

Named Entity Recognition in Query

Jiafeng Guo¹, Gu Xu², Xueqi Cheng¹, Hang Li²

¹*Institute of Computing Technology, CAS, China*

²*Microsoft Research Asia, China*

Outline

- Problem Definition
- Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

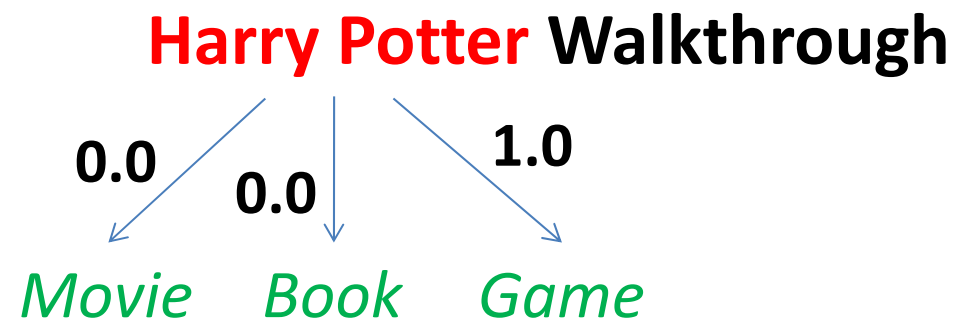
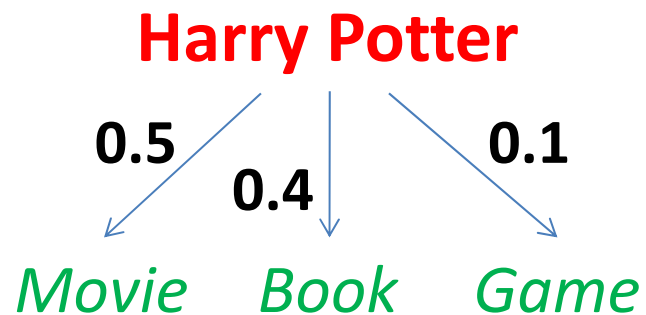
Outline

- Problem Definition
- Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

Problem Definition

Named Entity Recognition in Query (NERQ)

Identify Named Entities in Query and Assign them into Predefined Categories with Probabilities



Outline

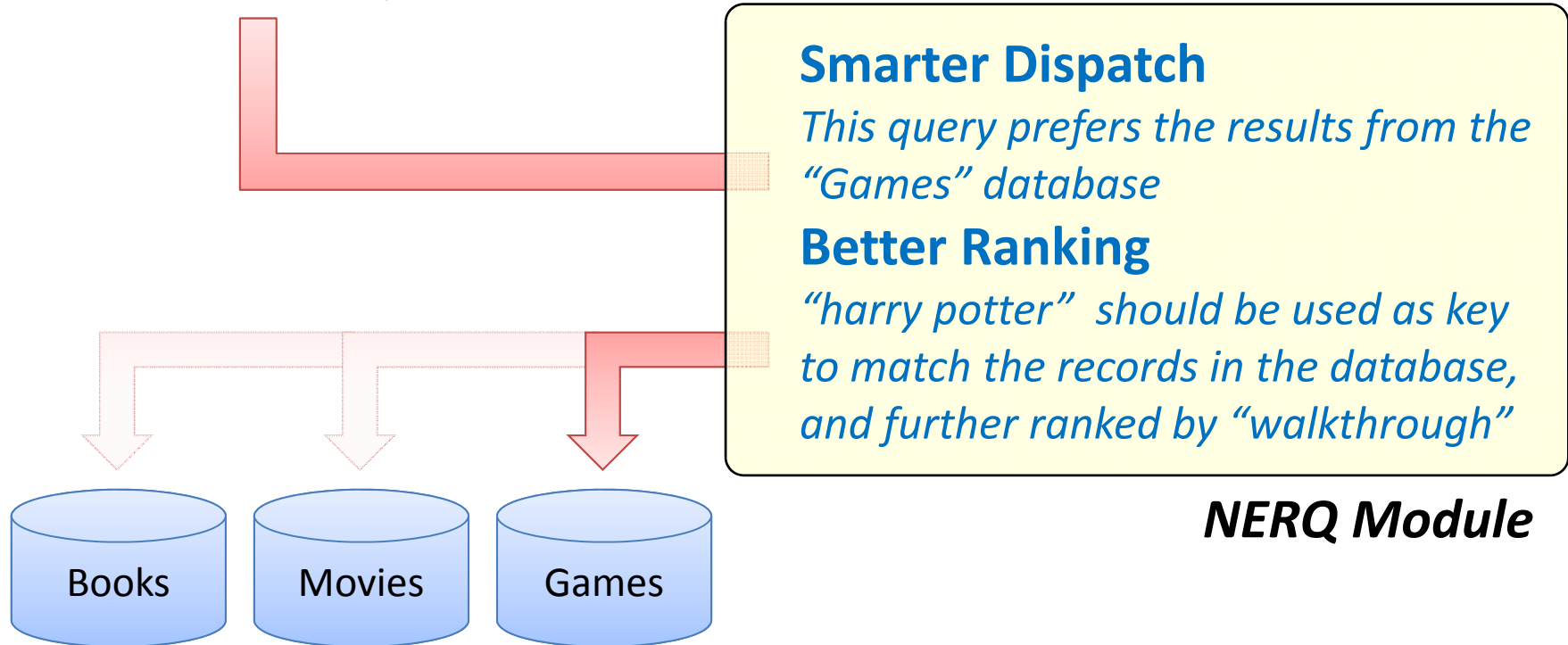
- Problem Definition
- Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

