User Model User-Adap Inter (2014) 24:35–65
DOI 10.1007/s11257-012-9135-y

ORIGINAL PAPER

Comparing context-aware recommender systems in terms of accuracy and diversity

Umberto Panniello · Alexander Tuzhilin · Michele Gorgoglione

Received: 22 March 2012 / Accepted in revised form: 30 October 2012 /
Published online: 27 December 2012

Abstract Although the area of context-aware recommender systems (CARS) has made a significant progress over the last several years, the problem of comparing various contextual pre-filtering, post-filtering and contextual modeling methods remained fairly unexplored. In this paper, we address this problem and compare several contextual pre-filtering, post-filtering and contextual modeling methods in terms of the accuracy and diversity of their recommendations to determine which methods outperform the others and under which circumstances. To this end, we consider three major factors affecting performance of CARS methods, such as the type of the recommendation task, context granularity and the type of the recommendation data. We show that none of the considered CARS methods uniformly dominates the others across all of these factors and other experimental settings; but that a certain group of contextual modeling methods constitutes a reliable "best bet" when choosing a sound CARS approach since they provide a good balance of accuracy and diversity of contextual recommendations.

Keywords Context-aware recommender systems · CARS · Pre-filtering · Post-filtering · Contextual modeling · Accuracy · Diversity · Performance measures

U. Panniello · M. Gorgoglione (✉)
Department of Mechanics, Management and Mathematics, Politecnico di Bari, Viale Japigia 182, 70126, Bari, BA, Italy
e-mail: m.gorgoglione@poliba.it

A. Tuzhilin
Stern School of Business, New York University, 44 West 4th Street, Room 8-92, New York, NY 10012, USA

1 Motivation and introduction

The importance of the contextual information in Recommender Systems (RS) has been recognized for some time (Adomavicius and Tuzhilin 2001), and as a result, the Context-Aware Recommender System (CARS) field has been formed. Although there exist several different approaches to incorporating context into the recommendation process, the majority of the CARS papers focus on the *representational* view (Dourish 2004) that assumes that the context is a priori known and is defined by several contextual factors having a known hierarchical structure that does not change significantly over time (Adomavicius and Tuzhilin 2011).

In Adomavicius and Tuzhilin (2008) and Adomavicius and Tuzhilin (2011) different representational approaches were categorized into pre-filtering, post-filtering and contextual modeling methods as follows:

1. *Contextual pre-filtering (PreF)* assumes that the contextual information is used to filter out irrelevant ratings *before* they are used for computing recommendations with standard (non-contextual) methods.
2. *Contextual post-filtering (PoF)* assumes that the contextual information is used *after* the standard non-contextual recommendation methods are applied to the recommendation data.
3. *Contextual modeling (CM)* assumes that the contextual information is used *inside* the recommendation-generating algorithms together with the user and item data.

Moreover, Adomavicius and Tuzhilin (2011) challenged the CARS community to study these three approaches further and also to compare them to determine which one outperforms the others and under which circumstances. Although there have been some initial studies on comparing these approaches, as described in Sect. 2, no systematic comparison has been done so far in order to determine which one dominates the others and under which circumstances. Therefore, the challenge of Adomavicius and Tuzhilin (2011) still remains pretty much open.

In this paper, we pursue this challenge and provide an extensive comparison of certain types of pre-filtering, post-filtering and contextual modeling methods in terms of the predictive performance and diversity measures in order to identify which of the CARS methods outperform the others and under which circumstances. We also show empirically that, although there are no clear winners among the CARS methods considered in this paper that uniformly outperform the alternative approaches, some CARS methods provide the best solutions in certain circumstances discussed in the paper.

The issue of comparing different approaches to CARS is important not only to the academic community, but also to the industry for several reasons. First, businesses operate in different and changing conditions, such as the channels through which recommendations are delivered to users or the goals on which recommendations are based. Therefore, incorporating contextual information into the recommendation process in the best possible manner is an important problem for some of them. For instance, Amazon (Linden et al. 2003) delivers product recommendations via its Web site where many products are displayed, and through the electronic newsletter emailed to the customers and containing only very few product recommendations because of

the limited space that does not permit to send a long list of recommendations. Therefore, Amazon, obviously, deploys two very different recommendation tasks across these two recommendation delivery mechanisms, i.e., show many good items in the order of their recommendation importance task ("find all good items") when delivering recommendations via its Web site, and show the "top-k" items task when using an email newsletter.

