urCF: User Review Enhanced Collaborative Filtering

Completed Research Paper

Zhenxue Zhang
University of Maryland, Baltimore County
zzhang3@umbc.edu

Dongsong Zhang
University of Maryland, Baltimore County
zhangd@umbc.edu

Jianwei Lai
University of Maryland, Baltimore County
jianwei1@umbc.edu

Abstract

Despite of success in both research and industry, traditional collaborative filtering (CF) based recommender systems suffer from a fundamental problem, which lies in its dependence on users' numeric ratings as its sole source of user preference information. User ratings are often unable to fully represent user preferences. As a result, it is difficult and error prone to identify genuinely similar users based on ratings only. On the other hand, online consumer product reviews have become a common source for consumers to share and acquire information about products, but there have been very few studies on how those text reviews can be analyzed and integrated with traditional CF approaches to improve the prediction of consumers' preferences. We propose a novel approach to memory-based collaborative filtering called urCF (User Review enhanced Collaborative Filtering) that integrates user text reviews and user numeric ratings in order to model users' preferences better and in turn improve the performance of CF-based recommender systems. This research extracts user opinions on individual item features from online reviews, and proposes a new weighting scheme by following the general idea of TF-IDF to measure the priority of item features in influencing users' overall opinions on different items. This study also explores and compares two different methods for integrating user opinion into user similarity measurement. The proposed urCF system is evaluated against existing approaches using a dataset collected from Yahoo! Movies. The results show that urCF significantly improves the performance of memory-based CF systems.

Keywords

Opinion Mining, Sentiment Analysis, Collaborative Filtering, Recommender Systems

Introduction

Collaborative filtering automates the word-of-mouth recommendation process, in which people share their preferences on items (e.g., products, movies) among friends to help each other find preferable items. In general, there are two major approaches to collaborative filtering, namely memory-based CF and model-based CF (Breese et al. 1998). Memory-based CF systems utilize an original, entire user-item rating matrix to generate predictions (Resnick et al. 1994), while model-based CF methods recommend items by first developing a descriptive model of user ratings based on a user-item matrix via different machine learning approaches such as Bayesian network and clustering. The generated model is then used for future prediction about user preferences (Breese et al. 1998). Although model-based CF methods overcome some shortcomings of memory-based counterparts, such as low scalability and high online computation overhead, some studies show that they are generally inferior to memory-based ones in terms of prediction accuracy (Adomavicius et al. 2008). This study focuses on improving memory-based CF.

Traditional CF based recommender systems suffer from a fundamental problem because of its dependence on users' numeric ratings as its sole source of user preference information. However, user ratings alone may not be able to fully reflect a user's actual preferences. For instance, two users in a

recommender system, Alice and David, both favor the movie *Titanic* and give it a high rating score. However, there may be different reasons behind their favorable ratings. Alice may like this movie because of its well-fabricated love story, while David prefers the movie due to its glamorous recreation of the ship and visual effects. As a result, CF systems may not be able to accurately identify similar users based solely on their ratings on items, resulting in potentially poor recommendations.

With the increasing popularity of Web 2.0, users have become more and more comfortable with expressing themselves and providing their opinions on the Internet using text (Chen et al. 2007). Such consumer reviews have potential to provide a system with more detailed, nuanced, and reliable user preference information. In other words, user text reviews can be used, in conjunction with numerical ratings, to augment the *word-of-mouth* recommendation process.

Recently, some approaches have been proposed to aggregate user preference information inferred from user item reviews for recommendation purposes (Jakob et al. 2009; Leung et al. 2006; Liu et al. 2005; Wang et al. 2012). Results of these prior studies have shown a positive impact of user text reviews on the performance of traditional CF systems. However, those approaches are model-based and adopt reviews for the purpose of creating user profiles, instead of directly integrating them into collaborative filtering itself (Jakob et al. 2009; Wang et al. 2012). urCF, our proposed method, extracts user opinions expressed in online user reviews and directly integrates them into memory-based CF, along with user item ratings, aiming to better identify genuinely similar neighbors and help generate more accurate recommendations.

There are several unique contributions of this research. First, to the best of our knowledge, this is the first study to integrate feature-oriented user text reviews into memory-based collaborative filtering. Second, we propose a novel user feature priority weighting scheme, called FF-IRF (Feature Frequency and Inverse Review Frequency), to reflect the importance of different item features to the overall opinion of a user toward an item. Third, this research also examines two methods to integrating user ratings and text reviews for user opinion identification at different levels of information comprehensiveness.

The remainder of this paper is organized as follows. We will first introduce related work, followed by a detailed description of the proposed urCF system. Next, we present how the proposed approach was evaluated and results. Finally, the paper concludes with discussions on the major findings, limitations of the study, and future research.

