

Dynamic Critiquing: An Analysis of Cognitive Load^{*}

Kevin McCarthy, Lorraine McGinty, and Barry Smyth

Adaptive Information Cluster, Smart Media Institute,
School of Computer Science and Informatics,
University College Dublin (UCD),
Belfield, Dublin 4, Ireland.
{kevin.mccarthy, lorraine.mcginthy, barry.smyth}@ucd.ie

Abstract. Conversational recommender systems solicit feedback from users in order to objectively inform the recommendation process. Ideally the user is presented with suitable products/services as promptly, as possible. Efficiency is key, and normally, this is measured in terms of the session length (i.e., the number of recommendation cycles with the user). In this paper we argue that it is also important to understand the effort required of the user during these interactions. Cognitive load refers to the level of effort associated with thinking and reasoning. We will look at the cognitive load implications, as measured by interaction time, of a critiquing conversational recommender which uses dynamically generated *compound critiques*. In particular, we find two interesting results. First, on a cycle-by-cycle basis the dynamic critiquing approach places a greater cognitive cost burden than that for the unit critiquing approach. Secondly, and arguably more importantly, the reverse is true when we look at overall session performance – that is, the dynamic critiquing approach outperforms the unit critiquing variation. We demonstrate these in relation to results obtained in a recent real-user trial.

1 Introduction

Recommender systems are important when it comes to helping users to navigate through complex product spaces [1–3]. In particular they provide useful assistance to users even when user requirements are initially unclear. For example, conversational recommender systems aim to refine a user’s initial requirements by presenting a sequence of recommendations and inviting the user to provide feedback on each suggestion. This feedback can be provided in different ways, and is often dependant on the recommendation setting in question. Previous work by [4] describes four distinct forms of feedback - value elicitation, critiquing, preference-based, and ratings-based feedback - and investigates how they rank in terms of reducing recommendation session length. However, this work does point out that recommendation session length is not the only influencing criterion

^{*} This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361.

when it comes to assessing recommender performance, and selecting appropriate feedback strategies. For instance, the *cognitive cost* to the user of providing feedback is another, albeit lesser investigated, issue. This is the issue under scrutiny in this paper.

In this paper we are interested in a particular form of feedback, called critiquing [5–7]. Critiquing-based recommender systems expect the user to provide feedback in the form of directional feature-level preferences for a product. For example, in a holiday recommender a user might indicate that they are looking for a holiday that is “*closer to home*”, this is a unit critique over the *location* feature of a holiday suggestion. In this way, the standard model of critiquing requires the user to apply a *single* critique to a *single* feature (*unit critiquing*), but recently this model has been extended to cover multiple features, by supporting dynamically generated *compound critiques*.

Section 2 describes the unit and compound critiquing ideas in further detail, and discusses some of the advantages we have previously reported in relation to the notion of dynamically generating compound critiques. A common question we have been faced with in the past has been in relation to the cognitive load impact associated with this idea. For example, how much of an increase in the user’s cognitive load is associated with choosing a compound critique over a unit critique? We evaluate this issue in Section 4, such that developers of critique-guided recommender systems can better understand the cost-benefit characteristics associated with using unit and compound critiques. We propose to estimate cognitive load by measuring the amount of time that it takes a user to interact with the recommender system, and we use this to compare the cognitive load characteristics of standard (unit) critiquing and dynamic compound critiquing. In particular, we demonstrate that while the dynamic critiquing approach presents with a higher cognitive load at the level of a single recommendation cycle, its ability to reduce overall average session length has a positive impact on the overall cognitive load at the level of the recommendation session. Importantly, we demonstrate these ideas using results obtained in a recent live-user trial.