LinkedIn constitutes another example of using different recommendation strategies in different contexts. When a user is actively looking for a job position, it makes sense for LinkedIn to present all the suitable job opportunities fitting her profile, even including less attractive job postings (because it is crucial for an active job seeker to have many job leads). When users are not active, however, it makes sense for LinkedIn to recommend only the very few best job opportunities to users in order to avoid bothering them with unattractive recommendations. Therefore, not only LinkedIn should deploy both recommendation tasks, but it should pursue different objectives that require different performance metrics, i.e., increasing recall when users are looking for any good opportunity and may accept less useful recommendations, and increasing precision when users do not want to be bothered with useless recommendations. Therefore, knowing how accurate and diverse CARS methods are in different settings can turn out to be crucial for companies for building effective and lasting relationships with their customers and increasing their competitiveness in the market.

Most of the comparisons in recommender systems are done for a single performance measure and this measure being predictive accuracy in many studies. However, other performance measures are also crucial in many settings, and it is often important for the companies using RS to consider other such measures, including diversity (Adomavicius and Kwon 2012) and determine how to balance accuracy of recommendations with their diversity. As an example, many online sellers of comic books email their users a weekly newsletter. If the newsletter displays very accurate recommendations, the users will be recommended books related to their favorite characters and thus will see books about the same set of characters over and over again. Such a newsletter will be accurate but, probably, very boring, and the user will lose interest and stop reading it after a few weeks. The solution to this problem is to increase the diversity of recommendations by including items in the newsletter related to different characters in order to stimulate users' curiosity. However, if these recommendations are quite inaccurate, then the risk is again that the users will lose their interests, leading to counterproductive results. The conclusion is that accuracy and diversity are very different performance measures which have to be carefully considered when comparing different recommender systems.

In this paper, we compared the performance of various pre-filtering, post-filtering and contextual modeling CARS methods in terms of their accuracy and diversity measures across various experimental conditions. We identified three key factors, including the type of the recommendation task, "Find-All" vs. "Top-k"; context granularity, coarse vs. fine granularity of the contextual information; and the type of the data set, each dataset being characterized by different levels of sparsity of ratings and heterogeneity of customers' behavior. These factors are described at length in Sect. 3.

2 Prior research

There has been much work done on CARS since the early publications on this topic, such as Adomavicius and Tuzhilin (2001), and most of this work is reviewed in Adomavicius and Tuzhilin (2011) and also in Adomavicius et al. (2011). Context-aware approaches have become popular in many areas, and applications have been recently explored in several fields, such as music (Reddy and Mascia 2006; Kaminskas and Ricci 2011; Hariri et al. 2012), movies (Said et al. 2011), travel and tourism (Cena et al. 2006; Baltrunas et al. 2011; Ge et al. 2011; Levi et al. 2012), mobile recommendations (Ricci 2011), personalized shopping assistants (Sae-Ueng et al. 2008), conversational and interactional services (Mahmood et al. 2010), learning-related services (Wang and Wu 2011), social rating services (Feng et al. 2012) and multimedia (Fagà et al. 2009). According to Adomavicius et al. (2011), various CARS approaches can be categorized based on what is known about the contextual factors and also how fast the available contextual information changes over time. One particularly important case is when the contextual information is fully observable and is static, i.e., does not change significantly over time. This case corresponds to the well-known representational view of contextual information introduced in Dourish (2004), and most of the CARS papers follow this representational approach. A different approach, called dynamic contextualization, is proposed by Hussein et al. (2014).

Furthermore, Adomavicius and Tuzhilin (2008) and Adomavicius and Tuzhilin (2011) categorized various representational approaches into pre-filtering, post-filtering and contextual modeling methods. Pre-filtering assumes that the contextual information is used to filter out irrelevant ratings before they are used for computing recommendations with standard (non-contextual) methods. One specific example of the pre-filtering approach is the "exact pre-filtering" (Gorgoglione and Panniello 2009; Adomavicius and Tuzhilin 2011) where the irrelevant ratings not corresponding exactly to the specific context of interest and *only* they are pre-filtered before the remaining relevant ratings are used for generating recommendations. Another example of the pre-filtering approach was proposed by Baltrunas and Ricci (2014) who introduced a pre-filtering technique called "item splitting" and studied it in different settings. Similarly to the item-splitting idea, Baltrunas and Amatriain (2009) introduce the idea of microprofiling, which splits the user profile into several (possibly overlapping) subprofiles, each representing the given user in a particular context. Post-filtering assumes that the contextual information is used after the standard non-contextual recommendation methods are applied to the recommendation data. One specific example of post-filtering is Filter-PoF (Panniello et al. 2009) when the ratings computed by the classical CF method are adjusted by filtering out those ratings having a low probability to be relevant in a given context. Post-filtering approach was also investigated in Bader et al. (2011) and Cremonesi et al. (2011) and compared to the uncontextual case, where it was shown that the proposed Post-filtering methods outperformed the uncontextual one. Finally, the contextual modeling (CM) approach assumes that the contextual information is used inside the recommendation-generating algorithms together with the user and item data (Adomavicius et al. 2005). One specific example of the contextual modeling approach based on the SVMs was presented and compared to the uncontextual case in Oku et al. (2006).

