Recommender Systems Evaluations: Offline, Online, Time and A/A Test

Gebrekirstos G. Gebremeskel[1] and Arjen P. de Vries[2]

[1] Information Access, CWI, Amsterdam,
gebre@cwi.nl
[2] Radboud University
arjen@acm.org

Abstract. We present a comparison of recommender systems algorithms along four dimensions. The first dimension is offline evaluation where we compare the performance of our algorithms in an offline setting. The second dimension is online evaluation where we deploy recommender algorithms online with a view to comparing their performance patterns. The third dimension is time, where we compare our algorithms in two different years: 2015 and 2016. The fourth dimension is the quantification of the effect of non-algorithmic factors on the performance of an online recommender system by using an A/A test. We then analyze the performance similarities and differences along these dimensions in an attempt to draw meaningful patterns and conclusions.

1 Introduction

Recommender systems can be evaluated offline or online. The purpose of recommender system evaluation is to select algorithms for use in a production setting. Offline evaluations test the effectiveness of recommender system algorithms on a certain dataset. Online evaluation attempts to evaluate recommender systems by a method called A/B testing where a part of users are served by recommender system A and the another part of users by recommender system B. The recommender system that achieves a higher score according to a chosen metric (for example, Click-Through-Rate) is chosen as a better recommender system, given other factors such as latency and complexity are comparable.

The purpose of offline evaluation is to select recommender systems for deployment online. Offline evaluations are easier and reproducible. But do offline evaluations predict online performance behaviors and trends? Do the absolute performances of algorithms offline hold online too? Do the relative rankings of algorithms according to offline evaluation hold online too? How do offline evaluation compare and contrast with online evaluations?

CLEF NewsREEL [5], a campaign-like news recommendation evaluation, provides opportunities to investigate recommender system performance from several angles. CLEF NewsREEL 2016 campaign, in particular, is focused on comparing recommender system performance in online and offline settings [9]. CLEF NewsREEL 2016 provides two tasks: Benchmark News Recommendations in a Living

Lab (Task 1) which enables evaluation of systems in a production setting [6], and Benchmarking News Recommendations in a Simulated Environment (Task 2) which enables the evaluation of systems in a simulated (offline) setting using dataset collected from the online interactions.

In 2015, we participated in Task 1. In 2016, we participated in both CLEF NewsREEL tasks. In this working notes, we report both offline and online evaluations and how they relate to each other. We also present the challenges of online evaluation from the dimensions of time and non-algorithmic causes of performance differences. On the time dimension, we specifically investigate online performances behaviors in 2015 and 2016, and on the dimension of non-algorithmic causes of performance differences, we employ an A/A test where we run two instances of the same algorithm to gauge the extent of performance difference resulting from non-algorithmic causes.

2 Tasks and Objectives

The objective of our participation this year is to investigate performance behaviors of recommender system along several dimensions. We are interested in differences and similarities in offline and online performances, the variations in performance over time, and the estimation of performance differences caused by non-algorithmic factors in the online recommender system evaluations. This work can be see as an extension of studies that have previously investigated the differences between offline and online recommender system evaluations [2, 1, 11].

In 2015, we participated in CLEF NewsREEL News Recommendations Evaluation, the task of Benchmark News Recommendations in a Living Lab [6]. We reported the presence of substantial non-algorithmic factors that cause two instances of the same algorithm to end up having statistically significant performance differences. The results are presented in [4]. This year, we run four of our 2015 recommender systems without change. This allows us to compare the performance of the systems in 2015 and 2016. In 2016, we participated also in Task 2, which allows us to evaluate the recommender systems in a simulated environment and then compare the offline performance measurements with the corresponding online performance measurements. In this report, we present the results of these evaluations along the four dimensions and highlight similarities, differences and patterns or the lack thereof.

