On Evaluating Session-Based Recommendation with Implicit Feedback

Fernando Diaz[1]

[1]Google, Montréal, Canada

Abstract

Session-based recommendation systems are used in environments where system recommendation actions are interleaved with user choice reactions. Domains include radio-style song recommendation, session-aware related-items in a shopping context, and next video recommendation. In many situations, interactions logged from a production policy can be used to train and evaluate such session-based recommendation systems. This paper presents several concerns with interpreting logged interactions as reflecting user preferences and provides possible mitigation to those concerns.

Keywords

session-based recommendation systems, evaluation

1. Introduction

Many production recommendation systems are designed using the abstraction of the session-based recommendation system (SBRS) [1]. In SBRS, system recommendation actions (e.g. rankings, slates, individual items) are interleaved with user choice responses (e.g. clicks, streams). This approach aligns well with production sequential interaction and data logging. Moreover, treating recommendation as a sequence of user interaction allows system designers to adopt sequential decision-making algorithms such as reinforcement learning.

In order to evaluate a SBRS, experimentation protocols (or teams, in the context of a production system) need to model ideal system behavior. Current practice suggests interpreting positive user responses to system recommendations (e.g. clicks, streams) as indications of positive reward (or, in some cases, negative reward). The major assumption underlying this approach is that user preferences and choices revealed *in situ* and are accurate reflections of item value. A new policy that accurately suggests items with a logged positive user response–with an off-policy correction–is preferable to one that does not.

This position paper explores the various ways in which implicit feedback present in interaction logs can deviate from unobserved ideal system labels. Although prior work has explored biases that may emerge in traditional ratings matrices [2, 3] and on-policy evaluation [4], we are interested in problems that result from the sequential nature of session data and associated

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021), September 25th, 2021, co-located with the 15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands

diazf@acm.org (F. Diaz)
https://841.io/ (F. Diaz)
0000-0003-2345-1288 (F. Diaz)

CEUR Workshop Proceedings (CEUR-WS.org)

interface constraints, user cognitive biases, and uncontrolled sequential dependencies. As a result, we believe that evaluation data gathered under current practices may be distorted and susceptible to misidentification of quality recommendation systems, *even when existing debiasing methods are employed.*

We propose addressing these concerns in several ways. First, we suggest that current reward-definition practices should be revisited and user behavior studied further. As part of this, we believe that there is an opportunity to adjust logging and interaction modeling practices to control for problematic behavior. Second, we suggest that system designers develop mechanisms and interface tools to gather evaluation data not prone to the biases discussed in Section 3.

2. Session-based Recommender Systems

Let $\mathcal{U}$ be the set of users and $\mathcal{A}$ the set of items. We define a *session prefix* $\rho = [a_1, a_2, a_{t-1}]$ as a sequence of items engaged with by a user $u \in \mathcal{U}$. Given a session prefix, each item a has an associated reward r_a^ρ reflecting the quality of the item, if we were to present it to user immediately after ρ. The next-item recommendation task is to, given a session prefix, produce a ranking π of $\mathcal{A}$ such that high quality items are above lesser-quality items. In practice, this ranking is truncated for display or efficiency reasons. Given a ranking π and r^ρ, we can evaluate performance using an information retrieval metric μ, which models user browsing behavior [5].

We are interested in how r is defined. One way to do this is to use an oracle to label the quality of all items given a prefix. This oracle would have access to the entire inventory, the internal state of the user, and sufficient time to assemble the ideal reward. In practice, we rarely have access to such an oracle, so we use logged interactions to infer r. Let τ be an observed length-ℓ session, from which we can extract $\ell - 1$ prefixes for evaluation. For a prefix $\rho = \tau_{[1,t)}$, we can define r_a^ρ as,

$$r_a^\rho = \begin{cases} 1 & a = \tau_t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Since, in our logs, we only observe one selected item for a prefix, this underestimates the reward for similar or substitutable items. To address this, we can introduce an assumption of 'in-session substitutability' which considers all selected items in the session substitutes,

$$r_a^\rho = \begin{cases} 1 & a \in \tau_{[t,\ell]} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

We can imagine more elaborate definitions inspired by reinforcement learning, but all are based on immediate implicit user feedback.

