Understanding User Behavior For Document Recommendation

Xuhai Xu
University of Washington
Seattle, WA
xuhaixu@uw.edu

Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite
Microsoft, Seattle, WA
{sdumais,hassanam,bogpop,robertro}@microsoft.com

Farnaz Jahanbakhsh
MIT
Cambridge, MA
farnazj@mit.edu

ABSTRACT

Personalized document recommendation systems aim to provide users with a quick shortcut to the documents they may want to access next, usually with an explanation about why the document is recommended. Previous work explored various methods for better recommendations and better explanations in different domains. However, there are few efforts that closely study how users react to the recommended items in a document recommendation scenario. We conducted a large-scale log study of users' interaction behavior with the explainable recommendation on one of the largest cloud document platforms office.com. Our analysis reveals a number of factors, including display position, file type, authorship, recency of last access, and most importantly, the recommendation explanations, that are associated with whether users will recognize or open the recommended documents. Moreover, we specifically focus on explanations and conduct an online experiment to investigate the influence of different explanations on user behavior. Our analysis indicates that the recommendations help users access their documents significantly faster, but sometimes users miss a recommendation and resort to other more complicated methods to open the documents. Our results suggest opportunities to improve explanations and more generally the design of systems that provide and explain recommendations for documents.

CCS CONCEPTS

• **Social and professional topics** **User characteristics**; • **Human-centered computing** **Human computer interaction (HCI)**.

KEYWORDS

Large Scale Log Analysis; User Behavior; Document Recommendation; Explanation

ACM Reference Format:
Xuhai Xu, Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding User Behavior For Document Recommendation. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366423.3380071

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380071

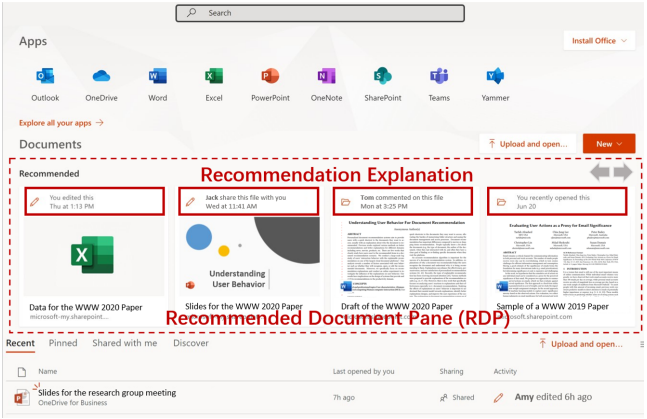


Figure 1: Interface when a user visits Microsoft Office 365. The documents in the Recommend Document Pane (RDP) are ranked by a machine learned recommendation model. Each document in the pane contains an explanation, a thumbnail or icon, and document title.

1 INTRODUCTION

Personalized recommendation is taking place in almost every aspect of our life. It offers users more exposure to what they may be interested in, and helps them save time finding what they need. For cloud-based document platforms such as Microsoft Office 365 and Google Drive, recommendations aim to provide users with quick shortcuts to the documents they may want to access next, alleviating the burden of memorizing folder structure and easing the document management and access processes. Document recommendation has important differences compared to other recommendations such as movies and shopping items. People typically know a lot about the documents (*e.g.*, the type of document, the author of the document and when they last interacted with it), and they often have a clear goal of finding or re-finding specific documents when they visit the document platform. There are also situations where users will want to open a document shared through collaborative work effort, even if they haven't seen the document before.

An accurate recommendation algorithm is important for the success of a document recommendation system. Additionally, explanations of why a document was recommended helps users recognize the document. Explanations can enhance the effectiveness, persuasiveness, and user satisfaction of personalized recommendation systems [29, 30]. Recently, the topic of explainable recommendation has received increasing attention [31]. Various methods were proposed to provide explanations of the recommendation results (*e.g.*, [2, 13, 23]). However, there is less work that specifically focuses

on users' interactions with explanations and their effectiveness in document recommendation. Studying the effects of explanations on users' behavior is important to understand how they perceive explanations, identify better explanations designs, and improve the overall user experience.

