Annotation-based Document Retrieval with Four-Valued Probabilistic Datalog

Ingo Frommholz, Ulrich Thiel, Thomas Kamps
Fraunhofer IPSI
Darmstadt, Germany
{frommholz|thiel|kamps}@ipsi.fraunhofer.de

Annotations…

…in our case: free text annotations

…support several tasks in digital libraries, for instance,

- Interpretation of and comments on document material
- Effective use of documents
- Authoring

… are a special kind of metadata

- They are connected to the object they annotate

… may form a hypertext

… establish a *document context*

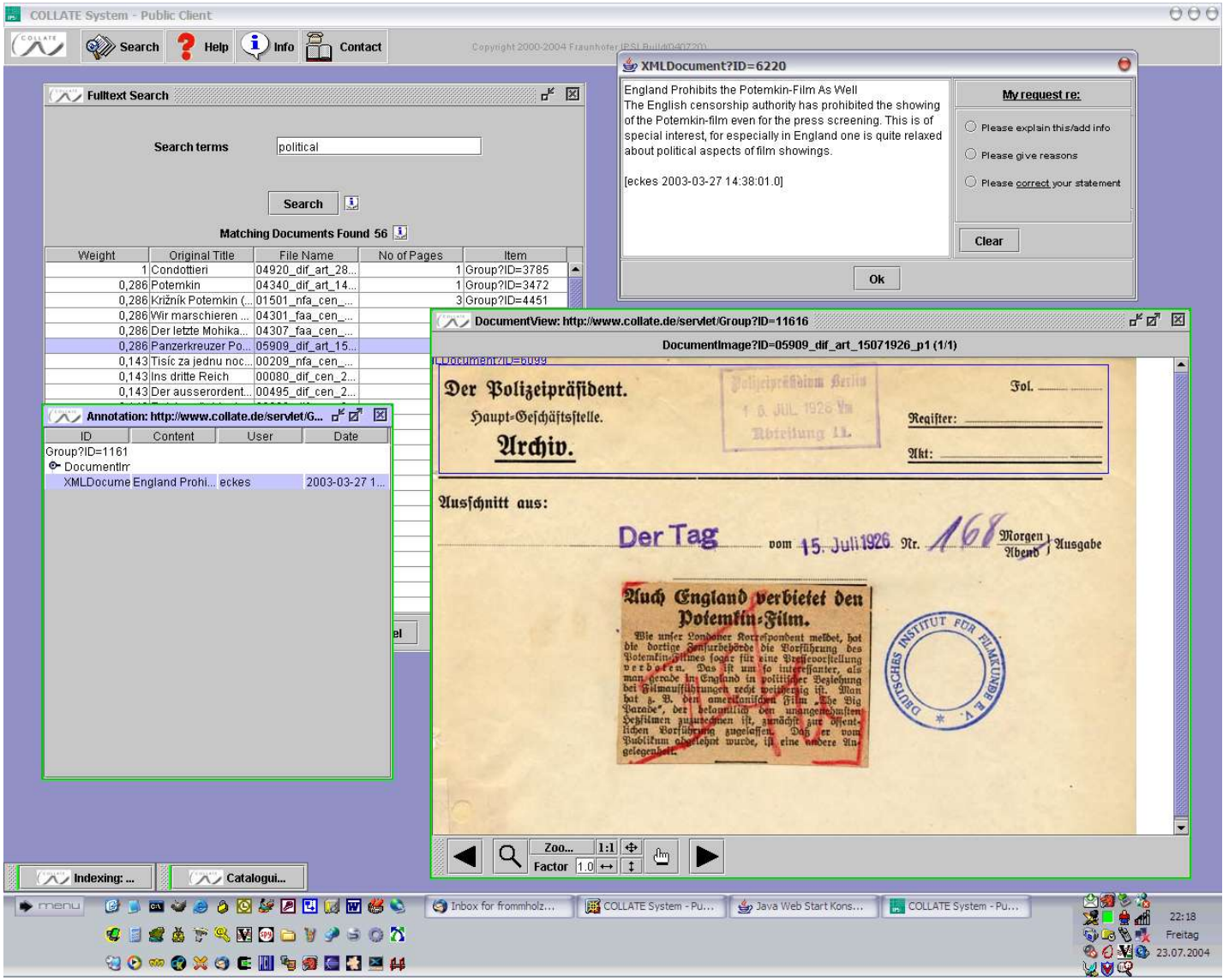
Annotations and Information Retrieval

- Address *vocabulary problem* (e.g., content summarised with other words)
- Reconstitute original document context (interpretations)
- Controlled context, immediately available
- Contain additional information

Our Scenario:

- COLLATE – Collaboratory for annotation, indexing and retrieval of digitized historical archive material
- Free text annotations as discussions about historic film material…how can we employ the results of such discussions?

