Gianluca Demartini

Ph.D. Defense

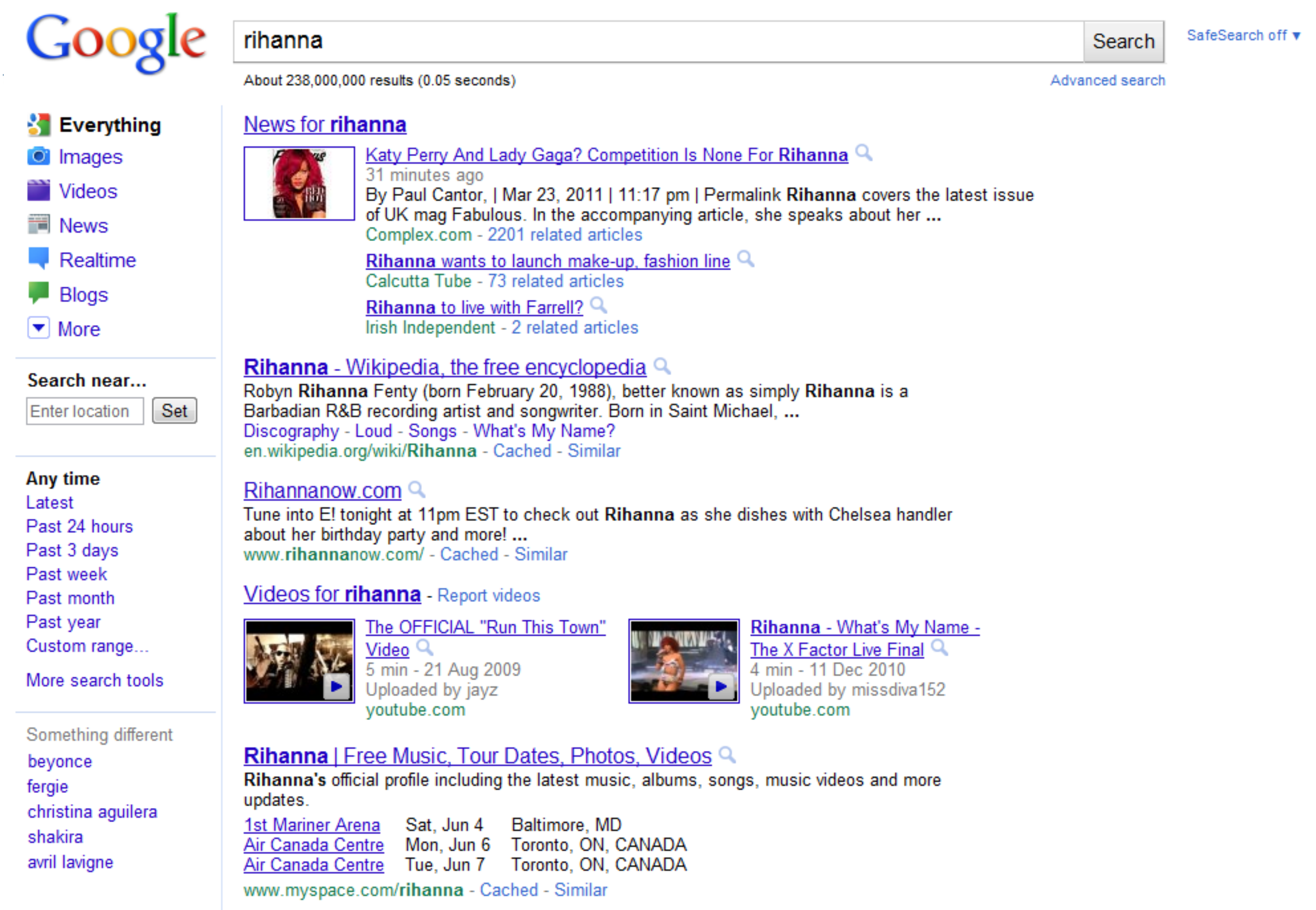
Hannover, April 6th, 2011

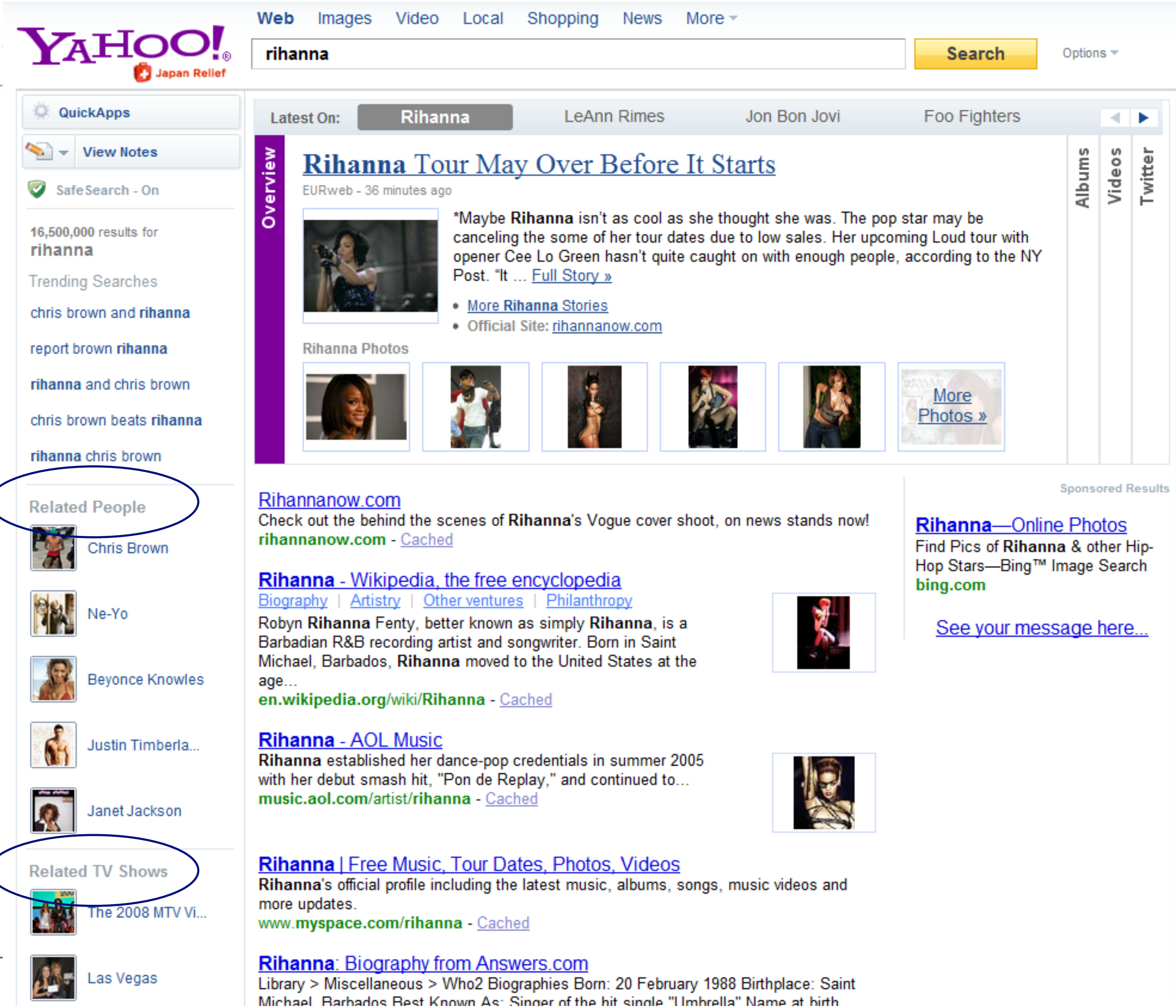
FROM PEOPLE TO ENTITIES: TYPED SEARCH IN THE ENTERPRISE AND THE WEB

Motivation

As the Web grows, Search Engines are the main entry points

- Search Engines are good for document ranking
- Provide more than just documents (e.g., pictures, videos, maps, etc.)


Google
rihanna
Search
SafeSearch off
About 238,000,000 results (0.05 seconds)
Advanced search
Everything
Images
Videos
News
Realtime
Blogs
More
Search near...
Enter location
Set
Any time
Latest
Past 24 hours
Past 3 days
Past week
Past month
Past year
Custom range...
More search tools
Something different
beyonce
fergie
christina aguilera
shakira
avril lavigne
News for rihanna
Katy Perry And Lady Gaga? Competition Is None For Rihanna
31 minutes ago
By Paul Cantor, | Mar 23, 2011 | 11:17 pm | Permalink Rihanna covers the latest issue of UK mag Fabulous. In the accompanying article, she speaks about her ...
Complex.com - 2201 related articles
Rihanna wants to launch make-up, fashion line
Calcutta Tube - 73 related articles
