Leveraging Knowledge Graphs for Web Search

Part 3 - Searching for Entities

Gianluca Demartini
University of Sheffield
gianlucademartini.net

Course Outline

- Part I – **Introduction to Knowledge Graphs**
- Part II – **Named Entity Recognition and Linking to Knowledge Graphs**
- Part III – **Searching for Entities**
- Part IV – **Crowdsourcing for Knowledge Graphs**
- **Slides here: gianlucademartini.net/kg**

Outline

- Expert Finding
- Entity Ranking
- Ad-hoc Object Retrieval
- Evaluation Collections
- Open Challenges

Entity Oriented Search

- All those search tasks that aim at retrieving as answer to a user query an *entity* instead of a document
  - *People, Countries, Movies, Restaurants, etc.*

Google tom cruise

Web Images Maps Shopping News More Search tools

6 personal results. 188,000,000 other results.

[Tom Cruise - IMDb]
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Top Gun. If you had told 14 year old Franciscan seminary student Thomas Cruise Mapother IV that one day in the not too distant future he ...
Filmography by year - Biography - Rock of Ages - All You Need Is Kill

[Tom Cruise - Wikipedia, the free encyclopedia]
https://en.wikipedia.org/wiki/**Tom_Cruise**
Thomas Cruise Mapother IV (/ˈtɒməs ˈkruːz ˈmeɪpɒθər/; born July 3, 1962), widely known as **Tom Cruise**, is an American film actor and producer. He has ...
Tom Cruise filmography - Katie Holmes - Mimi Rogers - List of awards and ...

[Official Tom Cruise: Oblivion, Movies, Video, Biography, News ...]
www.**tomcruise**.com/
Official **Tom Cruise** site: Get the latest Rock of Ages trailer, info & downloads! Watch career movie trailers, videos, and retrospective. Read the **Tom Cruise** ...

[TomCruise.com (TomCruise) on Twitter]
https://twitter.com/**TomCruise**
The latest from **TomCruise**.com (@**TomCruise**). Official http://**TomCruise**.com TeamTC tweets. Does Tom Tweet? Sometimes between family & movies & its ...

[Tom Cruise | Facebook]
https://www.facebook.com/official**tomcruise**
Tom Cruise. 3883109 likes · 76956 talking about this. Welcome to the Official www.**TomCruise**.com team Facebook page! **Tom Cruise** news, events, pics & video ...

Tom Cruise

Follow

Actor

Thomas Cruise Mapother IV, widely known as Tom Cruise, is an American film actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Taps. Wikipedia

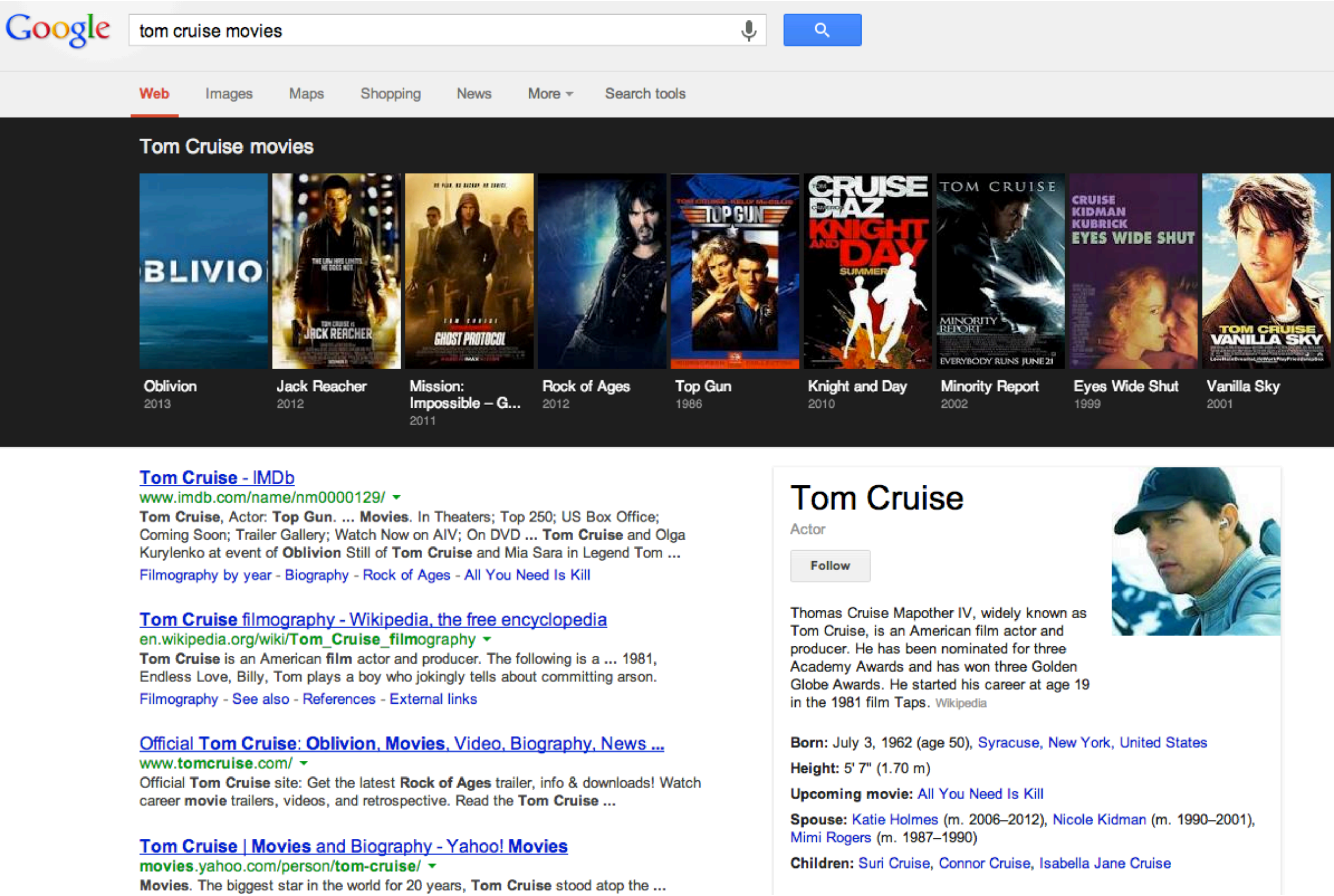
Born: July 3, 1962 (age 50), Syracuse, New York, United States

