Integrated Movie Database

Group 9:
Muhammad Rizwan Saeed
Santhoshi Priyanka Gooty Agraaram
Ran Ao
Outline

• Background
• Project Description
• Demo
• Conclusion
Background

• Semantic Web
  – Gives meaning to data

• Ontologies
  – Concepts
  – Relationships
  – Attributes

• Benefits
Background

• Semantic Web
  – Gives meaning to data
• Ontologies
  – Concepts
  – Relationships
  – Attributes
• Benefits
  – Facilitates organization, integration and retrieval of data
Project Description

- Integrated Movie Database
Project Description

• Integrated Movie Database
  – Project Phases
    • Data Acquisition
    • Data Modeling
    • Data Linking
    • Querying
Data Acquisition

• Datasets
  – IMDB.com
  – BoxOfficeMojo.com
  – RottenTomatoes.com
  – Wikipedia.org
  – GoodReads.com

• Java based crawlers using jsoup library
Data Acquisition: Challenge

• Crawlers require a list of URLs to extract data from.
• How to generate a set of URLs for the crawler?
  – User-created movie list on IMDB
    • Can be exported via CSV
  – DBpedia
    1. Using foaf:isPrimaryTopicOf of dbo:Film or schema:Movie classes to get corresponding Wikipedia Page.
    2. Crawl Wikipedia Page to get IMDB and RottenTomatoes link
Data Acquisition: IMDB

- Crawled 4 types of pages
  - Main Page
    - Title, Release Date, Genre, MPAA Rating, IMDB User Rating
  - Casting
    - List of Cast Members (Actors/Actresses)
  - Critics
    - Metacritic Score
  - Awards
    - List of Academy Awards (won)
- Records generated for 36,549 movies
- Casting Records generated: 856,407
Data Acquisition: BoxOfficeMojo

• BoxOfficeMojo provides an index of all the movies on their website
  – Total movie links found: 16,945

• Main Movie Page
  – Title, Release Date, Genre, Run time, Domestic Gross, Worldwide Gross, Budget
Data Acquisition: Others

• RottenTomatoes
  – Extracted another score based on % of positive reviews for the movie: (Scale: 0-100)
  – Records generated: 10,000

• GoodReads.com
  – Crawled user-generated lists of books which were adapted for movies
  – Records generated: 3,000
Data Modeling

• Integrated Movie Database Ontology
  – Ontology creation using Protégé
  – Automatic conversion of CSV data into RDF using Apache Jena API
Data Linking

• Why do we need to link data?
Data Linking

• Why do we need to link data?
    • http://www.imdb.com/title/tt0482571/
    • http://www.boxofficemojo.com/movies/?id=prestige.htm
    • http://www.rottentomatoes.com/m/prestige/
    • http://www.goodreads.com/book/show/239239.The_Prestige
Data Linking

• Connect independently modeled data sources

IMDB Movie

The Prestige

The Prestige

Box Office Mojo

2006

130 min

8.5

$109m

$40m

2006

http://www.imdb.com/title/tt0482571/

http://www.boxofficemojo.com/movies/?id=prestige.htm
Data Linking

• Connect independently modeled data sources

2006 IMDB Movie 130 min

2006

The Prestige

The Prestige

Box Office Mojo

2006 $109m

$40m

The Prestige

http://www.imdb.com/title/tt0482571/

http://www.boxofficemojo.com/movies/?id=prestige.htm

The Prestige

$109m

$40m

2006
Data Linking

• FRIL: Fine-grained Record Integration and Linkage Tool
  – Allows record linkage based on combination of similarity metrics
  – Linkage was fine-tuned based on multiple trials
Data Linking

![Diagram of FRIL: A Fine-Grained Record Linkage Tool v. 2.1.5]

- **Data source**: Name: Mojo, Status OK
- **Filter**
- **Deduplication**

- **Result savers**: Status OK
  - Deduplication

- **Linkage mode**: Status: idle

No linkage/deduplication was performed

Memory: Used 24MB out of 533MB
Data Linking
Data Linking

• Matching criteria:
  – Movies vs Movies
    • Match based on (similar) Title and same Release Year
    • \( \text{simScore} = 50 \times \text{Edit Distance}(\text{Titles}) + 50 \times \text{Equals}(\text{Release Years}) \)
Data Linking

• Edit Distance:
  – Number of modifications (insertion, deletion, modification) that make string A equal to string B
    • $f(\text{Max, May}) = 1$
    • $f(\text{Ma, May}) = 1$
Data Linking

• Matches found despite
  – Punctuation Difference
  – Textual Difference

<table>
<thead>
<tr>
<th>Crank 2: High Voltage</th>
<th>Crank: High Voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love Wedding Marriage</td>
<td>Love, Wedding, Marriage</td>
</tr>
<tr>
<td>The Hills have Eyes II</td>
<td>The Hills have Eyes 2</td>
</tr>
</tbody>
</table>

– High Precision
– Recall?
Data Linking

• Matching criteria:
  – Movies vs Books
  • Match based on (similar) Title and Book Publication Year <= Movie Release Year

<table>
<thead>
<tr>
<th>Avatar</th>
<th>Avatar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amnesia</td>
<td>Amnesiac</td>
</tr>
</tbody>
</table>
Data Linking

• Matching criteria:
  – Movies vs Books
    • Match based on (similar) Title and Book Publication Year <= Movie Release Year

<table>
<thead>
<tr>
<th>Avatar (Movie)</th>
<th>Avatar (The Last Air Bender)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amnesia</td>
<td>Amnesiac</td>
</tr>
</tbody>
</table>

• Lower precision
• Required manual cleaning
Querying

• Data hosted using Openlink Virtuoso
  – RDF Dataset created contains 2.3 million triples

• Demo: SPARQL Queries
  – Enrich user experience
  – Link previously disconnected data
    • e.g. Which author’s books have been most profitable for the Movie Industry?
  – Path Query
    • e.g. Finding collaboration
Querying: Path Queries

• Finding Collaboration or Degree of Separation
  – Show Business
    • (Kevin) Bacon Number
  – Research Community
    • Erdős Number, Einstein Number
  – Social Media
    • LinkedIn Connections
Demo
Conclusion

• Integrated data allows to cross reference information that previously required accessing multiple web pages.

• The datasets can be augmented and used for applying ML and Social Media Analysis techniques e.g.
  – Calculating Influence
  – Similarity between Entities (e.g. Movies, Actors)
Thank you!