Summarizing Cluster Evolution in Dynamic Environments

ICCSA 2011, Santander

Eirini Ntoutsi[1,2], Myra Spiliopoulou[3], Yannis Theodoridis[1]

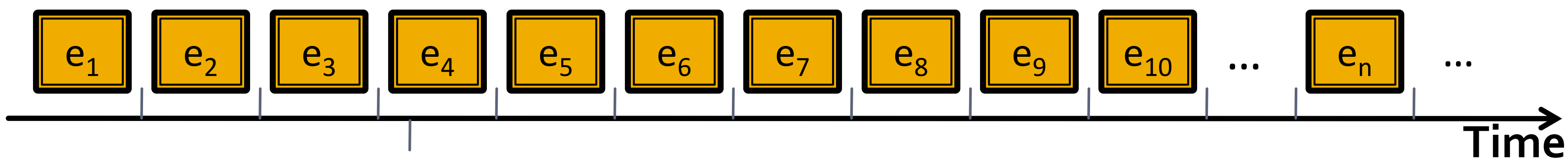
[1] Institute for Informatics, LMU, Germany
[2] Dept of Informatics, Uni of Piraeus, Greece
3 School of Computer Science, Uni of Magdeburg, Germany

Outline

- Motivation
- The evolution graph
- The FINGERPRINT of evolution
- Experiments
- Conclusions and outlook

Dynamic data/ data streams

- More and more data are produced nowadays:
  - Telcos, Banks, Health care systems, Retail industry, WWW …
- Modern data are dynamic
  - A special category is data streams: possible infinite sequence of elements arriving at a rapid rate



- Data Mining over such kind of data is even more challenging:
  - Huge amounts of data → only a small amount can be stored in memory
  - Arrival at a rapid rate → need for fast response time
  - The generative distribution of the stream might change over time → adapt and report on changes

Clustering over dynamic/stream data

- Traditionally clustering is applied over static data
- Lately there are approaches that deal with modern data

Dynamic Clustering Approaches

Adaptation

Adapt clusters to reflect current state of the population.

- CluStream [Aggrawal et al, VLDB’03]
- DenStream [Cao et al, SDM’06]
- Dstream [Chen and Tu, KDD’07]

Tracing & Understanding

Trace changes and reason on them so as to gain insights on the population.

- FOCUS [Ganti et al, PODS’99]
- PANDA [Bartolini et al, KDE’09]
- MONIC [Spiliopoulou et al, KDD’06]

Our contribution

- Although there exist methods for:
  - online cluster adaptation as the stream proceeds and
  - change detection between clusterings extracted at different time points
- they do not deal with the *efficient long-term maintenance of the changes over an infinite stream of data*