Furthermore, Adomavicius and Tuzhilin (2008) and Adomavicius and Tuzhilin (2011) challenged the researchers to compare these three CARS approaches to determine which one outperforms the others and under which circumstances. This challenge was taken in Panniello et al. (2009), where the pre- and the post-filtering approaches were compared to each other and to the uncontextual case, and it was shown that this comparison depends, to a large extent, on the type of the post-filtering method used. This initial study was further extended in Panniello and Gorgoglione (2012) where the contextual modeling approach was added to the study, and the three methods were compared among themselves and to the uncontextual case. It was shown that the pre-filtering and contextual modeling methods slightly outperform the uncontextual case while the post-filtering method outperforms the uncontextual one depending on how the post-filtering method is implemented. In particular, it was shown that when the post-filtering method is realized in the right way, it constitutes the best-of-breed contextual method. On the contrary, if it is realized in a poor way, it can be the worst contextual method. Furthermore, Panniello and Gorgoglione (2012) proposed an effective way of selecting the best alternative method between various CARS approaches and an uncontextual one.

Although Panniello et al. (2009, 2012) shed some light on the tradeoffs between the contextual pre-filtering, post-filtering and contextual modeling approaches, this was still an initial type of work that was limited in the following sense. It (a) provided only the marginal analysis and did not identify the regions where one approach outperforms the others; (b) compared the three approaches only in terms of accuracy and did not consider any diversity measures; (c) made a comprehensive comparison between the CARS methods and the uncontextual method, while the comparison among different CARS methods was fairly basic; (d) did not make any statements whether the observed differences in predictive accuracy where statistically significant or not. In other words, Panniello et al. (2009) and Panniello and Gorgoglione (2012) provided only the first attempts to compare the pre-filtering, post-filtering and contextual modeling methods and did not fully explain when a CARS approach outperforms the others and under which circumstances. Therefore, the challenge reported in Adomavicius and Tuzhilin (2011) remains pretty much open.

In this paper, we pursue the challenge of Adomavicius and Tuzhilin (2011) further and strive to provide a much more comprehensive comparison across various contextual pre-filtering, post-filtering and contextual modeling approaches in order to develop a deeper understanding of their tradeoffs. In particular, in this paper we compare the three approaches not only in terms of the predictive accuracy, but also in terms of diversity of recommendations and do this on a significantly more comprehensive data, using a much better "regional" comparison method (vis-a-vis a limited version of marginal comparison, as was done in Panniello and Gorgoglione (2012)). We do this comparison in a statistically more rigorous fashion. Moreover, after comparing CARS methods in terms of, separately, accuracy and diversity, we also compare them by combining the accuracy and diversity measures. The goal is to identify the CARS methods that provide the better balance of the two performance measures, which we believe is very important issue for industrial applications.

Comparing recommender systems in terms of diversity is not new, and it has been done in prior research including Mcginty and Smyth (2003), Ziegler et al. (2005),

Zhang and Hurley (2008), Adomavicius and Kwon (2009), Hu and Pu (2011), Adomavicius and Kwon (2012), and De et al. (2012). A very recent analysis of diversity in Time-Aware Recommender Systems is given by Campos et al. (2014). Typical approaches would replace items in the derived recommendation lists to minimize similarity between all items or remove "obvious" items from the list of recommendations, as was done in Billsus and Pazzani (2000). Adomavicius and Kwon (2009, 2012) present the concept of aggregated diversity as the ability of a system to recommend across all users as many different items as possible over the whole population while keeping accuracy loss to a minimum, which is achieved by a controlled promotion of less popular items towards the top of the recommendation lists. Furthermore, a trade-off between accuracy and diversity was established in Adomavicius and Kwon (2009) and further confirmed in Gorgoglione et al. (2011), where it was shown that ranking recommendations according to the predicted rating values provides good predictive accuracy but it tends to perform poorly with respect to recommendation diversity. Moreover, Hu and Pu (2011) investigated design issues that can enhance users' perception of recommendation diversity and improve users' satisfaction.

