CLEF/NTCIR/TREC REPRODUCIBILITY CENTRE

http://www.centre-eval.org/

Twitter: @_centre_

CENTRE@NTCIR-14
http://www.centre-eval.org/ntcir14/

Nicola Ferro
Maria Maistro
Tetsuya Sakai
Ian Soboroff
Zhaohao Zeng
centre-org@list.waseda.jp

Version 20181005

Motivation

- Researcher A publishes a paper; says

Algo X > Algo Y on test data D.

⇒ Researcher B tries Algo X and Y on D but finds

Algo X < Algo Y. A replicability problem.

- Researcher A publishes a paper; says

Algo X > Algo Y on test data D.

⇒ Researcher B tries Algo X and Y on D' but finds

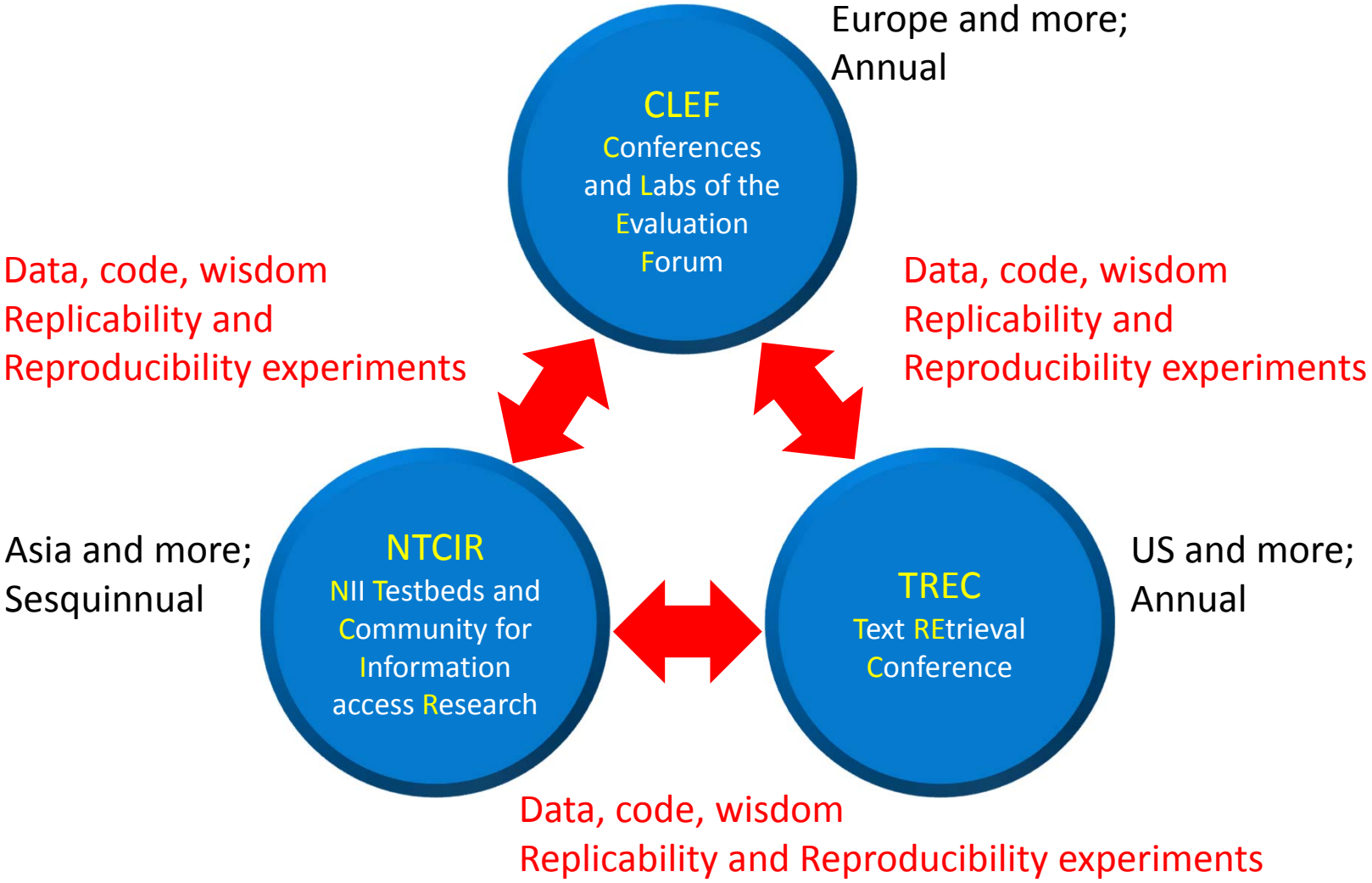
Algo X < Algo Y. A reproducibility problem.

