Gianluca Demartini
demartini@L3S.de

L3S Research Center
Hannover, Germany

RETRIEVING ENTITIES IN WIKIPEDIA AND IN NEWS APPLICATIONS

Research Interests

Information Retrieval

- Entity Retrieval
  - In Wikipedia [IRJ10, ECIR11]
  - Over Time [SIGIR10a, CIKM10]
  - Using Query Logs [SIGIR10b, ECDL10]
- Evaluation
  - Measures and Initiatives [INEX08, INEX09]
- Search Dimensions
  - Sentiment and Diversification [SemSearch10, ECIR11demo]

Semantic Web

- Desktop Search [JWS10]
- Entity Identifiers and Matching [ESWC09, iiWAS10]

Outline

Research Interests

Entity Retrieval

- In Wikipedia
  - Task and INEX XER
  - A Semantic approach to ER
- Over Time
  - Motivation and Task
  - Dataset and Data analysis
  - Approaches to TAER
- Conclusions

Entity Retrieval in Wikipedia

[IRJ10]

Ranking People

Expert Finding in TREC-ENT (Enterprise Track)

Collection:

- Corpus: crawl of *.w3.org sites
- People: names of 1092 people who may be experts

Query:

- ‘`information retrieval`’

Results:

- A **list of people** who know about information retrieval

Ranking Actors

Queries are lists of actors on the Web, e.g.

- Query: 1930s
  - Answers: Fred Astaire, Charlie Chaplin, W.C. Fields, Errol Flynn, Clark Gable, Greta Garbo, etc
- Query: action
  - Answers: Arnold Schwarzenegger, etc

Ranking…

People

Actors

… Car companies

[i.e., insert your fav entity type here]

Entity Ranking!!!

Example Entity Ranking Scenarios

- Impressionist art museums in Holland
- Countries with the Euro currency
- German car manufacturers
- Artists related to Pablo Picasso
- Countries involved in WWI
- Actors who played Hamlet
- English monarchs who married French women

Entity Ranking

Topical query Q

Entity (result) type T_X

A list of entity instances Xs

An entity is represented by its Wikipedia page

Systems employ categories, structure, links

Topic 116

Q

Title
Italian Nobel prize winners

Xs

Entities
Dario Fo (#176791)
Renato Dulbecco (#744909)
Carlo Rubbia (#44932)

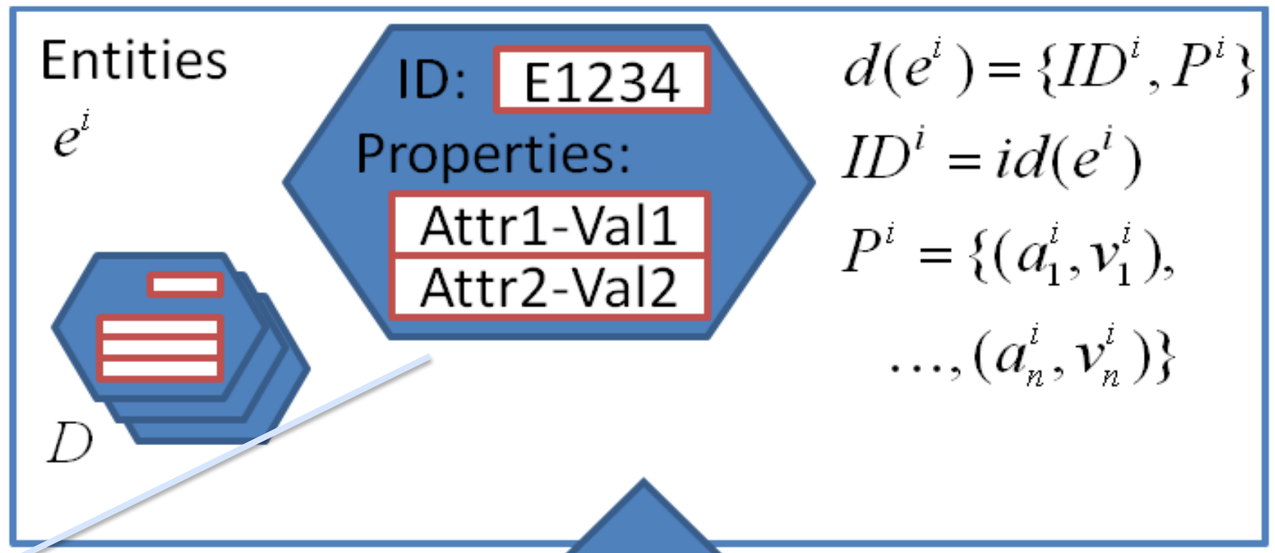
T_X

Categories
Nobel laureates (#924)

