

Εντοπισμός και Παρακολούθηση των Χαρακτηριστικών Διαφόρων Προϊόντων και της Πολικότητας των Απόψεων σχετικά με αυτά σε ένα Ρεύμα Απόψεων.

Max Zimmermann¹, Eirini Ntoutsi²,
Myra Spiliopoulou¹

*Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων
Ιούλιος 2014*

1



FAKULTÄT FÜR
INFORMATIK


2








Institute for Informatics, Ludwig-
Maximilians-Universität (LMU)
München, Germany

Opinionated streams

- Opinions from TripAdvisor on Vienna Marriott Hotel

2/5/2014: Great hotel, very nice **rooms**, perfect **location**, very nice **staffs** €  ot for a mid-aged female receptionist who tried to charge me extra for wifi fees when checking out. It was waived at the desk when I checked-in. And she started treating me with an attitude after she found out that I got a great deal through priceline.com.

26/1/2014: Spent a long weekend here. **Rooms** clean and functional without being spectacular and a nice pool etc. **Staff** in pool weren't Good and I found them actually quite r  . Executive **lounge** was ok and not busy but **selection of wine and beer** wasn't great. The **reception** has many shops and a bar at the end which kind of r  s it feel like a shopping centre. Overall great for business travel but not sure id come again for leisure.

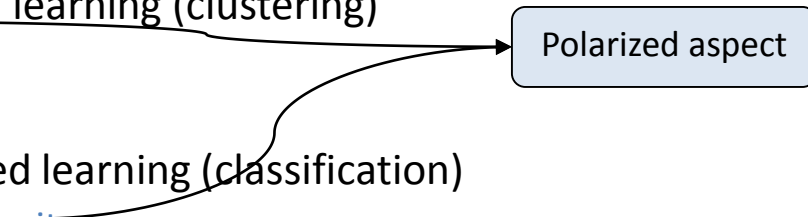
7/5/2013: The Vienna Marriott has all you expect; no frills, but solid service and they get all the basic stuff done  ight.
It's in a fine **location**, maybe 10 minute walk from the major city attractions while being in a quiet **area**. **Breakfast** buffet exceptional!  good **fitness center**. Very helpful and happy **staff**. 
Lobby **lounge** just okay. Not a good **wine selection** and the Sinatra-like **singer** adds nothing. Maybe just a little more expensive than it should be, too.

Outline

- Opinionated streams
- Opinion stream mining & the OPINSTREAM framework
- Extracting an initial opinionated hierarchy of product (sub)features
- Online (sub)feature hierarchy maintenance
- Online opinion classifier maintenance
- Experiments
- Summary

Opinion stream mining

- In a static setting, i.e., given is a collection of opinionated documents:
 - How to extract the different, *ad hoc aspects* discussed in the collection?
 - How to associate a dominant polarity to each aspect?
- When the set becomes a stream
 - How to extract the ad hoc aspects over time?
 - How to update their sentiment over time?
- The OPINSTREAM framework
 - Feature discovery → adaptive unsupervised learning (clustering)
 - Features are defined as **cluster centroids**
 - Features evolve over time
 - Polarity learning → adaptive semi-supervised learning (classification)
 - Majority voting within the cluster for the **polarity**
 - Learning under concept drift
 - Learn with a small amount of labeled reviews



Polarized aspect

Basic concepts

- Reviews arrive at distinct timepoints $t_0, t_1, \dots, t_i, \dots$ in *batches* of fixed size.
- We don't receive class labels from the stream; the only sentiment-related information we assume is an initial seed set S of labeled reviews.
- Reviews are subject to **ageing** according to the exponential ageing function
 - Recent reviews are more important for learning
 - Gradual decrease of the weight of a review over time
 - Ageing affects both clustering (feature extraction) and classification (sentiment learning)
- The **importance of a review** r with respect to a set of reviews R , is defined as the number of reviews in R that have r among their kNN, also considering age.
 - For the kNN, cosine similarity is employed

$$importance(r, R) = \sum_{r_i \in R} age(r_i) \cdot isRevNeighbour(r, r_i, R)$$

- Important reviews R_β : those exceeding the review importance threshold β .

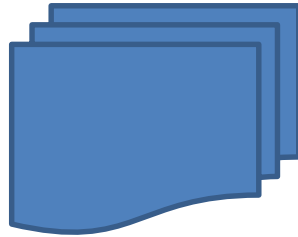
Outline

- Opinionated streams
- Opinion stream mining & the OPINSTREAM framework
- Extracting an initial opinionated hierarchy of product (sub)features
- Online (sub)feature hierarchy maintenance
- Online opinion classifier maintenance
- Experiments
- Summary

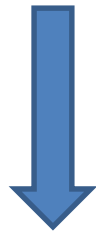
Extracting an initial hierarchy of polarized features

Global/ 1st level clusters

Initial seed set S (labeled)

