Interaction and Personalization of Criteria in Information Retrieval Models

Shawn R. Wolfe[1,2] and Yi Zhang[1]

[1] School of Engineering, University of California Santa Cruz,
Santa Cruz CA 95064, USA,
{srwolfe,yiz}@soe.ucsc.edu
[2] NASA Ames Research Center, Moffett Field CA 94035, USA

Abstract. A user's informational need and preferences can be modeled by criteria, which in turn can be used to prioritize candidate results and produce a ranked list. We examine the use of such a criteria-based user model in the context of two very different recommender scenarios: news article recommendations and product recommendations. We ask the following: are there nonlinear *interactions* among the criteria; and should the models be *personalized*? We assume that that user ratings on each criterion are available, and use machine learning to infer a user model that combines these multiple ratings into a single overall rating. We found that the ratings of different criteria have a nonlinear interaction in some cases, for example, article novelty and subject relevance often interact. We also found that these interactions vary from user to user.

1 Introduction

Choosing one or more items among many candidates often requires an evaluation on multiple criteria. For instance, the search for habitable planets may involve evaluating candidate observations on the basis of star type, observation quality and telescope time. In other cases, it may be necessary to trade off competing interests. For example, when evaluating potential Mars landing sites, it is necessary to balance the priorities of various science teams, as well as the likelihood of mission success and telemetric potential. Criteria-based models, which capture multiple, potentially competing aspects of a user's need, have been developed and used in operations research [18]. In commercial applications, a better understanding of the involved criteria for a particular user need could also lead to better marketing and product development opportunities.

A previous study showed that using a linear combination of multiple criteria to model the user's need can improve information retrieval results [20]. This paper goes beyond a linear combination to model the interactions among the criteria. Specifically, we seek answers to the following questions:

1. Is there evidence that some criteria interact in the decision/rating process?
2. If so, are there discernible patterns to these interactions?
3. Given interactions, are these interactions consistent across users?

To answer these questions, we perform our study within the context of two very different recommendation tasks: news article recommendations — a representative task for adaptive filtering; and product recommendations (for flat panel televisions) — a representative task for collaborative filtering. We expect certain interactions might exist. For example, low ratings on certain criteria might negate higher ratings in other criteria. Is respect for an article's author still important when the article is not of the desired topic? Is the durability of a television a factor if it has a poor picture? Or it may be that certain high ratings limit the impact of other criteria. For instance, considering two articles with the same breaking news, does the readability of the article make much difference? It is these sorts of interactions that we are searching for. On the two tasks, we test for the presence of interactions by comparing the root mean squared error (RMSE) of learned linear and nonlinear user models for predicting the overall item rating or recommendation.

The rest of the paper is organized as follows. In Section 2, we review related work. In Section 3, we describe our recommender datasets. In Section 4, we detail our approach to represent criteria interactions and to select the best model. We present our experimental results in Section 5 and our conclusions in Section 6.

2 Related Work

In information retrieval, the limited adoption of criteria-based user models has been mostly restricted to enhancing standard relevance-based models with novelty. Researchers have studied criteria such as information-novelty for search [6], summarization [4], filtering [22] and topic detection and tracking [1]. Prior research on a user's perception/criteria have found that a wide range of factors (such as personal knowledge, topicality, quality, novelty, recency, authority and author qualitatively) affect human judgments of relevance [2][10][17][19][16].

Most of the research in the information retrieval community that uses multiple criteria has been in information filtering. Manouselis and Costopoulou categorize 37 recommender systems that implicitly use some multi-criteria aspect in their operation [11]. These systems primarily use only the weighted sum (i.e., linear combination) model. Of the information filtering systems we are aware of, PENG [13] is the most similar in application to our experimental domain. PENG is a multi-criteria news bulletin filtering system that utilizes several criteria, including content, coverage, reliability, novelty and timeliness. A later evaluation of PENG, using only content and coverage, showed comparable or superior performance to other approaches [3]. Farah and Vanderpooten have explored the use of multiple criteria in search using rank-based methods. In their work [8], the user provides query terms as the only input (and thus criterion) for the search process. From this, additional criteria are formed from elements of the web page, such as text, keywords, anchor text and incoming links. Later work [9] expanded on this notion by using the rankings produced by several high performing algorithms, with each algorithm essentially acting as a criterion or critic.

