A Ranking Framework for Entity Oriented Search using Markov Random Fields

Hadas Raviv, David Carmel
IBM Research – Haifa Lab, Israel

Oren Kurland
Faculty of IE and Management Technion, Israel

Hadas

David

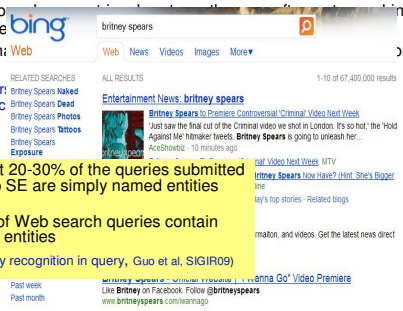
Oren

Outline

- Entities Oriented Search
  - Popular Approaches
- MRF for information retrieval
- MRF for Entity Oriented search
  - Entity Document Scoring
  - Entity Type Scoring
  - Entity Name Scoring
- Evaluation
  - Benchmarks: INEX entity tracks 2007 -2009
  - Experimental Results
- Summary and future work

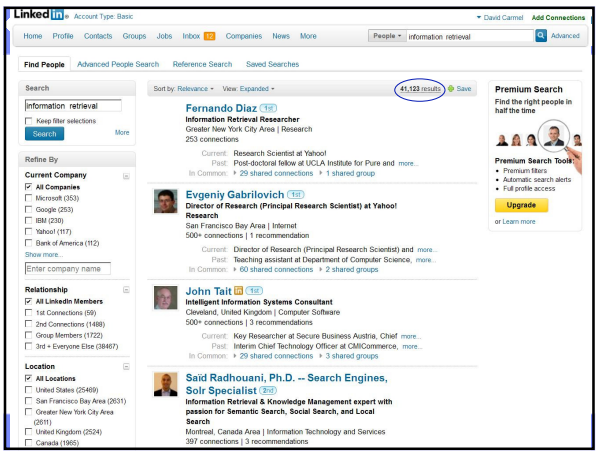
Entity Oriented Search (EoS)

- When p… ng for docume…
- Often n… ormation needs
  - per… loc…



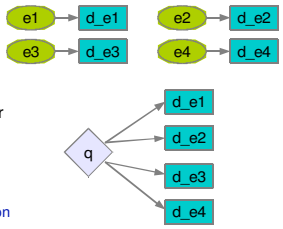
- At least 20-30% of the queries submitted to Web SE are simply named entities
- ~71% of Web search queries contain named entities

(Named entity recognition in query, Guo et al, SIGIR09)



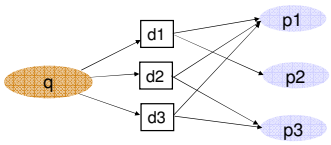
EoS: Profile based Approach (Craswell et al 2001):

- Represent each entity by a virtual document (a profile)
- e.g.
  - Entity home-page
  - Concatenating passages mentioning the entity
- Rank those profiles according to their relevance to the query
  - Using standard IR ranking techniques
- Difficulties:
  - Co-resolution and name disambiguation
  - Profiling is not an easy task



EoS: Voting approach (Balog06, MacDonald09)

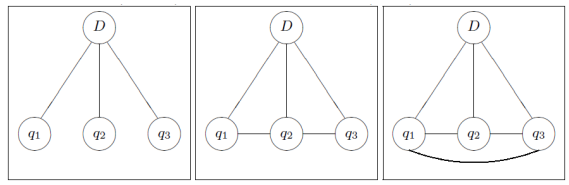
- Any relevant document is a "voter" for the entity mentioned within its content



$$Score(p,q) = \sum_d Score(d,q) * Score(p,d)$$

- What is the ratio behind?
  - An entity mentioned many times in relevant (top retrieved) docs is more likely to be relevant on the given topic?

