Universität Konstanz

Discovering OLAP Dimensions in Semi-Structured Data

Svetlana Mansmann, Nafees Ur Rehman, Andreas Weiler, Marc H. Scholl

Database & Information Systems (DBIS)

Dept of Computer Science, University of Konstanz, Germany

Universität Konstanz

Outline

- Introduction & Motivation
  - Social Networks and Big Data
  - OLAP and Data Mining for “Big Data”
- Acquiring Facts and Dimensions
  - Data Transformation
  - Discovering New Elements
- Modeling Discovered Elements
- Usage & Maintenance of Dynamic Elements
- Conclusion

Introduction & Motivation

- Social Networks
  - Growing popularity
  - Huge data volumes
  - High data generation rate
  - Heterogeneity
- "Big Data"

Universität Konstanz

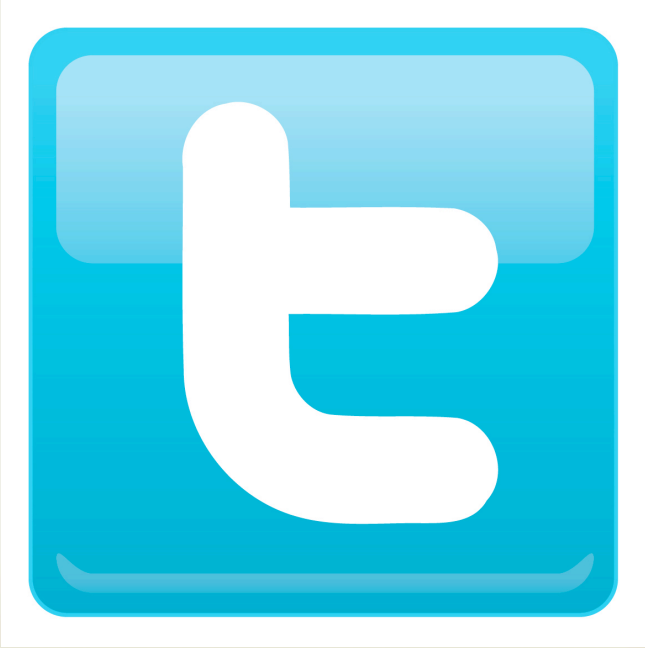
Introduction & Motivation

- Data Warehouse vs. noSQL
  - Established and mature technology
  - Standardized for interchangeability
  - Integration with Data Mining
  - Abundance of tools for various tasks
- Challenges
  - Heterogeneous and semi-structured content
  - Dynamic data, changing dimensions
  - High data arrival rate
  - Non-trivial analysis tasks

Universität Konstanz

Twitter: A motivational scenario

- Why Twitter?
  - News broadcast & Information exchange platform
  - Active Users
    - > 140 million
  - Daily Tweets
    - > 340 million
  - Set of configurable APIs:
    - Search, Rest, Stream