Height: 5' 7" (1.70 m)

Upcoming movie: All You Need Is Kill

Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

Children: Suri Cruise, Connor Cruise, Isabella Jane Cruise


Google
tom cruise movies
Web
Images
Maps
Shopping
News
More
Search tools
Tom Cruise movies
Oblivion 2013
Jack Reacher 2012
Mission: Impossible – G... 2011
Rock of Ages 2012
Top Gun 1986
Knight and Day 2010
Minority Report 2002
Eyes Wide Shut 1999
Vanilla Sky 2001
Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Top Gun. ... Movies. In Theaters; Top 250; US Box Office; Coming Soon; Trailer Gallery; Watch Now on AIV; On DVD ... Tom Cruise and Olga Kurylenko at event of Oblivion Still of Tom Cruise and Mia Sara in Legend Tom ...
Filmography by year - Biography - Rock of Ages - All You Need Is Kill
Tom Cruise filmography - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Tom_Cruise_filmography
Tom Cruise is an American film actor and producer. The following is a ... 1981, Endless Love, Billy, Tom plays a boy who jokingly tells about committing arson.
Filmography - See also - References - External links
Official Tom Cruise: Oblivion, Movies, Video, Biography, News ...
www.tomcruise.com/
Official Tom Cruise site: Get the latest Rock of Ages trailer, info & downloads! Watch career movie trailers, videos, and retrospective. Read the Tom Cruise ...
Tom Cruise | Movies and Biography - Yahoo! Movies
movies.yahoo.com/person/tom-cruise/
Movies. The biggest star in the world for 20 years, Tom Cruise stood atop the ...
Tom Cruise
Actor
Follow
Thomas Cruise Mapother IV, widely known as Tom Cruise, is an American film actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Taps. Wikipedia
Born: July 3, 1962 (age 50), Syracuse, New York, United States
Height: 5' 7" (1.70 m)
Upcoming movie: All You Need Is Kill
Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)
Children: Suri Cruise, Connor Cruise, Isabella Jane Cruise

Entities in SERP

rihanna concerts

Circa 66'300'000 risultati (0,46 secondi)

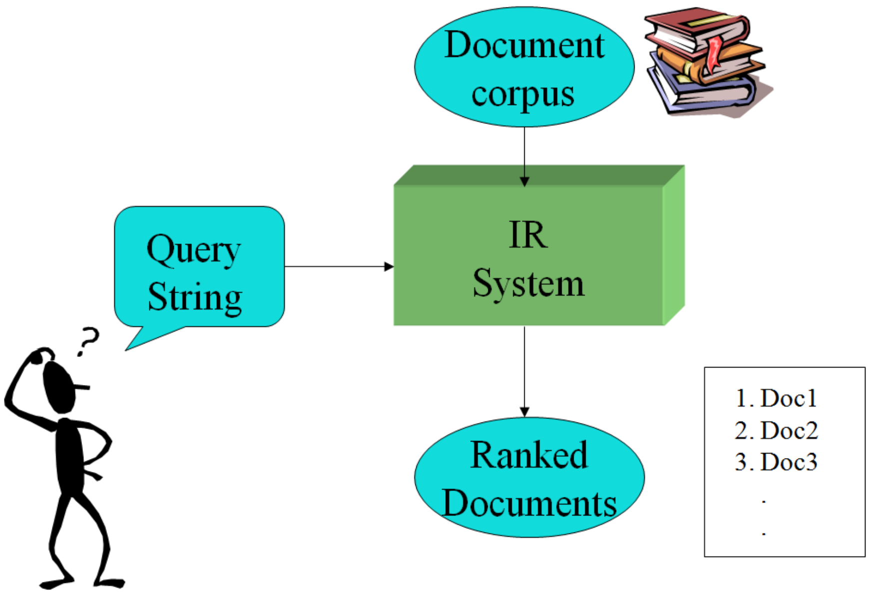
Rihanna Tour Dates 2012 — **Rihanna Concert** Dates and Tickets ...
www.songkick.com/artists/139648-**rihanna** - Traduci questa pagina
Find **Rihanna** live **concert** tour dates, tickets, reviews, and more on Songkick. Be the first to know when **Rihanna** is playing live in your town!
↳ 3 upcoming concerts - With Ke$ha and Travie McCoy - Media

