Pseudo-online Measurement of Retrieval Recall for Job Recommendations - A case study at Indeed

Liyasi Wu, Yi Wei Pang and Warren Cai

Indeed.com

Abstract

A typical large-scale recommender system in production involves several key stages: retrieval, filtering, scoring, and ordering. As the process unfolds, the quantity of recommendations decreases, ideally enhancing their quality. While many metrics such as NDCG, recall@k, and precision@k have been employed in offline evaluations, there have been observations that improvements in such offline metrics do not lead to gains in common online metrics, such as click-through and conversion rates, especially at the early stages of the recommender system. In this paper, we introduce a case study at Indeed where we designed a pseudo-online metric adapted from the traditional recall@k to measure the effectiveness of the retrieval stage within our job recommender system.

Keywords

Recommendation system, Retrieval efficiency, Recall, Funnel analysis

1. Introduction

Indeed sends millions of job recommendations daily to job seekers through various channels (e.g. onsite job recommendation feed, job recommendation emails). These job matches are generated by leveraging job seekers' profile, behavioral data on Indeed, and job data to predict the most relevant jobs for each individual. Following an industry-wide pattern, our recommendation system utilizes a multi-stage process that includes retrieval, filtering, scoring, and ordering [1]. The retrieval stage consists of multiple services, termed as match providers, each employing different strategies to efficiently retrieve matches. Matches retrieved are processed and filtered through layers of business logic before the final ranking stage where they are scored and ranked before the top N matches are selected to be sent to the job seeker.

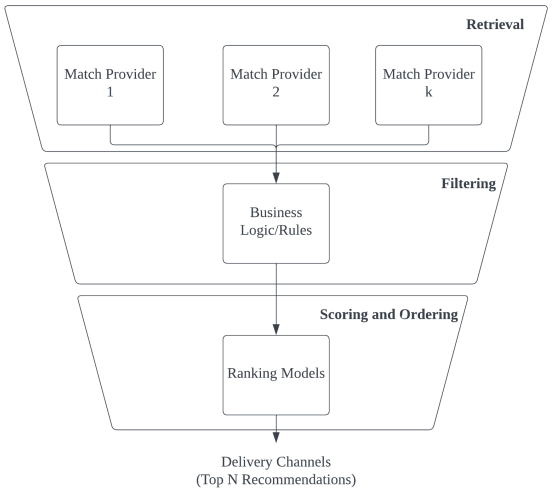

Figure 1: Multi-stage Recommender System

To continuously measure the effectiveness of our job recommendation system, business metrics such as click-through rate (apply button clicks) and conversion rate (job applications submitted) have been used for performance tracking and A/B experiments in online settings. While simple and effective, they do not consider intermediate steps in the recommendation process, and are often heavily influenced by the final ranking model. Improvements made at the retrieval stage could be undermined due to biasness of the final ranking models and as observed in other studies [2, 3], we have also encountered scenarios where offline evaluation results of individual retrieval strategies do not translate to improvement in online performance. It is imperative to find better ways to evaluate and measure the effectiveness of our retrieval strategies.

Recall@k is a popular metric used to measure predictive quality in offline evaluations [4] where it considers the proportion of relevant matches produced by a recommendation algorithm. It is suitable for measuring early stages in the recommendation system where the main focus is to ensure that the most relevant matches are selected from a large pool of items [5]. We present a case study where *recall@k* was adapted and employed as a business metric for measuring the efficiency of our retrieval strategies in an online setting and discuss how it has provided us with better explainability and insights into our recommendation system. As our approach involves performing some post aggregation and processing of collected online signals, we describe it as a pseudo-online measurement of recall.

2. Definition of Retrieval Recall and its Variants

The concept of recall we explore in this paper is an adaptation of the traditional *recall* metric commonly used in the field of recommender systems.

$$Recall = \frac{No.\ of\ recommended\ relevant\ matches}{Total\ number\ of\ relevant\ matches} \quad (1)$$

In our job recommendation system, a relevant match refers to a job match that receives a positive feedback from the job seeker such as apply button click across Indeed. A recommended match refers to a job match that was retrieved by a match provider or one that passes a certain step of the recommendation pipeline. We note that relevant matches

RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems, October 14–18, 2024, Bari, Italy
lwu@indeed.com (L. Wu); ywpang@indeed.com (Y. W. Pang); warren@indeed.com (W. Cai)
0009-0004-4435-2136 (L. Wu); 0009-0000-4557-3580 (Y. W. Pang); 0009-0000-0507-3928 (W. Cai)

on Indeed do not solely come from our recommendation system, but also include other channels, such as external links, searches, social media, etc. Indeed is ranked as the #1 job site worldwide[1] and attracts over 350 million unique visitors globally each month[2]. Given this scale, Indeed's application data serves as a reasonably representative proxy for relevant matches in general.

