Match, Forrest, Match

Tanja Bergmann, Stefan Bunk, Tim Draeger, Dominik Müller, Ricarda Schüler

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Match, Forrest, Match creates a unified, linked open data set about movies by combining different data sources. It enables the user to query the resulting data set using linked open data techniques.

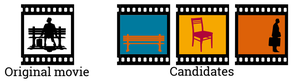
MATCHING

Problems matching only by name:

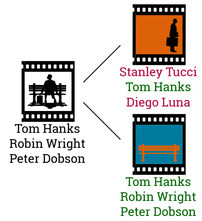
- Different movies, same name:
  The Avengers (2012) vs. The Avengers (1998)
- Same movie, different names:
  Batman vs. Badman
  The Internship vs. Prakti.com
  The Italian Job vs. Italian Job, The

Solution: Actor overlap with threshold

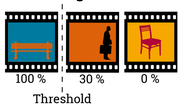
1. Find candidates



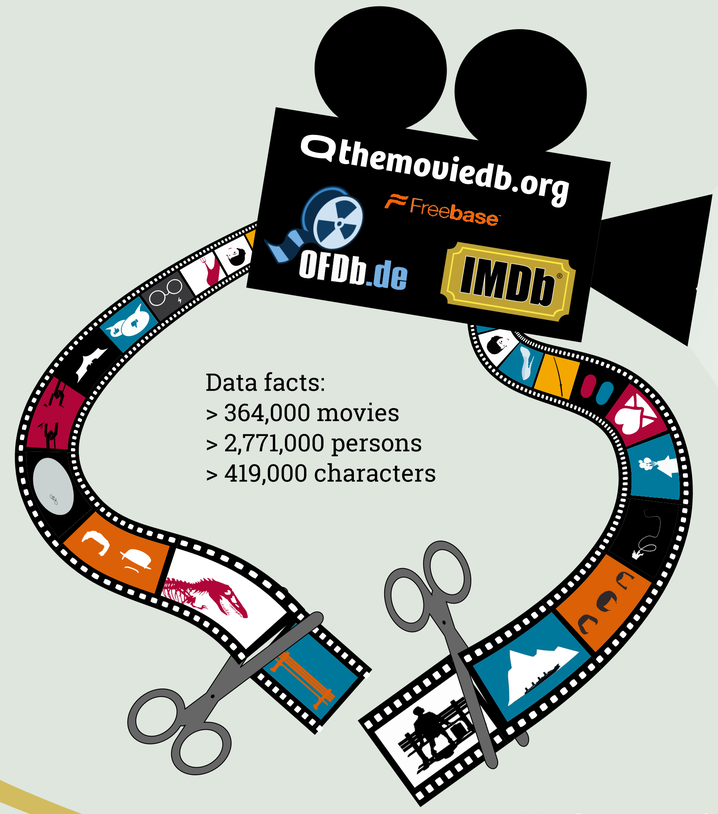
2. Calculate overlap



3. Take highest score over threshold

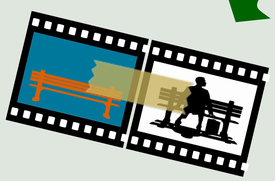


Several overlaps (actor, director, producer, writer) are combined with certain weights. To refine the matching, name and year similarity are used additionally.

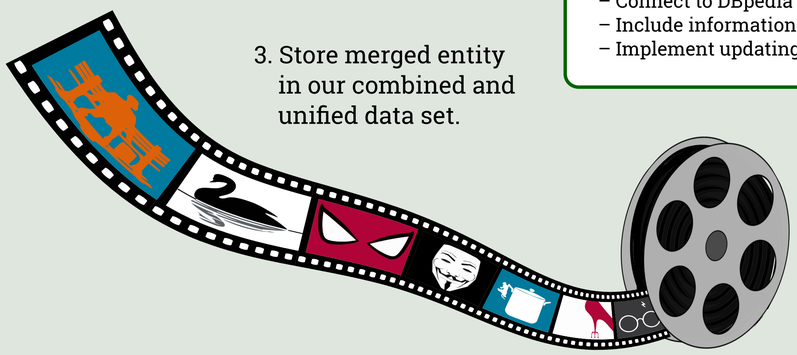


Data facts:
> 364,000 movies
> 2,771,000 persons
> 419,000 characters

1. Match same entities from different data sources.



2. Merge data from both data sets to create one unique view of the resource.



3. Store merged entity in our combined and unified data set.

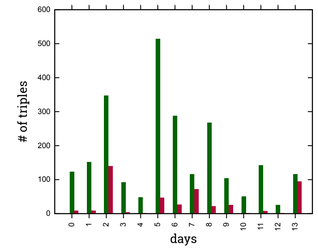
UPDATING

Scheduler

Creates new tasks to crawl and triplify IMDB movies, which should be updated.

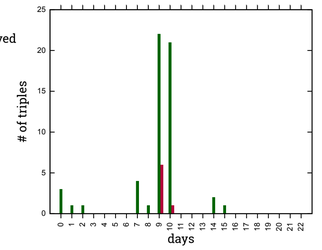
upcoming movies	daily	5 year old movies	weekly
a year old movies	monthly	5 - 25 year old movies	yearly

How often does a new released movie change in the first days on IMDB?



Often changing triples:
cast, links to images and videos, IMDB rating, characters, taglines, release infos, alternative names

How many new upcoming movies are released per day on IMDB?



Most movies are released between Thursdays to Sundays.

EVALUATION

Evaluation measures:

Test set:
Movies from TMDB, which have an IMDB id.

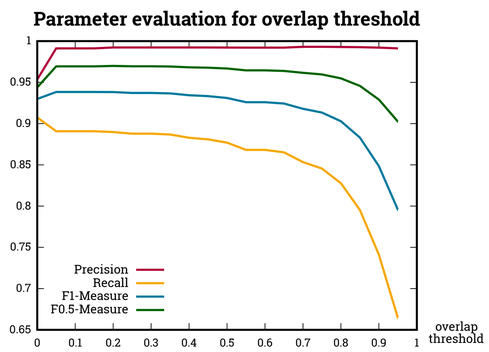
Of 2028 movies, we
matched 1829 correctly,
matched 10 incorrectly and
did not match 189.

Precision:	99,46 %
Recall:	90,19 %
F1-Measure:	94,60 %
F0.5-Measure:	97,45 %

Current error analysis:

Not in candidates	67
No candidates	95
Threshold to high	27

Parameter estimation:



Determining the best threshold for the overlap, when a match is recognized. Tests in 5 % steps. You can see, that a threshold of ~ 70 % yields the best results.

OUTLOOK

- Improve matching algorithm
- Improve merging of movies
- Include more data sources
  (e.g. data sources, which contains Indian movies)
- Connect to DBpedia
- Include information about TV films and TV series
- Implement updating algorithm for all data sources