Rihanna tickets, **concerts** and tour dates. Official Ticketmaster site.
www.ticketmaster.co.uk/**Rihanna**.../1013826 - Traduci questa pagina
Results 1 - 7 of 7 – Find and buy **Rihanna** tickets at Ticketmaster.co.uk.

dom 8 lug Barclaycard Wireless - Rihanna - Day ... - Hyde Park London, GB
dom 8 lug Barclaycard Wireless - Rihanna ... - Hyde Park London, GB
dom 8 lug Barclaycard Wireless 2012 - Disabled ... - Hyde Park London, GB

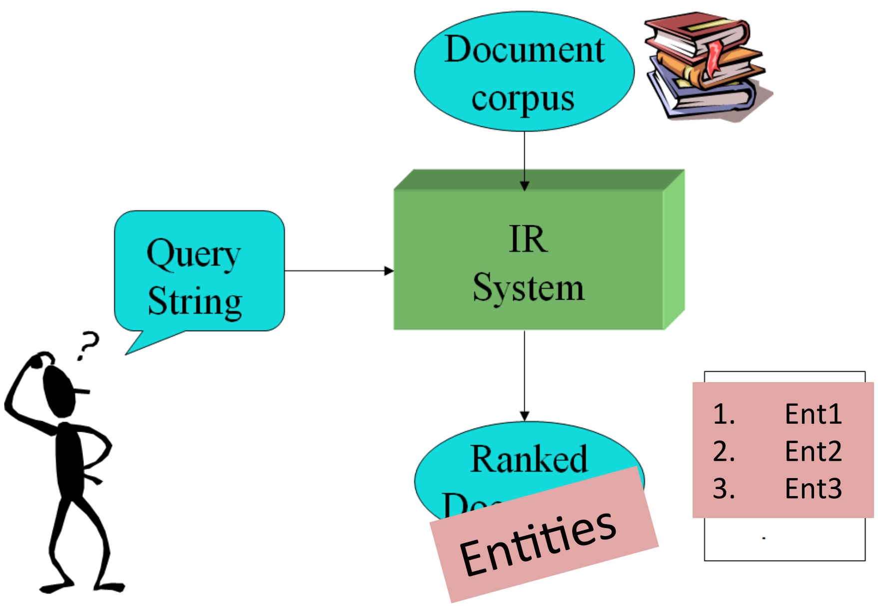
From Documents to Entities

- Document Search



From Documents to Entities

- Entity Search



Entity Search Tasks

- Expert Finding
- Entity Ranking, List Completion
- Related Entity Finding
- Ad-hoc Object Retrieval

Expert Finding

Expert Finding - Motivation

- Scenario
  - In large companies competencies and skills are spread
  - Executives need to create a team for a new project: find staff with the right expertise
  - Someone needs to solve a problem
  - Example: I need an expert on ontology engineering

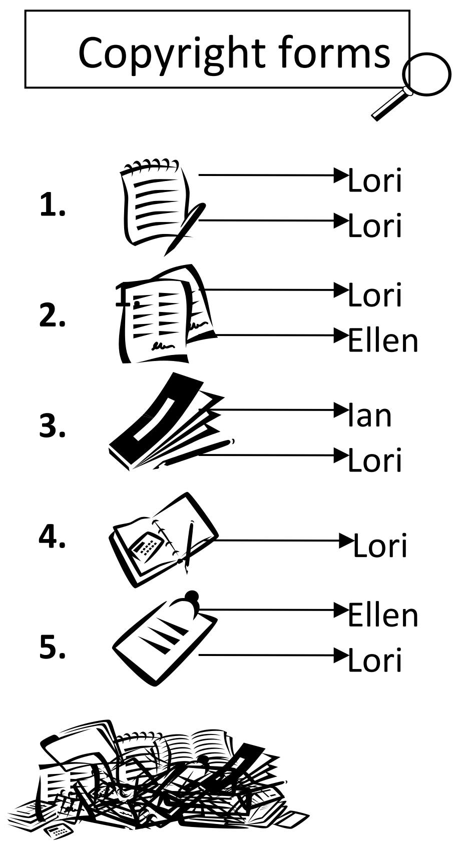
Expert Finding - Motivation

- Goal
  - Use the digital content available in the enterprise
  - Create a ranking of people who are experts in the given topic