NERQ in Searching Structured Data

harry potter walkthrough



Unstructured Queries



Structured Databases

(Instant Answers, Local Search Index, Advertisements and etc)

NERQ in Web Search

21 movie



[Home | 21 Movie - Official Site](#)

Ben Campbell (Jim Sturgess) finds himself quietly recruited by MIT's most gifted students in a daring plot to break Vegas. With the help of a brilliant ...

www.sonypictures.com/homevideo/21/ - 11k - [Cached](#) - [Similar pages](#) - [Note this](#)



[21 \(2008 film\) - Wikipedia, the free encyclopedia](#)

21 (referred to in advertising as "**21: The Movie**") is a 2008 drama film from Columbia Pictures. It is directed by Australian director Robert Luketic ...

[en.wikipedia.org/wiki/21_\(2008_film\)](http://en.wikipedia.org/wiki/21_(2008_film)) - 58k - [Cached](#) - [Similar pages](#) - [Note this](#)



[MOVIE 21](#)

2008 **21 MOVIE**, **21 MOVIE** COMING SOON! ... release date: Friday March 28, 2008
genre: Action director: Robert Luketic

movie-21.com - [Cached page](#)



[21 – Get wallpapers and ringtones for your mobile phone ...](#)

Download **21**'s Ringtones, Mobile Game, and Wallpapers for your Cell Phone. Join Now and Get 10 Bonus Downloads! Save on Alltel, AT&T, Nextel, Sprint, T-Mobile, Verizon Wireless ...

www.playphone.com/21-Movie - [Cached page](#)

*Search results can be better if we know that
“**21 movie**” indicates searcher wants the movie named **21***

