Relevant Context in a Movie Recommender System: Users' Opinion vs. Statistical Detection

Ante Odić*
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
ante.odic@ldos.fe.uni-lj.si

Marko Tkalčič
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
marko.tkalcic@ldos.fe.uni-lj.si

Jurij F. Tasič
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
jurij.tasic@ldos.fe.uni-lj.si

Andrej Košir
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25
Ljubljana, Slovenia
andrej.kosir@ldos.fe.uni-lj.si

ABSTRACT

Context-aware recommender systems help users find their desired content in a reasonable time, by exploiting the pieces of information that describe the situation in which users will consume the items. One of the remaining issues in such systems is determining which contextual information is relevant and which is not. This is an issue since the irrelevant contextual information can degrade the recommendation quality and it is simply unnecessary to spend resources on the acquisition of the irrelevant data. In this article we compare two approaches: the relevancy assessment from the user survey and the relevancy detection with statistical testing on the rating data. With these approaches we want to see if it is possible for users to predict which context influences their decisions and which approach leads to better detection of the relevant contextual information.

Keywords

context-aware, recommender systems, user modeling

1. INTRODUCTION

Incorporating contextual information in recommender system (RS) has been a popular research topic over the past decade. Contextual information is defined as information that can be used to describe the situation and the environment of the entities involved in such a system [6], and was proved to improve the recommendation procedure in context-aware recommender systems (CARS) [2], as well as other personalized services [14]. However, the question remains, which contextual information to use, or in other words which situation parameters influence users' decisions in a specific service? As the authors in [4] state, contextual information that does not have a significant contribution to explaining the variance in users' decisions could degrade the prediction, since it could play the role of noise. In addition, it is unnecessary to spend resources on the acquisition of irrelevant data.

1.1 Determining Relevant and Irrelevant Contextual Information

It is not always easy to predict which pieces of contextual information are important for a specific service. There are many pieces of contextual information that can influence users' decisions in a more (location, social, working day/weekend) or less (weather, temperature) intuitive way. The authors in [1] used the paired t-test to detect which pieces of contextual information are useful in their database. The χ^2 test was used for the detection in [11]. In [4] the authors conducted a context-relevance assessment to determine the influence of some pieces of contextual information on users' ratings in the tourist domain, by asking users to imagine a given situation and evaluate the influence of that contextual information. However, as they state, such an approach is problematic, since users rate differently in real and supposed contexts [12].

We can therefore identify two different approaches to the determination of context relevancy: *the assessment from user-survey* and *the detection from the rating data*. In the rest of the text we will simply refer to these two approaches as the assessment and the detection.

The assessment and the detection approaches are very different in terms of requirements and approach. The assessment does not require any real rating data, while for the detection we need a substantial number of ratings with associated context. The assessment could therefore be a valuable tool for determining relevant pieces of contextual information before data acquisition, or in other words, during the phase of designing a CARS. Furthermore, while the detection is made on real situation data, assessment is obtained

*Corresponding author.

CARS-2012, September 9, 2012, Dublin, Ireland.

.

from hypothetical situations described in survey questions. Apart from the fact that users do not necessarily know what really influences their decisions, quality of the assessment could also be influenced by users' ability to conceptualize a hypothetical situation. Finally, the assessment is intrusive and requires users to spend their time on an additional task. As we found out, only a minority of users in our system, that we will describe later in the article, were willing to participate in the survey as they did not see the immediate benefit from it (in the contrast with rating items to improve their profile). The detection, on the other hand, is done without the need for any additional effort on users' part.

In Table 1 we list pros and cons of each method.

Table 1: Pros and cons of assessment and detection of relevant context.

	Assessment	**Detection**
Pros	No rating data needed.	Real situation. Unintrusive.
Cons	Hypothetical situation. Intrusive.	Rating data needed.

1.2 Problem Statement

The assessment and the detection differ in the aspects of when each could be used, what information is needed and whether they rely on real or hypothetical situation. However, to the best of our knowledge, several other questions remain that we try to answer in this study.

- How well do the outcomes from these approaches match?
- Which approaches is better in determining relevant context?
- Are users aware of what influences their decisions?