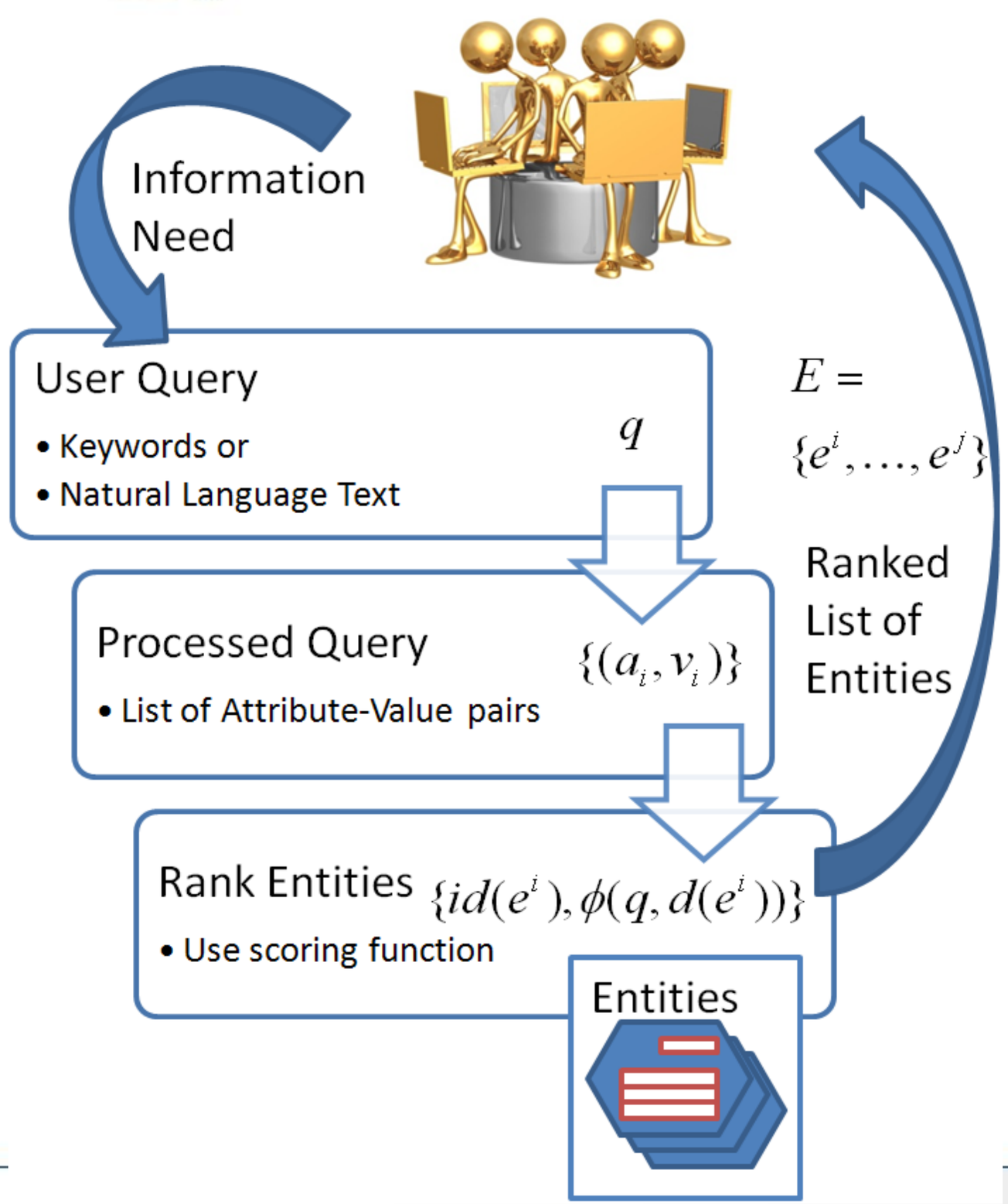
Description
I want all the Italian people who won the Nobel prize.

Narrative
I want a list of people who were Nobel prize laureates in any field and have Italian nationality.

