Entity Linking and Retrieval

Edgar Meij – @edgarmeij
Yahoo! Labs

(Krisztian Balog, Daan Odijk)

Minnesota Children's Museum

Deloitte.

GUIDANT

NORTHWEST AIRLINES

ECOLAB

Valspar

Donaldson
Filtration Solutions

BEST BUY

Carlson Companies

AMS
Solutions for Life

WELLS FARGO

Medtronic
Alleviating Pain · Restoring Health · Extending Life

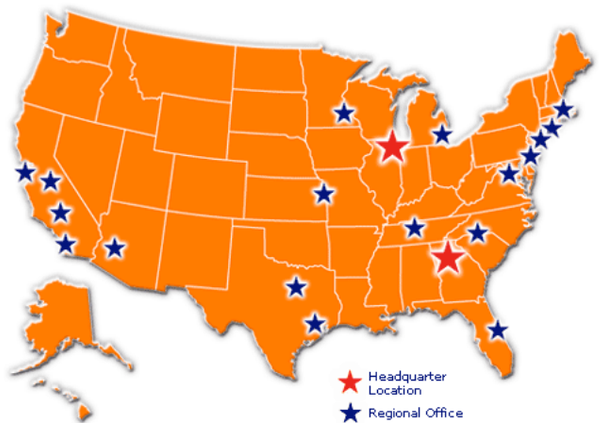
Thrivent Financial for Lutherans

MARVIN
Windows and Doors

3M Worldwide

POLARIS
The Way Out.

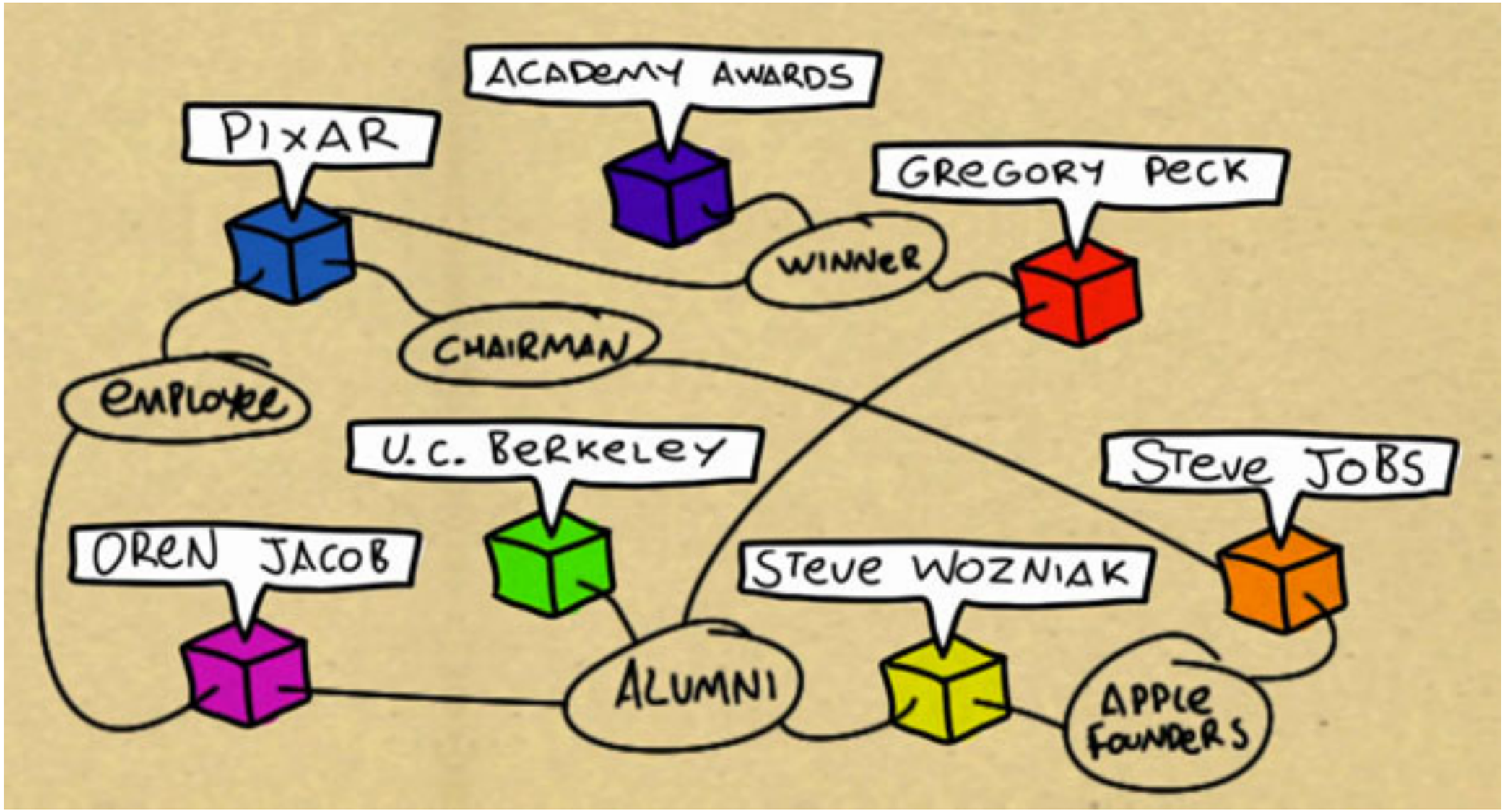
GENERAL MILLS



Entity?

- Uniquely identifiable “thing” or “object”
  - “A thing with a distinct and independent existence”
- Properties:
  - ID
  - Name(s)
  - Type(s)
  - Attributes (/Descriptions)
  - Relationships to other entities

Entity?


ACADEMY AWARDS
PIXAR
GREGORY PECK
WINNER
CHAIRMAN
EMPLOYEE
U.C. BERKELEY
STEVE JOBS
OREN JACOB
STEVE WOZNIAK
ALUMNI
APPLE FOUNDERS

Entity Linking?

Iranian POW negotiator holds talks with Iraqi ministers

The head of **Iran**'s **prisoner of war** commission met with two **Iraqi** Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister **Mohammed Saeed al-Sahhaf** told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file,'' INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the **1980**-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The **Associated Press** that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 fied as civil law detainees in the largest exchange since **1990**.

More than 1 million people w

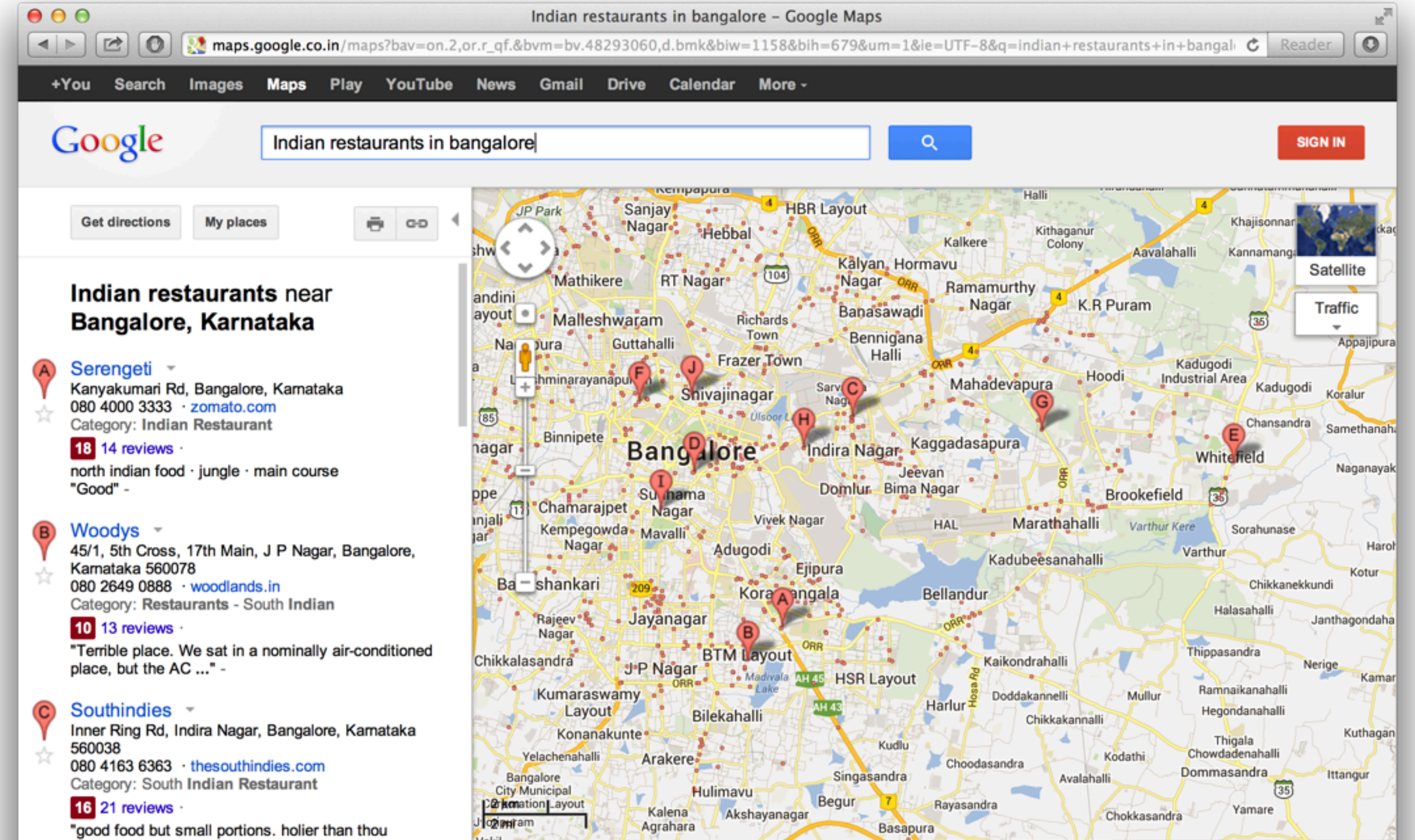
Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

open in wikipedia

Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

Entity Retrieval?



Menu

- Introduction
- Part 1 – Entity Linking
- Part 2 – Entity Retrieval

See http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/ or http://bit.ly/yahoosummerschool for the slides.

Menu

- 13:45 - 14:45
  - Part 1 – Entity Linking
    - Introduction and methods
- 14:45 - 15:00: Break
- 15:00 - 15:30
  - Part 1 – Entity Linking
    - Evaluation, hands-on, and open challenges
- 15:30 - 16:00: Coffee break
- 16:00 - 17:00
  - Part 2 – Entity Retrieval

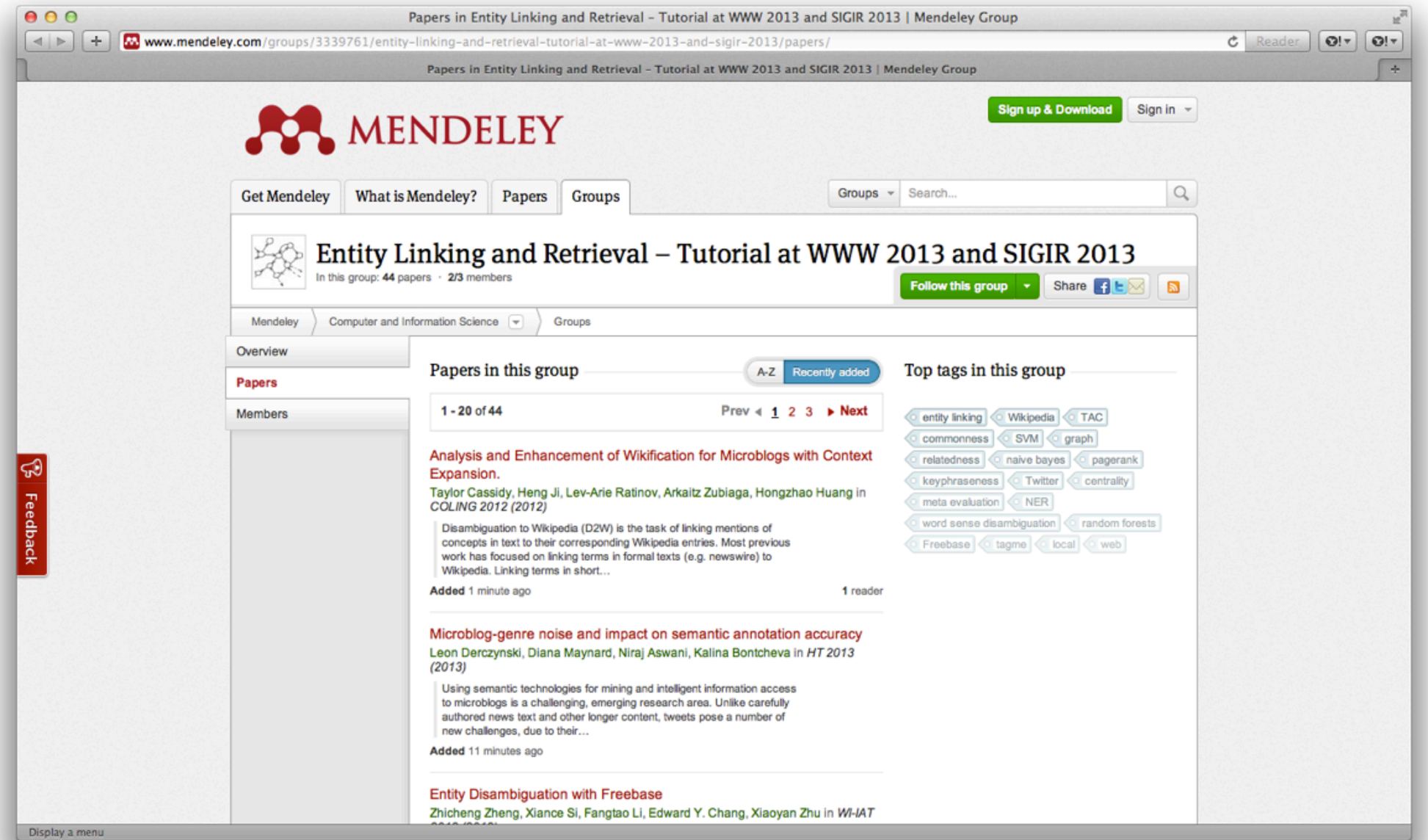
See http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/ or http://bit.ly/yahoosummerschool for the slides.

References

---

http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/

References



http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/

Outline

- Introduction
- Part 1 – Entity Linking
- Part 2 – Entity Retrieval

Part I

Entity Linking

Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - hands-on
  - open challenges

Introduction

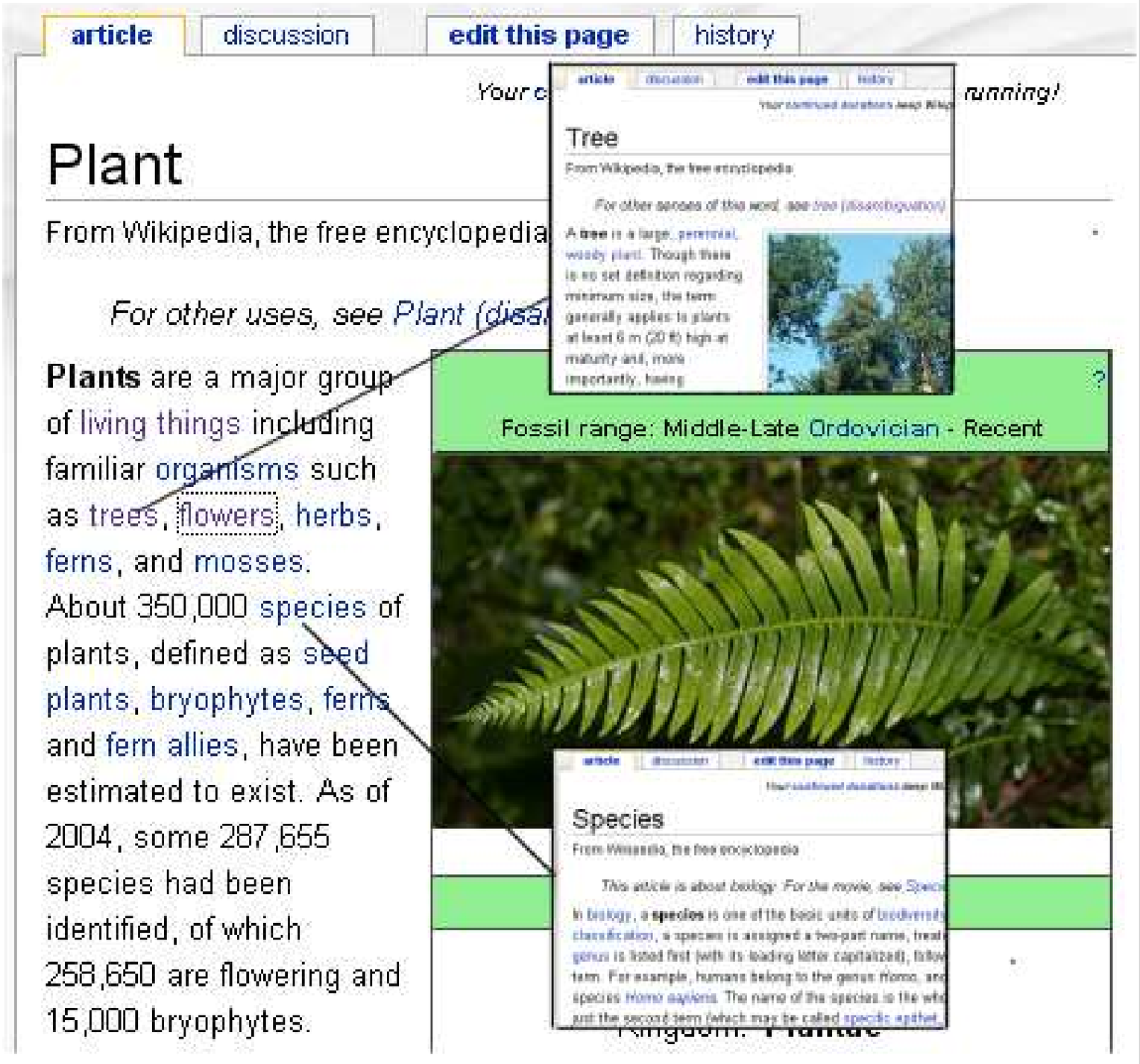


Image taken from Mihalcea and Csomai (2007). **Wikify!: linking documents to encyclopedic knowledge**. In CIKM '07.

Let’s learn something about Spin-Optical Metamaterial

Article Views
- Abstract
- Full Text
- Full Text (PDF)
- Figures Only
- Supplementary Materials

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints

Related Content
Similar Articles In:
- Science Magazine

Search Google Scholar for:
- Articles by Shitrit, N.

Science 10 May 2013:
Vol. 340 no. 6133 pp. 724-726
DOI: 10.1126/science.1234892

‹ Prev | Table of Contents | Next ›
Read Full Text to Comment (0)

REPORT

Spin-Optical Metamaterial Route to Spin-Controlled Photonics

Nir Shitrit, Igor Yulevich, Elhanan Maguid, Dror Ozeri, Dekel Veksler, Vladimir Kleiner, Erez Hasman*

± Author Affiliations

*Corresponding author. E-mail: mehasman@technion.ac.il

ABSTRACT | EDITOR'S SUMMARY

Spin optics provides a route to control light, whereby the photon helicity (spin angular momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of spin optics and metamaterials offers the dispersion engineering of a structured matter in a polarization helicity-dependent manner. We show that polarization-controlled optical modes of metamaterials arise where the spatial inversion symmetry is violated. The emerged spin-split dispersion of spontaneous emission originates from the spin-orbit interaction of light, generating a selection rule based on symmetry restrictions in a spin-optical metamaterial. The inversion asymmetric metasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled nanophotonic applications based on the design of the metasurface symmetry properties.

Received for publication 7 January 2013.
Accepted for publication 13 March 2013.

REPORT

Spin-Optical Metamaterial Route to Spin-Controlled Photonics

Nir Shitrit, Igor Yulevich, Elhanan Maguid, Dror Ozeri, Dekel Veksler, Vladimir Kleiner, Erez Hasman*

± Author Affiliations

*Corresponding author. E-mail: mehasman@technion.ac.il

ABSTRACT | EDITOR'S SUMMARY

Spin optics provides a route to control light, whereby the photon helicity (spin angular momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of spin optics and metamaterials offers the dispersion engineering of a structured matter in a polarization helicity-dependent manner. We show that polarization-controlled optical modes of metamaterials arise where the spatial inversion symmetry is violated. The emerged spin-split dispersion of spontaneous emission originates from the spin-orbit interaction of light, generating a selection rule based on symmetry restrictions in a spin-optical metamaterial. The inversion asymmetric metasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled nanophotonic applications based on the design of the metasurface symmetry properties.

Received for publication 7 January 2013.

Developed by Paolo Ferragina and Ugo Scaiella at A³ Lab
Dipartimento di Informatica, University of Pisa.

Input Text

Italiano English

momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of spin optics and metamaterials offers the dispersion engineering of a structured matter in a polarization helicity–dependent manner. We show that polarization-controlled optical modes of metamaterials arise where the spatial inversion symmetry is violated. The emerged spin-split dispersion of spontaneous emission originates from the spin-orbit interaction of light, generating a selection rule based on symmetry restrictions in a spin-optical metamaterial. The inversion asymmetric metasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled nanophotonic applications based on the design of the metasurface symmetry properties.

Many links

Few links

Reset

TAGME!

Tagged text Topics

Spin optics provides a route to control light, whereby the photon helicity (spin angular momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of sp… dispersion engineering of a structured matter in a p…r. We show that polarization-controlled optical mode…patial inversion symmetry is violated. The emerged spir…ssion originates from the spin-orbit interaction of …ed on symmetry restrictions in a spin-optical meta…etasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled

Degenerate energy levels

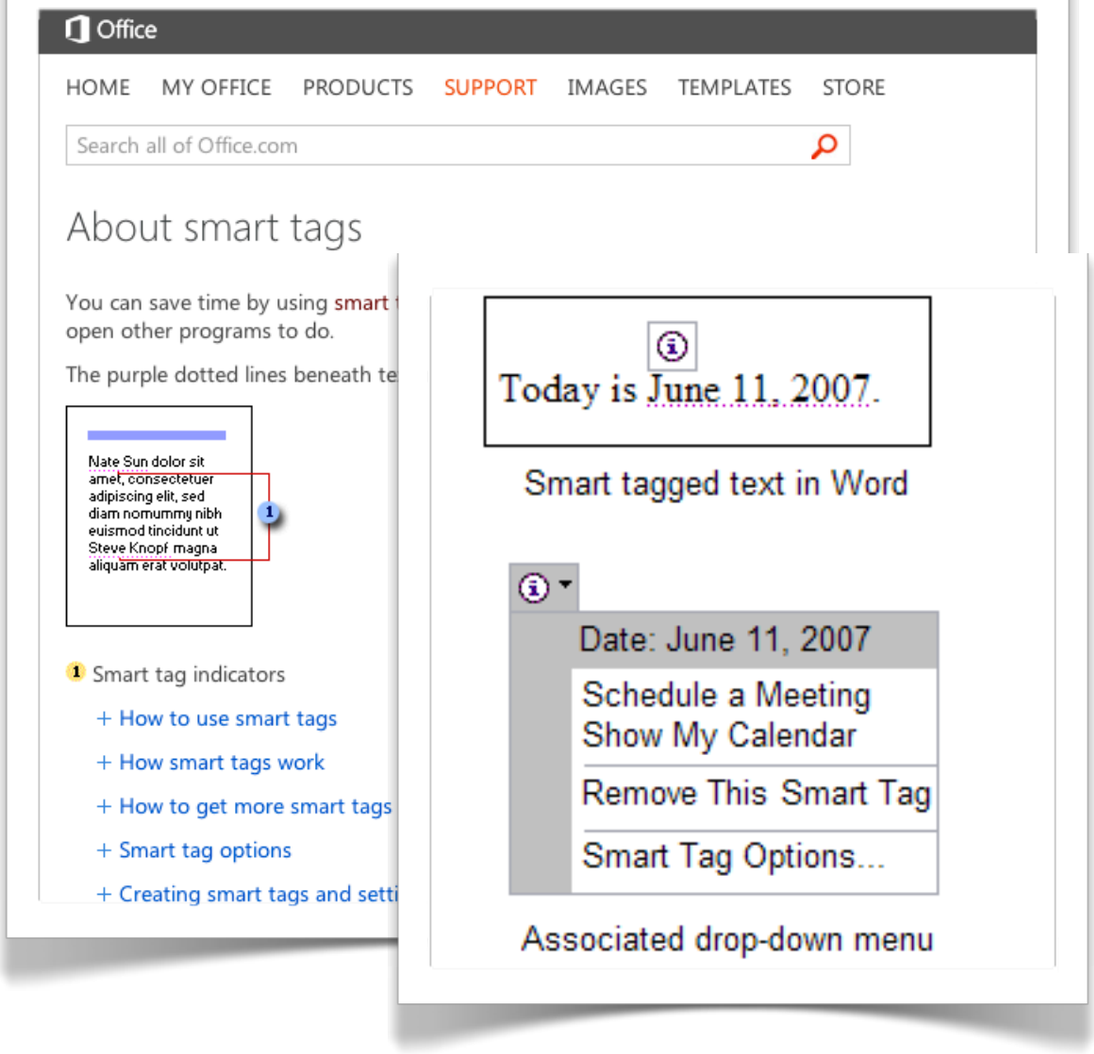
In physics, two or more different quantum states are said to be degenerate if they are all at the same energy level. Statistically this means that they are all equally probable of being filled, and in...

