Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 54(2), 252–262(2013)

Ante Odić, Marko Tkalčič, Jurij F. Tasič, Andrej Košir

Impact of the Context Relevancy on Ratings Prediction in a Movie-Recommender System

DOI 10.7305/automatika.54-2.258
UDK 004.5:004.89
IFAC 2.8; 1.2.1

Original scientific paper

Recommender systems are a popular and a highly researched way of helping users get to their desired content in the huge amount of available data, and services online. Understanding the situation in which users consume the items was shown to improve the recommendation process. For that reason, context-aware recommender system (CARS) employs contextual information in order to enhance the user's model and to improve the recommendations. An issue that is still open is how to decide which pieces of contextual information to acquire and how to incorporate them into CARS, since using irrelevant piece of contextual information could have a negative impact on the recommendations. We propose a methodology for detecting which pieces of contextual information contribute to explaining the variance in the ratings, based on statistical testing. We also inspect the impact of the detected relevant pieces of contextual information on the ratings prediction based on the matrix-factorization algorithm. The experiment was conducted on the *MovieAT* database. The results showed a significant difference in the ratings prediction using the relevant and the irrelevant pieces of contextual information. We also confirmed the positive impact of the relevant, and negative impact of the irrelevant pieces of contextual information with respect to the uncontextualized model.

Key words: Personalization, Recommender systems, Context-awareness

Utjecaj relevantnosti konteksta na predviđanje ocjena u sustavu za preporuke filmova. Sustavi za preporuke (eng. recommender systems) predstavljaju čest i vrlo istražen način pružanja pomoći korisnicima u svrhu pronalaska željenog sadržaja u velikoj količini dostupnih podataka i usluga. Pokazalo se da uvid u situaciju u kojoj korisnici koriste sadržaj doprinosi kvaliteti preporuka. Zbog toga, *konteksta svjesni sustavi za preporuke* (eng. *context-aware recommender systems CARS*) koriste kontekstne informacije kako bi poboljšali model korisnika i time kvalitetu preporuka. Jedan od neriješenih problema je kako odlučiti koje kontekstne informacije je potrebno sakupiti i kako ih upotrijebiti u CARSu, budući da upotreba nebitnih kontekstnih informacija može imati negativan utjecaj na kvalitetu preporuka. Mi predlažemo metodologiju za otkrivanje onih kontekstih informacija koje doprinose objašnjavanju varijabilnosti ocjena za sadržaje, utemeljenu na statističkom testiranju. Također, istražujemo utjecaj otkrivenog bitnog konteksta na predviđanje ocjena utemeljeno na algoritmu faktorizacije matrica. Eksperiment je proveden na bazi podataka *MovieAT*. Rezultati su pokazali znatnu razliku u predviđanju ocjena prilikom korištenja bitnog i nebitnog konteksta. Ujedno je potvrđen i pozitivan utjecaj bitnog, odnosno negativan utjecaj nebitnog konteksta, u odnosu na sustav koji ne koristi kontekst, što upućuje na važnost i kvalitetu detekcije.

Ključne riječi: personalizacija, sustavi za preporuke, kontekst

1 INTRODUCTION

The amount of media content online makes it difficult for users to find their desired content (e.g., movies, books, music, tourist destinations) in a reasonable time. One way of solving this problem is by using personalized services, defined as tailored services according to each user's characteristic and preferences [1], to adapt to each specific user, and to provide personalized content recommendations. Recommender systems (RS) are thus becoming a more and more common part of online media providers. Their main goal is to predict user's ratings for items that have not been seen by that user [2].

Understanding the situation in which users consume the items was shown to improve the recommendation process, since users' preferences can be situation dependent. For that reason, improving recommender system with contextual information, defined as information that can be used to describe the situation and the environment of the entities involved in such a system [3], has been a popular research topic over the past decade. Furthermore, contextual information was found useful in a number of different services [4]. However, there are still a number of issues concerning the definition, acquisition, detection and modeling of this dynamic information [5]. It is not easy to decide which contextual information to use. For some systems the weather could be important (e.g., when recommending a tourist destination), while for some it might be irrelevant [6]. Contextual information that does not have a significant contribution to explaining the variance of the ratings could degrade the prediction, since it could play the role of noise [7]. For that reason, we need to be able to detect whether a specific piece of information should be acquired and used or not.

In this article we propose a methodology for the detection of the relevant pieces of contextual information that contribute to explaining the variance in the ratings in a movie recommender system. We validate this methodology on different models for ratings prediction based on the contextualized matrix-factorization algorithm proposed in [7]. The experiments were conducted on an existing, well known, database acquired by the authors in [8].

1.1 State of the Art

The authors in [3, 6] proposed the definition of contextual information, which is being commonly cited, as: *"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves"*. However, this dynamic information should not be confused with the items' metadata (which describes the items involved in the service) and general user information (which describes the users) [9].

There is a variety of methods developed for context-aware RS (CARS). In [8], the authors presented the multidimensional recommendation model that adds context as an additional dimension to the classical $users \times items$ paradigm. The multidimensional approach was also used in [10]. The article also explains how to create a two-dimensional space by determining usage patterns under different contexts. A Bayesian-network based RS was proposed in [11]. The authors also tackled the problem of missing and erroneous context by harnessing the causal dependencies among the context variables. Three common ways of using the contextual information in a RS, i.e., context pre-filtering, post-filtering and context modeling, were explained by [12]. [13] used a pre-filtering method for the time context in order to improve music recommendations. The authors also tackled the problem of determining meaningful time partitions for pre-filtering, since the users can have different interpretations of time (morning, afternoon, etc.). They compared different methods for determining these partitions. In [14], the contextual information was used for pre-filtering to produce a submatrix of the rating matrix, with only the most relevant users and items with similar context. Context similarity was calculated by [15] to successfully tackle the problem of the increased data sparsity due to the context-reduction method or context filtering. The Random Walks method was used to utilize social context in a recommender system in [16]. In [17], the authors compared three approaches to a CARS: k-NN, linear-regression-based k-NN and inductive learning programming (ILP).