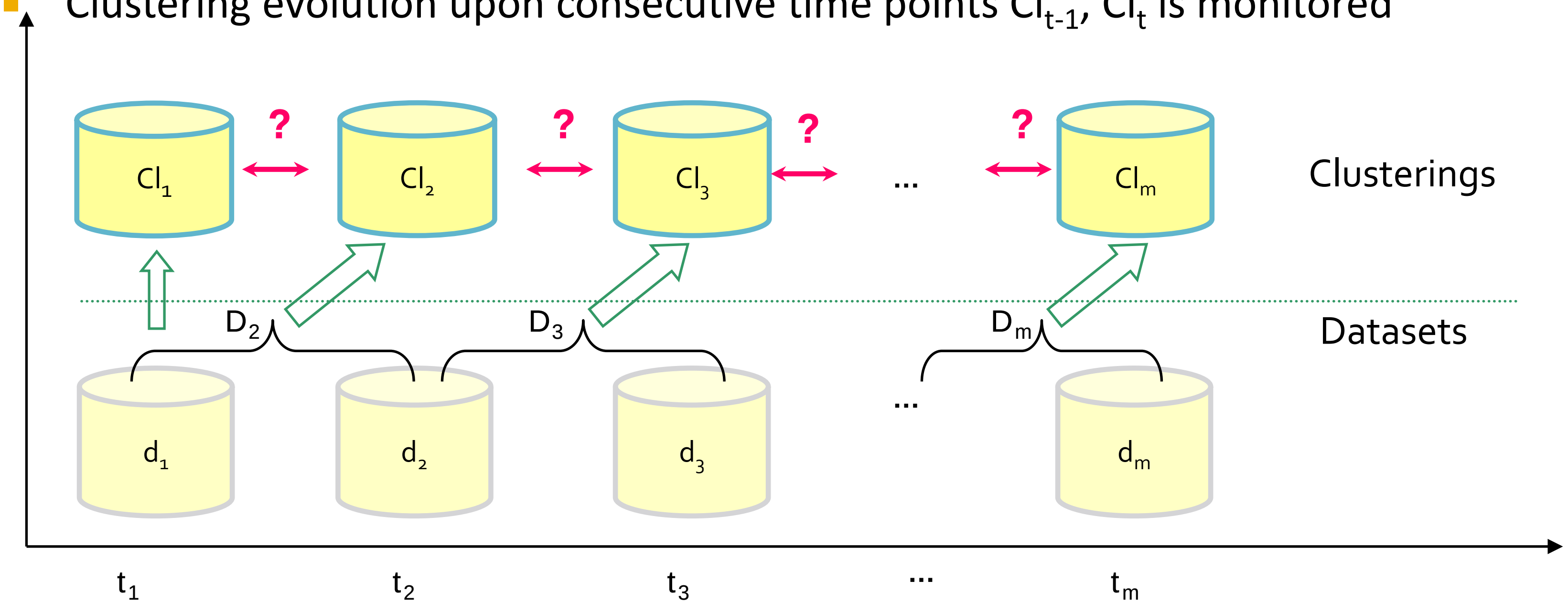
- To this end, we propose:
  i. A graph representation of cluster changes/ transitions, and
  ii. methods for condensing this graph into a FINGERPRINT
- The FINGERPRINT is a summary structure where similar clusters are efficiently summarized, subject to an information loss function.

Outline

- Motivation
- The evolution graph
- The FINGERPRINT of evolution
- Experiments
- Conclusions and outlook

Problem settings

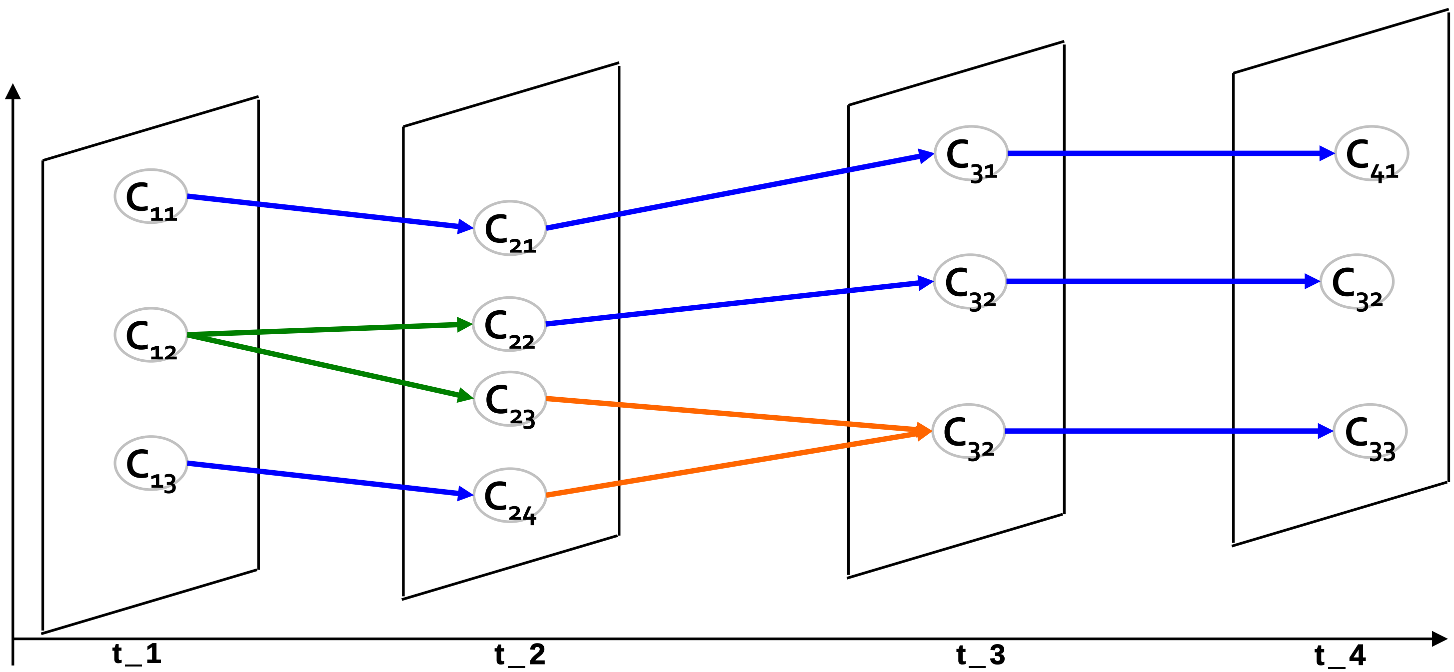
- We consider a period of observation $t_1, t_2, \ldots, t_m, \ldots$
- New records arrive over time and old records are subject to ageing according to a window size parameter.
  - Under these settings, we create at each time point t the dataset D_t
- At each t, we get a clustering Cl_t
  - Cl_t might be the result of i) complete reclustering at t or ii) cluster adaptation from Cl_{t-1}
- Clustering evolution upon consecutive time points Cl_{t-1}, Cl_t is monitored



