Enriching Topic-Based Publish-Subscribe Systems with Related Content

Rubi Boim  Tova Milo

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

Dynamic Distributed Clustering
- A topic/message is represented by a profile
- RM(t1,t2) – 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

$$F(T',C) = \sum_{i\in C'} \sum_{t\in T'} RM(t,t') - \frac{W}{\sum_{C_{t'}} \sum_{t\in T'} (1 - RM(t,t'))}$$

Relevance ratio Irrelevance ratio
$T$=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
- Massage profile is similarly extracted
- Compared to messages profile within the cluster

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

Experimental Results
RMFinder vs. K-Means
Adaptive clustering
Survey results