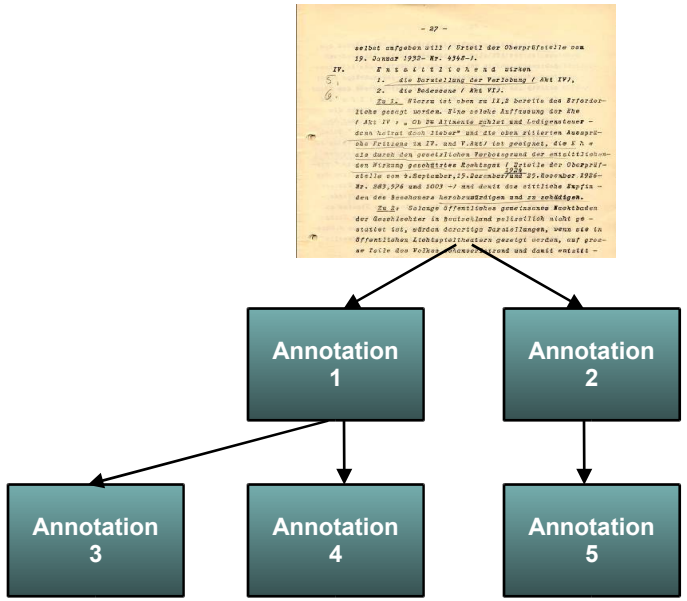
COLLATE


COLLATE System - Public Client
Search
Help
Info
Contact
Fulltext Search
Search terms
political
Search
Matching Documents Found 56
Weight
Original Title
File Name
No of Pages
Item
1 Condottieri 04920_dif_art_28... 1 Group?ID=3785
0,286 Potemkin 04340_dif_art_14... 1 Group?ID=3472
0,286 Križník Potemkin (... 01501_nfa_cen_... 3 Group?ID=4451
0,286 Wir marschieren ... 04301_faa_cen_...
0,286 Der letzte Mohika... 04307_faa_cen_...
0,286 Panzerkreuzer Po... 05909_dif_art_15...
0,143 Tisíc za jednu noc... 00209_nfa_cen_...
0,143 Ins dritte Reich 00080_dif_cen_2...
0,143 Der ausserordent... 00495_dif_cen_2...
Annotation: http://www.collate.de/servlet/G...
ID
Content
User
Date
Group?ID=1161
DocumentIm
XMLDocume
England Prohi...
eckes
2003-03-27 1...
XMLDocument?ID=6220
England Prohibits the Potemkin-Film As Well
The English censorship authority has prohibited the showing of the Potemkin-film even for the press screening. This is of special interest, for especially in England one is quite relaxed about political aspects of film showings.
[eckes 2003-03-27 14:38:01.0]
My request re:
Please explain this/add info
Please give reasons
Please correct your statement
Clear
Ok
DocumentView: http://www.collate.de/servlet/Group?ID=11616
DocumentImage?ID=05909_dif_art_15071926_p1 (1/1)
Der Polizeipräsident.
Haupt-Geschäftsstelle.
Archiv.
Fol.
Register:
Akt:
Ausschnitt aus:
Der Tag vom 15. Juli 1926 Nr. 168 Morgen Abend Ausgabe
Auch England verbietet den Potemkin-Film.
Zoo...
1:1
Factor 1.0
Indexing: ...
Catalogui...
Inbox for frommholz...
COLLATE System - Pu...
Java Web Start Kons...
COLLATE System - Pu...
22:18
Freitag
23.07.2004