When considering the *total number of relevant matches*, appropriate time bounds for determining the pool of relevant matches have to be set. There are two possible definitions - forward, or backward recall.

For simplicity, we introduce the definition of forward and backward recall in context of match providers, but these definitions can be easily generalized to other steps of the recommendation pipeline, such as filtering, scoring and ordering. We will see it as an example in section 4.3.

2.1. Forward Recall

In this definition, we aim to measure the immediate impact of the recommendations generated by the match provider. Starting with the job matches recommended by a match provider on a particular day, we check them against all relevant job matches in the subsequent few days:

$$\frac{\textit{No. of relevant matches recommended on a single day}}{\textit{Total number of relevant matches in the next x days}} \quad (2)$$

This approach allows us to assess how effectively each match provider can capture potential relevant matches shortly after their generation.

2.2. Backward Recall

Conversely, backward recall assesses a match provider's historical ability to predict and generate relevant job matches. Starting from the set of relevant matches recorded on a particular day, we check for the matches that have been recommended by each match provider in the prior few days.

$$\frac{\textit{No. of relevant matches recommended in the past x days}}{\textit{Total number of relevant matches on a single day}} \quad (3)$$

This measurement provides insights into the lasting relevance of the recommendations made by the match provider.

2.3. Comparison

The main difference between the above two definitions lies in the how samples for the numerator and denominator are obtained for recall calculation. We note that each approach will yield a different value of recall and the results from different definitions should not be compared with one another. However, in our practice, when we rank match providers by recall, those with high recall in one method, whether forward or backward, also rank high in the other method, leading to the same order.

Additionally, the parameter x is used to limit the amount of data for consideration and its main purpose is to account for potential delays between the time a job match is generated to the time a job seeker might respond to the recommendation, which could take a few days especially for our email recommendation channel. We chose 7 days (a week) in our implementation.

[1]Comscore, Total Visits, June 2023

[2]Indeed Internal Data, average monthly Unique Visitors April – July 2023

3. Implementation

3.1. Building the Dataset for Retrieval Recall

To measure the retrieval recall in our recommendation system, we made use of the backward recall definition above. On a given day, relevant matches logged on the Indeed site (i.e. matches that received an apply button click) were picked as the base event and a left join was done with the retrieved matches logged in the previous 7 days.

The resulting dataset was made available as a table on our internal data analytics platform, and could be easily queried for analysis on retrieval strategies. The key columns in our table for analyzing retrieval recall in our job seeker to job recommendation system are as follows:

- `Job and job seeker ID`
- `Metadata`: Additional information relating to the job or job seeker which could be useful in filtering the data for segmented analysis (e.g. locale information)
- `Apply button click flag`: Boolean flag that indicates that the match received an apply button click by the job seeker. With the existing definition of a relevant match, this will be *true* for all records. However, this could change if the definition of a relevant match expands to incorporate other user signals in the future.
- `Match providers covering the match`: An array of match providers that were able to generate the particular relevant match in the previous 7 days.

3.2. Delayed Signals

Indeed collects several other delayed positive signals along the job application funnel; after the initial apply button click (i.e. job application submitted, positive response from employers, etc). We incorporated these signals into our dataset as well by joining the resulting dataset with the corresponding event logs. This enabled tracking and measuring recall of additional user signals associated with job recommendations. We will introduce some delayed signals in section 4.1.

4. Real-world Applications

4.1. Example 1: Better Indication of Retrieval Stage Performance in an Online Setting

Common business metrics (i.e. click-through rate, conversion rates) used in online experiments for recommender systems only capture the impact of matches that were actually delivered to the user. In many cases, we may not be able to see the actual improvement of a positive change made to a match provider, possibly due to business logic, filtering and ranking stages that occur subsequently. In one of our online experiments, we tested a product change of an embedding based match provider supporting Indeed's job recommendation email channel. Note that there are several

Metric name	Change	Significant
Number of applications	+2.06%	Yes
Number of positive connections	-0.6%	No
Number of positive outcome	+0.59%	No

Table 1
A/B test result of metrics on matches delivered through a channel

Metric name	Change	Significant
Recall of applications	+0.85%	Yes
Recall of positive connections	+0.76%	Yes
Recall of positive outcome	+0.71%	Yes

Table 2
A/B test result of recall metrics

other match providers also generating job matches for this channel.