In this paper, we focus on online document platforms, where recommendations and explanations are generated based on the documents users can access, their interaction history with the documents, and their network of collaborators. The study aims to answer three main research questions. Our first question aims to understand user behavior towards recommendations (*RQ1*): *what are the characteristics of users' interaction with recommended documents on a cloud document platform?* Beyond the basic characterization, we are particularly interested in the relationship between recommendation explanations and users' behavior, which leads to our second question (*RQ2*): *how are explanations that reflect various interaction histories associated with user behavior for the recommended documents?* As correlation does not indicate causality, knowing their association does not inform us of what effect do explanations have on users. We further examine a third question (*RQ3*): *how is user behavior influenced by different explanations*?

To answer these questions, we used large-scale log data from users' interactions with a major document platform, Microsoft Office 365. Figure 1 shows the interface on the initial page of the main website office.com, with a Recommended Document Pane (RDP) in the middle. Our observational log study characterizes users' interaction behavior towards the RDP, which answers the *RQ1*. We further conducted an online randomization study on explanations to better characterize the influence of explanations on user behavior, which answers the *RQ2* and *RQ3*. Our results reveal interesting characteristics of user behavior towards various factors. The RDP helps users access their documents significantly faster. But there are also opportunities to improve explanations, *e.g.*, users sometimes missed the document in the RDP and resorted to other more complicated methods to find the file. Our findings shed light on better designs of the recommendation explanations.

Our contributions of this paper are threefold:

- Using large-scale observational log analysis, we provide the first characterization of users' behavior towards document recommendations in an online document platform.
- We examine, in detail, how explanations that reflect different interaction histories are correlated with user interaction with the recommended documents.
- Using an online randomization study, we investigate the impact of different explanations on users' behavior.

2 RELATED WORK

2.1 Characterizing User Behavior using Log Data

The development of centralized computing and Internet makes it possible to capture users' interaction with web service at a tremendous scale [12]. Large-scale log analysis enables researchers to understand and characterize user behavior in a wide range of scenarios, such as search engine [15, 24, 26], web browsing [1, 25], and email [3, 4, 11]. There are two major types of log studies [12]: 1) observational log studies, where massive amounts of log data is observed and collected to provide a descriptive overview of user behavior, such as [3, 4, 27, 28], and 2) experimental log studies, where in situ experiments are conducted and log data is collected and compared between the experiment group(s) and a control group. We conduct both types of log studies in this paper, involving over a million users. To the best of our knowledge, we are the first to deeply investigate user behavior towards recommendation explanations with such large-scale log analysis.

2.2 Explainable Recommendation

Explainable recommendation refers to recommendation systems that provide an explanation of why an items is recommended [30]. Two main strategies are used to generate explanations: one line of research focused on the interpretability of the recommendation model, such as topic modeling [19], matrix factorization [13], and deep learning [22], *etc.* Another strategy is through post-hoc analysis, where the recommendation model is treated as a black-box and separate methods are used to generate explanations. Examples include Markov logic networks [7], associate rule mining [21], *etc.* Since our focus is user behavior towards explanations rather than explanations generation, we treat our recommendation algorithm as a black box and employ a post-hoc heuristic explanation annotator. Explanations can be expressed in different styles, such as content-based [16], and context-based [17]. Moreover, explanations can be displayed in different ways, *e.g.*, text sentences [9] 1and graphics [8]. We refer readers to [30] for a comprehensive review of explainable recommendation. We display explanations with natural language based on users' actions on the documents and their collaboration network.

2.3 User Reactions Towards Explanations

The effect of recommendation explanations needs to be evaluated with real users [14]. Existing works usually ask participants to answer surveys after the explanations are displayed. The metrics include participants' subjective ratings on quality, trust, satisfaction, efficiency, *etc.* [9, 10, 23] However, these evaluations usually happen under an experiment setting such as Amazon MTurk that does not reflect real user behavior. Only a few studies evaluate explanations' influence under real situations. Zhang et al. [31] evaluated their explanations on an online shopping platform using customers' click-through rate (CTR) and purchase rate. McInerney et al. [20] employed the rate of service users' playing at least one song from the recommended playlist on a music platform to evaluate the explanations. The metrics in both works are some forms of *click rate*, while the interactions with online documents are much richer. In this paper, we investigate various behavior metrics to characterize user behavior, including searching, recognizing and clicking behavior. To our knowledge, we are the first to investigate rich user behavior towards recommendation explanations.

3 ANALYSIS SCOPE AND LOG DATA

We first introduce our log data and analysis scope. More importantly, we introduce the concept of users' intent to open a document to pinpoint our focus on the right population.

