We were in the clear for a bit and then hit another cloud bank ...	
	d_4	... so that if we saw the ARPT, we could land ...	(land 2.1, rule 0.9, ...), 2
	d_5	... we were in class G and the IFR rules tell us to land ...	

Goal: Improve online efficiency



Offline: Progressive Materialization

Goal: Save Storage



If larger than **M**, re-cluster into **K** bigger micro-clusters

Total number of micro-clusters from sub-cells is smaller than **M**

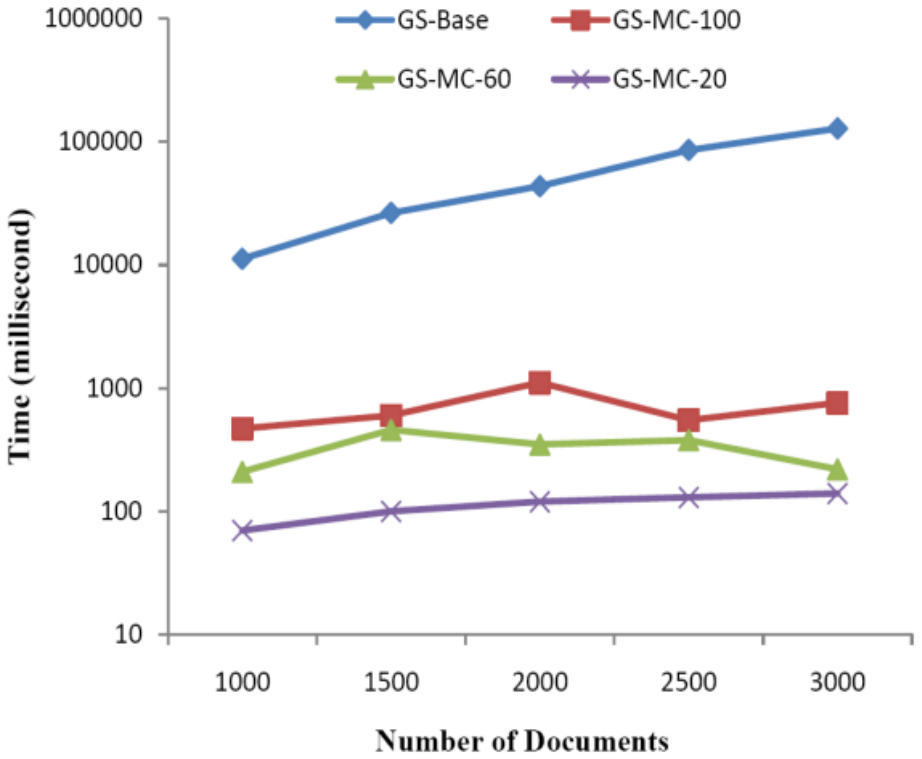
Each cell stores **K** micro-clusters

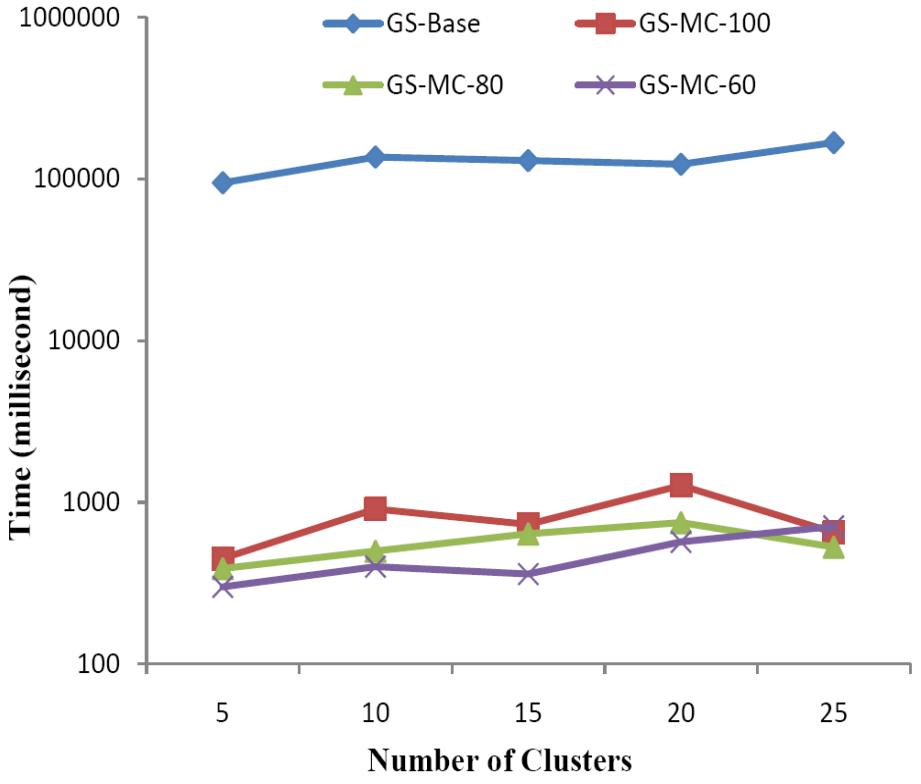
Online: Applications

- Neutral Cell Summarization
  - cluster documents into groups and select representative documents from each group
- Query-Specific Cell Summarization
  - select documents that are both representative and relevant to a query