The COLLATE Annotation Model

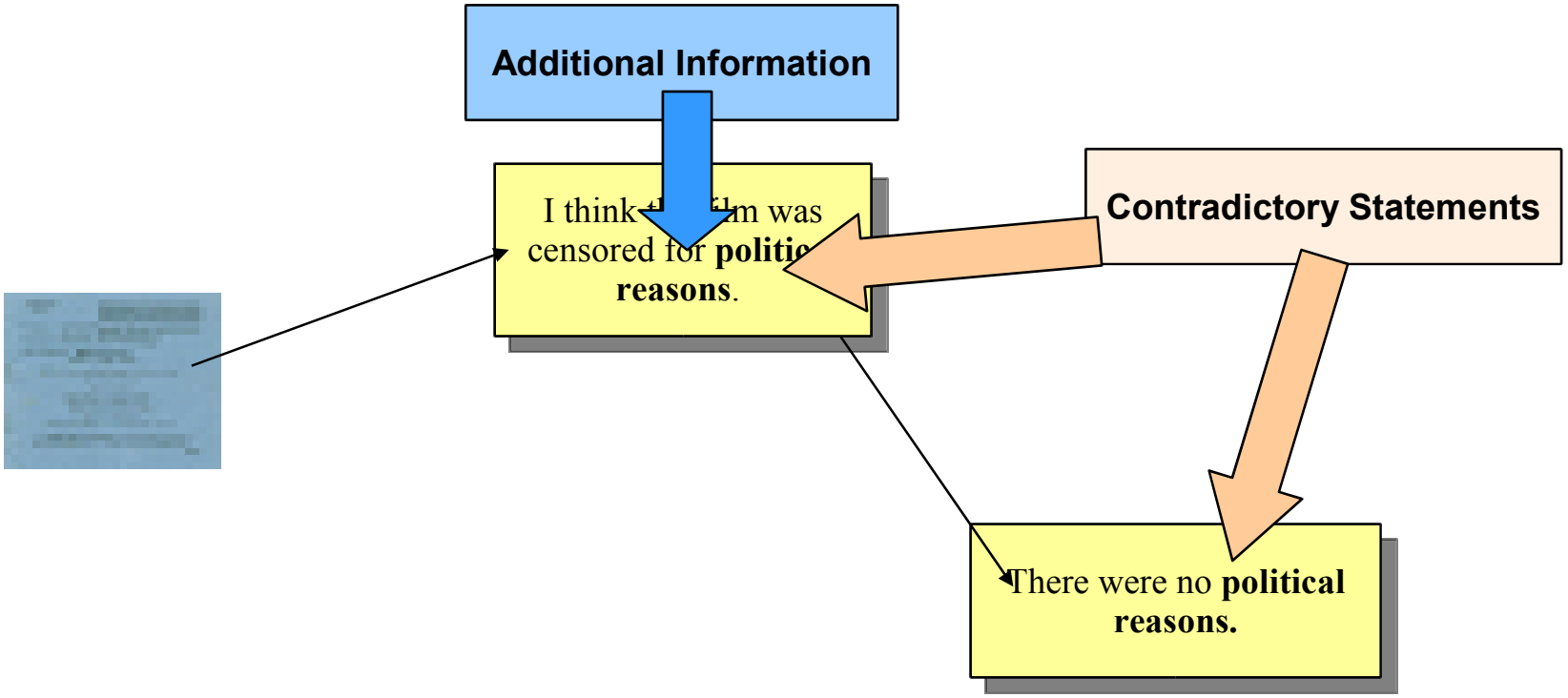
- **Annotation Thread** (nested annotations) constitutes scientific discourse

- Links are typed with **discourse structure relations**
  - Elaboration
  - Analogy
  - Difference
  - Cause
  - Background Information
  - Interpretation
  - Support Argument
  - Counterargument



