The Degree of Randomness in a Live Recommender Systems Evaluation

Gebrekirstos G. Gebremeskel and Arjen P. de Vries

Information Access, CWI, Amsterdam, Science Park 123, 1098 XG Amsterdam, Netherlands gebre@cwi.nl arjen@acm.org

Abstract. This report describes our participation in the CLEF NEWS-REEL News Recommendation Evaluation Lab. We report and analyze experiments conducted to study two goals. One goal is to study the effect of randomness in the evaluation of algorithms. The second goal is to see whether geographic information can help in improving the performance of news recommendation.

1 Tasks and Objectives

In this report we describe the results of experiments we conducted during our participation in the CLEF NEWSREEL News Recommendations Evaluation Lab, the task of Benchmark News Recommendations in a Living Lab [5]. The experiments have the following goals. The first goal is to investigate the effect of system and/or user behavior randomness on recommender systems evaluations. This randomness refers to the selection of algorithms for providing recommendation request by the Plista framework and the user click behavior which can be influenced by the situation of the user. The second is to study whether geographic information can play a role in news recommendation systems.

The literature on recommender systems shows that offline and online recommender system evaluations do not concur with each other [3, 1, 6]. This is to say that recommender systems behave differently in offline and online evaluations, both in terms of absolute and relative performance. This has a serious implication for recommender system research, because the whole point of offline evaluation is the assumption that at least the relative performance of recommender systems is indicative of their relative online performance and thus an important step for selecting algorithms that can be deployed in a live recommendation setting.

Many reasons can be mentioned for the disparate performances of offline and online evaluations. The three most important papers [3, 1, 6] that compare online and offline evaluations use different datasets for offline evaluation and online evaluation. It is possible that the difference in performance may be attributed to this difference in dataset. The literature lists some reasons for the disparate performance of recommender systems in offline and online evaluations [6, 7]. The first reason is that offline evaluations can measure only accuracy; they do not take the user behavior into account. The second reason is that offline data-sets are incomplete and imperfect, and thus recommender systems are being evaluated based on incomplete dataset.

The first reason, that is, that offline evaluations can measure only accuracy and thus can not take user behavior into account can be understood in two ways. One way is that online evaluations can be influenced by adapted implementations of the algorithms that were used in offline evaluation, to take user behavior into account, in which case it becomes a different implementation. The other way to understand it is that online evaluations happen in an environment that involves user behavior and thus their performances can be influenced by the system and/or user behavior "randomness". The first interpretation implies that offline and online algorithms are differently implemented. The second interpretation implies that the difference in offline and online evaluations can be due to the system and/or user behavior randomness that the online evaluations are subjected to. We attempt to explore this randomness in our participation

The motivation for the second goal comes from a descriptive study we conducted on a Plista dataset we collected in our previous participation. In the study, there were two findings [4]. One is that there is a substantial difference in the geographical distribution of the readerships of traditional news portals, and the second is that within the same portal, the geographical distribution of the readerships of the local news category and the rest of the categories shows substantial difference. In our experiments, we tried to exploit the second finding to improve news recommendation.

2 Approach

For the study of the effect of system and/or user behavior randomness on recommender system evaluations, we run two instances of the same news recommender algorithm with the view to measuring the effect of "randomness" involved in news recommendation clicks. For exploiting geographical information for news recommendation, we employ a recommendation diversification approach. Specifically, we include geographically relevant recommendation into the recommendation list, thus diversifying the recommendations in favor of geographic relevance

2.1 Algorithms

We experimented with five algorithms, all of them modifications of the recency algorithm. The recency algorithm takes into account recency and popularity of an item, and it has been shown to be a strong baseline in previous online evaluations. The algorithmic variations that we experimented with are listed below.

Recency: This algorithm keeps the 100 most recently viewed items for each publisher in consideration for being recommended to the user. For any recommendation request, the items that are , at the time of the recommendation request, the most recently read (clicked)) are the ones that are recommended.

We run two instances of this algorithm to get a sense of the randomness involved in the selection of algorithms by the Plista framework [2] and/or clicks on recommendations by users.

GeoRec: The geographical recommender takes the geographical region (states to be specific) of users and the local category of news items into account when generating recommendations. We generate two sets of recommendations, one by the recency recommender and one by a purely geographical recommender. For the purely geographical recommender, we take the 100 most recently viewed items and sort them according to their geographic conditional likelihood scores generated by Equation 1.

$$r_{u_a,i_k} = P(c_{i_k}|g_{u_a}) \tag{1}$$

Where c_{i_k} is the local category of item i_k . An item is either in the local category, or not. g_{u_a} is the state-level geographical information of the user u_a , that is, the state the user belongs to. Then, we take top twice the number of requested recommendations as recommendation of the geographic recommender. For the recency, we take the requested number of recommendations from the from the most recently viewed items. Then we intersect the two sets of recommendations. If the number of elements in the intersection is not equal with the requested number of recommendations, we get half - 1 from purely geographic recommender.

GeoRecHistory: This is a modification of the GeoRec recommender; it excludes from the recommendation list those items that the user has already visited.

RecencyRandom: This recommender recommends items randomly selected from the circular-buffer. This was started a bit late, and we used it a baseline algorithm.

The two instances of Recency and the GeoREc algorithms were run for a consecutive period of 53 days, from 2015-04-12 to 2015-06-03. The RecencyRandom algorithm was started 12 days later on 2015-04-24.

