Journal of the Association for Information Systems

JAIS

Research Article

Generating Effective Recommendations Using Viewing-Time Weighted Preferences for Attributes

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Paul Ralph
Lancaster University
paul@paulralph.name

Abstract

Recommender systems are an increasingly important technology and researchers have recently argued for incorporating different kinds of data to improve recommendation quality. This paper presents a novel approach to generating recommendations and evaluates its effectiveness. First, we review evidence that item viewing time can reveal user preferences for items. Second, we model item preference as a weighted function of preferences for item attributes. We then propose a method for generating recommendations based on these two propositions. The results of a laboratory evaluation show that the proposed approach generated estimated item ratings consistent with explicit item ratings and assigned high ratings to products that reflect revealed preferences of users. We conclude by discussing implications and identifying areas for future research.

Keywords: *Recommender System, Implementation, Psychology, Positivist, Design Science, Empirical, Experiment.*

* Choon Ling Sia was the accepting senior editor. This article was submitted on 11th May 2012 and went through three revisions.

Volume 15, Issue 8, pp. 484-513, August 2014

1. Introduction

A recommender system is an information system “that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options” (Burke, 2002, p. 6). Recommenders have become increasingly important in e-commerce because they can increase sales (Pathak, Garfinkel, Gopal, & Venkatesan, 2010; Schafer, Konstan, & Riedl, 2001), improve customer loyalty (Schafer et al., 2001), generate competitive advantage (Murthi & Sarkar, 2003), and reduce information overload (Liang, Lai, & Ku, 2007). The marketing power of recommenders is widely recognized (Gladwell, 1999; Vrooman, Riedl, & Konstan, 2002), and they are used commercially by online retailers including Amazon and Netflix. The popularity of the Netflix Prize competition (http://www.netflixprize.com/) to improve recommender accuracy exemplifies the intense level of interest among e-commerce vendors in improving recommendation quality.

Recommenders replace or augment other online navigation methods. In situations of information overload (cf. Edmunds & Morris, 2000), search engines and both expert- and user-generated taxonomies (Macgregor & McCulloch, 2006; Welty, 1998) encounter several problems. First, search engine effectiveness requires that the search terms appear in the information source. Second, expert taxonomies suffer from the difficulty in arriving at a single, correct set of classes for describing a particular domain (Sacco, 2006). Third, user-generated taxonomies require shared vocabularies (Mathes, 2004), which are difficult to guarantee when users independently tag resources. These limitations provide opportunities to improve navigation using recommenders.

Recommender effectiveness depends on recommendation accuracy, and considerable research attention has been given to designing and evaluating systems to generate accurate recommendations. As we describe below, existing recommenders typically rely on explicit indicators of preference for some items (e.g., ratings) or implicit indicators of preference (e.g., user profiles) to recommend new items. Such indicators are obtrusive and may require users to invest a considerable amount of time and effort before a system can make effective personalized recommendations.

In addition, some evidence suggests that accuracy improvements have stalled (Herlocker, Konstan, Terveen, & Riedl, 2004). One possible explanation for this is that the popular user-to-user comparison recommendation strategy is approaching its theoretical or practical limits. More generally, increasing the amount and diversity of both the information exploited and heuristics used by recommenders may produce greater accuracy improvements than refining strategies based on existing, information-starved heuristics (Bell, Koren, & Volinsky, 2007; Ralph & Parsons, 2006). This suggests that further progress may be made by identifying data that is theoretically linked to relevant constructs (e.g., preference, interest) and readily available to, but rarely used by, recommenders.

In view of these two issues, we sought a type of information that can be obtained unobtrusively, has been linked to preference by previous psychological research, and can be incorporated in an IT artifact to generate effective recommendations. The artifact addresses both the problem of obtrusiveness associated with existing artifacts and the call for increasing the diversity of information used to determine preferences and generate recommendations.

Specifically, one such type of information is viewing time. Viewing time is the period for which a user looks at an object or description of an object. Viewing time is theoretically interesting because numerous psychological studies have linked it to interest, preference, and related constructs in browsing and reading contexts (see Section 4); however, whether the viewing time / preference relationship in a shopping context is strong enough to inform recommendations remains unknown. Viewing time is practically interesting for recommender development because it is causally linked to interest and preference, readily calculable using client-side scripts, and rarely used directly by existing recommenders (see below).

Although viewing time may serve as an indicator of preference for items (and as a new source of information in recommenders) and can be collected and used unobtrusively in an online setting, it is not known whether the viewing time/preference relationship can be extracted from a “noisy” context in which other factors can influence viewing time and whether this extracted information can be exploited effectively to predict preferences for unseen (new) items. Therefore, we pose the following research question:

Research Question: *Is the relationship between viewing time and preference sufficiently robust that it can be incorporated in an artifact to recommend items that reflect user preferences?*

To address this question, in Section 2, we first review existing literature on recommender systems. We then examine specific challenges in recommender evaluation (Section 3) and review relevant psychological literature on viewing time (Section 4). In Section 5, we propose a recommendation heuristic that: 1) uses viewing times of seen items to estimate ratings of these items, 2) models item preference as a weighted function of preferences for item attributes, 3) uses the attribute preference / item preference relationship to rate unseen items, and 4) recommends highly rated unseen items. In Section 6 describes an experimental study to determine whether the proposed heuristic can leverage viewing time data to produce good recommendations. Next, we present out results, which indicate that viewing time data can be used to predict preferences and thereby generate good recommendations (Section 7). Finally, in Section 8, we discuss our study’s contributions, which are twofold. From a design perspective, we demonstrate that the psychological relationship between viewing time and preference may guide the design of an artifact to recommend items that match user preferences. From a practical perspective, we provide a recommendation heuristic that can incorporate new data into ensemble recommenders, especially in e-commerce contexts where obtrusive or collaborative recommenders are impractical.

2. Milestones in Recommender Systems Research

The first automated recommender system was Tapestry, which allowed users to rate emails and create queries based on other users’ ratings (Goldberg, Nichols, Oki, & Terry, 1992). This spawned a wave of development of standalone recommender systems, some collaborative (like Tapestry), others content-based.

Collaborative filtering systems “try to predict the utility of items for a particular user based on the items previously rated by other users” (Adomavicius, Sankaranarayanan, & Tuzhilin, 2005, p. 737). Using diverse methods of computing user similarity, collaborative systems more generally make predictions based on a user’s similarity to others. For instance, Konstan et al. (1997) compared users based on their explicit item ratings, while Mobasher, Dai, Luo, Sun, and Zhu (2000) compared users’ navigation patterns. Collaborative recommenders have been successful in academic environments (e.g., Mobasher, Dai, Luo, & Nakagawa, 2001; Shahabi & Chen, 2003; Shahabi, Banaei-Kashani, Chen, & McLeod, 2001) and commercial environments including Amazon.com and IMDb.com. However, such recommenders always assume that similar users have similar goals (Kohrs & Merialdo, 2001) and sometimes require users to rate items explicitly—an obtrusive and time-consuming task (Perkowitz & Etzioni, 2000). Some also require coincidence of ratings, such that performance degrades with sparse ratings data (Konstan et al., 1997; Sarwar et al., 1998). Despite methods proposed to overcome sparsity (e.g., Mobasher, Dai, Luo, & Nakagawa, 2002; Sarwar et al., 1998), lack of data remains problematic (Schafer, Frankowski, Herlocker, & Sen, 2007).

Content-based recommenders “recommend an item to a user based upon a description of the item and a profile of the user’s interests” (Pazzani & Billsus, 2007, p. 325). These vary on three primary dimensions: how items are represented, how user interests are represented. and how both representations are compared. Some systems model items as keyword/frequency vectors, model users as pseudo keyword/frequency vectors constructed from ratings, and then use the angle between user and item vectors as a similarity measure (Tai, Ren, & Kita, 2002; Zhao & Grosky, 2002). Others adopt a machine learning approach (Pazzani & Billsus, 1997) or allow users to navigate the item space directly (Burke, 1999, 2000). Content-based recommenders require sufficient information to both determine user preferences (Adomavicius et al., 2005) and differentiate liked and disliked items (Pazzani & Billsus, 1997).

The limitations of content-based and collaborative recommenders inspired hybrid recommenders, standalone systems that combine content-based and collaborative aspects (e.g., Basu, Hirsh, & Cohen, 1998). More generally, recognizing that results from heterogeneous recommender systems can be combined without degrading accuracy (Ralph & Parsons, 2006), more recent research has shifted toward ensemble approaches (Jahrer, Töscher, & Legenstein, 2010). An ensemble recommender combines the results of three or more predictors (recommendation heuristics) post hoc and is defined by the set of predictors and the method of blending their results. Ensemble recommenders have experienced much success (e.g., "BellKor's Pragmatic Chaos" ensemble recommender won the Netflix grand prize by combining 24 heuristics (Koren, 2009)).

In addition to the artifact construction research stream described above, a behavioral stream of research has emerged focusing on the interplay between recommender features, use, and effects on users' decision process and evaluations (Xiao & Benbasat, 2007). Among the key findings of this line of research are that use of recommenders increases decision quality while decreasing decision effort (Häubl & Murray, 2006; Pereira, 2001) and that these relationships are modified by characteristics of both the recommender and the product (e.g., product complexity) (see Xiao & Benbasat, 2007 for a summary). In addition, these studies have examined the effect of recommender characteristics on measures of trust, ease of use, perceived usefulness, and satisfaction, (e.g., Liang et al., 2007).

Furthermore, the construction and behavioral research streams have been supplemented by an evaluation stream, to which we turn in Section 3.

3. The Shifting Focus in Recommender Evaluation

A recommender may be evaluated by testing it against a neutral baseline (e.g., a null hypothesis) or a competing artifact (e.g., another recommender). Which is more appropriate depends on the type of recommender being studied. "Recommender" is commonly used to refer to three types of artifacts:

1. Heuristics: algorithms that predict user ratings on some dimension,

2. Ensembles: collections of heuristics blended to maximize cumulative predictive accuracy, and

3. Systems: applications that draw on a heuristic or ensemble to "guid[e] the user in a personalized way to interesting or useful objects in a large space of possible options" (Burke, 2002, p. 331).

In Section 3.1, we explain why it is more appropriate to evaluate heuristics (including our proposed heuristic) against a neutral baseline, and why it is more appropriate to evaluate ensembles (including the BellKor system described above) against a competing artifact.

3.1. Tiered Architecture in Recommender Design

In the 1990s, researchers often developed a heuristic (e.g., nearest neighbor collaborative filtering) and implemented a system that simply ran the heuristic and displayed the highest rated items (e.g., Konstan et al., 1997). Modern recommenders, however, exhibit a tiered architecture with a pool of heuristics on the bottom, a blending algorithm (ensemble) in the middle, and a graphical interface (recommender system) on top. Consequently, recommender research can be divided into these three tiers.

