User Fatigue in Online News Recommendation

Hao Ma
Microsoft Research
One Microsoft Way
Redmond, WA 98052
haoma@microsoft.com

Xueqing Liu
University of Illinois at
Urbana-Champaign
Urbana, IL 61801
xliu93@illinois.edu

Zhihong Shen
Microsoft Research
One Microsoft Way
Redmond, WA 98052
zhihosh@microsoft.com

ABSTRACT

Many aspects and properties of *Recommender Systems* have been well studied in the past decade, however, the impact of *User Fatigue* has been mostly ignored in the literature. User fatigue represents the phenomenon that a user quickly loses the interest on the recommended item if the same item has been presented to this user multiple times before. The direct impact caused by the user fatigue is the dramatic decrease of the Click Through Rate (CTR, i.e., the ratio of clicks to impressions).

In this paper, we present a comprehensive study on the research of the user fatigue in online recommender systems. By analyzing user behavioral logs from *Bing Now* news recommendation, we find that user fatigue is a severe problem that greatly affects the user experience. We also notice that different users engage differently with repeated recommendations. Depending on the previous users' interaction with repeated recommendations, we illustrate that under certain condition the previously seen items should be demoted, while some other times they should be promoted. We demonstrate how statistics about the analysis of the user fatigue can be incorporated into ranking algorithms for personalized recommendations. Our experimental results indicate that significant gains can be achieved by introducing features that reflect users' interaction with previously seen recommendations (up to 15% enhancement on all users and 34% improvement on heavy users).

Keywords

Recommender Systems, User Fatigue, News Recommendation, Click Prediction, User Modeling

1. INTRODUCTION

In the past few years, recommender systems have become extremely important and are applied in a variety of online applications. Industry-wise, recommendation algorithms are now powering many popular online services, including but not limited to movie recommendation at Netflix, news recommendation at Yahoo!, music recommendation at Spotify, game recommendation at XBox, jobs recommendation at LinkedIn, etc. Academically, almost all the important properties of a typical recommender system have been extensively studied, including context-aware, temporal dynamic, diversity, serendipity, social-aware, privacy preserving, etc.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2874813.

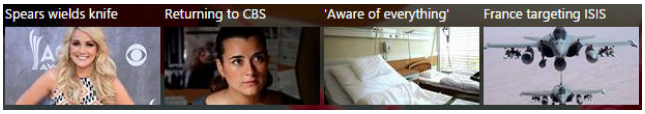


Figure 1: Bing Now Snapshot

In this paper, we study an interesting problem, i.e., the user fatigue in online recommendations, which attracted much fewer attention in the previous research work.

User fatigue is a ubiquitous phenomenon in online recommender systems, and it represents the factor that a user could quickly lose the interest on the repeatedly recommended items. In this paper, in order to find out what factors can cause user fatigue and how they influence users' browsing and clicking behaviors, we conduct a comprehensive study on the *Bing Now* news recommendation service.

As shown in Figure 1, when a user is visiting the home page of Bing.com, the Bing Now news recommendations are shown on the bottom of the page. In this specific example, a few short news titles with corresponding pictures are presented to the users. If a user clicked on any of them, such as "France targeting ISIS", Bing takes the user to the search results page in which several news articles related to "France targeting ISIS" will be displayed on the top. The user can then browse and read those full news articles based on his/her needs.

Normally, at any given time, the recommendation algorithm[1] in Bing Now system picks and shows 15 short news titles from an active set. New or fresh news titles created by the editors will be periodically pushed to replace some old ones. This iterating process helps the Bing Now service not only keep up with novel and important stories, but also exclude the stale and fading ones.

Due to the nature of the Bing Now application, it is almost inevitable that as a user frequently visits Bing's home page, many repeated recommendations will be presented to this user. In order to quantify how often the online users experience repeated items in the Bing Now recommendation service, we calculate the average number of overlapped news items between users' two consecutive visits. As shown in Figure 2, when a user's two consecutive impressions happen within the (0, 2] hours range, on average there are a total of 11 overlapped news items between these two impressions.

[1]The ranking of the items is mainly determined by the global click through rate of each item at the impression time.

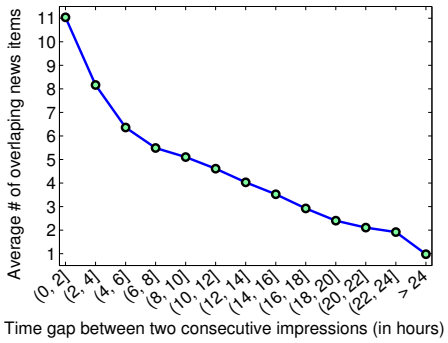


Figure 2: Item Overlaps between Two Consecutive Impressions

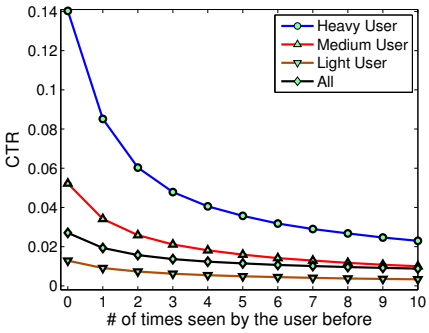


Figure 3: Same Item Fatigue

This number keeps dropping as the time gap increases. We notice that even a user's last impression happened more than 24 hours ago, on average this user's current impression still shares approximately 1 news item with the previous one.

From the above statistics, we can see that it is necessary to understand users' interactions with the repeated recommendations since those repetitions are prevalent in the system. In this paper, the fundamental research questions we are trying to explore are:

- Do users become fatigued when they see repeated recommendations?
- If yes, what are the major factors that greatly affect users' fatigue?

Aiming at addressing the above research questions, we conduct several in-depth analyses on 4 weeks of Bing Now log data. We evaluate a wide range of factors that can possibly impact users' fatigue on repeated items, including number of same item views, number of same item clicks, user demographics information, item topic category, item age, item position as well as various temporal related factors. We observe that many of these factors can significantly impact users' behaviors when repeated recommendations are presented.

