Effective Structured Query Formulation for Session Search

Dongyi Guan dg372@georgetown.edu

Grace Hui Yang huiyang@cs.georgetown.edu

Nazli Goharian nazli@cs.georgetown.edu

Department of Computer Science
Georgetown University

Problem

- A search session contains
  - Interactions with Previous Queries (Snippets, Search Results, Clicked Information)
  - Current Query
- Given a session, the tasks are to retrieve relevant documents to the current query.
- Four subtasks
  - RL1: only use the current query q_n
  - RL2: uses the previous queries $q_1, q_2, \dots, q_{n-1}$ and the current query q_n
  - RL3: provides top retrieved documents for previous queries
  - RL4: and information about which top results are clicked by users

Structure Query Formulation

- Identify Nugget: substrings in a query, similar to phrase

  servering spinal cord paralysis ⇨ servering #1(spinal cord) paralysis

- Add positional information in snippets

  ...preservation uk **spinal cord** injury care in egypt cky **cord**
  1 2 **3** **4** 5 6 7 8 9 **10**
  noats guitar viewsonic power cord malfunction **spinal cord**
  11 12 13 14 15 16 **17** **18**
  stimulator...injury **spinal cord** dell...extension **cord**
  19 20 **21** **22** 23 24 **25**
  nylon...**cord paralysis** vocal vegas...
  26 **27** **28** 29 30

  - Strict Method

$$\frac{\text{count}(w_i w_{i+1};\text{Snippet})}{\min(\text{count}(w_i w_{i+1};\text{Snippet}),\text{count}(w_i w_{i+1};\text{Snippet}))} \geq \theta \quad \Rightarrow \quad w_i w_{i+1} \text{ is nugget}$$

  - Relaxed Method

$$nugget = \begin{cases} \#1(w_i\ w_{i+1}) & |\bar{x}(w_i) - \bar{x}(w_{i+1})| \leq 5 \\ \#2(w_i\ w_{i+1}) & 5 < |\bar{x}(w_i) - \bar{x}(w_{i+1})| \leq 10 \\ \phi & |\bar{x}(w_i) - \bar{x}(w_{i+1})| > 10 \end{cases}$$

$$\bar{x}(w_i) = \frac{1}{k} \cdot \sum_{t=1}^{k} \frac{\sum_j x_j(w_i; S_t)}{\text{count}(w_i; S_t)}$$

Re-ranking by Dwell Time

Dwell time: the elapsed time that user stays in the page

$$\Delta t = t_{end} - t_{start}$$

Clicked documents: $\{c_1, c_2, \cdots, c_k\}$

Associated dwell time: $\{\Delta t_1, \Delta t_2, \cdots, \Delta t_k\}$

Re-ranking the returned documents $\{d_j\}$ by the score:

$$s(d_j) = \sum_{i=1}^{k} \text{Sim}(d_j, c_i) \cdot \Delta t_i$$

Query Expansion

- with previous queries

#weight(
λ_1 #combine($nugget_{11}\ nugget_{12} \cdots nugget_{1m}\ w_{11}\ w_{12} \cdots w_{1r}$)
λ_2 #combine($nugget_{21}\ nugget_{22} \cdots nugget_{2m}\ w_{21}\ w_{22} \cdots w_{2r}$)
…
λ_n #combine($nugget_{n1}\ nugget_{n2} \cdots nugget_{nm}\ w_{n1}\ w_{n2} \cdots w_{nr}$)
)

$$\lambda_k = \begin{cases} \lambda_p & k = 1,2,\cdots n-1 \\ 1-\lambda_p & k = n \end{cases} \qquad \lambda_p = 0.4$$

- with anchor texts
  - Generate anchor log by *harvestlinks* in Lemur toolkit
  - Extract the top 5 frequent anchor text in the previous results
  - Weights are proportional to the frequency of anchor text
  - Append to the structured query expanded with previous queries

$$\beta\omega_1 \#\text{combine}(e_1)\ \beta\omega_2 \#\text{combine}(e_2) \cdots \beta\omega_5 \#\text{combine}(e_5)$$

frequency anchor text

e.g. servering spinal cord consequenses ⇨ #weight(1.0 #1(spinal cord) 0.6 consequenses 0.4 paralysis 1.0 servering 0.38 #combine(type of paralysi) 0.0048 #combine(quadriplegia paraplegia) 0.0048 paraplegia 0.0048 #combine(spinal cord injury) 0.0024 #combine(quadriplegic tetraplegic))

Duplicated Queries

- Duplicates between a previous query and the current query, use the current query only
- Duplicates among previous queries, remove the duplicated queries

Queries	Without removing duplicate		Removing duplicate	
	Structured query	nDCG@10	Structured query	nDCG@10
shoulder joint pain	#weight(1.4 joint 0.4 nhs 1.4 pain 1.4 shoulder 0.036 #1(shoulder pain) 0.036 #1(frozen shoulder) 0.01 #1(shoulder pain causes) 0.006 #1(bursitis) 0.006 #1(painful shoulder conditions))	0.55	#weight(1.0 joint 1.0 pain 1.0 shoulder)	0.64 (+16.36%)
shoulder joint pain nhs				
shoulder joint pain				

Results

nDCG@10 for TREC 2011 runs

Method \ Baseline = 0.34			anchor text				nDCG@10 in 2011	
			all documents		clicked documents		Best run	Median run
			nDCG@10	%chg	nDCG@10	%chg	RL3	RL3
all queries			**0.47**	38.24%‡	**0.47**	38.24%‡	0.43	0.33
remove duplicated queries			**0.48**	41.18%‡	**0.45**	32.35%‡	RL4	RL4
re-rank by dwell time (RL4 only)			0.44	29.41%‡	N/A		0.45	0.34
Method	original query		strict		relaxed		RL1	RL1
RL1	0.34		0.38	11.76%†	**0.40**	17.65%‡	0.38	0.32
Method	uniform		previous vs current		distance-based		RL2	RL2
RL2	0.45	32.35%‡	**0.46**	35.30%‡	**0.44**	29.41%‡	0.43	0.32

nDCG@10 for TREC 2012 runs

run	Original q	guphrase1	guphrase2	gurelaxphr	Mean of the median
RL1	0.25	0.23	0.23	0.23	0.17
RL2		**0.29**	0.28	0.28	0.19
RL3		**0.30**	**0.30**	**0.30**	0.2
RL4		**0.30**	**0.30**	0.29	0.23

Conclusions

- Terms from previous queries can boost the search performance significantly
- Removing duplicated queries improves the search performance
- Formulating structured queries by grouping terms into nuggets is very effective for TREC2011 data.