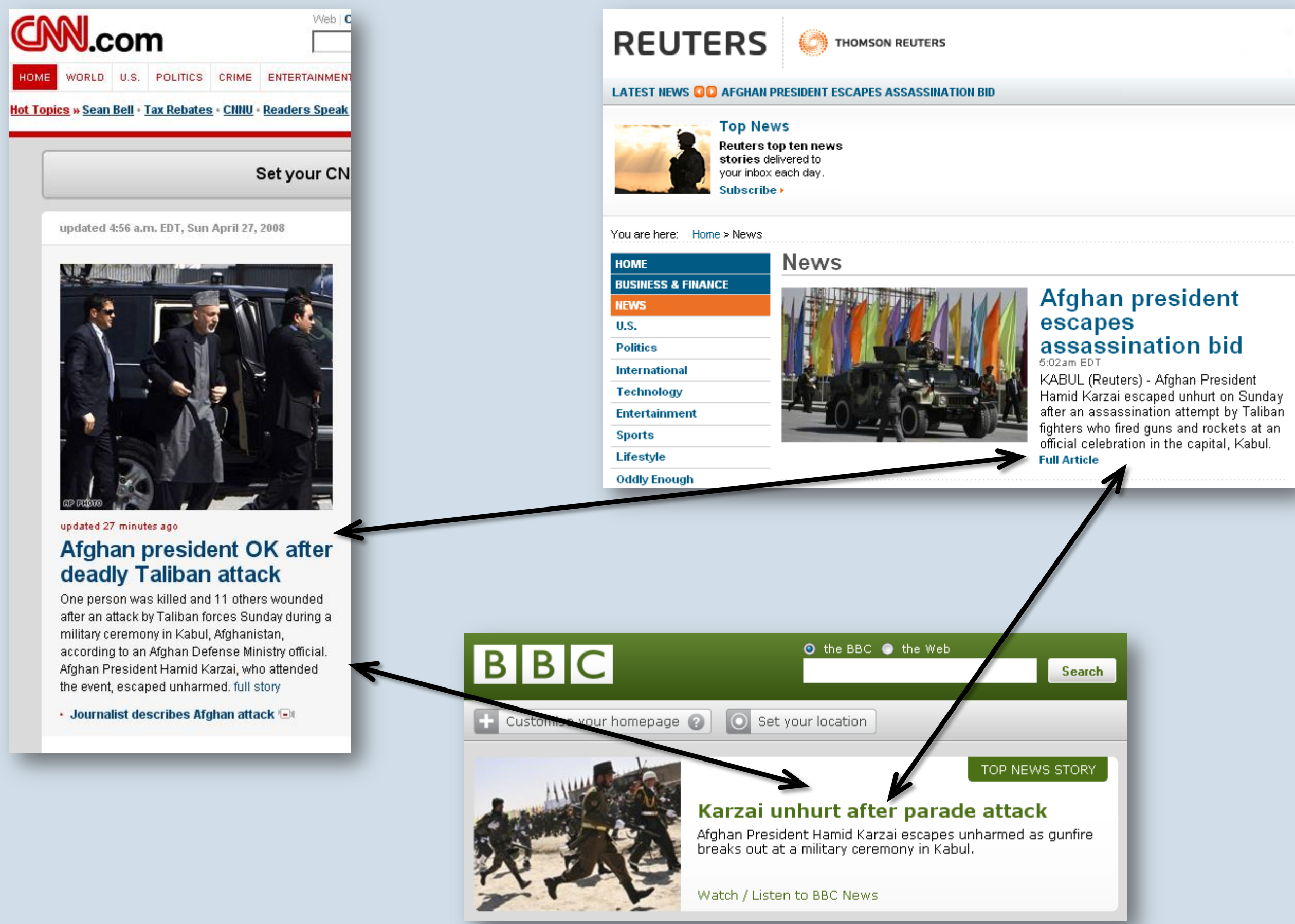


# Enriching Topic-Based Publish-Subscribe Systems with Related Content

Rubi Boim Tova Milo

## Up-To-Date Related News



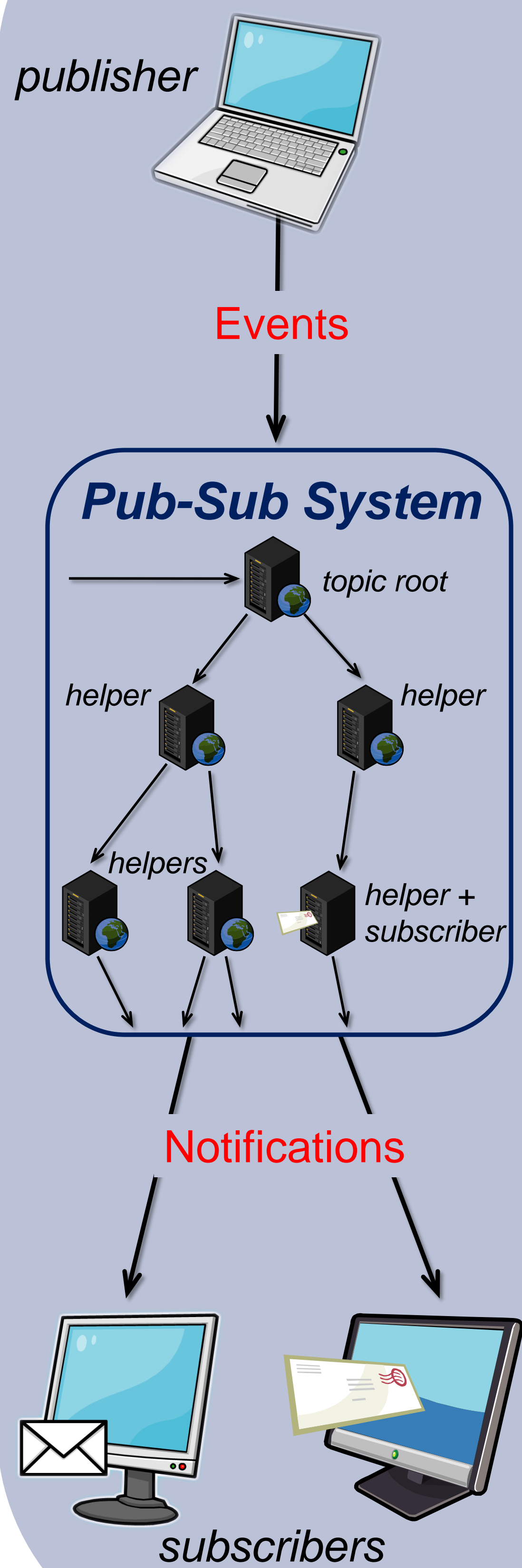
## How To Enrich?

- Impractical to query all topics (too many, unknown)
- Prior, cluster similar topics together
- Post, query only cluster topics

## Challenges

- **Adaptivity:** the content is constantly changing
- **Distribution:** P2P network – decentralized nature
- **Low Overhead:** minimize the load on each peer

## Topic-Based Pub-Sub



### Pub-Sub

- Publishers send messages to subscribers

### Topic-Based

- Subscription by topic name (CNN Sports, BBC News..)
- Common example – RSS Today, RSS is by Pull P2P Push → Scalability

### Problem

- Related (interesting) messages of other topics are not received

### Enriched Topic-Based

- Prior to notification, each message is enriched with related messages



## Dynamic Distributed Clustering

- A topic/message is represented by a profile
- $RM(t_1, t_2)$  – the likelihood of the topics to publish related messages
- A formula  $F$  estimates the clustering quality
- Local updates are performed when estimated to increase  $F$ 's value

$$F(T, C) = \sum_{t \in T} \left[ \underbrace{\frac{\sum_{t' \in T_t} RM(t, t')}{\sum_{t' \in T} RM(t, t')}}_{\text{Relevance ratio}} - w \underbrace{\frac{\sum_{c \in C_t} \sum_{t' \in c} (1 - RM(t, t'))}{\sum_{t' \in T} (1 - RM(t', t))}}_{\text{Irrelevance ratio}} \right]$$

