On the Evaluation of Dynamic Critiquing: A Large-Scale User Study *

Kevin McCarthy, Lorraine McGinty, Barry Smyth and James Reilly
Adaptive Information Cluster, School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland.
E-mail: {kevin.mccarthy, lorraine.mcginty, barry.smyth, james.d.reilly}@ucd.ie

Abstract

Critiquing is an important form of feedback in conversational recommender systems. However, in these systems the user is usually limited to critiquing a single product feature at a time. Recently *dynamic critiquing* has been proposed to address this shortcoming, by automatically generating compound critiques over multiple features that may be presented to the user at recommendation time. To date a number of different versions of *dynamic critiquing* have been evaluated in isolation, and with reference to artificial users. In this paper we bring together the main flavors of dynamic critiquing and perform a large-scale comparative evaluation as part of an extensive real-user trial. This evaluation reveals some interesting facts about the way real users interact with critique-based recommenders.

Introduction

Product recommender systems are designed to help users to find the products (e.g., books, computers, furniture, travel packages, etc.) they want faster. So-called *conversational* recommender systems have been especially successful as e-commerce solutions, helping prospective buyers to quickly locate suitable products by facilitating the incremental construction of a more accurate picture of their requirements.

Conversational recommender systems typically make a sequence of recommendations, offering users an opportunity to provide feedback in order to influence future recommendation cycles. Different types of feedback (eg. *value elicitation*, *ratings-based* and *preference-based* feedback) vary in terms of their recommendation-efficiency/cognitive-load characteristics. One feedback technique that achieves a useful balance is *critiquing*, which allows users to constrain a product feature, without requiring them to provide a specific value (Burke, Hammond, & Young 1996). For example, a user of a digital camera recommender might indicate that they are looking for something "less expensive" by critiquing the $Price$ feature of a suggestion. This standard approach to critiquing focuses on the use of so-called *unit critiques* that constrain a single feature at a time.

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

Recently, researchers have explored the possibility of critiquing multiple features simultaneously in order to facilitate larger jumps through a product space. For example, the *dynamic critiquing* family of algorithms by (McCarthy *et al.* 2004) demonstrates how the automatic generation of these so-called compound critiques during each recommendation cycle can significantly reduce recommendation session lengths.

In this paper we build on recent work in this area and, describe a recommendation approach that combines unit and compound methods of critiquing. We describe a number of approaches to generating compound critiques, one of which looks at ways to improve the diversity of compound critiques and another that allows a user's session critiquing history to influence recommendation. In a previous paper (McCarthy *et al.* 2005b), we reported on a small real-user study which we performed to test our prototype system and we reported initial results of our basic dynamic critiquing strategy. In this paper we present a new large-scale, real-user evaluation of dynamic critiquing comparing the effectiveness of 4 versions of the approach across more than 1000 real-user recommendation sessions. The results demonstrate the significant benefits that are possible with dynamic critiquing, and also clarify a number of outstanding issues in relation to the way in which real users interact with a unit and compound critiquing interface.

Dynamic Critiquing

Figure 1 shows a screen-shot of a conversational recommender system that we have developed to showcase and evaluate dynamic critiquing. 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 its current guise our system presents users with an interface that combines static and dynamic elements according to the principle of *partitioned dynamicity*; see (Weld *et al.* 2003). Accordingly the static interface elements (which include the camera description and unit critiques) are augmented by a separate panel for the dynamic, compound critiques. In this way it was possible to provide a level of interface adaptability (changing and context-sensitive compound critiques) without obscuring the core functionality provided by the primary interface elements.

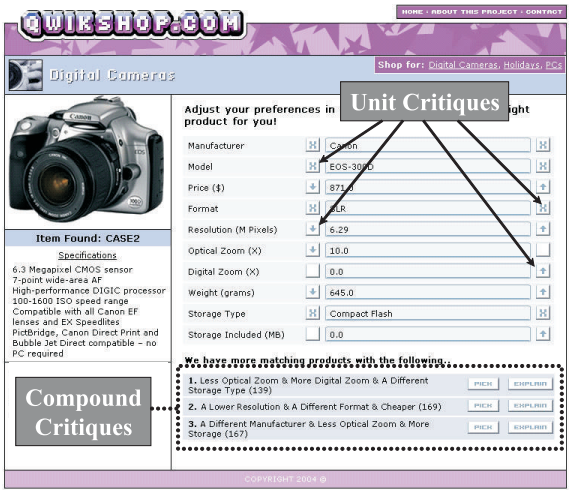

Figure 1: A digital camera recommender system that implements unit and compound critiquing.