Related Work

A memory-based CF system involves three steps: user similarity measurement, neighborhood selection, and prediction generation (Resnick et al. 1994). Pearson's Correlation Coefficient (PCC) is the most widely used technique for measuring user similarity (Resnick et al. 1994). It is usually used in a general form as follows (Resnick et al. 1994):

$$w_{x,y} = \frac{\sum_{i \in I_{x,y}} (r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i \in I_{x,y}} (r_{x,i} - \bar{r}_x)^2 \sum_{i \in I_{x,y}} (r_{y,i} - \bar{r}_y)^2}}, \quad (1)$$

where $I_{x,y}$ denotes the set of all items that both users *x* and *y* have rated; $r_{x,i}$ stands for the rating cast by a user *x* for an item *i*; $\bar{r}_x$, the mean rating of user *x*, is defined as:

$$\bar{r}_x = \frac{1}{|I_x|} \sum_{i \in I_x} r_{x,i}, \quad (2)$$

where I_x denotes the set of all items that user x has rated.

Once user similarity weights are calculated, a subset of similar users (i.e., the neighborhood) will then be selected for generating final prediction for an active user (i.e., the user that the system intends to generate a prediction for). Two methods are commonly used for neighborhood selection: similarity thresholding and best-*N*-neighbors. Similarity thresholding uses a certain threshold value *L* to filter out users whose similarity with the active user is lower than a threshold value *L* (Shardanand and Maes 1995). The best-*N*-

neighbors method (Herlocker et al. 1999) considers the top *N* users that are most similar to an active user for generating predictions.

After the neighborhood of the active user is identified, a prediction of preference for the active user *a* on an item *i*, $pred_{a,i}$, is computed as the weighted average of deviations from his neighbor's mean rating (Resnick et al. 1994):

$$pred_{a,i} = \overline{r}_a + \frac{\sum_{x \in N} w_{a,x} \times (r_{x,i} - \overline{r}_x)}{\sum_{x \in N} |w_{a,x}|}, \quad (3)$$

where *N* stands for the set of users who are selected as the neighbors of the active user.

There have been several approaches proposed to integrate user reviews into recommender systems (Jakob et al. 2009; Leung et al. 2006; Levi et al. 2012; Moshfeghi et al. 2011; Wang et al. 2012). The first research effort that used user text reviews in a CF system was made by (Leung et al. 2006). They proposed a probabilistic rating inference model that estimates both sentimental orientation and opinion intensity of movie reviews. The output of the rating inference model is then used in a CF system to replace original user ratings in order to test the performance of their proposed model (Leung et al. 2006). The study focused on how to extract user opinion information from online user reviews instead of integrating user feature opinions into CF.

Similar to this research, Jakob et al. (2009) proposed a model-based hybrid CF system that builds user profiles in a multi-relational model based on user opinions on item features extracted from text reviews. User profiles in their system are created based on an entity relationship model to reflect the interactions between users' feature opinion orientations and specific movies. A matrix factorization approach, MRMF (Multi-Relational Matrix Factorization), is used to decrease the dimensionality of the multi-relationship matrix resulting from the user profiling step of the system (Jakob et al. 2009). Finally, a machine learning method is used to generate final predictions. They evaluated the system using data collected from IMDB and achieved the best performance improvement of 2.45% over the baseline. (Wang et al. 2012) applied the same user modeling approach proposed in (Jakob et al. 2009) without the assumption that the relationship between user ratings and feature opinions has to be linear. Wang et al. adopted tensor factorization for matrix decomposition.

Another two recent studies have also proposed to use text reviews to address the cold start problem (Levi et al. 2012; Moshfeghi et al. 2011). Moshfeghi et al. used user feature opinion information extracted from movie plot summaries and reviews from IMDB to create user profiles based on their ratings on different movies (Moshfeghi et al. 2011). User opinions on three specific features (actor, director and genre) are used for user profiling. LDA (Latent Dirichlet Allocation) and gradient boosted trees are used for dimensionality reduction and prediction generation purposes, respectively (Moshfeghi et al. 2011). Their system was evaluated using the MovieLens dataset. In (Levi et al. 2012), a hotel recommender system was proposed to match users based on their intent (e.g., business trip, vacation, etc.) and nationality. User opinions extracted from text reviews are then used to generate recommendations for users with the same context (i.e., intent and nationality). Researchers evaluated the system empirically and received positive feedback from users (Levi et al. 2012). Those two studies intended to solve the cold start problem by incorporating user text reviews into CF, while this research intends to adopt user reviews as an extra source of information to help enhance the performance of CF systems. The results from existing approaches show positive support that user text reviews can be used to improve the prediction accuracy of collaborative filtering.