See http://tagme.di.unipi.it

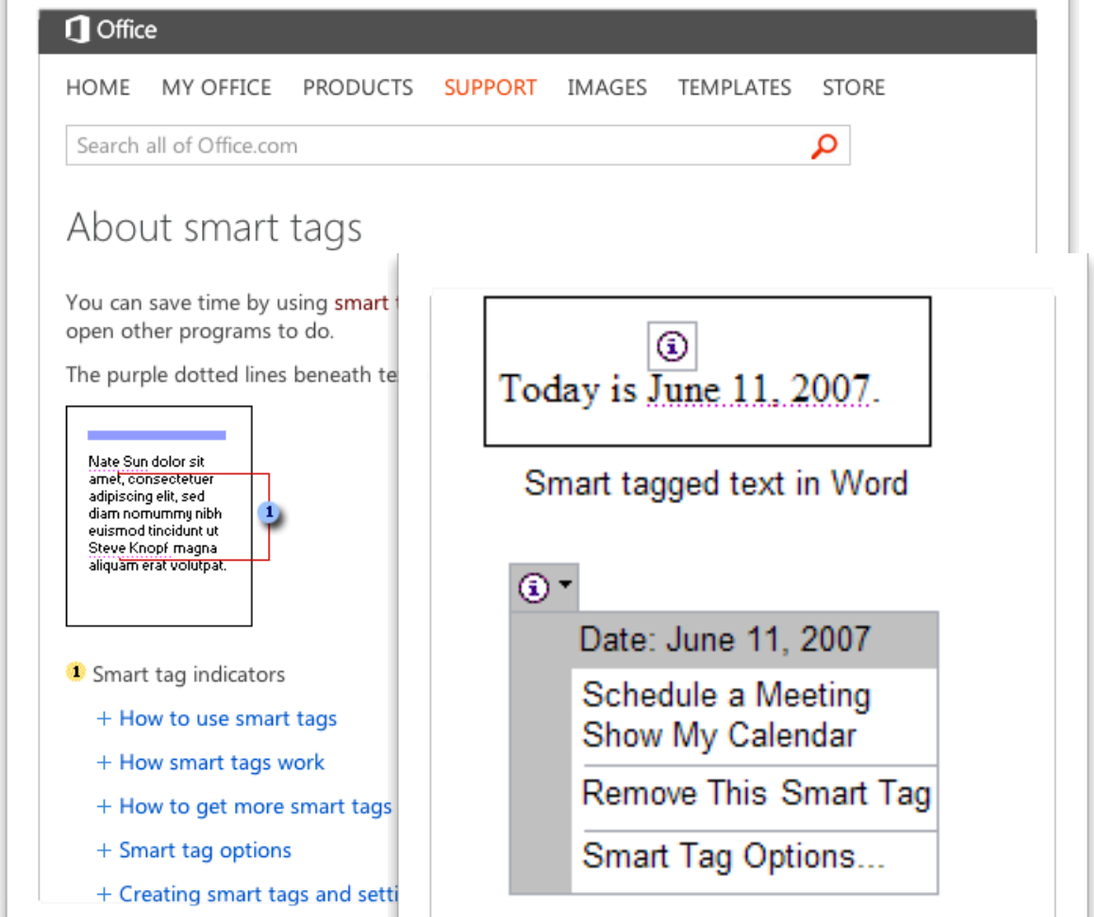
TAG·ME

See http://tagme.di.unipi.it

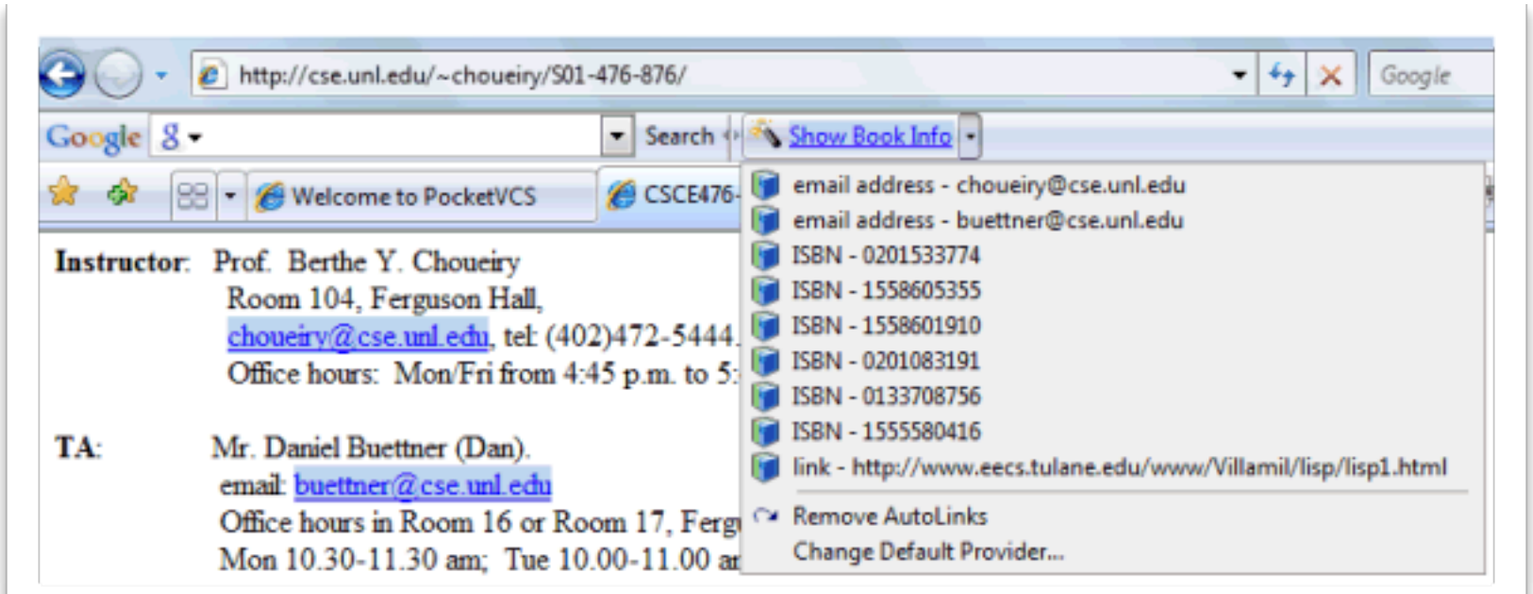
Microsoft Smart Tags



Microsoft Smart Tags



Google toolbar



Treaty of Versailles – Wolfram|Alpha

http://www.wolframalpha.com/input/?i=Treaty+of+Versailles&random=true

Google

WolframAlpha™ computational… knowledge engine

Treaty of Versailles

Examples Random

Assuming "Treaty of Versailles" is a historical event | Use as a word instead

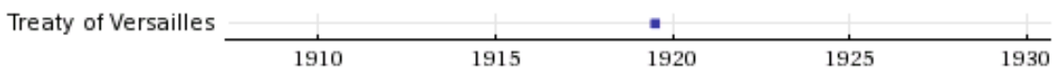
Input interpretation:

Treaty of Versailles

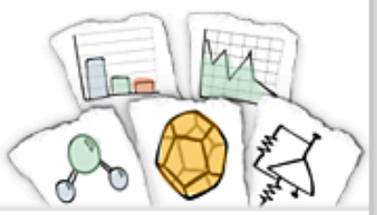
Basic information:

date	28 June 1919
city involved	Versailles, Ile-de-France, France
countries involved	French Third Republic | Italy | Japan | United Kingdom of Great Britain and Ireland | United States | German Empire
people involved	David Lloyd George | Georges Clemenceau | Woodrow Wilson

Timeline: Include today



Computed by Wolfram Mathematica Source information » Download as: PDF | Live Mathematica



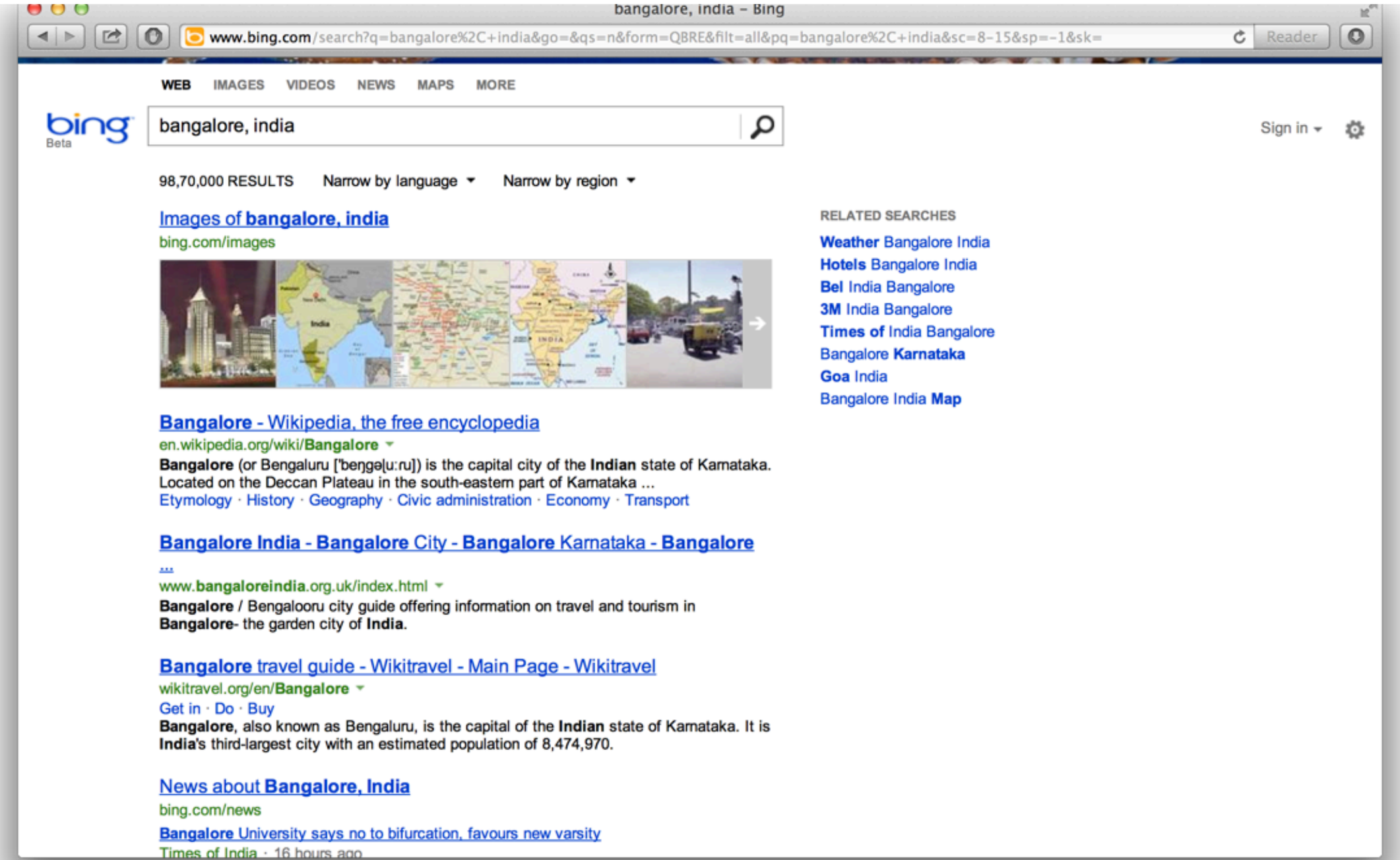
New to Wolfram|Alpha?

TAKE THE TOUR »

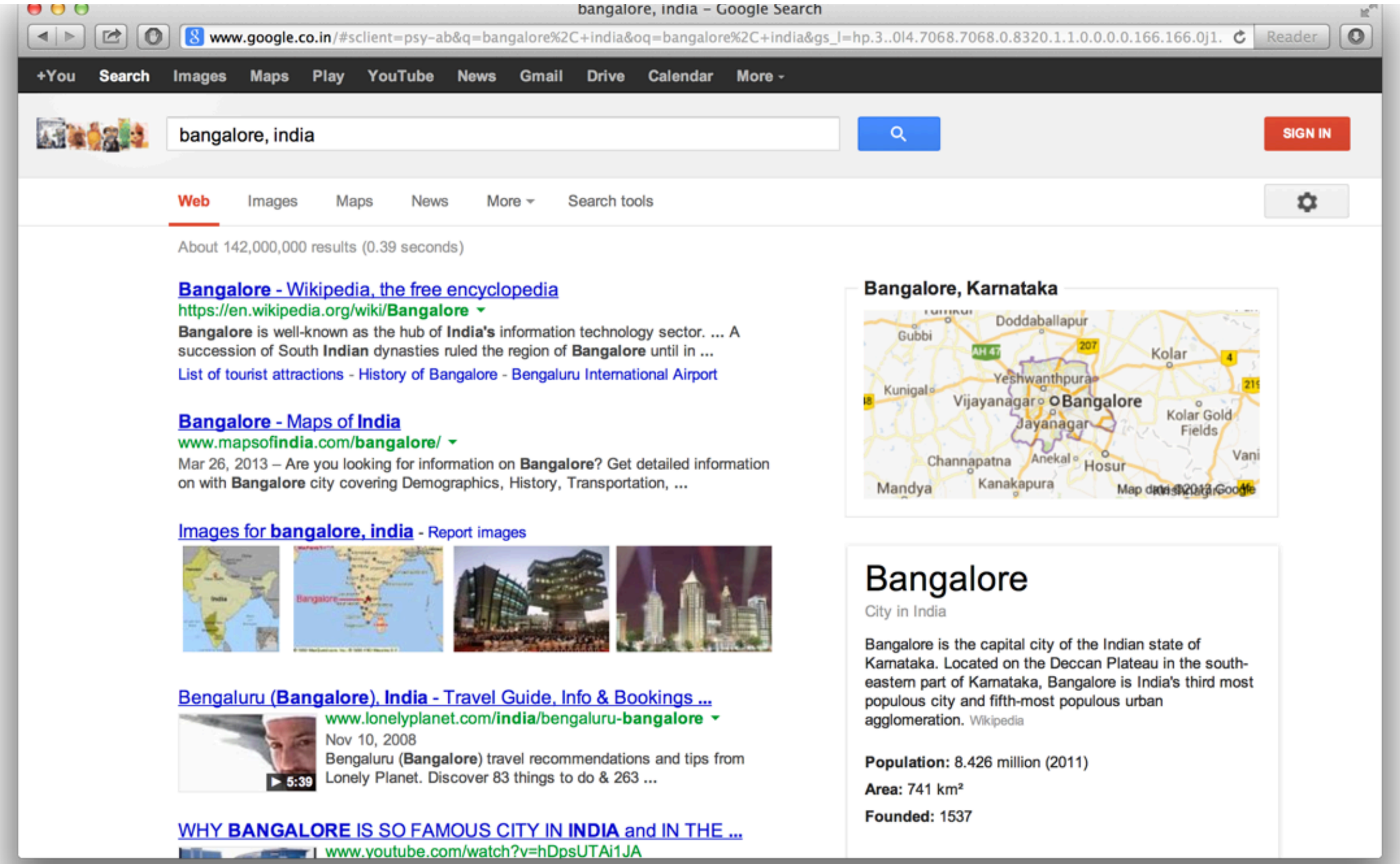
Serving up funky, fresh fun facts on the daily

Follow the fun: @WolframFunFacts

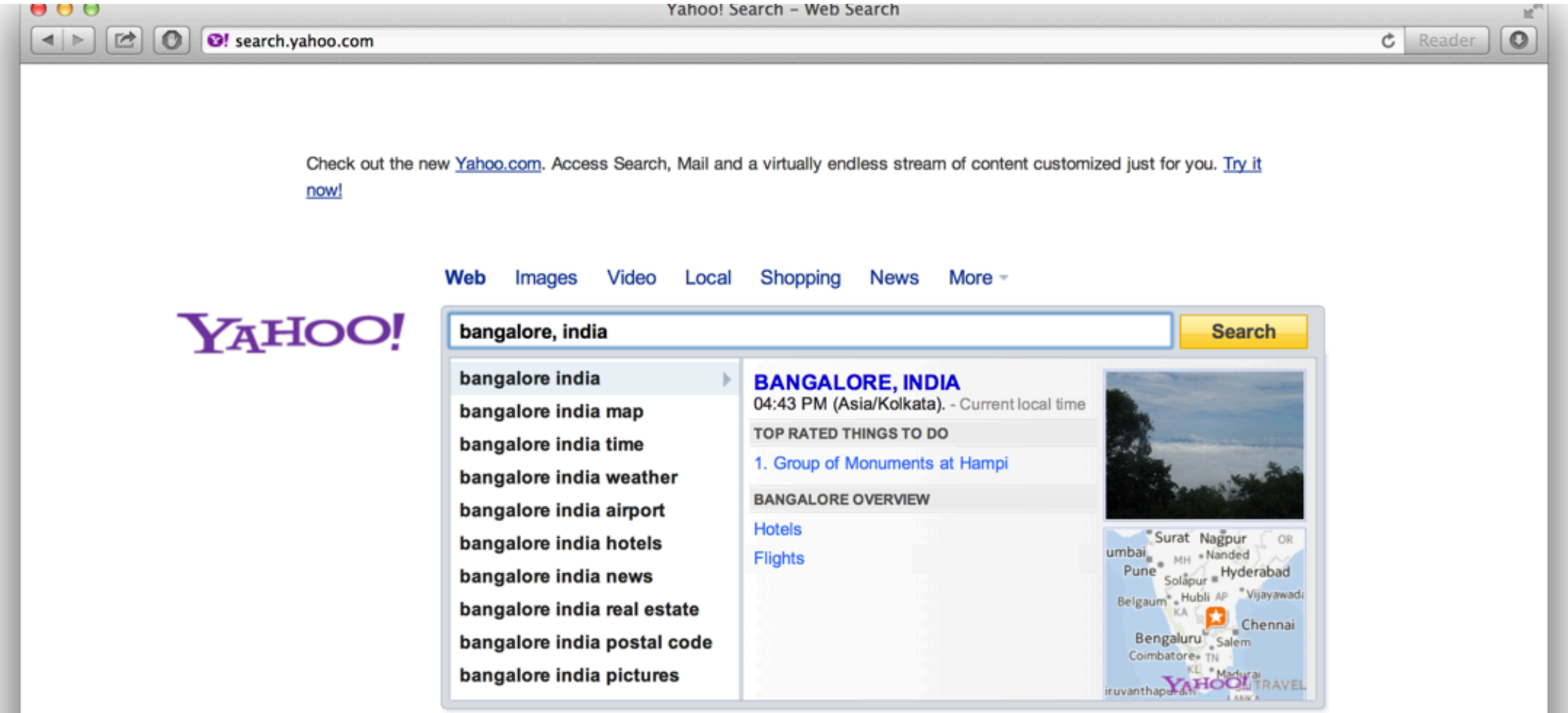
Bing


bangalore, india - Bing
www.bing.com/search?q=bangalore%2C+india&go=&qs=n&form=QBRE&filt=all&pq=bangalore%2C+india&sc=8-15&sp=-1&sk=
Reader
WEB IMAGES VIDEOS NEWS MAPS MORE
bing
Beta
bangalore, india
Sign in
98,70,000 RESULTS
Narrow by language
Narrow by region
Images of bangalore, india
bing.com/images
Bangalore - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bangalore
Bangalore (or Bengaluru ['beŋgəlu:ru]) is the capital city of the Indian state of Karnataka. Located on the Deccan Plateau in the south-eastern part of Karnataka ...
Etymology · History · Geography · Civic administration · Economy · Transport
Bangalore India - Bangalore City - Bangalore Karnataka - Bangalore
...
www.bangaloreindia.org.uk/index.html
Bangalore / Bengalooru city guide offering information on travel and tourism in Bangalore- the garden city of India.
Bangalore travel guide - Wikitravel - Main Page - Wikitravel
wikitravel.org/en/Bangalore
Get in · Do · Buy
Bangalore, also known as Bengaluru, is the capital of the Indian state of Karnataka. It is India's third-largest city with an estimated population of 8,474,970.
News about Bangalore, India
bing.com/news
Bangalore University says no to bifurcation, favours new varsity
Times of India · 16 hours ago
RELATED SEARCHES
Weather Bangalore India
Hotels Bangalore India
Bel India Bangalore
3M India Bangalore
Times of India Bangalore
Bangalore Karnataka
Goa India
Bangalore India Map

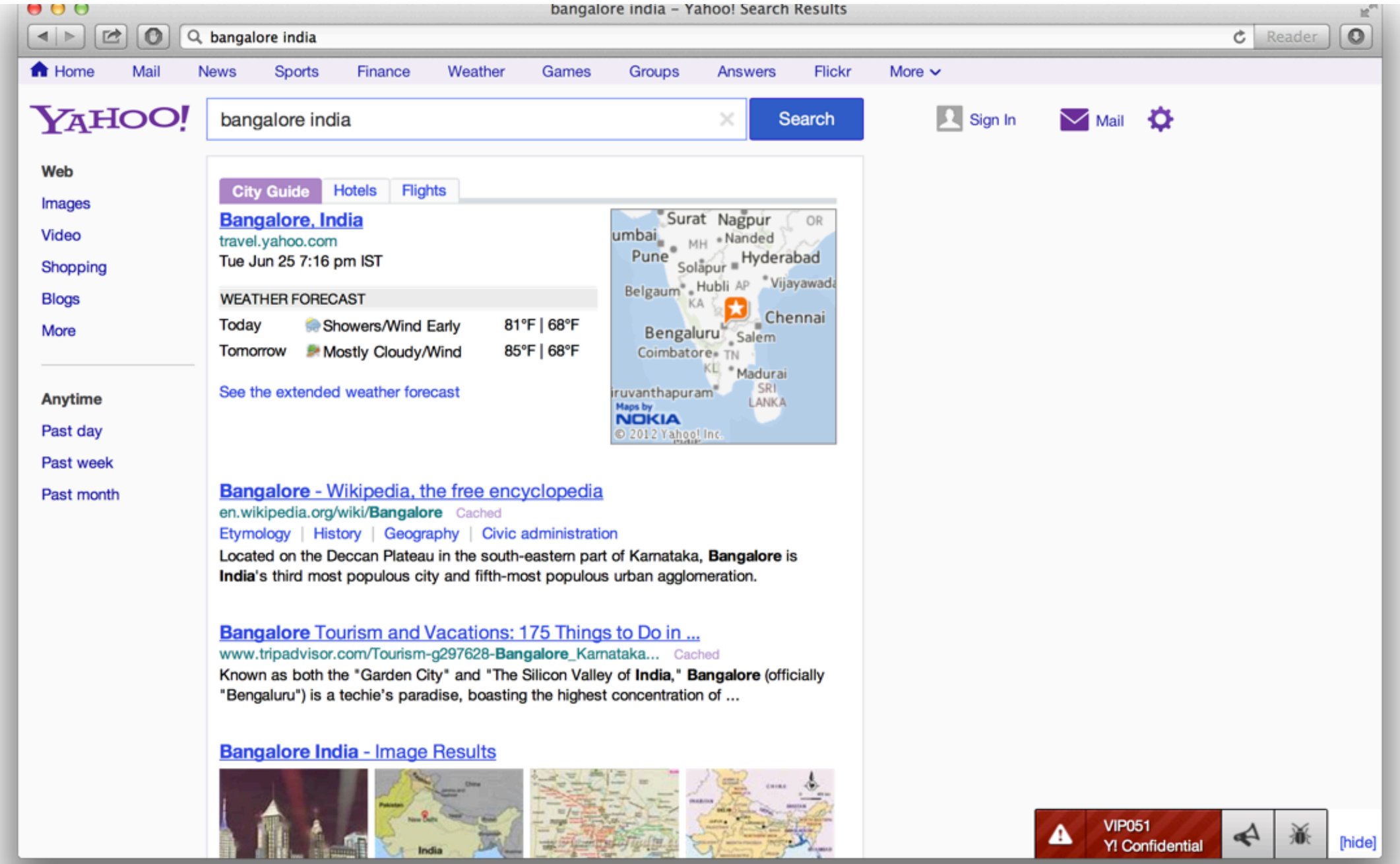
Bing


bangalore, india - Google Search
www.google.co.in/#sclient=psy-ab&q=bangalore%2C+india&oq=bangalore%2C+india&gs_l=hp.3..0l4.7068.7068.0.8320.1.1.0.0.0.0.166.166.0j1.
Reader
+You Search Images Maps Play YouTube News Gmail Drive Calendar More
bangalore, india
SIGN IN
Web Images Maps News More Search tools
About 142,000,000 results (0.39 seconds)
Bangalore - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Bangalore
Bangalore is well-known as the hub of India's information technology sector. ... A succession of South Indian dynasties ruled the region of Bangalore until in ...
List of tourist attractions - History of Bangalore - Bengaluru International Airport
Bangalore - Maps of India
www.mapsofindia.com/bangalore/
Mar 26, 2013 – Are you looking for information on Bangalore? Get detailed information on with Bangalore city covering Demographics, History, Transportation, ...
Images for bangalore, india - Report images
Bengaluru (Bangalore), India - Travel Guide, Info & Bookings ...
www.lonelyplanet.com/india/bengaluru-bangalore
Nov 10, 2008
Bengaluru (Bangalore) travel recommendations and tips from Lonely Planet. Discover 83 things to do & 263 ...
5:39
WHY BANGALORE IS SO FAMOUS CITY IN INDIA and IN THE ...
www.youtube.com/watch?v=hDpsUTAi1JA
Bangalore, Karnataka
Doddaballapur
Gubbi
Kolar
Yeshwanthpura
Kunigal
Vijayanagar
Bangalore
Jayanagar
Kolar Gold Fields
Channapatna
Anekal
Hosur
Mandya
Kanakapura
Bangalore
City in India
Bangalore is the capital city of the Indian state of Karnataka. Located on the Deccan Plateau in the south-eastern part of Karnataka, Bangalore is India's third most populous city and fifth-most populous urban agglomeration. Wikipedia
Population: 8.426 million (2011)
Area: 741 km²
Founded: 1537

