HITSZ-ICRC at NTCIR-11 Temporalia Task

Temporal Information Retrieval

Yongshuai Hou, Cong Tan, Jun Xu, Youcheng Pan,
Qingcai Chen and Xiaolong Wang

Intelligent Computing Research Center
Key Laboratory of Network Oriented Intelligent Computation
Harbin Institute of Technology Shenzhen Graduate School

2014/12/11

Outline

- Introduction
- Temporal Information Retrieval
  – Candidate document retrieval
  – Temporal relevant judging
  – Search subtopic classification
  – Relevant document re-ranking
- Results Evaluation
- Summary

Introduction

- Who are we?
  - HITSZ-ICRC group: we are from Intelligent Computing Research Center (ICRC), Harbin Institute of Technology Shenzhen Graduate School (HITSZ), Shenzhen, China.
- Which task/subtask did we participate?
  - Temporal Query Intent Classification (TQIC)
  - Temporal Information Retrieval (TIR)

Introduction

- Temporal Information Retrieval (TIR)
  - For a given topic, participants retrieve relevant documents for different kinds of temporal subtopics
  - A given search topic contains ***title***, ***description***, ***query date***, and **4 different class subtopics** including class information (*atemporal, past, recency, future*)

Topic for TIR subtask

Title	Girl with the Dragon Tattoo
Description	I've recently watched a film called Girl with the Dragon Tattoo, and really liked it. Therefore, I would like to gather information about the movie.
Past question	How did the casting of the film develop?
Recency question	What did the recent reviews say about the film?
Future question	Is there any plan about its sequel?
Atemporal question	What are the names of main actors and actresses of the film?
Search date	<u>28 Feb 2013 GMT+0:00</u>

