



YAHOO!

Understanding Queries through Entities

Peter Mika, Sr. Research Scientist, Yahoo Labs | November 7, 2014

Why Semantic Search?

- Improvements in IR are harder and harder to come by
 - › Basic relevance models are well established
 - › Machine learning using hundreds of features
 - › Heavy investment in computational power, e.g. real-time indexing and instant search
- Remaining challenges are not computational, but in modeling user cognition
 - › Need a deeper understanding of the query, the content and the relationship of the two

Real problem

Home Mail News Sports Finance



YAHOO! roi blanco

Web
Images
Video
News
Local
Answers
Shopping
More

Anytime
Past day
Past week
Past month

Rio Blanco County, CO
www.co.rio-blanco.co.us
Official website of Rio Blanco County, CO. Home of officials and offices, social media, and more.
[Employee Login](#)
[Assessor](#)
[View Our Webcam](#)

Roi Blanco | Yahoo! Labs
labs.yahoo.com/author/roi-blanco/
Roi is a senior researcher in the area of language and search. He received his Ph.D. degree from the University of California, Berkeley.

Roi Blanco - Image Results


[More Roi Blanco images](#)

Roi Blanco - HomePage
www.dc.fi.udc.es/~roi/
Roi Blanco, B. Barla Camblor, in Web Search ISWC 2013;

Intelius
Live in the know.™

Like 21k +1 14k Help | Sign In
★ Bookmark this Site

People Search Background Check Criminal Records Reverse Lookup Intelius Premier Identity Protection Employee Screening

People Search | Email Lookup | Social Network Search | Property Records | 24-Hour People Search Pass

Search results for Roi Blanco in the United States

We found 2 people that match **Roi Blanco** in the United States

- Get the report on** Roixy E Blanco , age 39 [Get more details](#)

Has lived in	DOB	Phone	Address	Related to
Hollywood, FL Fort Lauderdale, FL Miami, FL	✓	✓	✓	Jorge Blanco Iris Matta
- Get the report on** Roicio Blanco , age 24 [Get more details](#)

Has lived in	DOB	Phone	Address	Related to
Lake Havasu City, AZ	✓		✓	Jose Blanco Estela Blanco Annette Blanco

Related People Searches

- Roy Blanco
- Royce Blanco
- Royston Blanco
- R Blanco

dblp: Roi Blanco - Uni Trier
www.informatik.uni-trier.de/~a-tree/b/Blanco:Roi.html Cached
Roi Blanco, Christina Lioma: Mixed monolingual homepage finding in 34 languages: the role of language script and search domain. Inf. Retr. 12(3): 324-351 (2009)

Rio Blanco County, Colorado - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rio_Blanco_County,_Colorado Cached

What it's like to be a machine?

[Roi Blanco - HomePage](#)

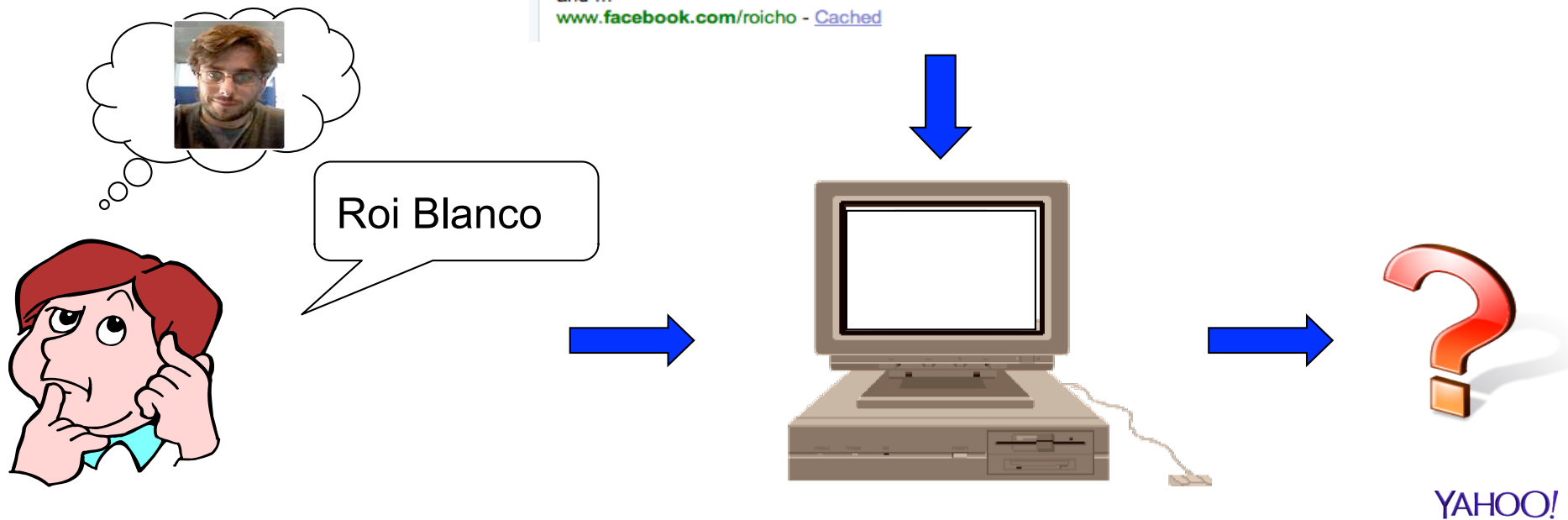
Phone contact: Voice: +34 981 167 000 ext. 1276 Fax: +34 981 167 160
www.dc.fi.udc.es/~roi - [Cached](#)

[Roi Blanco - Spain | LinkedIn](#)

Experience: Researcher, Yahoo! Research; Visitor, University of Glasgow
www.linkedin.com/pub/roi-blanco/8/58/2b0 - [Cached](#)

[Roi Blanco | Facebook](#)

Roi Blanco is on Facebook. Join Facebook to connect with **Roi Blanco** and others you may know. Facebook gives people the power to share and makes the world more open and ...
www.facebook.com/roicho - [Cached](#)



What it's like to be a machine?