- Indexing
  - Entities
  - Data Sources



"Alexandre Pato"
ID: ap12dH5a
(born in; 1989)
(playing with; acm15hDJ)



- Searching
  - Users' Information Need
  - Entity Ranking System

Approaches to ER in Wikipedia

Exploit and refine the category structure

- WordNet to find entity types (e.g., a professor is a person)

Extend the query

- Synonyms and related words (Wordnet synsets)

Exploit the link structure

- Links in Wikipedia are usually entities
- Search Keywords also in anchor text of outLinks

Predict topic difficulty

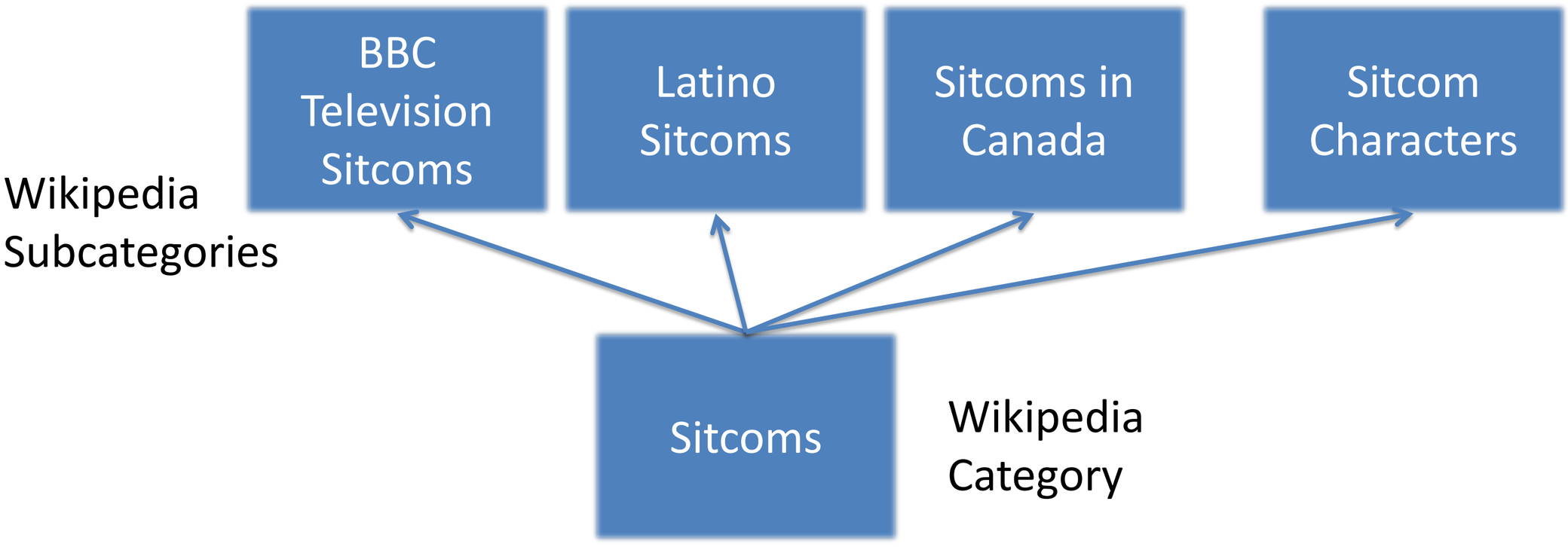
- Adapt model parameters

Category Based Search

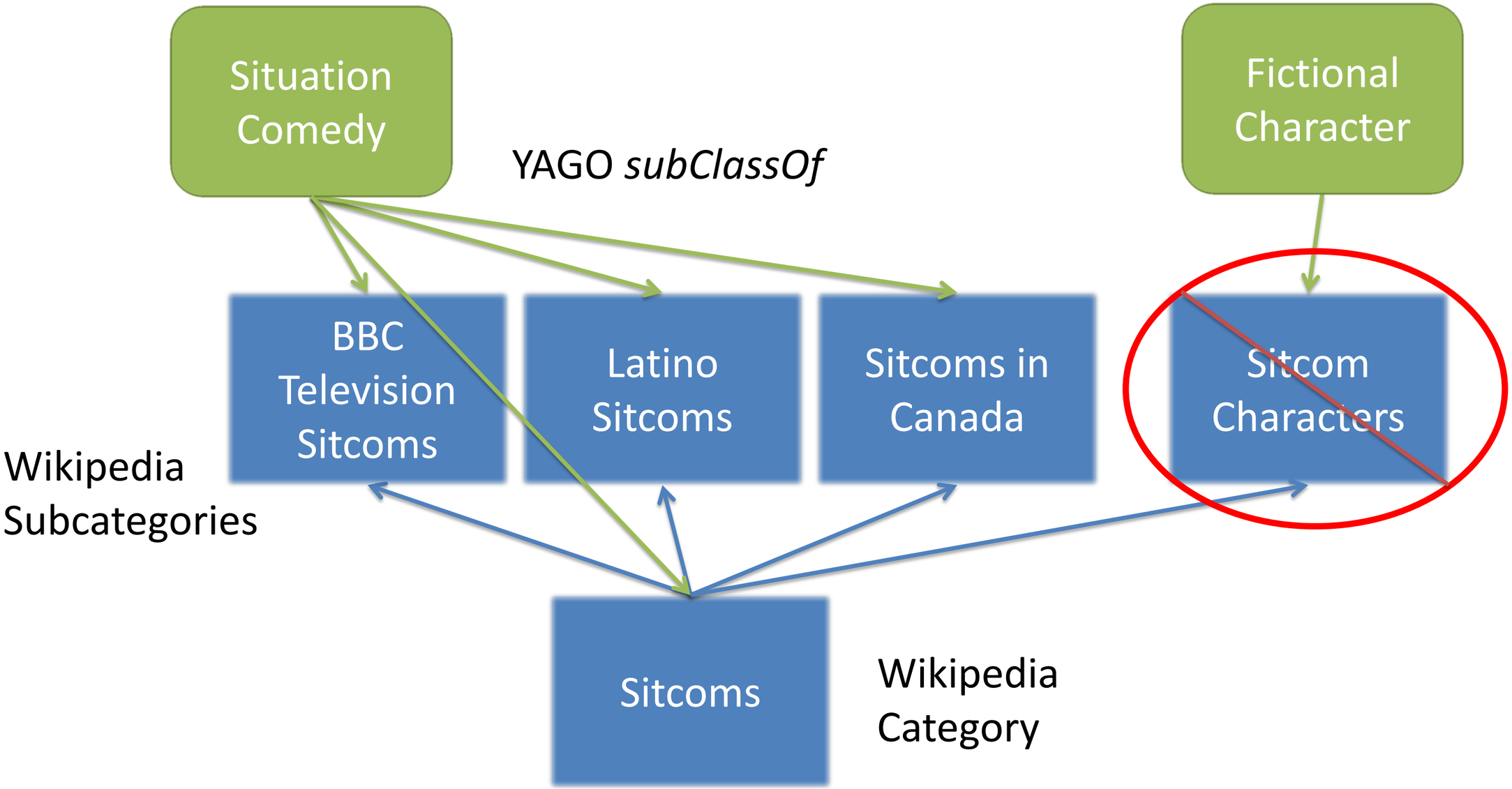
Query expansion by modifying category information

- Subcategories
  - Extracted from Wikipedia
- “Children” Categories
  - Filtered using the YAGO subClassOf relation
- “Sibling” Categories
  - Extracted from Wikipedia
  - Having with the same YAGO type