The Standard Dynamic Critiquing Approach

In this section we outline how these compound critiques are generated and discuss a number of techniques for improving their quality. The standard approach to dynamic critiquing has been previously described by (McCarthy *et al.* 2004). The basic idea is that during every recommendation cycle, in addition to selecting a new case to recommend to the user, the system should also present a set of compound critiques that characterise the remaining cases. For example, in Figure 1 we see an example of a compound critique leading to 169 cases with *less resolution* and a *different format* for a *cheaper* price. Generating these compound critiques involves 3 basic steps:

STEP 1 - Generating Critique Patterns: For a given recommendation cycle each of the remaining cases is re-described by a *critique pattern* which captures the relationship between each case and the current recommended case. For example, Figure 2 shows an example from the Digital

	Current Case	Case c from CB	Critique Pattern
Manufacturer	Canon	Sony	!=
Model	Powershot S500	DSC-V1	!=
Format	Ultra Compact	Ultra Compact	=
Resolution (M Pixels)	5.1	5.0	<
Optical Zoom (X)	3	4	>
Digital Zoom (X)	4.1	4	<
Weight (grams)	215	298	>
Storage Type	Compact Flash	Memory Stick	!=
Storage Included (MB)	32	16	<
Price (Euro)	443.00	455.00	>

Figure 2: Generating a critique pattern.

Camera domain. The resulting critique pattern reflects how a case c differs from the current case in terms of individual feature critiques. For example, the critique pattern shown includes a "$<$" critique for Resolution— we will refer to this as $[Resolution <]$—because the comparison case has less pixels than the current recommendation.

STEP 2 - Mining Compound Critiques: This step involves identifying recurring patterns of unit critiques within the current set of critique patterns. This is similar to the market-basket analysis task where the well-known Apriori algorithm (Agrawal *et al.* 1996) has been used to characterize recurring itemsets as association rules of the form $A \rightarrow B$. During this stage, Apriori is applied during each cycle to the remaining product cases in order to identify groups of recurring unit critiques; we might expect to find the co-occurrence of unit critiques like $[Resolution >]$ infers $[Price >]$. Apriori returns lists of compound critiques of the form $\{[Resolution >], [Price >]\}$ along with their *support* values (the percentage of critique patterns for which the compound critique holds).

STEP 3 - Grading Compound Critiques: Of course it is not practical to present large numbers of different compound critiques as user-feedback options in each cycle. A filtering strategy is used to select the k most useful critiques for presentation purposes based on their support values; compound critiques with low support values have the ability to eliminate many product cases from consideration if chosen. The work of (McCarthy *et al.* 2004) has looked at a number of ways to filter critiques, concluding that preferring critiques with low support values has the potential to offer the best recommendation efficiency benefits.

Improving Critique Diversity

Sometimes the critique generation strategy can produce *poor quality* compound critiques that are very similar to each other, limiting their applicability. One potential solution to this problem is to actively improve the diversity of compound critiques by including a direct measure of diversity during critique selection. Similar issues have been investigated by (McGinty & Smyth 2003; Shimazu 2001), albeit in a different recommender context. For example the diversity enhancing algorithm described by (Smyth & McClave 2001) selects items on the basis of a quality metric that maximizes similarity to some target item, while minimizing average similarity to the items selected so far. In (McCarthy *et al.* 2005a) a similar quality metric is defined for compound critiquing as shown by Equation 1, and a similar selection strategy is used during critique selection; note cc stands for the current compound critique being considered, C represents the set of critiques so far selected and $U(C)$ is the set of unit critiques that make up C. Priority is given to compound critiques with low support scores, that are diverse relative to any compound critiques that have been so far selected for presentation to the user.

$$DQual(cc, C) = Support(cc) \ * \ 1 - \frac{|U(\{cc\}) \cap U(C)|}{|U(\{cc\}) \cup U(C)|} \quad (1)$$