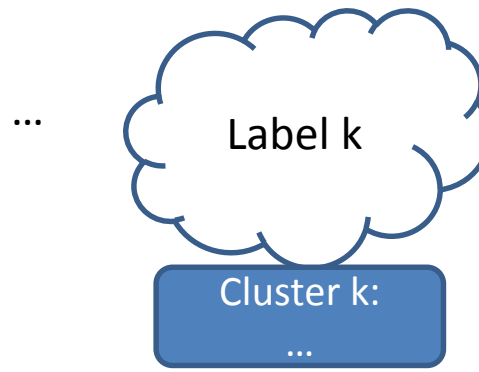


Extract clusters



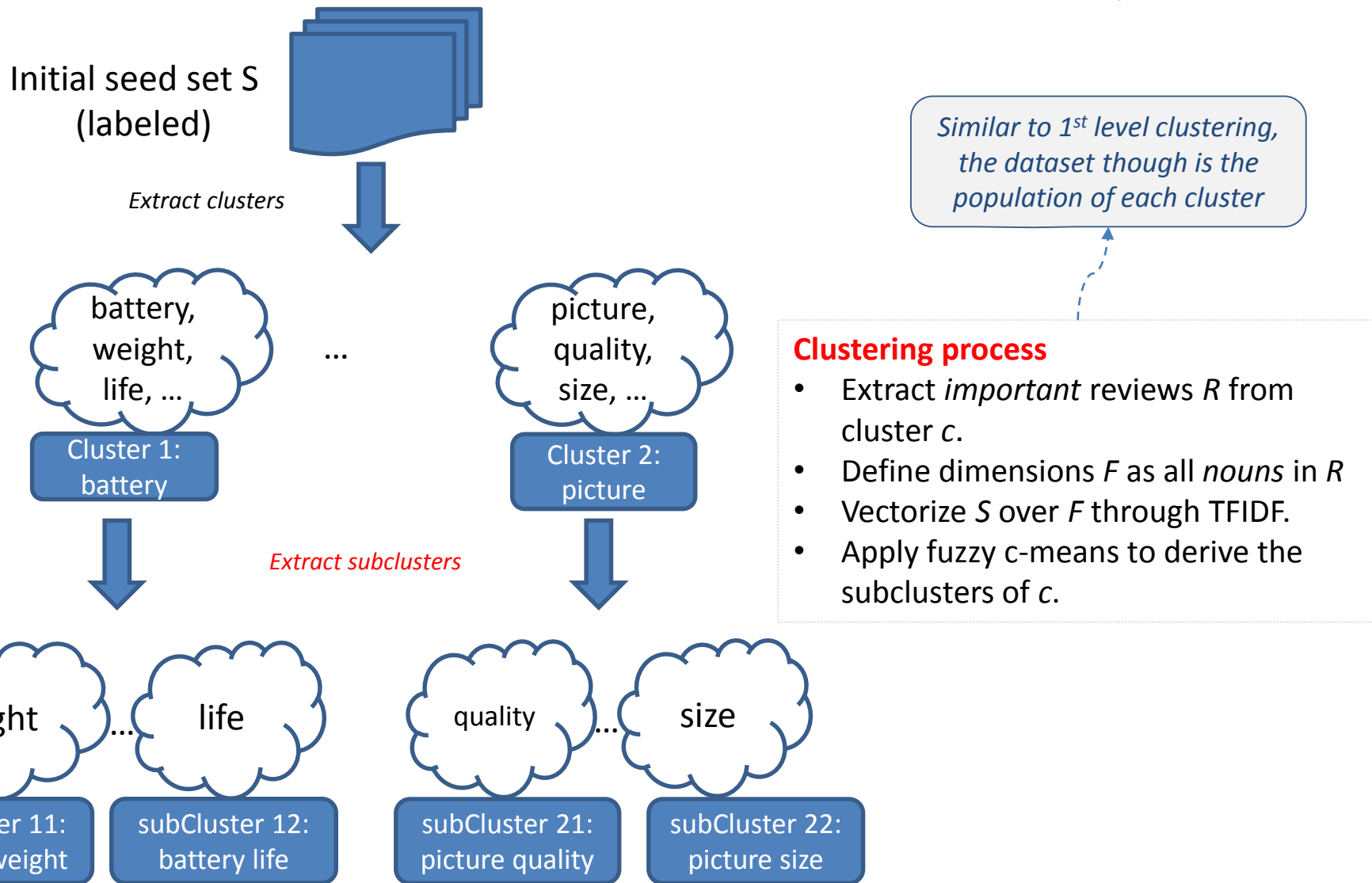
Clustering process

- Extract *important* reviews R from S.
- Define dimensions F as all *nouns* in R
- Vectorize S over F through TFIDF.
- Apply fuzzy c-means to derive the 1st level clusters.



Extracting an initial hierarchy of polarized features

Local/ 2nd level clusters



Polarized features & topic-specific classifiers

Definition: Polarized feature

The polarized feature represented by a cluster c defined in a feature space F_R consists of:

- The centroid vector $\langle w_1, \dots, w_{|F_R|} \rangle$, w_i is the avg TF-IDF of word i in c .
- The polarity label $c^{polarity}$, is the majority class label of reviews in c .