First, recommender heuristic researchers theorize about possible relationships a heuristic might exploit, design heuristics that exploit those relationships, and evaluate heuristic accuracy (e.g., Jin & Mobasher, 2003). When developing an incremental improvement of an existing heuristic, the new heuristic may be evaluated against the existing heuristic to quantify the accuracy improvement.

However, comparing a novel heuristic against a dissimilar existing heuristic is uninformative because the relative accuracy of dissimilar recommenders is irrelevant to their practical use. Therefore, novel heuristics may be compared against a neutral baseline representing the null hypothesis that the heuristic is ineffective and performs no better than a random recommender. Successful heuristics are not used directly; rather, they are made available for use in ensembles.

Second, recommender ensemble researchers iteratively select heuristics (from the pool of known heuristics) to maximize predictive accuracy in a specific domain (e.g., Jahrer et al., 2010). The results of diverse heuristics are blended such that adding a heuristic cannot reduce accuracy. This process (similar to a stepwise regression) illuminates why the relative accuracy of dissimilar recommenders is practically irrelevant: ensembles combine heuristics rather than choose between them. Here, evaluating against existing ensembles seems preferable—a new ensemble is innovative if it significantly outperforms the best available alternative ensemble for the domain of interest.

Third, recommender system researchers devise ways of displaying recommendations and (possibly) collecting data to improve diverse utility dimensions including ease of use, conversion rates, and consumer trust in recommendations (cf. Xiao & Benbasat, 2007). Recommender systems may be evaluated against existing systems or neutral baselines depending on the research question.

3.2. Systemic Challenges in Evaluating Recommender Heuristics

It is often assumed that novel heuristics should be evaluated by comparing them against existing heuristics. This section analyzes this assumption to convey its problems and examine the merit of evaluating novel heuristics using null hypothesis testing.

Suppose we have new heuristic d and an existing heuristic b. Further suppose d estimates “like” ratings (i.e., ratings of items on numerical like/dislike scales) from item viewing times and b estimates “like” ratings using similar users’ previous explicit “like” ratings (e.g., nearest neighbor collaborative filtering). A common method of testing d would involve comparing it to b experimentally: if d outperforms b ($d > b$), we would accept d as an innovation, while, if d fails to outperform b ($d \leq b$), we would reject d as non-innovative. This comparative logic is used in many recommender studies (e.g., Basu et al., 1998; Jin & Mobasher, 2003; Mobasher et al., 2001, 2002; Sarwar et al., 1998; Shahabi & Chen, 2003). For example, Jin and Mobasher (2003) compare a basic user-based collaborative filtering algorithm to one enhanced with a semantic similarity algorithm.

However, judging d by comparing it to b is problematic in at least three ways: 1) if b and d use different data, d may be useful in domains where b cannot be applied at all; 2) because recommender performance is domain-dependent (Herlocker et al., 2004), d may outperform b in some domains (e.g., books, movies) but not others (e.g., cameras, computers); and 3) practically speaking, the comparative performance of b and d is irrelevant because we can simply run both heuristics and blend their results to further increase accuracy.

With the advent of ensemble recommenders (Jahrer et al., 2010) and methods of combining the results of any set of heuristics such that adding heuristics cannot decrease overall accuracy (Ralph & Parsons, 2006), it is tempting to judge d by investigating whether adding it to an ensemble of existing heuristics improves overall accuracy. For example, suppose we have a pool of heuristics ($X=x_1, x_2,..., x_n$). Further suppose that previous research with these heuristics has determined the optimal ensemble for the domain of movie recommendations is $B = f(x_1, x_2, x_3)$, where f is the best identified blending function for these three predictors. To test a new heuristic, d, we append d to the baseline heuristic and empirically compare $D = g(x_1, x_2, x_3, d)$ against B, where g is the best identified blending function for these four predictors. Mimicking the logic of stepwise regression, if D significantly outperforms B, d is innovative, otherwise it is not.

However, judging d by comparing D to B is problematic in at least four ways. First, because not all heuristics can operate in all domains, d may be available in domains where x_1, x_2, or x_3 are not (e.g., if it uses different information to generate recommendations). In such domains, adding d to the remaining heuristics may produce significant accuracy gains. Second, because the relative performance of heuristics is domain dependent (Herlocker et al., 2004), D may outperform B in some domains but not others. Third, as the relative performance of D and B depends on the blending algorithm used, D may outperform B with some blends but not others. This is especially problematic when comparing D to a proprietary ensemble where the blending function may be unknown. Fourth, ensemble recommenders are often constructed using linear blending methods (Jahrer et al., 2010), which have several well-known problems including inflating the risk of overfitting models and an inability to guarantee optimality in the presence of redundant predictors (Judd & McClelland, 1989). Therefore, if we find that D outperforms B, d may be capturing only coincidental data features.

Additionally, the lack of large publicly available datasets in diverse domains (Herlocker et al., 2004) impedes cross-domain evaluation. Moreover, the lack of necessary data (e.g., viewing times) in these datasets impedes testing novel recommenders against ensembles of traditional predictors.

In summary, testing a new heuristic by comparing it to an existing heuristic is problematic because results are domain dependent and, in practice, recommenders blend numerous heuristics. Moreover, testing a new heuristic in comparative studies of ensemble recommenders is also problematic due to domain dependence, confounding effects of blending methods, and overfitting. The alternative is to evaluate heuristics against a neutral baseline (below).

3.3. Suggestions for Overcoming Challenges in Heuristic Evaluation

A recommender heuristic produces a set of estimated item ratings. The accuracy of these ratings can be evaluated using numerous measures of average error (lower being better). These error terms have no absolute meaning: one error term is only meaningful relative to other error terms. However, comparing a novel heuristic to an existing one may be misleading (above). Therefore, to test the hypothesis that heuristic d is accurate, researchers may construct a neutral baseline; that is, a set of estimated ratings that operationalizes the null hypothesis that d is inaccurate. The average error of d can then be compared to the average error of the baseline. Researchers may then evaluate the probability that the difference is due to chance using inferential statistics and the practical significance of the difference using measures of effect size.

Additionally, because heuristic performance is domain dependent, heuristics should be evaluated across several item classes. Moreover, the innovativeness of a heuristic should be judged based on how different it is from existing heuristics in terms of inputs, requirements, structure, and processing in addition to predictive accuracy (Bell et al., 2007).

In Section 4, we propose a novel source of information that can serve as a basis for generating recommendations.

4. Viewing Time as an Indicator of Preference

Bell et al. (2007) claim that “the success of an ensemble approach depends on the ability of its various predictors to expose different, complementing aspects of the data” (p. 6). One type of data readily estimable from weblogs and client-side scripts, but rarely utilized in recommender systems, is viewing time: the period for which a user looks at content associated with a particular item. Viewing time is interesting for several reasons.

First, viewing time data can be used in situations that are commonly problematic for existing content-based and collaborative recommenders. For example, some recommenders require: 1) users to explicitly rate items—an obtrusive and time-consuming task (Perkowitz & Etzioni, 2000); 2) extensive explicit or implicit ratings from other users (i.e., the cold-start problem) (Schafer et al., 2007); or 3) coincidence of ratings, such that performance degrades with sparse data (Konstan et al., 1997; Sarwar et al., 1998). In contrast, viewing time can be collected from user behavior automatically and unobtrusively. If recommendations can be generated directly from a user’s viewing time data (without comparing to other users), such a recommender could be used in situations that preclude many other recommendation approaches (e.g., by an electronic catalog that lacked sufficient purchase history for related collaborative recommendations). Furthermore, a recommender that uses a single session of viewing times may overcome problems with temporary interests (e.g., when one shops for a gift, preferences indicated by one's purchase history may not apply).

Second, a variety of psychological research shows a positive relationship between viewing time and preferences (or related constructs). Day (1966) reports that participants in a study looked longer at images rated "liked"; Faw and Nunnally (1967) found a positive correlation between "pleasant ratings" and viewing time; Oostendorp and Berlyne (1978) found that subjects viewed objects longer when they engendered pleasurable emotions.

Third, more recent work in the online context corroborates the viewing time / preference relationship. An online shopping simulation (Parsons, Ralph, & Gallagher, 2004) found a positive correlation between viewing time and items shoppers placed in carts. Time spent reading Usenet news was found to be positively related to explicit ratings (Konstan et al., 1997) and reader interest (Morita & Shinoda, 1994), as was webpage viewing time (Claypool, Le, & Brown, 2001). Several studies support a positive relationship between viewing time and relevance (Cooper & Chen, 2001; Miller, Riedl, & Konstan, 2003; Seo & Zhang, 2000).

Theoretically speaking, the causal relationship between preference and viewing time is complex. Preference is one of several possible antecedents of viewing time (cf. Heinrich, 1970), which Figure 1 summarizes. Other possible antecedents might attenuate the relationship between viewing time, and preference. However, to the extent viewing time is useful in predicting preference, the relationship is robust to such attenuation.

However, people tend to express unwarranted preference for more familiar items—a psychological phenomenon known as the mere exposure effect or familiarity principle (Bornstein & Carver-Lemley, 2004). Therefore, a user's preference for an item may cause longer viewing times, which may, in turn, increase the user's preference over time. Practically speaking, however, because prediction relies on correlation rather than causation, a substantial covariance of preference and viewing time may be useful for generating recommendation regardless of causal direction. This research examines the extent to which this relationship can be used to infer preferences based on viewing time and recommends items that best match these inferences.

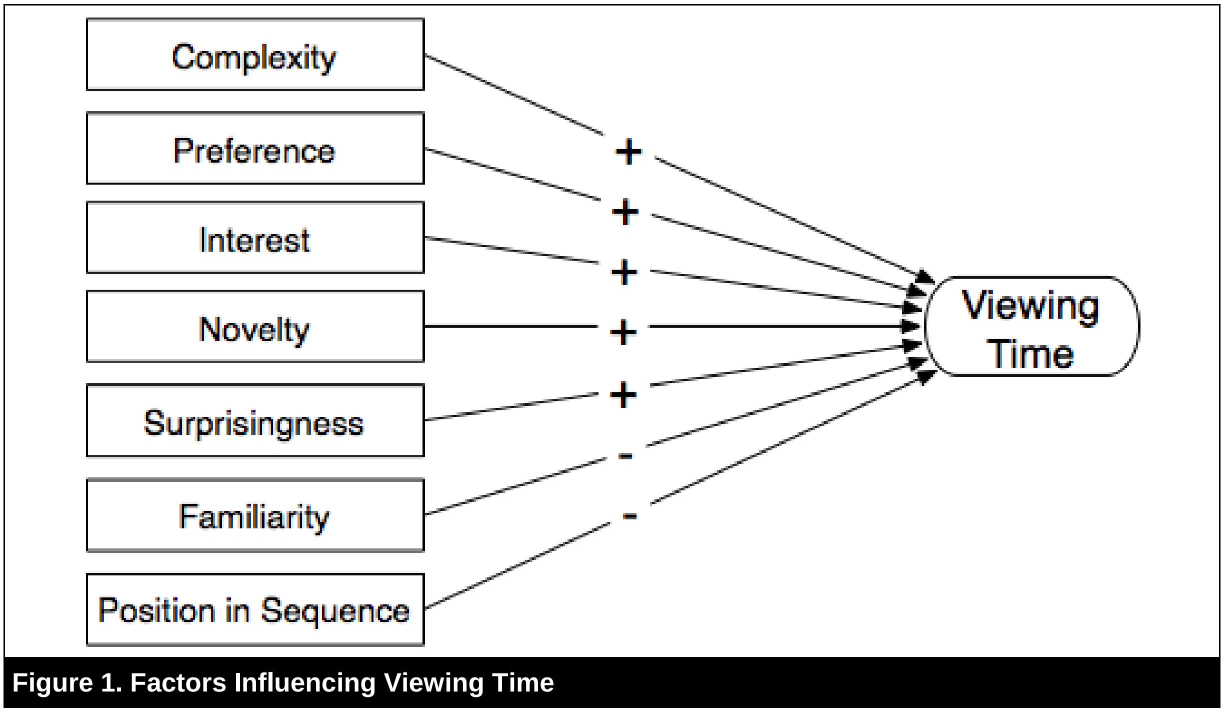


Figure 1. Factors Influencing Viewing Time

Viewing time has already been used in some recommenders, for example, as part of the user similarity calculations in collaborative filtering (e.g., Lee, Park, & Park, 2008; Mobasher et al., 2001) and information filtering (cf. Oard & Kim, 2001). However, we are not aware of any recommender research on directly extracting user preferences for items from viewing time data. Given the psychological basis for hypothesizing that useful preference information can be extracted from viewing time data, we created a content-based recommender to exploit this relationship.

