Sequential Clustering and Contextual Importance Measures for Incremental Update Summarization

TECHNISCHE UNIVERSITÄT DARMSTADT
AIPHES

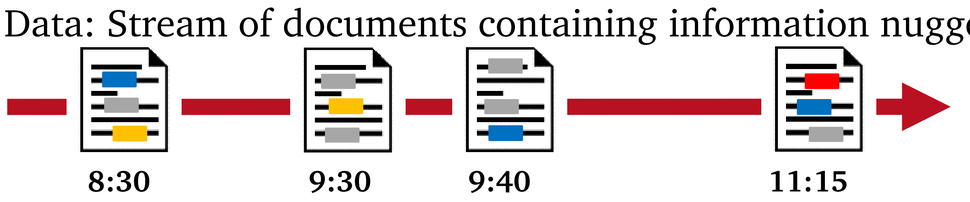
Markus Zopf, Eneldo Loza Mencía, Johannes Fürnkranz | Research Training Group AIPHES

SeqCluSum in a nutshell

- incremental update summarization
- combine benefits of pipeline and clustering approaches
- model importance and redundancy jointly with contextual cluster ranking
- best results on TREC-TS dataset

Incremental Update Summarization

Data: Stream of documents containing information nuggets

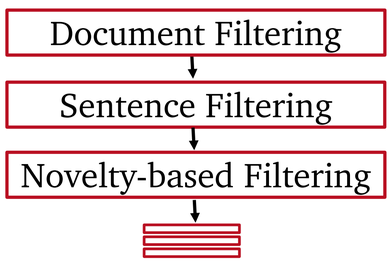


= On 16 January 2013,at around 0800 GMT
= A helicopter crashed into a construction crane in Vauxhall
= Captain Pete Barnes and a pedestrian, Matthew Wood, died

Task: Find new & important information nuggets as soon as possible
Application: developing events such as accidents, natural disasters, elections, etc.

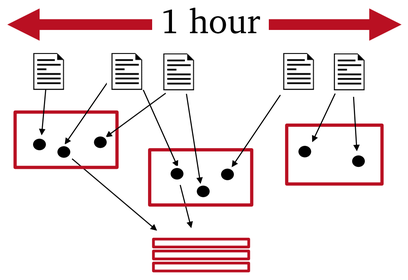
Clustering Information with Sequential Clustering

Pipeline



+ timely & - poor decisions

Clustering



- delayed & + good decisions

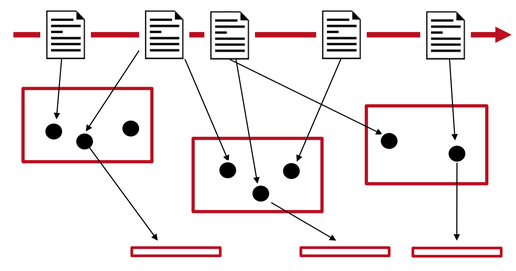
prior work used either pipeline or clustering

Idea: combine benefits of both approaches

1. be able to publish information at any time
2. use clustering to detect new and important information

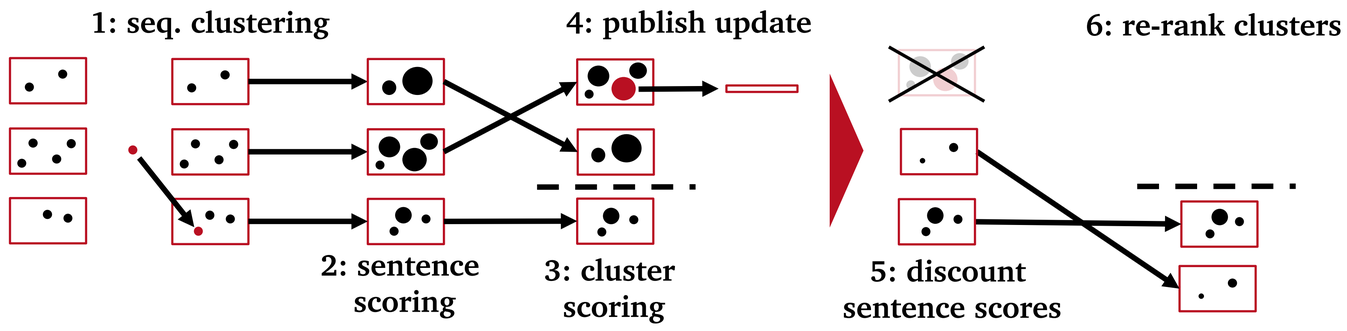
→ natural fit for IUS

Sequential Clustering



+ timely & + good decisions

Publishing Updates with Contextual Cluster Ranking

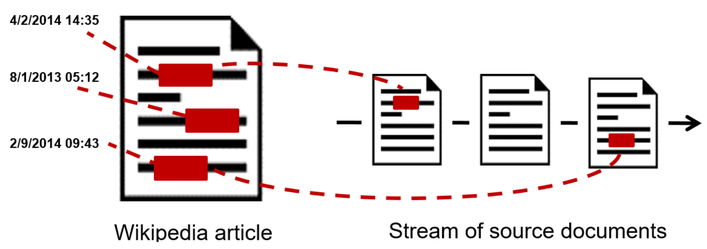


new document/sentence is available in the web:

1. **seq. clustering**: assign sentence to nearest cluster or create new cluster if distance is too big
2. **sentence scoring**: estimate sentence importance (high temporal TF-IDF → more important)
3. **cluster scoring**: cluster score based on sentence scores (bigger cluster → more important)
4. **publish update**: if a cluster is important enough → emit best sentence from cluster
5. **discount sentence scores** for redundancy avoidance
   (sentence similar to published sentence → larger discount)
6. **re-rank clusters** according to new sentence scores

Evaluation Dataset

- dataset from TREC-TS 2015 shared task
- based on webpage snapshots + Wikipedia revision history



- trace driven simulation to evaluate systems
- only small subset of sentences labeled due to high annotation effort
  → only lower bound evaluation ☹

Results

Type	System	$\mathcal{H}$	Timeliness	F_1(Prec, Rec)	Precision	Recall
Seq. clustering	SeqCluSum (lower bound)	**0.1526**	**0.8013**	**0.1842**	0.1485	0.2426
Pipeline	Raza, 2015; 1st@TREC-TS '15	0.0853	0.3983	0.1773	0.1840	0.1710
Clustering	McCreadie, 2015; 3rd@TREC-TS '15	0.0639	0.5335	0.1189	0.0667	0.5459
Pipeline	McCreadie, 2015	0.0508	0.6741	0.0758	0.0402	0.6590

$\mathcal{H}$: combination of Timeliness, Precision, and Recall
Timeliness: estimates timeliness of updates; 1=on time with ground truth, 0=outdated
Precision: fraction of important & novel system updates
Recall: fraction of information nuggets covered by system updates

Conclusions

→ SeqCluSum achieves **high $\mathcal{H}$** result derives from **both high timeliness** and **high F1 scores**

→ **combination** of **pipeline** and **clustering** works well for incremental update summarization

→ could be further improved with **prior knowledge** about information importance

This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1

www.aiphes.tu-darmstadt.de