Topic-specific classifiers

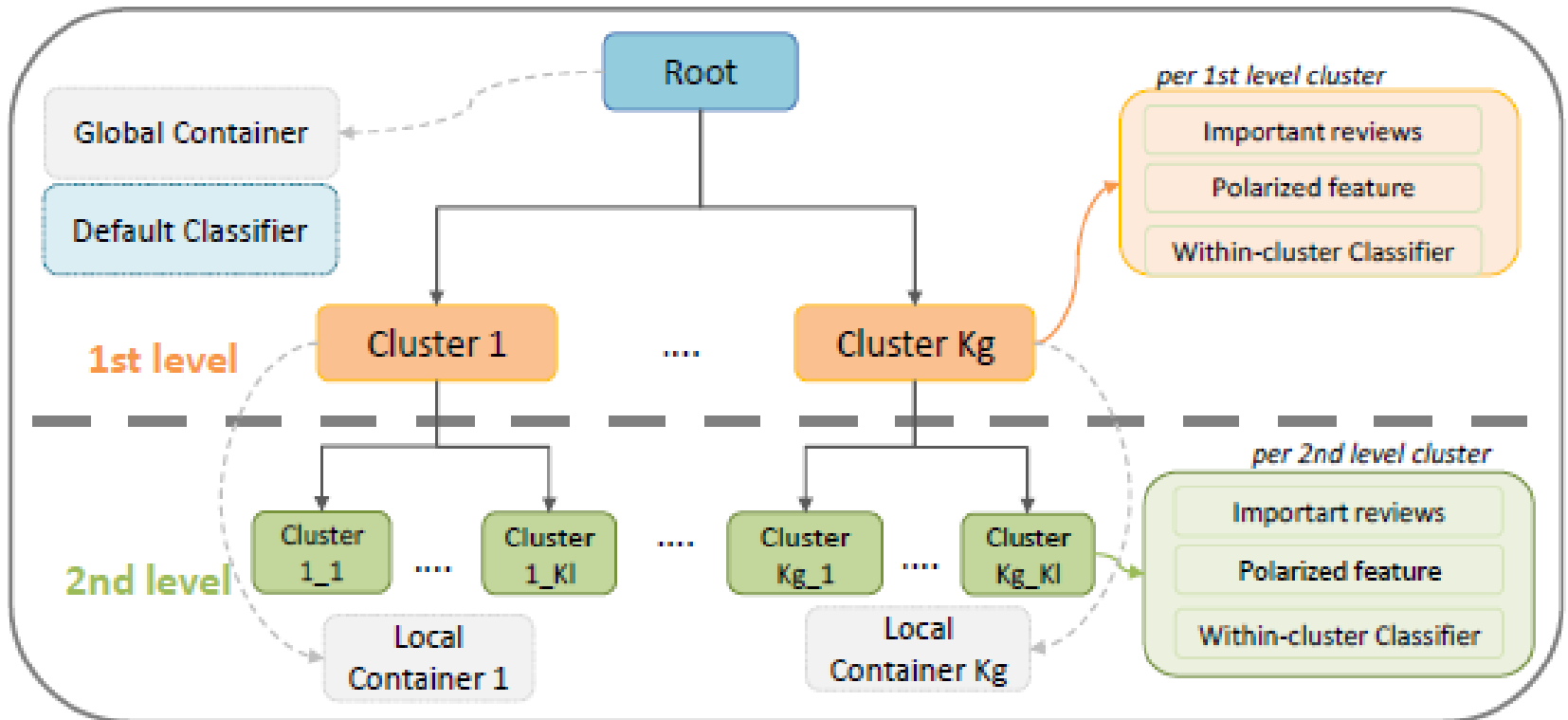
- Within each (sub)cluster, a topic-specific classifier is trained based on the corresponding instances from the initial (labeled) seed set S .
- We focus on adverbs, adjectives as sentimental words.
- Multinomial Naïve Bayes (MNB) sentiment learner
 - Laplace correction factor for unknown words