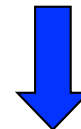
⌂⊕Θ♫♯ğ🍏√∞↑▪®ÇĤĪ⊕★⊙♫□✓✓
ğ🍏★⊙⊕→↑↑↑↑↑↑

◆⊕⊕⊕ΔTOÛŸİĈÊωυτρSM≠1⁄8⊠⊙Γ
≠=5%🍏©§★✓♫BΓE↑↑⊙SM

🍏⊙×Γ♫3⁄4±①↶↷⚡⊕□ğğğğμλκσςτ■
■■■◆↶↑↶°¶§🍏ΥΦΦΦΧΧ□⊕◆◆◆◆◆◆



↶⚡⊕□ğ



YAHOO!

Semantic Search

- **Def.** Semantic Search is any retrieval method where

- › User intent and resources are represented in a *semantic model*
 - A set of concepts or topics that generalize over tokens/phrases
 - Additional structure such as a hierarchy among concepts, relationships among concepts etc.
- › Semantic representations of the query and the user intent are exploited in some part of the retrieval process

- As a research field

- › Workshops
 - ESAIR (2008-2014) at CIKM, Semantic Search (SemSearch) workshop series (2008-2011) at ESWC/WWW, EOS workshop (2010-2011) at SIGIR, JIWES workshop (2012) at SIGIR, Semantic Search Workshop (2011-2014) at VLDB
- › Special Issues of journals
- › Surveys
 - Christos L. Koumenides, Nigel R. Shadbolt: Ranking methods for entity-oriented semantic web search. JASIST 65(6): 1091-1106 (2014)

Semantic search: implicit vs. explicit semantics

- Implicit/internal semantics

- › Models of text extracted from a corpus of queries, documents or interaction logs
 - Query reformulation, term dependency models, translation models, topic models, latent space models, learning to match (PLS), word embeddings
- › See
 - Hang Li and Jun Xu: Semantic Matching in Search. Foundations and Trends in Information Retrieval Vol 7 Issue 5, 2013, pp 343-469

- Explicit/external semantics

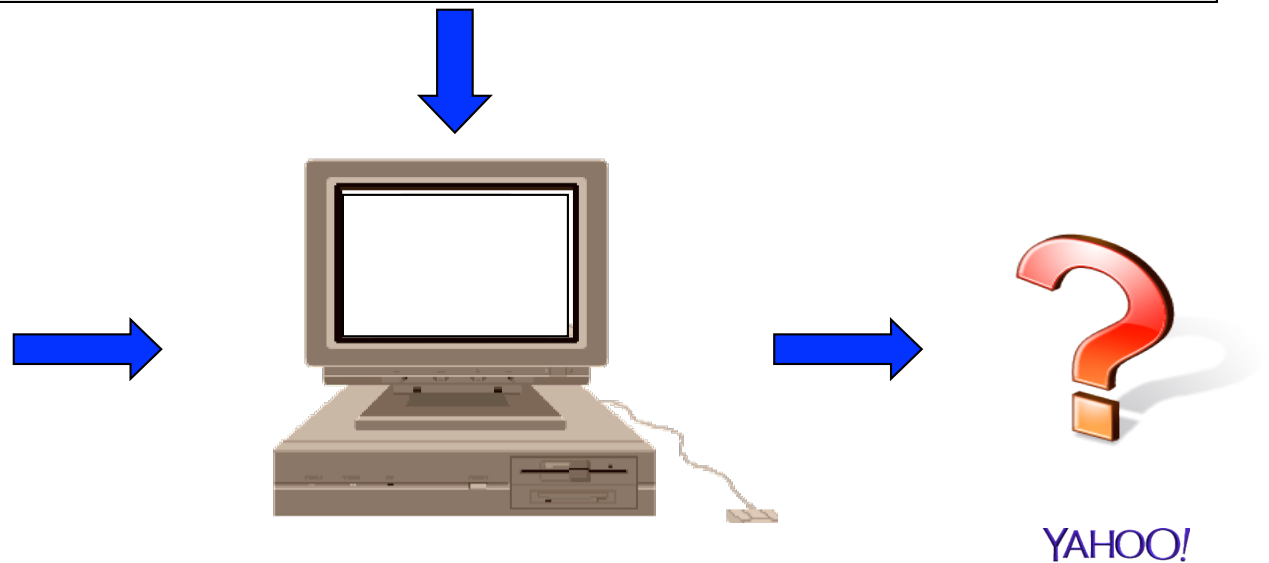
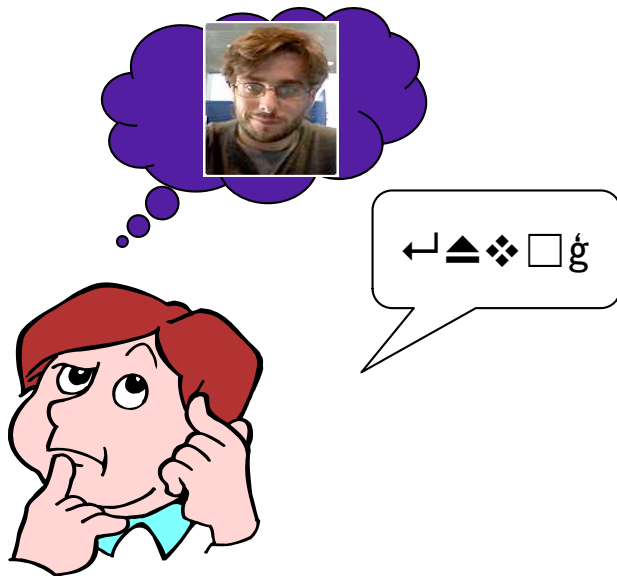
- › Explicit linguistic or ontological structures extracted from text and linked to external knowledge
- › Obtained using **IE techniques** or acquired from **Semantic Web markup**

What it's like to be a machine?