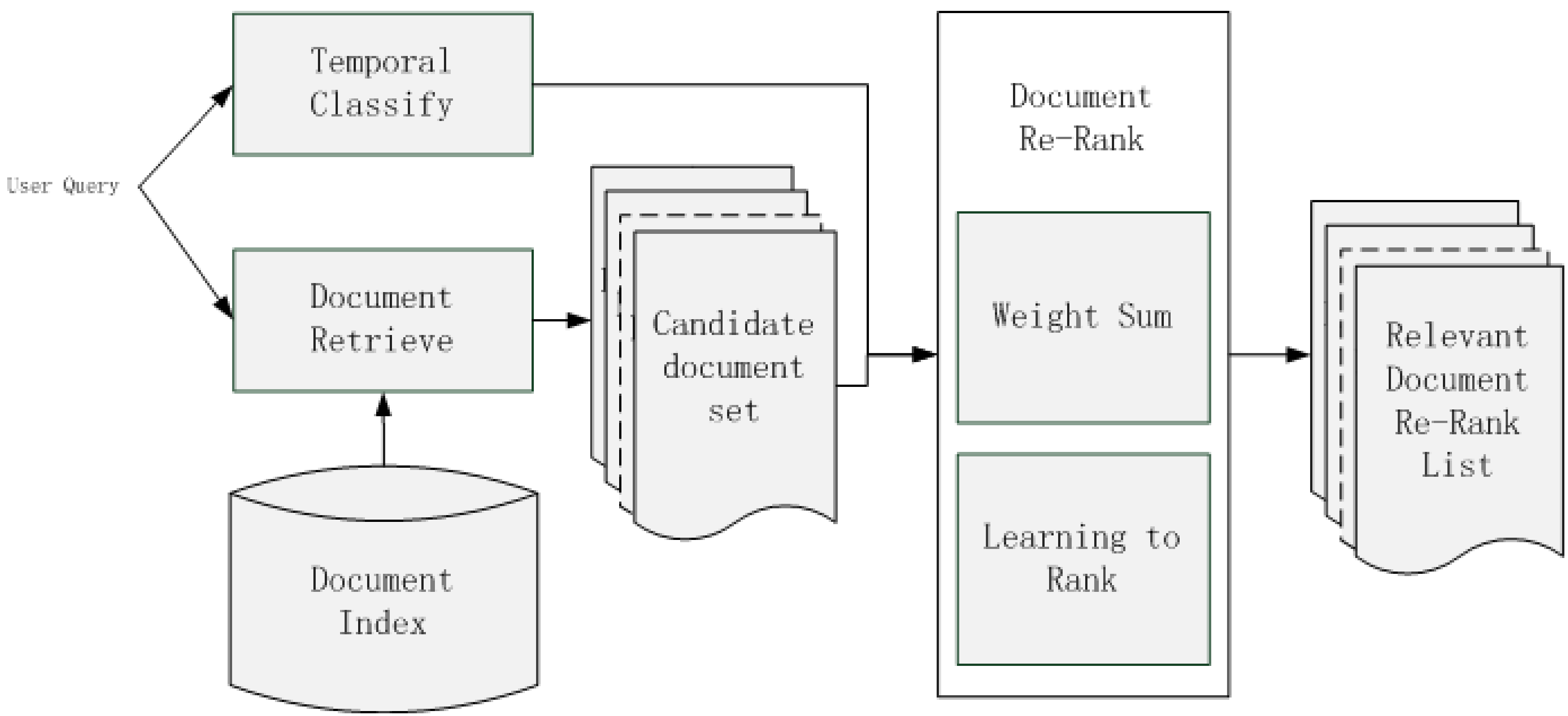
Document Collection

- Supplied “LivingKnowledge news and blogs annotated sub-collection” corpus, contains 3.8M documents from blogs and news sources.
- Available information of a document in Corpus
  - Document create time
  - Named entity tags in content
  - Time expression tags in content
  - Normalized value for time expression
  - …