Rihanna to live with Farrell?
Irish Independent - 2 related articles
Rihanna - Wikipedia, the free encyclopedia
Robyn Rihanna Fenty (born February 20, 1988), better known as simply Rihanna is a Barbadian R&B recording artist and songwriter. Born in Saint Michael, ...
Discography - Loud - Songs - What's My Name?
en.wikipedia.org/wiki/Rihanna - Cached - Similar
Rihannanow.com
Tune into E! tonight at 11pm EST to check out Rihanna as she dishes with Chelsea handler about her birthday party and more! ...
www.rihannanow.com/ - Cached - Similar
Videos for rihanna - Report videos
The OFFICIAL "Run This Town" Video
5 min - 21 Aug 2009
Uploaded by jayz
youtube.com
Rihanna - What's My Name - The X Factor Live Final
4 min - 11 Dec 2010
Uploaded by missdiva152
youtube.com
Rihanna | Free Music, Tour Dates, Photos, Videos
Rihanna's official profile including the latest music, albums, songs, music videos and more updates.
1st Mariner Arena Sat, Jun 4 Baltimore, MD
Air Canada Centre Mon, Jun 6 Toronto, ON, CANADA
Air Canada Centre Tue, Jun 7 Toronto, ON, CANADA
www.myspace.com/rihanna - Cached - Similar