Despite all this research on recommendation diversity, few of the prior publications study diversity of recommendations in the context of CARS. One example of such work is presented in Gorgoglione et al. (2011) where it was demonstrated that CARS can increase diversity while preserving accuracy. It was also argued in Gorgoglione et al. (2011) that just focusing on accuracy alone is not enough, and it is also important to use other measures, such as diversity when studying CARS. In this paper, we pursue this idea further and compare pre-filtering, post-filtering and contextual modeling methods in terms of both accuracy *and* diversity measures.

3 Methodology

As explained before, in this paper we conduct an extensive empirical comparison of the pre-, post-filtering and contextual modeling approaches. As a pre-filtering method, we selected the Exact contextual Pre-Filtering (EPF) (Adomavicius and Tuzhilin 2011) that uses contextual information for filtering out all the ratings not corresponding exactly to the specified context before the recommendation method is launched. As a post-filtering method, we have chosen two approaches, i.e., the Filter Post-Filtering (Filter PoF) and the Weight Post-Filtering (Weight PoF) methods (Panniello et al. 2009). In both of these methods, the recommendations are first generated by using the standard uncontextual recommendation methods on the $User \times Item$ matrix without any references to the contextual information. Then the computed uncontextual ratings are contextualized by estimating the probability with which a user chooses a certain item in a given context. The contextual probability $P_c(i, j)$, with which user i selects item j in context c, is computed as the number of neighbors who selected the same item in the same context, divided by the total number of neighbors in the neighborhood. While Filter PoF method contextualizes recommendations by filtering out those ratings r_{ij} having probability $P_c(i, j)$ below a certain threshold, the Weight PoF re-computes new contextualized ratings as $r'_{ij} = r_{ij} * P_c(i, j)$ and adjusts contextualized recommendations based on ratings r'_{ij}. We have selected these two post-filtering

approaches because they have been previously proposed in the literature and demonstrated reasonable performance (Panniello et al. 2009), and also because they perfectly fit our empirical studies without any need for adjustments or modifications. Also note that the EPF method is unique for the exact pre-filtering; therefore, we did not need to consider any alternatives for EPF.

We also consider four types of the Contextual Modeling (CM) approach, i.e., Mdl_1, Mdl_2, Mdl_3, Mdl_4 (Panniello and Gorgoglione 2012). For these CM methods, we first build a contextual profile *Prof(i,c)* for the ith user in context c, and then use the contextual profiles of all the users to find the N nearest neighbors of the ith user in context c. The four types of the CM approaches vary in the constraints by which the neighbors are selected. In Mdl_1 there is no constraint in the selection of the N neighbors which can be found in any context at any level of the hierarchy. In Mdl_2 we select an equal proportion of neighbors from each context c regardless of the context hierarchy. In Mdl_3 we select N neighbors from each context c and each level of the context hierarchy. In Mdl_4 we select an equal proportion of neighbors from each context c at the same level of context hierarchy. We selected the Mdl_1, Mdl_2, Mdl_3, Mdl_4 methods to represent the CM approach because they have been proposed and studied before Panniello and Gorgoglione (2012) and because they produced good performance results vis-à-vis other methods considered in the paper, as will be shown in Sect. 4. Finally, we focused on collaborative filtering in this paper since it is a very popular approach in recommender systems and since several CARS methods have been already developed for it.

We compare all the three described CARS approaches across a broad set of experimental conditions. In the next section, we describe the datasets used in our study.

3.1 Datasets

We used three dataset from three different e-commerce Web sites in our experiments. The first dataset (DSet 1) is taken from the study described in Palmisano et al. (2008). First, a special purpose browser was developed to help users navigate Amazon.com website and purchase products on its site. This browser was made available to a group of students who were asked to navigate and simulate purchases on Amazon.com during a period of 4 months based on the incentive scheme developed for this study. While navigation was real on Amazon.com, purchasing was simulated. Once a product was selected by a student to be purchased, the browser recorded the selected item, the purchasing price and other useful characteristics of the transaction and this information was stored in the database. In addition, the student was asked at the beginning of each browsing session to specify its context, what was the intent of a purchase in our case, i.e., whether the purchase would be intended for personal purpose or as a gift, for which specific personal purpose, and for whom the gift was intended. The structure of this contextual variable IntentOfPurchase is presented in Fig. 1a. Further, the data was pre-processed by excluding the students who made >40 transactions and eliminating the students who had any kind of misleading or abnormal behavior. The resulting number of students was 556, and the total number of purchasing transactions for the students was 31,925.