Document Collection


```xml
<?xml version="1.0" encoding="UTF-8"?>
<doc id=20111004040101_5171>
<meta-info>
    <tag name="host">latimesblogs.latimes.com</tag>
    <tag name="date">2011-10-04</tag>
    <tag name="url">
    http://latimesblogs.latimes.com/the_big_picture/2011/09/the-new-oscar-rule-book-can-the-academy-really-curtail-awards-season-excess.html?utm_source=feedburner
    &amp;utm_medium=feed&amp;utm_campaign=Feed%3A+PatrickGoldstein+%28L.A.+Times+-+Patrick+Goldstein%29</tag>
    <tag name="sourcerss">http://feeds.latimes.com/PatrickGoldstein/</tag>
    <tag name="title">New Oscar rules: Can the Academy curtail awards season excess?</tag>
    <tag name="source-encoding">UTF-8</tag>
    <tag name="rsscategory">Patrick Goldstein</tag>
</meta-info>
<text><SE><E type="E:ORGANIZATION:CORPORATION">New Oscar</E> rules: Can the <E type="E:ORGANIZATION:GOVERNMENT">Academy</E> curtail awards <E type="T:DATE:DATE">
season</E> excess?</SE>
 <SE>The <E type="T:DATE:DATE">Oscar silly season</E> has officially begun.</SE>
 <SE>That's the only way to look at the new <E type="E:FAC:BUILDING">Motion Picture Academy</E> rules governing how <E type="E:ORG_DESC:CORPORATION">studios</E> and
 <E type="E:PER_DESC">filmmakers</E> can promote their movies during <E type="T:DATE:DATE">Oscar season</E>, a period that <E type="T:DATE:DATE">these days lasts</E>
 longer than <E type="T:DATE:DATE">winter</E> in <E type="E:GPE:CITY">Siberia</E>.</SE>
 <SE>Being a sports <E type="E:PER_DESC">fan</E>, <E type="E:ORGANIZATION:CORPORATION">I've</E> always thought that it was impossible for any <E type=
 "E:ORG_DESC:OTHER">organization</E> to have more arcane rules than the <E type="E:ORGANIZATION:OTHER">NCAA</E>, but the <E type="E:ORG_DESC:EDUCATIONAL">academy</E>
 has easily topped that <E type="E:PER_DESC">body</E>.</SE>
 <SE>Its new regulations are intended to stop <E type="E:ORGANIZATION:CORPORATION">Oscar-season</E> <E type="E:ORG_DESC:CORPORATION">excess</E>, but many believe
 they could easily lead to more over-the-top campaigning than ever.</SE>
 <SE>When it comes to excess, nothing can really top an <E type="E:PERSON">Oscar</E> <E type="E:ORG_DESC:CORPORATION">shindig</E> like the <E type="N:CARDINAL">one
 </E> <E type="E:ORGANIZATION:CORPORATION">Arianna Huffington</E> threw <T val="201102">last February</T> at her <E type="E:FAC_DESC:BUILDING">house</E> for <E type=
 "E:ORGANIZATION:CORPORATION">Harvey Weinstein's "The King's Speech</E>," which featured not just the A-list <E type="E:PER_DESC">cast</E> and <E type="E:PER_DESC">
 filmmakers</E> from the movie, but real <E type="E:NORP:NATIONALITY">British</E> <E type="E:PER_DESC">royalty</E>, notably <E type="E:PERSON">Earl Charles Spencer
 </E>, <E type="E:PER_DESC">brother</E> of the late <E type="E:PER_DESC">Princess</E> <E type="E:PERSON">Diana</E>.</SE>
 <SE>The <E type="E:ORG_DESC:POLITICAL">party</E> generated <E type="N:QUANTITY:WEIGHT">tons</E> of <E type="E:PER_DESC">press</E> and publicity, and was clearly
 designed to create buzz for the film, which ended up winning the <E type="E:PERSON">Oscar</E> for best picture.</SE>
 <SE>According to the new rules, a similar <E type="E:ORG_DESC:POLITICAL">party</E> <T val="2011">this year</T> could offer <E type="N:MONEY">just as much</E> pomp
 and circumstance, <E type="T:TIME">just as long</E> as it happened <E type="T:DATE:DATE">two weeks earlier</E>, before the nominations were announced.</SE>
 <SE>Because <E type="E:ORGANIZATION:CORPORATION">"The King's Speech"</E> was already the <E type="E:PER_DESC">favorite</E> to win best picture even before the
 nominations, it seems clear that the <E type="E:ORG_DESC:POLITICAL">party</E> would have had <E type="N:MONEY">just as</E> much impact if it had been held in <T val=
 "201101">mid-January</T> instead of <T val="201102">early February</T>.</SE>
```

TIR Process



User Query
Temporal Classify
Document Retrieve
Document Index
Candidate document set
Document Re-Rank
Weight Sum
Learning to Rank
Relevant Document Re-Rank List

Candidate Document Retrieval

- Create an index on the document collection and retrieve for each subtopic using Lucene toolkit with BM25 language model
- Save top 500 retrieval results in result list as candidate documents for each subtopic
- Save BM25 score for each candidate document as content relevant score ***CR***

Temporal Relevant Judging

- Judge temporal relevant between subtopic and candidate document
  - Calculate the date distance between search date and each time expression in document
  - Classify each time expression into class *past, recency or future*

$$dis_i = Dq - DX_i$$

$$C_i = \begin{cases} future, & if\, dis_i < 0 \\ past, & if\, dis_i > B_p \\ recency, & if\, 0 \le dis_i \le B_r \end{cases}$$

Where Dq is search date of the topic, DX_i is normalized time expression in document, B_p is the classification boundary for *past* class time expression, B_r is the classification boundary for *recency* class time expression. $\boldsymbol{B_p}$**=**$\boldsymbol{B_r}$**=300** (days) here.