The change would enable the match provider to retrieve jobs that better align with job seekers' activity history and we hypothesized that it would result in improvements in the following user signals collected on Indeed:

- `Application`: It shows whether the job seeker applied for the job.
- `Positive connection`: It shows whether the job seeker and employer behind the job have positive interactions, such as messaging.
- `Positive outcome`: It shows whether the job seeker and employer behind the job have positive interactions which could lead to hiring, such as interviews.

Table 1 shows the change in the above signals when comparing the test bucket against the control bucket. Only the *number of applications* had significant improvement.

Using the definition described above, we also measured the recall for each of the signals for the entire retrieval stage (i.e. the combined recall of all match providers supporting the email job recommendation channel). This would allow us to capture relevant matches that were not actually delivered through the email system but could have received the user signals from other channels on Indeed.

Table 2 shows the change in the signals when comparing the test bucket against the control bucket. We see that the *recall of application, positive connection, and positive outcome* are all improved and have achieved significance. This shows that by measuring recall, we are able to isolate measurement of the impact of our change to the retrieval stage. Besides giving insights that we were able to improve the retrieval stage of our recommender system, it also hints at opportunities that could be worked on at subsequent stages of the system.

4.2. Example 2: Tracking Match Providers' Recall Change Over Time

Apart from analyzing the performance of the entire retrieval stage, we also made use of retrieval recall to track improvements for individual match providers.

For example, we had implemented multiple technical enhancements in one of our match providers, denoted by M. Each roll out was backed by online A/B experiments, and we anticipated seeing the actual improvements once all enhancements were deployed to production. We observed that our measurement of retrieval recall was able to better capture the impact of these improvements.

To illustrate this, we compare two different methods (figure 2):

1. Application Share Change Over Time: Application share is defined as the number of applications in job recommendation emails and attributed to M divided by the total number of applications in job recommendation emails. Despite the multiple enhancements, the share of applications of M did not increase. This can be attributed to other changes being made across the recommender system, such as the addition of new match providers whose matches might replace those retrieved by M; given the limits on the number of recommendations we can deliver to each user.
2. Recall Change Over Time: Recall is defined as the number of applications covered by M divided by the number of total applications in Indeed. We observe a clear increasing trend in M's recall. This indicates that the match provider is effectively identifying a higher number of relevant matches over time, even if this improvement is not reflected in the application share.

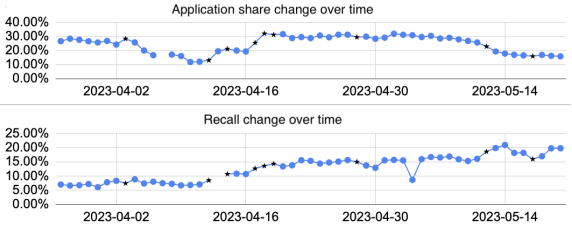

Figure 2: A period containing multiple product improvements, with exact dates of changes marked by stars

4.3. Example 3: Job-Recommendation Email Funnel's Effectiveness at Each Stage

As our recall measurements covers the first stage of the recommendation pipeline, we can further analyze the drop off of relevant retrieved matches through the subsequent stages. Ideally, while recall decreases through the recommendation funnel, precision should improve. By calculating the recall at each stage and comparing it with precision, we can identify improvement opportunities of the recommendation process.

In this example, we focus on job matches with positive outcomes in Indeed. Positive outcomes (PO), defined as per section 4.1, is a composition of multiple types of positive interactions between jobs and job seekers, such as interviews. The total number of matches with a positive outcome returned by our match providers serves as the numerator for the recall metric. Under this definition, the recall at each filter stage is 31%, 10%, 9%, and 3%, respectively.

By combining this analysis with quality improvement at each stage, we gain a clear understanding of the funnel's effectiveness. We can define the quality of each stage as precision, which is the number of relevant matches retrieved at this stage divided by the total number of matches at that stage. While we expected precision should improve as recall decreases through the funnel, we observed that the precision drops from 0.04% to 0.02% at step 1 (figure 4), while recall drops to 31% (figure 3). This suggests the need for further investigation and quality improvement at this stage.