Outline

- Problem Definition
- Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

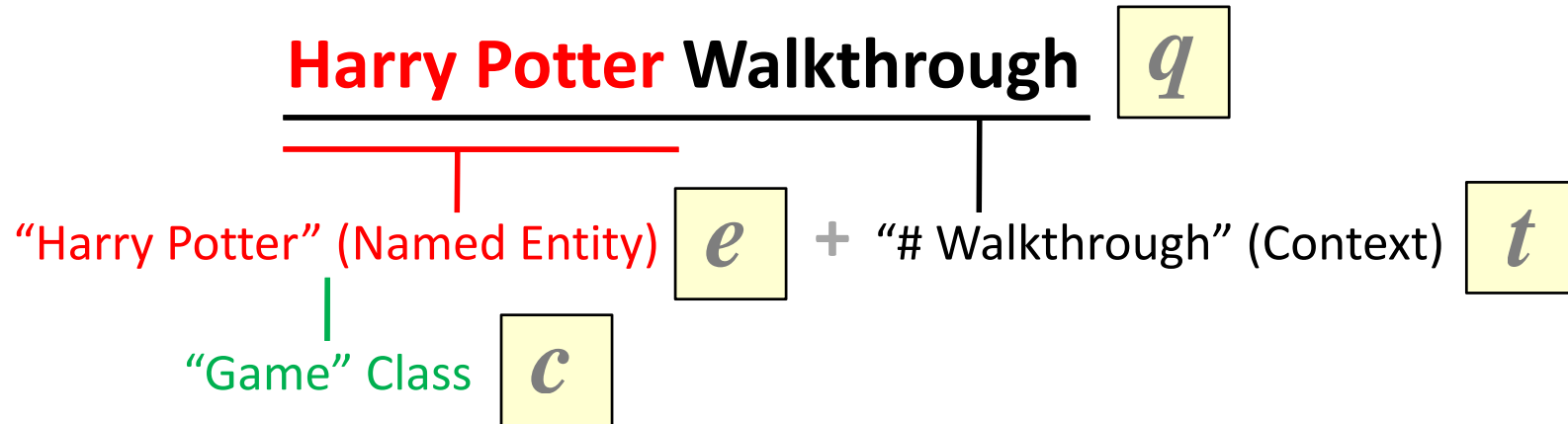
Challenges

- NER (Named Entity Recognition)
 - Well formed documents (e.g. news articles)
 - Usually a supervised learning method based on a set of features
 - Context Feature: whether “Mr.” occurs before the word
 - Content Feature: whether the first letter of words is capitalized
- NERQ
 - Queries are short (2-3 words on average)
 - Less context features
 - Queries are not well-formed (typos, lower cased, ...)
 - Less content features

Outline

- Problem Definition
- Motivation and Potential Applications
- Challenges
- **Our Approach**
- Experimental Results
- Summary

Our Approach to NERQ



- Goal of NERQ becomes to find the best triple $(e, t, c)^*$ for query q satisfying

$$\begin{aligned}(e, t, c)^* &= \arg \max_{(e,t,c) \in G(q)} p(e, t, c) \\ &= \arg \max_{(e,t,c) \in G(q)} p(e)p(c|e)p(t|c)\end{aligned}$$

Training With Topic Model

- Ideal Training Data $T = \{(e_i, t_i, c_i)\}$

$$\max \prod_i p(e_i, t_i, c_i)$$

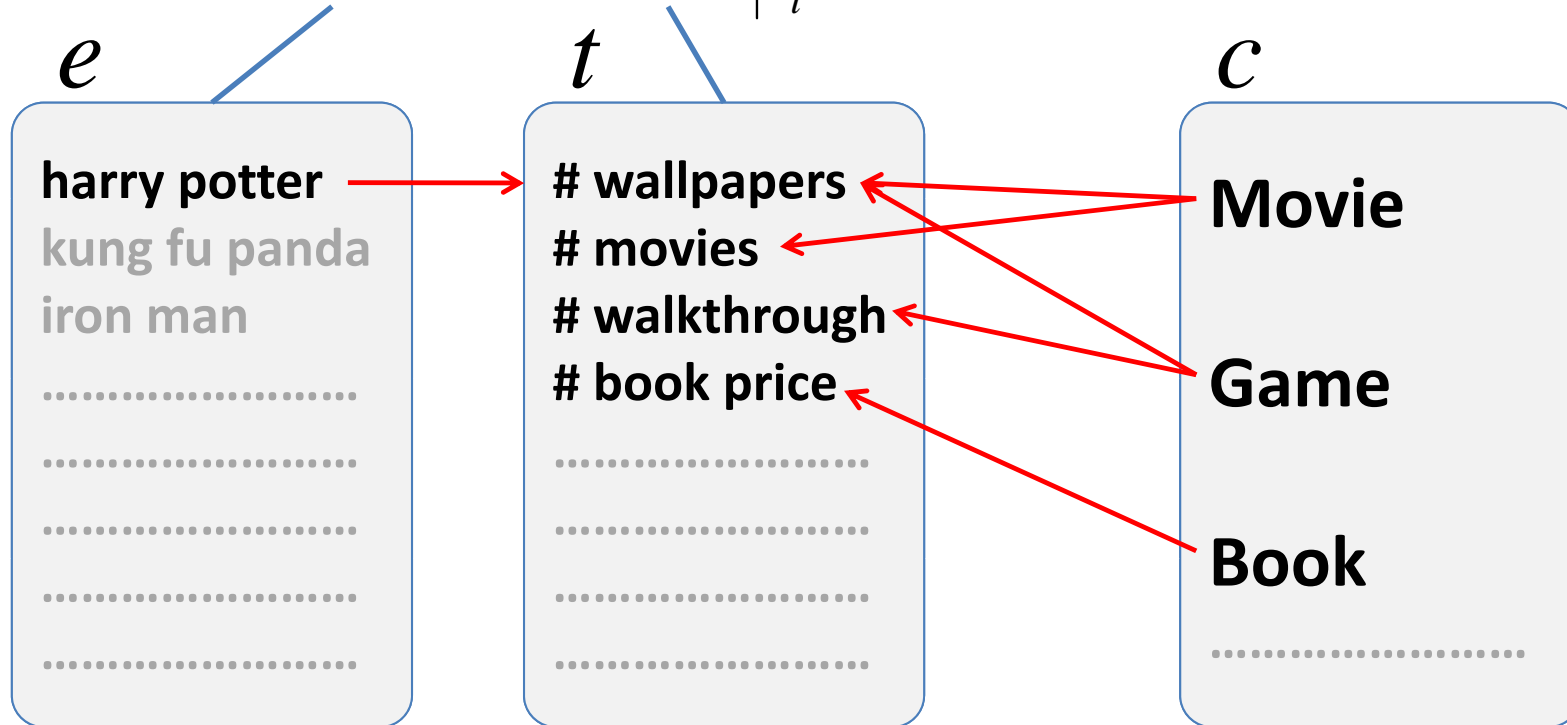
- Real Training Data $T = \{(e_i, t_i, *)\}$