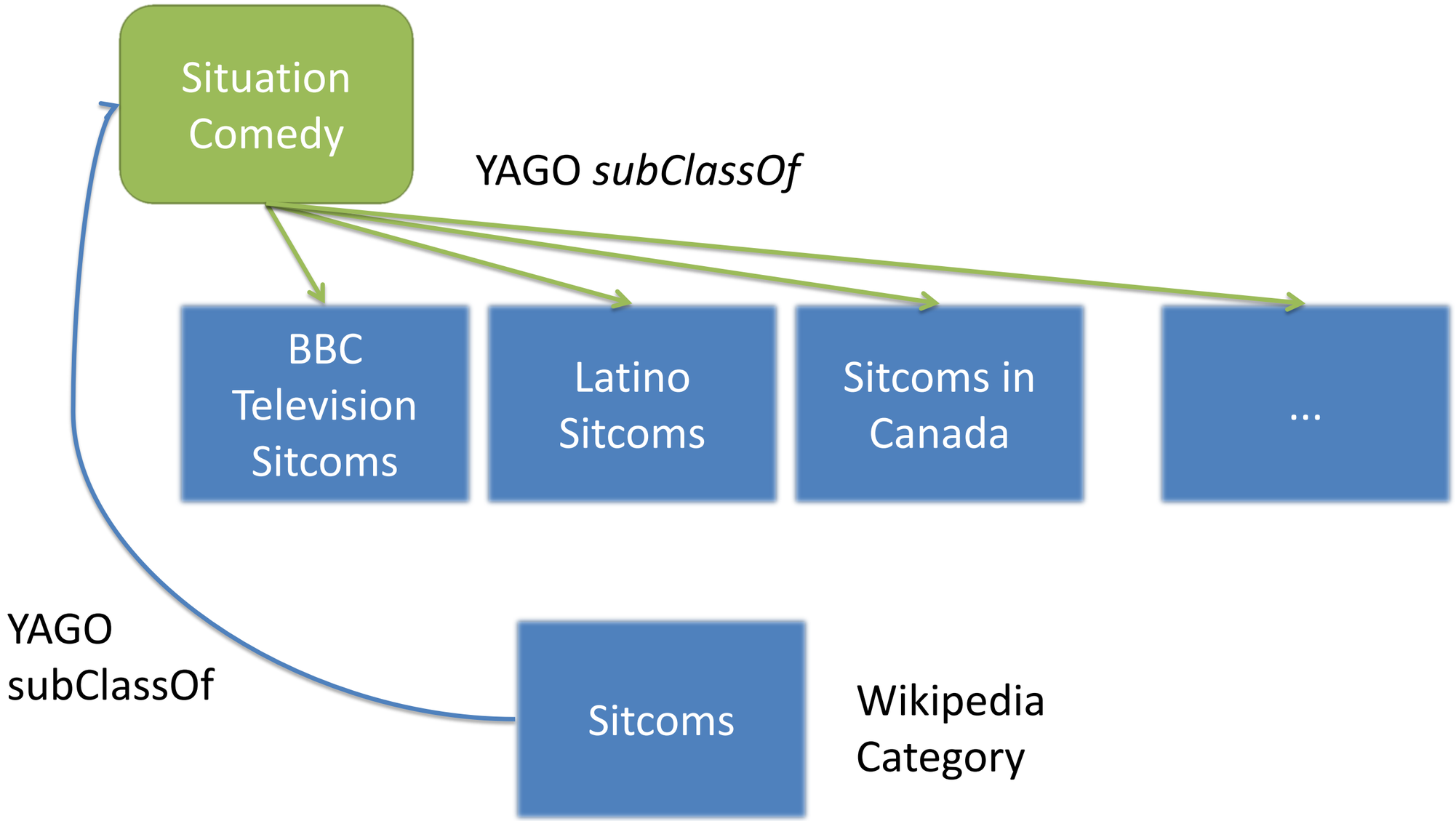
Subcategories



“Children” Categories



“Sibling” Categories



NLP techniques for ER

- INEX query
  - Keywords, Category
- Synonyms and Related Words
  - Query extension: synonyms of nouns in the Keywords + Word Sense Disambiguation for the correct meaning
- Core Characteristics
  - Clean the Keywords removing terms (and synonyms) appearing in Category
  - Keep only nouns and adjectives in Keywords (lexical compounds)
- Named Entities
  - Use only NE (i.e., organizations, locations, persons) from Keywords

Title	Tom Hanks movies where he plays a leading role.
Category	Films
Synonyms	Tom ”Uncle Tom” Hanks ”Thomas J. Hanks” movies film flick ”motion picture” ”motion-picture show” ”moving picture” pic picture ”picture show” ”moving-picture show” where he plays a leading role
Related Words	**Synonyms** plus 50 additional concepts related mainly to motion pictures
Core Characteristics	Tom Hanks leading role
Named Entities	Tom Hanks

Evaluating ER in Wikipedia

INEX Entity (XER) track 2007-2009

- http://www.inex.otago.ac.nz/tracks/entity-ranking/entity-ranking.asp

Standard test collection using

- Wikipedia dump from 2006
- Wikipedia dump from 2009 + extracted entities and types from Wordnet

Queries and manual relevance judgements

Evaluation measures to compare sytems

Table 8 Average precision and precision for the first 10 results for NLP based techniques for the XER task

Nr	Query; $q = \{category, W^C\} \cup \ldots$	xInfAP	P@10
1	$\{text, W^T\}$	0.2350	0.3057
9	$\{text, W^T\}, \{outLinks, W^T\}$	0.2556*	0.3371*
10	$\{text, W^T\}, \{outLinks, CC(W^T)\}$	0.2511	0.3114
11	$\{text, W^T\}, \{outLinks, NE(W^T)\}$	0.2504*	0.3171
12	$\{LexComp(W^T)\}$	0.2284	0.2971
13	$\{text, W^T \cup LexComp(W^T)\}$	0.2506	0.3257
14	$\{text, W^T \cup LexComp(W^T)\}, \{outLinks, W^T \cup LexComp(W^T)\}$	0.2616	0.3457
15	$\{text, W^T \cup SY(W^T)\}$	0.2439*	0.3257
16	$\{text, W^T \cup RW(W^T)\}$	0.2398	0.3199
17	$\{text, W^T \cup CC(W^T)\}$	0.2509*	0.3257
18	$\{text, W^T \cup NE(W^T)\}$	0.2530*	0.3257
19	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$	0.2705*	0.3571*
20	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}, \{outLinks, CC(W^T)\}$	0.2682*	0.3599*
21	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}, \{category, W^T\}$	**0.2909***	**0.3971***