User Review Enhanced Collaborative Filtering

Different from these two studies, we propose urCF, a new method that incorporates user reviews into a memory-based collaborative filtering system. Memory-based CF methods are the best candidate for exploring different methods for integrating user text reviews into CF, because they are easy to predict, control, and fine-tune. Figure 1 shows the architecture of the proposed urCF system. In the remainder of this section, each component of the system will be described in detail.

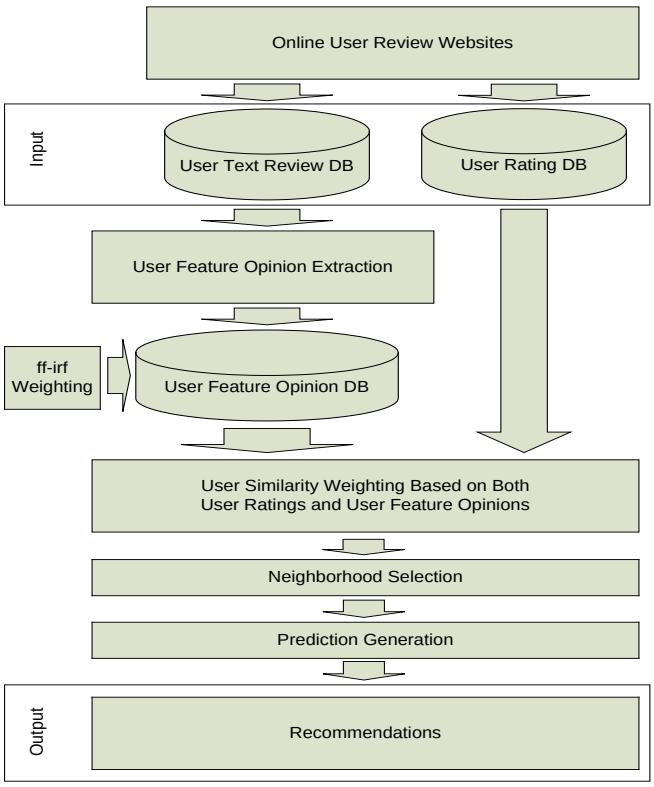

Figure 1. urCF System Architecture

System Input

A traditional memory-based CF system has a user-item rating matrix as its only input. In urCF, user text reviews on items are also included as part of the system input. Text reviews, which can be collected from specific web sites by an Internet crawler, are stored alongside corresponding user ratings. Therefore, a rating matrix in a traditional CF system will be expanded into a new matrix containing both user ratings and reviews. Only data of users who have provided both ratings and reviews for the same items are used.

User Feature Opinion Extraction

Since user item reviews are in free text and unstructured in nature, they cannot be used directly in CF systems. There needs a process to transform a text review $R_{x,i}$ from user x on item i to a structured representation. By adapting common practices (Liu et al. 2005; Zhou and Chaovalit 2008), we extracted a user's opinion orientations on item features from his/her text reviews at a sentence level in this research.

Because the goal of this study is not to improve existing text mining techniques, but to study effective ways to integrate user text reviews with traditional memory-based CF to improve its performance, we adopted a semi-automated approach by allowing human coders to perform feature extraction with a tool.

By adapting the movie review ontology proposed in (Zhou and Chaovalit 2008), we developed a movie review class consisting of 32 features (shown in Table 1) to guide the review coding process and organize identified user opinions on item features. Two human coders were recruited and trained to extract user feature opinions from text reviews based on the feature list shown in Table 1. User reviews were first pre-processed for misspelling correction and sentence segmentation through a user interface. The coders were then asked to extract item features discussed in each review sentence and user opinion orientations (positive or negative) expressed on these features using the interface. The output of this step is a matrix with every user's feature opinion orientation of each item on which he/she has written a review. Every entry in the matrix stores the opinion orientation $O_{x,s,f_{n,i}}$ of user x on the n^{th} feature $f_{n,i}$ of item i in sentence s of his/her review.

General	Actor	Acting	Character	Cast
Director	Directing	Plot/Story	Theme	Writer
Writing	Dialog	Adaptation	Editing	Pacing
Length	Scene	Cinematography	Visual Effects	Sound Effects
Music/Score	Soundtrack	Original Music Composer	Action	Costume
Makeup	Effect on Viewer	Recommendation	Animation	Genre
Trailer	Script			

Table 1. Movie Review Features Used in urCF

User Feature Priority Weighting

Opinions on different features of an item may have different influences on a user's overall attitude toward the item. In order to model users' opinion and distinguish users more effectively, we propose a user feature priority weighting method called FF-IRF (Feature Frequency–Inverse Review Frequency) based on the TF-IDF weighting scheme widely used in Information Retrieval (IR) literature (Salton and McGill 1983). The underlying notion of the proposed weighting method is that if a user uses a large portion of his/her review to express his/her opinion on a specific feature of an item, it implies that the commented feature plays a significant role in forming the user's overall perception of the item and, therefore, should receive a higher importance weight when the information is used to distinguish users. On the other hand, features of an item that are widely discussed by many users in their reviews should be less useful in separating users apart than features appearing only in a few reviews. In other words, the fact that two users agree on a rarely discussed feature of an item is more helpful in grouping the two together than the situation where two users agree on popular features. Hence, weights for features of an item that are mentioned by fewer users should be increased. FF-IRF consists of two parts: Feature Frequency (FF) and Inverse Review Frequency (IRF).