  - Judge temporal relevant according to whether the document contains same class time expressions as subtopic class
  - If have, the temporal relevant score ***TR*** for the document is 1, otherwise, the score ***TR*** is 0

Document Re-Ranking

- Relevant score weight sum
  - Sum content relevant score and temporal relevant score as final relevant score for candidate document ranking

$$R = \alpha R_c + (1-\alpha) R_t$$

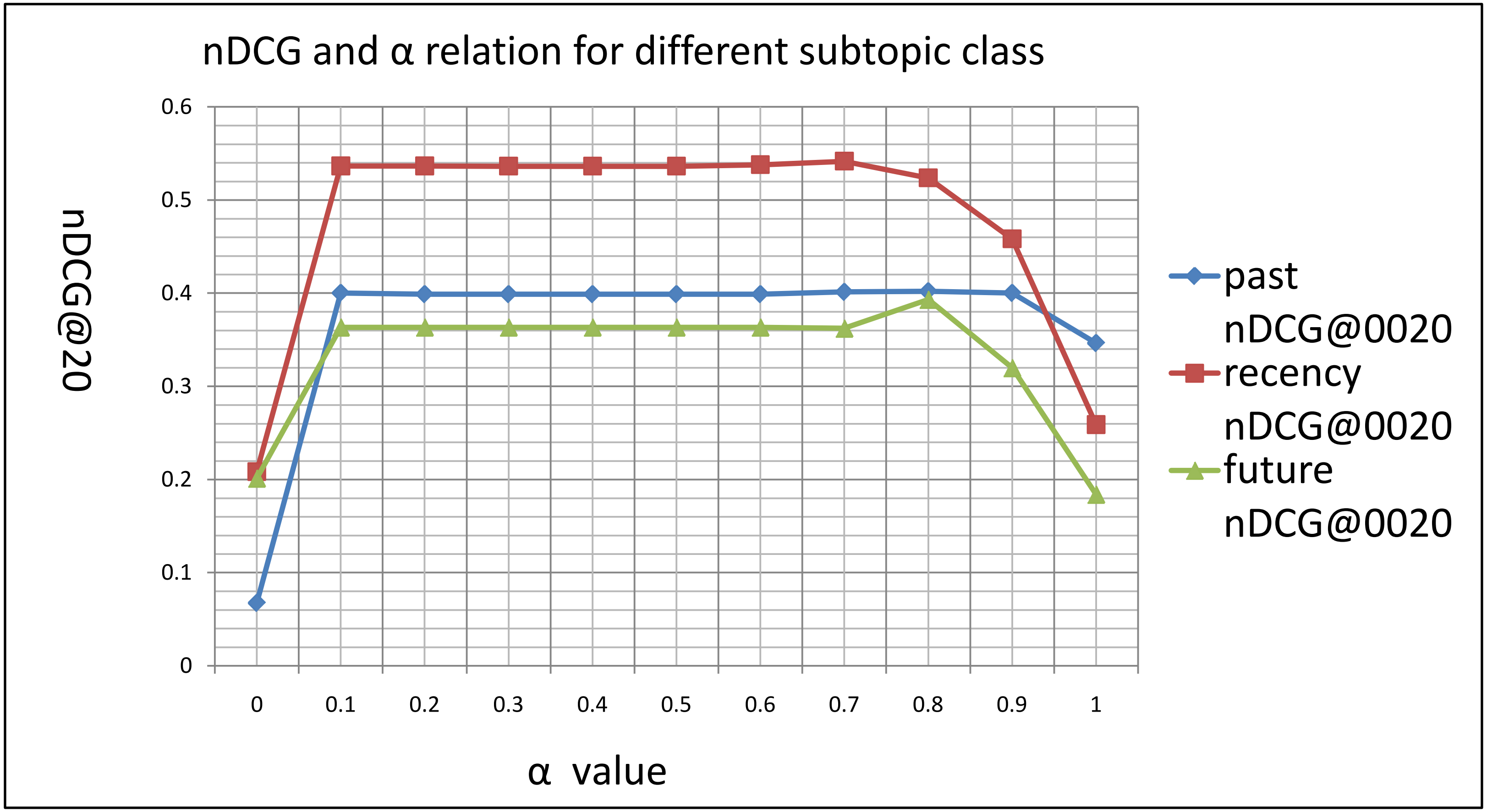
Where R is the document final relevant score to the search subtopic, R_c is the content relevant score, R_t is the temporal relevant score, α is the weight coefficient and α≥0, α≤1.