Learning user models based on multiple criteria (as opposed to content alone) is not common in information retrieval. Naïve Bayesian classifiers were used to learn content-based user profiles for movie search [7]. A more complicated scheme was used to predict whether a user would watch television programs [14], first by building a model of what genres a user likes, and then classifying each show based on its genres by means of a support vector machine. DIVA [12] uses a somewhat similar approach to recommend movies, using the C5.0 algorithm to classify each movie based on its metadata.

Outside the information retrieval community, general additive independence models have gained some popularity, and are akin to our current approach. One method for estimating generalized additive independent utility functions is to treat criteria as random variables and use Bayesian techniques to estimate them [5]. This same utility decomposition concept was later applied to the multi-issue negotiation task, by representing the utility of a buyer in a utility graph [15].

3 Datasets

We used two recommendation datasets for our research. Each dataset had four criteria and one overall rating defined. The range of these ratings are different for different criteria, as the data were originally collected for other research. For consistency, we have rescaled all ratings to have minimum and maximum values of 0 and 1, respectively. After this rescaling, the ratings were either binary (0 or 1) or five-valued (0.0, 0.2, 0.4, 0.6, 0.8 or 1.0). For both data sets, we restrict ourselves to user-item pairs with complete ratings (i.e., any items with missing ratings were excluded from our study).

3.1 News Recommendation

Our news recommendation data were provided by the University of California, Santa Cruz and Carnegie Mellon University [21]. The data were previously collected in a user study performed on the Yow-now news filtering system. Yow-now was an information filtering systems that delivered news articles to users from various RSS feeds. Approximately twenty-five users used the Yow-now system for about a month, reading news for at least one hour each day, rating approximately 9000 articles in all, with an average of 383 articles rated per user (with a standard deviation of 252.8). This allowed us to explore creating personalized user models with the Yow-now dataset.

The users rated each article according to the following four criteria:

Authoritative : how authoritative the article appeared (binary).
Novel : the novelty of the article (five-valued).
Readable : the ease of reading the article (binary).
Relevant : the degree to which the article was relevant to the general subject category of the article (five-valued).

The overall user rating of the article was given on a five-point scale.

3.2 Product Recommendation

Our product recommendation data came from a crawl of the Epinions.com review site. Our dataset is restricted to flat panel television reviews. Approximately 1100 users reviewed 1200 items, with an average of 1 review per user (with a standard deviation of 0.29). With such a small number of reviews per user, it was clearly not possible to build personalized user models with this dataset.

The users rated each product according to the following four criteria:

Sound : The sound quality of the television (five-valued).
Ease of Use : Ease of use of the various features and menus (five-valued).
Picture Quality : All visual aspects of the television's picture (five-valued).
Durability : Durability of the television set (five-valued).

The overall user rating of the article was given on a five-point scale.

4 Approach

To test for interactions among criteria in the final decision/rating process, we compared the performance of two sets of models on a rating prediction task. The first model is a linear combination of ratings on the criteria, which makes the assumption that the criteria do not interact in the user decision process. The second set of models are nonlinear combinations that explicitly represent interactions among pairs of criteria, assuming that such interactions occur in in the user decision process. Both models take the user's item rating on each criterion as input, and output a prediction of the item's overall rating.

In our experiment, we first used machine learning to estimate the model parameters from training data. We then compared the prediction accuracy of the two sets of models on testing data. If the nonlinear model performed better, then we would have expected similar results in practice under conditions comparable to our study. On the other hand, if there were no such interactions in practice, the nonlinear model should have performed no better than the linear model. As mentioned earlier, we used RMSE as our evaluation measure, as is commonly done for recommender systems.

