Overview of the NTCIR-14 We Want Web Task

Jiaxin Mao[1], Tetsuya Sakai[2], Cheng Luo[1]
Peng Xiao[2], Yiqun Liu[1], and Zhicheng Dou[3]

[1] Tsinghua University, P.R.C. yiqunliu@tsinghua.edu.cn
[2] Waseda University, Japan tetsuyasakai@acm.org
[3] Renmin University of China, P.R.C. dou@ruc.edu.cn

Abstract. In this paper, we provide an overview of the NTCIR-14 We Want Web-2 (WWW-2) task, which includes the Chinese and the English subtasks. The series of WWW tasks are classical ad-hoc textual retrieval tasks. The WWW-2 task received 10 runs from 2 teams for the Chinese subtask, and 18 runs from 4 teams for the English subtask. In this overview paper, we not only describe the task details, data and evaluation methods, but also show the report on the official results.

Keywords: Web Search, Evaluation, Test Collections

1 Introduction

While there are a variety of novel tracks and tasks at NTCIR, TREC, CLEF, and etc., the ad hoc Web search is still an unsolved problem with utmost practical importance. As the emergence of deep neural networks, a number of studies start to tackle this problem with neural network methods and have reported some promising results. Therefore, we believe it is necessary to provide an evaluation forum to monitor the development of ad hoc retrieval models, especially for the newly proposed neural IR models. As ad hoc web search tasks had disappeared from NTCIR and TREC, we started to run an ad hoc evaluation task named We Want Web (WWW) in NTCIR 13, and planned to run it for at least three runs (NTCIR 13-15). This overview paper describes the task settings, data, and evaluation methods of the second rounds of the tasks (NTCIR 14 WWW-2), and report the official results.

As we planned, WWW tasks are traditional ad hoc tasks. The participants need to build their ranking systems on a given corpus. Then they are required to submit several runs for a given topic set. Similar to the last round of WWW (NTCIR-13), in WWW-2 (NTCIR-14) we have the Chinese subtask and English subtask. The two subtasks adopt the similar task settings with different data (see Section 3). There are some overlaps between the two query sets, which can support potential cross-language IR studies. Different from WWW (NTCIR-13), in this round we create a short task description for each topic, which should make the relevance assessment more reliable. For the Chinese subtask, we also provide a new dataset, Sogou-QCL. This dataset contains large-scale weak relevance labels generated by click models, which is ideal for training complex ranking models, such as deep neural networks. More details about the task definition will be presented in Section 2. The performance of the retrieval systems will be evaluated in classical TREC ways. We presented the details of relevance judgments in Section 4.1, and official results in Section 6 and 7.

The schedule of WWW-2 in NTCIR-14 is presented in Table 1. Although there are quite a few teams registered for our tasks, finally we only received 10 Chinese runs from 2 teams, and 18 English runs from 4 teams. Compared to WWW in NTCIR-13, where we had 19 Chinese runs from 4 teams and 13 English runs from 3 teams, the number of the teams that participate in the Chinese subtask decreased but the number of teams in the English subtask increased.

Table 1. Schedule of WWW-2 at NTCIR-14

Time	Content
July 1, 2018	Test Topics and SogouQCL released
Aug 31, 2018	WWW2 task registrations due
Oct 10 2018	Run submissions due
Oct-Dec 2018	Relevance assessments
Jan 2019	failure analysis workshop in Beijing
Feb 1, 2019	Evaluation Results and draft overview released
Mar 15, 2019	Submission due of participant papers
May 1, 2019	Camera-ready participant paper due
Jun, 2019	NTCIR-14 Conference & EVIA 2019 in NII, Tokyo

2 Task Definition

The main task of WWW-2 is a classical ad hoc search task. The organizers provide a corpus, which contains a large number of documents (web pages) and a query set. Then the participants need to construct their own ranking systems on the corpus. Retrieval results for each query will be submitted in the form of a ranked list. After receiving the runs from participants, the organizers will first construct a result pool by aggregating the top k results from all the runs. The depth of the pool determines how many results will be taken into consideration when comparing the performance of different submissions. For example, if we use 20 as our pooling depth, we can calculate the metrics with a cutoff smaller than 20. The depth of pooling is also limited by the cost for relevance judgments, in terms of time and money. Relevance judgments are conducted on the result pool. We adopt the typical TREC relevance judgment setting in WWW-2. Once the relevance judgments are collected, the organizers use various evaluation metrics (such as nDCG, Q-measure, nERR and etc.) to compare the performance of different submitted runs.