  - Different class use different coefficient value to calculate relevant score

Subtopic Class	Coefficient
past	0.85
recency	0.73
future	0.76
atemporal	1

Document Re-Ranking

– Coefficient choosing



Document Re-Ranking

  - Deficiency of relevant score weight sum method
    - Difficult to get the best coefficient for each subtopic
    - Subtopic class information is necessary
- Learning to rank
  - Features
    - content relevant feature
    - Temporal relevant feature

Document Re-Ranking

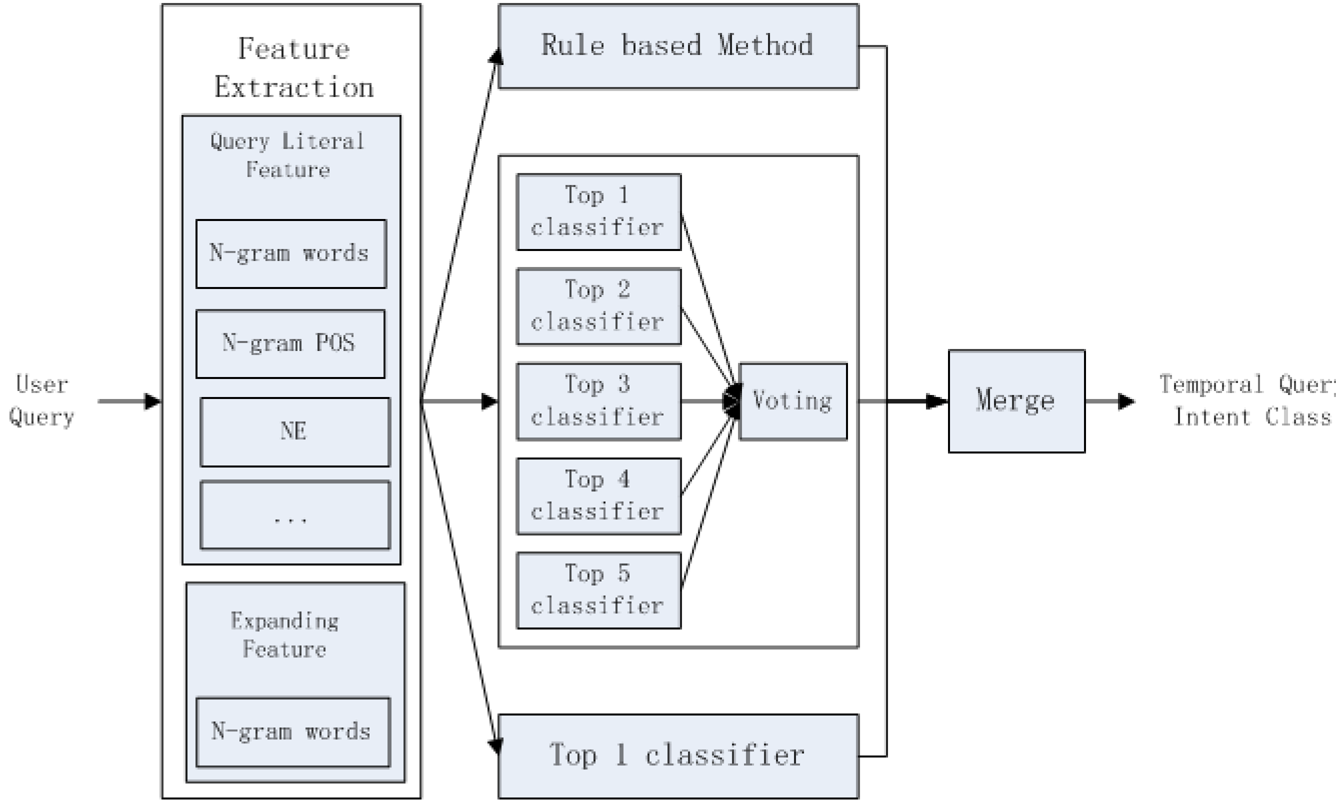
- Learning to rank
  - Feature List for learning to rank
    - ➢similarity between search topic and document title
    - ➢similarity between search topic and document content
    - ➢similarity between search subtopic and document title
    - ➢similarity between search subtopic and document content
    - ➢BM25 relevant score between search topic and document
    - ➢BM25 relevant score between search subtopic and document
    - ➢temporal relevant