$T$ =all topics,  $C$ =all clusters,  $C_t$ = $t$ 's clusters,  $T_t$ =all topics in  $C_t$

## Topic Profile

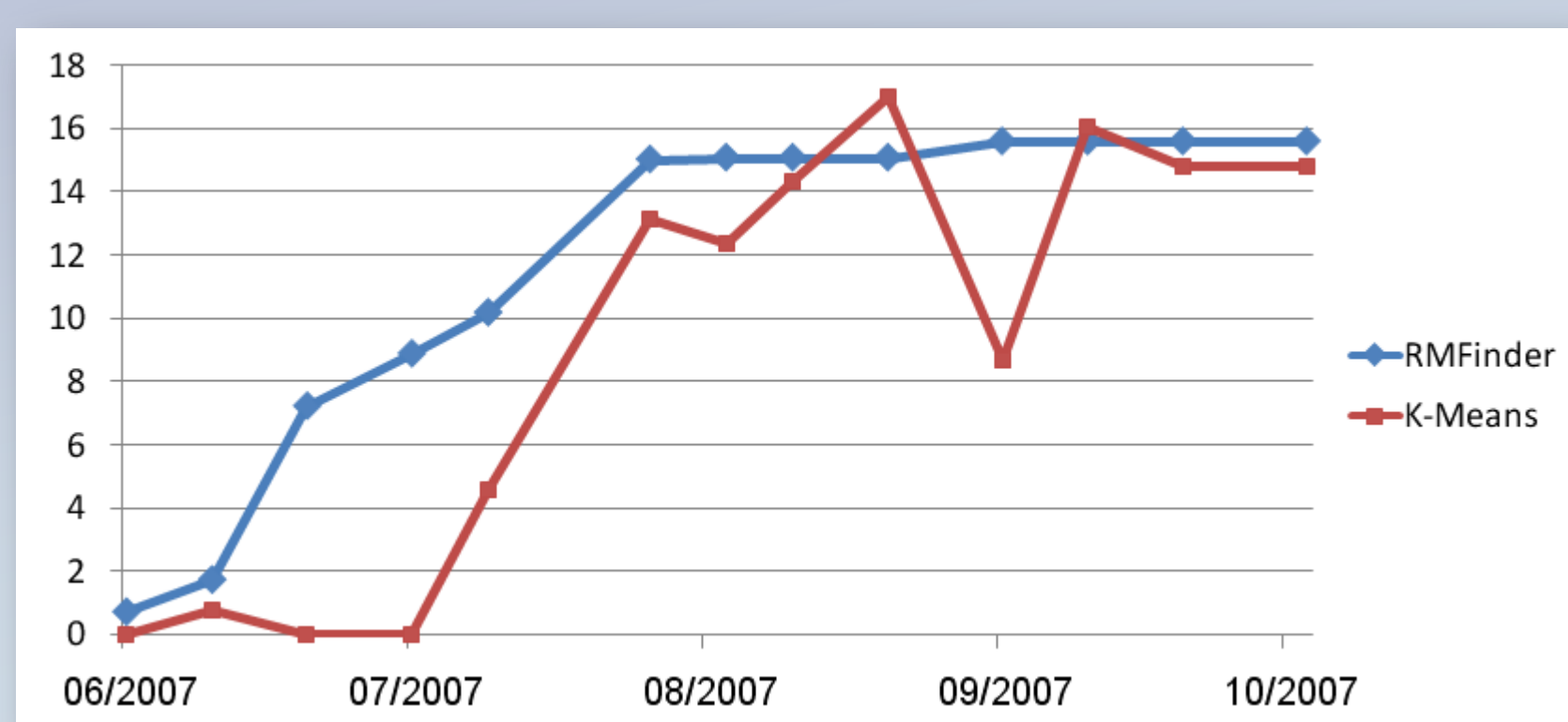
A set of features  $S$  (of size  $k$ ) constructed by:

- Features Extraction
- Sliding Window
- Filtering

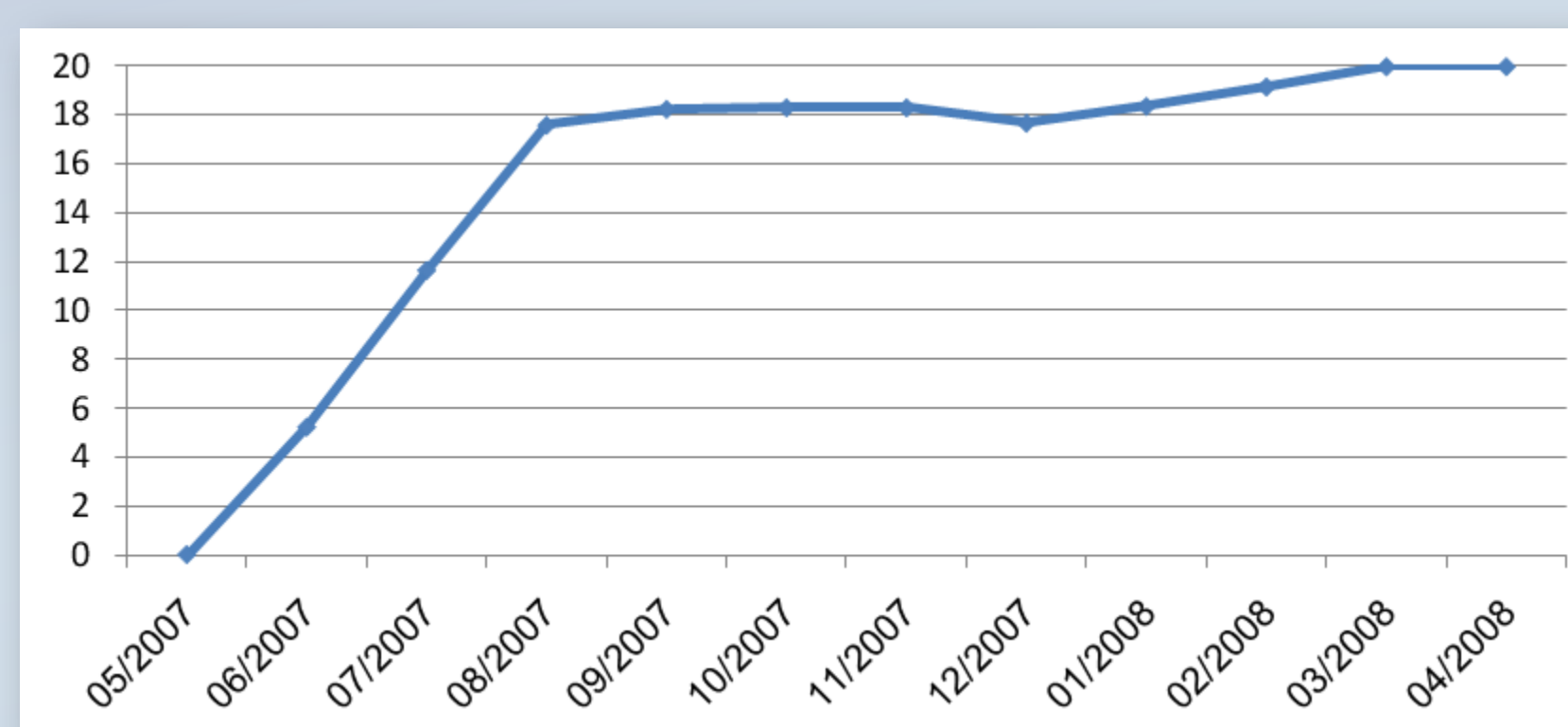
## Finding Related Messages

- Message profile is similarly extracted
- Compared to messages profile within the cluster