Considering that building an indexed retrieval system on a large corpus might be challenging and time-consuming, we provide a baseline ranking so that the participants could directly use their own algorithm to rerank it. More specifically, for each query, we provide the top 1,000 results retrieved by the BM25 model, as well as the corresponding BM25 scores and the original HTML files.

Similar to the last round of the WWW task, in WWW-2 task, we also release Chinese subtask and English subtask. They basically share the same task settings. For Chinese subtask, we provide an additional training dataset, Sogou-QCL, which contains 0.54 million queries and more than 9 million corresponding documents. For each query-doc pair, we provide 5 kinds of click labels generated by different click models, UBM, DBN, TCM, PSCM, and TACM. These click models utilize rich users' behavior such as click, skip and dwell time, to generate click labels that can be used as weak relevance labels.. The generated click labels can be used as weak relevance labels. We hope that this large-scale dataset can enable the training of some more complex retrieval models. Unfortunately, for English subtask, we do not have a similar dataset.

3 Data

3.1 Corpora

For the Chinese Subtask, we adopt the SogouT-16 as the document collection. SogouT-16 contains about 1.17 billion Web pages, which are sampled from the index of Sogou.com, the second largest commercial search engine in China. Considering that the original SogouT might be a little bit difficult to handle for some research groups (almost 80TB after decompression), we prepare a "Category B" version of SogouT-16, which is denoted as " SogouT16 B" . This subset contains about 15% webpages of SogouT-16. This Chinese corpus is free for research purpose. You can contact us to apply for it. We also implement a free online retrieval/page rendering service for this corpus. The online retrieval system is based on Solr[4], with the default parameter settings. You will get an account to use the service after the application is approved.

For the English Subtask, we adopt the ClueWeb12-B13 as the document collection. This corpus is also free for research purpose. You only need to pay for the disks and the shipment. More information can be found at Clueweb-12's homepage. ClueWeb12 also has a free online retrieval/page rendering service, which can be used after signing an agreement.

3.2 Topic Set Size design

We determined the size of the WWW-2 topic sets using the method proposed by Sakai [4] as follows. Table 2 shows the common variances $\hat{\sigma}^2$ obtained from the NTCIR-13 WWW-1 topic-by-run matrices of the three official evaluation measures [1], and the pooled variances obtained by consolidating the statistics

[4] http://lucene.apache.org/solr/

from the two subtasks. The pooled variances were used with a topic set size design tool[5] with $\alpha = 0.05, \beta = 20, m = 2$ to produce Table 3. The above setting means that we want to achieve 80% statistical power for any true difference larger than or equal to *minD* (*minimum detectable difference*) whenever we compare two systems with a t-test at the 5% significance level[6].

To determine the number of topics for WWW-2, we focused on our most unstable measure, nERR, since if we ensure that a certain statistical requirement is met for this measure, the same requirement will also be met by nDCG and Q. The bolded value in Table 3 means: "under Cohen's five-eighty convention (i.e., $\alpha = 0.05, \beta = 0.20$), any true difference that is at least 0.10 in terms of nERR@10 can be detected with 80% statistical power if we have 76 topics or more." Hence we decided to have 80 topics for both Chinese and English subtasks of WWW-2. Thus, the WWW-2 Chinese and English topic sets are expected to satisfy the above statistical requirement for all three measures.

Table 2. Variances from the WWW-1 data.

Measure	Chinese	English	Pooled
nDCG@10	0.028	0.030	0.029
Q@10	0.032	0.036	0.034
nERR@10	0.047	0.052	0.049

Table 3. Topic set size design results based on [4] and the pooled variances shown in Table 2 ($\alpha = 0.05, \beta = 0.20, m = 2$).

minD	nDCG@10	Q@10	nERR@10
0.05	178	209	301
0.10	45	53	**76**
0.20	6	7	9

3.3 Topic Sampling

The queries for Chinese subtask are sampled from Sogou's query logs in one day of November 2017. 55 queries are torso queries, which means that their frequencies are between 10 to 1,000 one day. 12 queries are tail queries which appeared only once in one day's log and the remaining 13 queries are hot queries which have a frequency larger than 1,000. We include more torso queries in the topic set because we believe that they are most appropriate for an ad hoc task. The

[5] `http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx`

[6] The excel tool is based on ANOVA, but a one-way ANOVA with $m = 2$ systems is equivalent to a two-sample t-test; a paired t-test generally requires fewer topics than a two-sample t-test. See Sakai [4] for details.

content of the queries, the intent types (navigational/information & transactional) and whether the queries are shared by English subtask are presented in Table 4.

