Inferring Users' Critiquing Feedback on Recommendations from Eye Movements

Li Chen(✉), Feng Wang, and Wen Wu

Department of Computer Science, Hong Kong Baptist University,
224 Waterloo Road, Kowloon, Hong Kong
{lichen,fwang,cswenwu}@comp.hkbu.edu.hk

Abstract. In recommender systems, *critiquing* has been popularly applied as an effective approach to obtaining users' feedback on recommended products. In order to reduce users' efforts of creating critiquing criteria on their own, some systems have aimed at suggesting critiques for users to choose. How to accurately match system-suggested critiques to users' intended feedback hence becomes a challenging issue. In this paper, we particularly take into account users' eye movements on recommendations to infer their critiquing feedback. Based on a collection of real users' eye-gaze data, we have demonstrated the approach's feasibility of implicitly deriving users' critiquing criteria. It hence indicates a promising direction of using eye-tracking technique to improve existing critique suggestion methods.

Keywords: Recommender systems · Critiquing feedback · Eye movements · Fixation metrics · Feedback inference

1 Introduction

In current online environments, recommender systems have been widely applied in various scenarios to support users to make product choices (e.g., e-commerce, social media, tourism, finance). Especially, in the situations where it is difficult to obtain users' historical records like ratings for collaborative filtering, case-based or preference-based methods have mainly been used to generate recommendations by retrieving items that are similar to users' queries/preferences [2,8]. In these systems, one popular approach to obtaining users' feedback on recommendations is *critiquing*, which has become the core feedback mechanism in so called conversational recommenders [17,22] and critiquing-based recommender systems [10]. Specifically, the *critiquing* allows users to critique a recommended product in terms of its attribute values (e.g., "*I would like to see some laptops with different manufacture and higher processor speed*"), based on which the system will return new recommendations that satisfy their critiquing criteria. Previous studies show that this critiquing process is effective to assist users in exploring the product space, refining their requirements, and making more confident decisions [6,18,19].

A. Goel et al. (Eds.): ICCBR 2016, LNAI 9969, pp. 62–76, 2016.
DOI: 10.1007/978-3-319-47096-2_5

So far, there are two major methods of acquiring users' critiquing feedback. One is *user-initiated critiquing* that requires users to specify critiquing criteria on their own [10]. For example, Fig. 1.a shows the screenshot of Example Critiquing interface [5], where users need to indicate which attributes to "keep" (keeping its existing value), "improve" (improving its existing value, e.g., cheaper), or "take any suggestion" (i.e., "compromise", accepting a compromised value). The advantage of this elicitation approach is that it can give users maximal freedom of creating any critiques they wish and stimulate them to make tradeoffs among attributes (i.e., sacrificing values on less important attributes for guaranteeing the intended improvements on more important ones), but it unavoidably demands extra user efforts and might hence be limited in real applications. Another method is *system-suggested critiquing* that proposes a set of critiques for users to choose [3,19]. For instance, the Dynamic Critiquing system generates several compound critiques (each operating over multiple attributes, e.g., "*Less Optical Zoom & More Digital Zoom & A Different Storage Type*") according to remaining product cases' availability (see Fig. 1.b) [22]. Intuitively, the system-suggested critiquing method could reduce users' critiquing efforts, but when the suggested critiques cannot precisely match users' intended feedback, it is likely that users will be involved in longer interaction session by pursuing other ways to locate their target choice [6].

In this paper, we focus on investigating users' eye-movements on recommendations to infer their critiquing feedback, so as to potentially improve system-suggested critiquing. Based on a collection of real users' eye-gaze data and their true critiques, we have empirically verified two hypotheses: one is the feasibility of inferring what product a user tends to critique (i.e., *critiqued product*) from her/his fixations laid on different products; another is about deriving the user's concrete *critiquing criteria* for the product's attributes. That is, what attributes s/he may be inclined to keep, improve, or compromise. Furthermore, we have compared different fixation metrics, including *fixation count*, *total fixation duration*, and *average fixation duration*, in terms of their inference accuracy.

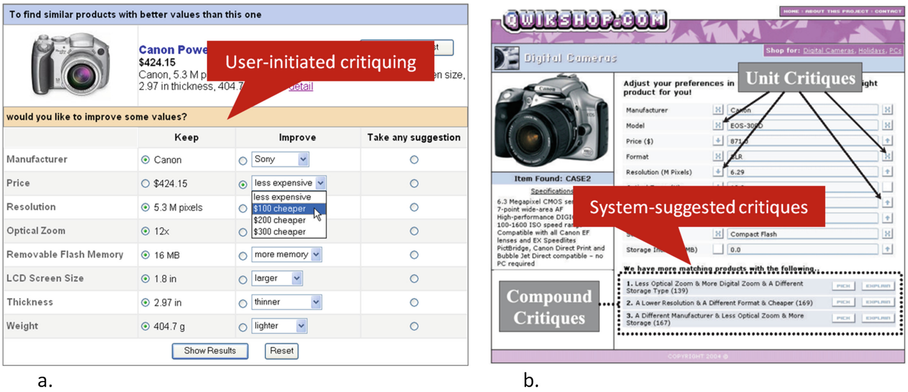
In the following, we first introduce related work (Sect. 2), and then give our research statement and hypotheses (Sect. 3). The experiment for data collection is described in Sect. 4, and in Sect. 5 we show the results. At the end, we conclude our findings and indicate future directions (Sect. 6).