Can the IR community do better?

Research questions

- Can results from CLEF, NTCIR, and TREC be replicated/reproduced easily? If not, what are the difficulties? What needs to be improved?
- Can results from (say) TREC be reproduced with (say) NTCIR data? Why or why not?
- [You may be motivated with your own research questions to participate in the task]

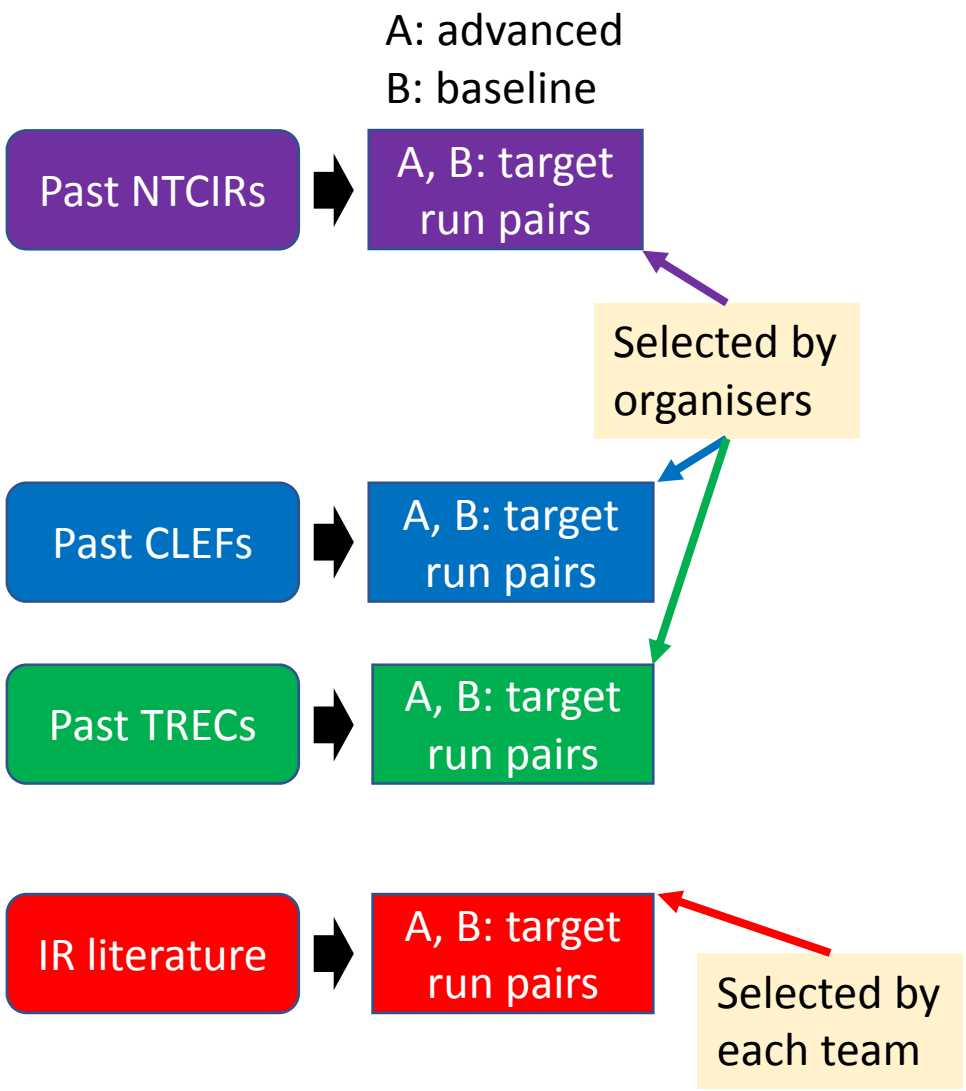
CENTRE = CLEF/NTCIR/TREC (replicability and) reproducibility