2 A Review of Critiquing-Based Recommender Systems

Critiquing, as a form of feedback, is best known by association with the FindMe recommender systems [1, 5]. The original motivation for critiquing included the need for a type of feedback mechanism that was simple for users to understand and apply, and yet informative enough to focus the recommender system. For instance, the Entrée restaurant recommender [1] presents users with a fixed set of directional critiques in each recommendation cycle. In this way users can easily request to see further suggestions that are different from the current recommendation, in terms of some specific feature. For example, the user may request another restaurant that is *cheaper* by critiquing its *price*. Critiques of this type are what we call *unit* critiques. This ultimately limits the ability of the recommender to narrow its focus, because it is guided by only single-feature

preferences from cycle to cycle. An alternative strategy is to consider critiques that operate over multiple features, what we call *compound* critiques. This idea of compound critiques is not novel; the seminal work of Robin Burke [5] refers to critiques for manipulating multiple features. For instance, Car Navigator, a car recommender from the FindMe family, offers both static unit critiques and static compound critiques to the user. An example of one such compound critique is the *sportier* critique, which operates over a number of different car features; *engine size*, *acceleration* and *price* are all increased. Obviously compound critiques have the potential to improve recommender efficiency because they allow the recommender to focus on multiple feature constraints in each cycle. However, in the past when compound critiques have been used they have been hard-coded by the system designer so that the user is presented with a fixed set of compound critiques in each recommendation cycle. These compound critiques may, or may not, be relevant depending on the cases that remain at a given point in time.

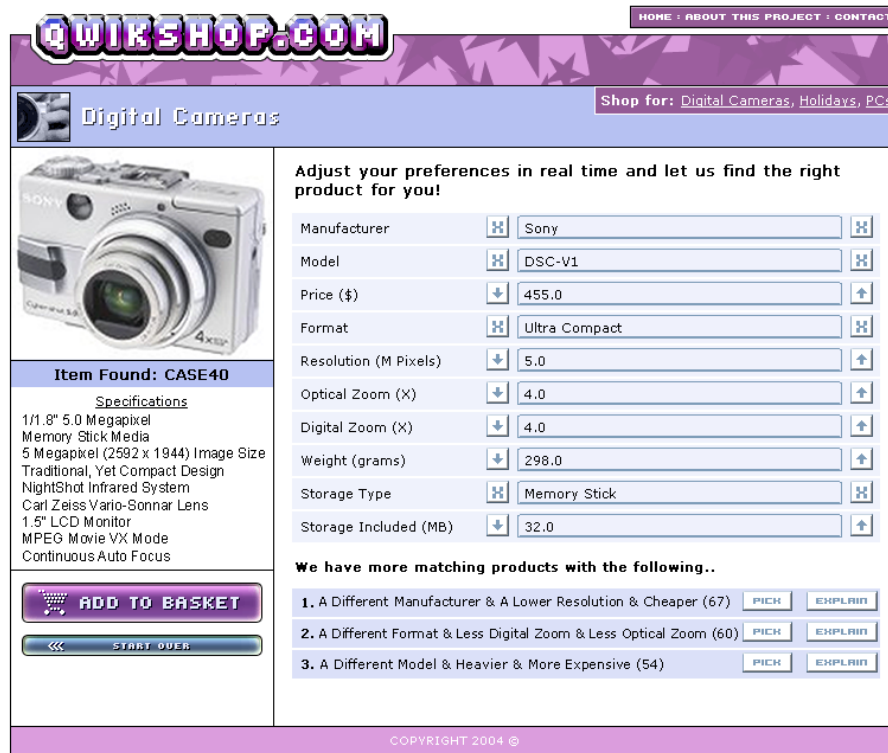


Fig. 1. Screenshot of our prototype dynamic critiquing recommender.

Recently, we have argued the need for a more dynamic approach to critiquing whereby compound critiques are generated, *on-the-fly*, for each recommendation cycle [8, 9]. This *dynamic critiquing* approach generates a set of compound critiques to present to the user in each recommendation cycle. Because these compound critiques are informed by the remaining products, the recommender system will *always* return a recommendation for one of these. In previous work [8, 9] we have described how these compound critiques are generated, selected and presented to the user. The compound critiques generated by the dynamic critiquing approach are always applicable to the current recommendation. We have also developed methods such as critique profiling and introduced diversity in order to enhance the beneficial effects of our dynamically generated compound critiques. Figure 1 shows a screenshot of a digital camera recommender system that we have developed to showcase and evaluate the dynamic critiquing approach. It shows a recommended case, its unit critiques and three relevant compound critiques. From here the user can select a critique to inform the next recommendation, terminating their session when they see a satisfactory camera.

