PageRank without hyperlinks:

Structural re-ranking using links induced by language models

Oren Kurland and Lillian Lee

Presentation by Yang Yu

Related Work

- Query-dependent clustering
  - Do not directly induces an obvious ranking of doc.
- Techniques based on graphs
  - Similarly used in text summarization

Main Idea

- Using language model to improve IR performance
  - *Generation links*, which are based on the probability assigned by the language model induced from one document
- Structural re-ranking
  - The analogy between hyperlinks and generation links is not perfect.
  - Compute an ordering not of the entire corpus but of a set.

Structural Re-ranking

- Generation Graphs
  - *pd*(·) denotes the smoothed unigram language model induced from *d*
  - Example
    - *d*1: Toronto Sheffield Salvador
    - *d*2: Salvador Salvador Salvador
  - Definition 1:
    - *The* top α generators *of a document d* $\in$ *D*init
  - Definition 2:
    - *The* offspring *of a document d* $\in$ *D*init
      $\{o \in D\text{init} : d \in TopGen(o)\}$

Structural Re-ranking

- Generation Graphs (cont.)
  - Weights on edges

$$wt_U(o \to g) = \begin{cases} 1 & \text{if } g \in TopGen(o), \\ 0 & \text{otherwise;} \end{cases}$$

$$wt_W(o \to g) = \begin{cases} p_g(o) & \text{if } g \in TopGen(o), \\ 0 & \text{otherwise.} \end{cases}$$

  - Definition 3:
    - Weights on smoothed graph

$$wt^{[\lambda]}(o \to g) = (1-\lambda) \cdot \frac{1}{|\mathcal{D}_{\text{init}}|} + \lambda \cdot \frac{wt(o \to g)}{\sum_{g' \in \mathcal{D}_{\text{init}}} wt(o \to g')},$$

Structural Re-ranking

- Computing Graph Centrality

$$Cen_I(d;G) \stackrel{def}{=} \sum_{o \in \mathcal{D}_{\text{init}}} wt(o \rightarrow d). \qquad (1)$$

  - ***Uniform Influx***
  - ***Weighted Influx***

$$Cen_{RI}(d;G) \stackrel{def}{=} \sum_{o \in \mathcal{D}_{\text{init}}} wt(o \rightarrow d) \cdot Cen_{RI}(o;G), \qquad (2)$$

  - ***Recursive Uniform Influx***
  - ***Recursive Weighted Influx***

Structural Re-ranking

- Incorporating initial scores

$$Cen(d;G) \cdot p_d(q), \qquad (3)$$

  - ***Uniform Influx+LM***
  - ***Weighted Influx+LM***,
  - ***Recursive Uniform Influx+LM***
  - ***Recursive Weighted Influx+LM***.

Structural Re-ranking

- Generation Probabilities:
  - *Maximum-likelihood estimate* (MLE) of *w* with respect to *x*

$$\widetilde{p}_x^{MLE}(w) \stackrel{def}{=} \frac{\mathrm{tf}(w \in x)}{\sum_{w'} \mathrm{tf}(w' \in x)}.$$

  - *Dirichlet-smoothed* version*:*

$$\widetilde{p}_x^{[\mu]}(w) \stackrel{def}{=} \frac{\mathrm{tf}(w \in x) + \mu \cdot \widetilde{p}_C^{MLE}(w)}{\sum_{w'} \mathrm{tf}(w' \in x) + \mu};$$

Structural Re-ranking

- Generation Probabilities (cont.)

$$p_x^{MLE}(w_1 w_2 \cdots w_n) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_x^{MLE}(w_j);$$

$$p_x^{[\mu]}(w_1 w_2 \cdots w_n) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_x^{[\mu]}(w_j).$$

  - Kullback-Leibler divergence *D*

$$p_d^{KL,\mu}(s) \stackrel{def}{=} \exp\left(-D\left(\widetilde{p}_s^{MLE}(\cdot) \,\middle\|\, \widetilde{p}_d^{[\mu]}(\cdot)\right)\right). \qquad (4)$$

$$p_d^{KL,\mu}(s) = \underbrace{(p_d^{[\mu]}(s))^{\frac{1}{|s|}}}_{\text{term A}} \cdot \underbrace{\exp(H(\widetilde{p}_s^{MLE}(\cdot)))}_{\text{term B}},$$