For the study of the effect of non-algorithmic factors on online recommender system performances, we run two instances of the same news recommender algorithm with the view to quantifying the extent of performance differences. To compare the online and offline performance behaviors, we conduct offline evaluations on a snapshot of a dataset collected from the same system. To investigate performance in the dimension of time, we rerun last year's recommender systems. This means that we can compare the performance of the recommender systems in 2016 with their corresponding performance in 2015.

The four recommender systems are two instances of **Recency**, one instance of **GeoRec** and one instance of **RecencyRandom**. **Recency** keeps the 100

most recently viewed items for each publisher, and upon recommendation request, the most recently read (clicked) are recommended. **GeoRec** is a modification of the Recency recommender to diversify the recommendations by taking into account the users' geographic context and estimated interest in local news. **RecencyRandom** recommends items randomly selected from the *100* most recently viewed items. For a detailed description of the algorithms, refer to [4].

3 Results and Discussions

We present the results and analysis from the different dimensions here. In 2015, the recommender systems ran from 2015-04-12 to 2015-07-06, a total of 86 days. RecencyRandom started 12 days later in 2015. In 2016, the systems ran from 2016-02-22 to 2016-05-21, a total of 70 days. We present three types of results for 2016: the daily performances, the incremental performances, and the cumulative performances. The plot for the daily performance is presented in Figure 1. From the plot, we observe large variations between the maximum and minimum performance measurements of the tested recommender systems; the minimum value equals 0 in every test, while the maximum varies between 12.5% for Recency2, 5.6% for GeoRec, 4.3% for RecencyRandom, and 4.2% for Recency. The highest performance measurements all occurred between the 18^{th} day from the start of our participation (2016-03-10) and the 31^{st} day. The highest scores of Recency2, and GeoRec occurred on March 21^{nd}, for RecencyRandom on March 20^{th} and GeoRec on March 10^{th}. We do not have a plausible explanation why the evaluation resulted in increased performance during that period, nor why the highest scores for two systems occurred on those two days in March. We did however observe quite a reduction in the number of recommendation requests issued in the period when the systems showed increased performance scores (such that minor variations would lead to larger normalized performance differences than in the rest of the evaluation period). Some of the systems have no reported results between 2016-03-24 and 2016-04-05; if this reduction in the number of recommendation requests is the same for all teams and systems who participated, the lower number of recommendations paired by an increase in CTR could indicate that users are more likely to click on recommendations when recommendations are offered sparsely. If this is the case, it might suggest further investigation into the relationship of the number of recommendations and user responses.

Figure 2 for 2015 and in Figure 3 for 2016 plot the performance measurements as the systems progress on a daily basis, which we call incremental performance. The cumulative number of requests, clicks and CTR scores of the systems in both years are presented in Table 1. The cumulative performance measurements remain below 1% for all systems. The maximum performance differences observed between the systems equal 0.16% in 2015 and 0.07% in 2016.

From the plots in Figure 2 and Figure 3 and the cumulative performance measurements in Table 1, we observe that the performance measurements of the different systems vary. Are these performance variations between the different systems also statistically significant? We look at statistical significance on a