Incremental Critiquing

Another problem with the standard approach stems from the fact that when it comes to generating critiques for the current cycle, no direct consideration is given to feedback previously provided by users during a session. Many users may not have a clear understanding of their requirements at the beginning of a session and rely on the recommender as a

means to educate themselves about features of a product-space (Sherin & Lieberman 2001; McCarthy *et al.* 2004; Pu & Kumar 2004). Accordingly, users may select apparently incompatible critiques as they explore different areas of the product space in order to build up a clearer picture of what is available. For example, we might find a prospective camera owner seeking a camera that is *cheaper* than the current 500 euro recommendation, but later asking for a recommendation that is *more expensive* than another 500 euro option once they recognize the compromises that are associated with lower priced cameras (McSherry 2003). We believe that it is important to consider such interaction histories when producing future recommendations.

Incremental Critiquing (Reilly *et al.* 2004) addresses these problems by considering past critique selections in each recommendation cycle. It does this by maintaining a session-based user model made up of those critiques chosen by the user so far. This model is given by $U = \{U_1, ..., U_n\}$, where U_i is a single unit critique. During recommendation this model is used to influence the choice of a new product case, along with the current critique. After a critique has been selected the model is updated accordingly, with compound critiques separated into their constituent unit parts.

Maintaining an accurate user model, however, is not quite as simple as storing a list of previously selected critiques. As we have mentioned above, some critiques may be *inconsistent* with earlier critiques. For example, in the case of a Digital Camera recommender, a user selecting a critique for *more digital zoom*, beyond the *5x* of the recommended case, may later *contradict* themselves by indicating a preference for *less digital zoom* than the *3x* offered by a subsequent case. In addition, a user may *refine* their requirements over time. They might start by indicating a preference for more than *3.1M Pixels* resolution. Later they might indicate a preference for more than *5.0M Pixels*. The incremental critiquing strategy updates the user model by adding the latest critique only after pruning previous critiques so as to eliminate these sorts of inconsistencies. Specifically, prior to adding a new critique all existing critiques that are inconsistent with it are removed, as are all existing critiques for which the new critique is a refinement.

$$Compat(c', U) = \frac{\sum_{\forall i} satisfies(U_i, c')}{|U|} \quad (2)$$

$$IQual(c', c, U) = \alpha * Compat(c', U) + (1-\alpha) * Sim(c', c) \quad (3)$$

The theory behind the user model is that it should be used to influence the recommendation process, prioritising those product cases that are compatible with its critiques. The standard approach to recommendation, when using critiquing, is a two step one. First, the remaining cases are filtered by eliminating all of those that fail to satisfy the current critique. Next, these filtered cases are rank ordered according to their similarity to the current recommendation. *Incremental critiquing* makes one important modification to this procedure. Instead of ordering the filtered cases on the basis of their similarity to the recommended case, a compatibility score for each candidate case is also computed. This compatibility score is essentially the percentage of critiques in the user model that this case satisfies (see Equation 2 and note that $satisfies(U_i, c')$ returns 1 when the critique, U_i satisfies the filtered case, c, and returns 0 otherwise). Thus, a case satisfying 3 out of 5 critiques in a user model gets a compatibility score of 0.6.

This compatibility score is then combined with the candidate's (c') similarity to the recommended product case, c, in order to obtain an overall quality score as in Equation 3; by default $\alpha = 0.75$. This quality score is used to rank order the filtered cases prior to the next recommendation cycle and the case with the highest quality is then chosen as the new recommendation. This approach allows us to prioritise those candidate cases that: (1) satisfy the current critique; (2) are similar to the most recently recommended case; and (3) satisfy many previous critiques. In so doing we are implicitly treating the past critiques in the user model as *soft constraints* for future recommendation cycles; it is not essential for future recommendations to satisfy all of the previous critiques, but the more they satisfy, the better they are regarded as recommendation candidates. Moreover, given two candidates that are equally similar to the previously recommended case, the algorithm prefers the one that satisfies the greater number of recently applied critiques.