The monitoring process (window size = 2)

The Evolution Graph

- We model the history of the population evolution in a graph structure, the Evolution Graph EG ≡ G(V,E), that spans the whole period of observation
  - $V = \{Cl_1, Cl_2, \ldots, Cl_n\}$, $Cl_i = \{C_1, C_2, \ldots, C_{|Cli|}\}$, $1 \leq i < n$
  - $E = \{e=(X, Y) : X \in Cl_i, Y \in Cl_{i+1}\}$, $1 \leq i < n$



Semantics of the Graph Nodes

- A node $v \in V$, represents a cluster c found at timepoint t_i, i.e. belonging to clustering Cl_i.
- Each node/ cluster is adorned with a label $\widehat{c}$ that summarizes its members in some intensional form.
- We work with 2 types of labels:
  - Cluster centroids, for clusters over *numerical* data
  - The set of most frequent important keywords, for clusters over *text* data

Semantics of the Graph Edges

- An edge $e=(X, Y) \in E$, denotes that a cluster $X \in Cl_i$ found at t_i has been succeeded by a cluster $Y \in Cl_{i+1}$ at t_{i+1}.
- Our notion of succession comes from our MONIC framework [Spiliopoulou et al, KDD'06] and is based on the notions of cluster overlap and cluster matching.
- The cluster overlap of X to Y denotes the members of X that still exist in Y:

$$overlap(X, Y) = \frac{|X \cap Y|}{|X|}$$

- Since X might overlap with more than one clusters in Cl_{i+1}, we use the notion of best cluster match or simply cluster match:

$Y = match(X, Cl_{i+1})$ iff:

1) $overlap(X, Y) = max_{Z \in Cl_{i+1}}(\text{overlap}(X, Z))$

2) $overlap(X, Y) \geq \tau_{survival} > 0.5$

Cluster transitions

(External) transitions of cluster X in clustering Cl_1 towards Cl_2:

- survival
  The best match of X in Cl_2 is not a match for any other cluster in Cl_1. $X \rightarrow Y$
- absorption
  There is a Y in Cl_2 that is a match for X AND for one more cluster in Cl_1. $X \xrightarrow{\subset} Y$
- split
  There are Y[1],…,Y[p] in Cl_2 that *together* match X
  AND the overlap of each one with X is at least τ_{split}. $X \xrightarrow{\subset} Y[1],\ldots,Y[p]$
- disappearance
  X is not absorbed and not split and has not survived. $X \rightarrow \otimes$

AND

- new cluster appearance
  Y in Cl_2 is not involved in the external transitions of any X. $\otimes \rightarrow Y$

Evolution Graph (EG) Construction

- EG is built incrementally as new clusterings arrive at t_1, t_2, …
- Whenever a new clustering Cl_i arrives at t_i:
  - Clusters of Cl_i are added as nodes to the EG and their labels are computed
  - We detect the cluster transitions w.r.t. Cl_{i-1} and an edge is added to the EG for each detected transition between clusters in Cl_{i-1} and Cl_i
  - Bookkeeping: The cluster members in Cl_{i-1} are discarded, whereas the members of the clusters in Cl_i are retained till the next time point t_{i+1}
    - We need these members to decide latter on the cluster transitions between Cl_i and Cl_{i+1}