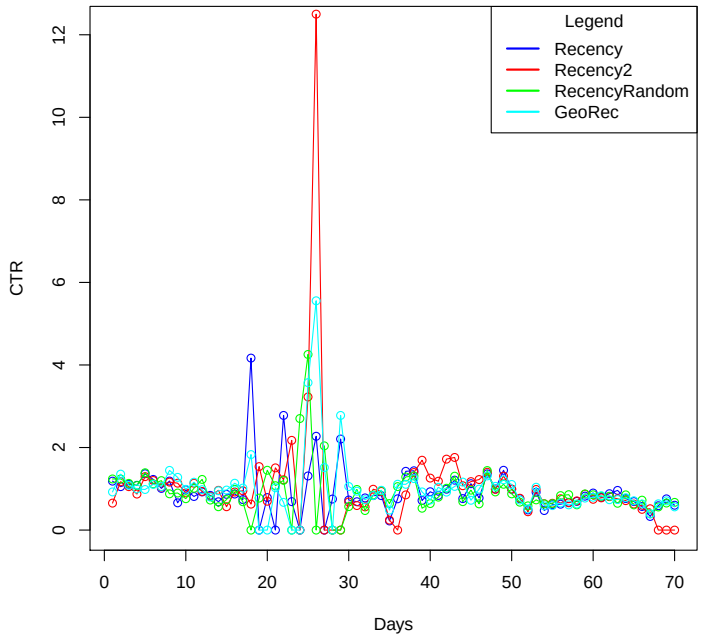

Fig. 1. Daily CTR performance measurements of the four online recommender systems in 2016. Notice the large differences between the days and the unusual increase in CTR between the 18^{th} and 31^{st} day.

Table 1. Number of requests, number of clicks and CTR scores of four systems in 2015 and 2016.

	2015			2016		
Algorithms	Requests	Clicks	CTR(%)	Requests	Clicks	CTR(%)
Recency	90663	870	0.96	450332	3741	0.83
Recency2	88063	810	0.92	398162	3589	0.90
RecencyRandom	73969	596	0.80	438850	3623	0.83
GeoRec	88543	847	0.96	448819	3785	0.84

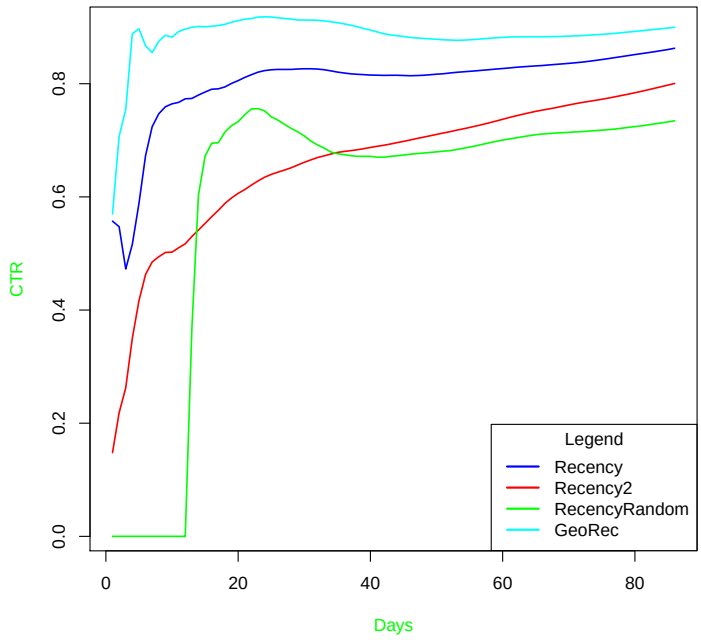

Fig. 2. CTR performance of the four online recommender systems (2015).

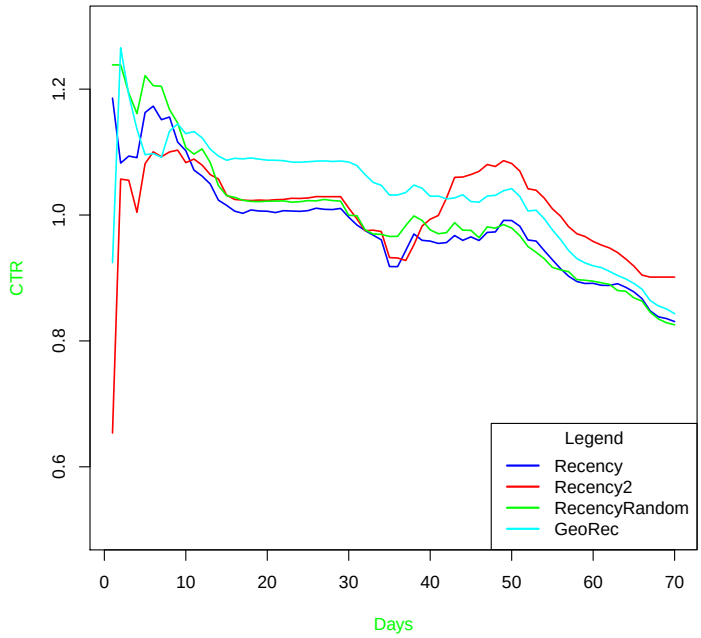

Fig. 3. Cumulative CTR performance measurements of the four online systems, as they progress on a daily basis in 2016.