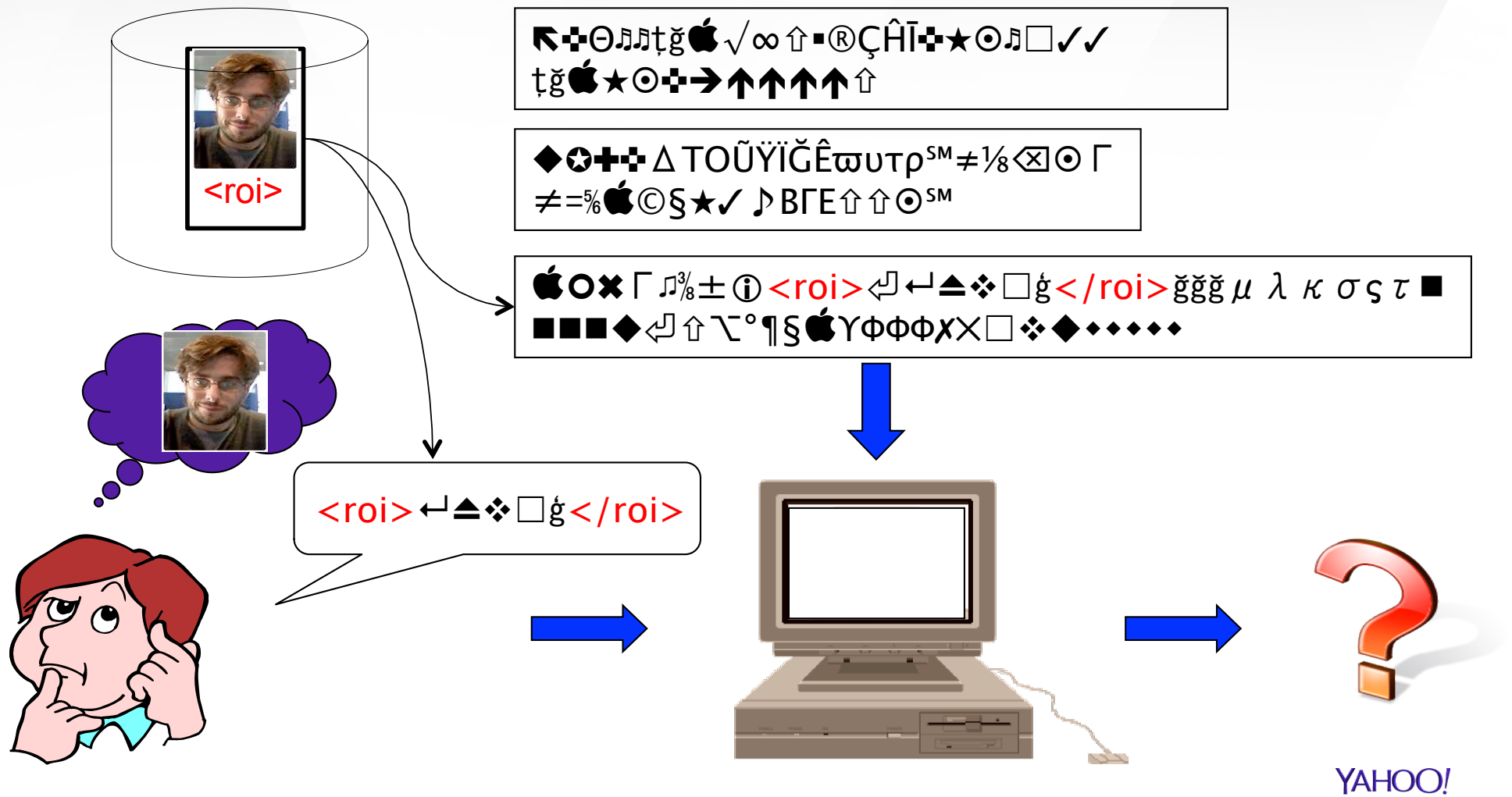
⌂+Θ♫♫ğ🍏√∞↑▪®ÇĤĪ÷★⊙♫□✓✓
ğ🍏★⊙÷→↑↑↑↑↑↑

◆☆+÷ΔTOÛŸİĈÊωυτρSM≠1/8⊠⊙Γ
≠=5%🍏©§★✓♫BΓE↑↑⊙SM

🍏○×Γ♫3/4±①↶↷↶↷◊□ğğğğμλκσςτ■
■■■◆↶↑↶°¶§🍏ΥΦΦΦΧΧ□◊◆◆◆◆◆



What it's like to be a machine?



Semantic understanding

- Documents

- › Text in general
 - Exploiting natural language structure and semantic coherence
- › Specific to the Web
 - Exploiting structure of web pages, e.g. annotation of web tables

- Queries

- › Short text and no structure... nothing to do?

Semantic understanding of queries

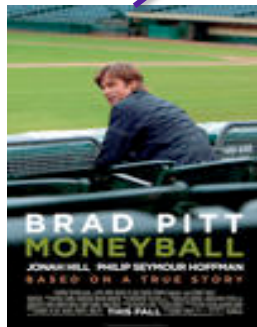
- Entities play an important role
 - › [Pound et al, WWW 2010], [Lin et al WWW 2012]
 - › ~70% of queries contain a named entity (*entity mention queries*)
 - brad pitt height
 - › ~50% of queries have an entity focus (*entity seeking queries*)
 - brad pitt attacked by fans
 - › ~10% of queries are looking for a class of entities
 - brad pitt movies
- Entity mention query = <entity> {+ <intent>}
 - › Intent is typically an additional word or phrase to
 - Disambiguate, most often by type e.g. *brad pitt actor*
 - Specify action or aspect e.g. *brad pitt net worth, toy story trailer*

Entities and Intents

Object of the query (entity)