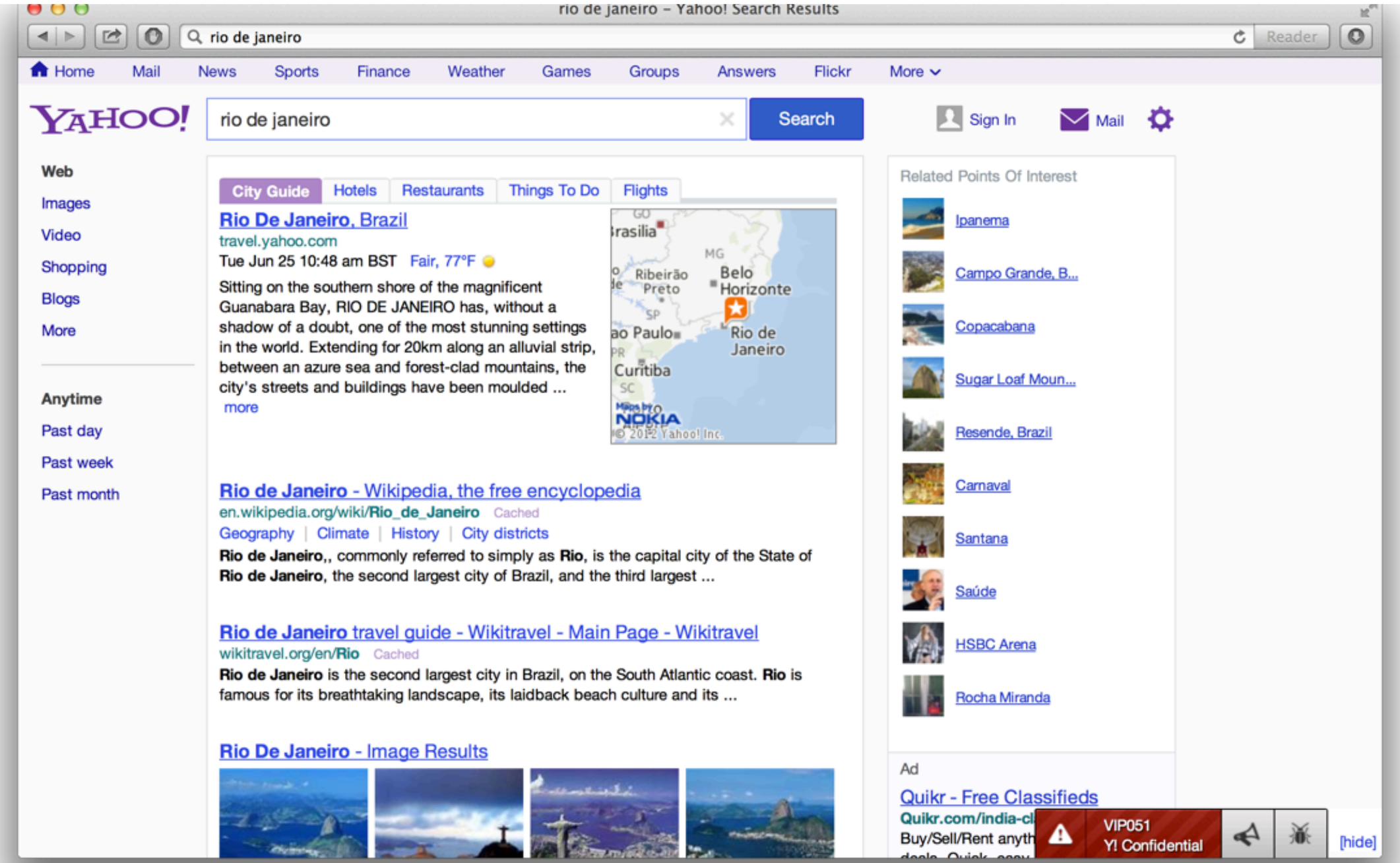
Google


Yahoo! Search - Web Search
search.yahoo.com
Reader
Check out the new Yahoo.com. Access Search, Mail and a virtually endless stream of content customized just for you. Try it now!
YAHOO!
Web Images Video Local Shopping News More
bangalore, india
Search
bangalore india
bangalore india map
bangalore india time
bangalore india weather
bangalore india airport
bangalore india hotels
bangalore india news
bangalore india real estate
bangalore india postal code
bangalore india pictures
BANGALORE, INDIA
04:43 PM (Asia/Kolkata). - Current local time
TOP RATED THINGS TO DO
1. Group of Monuments at Hampi
BANGALORE OVERVIEW
Hotels
Flights
Surat
Nagpur
Nanded
Pune
Hyderabad
Solapur
Hubli
Belgaum
Vijayawada
Chennai
Bengaluru
Salem
Coimbatore
Madurai
TRAVEL

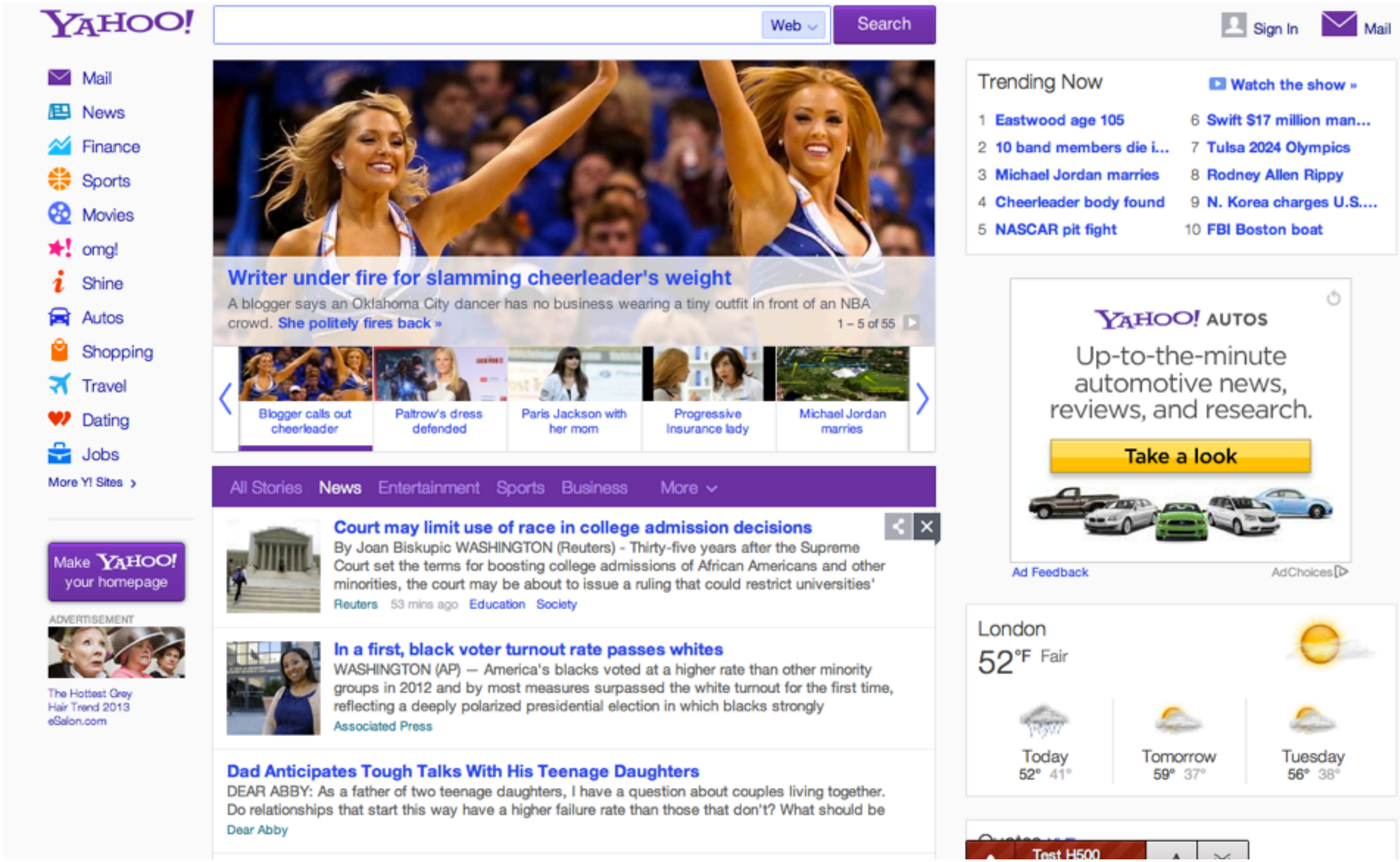
Yahoo!


bangalore india – Yahoo! Search Results
bangalore india
Reader
Home Mail News Sports Finance Weather Games Groups Answers Flickr More
YAHOO!
bangalore india
Search
Sign In
Mail
Web
Images
Video
Shopping
Blogs
More
Anytime
Past day
Past week
Past month
City Guide
Hotels
Flights
Bangalore, India
travel.yahoo.com
Tue Jun 25 7:16 pm IST
WEATHER FORECAST
Today Showers/Wind Early 81°F | 68°F
Tomorrow Mostly Cloudy/Wind 85°F | 68°F
See the extended weather forecast
Surat
Nagpur
Nanded
Pune
Hyderabad
Solapur
Belgaum
Hubli
Vijayawada
Chennai
Bengaluru
Salem
Coimbatore
Madurai
Maps by NOKIA
© 2012 Yahoo! Inc.
Bangalore - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bangalore Cached
Etymology | History | Geography | Civic administration
Located on the Deccan Plateau in the south-eastern part of Karnataka, Bangalore is India's third most populous city and fifth-most populous urban agglomeration.
Bangalore Tourism and Vacations: 175 Things to Do in ...
www.tripadvisor.com/Tourism-g297628-Bangalore_Karnataka... Cached
Known as both the "Garden City" and "The Silicon Valley of India," Bangalore (officially "Bengaluru") is a techie's paradise, boasting the highest concentration of ...
Bangalore India - Image Results
India
VIP051
Y! Confidential
[hide]

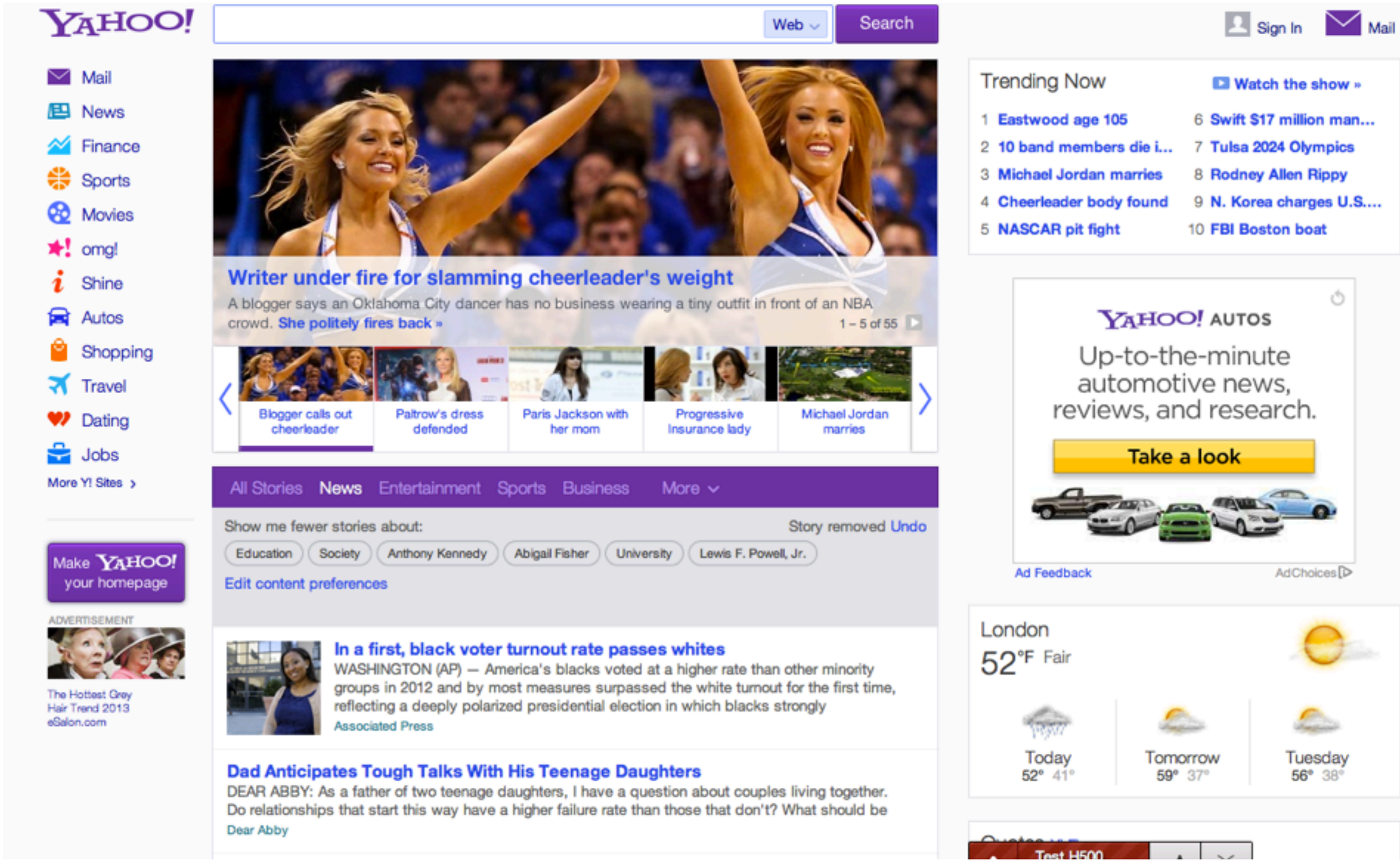
Yahoo!


rio de janeiro – Yahoo! Search Results
rio de janeiro
Reader
Home Mail News Sports Finance Weather Games Groups Answers Flickr More
YAHOO!
rio de janeiro
Search
Sign In
Mail
Web
Images
Video
Shopping
Blogs
More
Anytime
Past day
Past week
Past month
City Guide
Hotels
Restaurants
Things To Do
Flights
Rio De Janeiro, Brazil
travel.yahoo.com
Tue Jun 25 10:48 am BST Fair, 77°F
Sitting on the southern shore of the magnificent Guanabara Bay, RIO DE JANEIRO has, without a shadow of a doubt, one of the most stunning settings in the world. Extending for 20km along an alluvial strip, between an azure sea and forest-clad mountains, the city's streets and buildings have been moulded ...
more
Brasilia
MG
Ribeirão Preto
Belo Horizonte
SP
São Paulo
Rio de Janeiro
PR
Curitiba
SC
NOKIA
© 2012 Yahoo! Inc.
Rio de Janeiro - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Rio_de_Janeiro Cached
Geography | Climate | History | City districts
Rio de Janeiro,, commonly referred to simply as Rio, is the capital city of the State of Rio de Janeiro, the second largest city of Brazil, and the third largest ...
Rio de Janeiro travel guide - Wikitravel - Main Page - Wikitravel
wikitravel.org/en/Rio Cached
Rio de Janeiro is the second largest city in Brazil, on the South Atlantic coast. Rio is famous for its breathtaking landscape, its laidback beach culture and its ...
Rio De Janeiro - Image Results
Related Points Of Interest
Ipanema
Campo Grande, B...
Copacabana
Sugar Loaf Moun...
Resende, Brazil
Carnaval
Santana
Saúde
HSBC Arena
Rocha Miranda
Ad
Quikr - Free Classifieds
Buy/Sell/Rent anyth
VIP051
Y! Confidential
[hide]

Yahoo!



Yahoo! Homerun



Yahoo! Homerun

Writer under fire for slamming cheerleader's weight

A blogger says an Oklahoma City dancer has no business wearing a tiny outfit in front of an NBA crowd. **She politely fires back »**

1 – 5 of 55


Blogger calls out cheerleader


Paltrow's dress defended


Paris Jackson with her mom


Progressive Insurance lady


Michael Jordan marries

All Stories | News | Entertainment | Sports | Business | More

Show me fewer stories about: Story removed Undo

Education | Society | Anthony Kennedy | Abigail Fisher | University | Lewis F. Powell, Jr.

Edit content preferences

Trending Now

1. Eastwood age 105
2. 10 band members die i...
3. Michael Jordan marries
4. Cheerleader body found
5. NASCAR pit fight

London

Yahoo! Homerun

Goals of part I

- Learn entity linking basics
- Get familiar with
  - terminology and essentials
  - seminal papers/methods
  - evaluation and datasets
- Obtain experience with
  - (publicly available) toolkits
  - evaluation

Why do we need entity linking?

- (Automatic) document enrichment
  - go-read-here
  - assistance for (Wikipedia) editors
  - inline (microformats, RDFa)

Why do we need entity linking?

- “Use as feature”
- to improve
  - classification
  - retrieval
  - word sense disambiguation
  - semantic similarity
  - ...
- dimensionality reduction (as compared to, e.g., term vectors)

Why do we need entity linking?

- Enable
  - semantic search
  - advanced UI/UX
  - ontology learning, KB population
  - ...

A bit of history

- Text classification
- NER
- WSD
- NED/NEN
  - {person name, geo, movie name, ...} disambiguation
  - (Cross-document) coreference resolution
  - Automatic link generation
- Entity linking

Entity linking?

- NE normalization / canonicalization / sense disambiguation
- DB record linkage / schema mapping
- Knowledge base population
- Entity linking
  - D2W
  - Wikification
  - Semantic linking

Entity Linking: main problem

- Linking free text to *entities*
  - Entities (typically) taken from a knowledge base
    - Wikipedia
    - Freebase
    - ...
  - Any piece of text
    - news documents
    - blog posts
    - tweets
    - queries
    - ...

Typical steps

1. Determine “linkable” phrases
   - mention detection – **MD**
2. Rank/Select candidate entity links
   - link generation – **LG**
   - may include NILs (null values, i.e., no target in KB)
3. (Use “context” to disambiguate/filter/improve)
   - disambiguation – **DA**

Methods

Preliminaries

- Wikipedia
- Wikipedia-based measures
  - commonness
  - relatedness
  - keyphraseness

Wikipedia

- Basic element: article (proper)
- But also
  - redirect pages
  - disambiguation pages
  - category/template pages
  - admin pages
- Hyperlinks
  - use “unique identifiers” (URLs)
    - [[United States]] or [[United States|American]]
    - [[United States (TV series)]] or
      [[United States (TV series)|TV show]]

Wikipedia style guidelines

- “the lead contains a quick summary of the topic's most important points, and each major subtopic is detailed in its own section of the article”
  - “The lead section (also known as the lead, introduction or intro) of a Wikipedia article is the section before the table of contents and the first heading. The lead serves as an introduction to the article and a summary of its most important aspects.”

See http://en.wikipedia.org/wiki/Wikipedia:Summary_style

Disambiguation pages

- Senses of a phrase
- Short description
- (Possible) categorization
- Non-exhaustive

Some statistics

- WordNet
  - 80k entity definitions
  - 115k surface forms
  - 142k senses (entity - surface form combinations)
- Wikipedia (only)
  - ~4M entity definitions
  - ~12M surface forms
  - ~24M senses

Wikipedia-based measures

Wikipedia-based measures

- keyphraseness(w) **[Mihalcea & Csomai 2007]**

$$\frac{\mathrm{CF}(w_l)}{\mathrm{CF}(w)}$$

Wikipedia-based measures

- keyphraseness(w) **[Mihalcea & Csomai 2007]**

$$\frac{\mathrm{CF}(w_l)}{\mathrm{CF}(w)}$$

Collection frequency
term w as a link to another Wikipedia article

Collection frequency
term w

Wikipedia-based measures

- commonness(w,c) [Medelyan et al. 2008]

$$\frac{|L_{w,c}|}{\sum_{c'} |L_{w,c'}|}$$

Number of links
with target c' and anchor text w

Commonness and keyphraseness

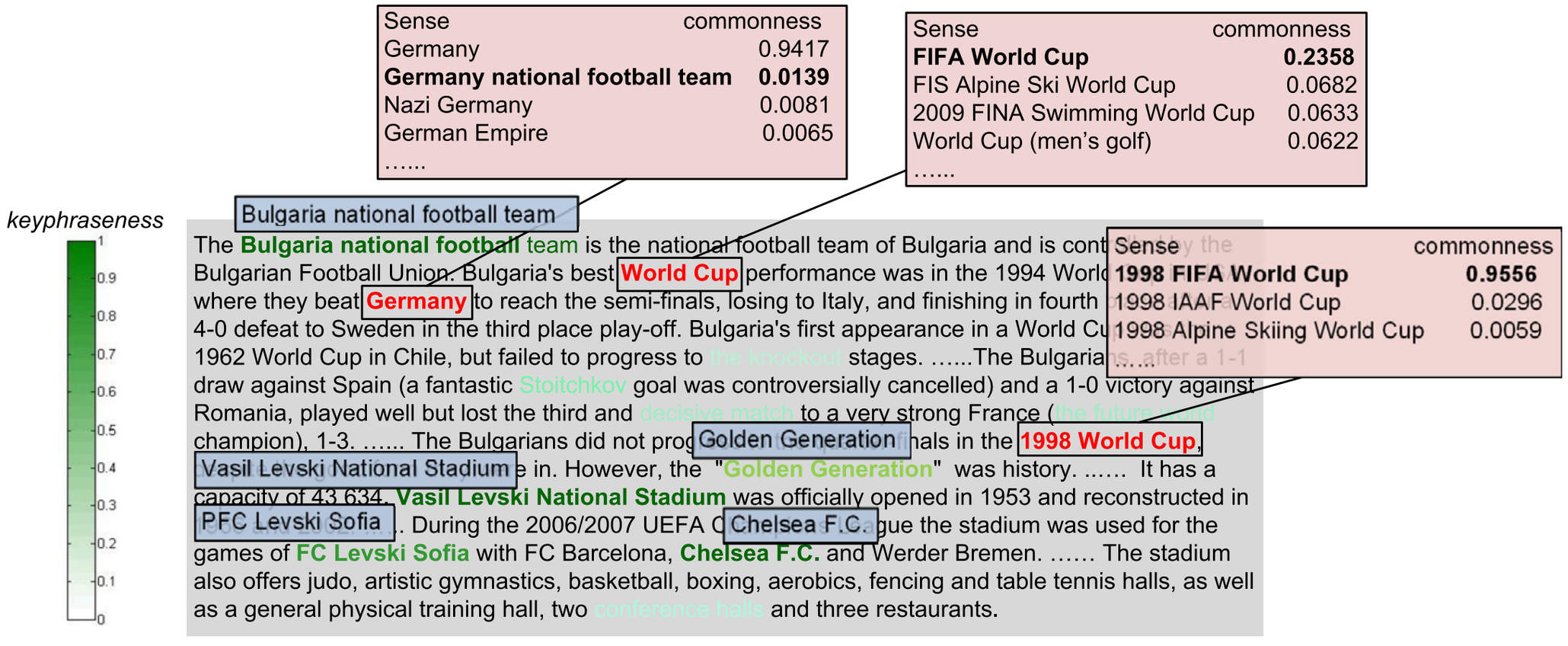


Image taken from Li et al. (2013). **TSDW: Two-stage word sense disambiguation using Wikipedia**. In JASIST 2013.

Wikipedia-based measures

- relatedness(c, c') **[Milne & Witten 2008a]**

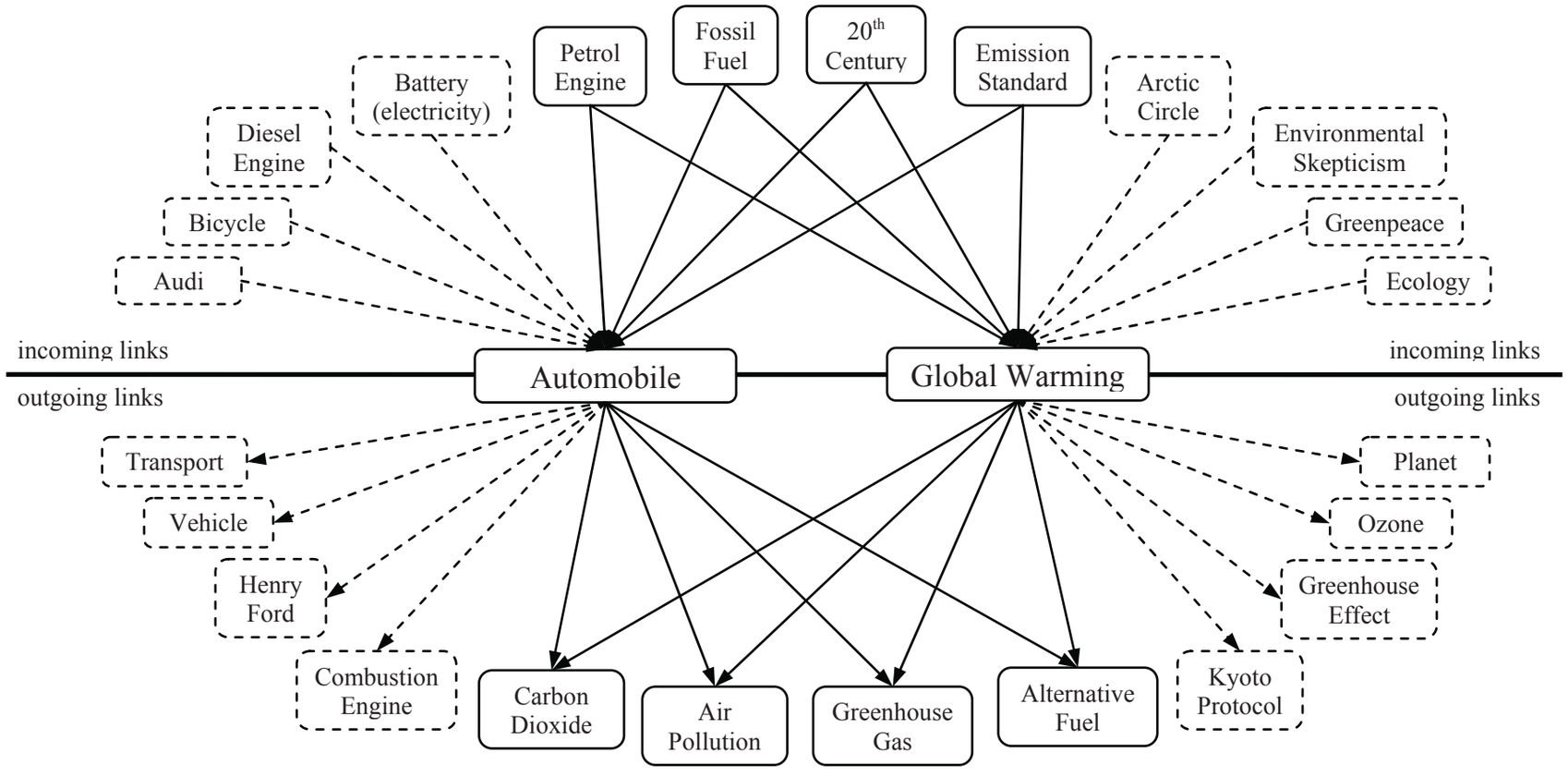


Image taken from Milne and Witten (2008a). **An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links**. In AAAI WikiAI Workshop.

Wikipedia-based measures

- relatedness(c, c') **[Milne & Witten 2008a]**

Number of links with target c

Intersection of inlinks with target c and c'

$$\frac{\log(\max(|L_c|, |L_{c'}|)) - \log(|L_c \cap L_{c'}|)}{\log(|WP|) - \log(\min(|L_c|, |L_{c'}|))}$$

Total number of Wikipedia articles

Baseline methods

Recall the steps

1. mention detection – **MD**
2. link generation – **LG**
3. (disambiguation) – **DA**

Large-Scale Named Entity Disambiguation Based on Wikipedia Data

[Cucerzan 2007]

- Key intuition: leverage context links
  - '''**Texas**''' is a [[pop music]] band from [[Glasgow]], [[Scotland]], [[United Kingdom]]. They were founded by [[Johnny McElhone]] in [[1986 in music|1986]] and had their performing debut in [[March]] [[1988]] at ...
- Prune the candidates, keep only:
  - appearances in the first paragraph of an article, and
  - reciprocal links

Large-Scale Named Entity Disambiguation Based on Wikipedia Data

[Cucerzan 2007]

- MD
  - NER; rule-based; co-ref resolution
- LG
  - Represent entities as vectors
    - context, categories
  - Same for all candidate entity links
  - Determine maximally coherent set

Wikify!