Document Re-Ranking

- Learning to rank
  - Training data: 15 search topic each with 4 subtopics and relevant documents in *qrels* file in task dry run step
  - Algorithm: LambdaMART algorithm in RankLib toolkit
  - Model
    - Train independent model for each subtopic class
  - Dificiency: Lack of training data

Subtopic classification

- Use method in TQIC subtask to classify search subtopic class



Subtopic classification

- Classification accuracy of method in TQIC subtask is poor
- accuracy on TIR subtopic classifying is about 80%

Accuracy of each class in TQIC formal runs

runID	atemporal	future	past	recency
PrW	**70.67%**	64.00%	78.67%	**62.67%**
PrWsQW	69.33%	66.67%	77.33%	57.33%
qRPrHNB	57.33%	**68.00%**	**81.33%**	61.33%

- Training one rank model for all the 4 subtopic classes to avoid the subtopic classification step

Results Evaluation

- We submit 3 runs: *Run BW (run1), Run BWCC (run2), Run LTRNC2 (run3)*
- Run description
  - All 3 runs used **search title**, **subtopic** domain as input information
  - *run1*: used the relevant score weight sum method, and used the original class information of search subtopic.
  - *run2*: used the relevant score weight sum method, used the our classifiers in TQIC to get subtopic class, and did not use the original class information of search subtopic.
  - *run3*: used 1 model for 4 subtopic classes learning to rank method (LambdaMART algorithm here) , and did not use the original class information

Formal Results Evaluation

- *Run run3* is better than *run1* and *run2*
- No significant difference between *run1* and *run2*
  - Use time factor can improve re-rank result, no matter which class it is.
  - Same temporal feature for each class is not much suitable

Results evaluation of TIR subtask runs

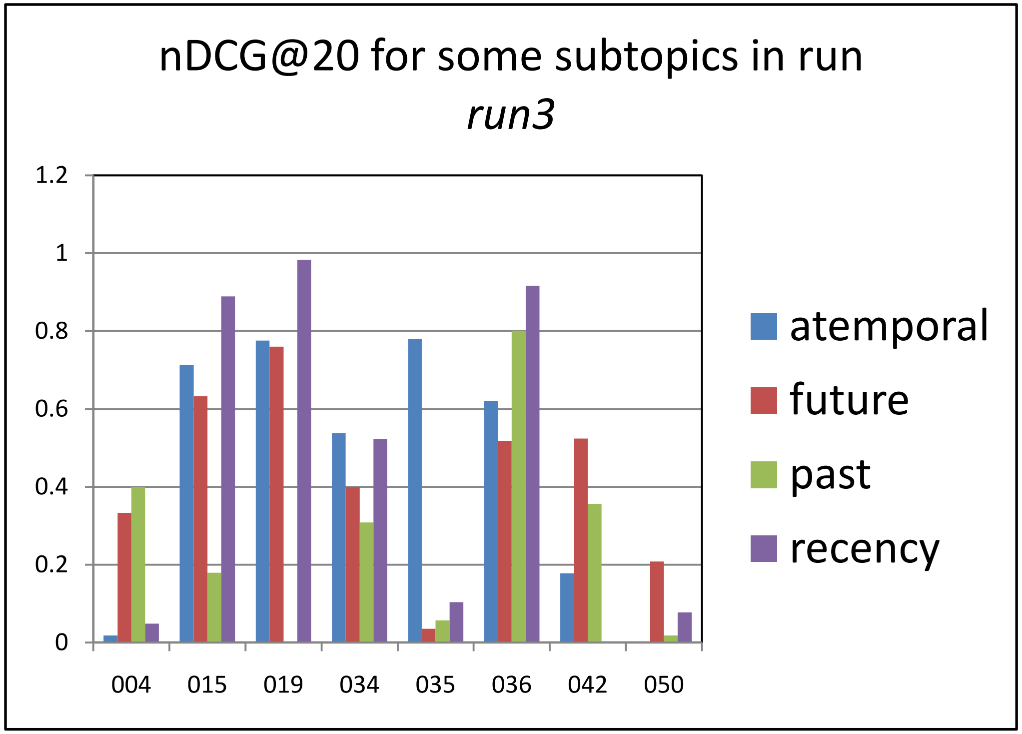
runID	nDCG @20	AP @20	P@ 20	nERR @20
BW (run1)	0.4544	0.4587	0.5895	0.6056
BWCC(run2)	0.4554	0.4599	0.5902	0.6064
LTRNC2(run3)	**0.4768**	**0.483**	**0.6018**	**0.6313**

