THUIR at NTCIR-9 INTENT Task

IR Group of Tsinghua University

Yufei Xue, Fei Chen, Tong Zhu, Chao Wang, Zhichao Li,
Yiqun Liu, Min Zhang, Yijiang Jin, Shaoping Ma

Outline

- Overview
- Subtopic Mining
  - Extracting Subtopics from Web Resrouces
  - Mining Subtopics from Clickthrough Data
  - Re-ranking Based on Clicked Titles and Snippets
  - Removing reduplicate subtopics
- Document Ranking
  - Retrieval Models
  - Result Re-ranking with HITS
  - Documents Duplication Elimination
  - Novelty-Result Selection algorithm
  - D#-nDCG-based Selection algorithm

Overview

- THUIR’s first experience of NTCIR
- Participate in INTENT task
  - Both Subtopic Mining and Document Ranking
  - Focus on Chinese topics only
- Methods
  - Minig subtopics from different resources (search engines, Wikipedia, Clickthrough data)
  - Diversifying search results with different algorithms (traditional and D#-measure oriented diversifying methods)

Subtopic Mining

- 5 runs submitted

Runs	Data
THU-S-C-1	Web resources: Query recommendations of search engines & Wikipedia items
THU-S-C-2	
THU-S-C-3	
THU-S-C-4	Clickthrough data: SogouQ
THU-S-C-5	Clickthrough data: Sogou web search log in about 1 year.

Minig subtopics from search engines

- Commercial search engines usually suggest related search queries in SERPs.
- Most of the suggested queries are specializations of the previous query.

Minig subtopics from search engines

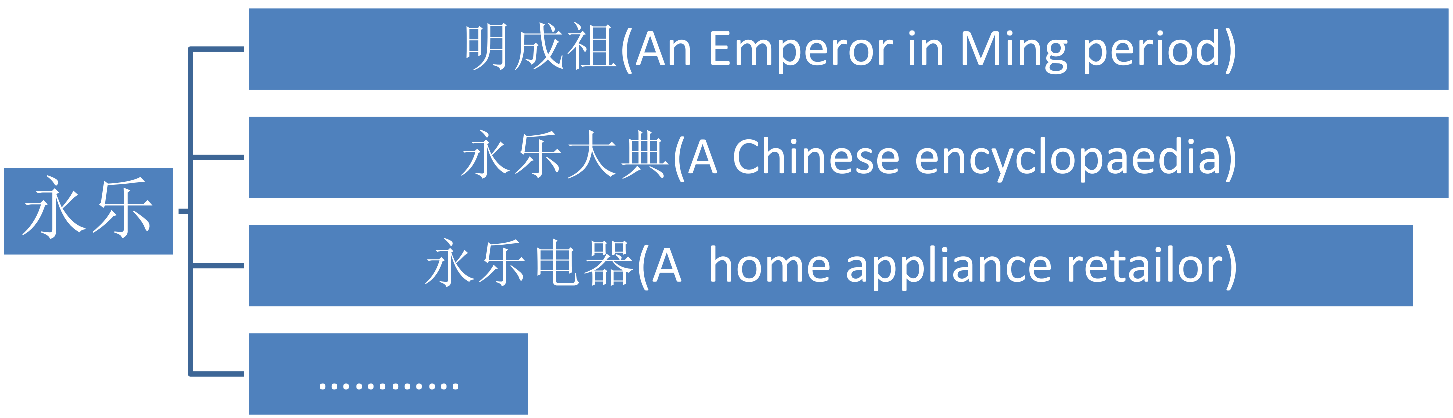
- Crawl the suggested related queries of each topic in different search engines.
- Use the search engines to vote for all related queries and get a ranked subtopic list.

Table 1: The search engines and their weights

Search Engine	Weight
Google	1
Baidu	1
Bing	1
Sogou	1
Soso	0.5
Youdao	0.5

Mining subtopics from Wikipedia

- Disambiguation pages in Wikipedia



Mining subtopics from Wikipedia

- Besides disambiguation pages, we can get more subtopics from the items in Wikipedia.
- We extract all the items with an INTENT subtopic as its substring.
- For the topic “巧克力”(Chocolate), there are items
  - 白巧克力(White chocolate)
  - 热巧克力(Hot chocolate)
  - 巧克力棒(Chocolate candy bar)
  - ……

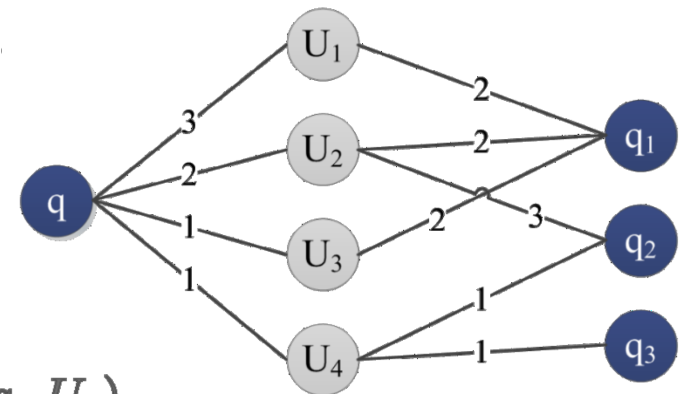
Mining subtopics from Wikipedia

- Make Wikipedia join the vote of search engines.