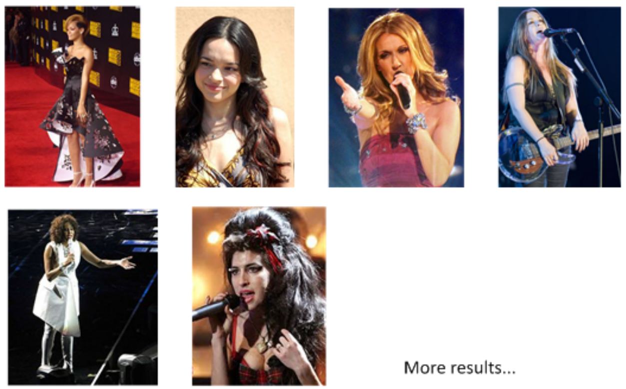

Web Images Video Local Shopping News More
YAHOO!
Japan Relief
rihanna
Search
Options
QuickApps
View Notes
SafeSearch - On
16,500,000 results for rihanna
Trending Searches
chris brown and rihanna
report brown rihanna
rihanna and chris brown
chris brown beats rihanna
rihanna chris brown
Related People
Chris Brown
Ne-Yo
Beyonce Knowles
Justin Timberla...
Janet Jackson
Related TV Shows
The 2008 MTV Vi...
Las Vegas
Latest On: Rihanna LeAnn Rimes Jon Bon Jovi Foo Fighters
Overview
Albums
Videos
Twitter
Rihanna Tour May Over Before It Starts
EURweb - 36 minutes ago
*Maybe Rihanna isn't as cool as she thought she was. The pop star may be canceling the some of her tour dates due to low sales. Her upcoming Loud tour with opener Cee Lo Green hasn't quite caught on with enough people, according to the NY Post. "It ... Full Story »
More Rihanna Stories
Official Site: rihannanow.com
Rihanna Photos
More Photos »
Rihannanow.com
Check out the behind the scenes of Rihanna's Vogue cover shoot, on news stands now!
rihannanow.com - Cached
Rihanna - Wikipedia, the free encyclopedia
Biography | Artistry | Other ventures | Philanthropy
Robyn Rihanna Fenty, better known as simply Rihanna, is a Barbadian R&B recording artist and songwriter. Born in Saint Michael, Barbados, Rihanna moved to the United States at the age...
en.wikipedia.org/wiki/Rihanna - Cached
Rihanna - AOL Music
Rihanna established her dance-pop credentials in summer 2005 with her debut smash hit, "Pon de Replay," and continued to...
music.aol.com/artist/rihanna - Cached
Rihanna | Free Music, Tour Dates, Photos, Videos
Rihanna's official profile including the latest music, albums, songs, music videos and more updates.
www.myspace.com/rihanna - Cached
Rihanna: Biography from Answers.com
Library > Miscellaneous > Who2 Biographies Born: 20 February 1988 Birthplace: Saint
Sponsored Results
Rihanna—Online Photos
Find Pics of Rihanna & other Hip-Hop Stars—Bing™ Image Search
bing.com
See your message here...

Problem Statement

- Many queries (50%) search for specific entities instead of documents [Kumar&Tomkins09]
- Web Search Engines do not satisfy users needing specific entities
- Retrieve a list of entities instead of a list of documents

Query: «Female singers who won the Grammy award»

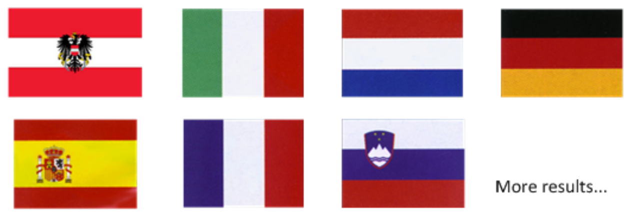
Results:



More results...

Query: «Countries where I can pay in Euro»

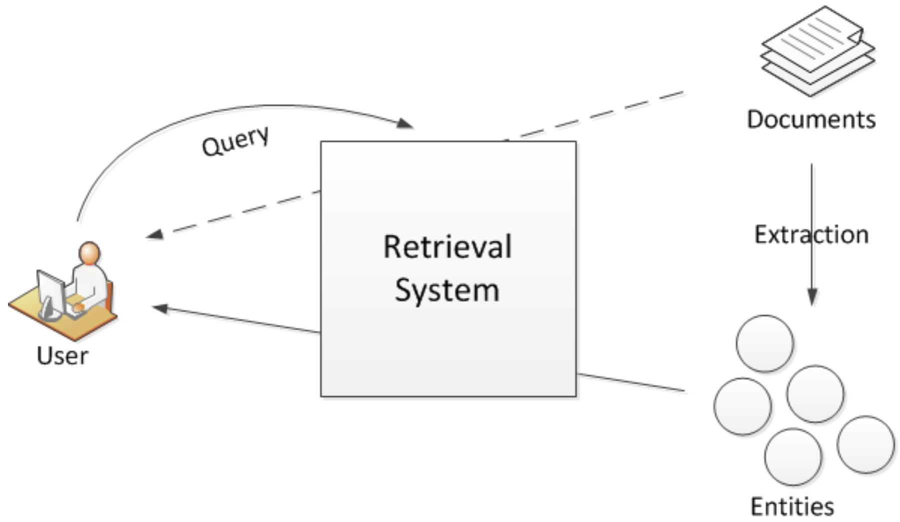
Results:



More results...

Ph.D. Overview

Entity Retrieval



Outline

Motivation

Part I - Retrieving Entities in the Enterprise

- A Vector Space Model for Ranking Entities
- Application to Expert Finding

Part II - Retrieving Entities in Wikipedia

- Evaluation of Entity Retrieval in Wikipedia
- Structure-based techniques
- NLP-based techniques

Part III - Additional Dimensions of Entity Retrieval: Time and Opinion

- Time-Aware Entity Retrieval
- Mining Opinions about Entities on the Web

Part IV - Test Collections for Entity Retrieval

Conclusions

PART I
RETRIEVING ENTITIES IN THE ENTERPRISE

Expert Finding - Motivation

Scenario

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

Goal

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

Related Work

- P@noptic Expert [Craswell et al. Ausweb01]
- Balog’s Model 1 [Balog et al. SIGIR06]
- Voting Model [Macdonald and Ounis CIKM06, ECIR07, ECIR08]
- Experise evidence [Macdonald et al. ECIR08]
- Topic drift: ProjSim allows multiple expertises

Gianluca Demartini, Julien Gaugaz, and Wolfgang Nejdl. **A Vector Space Model for Ranking Entities and Its Application to Expert Search**. In: 31st European Conference on Information Retrieval (ECIR 2009), Toulouse, France, April 2009.

ECIR 2009

A Vector Space Model for Entity Retrieval

A general model for ranking entities in a document collection

- Allowing integration of known techniques
- For any type of entity
- An application to the expert finding task