2 Related Work

2.1 Critiquing-Based Recommender Systems

As mentioned before, existing critiquing-based recommender systems can be classified into two categories [10]: *user-initiated critiquing*, with ATA [16] and Example Critiquing [5] as representative systems; *system-suggested critiquing* that has been adopted in FindMe [3], Dynamic Critiquing [22], MAUT-based compound critiquing [28], and preference-based Organization [8].

Take Example Critiquing system as an example to illustrate user-initiated critiquing process [5]: it first presents some products to a user that best match



To find similar products with better values than this one

Canon Powe
$424.15
Canon, 5.3 M p … en size,
2.97 in thickness, 404.7 … detail

would you like to improve some values?

	Keep	Improve	Take any suggestion
Manufacturer	● Canon	○ Sony	○
Price	○ $424.15	● less expensive	○
Resolution	● 5.3 M pixels	○	○
Optical Zoom	● 12x	○	○
Removable Flash Memory	● 16 MB	○ more memory	○
LCD Screen Size	● 1.8 in	○ larger	○
Thickness	● 2.97 in	○ thinner	○
Weight	● 404.7 g	○ lighter	○

QWIKSHOP.COM
HOME : ABOUT THIS PROJECT : CONTACT
Digital Cameras
Shop for: Digital Cameras, Holidays, PCs

Item Found: CASE2

Specifications

6.3 Megapixel CMOS se
7-point wide-area AF
High-performance DIGI
100-1600 ISO speed ra
Compatible with all Canon EF lenses and EX Speedlites
PictBridge, Canon Direct Print and Bubble Jet Direct compatible – no PC required

Adjust your preferences in … ight product for you!

Manufacturer	Canon
Model	EOS-300D
Price ($)	871.0
Format	SLR
Resolution (M Pixels)	6.29
	Compact Flash
Storage In… (MB)	0.0

We have more matching products with the following..

1. Less Optical Zoom & More Digital Zoom & A Different Storage Type (139)
2. A Lower Resolution & A Different Format & Cheaper (169)
3. A Different Manufacturer & Less Optical Zoom & More Storage (167)

Fig. 1. a. User-initiated critiquing in Example Critiquing system [5]; b. System-suggested critiquing in Dynamic Critiquing system [19].

her/his initially specified preferences; then it stimulates the user to select a near-satisfactory product and critique it in terms of its attribute values; in the next recommendation cycle, the system will return a new set of recommendations according to the user's critiquing feedback. Experiments show that for a user to reach his/her target choice, a number of critiquing cycles are usually required [7].

As for system-suggested critiquing, some systems like FindMe pre-design some static critiques for users to pick, but since those fixed critiques cannot reflect available products' realistic status, other systems attempt to dynamically generate critique suggestions being adaptive to remaining product cases' characteristics [19,22] or users' attribute preferences [8,28]. However, an empirical user evaluation on a hybrid critiquing interface, which combines both user-initiated critiquing support and adaptive critique suggestions, shows that users more frequently created critiques on their own than choosing the suggested critiques, implying the latter approach's limited accuracy [6].

In order to save users' critiquing efforts, some researchers have also endeavored to adopt speech-based critique input interface [14], or harness other users' critiquing histories to guide the current user [20,26]. It shows though these methods are capable of enhancing system efficiency, the limitation of system-suggested critiques is still not well resolved.

2.2 Eye Tracking Studies in Recommender Systems

The development of eye tracking technology has enabled academic and commercial sectors to apply it in various interaction designs [21]. In recommender systems, it has mainly been adopted for two purposes. One was to evaluate the recommendation interface's usability. For instance, one user experiment measured the effect of interface layout design on users' visual searching pattern [9]. It shows users tend to fixate more on the top area if recommended products are displayed in a list layout, but will be directed to view more products if the recommendations are arranged in a category structure. Another experiment investigated whether users would gaze at recommendations during their entire product searching process [4]. Its results verify the important role of recommendations in users' purchase decision.

As the second purpose, some researchers have exploited eye-movement metrics to elicit users' implicit *relevance feedback* on recommendations, i.e., "positive" or "negative" (or called "like" or "dislike"). For instance, in [13], the documents that users consumed higher amount of fixations and longer average fixation time were regarded with "positive" feedback. They then used clustering and content based techniques to retrieve similar documents and recommended them to the user. Some studies emphasized developing algorithms to incorporate eye-based relevance feedback, such as interactive genetic algorithm [11], evolutionary programming [15], and attention prediction method [27].

However, little work attempts to infer users' critiquing criteria for product attributes (i.e., *critiquing feedback*) through eye tracking, which is in nature more challenging than that for relevance feedback.

3 Research Statement and Hypotheses

What a person is looking at is assumed to indicate her/his thought "on top of the stack" of cognitive processes. This "eye-mind" hypothesis means that eye-movement recordings can provide a dynamic trace of where a person's attention is directed in relation to a visual display. In practice, the process of inferring useful information from eye-movement recordings involves defining "Areas Of Interest" (AOI) over certain part of a display or interface under evaluation, and analyzing the eye movements that fall within those areas [21]. In our work, we define AOI at two levels (see Fig. 2): *product level* and *attribute level*. At product level, all descriptions about a recommended product, including its title, image, and major attributes' values (e.g., a laptop's price, operating system, processor speed, etc.), are comprised in one area. At attribute level, each attribute of the recommended product is treated as a specific area.