Resource		Weight of vote
Query recommendation of	Google	1
	Bing	1
	Baidu	1
	Sogou	1
	Youdao	0.5
	Soso	0.5
Wikipedia Item from	Disambiguition Page	0.9
	Other	0.4

Mining subtopics from clickthrough data

- The queries and clicked URLs are often presented in a bipartite gr



- For *q* and $q_{i\ ,}$ define

$$Score(q, q_i) = \sum_j \frac{W(q, U_j)}{\sum_k W(q, U_k)} \times \frac{W(q_i, U_j)}{\sum_k W(q_i, U_k)}$$

- Score(q, qi) is the probability that user clicks the same URL when searching different query *q* and q_i.
- Show the relevance of two queries.

Mining subtopics from clickthrough data

- To ensure q_i is a subtopic of given query, we filter out a query q_i if:
  - q is a substring of q_i, or
  - q_i has no common items with q

Submitted Runs

Runs	Data	Postprocessing	D#-nDCG@10
THU-S-C-1	Query recommendations & Wikipedia items	Removing duplicate ones	0.5921
THU-S-C-2		Removing duplicate ones; Re-ranking based on snippets	0.5993
THU-S-C-3			0.5967
THU-S-C-4	Clickthrough data: SogouQ		0.3347
THU-S-C-5	Clickthrough data: Sogou web search log in about 1 year.		0.3672

Re-ranking Based on Clicked Titles and Snippets

- Titles and snippets are the only contents that users can see before they click on search results.
- Clicked titles and snippets contain a lot of information about users' needs and intents.
- With the clicked titles and snippets, we try to find which subtopics are more important.

Re-ranking Based on Clicked Titles and Snippets

- Step 1:
  For a topic, analyze all the clickthrough data of the topic and gather the title and snippet texts of each click into a “snippet document”

[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因 实力和计划 战争过程 战争的结局及对影响 [/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因 实力和计划 战争过程 战争的结局及对影响 [/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因 实力和计划 战争过程 战争的结局及对影响 [/snippet]

[title]日俄战争在线观看-搜狗视频[/title]
[snippet]在搜狗视频中约有848个相关结果。在线观看 日俄战争 sohu.com 日俄战争 youku.com 4.日俄战争 56.com 4.日俄战争 56.com 4.日俄战争 56.com ④日俄战争 youku.co...[/snippet]

[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]
[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]

Re-ranking Based on Clicked Titles and Snippets

- Step 2: Get term frequencies of snippet document.

[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]

[title]日俄战争在线观看-搜狗视频[/title]
[snippet]在搜狗视频中约有848个相关结果。在线观看 日俄战争 sohu.com 日俄战争 youku.com 4.日俄战争 56.com 4.日俄战争 56.com 4.日俄战争 56.com ④日俄战争 youku.co...[/snippet]

[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]
[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]

→

起因	6
过程	3
结局	3
影响	3
资料	2
图片	2
…	…

Re-ranking Based on Clicked Titles and Snippets

- Step 3: Assign a weight to each term according to the rank.

[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]
[title]日俄战争_百度百科[/title]
[snippet]《日俄战争》是指日本与沙皇俄国为了侵占中国东北和朝鲜，在中国东北的土地上进行了一场帝国主义战争。以沙皇俄国的失败而告终。 战争起因战争概述 19世纪末，沙皇俄...
战争起因　　实力和计划　　战争过程　　战争的结局及对影响　　[/snippet]

[title]日俄战争在线观看-搜狗视频[/title]
[snippet]在搜狗视频中约有848个相关结果。在线观看 日俄战争 sohu.com 日俄战争 youku.com 4.日俄战争 56.com 4.日俄战争 56.com 4.日俄战争 56.com ④日俄战争 youku.co...[/snippet]

[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]
[title]日俄战争 - 铁血网[/title]
[snippet]铁血网为您提供最新的日俄战争资料，同时包括最新的日俄战争介绍、日俄战争图片，以及日俄战争新闻等有关日俄战争的最新最全的信息。[/snippet]

→

起因	6
过程	3
结局	3
影响	3
资料	2
图片	2
…	…

→

Rank	Term	Freq.	Weight
1	起因	6	1
2	过程	3	0.98
2	结局	3	0.98
2	影响	3	0.98
5	资料	2	0.92
…			

Re-ranking Based on Clicked Titles and Snippets

- Step 4: Look back into the subtopic list. If any term in the ranked term list appears in a subtopic, we add the weight of the term to the score of the subtopic.