Feature Frequency (FF): According to the assumption explained earlier, in a user review, the length of the text that describes a user's opinion on a specific feature implies how important the feature is in shaping his/her overall attitude toward the item. However, the orientation of the opinion that a user expresses has to be considered as well. Since a user's opinion on an item feature expressed in different parts of a review may not always have the same orientation, feature frequency is designed based on a cumulative method to accommodate both the length of the review and the orientation of the opinion.

Since user feature opinion extraction is performed on each review sentence, we first define word count $W_{x,s}$ of a review sentence as the number of non-stop words in a sentence s of his/her review $R_{x,i}$. A review sentence may discuss multiple item features. Therefore, the word count $W_{x,s,f_{n,i}}$ for n^{th} feature $f_{n,i}$ of item i in sentence s is defined as the average word count of all item features on which the user expressed his/her opinions in that sentence and can be calculated as follows:

$$W_{x,s,f_{n,i}} = \frac{W_{x,s}}{\left|F_{x,s}\right|}, \tag{4}$$

where $F_{x,s}$ denotes the set of item features, on which user x commented in sentence s. Therefore, $\text{ff}_{x,f_{n,i}}$ from user x on feature $f_{n,i}$ in a review $R_{x,i}$, is defined as follows:

$$\text{ff}_{x,f_{n,i}} = \sum_{s \in R_{x,i}} (W_{x,s,f_{n,i}} \times O_{x,s,f_{n,i}}), \tag{5}$$

where $O_{x,s,f_{n,i}}$ is defined as the opinion orientation on review feature $f_{n,i}$ of item i discussed in review sentence s.

Inverse Review Frequency (IRF): Similar to term weighting in IR research, not all item features should be treated equally when it comes to distinguishing users, even though they may have the same FF

values. Features of an item that are mentioned only in a few reviews should have a higher weight because they are more useful in differentiating users, while those that are widely discussed should have a lower weight. The weighting scheme IRF is defined as follows:

$$\text{irf}_{f_{n,i}} = \left(\frac{N_i}{\text{rf}_{f_{n,i}}} \right)^{\alpha}, \tag{6}$$

where N_i denotes the total number of reviews on an item i in the user-item review matrix and $\text{rf}_{f_{n,i}}$ is defined as the number of reviews on i in the matrix that have commented on feature $f_{n,i}$.

In IR, a logarithm is used in idf calculation (Furner 2002). However, we apply an exponentiation treatment in calculating IRF because the scale of $\frac{N_i}{\text{rf}_{f_{n,i}}}$ ratio is very different from that of $\frac{N}{\text{df}_t}$ in IR, where N stands for the total number of documents in a document collection and df_t is the number of documents in the collection that contain a term t. In IR, the total number of documents N is usually in the millions and a typical $\frac{N}{\text{df}_t}$ ratio is in the order of tens of thousands (Furner 2002).

Therefore, the logarithm treatment used in idf is to compress the value non-linearly to increase its efficiency (Furner 2002). However, in CF domain, N_i is the number of reviews on a specific item i, not the whole collection of reviews in the dataset. N_i is typically in the hundreds or less. In addition, the IRF value needs to be in a range that is comparable with the magnitude of FF, so that the resulted FF-IRF weighting will not be dominated by either FF or IRF. Exponent α in Equation (6) will be determined empirically based on the distributions of both $\left|\text{ff}_{x,f_{n,i}}\right|$ and $\frac{N_i}{\text{rf}_{f_{n,i}}}$ in the dataset used for evaluation.

Finally, feature frequency and inverse review frequency are combined to generate a composite weight $\text{ff-irf}_{x,f_{n,i}}$ for each feature in each review. Therefore, user feature opinion $p_{x,f_{n,i}}$ can be shown as follows:

$$p_{x,f_{n,i}} = \text{ff-irf}_{x,f_{n,i}} = \text{ff}_{x,f_{n,i}} \times \text{irf}_{f_{n,i}}. \tag{7}$$

Once $p_{x,f_{n,i}}$ is calculated for each feature discussed in all reviews on all items in the system, a user-item feature opinion matrix will be generated. This is a $U \times F$ matrix in which rows represent individual users and columns corresponding to the collection of all features of every item. This user-item feature opinion matrix can be combined with the user-item rating matrix because user ratings and feature opinion weights share the same user-item pairs. A simplified example of the combined matrix is show in Table 2.