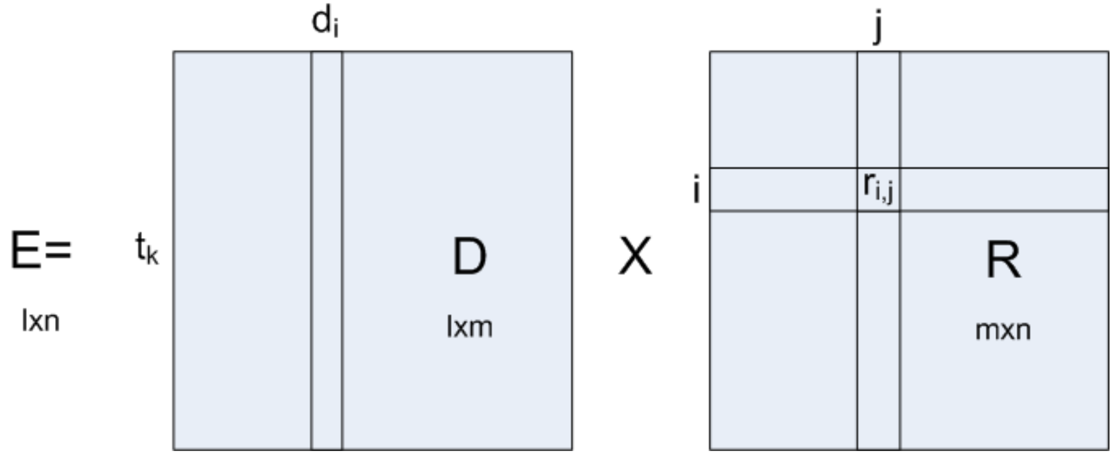
A Vector Space Model for Entity Retrieval

1. Send entities into the space (*l*-dimensional)
   a) Documents as vectors
   b) Relationship between (*n*) entities and (*m*) documents
   c) Entities as vectors

$$d_i = d_{1,i}t_1 + \ldots + d_{l,i}t_l$$

$$f : (d_i, e_j) \rightarrow r_{ij}$$

2. Get them back
   a) Query as vector
   b) Similarity measures
   c) Rank entities



$$e_j = \sum_{k=1}^{l} \left(\sum_{i=1}^{m} d_{k,i} r_{i,j} \right) t_k$$

Extensions of the Model

A lot of expertise evidence may be available in the Enterprise

Document dependent

$$E = D \times (diag(x) \times R)$$

- Importance of each document (e.g., doc type)
- diag(x) is m x m with x_{ii} is the weight for d_i

Entity dependent

$$E' = E \times diag(cf)$$

- Cost function for entities (e.g., salary)
- diag(cf) is n x n and cf_{jj} is the cost of e_j

Experiments - Dataset

TREC

- Evaluation initiative for text retrieval
- Document collections
- Queries
- Relevance judgements

<num> Number: EX52

<title>ontology engineering</title>

<desc> Description:
Find individuals with expertise regarding ontology engineering.
</desc>

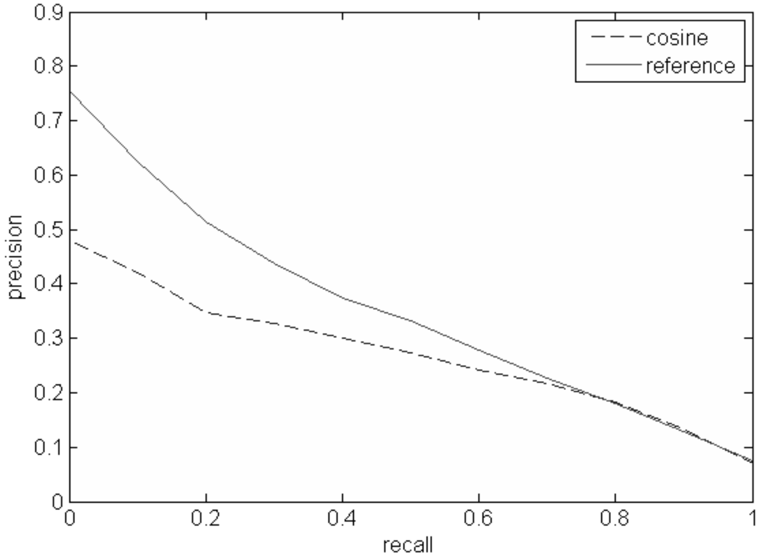
<narr> Narrative:
This topic attempts to find individuals with expertise regarding to ontology engineering. Ontology engineering concerns the whole life-cycle of ontologies, such as ontology construction, ontology learning, ontology mapping, and ontology evolution. We want people with expertise about ontology engineering rather then other things related to ontology.
</narr>

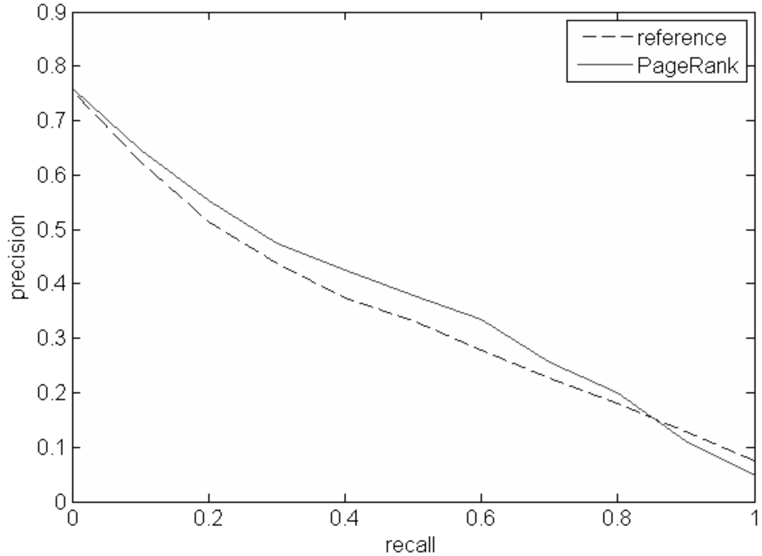
Experiments - Dataset

TREC Enterprise track 2006

- http://www.ins.cwi.nl/projects/trec-ent/
- W3C website (crawl of w3.org sites in June 2004)
- 331k documents (mailinglists, homepages, developers, etc.)
- 1092 (official) candidate experts
- 55 topics with manual relevance assessments

Results





- Cosine sim normalizes doc length
- "Long" entities are good

$$proj\,Sim(\vec{q},\vec{e}) = \cos\theta \|\vec{e}\| = \frac{\vec{q}\cdot\vec{e}}{\|\vec{q}\|}$$

- Reputation of documents
- PageRank

$$E = D \times (diag(x) \times R)$$

Discussion – Part I

We presented a model for Entity Ranking

- It is based on the Vector Space Model
- Can be applied where entities are available
- Can be extended with different types of evidence

Main observations for task of Expert Finding

- Occurences in the author field are more important
- Our similarity measure performs better than cosine similarity
- Document dependent extensions improve effectiveness

Open Issue

- It only considers a single entity type (i.e., people)

PART II
RETRIEVING ENTITIES IN WIKIPEDIA

Wikipedia

Encyclopedia

- Multilingual
- Web-based
- Free-content
- Openly-editable: errors are promptly corrected

Categories / sub-categories

Links, anchor text (Germany → Albert Einstein)

Entities in Wikipedia

Art museums and galleries

Countries

Famous people (Actors, Singers)

Monarchs

Artists

Magicians

...