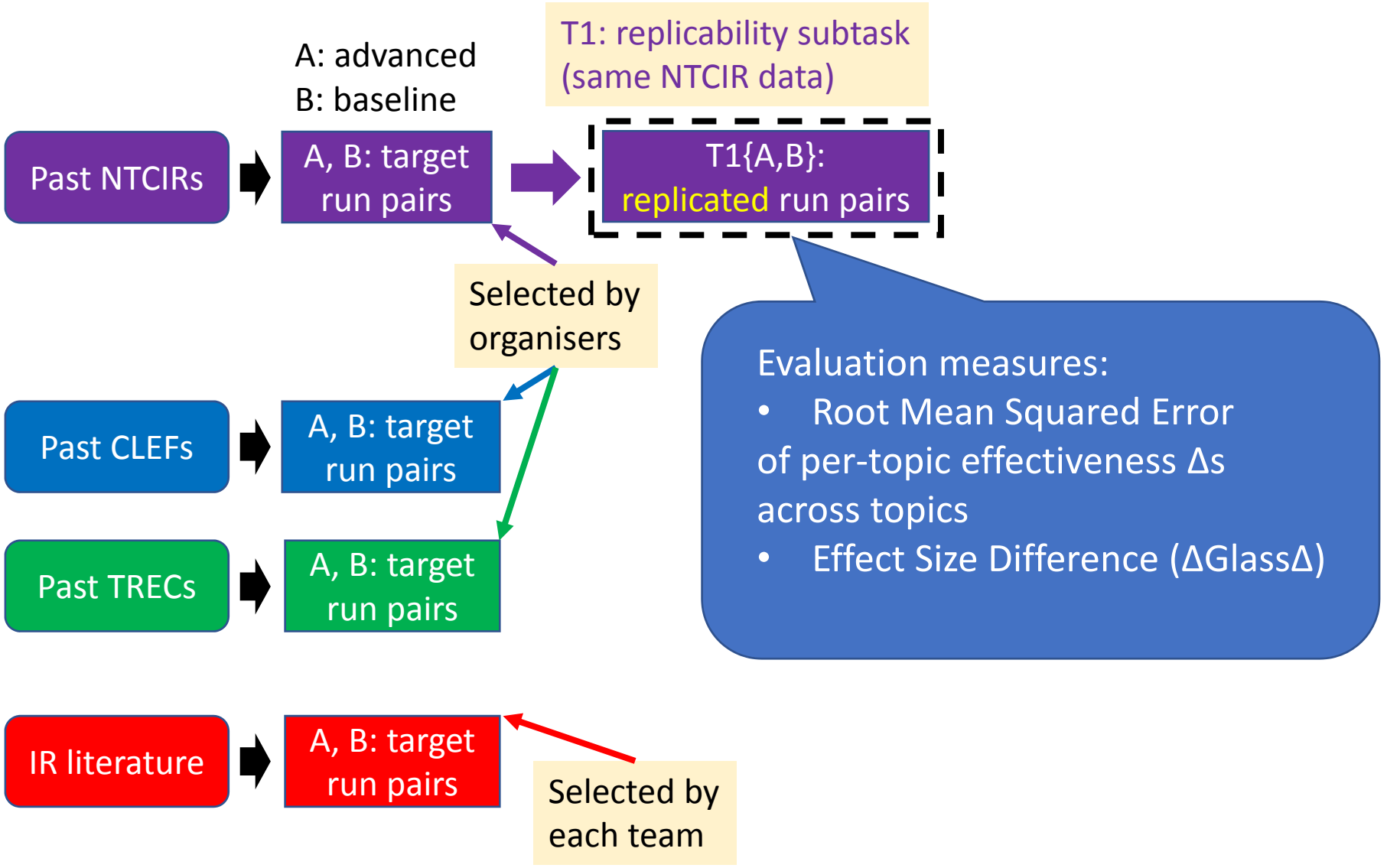
Who should participate in CENTRE?