For the purpose of our discussion, we assume that we have access to the oracle r^ρ. This allows us to compute, for each session, the set of optimal sequences. Such an oracle has access to the full catalog of items and understands any potential interactions between items that may affect their utility.

In addition, we will consider the protocol for gathering session data described in Algorithm 1. This covers a large class of production recommendation system workflows. Radio-style

recommendation is a special case where $|\pi| = 1$ and the user has an additional `SKIP` reaction, which does not get appended to τ. Furthermore, we are agnostic about the logging policy and include those that incorporate randomization.

Algorithm 1 Session Data Collection

```
function SessionCollect(u)                          ▷ gather session τ from a specific user
    τ ← []                                          ▷ initialize session sequence
    do
        π ←GetRanking(u, τ)  ▷ get slate from logging policy based on the current prefix
        a ←GetFeedback(u, π)                        ▷ observe item selected by user
        τ ← τa                                      ▷ append selected item to prefix
    while a ≠ EOS                                   ▷ terminate sequence if the user abandons
    return τ                                        ▷ return the session sequence
end function
```

3. Problems with Current SBRS Evaluation

We now turn the potential issues with this method of collecting labeled data to evaluate SBRS. Our claim is not that all of these concerns are present in all SBRSs, although we suspect that many are.

First, consider the impact of the user selecting items under incomplete information. Due to system constraints, the user is only ever presented with a ranking of a subset of the catalog. Moreover, because users scan rankings from top to bottom, with an increasing probability of abandoning the scan (i.e. position bias), a choice will often be made amongst the top-ranked items. As a result, sequential choices are made with severely limited options and information. This is referred to as *choice bracketing* [6] and we depict it in Figure 1a. There are several implications of choice bracketing. First, *limited options* can result in selecting a suboptimal item in the sequence, since a user may not see superior options. Moreover, these unexamined, superior options disappear in future rankings when a recommendation system removes previously-recommended items to reduce the perception of redundancy. Second, choice bracketing potentially narrows and distorts a user's decision context (i.e. inspected relevant and non-relevant items) leading to priming and potentially inaccurate choices [7]. A new system that presents different rankings will bracket choices differently and potentially result in different optimal choices.

Second, we turn to the problem of label sparsity in SBRS. For a given prefix, r_a^ρ will be incomplete, especially if we only consider the next observed item τ_t as relevant (Equation 1). Even if adopt the 'in-session substitutability' assumption (Equation 2), a user will rarely exhaust the set of relevant items in a session. We refer to this as the problem of *incomplete substitutes* and depict it in Figure 1b. Here, because we are only selecting one item from the ranking at a time, potential substitutes are unobserved and considered nonrelevant. Such sparsity issues have resulted in distorted evaluation in both recommendation system [8] and information retrieval contexts [9]. And, although this can be mitigated by algorithms based on equal exposure [10],

from an evaluation perspective, collecting a large set of equally effective trajectories is unlikely for tail interests.

Third, in many sequential recommendation tasks, there are sequential dependencies between items. For example, users may not want to hear two songs from the same musician one after the other, even though they find the two songs enjoyable otherwise. Unfortunately, the 'in-session substitutability' assumption can overestimate the value of a substituted item from $t' > t$ if, for example, the item degrades in value when recommended immediately after τ_{t-1}. Similarly, an *unselected* item at time $t' > t$ may be underestimated in value if it is substitutable with τ_t but was not selected at time t' because of its recommendation immediately after $\tau_{t'-1}$. The contextual utility of items, especially for entertainment goods, could be due to satiation [11] or other order effects [12]. We refer to this as the problem of *inter-item dependency* and depict it in Figure 1c, where the oracle preferences at time t depend on the choice made at time $t-1$.