$$\hat{P}(w_i|c) = \frac{N_{ic} + 1}{\sum_{j=1}^{|V|} N_{jc} + |V|}$$

Laplace correction for unknown words

$$N_{ic} = \sum_{d=1}^{|S|} f_{ic}^d$$

The polarized (sub)feature hierarchy



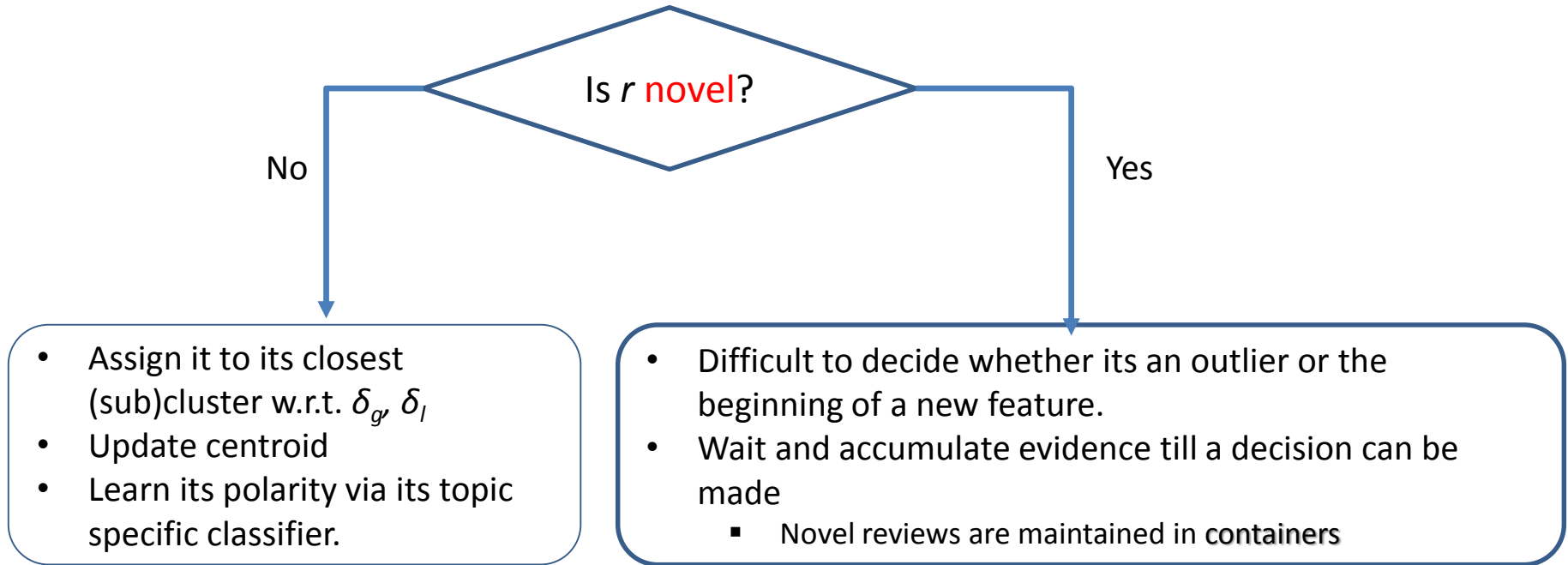
Here is how the hierarchy looks like (ignore the containers for the moment)

Outline

- Opinionated streams
- Opinion stream mining & the OPINSTREAM framework
- Extracting an initial opinionated hierarchy of product (sub)features
- Online (sub)feature hierarchy maintenance
- Online opinion classifier maintenance
- Experiments
- Summary

Hierarchy update

- So far, we discussed the initialization of the polarized hierarchy.
- How do we place a new coming review r (unlabeled) in the hierarchy?



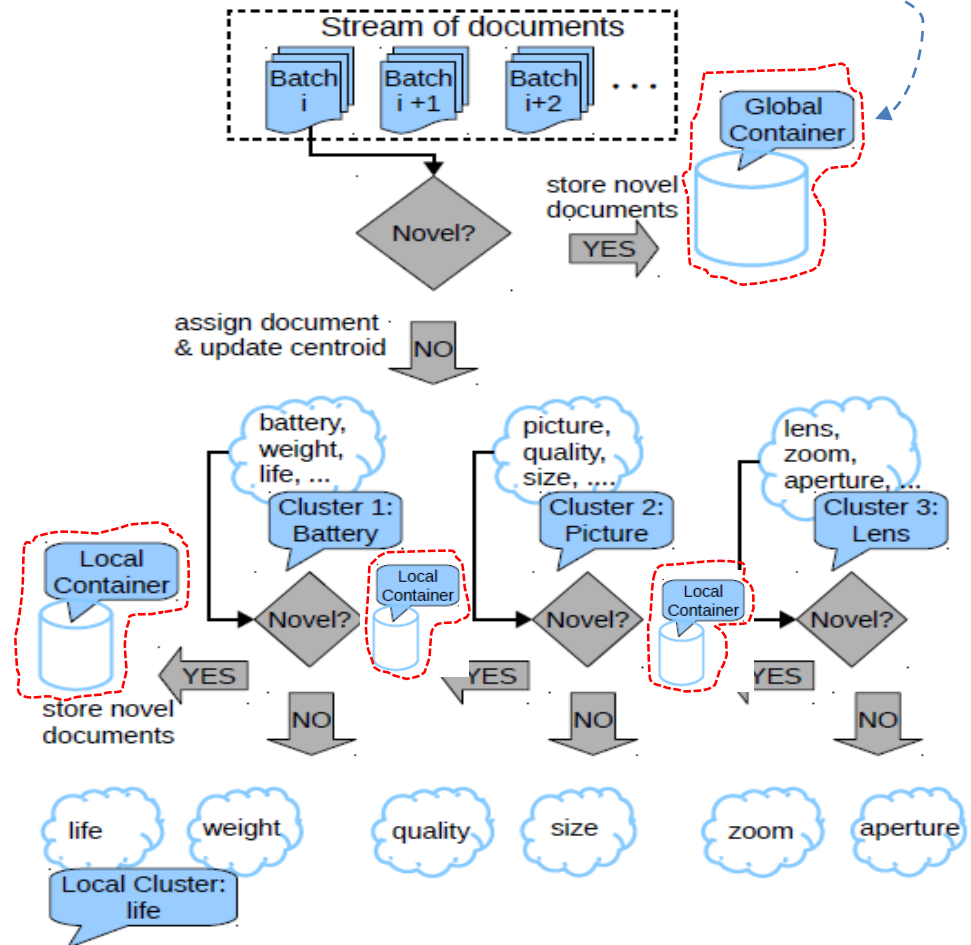
Definition: Review novelty

A review r is novel w.r.t. a set of clusters Θ if its cosine similarity to the closest cluster centroid in Θ is less than δ ($0 \leq \delta \leq 1$).