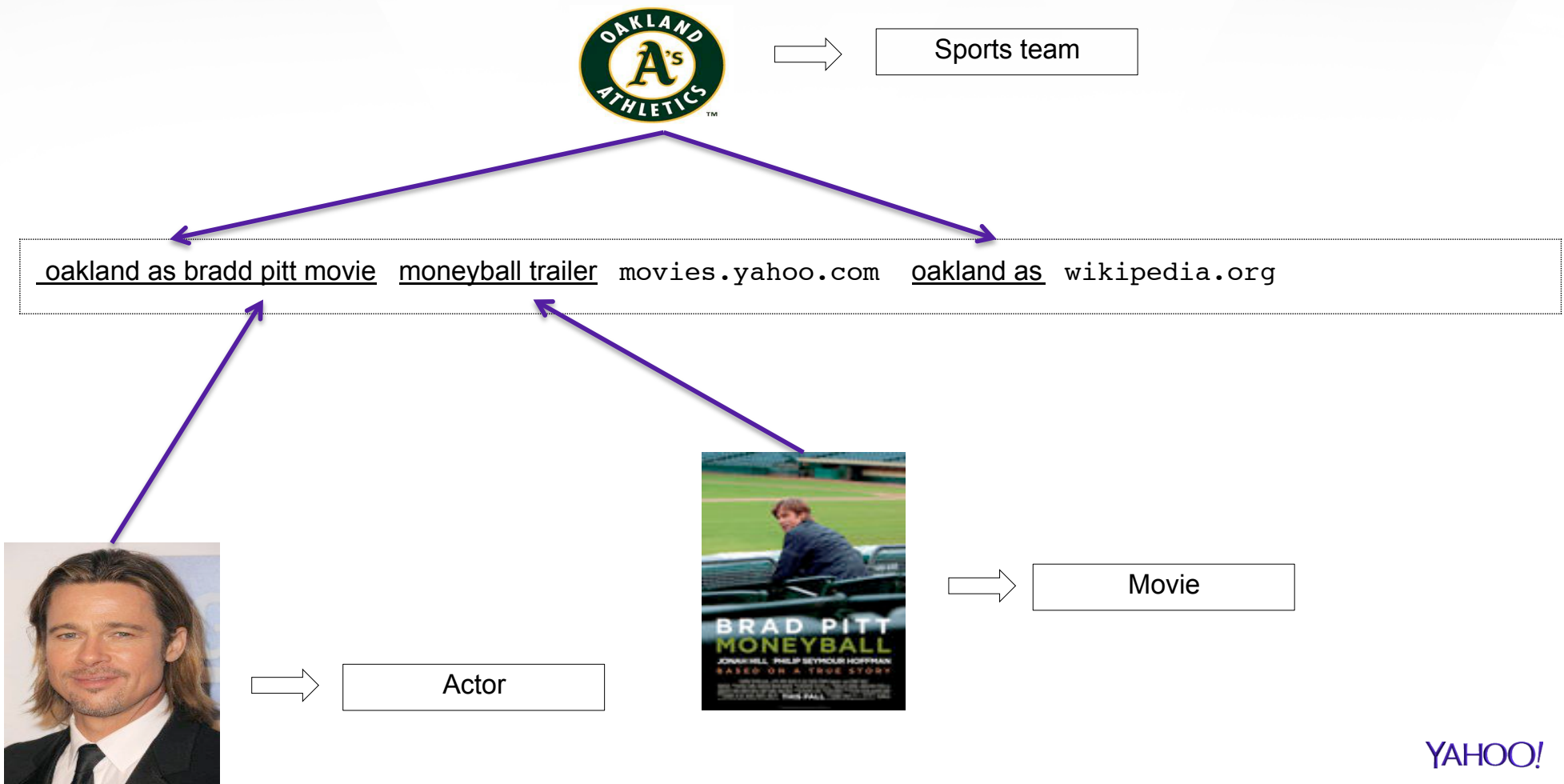
what the user wants to do with it (intent)