4.1 Lower Bound of Root Mean Squared Error

Although not necessary to determine if interactions among criteria exist, we defined a lower bound on RMSE to give our findings context. We defined an oracle who makes the optimal prediction for each set of values, but without knowing the actual overall user rating. Users were not entirely consistent when rating items; not only did different users give identical ratings on each criterion but a different overall rating, but single users also gave different overall ratings on two items that were otherwise rated identically. Such differences may have been due to some random variability in the ratings (from difficulty in estimating or user changes over time), or may also have been due to other factors, such as the coarseness of the ratings or from other criteria excluded from the study.

In any case, given that the overall rating was not a single-valued function of the ratings over the criteria, it was not possible to entirely eliminate prediction error. Instead, our oracle minimized RMSE over the entire dataset in the following way: given a set of ratings on each criterion, it predicted the overall mean rating from the entire dataset (using *both* training and test sets), as the mean produces the minimum RMSE. In this sense, the oracle knew the distribution of the entire dataset, but made the best possible prediction based only on the ratings over criteria, and without differentiating between the test and training set. We stress that the oracle was a lower bound which is unlikely to be found in all cases by any particular generalization algorithm.

4.2 Linear Model

The linear model is simply a linear combination over the ratings for each criterion; the independent variables are the ratings on the criteria, plus a bias term, and the dependent variable is the overall rating. If it was possible to select the best nonlinear model in every case, the RMSE of the linear model would serve as a upper bound on RMSE, as the linear model is a special case of nonlinear models described below. However, due to overfitting, it is possible to select a nonlinear model that is suboptimal and worse than the linear model. The RMSE achieved by the linear model is our baseline and a failure to improve upon it would indicate a lack of evidence for the criteria interactions the nonlinear model tries to capture. The linear model is simply:

$$P_L = \sum_{i=1}^{m} w_i v_i \tag{1}$$

where P_L is the predicted overall rating, v_i is the item rating on the ith criterion, and w_i are the coefficients to be learned.

4.3 Nonlinear Model

The general class of nonlinear models allows for any consistent prediction of overall rating based on the ratings on each criterion. However, this introduced too many possible models to effectively choose from, given the small amount of data, and exacerbated by inconsistency in the overall ratings (as noted earlier). Therefore, we limited ourselves to interactions between pairs of criteria. Observing interactions on this restricted set would be sufficient to show that criteria interactions existed, though we may not have found the optimal nonlinear model. Conversely, a failure to observe interactions would not have indicated that interactions do not exist, as the interaction may have been on several criteria.

We modeled interactions among pairs of criteria by creating derived binary features that correspond to specific ratings on criteria in a linear combination:

$$P_{ab} = \sum_{i=1}^{m} w_i v_i + c_{ab} \sum_{x \in A}^{m} \sum_{y \in B}^{m} I((v_a = x), (v_b = y)) \tag{2}$$

where P_{ab} is the predicted overall rating, a and b are the selected criteria pair, A is the set of possible values for criterion a, B is the set of possible values for criterion b, I is an indicator function that returns 1 when the arguments hold, 0 otherwise, v_i is the item rating on the ith criterion, and w_i and c_{ab} are the coefficients to be learned. Note that the first summation is simply Equation 1, and the second summation is simply a linear combination over a new set of (derived) features. In other words, we have created new binary features for each possible pair of ratings on criteria a and b. For example, when combining *authority* and *readability* (two binary criteria), $2 * 2 = 4$ new binary features are created; when combining *authority* and *novelty* (a binary and a five-valued criteria), $2 * 5 = 10$ new binary features are created. One can think of these induced binary features as correction factors, and as such, any nonlinear combination involving only these two features can be represented.

Since both datasets have four criteria, this gives us $C_4^2 = 6$ pairs of criteria to choose from. We also added a seventh nonlinear form (*all-pairs*) which uses all six pairwise combinations. We are further aided by the fact our criteria are discrete and take on a small set of values; for our data, the number of pair values for a criteria pair ranges from four to twenty-five. Table 1 shows the number of unique pairs of ratings observed for each criteria pair; a binary feature is created for each unique pair of ratings.

