An Eye-Tracking Study: Implication to Implicit Critiquing Feedback Elicitation in Recommender Systems

Li Chen and Feng Wang
Department of Computer Science
Hong Kong Baptist University, China
{lichen,fwang}@comp.hkbu.edu.hk

ABSTRACT

The *critiquing-based recommender system* (CBRS) stimulates users to critique the recommended item in terms of its attribute values. It has been shown that such *critiquing feedback* can effectively improve users' decision quality, especially in complex decision environments such as e-commerce, tourism, and finance. However, because its explicit elicitation process unavoidably demands extra user efforts, the application in real situations is limited. In this paper, we report an eye-tracking experiment with the objective of studying the relationship between users' eye gazes as laid on recommended items and their critiquing feedback. The results indicate the feasibility of inferring users' feedback based on their eye movements. It hence points out a promising roadmap to developing unobtrusive eye-based feedback elicitation for recommender systems.

CCS Concepts

•Human-centered computing → Empirical studies in interaction design; *User models; User studies;* **•Information systems →** *Recommender systems;*

Keywords

Recommender systems; feedback elicitation; critiquing; eye tracking; user study

1. INTRODUCTION

During the past decade, recommender systems have popularly been applied in various online scenarios to aid users in confronting overwhelming information and making effective decisions. It has been shown that the existing techniques such as collaborative filtering and content-based approaches are capable of estimating users' preferences based on their historical data like ratings [20]. However, in practical situations, especially in complex decision environments (e.g., e-commerce, tourism, and finance domains), where users have left few transaction records, it is difficult to adopt these techniques to predict user preferences and generate recommendation. In order to resolve this problem, the *critiquing-based recommenders system (CBRS)* has emerged, which distinguishes itself in feedback elicitation [1, 16, 4, 7]. Concretely, it involves users in a conversational dialog with the system so as to elicit their feedback and construct their preference model on site. Figure 1 shows the workflow of a representative CBRS called Example Critiquing [4]. It first presents some example products to a user according to her/his initially specified preferences. It then stimulates the user to select a near-satisfactory product and critique it in terms of its attribute values (such as "*I would like to see some laptops with different manufactures and higher processor speed*"). The system will refine its understanding of the user's preferences based on her/his critique and generate a new set of recommendations in the next interaction cycle. For a user to reach his/her target choice, a number of critiquing cycles are usually required. Prior work states that a typical user has many constraints and preferences, but s/he can only become aware of these latent preferences when some solutions are proposed [3]. Obtaining their critiques to recommendation has hence been regarded as an effective mechanism to disclose their latent preferences and help them to improve decision quality [7].

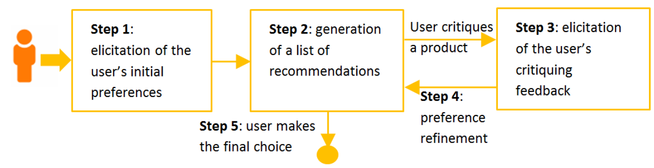

Figure 1: Workflow of a representative critiquing-based recommender system called Example Critiquing [4].

However, the applicability of existing CBRSs is limited as they mostly require users to *explicitly* specify their critiques in each recommendation cycle. As shown in previous studies, some users are subject to avoid making critiques due to the extra efforts it causes [5, 12]. The challenging question that CBRS faces is: *Is it possible to elicit users' critiquing feedback through implicit and unobtrusive way?*

In this paper, we are interested in investigating the relationship between users' eye-gaze behavior when they view recommendations and their critiquing feedback. The results could thus be suggestive for developing eye-based feedback elicitation in CBRS. Indeed, the eye and its movements, being "a window to the mind", are tightly coupled with human

UMAP '16, July 13-17, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4370-1/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2930238.2930286

cognitive processes [15]. Given that advanced eye-tracking instrument makes it feasible to identify how a user's attention is directed in relation to an interface, we may treat the eye measures as implicit feedback to sense the user's interest and intention. With this objective, we have performed an eye-tracking experiment that in depth examines users' eye-gaze behavior at both *product level* and *attribute level*.

