Master Thesis – MovieSearch

Building semantic search queries with suggestions

Tobias Sommer

Albert-Ludwigs-University Freiburg
Faculty of Engineering

Chair of Algorithms and Data Structures

16. November 2016

Goals

- Provide **user-friendly** Interface for Semantic Search in the domain movie
- Utilize plot and facts
- Support 3-ary relations

Use cases

Example tasks to fulfill:

- *Find movies made by Jerry Bruckheimer.*
  Explore data, e.g. relation names.
- *Find movies where Frodo was played by Elijah Wood.*
  Use and connect **3-ary relations**.
- *Find an action movie with Arnold Schwarzenegger where he fights with a sword.*
  Query conditions: **plot snippets and facts**.

Use cases

Example tasks to fulfill:

- *Find movies made by Jerry Bruckheimer.*
  Explore data, e.g. relation names.
- *Find movies where Frodo was played by Elijah Wood.*
  Use and connect **3-ary relations**.
- *Find an action movie with Arnold Schwarzenegger where he fights with a sword.*
  Query conditions: **plot snippets and facts**.

Use cases

Example tasks to fulfill:

- *Find movies made by Jerry Bruckheimer.*
  Explore data, e.g. relation names.
- *Find movies where Frodo was played by Elijah Wood.*
  Use and connect **3-ary relations**.
- *Find an action movie with Arnold Schwarzenegger where he fights with a sword.*
  Query conditions: **plot snippets and facts**.

Splitting plot and facts

Consider use case:

Find an action movie
with Arnold Schwarzenegger
where he fights with a sword.