1.3 Experimental Design

In this study we compare two approaches for relevant context determination: (i) the assessment from users' survey and (ii) the detection from rating data. The determination of contextual information relevancy for the contextualized recommendations is a binary decision. The piece of contextual information is either relevant (i.e., it contributes to explaining the variance of user's decision/rating) or irrelevant (i.e., it does not contribute to explaining the variance of user's decision/rating).

We will use the detection and the assessment as two different methods to classify each piece of contextual information in one of the two classes: relevant and irrelevant. Since the ground truth is unknown (i.e., we do not know which piece of contextual information is actually helpful), we will use the contextualized matrix-factorization algorithm, with each piece of contextual information separately, to determine the influence of each piece on the rating prediction. The idea is that the relevant pieces of contextual information will lead to better results than the irrelevant ones. The success of the contextualized rating prediction will be evaluated by the *root mean square error* (RMSE) measure. Once the RMSE for each piece of contextual information is achieved the detection and the assessment can each be evaluated based on the number of times a piece of contextual information determined as relevant performed better than the irrelevant one. Finally, we inspect how well do the results from these approaches match, how well do they perform and which approach is better.

2. MATERIALS AND METHODS

In this section we provide the description of the data used in the study and the methods used to answer the questions established in the problem statement.

2.1 Dataset

In order to be able to compare the methods for determining context relevancy, we needed a context-rich RS database. Unfortunately, commonly used databases such as Moviepilot and Yahoo! Music, contain only that context which can be derived from the timestamps. Other information available in these databases is general user information that describes users (age, sex, etc.), does not vary for a fixed user, and thus cannot be used as a contextual information.

Since we were interested in inspecting more different contextual variables we decided to create a database containing several potential pieces of contextual information. Since users have a tendency to rate items differently in real and supposed contexts [12], we decided to obtain a database in such a way that each rating and associated context is provided by a user after a real user-item interaction. The problem of this approach is that it takes a long time to create a database, since we cannot simply ask users to rate, for example, 30 items, but to enter the rating after each time they consume an item.

We created an online application for rating movies which users are employing to track the movies they watched and obtain the recommendations (`www.ldos.si/recommender.html`). Users are instructed to log into the system after watching a movie, enter a rating for a movie and fill in a simple questionnaire created to explicitly acquire the contextual information describing the situation during the consumption. Users are instructed to provide the rating and contextual information immediately after the consumption, so that we can make sure that the ratings are not influenced by any other factors (e.g., discussing the movie with others, observing the average movie score on the Internet, etc.) between the consumption and rating. The users' goal for rating movies is to improve their profiles, express themselves and help others, according to [7].

We collected 1611 ratings from 89 users to 946 items. Average users' age is 27. Users are from 6 countries and 16 different cities. The maximum number of ratings per user is 220 and the minimum is one. The contextual variables that we collected are listed in Table 2. The decision which pieces of contextual information to acquire was made according to the definition in [6] and the specificity of our system. The benefit from the affective metadata was proved in [13], in this study we decided to use the emotional state as a contextual information. Additional information about our *Context Movies Database* (LDOS-CoMoDa) can be found in [9].

2.2 Context Relevancy Detection

The relevancy of each contextual variable in the LDOS-CoMoDa database was tested by hypothesis testing to determine the association between each contextual variable and the ratings. The null hypothesis of the test was that the

Table 2: Contextual variables in LDOS-CoMoDa database.

Contextual variable	Description
time	morning, afternoon, evening, night
daytype	working day, weekend, holiday
season	spring, summer, autumn, winter
location	home, public place, friend's house
weather	sunny/clear, rainy, stormy, snowy, cloudy
social	alone, partner, friends, colleagues, parents, public, family
endEmo	sad, happy, scared, surprised, angry, disgusted, neutral
dominantEmo	sad, happy, scared, surprised, angry, disgusted, neutral
mood	positive, neutral, negative
physical	healthy, ill
decision	user picked the item, item suggested by other
interaction	first, n-th

contextual variable and ratings 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 ratings are dependent and thus that piece of contextual information is relevant.