The metrics used to analyze eye-movement data are commonly related to *fixation*. Specifically, each fixation is a spatially stable gaze point, during which most information acquisition and processing occur. Its minimum duration is usually set as 200 ms [23]. We concretely adopt three popular fixation-derived metrics in our work, because they can represent users' relative engagement with the interface object [12,21,25]:

- **Fixation Count (FC)** - the number of times a user fixates on an AOI;
- **Total Fixation Duration (TFD)** - the sum of the duration of all fixations a user has laid in an AOI;
- **Average Fixation Duration (AFD)** - the average duration of a fixation in an AOI.

Given a user's fixation values at both product and attribute levels, the question we are interested in answering is whether they could be utilized to infer the user's critiquing feedback. At the first step, it is to infer what product within a set of N recommendations the user would take as near-satisfactory and critique. Intuitively, we may hypothesize that the product with higher fixation values would be more likely to be selected, since more fixations on an object suggest that it is more important and engaging in some way [21,25].

Hypothesis 1: *Within a set of N recommendations, users would tend to critique the product for which they have consumed higher fixation values.*

The second step is to infer the user's critiquing criterion for each attribute of the selected product. According to [5], there are three critique options: *keep*, *improve*, and *compromise* (as mentioned in Sect. 1). If a user *keeps* an attribute's existing value, it indicates s/he is satisfied with it, so we may assume the user would have spent certain fixations on this attribute when s/he evaluated the whole set of recommendations. If the user chooses to *improve* its value, it also implies the user has fixated on this attribute. The duration may be even longer than that on attributes for keeping, since the user might have compared different values of the attribute across different products and finally chosen one that is the best among all but still not fully satisfying her need. On the contrary, the attribute the user *compromises* may be of the fewest fixations, as it is what the user tends to sacrifice and hence less important than attributes for keeping or improving. Therefore, we can have the following hypothesis:

Hypothesis 2: *At attribute level, for the attributes of which users have consumed higher fixation values, they would be likely to improve, followed by some they may keep, and then the others with fewer fixations to compromise.*

4 Data Collection

With the purpose of verifying our hypotheses and comparing different fixation metrics (i.e., FC, TFD, and AFD), we conducted an experiment to collect users' eye movements on recommendations and their true critiquing criteria. In this section, we first introduce the experimental system, and then experimental procedure, participants, and data analysis results.

4.1 Experimental System

We chose Example Critiquing [5] as the experimental system, because its user-initiated critiquing support allows us to obtain users' true critiquing criteria for product attributes (see Fig. 1.a). To be specific, we adopted one of its prototypes, Laptop Finder, for conducting our experiment. Its laptop catalog was extracted from a commercial e-commerce website, each product described by 10 primary attributes (i.e., manufacturer, price, operating system, battery life, display size, hard drive capacity, installed memory, processor class, processor speed, and weight). There are four major steps during users' interaction with this system:

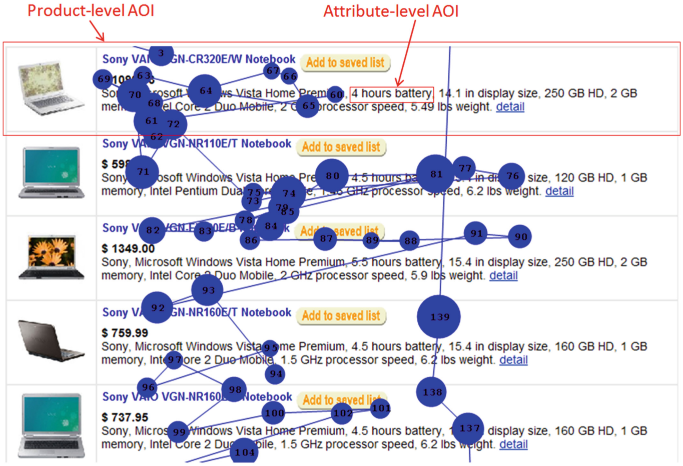


Fig. 2. Product-level and attribute-level Areas of Interest (AOI) for fixation analysis, and an example of a user's eye-gaze plot on recommended products (each fixation is represented by a blue circle). (Color figure online)

Step 1: Initial Preference Elicitation. The system first obtains a user's initial preferences by asking her/him to enter a product as query, or to state some specific preferences for product attributes. The system then builds a preference model for the user, which is formally represented as $Pref(u) = \{< V_i, W_i > | 1 \leq i \leq A\}$, where V_i denotes the user u's value preference for attribute a_i and W_i is a_i's relative weight.

Step 2: Recommendation Generation. Then, the system returns a set of N products (e.g., $N = 25$ in Laptop Finder) that are most relevant to $Pref(u)$. Formally, a utility score is computed for each product to indicate its relevance: $U(P_j) = \sum_{i=1}^{A} W_i \times V_i(x_i)$, where a product P_j contains attribute values $\mathbf{x} = \{x_i\}_{i=1}^{A}$. The products with higher utility scores are recommended.

In the recommendation interface, each product is described by three blocks of information (see Fig. 2): title, image, and ten primary attributes' values.