Annotation-based Document Retrieval: An Example

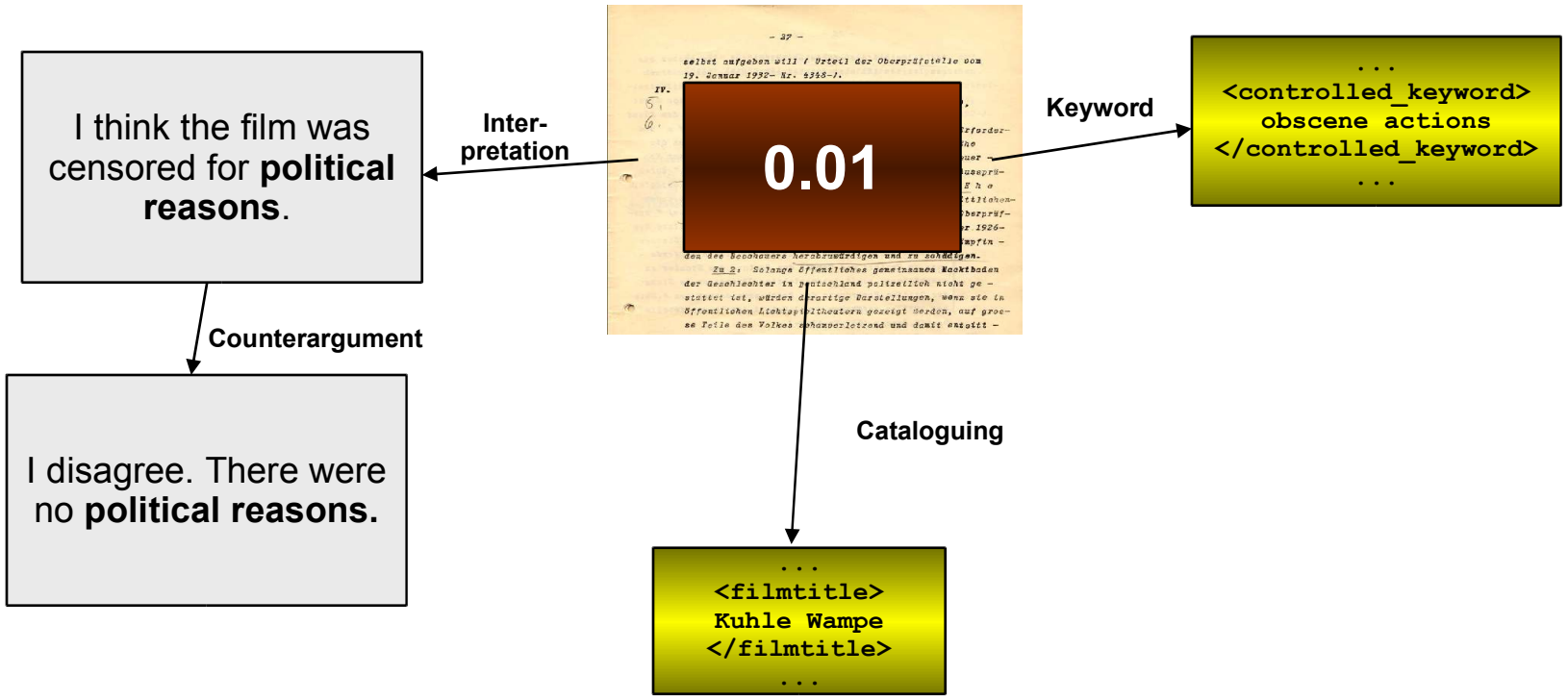
- What did the experts discuss about the document?
- In-depth analysis of annotation thread



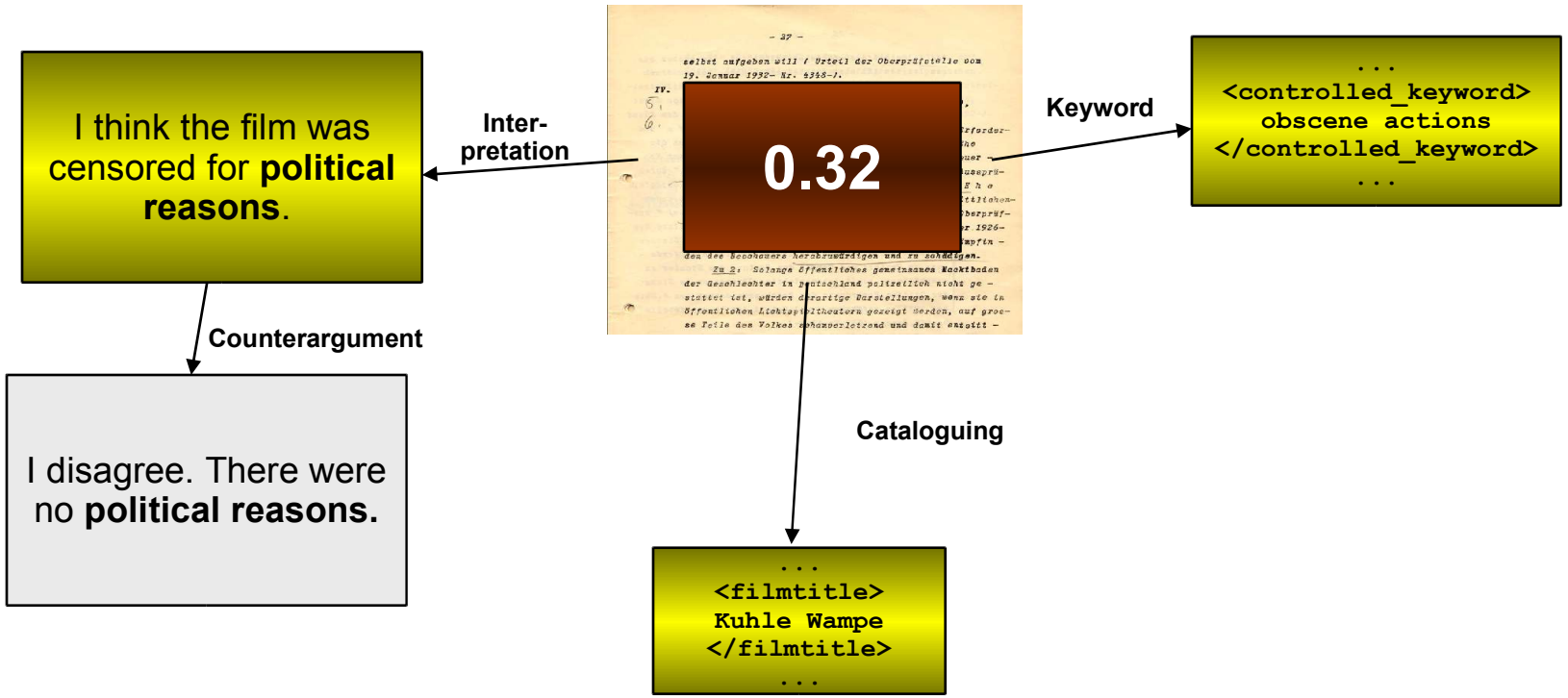
NARA: Nested Annotation Retrieval Approach