Since all the variables in the LDOS-CoMoDa database are categorical we decided to use the Freeman-Halton test, which is the Fisher's exact test extended to $n \times m$ contingency tables [3]. The significance level of our test was $\alpha = 0.05$. An a-priori power analysis was conducted and the results showed that the sample size is large enough for the statistical testing.

2.3 Online Survey for Context Relevancy Assessment

In order to acquire users' opinion on which contextual information is relevant we created an online survey. It contained 12 questions, one for each contextual information in the LDOS-CoMoDa database. All questions were presented in the same manner. For example, for the *day type* context: "Do you think you would rate/select a movie differently if you watched it: on a working day, weekend or holiday?". Available answers for each subject to select for each question were: *No*, *Probably not*, *Maybe*, *Probably yes* and *Yes*. All pieces of contextual information were explained and the questions presented in the participants' mother tongues.

The survey was answered by 72 subjects, from which 27 were also users in the LDOS-CoMoDa database.

2.4 Context Relevancy Assessment

Once the survey data was acquired we needed to assess which piece of contextual information is relevant and which is irrelevant, from the subjects' opinions. Since the LDOS-CoMoDa database is still small the detection could not be achieved for each user individually. Therefore, both the detection and the assessment were done for the entire population of users, so that the results can be compared. For each contextual information the assessment score was calculated as: $s = \sum_{i=1}^{5} \omega_i n_i$, where n_i is the amount of answers i and ω_i is the weight appointed to the answer i, where i goes from 1 (answer "No") to 5 (answer "Yes").

The weights were determined according to the following rules:

1. Answer "Maybe" was set as neutral. Weight for "Maybe" was set as $\omega_3 = 0$.
2. Answer "Yes" and "No", have more weight than "Probably yes" and "Probably no", respectively. Therefore we decided that $\omega_5/\omega_4 = 2$ and $\omega_1/\omega_2 = 2$.
3. Answers "Yes" and "No", and answers "Probably yes" and "Probably no" are exact opposites. $\omega_5/\omega_1 = -1$ and $\omega_4/\omega_2 = -1$

If the calculated score is $s \geq 0$ we assess that the piece of contextual information is relevant, otherwise it is irrelevant.

2.5 How Well do They Match?

The problem of determining how well the results from both approaches match, is basically the problem of determining the inter-annotator agreement. In this case we have two annotators: *the assessment* and *the detection*; five categorical classes form survey answers as the annotations: from *No* to *Yes*; and twelve pieces of contextual information to annotate. For the task we used the *Cohen's* κ coefficient, which is a statistical measure of inter-annotator agreement for categorical items [5]. This measure takes into account the agreement occurring by chance. κ is calculated by the equation:

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is the relative observed agreement among annotators and p_e is the hypothetical probability of chance agreement, $\kappa \in [-1, 1]$. The authors in [10] characterized $\kappa < 0$ as no agreement, $0 \leq \kappa \leq 0.2$ as slight, $0.21 \leq \kappa \leq 0.4$ as fair, $0.41 \leq \kappa \leq 0.6$ as moderate, $0.61 \leq \kappa \leq 0.8$ as substantial, and $0.81 \leq \kappa < 1$ as almost perfect agreement. In the case of the perfect agreement $\kappa = 1$.

We will use the same approach to test how similar the assessment from two groups of survey subjects is, one being only those subjects that are users in LDOS CoMoDa database, and the other being those subjects that are not users in the database. We will test this to see if there is a difference in opinions between the users that are already using a CARS and the ones that are not.

2.6 Rating Prediction

In order to evaluate how well each method determined relevant and irrelevant pieces of contextual information we calculated context-dependent ratings predictions. Predictions were made by the matrix factorization as a collaborative-filtering algorithm described and used in [8, 4]. We used the following equation and notations for the matrix factorization:

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

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, $\vec{q}_i$ is an item's latent feature vector, and $\vec{p}_u$ is a user's latent feature vector. $\hat{r}$, μ, b_u and b_i are scalars, and $\vec{q}_i$ and $\vec{p}_u$ are vectors. The contextual variable in the following equations will be denoted by c. We calculated the

users' and items' feature vectors using the gradient descent method [8].

