Learning to Diversify Search Results via Subtopic Attention

Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou,
Wayne Xin Zhao, Jian-Yun Nie, Ming Yue

Renmin University of China

rucjzb@163.com

Aug 10, 2017

Outline

1. Backgrounds
   - Problem
   - Related Work
2. Our Approach
   - Framework
   - Implementation
   - Training & Prediction
3. Experiment
   - Experiment Settings
   - Experimental Results
   - Case Study
4. Conclusion

Problem

Search **R**esult **D**iversification (SRD): aims to retrieve diverse results to satisfy as many different intents as possible.

"Java"

"Apple"

Related Work

	implicit approaches doc should be different (novel) form former ones → doc similarity	explicit approaches doc should cover intents (subtopics) unsatisfied by former ones → varied subtopic weights
unsupervised	MMR	IA-Select, xQuAD, PM2, TxQuAD, TPM2, HxQuAD, HPM2, 0-1 MSKP
supervised	SVM-DIV, R-LTR, PAMM, NTN	★

- explicit approaches > implicit approaches
  - using subtopic is a more natural way to handle diversity
  - document similarity (based on "word") $\neq$ document intent similarity
- supervised > unsupervised
  - avoid heuristic and handcrafted functions and parameters
  - directly optimizing objective function

★ modeling subtopics using supervised learning for SRD

Notations Definition

Table: Notations.

Notation	Definition
d_t, q, i_k, r	the t-th document, query, k-th subtopic, a ranking
$\boldsymbol{v}_{d_t}$	representation of the document at the t-th position.
$\boldsymbol{v}_q$, $\boldsymbol{v}_{i_k}$	representation of the query and the k-th subtopic.

Motivation

unsupervised explicit methods can be formulated as:

$$\begin{aligned}
&S_{\text{unsupervised}}(q, d_t, \mathcal{C}_{t-1}, \mathcal{I}_q) = \\
&(1-\lambda)S^{\text{rel}}(d_t, q) + && \Rightarrow \text{relevance} \\
&\lambda \sum_{i_k \in \mathcal{I}_q} S^{\text{div}}(d_t, i_k) \underbrace{A(\mathcal{C}_{t-1}, \mathcal{I}_q)_k}_{\text{subtopic weights}}, && \Rightarrow \text{diversity}
\end{aligned} \tag{1}$$

- xQuAD: $A(\mathcal{C}_{t-1}, \mathcal{I}_q)_k = P(i_k|q) \prod_{d_j \in \mathcal{C}_{t-1}} (1 - P(d_j|i_k))$
- PM2: mimics seats allocation of competing political parties

drawbacks:

- don't model the selected documents as a sequence
- the functions and parameters are heuristically defined

unsupervised explicit methods can be formulated as:

$$S_{\text{unsupervised}}(q, d_t, \mathcal{C}_{t-1}, \mathcal{I}_q) =$$
$$(1-\lambda)S^{\text{rel}}(d_t, q) + \qquad \Rightarrow \text{relevance}$$
$$\lambda \sum_{i_k \in \mathcal{I}_q} S^{\text{div}}(d_t, i_k) \underbrace{A(\mathcal{C}_{t-1}, \mathcal{I}_q)_k}_{\text{subtopic weights}}, \qquad \Rightarrow \text{diversity} \tag{2}$$

Document Sequence with Subtopic Attention Framework (DSSA):

$$S_{\text{DSSA}}(q, d_t, \mathcal{C}_{t-1}, \mathcal{I}_q) = s_{d_t} =$$
$$(1-\lambda)\mathcal{S}^{\text{rel}}(\mathbf{v}_{d_t}, \mathbf{v}_q) + \qquad \Rightarrow \text{relevance}$$
$$\lambda \mathcal{S}^{\text{div}}\left(\mathbf{v}_{d_t}, \mathbf{v}_{i_{(\cdot)}}, \underbrace{\mathcal{A}\left(\mathcal{H}([\mathbf{v}_{d_1}, \ldots, \mathbf{v}_{d_{t-1}}]), \mathbf{v}_{i_{(\cdot)}}\right)}_{\text{subtopic attention}}\right), \qquad \Rightarrow \text{diversity} \tag{3}$$

3 components

- document sequence representation: $\mathbf{h}_{t-1} = \mathcal{H}([\mathbf{v}_{d_1}, \ldots, \mathbf{v}_{d_{t-1}}])$
- subtopic attention: (similar to attention mechanism in NMT) $\mathbf{a}_{t,(\cdot)} = \mathcal{A}(\mathbf{h}_{t-1}, \mathbf{v}_{i_{(\cdot)}})$
- scoring: $\mathcal{S}^{\text{rel}}(\mathbf{v}_{d_t}, \mathbf{v}_q)$ and $\mathcal{S}^{\text{div}}(\mathbf{v}_{d_t}, \mathbf{v}_{i_{(\cdot)}}, \mathbf{a}_{t,(\cdot)})$

