Uncertainty in Querying and Monitoring Streams

Themis Palpanas
University of Trento
Data Evolution

- structured
- precise
- static (almost)
Data Evolution

- non-structured
- uncertain
- streaming
Things to Come

- big data
  - very large collections (i.e., terabytes to exabytes)
  - scientific, business, user generated
Things to Come

• big data
• streaming data
  ▫ sensors, feeds, continuous analytics
  ▫ response times in seconds to nanoseconds
Things to Come

• big data
• streaming data
• heterogeneous data
  ▫ structured, non-structured, text, multimedia
  ▫ variety of sources, schemas, representations, models
Things to Come

- big data
- streaming data
- heterogeneous data
- uncertain data
  - imprecision, inconsistencies, incompleteness, ambiguities, latency, deception, approximations, privacy preserving transformations
  - process/data/model uncertainty
Things to Come: Uncertain Data

- enterprise data are (somewhat) precise
- all other data have some degree of uncertainty
- by 2015 80% of all data will be uncertain
**uncertain data**
data series similarity

work with:
Michele Dallachiesa, Besmira Nushi, Katsiaryna (Katya) Mirylenka

*PVLDB’12, QUEST’11*
Data series

- Sequence of points ordered along some dimension
Data series

- Sequence of points ordered along some dimension

Wind speed
From ocean observing node project, http://bml.ucdavis.edu/boon/wind.html
Data series

• Sequence of points ordered along some dimension

Historical stock quotes
Data series

- Sequence of points ordered along some dimension

Trajectories from GPS logs
From http://www.flickr.com/photos/kitepuppet/3604115258
Data series

- Sequence of points ordered along some dimension

Projectile points (arrowhead) converted into time series

From “Detecting Time Series Motifs Under Uniform Scaling”, Dragomir Yankov et al., 2007
Uncertain Data Series

- Point value at each timestamp uncertain
Uncertain Data Series

• Point value at each timestamp uncertain

Tumor detection from noisy brain scan images
From “DUST: a generalized notion of similarity between uncertain time series”, Saret et al., 2010
Uncertain Data Series

- Point value at each timestamp uncertain

Ocean temperature at multiple depths

From UpTempO project, http://psc.apl.washington.edu/UpTempO/
Uncertain Data Series

- Point value at each timestamp uncertain

Stock price changes during one year
Similarity queries

- Find all objects in a dataset similar to given query object
Similarity queries

• base for several analysis and mining tasks
  ▫ pattern recognition
  ▫ clustering
  ▫ classification
  ▫ motif discovery
  ▫ outlier identification
  ▫ ...
Probabilistic range queries

- Similarity queries for uncertain data:

\[
\text{Find all uncertain time series } T \text{ s.t. } \Pr(\text{dist}(Q,T) \leq \varepsilon) \geq \tau
\]

- \textbf{Q} query sequence
- \textbf{T} candidate match
- \textbf{\varepsilon} distance threshold
- \textbf{\tau} probabilistic threshold
Outline

- Review state-of-the-art

<table>
<thead>
<tr>
<th>MUNICH</th>
<th>PROUD</th>
<th>DUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Probabilistic Similarity Search for Uncertain Time Series”</td>
<td>“PROUD: a probabilistic approach to processing similarity queries over uncertain data streams”</td>
<td>“DUST: a generalized notion of similarity between uncertain time series”</td>
</tr>
<tr>
<td>Åsfalq et al., SSDBM ’09</td>
<td>Yeh et al., EDBT ‘09</td>
<td>Sarangi et al., KDD ‘10</td>
</tr>
</tbody>
</table>

- Analytical comparison
- Experimental evaluation
- Neighborhood-aware models
- Guidelines for practitioners
Uncertainty models for time series

- Sequence of independent random variables:
  \[ X = \langle x_1, \ldots, x_n \rangle \]
  - \( x_i \) models point at timestamp \( i \)
  - independency as simplifying assumption