daily basis after the 14^{th} day, which is considered the average time within which industry A/B tests are conducted. To compute statistical significance, we used Python module of Thumbtack's Abba, a test for binomial experiments [7] (for a description of the implementation, please refer to [8])

We perform statistical significance tests on a daily basis to simulate the notion of an experimenter checking whether one system is better than the other at the end of every day. In testing for statistical significance on a daily basis, we seek an answer to the question: 'On how many days would an experimenter seeking to select the better system find out that one system is significantly different from the chosen baseline?' We investigate this under two baselines: Recency2, and RecencyRandom. Tables 2 and 3 present the actual number of days and the percentage of days on which significant performance differences were observed.

Table 2. Statistical significance when comparing Recency2 to the baseline.

Algorithms	2015		2016	
	No Sig Results	%	No Sig Results	%
Recency	2	2.7	27	47.4
GeoRec	25	34.3	8	14

Table 3. Statistical significance when comparing RecencyRandom to the baseline.

Algorithms	2015		2016	
	No Sig Results	%	No Sig Results	%
Recency	20	27.4%	0	0
GeoRec	41	56.2	5%	8.8

Next, we looked into the error notifications received by our recommender systems in the 2016 period. The error types and counts for each system are presented in Table 4. Three types of errors occurred, the highest number for the RecencyRandom recommender. According to the ORP documentation[3], error code 408 corresponds to connection timeouts, error code 442 to invalid format of recommendation responses, while error 455 is not described.

We aggregated error messages by day. Out of the 70 days, Recency received error notifications on 16 days, Recency2 on 19 days, GeoRec on 24 days, and RecencyRandom on 51 days. All systems received high number of error messages on specific days, especially on 2016-04-07 and 2016-04-08. While we do not know the explanation for errors on especially those days, we did observe that most of the high-error days seem to be those that correspond to the beginning of the start of the systems, or at the beginning of a change of load (from low to high).

[3] http://orp.plista.com/documentation/download

Why the RecencyRandom recommender received a high number of 'invalidly formatted' responses is not clear, because the format is the same as for other systems. The main difference between RecencyRandom and the other systems we deployed is that it has a lower response time, and we suspect the high number of errors to be related to its lower response rate.

Table 4. Count of errors messages received by our recommender systems in 2016. Error code 408 is for connection timeout, error code 442 is for invalid format of recommendation response and Error 455 is not described. RecencyRandom has the highest number of errors.

.

Error Types	Recency	Recency2	RecencyRandom	GeoRec
408 (Connection Timeout)	40	118 40	14	159
442 (Invalid-Format)	1377	1390	26608	1360
455	281	217	348	252
Total	1698	1725	26970	1771

3.1 Online Evaluation

In the online evaluation, or Benchmark News Recommendations in a Living Lab (Task 1) as it is called in CLEF NewsREEL, we investigate recommender systems in two dimensions. One dimension is time where we compare and contrast the performances of our systems in 2015 and 2016. The second dimension is an A/A test where we attempt to study non-algorithmic effects on the performance of systems. Each of the dimensions are discussed in the following subsections.

Time Dimension: Performances in 2015 and 2016 Participation in the Lab in 2015 and in 2016 gives us the opportunity to study the evaluation results from a time dimension. We compare the systems both in terms of their absolute and relative performance (in terms of their rankings). To compare the absolute performances, we used the 2015 instances of the recommender systems as baselines, and the corresponding 2016 instances as alternatives. The performance measurements of the Recency and GeoRec instances of 2016 were significantly different from the performance measurements of the Recency and GeoRec instances of 2015 with a P-values of 0.0001 and 0.0009 respectively. The 2015 instances of Recency2 and RecencyRandom were not significantly different from their corresponding instances in 2016.