Discussion

Ranking Entities in Wikipedia

- may help answering complex user queries
- can be done by exploiting the structure in Wikipedia
- must deal with the poor quality of category assignement

Combining Links,NLP,NER techniques we achieve 35% (MAP) and 53% (P10) improvement over normal search

Recent work deals with entities on the Web

- TREC Entity aims at evaluating “Related entity finding”
- Most successful systems exploit Wikipedia

Time Aware Entity Retrieval

[SIGIR10a, CIKM10]

Motivation

Going beyond document retrieval

Finding entities relevant to a query in a document collection (e.g., Wikipedia)

In collections of documents **over time**

- Decide about relevance at document level (TAER)
- Analyse and exploit relevance evolution

Scenario

An event

- Charles Schulz dies

Get Relevant Docs

Entities

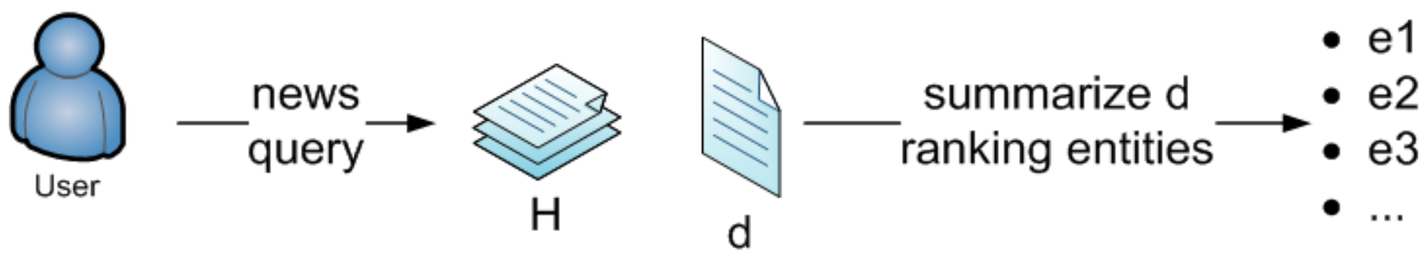
- Peanuts, his wife, media companies, hometown, other cartoonists, ...

Timeline of relevant news:

- 10/1999-09/2000:
  - 11/99 cancer diagnosted
  - 12/99 he retires
  - 02/00 he dies
  - 03/00 peanuts future discussed
  - ... Honors, museums, statues, airports, ...

Time Aware Entity Retrieval (TAER)

Find the set of entities e_i that best describe document *d* wrt a query *q* **given history d_i<*d***



Charles Schulz Dies Search

```
Important Entities:
- Charles_Schulz
- Congressional_Gold_Medal
- Santa_Rosa
- Peanuts
```

AP Online
02-15-2000
House Honors 'Peanuts' Creator

WASHINGTON (AP) -- ``Peanuts'' creator Charles Schulz was remembered today as a genius who touched the lives of millions of Americans as the House adopted a resolution to award him a Congressional Gold Medal.

The 77-year-old cartoonist died in his sleep Saturday at his Santa Rosa, Calif., home, a day before Schulz's last strip featuring Snoopy and the gang was published. He had announced in November he

Dataset

TREC Novelty Track 2004

- Sentence retrieval
- 25 event topics
- 779 **relevant** news

Entity annotations (7481 entities)

- Persons (26%), Locations (10%), Organizations (57%), Products (7%)

Relevance judgements

- Of each entity wrt to topic in this current news
- 21,213 judgements on 3 levels
- Cohen’s Kappa 0.59

Data Analysis

- P(e is Rel) 0.411 [0.404-0.417]
- P(e is NotRel) 0.168 [0.163-0.173]

How useful is to look at the past?