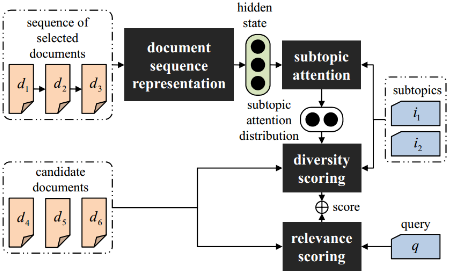


Figure: Illustration of DSSA framework.

- document representation $\boldsymbol{v}_{d_t}$
  - distributed representations (SVD, LDA, and doc2vec): $\boldsymbol{e}_{d_t} \in \mathbb{R}^{E_d}$
  - relevance features (BM25, TF-IDF): $\boldsymbol{x}_{d_t,q}$ and $\boldsymbol{x}_{d_t,i_k} \in \mathbb{R}^{R}$

  total size: $E_d + R + KR$
- query and subtopic representation $\boldsymbol{v}_q$, $\boldsymbol{v}_{i_k}$
  - distributed representations based on retrieved top Z documents: $\boldsymbol{e}_q$ and $\boldsymbol{e}_{i_k} \in \mathbb{R}^{E_q}$

  total size: E_q

- document sequence representation
  using RNN to encode document sequence (only based on document distributed representation).

$$\boldsymbol{h}_t = \tanh(\boldsymbol{W}^n[\boldsymbol{h}_{t-1}; \boldsymbol{e}_{d_t}] + \boldsymbol{b}^n), \tag{4}$$

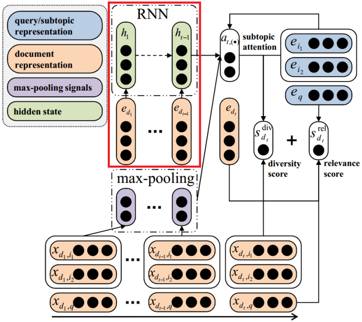


Figure: Implementation of DSSA using RNN and max-pooling.

- subtopic attention

  1. hidden state: the more similar $\boldsymbol{h}_{t-1}$ and $\boldsymbol{e}_{i_k}$ are, the less attention is will acquire.

$$\mathcal{A}^{'}(\boldsymbol{h}_{t-1}, \boldsymbol{e}_{i_k}) = \begin{cases} \boldsymbol{h}_{t-1}^{\mathsf{T}} \boldsymbol{W}^{a} \boldsymbol{e}_{i_k}, & \text{(general)} \\ -\boldsymbol{h}_{t-1}^{\mathsf{T}} \cdot \boldsymbol{e}_{i_k}, & \text{(dot)} \end{cases} \tag{5}$$

  2. relevance feature: use max-pooling to select the most significant signal.

$$\mathcal{A}^{''}(\boldsymbol{x}_{d_1,i_k}, \dots, \boldsymbol{x}_{d_{t-1},i_k}) = \max([\boldsymbol{x}_{d_1,i_k}^{\mathsf{T}} \cdot \boldsymbol{w}^{p}, \dots, \boldsymbol{x}_{d_{t-1},i_k}^{\mathsf{T}} \cdot \boldsymbol{w}^{p}]), \tag{6}$$

  3. softmax: obtain probability distribution.

$$a_{t,k}^{'} = \mathcal{A}^{'}(\boldsymbol{h}_{t-1}, \boldsymbol{e}_{i_k}) + \mathcal{A}^{''}(\boldsymbol{x}_{d_1,i_k}, \dots, \boldsymbol{x}_{d_{t-1},i_k}),$$
$$a_{t,k} = \frac{w_{i_k} \exp(a_{t,k}^{'})}{\sum_{j=1}^{K} w_{i_j} \exp(a_{t,j}^{'})} \quad (w_{i_j} \geq 0, \forall j). \tag{7}$$

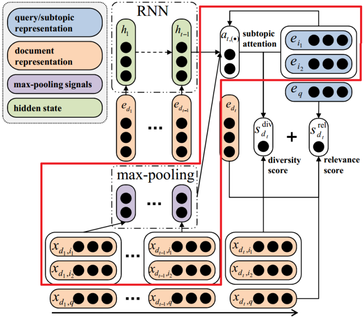


Figure: Implementation of DSSA using RNN and max-pooling.