In 2015, Recency2 ranked third, but in 2016, it ranked first. In 2015, almost all systems started from a lower CTR performance, and slowly increased towards the end where the performance measurements stabilized (see Figure 3). In 2016, however, the evaluation results of the systems reached its high at the beginning, and then decreased steadily towards the end, except for recommender Recency2, which showed an increase after the first half of its deployment and then decreased

(see Figure 3). In 2016, the performance measurements seemed to continue to decrease, and not converge to a stable result like in 2015.

When we compare the number of days for which the results are significantly different according to the statistical test (see Table 2 and Table 3), we observe that there is no consistency. In 2015, there were two days (2.7%) on which significant performance differences were observed between Recency and Recency2 while there are 25 days (34.3%) on which significant performance difference between GeoRec and Recency2. In 2016, Recency has shown 47.4% of the time significant performance, and GeoRec only 14%. When using RecencyRandom as a baseline, Recency has registered significant performance differences 27.4% of the time in 2015, and 0% in 2016. GeoRec has 56.2% in 2015 and 8.8% in 2016.[4]

We conclude that it is different to generalize the performance measurements over time. The patterns observed in 2015 and 2016 vary widely, both in terms of absolute and relative performance, irrespective of the baseline considered. The implication is that one can not rely on the absolute and relative rankings of recommender systems at one time for a similar job in another time. The systems have not changed between the two evaluations. The differences in evaluation results, therefore, can only be attributed to the setting in which the systems are deployed. It is possible that the presentation of recommendation items by the publishers, the users and content of the news publishers might have undergone changes which can then affect the performances in the two years, but we cannot be certain without more in-depth analysis.

A/A Testing In both 2015 and 2016, two of our systems were instances of the same algorithm. The two instances were run from the same computer; the only differences between them were the port numbers by which they communicated with CLEF NewsREEL's ORP[5]. The purpose of running two instances of the same algorithm is to quantify the level of performance differences due to non-algorithmic causes. From the participant's perspective, performance variation between two instances of the same algorithm can be seen pure luck. The extent of performance differences between the instances can be seen as also happening between the performances of the other systems. We can consider that he performance difference due to the effectiveness of the algorithms is therefore the overall performance minus the maximum performance difference between the performances of the two instances.

The results of the two instances (Recency and recency2) can be seen in 1, the incremental plots (Figure 2 and Figure 3. The cumulative performances on the 86^{th} day of the deployment in 2015 showed no significant difference. In 2016, however, Recency2 showed a significant performance over Recency with a P-value of 0.0005. Checking for statistical significance on a daily basis after the 14^{th} day (see 2), in 2015, there were 2 days (2.7%) on which the two instances

[4] We would like to mention a correction here over the reported statistical significance score of GeoRec in 2015. It was reported that Georec did not achieve any significant performance over Recency2 [3], which was an error in calculation.

[5] http://orp.plista.com/

differed significantly. In 2016, however, the number of days was extremely higher, a total of 27 days (47.4%). This is interesting for two reasons: 1) the fact that two instances can end up having statistically significant performance differences and 2) that the significant difference occurred. In 2016, one instance achieved significant performance differences over the other instance for almost half of the time.

3.2 Offline Evaluation

We present evaluations conducted offline, or in Benchmarking News Recommendations in a Simulated Environment (Task 2), as it is called in CLEF NewsREEL. Evaluation in Task 2 differs from other offline evaluation setups in that Task 2 actually simulates the online evaluation setting for each of the systems. Usually, systems are selected on the basis of offline evaluation and deployed online. Other things such as complexity and latency being equal, there is this implicit assumption that the relative offline performances of systems holds online too. That is that if System one has performed better than system two in an offline evaluation, it is assumed that the same rank holds when the two algorithms are deployed online. In this section, we investigate whether this assumption holds by comparing the offline evaluation results of the algorithms in Task 1 with their online results.