- What do we know about \( x_i \)?
Possible-world semantics

- Samples drawn from $x_i$
Distribution-aware model

- *A priori knowledge of $x_i$* distribution
Possible-world semantics
Distance samples from possible world instantiations
\( \Pr(\text{dist}(Q,T) \leq \varepsilon) \) as frequency of matching distances
PROUD

- Known mean and variance
- Distance $\text{Dist}(Q,T)$ as sum of random variables
- Central Limit Theorem: $\text{Dist}(Q,T) \sim N(...)$
DUST

• Full knowledge of **point** and **value** distributions
• \[ \text{dust}(x_i,y_i) = \sqrt{- \log( \Pr(\text{dist}(x_i,y_i) = 0))} \]
• \[ DUST(X,Y) = \sqrt{\Sigma_i \text{dust}(x_i,y_i)^2} \]
• New distance measure
## Analytical comparison

<table>
<thead>
<tr>
<th>Point independence</th>
<th>MUNICH</th>
<th>PROUD</th>
<th>DUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertainty model</td>
<td>Possible-world semantics</td>
<td>Distribution-aware model</td>
<td>Distribution-aware model</td>
</tr>
<tr>
<td>Knowledge assumptions</td>
<td>Samples</td>
<td>Mean and variance</td>
<td>Value and error distributions</td>
</tr>
<tr>
<td>Mixed error distribution</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance measures</td>
<td>Euclidean, DTW</td>
<td>Euclidean</td>
<td>DUST, DUST@DTW</td>
</tr>
<tr>
<td>Similarity queries</td>
<td>Probabilistic range queries</td>
<td>Probabilistic range queries</td>
<td>Distance measure</td>
</tr>
<tr>
<td>Quality guarantees</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experimental setup

- Techniques implemented in C++
- Datasets loaded in main memory
- 17 real datasets from UCR time series archive
  - on average 500 data series, 300 points each
- Perturbation models uncertainty
  - error in measurements
Probabilistic range queries

- Similarity queries for uncertain data

\[
\text{Find all uncertain time series } T \text{ s.t. } \Pr(\text{dist}(Q,T) \leq \varepsilon) \geq \tau
\]

- Q query sequence
- T candidate match
- \(\varepsilon\) distance threshold
- \(\tau\) probabilistic threshold

Distance and probability thresholds

- **Distance thresholds** $\varepsilon$ for query $Q$:
  - $C$ = 10th nearest neighbor to $Q$ on original dataset
  - $\varepsilon_{\text{Eucl}} = \text{EuclideanDist}(Q,C)$ on uncertain dataset
  - $\varepsilon_{\text{DUST}} = \text{DUST}(Q,C)$ on uncertain dataset

- **Probabilistic threshold** $\tau$
  - brute-force tuning to maximize accuracy
Experimental methodology

- Evaluating accuracy
  - original datasets used as ground truth
  - techniques evaluated on perturbed datasets

- All time series used as queries on respective datasets

- MUNICH high computational cost
  - evaluation on single truncated dataset
  - 60 time series, length 6. Each point 5 samples.
Results on time performance

- CPU time per query varying normal error
- error variance slightly affects DUST performance
Results on quality performance

- Truncated dataset, accuracy varying normal error
- **MUNICH penalized**: constant low number of samples
Results on quality performance

- Accuracy with mixed normal error: 20% stddev=1.0, 80% stddev=0.4
- DUST performs better than PROUD, Euclidean
Impact of perturbation on distances

- Accuracy with mixed normal error: 20% stddev=1.0, 80% stddev=0.4
- DUST performs better than PROUD, Euclidean

Why such a large difference?
Impact of perturbation on distances

- Histogram of Euclidean distances in original and perturbed datasets
- Adiac Euclidean distances affected more by perturbation
Neighborhood-aware models

• Neighboring points can be used for noise reduction.

• Inspired by moving average.