[Mihalcea & Csomai 2007]

- MD
  - tf.idf, X^2, keyphraseness
- LG
  1. Overlap between definition (Wikipedia page) and context (paragraph) **[Lesk 1986]**
  2. Naive Bayes **[Mihalcea 2007]**
     - context, POS, entity-specific terms
  3. Voting between (1) and (2)

Topic Indexing with Wikipedia

[Medelyan et al. 2008]

- MD
  - keyphraseness **[Mihalcea & Csomai 2007]**
- LG
  - combination of average relatedness & commonness
- LG/DA
  - Naive Bayes
    - TF.IDF, position, length, degree, weighted keyphraseness

Learning to Link with Wikipedia

[Milne & Witten 2008b]

- Key idea: disambiguation informs detection
  - compare each possible sense with its *relatedness* to the context sense candidates
  - start with unambiguous senses
  - So, first LG, then base MD on these results

Learning to Link with Wikipedia

[Milne & Witten 2008b]

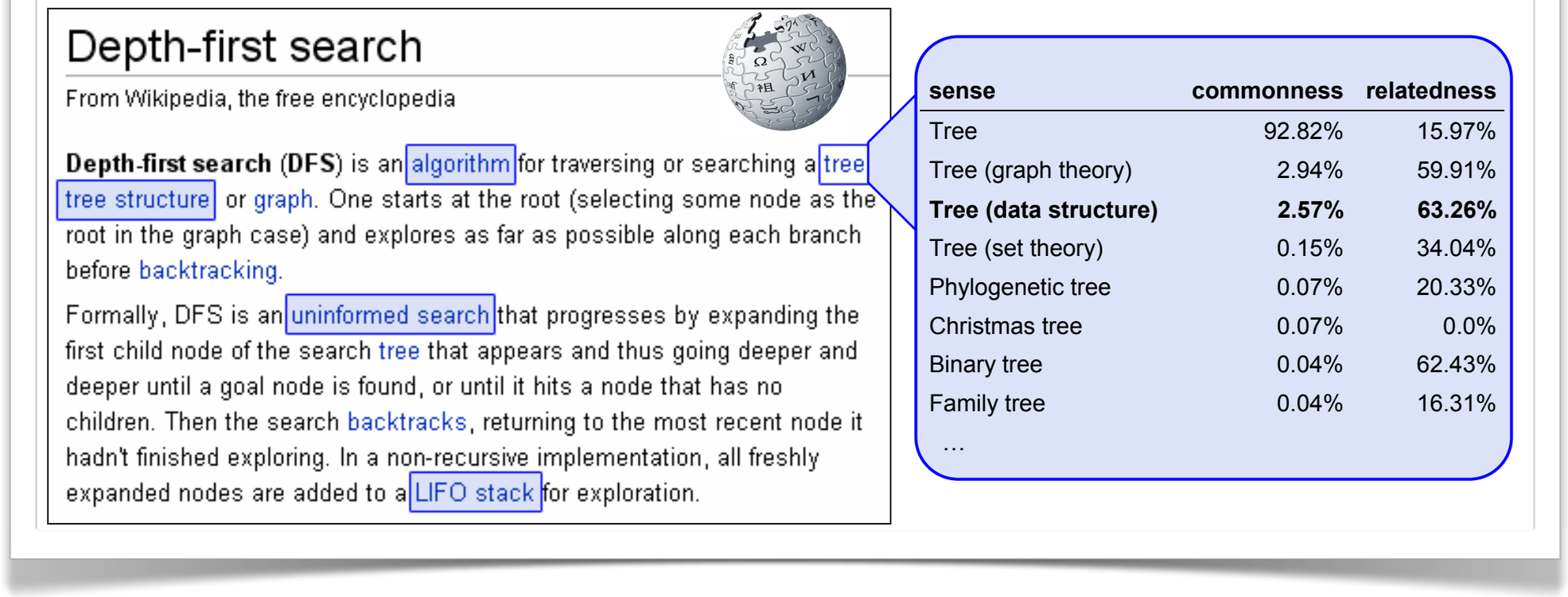


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

Learning to Link with Wikipedia

[Milne & Witten 2008b]

- Filter non-informative, non-ambiguous candidates (e.g., “the”)
  - based on keyphraseness, i.e., link probability
- Filter non-central candidates
  - based on average relatedness to all other context senses
- Combine

Learning to Link with Wikipedia

[Milne & Witten 2008b]

- MD
  - ...
- LG
  - Machine learning
    - keyphraseness, average relatedness, sum of average weights

Learning to Link with Wikipedia

[Milne & Witten 2008b]

- MD
  - Machine learning
    - link probability, relatedness, **confidence of LG**, generality, frequency, location, spread
- LG
  - Machine learning
    - keyphraseness, average relatedness, sum of average weights

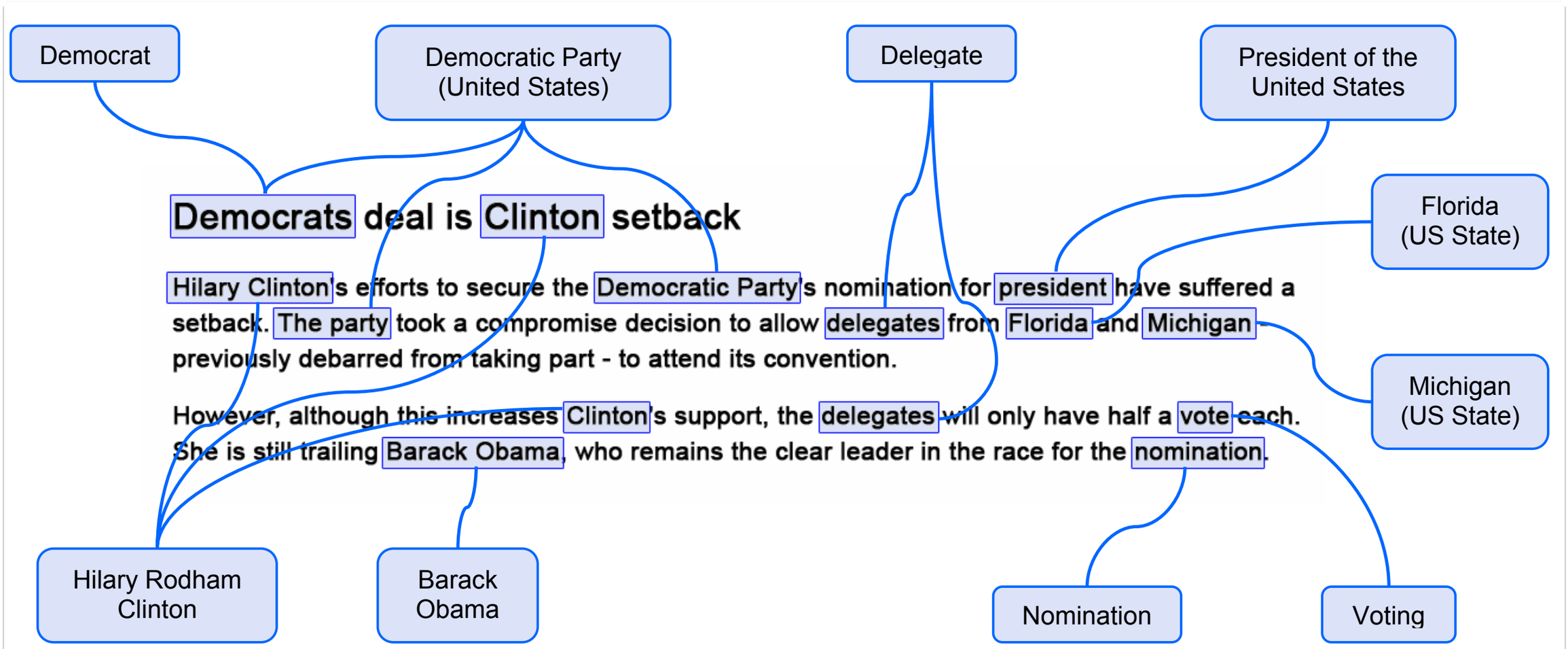


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

Local and Global Algorithms for Disambiguation to Wikipedia

[Ratinov et al. 2011]

- Explicit focus on *global* versus *local* algorithms
  - "Global," i.e., disambiguation of the candidate graph
  - NP-hard
- Optimization
  - reduce the search space to a "disambiguation context," e.g.,
    - all plausible disambiguations **[Cucerzan 2007]**
    - unambiguous surface forms **[Milne & Witten 2008b]**

Local and Global Algorithms for Disambiguation to Wikipedia

[Ratinov et al. 2011]

- Main contribution, in steps
  1. use “local” approach (e.g., commonness) to generate a disambiguation context
  2. apply “global” machine learning approach
     - relatedness, PMI
       - {inlinks, outlinks} in various combinations (c and c’)
       - {avg, max}

- Finally, apply another round of machine learning

TAGME: On-the-fly Annotation of Short Text Fragments

[Ferragina & Scaiella 2010]

- MD
  - keyphraseness **[Mihalcea & Csomai 2007]**
- LG
  - use “local” approach to generate a disambiguation context, similar to **[Ratinov et al. 2011]**
  - Heavy pruning
    - mentions; candidate links; coherence
- Accessible at http://tagme.di.unipi.it

Adding semantics to microblog posts

[Meij et al. 2012]

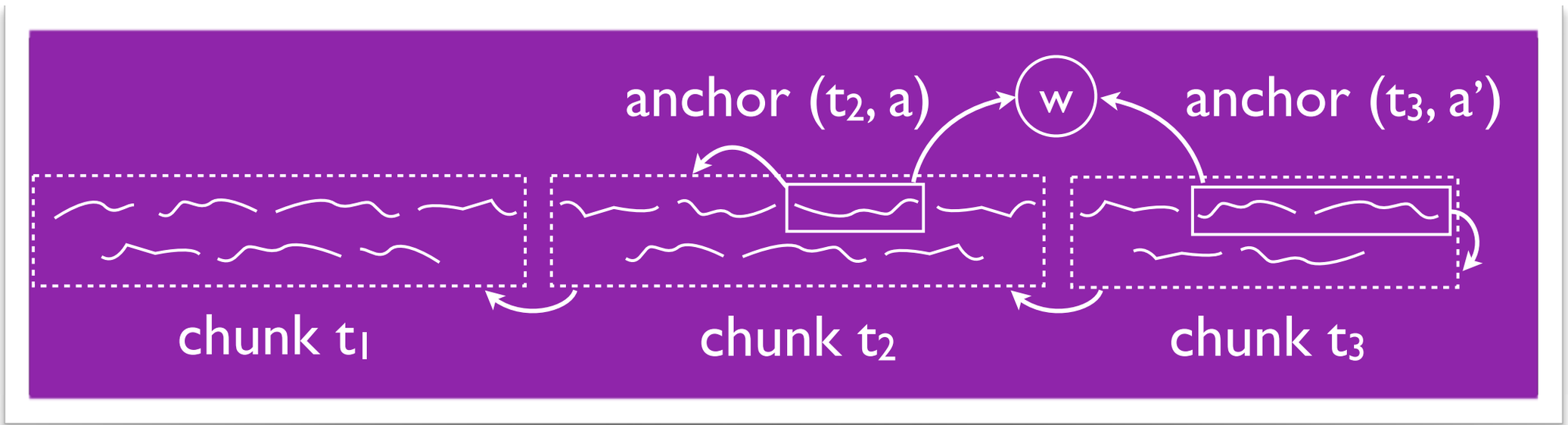
- MD
  - commonness (and others)
  - idea: obtain ranked list of **all** candidate entity links
- LG
  - use machine learning to determine which of the links to keep
    - “point-wise LeToR”
    - ..., random forests, GBRT
    - big set of {text, entity, text+entity, context} features

Graph-based methods

Feeding the Second Screen: Semantic Linking based on Subtitles

[Odijk et al. 2013]

- Setting: entity linking on closed captions
  - streaming, high-precision, real-time
- Graph information as additional features
  - Idea: maintain a (coherent) tripartite context graph
    - entities
    - chunks
    - anchors



Feeding the Second Screen: Semantic Linking based on Subtitles

[Odijk et al. 2013]

Context features	
$DEGREE(w, G)$	Number of edges connected to the node representing Wikipedia article w in context graph G.
$DEGREE-CENTRALITY(w, G)$	Centrality of Wikipedia article w in context graph G, computed as the ratio of edges connected to the node representing w in G.
$PAGERANK(w, G)$	Importance of the node representing w in context graph G, measured using PageRank.

A Graph-based Method for Entity Linking

[Guo et al. 2011]

- MD
  - rule-based; prefer longer links
  - generate a disambiguation context
- LG
  - (weighted interpolation of) in- and outdegree in disambiguation context to select entity links
    - edges defined by wikilinks
- Evaluation on TAC KBP

Graph-based named entity linking with Wikipedia

[Hachey et al. 2011]

- MD
  - generate disambiguation context
    - based on unambiguous entity links
    - edges defined by wikilinks (articles & categories)
      - max step size: 2 (articles), 3 (categories)
- LG
  - use degree centrality and PageRank to reweigh cosine-based similarity scores
- Evaluation on TAC KBP

Recap

- Essential ingredients
  - MD
    - commonness
    - keyphraseness
  - LG
    - commonness
    - machine learning
  - DA
    - relatedness
    - machine learning

Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - hands-on
  - open challenges

Evaluation

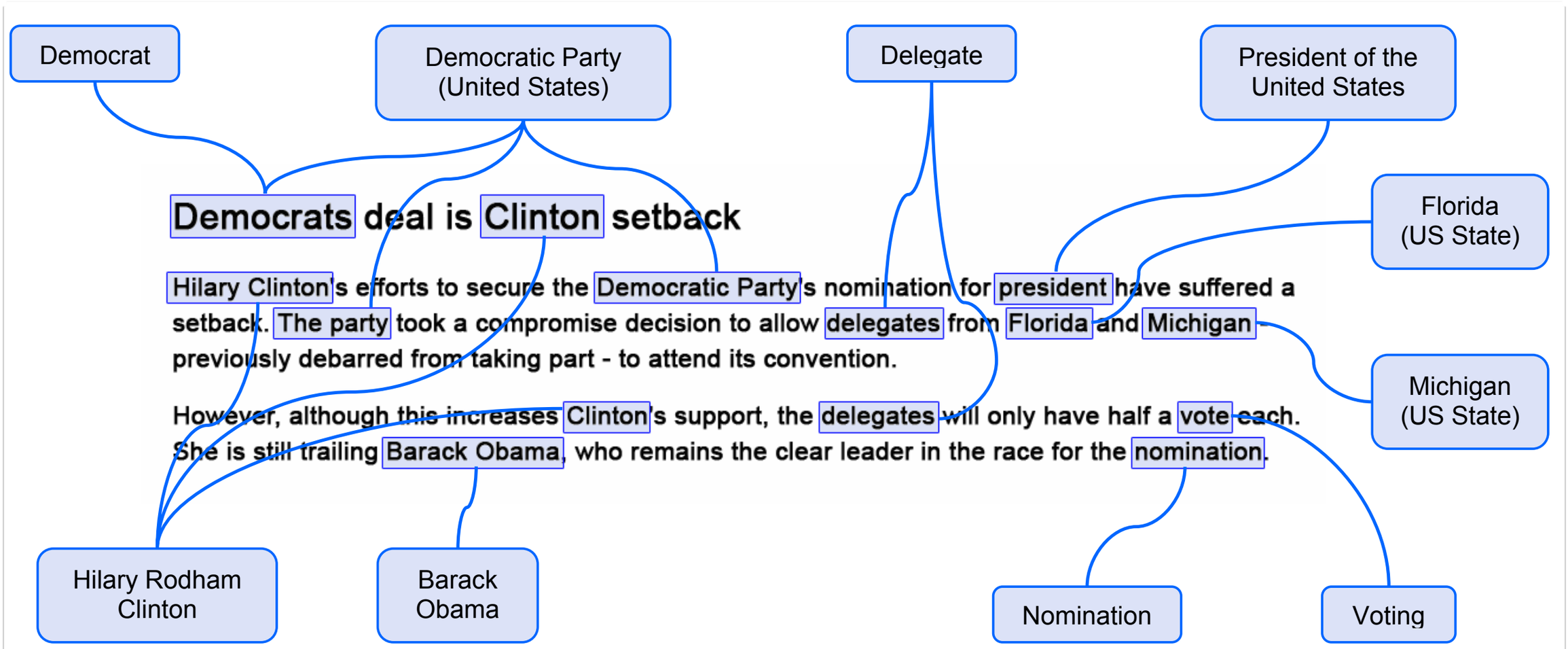


Image taken from Milne and Witten (2008b). **Learning to Link with Wikipedia**. In CIKM '08.

DIY Entity Linking

- Ingredients
  - target KB (e.g., Wikipedia)
  - test collection
  - evaluation metrics

DIY Entity Linking

- Ingredients
  - Target KB (Wikipedia)
    - wikipedia-miner
    - Google's Dictionaries for Linking Text, Entities and Ideas
  - Test collection
  - Evaluation metrics

Evaluation metrics

- Set-based (similar to WSD)
  - “How many correct links were retrieved?”
  - macro/micro
    - precision, recall, F-measure
- Rank-based

Common set-based metrics

- Accuracy

$$A = \frac{|\{\mathcal{C}_{i,0}|\mathcal{C}_{i,0} = \mathcal{G}\}|}{N}$$

- Precision

$$P_{\mathcal{C}} = \frac{|\{\mathcal{C}_i|\mathcal{C}_i \neq \emptyset \wedge \mathcal{G}_i \in \mathcal{C}_i\}|}{|\{\mathcal{C}_i|\mathcal{C}_i \neq \emptyset\}|}$$

- Recall

$$R_{\mathcal{C}} = \frac{|\{\mathcal{C}_i|\mathcal{G}_i \neq \text{NIL} \wedge \mathcal{G}_i \in \mathcal{C}_i\}|}{|\{\mathcal{G}_i|\mathcal{G}_i \neq \text{NIL}\}|}$$

N	Number of queries in data set
$\mathcal{G}$	Gold standard annotations for data set ($\|\mathcal{G}\| = N$)
$\mathcal{G}_i$	Gold standard for query i (KB ID or NIL)
$\mathcal{C}$	Candidate sets from system output ($\|\mathcal{C}\| = N$)
$\mathcal{C}_i$	Candidate set for query i
$\mathcal{C}_{i,j}$	Candidate at rank j for query i (where $\mathcal{C}_i \neq \emptyset$)

Common rank-based metrics

- Recall @ k
- Precision @ k
- R-precision
- Mean average precision
- Mean reciprocal rank

Test collections

Entity linking test collections

- Wikipedia
- MSNBC
- AQUAINT
- ACE
- Twitter
- AIDA (CoNLL)
- IITB (web data)
- INEX link-the-wiki
- TREC knowledge base acceleration (KBA)
- TAC knowledge base population (KBP)

Wikipedia (for evaluation)

- Widely used
- Pros
  - cheap and easy; the links are already provided
- Cons
  - biased (style guides!)
  - specific scenario
  - unbalanced

MSNBC

[Cucerzan 2007]

- 20 news articles
  - linked to EN Wikipedia from 2006
  - 756 total links; 127 of these are NIL
- Focus: disambiguate entities after NER and co-reference resolution
  - all mentions of all the detected entities are linked
- Con
  - collected by correcting the output of Cucerzan's system

See http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/

AQUAINT

[Milne & Witten 2008]

- 50 news articles
  - 449 links, obtained using Amazon mechanical turk
- Subset of AQUAINT newswire, annotated to mimic Wikipedia hyperlink structure
  - only first mentions of “important” titles were linked
  - uninteresting and redundant mentions of the same title not linked

---

See http://cogcomp.cs.illinois.edu/page/resources/data

ACE

[Ratinov et al. 2011]

- Subset of ACE co-reference data set
  - mentions and their types are given
  - co-references resolved
- First nominal mentions of each co-reference chain are linked
  - Amazon mechanical turk
  - accuracy of majority vote ~85%
  - manually corrected

See http://cogcomp.cs.illinois.edu/page/resources/data

Twitter

[Meij et al. 2012]

- Tweets taken from “verified accounts,” so relatively clean
- ~500 tweets, manually linked to Wikipedia
  - ~2 entity links per tweet on average

See http://edgar.meij.pro/dataset-adding-semantics-microblog-posts/

Task	Name	Year	Source	All Mentions	Instances
CDCR	John Smith	1998	News	✘	197
CDCR	WePS 1	2007	Web	✘	3,489
CDCR	Day et al.	2008	News	✔	3,660
CDCR	WePS 2	2008	Web	✘	3,432
CDCR	WePS 3	2009	Web	✘	31,950
Wikify	Mihalcea	2007	Wiki	✔	7,286
Wikify	Kulkarni	2009	Web	✔	17,200
Wikify	Milne	2010	Wiki	✔	11,000
NEL	Cucerzan	2007	News	✔	797
NEL	TAC 09	2009	News	✘	3,904
NEL	Fader	2009	News	✘	500
NEL	TAC 10	2010	News, Blogs	✘	3,750
NEL	Dredze	2010	News	✘	1,496
NEL	Bentivogli	2010	News, Web, Transcripts	✔	16,851
NEL	Hoffart	2011	News	✔	34,956

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

TAC

[McNamee et al. 2010]

- Target: KB – from Wikipedia (~800k instances)
  - infoboxes; article text; type
- “Query”
  - document ID (news, web, blog)
  - mention string (occurring at least once in that doc)
- Focus on ambiguous mentions
  - collected by cherry-picking ‘interesting’ mentions, rather than systematically annotating all mentions
- Explicit NILs (> 50% of the queries)

	TAC 2009 test		TAC 2010 train		TAC 2010 test	
$\|\mathcal{Q}\|$	3,904		1,500		2,250	
KB	1,675	(43%)	1,074	(72%)	1,020	(45%)
NIL	2,229	(57%)	426	(28%)	1,230	(55%)
PER	627	(16%)	500	(33%)	751	(33%)
ORG	2710	(69%)	500	(33%)	750	(33%)
GPE	567	(15%)	500	(33%)	749	(33%)
News	3904	(100%)	783	(52%)	1500	(67%)
Web	0	(0%)	717	(48%)	750	(33%)
Acronym	827	(21%)	173	(12%)	347	(15%)
$\|\mathcal{E}\|$	560		–		871	
KB	182	(33%)	462	(–)	402	(46%)
NIL	378	(67%)	–	(–)	469	(54%)
PER	136	(24%)	–	(–)	334	(38%)
ORG	364	(65%)	–	(–)	332	(38%)
GPE	60	(11%)	–	(–)	205	(24%)

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

Meta-evaluations

- [Hachey et al. 2013]
- [Cornolti et al. 2013]

Evaluating Entity Linking with Wikipedia

[Hachey et al. 2013]

- Named entity linking, a.k.a., “NEL”
  - include NILs
  - Wikipedia articles not always named entities
- Explicit focus on separating “search” (LG) and “disambiguation” (DA)
- Reimplement and evaluate three NEL systems
  - **[Bunescu & Pasça 2006]**
  - **[Cucerzan 2007]**
  - **[Varna et al. 2009]** (TAC system paper)

System	Extractor	Searcher: Condition	Title	Redirect	Link	Truncated	Bold	DABTitle	Filter	Disambiguator
Bunescu and Paşca (2006)	NER	NA	✔	✔				✔	NA	SVM rank over cosine and mention context word×category features
Cucerzan (2007)	NER, coreference expansion	NA	✔	✔	✘	✔		✔	NA	Scalar product between candidate category/term vector and document-level vector
Varma et al. (2009)	NER, acronym expansion	**if** acronym								Cosine between candidate article term vector and mention context vector
		if expandable	✔						in KB	
		else	✔	✔			✔	✔	NA	
		else								
		search 1	✔						in KB	
		if no candidates	✔	✔			✔	✔	NA	

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

Alias Source	$\langle C \rangle$	$P_{\mathcal{C}}^{\infty}$	$R_{\mathcal{C}}^{\infty}$	$P_{\emptyset}$	$R_{\emptyset}$
Title	0.2	**83.5**	37.2	68.1	96.5
Redirect	0.1	74.6	20.0	62.1	96.2
Link	4.2	55.7	**80.1**	**88.6**	59.5
Bold	1.6	45.1	48.8	71.7	67.2
Hatnote	0.0	42.6	1.2	57.7	**99.9**
Truncated	1.2	37.8	24.5	62.2	78.6
DABTitle	3.5	34.2	29.3	58.7	65.1
DABRedirect	2.7	34.0	18.9	57.9	77.3

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

System	A	A_C	$A_\emptyset$
NIL Baseline	57.1	0.0	100.0
Title Baseline	71.0	37.2	96.5
+ Redirect Baseline	76.3	54.6	92.6
Bunescu and Paşca	77.0	67.8	83.8
Cucerzan	78.3	71.3	83.5
Varma et al. Replicated	80.1	72.3	86.0
TAC 09 Median	71.1	63.5	78.9
TAC 09 Max (Varma)	82.2	76.5	86.4

Table taken from Hachey et al. (2013). **Evaluating Entity Linking with Wikipedia**. In AI '13.

Meta-evaluations

- [Hachey et al. 2013]
- [Cornolti et al. 2013]

A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- Compare five publicly available entity linkers
  - **[Hoffart et al. 2007]** (AIDA)
  - **[Ratinov et al. 2011]**
  - **[Ferragina & Scaiella 2010]** (TAGME)
  - **[Milne & Witten 2008]** (wikipedia-miner)
  - DBpedia Spotlight
- And also investigate parameter/cut-off settings

See http://acube.di.unipi.it/ (not available yet).

