Improving Web Page Retrieval using Search Context from Clicked Domain Names

Rongmei Li

University of Twente
Enschede - The Netherlands

Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions

Motivation

In the setting of ad-hoc Web page retrieval

Common problems: $\Rightarrow$ query ambiguity

- short keyword query for specific topic of interest
- different vocabularies for the same topic

Traditional Solution:

- explicit user feedback
  - understand search context
  - bridge semantic gap
- implicit user feedback
  - pseudo relevance feedback

Motivation

In the setting of ad-hoc Web page retrieval

Common problems: $\Rightarrow$ query ambiguity

- short keyword query for specific topic of interest
- different vocabularies for the same topic

Traditional Solution:

- explicit user feedback
  - understand search context
  - bridge semantic gap
- implicit user feedback $\Rightarrow$ query log
  - pseudo relevance feedback

Motivation

Query log contains users' search history:

- query terms: Prostate cancer treatments
- retrieved documents:
  http://appliedresearch.cancer.gov/areas/monitoring.html
  http://appliedresearch.cancer.gov/accessibility/
  http://cancercontrol.cancer.gov/hcirb/ceccr/
- clicked documents:
  http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- user identifier: 2178

Motivation

Query log contains users' search history:

- specifying query terms: DEXA 09 Sommerhaus
- retrieved documents:
  http://appliedresearch.cancer.gov/areas/monitoring.html
  http://appliedresearch.cancer.gov/accessibility/
  http://cancercontrol.cancer.gov/hcirb/ceccr/
- clicked documents:
  http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- an anonymous identifier: 2178

Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions

Related Work

Language modeling framework:

- combine ranking results of past queries
- build query model as the average of all past query models
- smooth the current query using past queries and clicked document summaries
- smooth the current query using past queries, retrieved documents, and clicked documents

Query expansion with terms from click-throughs:

- top-ranked documents of the initial retrieval result
- explicitly judged documents

Related Work

Re-ranking top ranked documents:

- re-score documents by statistical distribution of similar queries within a search session or all sessions
- adjust original rank with ranks generated by log queries that contain or are associated with the original query
- learn new rank from user interaction features

Outline

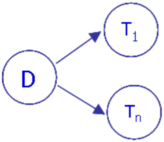
- Motivation
- Related work
- Our work
- Experiments
- Conclusions

- query terms: Prostate cancer treatments
- retrieved documents:
  http://appliedresearch.cancer.gov/areas/monitoring.html
  http://appliedresearch.cancer.gov/accessibility/
  http://cancercontrol.cancer.gov/hcirb/ceccr/
- clicked documents:
  http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- an anonymouse identifier: 2178

Research Questions

- Can we extract the common topical context from such data for a query?
- Can we use this knowledge to improve retrieval performance effectively?

Standard Language Models



$$P(t_1, ..., t_n|D) = \prod_{i=1}^{n} P(t_i|D), \quad P_{ml}(t_i|D) = \frac{tf(t_i, D)}{|D|}$$

Jelinek-Mercer Smoothing:

$$P(t|D) = \lambda P_{ml}(t_i|D) + (1 - \lambda) P_{ml}(t_i|C)$$

Parsimonious Language Models

Use EM estimator to compute $P(t_i|D)$

$$\text{E-step}: \quad e_t = tf(t,D) \cdot \frac{\mu P(t|D)}{\mu P(t|D) + (1-\mu) P_{ml}(t|C)}$$

$$\text{M-step}: \quad P(t|D) = \frac{e_t}{\sum_t e_t}, \text{i.e. normalizing+prunning process}$$

Application: select topical terms from restored Web pages for query expansion

Document Ranking by Cross-entropy Score

$$P_{ml}(t_i|Q) = \frac{tf(t_i, Q)}{|Q|}$$

$$P'(t_i|D) = \lambda P_{ml}(t_i|D) + (1-\lambda)P_{ml}(t_i|C)$$

$$Score(D) = \sum_{i=1}^{l}[P_{ml}(t_i|Q) \cdot \log(P'(t_i|D))]$$

Query Log Modeling (1)

- query terms: Prostate cancer treatments
- retrieved documents:
  http://appliedresearch.cancer.gov/areas/monitoring.html
  http://appliedresearch.cancer.gov/accessibility/
  http://cancercontrol.cancer.gov/hcirb/ceccr/
- clicked documents:
  http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- an anonymouse identifier: 2178

Query Log Modeling (2)

Choose subsets of Web page collection from the Internet

- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels

- **domain level:**
  http://www.cancer.gov ⇒ http://seer.cancer.gov/***

Query Log Modeling (2)

Choose subsets of Web page collection from the Internet

- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels

- **domain level:**
  http://www.cancer.gov ⇒ http://seer.cancer.gov/***
- **server level:**
  http://www.cancer.gov ⇒ http://www.cancer.gov/***

Query Log Modeling (2)

Choose subsets of Web page collection from the Internet

- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels

- **domain level:**
  http://www.cancer.gov ⇒ http://seer.cancer.gov/***
- **server level:**
  http://www.cancer.gov ⇒ http://www.cancer.gov/***
- **URL level:**
  http://www.cancer.gov ⇒ http://www.cancer.gov