Evaluation

	AP89			AP			WSJ			TREC8		
	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR
upper bound	63.7	53.1	75.5	87.6	78.8	93.0	89.6	80.0	100.0	94.4	85.0	98.0
init. ranking	28.3	26.5	52.3	45.7	43.2	59.6	54.8	48.4	76.2	50.0	45.6	69.1
opt. baselines	30.0	27.4	**54.3**	46.5	43.9	63.5	56.0	49.4	77.2	51.2	46.4	**69.6**
U-In	29.6	*27.8*	39.5 $_{o}$	*50.9*	*49.0* $^{*}_{o}$	**66.3**	50.0	46.6	66.7	50.0	45.0	62.0
W-In	*31.3*	*29.6*	46.8	*51.3*	*48.7* *	*64.4*	52.0	47.8	63.3 $_{o}$	49.2	43.4	63.7
U-In+LM	**33.5**	27.0	46.5	*51.3* *	**49.4** $^{*}_{o}$	63.2	*56.4*	49.2	73.6	*52.8*	**52.0** $^{*}_{o}$	66.6
W-In+LM	*31.7*	*27.6*	48.4	*51.1* *	*48.4* $^{*}_{o}$	63.0	*57.2*	*50.0*	77.2	*51.6*	*49.6* *	64.5
R-U-In	*31.3*	*28.9*	46.4	*51.5*	*48.9* *	63.4	53.6	*49.6*	68.5	*52.0*	44.6	66.5
R-W-In	*32.2*	*29.6*	40.5 $_{o}$	*52.1* *	*49.1* $^{*}_{o}$	*63.9*	54.0	49.2	70.2	*52.4*	44.6	66.5
R-U-In+LM	*33.0*	*29.3*	45.8	*52.1* $^{*}_{o}$	*49.2* $^{*}_{o}$	*64.3*	**58.8** *	**51.0** *	**78.6**	*55.6*	46.0	68.4
R-W-In+LM	**33.5**	**29.8**	46.0	**52.9** $^{*}_{o}$	*49.0* $^{*}_{o}$	62.6	**58.8** *	*50.6*	**78.6**	**56.0**	45.8	67.6

7 5 10 7 29/48

8 8 0 8 8 8 4 6 3 7 4 0 64/96

U-In+LM > U-In = 10

W-In+Lm > W-In = 8

R-U-In+LM > R-U-In = 11

R-W-In+LM > R-W-In =10

39/48

Evaluation

- *Links based on the vector-space model*

		U-In	W-In	U-In+LM	W-In+LM	R-U-In	R-W-In	R-U-In+LM	R-W-In+LM
AP89	prec @5	□	□				□		
	prec @10		♦						♦
	MRR	□	□	□		□	□	□	
AP	prec @5	♦	♦	♦	♦	♦	♦	♦	♦
	prec @10	♦	♦	♦	♦	♦	♦	♦	♦
	MRR	♦	♦		♦		♦	♦	
WSJ	prec @5						♦		
	prec @10					♦	♦		
	MRR		□						
TREC8	prec @5	♦	♦			♦	♦	♦	♦
	prec @10	♦	♦		♦		♦		
	MRR	□			□				

Evaluation

- Non-structural re-ranking

	AP89			AP			WSJ			TREC8		
	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR
uniform (= init)	28.3	26.5	52.3	45.7	43.2	59.6	54.8	48.4	76.2	50.0	45.6	69.1
W-ln	*31.7*	*27.6*	48.4	*51.1∗*	*48.4∗*	**63.0**	*57.2*	*50.0*	*77.2*	*51.6*	**49.6∗**	64.5
R-W-ln	**33.5**	**29.8**	46.0	**52.9∗**	**49.0∗**	*62.6*	**58.8∗**	**50.6**	**78.6**	**56.0**	*45.8*	67.6
length	*29.1*	24.3	50.8	41.6	41.4	55.3	44.4∗	42.4∗	64.6∗	47.2	41.4	64.2
log(length)	*30.4*	*27.0*	*52.5*	45.3	43.2	*60.6*	*57.2*	*49.0*	69.8∗	49.6	*46.8*	*69.2*
entropy	*30.0*	26.5	**52.6**	*46.1*	42.5	*60.8*	*56.8*	*48.6*	71.1∗	49.6	*46.8*	**71.7∗**
uniqTerms	27.4	24.8	52.3	42.0	41.3	56.2	50.0	44.6	68.8	49.2	44.2	*71.2*
log(uniqTerms)	*30.4*	*27.0*	*52.5*	*45.9*	42.3	*60.8*	*57.2*	*49.0*	70.0∗	49.6	*47.2*	*70.0*

Questions and Comments

Thank You !