The queries for English subtasks come from two sources. The first part (30 queries) are translated from the queries for Chinese subtask. The rest (50 queries) are sampled from the AOL query logs. In total, there are 11 hot queries, 57 torso queries, and 12 tail queries in the English topic set. Then content of the queries, the intent types, and whether the queries are shared by Chinese subtask are presented in Table 5.

For both English and Chinese query set, we did not use a lot of navigational queries. Since both SogouT and Clueweb are small subsets of the entire Web, it is very likely that the perfect answer for a navigational query is not in the corpus.

Different from the last WWW (NTCIR-13), in WWW-2, we further provide a task description for each topic, which will be shown to the assessors during relevance judgment. Providing a task description should make the relevance judgment more reliable and can enable the manual runs with varied queries formulated by the participants. Table 6 shows some examples for the task descriptions.

Table 4. Chinese query set (Int. indicates the intent types: we only point out the navigational queries while the remaining ones are informational or transactional; Trans. indicates whether the query is translated to English)

qid	Query	Int.	Trans.	qid	Query	Int.	Trans.	qid	Query	Int.	Trans.
0001	万圣节图片		Y	0028	腹胀			0055	三千越甲可吞吴		
0002	天秤座		Y	0029	宠物猪		Y	0056	心悸的症状		Y
0003	日历			0030	醴			0057	生日蛋糕图片创意设计		Y
0004	小米官网	NAV		0031	家常菜谱		Y	0058	国际金价		Y
0005	qq	NAV		0032	组装电脑			0059	直辖市有哪些		
0006	cba			0033	爱情 36 计			0060	EMS		
0007	汽车之家	NAV		0034	几个月有胎动			0061	江西会计网	NAV	
0008	上证指数			0035	谐音			0062	串联和并联的区别		Y
0009	vivo			0036	植物人			0063	小米粥怎么煮		
0010	网名大全			0037	哪天是万圣节		Y	0064	虎扑体育论坛	NAV	
0011	科目三通过率下降			0038	防空警报响了什么意思		Y	0065	唯品会官网	NAV	
0012	红烧肉做法			0039	职业生涯规划		Y	0066	梅德韦杰夫		Y
0013	快递查询单号			0040	地铁 2 号线线路图			0067	百度学术官网	NAV	
0014	小米分期怎么开通			0041	支链氨基酸		Y	0068	黄金时代电影		
0015	拼音音调怎么标			0042	卡农		Y	0069	荒缪是什么意思		Y
0016	争先恐后的意思			0043	鹿晗图片			0070	天津市国家税务局稽查局		
0017	南昌工学院			0044	硬拉的标准动作		Y	0071	环湖公路高邮		
0018	手机系统升级包			0045	复合函数求导		Y	0072	桉木木材		
0019	冬季女装		Y	0046	最强大脑			0073	世界上最大的瀑布叫什么		Y
0020	疝气是什么			0047	投屏			0074	南京海底捞营业时间		
0021	初级会计考试			0048	维生素 A		Y	0075	佳能 IP1188 驱动		Y
0022	肠胃痉挛		Y	0049	中国光大银行信用卡			0076	吃完柿子能喝酸奶吗		Y
0023	朱芳雨			0050	优秀毕业生主要事迹 300 字			0077	坐着看小说网	NAV	
0024	中国邮政集团网上营业厅	NAV		0051	宝马 z4 报价		Y	0078	车漆指甲划痕		Y
0025	墙面装饰		Y	0052	社保			0079	电脑显示屏怎么区分好坏		Y
0026	查理九世全集			0053	克里斯蒂亚诺 · 罗纳尔多		Y	0080	红珊瑚图片大全		Y
0027	绝地求生手游下载			0054	一个鸡蛋的热量		Y				

3.4 Chinese Training Data

For Chinese subtask, in this round of WWW-2, we provide a new training set, Sogou-QCL. Sogou-QCL contains two kinds of training sets:

- The first set contains traditional relevance assessments. It is made of 1000 Chinese queries and for each query, Sogou-QCL contains about 20 query-