2. RELATED WORK

The development of eye tracking technology has enabled academic and commercial sectors to apply it in various interaction designs [14, 10, 17]. In recommender systems, it has mainly be adopted for two purposes. One is to evaluate the usability of a recommendation interface. For instance, one user experiment measured the effect of interface layout on users' visual searching pattern [6]. It shows that users tend to fixate more on the top area if recommended items are displayed in a list layout, but will be directed to view more items if they are arranged in a category structure. Another experiment investigated whether users would gaze at recommendation during their entire product searching process [2]. Its results clarify the important role of recommendation in users' purchase decision.

As the second purpose, some researchers have exploited eye-gaze metrics to elicit users' implicit *relevance feedback* on recommendation, i.e., "positive" or "negative" (or called "like" or "dislike"). For instance, in [11], the documents that users consume higher number of fixations and longer average fixation time are regarded with "positive" feedback. They then use clustering and content based techniques to retrieve similar documents and recommend them to the user. Some studies aim at developing algorithms, e.g., interactive genetic algorithm [8], evolutionary programming [13], and attention prediction method [23], to incorporate the eye-based *relevant feedback* into the process of inferring users' preferences for images, documents, videos, or e-commerce products.

However, little work has exerted to elicit users' specific feedback to product attributes through eye tracking. The eye-based application in CBRS is even rare. As mentioned before, CBRS aims to obtain users' *critiquing feedback*, which contains not only the user's preference for a product to critique, but also her/his multi-type critiquing criteria for the product's attributes. Its elicitation procedure is hence more challenging than that for relevance feedback.

3. EXPERIMENT SETUP

3.1 Materials and Participants

We choose Example Critiquing as the experiment system to obtain users' explicit critiques. Its laptop catalog was extracted from a commercial e-commerce website. During each recommendation cycle, 25 laptops that best match the user's current preferences are returned. Each laptop is described by three blocks of information in the recommendation interface: title (e.g., "Apple 15 MacBook Pro Notebook"), image, and ten major attributes' values (i.e., manufacturer, price, operating system, battery life, display size, hard drive capacity, installed memory, processor class, processor speed, and weight). Within the set of recommendations, if the user cannot locate her/his target choice, s/he can select one product that is near-satisfactory and provide critiquing feedback on it. Specifically, for each attribute of the selected product, the user can make one of the following three critiques: "**Keep**" - keeping the attribute's existing value (default choice); "**Improve**" - improving the attribute's value, e.g., "cheaper", "bigger size"; "**Compromise**" - accepting a compromised value. Essentially, the critique that involves both "improve" and "compromise" is a kind of tradeoff decision, i.e., accepting an outcome that is undesirable in some respects but advantageous in others [19].

The experiment is in form of a controlled lab study. A Tobii 1750 eye-tracker that is integrated with a 17" TFT screen is used to record each subject's eye movements when s/he views recommended products. Its resolution setting is 1290x1024 pixels, and can sample the position of a user's eyes by every 20ms. The monitor frame has near infra-red light-emitting diodes, which allow for natural eye tracking without placing many restrictions on the user.

We recruited 18 participants (2 females) to join the study (according to [9], this scale is acceptable for an eye tracking study), who were interested in buying a laptop at the time of 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 in the university.

3.2 Experiment Procedure and Measurement

The user task was to "*find a product you would purchase if given the opportunity by using the Example Critiquing 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 Example Critiquing system's interfaces during a warm-up period. Afterwards, the eye-tracker calibration was performed, and the participant started to use the system to accomplish the given task. In the mean time, her/his eye-gaze behavior and mouse clicking actions were automatically recorded by the eye-tracker.

The process of deriving useful information from eye-gaze recordings is usually to analyze users' fixations. Each fixation is a spatially stable gaze point, during which most information acquisition and processing occur. We set its minimum duration as 200ms according to [21]. We performed two levels of fixation analysis: *product level* and *attribute level*. At product level, any fixations that fall inside the boundary of a product that contains its title, image and major attributes' values are treated equally. At attribute level, fixations laid on different attributes (e.g, price and operating system) are analyzed individually.