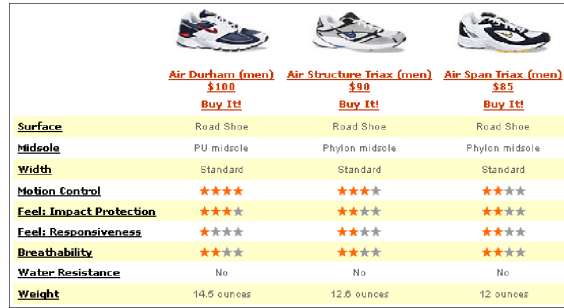
In previous publications we have looked at improving recommender efficiency by reducing the number of conversational cycles. We found in both artificial and on-line evaluations [8–10] that session reductions of up to 69% can be obtained when the user chooses dynamically generated compound critiques. However, the compound critiques we generate change from cycle to cycle, unlike their *unit* counterparts. For example, Entrée uses 7 fixed unit critiques in each cycle and the same critique options appear in the same positions every cycle. Our approach presents unit critiques in a similar way to Entrée, however the compound critiquing part of the interface changes dynamically during each cycle as do the presented compound critiques. How much of an increase in the user’s cognitive load is associated with choosing a compound critique over a unit critique, when we consider that the compound critiques are dynamically generated? We will try to answer this question in evaluating the results of a recent live-user trial.

3 Cognitive Load in Conversational Recommenders

The issue of cognitive load in the context of conversational recommender systems is one that has not been explored in great detail to date. The vast majority of recommender systems make cognitive demands on users in some way. Limiting factors include: (1) users rarely have a clear understanding of what product they are looking for, (2) users may not be proficient in the required product description knowledge, (3) users may be unable to effectively compare presented suggestions, and/or (4) are often incapable of retaining learned information in short-term memory.

Assessing cognitive load implications associated with recommendation scenarios of this kind is a very difficult task. Related research has investigated the prospects of using physiological methods, such as pupil size, to measure a user’s cognitive load [11]. Despite early optimism, empirical evaluations have demon-

stated that behavioural indicators, such as reading speed, are actually more accurate.



	Air Durham (men) \$100 Buy It!	Air Structure Triax (men) \$90 Buy It!	Air Span Triax (men) \$85 Buy It!
Surface	Road Shoe	Road Shoe	Road Shoe
Midsole	PU midsole	Phylon midsole	Phylon midsole
Width	Standard	Standard	Standard
Motion Control	★★★★	★★★★	★★★★
Feel: Impact Protection	★★★★	★★★★	★★★★
Feel: Responsiveness	★★★★	★★★★	★★★★
Breathability	★★★★	★★★★	★★★★
Water Resistance	No	No	No
Weight	14.5 ounces	12.6 ounces	12 ounces

Fig. 2. Screenshot from the Nike Advisor site, Nike.com, illustrating the cognitive cost burden associated with the case evaluation level.

Accordingly, evaluations of the cognitive load impact of recommender techniques and technologies have tended to examine a user’s willingness to interact and speed of interaction as a measure of this cost. In a typical recommendation cycle there are two cognitive task levels that contribute to this cost. They are, (1) cognitive load the user experiences at the *case evaluation level*, and (2) cognitive load they tolerate at the *feature interaction level*. The case evaluation level of cognitive load is best demonstrated when k cases are presented to the user as in the example in Figure 2. There is a comparative cognitive cost because the user has to compare the presented cases to each other before giving feedback.