Table 5. English query set (Int. indicates the intent types: we only point out the navigational queries while the remaining ones are informational or transactional; First 30 queries (0001–0030) are translated from Chinese)

qid	Query	Int.	qid	Query	Int.
0001	Halloween picture		0041	grips for 1911 pistol	
0002	calendar		0042	latest news on angelina jolie	
0003	women's clothing winter		0043	jobs with a finance degree	
0004	Gastrointestinal fistula		0044	louisvilleslugger.com	Nav
0005	Wall decoration		0045	free e-mails greetting cards	
0006	Pet pig		0046	peanut	
0007	Homemade recipes		0047	deaths 1927 cleveland ohio	
0008	Which day is Halloween		0048	dining room chairs	
0009	what does Civil defense siren mean?		0049	flemish painter	
0010	career plan		0050	white chocolate macadamia nut cookies	
0011	Branched chain amino acid		0051	prednisone	
0012	Canon		0052	movie trailer	
0013	deadlift proper form		0053	wake board towers	
0014	Derivatives of Composite Functions		0054	the friedman test	
0015	Vitamin A		0055	blue note	
0016	BMW z4 offer		0056	yosemite national park	
0017	Cristiano Ronaldo		0057	walker baskets	
0018	the calories of an egg		0058	pictures of smiles and teeth	
0019	palpitation symptoms		0059	boat repair	
0020	Birthday cake design picture		0060	car-parts.com	Nav
0021	International gold price		0061	old english fonts	
0022	difference between series and parallel circuit		0062	density	
0023	Medvedev		0063	www.dps.com	
0024	what does ridiculous mean		0064	liver diease	
0025	What is the biggest waterfall in the world?		0065	avril lavigne lyrics	
0026	Canon IP1188 driver		0066	carribean cruises	
0027	Can you drink yogurt after eating persimmon?		0067	tiki artist	
0028	nail scratches on car paint		0068	hp service	
0029	How to pick computer monitors		0069	snow crab legs	
0030	Red coral picture		0070	maps of costa rica	
0031	disney		0071	floor tile	
0032	history		0072	american tourister luggage	
0033	dictionary		0073	lululemon	
0034	southwest airline		0074	www.mbusa.com	Nav
0035	yahoo chat	Nav	0075	things to do & see in croatia	
0036	www.chase.com	Nav	0076	pogo mah jong	
0037	cheap flights		0077	snow white costume	
0038	priceline.com	Nav	0078	new movies for kids	
0039	song lyrics		0079	alcohol and its negative affects on society	
0040	recreation warehouse		0080	www.gardenburger.com	Nav

doc relevance judgments. Each pair is annotated by three trained assessors. Sogou-QCL also provides title and content extracted from raw HTMLs.

- The second set consists of click labels generated by click models. Releasing the original click logs could possibly harm user's privacy because it may contain personally identifiable information. Therefore we provide the relevance scores estimated based on group of users' behaviors. More specifically, for each query-doc pair, we provide five kinds of weak relevance label computed by five popular click models: UBM, DBN, TCM, PSCM, and TACM. These click models utilize rich users' behavior, such as click, skip, and dwell time, to estimate the relevance of the query-document pairs in click logs. Sogou-QCL contains more than half a million queries and more than 9 millions of documents. To the best of knowledge, this is so far the largest free training collection for Chinese ranking problem.

Handling the raw HTML content can be difficult. Therefore we also provide the extracted content with professional tools of Sogou.com. We hope it will reduce some effort for our participants and help them focus on the design of ranking models.

Table 6. Some examples of the task descriptions in both the Chinese subtask and English subtask.

Subtask	Query	Task description
Chinese	万圣节图片	你想向其他人介绍一下万圣节，所以需要查找和万圣节相关的图片。
Chinese	科目三通过率下降	你正准备考驾照，听说近期科目三考试通过率下降，想了解一下相关信息。
English	Halloween picture	You want to find some pictures about Halloween to introduce it to your children.
English	career plan	You are an undergraduate student who is about to graduate. You want to search some information about how to plan your career.

4 Runs and Relevance Assessments

4.1 Received Runs

Table 7 summarizes our run statistics. In this round, although there are quite a few teams registered for the tasks, we only received 10 Chinese runs from 2 teams, and 18 English runs from 4 teams, plus a few runs from the organisers' team (ORG). The Chinese subtask included the Chinese baseline run in the pool as a special run from the organizers' team and measured its performance. In the English subtask, we also had one baseline run and one manual run from the organizers' team. In the manual run, the queries were manually formulated by looking at the topic descriptions and BM25-based reranking was applied to the baseline English run.

Table 7. Run statistics including the runs from the organizers (ORG).