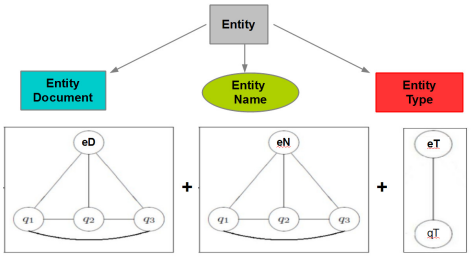
Markov Random Fields for IR (Metzler & Croft 2005)



Full Independence	Sequential dependence	Full dependence
$C_T=(q_i,D)$	$C_O=(\#1(q_i..q_{i+k}),D)$	$C_U=(\#uwN\{q_i..q_j\},D)$

$$P(D \mid Q) \triangleq \sum_{c \in C(G)} \lambda_c f(c)$$

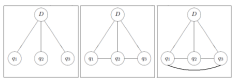
MRF for EoS



$$P(E \mid Q) \triangleq \sum_{P \in \{D,T,N\}} \lambda_{E_P} P(E_P \mid Q) \qquad Q = <\{q_1 ... q_n\}, T>$$

MRF based Entity Document Scoring $P(E_D|Q)$

- We consider cliques of the three types
  - Full Independent (C_T)
  - Sequential dependent (C_O)
  - Full dependent (C_U)



- The feature functions $f^I_D(c)$ over clique of type I (I *in {T,O,U}*)
  - measures how well the clique's terms represent the entity document
  - Based on Dirichlet smoothed language model

$$f_D^T(q_i, E_D) \triangleq \log\left[\frac{tf(q_i, E_D) + \mu \cdot cf(q_i)/|C|}{|E_D| + \mu}\right]$$

  - For C_O and C_U we replace q_i with $\#1(q_i..q_{i+k})$ and $\#uwN(\{q_i,.. q_j\})$ respectively
- The entity document scoring function aggregates the feature functions over all clique types

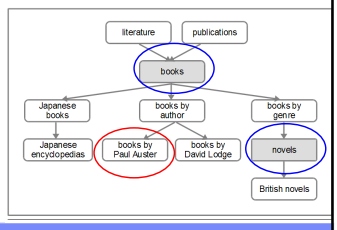
$$P(E_D \mid Q) \triangleq \sum_{I \in \{T,O,U\}} \lambda_{E_D}^I \sum_{c \in I_{E_D}} f_D^I(c)$$

Entity type Scoring $P(E_T|Q)$

- $f_T(c)$ is defined over a single clique composed of E_T and Q_T

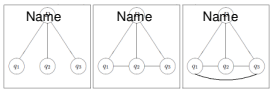
$$P(E_T \mid Q) = f_T(c) \triangleq \log\left[\frac{e^{-\alpha d(Q_T, E_T)}}{\sum_{E' \in R} e^{-\alpha d(Q_T, E'_T)}}\right]$$

- *d(QT,ET)* - the type distance, is domain dependent
- In our experiments we measured the distance in the Wikipedia category graph
  - The minimal path length between all pairs of the query and the entity's page categories



Entity Name Scoring $P(E_N|Q)$

- Clique types:
  - Query terms independent – S_{EN} - a single node clique containing the entity name alone
    - Equivalent to the voting approach
  - Query terms dependent – Consider proximity with the query terms
    - Does the entity is usually mentioned in proximity to query terms
    - In analogy to document scoring
      - T_{EN} Full independent
      - O_{EN} Sequential dependent
      - U_{EN} Full dependent



Entity Name Scoring $P(E_N|Q)$ (cont.)

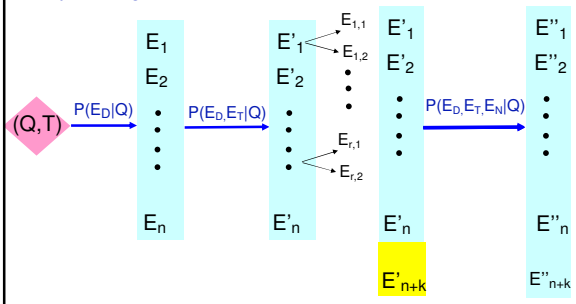
- Local approach
  - Measure the relationship (e.g. proximity) between the query terms and the entity name in the top retrieved documents
- Global approach
  - Measure the PMI between the query term(s) and the entity name in the whole collection
    - PMI – the pointwise mutual information – the likelihood of finding one term in proximity to another term

$$P(E_N \mid Q) = \sum_{X \in A} \lambda_{E_N}^{X} \sum_{c \in X_{E_N}} f_N^X(c)$$

$$A = \{S, T, O, U, PMI_T, PMI_O, PMI_U\}$$

local global

Entity Scoring Process



Evaluation

- The INEX Entity Ranking track 2007-2009
  - The Collection – Wikipedia articles
  - A retrievable entity must have a Wikipedia page
    - No need for a third-party named-entity extraction tools
    - Each entity has a unique name, document and type (WP categories)
  - INEX topics perfectly fit our model looking for relevant entities to a given topic
  - Metrics: MAP for 2007, infAP for 2008-2009