nDCG@20 of each class in TIR formal runs

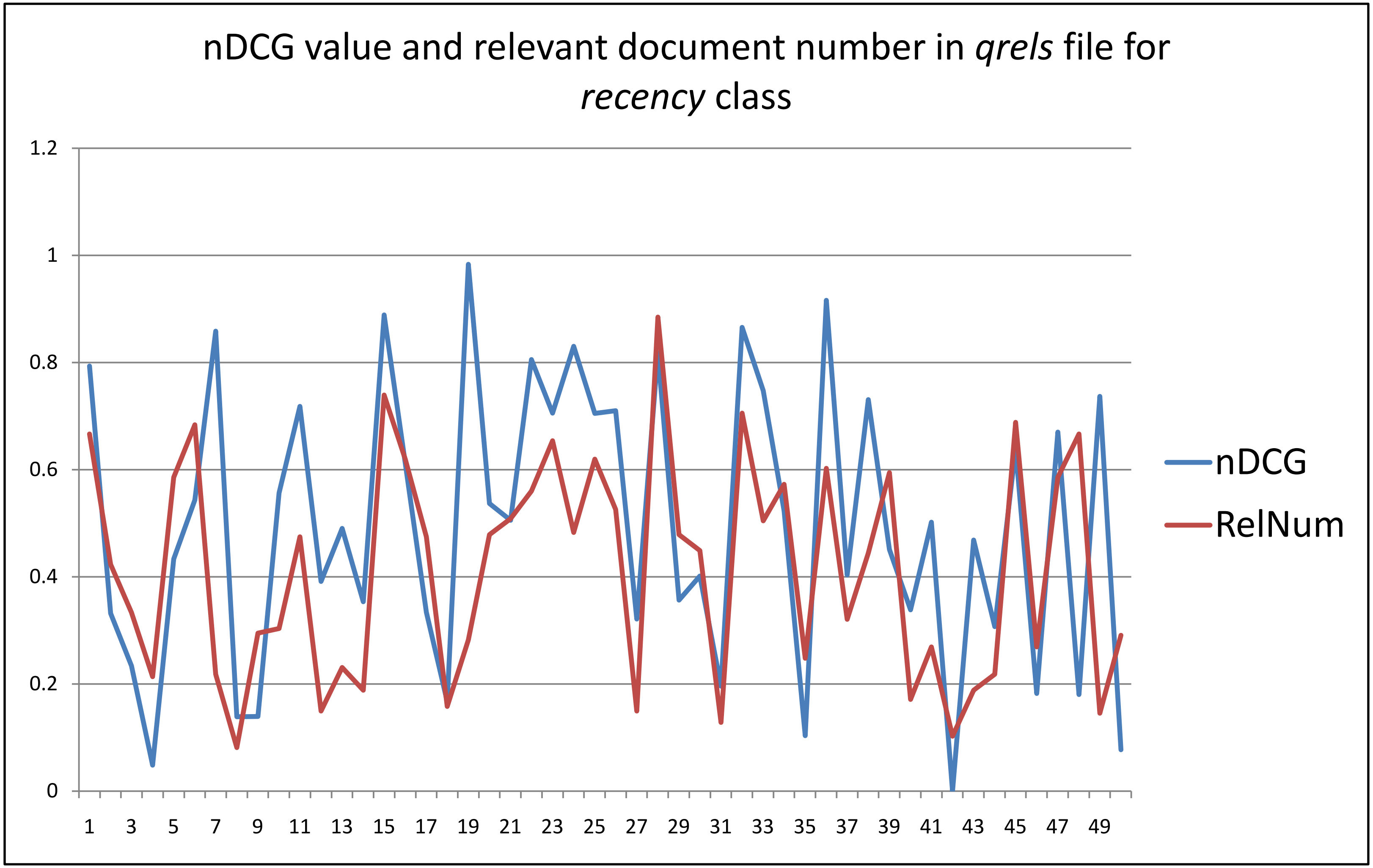
runID	atempor al	future	past	recency
BW (run1)	0.4669	0.4607	0.4005	0.4897
BWCC(run2)	0.4678	0.4593	0.403	0.4915
LTRNC2(run3)	**0.5092**	**0.4804**	**0.4227**	**0.495**