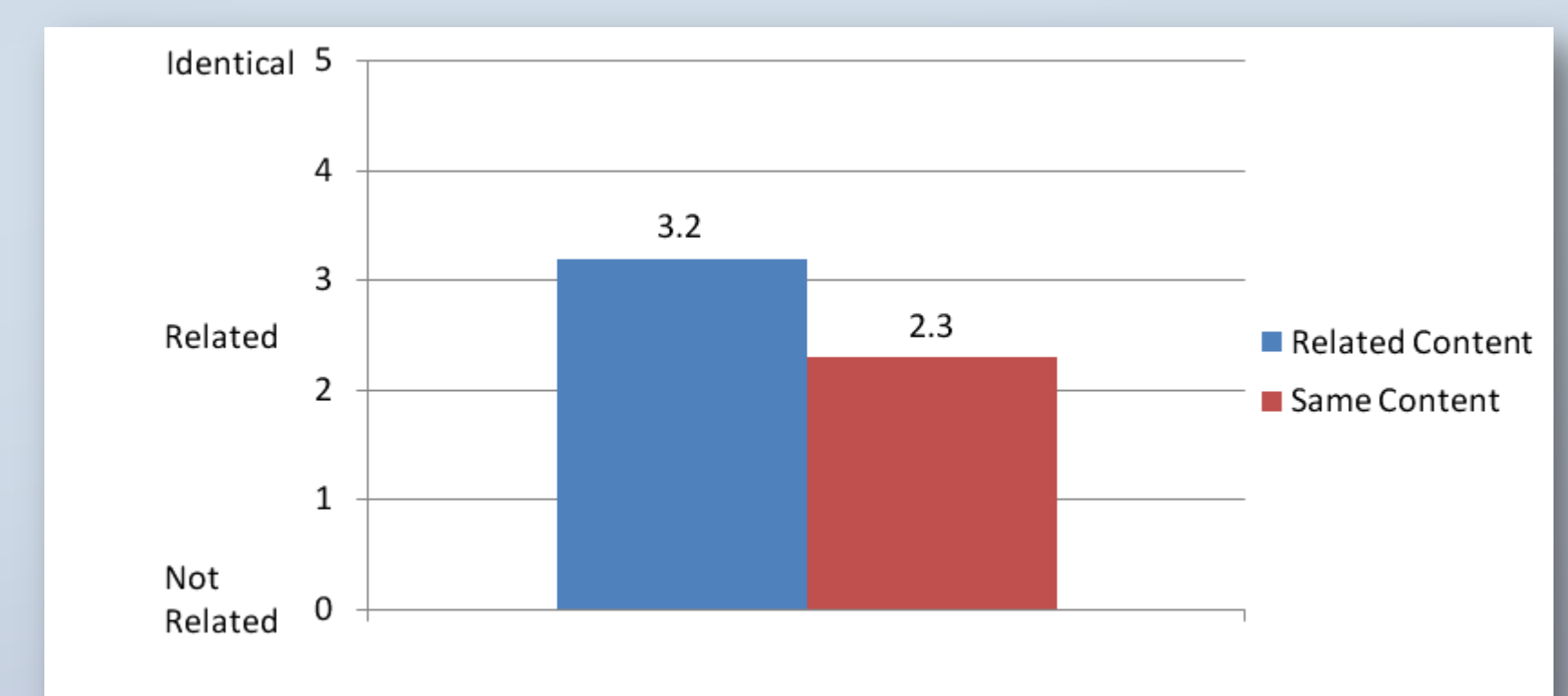
## Experimental Results



RMFinder vs. K-Means



Adaptive clustering



Survey results