Regarding fixation metrics, we adopted three commonly used measures [18, 9, 22]: **Fixation Count (FC)** - the number of times the user fixates on a product or an attribute; **Total Fixation Duration (TFD)** - the sum of the duration of all fixations the user has laid on a product or an attribute; **Average Fixation Duration (AFD)** - the average duration of a fixation on a product or an attribute. These three metrics generally represent users' relative engagement with the interface object [18, 22]. More fixations on an object suggest that it is more noticeable and important. A longer duration may indicate that the fixated object is more engaging in some way.

From users' clicking actions, we retrieved their actual **critiquing feedback** in each recommendation cycle, which includes the *critiqued product* (i.e., the product selected for critiquing) and the user's *critiquing criteria* (i.e., "keep", "improve", or "compromise") for the product's attributes.

4. RESULTS ANALYSIS

4.1 Critiquing Application

The results show that each user provided at least one critiquing feedback before s/he made the final choice. The total number of critiques made by all 18 users is 38 (mean = 2.11, st.d. = 1.45, min = 1, max = 6). Moreover, the number of *improvement-based critiques* (that "improve" some attribute values) is largely higher than that of *similarity-based critiques* (that "keep" all attribute values of the critiqued product) (36 vs. 2). Among those improvement-based critiques, 88.9% (32 out of 36) involve multiple attributes to "improve" (average 2.69 attributes) and/or "compromise" (average 1.94 attributes) (that are called *compound critiques* in [16]). Through computing conditional probability (Equation (1)), we find $P(\text{"}improve\text{"}|\text{"}compromise\text{"}) = 1$, whereas $P(\text{"}compromise\text{"}|\text{"}improve\text{"}) = 0.72$, which indicates that the appearance of "compromise" in a compound critique is always contingent on that of "improve", but not vice versa. It hence suggests that users are inclined to *improve* certain attribute values of a product in their critiques, which will (but not always) be at the cost of *compromising* some of other attributes' values for the purpose of tradeoff.

$$P(h|e) = \frac{N(h \wedge e)}{N(e)} \tag{1}$$

where $N()$ denotes the number of observations within all compound critiques.

4.2 Product-Level Fixation Analysis

Figure 2 shows the example of a user's gaze plot on recommended products, where each fixation is illustrated with a blue circle and its radius represents the duration of the fixation. Because the eye-tracker we used cannot automatically map a fixation onto the specific product or attribute that is displayed on the recommendation interface, we did the mapping manually. Concretely, two researchers first independently examined each fixation point for corresponding it to the actual information shown on the interface. If it fell into a product-level area, they associated it with that product's ID; if it was placed on an attribute's value, 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 (see next section for the attribute-level fixation analysis results).

More specifically, within the set of 25 products recommended to the user in each cycle, we find on average 9.87 products (st.d. = 5.73) were viewed. 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. It shows for every viewed product the mean values of FC-p, TFD-p, and AFD-p are respectively 6.57 (st.d. = 5.59), 2,308.87msec (st.d. = 2,011.55), and 345.43msec (st.d. = 50.95).

We then compute *Hit-Ratio@N* (shorted as *H@N*) (Equation (2)) and *Mean Reciprocal Rank (MRR)* (Equation (3)): 1) *Hit-Ratio@N* measures whether the user's critiqued

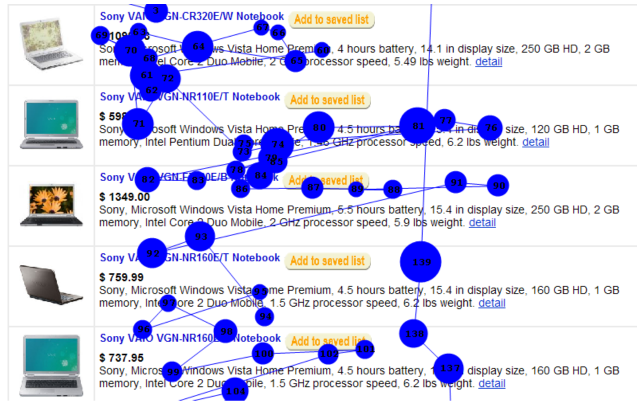

Figure 2: A user's eye-gaze plot on recommended products.

product appears in the top-N viewed products as ranked in descending order of FC-p, TFD-p, or AFD-p values, and 2) *MRR* denotes the critiqued product's position in this ordering.

$$H@N = \frac{\sum_{c \in C} 1_{rank(p_c) \leq N}}{|C|} \tag{2}$$

$$MRR = \frac{\sum_{c \in C} \frac{1}{rank(p_c)}}{|C|} \tag{3}$$

where $|C|$ is the total number of critiquing cycles by all users, and $rank(p_c)$ gives the ranking position of the critiqued product p_c within the top-N viewed products (in cycle c) as ranked by FC-p, TFD-p, or AFD-p.