Global and local containers for novelty accumulation

Global container:
accumulates reviews that are novel w.r.t. global clusters

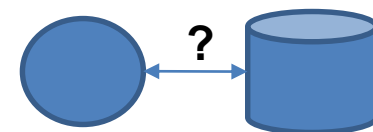
Local container (1 per global cluster):
accumulates reviews that are close to its cluster centroid but far away from all centroids of its subclusters



Adapting the feature hierarchy based on novelty

- Adapt the hierarchy when enough novelty has been accumulated
 - **Questions 1:** How to define adequate novelty?
 - **Questions 2:** How to incorporate novel reviews in the hierarchy?
- **Always recluster approach:** if the size of a container exceeds a threshold, reorganize from scratch its corresponding local/global clusters (*reclustering*)
 - Results in a lot of local/global reclusterings
 - Lot of effort for the end user to re-comprehend the hierarchy
- **Merge or recluster approach:** instead of a drastic change like reclustering, try first to merge containers to their corresponding clusters.
 - Rationale: Clusters and containers might “start moving towards each other” due to ageing.
 - Merge option 1: Merge b and c w.r.t. the feature space of c
 - Merge option 2: Merge b and c w.r.t. a new feature space derived from $b \cup c$

Such drastic changes should be avoided, unless necessary.



When merging is beneficial?

- Check model quality before and after the merge
 - Cluster description length as quality indicator
- The **cluster description length (CDL)** of a cluster c defined over dimensions D_c :

$$CDL(c, D_c) = - \sum_{r \in c} \log_2 P(r|c, D_c)$$

- **Merge strategy 1:** Do we gain in quality if we merge c with b , while retaining the feature space of c ?

$$\text{Conditional_I} : CDL(c|D_c) + CDL(b|D_b) - CDL(c \cup b|D_c) > 0$$

- **Merge strategy 2:** Do we gain in quality if we merge c with b , using a new set of dimensions derived from b and c ?

$$\text{Conditional_II} : CDL(c|D_c) + CDL(b|D_b) - CDL(c \cup b|D_{c \cup b}) > 0$$

This results actually in a new cluster.

To many rebuilds are annoying for the end user

- We should keep the number of rebuilds low to minimize the effort of the end user
 - The mental effort is modeled by fatigue.

Definition: Fatigue

Let $M(t)$ be the hierarchy at t and n the number of reviews contained in its clusters. Fatigue is defined as the percentage of reviews involved in rebuilt clusters:

$$fatigue = \frac{\sum_{c \in M(t) \setminus M(t-1)} |c|}{n}$$

← Sets of clusters involved in rebuilds

- At the end of each batch and for each global cluster
 - Try Merge Strategy I
 - If not satisfied, try Merge Strategy II
- Compute fatigue (based on all global clusters for which Merge Strategy II is true)
 - If fatigue < γ , go on with local reclusterings
 - otherwise, rebuild the whole hierarchy from scratch.

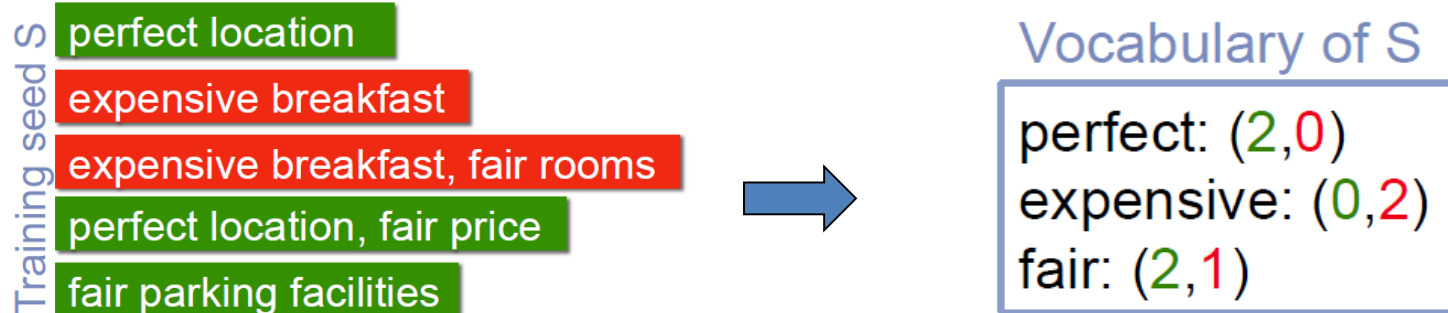
Outline

- Opinionated streams
- Opinion stream mining & the OPINSTREAM framework
- Extracting an initial opinionated hierarchy of product (sub)features
- Online (sub)feature hierarchy maintenance
- Online opinion classifier maintenance
- Experiments
- Summary

A polarity classifier over an opinionated stream

😊 or 😞 ?