Universität Konstanz

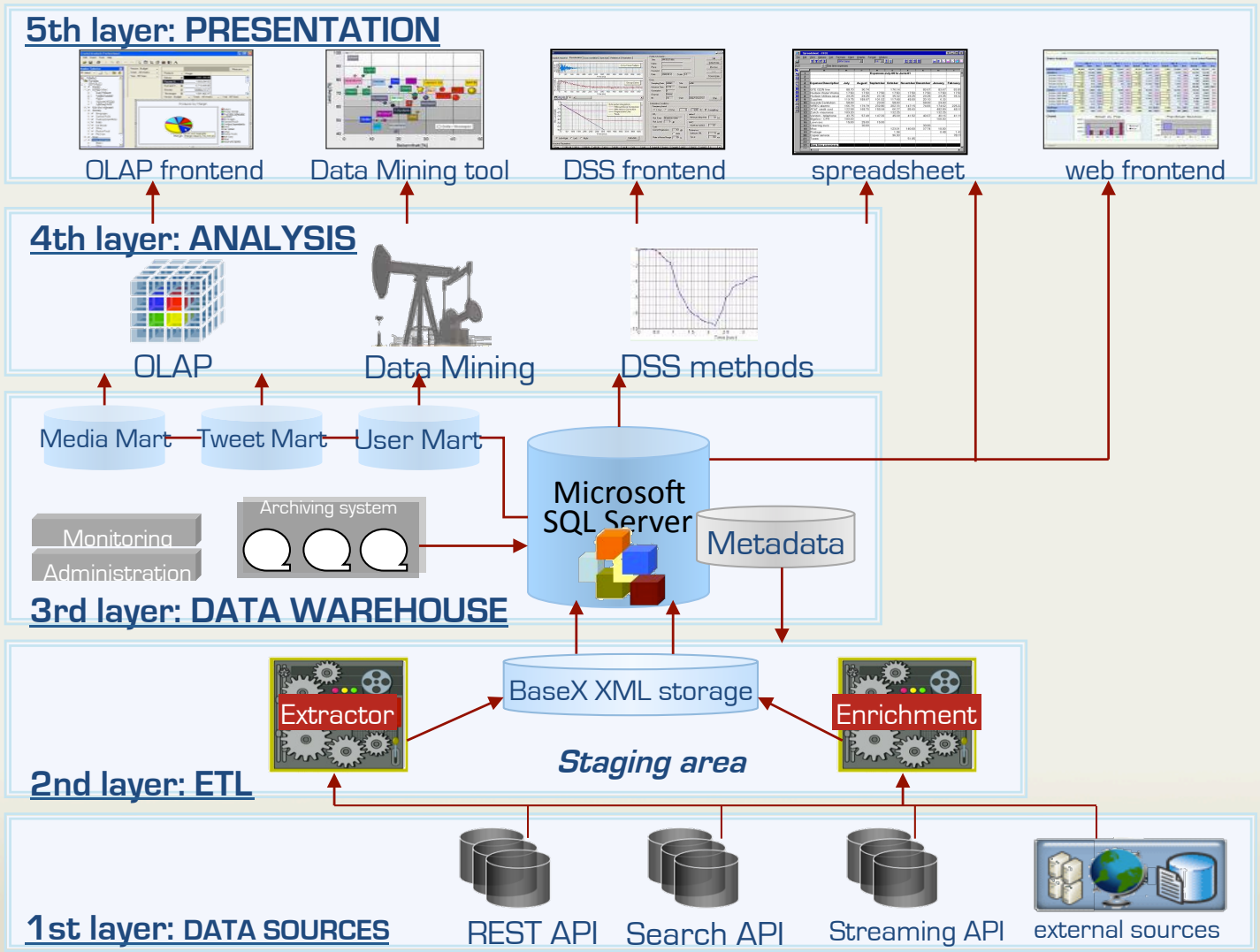
Twitter: Output Data Format

- Twitter APIs output the semi-structured data as JSON objects:
  - User data
  - Status (tweet) data
  - Timeline data
- Over 67 metadata fields
- 10% of the public stream is available

```
<tweet>
 <text>
    If you havent read about Mario Balotelli yet,
    you MUST before todays #EURO2012 final:
     http://t.co/2aFDjnsD
 </text>
    <truncated>true</truncated>
    <date>2012-01-07 18:36:05.000</date>
     <source>web</source>
     <retweeted>true</retweeted>
 <user>
   <name>Marcel***</name>
   <date>2011-08-01 06:06:34:12.000</date>
   <utc-offset>-18000</utc-offset>
   <language>en</language>
   <geo-enabled>False</geo-enabled>
   <statuses_count>1521</statuses_count>
   <followers_count>121</followers_count>
 </user>
</tweet>
```