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 locating the critiqued product. For example, when $N = 1$, the hit ratios of Rank-by-FC-p and Rank-by-TFD-p are around 0.5, showing that within about half of all critiquing cycles, the product with the highest fixation count or total fixation duration was the one that the user selected to critique. When N is increased to 5, the hit ratios of Rank-by-FC-p and Rank-by-TFD-p both achieve 0.868. As for Rank-by-AFD-p, its hit ratio is relatively low (maximum 0.605 at $N = 5$). MRR results again imply 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, we can infer they might be equivalent in terms of inferring users' critiquing intention at product level.

The above observations thus imply that if a user takes more times in viewing a product (with corresponding higher FC-p and TFD-p), the chance s/he selects it for critiquing will be higher than that of selecting others. In comparison, the average fixation duration (AFD-p), which mainly reflects a fixation's average dwell time, is less powerful to predict the critiqued product.

4.3 Attribute-Level Fixation Analysis

For the next step of analysis, we look into fixation data at attribute-level for identifying their relationship with users'

Table 1: Relationship between product-level fixations and critiqued products

	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: Relationship between attribute-level fixations and critiquing criteria (*note*: C for "Compromise", and the superscript indicates significant difference, i.e., $p < 0.05$)

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

critiquing criteria for product attributes. There are in total 1,227 fixation points associated with the 10 major attributes (e.g., manufacturer, price, operating system, battery life). On average, the number of distinct attributes viewed by a user within each set of recommendations is 7.13 (st.d. = 2.64), with mean FC per attribute (FC-a) 3.83 (st.d. = 3.15), mean TFD per attribute (TFD-a) 1,360.6msec (st.d. = 1,199.9), and mean AFD per attribute (AFD-a) 338.4msec (st.d. = 54.1).

In addition, it shows the differences among attributes that were respectively critiqued with "keep", "improve", and "compromise" are significant in terms of FC-a, TFD-a, and AFD-a by means of ANOVA test (see Table 2). Pairwise comparisons via paired samples T-test further 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 for total fixation duration (TFD-a) and average fixation duration (AFD-a). As for the difference between "keep" and "improve" attributes, it is just moderately significant regarding AFD-a (289.23msec vs. 340.35msec, $t = 1.75$, $p = 0.088$).

The results hence suggest that if a user's eyes fixate more on one attribute, s/he may tend to "keep" or "improve" it during critiquing, whereas for the attribute with fewer attentions, s/he may "compromise" it.

4.4 Discussion

The practical implication of this study is that: suppose we know a user's eye-gaze behavior on a recommendation interface, we can infer what product s/he is inclined to critique, and furthermore what attributes of the product s/he will be likely to "keep", "improve", or "compromise". The system could then suggest some critiques for the user to choose, instead of requiring the user to specify critique by her/himself. Moreover, the system could automatically refine the user's preference model for augmenting product recommendation simultaneously. For instance, the critiqued product's attribute values will become default acceptable value thresholds, and the weight of "kept" or "improved" attribute will be increased, while the weight of "compromised" attribute will be decreased. We may then adjust the utility computed for each candidate product to enhance products' ranking. By this way, we can not only reduce users' critiquing efforts, but also help them to locate the target choice earlier.

5. CONCLUSIONS AND FUTURE WORK

In conclusion, this work indicates the feasibility of inferring users' critiquing feedback from their eye movements on recommendations. There are two major findings: 1) The fixation count and total fixation duration at product level (i.e., FC-p and TFD-p) are helpful for estimating users' interest in a product for critiquing, since the one with higher FC-p or TFD-p was more frequently selected as critiqued product. 2) The differences among critiqued attributes in terms of their fixation values are significant, especially between "kept"/"improved" and "compromised" attributes. It suggests that attributes with higher FC-a/TFD-a/AFD-a are more likely to be "kept" or "improved", whereas those with lower values will be "compromised".