Team	Chinese	English	Total
MPII	-	5	5
RUCIR	5	5	10
SLWWW	-	3	3
THUIR	5	5	10
ORG	1	2	3
Total	11 (3 teams)	20 (5 teams)	31

4.2 Relevance Assessments

The Chinese relevance assessments were organized by Tsinghua University. We contacted to an annotation company named 小牛雅智. The relevance assessments were conducted in their company from Nov. 6th, 2018 to Nov. 30th, 2018.

For each query in the query set, we provided the query content, as well as the task description, which would help the assessors understand the search intent more specifically. We also provided the relevance assessment criteria for them, which are shown as follows:

GARBLED Garbled - The HTML page is shown to user with the *garbled* state.

NONREL Nonrelevant - It is *unlikely* that the user who entered this search query will find this page relevant.

MARGREL Marginally relevant - the user will get some relevant information from this page. However, she needs to browse more pages to satisfy her information needs.

REL Relevant - it is *possible* that the user who entered this search query will find this page relevant.

IGHREL Highly relevant - it is *likely* that the user who entered this search query will find this page relevant.

Finally, NONREL and GARBLED labels were mapped to zero; MARGREL labels were mapped to one; REL labels were mapped to two and HIGHREL labels were mapped to three. After dropping the garbled documents in the pooling set (which is easy to annotate), the Fleiss' κ of the left annotations is 0.5047, which indicates a moderate agreement in the different assessors.

The English relevance assessments were organised by Waseda University, using the PLY interface originally developed for the NTCIR-13 WWW task (WWW-1) [1]. Twenty international course students were hired as relevance assessors, and each student handled eight topics. Each topic was judged independently by two assessors. While we used a pool depth of 30 at WWW-1, we set it to 50 this time. Although we leveraged the relevance assessment process of WWW-2 to experiment on the effect of document ordering for the assessors, we shall report on the results of the analysis elsewhere.

The raw English relevance assessments were collected in the same way as WWW-1:

ERROR The right panel does not show any contents at all, even after waiting for a few seconds for the content to load.

H.REL Highly relevant - it is *likely* that the user who entered this search query will find this page relevant.

REL Relevant - it is *possible* that the user who entered this search query will find this page relevant.

NONREL Nonrelevant - it is *unlikely* that the user who entered this search query will find this page relevant.

Finally, **ERROR** and **NONREL** were mapped to zero, **REL** was mapped to one, and **H.REL** was mapped to two, and the relevance levels $L4$ through $L0$ were obtained by summing the judgments of the two assessors for each topic.

Table 8 summarizes our relevance assessment statistics.

Table 8. Relevance assessment statistics.

	Chinese	English
#topics	80	80
#assessors/topic	3	2
Pool depth	20	50
Total #docs pooled	12,271	27,627
Total L4-relevant	-	857
Total L3-relevant	1,961	2,332
Total L2-relevant	1,524	4,664
Total L1-relevant	2,401	6,469
Total L0	6,385	13,305

5 Evaluation Measures and Tools

For evaluation metrics, we used the NTCIREVAL tool[7] to compute nDCG@10 (Microsoft version of nDCG at cutoff 10), Q@10 (Q-measure at cutoff 10), and nERR@10 (normalised expected reciprocal rank at cutoff 10) [3]. Linear gain values were used, e.g., 3 for L3-relevant, 1 for L1-relevant.

The Discpower tool[8] was used to conduct randomised Tukey HSD tests, each with B = 10, 000 trials [3].

For the Chinese subtask, the evaluation method in NTCIR-14 WWW-2 is slightly different from that in last WWW task. In the last round of WWW (NTCIR-13), we used ten levels of relevance label, L0 to L9, to evaluate the runs we received. The relevance label was computed by summing all the three assessors' labels. For example, if the relevance labels on query-document pair received from three assessors label are MARGREL, REL, and REL, we first mapped them to the corresponding gain values, 1, 2, and 2. Then the final relevant label we used was computed by $1+2+2=5$, which means the document and the query is L5-relevant. However, in this round of WWW-2 (NTCIR-14), we find that if we continue to use the ten-level relevance labels, the Kendall's τs between three evaluation measures are not high enough. For example, the Kendall's τ between Q@10 and nERR@10 of the ten levels of relevance label method is 0.527, which means using different evaluation measures may lead to different conclusions. Therefore, we use the median of all the three assessors' labels, i.e. a four-level relevance labels, instead.