Example Entity Ranking Scenarios

Impressionist art museums in the Netherlands

German car manufacturers

Countries involved in WWI

Actors who played Hamlet

English monarchs who married French women

Harry Potter Quidditch Gryffindor character

Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, and Wolfgang Nejdl.
Why Finding Entities in Wikipedia is Difficult, Sometimes. In: "Information Retrieval" 13(5): 534-567, Springer, October 2010.

IR Journal 2010

Part II Overview

We proposed algorithms for ER on top of Wikipedia

It is possible to search for many different entity types with one system!

Main observations

- Link information is important
- Cleaning the category structure of Wikipedia is critical (YAGO)
- NLP-based techniques on the user query improve effectiveness

Open issues

- No temporal evolution of content is considered
- Wikipedia is meant to be objective

Outline

Motivation

Part I - Retrieving Entities in the Enterprise

- A Vector Space Model for Ranking Entities
- Application to Expert Finding

Part II - Retrieving Entities in Wikipedia

- Evaluation of Entity Retrieval in Wikipedia
- Structure-based techniques
- NLP-based techniques

Part III - Additional Dimensions of ER: Time and Opinion

- **Time-Aware Entity Retrieval**
- **Mining Opinions about Entities on the Web**

Part IV - Test Collections for ER

Conclusions

PART III
ADDITIONAL DIMENSIONS OF ER: TIME AND OPINION

Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, Hugo Zaragoza. **Entity Summarization of News Articles**. In: 33rd Annual ACM SIGIR Conference, Geneva, Switzerland, July 2010.
Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, Hugo Zaragoza. **TAER: Time Aware Entity Retrieval**. In: The 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, October 2010.

Motivation

Techniques for static document collections do not apply everywhere

News stories evolve over time and entities appear/disappear

- Analyse and exploit relevance evolution
- Decide about relevance at document level

SIGIR 2010
CIKM 2010

Scenario

An event

- Charles Schulz dies

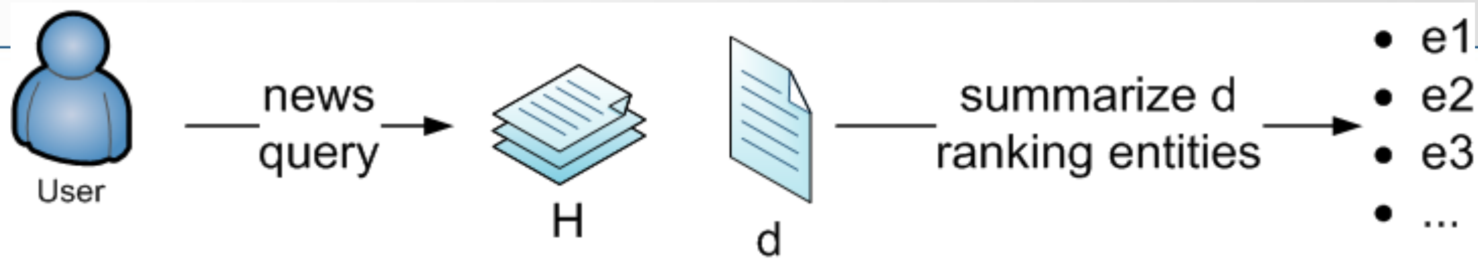
Get Relevant Docs

Entities

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

Timeline of relevant news:

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



Time-Aware Entity Retrieval

Find the entities e_i that best describe document *d* wrt a query *q*

Charles Schulz Dies | **Search**

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

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

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

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

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

Find e_i for *d* wrt a query *q* given history $d_i < d$

Related Work

Entity Retrieval on Wikipedia

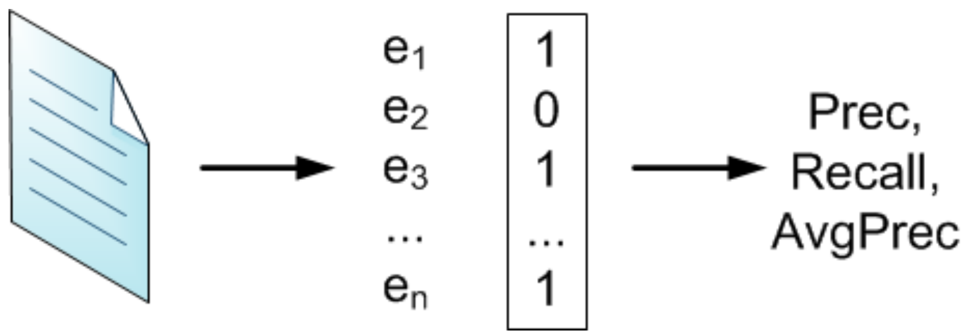
- Ranking entities in a static collection
- Links and structure is important

Finding related Entities

- Using Web content
- TREC Entity Track (Airlines that currently use Boeing 747 plane)

Time-based IR: doc search in collections over time

Experimental Setting



Dataset and Analysis

TREC Novelty Track 2004

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

Entity annotations (7481 entities)

Relevance judgements

How useful is to find relevant sentences?

