Creation of Heterogeneous XML Document Collections based on the Internet Movie Database

presented

by

Ivelina Stavreva

Content

- Goalrepresentation
- What is IMDB?
- Possible Sources of Heterogenity
- Examplediagram for Heterogeneous XML Documents
- Program run

Goal

- Motivation: Lack of large heterogeneous collections of XML data
- Until now: DBLP or INEX, large collections with homogeneous structure
- Problem: DBLP or INEX inadequate for similarity search
- Goal: heterogeneous collection of XML Document Collections

What is IMDB?

- a rich source of information about movies and people involved in the movie business (actors, directors, editors, producers, etc.)

- contains:
  - factual data (like birthday)
  - textual data (like biography)

Possible sources of heterogenity in XML docs for IMDB

- Applying information from IMDB

 - replace person (movie) names (titles) by their alternative names (titles).
 Example:

<movie id=„195067“>
 <title>Matrix Resolution, The</title>
 <alt_title>Matrix 3, The</alt_title>
</movie>

<movie id=„195067 “>
 <title>Matrix 3, The</title>
 <alt_title>Matrix resolution, The</alt_title>
</movie>

- replace tag <movie> by tags
 derived from genres (<thriller>, <drama>,etc.)

Possible sources of heterogenity in XML docs for IMDB

- Using different languages

- replace tags by their counterparts (e.g. <movie> by <film>)

```
<movie id=„195067“>
  <title>Matrix Resolution, The</title>
  <alt_title>Matrix 3, The</alt_title>
</movie>
```

```
<film id=„195067 “>
  <titel>Matrix resolution, The</titel>
  <alt_titel>Matrix 3, The</alt_titel>
</film>
```

Possible sources of heterogenity in XML docs for IMDB

- Different granularities

  - One XML document per year (location), listing some of the movies filmed then (there)

```
-<movie id=„195067“>
   <title>Matrix Resolution, The</title>
   <prod_year>2000<prod_year>
 </movie>

-<movie id=„195068“>
   <title>abc</title>
  <prod_year>2000</prod_year>
 </movie>
```

```
<year2000>
   <title>Matrix resolution, The</title>
   <title>abc</title>
</year >
```

Possible sources of heterogenity in XML docs for IMDB

- Different granularities

  - One XML document for all movies with the same director

```
-<movie id=„195067“>
   <title>xyz</title>
   <director>X<director>
 </movie>

-<movie id=„195068“>
   <title>abc</title>
   <director>X</director>
 </movie>
```

```
<personX>
    <title>xyz</title>
    <title>abc</title>
</personX>
```

Examplediagram for Heterogeneous XML Documents