- Static Multinomial Naïve Bayes



- Probability of class c for word w_i , estimation based on seed set S:

$$P(c|d) = \frac{P(c) \prod_{i=1}^{|d|} P(w_i|c)^{f_i^d}}{P(d)}$$

Laplace correction for unknown words

$$\hat{P}(w_i|c) = \frac{N_{ic} + 1}{\sum_{j=1}^{|V|} N_{jc} + |V|}$$

$$N_{ic} = \sum_{d=1}^{|S|} f_{ic}^d$$

Taking the ageing of the words into account

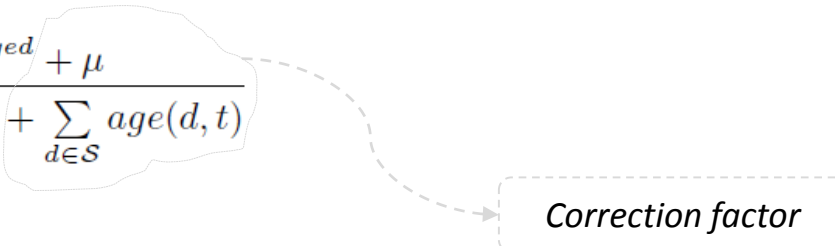
- **Backward adaptation version**

- Old reviews have gradually less effect on the classifier

- Effect of age in the word counts

$$N_{ic}^{aged} = \sum_{d=1}^{|\mathcal{S}|} f_{ic}^d * age(d, t)$$

- Effect of age in the probability estimates

$$\hat{P}(w_i|c)_{aged} = \frac{N_{ic}^{aged} + \mu}{\sum_{j=1}^{|\mathcal{V}|} N_{jc}^{aged} + \sum_{d \in \mathcal{S}} age(d, t)}$$


Correction factor

$$\mu = e^{-\lambda(\text{now}-t_0)}$$

Expanding the initial seed by useful documents

○ Forward adaptation version

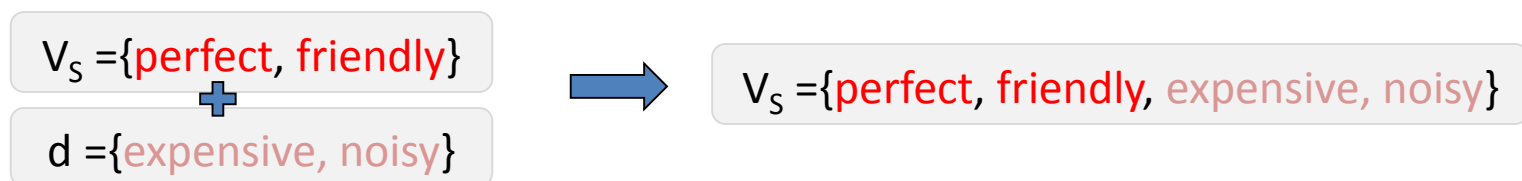
- Labels are expensive, can we extend the initial seed set S ?
- Yes, by adding *useful* reviews to S

○ Definition: Usefulness of a review

Let d be a new review, to which $\Delta(S)$ assigns the label c . The usefulness of d is:

$$Usefulness(d) = \sum_{w_i \in d} H(\mathcal{S}, w_i) - H(\mathcal{S} \cup d, w_i)$$

- A document is useful if its usefulness exceeds a threshold α in $(-1,0]$
 - ~ 0 values promote *smooth adaptation* (d should “agree” with existing model)
 - ~ -1 values promote *diversity*