- Our Approach
  - NARA: **N**ested **A**nnotation **R**etrieval **A**pproach
  - Using annotations for document retrieval
  - Nested annotations
  - Statements are seen in the context of other statements
  - Negative and contradictory statements

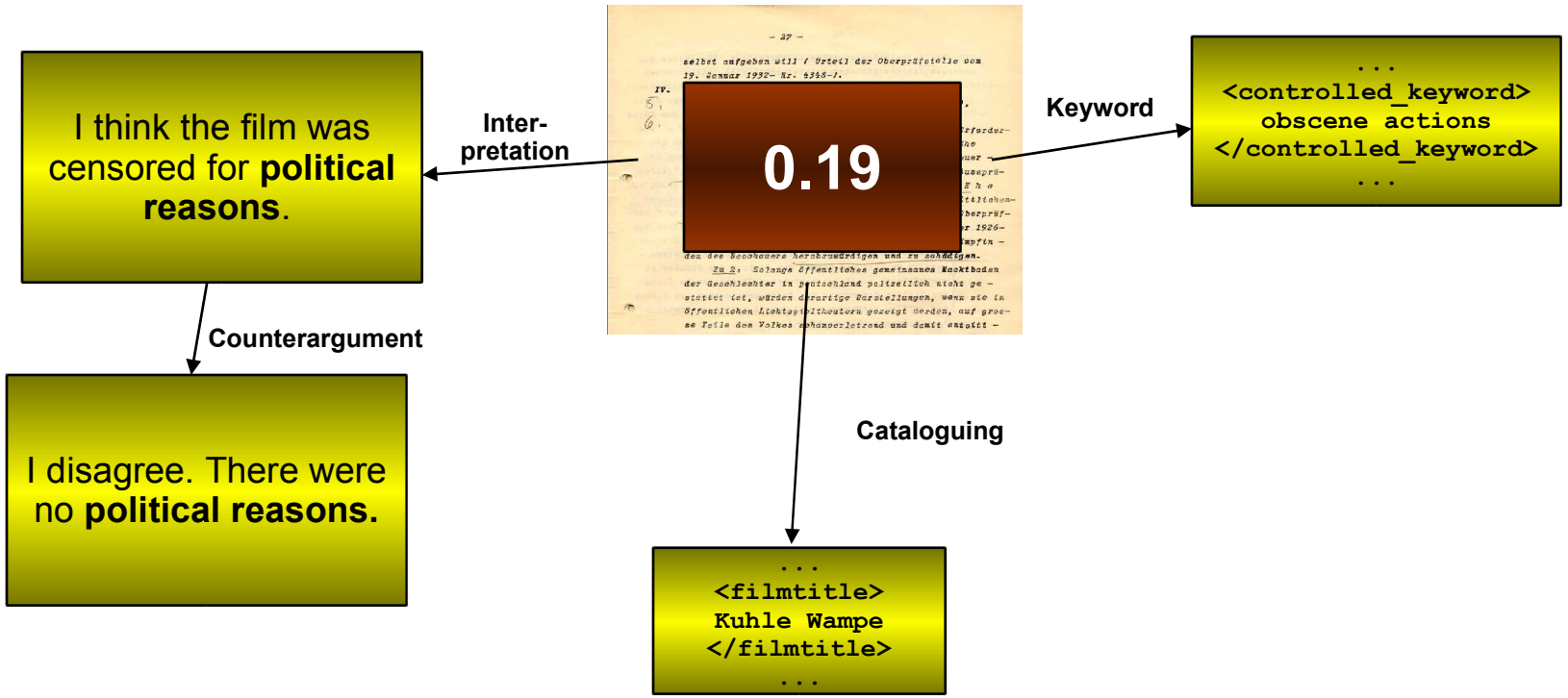
NARA: Example – Query for “political reasons”, only metadata