In [18] Koren proposed a new approach to context-aware recommendations. The author combined both the neighborhood and the latent factor model. Time as a piece of contextual information was then introduced in the form of temporal dependencies of the ratings in the matrix-factorization technique in [19]. This approach led to first place in the Netflix Prize Competition [20]. The authors in [21] proposed using the Pairwise Interaction Tensor Factorization (PITF), that was used also in [22], in tag recommendations, to predict which movies are rated in a specific time period. Matrix factorization was later extended with an additional parameter that models the interaction of the context and the items [7].

A lot of different pieces of contextual information are being exploited in different domains. For example, emotional context was exploited in [23]. Affective paramethers were also used in [24] as the affective metadata. The authors in [25] described ways that physical and social context can be used in a RS. Physical context was used to recommend mobile-device applications according to the location where they might be needed or that have been used by other users in a similar context. Social context was used to enhance the neighborhood creation in a collaborative RS. Social context was also used in [16]. Weather, mood and temperature, among others, were used in the tourist domain [26].

From all the possible contextual information that can be acquired, it is necessary to decide which is important for a specific service. In [8] the paired t-test was used to detect which contextual information is useful in their database; however, if certain assumptions about the data are not satisfied the t-test cannot be applied. The authors in [15] used the χ^2 test for the detection of the relevant context; however, this test could be inappropriate for small databases, i.e., for new systems and the cold-start problem, as we will explain later in the article. In [26] a context-relevance assessment was conducted to determine the influence of some contextual conditions on the users' ratings in the tourist domain, by asking users to imagine a given contextual condition and evaluate the influence of that condition. However, as they state, such an approach is problematic, since the users rate differently in real and supposed contexts [27].

1.2 Problem Statement

When creating a CARS by using multiple pieces of contextual information, the selection of the appropriate pieces of information to be used can be described as a feature selection problem. Relevant contextual information that contributes to explaining the variance in the ratings has to be identified from all the available pieces of contextual information, since the irrelevant information can degrade the performance of the CARS, and it is also unnecessary to spend the resources into acquiring the irrelevant information.

For that reason we propose a relevant-context detection method based on statistical hypothesis testing.

2 MATERIALS AND METHODS

In this section we describe the data and the methods that were used in this study. Figure 1 shows the experimental design. The experiment was conducted on the existing *MovieAT* dataset ([8]). We used and compared three statistical tests for the context-relevancy detection. Rating prediction was then contextualized by either using all pieces of contextual information, or using only the relevant or only the irrelevant pieces of contextual information. For the rating prediction we used and compared four different models, two of which were contextualized. The details about the data and methods are in the following subsections.

2.1 Database

We examined some of the available, existing RS databases that are commonly used in order to find adequate data for our experiment. First, we examined the Moviepilot and the Filmtipset databases that were used in the CAMRA challenge 2010 and 2011 (`http://www.dai-labor.de/camra2010/`). The only pieces of contextual information, in the parts of these databases that we could acquire, were the ones that could be derived from the timestamps. After examining the timestamps in these databases, we found that most users have all the timestamps from a relatively short period of time, usually most of them from the same day. This points to the fact that users provide most of the ratings at once. This means that these timestamps are from the moment when users rated the items and that we do not know when the users consumed the items. For this reason we decided not to use these databases since we are interested in finding how a specific item suited a specific user in a specific situation, i.e., what was the context during the consumption. We also examined the Yahoo music database used in the Yahoo KDD cup in 2011, however, once again the only pieces of contextual information were those that could be derived from the timestamps.

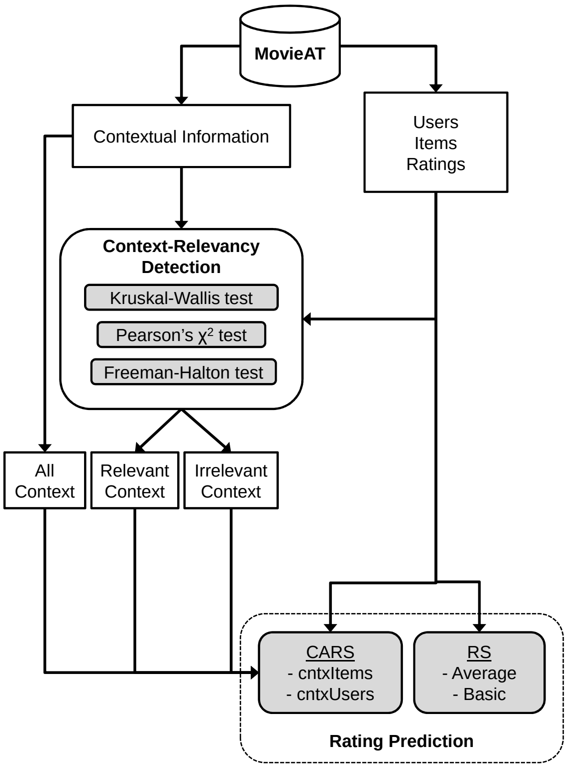

Fig. 1. Experimental design. The experiment was conducted on the data from the MovieAT database. Three statistical tests were used for the context-relevancy detection, and four models were used for the ratings prediction.