Our research looks at the cognitive burden placed on users of conversational recommender systems at the feature interaction level. As mentioned earlier, our digital camera recommender uses critiquing as its user feedback mechanism. Only one suggestion is presented in each recommendation cycle so there is no associated cost at the case evaluation level. As shown in Figure 1, the user first examines the relevant unit and compound critiques (i.e., each made up of 3 unit critiques), before providing their feedback. To date our evaluations have concentrated predominantly on the efficiency benefits of our dynamic critiquing approach, in terms of the session length reductions realised. While results have been positive we have been challenged as to whether there are significant cognitive load issues at the feature interaction level. That is, are there cost implications enforced on the user as a result of having to *breakdown* and *understand* the critique options offered. More importantly, these compound critiques change dynamically, and this could lead to longer overall session times. It could be the case that the cognitive cost (i.e., time spent) encountered at the feature interaction level far outweighs the benefit of shorter recommendation session lengths – for example, a unit critiquing user who finds their product in 10 cycles and

takes 2 minutes to do so is better off than a compound critiquing user who finds the same product in 5 cycles but takes 6 minutes to do so.

Obviously if the users in our experiment who use the compound critiques to reduce their system interactions take a considerably longer duration of time to be recommended the product they were looking for, then dynamic critiquing and the use of compound critiques is infeasible in a real system. In our evaluation in Section 4, instead of looking at the number of interactions with the system as a measurement of efficiency we analyse how the time taken for a user to respond with feedback can be used to determine how reductions in interaction cycles correlate with increases in cognitive load.

Some timing analysis has been performed in relation to recommendation efficiency. For example, Smyth and Cotter [12, 13] look at the efficiency of personalized mobile portals. They argued that the efficiency benefit comes not only from the fact that the user has fewer interaction cycles with the system (clicks and scrolls in this case), but also that the user ends up spending less time on unnecessary navigation. The fact that users spend less time navigating to what they want is very important in the mobile domain as users are often charged according to the amount of time they spend connected to the service.

Cognitive load analysis, as measured by interaction time, can only be carried out in the context of live-user evaluations. We have recently carried out such a trial but to date have only reported results that comment on the efficiency benefits of our dynamic critiquing approach (and its extensions) in terms of recommendation cycles. The key contribution of this paper is the cognitive analysis of this data to determine whether or not dynamic compound critiques place an unacceptable burden on the user in terms of their cognitive load.

4 Evaluation

We are especially interested in the response time of users utilizing the compound critiques in our system as a measure of feature interaction level cognitive load. Firstly, we report on the average overall session time and corresponding session lengths for low and high frequency compound critique users. We wish to see if high frequency users suffer from longer session times, therefore telling us if there is a cognitive load trade off by having a shorter number of conversational cycles. Then we look at a break down of the low and high frequency users' time results for both unit and compound critiques. This should give us some insight into how low and high frequency compound critique users differ in how they use the system with respect to cognitive time involved at the feature-interaction level.

4.1 Setup

Users for our trial were made up of both undergraduate and postgraduate students from the School of Computer Science and Informatics at University College Dublin. The participants were asked to use our Digital Camera recommender system (see Figure 1) during December 2004. They were asked to use the system

as if they were purchasing a new digital camera and they were presented with a very brief introductory tutorial to explain the interface and the different forms of feedback. In total 45 different users led to 55 unique recommendation sessions, which were closely monitored and logged. The average overall session length was just over 21 cycles with an average total elapsed time of just over 105 seconds. In addition, users tended to select compound critiques in about 26% of their cycles and unit critiques in the remaining cycles; these application frequencies are similar to those previously reported [8, 9].

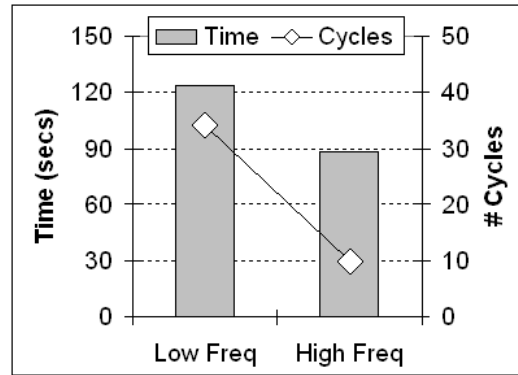


Fig. 3. Cognitive time with efficiency results.