A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- On five publicly available test collections
  - AIDA **[Hoffart et al. 2007]**
    - based on CoNLL 2003: noun annotations
    - 1393 Reuters newswire articles
    - hand-annotated all nouns with entities in YAGO2
  - AQUAINT **[Milne & Witten 2008]**
  - MSNBC **[Cucerzan 2007]**
  - IITB **[Kulkarni et al. 2010]** (web data)
  - Twitter **[Meij et al. 2012]**

A Framework for Benchmarking Entity-Annotation Systems

[Cornolti et al. 2013]

- Benchmarking framework
- “Fuzzy” evaluation measures
- Main findings
  - Different systems perform well in different scenarios
  - AIDA and TagMe seem to be the winners overall

Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - hands-on
  - open challenges

Hands-on

Public Toolkits and Web Services for Entity Linking

- Wikipedia Miner
- TagMe
- DBpedia Spotlight
- Illinios Wikifier
- AIDA
- (OpenCalais)

Wikipedia Miner

[Milne & Witten 2008b]

- Open source
- (Public) web service
  - Java
  - Hadoop preprocessing pipeline
- Lexical matching + machine learning
- See http://wikipedia-miner.cms.waikato.ac.nz

TagMe

[Ferragina & Scaiella 2010]

- Web service only (demo + API)
- Approach similar to Wikipedia Miner
- Voting for disambiguation
  - based on all possible bindings
  - heuristics to select best target
- Designed for short texts
- See http://tagme.di.unipi.it/

DBpedia Spotlight

[Mendes et al., 2011]

- Open source
- Public web service
- Disambiguation in local context
  - vector-space model using bag-of-words and cosine similarity
  - (actually, Lucene)
- See http://spotlight.dbpedia.org

Illinois Wikifier

[Ratinov et al. 2011]

- Local install + online demo
  - uses Illinois NER system
- Disambiguation as weighted sum of features
  - Textual similarity
  - Global coherence based on link structure
- See http://cogcomp.cs.illinois.edu/page/software_view/33

AIDA

[Yosef et al. 2011]

- Open source
  - uses Stanford NER system
- (Public) web service, API
- Links to YAGO2
- Disambiguation in 3 variants
  - PriorOnly: link to most common target
  - Local: disambiguate individual links with local features
  - CocktailParty: collective disambiguation maximizing coherence using iterative graph-based approach

OpenCalais

- Only on public content
  - does not keep a copy of content
  - keeps a copy of the metadata it extracts
- Free for up to 50,000 documents per day
- Early adopters:
  - CBS Interactive / CNET, Huffington Post, Al Jazeera, The White House
  - more than 30,000 developers && 50 publishers

	Programming Language	Service	Available Languages	Open Source
Wikipedia Miner	Java	Web API, Application	any WP	✔
TagMe	Java	Web API	EN, IT	✖
DBpedia Spotlight	Java	Web API, Application	EN + any WP	✔
Illinois Wikifier	Java	Application	EN	✔
AIDA	Java	Web API	EN	✔
OpenCalais	?	Web API	EN, FR, SP	✖

	Matching	Target KB	Context	Comment
Wikipedia Miner	Lexical	Wikipedia	ML on Relatedness	
TagMe	Lexical	Wikipedia	Vote on Relatedness	Focus on Short texts
DBpedia Spotlight	Lexical?	DBpedia	Cosine Similarity	Structure
Illinois Wikifier	NER	Wikipedia	Global Coherence	
AIDA	NER	YAGO2	Multiple	Structure
OpenCalais	?	Calais	?	

Code Academy

- Contains some (Javascript) coding examples for entity linking and retrieval
  - http://www.codecademy.com/courses/javascript-beginner-en-LkhDf/

Outline

- Part 1 – Entity Linking
  - introduction
  - methods
  - evaluation
  - test collections
  - hands-on
  - open challenges

Open challenges

Open challenges

- Difficulty prediction
  - similar to ambiguity, but not the same
  - dependent on context, candidate links, ...
- Cross-lingual entity linking **[Wang et al. 2013]**
- Cross-KB entity linking (“Freebase”)
  - what if there is no/little textual evidence?
  - directly
    - lexical matching
    - machine learning (if annotators/training data available)
  - use Wikipedia as pivot

Learning/Updating the KB

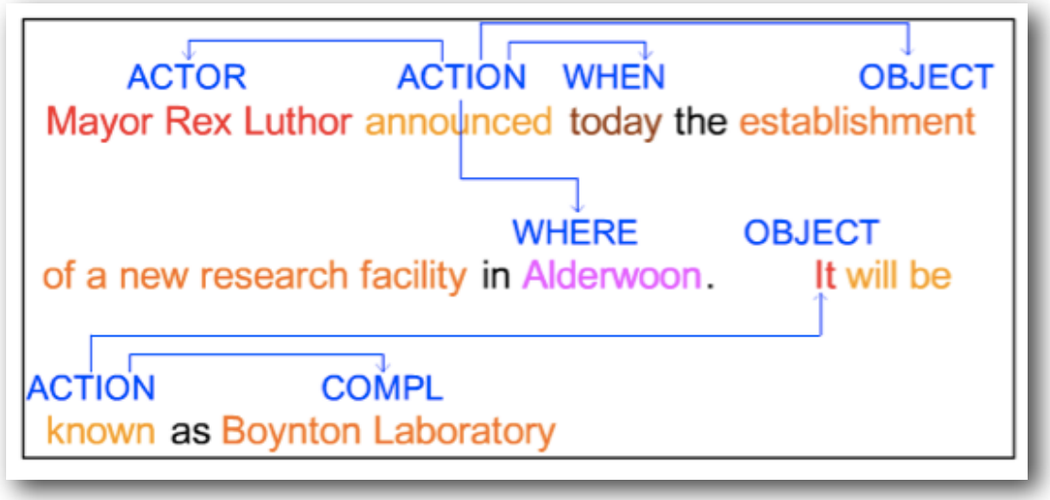
- Link parallel, continuous streams of items
  - news, tweets, blogs, status updates
  - queries, clicks
  - web pages, RDFa/schema.org
  - etc.

- Given an entity
  - “What is new?” What do I need to know now?”
  - Add: personal
  - Add: social

- TREC KBA/KBP/KBx, TREC TS

Learning/Updating the KB: ingredients?

- Accurate entity linking
  - real-time
  - cross-item
  - cross-genre
  - cross-vertical



- What is being said?
  - aspects, attributes, relations, events
- Correlate with already known facts
- Detect bursts, events

Open challenges

- Generic test collections
  - What's the task? User model? Evaluation?
    - TAC? set-based? ranking? known-item finding? top-*r*?
    - exhaustive linking? first mention only?
    - "aboutness"
- Moving beyond entities
  - events/news, concepts, relations
- Moving beyond "ad hoc" entity linking:
  - incorporate contextual evidence in the task (and evaluation)
  - {users, history, profile, social, trending, ...}

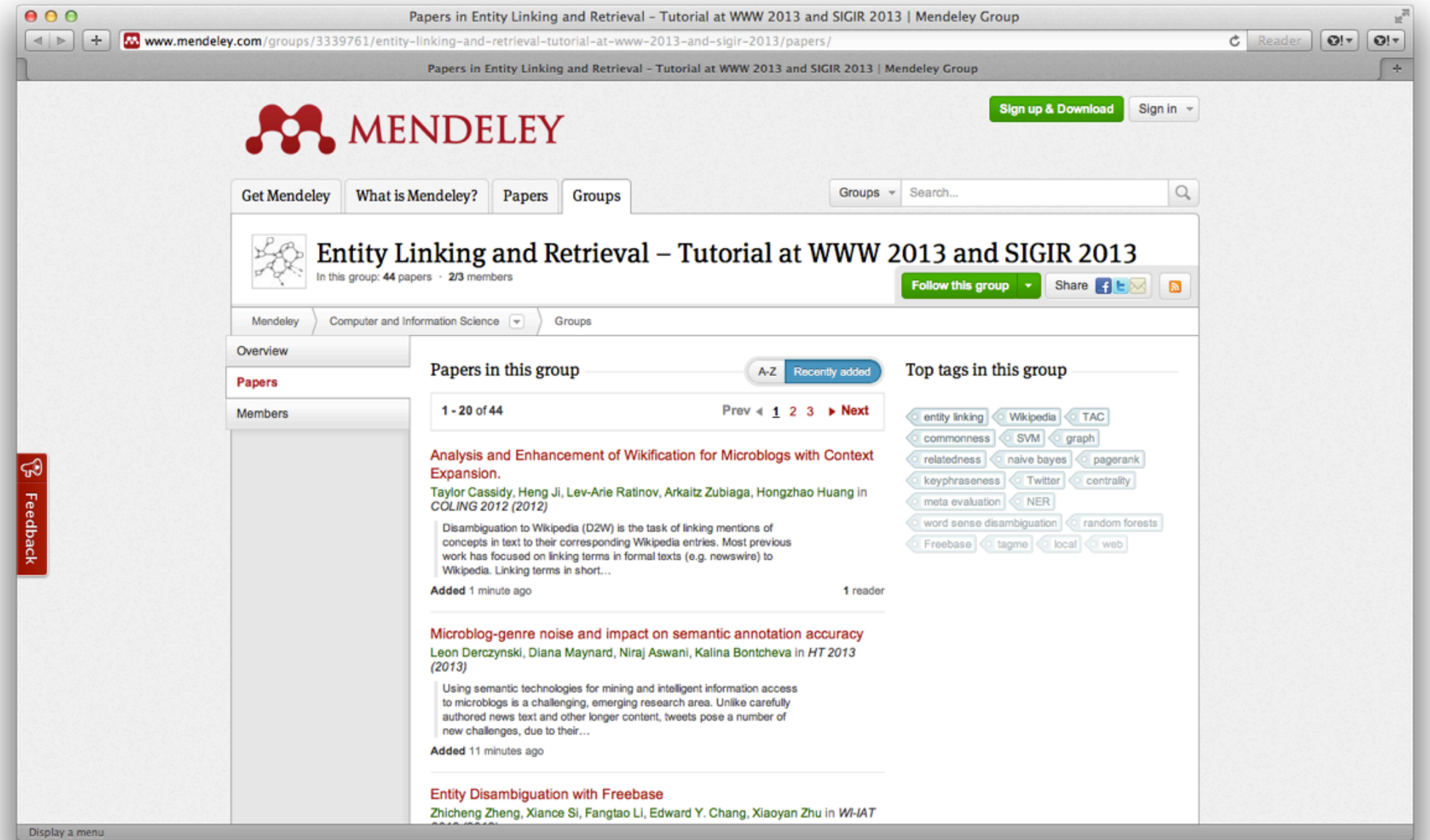
Follow-up reading

- Detecting unlinkable entities **[Lin et al. 2012a]**
- Linking entities to any database **[Sil et al. 2012]**
- Automatically generating Wikipedia articles **[Sauper & Barzilay 2009]**
- Scaling up to the web **[Lin et al. 2012b]**
- Serendipitous suggestions based on personalized entity links **[Bordino et al. 2013]**

References – Entity linking

http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/added/0/tag/entity+linking/

References – Entity linking



http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/added/0/tag/entity+linking/

Part II

Entity Retrieval

Introduction

Entity retrieval tasks

- Ad-hoc entity retrieval
- List completion
- Question answering
  - Factual questions
  - List questions
  - Related entity finding
- Type-restricted variations
  - People, blogs, products, movies, etc.

Indian restaurants in bangalore - Google Maps

maps.google.co.in/maps?bav=on.2,or.r_qf.&bvm=bv.48293060,d.bmk&biw=1158&bih=679&um=1&ie=UTF-8&q=indian+restaurants+in+bangal

Reader

+You Search Images **Maps** Play YouTube News Gmail Drive Calendar More

Google

Indian restaurants in bangalore

SIGN IN

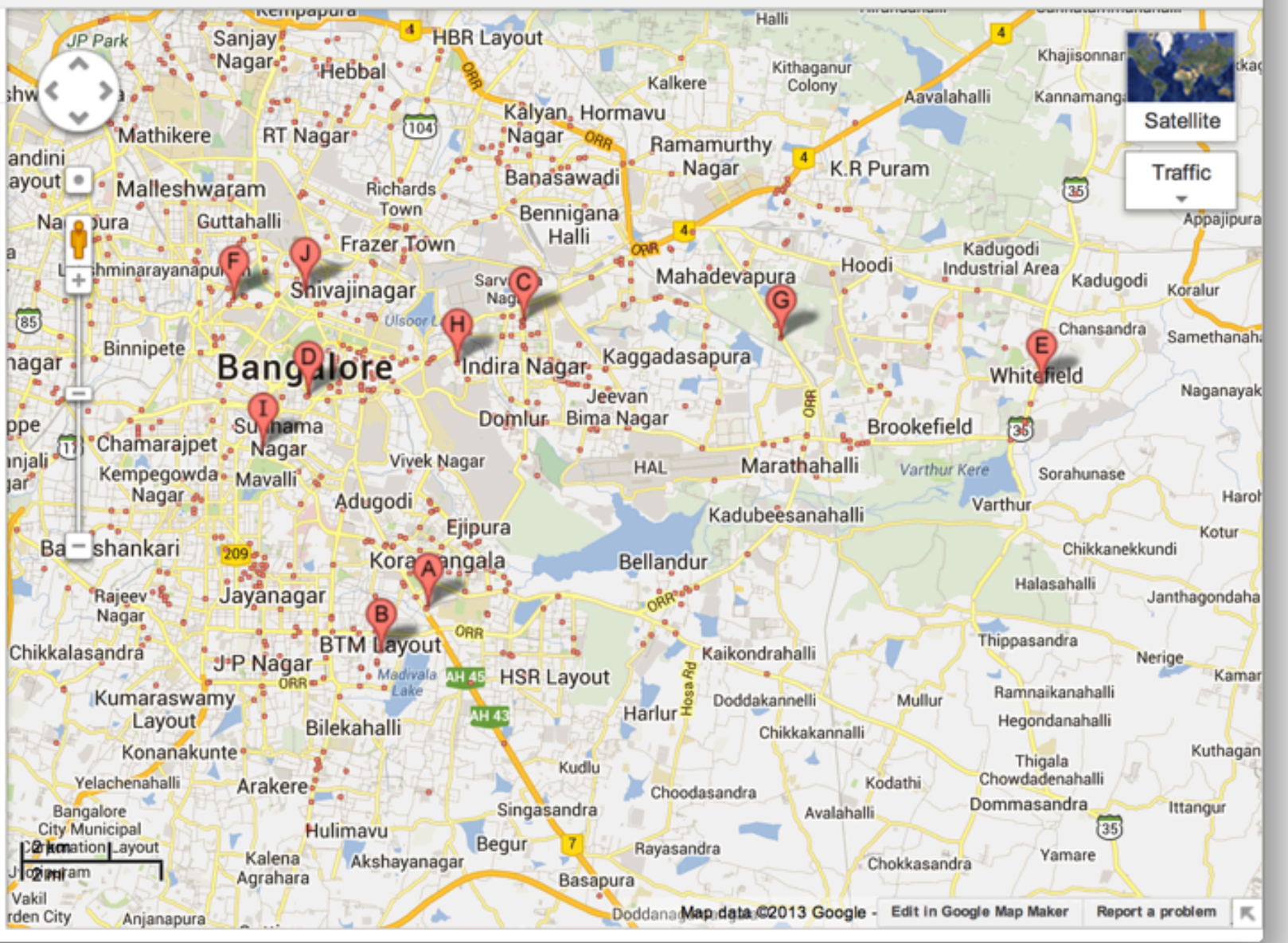
Get directions My places

Indian restaurants near **Bangalore, Karnataka**

A **Serengeti**
Kanyakumari Rd, Bangalore, Karnataka
080 4000 3333 · zomato.com
Category: **Indian Restaurant**
18 14 reviews ·
north indian food · jungle · main course
"Good" -

B **Woodys**
45/1, 5th Cross, 17th Main, J P Nagar, Bangalore, Karnataka 560078
080 2649 0888 · woodlands.in
Category: **Restaurants** - South **Indian**
10 13 reviews ·
"Terrible place. We sat in a nominally air-conditioned place, but the AC ..." -

C **Southindies**
Inner Ring Rd, Indira Nagar, Bangalore, Karnataka 560038
080 4163 6363 · thesouthindies.com
Category: South **Indian Restaurant**
16 21 reviews ·
"good food but small portions. holier than thou service. overpriced. i am not ..." -



expert finding

EXPERTS

language technology

Bogers, Drs. Toine M.	Arts
Bosch, Dr. Antal P. J. van den	Arts
Broeder, Dr. Peter	Arts
Canisius, Drs. Sander V.M.	Arts
Daelemans, Prof. dr. Walter M. P.	Arts
Geertzen, Jeroen	Arts
Keizer, Dr. ir. Simon	Arts
Marsi, Dr. Erwin C.	Arts
Reynaert, Dr. Martin W.C.	Arts
Sporleder, Dr. C.E.	Arts
Werf, Drs. Rintse van der	Arts

See also:

computer linguistics
language technology and computers

EXPERTS
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

EXPERTISE
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

expert profiling

Antal P. J. van den Bosch

Lecturer
Faculty of Arts
Language and information science

Room D 343
P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands

Phone +31 13 466 3117
Fax +31 13 466 2892
E-mail: Antal.vdnBosch@uvt.nl

Present

	mon	tue	wed	thu	fri
morning	✓	✓	✓	✓	
afternoon	✓	✓	✓	✓	

research
study guide
personal homepage

Expertise

My research is positioned in the intersection between artificial intelligence and linguistics. I am specialized in machine learning and language technology / computational linguistics. As for applications, I have professional experience with speech synthesis, the automatic syntactic and semantic analysis of text, text mining, dialogue systems, and spelling correction.

Subjects
artificial intelligence
computer linguistics

What's so special here?

- Entities are not always directly represented
  - Recognize and disambiguate entities in text
  - Collect and aggregate information about a given entity from multiple documents and even multiple data collections
  - ~ entity linking
- More structure than document-based IR
  - Types (from some taxonomy)
  - Attributes (from some ontology)
  - Relationships to other entities ("typed links")

In this Part

- Focus on the ad-hoc entity retrieval task
- Mainly probabilistic models
  - Specifically, Language Models

Basics

- Probability of an event

  $P(A)$

- Conditional probability

  $P(A|B)$

- Joint probability

  $P(A, B)$

Conditional dependence

- Independent events

$$P(A, B) = P(A) \cdot P(B)$$

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

- Conditionally dependent events

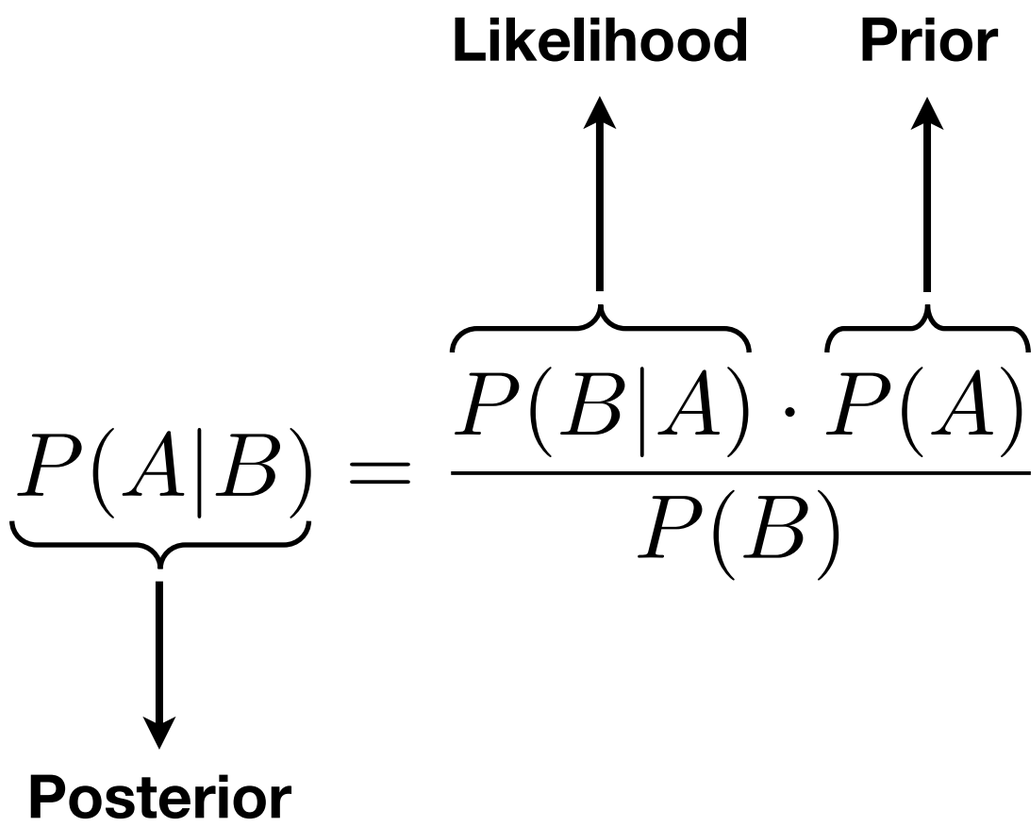
$$P(A, B) = P(A|B) \cdot P(B)$$

$$P(A, B|C) = P(A|B, C) \cdot P(B|C)$$

Bayes' theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayes' theorem



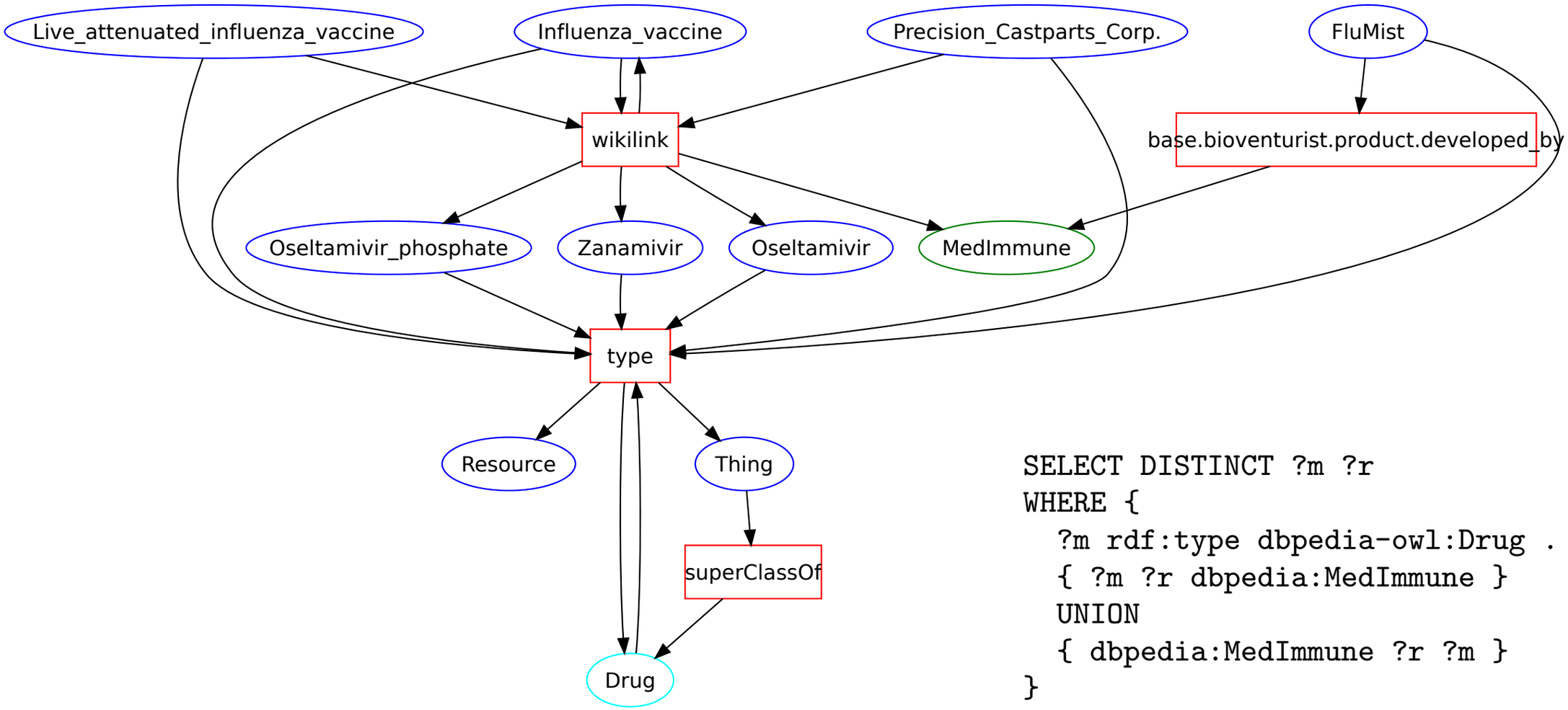
Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

Ad-hoc entity retrieval

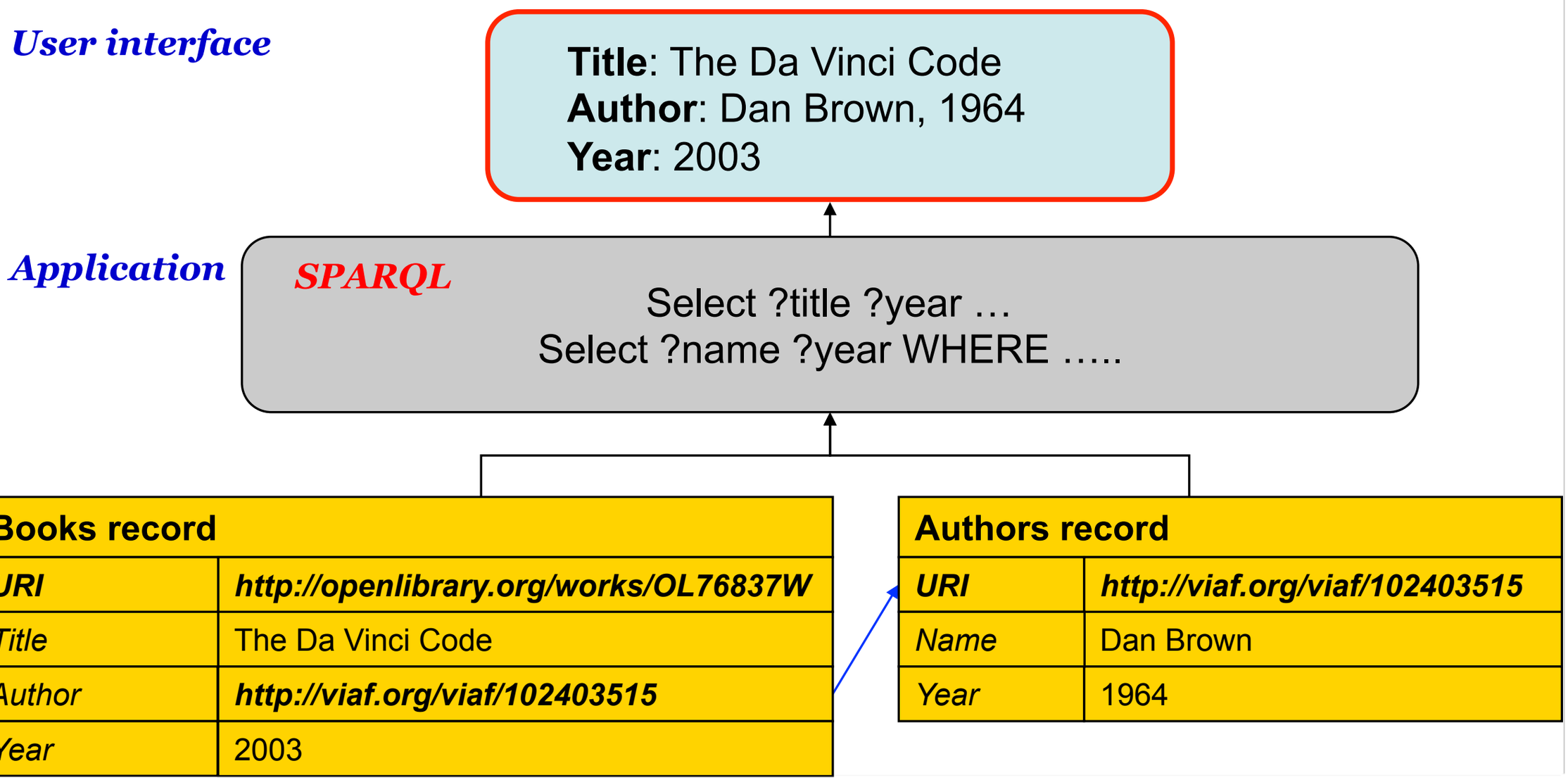
- **Input**: unconstrained natural language query
  - "telegraphic" queries (neither well-formed nor grammatically correct sentences or questions)
- **Output**: ranked list of entities
- **Collection**: unstructured and/or semi-structured documents

This is not...




```
SELECT DISTINCT ?m ?r
WHERE {
  ?m rdf:type dbpedia-owl:Drug .
  { ?m ?r dbpedia:MedImmune }
  UNION
  { dbpedia:MedImmune ?r ?m }
}
```

This is not...