We conjecture that a method which can leverage the fatigue related factors to automatically promote and demote news items could lead to better ranking results. We therefore design several effective features that reflect the recommendation repetitions as well as capture users' interactions with those repetitions. By incorporating these features into a ranking algorithm, we demonstrate that our method shows significant gains over competitive baselines.

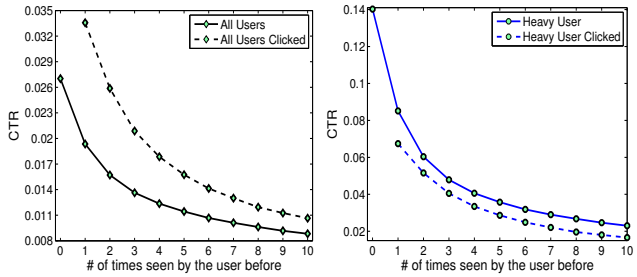
Based on the interaction logs we describe in Section 2, we conduct several in-depth analyses to reveal user fatigue phenomenon in Section 3. Section 4 presents our ranking approach, and describes our evaluation data, metrics as well as the experimental results. After presenting related work in Section 5, we state the concluding remarks in Section 6.

2. DATA DESCRIPTION

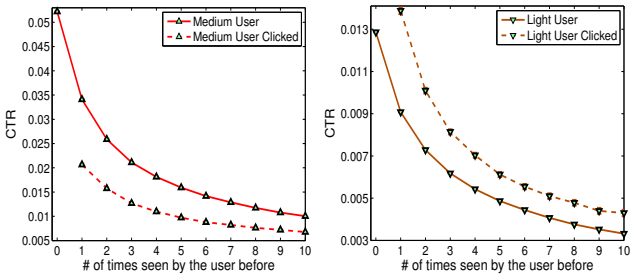
In this section, we describe what the data we analyze look like and how we further split the users into different categories in order to reveal more insights of user fatigue behaviors.

We randomly sampled 1 million users from 4 weeks of desktop Bing Now usage logs. Each user has at least one Bing Now impression and each impression contains exactly 15 news recommendations. We do not further filter the data in order to present the most accurate analysis in understanding users' behaviors. This indicates that even a user does not click any items, we will also include this user in the studies.

Like any other popular recommender systems, we observe that some users really enjoy clicking the recommended news, while some other users seldom click them. In order to find out whether different types of users share distinct fatigue levels, we further group users into 3 categories (i.e., heavy users, medium users and light users) based on users' click frequencies. The logic of grouping is the following: (1) We first so-319(o)5(3(l)93ofi10.23(o)5c)2(k)-2(r)-2(s)1(

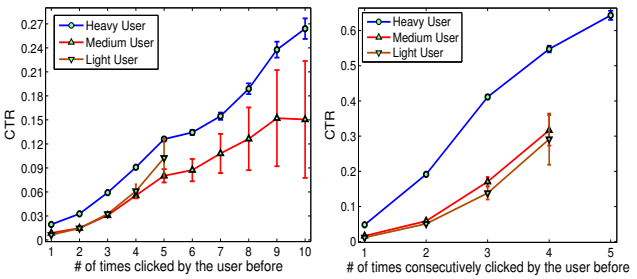


(a) Overall Users (b) Heavy User



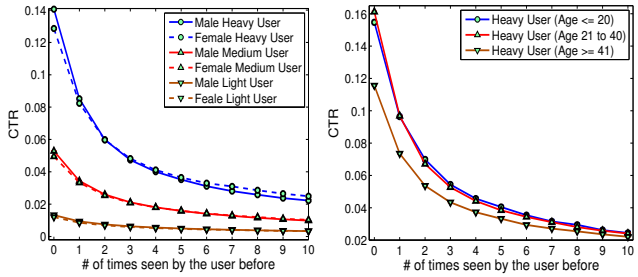
(c) Medium User (d) Light User

Figure 4: Clicked at least Once Before

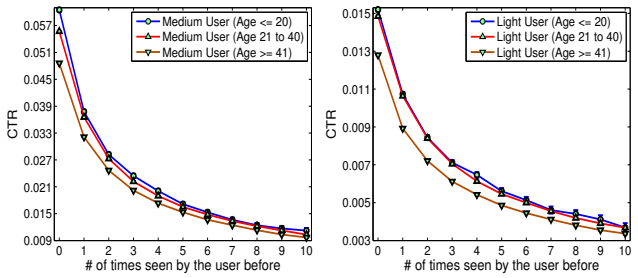


(a) Click Count (b) Consecutive Click Count

Figure 5: Clicked Multiple Times

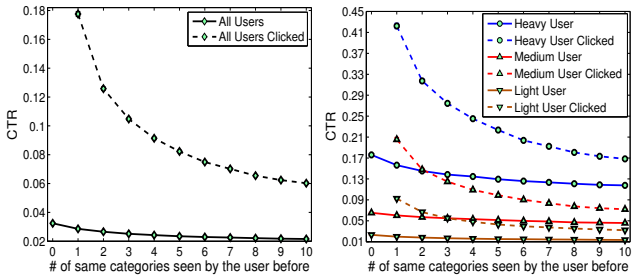


(a) Gender (b) Age - Heavy User



(c) Age - Medium User (d) Age - Light User

Figure 6: User Demographics



(a) All Users (b) Different User Categories

Figure 7: Same Category Fatigue

tion 3.2, user demographics in Section 3.3, same category fatigue in Section 3.4, item positional influence in Section 3.5, temporal related factors in Section 3.6. Moreover, in all the figures presented in this section, the error bars represent 95% confidence intervals. Also note that some errors may be very small, hence the corresponding error bars are barely visible.

3.2 Same Item Fatigue

The first analysis we explore in this paper is the same item fatigue situation. As shown in Figure 3, the key question we try to answer is what the CTR will look like if a user saw the same news item for multiple times.