- P(e is Rel) 0.411 [0.404-0.417]
- P(e is NotRel) 0.168 [0.163-0.173]
- P(e is Rel | s is Rel) 0.547 [0.534-0.559]
- Sentences:
  - 21727 total 1.46 entity occurences
  - 5122 relevant 1.88 entity occurences

Data Analysis

How useful is looking at the past?

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

Is useful to consider sentence co-occurence?

$P(e_1,e_2)$	Relevant	Related	NotRelevant	NotAnEntity
Relevant	0.24	0.08	0.03	0.07
Related		0.07	0.03	0.03
NotRelevant			0.07	0.05
NotAnEntity				0.04

Approach

Entity Ranking features for News articles

- Local Features
  - F(e,d)
  - FirstSenLen
  - FirstSenPos
  - F_{subj}
  - AvgBM25(q,s)
  - SumBM25(q,s)
- History Features

Feature combination

- Linear and Machine Learning

Local Features

Feature	P5	MAP
F(e,d)	**.56**	**.60**
FirstSenLen	.36	.45
FirstSenPos	.31	.43
F_{subj}	.44	.50
AvgBM25(q,s)	.30	.41
SumBM25(q,s)	.44	.52

Feature	P5	MAP
All Tied	.34	.42

Is the past useful?

Looking at previous documents

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

History Features

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

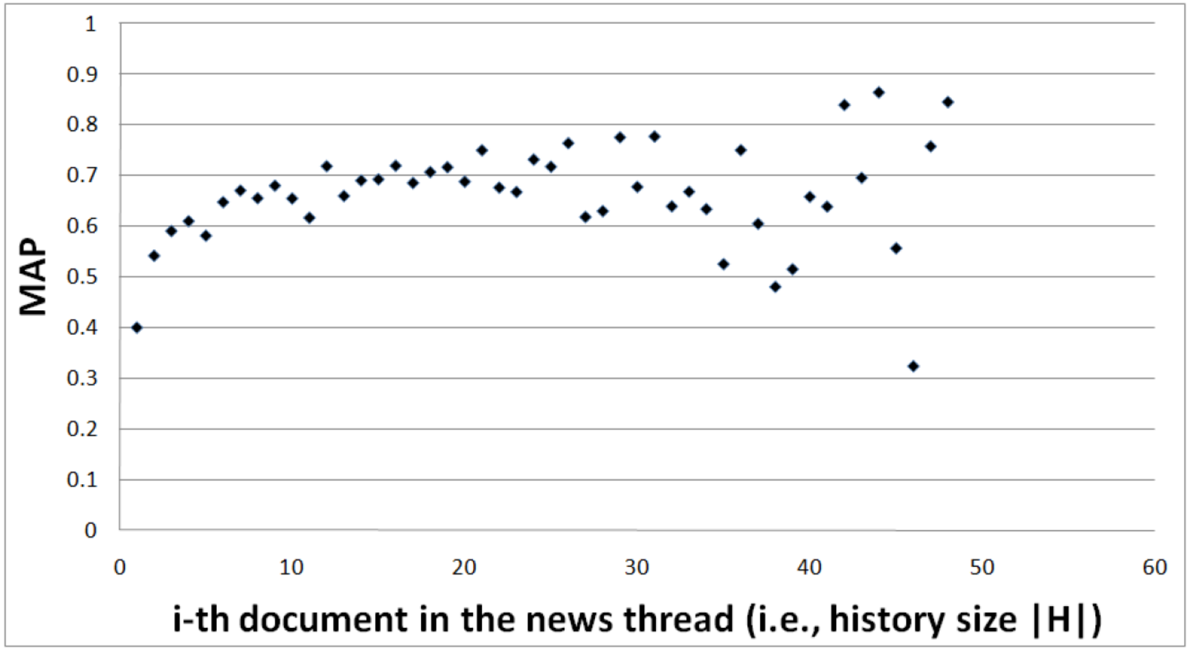
* t-test p value < 0.05 as compared with F(e,d)

** t-test p value < 0.01 as compared with F(e,d)

Using the History

Conclusion

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



Combining Features with ML

Logistic Regression for ranking entities

5-folds cross validation on 25 topics

Similar results for combinations of 2 features

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

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

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

Discussion

Defined new search task: Time-Aware Entity Retrieval

Constructed evaluation benchmark

Experimental Evaluation

- Investigated some features and combinations
- Information from the past helps most
- Obtain 15% improvement over F(e,d)

Gianluca Demartini, Stefan Siersdorfer, Sergiu Chelaru, and Wolfgang Nejdl. **Analyzing Political Trends in the Blogosphere**. To appear in: Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain, July 2011.

ICWSM 2011

Estimating public opinion about entities

Motivation

- Wikipedia is objective, while opinions are expressed on the Web
- Blogs as forum for sharing experiences and opinions
- Public opinion is usually estimated by surveys
  - Limited sample
  - High cost for companies and interviewed people

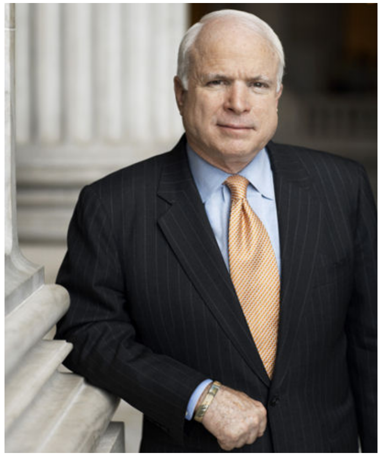
We propose models for estimating public opinions from the blogosphere

- mining opinions
- aggregating information over time
- exploiting Time Series Analysis