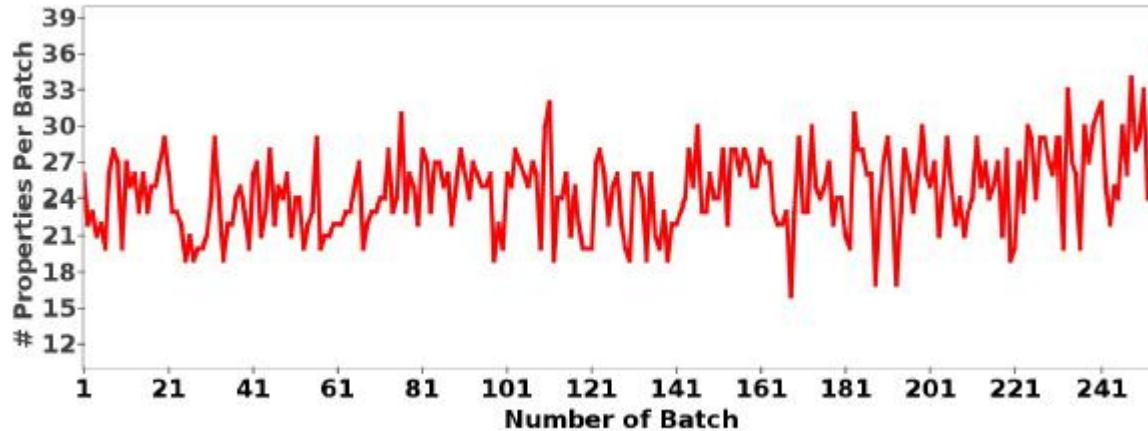


Outline

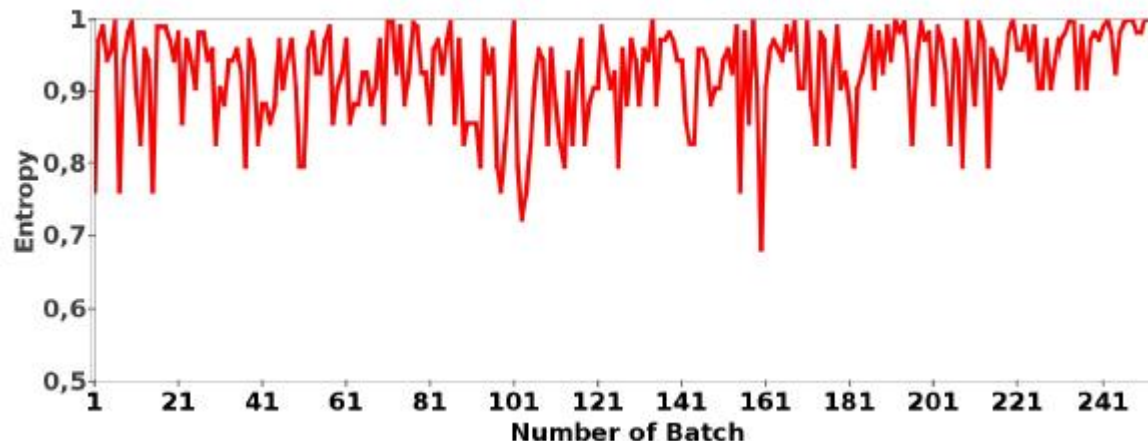
- Opinionated streams
- The OPINSTREAM framework
- Extracting an initial opinionated hierarchy of product (sub)features
- Online hierarchy maintenance
- Online opinion classifier maintenance
- Experiments
- Summary

Experiments

- Yu et al dataset: 12,825 reviews on 327 product features



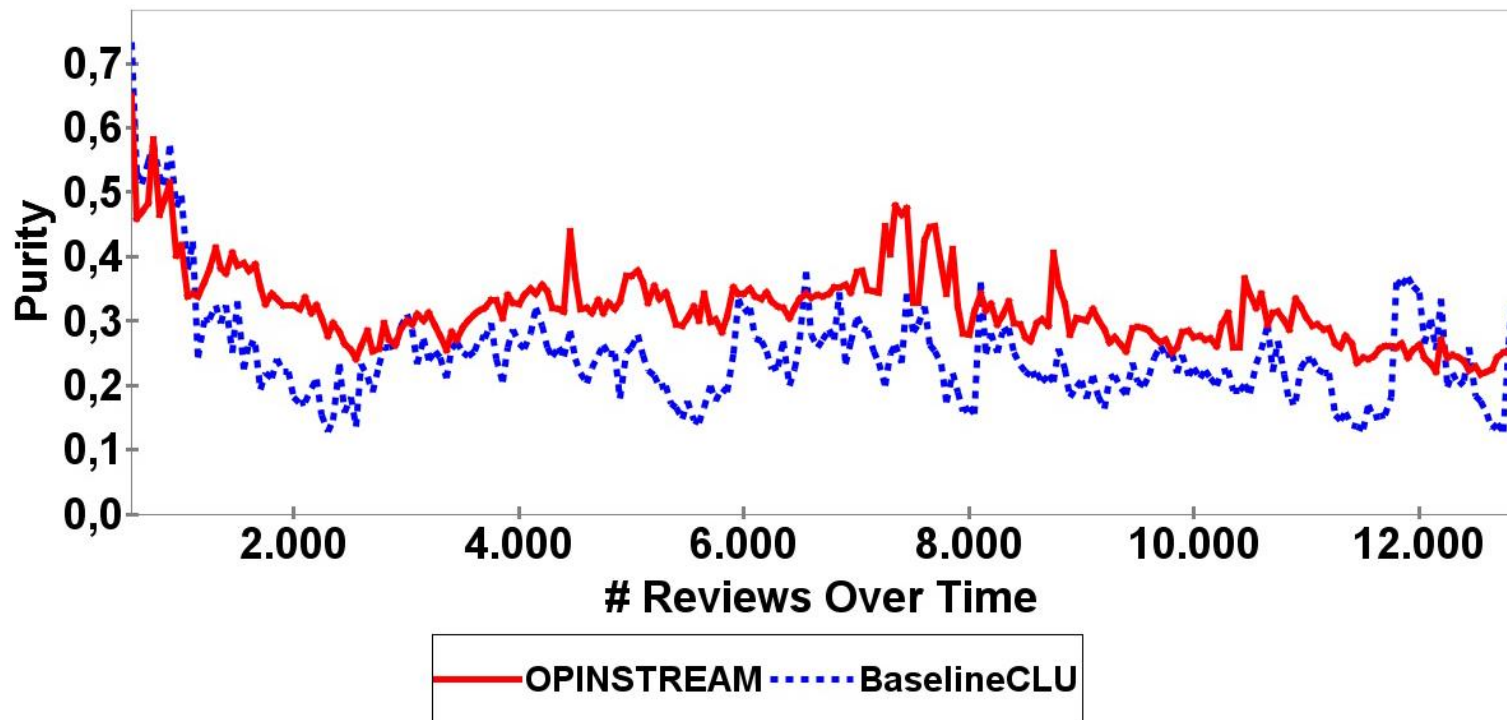
features strongly from one batch to the next making adaptation challenging