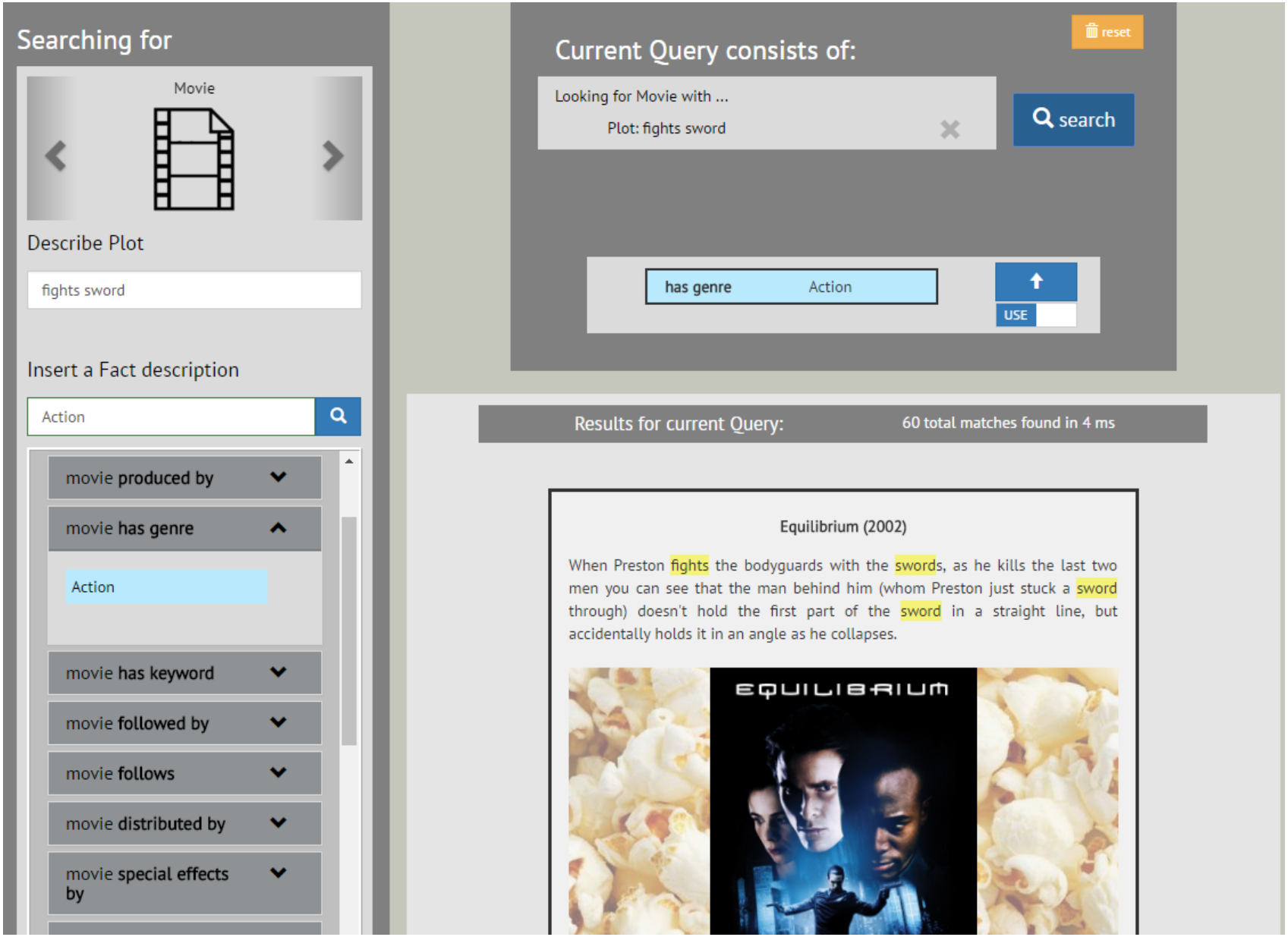
Plot

Text information $\rightarrow$ *... fights with a sword.*

Facts

Structured information $\rightarrow$ *... with actor A. Schwarzenegger.*

User Interface – Example (1/3)



User Interface – Example (2/3)

Searching for

Movie

Describe Plot

fights sword

Insert a Fact description

starring schwarzen

Click one of the following options:

show matches for

in movie as actor: schwarzen

show matches for

in movie as actor: schwarzen

and

character: ...

None of the above is what I want to search, repeat fact suggestion for

starring schwarzen

Current Query consists of:

reset

Looking for Movie with ...