Outline

- Motivation
- The evolution graph
- The FINGERPRINT of evolution
- Experiments
- Conclusions and outlook

Summarizing cluster evolution

- We summarize EG so as cluster transitions are reflected but redundancies are omitted.
  - To this end, we summarize traces (sequences of cluster survivals) into some condensed form, the fingerprint of the trace.
- For each emerged cluster c that appeared for the first time at t (i.e. a cluster with no incoming edges at t), we define its cluster trace as a sequence of cluster survivals:

$$trace(c) = \langle c_1, c_2, \ldots, c_m \rangle$$

- First, we introduce the virtual center as the summary of a (sub)trace
  - Let $trace(c) = \langle c_1, c_2, \ldots, c_m \rangle$. Let $X = \langle c_j, \ldots, c_{j+k} \rangle$ be a subtrace of it. The virtual center of X is a derived node composed of the averages of the labels of the nodes in X:

$$\widehat{X}[i] = \frac{1}{|X|} \sum_{c_i \in X} \widehat{c}[i]$$

  where [i] is the i-th dimension
  - To indicate that a cluster c has been mapped to a virtual center, we use the notation

$$c \mapsto \widehat{X}$$

The notion of summary for a trace

- Now, we define the summary of a trace
  - Let $T = <c_1, c_2, \ldots, c_m>$ be a trace. A sequence $S = <a_1, a_2, \ldots, a_k>$ is a summary of T iff
    a) $k \leq m$ and
    b) for each $c_i \in T$ there exists an $a_j \in S$ such that either $c_i = a_j$ or $c_i \rightarrow a_j$, i.e. c_i belongs to a subtrace that was summarized to the virtual center a_j.

- There are several possible summarizations of a trace, each one corresponding to a different partitioning of the trace into subtraces and consequently producing different virtual centers.
- We are interested in summarizations that achieve high space reduction while keeping information loss minimal

Summarization criteria

- Information Loss

$$ILoss\_trace(T, S) = \sum_{c \in T} ILoss\_cluster(c, a_c)$$

$$ILoss\_cluster(c, \widehat{X}) = dist(\widehat{c}, \widehat{X})$$

- Space Reduction

$$\begin{aligned} SReduction\_trace(T, S) &= \frac{(|T|-|S|)+(|T|-1-(|S|-1))}{|T|+|T|-1} \\ &= \frac{2\times(|T|-|S|)}{2\times|T|-1)} \approx \frac{|T|-|S|}{|T|} \end{aligned}$$

The FINGERPRINT of a trace

- Let T be a trace and S be a summary of T. S is a fingerprint for T iff:
  - **(C1)** For each node c ∈ T that has been replaced by a virtual center a ∈ S, it holds that:

$$dist(\widehat{c}, a) \leq \tau$$

  (C2) for each (sub)trace <c_1, c_2, … c_k> of T that has been summarized into a single virtual center a it holds that ∀ i=1, … ,k-1:

$$dist(\widehat{c_i}, \widehat{c_{i+1}}) \leq \tau$$

- Thus, S is a fingerprint of T if it has partitioned T into subtraces of clusters that are similar to each other (condition C_2) and each such subtrace has a virtual center that is close to all its original nodes (condition C_1).

Graph Summarization

- Once the traces are summarized into fingerprints, the evolution graph can be also summarized
- We propose 2 summarization strategies:
  - Incremental summarization of the graph
  - Batch summarization of the graph

Incremental summarization

- The traces are summarized incrementally as new clusterings arrive over time
- If a new clustering arrives, we check whether there is some cluster survival from the previous timepoint.
- Let x be a cluster that survives into a latter cluster y.
  - If $dist(x.label, \widehat{y})$ < τ, y is not added to the graph. Rather, x and y are summarized into a virtual center v and x is replaced in the graph by v.
  - Otherwise, the node y and the edge (x,y) are added to the graph

Batch summarization

- The summarization is performed over the whole trace based on two heuristics:
  - **Heuristic A** (deals with the violation of C2):