Universität Konstanz

Multi-Layered Architecture for Twitter Data Warehouse



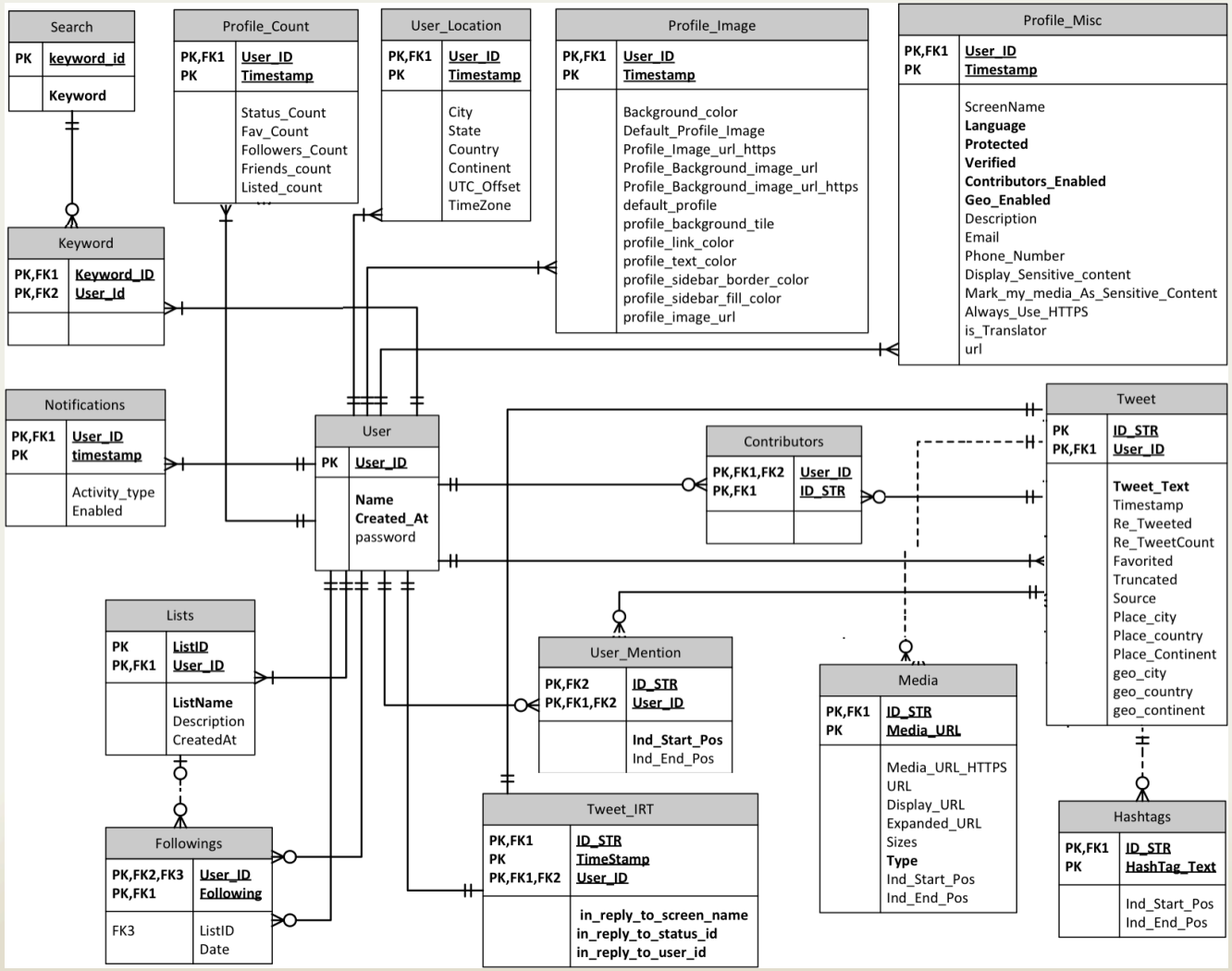
Universität Konstanz

Twitter stream - a Structured View

- Twitter data model:
  - Original model is not available
  - Streamed data is poorly documented
  - Relationships between fields are not obvious
- Reverse engineering of the data model
  - Related fields are grouped into classes
  - Relationships between classes are specified
  - Constraints are defined

Universität Konstanz

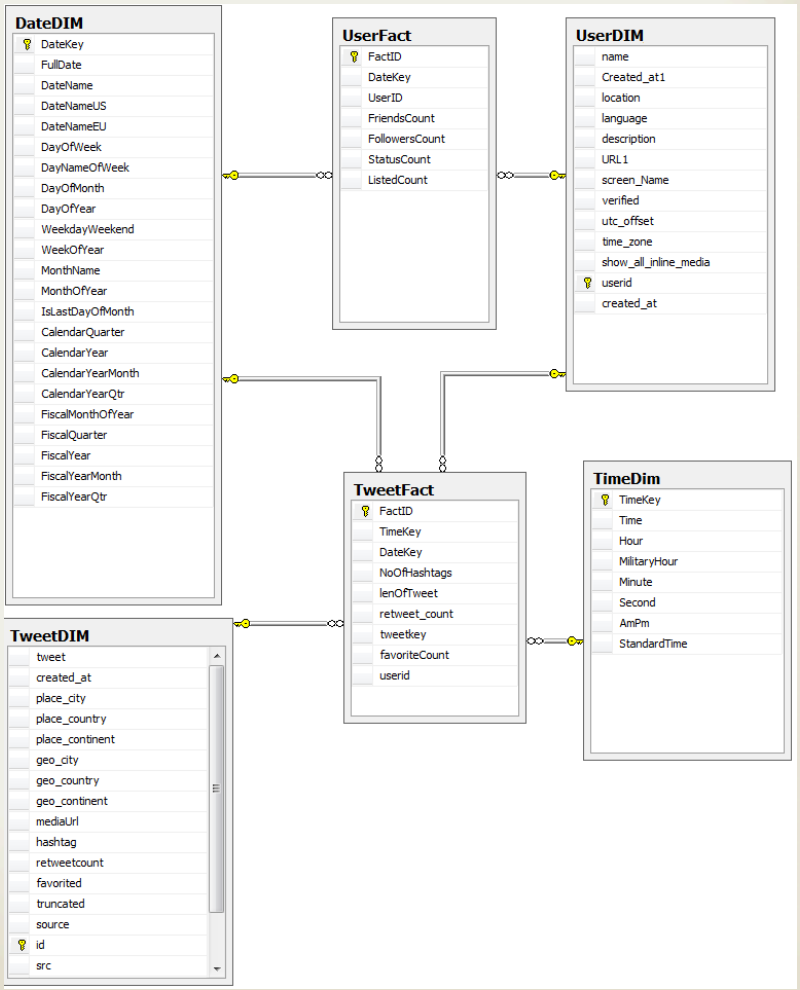
Twitter stream - a Structured View



Universität Konstanz

Acquiring Facts and Dimensions

- Cube candidates:
  - user-related data
  - tweet-related data
  - content elements
- Granularity levels:
  - user statistics
  - messaging statistics
  - topics & terms