Plot: fights sword

has genre Action

search

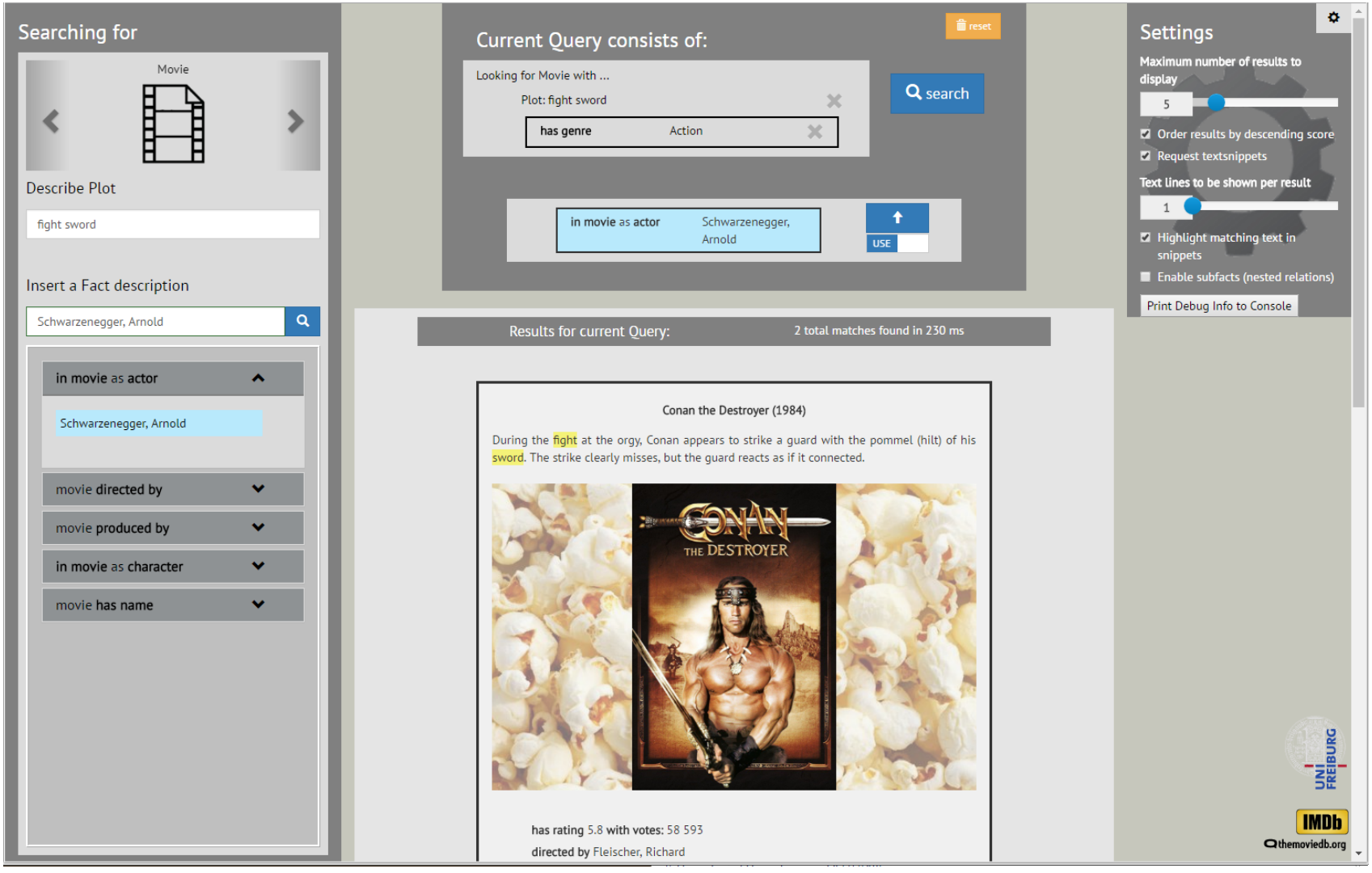
NOT

Results for current Query: 60 total matches found in 2 ms

Equilibrium (2002)

When Preston fights the bodyguards with the swords, as he kills the last two men you can see that the man behind him (whom Preston just stuck a sword through) doesn't hold the first part of the sword in a straight line, but accidentally holds it in an angle as he collapses.

User Interface – Example (3/3)



Facts

Store structured information as triples:

Example

(Conan, has-genre, Action)

In general

Fact := (Entity, relation, value)

Different kind of values

- word: (Inception, has-genre, Action)
- entity: (Inception, directed-by, Christopher Nolan)
- number: (Inception, has-budget, 160.000.000 $)
- date: (Inception, released, 29.07.2010)

Relations

Occurring relations:

- **Binary** relations

Example

as triple (Conan, has-genre, Action)

- **3-ary** relations

Example

from text *Mel Gibson plays William Wallace in Braveheart.*

as triples (cast-link$_1$, in-movie, Braveheart)
(cast-link$_1$, actor, Mel Gibson)
(cast-link$_1$, character, William Wallace)

Fact suggestions

Suggesting facts during input:

- ⇒ Discover names in unfamiliar data
- ⇒ Find connectable relations

How to find suggestions

Names: match description to relation and entity names

⇒ Inverted index of prefixes

Triples: find (relation, value) pairs

⇒ Facts graph from triples

Matching names – Inverted index

Example (Inverted index of prefixes)

ID	Entity names
0	"Barbara"
1	"Bill Berg"

$\Longrightarrow$

$\Longrightarrow$

Prefix		ID lists
[b]	$\rightarrow$	0, 1
[ba]	$\rightarrow$	0
[be]	$\rightarrow$	1
[bi]	$\rightarrow$	1

Matching names:

- Get ID lists of matching prefixes
- Intersect all (sorted) ID lists
- Further filter for contains from start