• Contribution of points to distance measure inversely proportional to their standard deviation.
  ▫ high variance \(\rightarrow\) low contribution
Uncertain Moving Average

• Point at timestamp $i$ defined as:

$$pm_i = \frac{\sum_{j=i-w}^{i+w} \frac{v_j}{s_j}}{2w + 1}$$

- $s_j$ standard deviation of point $x_j$
- $v_j$ sample of point $x_j$
Uncertain Exponential Moving Average

- Point at timestamp $i$ defined as:

$$ p e_i = \frac{\sum_{j=i-w}^{i+w} v_j \frac{e^{-\lambda |j-i|}}{s_j}}{\sum_{j=i-w}^{i+w} e^{-\lambda |j-i|}} $$

- $s_j$ standard deviation of point $x_j$
- $v_j$ sample of point $x_j$
- $\lambda$ decay constant
Results on quality performance

- Accuracy with mixed normal error: 20% stddev=1.0, 80% stddev=0.4
- UEMA/UMA accuracy 4-15% better than DUST
- UEMA better accuracy than UMA
Which technique should I choose?

- Constant perturbation over time: Euclidean
- Quality guarantees: MUNICH, PROUD
- DUST high accuracy and robust
  - does not require probabilistic threshold
  - requires full knowledge of error/value distributions
- No prior distribution knowledge: MUNICH
  - does not scale
- Highest accuracy: UEMA/UMA
  - not probabilistic measures
Summary

- Uncertain time series processing real problem in many application domains.

- Existing state-of-the-art techniques have limitations

- Simple neighborhood-aware models outperform more sophisticated techniques.
  - valuable information is conveyed in neighboring points.
Open problems

• Remove independency assumption
  ▫ modeling of sequential correlations in similarity measure

• What can we do in case of just few samples?

• Efficient data structures for processing and summarizing uncertain time series
uncertain data
existential uncertainty

work with:
Michele Dallachiesa, Gabriela Jacques-Silva, Buğra Gedik, Kun-Lung Wu
Motivating Example

- Continuous vibration measurements from machinery monitoring system
- Measurements from different sensors observing same device
- Measurements as samples from distribution of device vibration
Motivating Example

- Device vibration modeled by uncertain tuples:

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Device ID</th>
<th>Vibration</th>
<th>Existential probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-01-01 10:03:43</td>
<td>1234</td>
<td>0.203</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.252</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.180</td>
<td>0.34</td>
</tr>
</tbody>
</table>

- Stream operators may introduce existential uncertainty, e.g., filters: vibration < 0.2

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Device ID</th>
<th>Vibration</th>
<th>Existential probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-01-01 10:03:43</td>
<td>1234</td>
<td>0.180</td>
<td>0.34</td>
</tr>
</tbody>
</table>
How do we handle existential uncertainty in stream processing?

• In many application domains data sources can be seen as uncertain

• Need for handling data uncertainty in composition of stream operators

• We focus on sliding windows and similarity joins as driving use case
Possible world semantics

- Data stream $T$ growing sequence $<t_1, ..., t_n, ...>$
  - point $t_i =$ sequence of samples $t_{i,1}, ..., t_{i,s}$
  - sample $t_{i,j}$ occurs with probability $Pr(t_{i,j})$
  - point $t_i$ exists with probability $Pr(t_i) = \Sigma_j Pr(t_{i,j})$

- Strong model for uncertainty in databases
  - Trio system (Stanford)
  - Orion system (Purdue)
  - ...
Sliding windows on uncertain streams

- Given stream $T$, $W(T, w) = T[n-w+1, n]$
  - $n$ current timestamp
  - $w$ window width

"uncertain" sliding window

possible world #1

existentially uncertain points

$s_i$ exists
$s_j$ exists
Sliding windows on uncertain streams

- Given stream $T$, $W(T,w) = T[n-w+1, n]$
  - $n$ current timestamp
  - $w$ window width

existentially uncertain points

“uncertain” sliding window

$s_i$ exists
$s_j$ does not exist
Sliding windows on uncertain streams

- Given stream $T$, $W(T,w) = T[n-w+1, n]$
  - $n$ current timestamp
  - $w$ window width