Efficiency for Neutral Cell Summarization

Neutral Summarization: document *k*-means V.S. micro-cluster *k*-means





Effectiveness for Neutral Cell Summarization

Neutral Summarization: document *k*-means V.S. micro-cluster *k*-means

	$P=10$		$P=5$	
Method	Quality	Time	Quality	Time
Baseline	491.84	52.36	444.09	47.38
K80	445.59	0.57	408.02	0.50
K500	456.22	6.55	420.82	6.31
K1000	469.87	17.83	430.60	14.86
K80 + 1	463.88	3.53	422.35	2.77
K500 + 1	473.98	9.78	432.84	8.71
K1000 + 1	482.36	21.15	437.90	17.29
K80 + 2	468.11	6.46	427.01	4.98
K500 + 2	477.12	12.97	434.19	11.03
K1000 + 2	484.30	24.42	438.48	19.69

Quality: sum of similarities between a document vector and its cluster mean vector

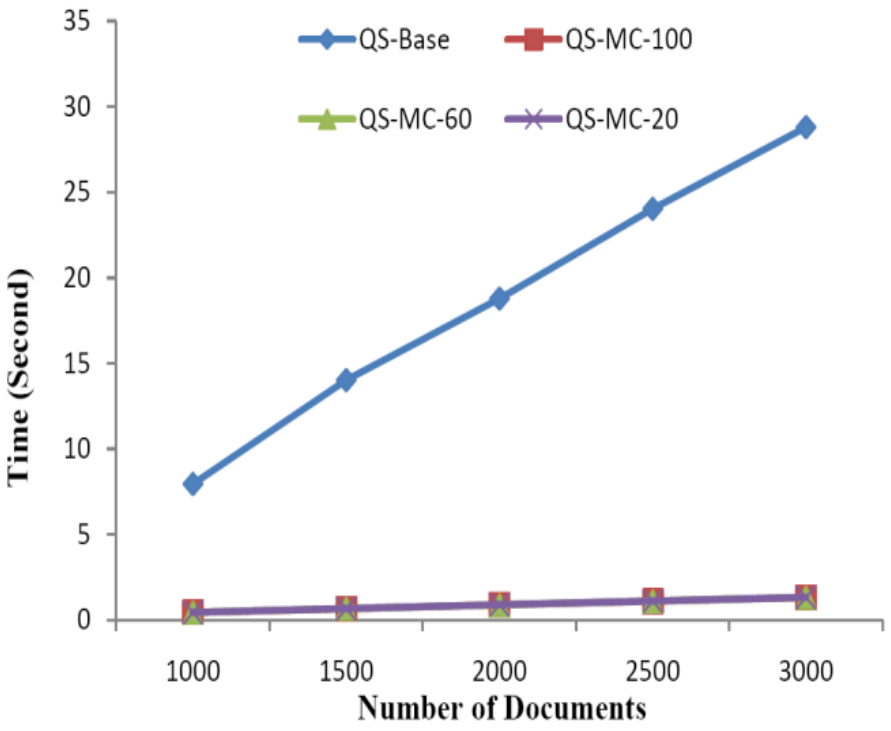
+1: after *k*-means on micro-clusters, do one more round on document unit