Universität Konstanz

Acquiring Facts and Dimensions

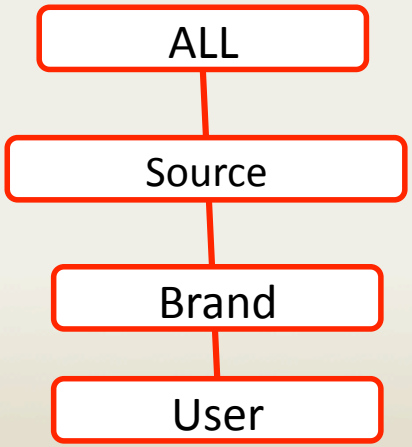
- Simple derivation / computation
- Including external data sources
  - geo-information, vocabularies
- Applying external functions (APIs)
  - language detection and translation
  - sentiment analysis
  - spam detection
  - …
- Data mining
  - hidden relationships, clustering, ranking

Universität Konstanz

Discovered Facts and Dimensions

- Simple Derivation
  - Fact/Measure Extraction
    - Length of Tweet : 64
    - Number of Hashtags: 1
  - Dimension
    - Source
      - Web, App, Phone
  - Hierarchy
    - Source

Watching #Euro final at British pub in Capitola while staring at the beach. Not what I expected but I,ll take it! Viva Italia



Universität Konstanz

Discovered Facts and Dimensions

- External Data Sources & APIs
  - Language
    - English
  - Entity Detection
    - Event
      - Euro (Championship)
    - Facility
      - British pub
    - Country
      - Italy
  - Topic
    - Sports
  - Tags: Sports, Fun, Eurocup
  - Sentiment: Positive

Watching #Euro final at British pub in Capitola while staring at the beach. Not what I expected but I,ll take it! Viva Italia

Universität Konstanz

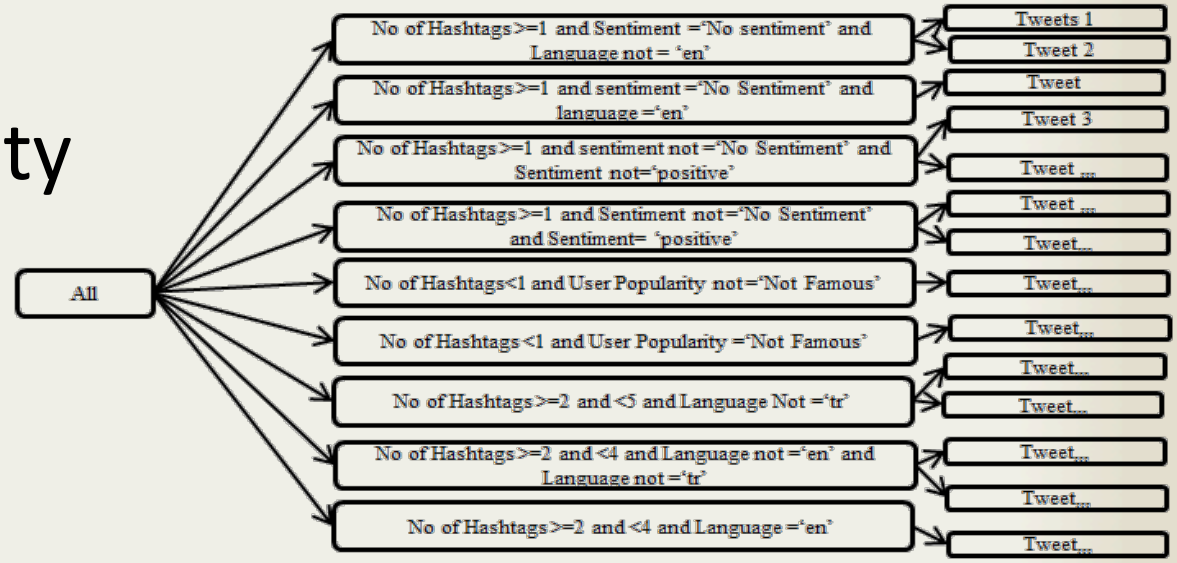
Discovered Facts and Dimensions

- Data Mining
  - Clusters of Users
    - Trending, Spam, Lifestyle, etc.
  - Clusters of Tweets
    - Popularity
  - Non-Trivial Relationships
    - What contributes to popularity & trending of
      - users
      - tweets