Diagrams illustrate the concept of "uncertain" sliding windows, where points $s_i$ do not exist and $s_j$ exist.
Sliding windows on uncertain streams

- Given stream $T$, $W(T,w) = T[n-w+1, n]$
  - $n$ current timestamp
  - $w$ window width

existentially uncertain points

$s_i$ does not exist
$s_j$ does not exist
Similarity join without uncertain data

- Given two input streams $S,T$ find similar points over time:
  - candidate matches identified by sliding windows
  - points $s_i, t_j$ similar if $\text{dist}(s_i, t_j) \leq \varepsilon$
Similarity join with uncertain data

- Uncertainty dimensions in similarity matches:
  - sliding window content (existential uncertainty)
  - points value (value uncertainty)
  - existential uncertainty can give rise to value uncertainty!

\[ S \text{ “uncertain” sliding window} \]
Summary

• existential uncertainty knocks on our door
  ▫ value uncertainty leads to existential uncertainty
  ▫ existential uncertainty leads to value uncertainty

• related to timestamp uncertainty
  ▫ (caused by clock/synchronization problems)

• we need streaming operators that can efficiently support it
non-structured data
subjectivity analysis

work with:
Mikalai (Nick) Tsytsarau, Sihem Amer-Yahia, Kerstin Denecke
[DMKD’12, DiversiWEb’11, WWW’10]
Motivation

- There are many opinion sources: blogs, wikis, forums, social networks, ...

- Sentiment analysis can be used to:
  learn customers attitude to a product (features)
  analyze people's reaction to some event

- Above problems require sentiment aggregation:
  diversity-preserving
  time-aware
  scalable
Some Examples

Google features opinion aggregation in product search

◦ displays the most representative diverse opinions about a product.
◦ summarizes ratings among different reviews
Subjectivity Analysis

- survey on Sentiment Analysis and Opinion Mining
  - Tsytsarau and Palpanas DMKD 2012
  - summarizes about 100 papers from 2001 to 2010
  - studies trends

- Subjectivity Analysis
- Sentiment Analysis
- Opinion Aggregation
- Contradiction Analysis
- Streaming Methods
- Scalability
- Sentiment Spam
- Some More...
Contradictions, what are they?

- **Contradictions** in text are situations where ‘two sentences are extremely unlikely to be true together’

- Contradictions may be of different types, for example:
  - antonymy: *hot* - *cold*, *light* - *dark*, *good* - *bad*
  - negation: *nice* - *not nice*
  - mismatches: *the solar system has 8 planets* - *there are 9 planets*
  - sentiments: *I like this book* - *this book is awful*

- **Sentiment Contradictions** may occur due to:
  - diversity of views (i.e., simultaneous contradiction)
  - change of views (i.e., sentiment shift)
Contradiction Detection Pipeline

- Text Extraction
- Topic Identification
- Opinion Extraction
- Opinion Aggregation
- Contradiction Extraction
**Sentiment Contradictions**

- Raw Sentiments \( S_i \)
- Aggr. Sentiment \( \mu_S = \frac{1}{n} \sum_{i=1}^{n} S_i \)
- Sentiment Variance \( \sigma_S^2 = \frac{1}{n} \sum_{i=1}^{n} (S_i - \mu_S)^2 \)
- Contradiction \( C = \frac{\vartheta \cdot \sigma_S^2}{\vartheta + (\mu_S)^2} W \)

- We calculate contradiction by combining Aggr. Sentiment and Sentiment Variance
- If Aggregated Sentiment close to 0, the contradiction is high
- The larger the variance, the higher the contradiction
- Detected contradictions not (always) possible to identify by visual inspection

Physical Structure of the Contradiction Storage

- Store only the values of first- and second-order moments of sentiments M1 and M2
- Include time-tree pointers (black dots •) and contradiction values (gray squares □)
- Add cross-links for adjacent pages to achieve a fast consequent reading (diamonds ◆)
- Reference the rest of the topics by a special pointer to a separate page (white dots ○)
Contradictions Tool