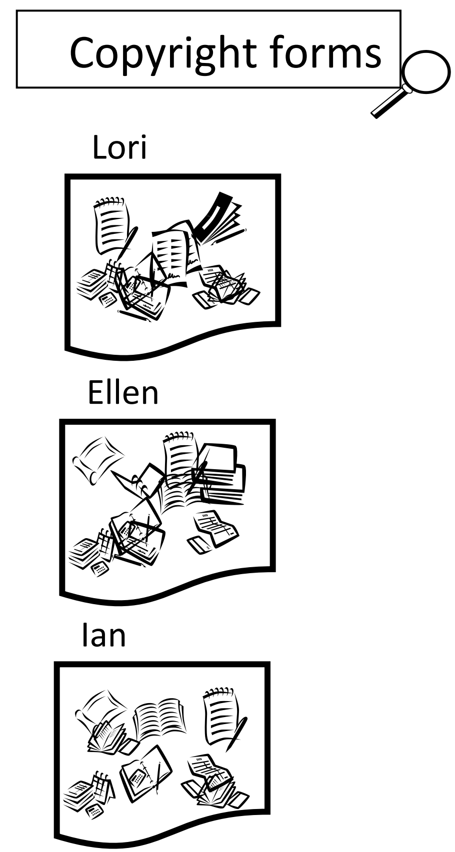
Two Basic Approaches

Who should I ask about the copyright forms?

- Document-based: rank docs, extract experts

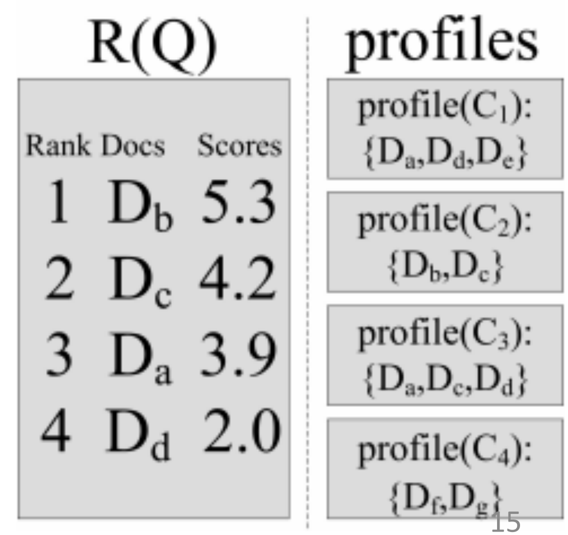


- Candidate-based: rank candidate profiles



Voting model

- Data fusion techniques
- Each ranked document represents a vote for the expertise of a candidate
- Vote aggregation:
  - Number of docs voting for each candidate
  - Scores of retrieved documents
  - Ranks of retrieved documents



Craig Macdonald, Iadh Ounis: Voting for candidates: adapting data fusion techniques for an expert search task. CIKM 2006: 387-396

User-Oriented Model

- Additional real-world constraints

- Distance between user and expert
  - User previous knowledge on the topic
  - Contact time (organizational hierarchy, geo location, collaboration)

Elena Smirnova, Krisztian Balog: A User-Oriented Model for Expert Finding. ECIR 2011: 580-592

Entity Ranking

Ranking…

- People
- Actors
- … Car companies

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

Entity Ranking!!!

Entities in Wikipedia

- Art museums
- Countries
- Actors, Singers
- Monarchs
- Artists
- Magicians
- ...

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

Approaches to ES 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

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, Springer, October 2010.

Entity Search over Wikipedia

- Search for many different entity types with one system!

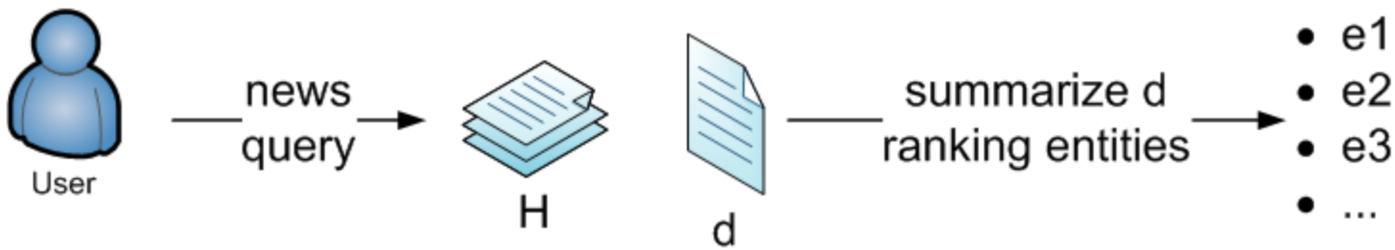
- Open issues
  - No temporal evolution of content is considered

Time-Aware Entity Retrieval

- In some cases the time dimension is available
  - News collections
  - Blog postings
- News stories evolve over time
  - Entities appear/disappear
  - Analyse and exploit relevance evolution
  - Decide about relevance at document level
- An Entity Search system can exploit the past to find relevant entities

Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, Hugo Zaragoza. TAER: Time Aware Entity Retrieval. CIKM 2010, Toronto, Canada.