Finally we decided to use the *MovieAT* database created and used in [8]. This database was suitable since it contains several different pieces of contextual information. Other research on this database was also done in [7].

MovieAT database contained 1465 ratings from 84 users for 191 items. Pieces of contextual information in the database are *month*, *year*, *social*, *day of the week*, *opening weekend* and *will recommend*. Table 1 contains information about which variables are in the database, what they describe, and the number of possible values for each variable.

Another important database property is the amount of ratings provided by each user and the amount of ratings provided for each item. If most of the users have provided only a small amount of ratings or if there is only a small amount of ratings per each item (also known as the *long tail* in the ratings distributions for items and users [28]) it is hard to train a model. This is especially true for some types of contextualized methods, like pre-filtering [12], since the

Table 1. The description of the variables in the MovieAT database.

Variable	Meaning	Range
userID	user	84
itemID	item	191
movieRating	rating	13
monthSeen	context	12
yearSeen	context	3
withWhom	context	4
dayOfWeek	context	3
openWeekend	context	3
willRecommend	context	4

number of ratings provided in a specific context is even lower than the overall number of ratings.

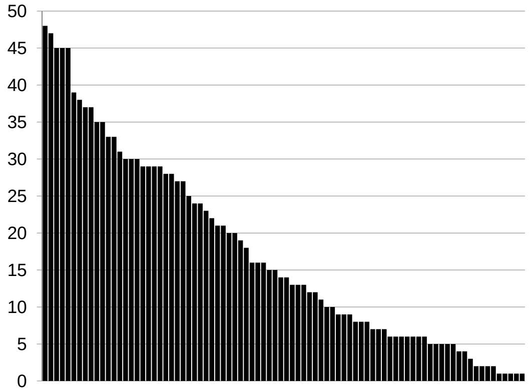

Fig. 2. Number of ratings per user. Heights of the bars represent the number of ratings provided by each user in the database.

Figures 2 and 3 show the amount of ratings provided by each user and the amount of ratings provided for each item, respectively.

2.2 Relevancy Detection

In this subsection we explain the basic reasoning and methods used for the context-relevancy detection.

2.2.1 Basic Notations and Reasoning

Assume that a user's decision (such as rating a user has assigned to a given content after the consumption) of a user $u \in \mathcal{U}$ on an item $h \in \mathcal{H}$ in the context $c \in \mathcal{C}$ is an additive function of contributions of the user model $\mathrm{UM}(u)$, content item metadata $\mathrm{MD}(h)$, context, and a random variable ε. Each of the listed contributions is estimated from several variables containing the corresponding information.

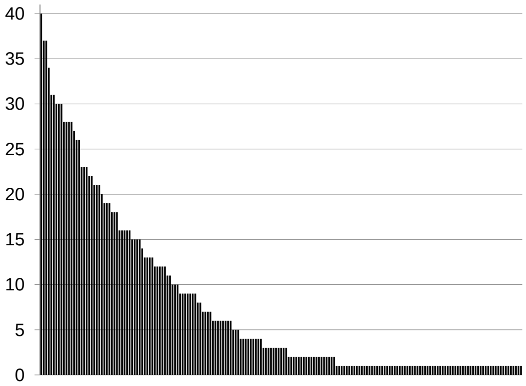

Fig. 3. Number of ratings per item. Heights of the bars represent the number of ratings provided for each item in the database.

The basic reasoning is that only a part of the user's decisions variability can be explained by the contributions from the user and item data. An unexplained variability is modeled by a random variable. The reason for this is the nature of the human decision-making process, which is very complex and dependent on many other factors besides the content item and user's past behavior recorded in his user model. The assumption is that an additional part of the user-decision variability can be explained by the context c, described by the contextual variable v_i^c. In this work we concentrate on explaining the variability of the ratings, however the reasoning is the same as for the decision making in general. How the contextual variable contributes to the rating prediction, and how it is employed in the model, we will describe in Section 2.3.

Clearly, a variable that has no variability for a given user, cannot provide any additional information about his decisions and cannot be contextual. On the other hand, the high variability can be inconsistent with the rating. Therefore, if the variability of the variable is low, it is not a relevant contextual information, if the variability is high it might or might not be a relevant contextual information. Furthermore, a variable that does not explain a significant part of the user-decision variability cannot provide any relevant contextual information.

Therefore, the two ways to identify context variables are:

(i) **Heuristic Scoring:** Measuring the variable's v_i^c variability;

(ii) **Significance-Test Scoring:** Measuring the part of the variability of the user decision that can be explained by the variable v_i^c.

The heuristic scoring approach gives a necessary condition for a variable to be contextual and the significance-test scoring approach gives a sufficient condition. Note that the heuristic scoring does not use a user-decision (i.e. rating in our case) variable as an input but only measures the variability of the variables, regardless of the users' decisions.

The application of both the above-listed approaches depends on the analyzed variables type. When they are of the interval or proportional type it is based on the Pearson correlation coefficient (factor analysis and regression analysis). But in the real world the candidate variables are typically categorical or ordinal, which requires a modified approach. The details are given in the following subsections.

2.2.2 Context-Variable Identification Using a Variable Variance (approach (i): Heuristic Scoring)

A variability of a numerical variable v is typically measured by its variance computed by a covariance formula $\eta(v) = Cov(v, v)$. Unfortunately, contextual variables are typically not numeric but categorical or at most ordinal. Besides, the straightforward adaptations of this formula using association coefficients among categorical variables such as contingency coefficients is also not applicable since it does not have all the required properties. Therefore, we have to use other variability measures. Some of them are presented in the subsequent paragraphs.