- Queries are ambiguous (**harry potter**, **harry potter** review)
- Training data are a relatively few

$$\begin{aligned} \max \prod_i \sum_c p(e_i, t_i, c) &= \max \prod_i p(e_i) \sum_c p(c|e_i) p(t_i|c) \\ &= \max \prod_e \prod_{i|e_i=e} p(e) \sum_c p(c|e) p(t_i|c) \end{aligned}$$

Training With Topic Model (cont.)

$$\max \prod_e p(e) \prod_{i|e_i=e} \sum_c p(c|e) p(t_i|c)$$

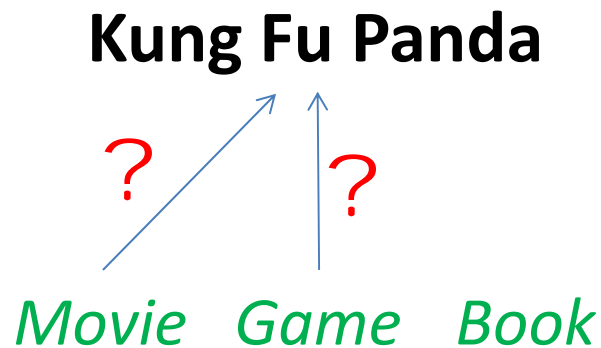


is a placeholder for name entity. # means "harry potter" here

Topics

Weakly Supervised Topic Model

- Introducing Supervisions
 - Supervisions are always better
 - Alignment between *Implicit Topics* and *Explicit Classes*
- Weak Supervisions
 - Label named entities rather than queries (doc. class labels)
 - Multiple class labels (Binary Indicator)