moneyball trailer

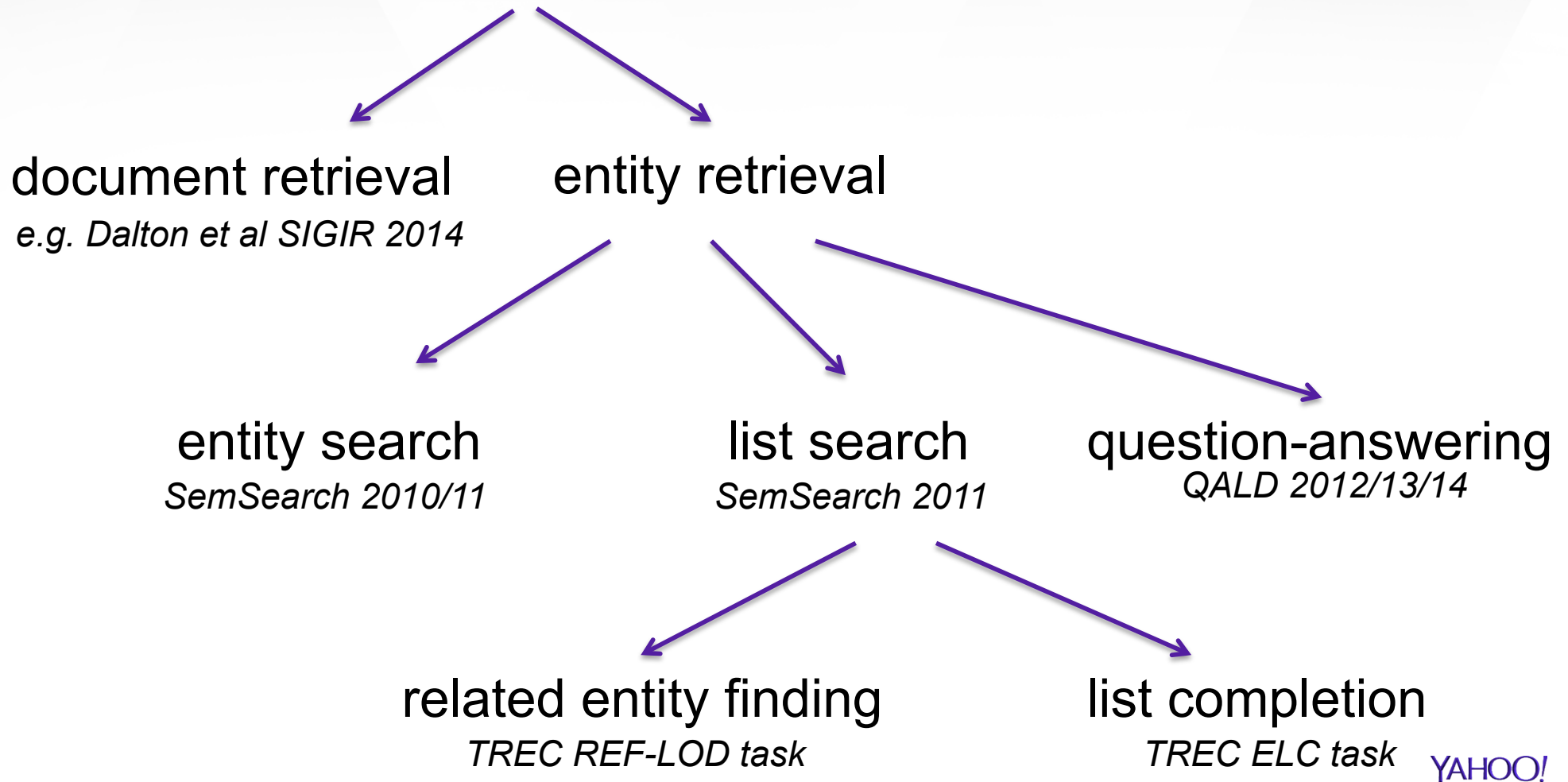


Movie

Annotation over sessions

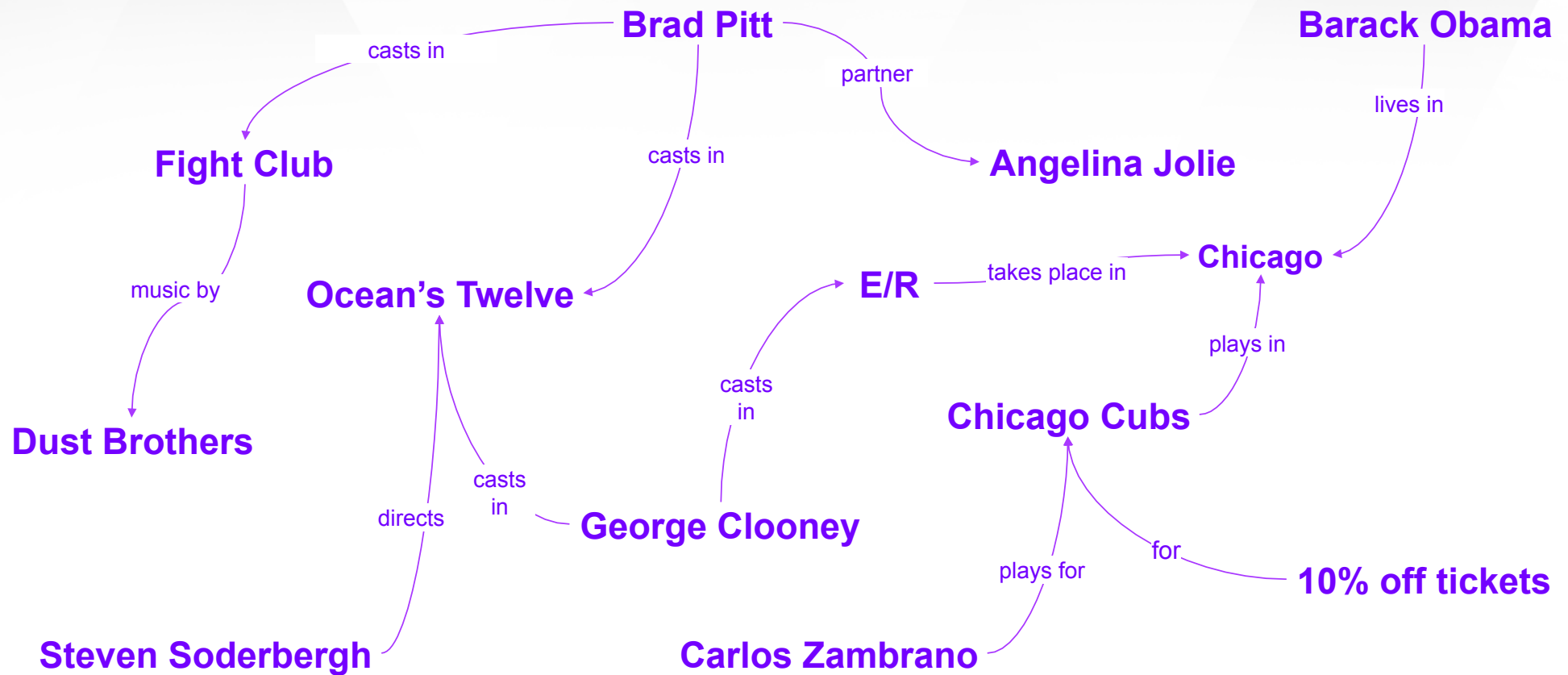


Common tasks in Semantic Search



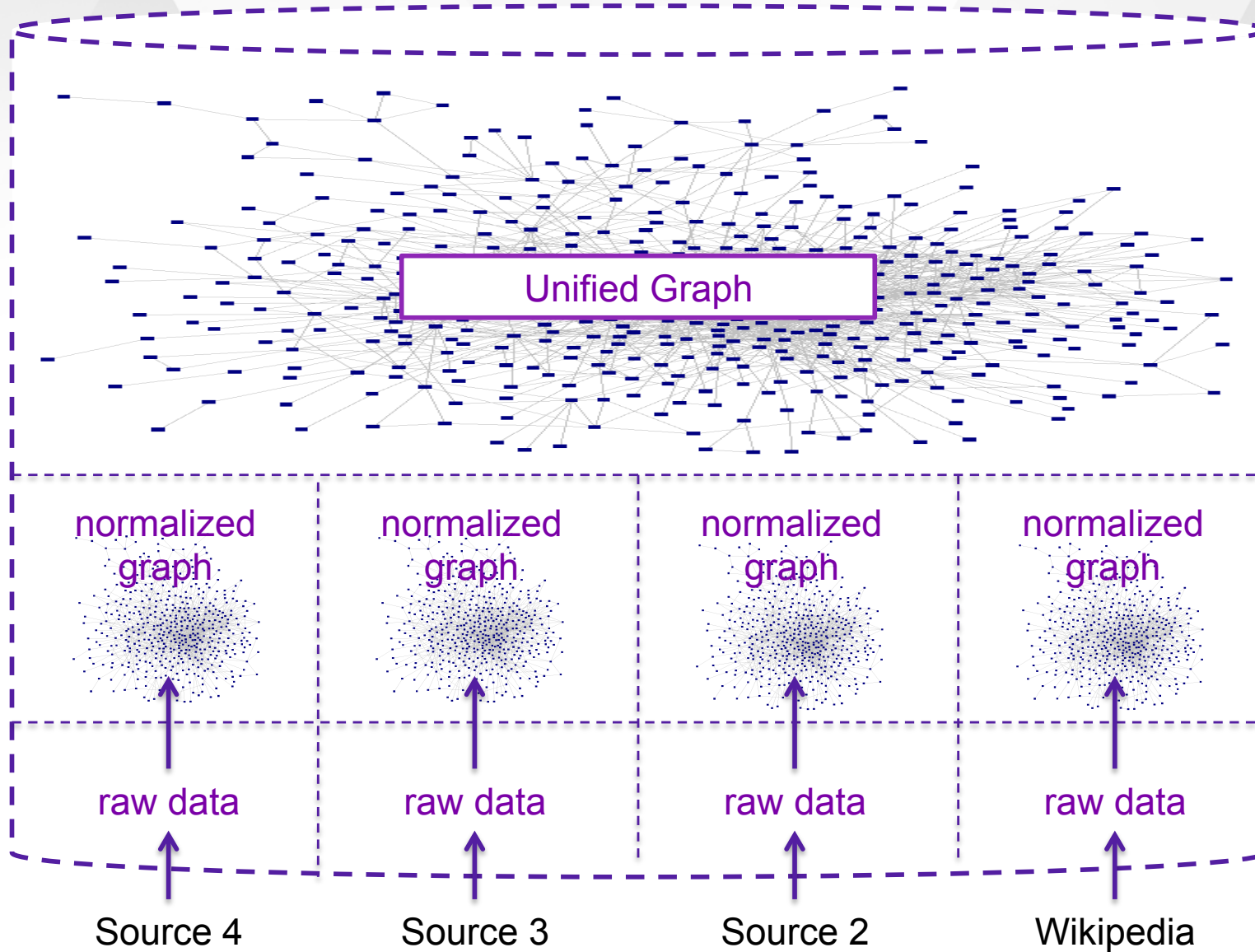
Query understanding

Yahoo's Knowledge Graph



Nicolas Torzec: [Yahoo's Knowledge Graph](#). SemTech 2014.

YAHOO!




Brad Pitt
according to
Yahoo

Brad Pitt
according to
Wikipedia

YAHOO!

Knowledge graphs...

- ... are not perfect
- Or: the importance of human editors



Ice cube
Product category

© Getty Images

An **ice cube** is a small, roughly **cube**-shaped piece of **ice** (frozen water), conventionally used to cool beverages. **Ice** cubes are sometimes preferred over crushed **ice** because they melt more slowly; they are standard in... [wikipedia.org](#)

Born: June 15, 1969 (age 44), [Los Angeles, California, USA](#)

Height: 5' 7" (1.73m)

Spouse: [Kimberly Woodruff](#) (m. 1992-present)

Parents: [Doris Benjamin](#), [Hosea Jackson](#)

Children: [Darrel Jackson](#), [O'Shea Jackson Jr.](#), [Shareef Jackson](#), [Karima Jackson](#), [Deja Jackson](#)

Feedback

Knowledge graphs...

- ... are not perfect
- Or: the importance of human editors



Vin Diesel
Actor

Vin Diesel , born as Mark Vincent, is an American actor, producer, director, and screenwriter. He came to prominence in the late 1990s, and first became known for appearing in Steven Spielbergs Saving Private Ryan... [wikipedia.org](#)

Born: July 18, 1967, [New York City, New York, USA](#)

Died: January 30, 2014, [TBA](#)

Height: 5' 11" (1.82m)

Partner: [Paloma Jiménez \(2008-2014\)](#)

Parents: [Irving Vincent](#), [Delora Vincent](#)

YAHOO!

Knowledge graphs...