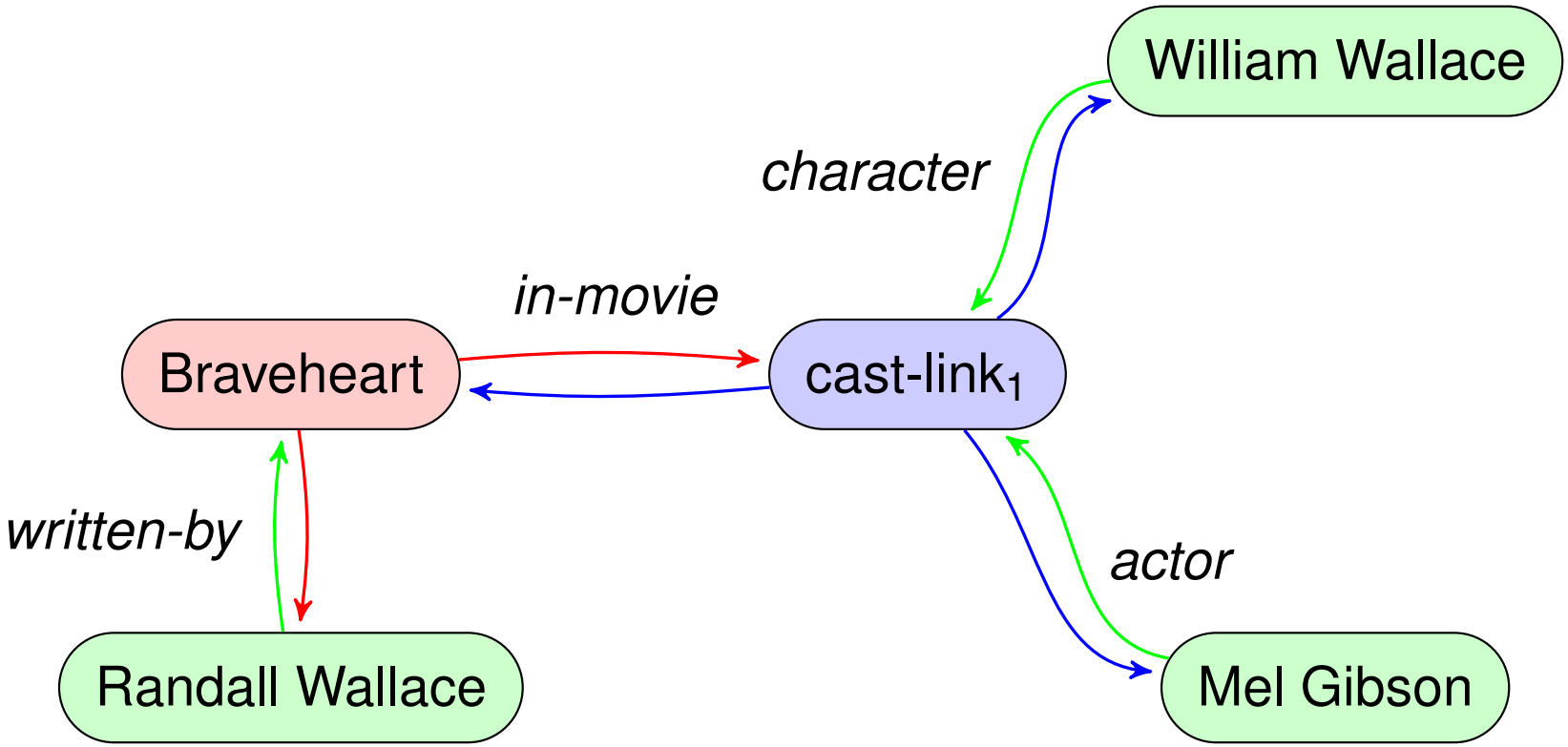
Connectable relations – Facts graph (1/2)

Build a graph from triples.