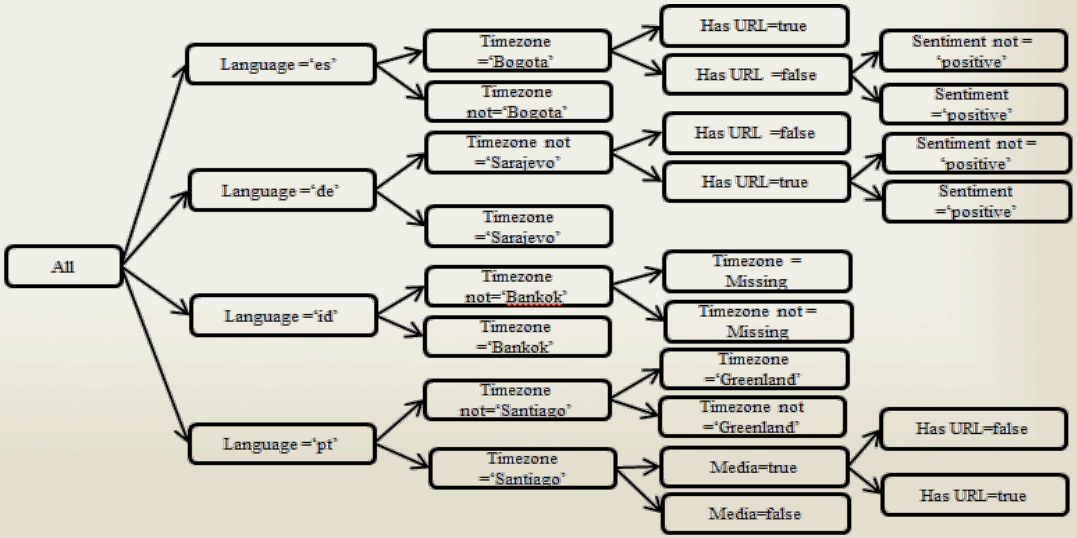
Universität Konstanz

Discovered Facts and Dimensions

- Tweet Popularity Classifier

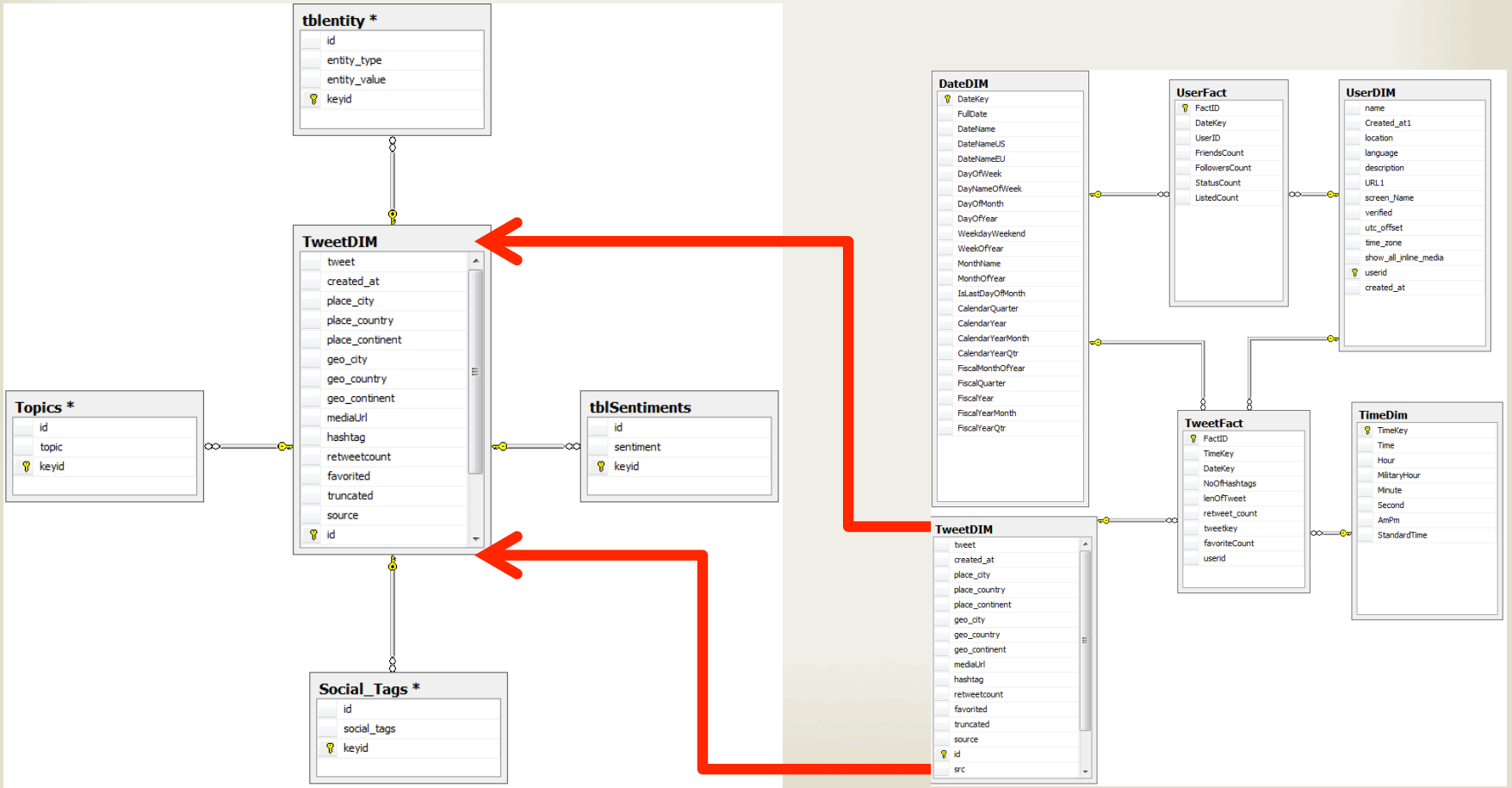


- User Popularity Classifier