Evaluation

We have previously reported on a small user study where we showed rudimentary efficiency results for our basic dynamic critiquing approach (McCarthy *et al.* 2005b). In this paper we are especially interested in understanding how users interact with the compound critiques that are generated by the *different* critiquing strategies. Firstly, we report on a direct comparison between the different critique generation strategies, comparing their overall application frequency and session length characteristics; in particular we analyse the session length implications of low versus high usage frequencies. Secondly, we perform a more fine-grained analysis in order to investigate *when* users tend to use compound critiques. We are especially interested in patterns of usage that might indicate a propensity to use compound critiques at the start of a session, in order to help the user to focus quickly in on a relevant region of the product-space, with unit critiques being used later in sessions, in order to fine-tune their requirements. Finally, we return to the issue of the *frequency-of-use* by asking whether there appears to be an optimal usage frequency that delivers the best efficiency gains. We answer these questions with reference to a large-scale real-user trial we conducted during December 2004.

Setup

Users for our trial were made up of both undergraduate and postgraduate students from the department of Computer Science at University College Dublin. Trial participants were invited to use our Digital Camera Recommender (see Figure 1) during December 2004. The trial consisted of two parts. In the first part, each trialist was asked to participate in a so-called *training* session so that they could become acquainted

with the critiquing mode of interaction. Here they were presented with a specific camera case as a starting point and then asked to *shop* from this case to locate their ideal camera. In the second part, they were presented with a fixed *start* case and *target* case, and were asked to locate the target case by using critiques (unit or compound) of their choice. There were 25 different start–target pairs and these were randomly assigned at the beginning of every session. Trialists were permitted to use the system as often as they liked. This setup ensured that any learning effect was minimised, which has been confirmed but space restrictions prohibit further discussion.

Here we report on results from part 2 of the trial, which generated 1092 user sessions from 76 unique users. The trialists were made up of 53% undergraduate and 47% postgraduate students with 61 male and 15 female participants. The different critiquing strategies; standard dynamic critiquing (*Standard*), dynamic critiquing with diversity (*Standard-Diversity*), incremental critiquing (*Incremental*) and incremental critiquing with diversity (*Incremental-Diversity*) were deployed during different days of this trial and a variety of details were logged for each user session (e.g., the camera cases presented, the critiques applied, etc.).

Overall Recommendation Efficiency

To be successful, recommender systems must be able to efficiently guide a user through a product-space and, in general, short recommendation sessions are to be preferred. Previously research by (McCarthy *et al.* 2004; 2005a; Reilly *et al.* 2004) has demonstrated how the different dynamic critiquing strategies are capable of delivering significantly shorter sessions than unit critiquing. However, to date the performance of these different strategies have not been compared directly. We do this now in Figure 3 where we present the overall average session length and compound critique application frequency for the different versions of our recommender system. The results clearly demonstrate that both

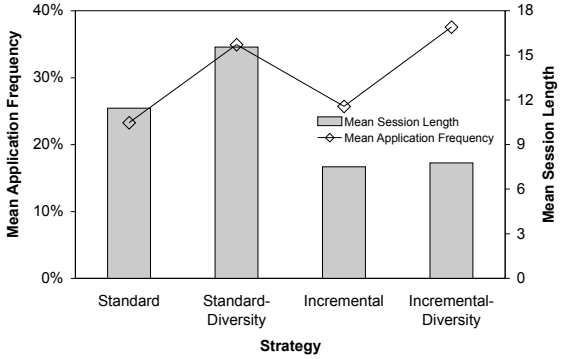

Figure 3: Overall recommendation efficiency results for the four different dynamic critiquing strategies.

versions of the incremental critiquing approach (*Incremental* and *Incremental-Diversity*) improve on both versions of the (non-incremental) standard dynamic critiquing approaches. For instance, the average session length for $Standard$ and *Standard-Diversity* is 13.5 cycles, compared to only 7.6 for the average of $Incremental$ and *Incremental-Diversity*; in other words, the incremental critiquing strategies deliver a 44% reduction in average session length. Interestingly, we see that the diversity-enhancing strategy has a negative impact on session length. *Standard-Diversity* leads to sessions that are 36% longer than $Standard$, and *Incremental-Diversity* delivers sessions that are almost 5% longer than $Incremental$. When we look at application frequency (i.e., the frequency with which compound critiques are selected by users during sessions) we see that the diversity-enhancing strategy results in the more frequent application of compound critiques by our trial users; an average application frequency of about 35% for the *Standard-Diversity* and *Incremental-Diversity* techniques. That is, users are selecting compound critiques in 35% of the cycles and unit critiques in the remaining 65% of cycles, compared to an average of 24% for the $Standard$ and $Incremental$ techniques.