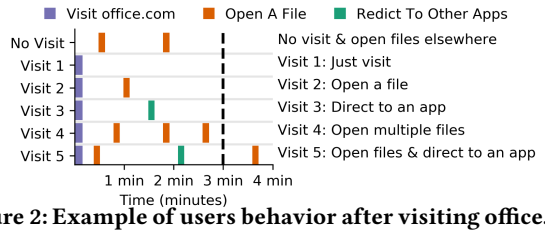


Figure 2: Example of users behavior after visiting office.com. Our analysis focuses on the cases where users open a document *somewhere* within 3 minutes after visiting the website.

3.1 Log Data

We analyze random samples of log data of the office.com web client in North America from two periods of time. The first period is for the observation log analysis (*RQ1*), with the range from May 1 to 31, 2019 involving millions of users. The second period is for the explanation-randomization study on 10% of these users (*RQ2*, *RQ3*), at the time period from August 19 to September 1, 2019.

As we focus on the RDP (see Figure 1), we only study users who have enough candidates in the RDP (*i.e.*, 4 or more), so that they could see a full RDP page when visiting the website) and clicked on the RDP at least once during the analysis period. From this subset, we sample approximately 800K users and their (millions of) visits to office.com in the first period, and randomly sample 10% of the users in the second period to receive the randomization treatment.

The log data contains two types of interactions: 1) Interactions on the cloud platform: interactions on items, apps, and other links on office.com. 2) Interactions with the document: open, edit, comment, *etc.*In addition to user behavior information, the logs also contain rich metadata of the documents in the RDP, including a unique document id, display position index of every document, the type of recommendation explanations, document size, *etc.* The logs do not contain any document or explanation content or personally identifiable information (PII). Note that we treat the recommendation model and the explanation generation model as black boxes and only log the documents recommended to the user.

3.2 Users' Intent on office.com

Users visit office.com (Figure 1) for a variety of reasons, including to find documents or to navigate to Office apps or sites. Figure 2 illustrates some examples of users' actions after visiting office.com. Sometimes users use the website as a hub to open a web app (*e.g.*, Outlook), sometimes they have the intent to find and open a document. In this paper, we are interested in cases where users have the intent to open a document when they visit office.com. To capture the intent, we examine the subset of visits where users open a document *somewhere* within 3 minutes *after* they visit office.com. We select 3 minutes as the threshold since this is the 99^{th} percentile of the interval between visiting and document opening according to our log data. It is noteworthy that *somewhere* includes all cases, such as RDP, the recent document list, *etc.* All of our analyses only involve these visits with users' intent to open a document.

According to the log data, the most common area that users resort to when having the intent to open a file is the RDP (65.5%). Moreover, the second common area, *i.e.*, recent document list (20.4%), is more transparent where the order is just based on the recency, thus less interesting. As such, we focus on understanding user behavior on the RDP in this paper.

4 USERS' INTERACTIONS WITH RDP

In this section, we provide a comprehensive analysis of log data to answer *RQ1* by investigating various aspects of users' behavior before and after opening the document. We study one of the most fundamental yet important factors: display position.

4.1 What Is Users' Click Behavior on The RDP?

To characterize users' click behavior on the RDP, we use a common metric **click through rate (CTR)** defined as follows.

$$\text{CTR} = \frac{\text{Number of Clicks}}{\text{Number of Visits}} \tag{1}$$

There are up to 16 candidates in total in the RDP, ranked by the recommendation model. When users visit office.com, the first four are shown and users can navigate to the other three pages (see Figure 1). We identify a few interesting findings from the figure.

- *Documents on the left side have higher CTR than those of the right side on each page.* The four pages share a similar pattern: the CTR decreases from the left to the right in one page. This can be caused by two factors: 1) ranking bias, the ranking order by the recommendation model, 2) interface bias, that users usually scan the RDP starting from the left to the right and may pay more attention to the documents at the beginning.
- *The CTR jumps up between two pages, especially from the first to the second page (position 4 to 5).* This reveals an interesting interface effect. If a user navigates to the next page, especially at the first navigation, it indicates that they notice the RDP and is leveraging it to find the document, thus leading to a higher CTR. Similar behavior is also observed in web search [6].