4.2 Results

To assess the relative impact of unit and compound critiques we divided the sessions into two equal groups: *low frequency* sessions included all those sessions whose compound critique application frequency was less than the median frequency, with *high frequency* sessions making up the remainder. On average users with low frequency sessions applied compound critiques only about 7% of the time compared to 43% of the time for high frequency users. Next we computed the average session length and time for these low and high frequency sessions. The results are presented in Figure 3 and clearly indicate an advantage to the high frequency sessions, which are seen to have 71% fewer cycles per session (10 *vs.* 34) and be 28% shorter in time (88.7 seconds *vs.* 123.8 seconds per session) when compared to the low frequency sessions. In other words, when users availed of compound critiques more frequently they benefited from fewer recommendation cycles and shorter recommendation session durations.

By looking at the timing of individual cycles within a session it was possible to gain an understanding of how long it took users to interpret and select between unit and compound critiques – the cognitive load at a feature interaction level.

To do this we compared the average time a user took during a cycle where a unit critique was selected, to the average time taken when a compound critique was selected. In Figure 4 we see that over all sessions, the average time for unit critique cycles is 3.99 seconds, compared to 11.75 seconds for cycles where compound critiques are chosen. This is consistent with the notion that compound critiques are likely to be more time consuming to interpret than the less complex unit critiques; typically, a compound critique is made up of about 3 separate unit critiques and we see here that cycles where compound critiques are selected are almost 3 times as long as unit critique cycles.

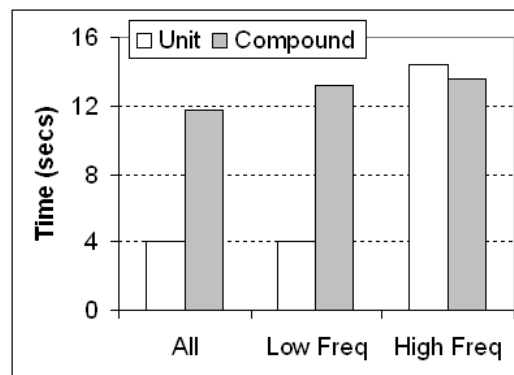


Fig. 4. Unit and compound critiques cognitive time results.

However, this is not the complete picture. In Figure 4 we also show the average time for unit and compound critique cycles within the low frequency and high frequency sessions. The low frequency results are similar to the overall results with unit critique cycles taking about 4 seconds and compound critique cycles taking about 13 seconds. However, the high frequency cycles are very different with unit critique cycles taking over 14 seconds. In fact they are more time consuming than the cycles where compound critiques are chosen. This can be explained by a fundamental difference in the behaviour of low frequency and high frequency users. Remember that the former had an average compound critique application frequency of only 7% (compared to more than 40% for the latter). The application data suggests that the low frequency users were very rarely selecting compound critiques and the timing data indicates that when they were selecting unit critiques they were doing so without due consideration of the compound critique options that were available; this is useful because it suggests that the average unit critique time is indicative of just the time taken to interpret unit critiques. In contrast, the average cycle time for the high frequency users is roughly the same regardless of whether a unit critique or a compound critique is finally chosen. This indicates that these users were at least giving

consideration to both unit and compound critiques during each cycle, and so even when a unit critique was chosen they were still taking the time to reflect on the compound critique options.

5 Conclusions

This paper has described the results of a recent live-user trial to evaluate the relative cognitive load merits of unit and compound critiquing in a conversational recommender system. For the first time we have focused not only on the number of cycles in a recommendation session, but also on the elapsed response time as a measure of cognitive load. Because of this we have been able to gain a better understanding of key differences in the way that users interact with a critiquing-based recommender system. We have clarified that there is a significant increase in the cognitive load associated with the use of compound critiques, when compared to users who focused almost exclusively on unit critiques. However this additional cognitive investment is worthwhile because it translates into sessions that are shorter, both in terms of their elapsed time (28% reduction) and their number of cycles (71% reduction).

It is also interesting to note how users use the system. The low frequency compound critique users very rarely utilize the compound critiques (only about 7% of the time). However, when they did take the time to use them, they responded in almost exactly the same time as the high frequency users for compound critiques. This shows us that the results show a consistent cognitive load across both groups of users for compound critiques. If we also look at the cognitive load times for unit critiques we see there is a huge difference, and as mentioned, this can be attributed to the fact that low frequency users are not even considering the compound critiques on offer.