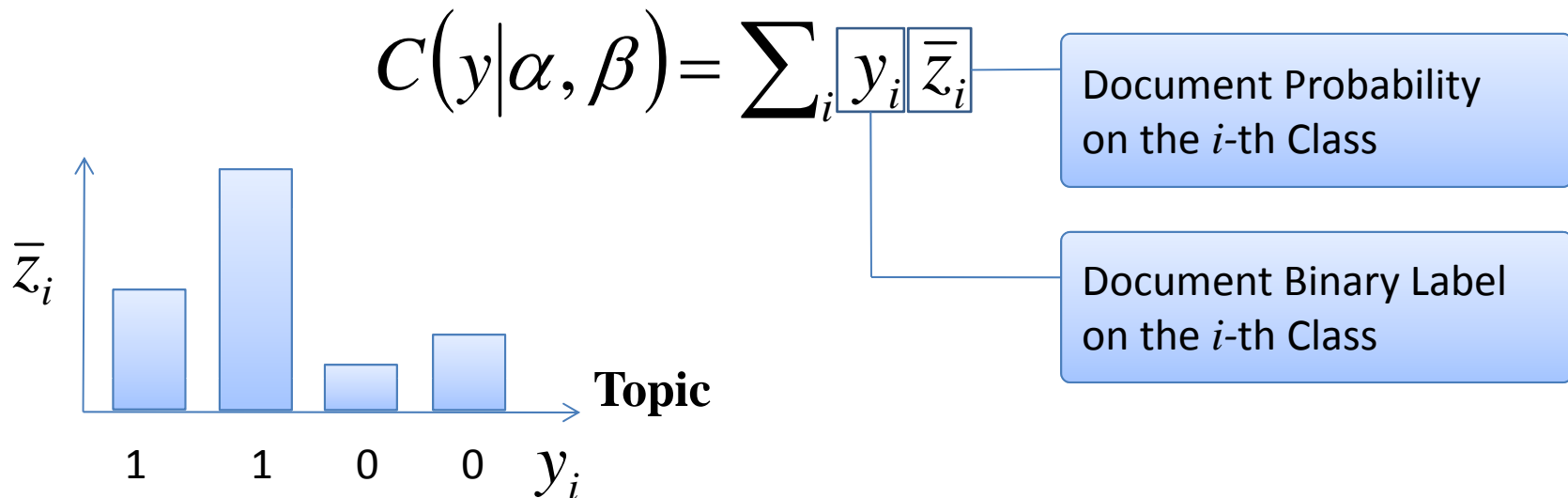


Weakly Supervised LDA (WS-LDA)

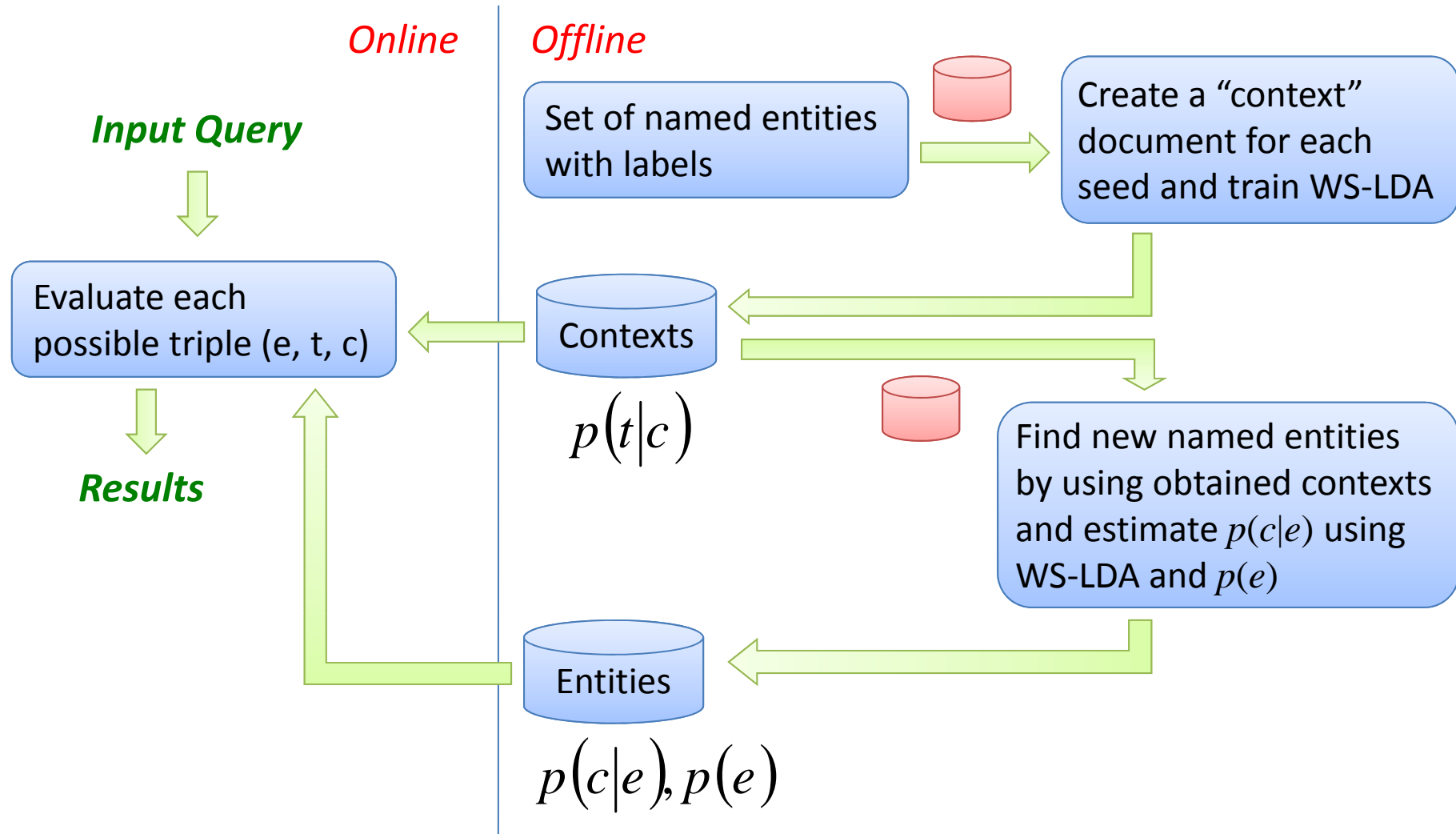
- LDA + Soft Constraints (w.r.t. Supervisions)