NARA: Example – direct annotation



NARA: Example – whole discourse



Assumption

- Atomicity of annotations w.r.t. their discourse relation

"There were no political but economical reasons"

(Counterargument plus interpretation)

"There were no political reasons"

(Counterargument)

&

"There were economical reasons"

(Interpretation)

NARAlog: Logic-based framework for implementing NARA

- Retrieval as probabilistic inference [van Rijsbergen 86]

$$P(d \rightarrow q)$$

- Open World Assumption:
  - Negative and positive evidence should be treated independently

- Based on four-valued probabilistic datalog (FVPD)

Four-Valued Probabilistic Datalog

[Roelleke/Fuhr 96]

- Knowledge Augmentation in Hypermedia Documents
- Four-Valued Probabilistic Datalog (FVPD):

  TRUE: {true}

  FALSE: {false}

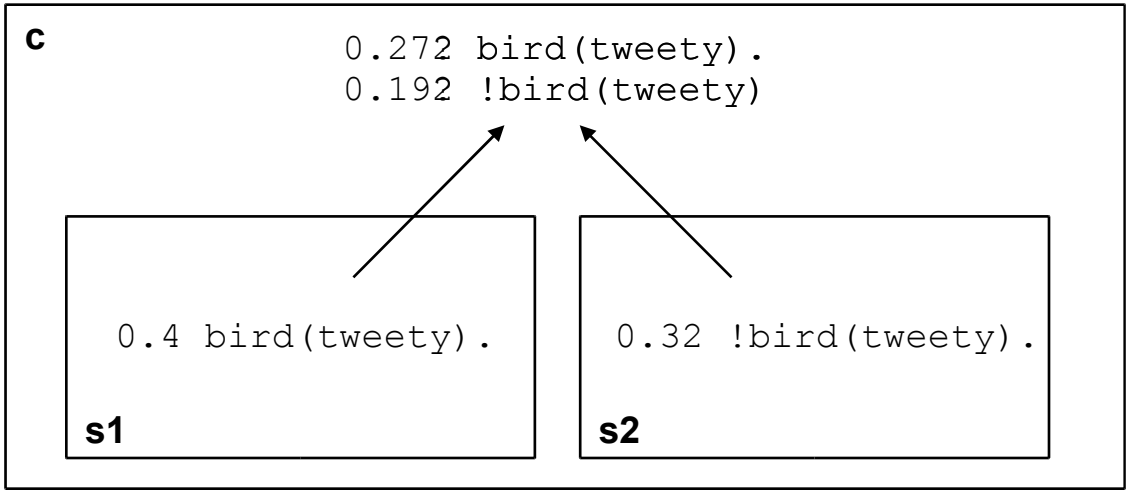
  INCONSISTENT: {true, false}

  UNKNOWN: {}

- Open World Assumption: the absence of an atom `p(a)` in the model does not imply `!p(a)`