Step 3: Critiquing Feedback Elicitation. Within the set of N recommended products, if the user cannot locate her/his target choice, s/he could select one product that is near-satisfactory and provide critiquing feedback on it.

Actually, users can initiate different types of critique. In terms of critiquing modality [5], there are two types: *similarity-based critique* (e.g., "*Find some products similar to this one*") by "keeping" all of the critiqued product's attribute values, and *improvement-based critique* (e.g., "*Find some products that are cheaper*") by "improving" some attribute values. For the latter, the user may even "compromise" the values of less important attributes. Regarding critiquing complexity [5], there are also two types: *unit critique* if the user "improves" or "compromises" only one attribute at a time, and *compound critique* if multiple attributes are involved in one critique (e.g., "*Find some products that are cheaper and bigger*").

Step 4: Preference Refinement. The system will update the user's preference model $Pref(u)$ according to her/his critique. For instance, the attribute's weight will be increased by β if it is "improved" or decreased by β if "compromised" ($\beta = 0.25$ in Laptop Finder).

Then, the system will go back to Step 2 to resume a new recommendation cycle (from Steps 2 to 4). This interaction process continues until the user accepts a product as her/his final choice.

4.2 Experimental Procedure and Participants

The experiment was in form of a controlled lab study. A Tobii 1750 eye-tracker that is integrated with a 17″ TFT screen was used to record each user's eye movements when s/he viewed recommended products. Its resolution setting is 1290 × 1024 pixels, and can sample the position of a user's eyes by every 20 ms. The monitor frame has near infra-red light-emitting diodes, which allow for natural tracking without placing many restrictions on the user.

The user task was to "*find a product you would purchase if given the opportunity by using the Laptop Finder system.*" An administrator was present in each experiment. She debriefed the experiment's objective to the participant and asked her/him to fill in a demographic questionnaire at the beginning. Then, the participant was prompted to get familiar with the system's interface during a warm-up period. Subsequently, after the eye-tracker calibration was performed, the participant started to accomplish the given task.

During each recommendation cycle, in addition to recording each participant's eye movements, we retrieved the product s/he selected to critique (i.e., *critiqued product*) and actual *critiquing criteria* (i.e., "keep", "improve", or "compromise") for the product's attributes from her/his clicking actions.

We recruited 18 participants (2 females) to join the study, who were interested in buying a laptop at the time of experiment. According to [12], this scale is acceptable to conduct eye tracking experiment. They are from nine different countries (e.g., China, Switzerland, Italy, Spain, India, USA, etc.), and most of them were students pursuing Master or PhD degree at the university.

4.3 Data Analysis

It shows that every participant posted at least one critique before s/he made the final choice. The total number of critiques by all users is 38 (mean = 2.11 per user, st.d. = 1.45, min = 1, max = 6), among which the number of *improvement-based critiques* is largely higher than that of *similarity-based critiques* (36 vs. 2). Within those improvement-based critiques, 88.9 % (32 out of 36) are *compound critiques*, with average 2.69 attributes selected for "improving" and 1.94 for "compromising" in one critique. Through computing conditional probability[1], we find $P(\text{“}improve\text{”} \mid \text{“}compromise\text{”}) = 1$, whereas

[1] $P(h|e) = N(h \wedge e)/N(e)$, where $N()$ denotes the number of observations within all compound critiques.

$P(\text{“}compromise\text{”}|\text{“}improve\text{”}) = 0.72$, which implies that the appearance of “compromise” is always contingent on that of “improve”, but not vice versa. All of the results hence indicate that users are inclined to *improve* certain attribute values of a product, which will (but not always) be at the cost of *compromising* some of other attributes' values.

As for users' eye movements on recommended products, Fig. 2 shows an example of a user's eye-gaze plot, where each fixation is represented by a blue circle with radius indicating its duration. Wish such eye-gaze plot, we are able to correspond each fixation point to the actual information shown on the interface. Specifically, two researchers first did the mapping independently. If it fell into a product-level AOI, they associated it with that product's ID; if it was placed on an attribute's value (attribute-level AOI), they associated it with both product ID and that attribute's name (e.g., price). They then met together to resolve any divergences. In this way, we identified 2,493 fixation points at product level, and 1,227 fixations associated with ten primary attributes.

Additional analysis shows that on average 9.87 products (st.d. = 5.73) were viewed per user within each set of 25 recommended products. The mean values of fixation count (FC), total fixation duration (TFD), and average fixation duration (AFD) on viewed products are respectively 6.57 (st.d. = 5.59), 2,308.87 ms (st.d. = 2,011.55), and 345.43 ms (st.d. = 50.95). As for attributes, there are 7.13 distinct attributes (st.d.= 2.64) viewed per user within each recommendation set, with mean FC per viewed attribute 3.83 (st.d. = 3.15), mean TFD 1,360.6 ms (st.d. = 1,199.9), and mean AFD 338.4 ms (st.d. = 54.1). Note that in this analysis the fixations on all values of an attribute in each recommendation set are counted together.

5 Inferring Critiquing Feedback

5.1 Inferring Critiqued Products

We are interested in first verifying Hypothesis 1 about the relationship between fixation values and critiqued products, for which *Hit-Ratio@K* (shortened as *H@K*) and *Mean Reciprocal Rank (MRR)*[2] are computed: (1) *Hit-Ratio@K* measures whether a user's critiqued product appears in the top-K products that s/he has viewed, as ranked in the descending order of FC-p, TFD-p, or AFD-p values[3], and (2) *MRR* denotes the critiqued product's ranking position in this order.