The authors in [29] proposed the unalikeability as a measure of how often observations differ from one another. The variability as the unlikability is calculated as:

$$\eta_u(v^c) = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n}$$

where x_i and x_j are the observations, *n* is the number of all the observations, and $c(x_i, x_j) = 0$ if $x_i = x_j$, and $c(x_i, x_j) = 1$ if $x_i \neq x_j$.

Another approach is to calculate the variability as the entropy of the observations, by the formula:

$$\eta_H(v^c) = -\frac{1}{\log_2 M} \sum_{i=1}^{M} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

where *M* is the number of the possible values of the contextual variable v^c, n_i is the number of observations with the value *i* and *n* in the number of all observations. $\frac{1}{log_2 M}$ is a normalizing constant which ensures that the variability is between zero and one.

The variability as a sample variance is calculated as:

$$\sigma^2(v^c) = \frac{1}{n} \sum_{i=1}^{M} n_i \left| v_i - \bar{v} \right|^2$$

where v_i is the value of the categorical class *i* and $\bar{v}$ is the mean of all the observations.

In order to establish whether the contextual variable should be accepted according to its variability, a threshold is needed. As a hypothesis, we assume that the variable for which a certain value occurs at least 95% percent of times is not relevant (for example, if users mostly consume items at home, the *location* is not informative contextual information). We use a significance-testing-based threshold determination. On the distribution of the variabilities for such a variable, for the risk level $\alpha = 0.05$, a critical value is calculated and used as a threshold. This threshold is not sample size dependent (i.e. does not depend on the number of ratings in the dataset), however it does depend on the number of possible values of the variable (M). For this reason, thresholds were calculated for different contextual variables according to the number of possible values of each variable. In our case, the calculated thresholds were 0.23 when $M = 3$, 0.19 when $M = 4$, and 0.14 when $M = 12$.

2.2.3 Context Variable Relevancy Detection Using the Explained Variance of a User Decision (approach (ii): Significance-Test Scoring)

When selecting a statistical test to detect the relevancy of the piece of contextual information, variable types and the assumptions about the data should be taken into consideration. Since the potentially relevant contextual variables under investigation are typically categorical or at most ordinal, the association among the contextual variables and ratings is measured using the association coefficients and pertained significance tests for the categorical variables. As further discussed in the following text, an a-priori and a post-hoc power analyses are of key importance here.

In [8] the paired t-test was used for the detection. In the case of the binary piece of contextual information (e.g., *daytype*: working day, weekend) ratings can be divided into two groups, one for each value of the contextual information (e.g., ratings during the working day, ratings during the weekend). If a contextual variable has more than two possible values, i.e., ratings divided in more than two groups, the analysis of variance (ANOVA) test can be used. In both cases we test whether or not the means of the groups are equal. However, for the t-test and the ANOVA, normality (i.e., the distributions of groups are normal) and homogeneity of variance (i.e., the variance of data in groups are the same) assumptions have to be met. To test the normality of the groups' distributions we will use the Shapiro-Wilk test [30]. If the distributions of the ratings in different groups are not normal, the ANOVA cannot be used.

The alternative for the ANOVA test, when the normality assumption is not met, is the Kruskal-Wallis test, often called an "ANOVA by Ranks" [31]. The Kruskal-Wallis test does not assume normality, however the homogeneity of variance assumption still has to be met. To test the homogeneity of variance the Levene's test can be used [32]. If the homogeneity assumption is not met, the Kruskal-Wallis test cannot be used.

When the assumptions for using the ANOVA and the Kruskal-Wallies test are not met, weaker tests should be employed. Since we mostly have categorical variables the Pearson's χ^2 test should be used, as in [15]. However, for the Pearson's χ^2 test the Cochran's rule has to be satisfied. It states that at least 80% of the expected cell count in a contingency table should be five or more, and that no expected cell count should be less than one. Unfortunately, due to the typical sparsity of the data in small databases (e.g., during the cold-start phase) and especially when one of the variables tested has a larger number of possible values, the Cochran's rule is not satisfied and thus the detection cannot be achieved. If the Cochran's rule is notsatisfied, we propose using the Freeman-Halton test, which is Fisher's exact test extended to $n \times m$ contingency tables, since the Fisher test does not depend on the sample size [33].

With the Freeman-Halton test the independence is tested between each contextual variable and the ratings. The null hypothesis of the test states that the two variables are independent. The alternative hypothesis states that they are dependent. If we successfully reject the null hypothesis we conclude that the contextual variable and the rating are dependent and thus the piece of contextual information represented by that contextual variable is relevant.

From the four possible outcomes of the hypothesis testing, two outcomes, i.e., the type-I and type-II errors, would mean that the relevancy of the contextual information was not detected correctly. A type-I error would mean that the contextual variable was irrelevant, but was detected as relevant. In other words, it would be used as relevant information when in fact it does not explain a part of the unexplained variance in the rating. A type-II error would mean that the contextual information was in fact relevant, but was detected as irrelevant. In other words, the information that does explain a part of the unexplained variance in the rating was overseen. For that reason we are especially interested in observing the probability of a type-II error in our hypothesis testing. The probability of a type-II error occurring is called the false-negative rate β. Therefore, the probability of obtaining a result from the statistical test that will allow the rejection of the null hypothesis, if the null is false, is called the statistical power and is equal to $1-\beta$.