User interface
Title: The Da Vinci Code
Author: Dan Brown, 1964
Year: 2003
Application
SPARQL
Select ?title ?year …
Select ?name ?year WHERE …..
Books record
URI
http://openlibrary.org/works/OL76837W
Title
The Da Vinci Code
Author
http://viaf.org/viaf/102403515
Year
2003
Authors record
URI
http://viaf.org/viaf/102403515
Name
Dan Brown
Year
1964

This is not...

User interface

Title: The Da Vinci Code
Author: Dan Brown, 1964
Year: 2003

Application

SQL

Select title, year from books
Select name, year from authors where books.author=authors.id

Database

Books record	
ID	***1289***
Title	The Da Vinci Code
Author	***456***
Year	2003

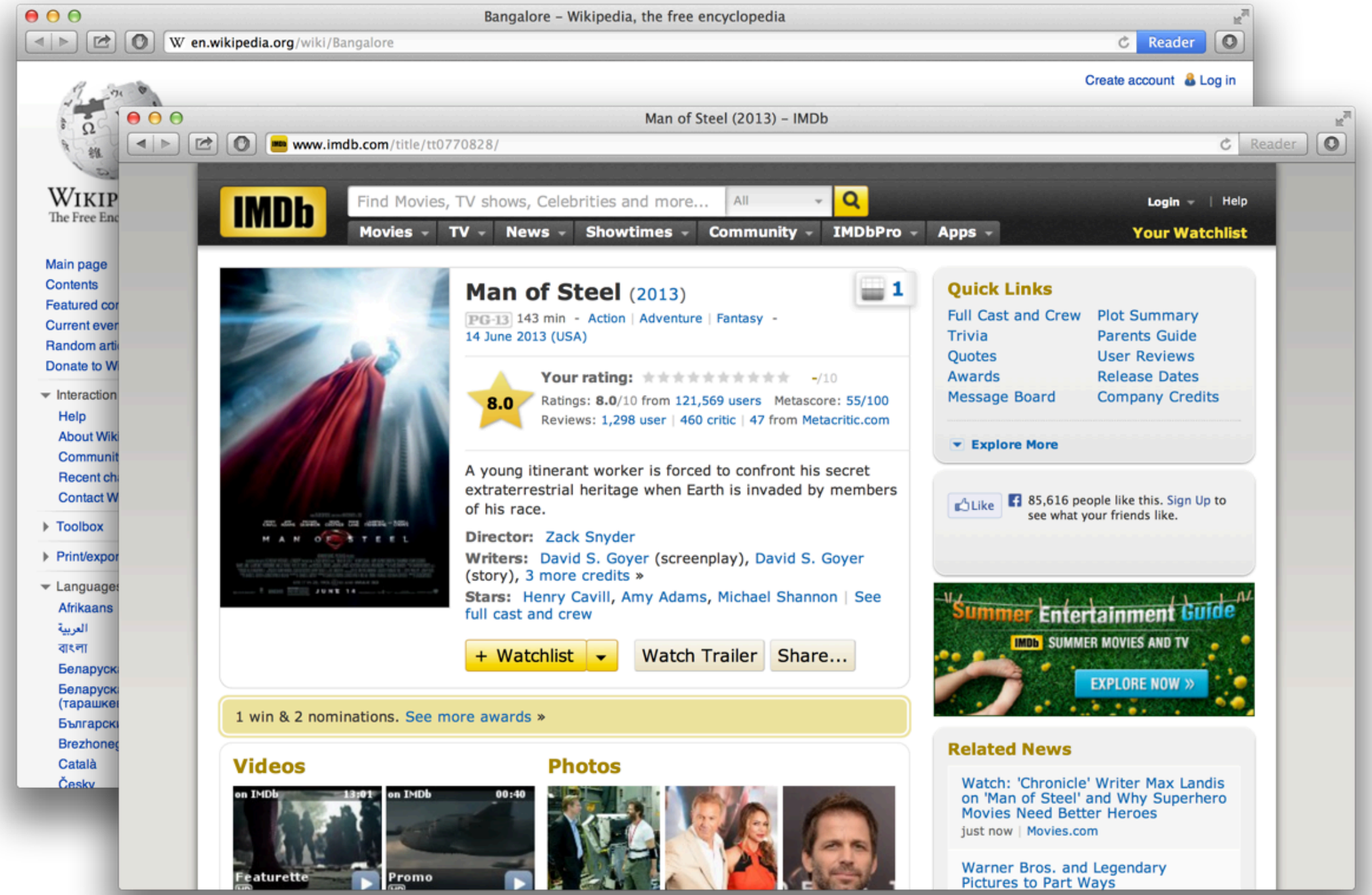
Authors record	
ID	***456***
Name	Dan Brown
Year	1964

Ranking with ready-made entity descriptions

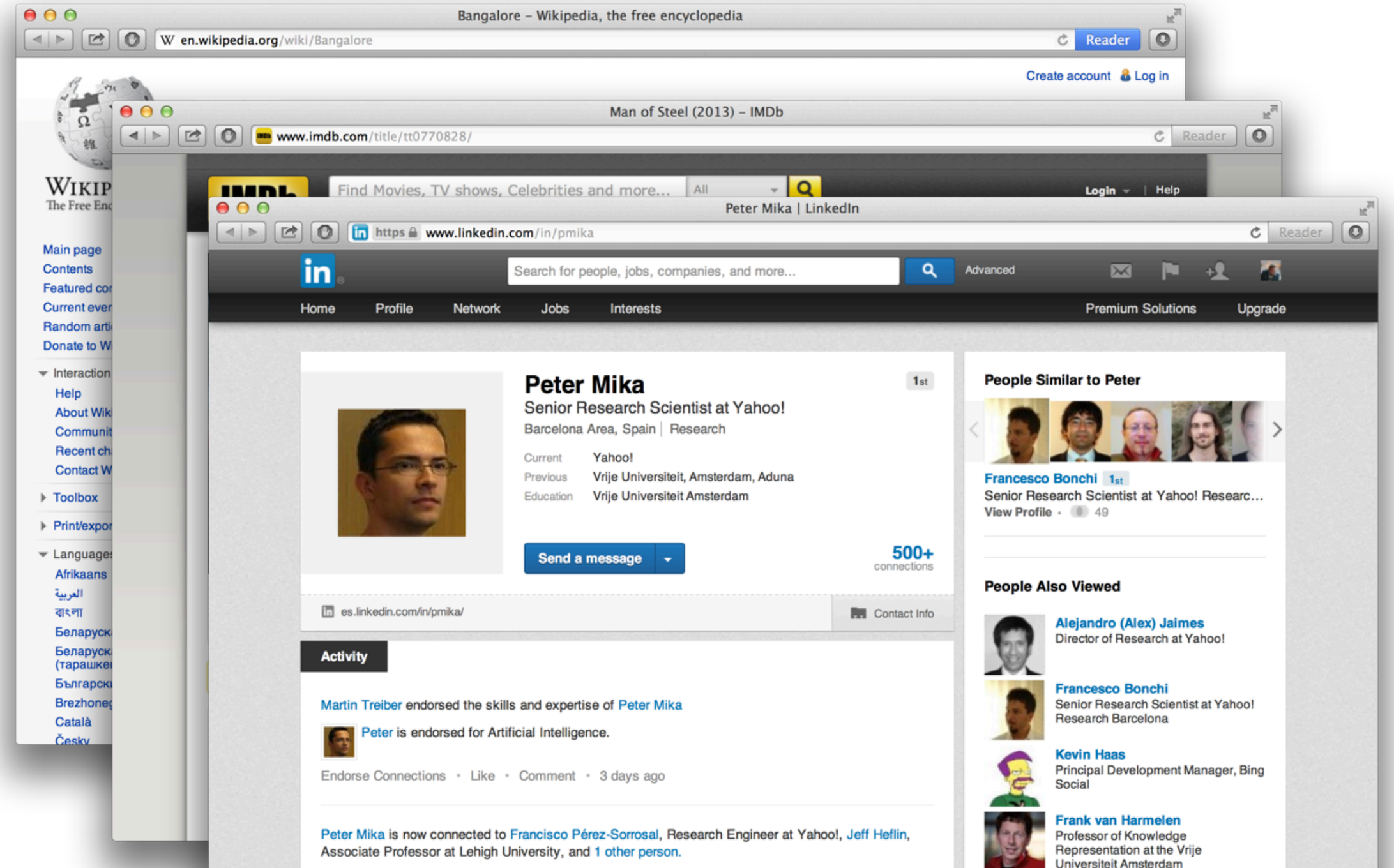
This is not unrealistic...

This is not unrealistic...

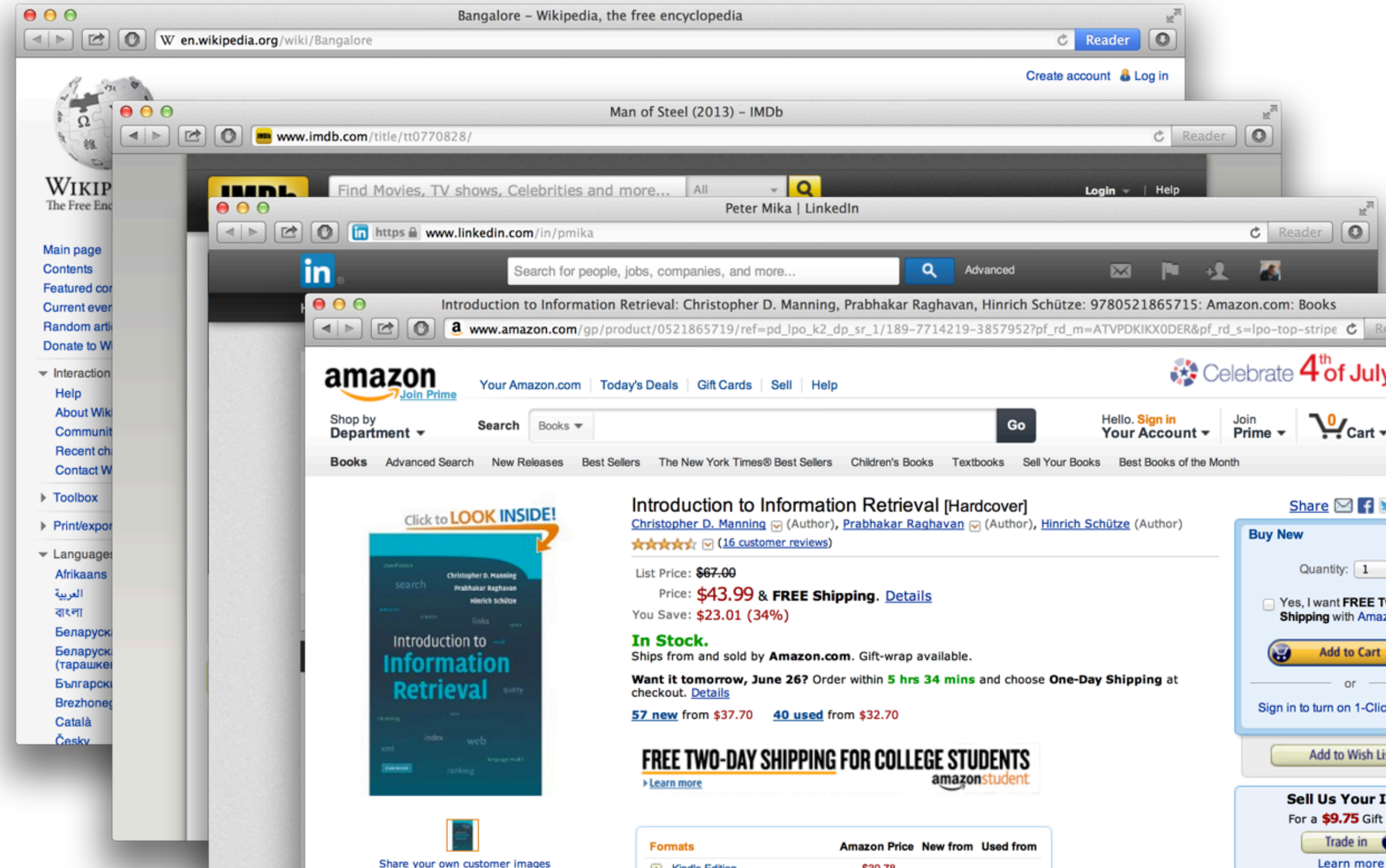
This is not unrealistic...



This is not unrealistic...



This is not unrealistic...



Document-based entity representations

- Most entities have a “home page”
- I.e., each entity is described by a document
- In this scenario, ranking entities is much like ranking documents
  - unstructured
  - semi-structured

Standard Language Modeling approach

- Rank documents d according to their likelihood of being relevant given a query q: $P(d|q)$

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

Standard Language Modeling approach

- Rank documents d according to their likelihood of being relevant given a query q: $P(d|q)$

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

Query likelihood
Probability that query q was "produced" by document d

Document prior
Probability of the document being relevant to *any* query

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

Standard Language Modeling approach (2)

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

Standard Language Modeling approach (2)

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

Number of times *t* appears in *q*

Document language model
Multinomial probability distribution over the vocabulary of terms

$$P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda P(t|C)$$

Standard Language Modeling approach (2)

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}$$

Number of times *t* appears in *q*

Document language model
Multinomial probability distribution over the vocabulary of terms

Smoothing parameter

$$P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda P(t|C)$$

Empirical document model

Maximum likelihood estimates

Collection model

$$\frac{n(t,d)}{|d|} \qquad \frac{\sum_d n(t,d)}{\sum_d |d|}$$

Here, documents==entities, so

$$P(e|q) \propto P(e)P(q|\theta_e) = P(e)\prod_{t \in q} P(t|\theta_e)^{n(t,q)}$$

Here, documents==entities, so

$$P(e|q) \propto P(e)P(q|\theta_e) = \underbrace{P(e)}_{\text{Entity prior}} \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Entity language model}}^{n(t,q)}$$

Entity prior
Probability of the entity being relevant to *any* query

Entity language model
Multinomial probability distribution over the vocabulary of terms

Semi-structured entity representation

- Entity description documents are rarely unstructured
- Representing entities as
  - Fielded documents – the IR approach
  - Graphs – the DB/SW approach

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
العربية
Беларуская
Български
Česky
Dansk
Deutsch

Audi A4

From Wikipedia, the free encyclopedia

The **Audi A4** is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group.

The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group.[2]

The Audi A4 automobile layout consists of a longitudinally oriented engine at the front, with transaxle-type transmissions mounted at the rear of the engine. The cars are front-wheel drive, or on some models, "quattro" all-wheel drive.

The A4 is available as a saloon/sedan and estate/wagon. The second (B6) and third generations (B7) of the A4 also had a convertible version, but the B8 version of the convertible became a variant of the Audi A5 instead as Audi got back into the compact executive coupé segment. The facebook fans of the Audi A4 page are more than 870,000.

Contents [show]

Audi A4

Manufacturer	Audi
Production	1994–present
Assembly	Ingolstadt, Germany Changchun, China[1] Tokyo, Japan (AMA; B5 only) Jakarta, Indonesia (Garuda Mataram Motor; B5 & B8) Solomonovo, Ukraine (Eurocar; B7 only) Aurangabad, India
Predecessor	Audi 80
Class	Compact executive car (globally)
Layout	front-engine, front-wheel-drive front-engine, four-wheel-drive
Platform	Volkswagen Group B

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
العربية
Беларуская
Български
Česky
Dansk
Deutsch

Audi A4

From Wikipedia, the free encyclopedia

The **Audi A4** is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group.

The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group.[2]

Audi A4

Manufacturer Audi

dbpedia:Audi_A4

foaf:name	Audi A4
rdfs:label	Audi A4
rdfs:comment	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
dbpprop:production	1994 2001 2005 2008
rdf:type	**dbpedia-owl:MeanOfTransportation** **dbpedia-owl:Automobile**
dbpedia-owl:manufacturer	**dbpedia:Audi**
dbpedia-owl:class	**dbpedia:Compact_executive_car**
owl:sameAs	**freebase:Audi A4**
is **dbpedia-owl:predecessor** of	**dbpedia:Audi_A5**
is **dbpprop:similar** of	**dbpedia:Cadillac_BLS**

Mixture of Language Models

[Ogilvie & Callan, 2003]

- Build a separate language model for each field
- Take a linear combination of them

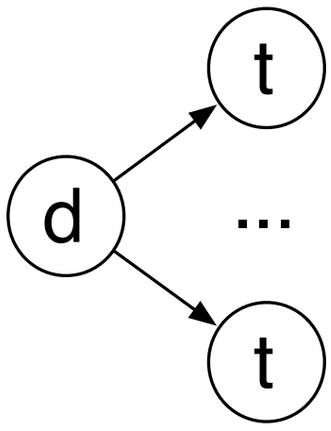
$$P(t|\theta_d) = \sum_{j=1}^{m} \underbrace{\mu_j}_{\text{Field weights}} P(t|\underbrace{\theta_{d_j}}_{\text{Field language model}})$$

Field language model
Smoothed with a collection model built from all document representations of the same type in the collection

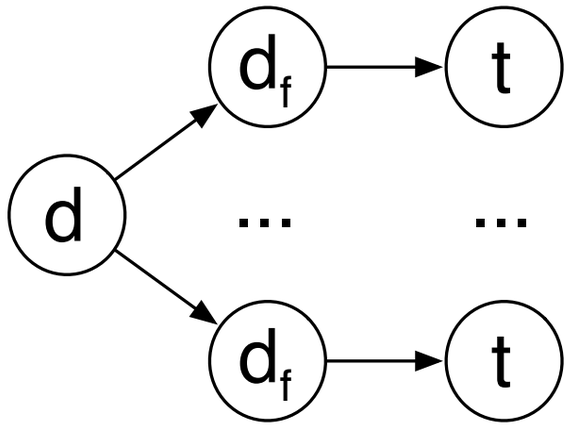
Field weights

$$\sum_{j=1}^{m} \mu_j = 1$$

Comparison of models



Unstructured document model



Fielded document model

Setting field weights

- Heuristically
  - Proportional to the length of text content in that field, to the field's individual performance, etc.
- Empirically (using training queries)
- Problems
  - Number of possible fields is huge
    - It is not possible to optimise their weights directly
- Entities are sparse w.r.t. different fields
  - Most entities have only a handful of predicates

Predicate folding

- **Idea**: reduce the number of fields by grouping them together
- Grouping based on (BM25F and)
  - type **[Pérez-Agüera et al. 2010]**
  - manually determined importance **[Blanco et al. 2011]**

Hierarchical Entity Model

[Neumayer et al. 2012]

- Organize fields into a 2-level hierarchy
  - Field types (4) on the top level
  - Individual fields of that type on the bottom level
- Estimate field weights
  - Using training data for field types
  - Using heuristics for bottom-level types

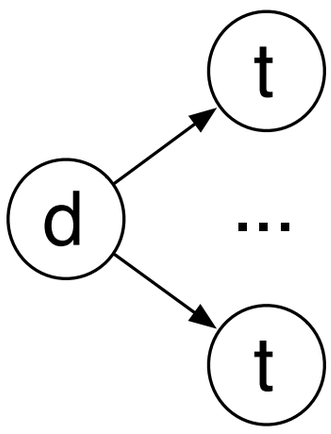
Two-level hierarchy

[Neumayer et al. 2012]

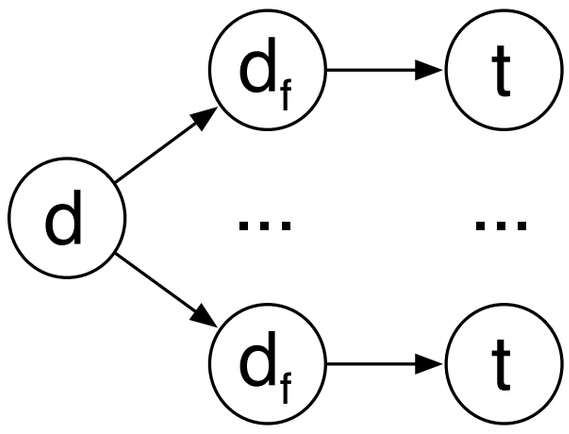
Name	**foaf:name**	Audi A4
	rdfs:label	Audi A4
Attributes	**rdfs:comment**	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
	dbpprop:production	1994 2001 2005 2008
Out-relations	**rdf:type**	**dbpedia-owl:MeanOfTransportation** **dbpedia-owl:Automobile**
	dbpedia-owl:manufacturer	**dbpedia:Audi**
	dbpedia-owl:class	**dbpedia:Compact_executive_car**
	owl:sameAs	**freebase:Audi A4**
In-relations	is **dbpedia-owl:predecessor** of	**dbpedia:Audi_A5**
	is **dbpprop:similar** of	**dbpedia:Cadillac_BLS**

Comparison of models

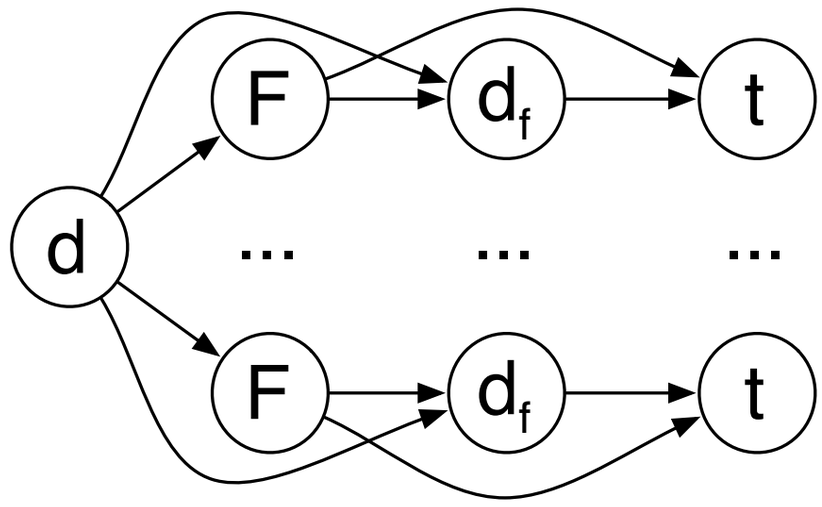
[Neumayer et al. 2012]



Unstructured document model



Fielded document model



Hierarchical document model

Probabilistic Retrieval Model for Semistructured data

[Kim et al. 2009]

- Extension to the Mixture of Language Models
- Find which document field each query term may be associated with

$$P(t|\theta_d) = \sum_{j=1}^{m} \mu_j P(t|\theta_{d_j})$$

Probabilistic Retrieval Model for Semistructured data

[Kim et al. 2009]

- Extension to the Mixture of Language Models
- Find which document field each query term may be associated with

$$P(t|\theta_d) = \sum_{j=1}^{m} \mu_j P(t|\theta_{d_j})$$

Mapping probability
Estimated for each query term

$$P(t|\theta_d) = \sum_{j=1}^{m} \overbrace{P(d_j|t)} P(t|\theta_{d_j})$$

Estimating the mapping probability

$$P(d_j|t) = \frac{P(t|d_j)P(d_j)}{P(t)}$$

Estimating the mapping probability

$$P(t|C_j) = \frac{\sum_d n(t, d_j)}{\sum_d |d_j|}$$

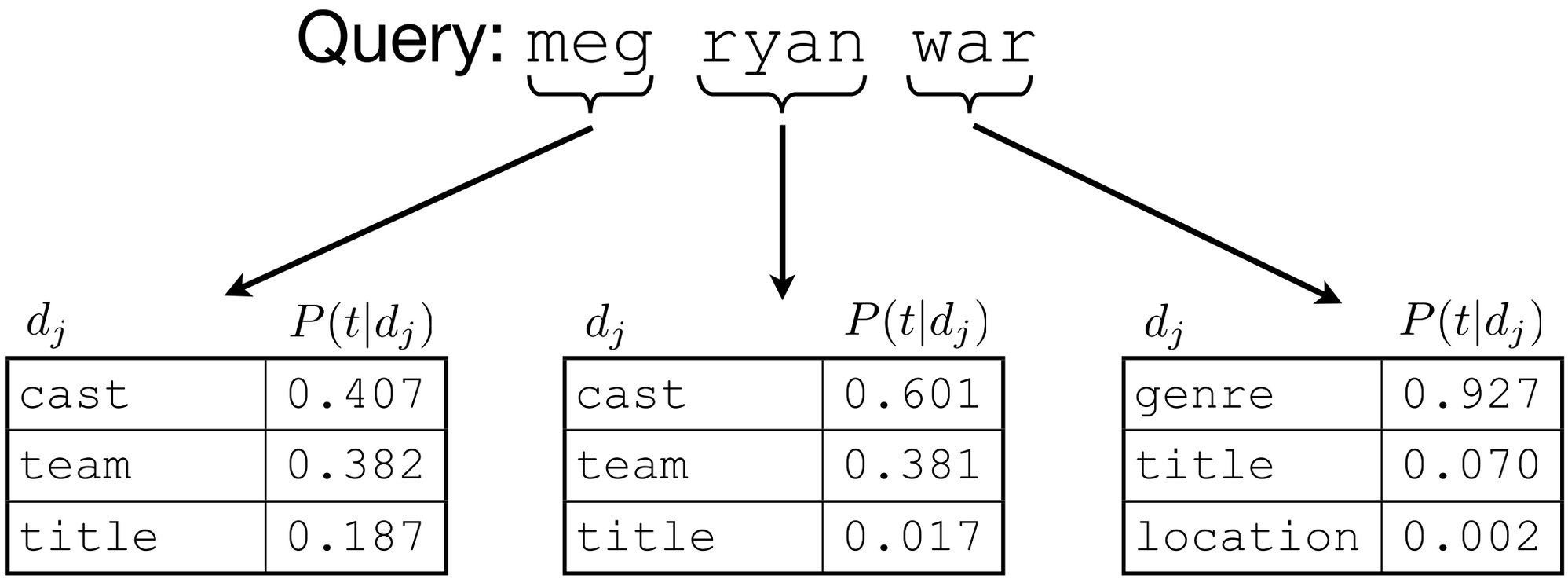
Term likelihood
Probability of a query term occurring in a given field type

Prior field probability
Probability of mapping the query term to this field before observing collection statistics

$$P(d_j|t) = \frac{\overbrace{P(t|d_j)}\overbrace{P(d_j)}}{\underbrace{P(t)}}$$

$$\sum_{d_k} P(t|d_k)P(d_k)$$

Example


Query: meg ryan war
d_j $P(t|d_j)$
cast 0.407
team 0.382
title 0.187
d_j $P(t|d_j)$
cast 0.601
team 0.381
title 0.017
d_j $P(t|d_j)$
genre 0.927
title 0.070
location 0.002

The usual suspects from document retrieval...