By applying this method, we can identify bottlenecks and opportunities for enhancement of the recommendation funnel, ensuring continuous improvement and higher satisfaction for job seekers.

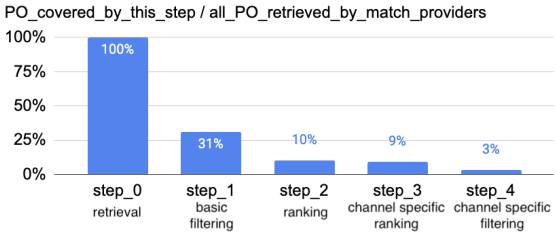

Figure 3: Change of recall through the funnel

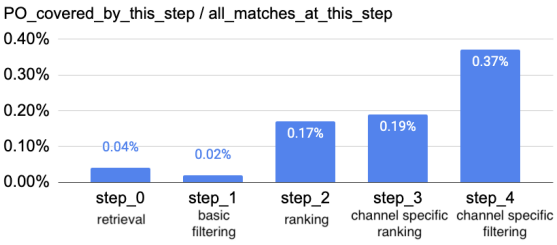

Figure 4: Change of precision through the funnel

4.4. Example 4: Identify Missing Opportunities

We are also able to look into 'missed recall' of our recommendation system. This represents relevant matches that were not delivered to job seekers through our recommendation system, but were found when the job seekers manually searched for jobs on Indeed. In a study conducted, we sampled the dataset of retrieval recall described in section 3.1 and analyzed 735 matches which were relevant but not successfully retrieved by the match providers in our email recommendation system. By analyzing the patterns of missed relevant matches (figure 5), we found opportunities to optimize our retrieval strategy across features such as job age and job title.

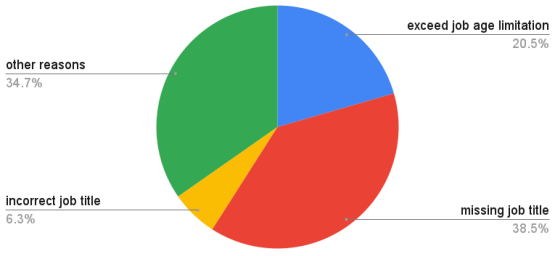

Figure 5: Analysis on missed relevant matches

5. Future Directions

In this section we list a few directions for future exploration.

- While forward and backward recalls are powerful tools for us to better understand the performance of retrieval stage and other intermediate steps, in practice we monitor not only these recall variants but also traditional online business metrics, such as click through rate, to make sure online business metrics are not harmed. Studying the correlation between recall and business metrics could be helpful to reduce the total number of metrics in an experiment and simplify the decision making process.
- As mentioned earlier, one important assumption behind this work is that relevant matches do not solely come from our recommendation system, but also include other channels, such as external links, searches, social media, etc. This assumption will be violated if relevant matches delivered from recommendation system dominate the pool of all relevant matches. It would be interesting to explore different methods of defining relevance of a match regardless of delivery.

6. Conclusion

In this work, we presented a case study on how we adapted the metric, *recall@k*, for use in an online setting to measure the effectiveness of the early stages within our recommender system for job recommendations. Compared to business metrics, which usually only capture the impact of recommendations that were actually delivered to users, this nuanced measurement of performance at each step ensures that enhancements to the system are accurately targeted. It also serves a diagnostic tool that identifies missed opportunities and bottlenecks within the recommendation funnel.

References

[1] K. Higley, E. Oldridge, R. Ak, S. Rabhi, G. de Souza Pereira Moreira, Building and deploying a multi-stage recommender system with merlin, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 632–635. URL: https://doi.org/10.1145/3523227.3551468. doi:10.1145/3523227.3551468.

[2] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, 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, RepSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 7–14. URL: https://doi.org/10.1145/2532508.2532511. doi:10.1145/2532508.2532511.

[3] K. Krauth, S. Dean, A. Zhao, W. Guo, M. Curmei, B. Recht, M. I. Jordan, Do offline metrics predict online performance in recommender systems?, 2020. URL: https://arxiv.org/abs/2011.07931. arXiv:2011.07931.

[4] L. Carnevali, Evaluation measures in information retrieval, 2023. URL: https://www.pinecone.io/learn/offline-evaluation/.

[5] P. Agrawal, Building a large-scale recommendation system: People you may know, 2024. URL: https://www.linkedin.com/blog/engineering/recommendations/building-a-large-scale-recommendation-system-people-you-may-know.