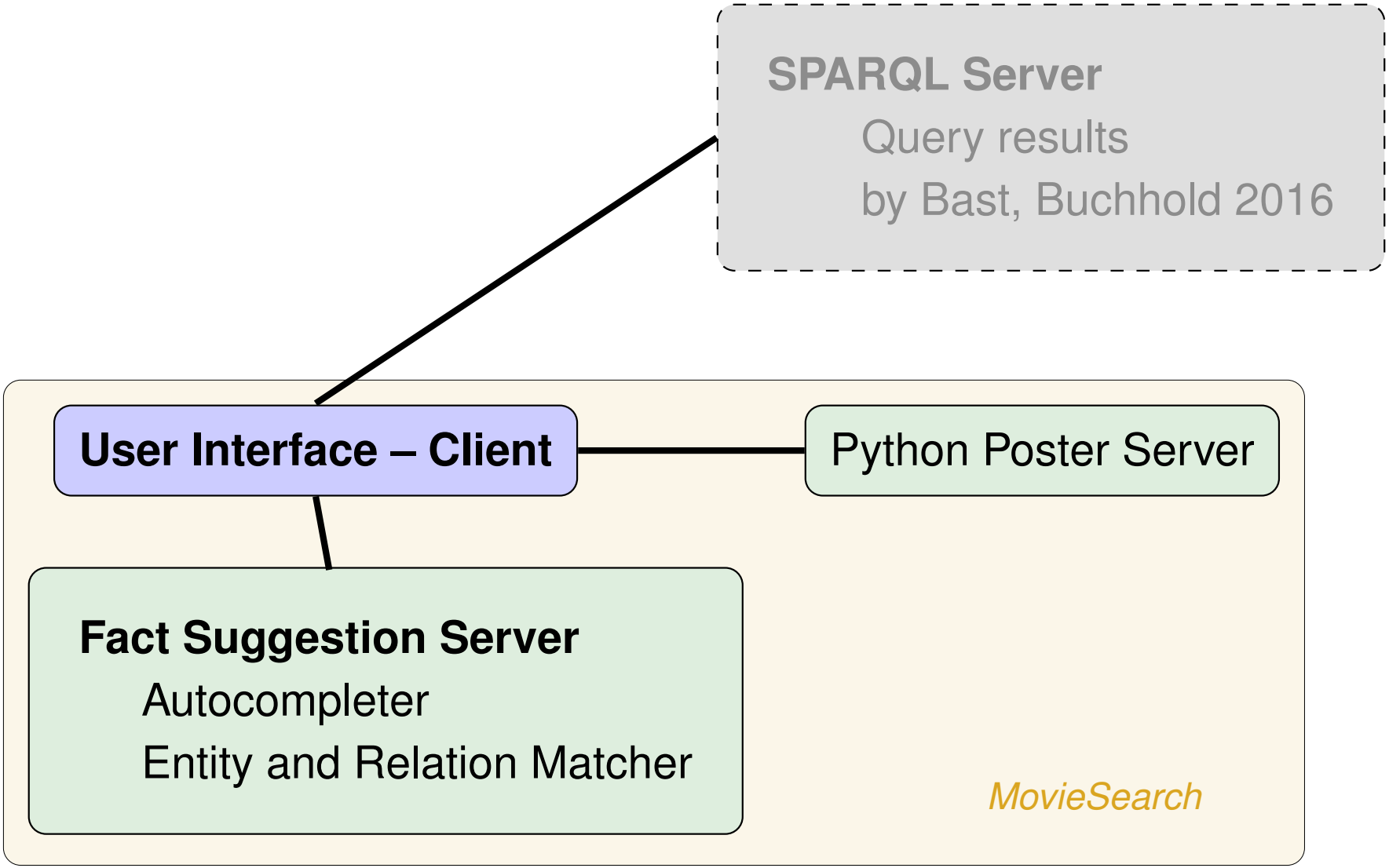
Example (Triples)

(Braveheart,	written-by,	Randall Wallace)
(cast-link$_1$,	in-movie,	Braveheart)
(cast-link$_1$,	actor,	Mel Gibson)
(cast-link$_1$,	character,	William Wallace)

Connectable relations – Facts graph (2/2)



Architecture



Usability evaluation

User study

User study

- 8 participants
- 21 tasks Q_i
- Build a query for each task

$\Rightarrow$ Evaluating usability

User study – Quality of query building

Evaluating the building process

Count number of text inputs needed to build a query.

	user study text inputs	minimum text inputs	relative user extra input
Avg. Q_i	**3.2**	**2**	**68%**

$\approx$ one extra input

User study – Input comparison

Q_{11} *"In which movies directed by Garry Marshall was Julia Roberts starring?"*.

Compare miniumum text inputs needed for Q_{11}:

Graph-based Systems	**Text inputs**
GoRelations	9
NotAnotherGoogleAnswer	6
SFC (Semantic Focused Crawler)	5
MovieSearch [minimum]	**2**
MovieSearch [study avg.]	2.8

[Styperek:2015] Evaluation of SPARQL-compliant semantic search UIs.

User study – Quality of the built queries

Results with the built queries in the user study:

Total query answers	168	
with expected results	159	**94.64%**
expected Result with expected Query	130	77.38%
expected Result with other Query	29	17.26%

Results of the queries: Quality and ranking (1/2)

MovieSearch vs. natural-language-based UI (Valossa)

- Usability $\rightarrow$ natural-language-based is main competitor
- Compare results for the 21 tasks Q_i:
  MovieSearch expected queries *vs.* Valossa task text input

- Regard Top 10 results
- Ranking quality via
  Discounted **C**umulative **G**ain, for $w_i \in \{0, 1\}$:

$$DCG_{10} := w_1 + \sum_{i=2}^{10} \frac{w_i}{\log_2 i}.$$

Results of the queries: Quality and ranking (2/2)

	avg. Recall	avg. Precision	avg. $NDCG_{10}$
MovieSearch	**66,60%**	**94,96%**	**95,81%**
Valossa	47,62%	35,00%	62,19%