- IR evaluation nerds
- Students willing to replicate/reproduce state-of-the-art methods in IR, so that they will eventually be able to propose their own methods and demonstrate their superiority over the state of the art
- Those who want to be involved in more than one evaluation communities across the globe
- Those who want to use data from more than one evaluation conferences

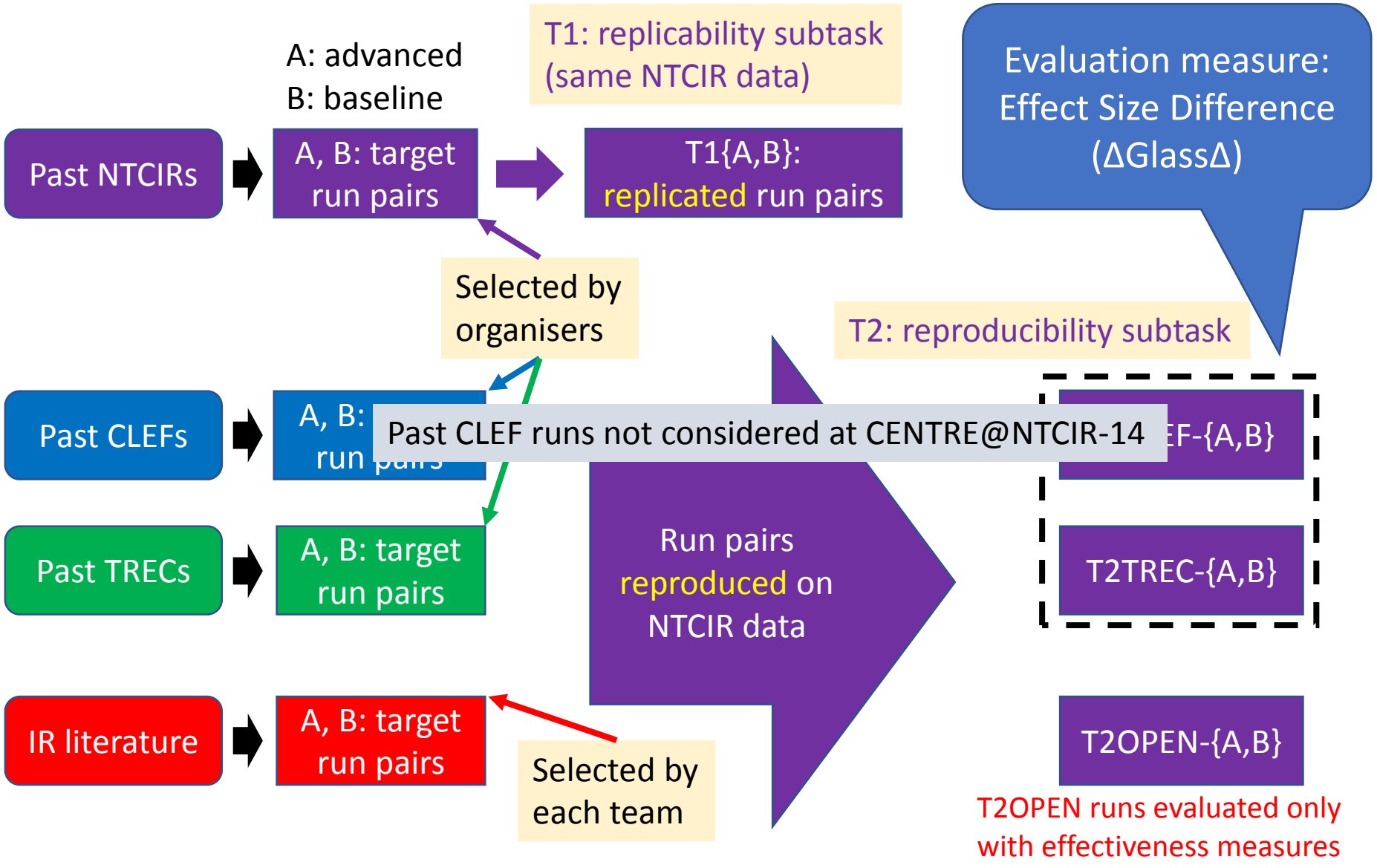
CENTRE@NTCIR-14: How it works (1)