    If T contains adjacent nodes that are in larger distance than τ from each other split T as follows: detect the pair (c_1, c_2) with the largest distance and split T into T_1, T_2 such that c_1 is the last node of T_1 and c_2 is the first node of T_2
  - **Heuristic B** (deals with the violation of C1):

    If T satisfies C2 but contains nodes that are in larger distance than τ from the virtual center vCenter(T), split T as follows: the node c that has the maximum distance to vCenter(T) is detected and T is partitioned into T1, T2 such that c is the last node of T1 and its successor is the first node of T2

Outline

- Motivation
- The evolution graph
- The FINGERPRINT of evolution
- Experiments
- Conclusions and outlook

Experiments

- We experiments with 3 datasets
  - The Network Intrusion dataset: contains TCP connection logs from 2 weeks of LAN network traffic
    - Numerical dataset
    - Rapidly evolving
  - The Charitable Donation dataset: contains information on people who have made charitable donations in response to direct mailings
    - Numerical dataset
    - Relatively stable
  - The ACM H2.8 dataset: the set of documents inserted in between 1997 and 2004 in the ACM Digital Library, category H2.8 on Database Applications
    - Text dataset
    - Evolves in an unbalanced way

Example from the ACM H2.8 dataset

- In 1998 we observe a new cluster on Information Systems which survives till 2000.

$$trace(c_{1998_2}) = \prec c_{1998_2} c_{1999_6} c_{2000_3} \succ$$

$$\widehat{c_{1998_2}} =< information(0.96), system(0.61) >,$$
$$\widehat{c_{1999_6}} =< information(0.88), system(0.74) > \text{ and}$$
$$\widehat{c_{2000_3}} =< information(0.76), system(0.78) >.$$

- The batch FINGERPRINT, condenses this trace into a single virtual center in 1 step:

$$\widehat{v} =< information(0.87), system(0.71) >$$

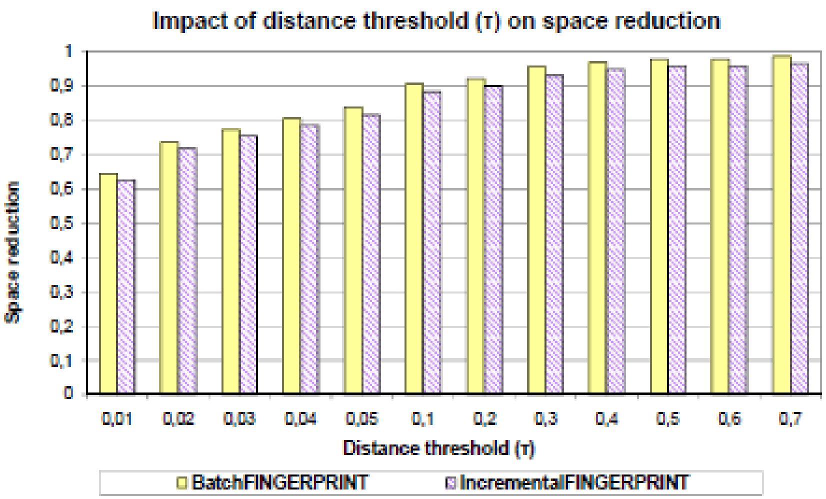
- The incremental FINGERPRINT does the same in 2 steps:
  - First summarizes c_{1998_2} and c_{1999_6} into a virtual center v_0

$$\widehat{v_0} =< information(0.92), system(0.68) >$$

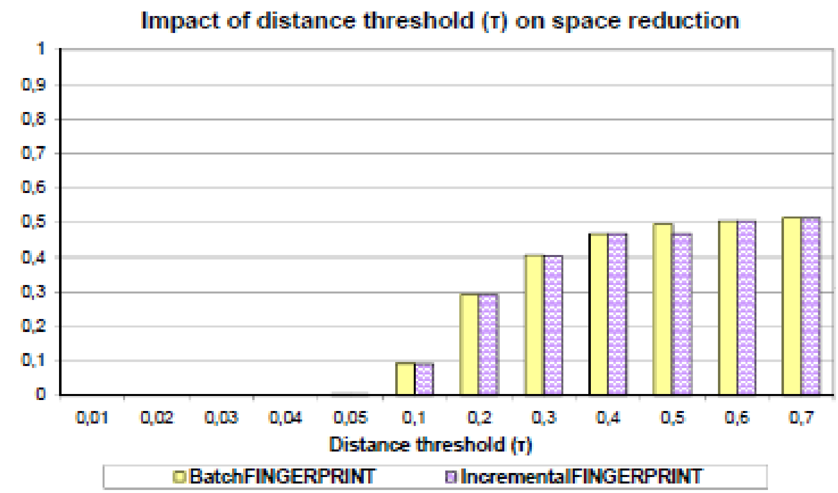
  - Then summarizes v_0 and c_{2000_3} into a new virtual center