- scoring
  both relevance and diversity score are calculated through distributed representations and relevance features.

$$s_{d_t} = (1-\lambda)s_{d_t}^{\text{rel}} + \lambda s_{d_t}^{\text{div}} \quad (0 \leq \lambda \leq 1), \tag{8}$$

  1. relevance score: calculate relevance to the query.

$$s_{d_t}^{\text{rel}} = \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_q) + \mathbf{x}_{d_t,q}^{\mathsf{T}} \cdot \mathbf{w}^r, \tag{9}$$

  2. diversity score: calculate relevance to all the subtopics and combine them using attention.

$$s_{d_t}^{\text{div}} = \mathbf{a}_{t,(\cdot)}^{\mathsf{T}} \cdot \begin{bmatrix} \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_1}) + \mathbf{x}_{d_t,i_1}^{\mathsf{T}} \cdot \mathbf{w}^r \\ \vdots \\ \mathcal{S}'(\mathbf{e}_{d_t}, \mathbf{e}_{i_K}) + \mathbf{x}_{d_t,i_K}^{\mathsf{T}} \cdot \mathbf{w}^r \end{bmatrix}, \tag{10}$$

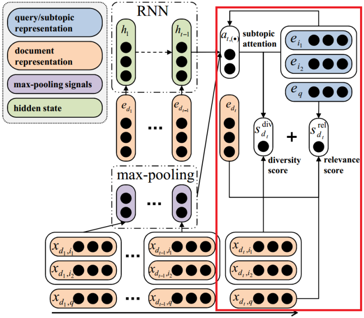


Figure: Implementation of DSSA using RNN and max-pooling.

Training

Given query q, subtopics $i_1, ..., i_K$, sequence of selected documents $d_1, ..., d_{t-1}$, we can calculate the score of a candidate document s_{d_t} using DSSA. How to train these parameters?

- existing optimization methods
  - MLE: maximize the probability of best rankings.
    Plackett-Luce model:
    $$P(r) = \prod_{i=1}^{|r|} \frac{\exp(s_{d_i}^{r[:i-1]})}{\sum_{j=i}^{|r|} \exp(s_{d_j}^{r[:i-1]})}. \tag{11}$$
  - PAMM: enlarges the probability margin between positive and negative rankings.
- drawbacks
  - MLE: the number of best rankings is small if we only have hundreds of queries.
  - PAMM: which positive and negative rankings should we sample?

- list-pairwise
  Sample pairs of rankings that vary only at the last position and treat it as binary claasification problem.
  - MLE: list-pairwise can generate abundant training samples.
  - PAMM: list-pairwise corresponds better to the decision-making situation in which we have to choose a document under a given context.

Table: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	√	√	×
d_2	√	√	×
d_3	×	×	√
d_4	×	√	×

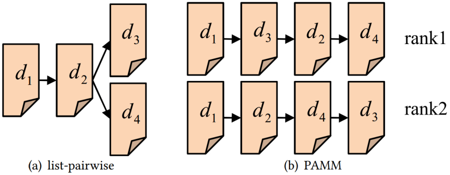


Figure: Pair sample examples of (a) list-pairwise and (b) PAMM.

For each query, we sequentially and greedily choose the document with the highest score and append it to the ranking list.

d_1, d_2, d_3, d_4

d_1

Table: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	$\surd$	$\surd$	$\times$
d_2	$\surd$	$\surd$	$\times$
d_3	$\times$	$\times$	$\surd$
d_4	$\times$	$\surd$	$\times$

Prediction

For each query, we sequentially and greedily choose the document with the highest score and append it to the ranking list.

$d_1, \overset{d_2, d_3, d_4}{d_3}$

Table: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	$\surd$	$\surd$	$\times$
d_2	$\surd$	$\surd$	$\times$
d_3	$\times$	$\times$	$\surd$
d_4	$\times$	$\surd$	$\times$

For each query, we sequentially and greedily choose the document with the highest score and append it to the ranking list.

$d_1, d_3, \overset{d_2, d_4}{d_2}$

Table: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	$\surd$	$\surd$	$\times$
d_2	$\surd$	$\surd$	$\times$
d_3	$\times$	$\times$	$\surd$
d_4	$\times$	$\surd$	$\times$

Prediction

For each query, we sequentially and greedily choose the document with the highest score and append it to the ranking list.