Movies	**Titanic**				**Star Wars III**				**Shrek**				**Harry Potter**				**Minority Report**			
Users	*Act.*	*Dir.*	*Plot*	*Vis.*	*Act.*	*Dir.*	*Plot*	*Vis.*	*Act.*	*Dir.*	*Plot*	*Vis.*	*Act.*	*Dir.*	*Plot*	*Vis.*	*Act.*	*Dir.*	*Plot*	*Vis.*
Alice	5								4				5				2			
	8		10								7	5.2		6	6.3		7	-3	-7	
Bob	4				5				5				1				5			
	-5.2	6.5		4		5.5	6.3	4			5	6.7	-5		-10		7	10	20	15
Clark	3				5				5				2							
	-5		-6.5	10	7		6.2				4	6	-4.5		-8	8.7				
David	5				4								4				4			
	6.5	7		8.7	-4		8.3	6.5					4		-6	7	-6		3	7
Eve	5				1								5				1			
	18		20				-15						15	6	20	10			-20	

Table 2. A Combined User-Item Matrix with Both Rating and Feature Opinion (Simplified)

User Similarity Weighting Based on Both User Rating and User Feature Opinion

Different methods have been proposed to improve accuracy of CF by applying weights to different items when users' similarity is measured (Breese et al. 1998; Herlocker et al. 1999). In this study, a similar

strategy is used to integrate user feature opinions into traditional CF at an item level by applying user similarity weights based on feature opinions as item weights in Pearson Correlation Coefficient (PCC) calculation.

The assumption behind the item weighting approach is that if two users' feature opinion similarity on a specific item coincides with their user rating similarity on the same item, then the users' rating similarity on the item is supported by the feature opinions expressed in their text reviews on the same item. The item, therefore, deserves a higher weight when it is used to compute the similarity between the two users. On the other hand, if the feature opinion similarity between two users differs from their rating similarity on an item, a lower weight will be used on the item, because their rating similarity is contradicted with the feature opinions discussed in their reviews on the same item.

In order to apply item weighting to user similarity measurement, we replace the regular PCC with a weighted version as follows:

$$w_{x,y} = \frac{\sum_{i \in I_{x,y}} \mu_{x,y,i}(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\sqrt{\sum_{i \in I_{x,y}} \mu_{x,y,i}(r_{x,i} - \overline{r}_x)^2 \sum_{i \in I_{x,y}} \mu_{x,y,i}(r_{y,i} - \overline{r}_y)^2}}, \tag{8}$$

where

$$\overline{r}_x = \frac{\sum_{i \in I_{x,y}} (\mu_{x,y,i} \times r_{x,i})}{\sum_{i \in I_{x,y}} \mu_{x,y,i}} \tag{9}$$

and

$$\mu_{x,y,i} \geq 0 . \tag{10}$$

In Equation (8), $\mu_{x,y,i}$ is the item weighting factor, which can be considered as an item frequency (Bills and Li 2005). In this study, to test the robustness of different types of user opinion information when being used to enhance performance of CF systems, we propose two different methods to calculate $\mu_{x,y,i}$ based on different information that can be derived from the user-item feature opinion matrix.

Method 1 (M1) In this method, only the orientation of user feature opinion $p_{x,f_{n,i}}$ is used in the calculation of $\mu_{x,y,i}$. First, an initial item weighting value $\mu'_{x,y,i}$ is calculated as follows:

$$\mu'_{x,y,i} = \sum_{f_{n,i} \in F_{x,y,i}} (O_{x,f_{n,i}} \times O_{y,f_{n,i}}) \times \frac{(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\left|(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)\right|}, \tag{11}$$

where $O_{x,f_{n,i}}$ and $O_{y,f_{n,i}}$ are the orientations (i.e., positive or negative signs) of user feature opinions $p_{x,f_{n,i}}$ and $p_{y,f_{n,i}}$, respectively. $F_{x,y,i}$ stands for the set of all features of *i* that both *x* and *y* discussed in their reviews on item *i*. The binary result (+1 or -1) of the multiplication of $O_{x,f_{n,i}}$ and $O_{y,f_{n,i}}$ indicates whether users *x* and *y* agree or disagree on their opinion orientations on feature $f_{n,i}$ of item *i*. In Equation (11), the $\frac{(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\left|(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)\right|}$ part returns either +1 or -1, which indicates whether ratings from *x* and *y* on *i* are in agreement or not, respectively, in their orientations relative to the means of their overall ratings. A positive (or negative) value of $\mu'_{x,y,i}$ indicates that reviews from *x* and *y* on item *i* are consistent (or inconsistent) with their ratings in terms of their opinion orientation, and the magnitude of $\mu'_{x,y,i}$ shows the strength of such consistency. A negative $\mu'_{x,y,i}$ means that the ratings do not at all represent genuine positive or negative correlation between users and will undermine the accuracy of the weighted PCC