Table 1. Number of unique ratings possible when combining pairs of criteria.

Yow-now	Authority	Novelty	Readability	(Subject) Relevance
Authority	n/a	10	4	10
Novelty		n/a	10	25
Readability			n/a	10
(Subject) Relevance				n/a
Epinions.com	**Sound**	**Ease of Use**	**Picture Quality**	**Durability**
Sound	n/a	24	22	24
Ease of Use		n/a	25	24
Picture Quality			n/a	24
Durability				n/a

4.4 Regularization

Since both sets of models take a linear form (as we have represented the nonlinear form as a linear model on a new feature space, as described above), we may use linear regression to find model parameters that minimize RMSE on the training data. However, our goal is to minimize RMSE on the unseen testing data, not the training data, and given the small training set size, some form of regularization is

needed to avoid overfitting. We use Tikhonov regularization, a special case of L2-norm regularization or ridge regression. The analytical solution to the minimize RMSE with regularization is:

$$\mathbf{W} = (\lambda \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} (\lambda \mathbf{W}_0 + \mathbf{X}^T \mathbf{Y}) \tag{3}$$

where an exponent of T indicates matrix transposition, λ controls the amount of regularization, $\mathbf{I}$ is the identify matrix, $\mathbf{X}$ is the instance matrix, $\mathbf{Y}$ is the vector of target values, $\mathbf{W}_0$ is the regularization vector we specify and $\mathbf{W}$ is the vector of coefficients we seek. Larger values of λ causes the solution to be closer to $\mathbf{W}_0$. We also added a constant term to our ratings representation to account for any bias in the overall rating.

For the linear model, we biased towards the following regularization vector:

$$\mathbf{W}_0 = \begin{bmatrix} 0.0 \; 0.25 \; 0.25 \; 0.25 \; 0.25 \end{bmatrix} \tag{4}$$

where the first position is the constant bias term and the remaining terms are the coefficients for the four criteria. We chose W_o such that all criteria would be weighted evenly, and the minimum (maximum) overall rating would be predicted when the minimum (maximum) rating was given on each criterion.

For the nonlinear models, we biased the model against interactions between criteria. The first five terms of the nonlinear regularization vector are the same as in the linear case, followed by zeros for each unique criteria pair value:

$$\mathbf{W}_0 = \begin{bmatrix} 0.0 \; 0.25 \; 0.25 \; 0.25 \; 0.25 \; 0.0 \; .... \; 0.0 \end{bmatrix} \tag{5}$$

Since the number of criteria pairs varies, the size of $\mathbf{W}_0$ also varies.

4.5 Tuning and Model Selection

The λ term in equation 3 controls the tradeoff between coefficients that minimize RMSE on the training set, and coefficients that are closer to the regularization vector ($\mathbf{W}_0$) described above. Higher values of λ moves the solution closer to the regularization vector, while allowing for higher RMSE; lower values of λ do the opposite. We automatically tuned the value of λ with ten-fold cross-validation on the training set alone. For a candidate value of λ and for each fold of the training data, we used the other 90% of the training data to learn the coefficients (using Equation 3); we used these coefficients to predict the overall ratings and record the RMSE. Starting with $\lambda = 0$, we tried successfully higher values of λ until the mean RMSE (i.e., the average over all ten folds) consistently increases. We then tried values of λ between the best two observed until no further reduction in RMSE is found. We did this in parallel for all seven nonlinear models, as well as the linear model, for eight models in all. From these eight models, we selected the one with the lowest mean RMSE across all the folds with the best corresponding value of λ. Note that we could select the linear model as the best model; we would run this model for comparison purposes in any case. Finally, the final coefficients were learned from the entire training set (i.e., no cross-validation) using this chosen model and value of λ.