From Table 1, we can see that Rank-by-FC-p and Rank-by-TFD-p are of higher accuracy than Rank-by-AFD-p and RAM (RAM refers to random ranking of viewed products), in terms of inferring critiqued products. For instance, when $K = 1$, the *Hit-Ratios* by Rank-by-FC-p and Rank-by-TFD-p are around 0.5,

[2] $H@K = \sum_{c \in C} \frac{1_{rank(p_c) \leq K}}{|C|}$ and $MRR = \sum_{c \in C} \frac{1}{rank(p_c)}$, where $rank(p_c)$ denotes the rank of critiqued product p_c (in cycle c) within the top-K viewed products as sorted by a certain fixation metric.

[3] We use FC-p, TFD-p, and AFD-p to respectively denote the measures of fixation count, total fixation duration, and average fixation duration at product level.

Table 1. Accuracy of inferring critiqued products based on different fixation metrics

	H@1	*H@2*	*H@3*	*H@4*	*H@5*	*MRR*
Rank by FC-p	0.474	**0.605**	**0.789**	0.842	**0.868**	0.628
Rank by TFD-p	**0.5**	**0.605**	0.711	**0.868**	**0.868**	**0.635**
Rank by AFD-p	0.184	0.368	0.447	0.526	0.605	0.378
RAM	0.316	0.342	0.342	0.553	0.5	0.36

Table 2. Relation between attribute-level fixations and critiquing criteria (*note*: C for "Compromise" and K for "Keep". The superscript indicates significant difference.)

	FC-a	TFD-a (ms)	AFD-a (ms)
"Keep" attr.	3.165^{C}	$1,088.92^{C}$	$289.23^{C,K}$
"Improve" attr.	2.64^{C}	$1,038.19^{C}$	340.35^{C}
"Compromise" attr.	1.42	448.42	143.96
ANOVA test	$F = 3.42, \mathbf{p = 0.036}$	$F = 4.045, \mathbf{p = 0.02}$	$F = 21.34, \mathbf{p < 0.001}$

showing that within about half of critiquing cycles, the product with the highest fixation count or total fixation duration was the one that the user selected to critique. When K is increased to 5, the *Hit-Ratios* of Rank-by-FC-p and Rank-by-TFD-p both reach at 0.868. As for Rank-by-AFD-p, its hit ratio is relatively low (maximum 0.605 at $K = 5$). MRR results again indicate that Rank-by-FC-p and Rank-by-TFD-p are more predictive than Rank-by-AFD-p and RAM (0.635 and 0.628, vs. 0.378 and 0.36). Moreover, as the differences between Rank-by-FC-p and Rank-by-TFD-p are not obvious across all measures, they may be equivalent in terms of inferring users' critiquing propensity towards products.

Therefore, the above analysis shows that users' fixation values are helpful for inferring what products they tend to critique. Moreover, fixation count and total fixation duration are more effective than average fixation duration in achieving this goal. Concretely, it suggests that if a user more frequently views a product (with corresponding higher FC-p) or spends totally higher duration on a product (with higher TFD-p), the chance s/he selects it for critiquing will be higher than that of selecting others.

5.2 Inferring Critiquing Criteria for Attributes

Our second hypothesis is about inferring users' critiquing criteria (i.e., keep, improve, or compromise) for product attributes. Formally, each critiquing feedback can be represented as $(p_i, \{\langle a_j, c_j \rangle\})$, where p_i is the critiqued product, $a_j \in A = \{a_1, ..., a_{10}\}$ (A is the set of attributes), and $c_j \in \{\text{"}keep\text{"}, \text{"}improve\text{"}, \text{"}compromise\text{"}\}$. By comparing the fixation values among the three categories of attributes that were respectively selected to "keep", "improve", and "compromise", we find their differences are significant in terms

Table 3. Accuracy of inferring attributes' critiquing criteria through two alternative inference rules (1. *high* => "improve", *medium* => "keep", *low* => "compromise"; or 2. *high* => "keep", *medium* => "improve", *low* => "compromise")

	Precision		Recall		F1		Hit-Ratio	
Classification by FC-a	0.282_1	0.38_2	0.301_1	0.366_2	0.291_1	0.373_2	0.253_1	0.297_2
Classification by TFD-a	0.291_1	0.392_2	0.324_1	0.393_2	0.306_1	0.392_2	0.255_1	0.303_2
Classification by AFD-a	$\mathbf{0.392_1}$	0.344_2	$\mathbf{0.417_1}$	0.416_2	$\mathbf{0.404_1}$	0.376_2	$\mathbf{0.355_1}$	0.289_2

of FC-a, TFD-a, and AFD-a[4] by means of *ANOVA* test (see Table 2). Pairwise comparisons via *paired samples T-test* reveal that the fixation values of "keep" and "improve" attributes are significantly higher than those of "compromise" attributes. Specifically, the mean fixation count (FC-a) of "keep" attributes is 3.165 and that of "improve" attributes is 2.64, against 1.42 of "compromise" attributes ("keep" vs. "compromise": $t = 2.36$, $p = 0.02$; "improve" vs. "compromise": $t = 3.01$, $p < 0.01$). Similar trends are observed regarding total fixation duration (TFD-a) and average fixation duration (AFD-a). As for the difference between "keep" and "improve" attributes, it is moderately significant w.r.t. AFD-a (289.23 ms vs. 340.35 ms, $t = 1.75$, $p = 0.088$).