- ... are not perfect
- Or: the importance of human editors

Michelangelo

Artist



Michelangelo di Lodovico Buonarroti Simoni , commonly known as Michelangelo, was an Italian sculptor, painter, architect, poet, and engineer of the High Renaissance who exerted an unparalleled influence on the... [wikipedia.org](https://en.wikipedia.org/wiki/Michelangelo)

Born: March 6, 1475, [Caprese Michelangelo](#)

Died: February 18, 1564, [Rome](#)

Parents: [Ludovico di Leonardo di Buonarotto Simoni](#), [Francesca di Neri del Miniato di Siena](#)

Feedback

Entity displays in Web search

- Entity retrieval
 - › Which entity does a keyword query refer to, if any?
 - › This talk
- Related entity recommendation
 - › Which entity would the user visit next?
 - › [Blanco et al. ISWC 2013]



Matthew Paige "Matt" Damon is an American actor, voice actor, screenwriter, producer, and philanthropist whose career was launched following the success of the drama film *Good Will Hunting* (1997) from a screenplay... [wikipedia.org](#)

Born: October 8, 1970 (age 43), [Cambridge, Massachusetts, USA](#)

Height: 5' 10" (1.78m)

Spouse: [Luciana Barroso \(m. 2005-present\)](#)

Partner: [Winona Ryder \(1998-2000\)](#)

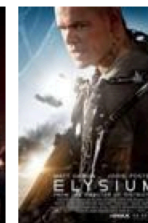
Parents: [Kent Damon](#), [Nancy Carlsson-Paige](#)

Children: [Isabella Damon](#), [Alexia Barroso](#), [Gia Zavala Damon](#), [Stella Damon](#)

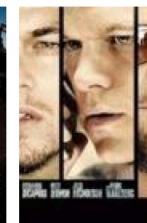
Movies & TV Shows



[The Zero Theorem](#)



[Elysium](#)



[The Departed](#)



[We Bought a Zoo](#)



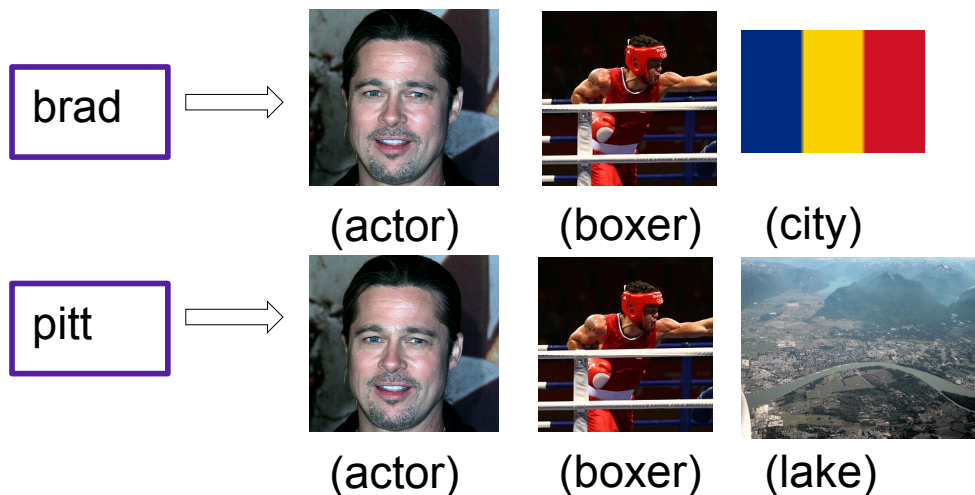
[Good Will Hunting](#)

[Feedback](#)

Two matching approaches

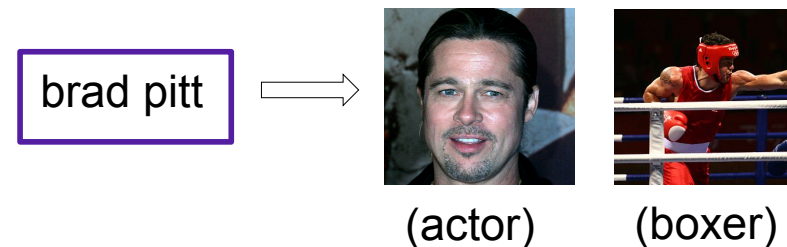
■ Match by keywords

- › Closer to text retrieval
 - Match individual keywords
 - Score and aggregate
- › e.g. [Blanco et al. ISWC 2013]
 - <https://github.com/yahoo/Glimmer/>



■ Match by aliases

- › Closer to entity linking
 - Find potential mentions of entities (spots) in query
 - Score candidates for each spot
- › This talk



YAHOO!

Entity linking

- Large-scale entity-alias mining
 - › From usage data, Wikipedia etc.
- Novel method for linking (FEL)
 - › Significantly improved relevance
 - › Completely unsupervised
 - › Efficient implementation
- See next... more details in:
 - › Roi Blanco, Giuseppe Ottaviano and Edgar Meij.
Fast and space-efficient entity linking in queries. WSDM 2015 (to appear)

Problem definition

- Given

- › Query q consisting of an ordered list of tokens t_i
- › Segment s from a segmentation \mathbf{s} from all possible segmentations S_q
- › Entity e from a set of candidate entities \mathbf{e} from the complete set E

- Find

- › For all possible segmentations and candidate entities
- › Select best entity for segment independently of other segments

$$\begin{aligned} & \underset{\mathbf{e} \in E, \mathbf{s} \in S_q}{\operatorname{argmax}} \max_{e \in \mathbf{e}, s \in \mathbf{s}} P(e|s) \\ & \text{s.t. } s \in \mathbf{s}, \bigcup_s \subseteq \mathbf{s}, \bigcap_s = \emptyset \end{aligned}$$

Intuitions

Assume: also given annotated collections c_i with segments of text linked to entities from E .

1. Keyphraseness

- › How likely is a segment to be an entity mention?

$$P(a_s = 1|c, s) = \frac{\sum_{s:a_s=1} n(s, c)}{n(s, c)}$$

- › e.g. how common is “in”(unlinked) vs. “in” (linked) in the text

2. Commonness

- › How likely that a linked segment refers to a particular entity?

$$P(e|a_s = 1, c, s) = \frac{\sum_{s:a_{s,e}=1} n(s, c)}{\sum_{s:a_s=1} n(s, c)}$$

- › e.g. how often does “brad pitt” refers to Brad Pitt (actor) vs. Brad Pitt (boxer)

See also [Entity Linking and Retrieval](#) (tutorial) by Meij et al.