Comparative Recommendation Efficiency

While the previous results tell us about the relative performance of our different critiquing strategies, they do not directly inform us about the value of dynamic critiquing per se (i.e., relative to unit critiquing). This question has been partly answered by the work of (McCarthy *et al.* 2004) through a direct comparison of dynamic and unit critiquing. An alternative is to look for efficiency differences between those sessions that contain few applications of compound critiques (*low frequency* sessions) compared to *high frequency* sessions. To do this, for each recommendation strategy, we partition its sessions into two groups according to the median application frequency for compound critiques observed. This gives a set of low frequency sessions (those that have application frequencies that are less than the median) and a set of high frequency sessions (where application frequency is greater than or equal to the median).

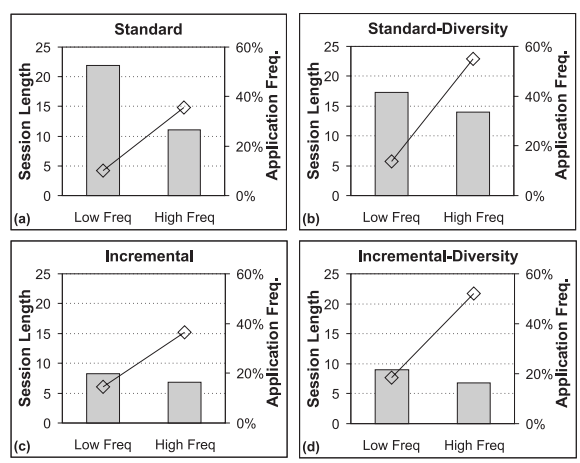

Figure 4: Comparative recommendation efficiency results for the four different dynamic critiquing strategies.

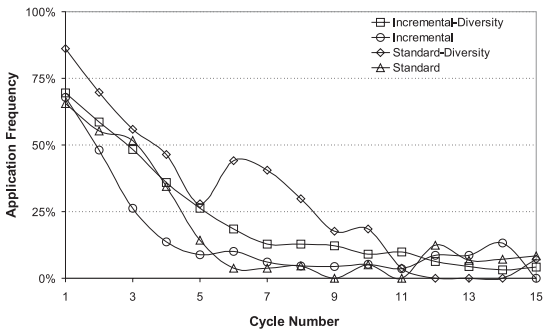

Figure 5: The application frequency results for the four different dynamic critiquing strategies on a per cycle basis.

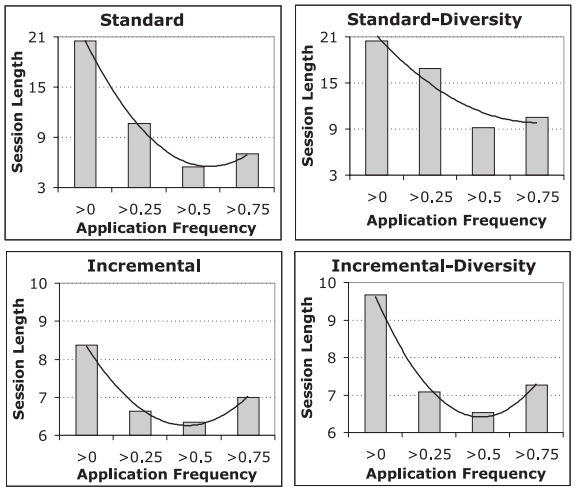

Figure 6: Average session length results for recommendation sessions with different degrees of application frequency for the four different dynamic critiquing strategies.