Context was incorporated in the matrix factorization in two ways, by contextualizing users' biases:

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

and users' latent features:

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

separately. We decided not to contextualize the item's biases due to the small number of ratings per item in LDOS-CoMoDa database and since the context filtering of items would increase sparsity in the ratings per items and degrade the results significantly. We contextualized users' biases and feature vectors separately to inspect how relevant and irrelevant context influences biases and feature vectors.

We used the *root mean square error* (RMSE) as the evaluation measure for the predicted ratings.

2.7 Which Approach is Better in Determining Relevant Context?

In order to determine which approach is better, we need to have a measure which tells us how good each approach is. This can be measured by comparing the list of relevant and irrelevant contextual information, given by each method, with the results of rating predictions for each contextual information.

We assume that rating prediction that utilizes the relevant context will result in better predictions, i.e., lower RMSE, than the one that utilizes the irrelevant one. This means that for each pair $(c^{(r)}, c^{(i)})$ we should have $\xi(c^{(r)}) < \xi(c^{(i)})$, where $c^{(r)}$ is the piece of contextual information detected as relevant, $c^{(i)}$ is the piece of contextual information detected as irrelevant and $\xi(c)$ is the root mean square error achieved with the matrix factorization utilizing the context c.

With this assumption in mind we now count the number of times the piece of contextual information determined as relevant lead to worse results than the irrelevant one by the equation:

$$\tau = \sum_{i=1}^{n^{(r)}} \sum_{j=1}^{n^{(i)}} \varphi_{ij},$$

where $n^{(r)}$ and $n^{(i)}$ are the number of pieces of contextual information detected as relevant and irrelevant respectively and

$$\varphi_{ij} = \begin{cases} 0 & , \quad \xi(c_i^{(r)}) < \xi(c_j^{(i)}) \\ 1 & , \quad \xi(c_i^{(r)}) \geq \xi(c_j^{(i)}) \end{cases}$$

Finally the measure of performance of each method can be calculated as:

$$\psi = 1 - \frac{\tau}{n^{(r)} n^{(i)}}, \psi \in [0,1],$$

where $n^{(r)}n^{(i)}$ is the number of all possible pairs $(c^{(r)}, c^{(i)})$. In the best case, when $\xi(c^{(r)}) < \xi(c^{(i)})$ for every pair $(c^{(r)}, c^{(i)})$, $\tau = 0$; in the worst case, when $\xi(c^{(r)}) \geq \xi(c^{(i)})$ for every pair $(c^{(r)}, c^{(i)})$, $\tau = n^{(r)}n^{(i)}$. Note that when calculating φ_{ij} we could use relative RMSE difference instead of 1 when $\xi(c_i^{(r)}) \geq \xi(c_j^{(i)})$, however, since in this study we are interested in the binary decision on the contextual information relevancy, we ignore the degree of difference between the RMSE scores.

3. RESULTS

On the survey data, using the assessment method described in Section 2.4 we assessed which contextual information is relevant and which is irrelevant. Similarly, on the data from the LDOS-CoMoDa database, using the detection method described in Section 2.2, we detected which contextual information is relevant and which is irrelevant. The assessment and the detection results are presented in Table 3.

Table 3: Assessment and detection results.

Assessment		Detection	
Relevant	**Irrelevant**	**Relevant**	**Irrelevant**
time		daytype	
location	day type	location	
social	season	endEmo	time
endEmo	weather	domEmo	season
domEmo	physical	mood	weather
mood	decision	physical	social
interaction		decision	
		interaction	

Between the assessment and the detection method the calculated Cohen's κ coefficient was $\kappa = 0.118$ which is characterized as slight agreement. Between the *users* and the *non users* groups of subjects the calculated coefficient was $\kappa = 0.833$ which is an almost-perfect agreement.

Figure 1 shows the average RMSE from both approaches i.e., contextualized users' biases and contextualized users' latent-feature vectors, for all the collected contextual variables. Rating for the items in the dataset are from one to five.