Ranking function

Probability of the segment generated
by a given collection

$$P(e|s) = \sum_{c \in \{c_q, c_w\}} P(c|s) P(e|c, s) \quad (3)$$

$$= \sum_{c \in \{c_q, c_w\}} P(c|s) \sum_{a_s = \{0, 1\}} P(a_s|c, s) P(e|a_s, c, s)$$

$$= \sum_{c \in \{c_q, c_w\}} P(c|s) \left[P(a_s = 0|c, s) P(e|a_s = 0, c, s) \right. \\ \left. + P(a_s = 1|c, s) P(e|a_s = 1, c, s) \right] \quad (4)$$

Commonness

Keyphraseness

Context-aware extension

Probability of segment and query are independent of each other

Estimated by word2vec representation

Probability of segment and query are independent of each other given the entity

$$P(e|s, q) = \frac{P(e)P(s, q|e)}{P(s, q)}$$

$$= \left(\frac{P(e)}{P(q)P(s)} \right) \cdot P(q|e)P(s|e) = P(e|s) \frac{P(q|e)}{P(q)} = P(e|s) \prod_i \frac{P(t_i|e)}{P(q)}$$

Query	FEL answer	FEL+Context
install roof insulation	Insolation	Building_insulation
inventor of gunpowder	Gunpowder	History_of_gunpowder
us political map	Map	Red_states_and_blue_states
dj jobs	Jobs_(film)	Disc_jockey
buy used car parts online	Automobile	Used_car
what is the longest running tv show	Television	The_Simpsons

Results: effectiveness

- Significant improvement over external baselines and internal system
 - › Measured on public [Webscope dataset Yahoo Search Query Log to Entities](#)

		P@1	MRR	MAP	R-Prec
A trivial search engine over Wikipedia	LM	0.0394	0.1386	0.1053	0.0365
Search over Bing, top Wikipedia result	LM-Click	0.4882	0.5799	0.4264	0.3835
	Bing	0.6349	0.7018	0.5388	0.5223
State-of-the-art in literature	Wikifier	0.2983	0.3201	0.2030	0.2086
	Wikipedia-miner	0.6450	0.7126	0.6110	0.5892
Our method: Fast Entity Linker (FEL)	Commonness	0.7336	0.7798	0.6418	0.6464
	FEL	0.7669	0.8092	0.6528	0.6575
FEL + context	FEL+Centroid	0.8035	0.8366	0.6728	0.6765
	FEL+LR	0.8352	0.8684	0.6912	0.6883

Results: efficiency

- Two orders of magnitude faster than state-of-the-art
 - › Simplifying assumptions at scoring time
 - › Adding context independently
 - › Dynamic pruning
- Small memory footprint
 - › Compression techniques, e.g. 10x reduction in word2vec storage

System	Average time	Data size
Wikifier	44.32 ms	6.8G
Retrieval (And)	13.63 ms	2.5G
Retrieval (Or)	210.13 ms	2.5G
FEL	0.14 ms	1.8G
FEL+Centroid	0.27 ms	2.3G
FEL+LR	0.40 ms	2.3G

Related and future work

- Session-level analysis
 - › Online
 - Entity retrieval using queries in the same session or entire history
 - › Offline
 - Dealing with sparseness, e.g. for usage mining
 - [Hollink et al. WWW 2013]
- Intents/tasks/actions
 - › Schema.org Actions as potential ontology of intents
- Personalization? Temporal?
 - › How often does the meaning of Brad Pitt change per user?
 - › How often does the meaning of Brad Pitt change?
- Interactive query construction

Q&A

- Many thanks to members of the Semantic Search team at Yahoo Labs Barcelona and to Yahoos around the world
- Contact
 - › pmika@yahoo-inc.com
 - › @pmika
 - › <http://www.slideshare.net/pmika/>



Leftover

Entity Retrieval

- Keyword search over entity graphs
 - › see Pound et al. WWW08 for a definition
 - › No common benchmark until 2010
- SemSearch Challenge 2010/2011
 - 50 entity-mention queries Selected from the Search Query Tiny Sample v1.0 dataset (Yahoo! Webscope)
 - Billion Triples Challenge 2009 data set
 - Evaluation using Mechanical Turk
 - › See report:
 - Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, Thanh Tran: Repeatable and reliable semantic search evaluation. J. Web Sem. 21: 14-29 (2013)

Glimmer: open-source entity retrieval engine from Yahoo

- Extension of [MG4J](#) from University of Milano
- Indexing of RDF data
 - › MapReduce-based
 - › Horizontal indexing (subject/predicate/object fields)
 - › Vertical indexing (one field per predicate)
- Retrieval
 - › BM25F with machine-learned weights for properties and domains
 - › 52% improvement over the best system in SemSearch 2010
- See
 - › Roi Blanco, Peter Mika, Sebastiano Vigna: Effective and Efficient Entity Search in RDF Data. International Semantic Web Conference (1) 2011: 83-97
 - › <https://github.com/yahoo/Glimmer/>

Other evaluations in Entity Retrieval

▪ TREC Entity Track

- › 2009-2011
- › Data
 - ClueWeb 09 collection
- › Queries
 - Related Entity Finding
 - Entities related to a given entity through a particular relationship
 - *(Homepages of) airlines that fly Boeing 747*
 - Entity List Completion
 - Given some elements of a list of entities, complete the list
 - *Professional sports teams in Philadelphia such as the Philadelphia Wings, ...*
- › Relevance assessments provided by TREC assessors

▪ Question Answering over Linked Data

- › 2011-2014
- › Data
 - Dbpedia and MusicBrainz in RDF
- › Queries
 - Full natural language questions of different forms, written by the organizers
 - Multi-lingual
 - *Give me all actors starring in Batman Begins*
- › Results are defined by an equivalent SPARQL query
 - Systems are free to return list of results or a SPARQL query