Fourth, many session-based recommendation systems include a default choice, usually the top-ranked item. In streaming media platforms, this often means automatically playing the default choice after some short period of time for the user to select an alternative. This functionality results in reinforcing the default option (or, more generally, the system ranking), even if the user compares it to alternatives [13]. We call this the problem of *default preferences* and depict it in Figure 1d. As with choice bracketing and incomplete substitutes, this can result in missing labels and, in cases where the default option is not relevant, the implicit feedback is incorrect.

Fifth, consider user interfaces that include a thumbnail summary for each recommended item. The visual attributes of this summary can vary across items and can cause users to inspect and select items that are more visually salient [14, 15]. Salient items can disrupt inspection by rank order, resulting in selection of inferior items. We refer to this as the problem of *presentation bias* and depict it Figure 1e. Given the similarity to choice bracketing, it results in the same problems.

Finally, in some cases, oracle preferences may be inconsistent with the observed preferences because preferences change at the moment of choice. Consider the example of choosing what to eat. Prior work has demonstrated that people often choose healthy options if there is some temporal delay between the choice of what to eat and when they actually eat their choice; those preferences reverse in favor of the less healthy option if the choice is made immediately before consumption [16]. Similarly, experiments have shown that people will select 'highbrow' movies if asked days before watching the film; their preferences will shift to 'lowbrow' movies if asked on the day of the watching the film [17]. In the context of SBRS, this means that observed choices made instantaneously and sequentially may be reversals of 'healthier' preferences expressed with foresight. We call this the problem of *immediacy effects* and depict it in Figure 1f. By construction, we defer to the oracle preferences when they disagree with immediate preferences, which may be susceptible to impulsive behavior.

In addition to these concerns with implicit labels from session data, *how* session data is gathered and segmented into evaluation prefixes can distort performance. Simple issues like over-representing sessions from active users are familiar to recommendation systems researchers. Sessions can introduce additional issues. Prefixes are often selected to include all but the final item in the session. This can hide under-performance at earlier points in the session, which is a problem when user preferences or behavior changes over the length of the session. Evaluation preparation that considers all prefixes in a session for evaluation can, for situations where

session lengths are not fixed, over-emphasize performance at the beginning of the session. Separately, evaluation trajectories themselves are biased by the data-gathering policy and may not be representative of the prefixes encountered when the evaluated SBRS is deployed [18].

The magnitude of these problems depends on the domain. For example, in radio-style music recommendation, users have an aversion to silence [19]; this may result in urgency and, as a result, amplify position bias and narrow choice brackets. In some shopping settings, order effects may be less pronounced. Text-only interfaces will be less susceptible to presentation bias. Furthermore, these problems can interact and compound effects. For example, visually salient items can increase impulsivity and potentially lead to immediacy effects [20].

The implications of label unreliability depends largely on the domain. In the context of search, cognitive biases in labels have been demonstrated to impact performance of learning-to-rank systems [21]. In the context of traditional recommendation system evaluation, failing to consider biased labeling can result in system under-performance in practice [2, 3].

4. Addressing the Problems with Current SBRS Evaluation

Since this is a position paper, we would like to conclude with possible next steps for the community to consider, given the potential issues with current SBRS evaluation. Specifically, a research program on SBRS preference elicitation could be built around the following two themes: (i) recognizing that a domain is susceptible to problems in Section 3, and (ii) mitigation strategies for those problems.

Some of these problems can be recognized by looking at out-of-session information indicating higher-level user satisfaction with the SBRS. One way to do this is to understand the relationship between in-session behavior and longer-term user retention, surveys, and other tools [22, Part 4]. Alternatively, smaller scale, controlled, laboratory experiments and qualitative research can also provide an indication of these problems and is especially effective when combined with larger scale log data [23, 24].

These problems, when detected, can be addressed in a variety of ways. In the context of search, there is a small body of work focused on extracting relevance information in the presence of click feedback under position bias [25]. Randomization and other off-policy evaluation techniques can be used to address some, although not all, of the concerns in SBRS [4]. Explicit models of 'unhealthy' items can also be used to guide recommendations toward healthier options [26].