4.2 Is the RDP Really Helpful?

The CTR only reflects the ratio of whether users click on the documents in the RDP. It does not indicate whether the RDP benefits users when they want to find a document. To capture this, we further define two metrics: the recognize rate and the time to open.

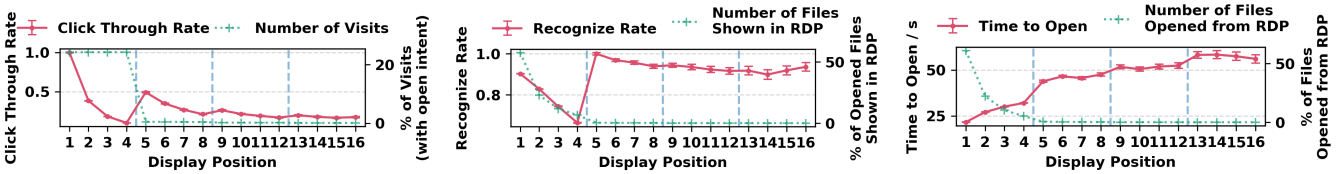
4.2.1 Recognize Rate. Given the recommendation algorithm successfully predicts the document that is eventually opened by the user and displays the documents in the RDP, will the user recognize it and open it from the RDP?

Our analysis indicates that the algorithm often does a good recommendation, *i.e.*, the documents that are eventually opened *somewhere* are recommended by the model and shown to users in the RDP. However, only in 73.4% of the cases users will recognize it and open it from the RDP. In the rest of the 26.6% cases, although the documents that are shown in the RDP and users see it, users miss the documents or rely on their habitual practice, and still open the documents elsewhere. Among these cases, 38.9% of them are from the recent document list and the rest of the 61.1% are opened elsewhere other than the direct access (*i.e.*, one click) on office.com, such as through email, browsing, *etc.*

We define the **recognize rate (RR)** as follows,

$$\text{RR} = \frac{\text{Docs Opened from the RDP}}{\text{Eventually-opened Docs Shown in the RDP}} \tag{2}$$

The RR is interestingly different from the CTR: the RR is based on an accurate recommendation and measures an interesting aspect



(a) CTR of display positions (b) RR of display positions (c) TTO of display positions

Figure 3: Click Through Rate (CTR), Recognize Rate (RR), and Time to Open(TTO) of 16 display positions in the RDP, with CTR and RR normalized by the maximum. Error bar indicates the bootstrap standard error. The vertical lines indicates pages. The green dashed line shows the number of the visits where the document at the position is shown on the website.

of "success rate" for users to recognize the right recommendation, while the CTR depicts the interaction frequency. Figure 3b shows the interesting effect of position.

- *Documents on the left side has higher RR than those of the right side on the first page.* On the first page, the RR is similar to the CTR in a way that the RR is decreasing from position 1 to 4.
- *Once users navigate to other pages, the RR remains very high.* After a big jump of the RR when users navigate to the second page, the RR remains at a high level, which is different from the CTR. This indicates that although not often (as indicated by the green dashed line), when users are actively looking for specific documents, they will maintain an active recognition behavior after navigating to later pages of the RDP, leading to the high RR.

4.2.2 Time to Open. The definition of **time to open (TTO)** is straightforward. It indicates the time needed by a user to locate and open the document after they visit office.com.

$$\text{TTO} = T(\text{Open A Doc } \textit{somewhere}) - T(\text{Visit office.com}) \quad (3)$$

Our results show that the RDP significantly shortens the time to open the document. It only takes 52.6% of the time compared to the cases when documents are in the RDP but opened elsewhere and 38.3% when documents are not shown in the RDP. Figure 3c shows the TTO on different positions for the documents opened from the RDP. The larger the position number, the longer it takes. Moreover, the increase of the time between pages is more significant than the increase within a page. This reflects the time needed for users to scan from left to the right, and to navigate to the next page.

In the rest of the analysis, we normalize the effect of display position (also plus file type) by dividing the marginal value.

5 HOW DO EXPLANATIONS ASSOCIATE WITH USERS' INTERACTION?

Given the basic characterization of the user behavior with the RDP, we answer the *RQ2* by investigating the association between the recommendation explanations and the three behavior metrics described in the preceding section. Moreover, users' perception of documents builds on their historical interactions with the documents, which may affect users' reaction to the explanations. Therefore, we further investigate the relationship between the explanations and two aspects of users' historical interactions: the authorship and the time since last-open.

