Exploiting Structure, Annotation, and Ontological Knowledge for Automatic Classification of XML-Data

Martin Theobald,
Ralf Schenkel, and Gerhard Weikum

Saarland University
Saarbrücken, Germany

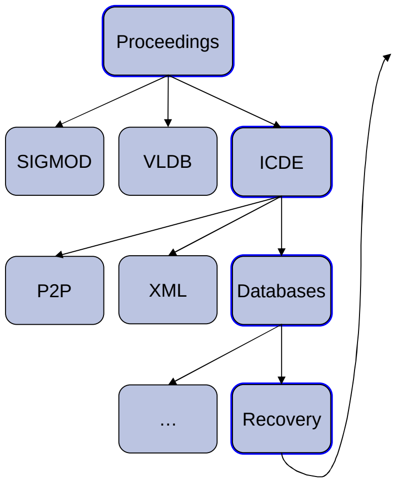
Limitations of XPath & XQuery in an Environment with Diverse Schemes

```
….
<inproceedings key="conf/icde/BargaLW02">
  <author>Roger S. Barga</author>
  <author>David B. Lomet</author>
  <author>Gerhard Weikum</author>
  <title>Recovery Guarantees
         for General Multi-Tier Applications.</title>
  <year>2002</year>
  <booktitle>ICDE</booktitle>
</inproceedings>
…
```

DBLP

- //proceedings[contains(., "icde")]/title[contains(.,“Recovery")]/parent::*
  → 0 Results.
- //title[contains(.,“Recovery")]/parent::*
  → 7.859 Results.

Automatic Classification helps.



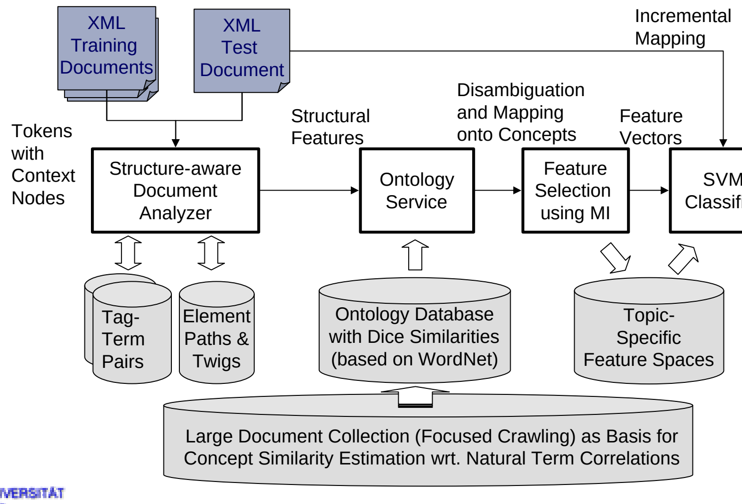
→ 12 Results:

…
<inproceedings key="**conf/icde/BargaLW02**">
<author>**Roger S. Barga**</author>
<author>**David B. Lomet**</author>
<author>**Gerhard Weikum**</author>
<title>**Recovery Guarantees for General Multi-Tier Applications.**</title>
<year>**2002**</year>
<booktitle>**ICDE**</booktitle>
</inproceedings>
…..

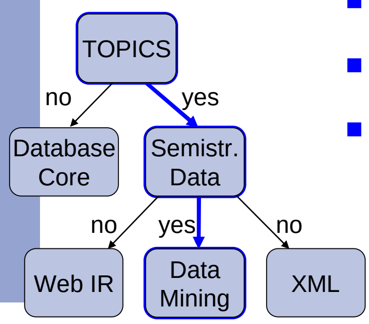
Challenges in XML Classification

- Exploit annotation and structure
- Exploit ontological knowledge on sparse and/or heterogeneous training data
- Mapping of tags (and text terms) to semantic concepts
- In-document word sense disambiguation
- Quantification of concept similarities

Using Structure and Ontological Knowledge for Classification



Feature-Selection & Term Weighting



- **Linear Support Vector Machines** for binary classifications in the topic tree
- **Topic-specific feature spaces** to support binary classification steps
- **Mutual Information (MI)** yields ranking for the most discriminating features per topic (aka. Kullback-Leibler-Divergence)

$$MI(X_i, c_j) := P[X_i \wedge c_j] \log_2 \frac{P[X_i \wedge c_j]}{P[X_i]P[c_j]}$$

- Term weights in classic TF*IDF
- IDF computed on element frequencies

Exploiting Annotation: Tag-Term Pairs

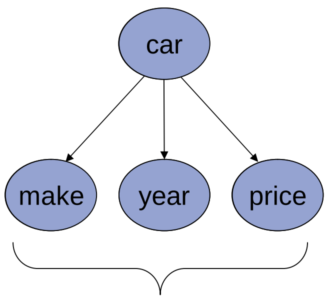
- Structure-aware features for more precise document representation
- Interpret **(tag, term) pairs** as **concept-value pairs** in the spirit of a database schema

```
<car>
  <make>Audi</make>
  <type>A4</type>
  <year>98</year>
  <price>10.000</price>
</car>
```