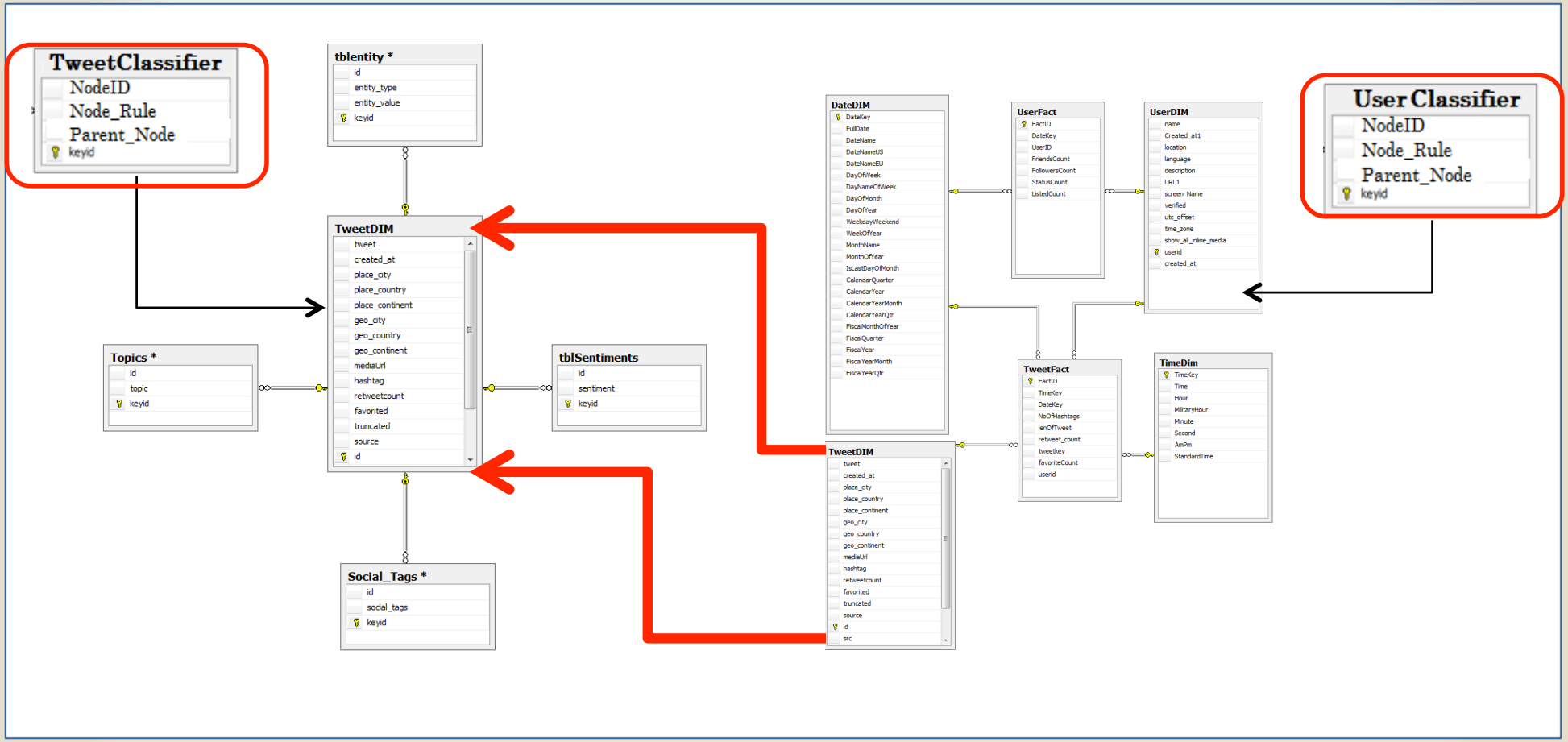
Universität Konstanz

Modeling Discovered Elements



Universität Konstanz

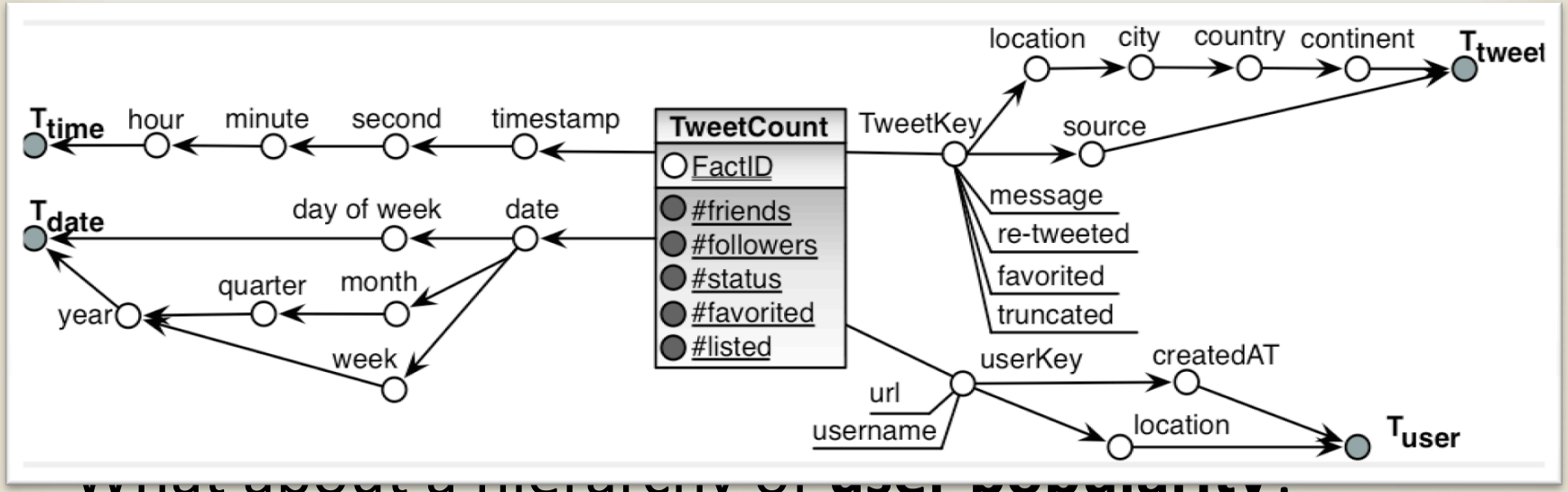
Modeling Discovered Elements



Universität Konstanz