5.1 Explanations and Randomization Study

There are 14 predefined explanation types in the generator (see Table 1). An explanation generation model (independent of the recommendation model) ranks them and the corresponding language is generated from a pre-defined template. Note that since a document can have different activities during its lifecycle, the same document can show up with different explanations at different times. For simplicity, we group the 14 explanation types into four action groups (edit, comment, open, and share). As editing is one of the most common actions. we further divide the edit action by the subject (me versus others), as summarized in Table 1.

To remove the bias of the explanation generator while maintaining the validity of the explanation, we randomly sampled a subset of users and conducted an explanation-randomization study for two weeks (from August 19 to September 1, 2019). When a user visits office.com, for each document recommended by the model, we select the top four explanations and randomly pick one as the explanation displayed to the user. To reduce the bias of the documents with fewer explanations, we excluded the documents that have less than four explanation candidates.

5.2 Behavior Metrics among Explanations

We investigate the associations between the five explanation groups and the three behavior metrics as defined in Section 4.

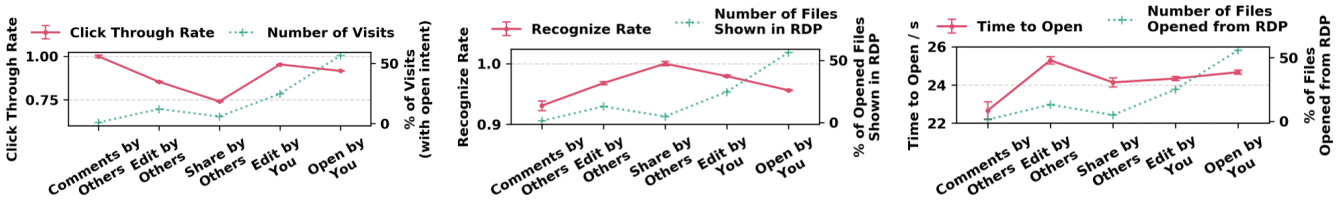
All ANOVAs (with Greenhouse-Geisser correction if there is a sphericity violation) and pairwise post hoc t-tests (with Holm's sequential Bonferroni procedure correction) show significance ($p < 0.05$), thus we omit these statistics in the rest of this section.

5.2.1 CTR. We highlight a few findings from the Figure 4a.

- *Among the collaborative explanation groups, Comment by Others has the highest CTR.* This reflects that compared to co-workers' editing and sharing action, commenting usually indicates feedback from collaborators, which requires more involvement and thus triggers more attention that leads to higher CTR.
- *For the individual explanation groups, Edit by You has higher CTR.* Although opening a document is the most frequent explanation, our results reveal that users may be more familiar with and react

Table 1: Five groups based on explanation types.

Exp Group	Exp Type	Property	%
Comment by Others	CommentBySingleOthers, CommentByMultipleOthers, CommentReplyToYou, CommentReplyByOthers, MentionBySingleOthers, MentionByMultipleOthers	Collaborative	1.4
Edit by You	EditByYou,EditByYou&Others	Individual	28.3
Edit by Others	EditBySingleOthers, EditByMultipleOthers	Collaborative	17.9
Share by Others	SharedWithMe	Collaborative	7.0
Open by You	FrequentlyOpen,Recently-Open,WeeklyOpen	Individual	45.4



(a) CTR of explanation groups (b) RR of explanation groups (c) TTO of explanation groups

Figure 4: Click through Rate (CTR), Recognize Rate (RR) and Time to Open (TTO) of the five explanation groups.

more actively to the documents they opened and edited than the documents they just opened and read.

- *Share by Others has the lowest CTR.* This indicates that users less frequently use the RDP to access shared documents compared to documents with other reasons.

5.2.2 RR. We notice two interesting explanation groups that have reversed results in the RR, as shown in Figure 4b.

- *Comment by Others has the highest CTR but the lowest RR.* Users are very likely to click on documents with *Comment by Others* explanation (high CTR). However, if a document with this explanation is shown in the RDP, users are also likely to miss it (low RR) and open this document elsewhere. This reflects that users not only frequently use the RDP for these documents but also resort to other methods such as email to open them.
- *Share by Others has the lowest CTR but the highest RR.* Users are less likely to click on the shared documents in the RDP (low CTR). However, if they eventually open a shared file after visiting office.com, most of the cases they access it through the RDP (high RR). This shows that the RDP works effectively for users to open shared documents once they pay attention to.