Task 2 of CLEF NewsREEL provides a reproducible environment for participants to evaluate their algorithms in a simulated environment that uses user-item interaction dataset recorded from the online interactions [10]. In the simulated environment, a recommendation is successful if the user has viewed or clicked on the recommendations. This is different from Task 1 (online evaluation) where a recommendation is a success only if the recommendation is clicked. The performances of our algorithms in the simulated evaluation are presented in Table 5. The plots as they progress on a daily basis are presented in Figure 4. In this evaluation, Recency leads followed by Georec and then RecencyRandom. Using RecencyRandom as a baseline, there was no significant performance difference in both Recency and GeoRec. Comparing the ranking with those of the systems in Task 2, there is no consistency. We conclude that the relative offline performance measurements do not generalize to those online, much less the absolute performance.

From Table 5, we observe that only RecencyRandom has invalid responses. We also observed that RecencyRandom has higher error messages and lower performance in Task 1. To understand why, we looked at the response times of the systems under extreme load. The mean, min, max and standard deviations of the response times of the three systems are presented in Table 6. We observe that RecencyRandom has the slowest response time followed by GeoRec. We have also plotted the number of recommendations within 250 milliseconds in Figure 5. Here too, we observe the lowest reponse times for RecencyRandom (attributed to the randomization before selecting recommendation items). When we look at the publisher-level breakdown of the recommendation response in Table 5, we see that RecencyRandom has invalid responses for two publishers, but for publisher

Table 5. The performances of our algorithms in simulated evaluation (Task 2). For each system, there are the number of correct clicks (clicks), the number of requests, and the CTR (clicks*1000/requests) and the number of invalid responses (Inv). Results for publishers http://www.cio.de (13554), http://www.gulli.com (694), http://www.tagesspiegel.de (1677), sport1 (35774) and All are shown in the table.

Publishers	Recency				RecencyRandom				GeoRec			
	Click	Request	Inv	CTR	Click	Request	Inv	CTR	Click	Request	Inv	CTR
13554	0	21504	0	0	0	12798	1451	0	0	21504	0	0
694	13	4337	0	2	3	4347	0	0	13	4337	0	2
1677	69	46101	0	1	0	0	7695	0	69	46101	0	1
35774	3489	518367	0	6	2297	519559	0	4	3445	518411	0	6
All	3571	590309	0	6	2300	536704	9146	3	3527	590353	0	5

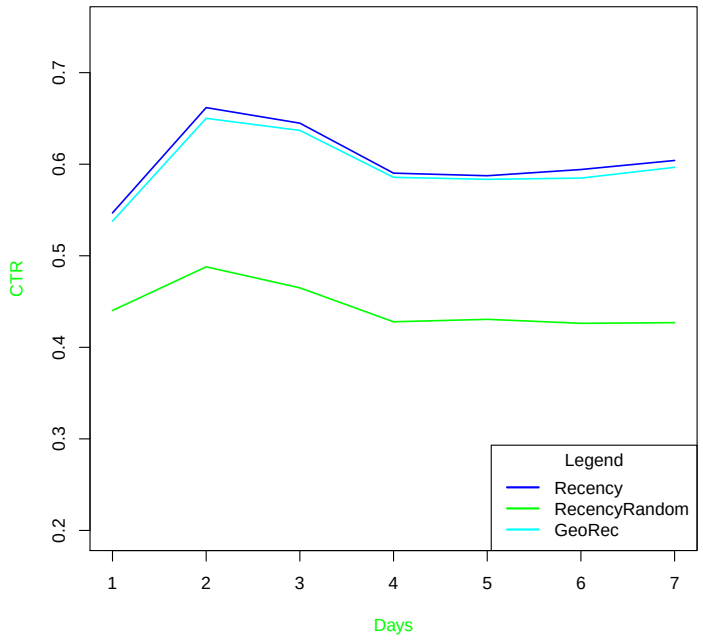

Fig. 4. CTR performance measurements of the three offline systems as they progress on a daily basis.