A different way to approach these problems is to change interface elements to support decision-making. For example, widgets like shortlists can help expand brackets [27]. In the context of exploratory search, assistive tools like note-taking devices can also improve long-term goals like task completion [28, 29].

A final option, in domains where the space of information needs is small, we can consider directly modeling the oracle. In the context of music recommendation, users often spend time manually-curating playlists for future consumption in specific situations [30]. As such, manually-curated playlists provide a rich source of 'gold standard' data for SBRS [31, 32]. Therefore, developing exploratory search and other tools to support curation, in music or other domains, can provide one way to crowd-source oracle data [33]. Similar methods have been used in the context of query autocompletion, which can be considered a character-level recommendation

task [34]. That said, there are some cases where asking a user to provide oracle decisions is unsuccessful because people can fail to consider important contextual information necessary for understanding the appropriate choice [35]. For example, one might select meals for a week but not consider the time pressures or exhaustion that may make the effort to prepare the healthiest meal not worth it. This tension between immediacy effects and inaccurate forecasting, then, comes to the fore.

5. Conclusion

In this position paper, we have argued that several of the current practices for gathering label and reward data from implicit feedback is susceptible to error and may impact evaluation of SBRSs. While several of these problems have been discussed in the context of algorithm design, we believe that moving the investigation to our practice of evaluation will add nuance to our understanding of how users interact with recommendation systems and, as a result, improve the design of these systems.

References

[1] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, D. Lian, A survey on session-based recommender systems, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3465401. doi:10.1145/3465401.

[2] B. M. Marlin, R. S. Zemel, Collaborative prediction and ranking with non-random missing data, in: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 5–12. URL: https://doi.org/10.1145/1639714.1639717. doi:10.1145/1639714.1639717.

[3] D. Liang, L. Charlin, J. McInerney, D. M. Blei, Modeling user exposure in recommendation, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, pp. 951–961. URL: https://doi.org/10.1145/2872427.2883090. doi:10.1145/2872427.2883090.

[4] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, E. H. Chi, Top-k off-policy correction for a reinforce recommender system, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, ACM, New York, NY, USA, 2019, pp. 456–464. URL: http://doi.acm.org/10.1145/3289600.3290999. doi:10.1145/3289600.3290999.

[5] B. Carterette, System effectiveness, user models, and user utility: a conceptual framework for investigation, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, ACM, New York, NY, USA, 2011, pp. 903–912. URL: http://doi.acm.org/10.1145/2009916.2010037. doi:10.1145/2009916.2010037.

[6] D. Read, G. Loewenstein, M. Rabin, Choice bracketing, Journal of Risk and Uncertainty 19 (1999) 171–197. URL: http://www.jstor.org/stable/41760959.

[7] L. Azzopardi, Cognitive biases in search: A review and reflection of cognitive biases in

information retrieval, in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, CHIIR '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 27–37. URL: https://doi.org/10.1145/3406522.3446023. doi:10.1145/3406522.3446023.

[8] P. Kouki, I. Fountalis, N. Vasiloglou, X. Cui, E. Liberty, K. Al Jadda, From the lab to production: A case study of session-based recommendations in the home-improvement domain, in: Fourteenth ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 140–149. URL: https://doi.org/10.1145/3383313.3412235. doi:10.1145/3383313.3412235.

[9] N. Arabzadeh, A. Vtyurina, X. Yan, C. L. A. Clarke, Shallow pooling for sparse labels, 2021.

[10] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, B. Carterette, Evaluating stochastic rankings with expected exposure, arXiv e-prints (2020) arXiv:2004.13157.

[11] J. Galak, J. P. Redden, The properties and antecedents of hedonic decline, Annual Review of Psychology 69 (2018) 1–25. URL: https://doi.org/10.1146/annurev-psych-122216-011542. doi:10.1146/annurev-psych-122216-011542, pMID: 28854001.

[12] M. Eisenberg, C. Barry, Order effects: A study of the possible influence of presentation order on user judgments of document relevance, Journal of the American Society for Information Science 39 (1988) 293–300.