5 Experimental Results

Table 2. Non-personalized models results over 1000 trials

Method	RMSE Mean	RMSE Std. Dev.	RMSE Median	Mean RMSE Reduction	Possible RMSE Reduction Achieved
			Yow-now		
$\mathbf{W}_0$ only	0.2507	0.00408	0.2508	-36.99%	-1281.63%
Lower Bound	0.1830	0.00352	0.1829	2.61%	100.00%
Linear	0.1879	0.00362	0.1880	0.00%	0.00%
Nonlinear	0.1853	0.00117	0.1852	1.38%	52.74%
			Epinions.com		
$\mathbf{W}_0$ only	0.2225	0.00911	0.2223	-10.09%	-83.27%
Lower Bound	0.1776	0.00665	0.1774	12.12%	100.00%
Linear	0.2021	0.00659	0.2021	0.00%	0.00%
Nonlinear	0.2008	0.00700	0.2005	0.64%	5.31%

We tested for interactions among criteria by contrasting the observed RMSE of our criteria interaction models with that of the linear model. To decrease the possibility for random misleading effects, we ran the experiment 1000 times (i.e., 1000 trials). The test set was randomly chosen from the full dataset each time, which means different trials will have different training sets and testing sets, and a single item is likely to serve as both training and test data (but in different trials; no testing data is ever included in training data). This is valid because all of our modeling choices (regularization tuning and learning model coefficients) are done solely on the basis of the training data.

Table 2 shows the RMSE results on both datasets without personalization. Four methods are reported: RMSE results using the regularization vector $\mathbf{W}_0$ only (equivalent to setting regularization parameter λ to infinity); the lower bound on RMSE; the learned linear combination; and the learned nonlinear combination. The mean RMSE reduction shows how much the RMSE decreased as a percentage of the learned linear combination. However, the lower bound is very close to the linear combination, so there is not much potential for RMSE reduction. The possible RMSE reduction shows how much of this potential RMSE reduction is achieved; by definition, it is always 100% at the lower bound.

The nonlinear model has a lower RMSE for both datasets, but the difference is very small. This is not surprising as the RMSE for the linear model is quite close to the lower bound. The possible RMSE tells a different story. For the Yow-now model, over half of the possible RMSE reduction was achieved with the nonlinear model. For the Epinions.com model, much less of the possible RMSE reduction was achieved. The smaller dataset size may have played a role,

Table 3. Personalized Yow-now model results over 1000 trials

User	Articles	Most Sel.	Most Pct	RMSE Mean	RMSE Std. Dev.	RMSE Median	Mean RMSE Reduction	Possible RMSE Reduction Achieved
u51	305	⟨2,4⟩	46%	0.1654	0.01761	0.1654	-1.45%	-18.38%
u56	362	⟨2,4⟩	53%	0.1277	0.01274	0.1272	1.86%	13.56%
u58	569	B	31%	0.2032	0.01204	0.2033	-1.25%	-19.77%
u59	358	C	66%	0.1280	0.00943	0.1281	0.41%	7.32%
u60	138	⟨1,2⟩	48%	0.1488	0.02475	0.1440	-1.23%	-6.45%
u62	161	B	42%	0.1065	0.01161	0.1065	-2.86%	-21.18%
u63	472	⟨2,4⟩	60%	0.1089	0.01522	0.1092	-1.89%	-15.59%
u65	607	⟨3,4⟩	76%	0.1347	0.01537	0.1329	4.53%	34.26%
u66	443	⟨2,4⟩	53%	0.1487	0.01723	0.1472	0.85%	4.72%
u67	590	⟨2,4⟩	96%	0.2344	0.01240	0.2345	5.23%	45.00%
u68	388	B	57%	0.1455	0.00932	0.1453	-1.11%	-21.74%
u69	848	C	82%	0.1772	0.00772	0.1770	1.01%	13.44%
u73	232	B	33%	0.1678	0.01700	0.1677	-0.68%	-5.95%
u74	14	⟨3,4⟩	33%	0.1969	0.07209	0.1810	1.42%	2.93%
u76	603	⟨1,2⟩	58%	0.0888	0.00834	0.0888	-1.14%	-17.68%
u80	218	⟨3,4⟩	55%	0.2795	0.03244	0.2802	0.00%	0.03%
u82	516	C	97%	0.1064	0.01385	0.1062	9.98%	55.37%
u83	1079	⟨2,4⟩	73%	0.0960	0.00473	0.0960	3.06%	38.19%
u84	426	⟨2,4⟩	43%	0.2318	0.01475	0.2321	-0.89%	-12.35%
u87	129	B	68%	0.1740	0.02131	0.1734	-0.99%	-7.49%
u88	54	⟨2,4⟩	56%	0.1704	0.03602	0.1689	-2.02%	-4.51%
u91	367	C	59%	0.2212	0.01646	0.2215	0.46%	3.79%
u92	310	B	21%	0.1557	0.01097	0.1557	-2.74%	-21.73%
micro				0.1539			1.21%	8.42%
macro				0.1616			0.46%	1.99%