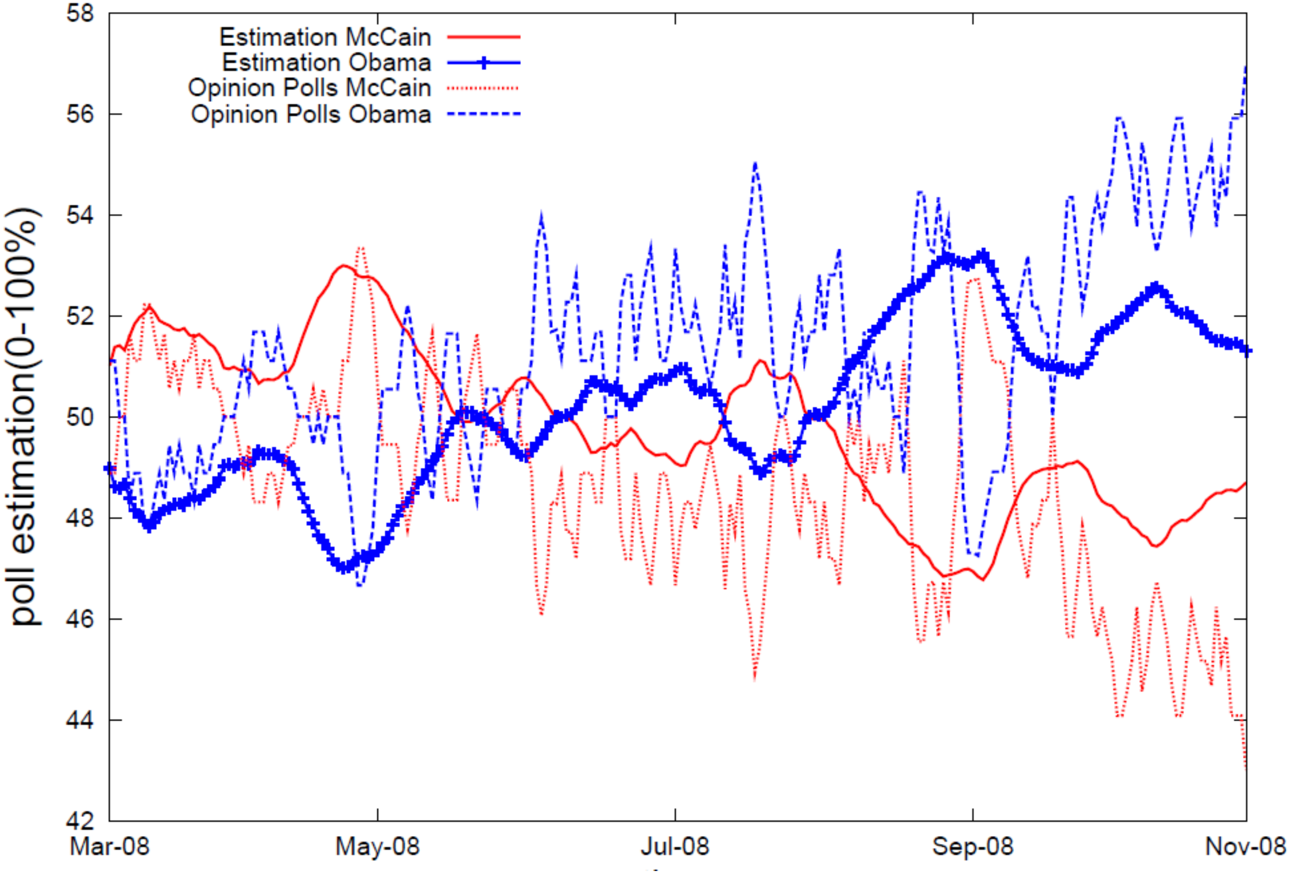
Scenario

US Presidential Elections 2008

- 2 competing entities: Obama vs McCain
- Estimate public opinion over time
- Ground truth:
  - Professional telephonic polls (Gallup)



Public Opinion Estimation



Related Work

Opinion Mining

- TREC Blog track
- Retrieve opinionated postings and their polarity
- Opinion aggregation over Twitter Data (O’Connor et al. 2010)

Time Series Analysis

- Only looks at data series and trends

Our Approach

Extracting opinions

- Lexicon of opinionated words
- Training a learning model using opinionated text

Aggregating opinions

- Unsupervised
  - Only blogs
- Supervised (using history of polls)
  - Learn model parameters
  - Learn prediction models