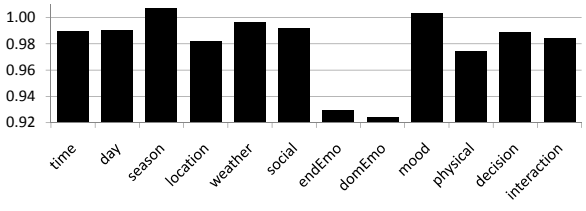

Figure 1: Average RMSE from both matrix-factorization approaches per contextual information. Ratings in the dataset are from one to five. These results are used as a ground truth for the detection and the assessment evaluation.

Performance of each method was evaluated by the method described in Section 2.7. For the assessment we obtained $\psi_a = 0.629$ and for the detection $\psi_d = 0.969$.

3.1 Discussion

The results presented in the previous section can help us answer the questions proposed in the problem statement (Section 1.2).

How well do the outcomes from these approaches match? There was only a slight agreement between the assessment and the detection approach in determining relevant and irrelevant contextual information ($\kappa = 0.118$). The difference between the assessment and the detection can also be seen in Table 3. This is due to the fact that the detection is made on the real ratings data and the assessment depends on the users' ability to imagine a hypothetical situation. This result agrees with the conclusions in [12].

Which approaches is better in determining relevant context? If we compare $\psi_a = 0.629$ and $\psi_d = 0.969$ we can conclude that for the determination of the relevant and the irrelevant context, the detection from the rating data performs better for the rating prediction task than the assessment from the users' opinion. In other words, using those pieces of contextual information that were detected as relevant will lead to better rating prediction than when using those that were assessed as relevant.

Are users aware of what influences their decisions? Almost-perfect agreement ($\kappa = 0.833$) between the survey subjects that were the users in our CARS and those that were not suggests that there is a sort of an overall population opinion on what context could influence their decisions regarding movies. This also means that there is no difference between the users that are familiar with using CARS and those that are not. However, only a slight agreement between the assessment and the detection, and the fact that the detection performs better, tells us that users are not entirely aware of what really influences their decisions (i.e., ratings) in a movie domain. This is important since the determination of context relevancy from the survey data can lead (as in this case), to using harmful pieces of information, and ignoring the relevant ones.

For the detection, assessment and rating prediction, in this article, we used each contextual information independently. We will inspect these effects on the multiple, combined context models in the future work.

3.2 Conclusion and Future Work

In this study we compared two approaches for determining relevant and irrelevant pieces of contextual information in a movie RS: the assessment from the users' survey and the detection from the rating data. We used a real rating data, that we collected in the LDOS-CoMoDa database, and a survey data to test how well these approaches match and which performs better for the rating prediction task. To evaluate each approach we used a contextualized matrix-factorization algorithm. The results showed that there is a difference between the outputs of these two approaches and that the detection performs better. This points to the fact that the users are not necessarily aware of what influences their decisions in a movie domain. Still, the assessment could be a valuable approach since it can be used a priori, i.e., before any rating data is collected. However, once the rating data is acquired, the detection should be employed since it will provide better insight into which piece of contextual information is relevant and should be used to improve the recommendations. Our future work consists of inspecting other statistical methods for the detection of relevant context, for different variable types. We are also interested in inspecting the detection on the user level, i.e., for each user separately.

4. REFERENCES

[1] 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)*, 23(1):103–145, 2005.

[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*, 17(6):734–749, June 2005.

[3] A. Agresti. A Survey -of Exact Inference for Contingency Tables. *Statistical Science*, 7(1):131–153, 1992.

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

[5] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] 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*, pages 304–307, 1999.

[7] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[8] 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*, pages 426–434. 2008.

[9] A. Košir, A. Odić, M. Kunaver, M. Tkalcic, and J. Tasic. Database for contextual personalization. *ELEKTROTEHNIŠKI VESTNIK*, 78(5):270–274, 2011.

[10] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):pp. 159–174, 1977.

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

[12] C. Ono, Y. Takishima, Y. Motomura, and H. Asoh. Context-Aware Preference Model Based on a Study of. In *User Modeling, Adaptation, and Personalization*, pages 102–113, 2009.

[13] M. Tkalčič, U. Burnik, and A. Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, Sept. 2010.

[14] C. Toutain, F. and Bouabdallah, A. and Zemek, R. and Daloz. Interpersonal Context-Aware Communication Services. *IEEE Communications Magazine*, (January):68–74, 2011.