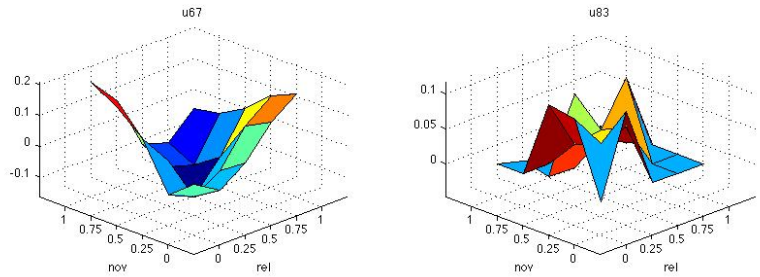

Fig. 1. Interactions of the ⟨novelty,relevance⟩ criteria pair for users *u67* and *u83*

as less data will tend to produce poorer learned models but also a lower lower bound (because there are less opportunities for inconsistent ratings).

Table 3 shows the results when a separate model is learned for each user (personalized models), as well as the microaverage and macroaverage. Results are generally poorer for users with less data. Performance varies a lot among users: mean RMSE ranges from 0.0888 to 0.2795; mean RMSE reduction ranges from -2.86% to 9.98%; and the percentage of possible RMSE reduction achieved ranges from -21.73% to 55.37%. In fact, a slight majority of users had negative results with respect to the baseline. Comparing Table 3 with Table 2, we can see that the microaverage over the personalized models is lower than even the lower bound on the non-personalized model. This shows that there was considerable differences among user models, and thus personalization reduced RMSE.

Table 3 also shows the most frequently selected nonlinear model (Most Sel.) for each user. Due to space limit, the criteria are numbered as 1 (authority), 2 (novelty), 3 (readability) and 4 (subject relevance). For example, our method selected the novelty and subject relevance pair (listed as ⟨2,4⟩) for user *u51* in 46% of the trials. Additionally, B indicates the basic linear model (no criteria interactions) and C indicates the *all-pairs* nonlinear model. The same model was not always selected for the same user on every trial, because it was dependent on the trial's randomly selected training set. Users that showed mostly a linear trend had an increase in RMSE because overfitting occurred when a nonlinear model was selected. Also, users that did not show a consistent preference for a particular form also had an increase in RMSE, for similar reasons.

From Table 3, we see that a variety of criteria interact in the personal models, and in fact each pair was selected at least once on some trial. However, some pairs tend to interact more than others. For the non-personalized models, the *all-pairs* nonlinear form was always selected for the Yow-now dataset, while the ⟨sound,picture quality⟩ pair was selected 82% of the time for the Epinions.com dataset. For the personalized Yow-now models, ⟨novelty,relevance⟩ was the most commonly selected pair, and indeed along with ⟨readability,relevance⟩ and the *all-pairs* nonlinear form accounted for all mean reductions in RMSE.

Figure 1 show the mean learned interactions for users *u67* and *u83*, who had some of the largest RMSE reductions. Though our method consistently selected ⟨novelty,relevance⟩ for both users, the learned interactions were quite different. The plot for *u67* has a smooth surface, with an upward adjustment for higher values on either of the criteria while the other criterion remains low. On the other hand, *u83*'s plot has no such easily interpretable pattern, which was also true for most users. More research is needed to understand these interactions.