Time-Aware Entity Retrieval



Charles Schulz Dies | Search

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

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

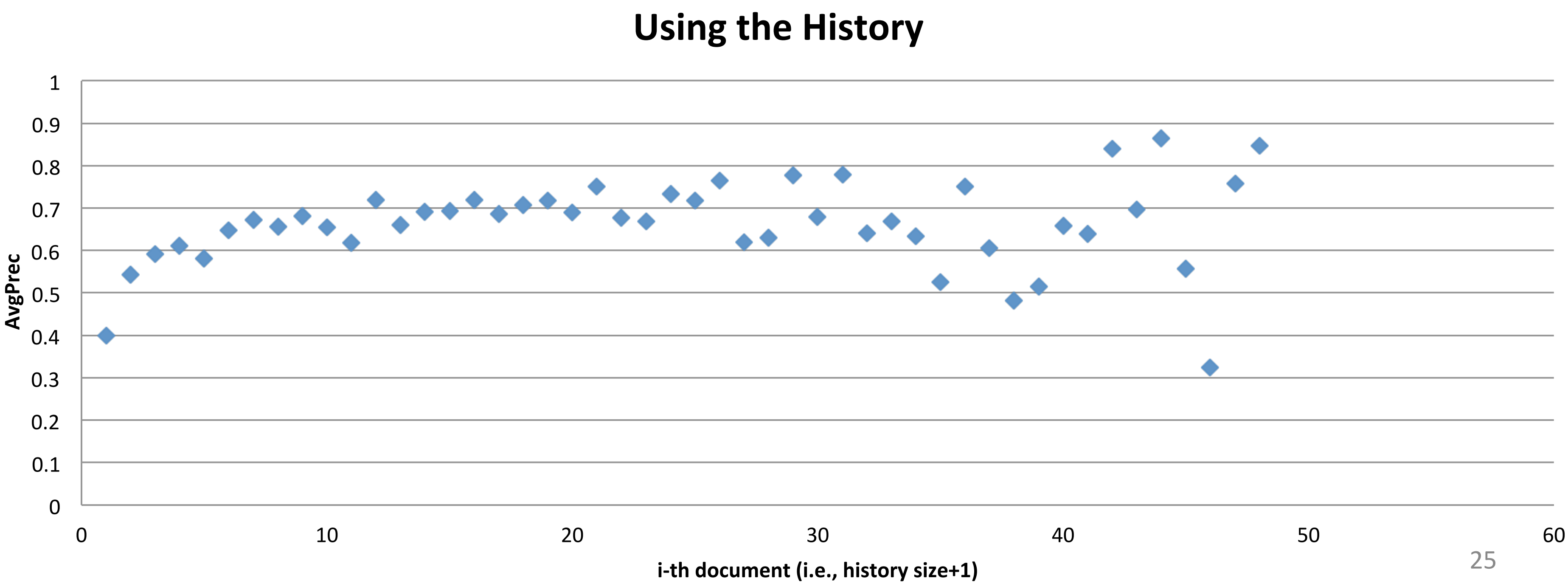
WASHINGTON (AP) -- ``Peanuts'' creator Charles Schulz was remembered today as a genius who touched the lives of millions of Americans as the House adopted a resolution to award him a Congressional Gold Medal.

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 would retire after being diagnosed with colon cancer.

``On Saturday night, millions of Americans lost their security blanket,'' said Rep. Lynn Woolsey, D-Calif. ``Life won't be the same without Charles …

Using the History

- Conclusion
  - Evidence from past documents is very important
  - Effectiveness should improve over time



Ad-hoc Object Retrieval

- Given a KG
- We want to rank them as answer to a query
- (Entity linking over search queries)
- AOR
  - Given the description of an entity
  - give me back its identifier
  - Input: query *q*, data graph *G*
  - Output: ranked list of URIs from *G*

Ad-hoc Object Retrieval

- Supporting end-users
  – Users who can not express their need in SPARQL
- Dealing with large-scale data
  – Giving up query expressivity for scale
- Dealing with heterogeneity
  – Users who are unaware of the schema of the data
  – No single schema to the data
    - Example: 2.6m classes and 33k properties in Billion Triples 2009

Indexing

- Search requires matching and ranking
  - Matching selects a subset of the elements to be scored
- The goal of indexing is to speed up matching
  - Retrieval needs to be performed in milliseconds
  - Without an index, retrieval would require scanning through the collection
- The type of index depends on the **types of data and queries** to be supported
  - DB-style indexing
  - IR-style indexing