$$O(w, y) = \underbrace{\log p(w|\alpha, \beta)}_{\text{LDA Probability}} + \underbrace{\lambda C(y|\alpha, \beta)}_{\text{Soft Constraints}}$$

- Soft Constraints



System Flow Chat



Outline

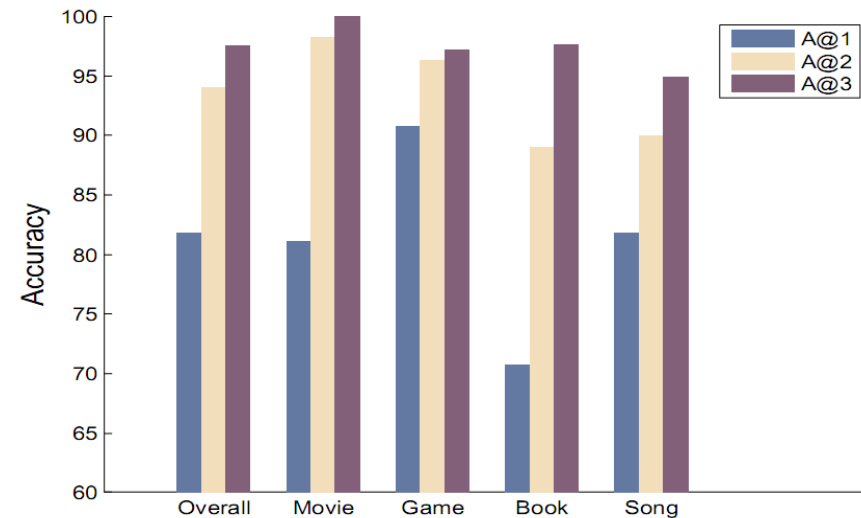
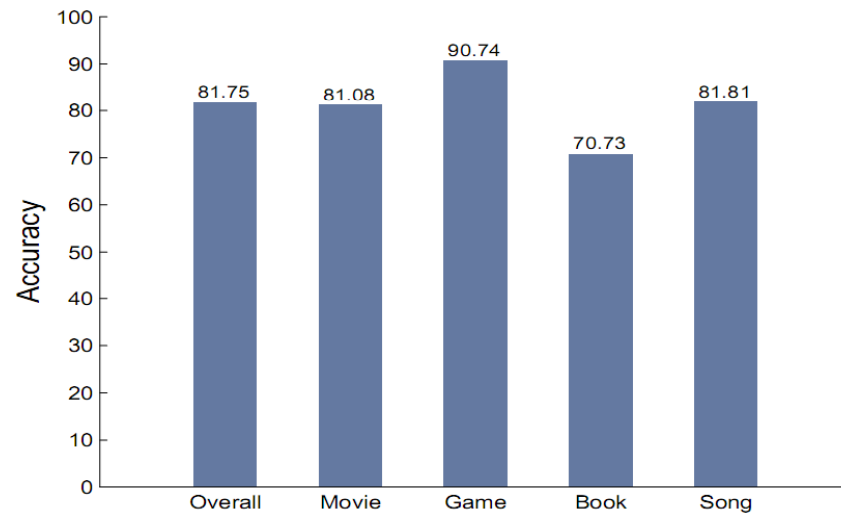
- Problem Definition
- Motivation and Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

Experimental Results

- Data Set
 - Query log data
 - Over 6 billion queries and 930 million unique queries
 - About 12 million unique queries
 - Seed named entities
 - 180 named entities labeled with four classes
 - 120 named entities are for training and 60 for testing

Experimental Results (cont.)