5. The Desire Recommender System

DESIRE (Desirability Estimator and Structured Information Recommendation Engine) is a content-based recommender system for predicting user preferences for unseen items in a catalog based on time spent browsing a small set of items from the catalog. The DESIRE algorithm is presented in the Appendix (for a complete technical exposition see Parsons and Ralph, 2010); this section overviews the general strategy used to generate recommendations. DESIRE is comprised of three components:

1. The user rating estimator computes a user's implicit ratings for seen items from the user's viewing time data

2. The user profile generator formulates a user's expected preferences based on the implicit ratings and the attributes of seen items

3. The recommendation engine predicts ratings for unseen items based on their attributes and the user's profile

While many content-based recommenders share a similar structure, DESIRE employs unique methods of generating the user profile and estimating ratings.

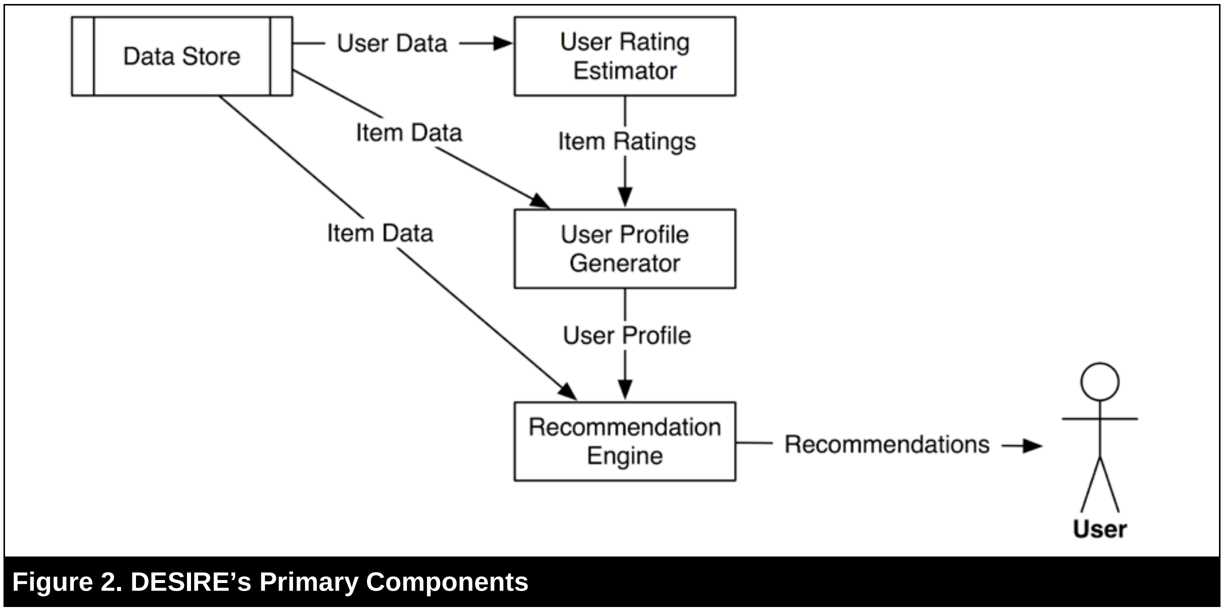


Figure 2. DESIRE's Primary Components

5.1. Algorithm Overview

5.1.1. Rating Estimator

The rating estimator converts a list of item/viewing time pairs into a list of item/rating pairs by calculating z-scores for the viewing times and normalizing them to a [-1,1] range. This conversion reduces the impact of outliers. The normalized viewing time is then used as the user's implicit rating of the seen item.

5.1.2. User Profile Generator

The user profile generator describes a user's modeled preferences in terms of the desirability of particular item characteristics, which is consistent with the additive value model from multi-attribute utility theory (Fishburn, 1970; Keeney, 1968) and conjoint analysis in the marketing literature (Green & Srinivasan, 1978, 1990). It converts item attribute data and implicit ratings into inferred attribute ratings. It uses different approaches for nominal or ordinal data (categorical attributes) and ratio or interval data (numeric attributes).

For each categorical attribute (e.g., color), the user profile generator will first construct a list of every value (e.g., red, blue, yellow) in the list of seen items. It then assigns a rating to each value equal to the mean rating of each seen item having that value. For instance, if the user has viewed two red bicycles, and the rating estimator estimates a ratio of 0.5 and -0.3 for those bicycles, the value "red" would be given a rating of 0.1. In this way, each value of each attribute of each seen item is given an estimated rating.

For each numeric attribute (e.g., price), the user profile generator estimates the user's ideal quantity for that attribute as a weighted average of values for that attribute in each liked item, where the weights are the item ratings. For example, given four items with prices $2, $4, $6, and $10, and ratings -0.4, 0.3, 0.8, and 0.2, respectively, the ideal price would be (0.3*4 + 0.8*6 + 0.2*10)/(0.3+0.8+0.2), or $6.15. Here, a liked item is a seen item with a positive rating. DESIRE ignores seen items with negative ratings to avoid biasing ideal value calculation. For example, if a user views a large number of more expensive items for a short period and a small number of less expensive items for a long period, including disliked items will inflate the ideal price estimate. While more sophisticated approaches are possible, ignoring disliked items for initial studies seemed reasonable.

In summary, a user profile consists of a set of ideal quantities for each numeric attribute and a set of ratings for each value of each categorical attribute.

5.1.3. Recommendation Engine

Given a user profile, the recommendation engine predicts the ratings the user would give to a set of unseen items. For each item, DESIRE must first compute an attribute/rating vector—a set of (attribute/rating) pairs where the rating indicates the similarity between the item and the user profile with respect to the attribute.

For categorical attributes having a single value (e.g., color = red) the similarity rating is equal to the inferred rating from the user profile. If the user profile has no rating for a value of categorical attribute in an item, that value (or that attribute if it has only one value) is omitted. Categorical attributes having multiple values may be handled in several ways. For example, suppose a user profile has color value ratings of 0.4 for black and 0.6 for brown. The color similarity rating for a pair of brown and black boots could be calculated using the mean (0.5), the minimum value (0.4), or the maximum value (0.6). For the purposes of this study, we used the mean.

Numeric attributes are assumed to have only one value. By examining the population of both seen and unseen items, z-scores are calculated for both the ideal quantity (from the profile) and the item values for each numeric attribute. The similarity rating for a particular attribute is then calculated as the absolute value of the difference between the z-score of the item value and z-score of the profile value.

The similarity ratings for both categorical and numeric attributes are then transformed to the range [-1,1]. This results in a set of attribute/rating pairs with each rating having the same [-1,1] range. A single value to represent the similarity between the item and the user profile can then be calculated as the mean of the ratings. However, not all attributes are equally important; therefore, a weighted average, where the weights indicate the relative importance of each attribute, is more appropriate. These weights would obviously vary by product category. They can be determined a priori using the procedure described in Section 6.1.

This process is repeated for each item. Once DESIRE has predicted ratings for all items, recommendations can be made in one of two ways. DESIRE can generate a recommendation set of a particular size or recommend all items exceeding a particular “recommendation threshold”. In both cases, the predicted ratings form the basis for including items in the recommendation set (i.e., the recommended items are those with the highest predicted rating). In either case, DESIRE’s recommendations may be combined with the results of other recommender heuristics in an ensemble system.

5.2. Conceptual Evaluation

DESIRE has several desirable properties relative to most existing recommender systems. First, it is unobtrusive—the recommendation set is generated without the user being aware of, or interacting with, the recommender system. For example, the user does not have to rate items in order to receive recommendations. Second, DESIRE maintains user independence—recommendations for user U do not require information about users other than U. In contrast, systems that require users to rate items explicitly and compare ratings from different users are neither transparent nor user independent. Third, DESIRE’s scalability is superior to nearest neighbor collaborative filtering (CF) algorithms because DESIRE’s complexity is linear in the number of items, while CF depends on the numbers of both items and users. Fourth, DESIRE can operate either on a single session’s browsing data (overcoming the gift-shopping problem mentioned above) or on a user’s entire history. Fifth, DESIRE consists of loosely coupled modules in the sense that its three components (rating estimator, profile generator and recommendation engine) can be individually replaced. For example, in a context where users are willing to explicitly rate items, the profile generator could use the explicit ratings directly.

However, DESIRE also has three primary limitations. First, DESIRE is only applicable where attribute data, the relative importance of item attributes, and user viewing times are available. Second, DESIRE assumes that the available attribute data are related to items’ value dimensions. Thus, if the available attribute data are based on objective dimensions such as price and size, but user preferences are based on intangible or qualitative dimensions such as fashionability, DESIRE is not expected to work well. We return to this issue in the empirical evaluation. Third, DESIRE is best suited for exploratory search contexts, especially hedonic browsing, rather than directed search contexts (cf. Hong, Tong, & Tam, 2005; Moe, 2003). Moe’s (2003) analysis of data from a nutrition products website found that hedonic browsing accounted for 66 percent of (non-shallow) website visits. In contrast, DESIRE would likely be ineffective in modeling a user who adopts satisficing search strategy (Simon, 1956) and a “single criterion” stopping rule (Browne, Pitts, & Wetherbe, 2007); that is, a user who evaluates items on a single dimension and stops when the first satisfactory item is found.