- Handling of negative and contradictory statements

FVPD Example

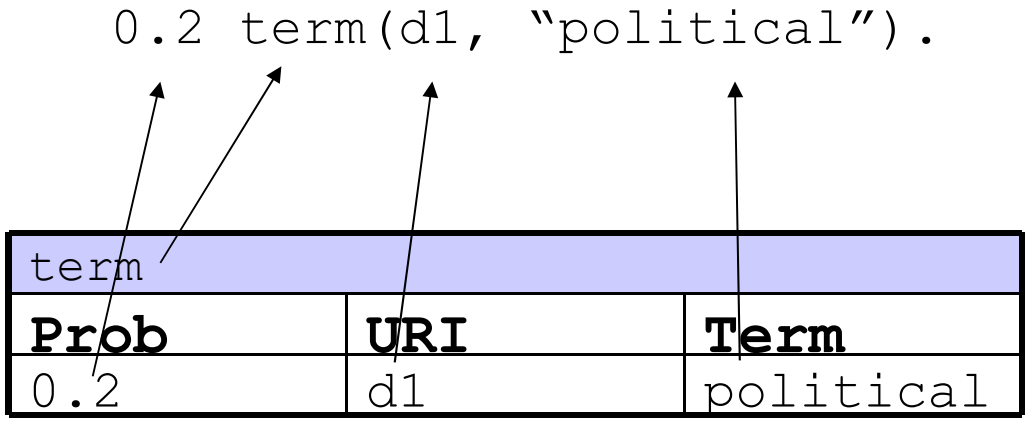


$$0.4 * (1 - 0.32) = 0.272$$
$$0.32 * (1 - 0.4) = 0.192$$

HySpirit

- HySpirit: Implementation of FVPD, combining database and IR technology

- Store and access probabilistic facts in relational database



NARAlog: Components

- Content-based indexing and retrieval
  - Documents, Annotations
  - Queries
- Context-based indexing and retrieval
  - Document and link types
  - Positive and negative links
  - Access probability

Content-based Indexing and Retrieval

- Indexing documents, annotations and queries as probabilistic facts

```
0.2 term(d1, "political").
0.3 term(d1, "censorship").

0.3 termspace("political").
0.5 termspace("censorship").

qterm("political").
qterm("reasons").
```

- Initial content-based retrieval

```
r_content(N) :- qterm(T) & termspace(T) & term(N,T).
```

Document and Link Types

- Document Types

```
document(d1).
annotation(a1).
annotation(a2).
```

- Link Types

```
interpretation(d1,a1).
counterargument(a1,a2).
```

- Positive and negative links: Categorise link types w.r.t. their effect on the retrieval weight of its source

```
pos_link(X,Y) :- interpretation(X,Y).
neg_link(X,Y) :- counterargument(X,Y).
```

Access Probability

- Probability that an annotation is actually accessed and considered

```
0.8 acc(d1,a1).
0.8 acc(a1,a2).
```

- Customisation w.r.t. user preferences

- Example: Emphasise one author's annotations, neglect another author's

Context-based Retrieval Function

- Annotation-based re-weighting
- Positive and negative evidence should be considered independently (→ Open World Assumption)
- Positive evidence in the node itself

```
r_nara(N1) :- r_content(N1).
```

- Positive evidence in successor nodes with positive link type

```
r_nara(N1) :- pos_link(N1,N2) & acc(N1,N2) & r_nara(N2).
```

- Negative evidence in successor nodes with negative link type

```
!r_nara(N1) :- neg_link(N1,N2) & acc(N1,N2) & r_nara(N2).
```

Example

```
0.2 term(d1, "political"). 0.3 term(d1, "censorship").
0.3 termspace("political"). 0.5 termspace("censorship").
qterm("political"). qterm("reasons").
document(d1). annotation(a1). annotation(a2).
```

Indexing of content, stored in database

```
r_content(X) :- qterm(T) & termspace(T) & term (T,X).
```

Initial content-based retrieval

```
interpretation(d1,a1).
counterargument(a1,a2).
0.8 acc(d1,a1).
0.8 acc(a1,a2).
pos_link(X,Y) :- interpretation(X,Y).
neg_link(X,Y) :- counterargument(X,Y).
```