- NERQ Precision



Experimental Results (cont.)

- Named Entity Retrieval and Ranking
 - class distribution
 - Aggregation of seed context distributions (Pasca, WWW07)
 - $p(t|c)$ from WS-LDA model
 - $q(t|e)$ as entity distribution
 - Jensen-Shannon similarity between $p(t|c)$ and $q(t|e)$

Table 6: Comparisons on Ranked Candidate Named Entities of each Class (P@N)

	Movie		Game		Book		Music		Average-Class	
	Determ	WS-LDA	Determ	WS-LDA	Determ	WS-LDA	Determ	WS-LDA	Determ	WS-LDA
P@25	0.92	1	0.98	1	0.84	1	0.96	1	0.92	1
P@50	0.9	1	0.96	1	0.82	1	0.92	1	0.905	1
P@100	0.85	1	0.93	0.98	0.79	0.98	0.89	1	0.865	0.99
P@150	0.82	1	0.92	0.953	0.767	0.98	0.833	1	0.835	0.983
P@250	0.724	0.988	0.896	0.928	0.732	0.968	0.76	0.984	0.778	0.967

Experimental Results (cont.)

- Comparison with LDA

– Class Likelihood of e : $\sum_{i=1}^K y_i p(c_i|e)$

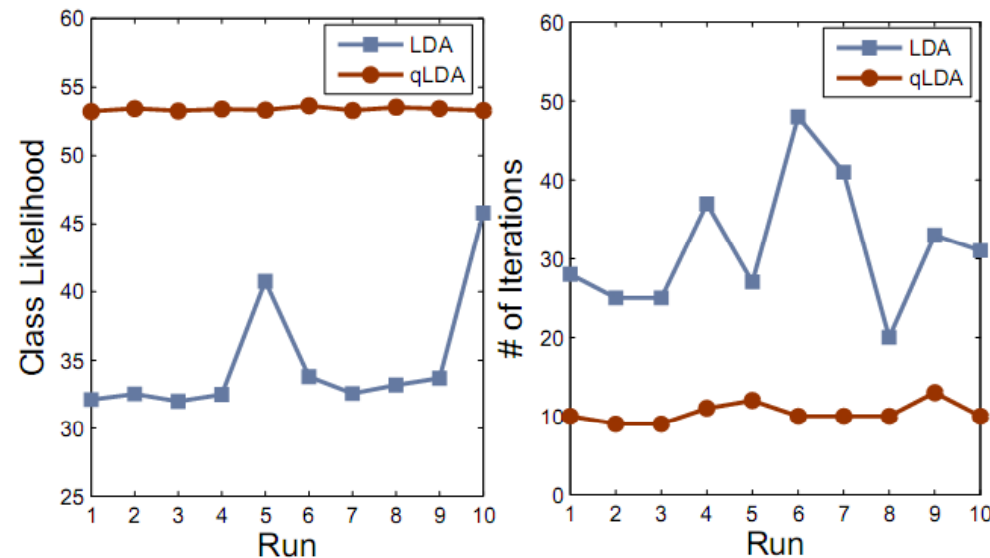


Figure 4: Comparisons between the WS-LDA and LDA approach on (a) Overall Class Likelihood on Testing Set, (b) Convergence Speed on Training Set

Outline

- Problem Definition
- Motivation and Potential Applications
- Challenges
- Our Approach
- Experimental Results
- Summary

Summary

- We first proposed the problem of named entity recognition in query.
- We formulized the problem into a probabilistic problem that can be solved by topic model.
- We devised weakly supervised LDA to incorporate human supervisions into training.
- The experimental results indicate that the proposed approach can accurately perform NERQ, and outperforms other baseline methods.

THANKS!