measures. Therefore, these ratings should be excluded from the PCC calculation. The following condition check is used to obtain the final item weight $\mu_{x,y,i}$ in Method 1:

$$\mu_{x,y,i} = \begin{cases} 0 & if\ \mu'_{x,y,i} < 0 \\ \mu'_{x,y,i} + 1 & if\ \mu'_{x,y,i} \geq 0 \end{cases}. \quad (12)$$

In Equation (12), default value 1 is added to $\mu'_{x,y,i}$ when it is larger or equal to zero. The treatment is used so that when $\mu'_{x,y,i}$ is zero, the ratings from the two users on the same item will still be used in computing the user similarity measure, though with the lowest weight.

Method 2 (M2) Although M1 has taken into consideration the consistencies between user reviews and ratings when calculating item weighting, the FF-IRF part of the user-item feature opinion matrix is not fully included in the calculation. In comparison, Method 2 uses user feature opinion in its complete form to compute $\mu_{x,y,i}$.

Before proposing Method 2, we first discuss user similarity measurement based on user feature opinions. In urCF, PCC is used to measure user similarity based on user feature opinions of a single item *i*. Therefore, similarity weight $\text{sim}^f_{x,y,i}$ between the feature opinions of user *x* and *y* on item *i* is defined as follows:

$$\text{sim}^f_{x,y,i} = \frac{\sum_{f_{n,i} \in F_{x,y,i}} (p_{x,f_{n,i}} - \overline{p}_x)(p_{y,f_{n,i}} - \overline{p}_y)}{\sqrt{\sum_{f_{n,i} \in F_{x,y,i}} (p_{x,f_{n,i}} - \overline{p}_x)^2 \sum_{f_{n,i} \in F_{x,y,i}} (p_{y,f_{n,i}} - \overline{p}_y)^2}}. \quad (13)$$

Value of $\text{sim}^f_{x,y,i}$ ranges from -1 to 1. A positive (or negative) value of $\text{sim}^f_{x,y,i}$ means a positive (or negative) correlation between *x* and *y* over their feature opinions. Similar to M1, $\text{sim}^f_{x,y,i}$ also needs to be compared to users' opinion orientations in order to check whether they are consistent with each other or not. First, an initial item weight value $\mu'_{x,y,i}$ is calculated as follows:

$$\mu'_{x,y,i} = \text{sim}^f_{x,y,i} \times \frac{(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\left|(r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)\right|}. \quad (14)$$

Then, the same condition check shown in Equation (12) will be used to determine the item weight $\mu_{x,y,i}$.

The aggregation function (Equation (3)) is used for prediction generation based on the similarity weight $w_{a,x}$ (calculated using Equation (8)) between the active user *a* and each user *x* who has rated and reviewed the same items.

Evaluation

Data

We chose movies as the domain for evaluating the proposed urCF system. Movie domain is rich with content. More importantly, the enthusiasm of the general population toward movies and their willingness to express their opinions on movies make movie domain one of the mostly used in CF research (Breese et al. 1998; Resnick et al. 1994). Yahoo! Movies was chosen as the data source in this research and we collected user movie ratings and text reviews dated from July 2003 to January 2012 via a Web crawler. The collected dataset was very sparse (the sparsity level is higher than 99.99%) and it cannot be directly used for evaluation purpose. Following the common practice in CF research (Sahoo et al. 2006), we first pre-processed the whole dataset and retained users who had rated at least 50 movies and movies that had received at least 50 ratings. The resulted dataset contained a total of 49,562 ratings/reviews on 735 different movies from 544 unique users. The sparsity of the dataset was 87.60%. Due to time constraint, a

smaller subset of 53 movies and 41 users were randomly selected from the preprocessed dataset to evaluate the proposed system. The subset has 354 reviews in total and an 83.71% sparsity level.

In the Yahoo! Movies dataset, movie ratings are on a 13-point scale (A+ to F). Following the common practice in existing CF research (Sahoo et al. 2006), we converted the Yahoo! Movies rating scale to a new scale ranging from 1 to 5, which is commonly used by other datasets (Sahoo et al. 2006).