make$Audi, type$A4, year$98, price$10.000

car$make$Audi, car$type$A4, car$year$98, car$price$10.000

Exploiting Structure: Element Paths and Twigs



car$make$year

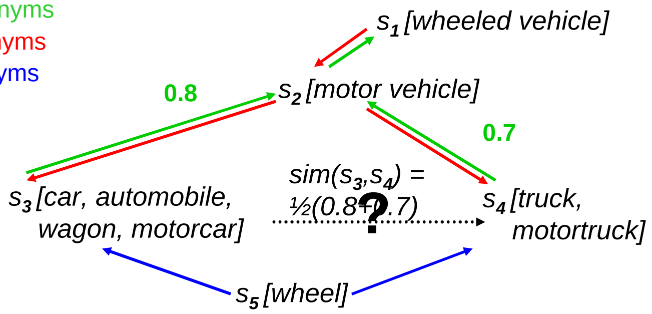
car$year$price

car$make$price

- Extension of the feature space by structural patterns → **Paths & Twigs**
- Preserve or disregard element ordering
- Different feature types (tag-term pairs & twigs) are mapped to distinct dimensions in the vector space
- **Scalability** and **noise reduction** through feature selection (MI) under an integrated SVM model

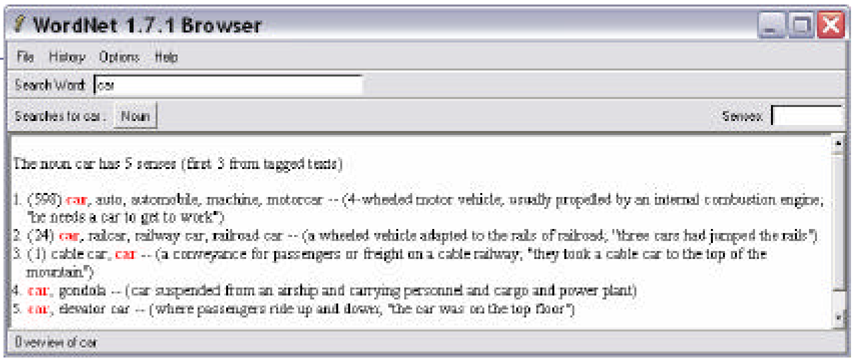
Exploiting Ontological Knowledge

- **WordNet:** Directed and weighted ontology graph capturing
  - Hypernyms
  - Hyponyms
  - Holonyms



- **Quantified relationships** based on estimated concept similarities:
  - Dice coefficient: $dice(s_1, s_2) = \dfrac{2\, df\left(senses(s_1) \cup senses(s_2)\right)}{df\left(senses(s_1)\right) + df\left(senses(s_2)\right)}$

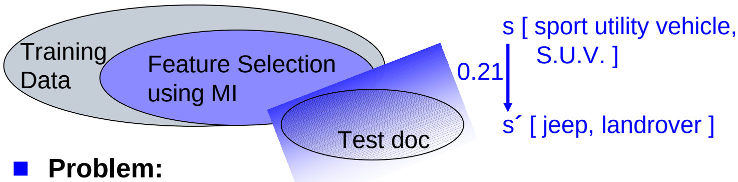
Word Sense Disambiguation



- Compare **term context *con(t_k)*** with **synset context *con(s_j)*** using cosine measure
- Synset context includes hypernyms, hyponyms, and holonyms plus WordNet descriptions
- → Infer semantics from current context rather than stipulate it

Incremental Mapping for Classification

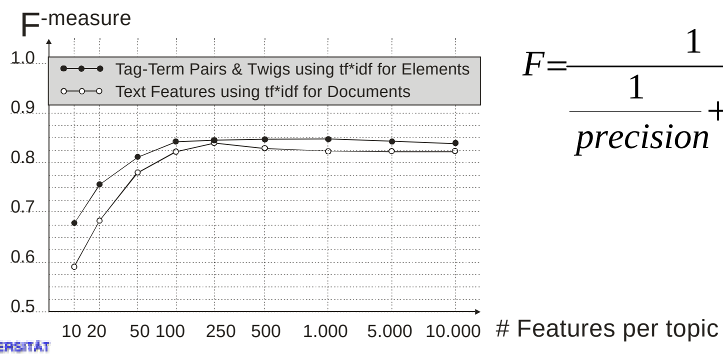
- For any unknown concept *s* in a test document *d* do:
  - Replace *s* with its closest match *s´* from the training data
  - Adjust term weight of *s* in *d* by concept similarity *sim(s, s´)*



- **Problem:**
  - Possible loss of feature correlations that the SVM has learned
  - No feature independency for SVM
- → Reconsider *dice(s, s´)* with restrictive threshold
- → Replace concept *s* only if *s´* is strongly correlated to *s* , otherwise skip *s*

Experimental Evaluation: Internet Movie Database (IMDB)

- Training with very view features for *Action* vs. *Western*
- Homogenous, but rich structure with varying amounts of content
- Tag-term pairs (95%) plus twigs (5%) using MI
- Ontology lookups on tags only



$$F = \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Summary

- *Concept-based classification boosts classification results*
  - Detection of synonyms
  - Incremental mapping of unknown concepts
- *Structure-aware features offer a more precise document representation for XML*
- *Application area:*
  - Training on small, user-specific topic directories, e.g., bookmarks
  - Classification of heterogeneous data sources

Future Work

- *More robust term-to-sense mapping*
  - Improved disambiguation of word senses
  - Better awareness of feature correlations (in incremental term-to-concept mapping)
  - Topic-specific ontologies
  - Is-instance-of relationships
- *Integration into large web applications, e.g., focused crawling*

Questions?