Indexing of structural context, stored in database

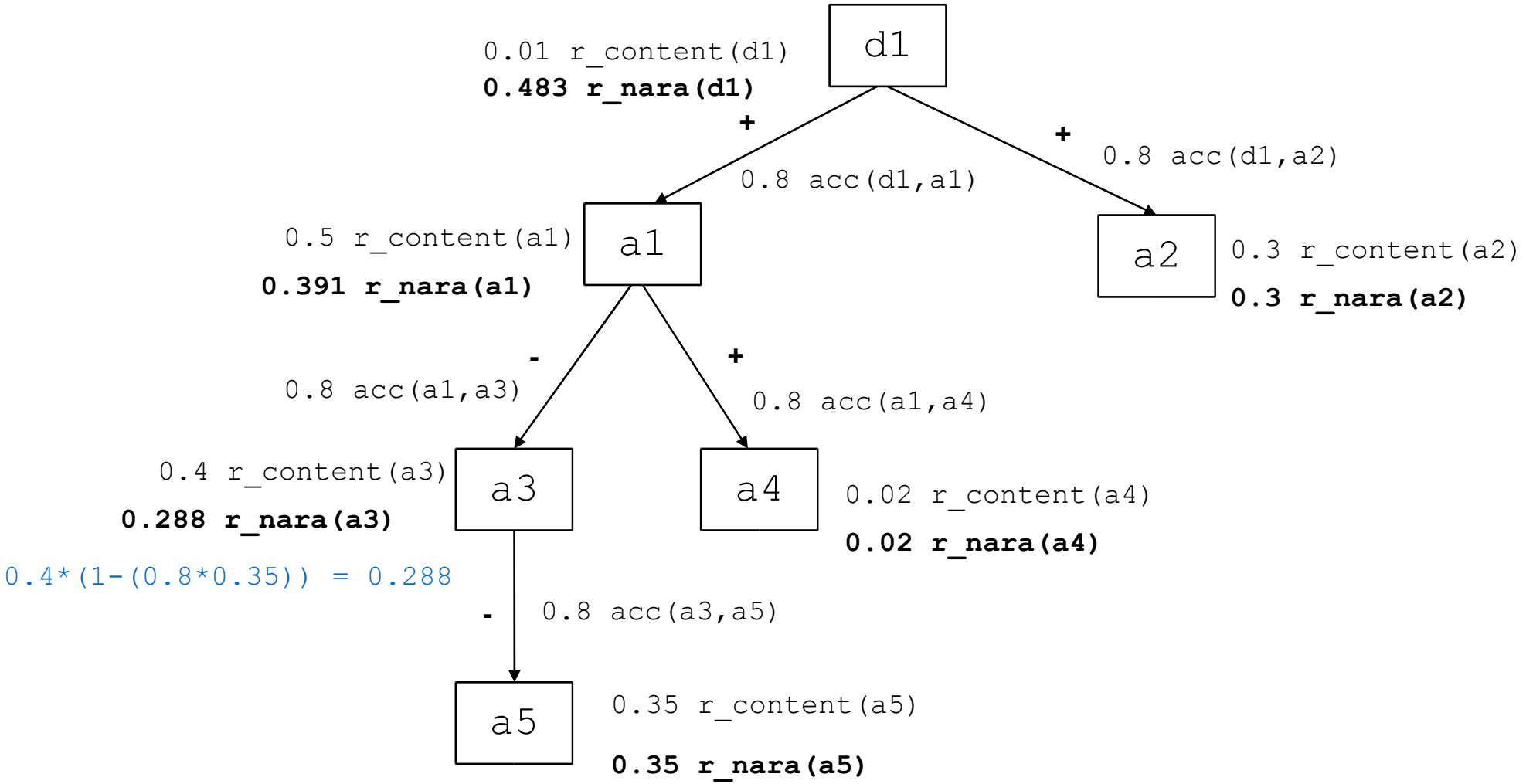
```
r_nara(X) :- r_content(X).
r_nara(X) :- pos_link(X,Y) & acc(X,Y) & r_nara(Y).
!r_nara(X) :- neg_link(X,Y) & acc(X,Y) & r_nara(Y).
```

Annotation-based re-weighting (context-based retrieval)

```
?- document(D) & r_nara(D)
```

$$P(d_1 \rightarrow q) = P(\texttt{r\_nara(d1)}) \cdot (1 - P(\texttt{!r\_nara(d1)}))$$

Example Execution



Future Work

- Lots of things to do!

- Experiments with the COLLATE collection
  →Evaluation of effectiveness and efficiency of the approach

- Atomicity of link types is not user-friendly
  →Methods to automatically recognise link types w.r.t. their effect (positive or negative)? [Marcu and Echihabi 02]
    →Probabilities for link types

- Suitable values for access probability `acc`
  →Assign a global value, but which one?

- Other possibilities to use annotations for information retrieval
  →Query expansion and relevance feedback? [Golovchinski et al. 99]
  →Newsgroups [Xi et al. 04]

Conclusion

- Annotations, as a certain kind of metadata, can be exploited for document retrieval
- NARA: Nested Annotation Retrieval Approach
  - Bias content-based RSV of documents
- NARAlog: Implementation of NARA based on FVPD
  - Analysis of annotation threads (i.e. nodes and typed links)
  - Can deal with contradictory and negative statements

Thank you for your attention!