In this figure, we have three major observations. First of all, under all the user categories, we notice that the CTR values drop quickly as the number of same item views increases, which indicates that this number could be a good predictor in re-ranking the results. Secondly, heavy users are more willing to click news items than medium users, who in turns contribute more clicks than light users. Thirdly, comparing with the medium and light users, the heavy users suffer the most from the fatigue since the slope of the curve is much sharper.

The overall trends we observe so far are intuitive: repeated recommendations have substantially lower CTR than those items never seen by the users before. Hence, we continue by studying how users' interaction with repeated recommendations could change the fatigue patterns.

First, we consider whether the users clicked at least once on the previous seen items, which is detailed in four subfigures in Figure 4. The dashed lines in these subfigures represent the trends for corresponding users clicked the same item at least once before, while the solid lines include all the users. When considering all the users together (Figure 4(a)), we find that the CTR values become higher once we know a user clicked at least once on previously seen results. However, when looking at the users separately (Figure 4(b), Figure 4(c) and Figure 4(d)), only the light users share similar patterns with the overall users, while the tendency of the heavy and medium users is on the opposite side. This phenomenon may suggest that heavy and medium users' click patterns are more diverse than the light users, and they tend to click different items. Light users seem to be more focused and once they clicked on some items, they are more likely to click them again. From these four figures, we can also see that the behaviors of the light users dominate the trends of overall users since as mentioned in Section 2, the number of the light users is much higher than the number of the heavy and medium users. This also indicates that it is necessary to group users into different categories in order to gain more insights of user fatigue behaviors and not to mention that the user category information could be another effective feature in improving the ranking of recommendations.

Next, we study the CTR trends when the users previously clicked the same item multiple times. Figure 5(a) and Figure 5(b) show the CTR changes based on the number of clicks and number of consecutive clicks, respectively. In both cases, we notice that as the number of click times increases the CTR rises substantially. If a heavy user consecutively clicked on an item 5 times before, the probability on clicking it again next time is 10 times more than if the user only clicked the item once before. This indicates that a user's previous click count of each news item could become a very strong signal in ranking them. One might ask why a user would click and read the same news item multiple times. The answer is Bing Now is more like a news event recommender system than a news article recommender system developed by Yahoo! or Google. As introduced in Section 1, what we show on Bing Now are short news titles or events. Once a user clicks on one of the titles, this user can then read a few news articles that are related to this title or event. It thereby becomes natural for a user to click the same title multiple times in order to follow the most recent developments of some breaking news, like "the missing of mh370".

3.3 User Demographics

In the above subsection, we draw the conclusion that the fatigue behaviors for different user categories are sometimes divergent. Hence, in this subsection, we are curious to know whether groups based on user demographics information[3] could be potentially useful in influencing users' fatigue.

Figure 6(a) reveals the user fatigue analysis on males and females, while Figure 6(b), Figure 6(c) and Figure 6(d) represent the trends for different age groups. The first conclusion we make is that there is basically little fatigue differences between male and female users. Secondly, the fatigue levels between three different age groups seem to bear lots of similarities. The only major difference is that the CTR values for the age group "Age >= 41" are a little bit lower than the other two age groups, i.e., "Age 21 to 40" and "Age <= 20". Based on the above observation, we conjecture that demographics group information might not be an effective factor in predicting users' fatigue level.

3.4 Same Category Fatigue

Thus far in this paper, all the fatigue studies are based on the users' interactions with the same item. In this section, we extend the research to the same category fatigue analysis.

In Bing Now news recommender system, each news item has been assigned a topic category by the editors, including Sports, Entertainment, Politics, Health, Business, etc. Hence, we are also interested in exploring users' fatigue levels conditioned on the same item category, as displayed in Figure 7. Again, the dashed lines in these subfigures represent the trends for corresponding users clicked the same item at least once before, while the solid lines do not have this constraint. From this figure, we can definitely notice the decrease of CTR as the number of items in the same categories viewed by a user increases, however, the dropping slopes for the solid lines are much more shallow than what we observe from the same item fatigue presented in Figure 3. Moreover, unlike what we found in Figure 4, as illustrated by the dashed lines in Figure 7, the patterns for heavy users, medium users, light users are consistent once the users clicked on at least one category before. All of the above observations suggest that category level fatigue is different from item level fatigue. An item's topic category is a higher level abstraction of item itself, thus, category is more reliable in capturing users' inherent interests. Hence, a user viewed/clicked the same category multiple times should not cause more fatigue than she viewed/clicked the same item many times. Based on this conclusion, in Section 4, we also design several effective features based on users' category level behaviors.

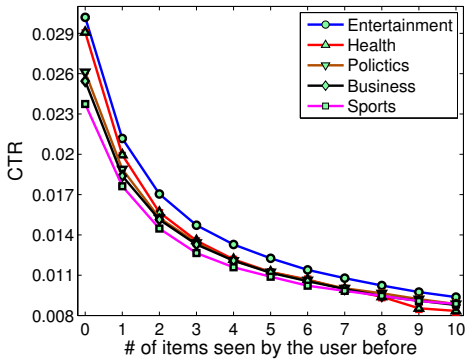


Figure 8: Overall Fatigue within the same Item Category

Since we have items' category information in our data, it is hence intuitive to further investigate if users' same item fatigue shares similar patterns within different item categories. Figure 8 displays overall users' CTR trends for 5 specific categories. The major difference we see from the figure is that users generally have higher CTRs in Entertainment category than some other categories, like Sports and Business. In terms of fatigue level, we can hardly find any differences between these 5 categories since the corresponding curves all drop quick as the number of same items user previously saw increases. This observation confirms that user fatigue is a ubiquitous and common phenomenon that exists in online recommender systems.

3.5 Positional Impact

It is a well-known factor that users are biased towards clicking on higher ranked results in online search engines and recommender systems [6, 14, 23]. This leads us to consider how an item's positions in the past could potentially change a user's behavior if the same item is presented to this user again.