Application Frequency

Figure 4(a-d) presents the average session length for these groupings of sessions, and their average compound critique application frequency, across all four recommendation strategies. For each of the recommendation strategies we see a significant (at the 99% confidence level) reduction in session length for sessions that tend towards more frequent use of compound critiques (ranging from 35% to 55% application frequency) compared to the low frequency sessions (with between 10% to 20% application frequency). For example, there is a 50% reduction in average session length (11 vs. 21.8) for $Standard$ (Figure 4(a)), a 20% reduction for *Standard-Diversity* and $Incremental$ and a 24% reduction for *Incremental-Diversity*.

Although these results strongly suggest that there is a significant benefit to be gained from the more frequent use of compound critiques, it seems reasonable to ask if a sustained use of compound critiques will *always* deliver shorter sessions, or is there typically an *optimal usage frequency*, beyond which the benefits of compound critiques tend to decline? Intuitively the latter seems more plausible. Compound critiques provide users with an opportunity to make large jumps through the product-space, but they are not well suited to taking more fine-grained steps. We attempt to answer these questions in this section.

Figure 5 graphs the application frequency of compound critiques for different cycles, for our 4 recommender systems. That is, we count the percentage of sessions where compound critiques are used in the first, second, third cycles, and so on. The graph demonstrates how compound critiques are more likely to be applied during the early cycles of a session. We see that for all strategies compound critiques are used in at least 50% of the first 3 cycles of a session; for instance, for the *Standard-Diversity* strategy we see users picking compound critiques in 86% of their first cycles.

To answer the question about whether there is an optimal application frequency for compound critiques, in Figure 6, we graph the average session length for sessions with different application frequency ranges (0%-25%, 25%-50%, 50%-75%, 75%-100%), for each of the strategies. The data reveals a common pattern for each of the recommendation strategies, consistent with our results. Sessions in which compound critiques are used infrequently (less than 25% of the time) tend to be long, and more frequent use of compound critiques can produce significant reductions in average session length. This reduction appears to be available only up to a certain point. In each case we find the shortest sessions to occur in the 50%-75% application frequency range. Beyond this, session length begins to increase.

These results are consistent with the view that the compound critiques are best applied during the early stages of a recommendation session and that they should not be overused. Their value is derived from helping the user to quickly focus in on a suitable region of the product space, but unit critiques are better suited to the finer adjustments that are needed during the final stages of recommendation.

Other Related Work

Our research builds directly on work originally described by Burke *et al.* and implemented by the "FindMe" systems (Burke, Hammond, & Young 1996). Although this past work initially concentrated on the usefulness of *unit* critiques, the FindMe systems also recognised the navigation benefits associated with allowing users to manipulate multiple features simultaneously. Some FindMe systems presented the user with *compound critique* options. However they adopted a fixed interaction style, presenting a *static* set of critiques to the user during each cycle. There are problems with this approach. The static compound critiques may not be particularly relevant to the user or they may refer to critiques that do not satisfy any remaining product cases; problems that do not arise with our approach.

"Apt Decision" acts as an advisor to a consumer by inferring general preference orderings from a history of explicit *unit* critiques in response to specific examples (Sherin & Lieberman 2001). User preference profiles are collected implicity by the system and used in subsequent recommen-

dation sessions. This shares some similarities with the incremental critique approach except that the this system maintains persistent critique profiles for each user whereas our approach uses short-term session-based profiles.

Finally, the *example-critiquing* approach, implemented as the "SmartClient" tool (Faltings *et al.* 2004; Pu & Kumar 2004), proposes a more user-centric mode of interaction. Users can critique examples (e.q., travel plans) by directly indicating preferences for individual or combinations of features (e.g., travel time and price); essentially the user can build their own compound critiques during each cycle. The approach is also incremental in nature, in the sense that recently applied critiques influence subsequent retrievals. However, a key difference with our work is that we proactively select a set of diverse compound critiques for the user — rather than just allowing the user to select combinations of critiques — that are adapted to their current cycle.

Conclusions

Critiquing is an important user-feedback strategy in conversational recommender systems. Recently researchers have extended the standard approach to critiquing to accommodate the use of dynamic, compound critiques that facilitate feedback on multiple features simultaneously. Past results indicate that these extensions have the *potential* to deliver significant advantages when it comes to recommendation efficiency and quality, but previous results have been limited by their reliance on artificial user data and the lack of any significant comparative analysis in order to evaluate the relative merits of the different flavours of dynamic critiquing.