5.2.3 TTO. Figure 4c indicates that *Comment by Others* requires significantly less time than documents with other explanations. This is in line with the findings that documents with *Comment by Others* usually require more engagement, thus faster reactions.

5.3 How's Authorship × Explanations?

Whether the user is the author of the document (*i.e.*, creator) will affect the user's reaction to the document. Understanding this is important to customize the explanations for documents with different authorship conditions. Figure 5 reveals several interaction that shows significance, as highlighted below.

- *Comment by Others and Edit by Others have higher CTR, RR, and lower TTO when the user is the author.* This shows that users are more likely to react actively to others' actions on the documents if they created these documents. They may be more interested in checking these activities since these documents are "theirs".
- *Share by Others have lower CTR, lower RR, and higher TTO when the user is the author of the file.* Shared documents have a reversed trend: users react to others' actions less actively if others share the documents that were originally created by themselves. Users initiated the documents and when the documents are shared by others back to them, they may feel they are already aware of the documents content , leading to less reaction.

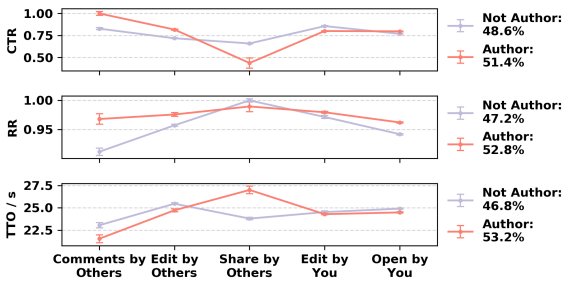


Figure 5: CTR, RR, and TTO of the five explanation groups across authorship conditions.

5.4 How's Last-Open Interval × Explanations?

Another interesting user-behavior factor is the interval between the last and the current open time, *i.e.*, time since last-open. We select four different intervals in Figure 6. The findings are summarized as follows:

- *Generally, the longer the time since last-open, the lower the CTR and the higher the TTO.* The older the documents are, the less likely users will interact with them. Our finding suggests that similar to emails, the lifecycle of documents is also quite short [4].
- *The RR is low if documents were opened earlier today. It becomes high once documents were opened earlier than yesterday.* We observe a reverse trend between the CTR and the RR. As the CTR decreases, the RR increases. This indicates that the RDP can "remind" users about the old documents and becomes the major channel to access them. However, when documents are recent (*i.e.*, opened earlier today), although users open them frequently through the RDP, users also use other methods to open the file.

6 HOW IS USER BEHAVIOR INFLUENCED BY DIFFERENT EXPLANATIONS?

We further answer *RQ3* by conducting a pairwise comparison between different explanation groups. Our results indicate the differences between explanation pairs: when two explanations are valid for a document, showing one explanation will trigger more active reactions than the other. This reveals that there are opportunities to improve explanations under different contexts.

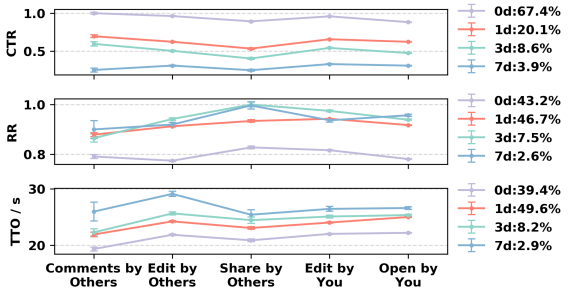
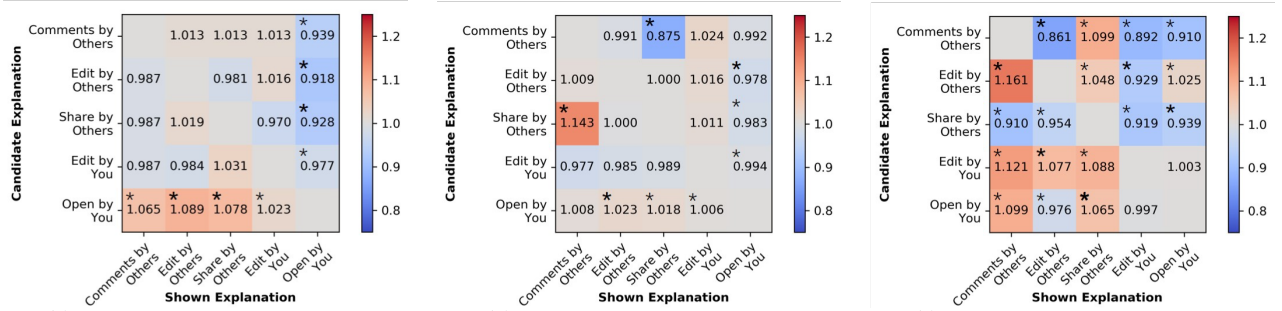