Discovered Hierarchy - Example

- Conceptual modeling of new elements
- Consider **user** dimension in the **TweetCount** fact type

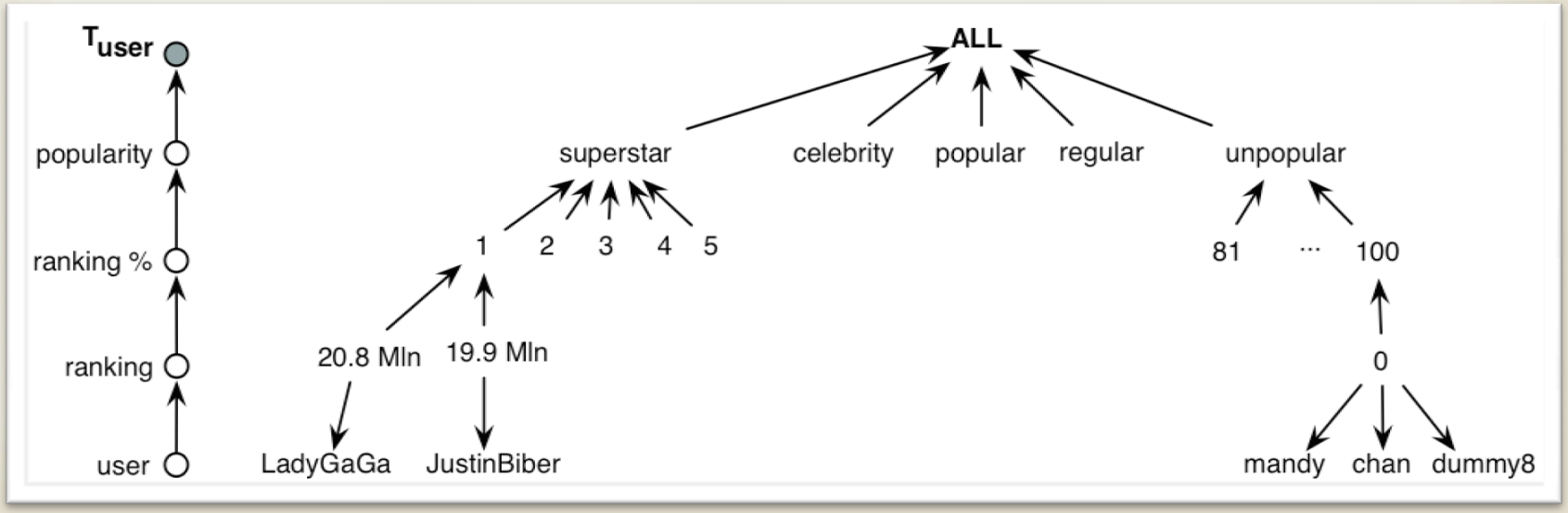


What about a hierarchy of **user popularity**?