In search engines, the ranked search results which suffer position bias problem are normally vertically presented to users, however, in our case, the recommendations are displayed horizontally. Hence, we first conduct an experiment to see if the position bias problem also exists in our Bing Now recommendation service. We use 1% of our Bing Now traffic to show random news titles to users, and collect users' behaviors towards those random news titles. By aggregating the CTRs on each position, we then draw the position bias graph in Figure 9. From this figure, we have several interesting observations: (1) First of all, we find that the position bias is very strong in our Bing Now horizontal representations, which indicates users normally scan the news from left to right. (2) Secondly, we notice that the CTR value drops sharply between position 8 and position 9. The reason behind this phenomenon is that, in Bing Now home page, normally only the first 8 news items are directly visible to the users[4]. The users need to scroll or swipe right to see all other items. This UI design impacts the CTR rates after position 8. (3) Lastly, the CTR of the last position is actually higher than a few neighbors in front of it. This indicates that users tend to click the last news item when they find there is no more news.

[3]The gender and age information we use in this paper is obtained from those users who opted-in to provide their demographics data for analysis purpose.

[4]Recently, Bing Now changed the UI design, and more than 8 news items are directly visible to the users. The total number of news items is also greater than 15. However, this change should not impact all the trends and conclusions we observe in this paper since the data we employ in this paper are collected before this change.

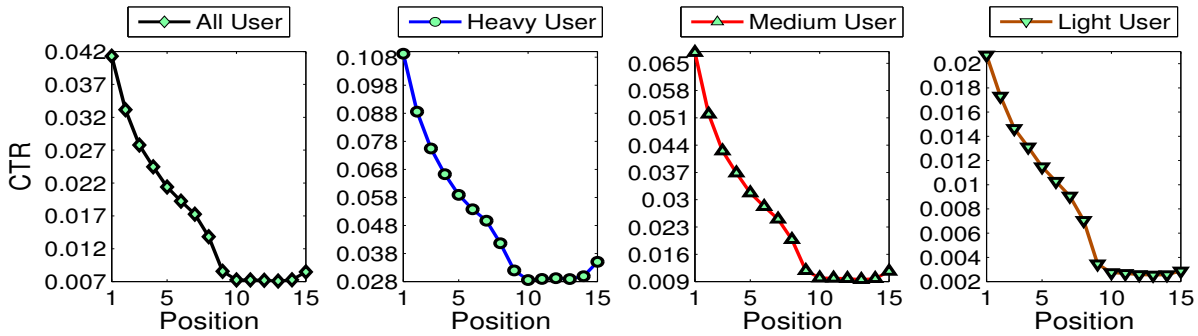
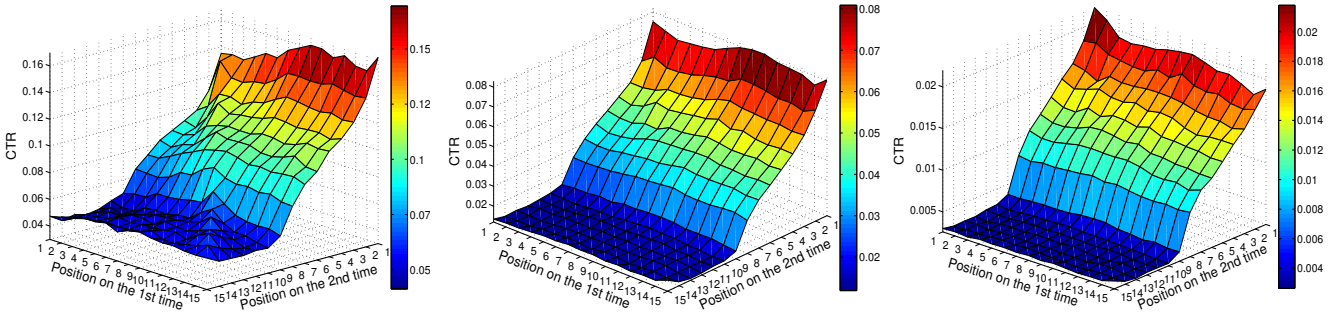


Figure 9: Position Bias



(a) Heavy (b) Medium (c) Light

Figure 10: Positional Impact

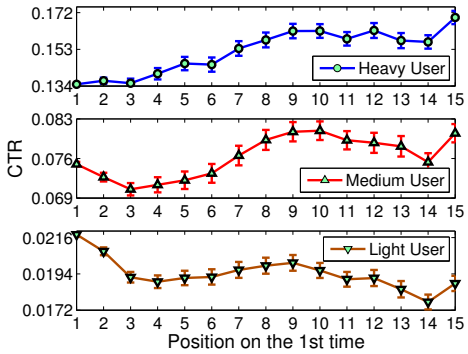


Figure 11: CTR Trends on the 1st Position at the Second Time

Next, we explore if the items' positions are critical in influencing users' fatigue degree.

More specifically, as shown in Figure 10, we are particularly interested in what is the CTR pattern if an item was firstly presented at position i and is displayed at position j in the second time. At the first glance, we find that an item's last time impression position is indeed an important factor in determining the next time CTR since the CTR values change significantly by varying the first time item position. In order to compare the trends for different user groups more intuitively, we take one special case from Figure 10 and plot them into Figure 11, which shows the CTR trend on the 1st position if the same item is displayed to the user at the second time.