Once the p value (p) and the statistical power ($1-\beta$) are calculated for a contextual variable, we can determine the probability that the contextual variable is relevant. If the $p < \alpha$, the probability is $P[A] = 1-\beta$. If the $p \geq \alpha$, the probability is $P[A] = \beta$. A stands for a **relevant contextual variable**, which means that it has a significant contribution in explaining the variance in the ratings, and α is the significance level of the test. In the case when $p \geq \alpha$, if the statistical power is great, we conclude that the contextual variable is not relevant; however, if the statistical power is small, we do not conclude that the context is not relevant. This principle is independent of the test that we use.

We also conducted an a-priori power analysis to compute the required sample size, given the significance level, desired power and effect size.

The a-priori power analysis and post-hoc power analysis were conducted according to [34, 35]. The post-hoc power analysis for the Freeman-Halton test was made using the Monte-Carlo method, and was implemented using the *R* software environment (`http://www.r-project.org/`).

2.3 Rating Prediction

Several different methods were compared in order to inspect the impact of contextual information on the ratings prediction.

First we used a simple *Average Movie Rating* method as a baseline predictor. Predicted rating is calculated as an average rating for the item, ignoring the user and context:

$$\hat{r}(u,i) = \mu_i$$

where $\hat{r}(i)$ is the predicted rating for the item i and μ_i is the average rating for that item.

Second method is the *Basic Matrix-Factorization* algorithm described in [18]. We used the following equation and notations for the matrix factorization, ignoring the context:

$$\hat{r}(u,i) = \mu + b_i + b_u + \mathbf{q}_i^T \cdot \mathbf{p}_u$$

where $\hat{r}(u,i)$ is the predicted rating from a user u for the item i, μ is a global ratings' bias, b_u is a user's bias, b_i is an item's bias, $\mathbf{q}_i$ is an item's latent feature vector, and $\mathbf{p}_u$ is a user's latent feature vector. $\hat{r}$, μ, b_u and b_i are scalars, and $\mathbf{q}_i$ and $\mathbf{p}_u$ are vectors. The system learning procedure is defined as an optimization problem:

$$\min_{p_*,q_*,b_*} \sum_{r\in K} \left[\left(r(u,i) - \mu - b_i - b_u - \mathbf{q}_i^T \cdot \mathbf{p}_u\right)^2 + \lambda \left(b_i^2 + b_u^2 + \|p_u\|^2 + \|q_i\|^2\right) \right]$$

where λ is the constant that controls the regularization, and $K = \{(u,i) \mid r(u,i) \text{ is known}\}$.

To incorporate contextual information in the matrix factorization, we used an algorithm described in [7, 26], and a slightly modified one. As an extension to matrix factorization the authors added parameters that model the interaction between context and the items:

$$\hat{r}\left(u,i\right) = \mu + b_i + b_u + \mathbf{q}_i^T \cdot \mathbf{p}_u + \sum_{j=1}^{k} b_{ic_j}$$

where b_{ic_j} is a modification of the item's i bias in context c_j, and k is the amount of the different pieces of contextual information incorporated in a model. We will refer to this model in the following text as *CARS-cntxItems*.

However, we decided to add parameters that model the interaction between the context and the users, to inspect the influence of the contextual information on the users' behavior. In this approach context dependent ratings are modeled as:

$$\hat{r}\left(u,i\right) = \mu + b_i + b_u + \mathbf{q}_i^T \cdot \mathbf{p}_u + \sum_{j=1}^{k} b_{uc_j}$$

where b_{uc_j} is a modification of the user's u bias in context c_j, and k is the amount of the different pieces of contextual information incorporated in a model. We will refer to this model in the following text as *CARS-cntxUsers*.

In all the models, the users' and the items' feature vectors were calculated using the gradual descent method [18]. As the evaluation measure for the predicting ratings we used the *root mean square error* (RMSE). The results were obtained by the 10-fold cross validation.

To test the impact of the relevant and the irrelevant pieces of contextual information on the contextualized rating prediction, we will test these models in the three ways: by incorporating all the available pieces of contextual information, by incorporating only the pieces of contextual information that were detected as relevant, and by incorporating only the pieces of contextual information that were detected as irrelevant.

3 RESULTS

In this section we present the experimental results for the detection of the relevant pieces of contextual information, and the ratings prediction.

3.1 Relevancy Detection Results

To detect which piece of contextual information is relevant in the *MovieAT* database, we tested each contextual variable with the methods described in Section 2.2. The significance level of the test was $\alpha = 0.05$.

For the a-priori power analysis for the Freeman-Halton test we used the effect size $\omega = 0.2$ and power $(1 - \beta) = 0.95$. For the variables with the lowest degrees of freedom $df = 24$ the required sample size calculated was 365 samples. For the variables with the highest degrees of freedom $df = 132$ the required sample size calculated was 710 samples.

Variable variability (approach (i)) did not clearly point to the irrelevancy of any contextual variable. Variabilities of all the contextual variables in the *MovieAT* database were high and thus all the variables were detected as potentially relevant by the heuristic scoring approach.

The Shapiro-Wilk test for the normality showed that the distributions of all the groups for every contextual variable in the database are not normal, thus the ANOVA test could not be used for the relevancy detection of any piece of contextual information.

Table 2 contains the p-values from the Lavene's test for homogeneity with the associated ruling whether the homogeneity assumption was met or not. The significance level of the Levene's test was $\alpha = 0.05$.

Table 2. Results of the Levene's test for homogeneity; p-value of the test and the ruling on the assumption.