- Priors
  - HITS, PageRank
  - Document link indegree **[Kamps & Koolen 2008]**
- Pseudo relevance feedback
  - Document-centric vs. entity-centric **[Macdonald & Ounis 2007; Serdyukov et al. 2007]**
    - sampling expansion terms from top ranked documents and/or (profiles of) top ranked candidates
  - Field-based **[Kim & Croft 2011]**

So far...

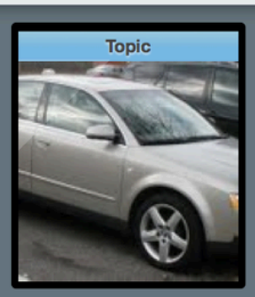
- Ranking (fielded) documents...
- What is special about entities?
  - Type(s)
  - Relationships with other entities

Entity types

`rdf:type`	`dbpedia-owl:MeanOfTransportation` `dbpedia-owl:Automobile`

Categories: Audi vehicles | Compact executive cars | Euro NCAP large family cars | Sedans | Station wagons | Convertibles | Vehicles with CVT transmission | All-wheel-drive vehicles | Front-wheel-drive vehicles | Vehicles introduced in 1994 | 1990s automobiles | 2000s automobiles | 2010s automobiles | Hybrid electric cars

Freebase | Find... | Browse | Query | Help | Sign In or Sign Up | English



Audi A4 en

Created by metaweb on 10/22/2006

id: /guid/9202a8c04000641f8000000000305a7c mid: /m/030qmx notable type: /automotive/model notable for: /automotive/model on the web: wikipedia.org

The Audi A4 is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group. The Audi A4 automobile layout consists of a longitudinally oriented engine at the front, with transaxle-type transmissions mounted at the rear of the engine. The cars are front-wheel drive, or on some models, "quattro" all-wheel drive. The A4 is available as a saloon/sedan and estate/wagon. The second and third generations of the A4 also had a convertible version, but the B8 version of the convertible became a variant of the Audi A5 instead as Audi got back into the compact executive coupé segment. Wikipedia [-]

Properties | I18n | Keys | Links

View and edit specific domains, types, or properties

Filter options: Show all domains and properties

Common /common — Freebase Commons

• Topic /common/topic x

Also known as /common/topic/alias

Types:

Common
Topic

Automotive
Automobile Model

Using target types

Assuming they have been identified...

- Constraining results
  - Soft/hard filtering
  - Different ways to measure type similarity (between target types and the types associated with the entity)
    - Set-based
    - Content-based
    - Lexical similarity of type labels
- Query expansion
  - Adding terms from type names to the query
- Entity expansion
  - Categories as a separate metadata field

Modeling terms and categories

[Balog et al. 2011]

$$P(e|q) \propto P(q|e)P(e)$$

$$P(q|e) = (1-\lambda)\underbrace{P(\theta_q^T|\theta_e^T)} + \lambda\underbrace{P(\theta_q^C|\theta_e^C)}$$

Modeling terms and categories

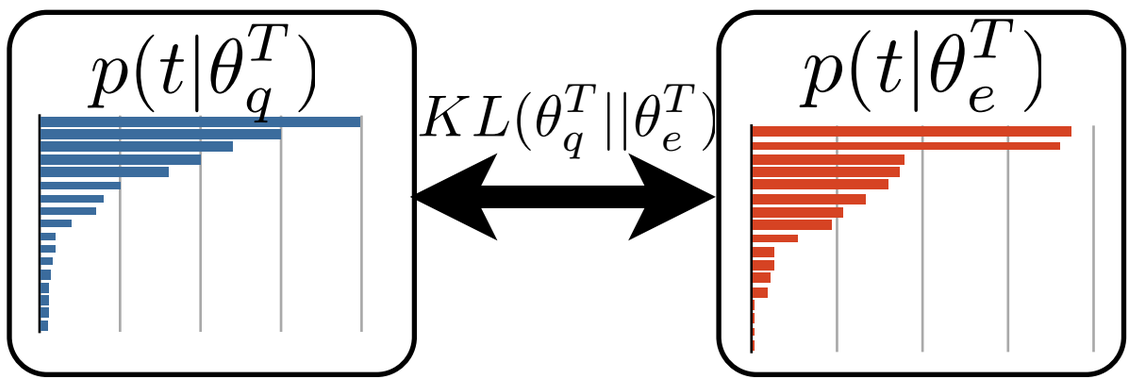
[Balog et al. 2011]

$$P(e|q) \propto P(q|e)P(e)$$

$$P(q|e) = (1-\lambda)\underbrace{P(\theta_q^T|\theta_e^T)} + \lambda\underbrace{P(\theta_q^C|\theta_e^C)}$$

Term-based representation

Query model | Entity model



Modeling terms and categories

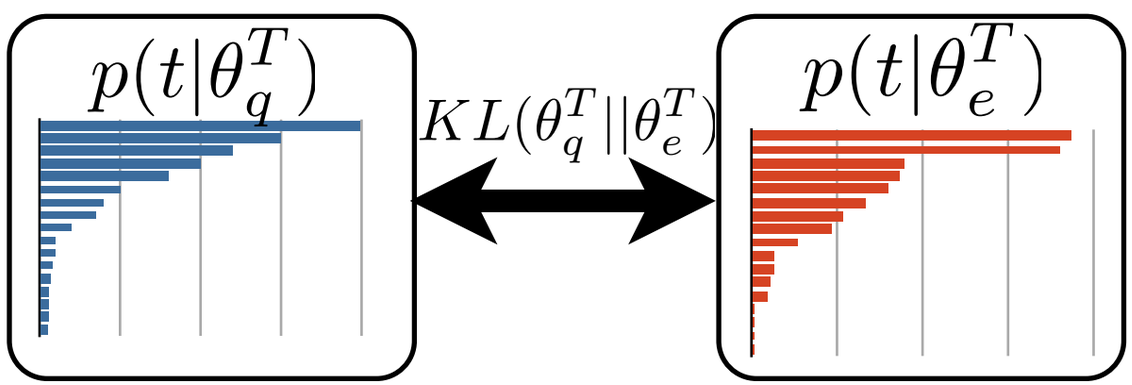
[Balog et al. 2011]

$$P(e|q) \propto P(q|e)P(e)$$

$$P(q|e) = (1-\lambda)\underbrace{P(\theta_q^T|\theta_e^T)} + \lambda\underbrace{P(\theta_q^C|\theta_e^C)}$$

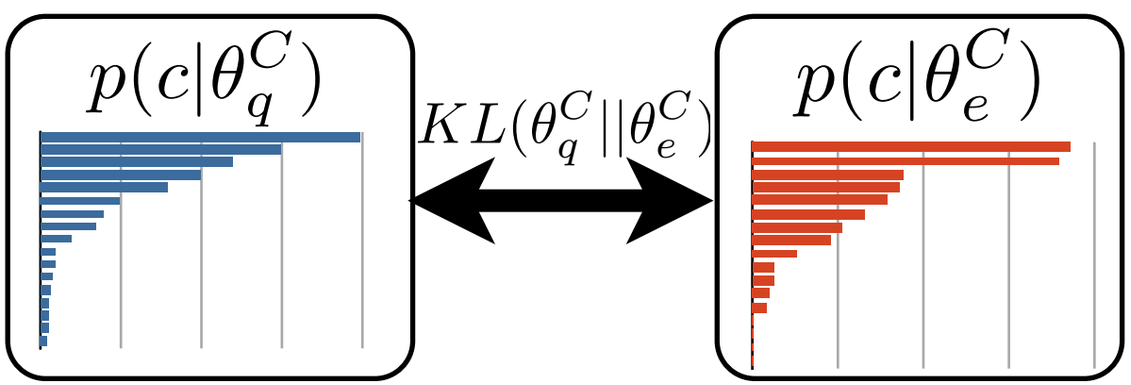
Term-based representation

Query model | Entity model



Category-based representation

Query model | Entity model



Identifying target types for queries

- Types of top ranked entities **[Vallet & Zaragoza 2008]**
- Direct term-based vs. indirect entity-based representations **[Balog & Neumayer 2012]**
- Hierarchical case is difficult... **[Sawant & Chakrabarti 2013]**

Expanding target types

- Pseudo relevance feedback
- Based on hierarchical structure
- Using lexical similarity of type labels

Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

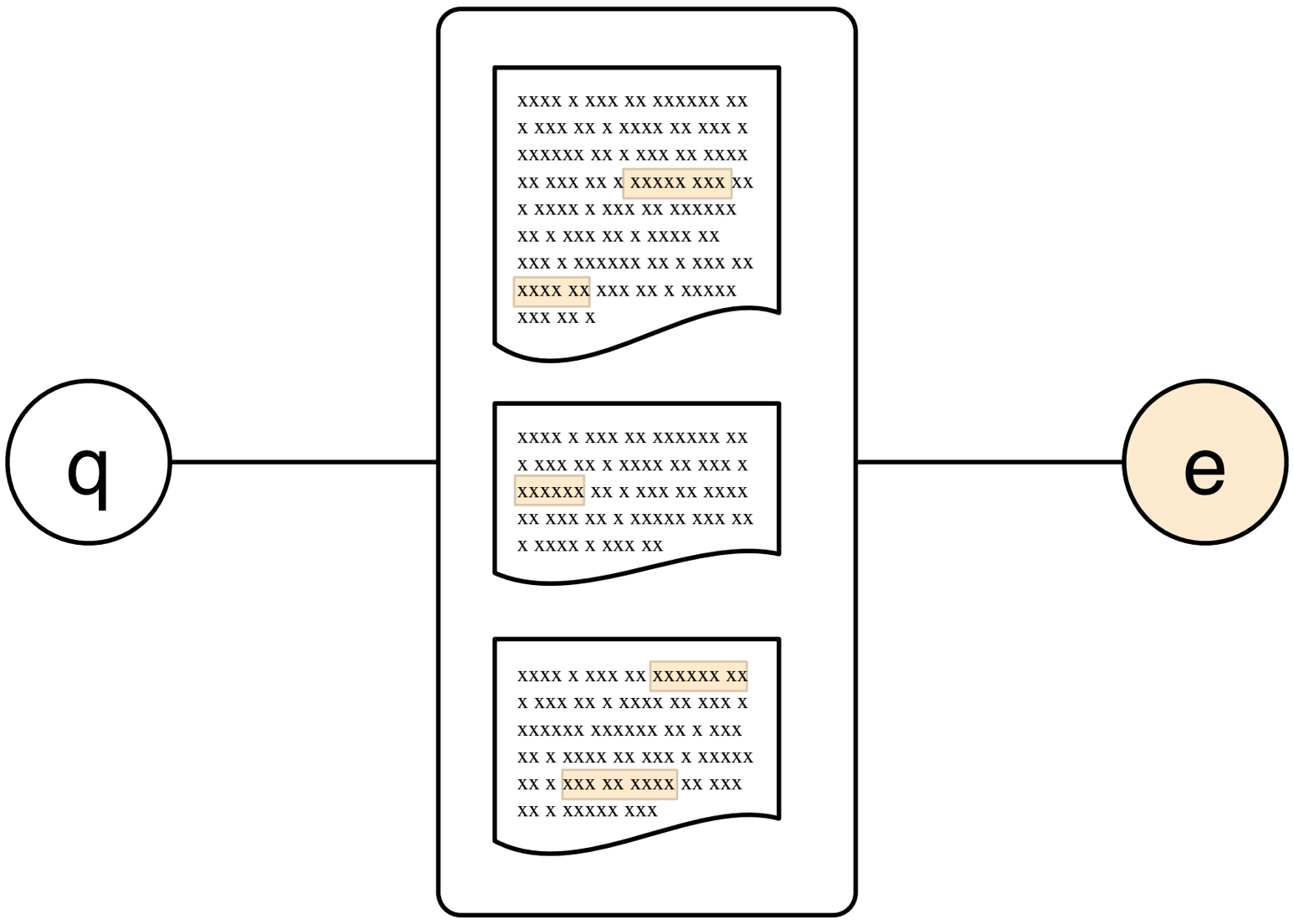
Ranking without explicit entity representations

Scenario

- Entity descriptions are not readily available
- Entity occurrences are annotated
  - manually
  - automatically (~entity linking)

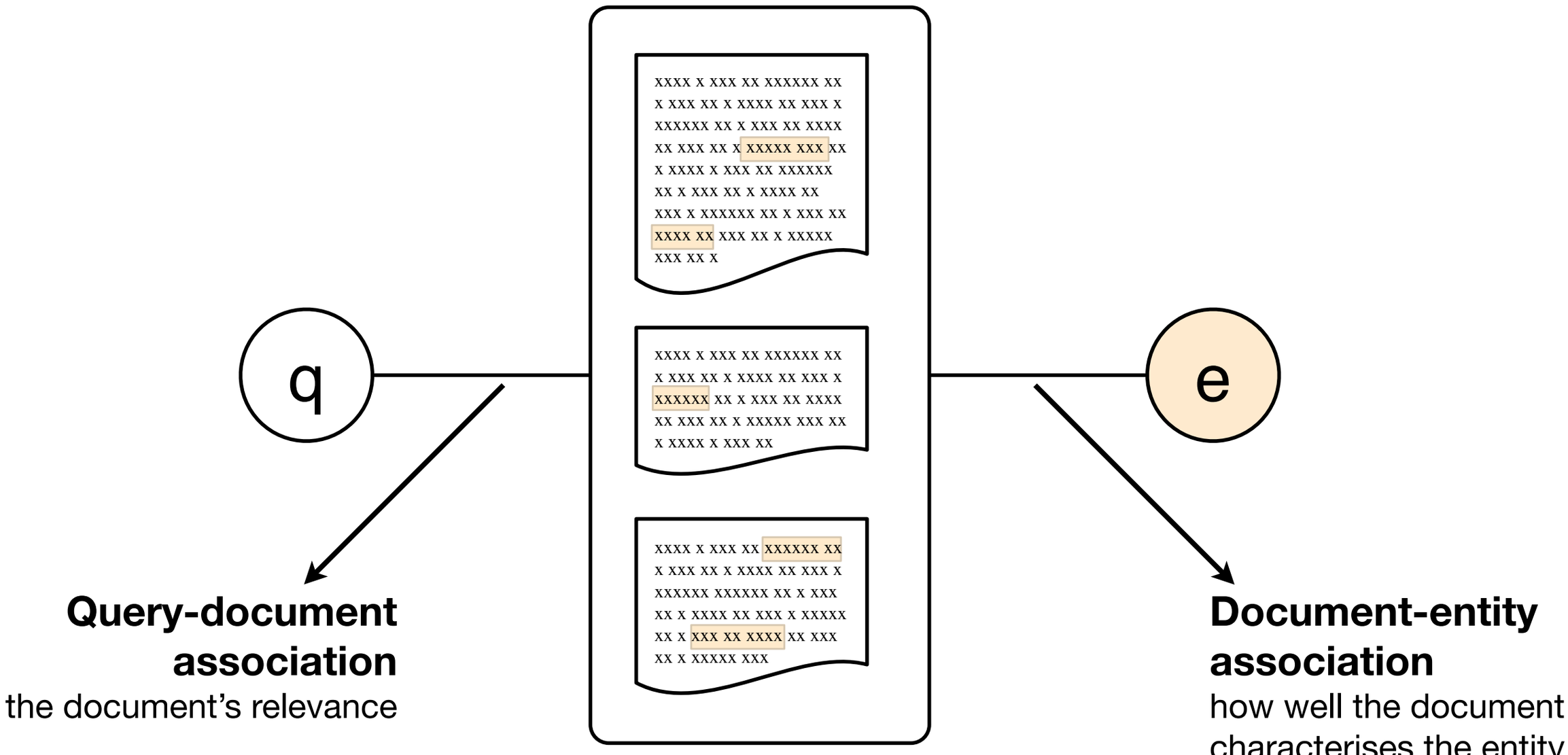
The basic idea

Use documents to go from queries to entities



The basic idea

Use documents to go from queries to entities



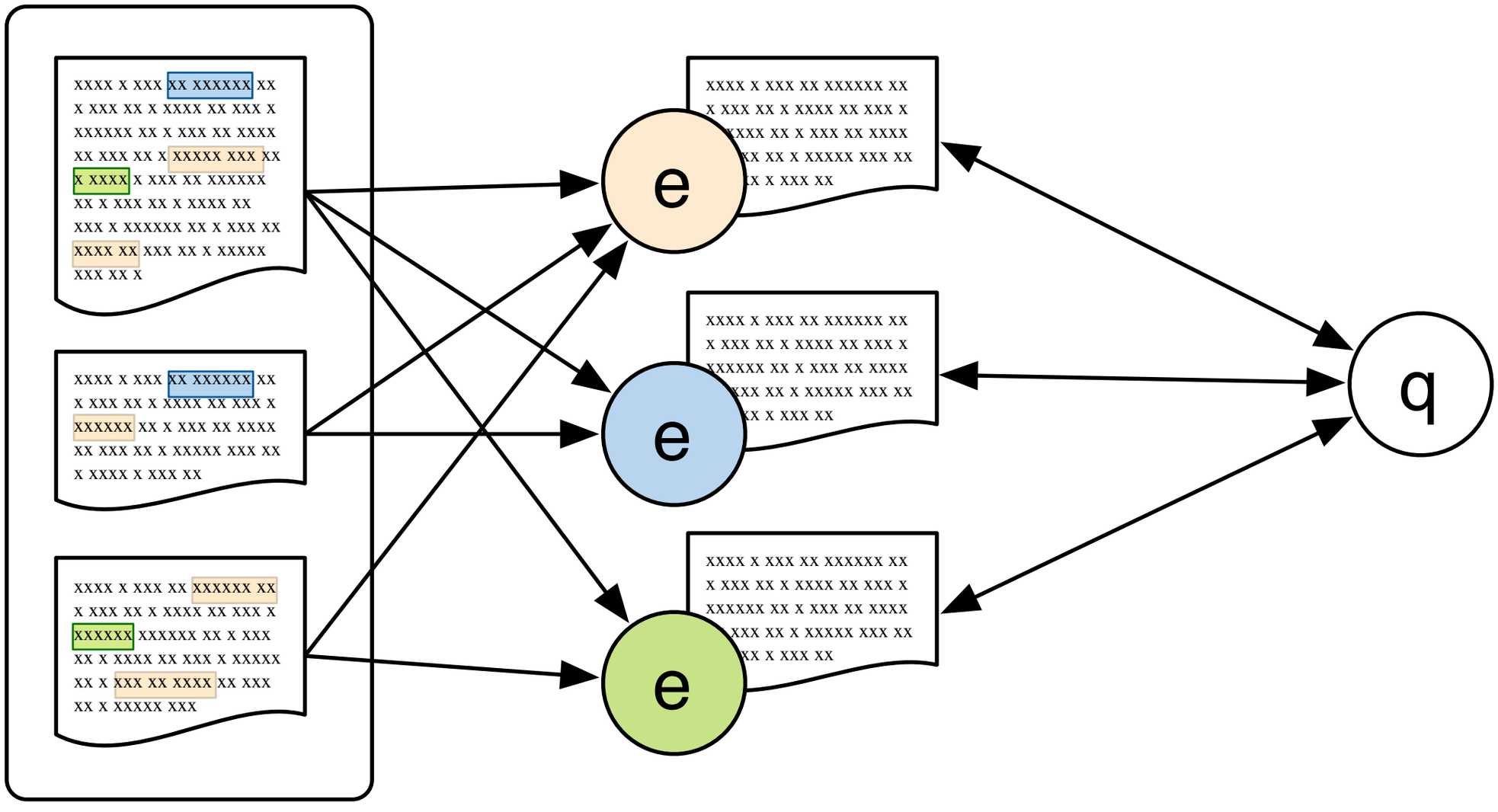
Query-document association
the document's relevance

Document-entity association
how well the document characterises the entity

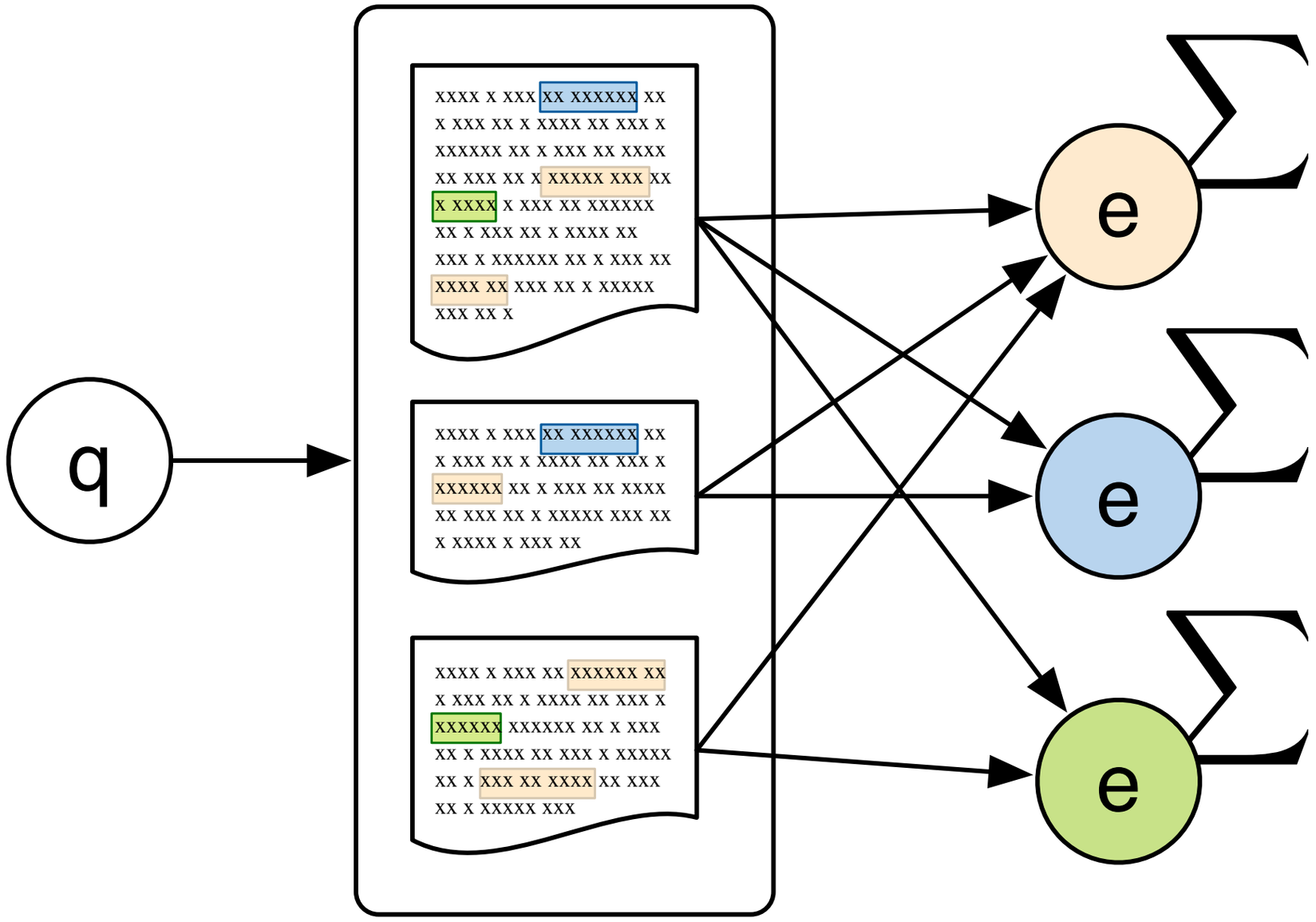
Two principal approaches

- **Profile-based** methods
  - Create a textual profile for entities, then rank them (by adapting document retrieval techniques)
- **Document-based** methods
  - Indirect representation based on mentions identified in documents
  - First ranking documents (or snippets) and then aggregating evidence for associated entities

Profile-based methods



Document-based methods



Many possibilities in terms of modeling

- Generative (probabilistic) models
- Discriminative (probabilistic) models
- Voting models
- Graph-based models

Generative probabilistic models

- Candidate generation models (*P(e|q)*)
  - Two-stage language model
- Topic generation models (*P(q|e)*)
  - Candidate model, a.k.a. Model 1
  - Document model, a.k.a. Model 2
  - Proximity-based variations
- Both families of models can be derived from the Probability Ranking Principle **[Fang & Zhai 2007]**

Generative probabilistic models

- Candidate generation models (*P(e|q)*)
  - Two-stage language model
- Topic generation models (*P(q|e)*)
  - Candidate model, a.k.a. Model 1
  - Document model, a.k.a. Model 2
  - Proximity-based variations
- Both families of models can be derived from the Probability Ranking Principle **[Fang & Zhai 2007]**

Candidate models ("Model 1")

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} P(t|\theta_e)^{n(t,q)}$$

Candidate models ("Model 1")

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} \underbrace{P(t|\theta_e)}^{n(t,q)}$$

Smoothing
With collection-wide background model

$$(1-\lambda)P(t|e) + \lambda P(t)$$

Candidate models ("Model 1")

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} \underbrace{P(t|\theta_e)}^{n(t,q)}$$

Smoothing
With collection-wide background model

$$(1-\lambda)\underbrace{P(t|e)} + \lambda P(t)$$

$$\hookrightarrow \sum_d P(t|d,e)P(d|e)$$

Candidate models ("Model 1")