6 Chinese Subtask Results

6.1 Overall Chinese Results

For the Chinese subtask, we find that 5 of 80 queries are inappropriate for the evaluation. They are “醴” (qid: 0030), “江西会计网” (0061), “百度学术官网”

[7] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

[8] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

(0067), "坐着看小说网"(0077), and "车漆指甲划痕" (0078). For the query "醴", we find that the rate of garbled HTML documents in our pooling set is 86.7%. The query is a Chinese character that is mostly used ancient Chinese literature but rarely used in modern Chinese. Most of the documents were retrieved because of the encoding errors of Chinese webpages. So we drop this query. For the other four queries, we find that all of the HTML documents in the pooling set are nonrelevant. Therefore, we drop them as well.

Table 9 shows the mean effectiveness scores for all Chinese runs. Table 10 summarizes the statistical significance test results. Randomized Tukey HSD p-values and effect sizes (i.e., standardized mean differences) based on two-way ANOVA (without replication) [4] are also shown. For example, the effect size for the difference between THUIR-C-CO-CU-Base-5 and baseline run in terms of nDCG@10 is given by $ES_{E2} = (0.4916 - 0.3545)/\sqrt{0.0346} = 0.737$.

From the official Chinese results with the three evaluation measures, it can be observed that:

- THUIR and RUCIR are not statistically significantly different from each other;
- THUIR and RUCIR both outperforms baseline run (using the BM25 model with the default parameter in Solr) significantly.

In Table 11, we compare the system rankings according to the three evaluation measures in terms of Kendall's τ, as well as their 95% confidence intervals. It can be observed that the three rankings are statistically equivalent.

Table 9. Official Chinese results.

Run	Mean nDCG@10	Run	Mean Q@10	Run	Mean nERR@10
THUIR-C-CO-CU-Base-5	0.4916	THUIR-C-CO-CU-Base-5	0.4610	THUIR-C-CO-CU-Base-5	0.6374
RUCIR-C-CO-PU-Base-2	0.4866	THUIR-C-CO-MAN-Base-2	0.4604	RUCIR-C-CO-PU-Base-2	0.6044
THUIR-C-CO-MAN-Base-2	0.4835	RUCIR-C-CO-PU-Base-2	0.4571	THUIR-C-CO-MAN-Base-1	0.6019
THUIR-C-CO-MAN-Base-1	0.4748	THUIR-C-CO-MAN-Base-1	0.4479	THUIR-C-CO-MAN-Base-2	0.5973
THUIR-C-CO-MAN-Base-3	0.4706	THUIR-C-CO-MAN-Base-3	0.4364	THUIR-C-CO-MAN-Base-3	0.5829
RUCIR-C-DE-PU-Base-1	0.4515	RUCIR-C-DE-PU-Base-1	0.4228	RUCIR-C-DE-PU-Base-1	0.5792
RUCIR-C-DE-PU-Base-4	0.4510	RUCIR-C-DE-PU-Base-4	0.4226	THUIR-C-CO-CU-Base-4	0.5663
RUCIR-C-DE-PU-Base-3	0.4503	RUCIR-C-DE-PU-Base-3	0.4223	RUCIR-C-DE-PU-Base-3	0.5630
THUIR-C-CO-CU-Base-4	0.4458	THUIR-C-CO-CU-Base-4	0.4189	RUCIR-C-DE-PU-Base-4	0.5619
baseline	0.3545	baseline	0.3080	baseline	0.4869
RUCIR-C-DE-PU-Base-5	0.2745	RUCIR-C-DE-PU-Base-5	0.2404	RUCIR-C-DE-PU-Base-5	0.3832

7 English Subtask Results

Table 12 shows the official WWW-2 English run results. Table 13 summarises the statistical significance test results. It can be observed that the runs from THUIR are the most effective, and that for all statistically significant differences, the effect sizes (standard mean differences) are worth over half a standard deviation. (We did not obtain any statistically significant differences in terms of nERR.)

Table 14 compares the system rankings according to the three official measures in terms of Kendall's τ, with 95%CIs.

Table 10. Statistical significance with the best Chinese run from each team (Randomised Tukey HSD test, $B = 10,000, \alpha = 0.05$).