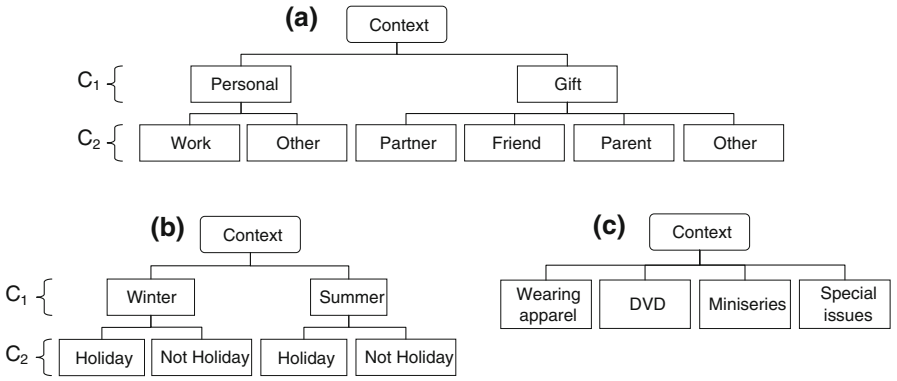

Fig. 1 Hierarchical structure of context: **a** IntentOfPurchase, **b** TimeOfTheYear and **c** Store

The second dataset (DSet 2) comes from an e-commerce website commercially operating in a certain European country which sells electronic products to approximately 120,000 users and contains about 220,000 purchasing transactions during an observation period of 3 years. For this dataset, we selected the time of the year as a contextual variable. Its hierarchical structure is presented in Fig. 1b. The classification into Summer or Winter and Holiday or Not Holiday is based on the experiences of the CEO of the e-commerce website that we used in our study. He defined June, July, August, April, May and September as "Summer". The first 3 months of this period are considered as "Holiday" while the remaining as "Not Holiday". Also he defined October, November, December, January, February and March as "Winter". The first 3 months of this period are considered as "Holiday" while the remaining as "Not Holiday". According to this definition, a purchase made, for example, on December 1 is labeled as "Winter Holiday". The CEO made these non-standard classifications of the seasons and the holidays because of the practical realities of his business and his extensive experiences with the merchandising process. The data was pre-processed by excluding about 80,000 customers who made only one single transaction (for these customers, it is hard to generate any meaningful recommendations due to the lack of preference data), around 500 customers who had any kind of abnormal behavior such as buying the same product for 1,000 times at the same time (this was probably a retailer), and around 38,000 customers who had transactions either only in the first 2 years or only in the third year. The reason for this last elimination is that we used the transactions in the first 2 years as training set and those in the third year as validation set, as explained below. The resulting dataset contained 1,633 users and 7,332 transactions.

The third dataset (DSet 3) comes from an e-commerce website which sells comics and comics-related products, such as T-shirts, DVDs and various gadgets. It contains about 22,845 transactions and 2,174 users. In this case, we used the store (i.e., the section in the Web site where products are bought), as a contextual variable, distinguishing whether the product is bought in "Wearing apparel", "DVD", "Miniseries" or "Special issues" section (store) of the website (see Fig. 1c). This contextual variable store specifies the immediate browsing activity in which the customer was engaged

just before the recommendation by identifying the location of the customer on the website. The importance of this contextual variable comes from the expectation that customers' behavior changes when navigating and buying products in different sections of the Web site. For instance, the same customer can behave differently when purchasing a comic book versus when purchasing other categories of products, such as a T-shirt. In a real-time recommender system, when a customer enters a specific store of the website, the system should use this context (the store type) to focus mainly on the recommendations pertinent to that store. Feedbacks from users are always implicit, representing the purchasing frequencies.

As shown in Fig. 1, we only consider context hierarchies with one or two levels, even though it is possible to know more detailed contextual information in some applications and, therefore, use more structured contextual hierarchies. Although more detailed contextual information can lead to a better knowledge about the CARS system, it can also have some negative implications. For example, obtaining this more detailed information may require a CARS system to bother the user with various questions when extracting this information from him/her. This can even upset the user in some cases. Therefore, research suggests that it is better to have contextual hierarchies of one or two levels (Adomavicius and Tuzhilin 2011) because such shallow hierarchies require significantly fewer questions for the user and because this shallow contextual information can carry a long way in terms of providing accurate recommendations, as is shown in Sect. 4 of the paper.