$$\widehat{v'} =< information(0.84), system(0.73) >$$

Space reduction



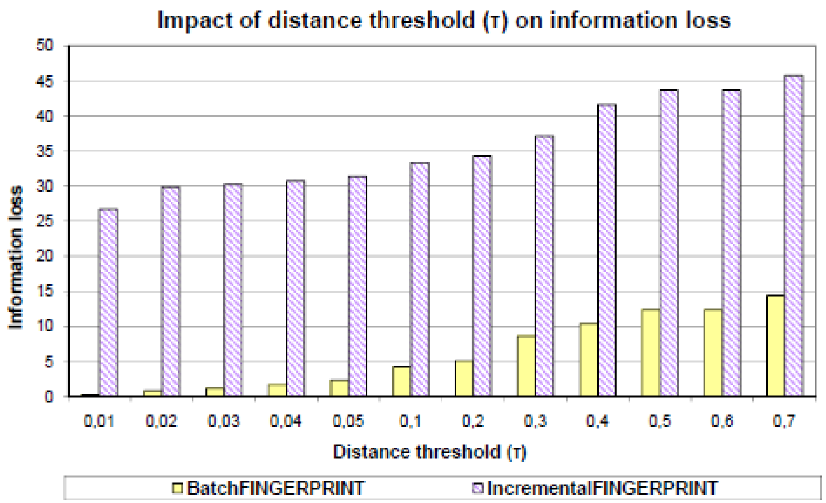
Network intrusion dataset



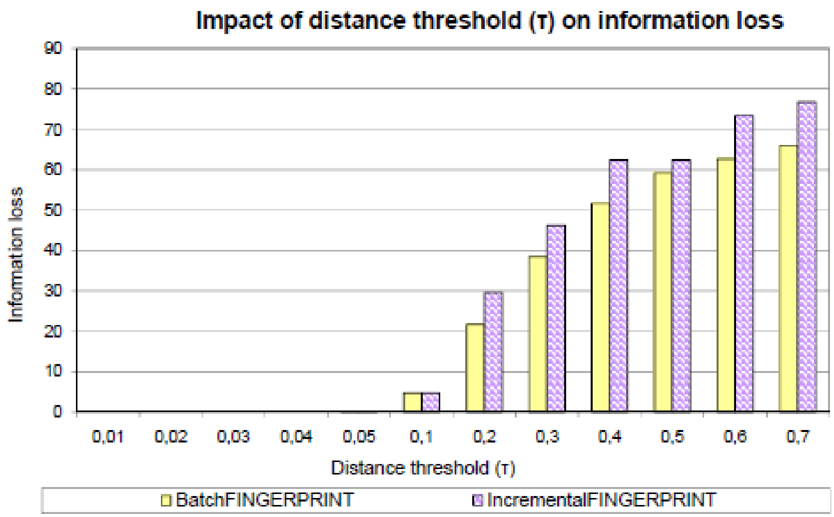
Charitable donation dataset

Impact of threshold τ on space reduction

Information loss



Network intrusion dataset



Charitable donation dataset

Impact of threshold τ on information loss

Outline

- Motivation
- The evolution graph
- The FINGERPRINT of evolution
- Experiments
- Conclusions and outlook

Conclusions & Outlook

- We presented the FINGERPRINT framework for summarizing cluster evolution in a dynamic environment subject to information loss and space reduction criteria

- Batch FINGERPRINT has better quality but requires the whole graph as input. Some hybrid method might be interesting
- So far we summarize only cluster survivals. What about splits and absorptions?
- The impact of clustering quality on the summarization

Questions?

Thank you
for your attention!

For further questions please contact me at:
ntoutsi@dbs.ifi.lmu.de

The speaker‘s attendance at this conference was sponsored by the Alexander von Humboldt Foundation

http://www.humboldt-foundation.de