Tagesspiegel (1677), all its recommendations are invalid. In the offline evaluation, invalid response means that the response generates an exception during parsing. We looked into the recommendation responses of RecencyRandom, and compared the response for publisher 694 and 1677. Almost all item responses for publisher 1677 were empty, which we assume to be related with the extreme load.

Table 6. Response times in milliseconds of the recommender systems. RecencyRandom has the lowest response time.

	Mean	Min	Max	stDev
Recency	9.057	0.0	2530.0	41.619
RecencyRandom	83.868	1.0	5380.0	319.463
GeoRec	11.549	1.0	2320.0	56.570

4 Discussion

Our systems are very similar to each other, in that they are slight modifications of each other. This means that it is expected that their performances do not vary much. We have analyzed the performance of our systems from the dimensions of online, offline, and time. We have also investigated the the extent of performance difference due to non-algorithmic causes in online evaluation by running two instances of the same algorithms.

We have observed substantial variation along the four dimensions. The performance measurements in both absolute and relative sense varied significantly in 2015 and in 2016. More surprisingly, the two instances of the same algorithm varied significantly both in the two years and within the same year. This is surprising and indicates how challenging it is to evaluate algorithms online. In the online evaluation, non-algorithmic and non-functional factors impact performance measurements. Non-algorithmic factors include variations in users and items that systems deal with, and the variations in recommendation requests. Non-functional factors include response times and network problems. The performance difference between the two instances of the same algorithms can be considered to reflect the impact of non-algorithmic and non-functional factors on performance. It can then be subtracted from the performances of online algorithms before they are compared with baselines and each other. This can be seen as a way of discounting the randomness in online system evaluation from affecting comparisons.

The implication of the lack of pattern in the performance of the systems across time and baselines, and more specially the performance differences between the two instances of the same algorithm highlights the challenge of comparing systems online on the basis of statistical significance tests alone. The results call for caution in the comparison of systems online where user-item dynamism, operational decision choices and non-functional factors all play roles

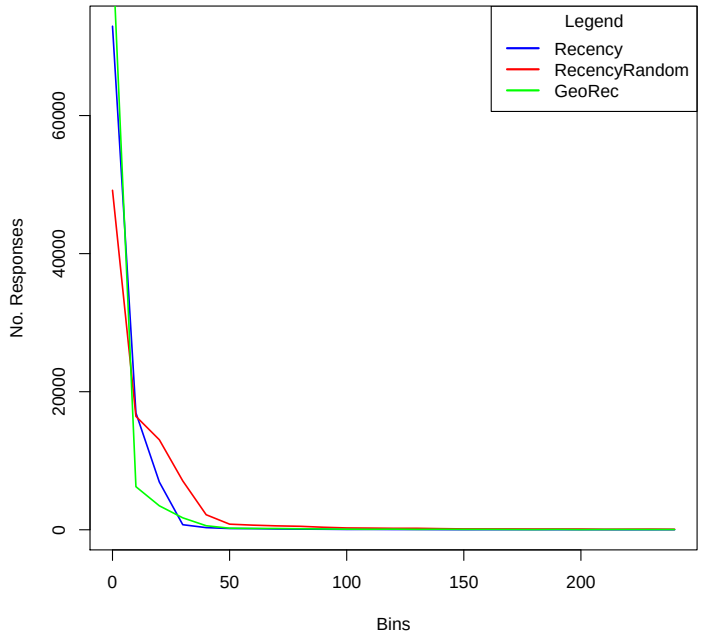

Fig. 5. Number of recommendation responses against response times in milliseconds, for the systems in Task 2.

in causing performance differences that are not due to the effectiveness of the algorithms.

4.1 Comparison With Other Teams

Let us also compare evaluation results for our systems to those of other teams that participated in 2016 CLEF NewsREEL's Task 1, based on the results over the period between 28 April and 20 May (provided by CLEF NewsREEL). The plot of the team ranking as provided by CLEF NewsREEL is provided in Figure 6. We examined whether the performance of the best performing systems from the teams that are ranked above us were significantly different from ours. Only the ABC's and Artificial Intelligence's systems were significantly different from Recency2 (our best performing system for 2016).