DB-style indexing

- B-trees, etc.
- Requires a structured query:
  - SQL
  - SPARQL
  - …

IR-style indexing

- Index data as text
  - Create virtual documents from data
  - One virtual document per subgraph, resource or triple
    - typically: resource
- Key differences to Text Retrieval
  - RDF data is structured
  - Minimally, queries on property values are required

Horizontal index structure

- Two fields (indices): one for terms, one for properties
- For each term, store the property on the same position in the property index
  - Positions are required even without phrase queries
- Query engine needs to support the alignment operator
- Dictionary is number of unique terms + number of properties

<uri1> <foaf:name> “peter mika”
<uri1> <foaf:age> “32”
<uri1> <vcard:location> “barcelona”

Field	p1	p2	p3	p4
token	peter	mika	32	barcelona
property	foaf:name	foaf:name	foaf:age	vcard:location

Vertical index structure

- One field (index) per property
- Positions are not required
  - But useful for phrase queries
- Query engine needs to support fields
- Dictionary is number of unique terms
- Number of fields could be a problem for merging, query performance

Field	p1	p2	p3	p4
foaf:name	peter	mika		
foaf:age	32			
vcard:location	barcelona			

BM25F Ranking

BM25(F) uses a term-frequency (tf) that accounts for the decreasing marginal contribution of terms

$$\tilde{tf_i} = \sum_{s=1}^{S} v_s \frac{tf_{si}}{B_s}$$

where

v_s is the weight of the field

tf_{si} is the frequency of term *i* in field *s*

B_s is the document length normalization factor:

$$B_s = \left((1 - b_s) + b_s \cdot \frac{l_s}{avl_s}) \right)$$

l_s is the length of field *s*
avl_s is the average length of *s*
b_s is a tunable parameter

Roi Blanco, Peter Mika, Sebastiano Vigna: Effective and Efficient Entity Search in RDF Data. International Semantic Web Conference 2011:83-97

BM25F ranking cont.

- Final term score is a combination of tf and idf

$$w_i^{BM25F} = \frac{tf}{k_1 + \tilde{tf_i}} \cdot w_i^{IDF}$$

where

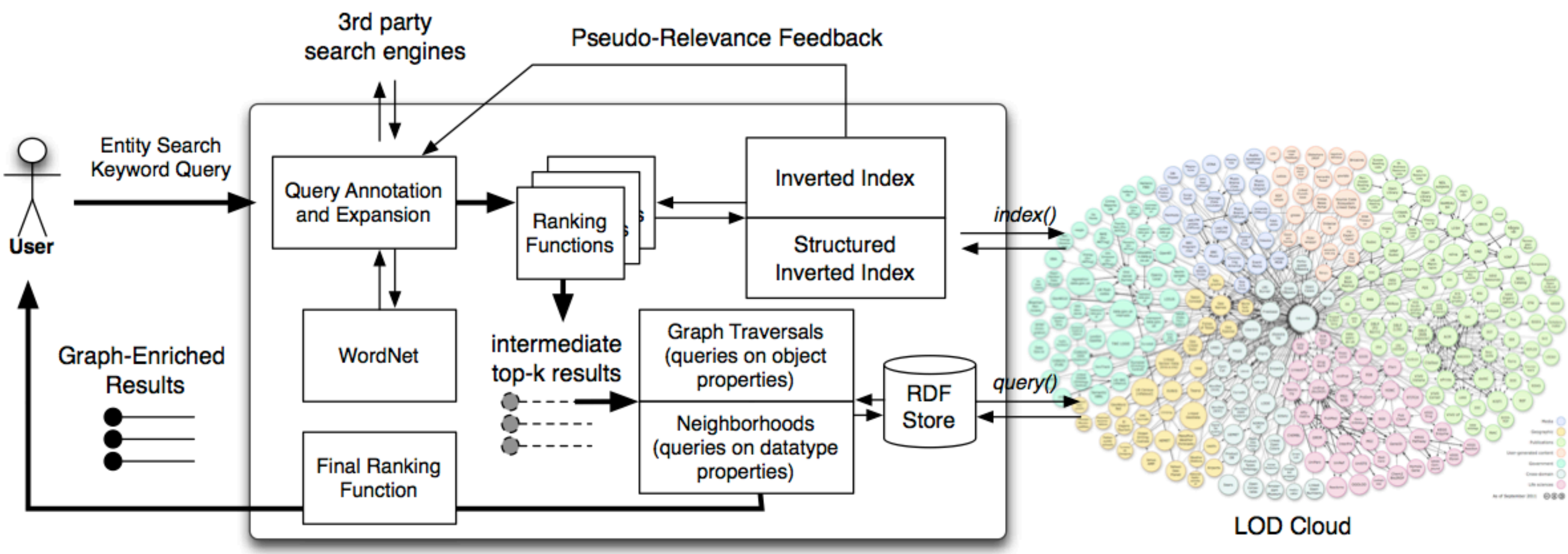
k_1 is a tunable parameter

w^{IDF} is the inverse-document frequency: $\log\left(\frac{D - n_i + 0.5}{n_i + 0.5}\right)$

- Finally, the score of a document D is the sum of the scores of query terms q

$$score^{BM25F}(Q, D) = \sum_{q \in Q} w_i^{BM25F}$$

Combining IR and DB indices



Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. Combining Inverted Indices and Structured Search for Ad-hoc Object Retrieval. In: 35th Annual ACM SIGIR Conference (SIGIR 2012), Portland, Oregon, USA, August 2012.