6. Empirical Evaluation of DESIRE Recommendations

Because DESIRE is a recommender heuristic, we designed a lab study to empirically evaluate it against a neutral baseline across several item classes. Accomplishing this requires a variety of data:

1. A data store of items and their attributes

2. The relative importance of each attribute

3. A data store of viewing time triples (user, item, viewing time), and

4. A data store of explicit rating triples (user, item, rating).

We therefore discuss the empirical evaluation in several steps. First, we describe the development of the item data store and relative attribute importance index using two pre-studies. Second, we present our hypotheses. Third, we describe the shopping simulation used to construct the viewing time data store. Fourth, we provide details of the explicit ratings exercise, from which the explicit ratings data store is constructed.

6.1. Pre-Studies

In the first pre-study, we asked a convenience sample of students in an MBA class at a mid-sized Canadian university: “List all the factors you take into account when buying from each product category listed below”. The categories were bicycles, boots, digital cameras, digital music players, DVD movies, notebook computers, winter coats, video games, jewelry, and winter gloves. We compiled the responses, during which we eliminated repeated responses and combined similar responses. We then investigated the extent to which readily available data matched the factors listed for each item class. Based on attribute data availability, we chose five categories: bicycles, boots, digital cameras, digital music players, and notebook computers. Table 1 lists the factors identified by participants for each of these categories.

Table 1. Product Characteristics Identified by Pre-study Participants

Bicycles	brake types*, brand*, color, comfort, frame material*, number of seats*, price*, suspension*, frame size*, speeds/gears*, tires*, tire size*, testimonials, type (e.g., mountain)*, warranty coverage, warranty term*, weight*
Boots	brand*, comfort, color*, fashionability, functionality, material*, maintenance required, price*, purpose*, sole type*, style, warmth, waterproofing*, weight
Digital cameras	ac adapter*, accessories*, appearance*, batteries included*, battery type*, brand*, card reader*, charger type, color*, digital zoom*, display type*, ease of connection to computer, features, internal memory*, macro lens*, maximum memory expansion*, resolution*, memory card included*, optical zoom*, price*, quality settings*, recharge time, service, size*, sound capability*, style, type of memory card*, warranty*, video capability*
Notebook computers	battery life*, brand*, capabilities, color*, compatibility, display size*, display quality, display resolution*, drives*, memory*, memory speed*, platform (mac/pc)*, ports*, price*, processor speed*, service, size*, weight*, warranty*
Digital music players	anti-skip protection, battery life*, battery type*, brand*, accessories*, expandable, features, formats played*, headphones*, max expanded memory, memory*, portable hard drive capability*, price*, recording capability*, sound quality*

* Attributes used in the subsequent study; others were dropped due to lack of data.

We then gathered attribute data corresponding to the factors identified by the participants. Some factors, such as “ease of connection to computer” for digital cameras, were eliminated due to lack of available data. This produced the item data store. We divided the item data store into a training set, used in the shopping simulation (see Section 6.3), and a holdout set, used for explicit item rating (see Section 6.4).

To determine the relative importance of these attributes, simply asking users to rank or rate each attribute’s influence on their preferences would be ineffective because self-reports of the relative importance of factors are poorly correlated with relative importance revealed implicitly through regression analysis (Fishbein & Ajzen, 1975). Therefore, we recruited a second convenience sample of 14 undergraduate business students from the same university. Participants viewed a series of products from each category and, for each product, rated each attribute and the product overall on a nine point scale from unsatisfactory to satisfactory. We performed a stepwise multiple regression on the results to determine which attributes predicted overall ratings for each product category. We then adopted the coefficients of the significant variables in the regression equations, shown in Table 2, as DESIRE’s relative attribute importance weights.

Table 2. Relative Importance of Item Attributes

Bicycles	Type of brakes	0.383
	Tire size	0.204
	Number of gears or speeds	0.173
	Warranty	0.165
Boots	Brand	0.493
	Price	0.308
	Purpose	0.191
Digital cameras	Size	0.372
	Price	0.242
	Digital zoom	0.214
	Brand	0.195
	Accessories	0.157
	Warranty	0.131
Digital music players	Price	0.348
	Brand	0.310
	Audio quality	0.198
	Accessories	0.152
Notebook computers	Battery life	0.354
	Platform	0.272
	CPU	0.226
	Brand	0.206
	Price	0.172
	Maximum screen resolution	-0.249

6.2. Hypotheses

We first hypothesize that DESIRE will predict explicit item ratings more accurately than a recommender that produces random recommendations. Operationalizing this hypothesis requires an accuracy measure and a procedure for generating random recommendations.

Herlocker et al. (2004) review several methods of computing recommendation accuracy. Two of the most common measures are mean absolute error (MAE) and root mean square error (RMSE), which we define as:

$$\text{MAE} = \sum \frac{|\,\textit{actual rating} - \textit{predicted rating}\,|}{\textit{number of ratings}} \qquad \text{RMSE} = \sqrt{\sum \frac{(\textit{actual rating} - \textit{predicted rating})^2}{\textit{number of ratings}}}\ .$$

MAE and RMSE measure the discrepancy between a set of predictions and a set of observations. An MAE (or RMSE) of zero indicates perfect accuracy. However, error terms other than zero are difficult to assess in isolation: they are most meaningful when compared with the MAE (or RMSE) of another set of predictions. RMSE penalizes larger errors more severely than MAE. However, because MAE is more commonly used and "has well studied statistical properties that provide for testing the significance of a difference between the mean absolute errors of two systems" (Herlocker et al., 2004, p. 21), we focus on this measure in the following analysis.

Theoretically, if item viewing times and user preferences are not related, DESIRE will predict ratings no more accurately than a random rating generator (RANDOM). However, recommendations may be randomly generated using many different distributions. For comparison purposes, we adopt two interpretations of "random". First, we can compare DESIRE's ratings to ratings generated on a uniform distribution. Second, we can compare DESIRE's ratings to ratings generated randomly from the distribution of user ratings obtained during the item rating phase (Section 6.4). While it is possible to evaluate DESIRE using random recommendations conforming to any number of other distributions, using the distribution derived from actual rating data provided the most conservative test of DESIRE's accuracy.

Consequently, Hypothesis 1 may be operationalized as follows:

H1: *MAE(DESIRE) < MAE(RANDOM)*.

Additionally, we can compare DESIRE ratings with users' revealed preferences. We argue that, by placing items in their virtual shopping baskets, online shoppers explicitly reveal their preference for these items. Although in practice shoppers do not always buy these "basket items" (e.g., shopping baskets may be abandoned), adding an item to a basket indicates preference for the item compared to items viewed but not added to the basket. Therefore, we expect DESIRE to rate these "basket items" higher than items not placed into the shopping basket ("non-basket items"). Because the shopping simulation (Section 6.3) includes a shopping basket feature, we can operationalize Hypothesis 2 as follows:

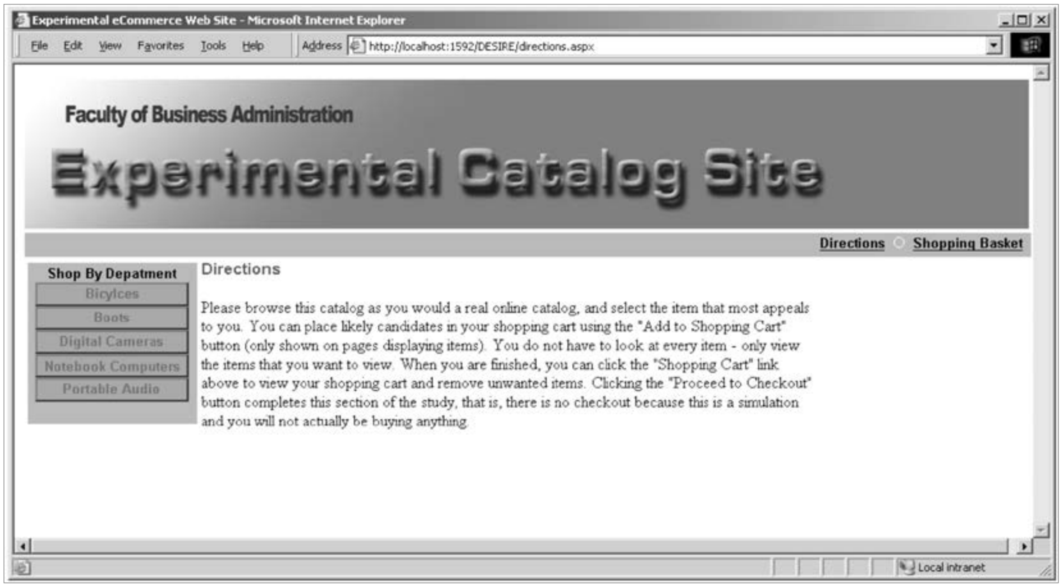
H2: *DESIRE's ratings of basket items will be higher than its ratings of non-basket items*.

Hypotheses 1 and 2 address predictive accuracy rather than recommendation quality per se. We follow the common assumption in recommender evaluation that predictive accuracy is the primary determinant of recommendation quality (cf. Herlocker et al. 2004),.

6.3. Shopping Simulation

We recruited a convenience sample of 67 participants from an undergraduate business course. We encouraged participation by giving students the option of participating in the study or completing another task of equivalent time commitment to receive a three percent bonus on their course grade. These students took part in a laboratory study in small groups of 10 to 20. Each subject participated individually at a pre-configured computer.

We told the participants that the study's purpose was to improve our understanding of electronic commerce and online catalogs. After reviewing the study purpose and agreeing to take part, participants viewed the study directions online (Figure 3). The directions were embedded in the shopping simulation website such that participants could return to them at any time. In short, participants were asked to pretend that they were shopping, look at whatever they wanted to for however long they wanted to, and add items they wanted to "buy" to their shopping baskets.