Levene's test		
context	**p-value**	**assumption ruling**
monthSeen	0.22	met
yearSeen	0.09	met
withWhom	0.59	met
dayOfWeek	0.29	met
openWeekend	< 0.001	not met
willRecommend	< 0.001	not met

Tables 3, 4 and 5 contain the results of the contextual information detection made by the Kruskal-Wallis test, the Pearson's χ^2 test, and the Freeman-Halton test, for each contextual variable, respectively. Pieces of the contextual information that were detected as relevant are: *withWhom*, *dayOfWeek*, *openWeekend* and *willRecommend*. The irrelevant pieces of contextual information are *monthSeen* and *yearSeen*.

3.2 Rating Prediction Results

This section contains the results from all the models described in Section 2.3.

On Fig. 4 we provide the boxplot of the results from the *Average* and the *Basic* model, as well as the results from the *CARS-cntxItems* model, using all the pieces of contextual information, using only the pieces of contextual information detected as relevant and using only the pieces of contextual information detected as irrelevant.

Table 3. Context-detection results for the Kruskal-Wallis test. Column homogeneity contains information whether the homogeneity assumption was met.

Kruskal-Wallis test			
context	**homogeneity**	**p-value**	**power**
monthSeen	yes	0.93	> 0.99
yearSeen	yes	0.64	> 0.99
withWhom	yes	< 0.01	> 0.99
dayOfWeek	yes	< 0.01	> 0.99
openWeekend	no	-	-
willRecommend	no	-	-

Table 4. Context-detection results for the Pearson's χ^2 test. Column Cochran contains information whether the Cochran's rule was satisfied.

Pearson's χ^2 test			
context	**Cochran**	**p-value**	**power**
monthSeen	no	-	-
yearSeen	no	-	-
withWhom	yes	0.01	> 0.99
dayOfWeek	yes	0.02	0.98
openWeekend	yes	< 0.01	> 0.99
willRecommend	yes	< 0.01	> 0.99

Table 5. Context-detection results for the Freeman-Halton test.

Freeman-Halton test		
context	**p-value**	**power**
monthSeen	0.15	> 0.99
yearSeen	0.10	0.94
withWhom	0.01	> 0.99
dayOfWeek	0.02	0.96
openWeekend	< 0.01	> 0.99
willRecommend	< 0.01	> 0.99

On Fig. 5 we provide the boxplot of the results from the *Average* and the *Basic* model, as well as the results from the *CARS-cntxUsers* model, using all the pieces of contextual information, using only the pieces of contextual information detected as relevant and using only the pieces of contextual information detected as irrelevant.

3.3 Discussion

With the results obtained from the detection and the ratings prediction we can make the following conclusions.

An a-priori power analysis confirmed that the sample size in the database is large enough to conduct the statistical testing for relevant-context detection.

The Kruskal-Wallis, the Freeman-Halton and the Pearson's χ^2 tests gave consistent results for all the contextual variables. However, the conclusion about the relevancy

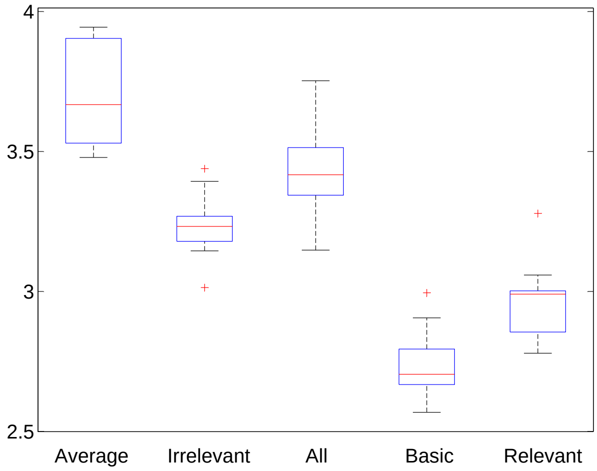

Fig. 4. Boxplot of the results for the CARS-cntxItems model. Vertical axis shows the RMSE value. These results are obtained by the 10-fold cross validation for the Average and the Basic models, and CARS-cntxItems model using All, Relevant and Irrelevant pieces of contextual information.

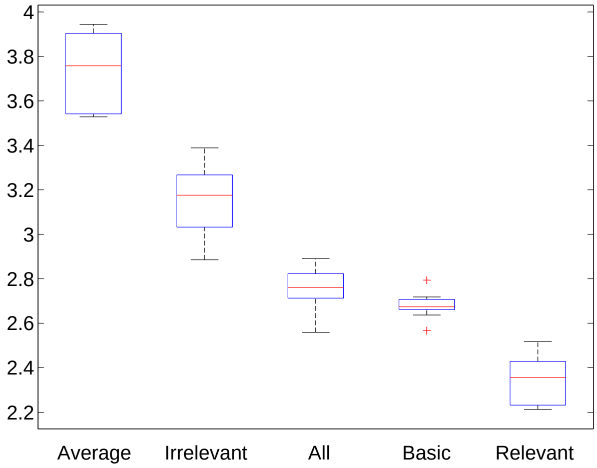

Fig. 5. Boxplot of the results for the CARS-cntxUsers model. Vertical axis shows the RMSE value. These results are obtained by the 10-fold cross validation for the Average and Basic models, and CARS-cntxUsers model using All, Relevant and Irrelevant contextual information.

could not be achieved for two variables (*openWeekend* and *willRecommend*) with the Kruskal-Wallies test due to the homogeneity assumption, and two variables (*monthSeen*, *yearSeen*) with the Pearson's χ^2 test due to the Cochran's rule. This points to the importance of the Freeman-Halton test if the data assumptions are not met, and in the case of the small databases. Since the detection of the relevant contextual information is especially important in the early stage of the system, while there is still a low number of ratings, Freeman-Halton test proves to be a better choice than the Pearson's χ^2 test.