As discussed earlier, the exponent value α in the IRF calculation is selected empirically based on the distributions of both $\left|\mathrm{ff}_{x,f_{n,i}}\right|$ and $\frac{N_i}{\mathrm{rf}_{f_{n,i}}}$ in the dataset used for evaluation. Therefore, we first ran the urCF system to record every $\left|\mathrm{ff}_{x,f_{n,i}}\right|$ and $\frac{N_i}{\mathrm{rf}_{f_{n,i}}}$ values for every item i and every feature $f_{n,i}$. It was found that a majority (95%) of $\left|\mathrm{ff}_{x,f_{n,i}}\right|$ and $\frac{N_i}{\mathrm{rf}_{f_{n,i}}}$ values ranged from 0 to 59 and 1 to 2.8, respectively. In order to amplify the majority of $\frac{N_i}{\mathrm{rf}_{f_{n,i}}}$ values to a range that is closest to that of $\left|\mathrm{ff}_{x,f_{n,i}}\right|$, we selected a value of 4 for α in the IRF calculation for the dataset.

Baseline

The proposed urCF system with two methods for calculating $\mu_{x,y,i}$ was evaluated by comparing its predictive performances versus those of the traditional memory-based CF algorithm as the baseline. A commonly used CF evaluation task named "all-but-one" was adopted for the evaluation (Breese et al. 1998; Herlocker et al. 1999). The testing was executed for each user-item pair in the dataset and predictions were generated by using the baseline method and the methods proposed in urCF. Due to the limited size of the dataset used for evaluation, all neighbors of the active user were used in prediction generation. In another word, similarity thresholding was used for neighborhood selection with value L set to zero.

Measures

Mean Absolute Error (MAE) was used as the metric to evaluate the accuracy of two approaches. The metric is defined as the average absolute deviation of the predictions generated by a system on how users would rate different items compared to the actual ratings on those items cast by users. If $pred_{x,i}$ is the predicted rating generated for user x on item i and $r_{x,i}$ is his/her actual rating, MAE can be calculated using the following formula:

$$S = \frac{\sum_{x,i \in D} \left| pred_{x,i} - r_{x,i} \right|}{\left| D \right|}, \tag{15}$$

where D is the whole dataset and $\left|D\right|$ is the number of items in the set.

Results

The performance of the two proposed methods based on item weighting approach is compared with that of the baseline method. Paired t-tests were used to compare the mean differences of the final MAE results obtained by applying different methods over the whole dataset. A summary of the results is listed in Table 3.

Both methods proposed outperformed the baseline method. Method 1 resulted in a 6.178% improvement over the baseline. The result shows that the performance of traditional CF-based recommenders can be improved by bringing in only the orientation aspect of user feature opinions. The result also indicates that

by taking into consideration the consistency between the opinions expressed in user ratings and reviews the system can elevate the performance of the system dramatically.

Methods	MAE	Improvement over baseline	P-value
Baseline	0.6342		
M1	0.5950	6.178%*	0.0216
M2	0.5819	8.235%**	0.0038

*P<0.05
**P<0.01

Table 3 Summary of MAE Results

Method 2 achieved a very promising 8.24% improvement over the baseline. This shows the proposed User Feature Priority Weighting (FF-IRF) scheme is able to bring in more detailed information about user opinions from user reviews and improve the performance of the system even more than M1.

Discussion

Since the start of collaborative filtering research in the mid 1990s, user ratings have been the sole direct source for user item preferences (Adomavicius et al. 2008). This study is the first attempt to bring feature-oriented user text review analysis into memory-based CF and integrate user opinion information discussed in these reviews into the system in order to enhance its performance. The positive results show that user text reviews are indeed a promising complementary source of user opinion information for CF systems. The study also shows that user text reviews can be integrated into memory-based CF using different methods based on different amounts of user opinion information extracted from user reviews.

The user feature priority weighting (FF-IRF) proposed in the study is proven to be a valid measurement to model user's feature priority based on their text reviews. The modeling tool provides an alternative to the direct intensity measurement of user opinion, which is not yet obtainable with high accuracy based on the existing text mining techniques. On the other hand, the proposed FF-IRF weighting scheme can be used alongside intensity measurement, because it measures a different aspect of user preferences and is, therefore, complementary to the intensity measurement. Another advantage of the FF-IRF scheme is that it is a measurement based on objective information that can be directly obtained from user text reviews without relying on extracting subjective information from reviews. This means the result of FF-IRF measurement tends to be stable and reliable when it is applied to different domains or systems.

The Item Weighting Approach with Method 1 and Method 2 proposed in the study illustrates the strength of user text reviews when they are adopted in CF to verify the user opinions expressed in ratings. The results of these two methods directly validated the main motivation behind this study; that is, the ineffectiveness of user ratings as the sole source for user opinion-inferring in CF. By illustrating the robustness of user text reviews when used as an additional source for user opinion information in CF, the results of the study should attract more efforts to pursue this special research direction in collaborative filtering.