Experimental eCommerce Web Site - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:1592/DESIRE/directions.aspx
Faculty of Business Administration
Experimental Catalog Site
Directions
Shopping Basket
Shop By Depatment
Bicylces
Boots
Digital Cameras
Notebook Computers
Portable Audio
Local intranet

Directions

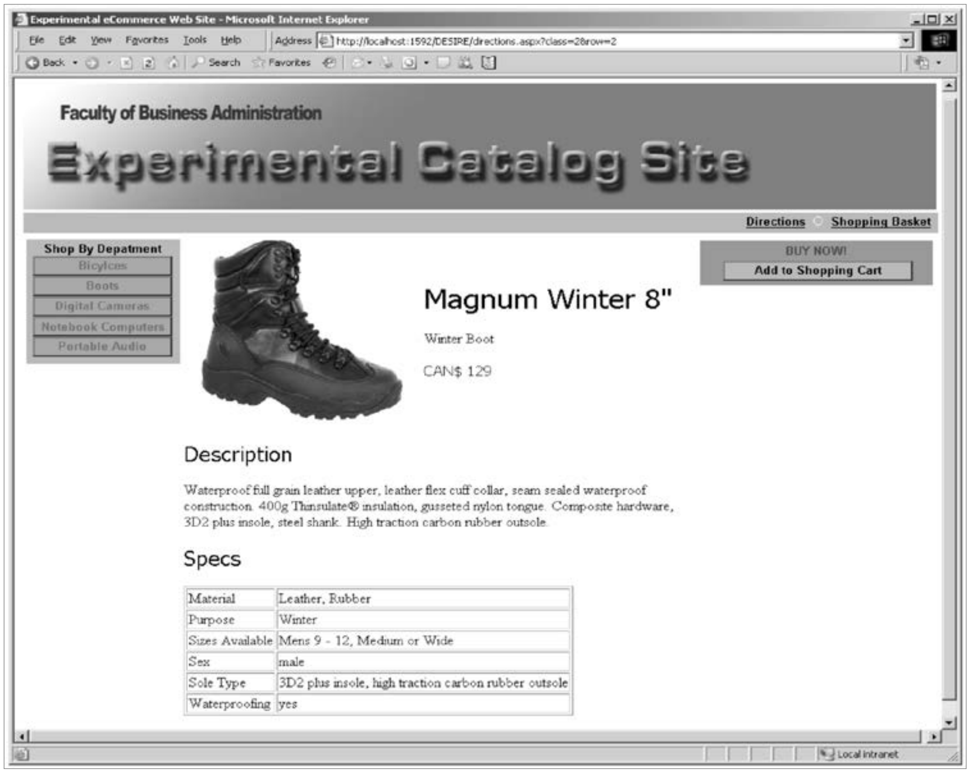
Please browse this catalog as you would a real online catalog, and select the item that most appeals to you. You can place likely candidates in your shopping cart using the "Add to Shopping Cart" button (only shown on pages displaying items). You do not have to look at every item - only view the items that you want to view. When you are finished, you can click the "Shopping Cart" link above to view your shopping cart and remove unwanted items. Clicking the "Proceed to Checkout" button completes this section of the study, that is, there is no checkout because this is a simulation and you will not actually be buying anything.

Figure 3. Front Page of Simulated Catalog

Participants then took part in a simulated shopping exercise. Participants could view as many or as few item pages as they chose for as little or as much time as they wished. Participants could return to a previously-viewed item page—of 3669 item page views in total, 428 (11.7%) were repeats. This low repeat rate is consistent with browsing (rather than searching) behavior.

Item pages were generated from the datastore (training set) created during the pretests. Figure 4 depicts a sample item page. Item pages contained all of the italicized attributes indicated in Table 1, not just the attributes used by DESIRE. All item pages in each product category were as similar as possible, with exactly the same attributes listed and exactly one picture. The system recorded the time each participant spent viewing each item page. Participants could navigate among the available items either by browsing “departments”; that is, listings by product category (Figure 5) or by keyword search. Participants could add any number of items to their shopping basket, which could be viewed at any time. Looking at the shopping basket (Figure 6), participants could remove items, “proceed to checkout”, or continue shopping. Clicking the “proceed to checkout” button ended the simulation, consistent with the directions provided.

This phase produced the data store of viewing time triples. Where a participant viewed the same item page repeatedly, total viewing time was used. No recommendations were generated or displayed during the shopping simulation.


Experimental eCommerce Web Site - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:1592/DESIRE/directions.aspx?class=2&row=2
Back Search Favorites
Faculty of Business Administration
Experimental Catalog Site
Directions
Shopping Basket
Shop By Depatment
Bicylces
Boots
Digital Cameras
Notebook Computers
Portable Audio
BUY NOW!
Add to Shopping Cart
Local intranet

Magnum Winter 8"

Winter Boot

CAN$ 129

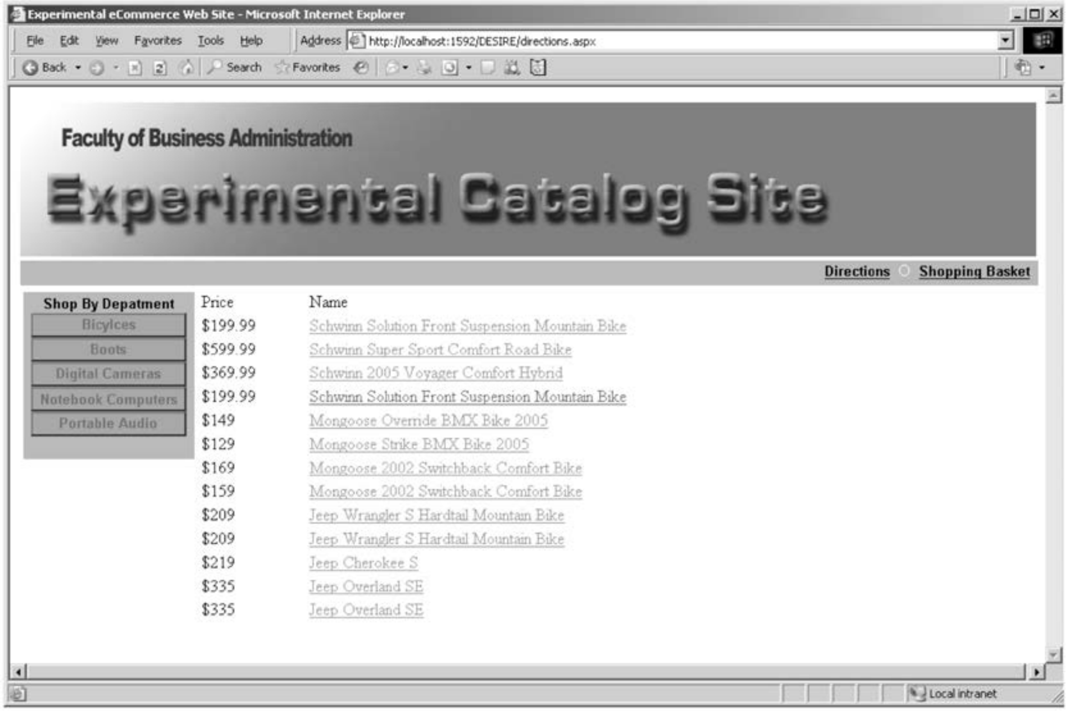
Description

Waterproof full grain leather upper, leather flex cuff collar, seam sealed waterproof construction. 400g Thinsulate® insulation, gusseted nylon tongue. Composite hardware, 3D2 plus insole, steel shank. High traction carbon rubber outsole.

Specs

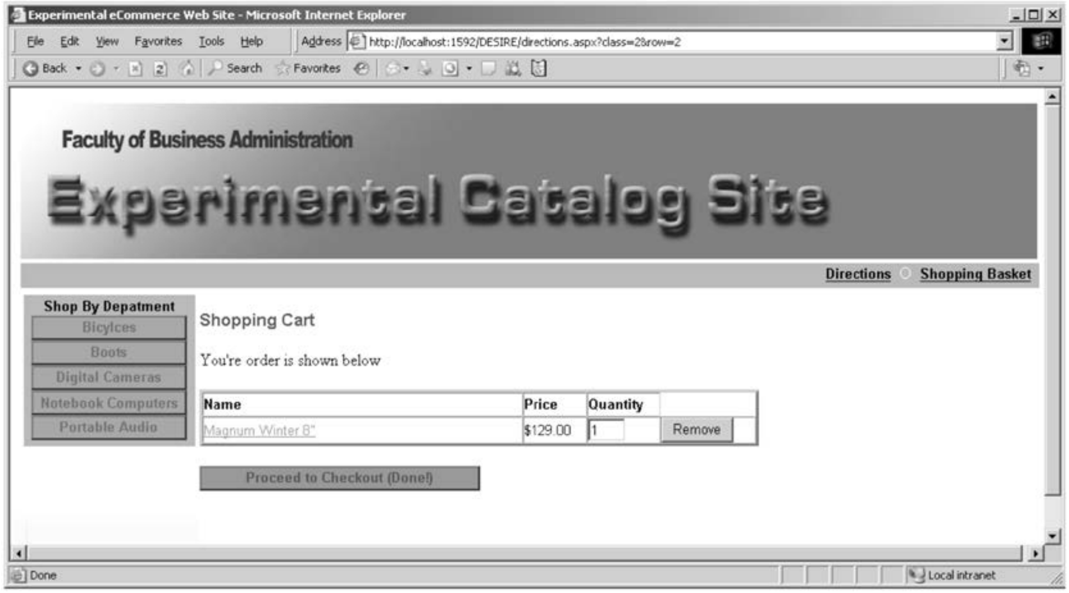
Material	Leather, Rubber
Purpose	Winter
Sizes Available	Mens 9 - 12, Medium or Wide
Sex	male
Sole Type	3D2 plus insole, high traction carbon rubber outsole
Waterproofing	yes

Figure 4. Sample Item Page


Experimental eCommerce Web Site - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:1592/DESIRE/directions.aspx
Back Search Favorites
Faculty of Business Administration
Experimental Catalog Site
Directions
Shopping Basket
Shop By Depatment
Bicylces
Boots
Digital Cameras
Notebook Computers
Portable Audio
Local intranet

Price	Name
$199.99	Schwinn Solution Front Suspension Mountain Bike
$599.99	Schwinn Super Sport Comfort Road Bike
$369.99	Schwinn 2005 Voyager Comfort Hybrid
$199.99	Schwinn Solution Front Suspension Mountain Bike
$149	Mongoose Override BMX Bike 2005
$129	Mongoose Strike BMX Bike 2005
$169	Mongoose 2002 Switchback Comfort Bike
$159	Mongoose 2002 Switchback Comfort Bike
$209	Jeep Wrangler S Hardtail Mountain Bike
$209	Jeep Wrangler S Hardtail Mountain Bike
$219	Jeep Cherokee S
$335	Jeep Overland SE
$335	Jeep Overland SE

Figure 5. Sample Product Listing by Category


Experimental eCommerce Web Site - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:1592/DESIRE/directions.aspx?class=2&row=2
Back Search Favorites
Faculty of Business Administration
Experimental Catalog Site
Directions
Shopping Basket
Shop By Depatment
Bicylces
Boots
Digital Cameras
Notebook Computers
Portable Audio
Remove
Proceed to Checkout (Done!)
Done
Local intranet

Shopping Cart

You're order is shown below

Name	Price	Quantity	
Magnum Winter 8"	$129.00	1	Remove

Figure 6. Sample Shopping Cart

6.4. Explicit Item Rating

After completing the shopping simulation, participants were asked to rate a series of items they had not previously seen (the holdout set, which was randomly selected from the set of all catalog items). Participants were shown an item page similar to the item pages in the shopping simulation but with a nine-point rating scale. Up to ten items were presented per product category; however, if a participant did not view any items from one or more categories in the shopping simulation, that participant was not asked to rate items from those categories. During this phase, an explicit ratings data store comprising 930 explicit product ratings was constructed. No recommendations were generated or displayed during the explicit item rating phase.

6.5. DESIRE Computation

With data collection complete, the DESIRE algorithm (Appendix A) was run. Input data included the item data store (including attribute data), the relative attribute importance index, and the item viewing times from the shopping simulation (but not the explicit item rating page). This produced a set of estimated ratings for holdout-set items in each product category in which the participant had viewed items during the simulation. We could then compute DESIRE's accuracy by comparing DESIRE's estimated ratings of holdout items to users' explicit ratings of the same holdout items.