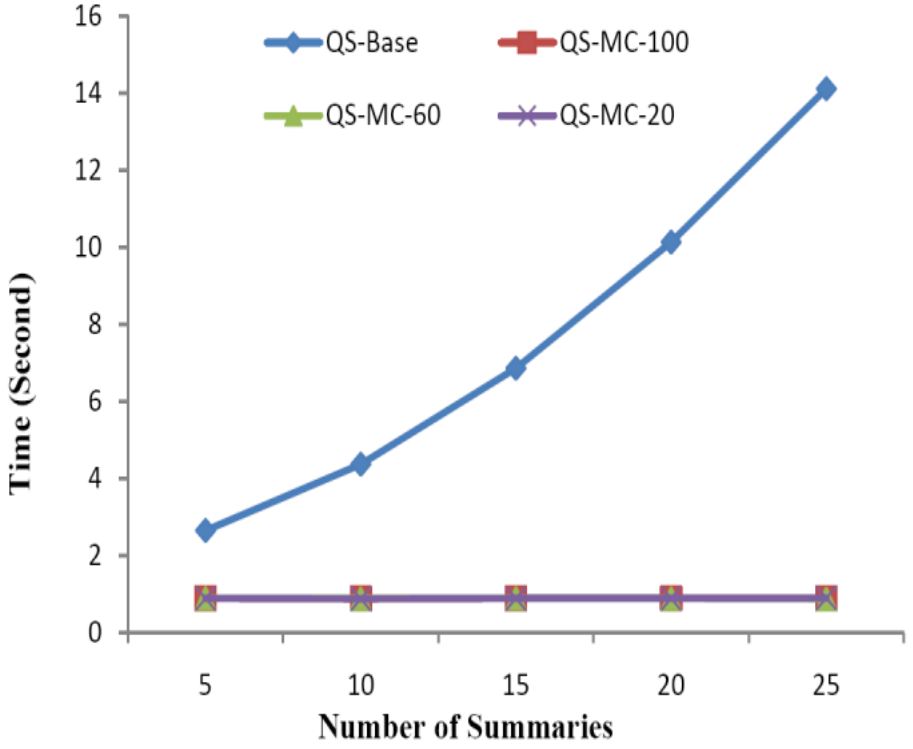
Efficiency for Query-Specific Summarization

Query-Specific Summarization:

MMR: $argmax_{D_i \in R \setminus S} \left[\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$

Micro-Cluster ranking: only the most relevant doc in each micro-cluster





Effectiveness for Query-Specific Summarization

Query-Specific Summarization:

MMR: $argmax_{D_i \in R \setminus S} \left[\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$

Micro-Cluster ranking: only the most relevant doc in each micro-cluster

λ	0.2	0.4	0.6	0.8	1
relevance	-283.27	-282.278	-282.278	-282.218	-282.21
coverage	258.28	257.919	257.919	259.707	259.707

K	10	20	30	50	100
relevance	-284.86	-284.0996	-282.8749	-282.6589	-282.5442
coverage	264.3687	269.9924	271.238	264.84	267.6433

Relevance: total similarity of selected documents with the query

Coverage: sum of similarity of each unselected document with its most similar selected document

Acknowledgement

- This work is supported in part by:
  - NASA NRA-NNH10ZDA001N
  - U.S. Air Force Offices of Scientific Research MURI award FA9550-08-1-0265
  - U.S. National Science Foundation grants IIS-0905215 and CNS-1028381

Thanks!