By comparing Fig. 4 and Fig. 5 it is clear that the *CARS-cntxUsers* model performed better than *CARS-cntxItems* model in each case (relevant, irrelevant, all context). Furthermore, *Basic Matrix-Factorization* model outperformed *CARS-cntxItems* model, even in the case of relevant context. It is shown that by using only the relevant contextual variables, we still outperform the models using all or the irrelevant pieces. However, in the *CARS-cntxItems* model, contextualization in general deteriorates the result of the basic model. This can be explained by the long tail in the amount of ratings per item (Fig. 3). Since there are a lot of items with a low number of associated ratings, and consequently even lower number of ratings per context values, contextualized model with items' biases could not be trained well, which deteriorated the results. Furthermore, since there are only a few ratings for some item-context pairs, the more contextual variables we use, the worse is the result, since many b_{ic_j} parameters are not sufficiently trained. Hence in the *CARS-cntxItems* model, employing all contextual variables leads to worse results then when employing only the irrelevant variables. This points to the high impact of the sparsity on the training.

In the case of *CARS-cntxUsers* model however, on Fig. 5, we can see the positive impact of the relevant contextual information on the results. When using only the pieces of contextual information detected as relevant the results are significantly better than with the uncontextualized *Basic Matrix-Factorization* model. In addition, *Basic Matrix-Factorization* performed significantly better than when using the irrelevant context. Another important result is that when all the available pieces of contextual information were used (i.e., no detection), the achieved results were worse than with the *Basic Matrix-Factorization* model, which means that incorporating irrelevant contextual information can in fact deteriorate the results. This directly points to the importance and the quality of the detection method.

4 CONCLUSION AND FUTURE WORK

In this work we addressed the problems of detecting and incorporating pieces of contextual information in a collaborative filtering recommender system. Experiments were conducted on the well known *MovieAT* database. Collaborative filtering rating predictions were computed with the matrix-factorization algorithm which was then contextualized by the multiple pieces of contextual information at once. We proposed the Freeman-Halton statistical test with power analysis for the detection of the relevant contextual information when other tests fail to reach the conclusion. We also proposed incorporating context in ratings prediction by adding parameters describing context-user interaction, based on the method in [7, 26].

Our detection method proved useful when the conclusion about the context relevancy could not be achieved by the Kruskal-Wallis test due to the data assumptions, and the Pearson's χ^2 test due to the Cochran's rule. In addition, the rating-prediction results confirmed the importance of the detection, since the incorporation of all the available pieces of contextual information as well as the irrelevant ones, in the matrix factorization, deteriorated the results. Furthermore, matrix factorization that utilizes relevant contextual information performs significantly better than the uncontextualized model.

In our experiment we also noticed the impact of the *long tail* on the different models results. This suggests that this information about the data should not be overlooked while designing a model.

Our future work consists of upgrading both the detection procedure and the context-aware matrix factorization algorithm. We would also like to test our methods on other databases and real applications. Our goal is also to evaluate these methods with other real-world measures like users' satisfaction, novelty, diversity, etc. Further more, we are interested in inspecting the possible solutions for the *long-tail* problem.

REFERENCES

[1] Y. Joung, M. E. Zarki, and R. Jain, "A user model for personalization services," *2009 Fourth International Conference on Digital Information Management*, pp. 1–6, Nov. 2009.

[2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 734–749, June 2005.

[3] A. Dey and G. Abowd, "Towards a better understanding of context and context-awareness," *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, pp. 304–307, 1999.

[4] F. Toutain, A. Bouabdallah, R. Zemek, and C. Daloz, "Interpersonal Context-Aware Communication Services," *IEEE Communications Magazine*, no. January, pp. 68–74, 2011.

[5] Z. Yujie and W. Licai, “Some Challenges for Context-aware Recommender Systems,” in *Computer Science and Education (ICCSE), 2010 5th International Conference on*, pp. 362–365, 2010.

[6] A. K. Dey, “Understanding and Using Context,” *Personal and Ubiquitous Computing*, vol. 5, pp. 4–7, Feb. 2001.

[7] L. Baltrunas, B. Ludwig, and F. Ricci, “Matrix Factorization Techniques for Context Aware Recommendation,” in *Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 301–304, 2010.

[8] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, “Incorporating contextual information in recommender systems using a multidimensional approach,” *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 1, pp. 103–145, 2005.

[9] W. Woerndl and J. Schlichter, “Introducing context into recommender systems,” *2007 Workshop on Recommender Systems in e-commerce*, pp. 138–140, 2007.

[10] M. Hosseini-pozveh, “A multidimensional approach for context-aware recommendation in mobile commerce,” *Journal of Computer Science*, vol. 3, no. 1, 2009.

[11] G.-E. Yap, A.-H. Tan, and H.-H. Pang, “Discovering and Exploiting Causal Dependencies for Robust Mobile Context-Aware Recommenders,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 977–992, July 2007.

[12] G. Adomavicius and A. Tuzhilin, “Context-aware recommender systems,” *Recommender Systems Handbook*, pp. 217–253, 2011.

[13] L. Baltrunas, X. Amatriain, and V. Augusta, “Towards Time-Dependant Recommendation based on Implicit Feedback,” in *Proceedings of the RecSys 2009 Workshop on Context-aware Recommender Systems*, pp. 1–5, 2009.

[14] J. Su, H. Yeh, P. Yu, and V. Tseng, “Music recommendation using content and context information mining,” *Intelligent Systems, IEEE*, vol. 25, pp. 16–26, 2010.