Universität Konstanz

Discovered Hierarchy - Example

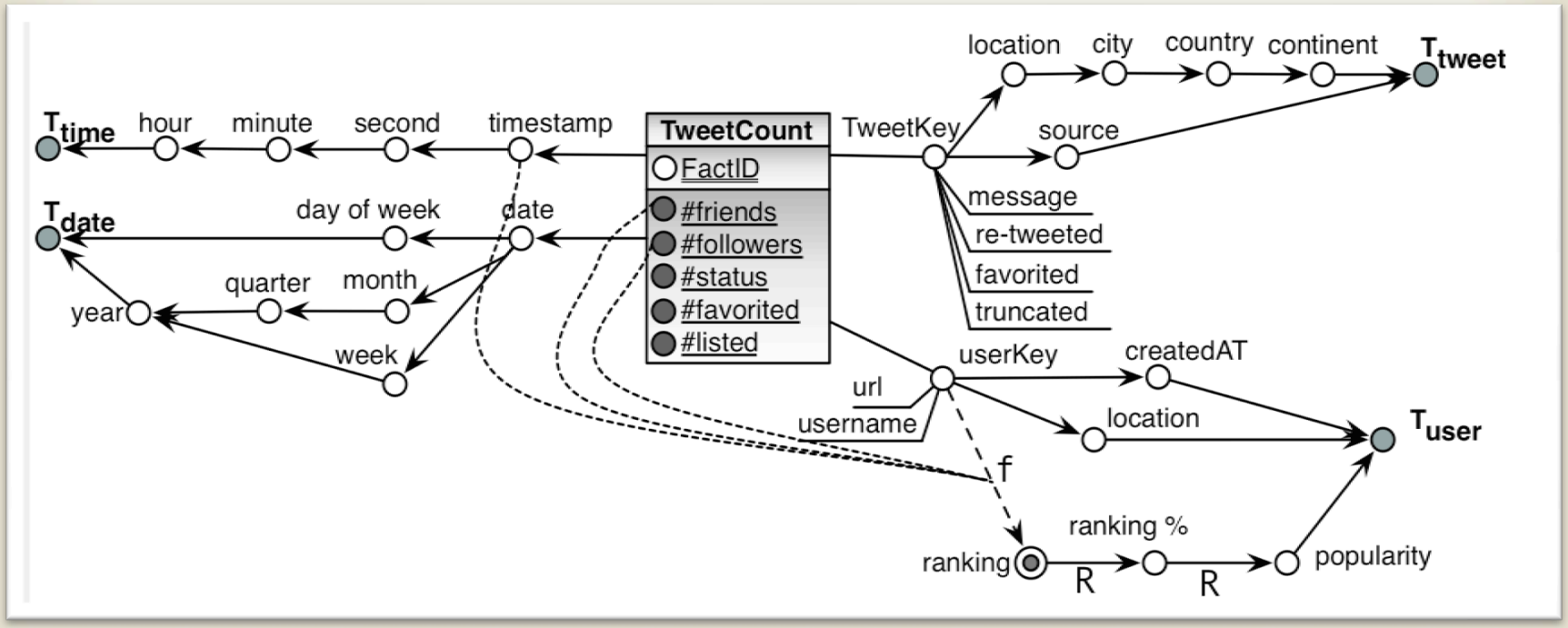
- Adopt some ranking function (e.g., based on the number of followers)
- Define higher-level groupings (e.g., based on percentages or thresholds)



Universität Konstanz

Discovered Hierarchy - Example

- Add new aggregation path to the fact schema
- Specify the computation formula for added elements



- Problem: the added hierarchy is **dynamic**

Universität Konstanz

Maintenance of Dynamic Elements

- Similar to *Slowly Changing Dimensions*
- Multi-versioning / historization
  - current version in the dimension table
  - Previous version in history table(s)
  - Temporal constraints for historical records

USER-DIM

userkey	url	name	createdAt	location	ranking	ranking%	popularity	...
1308331	-	wp-guru	2010-06-21	London, GB	171.2	79	regular	...

USER-HiSTORY

userkey	url	name	createdAt	location	ranking	ranking%	popularity	...	created
1308331	-	wp-guru	2010-06-21	London, GB	117.4	83	unpopular	...	2012-02-01

Universität Konstanz

Querying along Dynamic Elements

- OLAP queries with multi-versioned dimensions
  - correct aggregation by joining the fact entries with the matching versions of the dimension
  - "playing" with different versions for what-if analysis
- Examples
  - *retrieve the messages tweeted in 2009 by those users who are popular now (and not in 2009!)*
  - *retrieve recent tweets containing the hashtags which were in TOP 20 in 2008*

Universität Konstanz

Conclusion

- Proposed extraction of multi-dimensional data cube from semi-structured data.
- Extended the underlying dataset and model using DM and semantic enrichment methods.
- Adapted the DWH to deal with the changing/ dynamic data using concept of SCD.
- Enabled OLAP for recent and historic data analysis.

Universität Konstanz

Thank You

Questions