- $P(e|d_1)$ 0.893 [0.881-0.905]
- $P(e|d_{-1})$ 0.701 [0.677-0.726]

Local Features

Feature	P3	P5	MAP
F(e,d)	**.65**	**.56**	**.60**
FirstSenLen	.37	.36	.45
FirstSenPos	.31	.31	.43
F_{subj}	.49	.44	.50
AvgBM25s	.27	.30	.41
SumBM25s	.50	.44	.52

Feature	P3	P5	MAP
All Tied	.34	.34	.42

Exploiting the Past

Look at previous documents

- Entity occurences so far **F(e,H)**
- Docs where the entity appeared so far **DF(e,H)**
- Entity occurences in the previous doc **$F(e,d_{-1})$**
- Frequency of entity the first time? **$F(e,d_1)$**
- Number of other entities with which the entity co-occured so far **CoOcc(e,H)**

History Features

Feature	P3	P5	MAP
F(e,d)	.65	.56	.60
$F(e,d_1)$	.58	.53	.56
$F(e,d_{-1})$	.64	.56	.62*
F(e,H)	**.66**	**.59****	**.66****
CoOcc(e,H)	.62	.57	.65**
DF(e,H)	.63	.57*	.65**

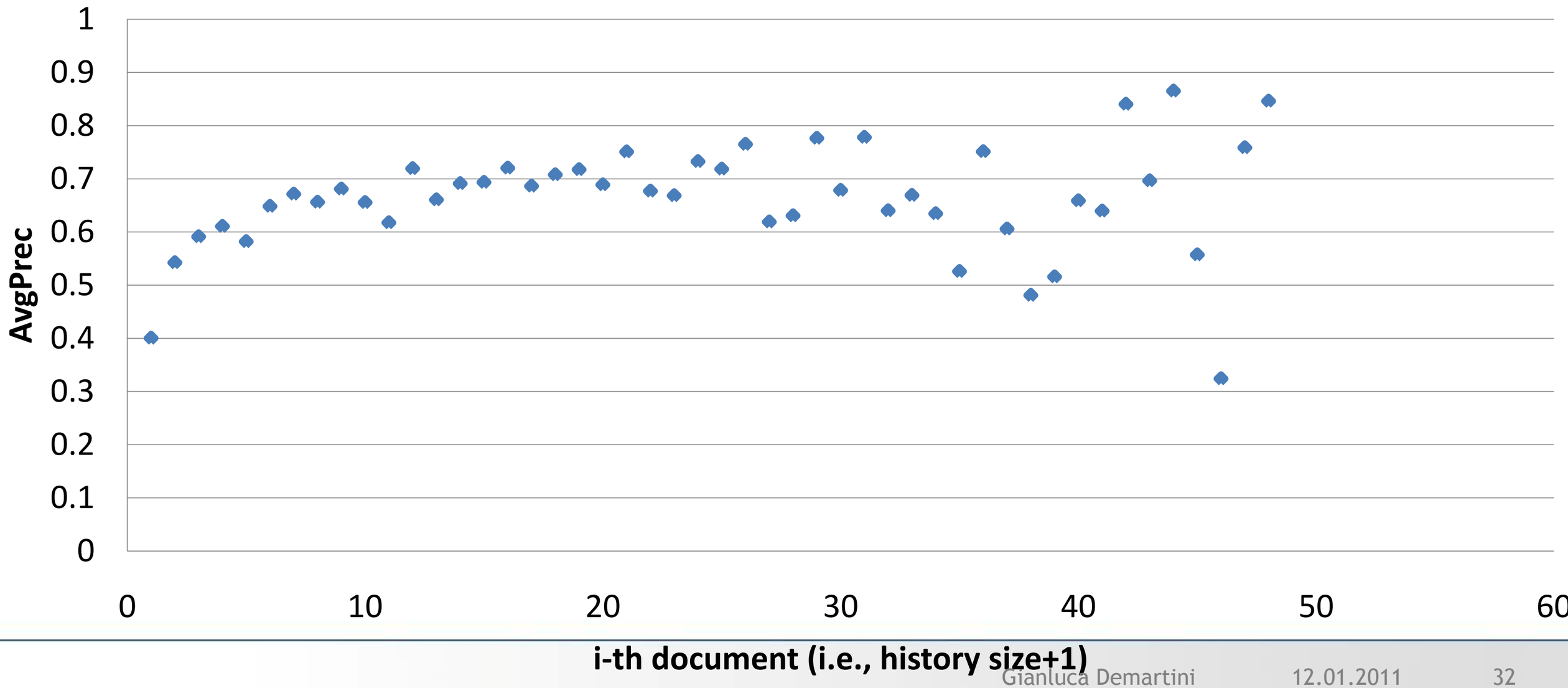
We also tried

- Weight history features with doc length
- Weight history features with BM25

Using the History

Conclusion

- Evidence from past documents is very important
- Effectiveness should improve over time (run F(e,H))