From the above results, we can derive two alternative inference rules: (1) *high* => "improve", *medium* => "keep", *low* => "compromise"; or (2) *high* => "keep", *medium* => "improve", *low* => "compromise". That is, suppose the fixation values on all attributes in one recommendation set are classified into three levels: *high*, *medium*, and *low*, so we may map each level to a specific critique criterion. For example, if the fixation count of an attribute is at relatively *high* level, we may infer the user would tend to "improve" it in her/his critique (*high* => "improve"). For this purpose, we applied 3-means clustering algorithm to automatically group 10 attributes into three clusters according to their fixation values in each recommendation cycle, and then classified the three clusters into *high*, *medium*, and *low* levels based on their centroids.

Next, we use *Precision*, *Recall*, *F1-measure*, and *Hit-Ratio*[5] to measure each inference rule's accuracy (see the results in Table 3). It shows as for FC-a and TFD-a, the second rule is more accurate in terms of all measures, and the classification by TFD-a achieves slightly higher accuracy than that by FC-a. In comparison, the accuracy of classification by AFD-a via the first rule is even higher, with the highest *Precision* 0.392, *Recall* 0.417, *F1* 0.404, and *Hit-Ratio*

[4] FC-a, TFD-a, and AFD-a respectively denote the measures of fixation count, total fixation duration, and average fixation duration at attribute level.

[5] $Precision = \sum_{k \in AC} \frac{|Pred(k) \cap R(k)|}{|Pred(k)|} / |AC|$, $Recall = \sum_{k \in AC} \frac{|Pred(k) \cap R(k)|}{|R(k)|} / |AC|$, $F1 = \sum_{k \in AC} \frac{2 \times Precision(k) \times Recall(k)}{Precision(k) + Recall(k)} / |AC|$, and $HitRaito = \frac{\sum_{k \in AC} |Pred(k) \cap R(k)|}{q}$, where AC denotes the set of three critique options {"keep", "improve", "compromise"}, $Pred(k)$ denotes the set of attributes that are inferred with critique k, $R(k)$ contains attributes that are actually critiqued with k, and q is the total number of attribute critiques (that is 380 in our data).

0.355 among all results. It hence suggests that, for inferring attributes' critiquing criteria, average fixation duration (AFD-a) behaves more effectively than the other two metrics fixation count (FC-a) and total fixation duration (TFD-a); and the attributes with relatively *high* level of AFD-a will be more likely to be "improved", followed by those at *medium* level to be "kept", and then the remainder at *low* level to be "compromised" (as per the first inference rule). Our Hypothesis 2 is thus verified.

5.3 Other Results: Refining Inference Rules

The derivation of inference rules in the previous section motivates us to consider more information to refine them. By matching each attribute fixation to its actual value (e.g., price $759.99), we can actually identify all values of an attribute the user had fixated (compared) across different recommended products before s/he made a critique. Therefore, in order to generate more precise inference rules, in this section, we investigate fixations on particular attribute values.

Specifically, for each attribute of the critiqued product, we can associate it with a *value comparison label* by comparing its value with the other values the user fixated. If it is a numerical attribute (e.g., price, processor speed, battery life), there are three possible labels: *Better than All* (in the case that the critiqued product's attribute value is better than or equal to the other viewed values in the same recommendation set), *Better than Some* (it is better than some of the other viewed values), and *Worse than All* (it is worse than the other viewed values). If it is a categorical attribute (e.g., manufacturer, processor class, operating system), there are two optional labels: *Equal to Others* (the critiqued product's attribute value is the same as the other viewed ones) and *Different from Others* (it is different from some of the other viewed values). If the user did not leave fixation on any of the attribute's values, it is labeled with *None*.

Figure 3 shows the distribution of all value comparison labels with respect to the three critiques "keep", "improve", and "compromise" that users posted to the corresponding attributes (the distribution is significant, i.e., $p < 0.01$, via *Pearson's Chi-square test*, relative to equal probabilities). Several phenomena can be observed from this figure: (1) *Better than All* and *Better than Some* attribute values appear more often in "keep" critiques. (2) Some *Better than All* attribute values are "compromised", implying that they may be less important to some users. (3) *Worse than All* values are more subject to be "improved" or "kept". These three observations imply that for a numerical attribute, if the user has viewed different values and then selected a product (to critique) that has better value than others, s/he may tend to "keep" it; otherwise, if its value is the worst, s/he may "improve" it. (4) *Different from Others* attribute values are more often "kept" or "improved", whereas *Equal to Others* attribute values are mostly "kept". This observation implies that if all fixations by a user regarding a categorical attribute (e.g., manufacturer) are laid on only one value (e.g., "Apple"), this value might be the user's target, so s/he will be likely to "keep" it. Otherwise, if s/he has fixated over different values of the attribute, s/he may either "keep" the value of her/his selected product (if it meets with