Query Log Modeling (3)

Our strategies:

- **strategy 1:** promote the restored Web pages to the top of the ranking list

a	b
b	d
c	a
d	c

- **strategy 2 & 3:** extract topical terms from restored Web pages or top ranked Web pages of strategy 2 to expand the original query

Outline

- Motivation
- Related work
- Our work
- **Experiments**
- Conclusions

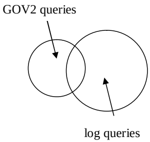
Experiments - data

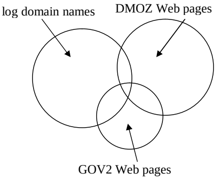
- Query log: 3 month search record containing 10 million queries and 20 million clicked domain names

2178	kbb	2006-03-19 16:33:54	1	http://www.kbb.com
2178	remax	2006-03-21 10:32:18	1	http://www.remax.com
2178	fidelity.com	2006-03-22 21:38:25	1	http://www.fidelity.com

- Reconstruction source: DMOZ (Open Directory Project) or GOV2 collection
- Test data
  - terabyte collection of Web pages with .gov domain name (GOV2 collection)
  - 150 associated queries

Experiments - pre-processing results





Web page restoration at domain level

Stripped URLs	URLs of GOV2 Web Pages	URLs of DMOZ Web Pages
whitehouse.gov	www.whitehouse.gov	www.whitehouse.gov
fws.gov	www.fws.gov	www.fws.gov
doi.gov	www.doi.gov	www.doi.gov

Experiments - topical terms

- original query terms: prostate cancer treatment
- expansion terms

topical term	term probability (EM)	term probability (ML)
cancer	0.0532	0.0248
patient	0.0150	0.0070
para	0.0125	0.0059
trial	0.0103	0.0048
cell	0.0100	0.0047
therapi	0.0097	0.0046
studi	0.0094	0.0044
...	...	...

Performance of strategy 1

models/levels	performance metrics			
	BPREF	improvement	P@10	improvement
baseline (JM)	0.4625	-	0.4455	-
url	0.4625	-	0.4636	+3.90%
server	**0.4814**	+3.93%	0.5182	+14.03%
domain	0.4768	+3.00%	**0.5364**	+16.95%

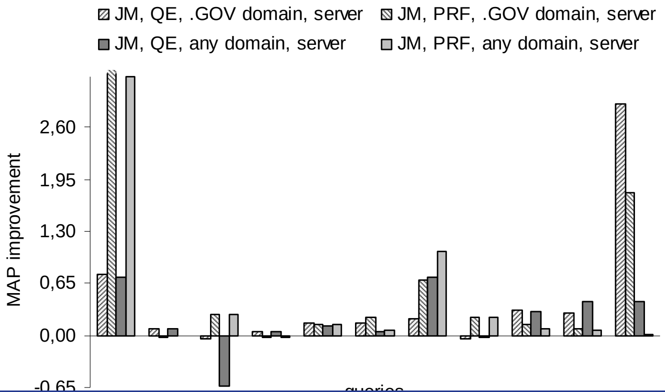
Performance of strategy 2

models/levels	performance metrics			
	MAP	improvement	P@10	improvement
baseline (JM)	0.3294	-	0.4455	-
url.ML	0.3484	+5.45%	0.5273	+15.51%
server.ML	0.3729	+11.67%	0.6091	+26.86%
domain.ML	0.3714	+11.31%	0.5909	+24.61%
url.EM	0.3509	+5.79%	0.5273	+15.51%
server.EM	**0.3849**	+14.42%	**0.6273**	+28.98%
domain.EM	0.3842	+14.26%	0.6091	+26.86%

Performance of strategy 3

models/levels	performance metrics			
	MAP	improvement	P@10	improvement
baseline (JM)	0.3470	-	0.4909	-
url.ML	0.3821	+9.19%	0.5818	+15.62%
server.ML	0.4226	+17.89%	**0.6818**	+28.00%
domain.ML	0.4232	+18.01%	0.6455	+23.95%
url.EM	0.3959	+12.35%	0.5727	+14.28%
server.EM	0.4386	+20.88%	0.6727	+27.03%
domain.EM	**0.4391**	+20.97%	0.6455	+23.95%

Improvement on individual queries



Implicit vs Explicit (true) Feedback

models/levels	performance metrics			
	MAP	difference	P@10	difference
baseline (JM11)	0.4734	-	0.8455	-
best run (11)	0.4391	7.25%	0.6455	23.65%
baseline (JM29)	0.3366	-	0.7034	-
best run (29)	0.3020	10.28%	0.5172	26.47%

Outline

- Motivation
- Related work
- Our work
- Experiments
- **Conclusions**

Conclusions

In this work, we:

- demonstrate how to restore stripped URLs to Web pages at three different levels from two collections
- present three strategies to integrate query and click-through information with the language modeling framework
- show that retrieval performance can be improved effectively

Thank you for your attention ...

Questions and comments?