[13] W. Samuelson, R. Zeckhauser, Status quo bias in decision making, Journal of Risk and Uncertainty 1 (1988) 7–59. URL: https://doi.org/10.1007/BF00055564. doi:10.1007/BF00055564.

[14] Y. Yue, R. Patel, H. Roehrig, Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data, in: Proceedings of the 19th international conference on World wide web, WWW '10, ACM, New York, NY, USA, 2010, pp. 1011–1018. URL: http://doi.acm.org/10.1145/1772690.1772793. doi:http://doi.acm.org/10.1145/1772690.1772793.

[15] F. Diaz, R. W. White, G. Buscher, D. Liebling, Robust models of mouse movement on dynamic web search results pages, in: Proceedings of the 22nd ACM conference on Information and knowledge management (CIKM 2013), Association for Computing Machinery, New York, NY, USA, 2013, pp. 1451–1460. URL: https://doi.org/10.1145/2505515.2505717.

[16] D. Read, B. van Leeuwen, Predicting hunger: The effects of appetite and delay on choice, Organizational Behavior and Human Decision Processes 76 (1998) 189–205. URL: https://www.sciencedirect.com/science/article/pii/S0749597898928035. doi:https://doi.org/10.1006/obhd.1998.2803.

[17] D. Read, G. Loewenstein, S. Kalyanaraman, Mixing virtue and vice: combining the immediacy effect and the diversification heuristic, Journal of Behavioral Decision Making 12 (1999) 257–273.

[18] K.-W. Chang, A. Krishnamurthy, A. Agarwal, H. Daume, J. Langford, Learning to search better than your teacher, in: Proceedings of The 32nd International Conference on Machine Learning, 2015, pp. 2058–2066.

[19] A. J. Lonsdale, A. C. North, Why do we listen to music? a uses and gratifications analysis, British Journal of Psychology 102 (2011) 108–134. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000712610X506831. doi:https://doi.org/10.1348/000712610X506831.

[20] B. V. den Bergh, S. Dewitte, L. Warlop, J. D. served as editor, B. S. served as associate editor for this article., Bikinis instigate generalized impatience in intertemporal choice, Journal of Consumer Research 35 (2008) 85–97. URL: http://www.jstor.org/stable/10.1086/525505.

[21] C. Eickhoff, Cognitive biases in crowdsourcing, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 162–170. URL: https://doi.org/10.1145/3159652.3159654. doi:10.1145/3159652.3159654.

[22] P. Chandar, F. Diaz, B. St. Thomas, Beyond accuracy: Grounding evaluation metrics for human-machine learning systems, in: Advances in Neural Information Processing Systems, 2020.

[23] P. Chandar, F. Diaz, C. Hosey, B. St. Thomas, Mixed method development of evaluation metrics, in: KDD '21: Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2021. URL: https://kdd2021-mixedmethods.github.io/.

[24] Q. Zhao, M. C. Willemsen, G. Adomavicius, F. M. Harper, J. A. Konstan, Interpreting user inaction in recommender systems, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 40–48. URL: https://doi.org/10.1145/3240323.3240366. doi:10.1145/3240323.3240366.

[25] A. Chuklin, I. Markov, M. d. Rijke, Click models for web search, Synthesis Lectures on Information Concepts, Retrieval, and Services 7 (2015) 1–115. URL: https://doi.org/10.2200/S00654ED1V01Y201507ICR043. doi:10.2200/S00654ED1V01Y201507ICR043.

[26] A. Singh, Y. Halpern, N. Thain, K. Christakopoulou, E. H. Chi, J. Chen, A. Beutel, Building healthy recommendation sequences for everyone: A safe reinforcement learning approach, in: FAccTRec Workshop, 2020.

[27] T. Schnabel, P. N. Bennett, S. T. Dumais, T. Joachims, Using shortlists to support decision making and improve recommender system performance, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, pp. 987–997. URL: https://doi.org/10.1145/2872427.2883012. doi:10.1145/2872427.2883012.