- Step 5: Re-rank the subtopics by the updated score.

Other Approaches

- Removing reduplicate subtopics
  - Analyze the clickthrough data of the subtopics.
    If 2 subtopics have more than 5 common clicked URLs, they are considered as reduplicate ones.
  - Merge reduplicate subtopics together, keep the one with higher rank in the list.
- Specificially, we try to recognize the topics with four kinds of common intents:
  - Online Music
  - Online Video
  - Online Novel
  - Encyclopedia.

Submitted Runs

Runs	Data	Postprocessing	D#-nDCG@10
THU-S-C-1	Query recommendations & Wikipedia items	Removing duplicate ones	0.5921
THU-S-C-2		Removing duplicate ones; Re-ranking based on snippets	0.5993
THU-S-C-3			0.5967
THU-S-C-4	Clickthrough data: SogouQ		0.3347
THU-S-C-5	Clickthrough data: Sogou web search log in about 1 year		0.3672

Table 5. Chinese Subtopic Mining runs ranked by D♯-nDCG@10.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THU-S-C-2	0.4801	0.7186	0.5993
MSINT-S-C-2	0.5130	0.6806	0.5968
THU-S-C-3	0.4828	0.7107	0.5967
THU-S-C-1	0.4946	0.6896	0.5921
ICTIR-S-C-1	0.5161	0.6434	0.5797
uogTr-S-C-5	0.4947	0.6598	0.5772
MSINT-S-C-4	0.4864	0.6604	0.5734
ICTIR-S-C-4	0.5035	0.6417	0.5726
ICTIR-S-C-2	0.4826	0.6576	0.5701
HITIR-S-C-5	0.4936	0.6449	0.5693
ISCAS-S-C-1	0.5022	0.6336	0.5679
ICTIR-S-C-3	0.4808	0.6530	0.5669
HITIR-S-C-1	0.4854	0.6453	0.5653
ISCAS-S-C-3	0.4910	0.6386	0.5648
MSINT-S-C-1	0.5002	0.6240	0.5621
NTU-S-C-2	0.4683	0.6546	0.5615
MSINT-S-C-5	0.4578	0.6543	0.5560
NTU-S-C-3	0.4807	0.6308	0.5558
HITIR-S-C-4	0.4738	0.6291	0.5514
HITIR-S-C-3	0.4738	0.6291	0.5514
HIT2jointNLPLab -S-C-2	0.4596	0.6407	0.5501

Document Ranking

- Retrieval Models
- Result Re-ranking with HITS
- Documents Duplication Elimination
- Novelty-Result Selection algorithm
- D#-nDCG-based Selection algorithm

Retrieval Models

- Improved Probabilistic Model

$$R(Q,D) = W_{BM25} + \alpha_1 \cdot W_{wp}$$

$$W_{BM25} = \sum_{i=1}^{m} log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

$$W_{wp} = \sum_{i=1}^{m} log \frac{N - n(q_i q_{i+1}) + 0.5}{n(q_i q_{i+1}) + 0.5} \cdot \frac{f(q_i q_{i+1}, D) \cdot (k_1 + 1)}{f(q_i q_{i+1}, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

N is the total number of documents, $n(q)$ is the number of documents contain q, k_1 and b are experimental parameters of BM25 ranking, $|D|$ is the length of document D, $avgdl$ is the average document length, $f\ (q, D)$ is the term frequency of q in D.

Result Diversification

- Result Re-ranking with HITS
  - Top m documents sorted by either Authority or Hub Value are placed up to the front.

$$R_{new} = R_{old} - R_{old} \times (Authority + Hub)$$

where R_{new} stands for the new rank of the document, and R_{old} is the old one.

Result Diversification

- Documents Duplication Elimination
  - Calculate the cosine similarity between the current document and the documents before, respectively.
  - If the similarity is greater than the threshold: 0.95, list it in the end.
  - This method is based on HITS.