Figure 6: CTR, RR, and TTO of the five explanation groups across differnet time since last-open.



(a) Pairwise Comparison on CTR (b) Pairwise Comparison on RR (c) Pairwise Comparison on TTO

Figure 7: Pair wise comparison between the explanation groups. Each cell shows the ratio between the shown explanation and the candidate on CTR, RR, and TTO, respectively. Thus the multiplication of two diagonal symmetric cells equals one. The ratio of TTO is reversed so that any ratio above 1.0 indicates more active reaction towards the explanation. T-test is used to measure the difference. * (bold) indicates statistical significance ($p < 0.05$) and * indicates marginal significance ($p < 0.1$).

For each pair among the five explanation groups (10 pairs in total), *e.g.*, *Comment by Others* and *Edit by Others*, we first narrow down the cases where both explanations are in the top four candidate explanation list and either of them is displayed. Then, we compare user behavior between two cases, one with *Comment by Others* shown in the RDP, the other with *Edit by Others* shown in the RDP. Note that there is bias introduced by the candidate explanation list, *i.e.*, the candidate list may reflect certain properties of the document. To remove the bias, we further normalize by dividing by the marginal value of each candidate explanation list.

We summarize the comparisons that lead to significantly different user behavior (based on t-test, with significance level at $p = 0.05$, and marginal level at $p = 0.1$.). We particularly focus on the results that are not in line with the results in Figure 4.

- *Although Comment by Others has the highest CTR in Figure 4a, the pairwise comparison indicates that its CTR is only marginal-significantly higher than that of Open by You and not higher than others.* When *Comment by Others* and other explanations are both in the candidate list, displaying which explanation won't significantly affect users' behavior. The CTR stays high.
- *The RR of Comment by Others and Share by Others have a reversed order.* In Figure 4a, *Comment by Others* (the lowest RR) and *Share by Others* (the highest RR) are at opposite positions. However, in Figure 7b they are reversed. This reveals that when documents are shared and have comments by others, users are more likely to recognize them from the RDP when they are displayed with the *Comment by Others*.
- *Open by You has significantly lower CTR and RR than all other explanations.* Although this explanation is the most frequent one in the candidate list (see Table 1), it contains the least information, leading to inactive reactions.
- *The TTO results can establish a "complete pairwise order" as Edit by You > Edit by Others $\widetilde{>}$ Open by You $\widetilde{>}$ Comment by Others $\widetilde{>}$ Share by Others ($>/\widetilde{>}$ indicates $p < 0.05 / 0.1$).* Note that essentially the pairwise comparison does not have transitivity. But this order can still provide a straightforward relationship between the explanations. The sharing/editing explanations take the shortest/longest time for users to click on the documents.

7 DISCUSSION

Our results, not only reveal the characteristics of user behavior towards explainable recommendations, but also suggest better designs of the explanations for document recommendation systems. These findings can potentially be generalized to other known-item and navigational recommendation systems, *e.g.*, [5, 18]. We summarize a few potential suggestions driven by our findings.

- Section 5.3 reveals that when the user is the author of the file, showing *Comment by Others* or *Edit by Others* explanations can trigger more active reactions than others (see Figure 5).
- Section 5.4 suggests that if the RDP is showing an old document that has not been opened for a long time, the explanation *Share by Others* can help users to better recognize the file and faster access the file (see Figure 6).
- As shown in the pairwise comparison, *Open by You* contains the least information and does not trigger a lot reactions. Whenever there is other explanations that are available, documents should be shown with other explanations.
- Comparison matrices in Figure 7 can serve as a good reference when deciding between two explanations, depending on designers' goal. For instance, if the recognition is the major concern, *Comment by Others* is preferred by *Share by Others*. If the time is the concern, then the preference order is reversed.