The results reported in this paper tell us that even if there is a larger cognitive load at the feature interaction level, associated with cycles where compound critiques are considered, the end result is that the cognitive increase is still outweighed by the recommendation cycle reduction possible. Users will find the product they are looking for more quickly and more efficiently with less overall cognitive effort and time consumed if they utilize compound critiques.

As a future work discussion, if we refer again to the screenshot shown in Figure 1, we can see that the unit critiques are placed right beside the feature values. When a user is considering a unit critique they can see the feature value they will be critiquing. It is inherently more difficult for a user to understand a compound critique as the user does not have the benefit of being able to see the feature values beside the unit critiques that make up the compound critique. Also it is a well-known fact in interface design that if toolbar icons, or in our case the unit critiques, do not change position or value from interaction to interaction, that the user will come to subconsciously know that they are available. This learning ability is not available to the compound critiques as they are generated dynamically during each cycle of the recommendation process. It

would be interesting to consider how these factors affect the cognitive load results presented here. However such an evaluation is beyond the scope of this paper.

References

1. Burke, R., Hammond, K., Young, B.: The FindMe Approach to Assisted Browsing. *Journal of IEEE Expert* **12**(4) (1997) 32–40
2. Schafer, J.B., Konstan, J.A., Riedl, J.: E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery* **5** (2001) 115–153
3. Shimazu, H.: ExpertClerk : Navigating Shoppers' Buying Process with the Combination of Asking and Proposing. In Nebel, B., ed.: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, Morgan Kaufmann (2001) 1443–1448 Seattle, Washington, USA.
4. Smyth, B., McGinty, L.: An Analysis of Feedback Strategies in Conversational Recommender Systems. In Cunningham, P., ed.: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Cognitive Science (AICS-2003)*. (2003) Dublin, Ireland.
5. Burke, R., Hammond, K., Young, B.: Knowledge-based Navigation of Complex Information Spaces. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press (1996) 462–468 Portland, OR.
6. Faltings, B., Pu, P., Torrens, M., Viappiani, P.: Designing Example-Critiquing Interaction. In: *Proceedings of the International Conference on Intelligent User Interface (IUI-2004)*, ACM Press (2004) 22–29 Funchal, Madeira, Portugal.
7. Sherin, S., Lieberman, H.: Intelligent Profiling by Example. In: *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2001)*, ACM Press (2001) 145–152 Santa Fe, NM, USA.
8. McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: On the Dynamic Generation of Compound Critiques in Conversational Recommender Systems. In Bra, P.D., ed.: *Proceedings of the Third International Conference on Adaptive Hypermedia and Web-Based Systems (AH-04)*, Springer (2004) 176–184 Eindhoven, The Netherlands.
9. Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Dynamic Critiquing. In Calero, P.A.G., Funk, P., eds.: *Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04)*, Springer (2004) 763–777 Madrid, Spain.
10. McCarthy, K., McGinty, L., Smyth, B., Reilly, J.: On the Evaluation of Dynamic Critiquing: A Large-Scale User Study. In Veloso, M., Kambhampati, S., eds.: *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, AAAI Press (2005) Pittsburgh, PA, USA.
11. Schultheis, H., Jameson, A.: Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In Bra, P.D., ed.: *Proceedings of the Third International Conference on Adaptive Hypermedia and Web-Based Systems (AH 2004)*, Springer (2004) 225–234 Eindhoven, The Netherlands.
12. Smyth, B., Cotter, P.: The Plight of the Navigator: Solving the Navigation Problem for Wireless Portals. In Bra, P.D., Brusilovsky, P., Conejo, R., eds.: *Proceedings of the Second International Conference on Adaptive Hypermedia and Web-Based Systems (AH 2002)*, Springer (2002) 328–337 Malaga, Spain.
13. Smyth, B., Cotter, P.: Personalized adaptive navigation for mobile portals. In van Harmelen, F., ed.: *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI 2002)*. (2002) 608–612 Lyon, France.