Novelty-Result Selection Algorithm

- Novelty-Result Selection Algorithm
  - Novelty function:

$$f(d_i, S) = \frac{|S|}{\sum_{d_j \in S} \alpha_j \cdot \frac{1}{|\omega_i, \omega_j|}}$$

the weight parameter $\alpha_j = \frac{1}{original\, rank\ of\ d_j}$

Novelty-Result Selection algorithm

- *Set* $S = \{d_0\}$
- *While(* $\left\| \omega_i \right\| > 0$ *&&* $|S| < k$ *)*
-     $\omega_i = argmax\ f(d_i, S)$
-     *Add* d_i *to the end of S*
- *End while*
- *For i from 1 to n*
- *If* d_i *are not in S*
-     *Add* d_i *to the end of S*
- *End for*

D#-nDCG-based Selection Algorithm

- Definition
  - Intent probability:

$$p(i|q) = \frac{w_i}{\sum_i w_i}$$

Where w_i standards for the weight of the ith intent.

  - Document gain:

$$g_i(d) = \begin{cases} 5, & r_d \epsilon [1,5] \\ 4, & r_d \epsilon [6,20] \\ 3, & r_d \epsilon [21,50] \\ 2, & r_d \epsilon [51,100] \\ 1, & r_d \epsilon [101,1000] \end{cases}$$

r_d is the original rank of the document. $g_i(d)$ is the gain of document d under intent i.

D#-nDCG-based Selection algorithm

- Given q, I, D, S
- if $|I| > 3$ Then
- for every $d\ in\ D$ *do*
- $GG(d) = \sum_i Pr(i|q)\ g_i(d)$
- $C_i(d) = p(i|q) \cdot \sum_{k=1}^{r} g_i(d)$
- end for
- while $|S| < 1000$ do
- for every $d\ in\ D$ do
- $IA - O(d) = \sum_i g_i(d) \cdot (1-\alpha)^{C_i(r-1)}$
- $D\#Value(d) = \gamma IA - O(d) + (1-\gamma)GG(d)$
- Add $max\{D\#Value(d)\}$ to S, then delete it in D.
- end for
- end while
- return S
- else
- return D
- Where I is the intents collection of q. D is the searching result of q, S is the re-ranked list. $IA - O(d)$ stands for the recall how documents in S cover the intents I.

Experiment Results

Run	Descriptions
THUIR-D-C-1	Documents Duplication Elimination.
THUIR-D-C-2	Novelty-Result Selection algorithm.
THUIR-D-C-3	$D\# - nDCG$-based Selection algorithm.
THUIR-D-C-4	$D\# - nDCG$-based selection + user search log.
THUIR-D-C-5	Result Re-ranking with HITS.
Baseline	The original search result.

	I-rec@10	D-nDCG@10	D#-nDCG@10
THUIR-D-C-1	0.6893	0.4542	0.5717
THUIR-D-C-2	0.6495	0.3853	0.5174
THUIR-D-C-3	0.5979	0.2598	0.4288
THUIR-D-C-4	0.6001	0.2569	0.4285
THUIR-D-C-5	0.6861	0.4573	0.5717
Baseline	0.5157	0.2967	0.4062

Submitted results

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
THUIR-D-C-5	0.6861	0.4573	0.5717
THUIR-D-C-1	0.6893	0.4542	0.5717
uogTr-D-C-5	0.6624	0.4374	0.5499
MSINT-D-C-1	0.7068	0.3854	0.5461
uogTr-D-C-2	0.6600	0.4316	0.5458
MSINT-D-C-4	0.7091	0.3822	0.5456
uogTr-D-C-4	0.6474	0.4423	0.5449
MSINT-D-C-2	0.7003	0.3783	0.5393
uogTr-D-C-3	0.6301	0.4480	0.5390
MSINT-D-C-5	0.6936	0.3783	0.5359
uogTr-D-C-1	0.6406	0.4252	0.5329
THUIR-D-C-2	0.6495	0.3853	0.5174
HIT2jointNLPLab -D-C-2	0.5794	0.3704	0.4749
NTU-D-C-1	0.6180	0.3314	0.4747
SJTUBCMI-D-C-2	0.6008	0.3317	0.4663
MSINT-D-C-3	0.5987	0.3222	0.4604
SJTUBCMI-D-C-3	0.5856	0.3288	0.4572
SJTUBCMI-D-C-5	0.6228	0.2816	0.4522
SJTUBCMI-D-C-4	0.6108	0.2756	0.4432
SJTUBCMI-D-C-1	0.6038	0.2654	0.4346
THUIR-D-C-3	0.5979	0.2598	0.4288
THUIR-D-C-4	0.6001	0.2569	0.4285
HIT2jointNLPLab -D-C-1	0.4716	0.3573	0.4144
III_CYUT_NTHU -D-C-1	0.4630	0.2040	0.3335

THANK YOU! QUESTIONS?