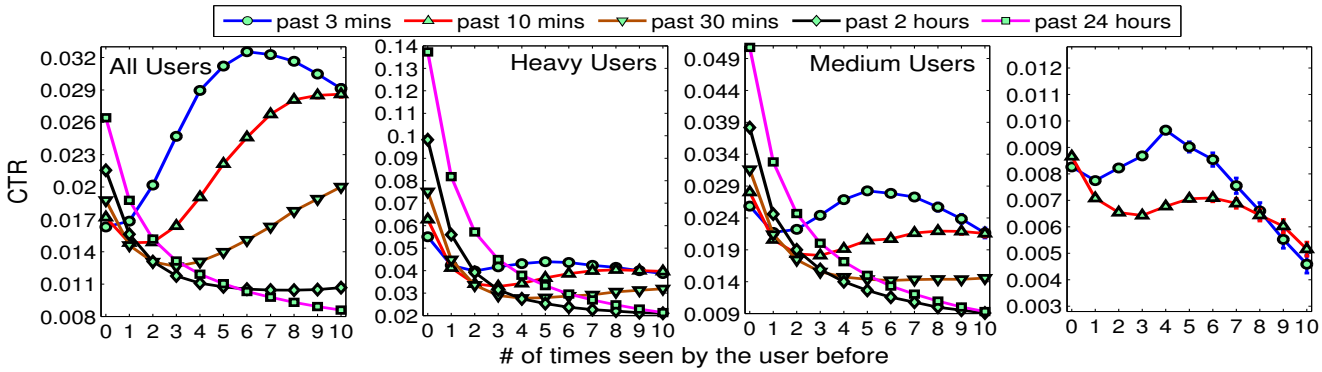
In Figure 11, we observe totally different patterns for the heavy and medium users when comparing with the light users. The CTR for the heavy and medium users increases as the last time position increases, while it keeps decreasing for light users. It seems that when an item is presented at the 1st place in the second time, the heavy users and medium users are more likely to click it if this item's last time position is at the 15th place. However, light users tend to click it more often if this item's last time position is at the 1st place. All these observations suggest that heavy users and medium users are more proactive. These two types of users are actively seeking different news articles to satisfy their evergrowing information needs. They are less interested in the same item if they already saw or consumed it previously in the top ranked positions. Instead, next time when these users are visiting Bing's homepage, they prefer to see different items showing on the top positions. The situation for light users are totally in the opposite side since they are more reactive comparing with the other two groups of users. It looks like displaying the same items multiple times on the top positions will reinforce the messages and will attract more attention from light users. From all the above analysis, we can see that the previous item's appearing position could become a very useful factor in reranking the recommendation results.

3.6 Temporal Factors

The last major analysis we conduct in this paper is related to the temporal dynamic information from both user side and item side.

The first investigation we perform is to explore the same item fatigue behaviors conditioned on different time periods by clustering users' interactions with the recommended items into five time periods, i.e., "past 3 minutes", "past 10 minutes", "past 30 minutes", "past 2 hours" and "past 24 hours". Figure 12 presents the fatigue level conditioned on the number of the same item views while Figure 13 analyzes the fatigue degree depending on the number of the same item clicks.

In Figure 12, we observe that users' behaviors during the past short time periods, such as "past 3 minutes" and "past 10 minutes", are quite different with other clusters. In these two time periods, after the first item view, as the increase of number of views, the CTR values increase in the beginning, and start dropping after passing certain threshold. Note that Bing Now recommendations are shown on the Bing.com home page, hence, the majority of users


past 3 mins
past 10 mins
past 30 mins
past 2 hours
past 24 hours
All Users
Heavy Users
Medium Users
CTR
# of times seen by the user before

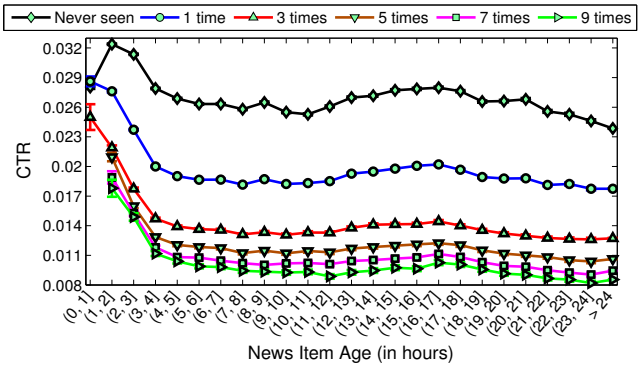


Figure 16: News Item Age

the same between different item age buckets, which indicates that news item age might not be a strong indicator in distinguishing user fatigues.

3.7 Other Factors

We also look into other factors that can potentially affect users' fatigue behaviors, like the number of overlapping items between two consecutive impressions and the content diversity level within one impression. We do not report the results here since we cannot find obvious connections between these factors and the user fatigue levels. We leave these further explorations into the future work.

4. EXPERIMENTAL ANALYSIS

In Section 3, we discovered several factors that are correlated with the user fatigue levels. We also showed that depending on the number of previous views, clicks, positions and different time spans, the click probability of repeated recommendations may vary significantly. In this section, we present the empirical results to demonstrate the effectiveness of the features we construct that reflect users' interactions with repeated recommendations.

4.1 Training and Testing Data

To evaluate our method quantitatively, we utilize the data described in Section 2 to generate training and testing data as well as calculating the corresponding feature values. We first removed all the users without any clicks from the dataset since we cannot evaluate those users at all as they did not show any preferences. We then randomly sampled around 60,000 users from the user pool, which results in 1.38 million impressions.

We split the sampled data into train-test sets in two different ways. In the first scenario, we use the last clicked impression of each user as the testing set and all previous impressions as the training set. The training to testing data size ratio is about 20 to 1. We call this "leave-one-out" dataset. In the second scenario, we use the first three weeks data as the training set and the last one week data as the testing set. The training to testing data size ratio is about 3 to 1. This scenario is a more realistic case in a production environment. We call this "last-week" dataset.

4.2 Evaluation Metrics

We use two popular evaluation metrics, i.e., Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), to measure the recommendation quality.

Let I be a set of impressions of online news recommendation. For each impression $i \in I$, there are K news items being selected for recommendation. In our experiment data set, $K = 15$. A ranking algorithm returns a ranked list of the K items, with descending order of the probability that the user will click on the item.

The Reciprocal Rank (RR) of a ranked list is the multiplicative inverse of the rank of the first hit in the list. The MRR is the average reciprocal rank obtained by the ranked lists given by an algorithm with respect to the set I.

$$MRR = \frac{1}{|I|}\sum_{i=1}^{|I|}\frac{1}{rank_i}, \quad (1)$$

where $rank_i$ is the rank of the first clicked item in the ranked list for the i-th impression.