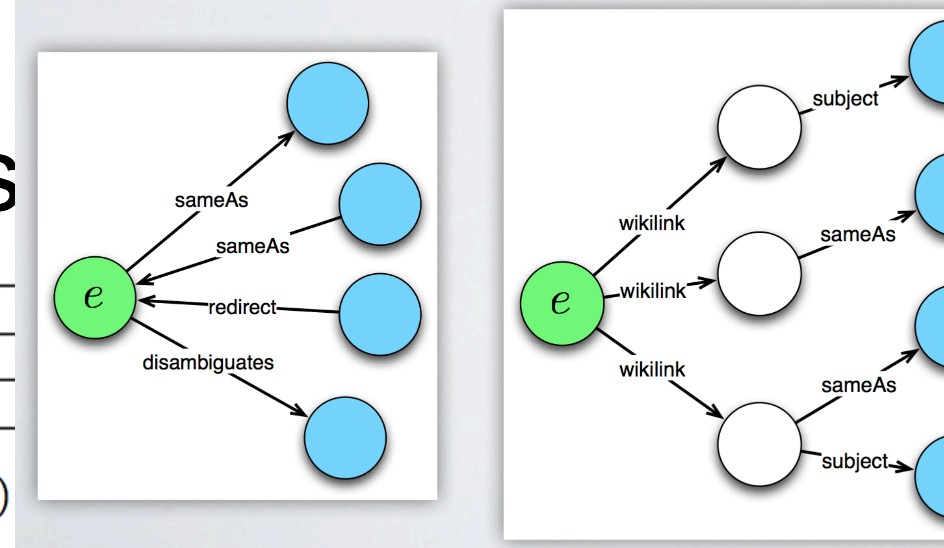
AOR Evaluation

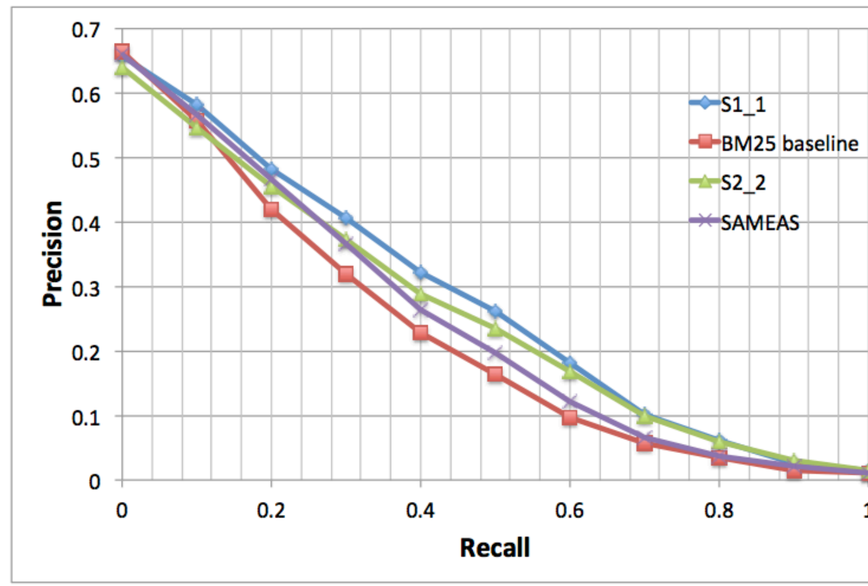
- 1.3 billions RDF triples from LOD cloud
- Crowdsourced relevance judgments
- 92 and 50 queries

- http://km.aifb.kit.edu/ws/semsearch10/
- http://km.aifb.kit.edu/ws/semsearch11/

Evaluation Results

	2010 Collection	
	MAP	P10
BM25	0.2070	0.3348
SAMEAS	0.2293* (+11%)	0.363* (+8%)
S1_1	**0.2586*** (+25%)	**0.3848*** (+15%)
S1_2	0.2305* (+11%)	0.3217 (-4%)
S1_3	0.2306* (+11%)	0.3217 (-4%)
S2_1	0.2118 (+2%)	0.3370 (+1%)
S2_2	0.2118 (+2%)	0.3370 (+1%)
S2_3	0.2113 (+2%)	0.3402 (+2%)