CENTRE@NTCIR-14: How it works (2)



CENTRE@NTCIR-14: How it works (2)



Target runs for CENTRE@NTCIR-14

- We focus on ad hoc monolingual web search.
- For T1 (replicability): A pair of runs from the NTCIR-13 We Want Web task

- Overview paper:

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf

- For T2 (reproducibility): A pair of runs from the TREC 2013 web track

- Overview paper:

https://trec.nist.gov/pubs/trec22/papers/WEB.OVERVIEW.pdf

- All of the above target runs use clueweb12 as the search corpus

http://www.lemurproject.org/clueweb12.php/

- Both the target NTCIR and TREC runs are implemented using Indri

https://www.lemurproject.org/indri/

Target NTCIR runs and data for T1 (replicability)

- Test collection:

NTCIR-13 WWW English (clueweb12-B13, 100 topics)

- Target NTCIR run pairs:

From the NTCIR-13 WWW RMIT paper [Gallagher+17]

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/02-NTCIR13-WWW-GallagherL.pdf

A: RMIT-E-NU-Own-1 (sequential dependency model: SDM)

B: RMIT-E-NU-Own-3 (full dependency model: FDM)

- SDM and FDM are from [Metzler+Croft SIGIR05]

https://doi.org/10.1145/1076034.1076115

Target TREC runs and data for T2 (reproduciblity)

- Test collection:

TREC 2013 Web Track (clueweb12 category A, 50 topics)

- Target TREC run pairs:

From the TREC 2013 Web Track U Delaware paper [Yang+13] :

https://trec.nist.gov/pubs/trec22/papers/udel_fang-web.pdf

A: UDInfolabWEB2 (selects semantically related terms using Web-based working sets)

B: UDInfolabWEB1 (selects semantically related terms using collection-based working sets)

- The working set construction method is from [Fang+SIGIR06]

https://doi.org/10.1145/1148170.1148193

CENTRE organisers' results on the A- and B-runs

(evaluation measures computed with NTCIREVAL
http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html
randomisation test with Discpower
http://research.nii.ac.jp/ntcir/tools/discpower-en.html)

Please avoid NTCIREVAL versions 161017 and 180312. They do not work properly for diversity search evaluation due to a bug (though it works ok for our evaluation).

NTCIR-13 WWW RMIT: A-run > B-run

	Mean difference	P-value
MSnDCG@0010	0.0809	0
Q@0010	0.0891	0
nERR@0010	0.0486	0.0509

TREC 2013 Web UDel: A-run > B-run

	Mean difference	P-value
MSnDCG@0010	0.0963	0.0033
Q@0010	0.0601	0.0659
nERR@0010	0.1536	0.0014