MAP is a single-figure measure of quality across recall levels and has been shown to have especially good discrimination and stability in the information retrieval domain. For a given impression i, we define C_i is the total number of clicks; $d_{i;k}$ is 1 when the item at position k is clicked, 0 otherwise; and $c_{i;k}$ is the total number of clicks until position k; the precision value of top k items can be expressed as $d_{i;k}\frac{c_{i;k}}{k}$; and finally, the average precision for impression i is $\frac{\sum_{k=1}^{K} d_{i;k}\frac{c_{i;k}}{k}}{C_i}$. We can then formally define MAP over the set of impressions as:

$$MAP = \frac{1}{|I|}\sum_{i=1}^{|I|}\frac{\sum_{k=1}^{K} d_{i;k}\frac{c_{i;k}}{k}}{C_i}. \quad (2)$$

4.3 Features

Inspired by the observation and analysis in Section 3, we have developed the following six groups of features. Among them, the second to fifth groups reflect the past user-item interactions and reveal the user fatigue effect.

1. Click through rate (CTR) : This feature represents an item's CTR across all users at the impression time.
2. Same item fatigue features : These are features that reflect how a user has previously interacted (viewed or clicked) with a given news item.
3. Same news category fatigue features: These features captures how a user has previously interacted (viewed or clicked) with news items in the same topic category of the given item.
4. Display positional features: These are features that represent an item's previous positional information when it was previously seen or clicked by the user.
5. Temporal features: These features are developed based on the temporal aspect of the past user-item interaction.
6. Context features: These are features that are not directly relevant to past user-item interaction, such as features only related to users: user demographics, user categories; and features only related to items: item's age and it's topic category.

We present the details of all 34 features we generate in Table 1.

4.4 Learning Model and Baseline

In the experiments, we use each impression (which contains 15 news recommendation items) as a ranking group and if an item is clicked by the user, we label it as positive while consider the other items as non-relevant. We choose LambdaMart [18] as our ranking algorithm for the preference learning since it has been successful over a number of information retrieval problems. This algorithm adopts the gradient-boosted decision tree approach for optimizing a variety of non-continuous ranking objective functions. We apply the default settings for key parameters, they are, number of trees = 100, number of leaves = 20, minimum documents per leaf = 10

Table 1: Complete Feature List

Category	Feature Names	Notes
CTR	CTR	CTR at the time of being presented to user
Same item fatigue	ViewsBefore, ClicksBefore	Number of times this user has seen or clicked this item before
Same category fatigue	CatViewsBefore, CatClicksBefore	Number of times this user has seen or clicked items belongs to the same topic category before
Positional	LastViewPosition, LastClickPosition, FirstViewPosition, FirstClickPosition, AvgViewPosition, AvgClickPosition	Position of the item when the user last, first and averagely saw or clicked it
Temporal	ViewsIn3,ViewsIn10,ViewsIn30,ViewsIn120,ViewsIn1440 ClicksIn3,ClicksIn10,ClicksIn30,ClicksIn120,ClicksIn1440	Number of times this item was viewed or clicked by this user in past 3, 10, 30, 120, 1440 minutes
	TimeSinceLastView, TimeSinceLastClick, TimeSinceFirstView, TimeSinceFirstClick,	Time elapsed since last (first) view or click in minutes
	CatTimeSinceLastView, CatTimeSinceLastClick, CatTimeSinceFirstView, CatTimeSinceFirstClick	Time elapsed since last (first) view or click on items in the same category in minutes
Context	UserCategory, Gender, Age, ItemAge, ItemTopicCategory	User groups, demographics and items attributes

Table 2: MRR Measure Summary

Dataset	User Category	Metrics	CTR	CTR + Context	FAR
Leave one out	Light	MRR	0.3911	0.3905	**0.4018**
		Improve	2.73%	2.90%	
	Medium	MRR	0.4197	0.4278	**0.4697**
		Improve	11.90%	9.78%	
	Heavy	MRR	0.4886	0.5096	**0.5933**
		Improve	21.43%	16.42%	
	All	MRR	0.3996	0.4012	**0.4203**
		Improve	5.17%	4.74%	
Last week	Light	MRR	0.3855	0.3854	**0.4025**
		Improve	4.41%	4.43%	
	Medium	MRR	0.3951	0.4032	**0.4513**
		Improve	14.21%	11.92%	
	Heavy	MRR	0.3976	0.4200	**0.5248**
		Improve	31.97%	24.94%	
	All	MRR	0.3920	0.4003	**0.4501**
		Improve	14.84%	12.44%	

Table 3: MAP Measure Summary

Dataset	User Category	Metrics	CTR	CTR + Context	FAR
Leave one out	Light	MAP	0.3757	0.3751	**0.3861**
		Improve	2.81%	2.94%	
	Medium	MAP	0.3920	0.4003	**0.4414**
		Improve	12.59%	10.26%	
	Heavy	MAP	0.4535	0.4737	**0.5595**
		Improve	23.43%	18.16%	
	All	MAP	0.3814	0.3832	**0.4019**
		Improve	5.38%	4.90%	
Last week	Light	MAP	0.3714	0.3713	**0.3883**
		Improve	4.56%	4.56%	
	Medium	MAP	0.3757	0.3839	**0.4314**
		Improve	14.83%	12.38%	
	Heavy	MAP	0.3742	0.3971	**0.5026**
		Improve	34.32%	26.58%	
	All	MAP	0.3736	0.3822	**0.4319**
		Improve	15.60%	13.01%	

and learning rate = 0.2. Our primary goal is to demonstrate the usage of a learning model on the fatigue-related features that we construct, to improve the click prediction accuracy on online news recommendation. Hence neither the choice of a learning algorithm nor parameters are of particular interests for this study, and we will not include the discussion about the algorithm comparison and parameter sensitivity analysis in this work.

In our experiments, we use two baselines. The first baseline involves no learning or training, it ranks the news items based on the descending order of the items' CTR in a given impression. We refer to this as CTR baseline. The CTR baseline can be considered as a very strong baseline since it incorporates click behaviors across all the users and it can accurately represent the trending items at the any given moment.