The results of this study also have implications for real-world recommender systems based on collaborative filtering techniques. The study serves as evidence to the practitioners of CF that by incorporating the vast amount of user text reviews that are available in the existing systems, they can improve the performance of the system dramatically. E-Businesses, like Amazon.com and Yelp.com have already accumulated a huge amount of user text reviews on various products and services. It is shown in this study that by leveraging these user reviews, e-Businesses can enhance the performance of their recommender systems significantly.

Due to limited resources and other constraints, this research is not without limitations. For example, in the data set used in this study, for each user rating in the system there is also a user text review available for the same user-item pair, which may not always be possible. There are usually many more users who provide ratings for items without a full-length text review. Therefore, the "sparsity" of user text review among real-world recommender systems will challenge the performance of any implementation of recommender systems based on online user text reviews.

In addition, methods employed in this study to integrate user feature opinions into CF could be optimized for fine-tuning system performance. As a future direction, more advanced methods can be developed to better integrate user opinion information into CF. More importantly, various parameters can be adapted in these methods to provide controls on different aspects of the modeling tool in order to further improve the system performance through empirically setting the parameters.

REFERENCES

Adomavicius, G., Huang, Z., and Tuzhilin, A. 2008. "Personalization and Recommender Systems," in *Tutorials in Operations Research 2008: State-of-the-Art Decision-Making Tools in the Information-Intensive Age,* Z.-L. Chen and S. Raghavan (eds.). INFORMS, pp. 55-107.

Bills, C.B., and Li, G. 2005. "Correlating Homicide and Suicide," *International Journal of Epidemiology* (34:4), August 2005, pp. 837-845.

Breese, J.S., Heckerman, D., and Kadie, C. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Fourteenth Conference on Uncertainty in Artificial Intelligence,* G.F. Cooper and S. Moral (eds.), University of Wisconsin Business School, Madison, Wisconsin, USA: Morgan Kaufmann, pp. 43-52.

Chen, P.-Y., Dhanasobhon, S., and Smith, M.D. 2007. "An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.Com," *International Conference on Information Systems (ICIS)*, Montréal, Québec, Canada: Association for Information Systems.

Furner, J. 2002. "On Recommending," *Journal of the American Society for Information Science and Technology* (53:9), pp. 747-763.

Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. 1999. "An Algorithmic Framework for Performing Collaborative Filtering," *Proceedings of the 22nd annual international conference on Research and development in information retrieval*, Berkeley, California, United States: ACM Press, pp. 230-237.

Jakob, N., Weber, S.H., Müller, M.C., and Gurevych, I. 2009. "Beyond the Stars: Exploiting Free-Text User Reviews to Improve the Accuracy of Movie Recommendations," *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, Hong Kong, China: ACM.

Leung, C.W.-k., Chan, S.C.-f., and Chung, F.-l. 2006. "Integrating Collaborative Filtering and Sentiment Analysis: A Rating Inference Approach," *The ECAI 2006 Workshop on Recommender Systems*, Riva del Garda, Italy, pp. 62–66.

Levi, A., Mokryn, O., Diot, C., and Taft, N. 2012. "Finding a Needle in a Haystack of Reviews: Cold Start Context-Based Hotel Recommender System," *Proceedings of the sixth ACM conference on Recommender systems*, Dublin, Ireland: ACM.

Liu, B., Hu, M., and Cheng, J. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proceedings of the 14th international conference on World Wide Web*, Chiba, Japan: ACM Press.

Moshfeghi, Y., Piwowarski, B., and Jose, J.M. 2011. "Handling Data Sparsity in Collaborative Filtering Using Emotion and Semantic Based Features," *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, Beijing, China: ACM.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. 1994. "Grouplens: An Open Architecture for Collaborative Filtering of Netnews," *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, Chapel Hill, North Carolina, United States: ACM Press, pp. 175-186.

Sahoo, N., Krishnan, R., Duncan, G., and Callan, J.P. 2006. "Collaborative Filtering with Multi-Component Rating for Recommender Systems," *The Sixteenth Annual Workshop on Information Technologies and Systems (WITS'06)*, Milwaukee, Wisconsin, USA.

Salton, G., and McGill, M.J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Shardanand, U., and Maes, P. 1995. "Social Information Filtering: Algorithms for Automating "Word of Mouth"," *Proceedings of the SIGCHI conference on Human factors in computing systems*, Denver, Colorado, United States: ACM Press/Addison-Wesley Publishing Co., pp. 210-217.

Wang, Y., Liu, Y., and Yu, X. 2012. "Collaborative Filtering with Aspect-Based Opinion Mining: A Tensor Factorization Approach," *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, Brussels, Belgium, pp. 1152-1157.

Zhou, L., and Chaovalit, P. 2008. "Ontology-Supported Polarity Mining," *Journal of the American Society for Information Science and Technology* (59:1), pp. 98-110.