Results Evaluation

- nDCG values of different subtopics are very different, 4 subtopics in one topic are also different



Results Evaluation



Results Evaluation

- The more number of relevant documents in answer file, the higher nDCG value for most of subtopics
  - The number of relevant document for subtopic 036r, 015r, 050r and 042r is 141, 173, 68 and 24, nDCG value for those subtopics is 0.9159 , 0.8887, 0.0772 and 0
  - Fewer relevant documents in data set means smaller chance to get relevant documents for the top 20 results

Summary

- Use date distance as temporal feature
- Tried relevant score weight sum method
- Tried learning to rank method
- Future work
  - Improve temporal feature extraction. Eg. using float data between 0 and 1 to indicate the temporal relevant instead of 0, 1 only
  - Using the NE tags in document content

References

- H. Joho, A. Jatowt, and B. Roi. NTCIR Temporalia: A Test Collection for Temporal Information Access Research. In *Proceedings of the 4th Temporal Web Analytics Workshop (TempWeb 2014)*, ACM Press, Seoul, Korea, pages 845-849, 2014.
- H. Joho, A. Jatowt, and B. Roi. Overview of NTCIR-11 Temporal Information Access (Temporalia) Task. In *Proceedings of the 11th Evaluation of Information Access Technologies (NTCIR),* 2014.
- H. Joho, A. Jatowt, and B. Roi. A survey of temporal web search experience. In P*roceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 1101-1108, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- Y. Hou, Y. Zhang, X. Wang, Q. Chen, Y. Wang and B. Hu. Recognition and Retrieval of Time-sensitive Question in Chinese QA System [J]. Journal of Computer Research and Development, 2013, 50(12), pages 2612-2620
- J. Xu, H. Li. AdaRank: A Boosting Algorithm for Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 391-398, New York(NY), USA: ACM, 2007.
- Q. Wu, C.J.C. Burges, K. Svore and J. Gao. Adapting Boosting for Information Retrieval Measures. Journal of Information Retrieval, 2007.
- C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* pages 55-60, 2014.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1, 2009.
- S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management,* CIKM ’04, pages 42-49, New York, NY, USA, 2004. ACM.
- T. Sakai, R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’11, 2011.

Thanks for your Attention
Q&A

Welcome offline discussion by sending emails to
houyongshuai@hitsz.edu.cn