The findings inspire us to conduct more studies in the future. We will attempt to identify which fixation metric, among FC-a, TFD-a, and AFD-a, would be more precise to infer users' critiquing criteria for attributes. We will also manage to recover users' decision process of comparing different attribute values of recommended products, by investigating their fixations on attributes' actual values and scanpath. Particularly, scanpath analysis can help detect users' *pairwise* value comparison behavior, as each scanpath shows a complete saccade-fixate-saccade sequence [22]. Eventually, we will develop an eye-based feedback elicitation and preference prediction model for critiquing-based recommender systems (CBRS), and perform more user studies to verify its practical performance.

6. ACKNOWLEDGMENTS

We thank participants who took part in the experiment. We also thank Hong Kong RGC and China NSFC for sponsoring the described research work (under projects RGC/HKBU12200415 and NSFC/61272365).

7. REFERENCES

[1] R. D. Burke, K. J. Hammond, and B. Young. The FindMe approach to assisted browsing. *IEEE Expert: Intelligent Systems and Their Applications*, 12(4):32–40, July 1997.

[2] S. Castagnos, N. Jones, and P. Pu. Eye-tracking product recommenders' usage. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 29–36. ACM, 2010.

[3] L. Chen. *User Decision Improvement and Trust Building in Product Recommender Systems*. PhD thesis, Ecole Polytechnique Federale De Lausanne (EPFL), Lausanne, Switzerland, August 2008.

[4] L. Chen and P. Pu. Evaluating critiquing-based recommender agents. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 157–162. AAAI Press, 2006.

[5] L. Chen and P. Pu. Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction*, 19(3):167–206, Aug. 2009.

[6] L. Chen and P. Pu. Eye-tracking study of user behavior in recommender interfaces. In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization*, UMAP '10, pages 375–380. Springer-Verlag, 2010.

[7] L. Chen and P. Pu. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1-2):125–150, 2012.

[8] S. Cheng, X. Liu, P. Yan, J. Zhou, and S. Sun. 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 '10, pages 94–101. ACM, 2010.

[9] C. Ehmke and S. Wilson. 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 - Volume 1*, BCS-HCI '07, pages 119–128. British Computer Society, 2007.

[10] C. Eickhoff, S. Dungs, and V. Tran. An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 13–22. ACM, 2015.

[11] D. Giordano, I. Kavasidis, C. Pino, and C. Spampinato. Content based recommender system by using eye gaze data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 369–372. ACM, 2012.

[12] G. Haübl and V. Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19(1):4–21, Jan. 2000.

[13] J. Jung, Y. Matsuba, R. Mallipeddi, H. Funaya, K. Ikeda, and M. Lee. Evolutionary programming based recommendation system for online shopping. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–4, Oct 2013.

[14] S. Kardan and C. Conati. Comparing and combining eye gaze and interface actions for determining user learning with an interactive simulation. In *Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization*, UMAP '13, pages 215–227, 2010.

[15] S. Liversedge, I. Gilchrist, and S. Everling. *The Oxford Handbook of Eye Movements*. Oxford University Press, 2011.

[16] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth. Experiments in dynamic critiquing. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, pages 175–182. ACM, 2005.

[17] Y. I. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 139–148. ACM, 2010.

[18] A. Poole and L. J. Ball. Eye tracking in human-computer interaction and usability research: Current status and future prospects. In *Prospectsą́, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction*. Pennsylvania: Idea Group, Inc, 2005.

[19] P. Pu and L. Chen. Integrating tradeoff support in product search tools for e-commerce sites. In *Proceedings of the 6th ACM Conference on Electronic Commerce*, EC '05, pages 269–278. ACM, 2005.

[20] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[21] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78. ACM, 2000.

[22] T. Tullis and W. Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[23] S. Xu, H. Jiang, and F. C. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 83–90. ACM, 2008.