In this paper we address these issues by evaluating 4 different versions of dynamic critiquing through a large-scale user trial. The results show that incremental critiquing offers significant efficiency benefits over the standard form of critiquing, backing up artificial user studies presented by (McCarthy *et al.* 2004). However, they also show that the diversity-enhancements lead to reductions in recommendation efficiency, contradicting the former studies.

Interestingly, our analysis reveals a number of new results in relation to the manner in which users interact with the compound critiques that are generated by the dynamic critiquing techniques. We see that the diversity-enhancing methods produce compound critiques that are more likely to be selected by users, even though these selections do not favour recommendation efficiency. We also find that users seem to use compound critiques frequently during the early stages of a session, in order to focus in on the right general region of a product-space, but rarely in the latter stages of a session, where they prefer to use unit critiques to fine-tune their requirements on a feature-by-feature basis.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast Discovery of Association Rules in Large Databases. *Advances in Knowledge Discovery and Data Mining* 307–328.

Burke, R.; Hammond, K.; and Young, B. 1996. Knowledge-based Navigation of Complex Information Spaces. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 462–468. Portland, OR.

Faltings, B.; Pu, P.; Torrens, M.; and Viappiani, P. 2004. Designing Example-Critiquing Interaction. In *Proceedings of the International Conference on Intelligent User Interface(IUI-2004)*, 22–29. Funchal, Madeira, Portugal.

McCarthy, K.; Reilly, J.; McGinty, L.; and Smyth, B. 2004. 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)*, 176–184. Springer. Eindhoven, The Netherlands.

McCarthy, K.; Reilly, J.; McGinty, L.; and Smyth, B. 2005a. An Analysis of Critique Diversity in Case-Based Recommendation. In *Proceedings of the Eighteenth International FLAIRS Conference (FLAIRS-05)*. AAAI Press. Clearwater Beach, Florida, USA.

McCarthy, K.; Reilly, J.; McGinty, L.; and Smyth, B. 2005b. Experiments in Dynamic Critiquing. In Riedl, J.; Jameson, A.; Billsus, D.; and Lau, T., eds., *Proceedings of the International Conference on Intelligent User Interfaces (IUI'05)*, 175–182. ACM Press. San Diego, CA, USA.

McGinty, L., and Smyth, B. 2003. The Role of Diversity in Conversational Systems. In Bridge, D., and Ashley, K., eds., *Proceedings of the Fifth International Conference on Case-Based Reasoning (ICCBR-03)*, 276–290. Springer. Troindheim, Norway.

McSherry, D. 2003. Balancing User Satisfaction and Cognitive Load in Coverage-Optimised Retrieval. In Coenen, F.; Preece, A.; and Macintosh, A., eds., *Research and Development in Intelligent Systems XX. Proceedings of AI-2003*, 381–394. Springer-Verlag. Cambridge, UK.

Pu, P., and Kumar, P. 2004. Evaluating Example-based Search Tools. In *Proceedings of the ACM Conference on Electronic Commerce (EC'04)*, 208–217. New York, USA.

Reilly, J.; McCarthy, K.; McGinty, L.; and Smyth, B. 2004. Incremental Critiquing. In M. Bramer, F. C., and Allen, T., eds., *Research and Development in Intelligent Systems XXI. Proceedings of AI-2004*, 101–114. Cambridge, UK.

Sherin, S., and Lieberman, H. 2001. Intelligent Profiling by Example. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2001)*, 145–152. ACM Press. Santa Fe, NM, USA.

Shimazu, H. 2001. 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)*, 1443–1448. Seattle, WA, USA.

Smyth, B., and McClave, P. 2001. Similarity vs Diversity. In Aha, D., and Watson, I., eds., *Proceedings of the International Conference on Case-Based Reasoning*, 347–361.

Weld, D.; Anderson, C.; Domingos, P.; Etzioni, O.; Lau, T.; Gajos, K.; and Wolfman, S. 2003. Automatically Personalizing User Interfaces. In *Proceedings of the 18th International Joint Conference Artificial Intelligence (IJCAI-03)*, 1613–1619. Morgan Kaufman. Acapulco, Mexico.