7. Results and Discussion

7.1. Comparing DESIRE to RANDOM Ratings

To test Hypothesis 1 (that DESIRE ratings are better than RANDOM ratings), we compared the MAE of DESIRE's ratings (versus the user-generated ratings described above) by product category to the MAE of RANDOM ratings.

Tables 3 and 4 contain the results of independent-sample t-tests comparing DESIRE's ratings with ratings generated from a uniform distribution and the observed distribution (from the product rating phase), respectively. The analysis in Table 3 demonstrates that the difference between DESIRE ratings and random (uniform) ratings was highly significant across all five product categories (using a Bonferroni-adjusted significance level of 0.01), with DESIRE performing substantially better than a system that generates ratings randomly. In addition, the effect size ranged from moderate to high in all categories except boots (the rightmost column of Table 3 lists the Cohen's-d statistic for each category tested, where we interpret 0.8 as high, 0.5 as moderate, and 0.2 as low).

Table 3. T-test Results of DESIRE versus RANDOM (Uniform Distribution) Recommendations

Category		DESIRE		UNIFORM				
	N	MAE	SD	MAE	SD	t	p	Cohen's d
Notebook computers	182	1.52	1.12	2.77	2.12	7.0442	<0.0001	0.74
Digital cameras	182	1.69	1.34	2.68	2.10	5.3398	<0.0001	0.56
Bicycles	185	1.74	1.39	3.12	2.02	7.6437	<0.0001	0.80
Digital music players	127	2.13	1.22	2.96	2.21	3.6846	0.0003	0.46
Boots	254	2.42	1.71	2.96	2.04	3.2512	0.0012	0.29

Table 4. T-test Results of DESIRE versus Random (Distribution from Data) Recommendations

Category		DESIRE		OBSERVED				
	N	MAE	SD	MAE	SD	t	P	Cohen's d
Notebook computers	182	1.52	1.12	2.37	1.97	5.1121	<0.0001	0.53
Digital cameras	182	1.69	1.34	2.29	1.82	3.5433	<0.0004	0.38
Bicycles	185	1.74	1.39	2.61	2.13	5.6557	<0.0001	0.48
Digital music players	127	2.13	1.22	2.23	1.78	0.4928	n.s.	n/a
Boots	254	2.42	1.71	3.07	2.24	3.6870	0.0003	0.32

Table 4 shows that the improvement of DESIRE ratings compared to random ratings (based on observed item ratings) was highly significant across four of the five product categories (using a Bonferroni-adjusted significance level of .01), with DESIRE performing better than a system that generates ratings randomly, except for the digital music players category. The effect sizes were moderate in all four categories.

7.2. Comparing DESIRE Rating of Basket and Non-basket Items

To test Hypothesis 2 (that DESIRE ratings of basket items will be higher than DESIRE ratings of non-basket items), we examined DESIRE's ratings only for the items that participants placed in shopping baskets because these items reflect participants' preferences as revealed by their simulated shopping choices. We expect user ratings of these items to be high (note that users were not asked to rate these items). The average DESIRE rating of shopping basket items on a scale from -1 to 1 was positive (mean = 0.32, $p < 0.01$), and all but two of the DESIRE ratings were positive. If we exclude boots—because the value dimensions for a fashion product such as boots may not match availability attribute data—the average DESIRE rating of items from the remaining categories was 0.38. Interestingly, the lowest eight ratings by DESIRE for shopping basket items were for "boots", which suggests that the system is limited in predicting user preferences for product categories that appear to be highly influenced by style rather than by measurable product attributes.

Table 5 compares the average DESIRE rating of shopping basket items to the average DESIRE rating of non-basket items across product categories. DESIRE ratings of basket items (reflecting revealed preferences of participants) were significantly higher than ratings of non-basket items for all five product categories (using a Bonferroni-adjusted significance level of .01). The effect size was large for all product categories. Collectively, these results provide strong support for H2: DESIRE did provide higher ratings for products that were preferred by participants than for items that were not preferred by participants.

Table 5. T-test Results of Basket versus Non-Basket DESIRE Ratings

Category	Mean basket (N)	SD basket	Mean non-basket (N)	SD non-basket	t	p	Cohen's d
Bicycles	0.434 (22)	0.157	0.252 (185)	0.221	3.748	0.0002	0.85
Boots	0.186 (20)	0.198	-0.172 (254)	0.318	4.955	<0.0001	1.15
Digital cameras	0.346 (6)	0.047	0.247 (182)	0.095	2.538	0.01	1.05
Notebook computers	0.361 (14)	0.114	0.253 (182)	0.169	2.347	0.01	0.65
Digital music players	0.312 (6)	0.164	0.045 (127)	0.186	3.451	0.0008	1.44

7.3. Comparing DESIRE Ratings Between Product Categories

As we describe in Section 5, DESIRE estimates user preference for an item based on the item's attributes. This assumes the attributes for which data are available correspond to the value dimensions of the item (Keeney & Raiffa, 1976)—the item characteristics that primarily determine preferences. For example, if purchase decisions for notebook computers are based entirely on color, keyboard comfort, and screen glare, whereas the only data available concern price, processor speed, and memory capacity, and the latter are unrelated to the former, DESIRE is unlikely to infer preferences accurately. Therefore, the quality of DESIRE's predictions should be positively related to the extent to which the available item data coincides with the value dimensions of the items.

To explore whether DESIRE rating quality will be higher for product categories for which preferences are based on quantifiable and easily measurable attributes, we conducted a one-way ANOVA comparing the MAEs of the five product categories. Because the pretest showed that boot preferences were significantly influenced by criteria that were not quantified or available for use by DESIRE (e.g., “comfort” and “style”), we expect that the MAE will be higher for boots than for the other product categories. To begin, we conducted an overall test for differences in MAE among the five product categories. The MAE differed significantly across the categories: F (4, 925) = 14.48, p < .001.

To determine whether the mean for boots was higher than for other categories, we conducted a post hoc test to pair-wise compare the means of the product categories. A Levene's test for homogeneity of variance showed that the variances of the groups differed (Levene (4, 925) = 13.435, p < .001). Therefore, we used Tamhane's test for differences in the group means. Table 6 shows the comparison of the MAE of boots with each of the other product categories. As we can see, the MAE of boots was significantly higher than that of bicycles, digital cameras, and notebook computers. While the MAE for boots was higher than for digital music players, the difference was not significant.

Table 6. Post Hoc Comparison of MAE of Boots With Other Product Categories

Category	Mean difference (boots category)	Std. error	p^1
Bicycles	0.681	0.148	< .001
Digital cameras	0.729	0.146	< .001
Notebook computers	0.905	0.135	< .001
Digital music players	0.287	0.153	0.465

[1]p-values are for Tamhane's test. Similar results hold for other tests.

To further explore the results for digital music players, Table 7 reports the comparison of the MAE of this category with the others. As we can see, the MAE for digital music players did not differ significantly from that of boots, but was significantly higher than the MAE for digital cameras and for notebooks computers, and was marginally higher than the MAE for bicycles.

Table 7. Post Hoc Comparison of MAE of Digital Music Players to Other Product Categories

Category	Mean difference (players category)	Std. error	p^1
Bicycles	0.393	0.149	0.084
Boots	-0.287	0.153	0.465
Digital cameras	0.442	0.147	0.029
Notebook computers	0.617	0.137	<.001

[1]p-values are for Tamhane's test. Similar results hold for other tests.

Initially, the result for digital music players appears surprising because these products can be evaluated based on quantifiable attributes such as memory size, sound quality, and battery life. However, on further consideration, it seems reasonable that—as with boots—style and fashion may have a major influence on preferences for these items. Indeed, a senior brand manager at Creative Technologies said of digital music players: “We understood the whole thing with these players can't be just functionality, that we always concentrated on...people were using them as fashion statements” (Marriott, 2004, p. 1). Our results therefore provide support for the conjecture that DESIRE generates more accurate recommendations when available item attributes coincide with user value dimensions.

7.4. Limitations

The above results should be understood in light of the empirical validation's limitations. The lab study used a mostly homogeneous group of university students. As a result, it is not clear whether the relationship between viewing time and preference, and the expression of preference for an item as an aggregation of preferences for its attribute values, will generalize to other demographic contexts including different cultures and age groups. Additionally, the same attribute weights were used for all participants. As the relative importance of attributes will vary among individuals, the accuracy of DESIRE could be improved by optimizing the attribute weights at an individual or demographic group level. Notwithstanding this, our empirical study demonstrates that useful recommendations can be generated even with global attribute weights given a relatively homogeneous population.

Additionally, as Figure 1 depicts, factors other than preference affected viewing time. Thus, it will not be possible to use viewing time as a perfect measure of preference. What we have shown instead is that indicators of preference can be extracted from viewing times and used to predict preferences for item attributes.

Furthermore, one could argue that the laboratory setting unduly favors DESIRE as real-world viewing time data would include significant noise from multitasking, distractions, etc. However, modern web technologies are capable of detecting inactivity (which is how instant messaging systems automatically change user status from “online” to “away”). More fundamentally, the purpose of the study was to determine the feasibility of using the relationship between preferences and viewing time to generate recommendations.

Moreover, from our observations during the study, many sources of noise were evident. Although participants were asked to focus on the experimental task, some participants talked to each other or interacted with their mobile phones. While this may increase the external validity of the study, such noise may introduce ill-understood variance in the results and therefore threaten internal validity. However, we felt that taking more draconian steps such as confiscating participants’ mobile phones and requesting silence would have increased the setting’s artificiality. Similarly, running individual sessions may have increased internal validity but would have required a substantial reduction in sample size.

Additionally, in the conceptual evaluation of DESIRE, we noted that DESIRE assumes that users’ value dimensions are reflected in item attribute data. However, the above results show that the approach may be somewhat useful for items (such as boots) where preferences are often understood at least partly in terms of intangible qualities.

Finally, our experimental evaluation considered only single session viewing times to infer preferences for item attributes. Thus, it was not possible in our study to distinguish enduring from situational preferences. These can clearly differ, depending on whether a person is shopping for oneself or for a gift for someone else.