$d_1, d_3, d_3, \overset{d_4}{d_4}$

Table: Subtopic relevance example.

doc\subtopic	i_1	i_2	i_3
d_1	$\surd$	$\surd$	$\times$
d_2	$\surd$	$\surd$	$\times$
d_3	$\times$	$\times$	$\surd$
d_4	$\times$	$\surd$	$\times$

Experiment Settings

- Dataset
  Web Track of TREC 2009-2012, totally 198 queries with binary relevance judgement at subtopic level.
- Evaluation metrics
  ERR-IA, α-nDCG, NRBP, D$\sharp$-measures, Precision-IA, and Subtopic Recall.
- Baselines
  - unsupervised: Lemur, xQuAD, PM2, TxQuAD, TPM2, HxQuAD, HPM2
  - supervised: ListMLE, R-LTR, PAMM, NTN
- Setting
  top 20 of Lemur for training; top 50 of Lemur for prediction. t-test p-value is 0.05.

Overall Results

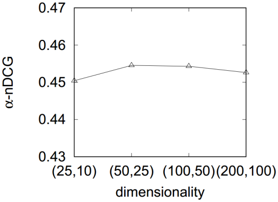
Table: Performance comparison of all methods. The best result is in bold. Statistically significant differences between DSSA and baselines are marked with various symbols. ★ indicates significant improvement over all baselines.

Methods	ERR-IA	α-nDCG	NRBP	D$\sharp$-nDCG	Pre-IA	S-rec
Lemur①	.271	.369	.232	.424	.153	.621
ListMLE❶	.287	.387	.249	.430	.157	.619
xQuAD②	.317	.413	.284	.437	.161	.622
TxQuAD③	.308	.410	.272	.441	.155	.634
HxQuAD④	.326	.421	.294	.441	.158	.629
PM2⑤	.306	.411	.267	.450	.169	.643
TPM2⑥	.291	.399	.250	.443	.161	.639
HPM2⑦	.317	.420	.279	.455	.172	.645
R-LTR❷	.303	.403	.267	.441	.164	.631
PAMM❸	.309	.411	.271	.450	.168	.643
R-LTR-NTN❹	.312	.415	.275	.451	.166	.644
PAMM-NTN❺	.311	.417	.272	.457	.170	.648
DSSA	**.356**★	**.456**★	**.326**★	**.473**★	**.185**★	**.649**①②❶

Effects of Different Settings

Table: Effects of different settings.

Methods	ERR-IA	α-nDCG	NRBP	D$\sharp$-nDCG	Pre-IA	S-rec
SVD	.348	.450	.315	.470	.184	.646
LDA	.356	.456	.326	.473	.185	.649
doc2vec	.351	.452	.318	.471	.184	.646
general	.356	.456	.326	.473	.185	.649
dot	.347	.450	.314	.470	.184	.647
vanilla	.354	.454	.322	.471	.184	.649
GRU	.357	.457	.326	.473	.185	.649
LSTM	.356	.456	.326	.473	.185	.649
DSSA-RNN	.342	.445	.306	.466	.172	.657
DSSA-RNNMP	.356	.456	.326	.473	.185	.649



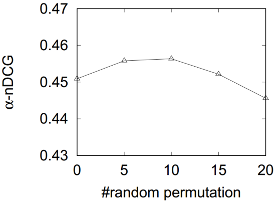


Figure: Performance tendency of different settings.

Case Study

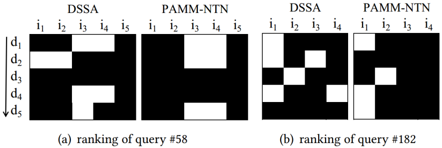


Figure: Case study for DSSA and PAMM-NTN. White means relevant and black means irrelevant.

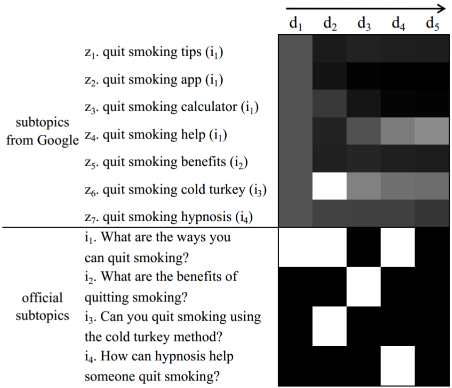


Figure: Subtopic attention variation of query #182. The top part is attention and the bottom part is relevance judgment.

Conclusion

- Contribution
  - first supervised search result diversification framework to model subtopics explicitly.
  - an implementation using RNN and max-pooling with attention mechanism which fits our task well.
- Furture work
  - more complex implementations on larger datasets.
  - learn representations simultaneously.

data and code available at:
`http://www.playbigdata.com/dou/DSSA/` or
`https://github.com/jzbjyb/DSSA`.

Question