Combining Features with ML

Logistic Regression for ranking entities

2-folds cross validation on 25 topics

Similar results for combinations of 2 features

Local Doc Features
F(e,d)
FirstSenLen
FirstSenPos
F_{subj}
AvgBM25s
SumBM25s

History Features
$F(e,d_1)$
$F(e,d_{-1})$
F(e,H)
CoOcc(e,H)
DF(e,H)

Features	P3	P5	AvgPrec
F(e,d)	.65	.56	.60
Local	.65	.56	.62
History	.66	.60	.67
All	.69	.62	.68

Discussion

Defined new task: TAER

Constructed evaluation benchmark

Investigated some features and combinations

Conclusions

- Information from the past helps most
- Obtain 15% improvement over F(e,d)

Conclusions

Returning Entities as answer to a user query

- lowers the user effort in looking for the answer
- enables new functionalities (e.g., “hot entities”)

Current approaches

- exploit the structure in Wikipedia
- deal with different collections (e.g., Web, news)

Future work could

- try to predict future relevance of entities
- consider events and entities together

Thanks to

Roi Blanco

Claudiu S. Firan

Tereza Iofciu

Ralf Krestel

Malik Muhammad Saad Missen

Wolfgang Nejdl

Hugo Zaragoza

References

[IRJ10] Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, and Wolfgang Nejdl. **Why Finding Entities in Wikipedia is Difficult, Sometimes.** In: "Information Retrieval" 13(5): 534-567, Special Issue on Focused Retrieval and Result Aggregation, Springer, October 2010.

[ECIR11] Tereza Iofciu, Gianluca Demartini, Nick Craswell, and Arjen P. de Vries. **ReFER: effective Relevance Feedback for Entity Ranking**. To appear in: 33rd European Conference on Information Retrieval (ECIR 2011), Dublin, Ireland, April 2011.

[SIGIR10a] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. **Entity Summarization of News Articles.** In: 33rd Annual ACM SIGIR Conference (SIGIR 2010 poster session), Geneva, Switzerland, July 2010.

[CIKM10] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. **TAER: Time Aware Entity Retrieval.** In: The 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, October 2010.

References

[SIGIR10b] Bodo Billerbeck, Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, and Ralf Krestel. **Exploiting Click-Through Data for Entity Retrieval.** In: 33rd Annual ACM SIGIR Conference (SIGIR 2010 poster session), Geneva, Switzerland, July 2010.

[ECDL10] Bodo Billerbeck, Tereza Iofciu, Gianluca Demartini, Claudiu S. Firan, and Ralf Krestel. **Ranking Entities Using Web Search Query Logs.** In: 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2010), Glasgow, Scotland, September 2010.

[INEX08] Gianluca Demartini, Arjen P. de Vries, Tereza Iofciu, and Jianhan Zhu. **Overview of the INEX 2008 Entity Ranking Track.** In: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008 Dagstuhl Castle, Germany, December, 2008.

[INEX09] Gianluca Demartini, Tereza Iofciu, and Arjen P. de Vries. **Overview of the INEX 2009 Entity Ranking Track**. In: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009 Brisbane, Australia, December 2009.

References

[SemSearch10] Gianluca Demartini and Stefan Siersdorfer. **Dear Search Engine: What's your opinion about...? - Sentiment Analysis for Semantic Enrichment of Web Search Results**. In: Semantic Search 2010 Workshop located at the 19th Int. World Wide Web Conference WWW2010, Raleigh, NC, USA, April 2010.

[ECIR11demo] Gianluca Demartini. **ARES: A Retrieval Engine based on Sentiments - Sentiment-based Search Result Annotation and Diversification**. To appear in: 33rd European Conference on Information Retrieval (ECIR 2011 - Demo), Dublin, Ireland, April 2011.

[JWS10] Enrico Minack, Raluca Paiu, Stefania Costache, Gianluca Demartini, Julien Gaugaz, Ekaterini Ioannou, Paul-Alexandru Chirita, and Wolfgang Nejdl. **Leveraging Personal Metadata for Desktop Search: The Beagle++ System.** In: Journal of Web Semantics, 8(1): 37-54, Elsevier, March 2010.

.

References

[ESWC09] Julien Gaugaz, Jakub Zakrzewski, Gianluca Demartini, and Wolfgang Nejdl. **How to Trace and Revise Identities.** In: 6th Annual European Semantic Web Conference (ESWC2009), Heraklion, Greece, June, 2009.

[iiWAS10] George Papadakis, Gianluca Demartini, Philipp Kärger and Peter Fankhauser. **The Missing Links: Discovering Hidden Same-as Links among a Billion of Triples**. In: The 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS2010), Paris, France, November 2010