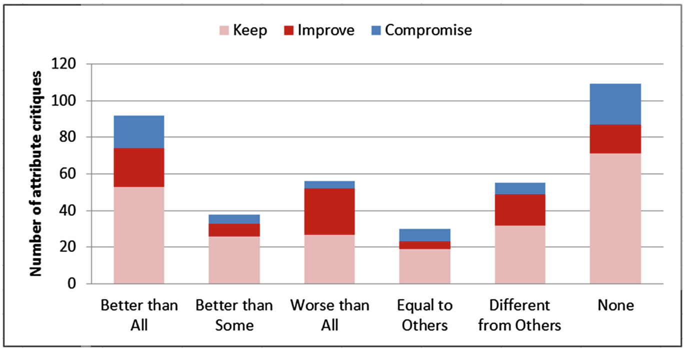


Fig. 3. Distribution of value comparison occurrences with respect to attribute critiques.

her/his requirement after comparison) or "improve" it (if none of viewed values are satisfactory). (5) The attributes without any fixations (labelled as *None*) are mostly "kept" or "compromised".

Then, we combine attribute fixations and value comparison labels to generate association rules. Concretely, we ran Apriori algorithm [1], which is a popular association rule mining tool, on the whole set of 380 attribute critiques, in order to derive high-confident rules in form of $\{attribute\ fixation$, *value comparison*$\}$ => $\{$*attribute critique*$\}$. As for *attribute fixation*, there are three levels, *high*, *medium*, and *low*, as obtained via AFD-a based classification (see the previous section). As for *value comparison*, there are in total six different categories (as described above). The *attribute critique* takes any of the three options, "keep", "improve", and "compromise".

Among the returned rules, we first select those with *Lift*[6] greater than 1, because $Lift > 1$ (also called *Interest*) suggests that the occurrences of antecedent and consequence are dependent on each other, making the rule useful for predicting consequence in other data sets [24]. The selected rules are then sorted in descending order by *Confidence* value. As a result, there are six rules with *Confidence* bigger than or equal to 0.5[7]:

1. $\{$*high AFD-a, Different from Others*$\}$ => "keep" ($Conf. = 0.857$);
2. $\{$*medium AFD-a, Better than Some*$\}$ => "keep" ($Conf. = 0.826$);
3. $\{$*medium AFD-a, Better than All*$\}$ => "keep" ($Conf. = 0.647$);
4. $\{$*Equal to Others*$\}$ => "keep" ($Conf. = 0.633$);
5. $\{$*high AFD-a, Worse than All*$\}$ => "improve" ($Conf. = 0.625$);
6. $\{$*low AFD-a, Better than Some*$\}$ => "compromise" ($Conf. = 0.5$).

The 1st rule implies if a user's average fixation duration (AFD-a) on one categorical attribute is relatively *high* and this attribute's value in the critiqued

[6] $Lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$, $Confidence(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$, where $supp(X)$ gives the proportion of transactions that contain X.

[7] We set 0.5 as *Confidence* threshold, as it indicates a high probability that at least half of transactions contain the antecedent leading to the consequence.

product is different from the other fixated values of the same attribute, the chance that the user will "keep" it is high (with over 85 % confidence). The 2nd and 3rd rules suggest if AFD-a on a numerical attribute is at *medium* level (relative to AFD-a of the other attributes in the same recommendation set) and its value in the critiqued product is better than at least some of the other viewed values, the user may also "keep" it (above 64 % confidence). The 4th rule is about categorical attribute, which, if its critiqued value is equal to the other viewed values, is likely to be "kept" (with 63.3 % confidence). The 5th rule indicates if AFD-a on a numerical attribute is relatively *high* and its critiqued value is the worst among all viewed values of the same attribute, there is around 62.5 % confidence that the user will "improve" it; whereas for an attribute with *low* AFD-a, though its critiqued value is better than some compared ones, the probability that the user will "compromise" it is higher than that of "keeping" or "improving" it (the 6th rule).

6 Conclusions and Future Work

In conclusion, this work verifies our hypotheses about inferring users' critiquing feedback from their eye movements on recommended products. There are three major findings: (1) Based on products' fixation values, we can infer what product the user is inclined to critique within a set of recommendations. In particular, fixation count (FC-p) and total fixation duration (TFD-p) are more accurate than average fixation duration (AFD-p) for achieving this goal. (2) At attribute level, we find the fixation values of attributes that users choose to "keep" or "improve" are significantly higher than those of attributes they "compromise". On the other hand, average fixation duration (AFD-a) performs more effectively than FC-a and TFD-a in terms of inferring users' critiquing criteria for attributes. (3) We further attempted to derive some precise inference rules by incorporating users' value comparison behavior based on their fixations on attribute values. As a result, several high-confident association rules are generated. The findings are thus constructive for improving existing system-suggested critiquing methods in recommender systems. In addition to making the critique suggestions representative of remaining products [19,22], we can make them more reflective of users' critiquing intentions so that the users will be more likely to accept them.

In the future, we will conduct more experiments to validate the association rules' inference accuracy. We will also investigate more fixation metrics, such as fixation spatial density, saccade/fixation ratio, and scanpath, in order to make the inference process more accurate. It is expected that we will eventually build a prediction model that can well unify all of the valuable eye-based metrics to infer users' critiquing feedback, which will enable the system to automatically adjust recommendations even without requiring users to explicitly make critiques.

Acknowledgments. We thank participants who took part in our experiment. We thank Dr. Weike Pan and Ms. Wai Yee Wong for assisting in data processing and analysis. We also thank Hong Kong RGC and China NSFC for sponsoring the described research work (under projects RGC/HKBU12200415 and NSFC/61272365).