The above measures are described in:
https://waseda.box.com/sakai14PROMISE

How participants are expected to submit T1 (replicability) runs

0. Obtain clueweb12-B13 (or the full data)

http://www.lemurproject.org/clueweb12.php/

and Indri https://www.lemurproject.org/indri/

1. Register to the CENTRE task and obtain the NTCIR-13 WWW topics and qrels from the CENTRE organisers

2. Read the NTCIR-13 WWW RMIT paper [Gallagher+17]

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/02-NTCIR13-WWW-GallagherL.pdf

and try to replicate the A-run and the B-run on the NTCIR-13 WWW-1 test collection

A: RMIT-E-NU-Own-1

B: RMIT-E-NU-Own-3

3. Submit the replicated runs by the deadline

Submitted runs will be evaluated with
Effect Size Difference and Root Mean Squared Error
(in nDCG, Q, nERR)

How participants are expected to submit T2TREC (reproducibility) runs

0. Obtain clueweb12-B13 (or the full data)

http://www.lemurproject.org/clueweb12.php/

and (optionally) Indri https://www.lemurproject.org/indri/

1. Register to the CENTRE task and obtain the NTCIR-13 WWW topics and qrels from the CENTRE organisers

2. Read the TREC 2013 Web Track U Delaware paper [Yang+13] :

https://trec.nist.gov/pubs/trec22/papers/udel_fang-web.pdf

and try to reproduce the A-run and the B-run on the NTCIR-13 WWW-1 test collection

A: UDInfolabWEB2

B: UDInfolabWEB1

3. Submit the replicated runs by the deadline

Submitted runs will be evaluated with Effect Size Difference (in nDCG, Q, nERR)

How participants are expected to submit T2OPEN (reproducibility) runs

0. Obtain clueweb12-B13 (or the full data)

http://www.lemurproject.org/clueweb12.php/

and (optionally) Indri https://www.lemurproject.org/indri/

1. Register to the CENTRE task and obtain the NTCIR-13 WWW topics and qrels from the CENTRE organisers

2. Pick any pair of runs A (advanced) and B (baseline) from the IR literature,

and try to reproduce the A-run and the B-run on the NTCIR-13 WWW-1 test collection

3. Submit the replicated runs by the deadline

nDCG, Q, nERR will be computed for the submitted runs, but the organisers will not evaluate them in terms of reproducibility

Why these target NTCIR runs?

- NTCIR-13 WWW is from the most recent NTCIR.
- The NTCIR-13 WWW English subtask used clueweb12-B13, which should be convenient for many IR researchers, even those outside NTCIR.
- RMIT was the top performer in NTCIR-13 WWW E.
- RMIT takes simple approaches using Indri, so they should be relatively easy to replicate.
- RMIT reports that SDM statistically significantly outperforms FDM (while randomised Tukey HSD in Overview says the difference is NOT statistically significant).

Submission instructions

- Each team can submit only one run pair per subtask (T1, T2TREC, T2OPEN). So at most 6 runs in total.
- Run file format: same as WWW-1; see

http://www.thuir.cn/ntcirwww/ (Run Submissions format). In short, it's a TREC run file plus a SYSDESC line.

- Run file names:

CENTRE1-<teamname>-<subtaskname>-[A,B]

e.g.

CENTRE1-WASEDA-T2TREC-A (advanced)

CENTRE1-WASEDA-T2TREC-B (baseline)

Where to submit

- As we have a small number of participating groups this year, please send a zipped file by email to centre-org@list.waseda.jp .
- Please include “NTCIR-14 CENTRE submission” in the email subject.
- Deadline: See final slide

Evaluation measure for T1 only:
Root Mean Squared Error of per-topic score Δs

$$\text{RMSE} = \sqrt{\sum_{j} (\Delta M_j \text{ (replicated)} - \Delta M_j \text{ (original)})^2}$$

where $\Delta M_j = M_j$ (advanced) – M_j (baseline)

and M_j is the effectiveness measure value for topic j.

(Thus we disregard the absolute performances of advanced/baseline.)