There are some important limitations of this work. First, the three behavior metrics only depict certain aspects of user behavior. Other behaviors such as collaborative actions and detailed editing actions will be included in future work. Second, we did not analyze user behavior in interacting with other aspects of the website. The recent document list is of special interest because it represents a sizeable proportion of how users access files. We will compare the RDP and the recent document list in future work. Third, in the explanation-randomization study, we did not experiment with a "no explanation" option since this could adversely affect users' experience. We hope to try a limited study of this baseline to understand the effectiveness of the explanations in future work.

8 CONCLUSION

In this paper, we conduct large-scale log studies to characterize user behavior towards explainable recommendations. Our analysis leverages the data from a major cloud document platform office.com.

We define three metrics to depict user behavior before opening the documents through the Recommended Document Pane (RDP). We first study one-month data involving millions of users to understand behavior characteristics in light of these metrics. Then, through an explanation-randomization study, we analyze two-week worth of data involving hundreds of thousands of users to understand the association between recommendation explanations and user behavior, as well as the influence of explanations on user behavior. Our results reveal a number of interesting findings that shed light on better explanation design in the future.

REFERENCES

[1] Eytan Adar, Jaime Teevan, and Susan T Dumais. 2008. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1197–1206.

[2] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms* 11, 9 (2018), 137.

[3] Qingyao Ai, Susan T Dumais, Nick Craswell, and Dan Liebling. 2017. Characterizing email search using large-scale behavioral logs and surveys. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1511–1520.

[4] Tarfah Alrashed, Ahmed Hassan Awadallah, and Susan Dumais. 2018. The lifetime of email messages: a large-scale analysis of email revisitation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 120–129.

[5] Vamsi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. Towards task recommendation in micro-task markets. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[6] Ricardo Baeza-Yates, Carlos Hurtado, Marcelo Mendoza, and Georges Dupret. 2005. Modeling user search behavior. In *Third Latin American Web Congress (LA-WEB'2005)*. IEEE, 10–pp.

[7] Roi Blanco, Diego Ceccarelli, Claudio Lucchese, Raffaele Perego, and Fabrizio Silvestri. 2012. You should read this! let me explain you why: explaining news recommendations to users. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1995–1999.

[8] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 35–42.

[9] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 175–182.

[10] Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 17–28.

[11] Susan Dumais, Edward Cutrell, Jonathan J Cadiz, Gavin Jancke, Raman Sarin, and Daniel C Robbins. 2016. Stuff I've seen: a system for personal information retrieval and re-use. In *Acm sigir forum*, Vol. 49. ACM, 28–35.

[12] Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. 2014. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*. Springer, 349–372.

[13] Yunfeng Hou, Ning Yang, Yi Wu, and S Yu Philip. 2019. Explainable recommendation with fusion of aspect information. *World Wide Web* 22, 1 (2019), 221–240.

[14] Farnaz Jahanbakhsh, Ahmed Hassan Awadallah, Susan T. Dumais, and Xuhai Xu. 2020. Effects of Past Interactions on User Experience with Recommended Documents. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver, BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 10. https://doi.org/10.1145/3343413.3377977

[15] Bernard J Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information processing & management* 42, 1 (2006), 248–263.

[16] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 487.

[17] Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 157–166.

[18] Yumao Lu, Fuchun Peng, Xin Li, and Nawaaz Ahmed. 2010. Techniques for navigational query identification. US Patent 7,693,865.

[19] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 165–172.

[20] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 31–39.

[21] Georgina Peake and Jun Wang. 2018. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2060–2069.

[22] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 297–305.

[23] Amit Sharma and Dan Cosley. 2013. Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1133–1144.

[24] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. In *ACm SIGIR Forum*, Vol. 33. ACM, 6–12.

[25] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1591–1600.

[26] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael AS Potts. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 151–158.

[27] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 116 (Sept. 2019), 33 pages. https://doi.org/10.1145/3351274

[28] Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2020. Recognizing Unintentional Touch on Interactive Tabletop. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1 (March 2020), 27. https://doi.org/10.1145/3381011

[29] Markus Zanker. 2012. The influence of knowledgeable explanations on users' perception of a recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 269–272.

[30] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).

[31] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 83–92.