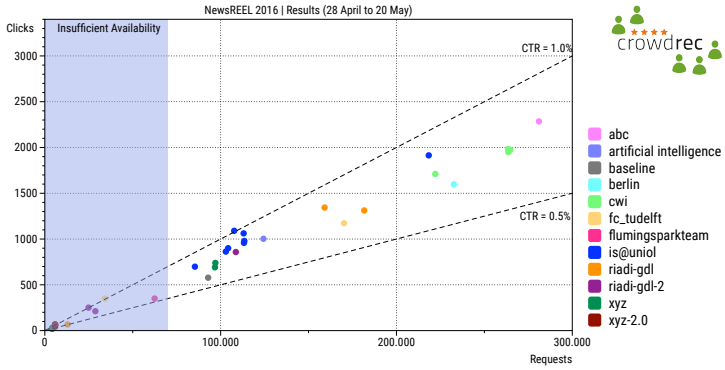

Fig. 6. The rankings of the 2016 teams that participated in the CLEF NewsREEL challenge. The plot was provided by CLEF NewsREEL.

5 Conclusions

We set out to investigate the performance of recommender system algorithms online, offline, and in two separate periods. The recommender systems' performances in different dimensions indicate that there is no consistency. The offline performances were not predictive of the online performances in both absolute and relative sense. Also the performance measurements of the systems in 2015 were not predictive of those in 2016, both in relative and absolute sense. Our systems are slight variations of the same algorithm, and yet the performances varied in all dimensions. We conclude that we should be cautious in interpreting the results of performance differences, especially considering the differences observed between the two instances of the same algorithm.

Acknowledgements

This research was partially supported by COMMIT project Infiniti.

References

1. J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 7–14. ACM, 2013.
2. F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo. ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176. ACM, 2014.
3. G. Gebremeskel and A. P. de Vries. The degree of randomness in a live recommender systems evaluation. In *Working Notes for CLEF 2015 Conference, Toulouse, France. CEUR*, 2015.
4. G. G. Gebremeskel and A. P. de Vries. Random performance differences between online recommender system algorithms. In N. Fuhr, P. Quaresma, B. Larsen, T. Goncalves, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, vora, Portugal, September 5-8, 2016*. Springer, 2016.
5. F. Hopfgartner, T. Brodt, J. Seiler, B. Kille, A. Lommatzsch, M. Larson, R. Turrin, and A. Serény. Benchmarking news recommendations: the clef newsreel use case. In *SIGIR Forum*, volume 49, pages 129–136. ACM Special Interest Group, 2015.
6. F. Hopfgartner, B. Kille, A. Lommatzsch, T. Plumbaum, T. Brodt, and T. Heintz. Benchmarking news recommendations in a living lab. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 250–267. Springer, 2014.
7. S. Howard. Abba 0.1.0. https://pypi.python.org/pypi/ABBA/0.1.0. Accessed: 2016-06-18.
8. S. Howard. Abba: Frequently asked questions. https://www.thumbtack.com/labs/abba/. Accessed: 2016-06-18.
9. B. Kille, A. Lommatzsch, G. Gebremeskel, F. Hopfgartner, M. Larson, J. Seiler, D. Malagoli, A. Sereny, T. Brodt, and A. de Vries. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In N. Fuhr, P. Quaresma, B. Larsen, T. Goncalves, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, vora, Portugal, September 5-8, 2016*. Springer, 2016.
10. B. Kille, A. Lommatzsch, R. Turrin, A. Serény, M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Stream-based recommendations: Online and offline evaluation as a service. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 497–517. Springer, 2015.
11. E. Kirshenbaum, G. Forman, and M. Dugan. A live comparison of methods for personalized article recommendation at forbes. com. In *Machine Learning and Knowledge Discovery in Databases*, pages 51–66. Springer, 2012.