Approach	IR time	RDF time	Total time
BM25 Baseline	285	-	285
Extension	580	-	580 (+104%)
Query Autoc.	1447	-	1447 (+408%)
PRF3	2670	-	2670 (+837%)
SAMEAS	285	30	315 (+11%)
S1_1	285	48	333 (+17%)
S1_2	285	84	369 (+29%)
S1_3	285	86	371 (+30%)
S2_1	285	1746	2031 (+613%)
S2_2	285	2192	2477 (+769%)
S2_3	285	105	390 (+37%)

Summary

- AOR = *"Given the description of an entity, give me back its identifier"*
- combining classic IR techniques + structured database storing graph data
- significantly better results (up to +25% MAP over BM25 baseline).
- overhead caused from the graph traversal part is limited

Latest AOR method

- "Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data", SIGIR 2015.
  - account for term dependencies in multi-field entity descriptions

Entity Search
Evaluation Initiatives

INEX Entity Ranking

- Topical query Q
- Entity (result) type T_X
- A list of entity instances Xs

Q

Xs

T_X

Topic 60

Title

olympic classes dinghy sailing

Entities

470 (dinghy) (#816578)

49er (dinghy) (#1006535)

Europe (dinghy) (#855087)

Categories

dinghies (#30308)

Description

The user wants the dinghy classes that are or have been olympic classes, such as Europe and 470.

Narrative

The expected answers are the olympic dinghy classes, both historic and current. Examples include Europe and 470.

INEX-XER

- INEX XML Entity Ranking Track
- Assumptions:
  - Entities (Xs) are represented as Wikipedia pages
  - Binary relevance

Examples of Wikipedia *Entities* (T_X)

- Art museums and galleries
- Countries
- Famous people
- Monarchs of the British Isles
- Artists
- Magicians

Tasks

- Entity Ranking (ER)
  - Given Q and T_X, provide Xs
- List Completion (LC)
  - Given Q and Xs[1..m]
  - Return Xs[m+1..N]

TREC (Web) Entity (Search)

- Related Entity Finding (REF)
- Topics:
  - Input Entity:
    Name + Homepage
  - Target Type:
    Person | Organisation | Product | Location
  - Narrative:
    Description of the relation in free text

Lessons Learned

- Not *that* many entities in ClueWeb B
  - Makes it difficult to define good topics, especially product topics
- Wikipedia/DBPedia dominate approaches and results

Entity Recognition and Disambiguation Challenge (at SIGIR 2014)

- Sample of Freebase KG
- Short text: web search queries from past TREC competitions
  - Winning approach: extract entities from search results for the query
- Long text: ClueWeb pages
  - Winning approach: supervised machine learning, training on Wikipedia

TREC Knowledge Base Acceleration

- Given
  - Incoming text stream (news and social media content)
    - First month w/ human-generated labels as training data
  - A target entity from a knowledge base (e.g.,: people, specified by their Freebase and Wikipedia entries)
- Score each item ("document") based on how "pertinent" it is to the target KB node

TAC Knowledge Base Population

- **T**asks related to extracting information about entities with reference to an external knowledge source (Wikipedia infoboxes)
- KBP 2011 had three tasks:
  - *entity-linking*: given an entity name (person, organization, or geopolitical entity) and a document containing that name, determine the KB node for that entity or add a new node for the entity if it is not already in the KB
  - *slot-filling*: given a named entity and a pre-defined set of attributes ("slots") for the entity type, augment a KB node for that entity by extracting all new learnable slot values from a large corpus of documents
  - *temporal slot-filling*: similar to the regular slot-filling task, but also requests time intervals to be specified for each extracted slot value.

Entity Search - Conclusions

- Historically:
- Expert Finding came first
- Generalized to Entity Search
  - First on Wikipedia (easier)
  - Then on the Web (harder)
- Over structured data
  - AOR
  - Relational Entity Search (e.g., airlines that use the Airbus A380)

Next Steps for Entity Search

- Improve effectiveness for existing tasks
  - Errors propagate along the pipeline
  - Improve individual components (extraction, linking, de-duplication, etc.)
  - Improve ranking models by considering additional evidence (as done for expert finding)
- Work on top of Entity Search
  - New User experiences (based on entities)
  - Exploratory Search

Next Steps for Entity Search

- Novel entity-oriented search tasks
  - Entity summaries: Select attributes
  - Entity Attribute Search ("At which age Nobel prize winners in physics died?") → Crowdsourcing for query understanding!

    Gianluca Demartini, Beth Trushkowsky, Tim Kraska, and Michael Franklin. CrowdQ: Crowdsourced Query Understanding. In: 6th Biennial Conference on Innovative Data Systems Research (CIDR 2013).

    - Slow Search, CACM Aug 2014.
  - Entity Popularity (rank Nobel laureates by popularity)
  - Tail Entities

References

- Gianluca Demartini, Peter Mika, Thanh Tran, Arjen P. de Vries. From Expert Finding to Entity Search on the Web - Tutorial at ECIR 2012