These runs are	significantly better than these runs in terms of nDCG@10
THUIR-C-CO-CU-Base-5	baseline ($p < 0.0001$, $ES_{E2} = 0.737$)
RUCIR-C-DE-PU-Base-2	baseline ($p = 0.0002$, $ES_{E2} = 0.710$)
These runs are	significantly better than these runs in terms of Q@10
THUIR-C-CO-CU-Base-5	baseline ($p < 0.0001$, $ES_{E2} = 0.786$)
RUCIR-C-DE-PU-Base-2	baseline ($p < 0.0001$, $ES_{E2} = 0.766$)
These runs are	significantly better than these runs in terms of nERR@10
THUIR-C-CO-CU-Base-5	baseline ($p = 0.0014$, $ES_{E2} = 0.658$)
RUCIR-C-DE-PU-Base-2	baseline ($p = 0.0077$, $ES_{E2} = 0.513$)

Table 11. Kendall’ s τ values with 95% CIs (11 Chinese runs).

	Mean Q@10	Mean nERR@10
Mean nDCG@10	0.964 [0.906, 0.986]	0.855 [0.654, 0.943]
Mean Q@10	-	0.818 [0.579, 0.928]

Table 12. Official English results.

Run	Mean nDCG@10	Run	Mean Q@10	Run	Mean nERR@10
THUIR-E-CO-MAN-Base-3	0.3536	THUIR-E-CO-MAN-Base-2	0.3391	THUIR-E-CO-MAN-Base-1	0.5048
THUIR-E-CO-MAN-Base-2	0.3512	RUCIR-E-CO-PU-Base-2	0.3352	THUIR-E-CO-MAN-Base-2	0.5026
RUCIR-E-CO-PU-Base-2	0.3489	MPII-E-CO-NU-Base-4	0.3265	RUCIR-E-CO-PU-Base-2	0.4917
THUIR-E-CO-MAN-Base-1	0.3444	THUIR-E-CO-MAN-Base-3	0.3256	THUIR-E-CO-MAN-Base-3	0.4805
MPII-E-CO-NU-Base-3	0.3413	MPII-E-CO-NU-Base-2	0.3255	baseline-eng-v1	0.4779
MPII-E-CO-NU-Base-2	0.3394	THUIR-E-CO-MAN-Base-1	0.3249	THUIR-E-CO-PU-Base-5	0.4779
MPII-E-CO-NU-Base-4	0.3336	MPII-E-CO-NU-Base-3	0.3183	MPII-E-CO-NU-Base-4	0.4723
THUIR-E-CO-PU-Base-4	0.3294	THUIR-E-CO-PU-Base-4	0.3161	THUIR-E-CO-PU-Base-4	0.4692
RUCIR-E-DE-PU-Base-4	0.3293	MPII-E-CO-NU-Base-5	0.3110	MPII-E-CO-NU-Base-3	0.4658
MPII-E-CO-NU-Base-5	0.3293	RUCIR-E-DE-PU-Base-4	0.3094	RUCIR-E-DE-PU-Base-4	0.4602
baseline-eng-v1	0.3258	baseline-eng-v1	0.3043	MPII-E-CO-NU-Base-2	0.4590
THUIR-E-CO-PU-Base-5	0.3258	THUIR-E-CO-PU-Base-5	0.3043	MPII-E-CO-NU-Base-5	0.4584
MPII-E-CO-NU-Base-1	0.3204	MPII-E-CO-NU-Base-1	0.3009	MPII-E-CO-NU-Base-1	0.4541
RUCIR-E-DE-PU-Base-3	0.3137	RUCIR-E-DE-PU-Base-3	0.2973	RUCIR-E-DE-PU-Base-3	0.4469
RUCIR-E-DE-PU-Base-1	0.3137	RUCIR-E-DE-PU-Base-1	0.2973	RUCIR-E-DE-PU-Base-1	0.4469
RUCIR-E-DE-PU-Base-5	0.2876	ORG-MANUAL	0.2685	ORG-MANUAL	0.4294
SLWWW-E-CO-NU-Base-1	0.2860	SLWWW-E-CO-NU-Base-1	0.2665	RUCIR-E-DE-PU-Base-5	0.4188
ORG-MANUAL	0.2844	RUCIR-E-DE-PU-Base-5	0.2659	SLWWW-E-CO-NU-Base-1	0.4071
SLWWW-E-CO-NU-Base-4	0.2775	SLWWW-E-CD-NU-Base-3	0.2499	SLWWW-E-CD-NU-Base-3	0.4034
SLWWW-E-CD-NU-Base-3	0.2767	SLWWW-E-CO-NU-Base-4	0.2498	SLWWW-E-CO-NU-Base-4	0.4015

Table 13. Statistical significance with the Randomised Tukey HSD test, $B = 10,000, \alpha = 0.05$. Statistically significant differences were not observed in terms of nERR@10. The effect sizes are based on the residual variance V_{E2} from two-way ANOVA without replication [4].