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD 1993, pp. 207–216. ACM, New York (1993)
2. Bridge, D., Göker, M.H., McGinty, L., Smyth, B.: Case-based recommender systems. Knowl. Eng. Rev. **20**(3), 315–320 (2005)
3. Burke, R.D., Hammond, K.J., Young, B.: The findme approach to assisted browsing. IEEE Expert Intell. Syst. Appl. **12**(4), 32–40 (1997)
4. Castagnos, S., Jones, N., Pu, P.: Eye-tracking product recommenders' usage. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 29–36. ACM, New York (2010)
5. Chen, L., Pu, P.: Evaluating critiquing-based recommender agents. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 1, pp. 157–162. AAAI Press (2006)
6. Chen, L., Pu, P.: Hybrid critiquing-based recommender systems. In: Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, pp. 22–31. ACM (2007)
7. Chen, L., Pu, P.: Interaction design guidelines on critiquing-based recommender systems. User Model. User-Adap. Inter. **19**(3), 167–206 (2009)
8. Chen, L., Pu, P.: Experiments on the preference-based organization interface in recommender systems. ACM Trans. Comput. Hum. Inter. **17**(1), 1–33 (2010)
9. Chen, L., Pu, P.: Eye-tracking study of user behavior in recommender interfaces. In: Bra, P., Kobsa, A., Chin, D. (eds.) UMAP 2010. LNCS, vol. 6075, pp. 375–380. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13470-8_35
10. Chen, L., Pu, P.: Critiquing-based recommenders: survey and emerging trends. User Model. User-Adap. Inter. **22**(1–2), 125–150 (2012)
11. Cheng, S., Liu, X., Yan, P., Zhou, J., Sun, S.: Adaptive user interface of product recommendation based on eye-tracking. In: Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction, EGIHMI 2010, pp. 94–101. ACM, New York (2010)
12. Ehmke, C., Wilson, S.: Identifying web usability problems from eye-tracking data. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI..But Not As We Know It, BCS-HCI 2007, vol. 1, pp. 119–128. British Computer Society, Swinton (2007)
13. Giordano, D., Kavasidis, I., Pino, C., Spampinato, C.: Content based recommender system by using eye gaze data. In: Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA 2012, pp. 369–372. ACM, New York (2012)
14. Grasch, P., Felfernig, A., Reinfrank, F.: Recomment: towards critiquing-based recommendation with speech interaction. In: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys 2013, pp. 157–164. ACM, New York (2013)
15. Jung, J., Matsuba, Y., Mallipeddi, R., Funaya, H., Ikeda, K., Lee, M.: Evolutionary programming based recommendation system for online shopping. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, pp. 1–4, October 2013
16. Linden, G., Hanks, S., Lesh, N.: Interactive assessment of user preference models: the automated travel assistant. In: Jameson, A., Paris, C., Tasso, C. (eds.) UM 1997. ICMS, vol. 383, pp. 67–78. Springer, Heidelberg (1997). doi:10.1007/978-3-7091-2670-7_9

17. Mahmood, T., Mujtaba, G., Venturini, A.: Dynamic personalization in conversational recommender systems. IseB **12**(2), 213–238 (2014)
18. McCarthy, K., McGinty, L., Smyth, B., Reilly, J.: A live-user evaluation of incremental dynamic critiquing. In: Muñoz-Ávila, H., Ricci, F. (eds.) ICCBR 2005. LNCS (LNAI), vol. 3620, pp. 339–352. Springer, Heidelberg (2005). doi:10.1007/11536406_27
19. McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Experiments in dynamic critiquing. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI 2005, pp. 175–182. ACM (2005)
20. McCarthy, K., Salem, Y., Smyth, B.: Experience-based critiquing: reusing critiquing experiences to improve conversational recommendation. In: Bichindaritz, I., Montani, S. (eds.) ICCBR 2010. LNCS (LNAI), vol. 6176, pp. 480–494. Springer, Heidelberg (2010). doi:10.1007/978-3-642-14274-1_35
21. Poole, A., Ball, L.J.: Eye tracking in human-computer interaction and usability research: current status and future prospects. In: Ghaoui, C. (ed.) Encyclopedia of Human-Computer Interaction. Idea Group Inc., Pennsylvania (2005)
22. Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Dynamic critiquing. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 763–777. Springer, Heidelberg (2004). doi:10.1007/978-3-540-28631-8_55
23. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, ETRA 2000, pp. 71–78. ACM, New York (2000)
24. Tan, P.-N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 32–41. ACM (2002)
25. Tullis, T., Albert, W.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann Publishers Inc., San Francisco (2008)
26. Xie, H., Chen, L., Wang, F.: Collaborative compound critiquing. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 254–265. Springer, Heidelberg (2014). doi:10.1007/978-3-319-08786-3_22
27. Xu, S., Jiang, H., Lau, F.C.: Personalized online document, image and video recommendation via commodity eye-tracking. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, pp. 83–90. ACM, New York (2008)
28. Zhang, J., Pu, P.: A comparative study of compound critique generation in conversational recommender systems. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) AH 2006. LNCS, vol. 4018, pp. 234–243. Springer, Heidelberg (2006). doi:10.1007/11768012_25