[15] L. Liu, F. Lecue, N. Mehandjiev, and L. Xu, “Using Context Similarity for Service Recommendation,” *2010 IEEE Fourth International Conference on Semantic Computing*, pp. 277–284, Sept. 2010.

[16] F. Díez, J. E. Chavarriaga, P. G. Campos, and A. Bellogín, “Movie Recommendations based in explicit and implicit features extracted from the Filmtipset dataset,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation*, pp. 45–52, 2010.

[17] H. Rahmani, B. Piccart, and H. Blockeel, “Three complementary approaches to context aware movie recommendation,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation*, pp. 57–60, 2010.

[18] Y. Koren, “Factorization Meets the Neighborhood : a Multifaceted Collaborative Filtering Model,” in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434, 2008.

[19] Y. Koren, “Collaborative filtering with temporal dynamics,” *Communications of the ACM*, vol. 53, pp. 89–97, Apr. 2010.

[20] Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems,” *Computer*, vol. 42, pp. 30–37, Aug. 2009.

[21] Z. Gantner, S. Rendle, and L. Schmidt-Thieme, “Factorization Models for Context- / Time-Aware Movie Recommendations Encoding Time as Context,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation*, pp. 14–19, 2010.

[22] S. Rendle, *Context-Aware Ranking with Factorization Models*. Springer-Verlag New York Inc, 2010.

[23] G. Gonzalez, J. L. de la Rosa, M. Montaner, and S. Delfin, “Embedding Emotional Context in Recommender Systems,” *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pp. 845–852, Apr. 2007.

[24] M. Tkalčič, U. Burnik, and A. Košir, “Using affective parameters in a content-based recommender system for images,” *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 279–311, 2010.

[25] W. Woerndl and G. Groh, “Utilizing Physical and Social Context to Improve Recommender Systems,” *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pp. 123–128, Nov. 2007.

[26] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci, “Context relevance assessment and exploitation in mobile recommender systems,” *Personal and Ubiquitous Computing*, pp. 1–20, June 2011.

[27] C. Ono, Y. Takishima, Y. Motomura, and H. Asoh, “Context-Aware Preference Model Based on a Study of,” in *User Modeling, Adaptation, and Personalization*, pp. 102–113, 2009.

[28] Y. Park and A. Tuzhilin, “The long tail of recommender systems and how to leverage it,” in *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 11–18, ACM, 2008.

[29] G. D. Kader and M. Perry, “Variability for Categorical Variables,” *Journal of Statistics Education*, vol. 15, no. 2, 2007.

[30] S. Shapiro and M. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[31] S. Maxwell and H. Delaney, *Designing experiments and analyzing data: A model comparison perspective*, vol. 1. Lawrence Erlbaum, 2004.

[32] M. Brown and A. Forsythe, “Robust tests for the equality of variances,” *Journal of the American Statistical Association*, pp. 364–367, 1974.

[33] A. Agresti, “A Survey -of Exact Inference for Contingency Tables,” *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.

[34] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 1988.

[35] F. Faul, E. Erdfelder, A. Buchner, and A. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, 2009.

Ante Odić is a junior researcher at the University of Ljubljana Faculty of Electrical Engineering. He is pursuing his PhD degree in the fields of personalization, user modeling, recommender systems and contextual information.

Marko Tkalčič received his PhD degree at the University of Ljubljana Faculty of Electrical Engineering (UL FE). In 1999 he received the student Prešeren award for his BSc thesis. Since 1999 he has been employed as a researcher at the Digital Signal Processing Laboratory (LDOS) at UL FE where he has been working in various research areas including human visual perception, colour management, peer-to-peer networking, web interfaces and mobile services. In the years 2006-2007 he served as the technical manager of the FP6 eTEN P2PME project. In 2010 he founded the Affective Computing Students Interest Group (ACSIG) at UL FE. He is currently investigating other aspects of emotions and personality in human-computer interaction systems, especially as contextual information.

Jurij F. Tasič is a professor of system theory and computing at University of Ljubljana, and has been associated with signal processing, advanced and adaptive algorithms and Parallel algorithms for many years. He developed the first former Yugoslav process computer system in 1976 and first singleboard computer system in 1979. On the basis of mentioned computers in 1976 he developed a computer based Spectral Analyser and from 1978 to 1980 he developed a set of chemical instruments as Flame Photometer, Refractometer, and Densitymeter.In 1980 he with his colleagues developed a microcomputer based system for production control of the Slovenian Electro-Energetic System, as well as first single-board CPM compatible computer in former Yugoslavia.

Andrej Košir PhD, is associate professor at the Faculty of Electrical Engineering, University of Ljubljana. He was awarded the Vidmar prize for his educational prowess. He is active in several research fields, including signal, image and video processing, optimization (numerical optimization, genetic algorithm), and user interfaces. He was a guest researcher at the University of Westminster, London, UK, at the University of Waterloo, Canada, and at the North Carolina State University, USA. He is currently leading projects from the field of multimedia, optimization methods and digital signal processing – especially in object recognition on digital images, intelligent networks, user interfaces and user modeling.

AUTHORS' ADDRESSES

Ante Odić
Marko Tkalčič, Ph.D.
Prof. Jurij F. Tasič, Ph.D.
Prof. Andrej Košir, Ph.D.
Digital Signal, Image and Video Processing Laboratory,
Faculty of Electrical Engineering,
University of Ljubljana,
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
email: {ante.odic, marko.tkalcic,andrej.kosir}
@ldos.fe.uni-lj.si, jurij.tasic@fe.uni-lj.si

Received: 2012-04-25
Accepted: 2013-01-24