Data set	Wikipedia Year	#docs	Train topics	Test topics
2007	2006	659,388	28	46
2008			74	35
2009	2009	2,666,190	-	55

INEX – Entity Ranking track

- Entity Ranking (XER),
  - Return entities that satisfy a topic described in natural language text
- List Completion (LC)
  - Complete a partial list of given answers
- Entities:
  - Must have a Wikipedia page
    - Entity type is determined by corresponding Wikipedia categories
      - "movies", "trees",
      - "Italian politicians"

```
<inex_topic>
    <title> circus mammals</title>
    <description>
        I want a list of mammals which have ever
        been tamed to perform in circuses.
    </description>
    <narrative>
        Each answer should contain an article about
        mammal which can be a part of any circus
        show.
    </narrative>
    <categories>
        <category id="138">mammals</category>
    </categories>
    <entities>
        <entity id="379035">Asian Elephant</entity>
        <entity id="4402">Brown Bear</entity>
    </entities>
</inex_topic>
```


Parameter tuning

- The following parameters are common for all the entity ranking scores
- Values were selected after extensive search over a wide range of values

Entity Property	Symbol	Parameter name	Optimal Value
E_D	N	Query term proximity window size	10
E_T	d_{max}	Max. distance in category graph	5
	alpha	Category score decay	3
E_N	R	#top-docs for voting score computation	500
	K	Entity name terms proximity window	3
	R_{init}	#top-docs for entity expansion	10

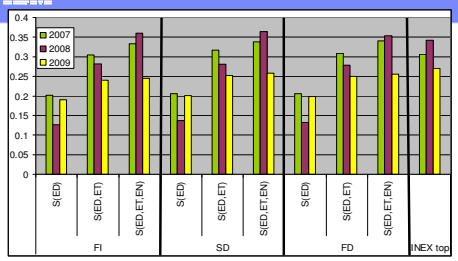
Parameter tuning (cont) – Coordinate Ascent

The parameters of the scoring function were tuned using the Coordinate Ascent algorithm for each benchmark

- Optimization process was done separately for each dataset, using the training topics
- Performance was estimated using the test topics
  - For 2009 we used Cross-Validation

Entity Score	Parameter symbol	2007	2008	2009
$S(E_D)$	μ	100	100	800
	$\lambda_{E_D}^{t}$	0.83	0.87	0.9
	$\lambda_{E_D}^{c}$	0.05	0.05	0.04
	$\lambda_{E_D}^{p}$	0.12	0.08	0.06
$S(E_D, E_T)$	λ_{E_D}	0.42	0.45	0.7
	λ_{E_T}	0.58	0.55	0.3
$S(E_D, E_T, E_N)$	λ_{E_D}	0.19	0.19	0.42
	λ_{E_T}	0.42	0.42	0.35
	$\lambda_{E_N}^{c}$	0.01	0.04	0
	$\lambda_{E_N}^{t}$	0.05	0.02	0.01
	$\lambda_{E_N}^{p} = \lambda_{E_N}^{r}$	0.005	0.01	0
	$\lambda_{E_N}^{PMI_Y}, (Y \in \{T, O, U\})$	0.11	0.11	0.07

Table 4: Parameters of the model which were tuned using the CA algorithm.



- Results improved significantly when type and name scoring were added
- Final Results are superior to top INEX 2007,2008, and comparable to 2009
- Dependence models (SD, FD) have not improved over Independence model (FI) ???
- Global based name scoring (PMI) outperforms local based name scoring

Summary

- In this work we presented an entity ranking model using the MRF framework which integrates
  - Profile approach: query E-document relationship
  - Voting approach: query E-name relationship
  - Type filtering approach: query E-type relationship
- Experiments over INEX benchmarks showed that
  - Performance is relatively high and comparative to leading INEX systems
  - Using dependence models did not result in significant improvement over Full Independence model.
  - Global based name scoring outperforms local based name scoring
- Future work
  - Explore this model with additional data collections, specifically, large web collections
  - Using additional entity properties, e.g. exploring the entity graph
  - Further investigation of the dependence models

Thank You!

Questions?