- Recall → tasks more hits than 10

  Q_2 *Movies with songs from Hans Zimmer.*

- Valossa: answers without any hit

  Q_{20} *Movie with Angelina Jolie and Brad Pitt where they have secrets.*

  Q_7 *Movie that is 111 minutes long and released at 11.11.2011.*

- MovieSearch: hard criteria with facts
  Tradeoff: (high Precision) ▷ (potential for almost hits)

Results of the queries: Quality and ranking (2/2)

	avg. Recall	avg. Precision	avg. $NDCG_{10}$
MovieSearch	**66,60%**	**94,96%**	**95,81%**
Valossa	47,62%	35,00%	62,19%

- Recall $\rightarrow$ tasks more hits than 10

  Q_2 *Movies with songs from Hans Zimmer.*

- Valossa: answers without any hit

  Q_{20} *Movie with Angelina Jolie and Brad Pitt where they have secrets.*

  Q_7 *Movie that is 111 minutes long and released at 11.11.2011.*

- MovieSearch: hard criteria with facts
  Tradeoff: (high Precision) $\triangleright$ (potential for almost hits)

Results of the queries: Quality and ranking (2/2)

	avg. Recall	avg. Precision	avg. $NDCG_{10}$
MovieSearch	**66,60%**	**94,96%**	**95,81%**
Valossa	47,62%	35,00%	62,19%

- Recall $\rightarrow$ tasks more hits than 10

  Q_2 *Movies with songs from Hans Zimmer.*

- Valossa: answers without any hit

  Q_{20} *Movie with Angelina Jolie and Brad Pitt where they have secrets.*

  Q_7 *Movie that is 111 minutes long and released at 11.11.2011.*

- MovieSearch: hard criteria with facts
  Tradeoff: (high Precision) $\triangleright$ (potential for almost hits)

Results of the queries: Quality and ranking (2/2)

	avg. Recall	avg. Precision	avg. $NDCG_{10}$
MovieSearch	**66,60%**	**94,96%**	**95,81%**
Valossa	47,62%	35,00%	62,19%

- Recall $\rightarrow$ tasks more hits than 10

  Q_2 *Movies with songs from Hans Zimmer.*

- Valossa: answers without any hit

  Q_{20} *Movie with Angelina Jolie and Brad Pitt where they have secrets.*

  Q_7 *Movie that is 111 minutes long and released at 11.11.2011.*

- MovieSearch: hard criteria with facts
  Tradeoff: (high Precision) $\triangleright$ (potential for almost hits)

Goals achieved?

Query-building worked in a lot of the cases from the user study.

- *Provide user-friendly Interface*
  94% succesful answered tasks in user study
- *Utilize plot and facts*
  Splitting tasks – only problem affecting results
  $\Rightarrow$ More help from UI would be good
- *Support 3-ary relations* $\Rightarrow$ Better awareness

$\oplus$ Partial value matching

Goals achieved?

Query-building worked in a lot of the cases from the user study.

- *Provide user-friendly Interface*
  94% succesful answered tasks in user study
- *Utilize plot and facts*
  Splitting tasks – only problem affecting results
  $\Rightarrow$ More help from UI would be good
- *Support 3-ary relations* $\Rightarrow$ Better awareness

$\oplus$ Partial value matching

Goals achieved?

Query-building worked in a lot of the cases from the user study.

- *Provide user-friendly Interface*
  94% succesful answered tasks in user study
- *Utilize plot and facts*
  Splitting tasks – only problem affecting results
  $\Rightarrow$ More help from UI would be good
- *Support 3-ary relations* $\Rightarrow$ Better awareness

⊕ Partial value matching

Goals achieved?

Query-building worked in a lot of the cases from the user study.

- *Provide user-friendly Interface*
  94% succesful answered tasks in user study
- *Utilize plot and facts*
  Splitting tasks – only problem affecting results
  $\Rightarrow$ More help from UI would be good
- *Support 3-ary relations* $\Rightarrow$ Better awareness

⊕ Partial value matching

SPARQL Backend at
`https://github.com/Buchhold/SparqlEngineDraft`

[Styperek:2015]
STYPEREK, Adam ; CIESIELCZYK, Michal ; SZWABE, Andrzej ; MISIOREK, Pawel:
Evaluation of SPARQL-compliant semantic search user interfaces.
In: Vietnam Journal of Computer Science 2 (2015), Nr. 3, pp. 191–199., ISSN 2196–8896

Thank you for your attention.