Retrieval effectiveness measures: nDCG, Q-measure, nERR (official WWW-1 measures)

Evaluation measures for T1 and T2TREC runs: Effect size difference (ΔGlassΔ)

ΔGlassΔ = GlassΔ(replicated) – GlassΔ(original)

GlassΔ =

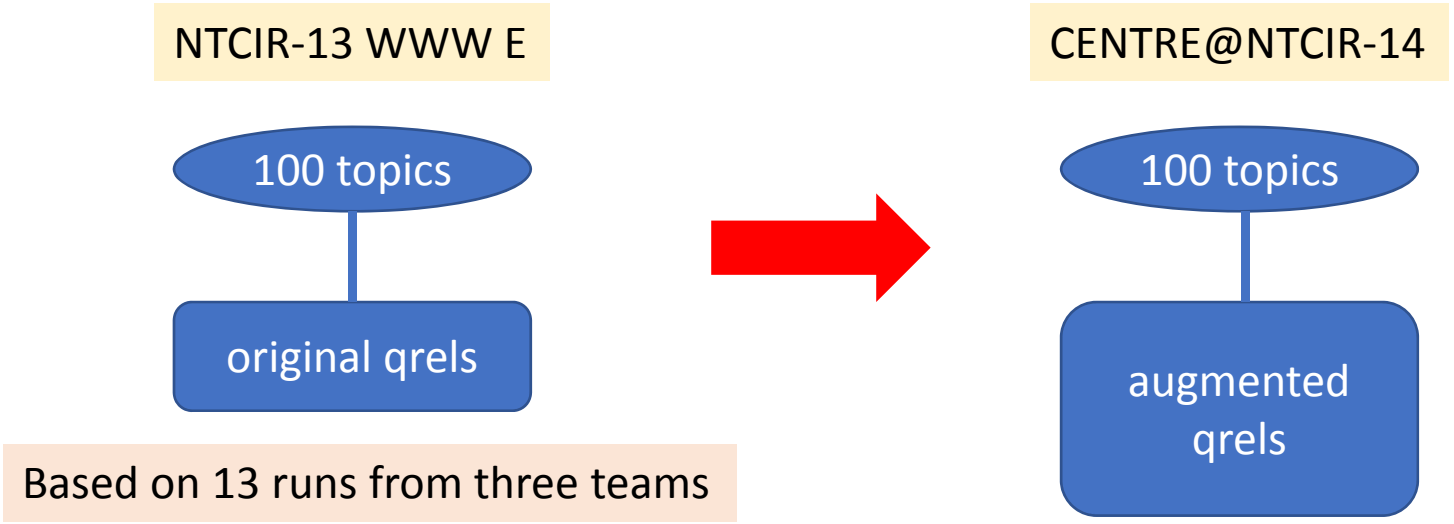
(MeanM(advanced) – MeanM(baseline)) / s_B

where s_B is the standard deviation of the original baseline run for effectiveness measure M

Retrieval effectiveness measures: nDCG, Q-measure, nERR (official WWW-1 measures)

Additional relevance assessments

- NTCIR-13 WWW E is probably an incomplete test collection – we will have to expand the qrels by pooling the submitted CENTRE runs (Nov-Dec).
- Depth-30 pools will be created, following WWW.



Timeline (incl. generic NTCIR-14 schedule)

- Jan-April 2018 Selection of official target runs from CLEF/TREC and selection of an existing NTCIR collection
- May 1, 2018 Announcement of the above selection – participants can start working on T1 and T2 experiments
- September 30, 2018 CENTRE@NTCIR-14 task registrations due

(Participants can select their own target runs in addition to the official target runs)

Nov 9 (Japan)

- ~~Oct 31, 2018~~ T1 and T2 runs due

Participants have 6 months for their experiments

- Nov-Dec, 2018 Additional relevance assessments
- Feb 1, 2019 Results summary and draft overview released
- Mar 15, 2019 Participant paper submissions due
- May 1, 2019 All camera-ready papers due
- Jun 2019 NTCIR-14 Conference@NII, Tokyo