For the second baseline, we train a ranker with CTR and context-related features (which are the first and last groups of features in Table 1: user category (light, medium or heavy), gender, age, item's age and item's topic category. We refer to this as CTR+Context baseline.

Finally, our method will use all six groups of features described in Section 4.3 to train a model to rank the given 15 items in each impression. We refer to this trained model as FAR (Fatigue-Aware Recommendation).

4.5 Comparison with Baselines

The result which compares FAR and two baselines is summarized in Table 2 and Table 3. We report the results on both overall population and on each light, medium and heavy category. The percentage numbers in the "Improve" cells represent the improvement of our FAR method over two baselines.

Our FAR model outperforms two baselines on both MRR and MAP metrics significantly on both datasets. We also perform paired t-test between FAR and two baselines separately with p-value < 0.01. It indicates that the improvement of FAR over two baselines is statistically significant.

For the first "leave-one-out" dataset, on the overall users, FAR improves the MRR and MAP by around 5.2-5.4% compared to the CTR baseline and by around 4.8-4.9% compared to the CTR+Context baseline. When we compare the metrics among three different user

Table 4: Comparison of Feature Group Effectiveness

Dataset	Model	MRR	Improve	MAP	Improve
Leave one out	CTR	0.3996	-	0.3814	-
	CTR+Context	0.4012	0.41%	0.3832	0.46%
	CTR+Category	0.4014	0.46%	0.3834	0.53%
	CTR+Position	0.4089	2.33%	0.3907	2.45%
	CTR+Item	0.4110	2.85%	0.3929	3.02%
	CTR+Temporal	0.4147	3.77%	0.3964	3.94%
	FAR	**0.4203**	**5.17%**	**0.4019**	**5.38%**
Last week	CTR	0.3920	-	0.3737	-
	CTR+Context	0.4003	2.13%	0.3822	2.29%
	CTR+Category	0.3947	0.69%	0.3762	0.69%
	CTR+Position	0.4351	11.00%	0.4169	11.57%
	CTR+Item	0.4389	11.98%	0.4208	12.63%
	CTR+Temporal	0.4407	12.43%	0.4224	13.04%
	FAR	**0.4501**	**14.84%**	**0.4319**	**15.60%**

categories, the MRR and MAP measure improves about 2.8-3.0%, 10-12% and 17-23% for light, medium and heavy categories over two baselines respectively. We also observe a few interesting patterns from these two tables. First, both MRR and MAP values for all the methods increase as users engages more with the Bing Now (a user shifts from light to heavy category) service. This phenomenon suggests that the more users click, the better we know the users. Second, the FAR model enhances the heavy users' MRR and MAP values about 8 times more than the improvement over light users. Since heavy users have much more clicks than light users, they are more likely to be impacted by the fatigue issue. Thus, when leveraging the fatigue related features into FAR model, it effectively helps alleviate the user fatigue problem. Third, the light users improvement results dominate the overall results on both MRR and MAP metrics. This is expected since in this "leave-one-out" data set, we split the training and testing data in such a way that each user only contributes one data point in the testing set. Based on the user distribution among different categories disclosed at the end of Section 2, there is no surprise to see the light users' improvement dominates the whole population result.

On the second "last-week" testing set, we observe even more significant improvement on both MRR and MAP measures. FAR improves the MRR and MAP about 15% over the CTR baseline and about 13% over the CTR+Context baseline respectively on the overall population, which represents three times improvement over the "leave-one-out" dataset. In the "last-week" testing set, as mentioned in Section 2, it should contain approximately one third clicks from the heavy users, one third clicks from the medium users and another one third clicks from the light users. The distribution of clicks is not dominated by the light users anymore. Hence, we can observe much larger gain comparing with the "leave-one-out" testing set.

In general, our FAR method obtains consistent and significant gains over two baselines, which demonstrates the effectiveness of the features we construct from users' fatigue behaviors.

4.6 Evaluating Different Feature Groups

To evaluate the effectiveness of each fatigue-aware feature group that described in Section 4.3, we report the MRR and MAP values in Table 4 based on models trained using CTR plus individual feature groups. The percentages in the "Improve" cells represent the improvement over the CTR baseline on the whole user population.

From this table, we can see that the most important feature groups are "Postional", "Same Item" and "Temporal" groups. When using together with the CTR baseline, they can all greatly enhance the recommendation results. This also coincides with all the conclusions and trends we observe in Section 3. Moreover, it seems that incorporating the "Temporal" features can achieve the most gains since as shown in Table 1, the "Temporal" feature group actually also takes advantages of "Same Item" and "Same Category" features which makes it the strongest feature group.

5. RELATED WORK

In this section, we review several research directions which are relevant to our work, including research on recommender systems in general, research on repeated search and repeated recommendation.

5.1 Recommender Systems in General

Due to its commercial values, many aspects and properties of a typical recommender system have been widely studied in the past decade, including but not limited to context-aware [1, 10], temporal dynamic [5, 8, 9, 20], diversity [11, 17], serendipity [24], social-aware [13], location-based [7], etc.

Our work analyzes and models the dynamic changing interaction between users and news items, which is closely related to temporal dynamics in recommendation system. Koren [8] demonstrated the fact that same user's interest in the same movie can drift over time, and model the temporal dynamics by spliting both item bias and user bias into a stationary part and a time changing part. Chen et al. [4] studied this problem in the context of a social trusted recommendation, by arguing that users' interests in recommended items can change due to the change of their social relationships. Xiong et al. [20] modeled temporary effect as the third dimension besides the user and item dimension, and in contrast to the family of existing collaborative filtering methods using matrix factorization, they proposed to solve their three dimensional problem using tensor factorization. Whereas Xiang et al. [19] found it not desirable to model the temporal dynamics as a global pattern in [20], and they used a random walk graph to find the local drifting pattern for each user. Wei and Park [5] incorporated temporal dynamics in news recommendation.