High entropy values indicating that there is no clear sentiment per batch

Evaluation of the feature extraction

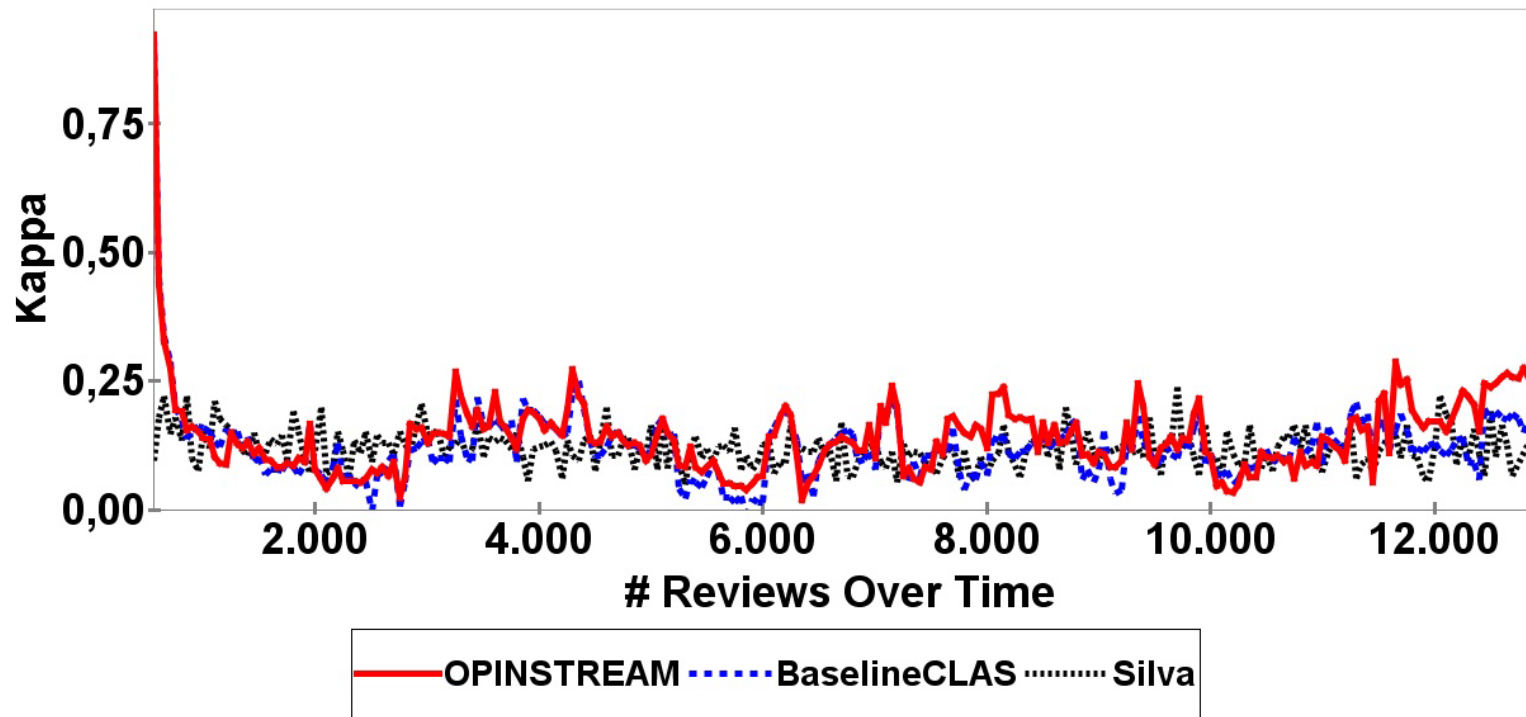
- $K_g=9, \delta_g=0.1, \sigma_g=500; K_l=9, \delta_l=0.2, \sigma_l=150;$
- *BaselineClu: is the container size-based approach*



→ *Smooth adaptation leads to clusters of better quality*

Evaluation of the polarity learner

- $K_g=6$, $\delta_g=0.2$, $\sigma_g=500$; $K_l=6$, $\delta_l=0.3$, $\sigma_l=150$;
- *BaselineClas*: non-adaptive semi-supervised approach
- *Silva*: the classification rule learner by Silva et al.



➔ Performance oscillates for all methods, OPINSTREAM and BaselineCLAS perform similarly and better than Silva.

Summary

- We presented OPINSTREAM, a framework for the extraction of *implicit features* and their *associated polarities* from a stream of reviews.
- Feature extraction → unsupervised task (stream clustering)
 - Accumulate novelty in containers
 - Cluster description length as an indicator of cluster's quality
 - Fatigue for quantifying the mental effort caused by rebuilds
- Sentiment learning -> supervised/semi-supervised task (stream classification)
 - Learn with only an initial seed of labeled documents
 - Backward adaptation to count for ageing
 - Forward adaptation to expand the seed set by incorporating useful documents
- Ongoing/future work
 - Simplifying the framework, parameter tuning, new datasets/ domains
 - Semi-supervised sentiment learning
 - Evaluation of backward propagation in full supervised learning
 - Evaluation of polarity and aspect learning simultaneously

Σας ευχαριστώ πολύ!!!

Ερωτήσεις?