Model parameters

Lag

- Time delay between telephonic polls and blogs

Bias

- Difference in opinions due to the biased sample

Scale

- "Strength" of opinions

$$poll(t, lag, bias, scale) = (poll(t + lag) + bias) \cdot scale$$

Smoothing

- Remove noise

$$poll(t, k) = \frac{\sum_{j=0}^{k-1} poll(t - j)}{k}$$

The collection

Blogs08

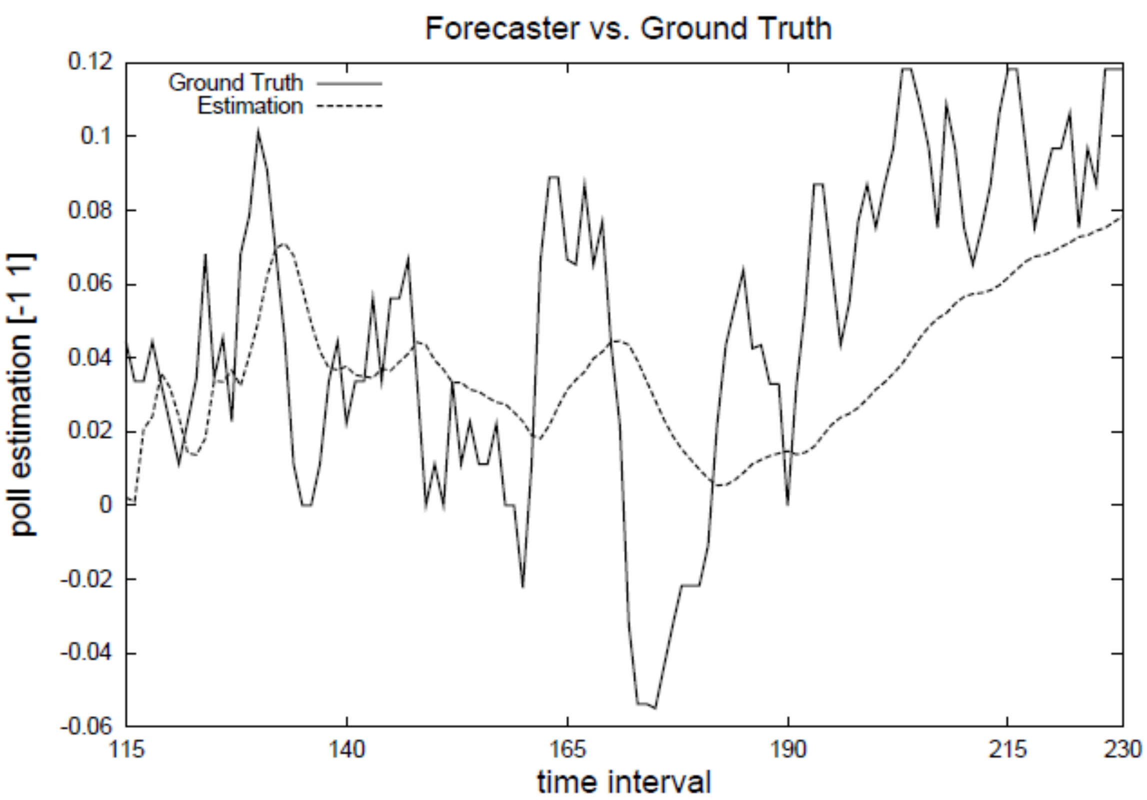
- 28M permalink documents
  - 65% are in English
  - 671K posts contain obama/mccain
- January 2008 - February 2009

Gallup polls

- 1500 US adults
- March-November 2008 (230 polls)
- Normalized scores to avoid “undecided voters”

Experimental Evaluation

Root Mean Square Error between estimation and telephonic polls



Experimental Evaluation

Method	RMSE	
	unsupervised	supervised
LexAvg	0.0642	0.0556
ClassifyAvg	0.0619	0.0608
LexCount	**0.0572**	0.0483
ClassifyCount	0.1980	**0.0482**
LexCountLinFor	**0.0394** ($\lambda = 0.2$)	

Discussion

Problem of estimate the public opinion about target entities

We proposed unsupervised and supervised methods exploiting the Blogoshpere based on

- Lexicon
- Text classification

Effective as compared to standard telephonic polls

STANDARD REUSABLE TEST COLLECTIONS

Comparing and evaluating IR Systems

Availability of standard and reusable test collections is critical

- For comparing systems and algorithms
- For evaluating a model in different settings/collections

We contributed to the creating of such collections by

- Organizing and evaluation initiative
- Creating a test collection for TAER

Gianluca Demartini, Arjen P. de Vries, Tereza Iofciu, and Jianhan Zhu. **Overview of the INEX 2008 Entity Ranking Track**. In: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008 Dagstuhl, Germany, December 2008.
Gianluca Demartini, Tereza Iofciu, and Arjen P. de Vries. **Overview of the INEX 2009 Entity Ranking Track**. In: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009 Brisbane, Australia, December 2009.

Evaluating ER in Wikipedia

INEX Entity (XER) track 2007-2009

- http://www.inex.otago.ac.nz/tracks/entity-ranking/entity-ranking.asp
- http://www.L3S.de/~demartini/XER08
- http://www.L3S.de/~demartini/XER09

INEX 2008
INEX 2009

Standard test collection using

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

Evaluating ER in Wikipedia

35+55 Queries and manual relevance judgements

<inex_topic topic_id="113">
<title>**Formula 1 drivers that won the Monaco Grand Prix**</title>
<description>I want a list of Formula 1 drivers. Each of them must have won at least once the Monaco Grand Prix held in Monte Carlo.</description>
<narrative>The Monaco Grand Prix (GP) is a Formula 1 competition held in Monte Carlo since 1929. I want to find all the drivers that won the GP over the years.</narrative>
</inex_topic>

Pooling techniques and evaluation measures to compare sytems

2008 and 2009 Overview papers: 15+5 citations

Dataset for Evaluating ER over Time

TREC Novelty Track 2004

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

Entity annotations (7481 entities)

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

Relevance judgements

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

http://www.L3S.de/~demartini/deert/

CONCLUSIONS


Ph.D. Contributions

We contributed in going beyond document retrieval in Search Engines

Extensive study of the Entity Retrieval problem

- in different settings: Enterprise, Wikipedia, News, Blogs

We addressed

- The problem of Expert Finding exploiting Enterprise content
- The problem of multi-type Entity Retrieval on top of Wikipedia
- The problem of entity relevance evolving over time
- The problem of different opinions expressed about entities

Created reusable evaluation collections for Entity Retrieval

- Organizing evaluation initiatives
- Releasing new standard datasets

Ph.D. Summary

Demartini et al. **A Vector Space Model for Ranking Entities and Its Application to Expert Search**. In: 31st European Conference on Information Retrieval (ECIR 2009), Toulouse, France, April 2009.

Demartini et al. **Why Finding Entities in Wikipedia is Difficult, Sometimes**. In: "Information Retrieval" 13(5): 534-567, Springer, October 2010.

Demartini et al. **Entity Summarization of News Articles**. In: 33rd Annual ACM SIGIR Conference, Geneva, Switzerland, July 2010.

Demartini et al. **TAER: Time Aware Entity Retrieval**. In: The 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, October 2010.

Demartini et al. **Analyzing Political Trends in the Blogosphere**. To appear in: Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain, July 2011.

Demartini et al. **Overview of the INEX 2008 Entity Ranking Track**. In: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008 Dagstuhl Castle, Germany, December 2008.

Demartini et al. **Overview of the INEX 2009 Entity Ranking Track**. In: 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009 Brisbane, Australia, December 2009.

40 publications

- 2 in Journals
- 10 papers and 8 posters in Conference proceedings
- 14 in peer-reviewed Workshops