[28] D. Donato, F. Bonchi, T. Chi, Y. Maarek, Do you want to take notes? identifying research missions in yahoo! search pad, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 321–330. URL: https://doi.org/10.1145/1772690.1772724. doi:10.1145/1772690.1772724.

[29] A. Crescenzi, Y. Li, Y. Zhang, R. Capra, Towards better support for exploratory search through an investigation of notes-to-self and notes-to-share, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1093–1096. URL: https://doi.org/10.1145/3331184.3331309. doi:10.1145/3331184.3331309.

[30] A. N. Hagen, The playlist experience: Personal playlists in music streaming services, Popular Music and Society 38 (2015) 625–645. URL: https://doi.org/10.1080/03007766.2015.1021174. doi:10.1080/03007766.2015.1021174.

[31] I. Kamehkhosh, D. Jannach, User perception of next-track music recommendations, in:

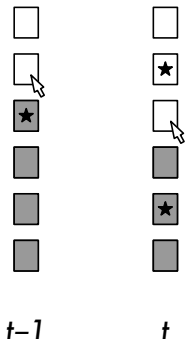
Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 113–121. URL: https://doi.org/10.1145/3079628.3079668. doi:10.1145/3079628.3079668.

[32] C.-W. Chen, P. Lamere, M. Schedl, H. Zamani, Recsys challenge 2018: Automatic music playlist continuation, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 527–528. URL: https://doi.org/10.1145/3240323.3240342. doi:10.1145/3240323.3240342.

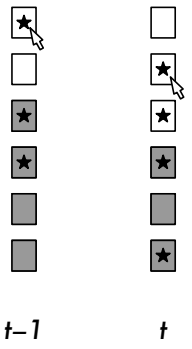
[33] K. Lukoff, U. Lyngs, H. Zade, J. V. Liao, J. Choi, K. Fan, S. A. Munson, A. Hiniker, How the Design of YouTube Influences User Sense of Agency, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445467.

[34] M. Shokouhi, Learning to personalize query auto-completion, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 103–112. URL: https://doi.org/10.1145/2484028.2484076. doi:10.1145/2484028.2484076.

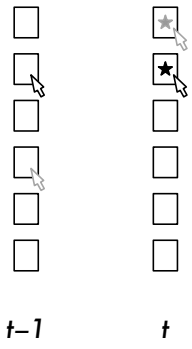
[35] T. D. Wilson, D. T. Gilbert, Affective forecasting, volume 35 of *Advances in Experimental Social Psychology*, Academic Press, 2003, pp. 345–411. URL: https://www.sciencedirect.com/science/article/pii/S0065260103010062. doi:https://doi.org/10.1016/S0065-2601(03)01006-2.



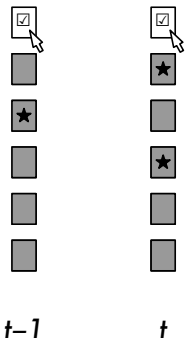
(a) Choice Bracketing.



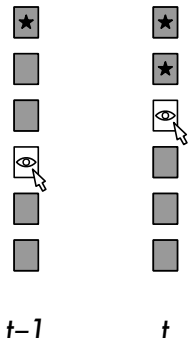
(b) Incomplete Substitutes



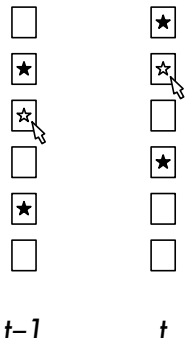
(c) Inter-item Dependencies



(d) Default Preferences



(e) Presentation Bias



(f) Immediacy

Figure 1: Unreliable Implicit Labels. All subfigures represent the system-generated rankings of items at times $t-1$ and t, with ★ representing oracle preferences. Shaded items represent *uninspected* items and user choices are indicated with the cursor. c: gray symbols represent changes in preferences under alternative user choice at time $t-1$, d: items indicated with '✓' are defaults, e: items indicated with an eye have higher relative visual salience, f: ☆ represents immediate preferences.