Just say "NO" ladies! My ARNP prescribed it and three months later: headaches 24/7, hair l ...
I have been taking Yaz for 2 1/2 years, it has been very effective in preventing pregnancy ...
The three months I took Yaz were probably the three worst months of my life. I tried Yaz a ...
I am almost 19 years old and Yaz was the first birth control I have ever been on. I am cur ...
Ladies, you MUST take a Vitamin B supplement with Yaz! (I take B- ...
I was on yasmine for one year and things were fine. My doctor switched me to yaz and I dec ...
I am with most of the other woman on how awful Yaz has made me feel. I am also taking Lexa ...
Summary

- Described framework for automatic contradiction detection
- Defined two types of contradictions
- Proposed effective criteria to calculate contradictions
- Proposed scalable, efficient, incremental storage structures
Open Questions

‣ How can we explain contradictions?

‣ What news caused them?

‣ Which demographics contribute to them?
non-structured data
entity resolution in data spaces

work with:
George Papadakis, Ekaterini Ioannou, Claudia Niederee, Wolfgang Nejdl
[TKDE’13, WSDM’12, JCDL’11]
Entity Resolution

• benefits:
  • identifies and aggregates the different objects describing the same entity
  • improves data quality and integrity, fosters re-use of existing data sources

• problems:
  • is an inherently quadratic problem: each object has to be compared with all others
Blocking for Entity Resolution

• benefits:
  • similar entities are grouped into blocks
  • comparisons are only executed inside blocks

• problems:
  • existing blocking techniques require a fixed, a-priori known schema
Highly Heterogeneous Information Spaces

- Web 2.0, Semantic Web, Dataspaces

- Voluminous, semi-structured datasets
  - *DBPedia 3.4*: 36.5 million triples and 2.1 million entities
  - *BTC09*: 1.15 billion triples and 182 million entities

- Users are free to insert not only attribute values but also attribute names → high levels of heterogeneity
  - *DBPedia 3.4*: 50,000 attribute names
  - *Google Base*: 100,000 schemata and 10,000 entity types

- Large portion of data stemming from automatic information extraction → noise, tag-style values
Problem Definition

given two duplicate-free heterogeneous information spaces (Clean-Clean ER), or a single one that contains duplicates in itself (Dirty ER):

• **Problem 1 (Effectiveness)**
  Develop *block building methods* that lead to blocks with low number of missed matches (i.e., high recall)

• **Problem 2 (Efficiency)**
  Develop *block processing methods* that reduce the number of required pair-wise (entity) comparisons
Conclusions

- confluence of very large, uncertain, heterogeneous, non-structured, streaming data

- fueling the INnovative QUErying of STreams
References

**Uncertain Data Series**

**Subjectivity Analysis**
- Mikalai Tsytserau, Themis Palpanas. *Survey on Mining Subjective Data on the Web*. Data Mining and Knowledge Discovery (DMKD) Journal, Special Issue on A Decade of Mining the Web, 24(3), 2012
- Mikalai Tsytserau, Themis Palpanas, Kerstin Denecke. *Scalable Detection of Sentiment-Based Contradictions*. International Workshop on Knowledge Diversity on the Web (DiversiWeb), in conjunction with the World Wide Web Conference (WWW), Hyderabad, India, March 2011

**Entity Resolution in Data Spaces**
- George Papadakis, Ekaterini Ioannou, Themis Palpanas, Claudia Niederee, Wolfgang Nejdl. *A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces*. IEEE Transactions on Knowledge and Data Engineering (TKDE), accepted for publication
- George Papadakis, Ekaterini Ioannou, Claudia Niederee, Themis Palpanas, Wolfgang Nejdl. *Beyond 100 Million Entities: Large-Scale Blocking-based Resolution for Heterogeneous Data*. ACM International Conference on Web Search and Data Mining (WSDM), Seattle, WA, USA, February 2012
dbTrento: http://db.disi.unitn.eu