This run is significantly better than	these runs in terms of nDCG@10	p-value	effect size ES_{E2} ($V_{E2} = 0.0126$)
THUIR-E-CO-MAN-Base-3	RUCIR-E-DE-PU-Base-5	0.034	0.587
	SLWWW-E-CO-NU-Base-1	0.025	0.601
	ORG-MANUAL	0.017	0.616
	SLWWW-E-CO-NU-Base-4	0.004	0.677
	SLWWW-E-CD-NU-Base-3	0.003	0.684
THUIR-E-CO-MAN-Base-2	SLWWW-E-CO-NU-Base-1	0.041	0.580
	ORG-MANUAL	0.029	0.595
	SLWWW-E-CO-NU-Base-4	0.008	0.656
	SLWWW-E-CD-NU-Base-3	0.007	0.663
THUIR-E-CO-MAN-Base-1	SLWWW-E-CO-NU-Base-4	0.028	0.596
	SLWWW-E-CD-NU-Base-3	0.023	0.603
RUCIR-E-CO-PU-Base-2	ORG-MANUAL	0.045	0.574
	SLWWW-E-CO-NU-Base-4	0.011	0.636
	SLWWW-E-CO-NU-Base-3	0.010	0.643
MPII-E-CO-NU-Base-3	SLWWW-E-CO-NU-Base-3	0.044	0.575

This run is significantly better than	these runs in terms of Q@10	p-value	effect size ES_{E2} ($V_{E2} = 0.0142$)
THUIR-E-CO-MAN-Base-2	ORG-MANUAL	0.030	0.593
	SLWWW-E-CO-NU-Base-1	0.021	0.609
	RUCIR-E-DE-PU-Base-5	0.018	0.615
	SLWWW-E-CD-NU-Base-3	0.001	0.749
	SLWWW-E-CO-NU-Base-4	0.001	0.750
THUIR-E-CO-MAN-Base-3	SLWWW-E-CD-NU-Base-3	0.011	0.635
	SLWWW-E-CO-NU-Base-4	0.011	0.636
THUIR-E-CO-MAN-Base-1	SLWWW-E-CD-NU-Base-3	0.013	0.629
	SLWWW-E-CO-NU-Base-4	0.013	0.630
RUCIR-E-CO-PU-Base-2	SLWWW-E-CO-NU-Base-1	0.043	0.576
	RUCIR-E-DE-PU-Base-5	0.038	0.581
	SLWWW-E-CD-NU-Base-3	0.002	0.715
	SLWWW-E-CO-NU-Base-4	0.002	0.717
MPII-E-CO-NU-Base-4	SLWWW-E-CD-NU-Base-3	0.009	0.643
	SLWWW-E-CO-NU-Base-4	0.008	0.644
MPII-E-CO-NU-Base-2	SLWWW-E-CD-NU-Base-3	0.011	0.635
	SLWWW-E-CO-NU-Base-4	0.011	0.636
MPII-E-CO-NU-Base-3	SLWWW-E-CD-NU-Base-3	0.045	0.574
	SLWWW-E-CO-NU-Base-4	0.044	0.575

Table 14. Kendall's τ values with 95% CIs (19 English runs).

	Mean Q@10	Mean nERR@10
Mean nDCG@10	0.869 [0.750, 0.989]	0.731 [0.537, 0.925]
Mean Q@10	-	0.745 [0.521, 0.969]

8 Conclusions

This overview summarizes the task settings, dataset, and evaluation methodology of NTCIR-14 WWW-2 task and report the official results of this task. In the Chinese subtask, all the teams outperforms the BM25 baseline run significantly but the difference between them are not significant. In the English subtask, the runs from THUIR are the most effective. Further discussions of the NTCIR-14 WWW-2 Task will be given in our Final Report [2].

References

1. Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C., Xu, J.: Overview of the NTCIR-13 we want web task. In: Proceedings of NTCIR-13. pp. 394–401 (2017), http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf
2. Mao, J., Luo, C., Liu, Y., Xiao, P., Sakai, T., Dou, Z.: Final report of the NTCIR-14 we want web task. In: LNCS. p. to appear (2019)
3. Sakai, T.: Metrics, statistics, tests. In: PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173). pp. 116–163 (2014)
4. Sakai, T.: Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power. Springer (2018), https://link.springer.com/book/10.1007/978-981-13-1199-4