Our study also shows time span can essentially influence on the degree of user fatigue. There is another branch of work that studies the influence of the time span factor in recommendation systems. Yin et al. [22] used the time a user spend on viewing an item as an implicit signal of voting. They demonstrated that the longer that time is, the more likely she is interested in the item. This implicit signal can help resolve the problem of data sparsity caused by the vacancy of most ratings. Yi et al. [21] studied dwelling time in the context of users' implicit feedback, i.e., clicking instead of rating. They incorporated the factor of dwelling time to collaborative filtering and observed lifting.

Although our work is related to previous temporal dynamic and time span research in recommender systems, the user fatigue problem studied in this paper is fundamentally different, which attracts little attention in the literature.

5.2 Repeated Search Results

In search engine or information retrieval research field, there are a few research papers studied the users' re-finding and re-visitation search behaviors. Teevan et al. [16] proposed a model for personal navigation, which monitored the user's long-term history and showed that when a user repeatedly issues the same query and clicks on the same single result over time, the target URL can be regarded as a personal navigation destination. Shokouhi et al. [15] studied the users' re-search and re-click behaviors, and presented how to design a method to improve the search results ranking. Our work is significantly different from the research on repeated search

results since our targeting recommendation domain is fundamentally different with the search domain where the latter needs users' search queries to drive users' click behaviors.

5.3 Repeated Recommendations

In a study and perhaps the most related to our work, Agarwal et al. [2] briefly mentioned the fatigue issue in an online news recommender system. The authors observed the CTR drop if an item was presented to a user for multiple times by examining a couple of factors, like the number of impression times and the number of minutes since the first impression. However, the main focus of this piece of research work was on how to estimate CTR, which is quite different with the theme of our paper. In a recent study detailed in [12], Lee et al. followed the work in [2] by experimenting with the LinkedIn person recommendation data. Instead of estimating CTR, the authors proposed to optimize for a different metric, i.e., Conversion Rate, by taking into consideration of the impression discounting factor. Again, in this paper, we study the user fatigue behavior from very different aspects in which some of them have never been studied by previous work before.

The user fatigue analysis studied in this paper is also related to the Frequency Capping concept in online advertising, which limits user exposure to an ad due to the advertiser budget constraints. More specifically, frequency capping prevents ads from being displayed repeatedly to the point where visitors are being overexposed and response drops. As shown in [3], a typical frequency capping scenario focuses on an optimization problem which maximizes an advertiser's value by fixing a user's frequency cap f_i and imposing some other constraints. As we can see, the problem setting of frequency capping is fundamentally different with the topic we study in this paper. Most importantly, the observations and conclusions we have in this paper can actually be very useful to help determine the specific cap f_i for each user.

6. CONCLUSION

In this paper, we study the user fatigue phenomenon that is rarely discussed by previous research work. By analyzing the user interactive logs of Bing Now news recommendation service, we identify several interesting factors that could potentially affect users' fatigue levels. We also design several intuitive fatigue related features based on our observation. By utilizing the learning power of LambdaMart algorithm, we demonstrate that our FAR method can greatly improve the recommendation results over competitive baselines, which proves the importance of user fatigue understanding in online recommender systems.

7. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender Systems Handbook*, pages 217–253. 2011.

[2] D. Agarwal, B. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of WWW 2009*, pages 21–30, 2009.

[3] N. Buchbinder, M. Feldman, A. Ghosh, and J. Naor. Frequency capping in online advertising. *Journal of Scheduling*, 17(4):385–398, 2014.

[4] W. Chen, W. Hsu, and M. L. Lee. Modeling user's receptiveness over time for recommendation. In *Proceedings of SIGIR 2013*, pages 373–382, 2013.

[5] W. Chu and S.-T. Park. Personalized recommendation on dynmic content using predictive bilinear models. In *Proceedings of WWW 2009*, pages 691–700, 2009.

[6] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of WSDM 2008*, pages 87–94, 2008.

[7] K. Kim, J. G. Park, and S. Cho. Correlation analysis and performance evaluation of distance measures for evolutionary neural networks. *Journal of Intelligent and Fuzzy Systems*, 22(2-3):83–92, 2011.

[8] Y. Koren. Collaborative filtering with temporal dynamics. In *Proceedings of KDD 2009*, pages 447–456, 2009.

[9] Y. Koren. Collaborative filtering with temporal dynamics. *Commun. ACM*, 53(4):89–97, 2010.

[10] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[11] N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal diversity in recommender systems. In *Proceeding of SIGIR 2010*, pages 210–217, 2010.

[12] P. Lee, L. V. Lakshmanan, M. Tiwari, and S. Shah. Modeling impression discounting in large-scale recommender systems. In *Proceedings of KDD 2014*, pages 1837–1846, 2014.

[13] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of WSDM 2011*, pages 287–296, 2011.

[14] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of WWW 2007*, pages 521–530, 2007.

[15] M. Shokouhi, R. W. White, P. N. Bennett, and F. Radlinski. Fighting search engine amnesia: reranking repeated results. In *Proceedings of SIGIR 2013*, pages 273–282, 2013.

[16] J. Teevan, D. J. Liebling, and G. R. Geetha. Understanding and predicting personal navigation. In *Proceedings of WSDM 2011*, pages 85–94, 2011.

[17] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of RecSys 2011*, pages 109–116, 2011.

[18] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, 2010.

[19] L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang, and J. Sun. Temporal recommendation on graphs via long and short-term preference fusion. In *Proceedings of KDD 2010*, pages 723–731, 2010.

[20] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of SDM*, 2010.

[21] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: Dwell time for personalization. In *Proceedings of RecSys 2014*, pages 113–120, 2014.

[22] P. Yin, P. Luo, W.-C. Lee, and M. Wang. Silence is also evidence: Interpreting dwell time for recommendation from psychological perspective. In *Proceedings of KDD 2013*, pages 989–997, 2013.

[23] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of WWW 2010*, pages 1011–1018, 2010.

[24] Y. C. Zhang, D. O. Seaghdha, D. Quercia, and T. Jambor. Aurallist: Introducing serendipity into music recommendation. In *Proceedings of WSDM 2012*, pages 13–22, 2012.