[Balog et al. 2006]

$$P(q|\theta_e) = \prod_{t \in q} \underbrace{P(t|\theta_e)}^{n(t,q)}$$

Smoothing
With collection-wide background model

$$(1-\lambda)\underbrace{P(t|e)} + \lambda P(t)$$

$$\sum_d \underbrace{P(t|d,e)}\underbrace{P(d|e)}$$

Term-candidate co-occurrence
In a particular document.
In the simplest case: $P(t|d)$

Document-entity association

Document models ("Model 2")

[Balog et al. 2006]

$$P(q|e) = \sum_d P(q|d,e)P(d|e)$$

Document models ("Model 2")

[Balog et al. 2006]

$$P(q|e) = \sum_d \underbrace{P(q|d,e)}\underbrace{P(d|e)}$$

Document relevance
How well document d supports the claim that e is relevant to q

Document-entity association

$$\prod_{t \in q} P(t|d,e)^{n(t,q)}$$

Document models ("Model 2")

[Balog et al. 2006]

$$P(q|e) = \sum_d \underbrace{P(q|d,e)}_{\text{Document relevance}} \underbrace{P(d|e)}_{\text{Document-entity association}}$$

Document relevance
How well document d supports the claim that e is relevant to q

Document-entity association

$$\prod_{t \in q} \underbrace{P(t|d,e)}_{P(t|\theta_d)}^{n(t,q)}$$

Simplifying assumption (t and e are conditionally independent given d)

Document-entity associations

- Boolean (or set-based) approach
- Weighted by the confidence in entity linking
- Consider other entities mentioned in the document

Proximity-based variations

- So far, conditional independence assumption between candidates and terms when computing the probability *P(t|d,e)*
- Relationship between terms and entities that in the same document is ignored
  - Entity is equally strongly associated with everything discussed in that document
- Let’s capture the dependence between entities and terms
  - Use their distance in the document

Using proximity kernels

[Petkova & Croft 2007]

$$P(t|d,e) = \frac{1}{Z}\sum_{i=1}^{N}\delta_d(i,t)k(t,e)$$

Using proximity kernels

[Petkova & Croft 2007]

$$P(t|d,e) = \frac{1}{Z} \sum_{i=1}^{N} \underbrace{\delta_d(i,t)}\underbrace{k(t,e)}$$

Normalizing contant

Indicator function
1 if the term at position i is t,
0 otherwise

Proximity-based kernel
- constant function
- triangle kernel
- Gaussian kernel
- step function

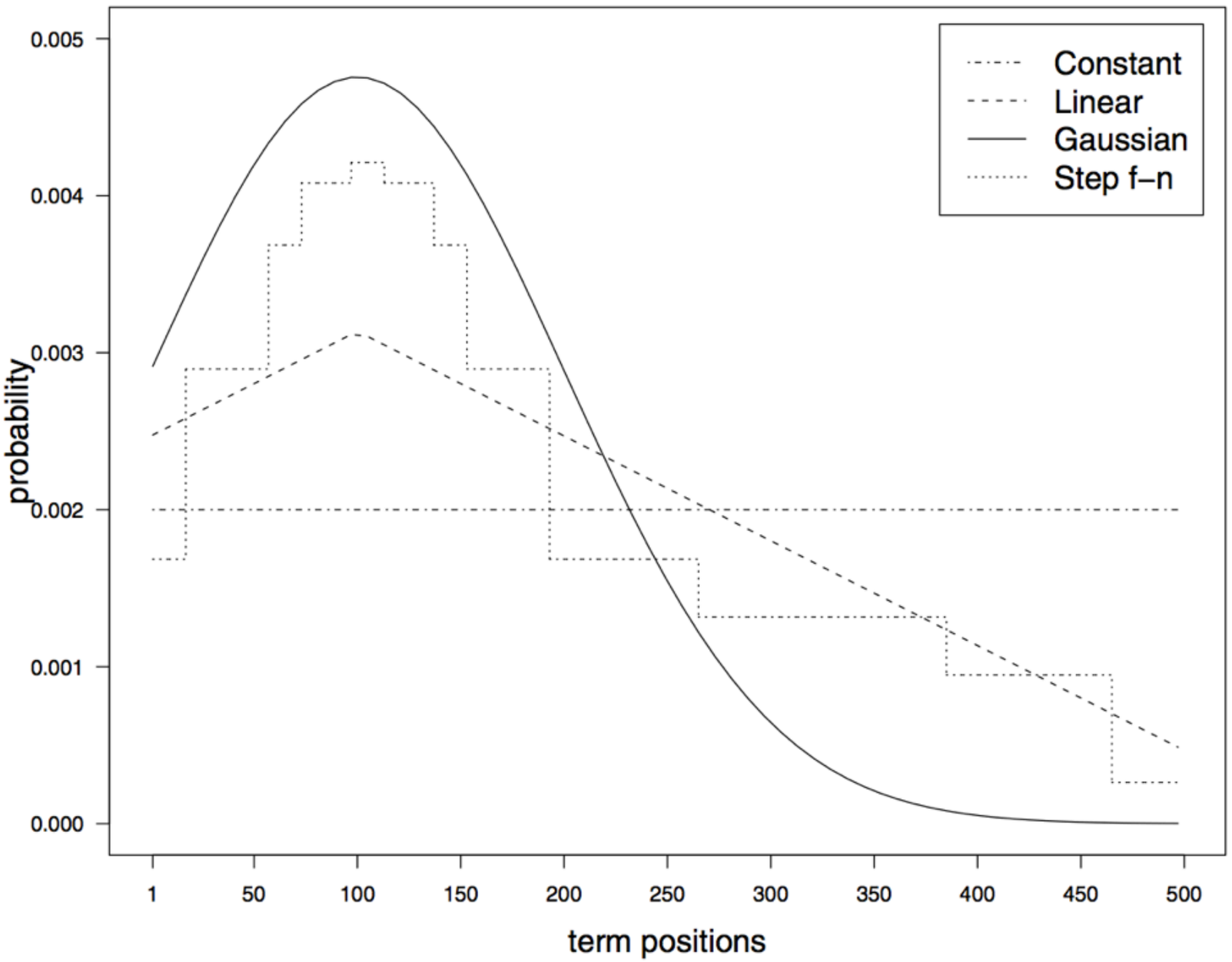


Figure taken from D. Petkova and W.B. Croft. **Proximity-based document representation for named entity retrieval**. *CIKM'07*.

Many possibilities in terms of modeling

- ~~Generative probabilistic models~~
- Discriminative probabilistic models
- Voting models
- Graph-based models

Discriminative models

- Vs. generative models:
  - Fewer assumptions (e.g., term independence)
  - “Let the data speak”
    - Sufficient amounts of training data required
- Incorporating more document features, multiple signals for document-entity associations
- Estimating *P(r=1|e,q)* directly (instead of *P(e,q|r=1)*)
- Optimization can get trapped in a local maximum/ minimum

Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]

$$P_\theta(r = 1|e, q) = \sum_d P(r_1 = 1|q, d)P(r_2 = 1|e, d)P(d)$$

Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]

$$P_\theta(r=1|e,q) = \sum_d \underbrace{P(r_1=1|q,d)}_{\text{Query-document relevance}} \underbrace{P(r_2=1|e,d)}_{\text{Document-entity relevance}} \underbrace{P(d)}_{\text{Document prior}}$$

Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]

$$P_\theta(r=1|e,q) = \sum_d \underbrace{P(r_1=1|q,d)}_{\text{Query-document relevance}} \underbrace{P(r_2=1|e,d)}_{\text{Document-entity relevance}} \underbrace{P(d)}_{\text{Document prior}}$$

logistic function over a linear combination of features

$$\sigma\Big(\sum_{i=1}^{N_f} \alpha_i f_i(q,d_t)\Big) \quad \sigma\Big(\sum_{j=1}^{N_g} \beta_j g_j(e,d_t)\Big)$$

Arithmetic Mean Discriminative (AMD) model

[Yang et al. 2010]

$$P_\theta(r=1|e,q) = \sum_d \underbrace{P(r_1=1|q,d)}_{\text{Query-document relevance}} \underbrace{P(r_2=1|e,d)}_{\text{Document-entity relevance}} \underbrace{P(d)}_{\text{Document prior}}$$

logistic function
over a linear
combination of features

$$\underbrace{\sigma}_{\text{standard logistic function}}\Big(\sum_{i=1}^{N_f} \underbrace{\alpha_i}_{\text{weight parameters (learned)}} \underbrace{f_i(q,d_t)}_{\text{features}}\Big) \qquad \sigma\Big(\sum_{j=1}^{N_g} \beta_j g_j(e,d_t)\Big)$$

Learning to rank && entity retrieval

- Pointwise
  - AMD, GMD **[Yang et al. 2010]**
  - Multilayer perceptrons, logistic regression **[Sorg & Cimiano 2011]**
  - Additive Groves **[Moreira et al. 2011]**
- Pairwise
  - Ranking SVM **[Yang et al. 2009]**
  - RankBoost, RankNet **[Moreira et al. 2011]**
- Listwise
  - AdaRank, Coordinate Ascent **[Moreira et al. 2011]**

Voting models

[Macdonald & Ounis 2006]

- Inspired by techniques from data fusion
  - Combining evidence from different sources
- Documents ranked w.r.t. the query are seen as “votes” for the entity

Voting models

Many different variants, including...

- Votes
  - Number of documents mentioning the entity

$$Score(e, q) = |M(e) \cap R(q)|$$

- Reciprocal Rank
  - Sum of inverse ranks of documents

$$Score(e, q) = \sum_{\{M(e) \cap R(q)\}} \frac{1}{rank(d, q)}$$

- CombSUM
  - Sum of scores of documents

$$Score(e, q) = |\{M(e) \cap R(q)\}| \sum_{\{M(e) \cap R(q)\}} s(d, q)$$

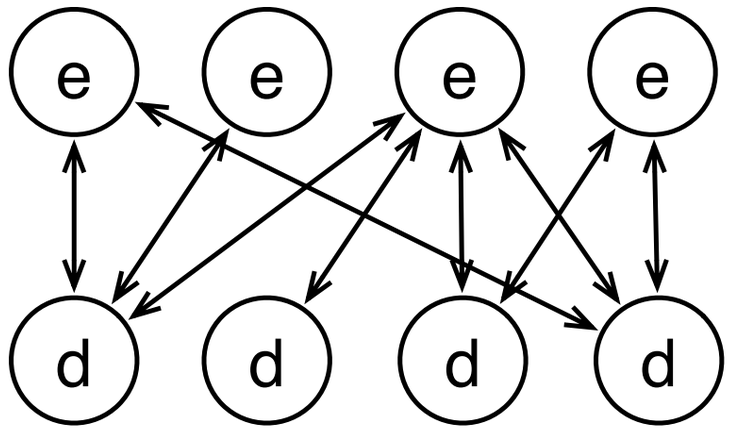
Graph-based models

[Serdyukov et al. 2008]

- One particular way of constructing graphs
  - Vertices are documents and entities
  - Only document-entity edges
- Search can be approached as a random walk on this graph
  - Pick a random document or entity
  - Follow links to entities or other documents
  - Repeat it a number of times

Infinite random walk

[Serdyukov et al. 2008]



$$
\begin{aligned}
P_i(d) &= \lambda P_J(d) + (1-\lambda) \sum_{e \to d} P(d|e) P_{i-1}(e), \\
P_i(e) &= \sum_{d \to e} P(e|d) P_{i-1}(d), \\
P_J(d) &= P(d|q),
\end{aligned}
$$

Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

Test collections

Test collections

Campaign	Task	Collection	Entity repr.	#Topics
TREC Enterprise (2005-08)	Expert finding	Enterprise intranets (W3C, CSIRO)	Indirect	99 (W3C) 127 (CSIRO)
TREC Entity (2009-11)	Rel. entity finding	Web crawl (ClueWeb09)	Indirect	120
	List completion			70
INEX Entity Ranking (2007-09)	Entity search	Wikipedia	Direct	55
	List completion			
SemSearch Chall. (2010-11)	Entity search	Semantic Web crawl (BTC2009)	Direct	142
	List search			50
INEX Linked Data (2012-13)	Ad-hoc search	Wikipedia + RDF (Wikipedia-LOD)	Direct	100 (’12) 144 (’13)

Test collections (2)

- Entity search as Question Answering
  - TREC QA track
  - QALD-2 challenge
  - INEX-LD Jeopardy task
- DBpedia entity search **[Balog & Neumayer 2013]**
  - synthesized queries and assessments, distilled from previous campaigns
  - from short keyword queries to natural language questions
  - 485 queries in total; mapped to DBpedia

Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

Hands-on

Public Toolkits and Web Services for Entity Retrieval

- YAGO
- Freebase
- DBpedia
- EARS
- Sig.ma
- Sindice & SIREn

YAGO

- Accuracy manually evaluated
  - confirmed accuracy of 95%
  - relation is annotated with its confidence value.
- Anchored in Time and Space
- Thematic domains (e.g. "music" or "science")
- Includes the WordNet class hierarchy
- See http://www.mpi-inf.mpg.de/yago-naga/yago/

Freebase

- Initially seeded from high-quality open data
- Now composed mainly by community
- Harvested from many sources
  - Wikipedia, MusicBrainz, and others.
- Acquired by Google in 2010
  - Google Knowledge Graph
- See http://www.freebase.com/

DBpedia

- Extract structured information from Wikipedia
- Crowd-sourced community effort
- Open source
  - written in Scala, Java and VSP
  - Virtuoso Universal Server Operating system
- See http://dbpedia.org/About

Sense of Scale

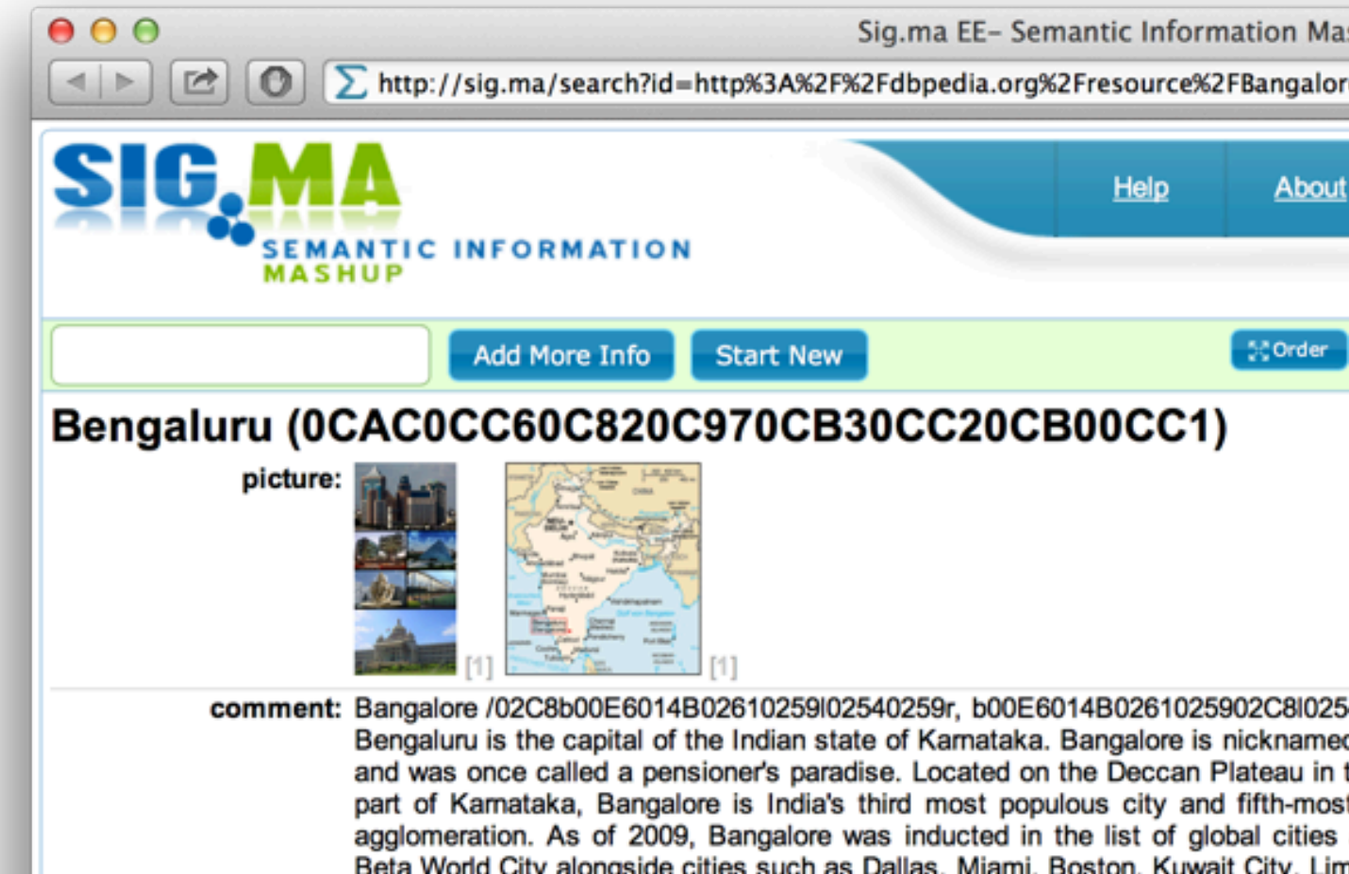
- YAGO: 10 million entities and 120 million facts
- Freebase: 37 million topics, 1,998 types, and more than 30,000 properties
- DBpedia: 3.77 million things
  - 2.35 million classified in Ontology, including:
    - 764,000 persons, 573,000 places,
    - 333,000 creative works, 192,000 organizations,
    - 202,000 species and 5,500 diseases.
  - 111 languages, together 20.8 million things

EARS

- Entity and Association Retrieval System
  - open source, built on top of Lemur in C++
    - not actively maintained anymore (but still works)
- Entity-topic association finding models
  - suited for other tasks, e.g. blog distillation
  - focuses on two entity-related tasks:
    - finding entities:
      - "Which entities are associated with topic X?"
    - profiling entities:
      - "What topics is an entity associated with?"
- See https://code.google.com/p/ears/

Sig.ma

- Search, aggregate, and visualize LOD data
- Powered by Sindice
- See http://sig.ma/



Sindice/SIREn

- Handling of semi-structured data
  - efficient, large scale
  - typically based on DBMS backends
  - uses Lucene for semi-structured search
- Open source
- Online demo, local install
- See http://siren.sindice.com/

Code Academy

- Contains some (Javascript) coding examples for entity linking and retrieval
  - http://www.codecademy.com/courses/javascript-beginner-en-LkhDf/

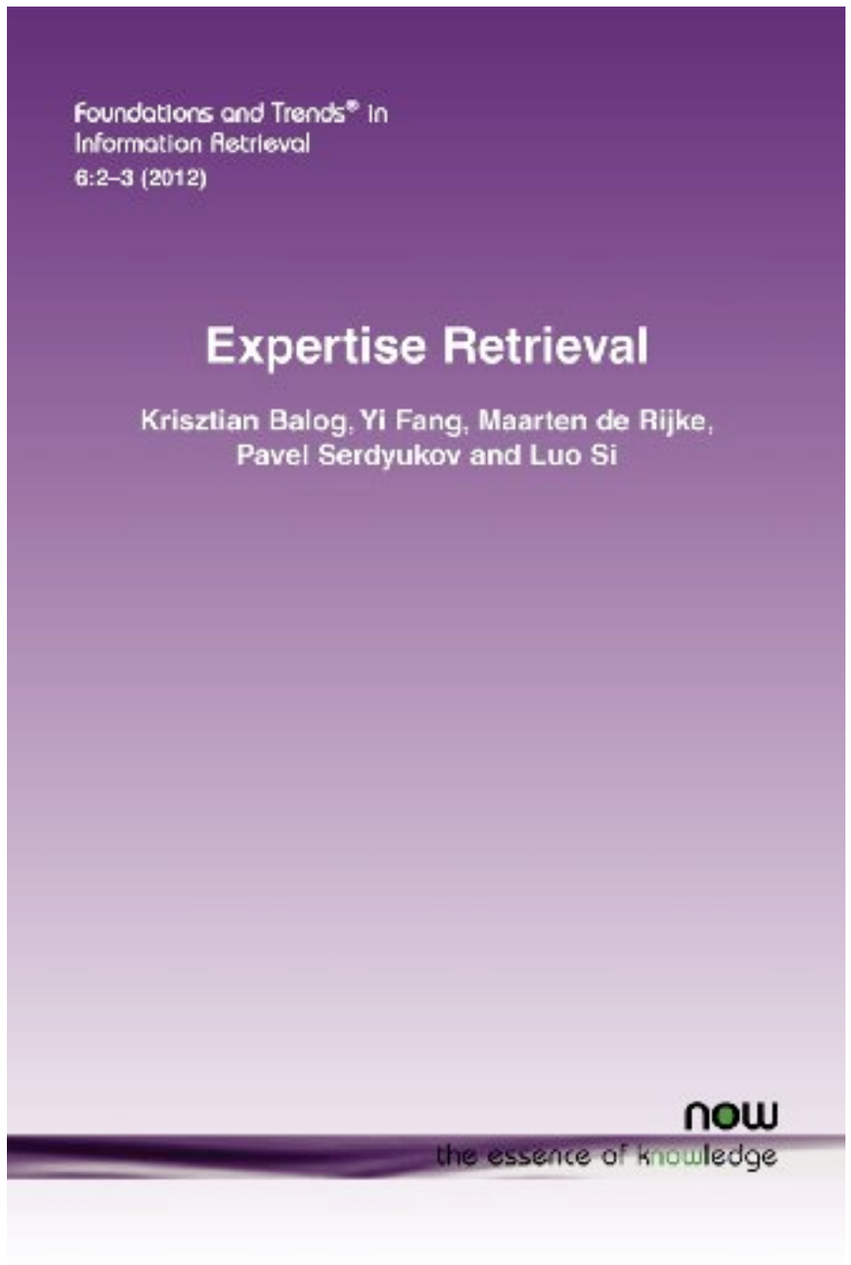
Outline

- Part 2 – Entity Retrieval
  - introduction
  - ranking with ready-made entity descriptions
  - ranking without explicit entity representations
  - test collections
  - hands-on
  - open challenges

Open challenges

- Combining text and structure
  - knowledge bases and unstructured Web documents
- Query understanding and modeling **[Sawant & Chakrabarti 2013]**
- UI/UX/Result presentation
  - how to interact with entities
- Hyperlocal
  - Siri/Google Now/...
  - recommendations

Follow-up reading

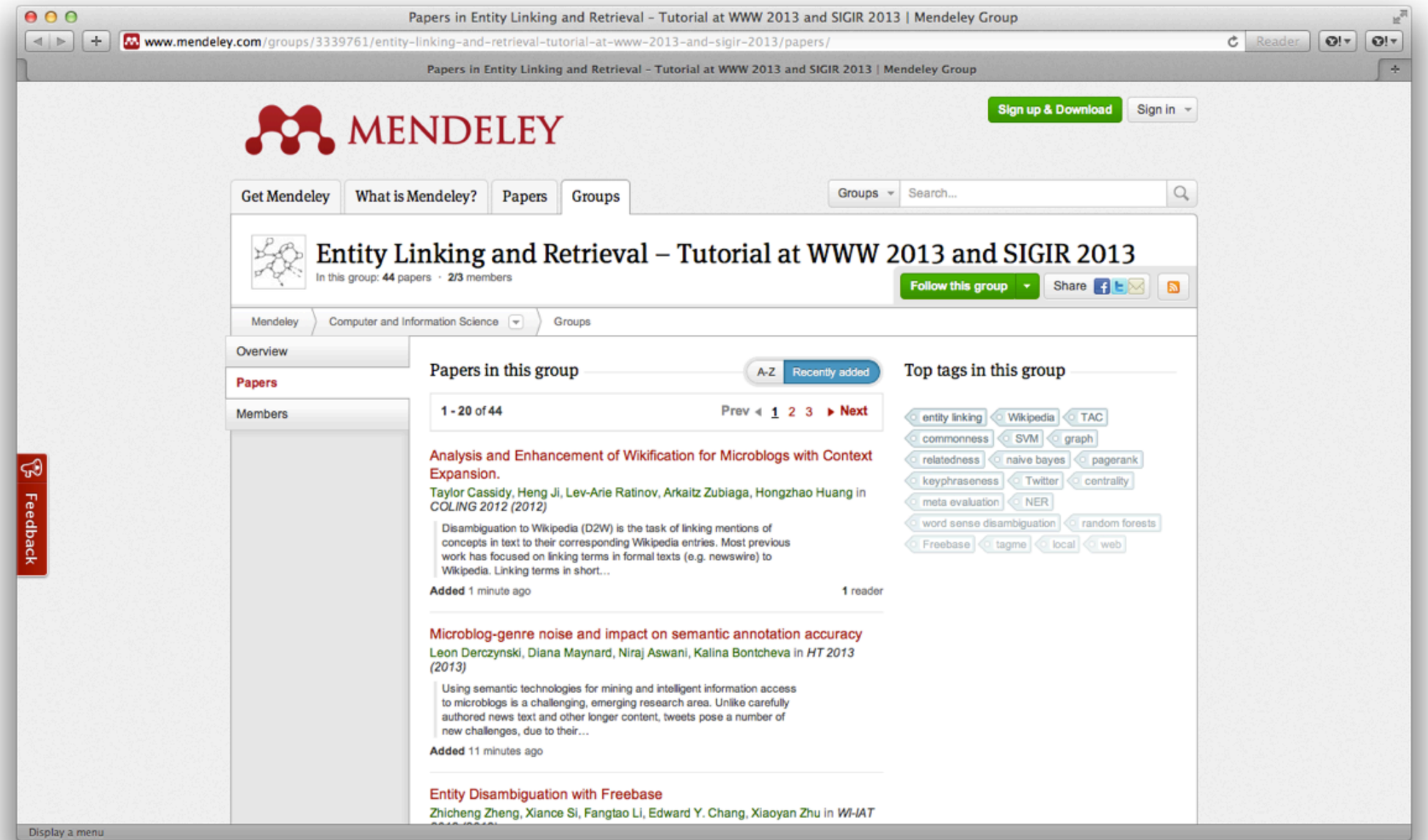


K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. **Expertise Retrieval**. *FnTIR'12*.

References – Entity retrieval

http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/added/0/tag/entity+retrieval/

References – Entity retrieval



http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/papers/added/0/tag/entity+retrieval/

Tutorial Resources

- Complete tutorial material

  http://bit.ly/yahoosummerschool

- Coding exercises

  see link above under “CodeAcademy Course”

- References

  http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval-tutorial-at-www-2013-and-sigir-2013/

- Contact
  - **emeij@yahoo-inc.com**
  - **@edgarmeij**
  - **http://edgar.meij.pro**