6 Conclusions and Future Work

Our results show that interactions among criteria exist in criteria-based information retrieval models, at least in some cases, as measured by an observed reduction in RMSE. We observed this reduction in both non-personalized and personalized models. However, the amount of RMSE reduced by exploiting interactions was slight in the datasets we used; in fact, it often *increased* RMSE, but the magnitude of the reduction for some users outweighed the increases for the rest. Personalization was more clearly beneficial. In terms of the interactions themselves, certain criteria had consistently stronger observed interactions than others, but we could not discern an interpretable pattern in these interactions.

Despite our use of regularization, overfitting remained a problem, as evidenced by the occasional *increase* in RMSE over the linear model. This could potentially be avoided by opting for the linear model when there is insufficient evidence for interactions (i.e., when the reductions are not consistently observed in the training data, or not large enough relative to the training set size). This could be expanded to a Bayesian framework, using prior probabilities to avoid selecting less probable models when there is not sufficient support in the data. Even without these improvements, in our experiments we were successful in reducing the overall mean RMSE by exploiting criteria interactions and personalization.

References

1. J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2003.
2. C. Barry. User-defined relevance criteria: An exploratory study. In *Journal of the American Society for Information Science, 45(3)*, 1994.
3. G. Bordogna and G. Pasi. A multi criteria news filtering model. In *Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6, 2008.
4. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference*, 1998.
5. U. Chajewska and D. Koller. Utilities as random variables: Density estimation and structure discovery. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 63–71, 2000.
6. H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06: Proceedings of the 29th annual international*

ACM SIGIR conference on Research and development in information retrieval, pages 429–436, New York, NY, USA, 2006. ACM Press.

7. M. de Gemmis, G. Semeraro, P. Lops, and P. Basile. A retrieval model for personalized searching relying on content-based user profiles. In *6th AAAI Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2008)*, pages 1–9, 2008.
8. M. Farah and D. Vanderpooten. A multicriteria paradigm of relevance for the web information retrieval problem. In *IEEE Sciences of Electronic, Technologies of Information and Telecommunications*, 2005.
9. M. Farah and D. Vanderpooten. An outranking approach for rank aggregation in information retrieval. In *30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 591–598, 2007.
10. K. Maglaughlin and D. Sonnenwald. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. In *Journal of the American Society for Information Science and Technology*, 2003.
11. N. Manouselis and C. Costopoulou. Analysis and classification of multi-criteria recommender systems. In *World Wide Web*, volume 10, pages 415–441, 2007.
12. H. Nguyen and P. Haddawy. The decision-theoretic video advisor. In *AAAI-98 Workshop on Recommender Systems*, pages 77–80, 1998.
13. G. Pasi, G. Bordogna, and R. Villa. A multi-criteria content-based filtering system. In *30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 775–776, 2007.
14. M. Pognačnik, J. Tasič, and A. Košir. Optimization of multi-attribute user modeling approach. *International Journal of Electronics and Communications*, 58(4):402–412, 2004.
15. V. Robu, D. J. A. Somefun, and J. A. La Poutré. Modeling complex multi-issue negotiations using utility graphs. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 280–287, New York, NY, USA, 2005. ACM.
16. L. Schamber and J. Bateman. User criteria in relevance evaluation: Toward development of a measurement scale. In *ASIS 1996 Annual Conference Proceedings*, October 1996.
17. A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *J. Am. Soc. Inf. Sci. Technol.*, 56(4):327–344, 2005.
18. E. Triantaphyllou. *Mutli-Criteria Decision Making Methods: A Comparative Study.* Kluwer Academic Publishers, 2000.
19. P. Wang. *A cognitive model of document selection of real users of IR Systems.* PhD thesis, University of Maryland, 1994.
20. S. R. Wolfe and Y. Zhang. User-centric multi-criteria information retrieval. In *32st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09).* ACM, 2009.
21. Y. Zhang. Yow user study data: Implicit and explicit feedback for news recommendation. http://www.soe.ucsc.edu/~yiz/papers/data/YOWStudy/, last visited in Jan. 2009.
22. Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.