3 Results and Discussions

We present two types of performance scores: cumulative and daily click-through rates (CTR). The cumulative CTR is presented in Table 1. We see that the performances are all close to each other. However, if we rank them, we see that the GeoRec recommender followed by Recency followed by GeoRecHistory end up as the first three best performing algorithms. The daily performances of the algorithms are shown in Figure 1, and the cumulative CTR as a function of the number of days is shown in figure 2.

From the daily (figure 1) and cumulative (figure 2) plots, we see that the performance of any of the algorithms varies greatly. In the cumulative plot, we see that Recency and Recency2 performed differently for a long time, but generally getting closer towards each other until some point after which they seem to stabilize. If one was continuously monitoring the performances of this

Algorithms	Requests (Clicks C	TR(%)					
Recency	$56,\!350$	478	0.85					
Recency2	$53,\!863$	420	0.78					
GeoRec	$54,\!338$	470	0.86					
GeoRecHistory	47,001	395	0.84					
RecencyRandom	39,616	283	0.71					
Table 1. Live performance of different algorithms								

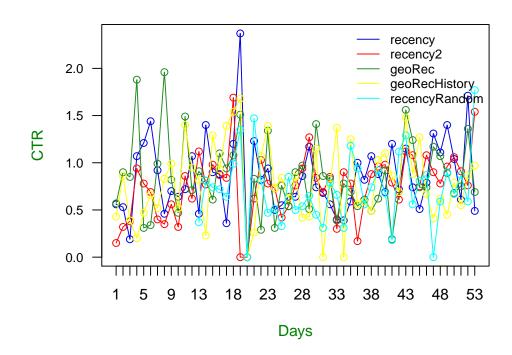


Fig. 1. The daily CTR performances of the five algorithms

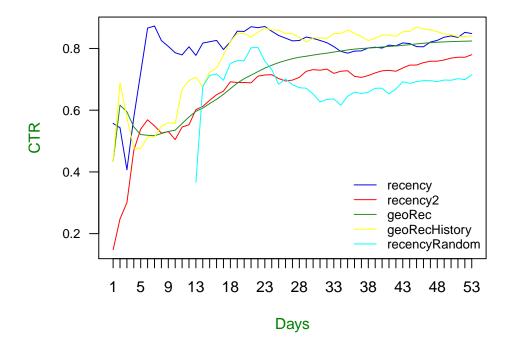


Fig. 2. The cumulative CTR performances of the five algorithms as they progress on a daily basis

two algorithms, then one would say that the Recency algorithm is better than Recency2 algorithm. This would happen, at least some times, even if one employs statistical significance tests.

Let's assume that the experimenter was peeking at the experiments everyday to make a decision on which algorithm is better. How many times would the experimenter declare statistically significant differences between the different algorithms? We examined this by using two baselines: the random recommender (RecencyRandom) and Recency2. The results, when using the RecencyRandom recommender as a baseline, are given in Table 2. Similarly, the results for the baseline of Recency2 are given in Table 3.

We see that, when RecencyRandom is used a baseline, Recency, GeoRec and GeoRecHistory achieve statistically significant performance almost more than half of the time. With Recency2 as baseline, we see that only Recency achieves a statistically significant result twice.

	Algorithm	Days of Significance Percentage ((%)
	recency	16 4	10.0
	geoRec	23 5	57.5
	geoRecHistory	24 6	0.0
2	Statistical sign	if comes over the baseline of Deer	m or Do

Table 2. Statistical significance over the baseline of RecencyRandom

	Algorithm	Days of Significance Percentage	e (%)
	recency	2	5
	geoRec	0	0
	geoRecHistory	1	2.5
~1	o 9 Statistical	cirrificance over the baceline of	f Dogon

 Table 3. Statistical significance over the baseline of Recency2

What is truly interesting is that the same algorithm (Recency) can end up achieving statistically significant performance over another instance of itself (Recency2). The two instances of the same algorithm show so big difference in performance that there is a big chance of concluding one is better than itself. The implication of this raises questions on improvement of performance of algorithms that involve live users in general. Specifically, to what extent is one algorithm's improvement over another a real improvement? It seems to us that statistical significance tests alone are not sufficiently indicative of performance improvements. Given the randomness involved in user's clicks on recommendations (and the system), a certain range of performance difference is not worth taking seriously. Studies that involve users must account for some level of randomness, in addition to statistical significance tests.

4 Conclusions and Perspectives for Future Work

We set out to study two factors in news recommendation: geographical information, and user and/or system randomness in news recommendation clicks. Although the geographical recommendation did not show any strikingly useful improvement over the recency algorithm, the user-system randomness seems to indicate that care must be taken to take into account some degree of randomness in recommender systems evaluation that involve users in a live setting, in addition to statistical significance tests. In the future, we would like to extend this work to include a better statistical testing system that can account for this randomness. We also would like to compare the offline and online evaluations, Specifically, we would like run one algorithm on the logs of another to see to what extent they can replace each other.

References

- 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.
- T. Brodt and F. Hopfgartner. Shedding light on a living lab: the clef newsreel open recommendation platform. In *Proceedings of the 5th Information Interaction* in Context Symposium, pages 223–226. ACM, 2014.
- 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.
- G. G. Gebremeskel and A. P. de Vries. The role of geographic information in news consumption. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015.
- 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.
- 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.
- S. M. McNee, N. Kapoor, and J. A. Konstan. Don't look stupid: avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th anniversary* conference on Computer supported cooperative work, pages 171–180. ACM, 2006.