7.5. Directions for Future Research

These limitations suggest several areas for future research. Because DESIRE’s viewing-time-based user profile generator and content-based recommendation engine are independent, they could be investigated in combination with alternative components. For example, implicit ratings from the user profile generator could be used as input for user-based collaborative filtering, and the recommendation engine could be used with explicit ratings. Furthermore, a key aspect of DESIRE’s recommendation approach is the construction of preferences based on item attributes. This approach supports an alternative to the traditional way of thinking about product catalogs in predetermined and fixed categories. Indeed, in specific situations, our approach can be used to make recommendations from one product category based on attribute preferences inferred from items in another category. In this way, it is possible to support multiple item classifications on an ad hoc basis (Parsons & Wand, 2008a, 2008b). Additionally, because DESIRE requires estimates of the relative importance of product attributes in influencing preferences, further research is needed to examine the role of enduring and situational factors in determining these estimates. Future work could incorporate longer-term preferences (reflected by repeated measures of viewing time for particular attributes) into user profiles, and factor these preferences into DESIRE recommendations. A more sophisticated ideal value calculation that incorporates “disliked” items (perhaps by aggregating multiple views of disliked items to arrive at an appropriate weight representing the extent to which these items should be avoided in recommendations) may increase DESIRE’s performance. Moreover, work is needed to better understand the effectiveness of the proposed approach for different kinds of items. We examined five specific categories of consumer products, but our understanding of the factors that influence preference for different kinds of products and other items is incomplete. Similarly, to fully understand the range of contexts to which DESIRE can be applied, the effectiveness of recommendations in other domains (e.g., information search) needs to be examined. Further research is also needed to examine the impact of other factors that influence viewing time (e.g., complexity, novelty, position in sequence) and control for these in DESIRE’s algorithm. By isolating the impact of preference from other factors that influence viewing time, it should be possible to produce further improvements in DESIRE’s accuracy.

Finally, note also that, in the experimental setting, DESIRE was not used to generate recommendations to be evaluated by participants. Instead, predicted ratings from DESIRE were compared to actual ratings of items provided by participants. Further research might use DESIRE to make recommendations for users, and the quality of such recommendations evaluated directly by participants. Such a study could be conducted in either a lab or field setting.

8. Summary and Conclusions

Recommender systems guide a user “in a personalized way to interesting or useful objects in a large space of possible options” (Burke, 2002, p. 6). Many modern recommenders blend an ensemble of predictors to improve recommendation quality. Bell et al. (2007) explain: “The success of an ensemble approach depends on the ability of its various predictors to expose different, complementing aspects of the data” (p. 6). Following this, we proposed and tested a novel predictor, DESIRE, which exploits the positive relationship between item viewing times and user preferences to make accurate predictions. The results of a laboratory experiment with DESIRE demonstrate not only that the viewing time / preference relationship can be used to predict preferences for unseen items without comparing users, but also that preference for an item can be modeled as an aggregation of preference ratings of its attributes.

DESIRE is unlike existing predictors in that it employs a rarely-used aspect of the available data (i.e., viewing time). Furthermore, it makes inferences about a user’s preferences directly, not by comparison to other users. Moreover, DESIRE achieves this without any explicit ratings or onerous user interaction. These characteristics make DESIRE applicable in many situations where collaborative predictors and more obtrusive content-based predictors are impractical or ineffective.

Therefore, the paper has three contributions. First, we presented the DESIRE algorithm (fully specified in the Appendix). Second, we presented a thorough analysis of recommender evaluation, and conclude that heuristics should be tested against a null hypothesis rather than a competing artifact. Third, we presented substantive empirical evidence that the DESIRE predictor can extract useful patterns from viewing time data, which suggests strong potential for its inclusion in ensemble recommenders across diverse domains. These contributions have implications for both research and practice.

From a theoretical perspective, DESIRE instantiates specific theoretical propositions regarding the relationships between viewing time and user preference and between preferences for items and preferences for their attributes. Demonstrating that DESIRE’s accuracy exceeds that of a random recommender (which instantiates the null hypothesis) by a statistically and practically (moderate to large effect sizes) significant margin supports both of these theoretical propositions. Specifically, we showed that viewing time of an item reveals useful information about user attitude toward that item, attitude toward an item can be modeled as a function of attitudes toward attributes of the item, and inferred preference for item attributes facilitates inferences about preferences toward other items. This contributes to both the psychological research on viewing time and the research on content-based recommender systems because DESIRE can be employed to determine preferences in an effective, unobtrusive manner based on observed browsing behavior.

From a design perspective, our research highlights the potential of psychological theories or research findings to guide design choices in constructing IT artifacts. Prior research on recommender design has emphasized abstract mathematical methods to calculate similarity (such as user-to-user similarity or item-to-item similarity) and has usually required large amounts of data to perform well. Such approaches lack a clear psychological basis for why similarity measures will translate to shared preferences. In contrast, DESIRE demonstrates that a well-supported psychological property (the relationship between viewing time and preference) can be used to predict preferences accurately (thereby forming a basis for recommendations) using relatively little data. This, in turn, can motivate a reanalysis of existing psychological literature with a view to identifying empirical findings that can guide IT artifact design.

Methodologically, our analysis of recommender evaluation contributes to evaluation methodology because systematically distinguishing between recommender systems and heuristics motivates a shift in evaluation practices across the field. This is a widely misunderstood issue, and we hope our exposition initiates further investigation of more appropriate evaluation practices. Venable, Pries-Heje, and Baskerville (2012) lay out a state-of-the-art framework for choosing artifact evaluation methods. One of the questions they ask is: “Determine the goal/purpose of the evaluation. Will you evaluate single/main artifact against goals? Do you need to compare the developed artifact against with other, extant artifacts?” (p. 434). However, they do not answer the question of when is it appropriate to compare the new artifact against existing alternative artifacts. We explore this issue in depth and find that artifacts should be evaluated against alternatives unless existing alternatives require different data or resources, or the proposed artifact embeds hypotheses (in our case, based on psychological theory) that one wants to test.

Practically speaking, our results illuminate the potential usefulness of viewing time data and attribute data for non-collaborative recommender heuristics. For companies that currently employ recommender systems, our results suggest incorporating viewing time data into recommendation calculation by, for example, adding DESIRE to their existing recommender ensembles. Firms that lack the data required by many existing predictors may be able to use or adapt DESIRE for their particular situations. For example, firms that have substantial item attribute data, but lack the explicit ratings or browsing history needed for collaborative filtering, may be able to deploy a recommender system based on DESIRE. In the simplest case, DESIRE can be implemented to operate in real-time setting. Users can be provided recommendations after browsing a small number (five to ten) of calibration items. Viewing times and attributes of the calibration items serve as input to DESIRE. The remaining items are then rated using DESIRE’s rating algorithm. The highest rated items can then be recommended. Moreover, ratings can be updated based on additional viewing time data gathered as users view additional items. Furthermore, firms may use techniques including browser cookies to retain item viewing history across user sessions, facilitating more accurate DESIRE recommendations. More generally, however, our research adds DESIRE to the pool of known useful predictors, from which an expert may compose an ensemble of predictors for a specific domain. Furthermore, developers of of-the-shelf recommender systems may benefit by adding DESIRE, which uses significantly different data and methods than existing approaches, as an additional heuristic to improve quality.

Acknowledgements

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada for funding that supported this research. The authors also thank the senior editor and the reviewers for detailed feedback that substantially improved the paper.

References

Adomavicius, G. R., Sankaranarayanan, S. S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems*, *23*(1), 103.

Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 714–720). Palo Alto, CA: AAAI Press.

Bell, R., Koren, Y., & Volinsky, C. (2007). The BellKor solution to the Netflix prize. *AT&T Labs.*

Bornstein, R., & Carver-Lemley, C. (2004). Mere exposure effect. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory* (pp. 215–234). Hove, UK: Psychology Press.

Browne, G. J., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, *31*(1), 89–104.

Burke, R. (1999). *Integrating knowledge-based and collaborative-filtering recommender systems*. Paper presented at the 1999 Workshop on AI and Electronic Commerce.

Burke, R. (2000). Knowledge-based recommender systems. In A. Kent (Ed.), *Encyclopedia of library and information systems* (Vol. 69, Suppl. 32). New York: Marcel Dekker.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331.

Claypool, M., Le, P., M, W., & Brown, D. (2001). Implicit interest indicators. *Proceedings of the 6th International Conference on Intelligent User Interfaces, USA*, 33–40.

Cooper, M. D., & Chen, H.-M. (2001). Predicting the relevance of a library catalog search*. Journal of the American Society for Information Science*, *52*(10), 813–827.

Day, H. (1966). Looking time as a function of stimulus variables and individual differences. *Perceptual & Motor Skills*, *22*(2), 423–428.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organizations: A review of the literature. *International Journal of Information Management*, *20*(1), 17–28.

Faw, T., & Nunnally, J. (1967). The effects on eye movements of complexity, novelty, and affective tone. *Perception & Psychophysics*, *2*(7), 263–267.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.

Gladwell, M. (1999). The science of the sleeper: How the information age could blow away the blockbuster. *The New Yorker*, *75*(29), 48–55.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*(2), 102–123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3–19.

Häubl, G., & Murray, K. B. (2006). Double agents: Assessing the role of electronic product recommendation systems. *MIT Sloan Management Review*, *47*(3), 8-12.

Heinrich, P. (1970). Free looking time: A method for determining preference. *Psychologie und Praxis*, *14*(2), 79–93.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5-53.

Hong, W., Thong, J. Y. L., & Tam, K. Y. (2005). The effects of information format and shopping task on consumers' online shopping behavior: A cognitive fit perspective. *Journal of Management Information Systems*, *21*(3), 149–184.

Jahrer, M., Töscher, A., & Legenstein, R. (2010). Combining predictions for accurate recommender systems. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 693–702). New York, NY: ACM.

Jin, X., & Mobasher, B. (2003). *Using semantic similarity to enhance item-based collaborative filtering*. Paper presented at the 2nd IASTED International Conference on Information and Knowledge Sharing.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. New York, USA: Harcourt Brace Jovanovich.

Keeney, R. L. (1968). Quasi-separable utility functions. *Naval Research Logistics Quarterly*, *15*, 551-565.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Kohrs, A., & Merialdo, B. (2001). Creating user-adapted websites by the use of collaborative filtering. *Interacting with Computers*, *13*(6), 695–716.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, *40*(3), 77-87.

Koren, Y. (2009). The BellKor solution to the Netflix grand prize. *Netflix*. Retrieved November 1, 2012, from http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

Lee, T. Q., Park, Y., & Park, Y.-T. (2008). A time-based approach to effective recommender systems using implicit feedback. *Expert Systems with Applications*, *34*(4), 3055-3062.

Liang, T.-P., Lai, H.-J., & Ku, Y.-C. (2007). Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, *23*(3), 45–70.

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review*, *55*(5), 291-300.

Marriott, M. (2004). And now for something slightly different. *The New York Times*. Retrieved from from http://www.nytimes.com/2004/12/16/technology/circuits/16musi.html

Mathes, A. (2004). Folksonomies—cooperative classification and communication through shared metadata. Retrieved July 2, 2012, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Miller, B. N., Riedl, J. T., & Konstan, J. A. (2003). GroupLens for Usenet: Experiences in applying collaborative filtering to a social information system. In C. Lueg & D. Fisher (Eds.), *From Usenet to CoWebs: Interacting With social information spaces* (pp. 206–231). London: Springer Press.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Improving the effectiveness of collaborative filtering on anonymous web usage data. *Proceedings of the Third International Workshop on Web Information and Data Management* (pp. 9–15).

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, *6*(1), 61–82.

Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000). Integrating web usage and content mining for more effective personalization. *Proceedings of the First International Conference on Electronic Commerce and Web Technologies*, *LNCS 187*5, 165–176.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, *13*(1), 29–39.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 272–281).

Murthi, B., & Sarkar, K. (2003). The role of the management sciences in research on personalization. *Management Science*, *49*(10), 1344–1362.

Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology* (pp. 38–45).

Oostendorp, A., & Berlyne, D. E. (1978). Dimensions in the perception of architecture II: Measures of exploratory behavior. *Scandinavian Journal of Psychology*, *19*(1), 83–89.

Parsons, J., & Ralph, P. (2010). *System and method for estimating user ratings from user behavior and providing recommendations*. United States Patent No. US7756879.

Parsons, J., & Wand, Y. (2008a). A question of class. *Nature*, *255*(7216), 1040–1041.

Parsons, J., & Wand, Y. (2008b). Using cognitive principles to guide classification in information systems modeling. *MIS Quarterly*, *32*(4), 839–868.

Parsons, J., Ralph, P., & Gallagher, K. (2004). *Using viewing time to infer user preference in recommender systems*. Paper presented at the AAAI Workshop in Semantic Web Personalization.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159–188.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, *27*(3), 313–331.

Pazzani, M., & Billsus, D. (2007). The adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (Vol. 4321, pp. 325–341). Heidelberg, Berlin: Springer-Verlag.

Pereira, R. E. (2001). Influence of query-based decision aids on consumer decision making in electronic commerce. *Information Resources Management Journal*, *14*(1), 31–48.

Perkowitz, M., & Etzioni, O. (2000)*. Towards adaptive web sites: Conceptual framework and case study.* Paper presented at the Proceedings of the Eighth World Wide Web Conference WWW8.

Ralph, P., & Parsons, J. (2006). *A framework for automatic online personalization*. Paper presented at the Presented at the 39th Annual Hawaii International Conference on System Sciences.

Sacco, G. (2006). Dynamic taxonomies and guided searches. *Journal of the American Society for Information Science and Technology*, *57*(6), 792–796.

Sarwar, B., Konstan, J., Brochers, A., Herlocker, J., Miller, B., & Riedl, J. (1998). Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 345–354).

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1-2), 115.

Schafer, J., Frankowski, D., Herlocker, J., & Sen, S. (2007). The adaptive web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (Vol. 4321, pp. 291–324). Heidelberg, Berlin: Springer-Verlag.

Seo, Y.-W., & Zhang, B.-T. (2000). A reinforcement learning agent for personalized information filtering. In *Proceedings of the 5th International Conference On Intelligent User Interfaces* (pp. 248–251). New Orleans LA: ACM.

Shahabi, C., & Chen, Y.-S. (2003). An adaptive recommendation system without explicit acquisition of user relevance feedback. *Distributed and Parallel Databases*, *14*(2), 173–192.

Shahabi, C., Banaei-Kashani, F., Chen, Y., & McLeod, D. (2001). *Yoda: An accurate and scalable web-based recommendation system.* Paper presented at the Proceedings of the Sixth International Conference on Cooperative Information Systems, Trento, Italy.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Tai, X., Ren, F., & Kita, K. (2002). An information retrieval model based on vector space method by supervised learning. *Information Processing and Management*, *38*(6), 749–764.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In K. Peffers, M. Rothenberger, & B. Kuechler (Eds.), *Proceedings of DESRIST* (LNCS 7286) (pp. 423–438). Berlin: Springer-Verlag.

Vrooman, E., Riedl, J., & Konstan, J. (2002). *Word of mouse: The marketing power of collaborative filtering.* New York: Warner Business Books.

Welty, C. A. (1998). The ontological nature of subject taxonomies. In *Proceedings of Formal Ontology in Information Systems* (pp. 317–327).

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137–209.

Zhao, R., & Grosky, W. (2002). Narrowing the semantic gap—improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, *4*(2), 189–200.

Appendix. DESIRE Algorithm

Rating Estimator

Description: For each viewed item, the rating estimator calculates an implicit rating equal to the standardized viewing time of that item in the range [-1,1].

Precondition: The user must have viewed at least one item from the item set

Input: A set of item/viewing time pairs

Output: A set of item/rating pairs

Algorithm:

```
1. For each item, i, set the rating of i, $r_i$, equal to the z-score of the viewing time of i

2. Limit outliers

    a. For each item rating such that $r_i > 3$, set $r_i = 3$

    b. For each item rating such that $r_i < -3$, set $r_i = -3$

3. For each $r_i$, set $r_i = r_i /3$
```

Discussion of variations: At least three variations on this method are possible. First, outliers may be deleted instead of limited to the closer boundary of the [-1,1] range. Second, viewing times can be transformed to fit or approximate any distribution (not just a normal distribution) as long as the range of the returned values is [-1,1]. Third, a more complex function of viewing time can be employed; for example, one that adjusts for the complexity of the item.

User Profile Generator

Description: Given a set of item/rating pairs, generate a description of the user's preferences in terms of the attributes of the items. This is computed differently for numeric attributes than categorical attributes.

Preconditions: The sets of item ratings and item attributes must not be empty; the positive example threshold must be ≥ 0.

Input: 1) A set of n seen items, I, 2) a rating for each item, 3) a set of attributes (e.g., color, price) for each item, 4) a set of values (e.g. red, $15), corresponding to the attributes, for each item, 5) the positive example threshold.

Output: A user preference profile, consisting of ideal values of numeric attributes and inferred ratings of categorical (non-numeric) attributes.

Algorithm:

```
1. Divide the attributes into two groups, numeric and categorical, as follows

    a. If the data associated with an attribute is not unique to the item (e.g., the ISBN of a
       book), and the attribute is nominal or ordinal, assign the attribute to the categorical
       attribute group
```

   b. If the data associated with an attribute is interval or ratio, assign the attribute to the numeric attribute group

2. Calculate Ideal Values of Numeric Attributes

   a. For each item, i, if the rating of i ≥ 'positive example threshold,' add i to the set of positive examples.

   b. For each numeric attribute found in one or more positive examples, calculate its "ideal value," as a weighted average of the attribute's value in positive examples, as follows. Here, r_i is the rating of the i^{th} positive example, v_{ai} is the value of attribute a in the i^{th} positive example and IV_a is the ideal value of attribute a.

$$IV_a = \frac{\sum_i r_i v_{ai}}{\sum_i r_i}$$

3. Calculate Ratings of Categorical Attributes

   For each value, v, of each categorical attribute of each item, calculate the rating of v as the mean rating of items having v. (e.g., if three items are red, the rating of red is the mean rating of those three items).

4. Return the User Profile, consisting of the ideal values of each numeric attribute and the ratings of each categorical attribute value.

Recommendation Engine

Description: Predict ratings of unseen items based on the user profile and recommend items with high predicted ratings.

Input: 1) the user profile, U, 2) the unseen-item data store, D, consisting of a set of unseen items and their attributes, 3) a list of "relative importance weights" that indicate the importance of each attribute in determining preference, and 4) the number of recommendations requested.

Preconditions: 1) The user profile must be non-empty; 2) The unseen-item datastore must be non-empty; 3) the intersection of attributes in the user profile and attributes of unseen items must be non-empty; 4) the list of "relative importance weights" must contain an entry corresponding to each attribute in both the user profile and datastore; 5) The number of recommendations requested must not exceed the number of unseen items.

Output: A set or recommended items

Algorithm:

1. Standardize numeric data. For each numeric attribute in the intersection of U and D:

   a. Replace all values of that attribute in D with their z-scores.

   b. Replace the ideal value of that attribute in U with its z-scores.

2. Let R be an empty list of item similarity/pairs.

3. For each unseen item, i_u:

  a. Let S be a list of attribute/similarity pairs

  b. For each categorical attribute, a, in i_u:

    i) Let s be the similarity of the value to the user profile

    ii) Set s as follows, where n is the number of values of a and $U(V_j)$ is the user profile rating of the jth value of a (i.e., if an item is red and white, average the rating of red and white from the user profile, then convert from a [-1,1] range to a [0,1] range).

$$s = \frac{\dfrac{\sum_{j=1}^{n} \left| U(V_j) \right|}{n} + 1}{2}$$

    iii) Add (a, s) to S

  c. For each numeric attribute, a, in i_u:

    i) Let s be the similarity of the value to the user profile

    ii) Set s as follows, where V_a is the value of a and U_a is the ideal value of a from the user profile (i.e., calculate dissimilarity as the absolute value of the difference between the ideal value and the actual value, then divide by six to convert to a [0,1] range, and subtract from one to get similarity).

$$s = 1 - \frac{\left| V_a - U_a \right|}{6}$$

    iii) Add (a, s) to S

  d. Let R_i be the overall similarity between the user profile, U and unseen item, i.

  e. Let W be the subset of relative importance weights corresponding to, and in the same order as, the attributes in S.

  f. Set R_i as follows, where W_a is the relative importance of attribute a, and S_a is the similarity of i to U in terms of a.

$$R_i = \frac{\sum_{a=1}^{|S|} (W_a S_a)}{\sum_{a=1}^{|S|} (W_a)}$$

  g. Add (i, R_i) to R.

4. Let R' be the subset of R having the n items with highest similarity, where n is the number of recommendations desired.

5. Return R'.

Discussion of variations: Many variations on this method are possible. First, numeric attributes could be fit to distributions other than the standard, normal distribution, as long as they can be transformed to a [0,1] range. Second, if items could have multiple values for a numeric attribute, one could apply the same averaging technique as used to account for multiple categorical values per attribute. Third, different techniques for accounting for multiple values per attribute could be used; for example, instead of averaging the similarity of different values, DESIRE could simply choose the highest or lowest similarity. Fourth, instead of returning a set number of recommendations, DESIRE could return all of items exceeding a given similarity threshold. Fifth, DESIRE can incorporate a forgetting function (i.e., ignoring viewing data beyond a certain age or giving additional weight to more recent viewing history) to account for changing preferences.

About the Authors

Jeffrey PARSONS is University Research Professor and Professor of Information Systems in the Faculty of Business Administration at Memorial University of Newfoundland. He holds a Ph.D. in Information Systems from the University of British Columbia. His research interests include conceptual modeling, data management, business intelligence, and recommender systems; he is especially interested in classification issues in these and other domains. His research has been published in journals such as *Nature*, *Management Science*, *MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Communications of the ACM*, *ACM Transactions on Database Systems*, and *IEEE Transactions on Software Engineering*. He has served in several editorial roles, both for journals and conferences.

Paul RALPH is a Lecturer in Information Systems at Lancaster University. He holds a Ph.D. in Information Systems from The University of British Columbia. His research centers on the theoretical and empirical study of software engineering and information systems development including projects, processes, practices, tools and designer cognition, socialization, productivity, wellbeing and effectiveness. His research has been published by IEEE, ACM, AIS, Springer, and Elsevier. He has served in several editorial roles, both for journals and conferences.