Each of these three datasets has unique properties, such as certain levels of sparsity of its ratings and heterogeneity of behavior of its customers. Therefore, we characterize each of these three datasets by the levels of its sparsity and customer heterogeneity as follows. In the first dataset (DSet 1) sparsity ranges from 52 % (uncontextual matrix) to 71 % (on average for the contextual matrices). In the second dataset (DSet 2) it ranges from 82 % (uncontextual matrix) to 86 % (on average for the contextual matrices). In the third dataset (DSet 3) it ranges from 98 % (uncontextual matrix) to 99 % (on average for the contextual matrices). Note that the sparsity levels vary across different CARS settings because sparsity is a measure characterizing the $User \times Item$ matrix, and each CARS uses its own $User \times Item \times Context$ matrix and therefore having its own level of sparsity corresponding to that matrix. To measure heterogeneity of customers' behavior for each dataset, we use the Shannon's Entropy, as defined in Sect. 3.2 below. The heterogeneity of customer behavior is measured for the particular "$User \times Item$" matrix and therefore it measures how much the customers behaved differently across items and contexts (on average) for each application. Therefore, it represents the "customers' behavior heterogeneity". It is computed by applying the Shannon's Entropy formula to each user's vector in the "known" $User \times Item$ matrix. The results are then averaged over all the customers. In the first dataset, entropy is 65.63 %, in the second dataset it is 29.50 %, while in the third dataset the entropy is 9.79 %. These statistics about the sparsity and heterogeneity properties of the three datasets are summarized in Table 1. The three very different characteristics of these three datasets are due to significant variations of customer's behavior across the three very different e-commerce applications. Customers in the first dataset buy often and buy various kinds of products. This behavior causes low level of sparsity and high level of heterogeneity (entropy). On the contrary, users in the third dataset buy rarely

Table 1 Type of data represented by sparsity and heterogeneity in the User-Item-Context matrix

Type of data	Sparsity (S)	Heterogeneity (H)
DSet 1	52–71 %	65.63 %
DSet 2	82–86 %	29.50 %
DSet 3	98–99 %	9.79 %

and tend to purchase the same or similar kinds of products. This behavior causes high level of sparsity and low level of entropy. The second dataset is somewhere in between the other two in terms of its levels of sparsity and entropy.

3.2 Performance measures

We used recommendation accuracy and diversity measures when comparing performance of pre-filtering, post-filtering and CM methods in our study. The recommendation accuracy is measured by Precision, Recall and F-measure (Herlocker et al. 2004). We computed Precision and Recall as follows. For the "find all good items" strategy, we set the threshold between relevant and irrelevant items equal to 1, thus, assuming that if an item is selected more than once by a customer, it is relevant ("good") for that customer, and we recommend it; otherwise, we did not. Then, we verified if the recommended item was actually selected in the validation set. If it was, we considered that as a "good" recommendation, otherwise as a "bad" one. For the "recommend top-k items" strategy, we determined the top-k items as "good" items to be recommended to a user. Then we compared those with the actual items selected by the user to compute Precision and Recall in a standard manner. Finally, we divided each dataset into the training and the validation sets, the training set containing 2/3 and the validation set 1/3 of the whole dataset. For the DSet 1 dataset, the first 2 years were the training set and the third year was the validation set. For the DSet 2 dataset, we randomly split it in 2/3 for the training set and the remaining 1/3 for the validation set (in this case, it was impossible to make a good temporal split because all the transactions were made within a couple of months). For the DSet 3 dataset, the first 9 months were the training set and the last 3 months were the validation set.

We measured the recommendation diversity in our experiments using the classification of diversity metrics in probability-based, logarithm-based and rank-based measures (McDonald and Dimmick 2003) and selecting popular measures from each of the three categories, i.e., the Simpson's diversity index, the Shannon's entropy and the Tidemann & Hall's index (McDonald and Dimmick 2003) respectively. The normalized Simpson's diversity index (D) is defined as:

$$D = \frac{1 - \sum_i p_i^2}{(1 - \frac{1}{k})}$$

where p_i is the proportion of recommended items in the ith category and k is the number of categories. The denominator of the formula is a normalization factor. Dividing by this factor is needed because we want to compare the diversity in three different

datasets, each one characterized by a different number of categories. In this case, the general Simpson's diversity index (the nominator in the previous formula) takes a different maximum value in each dataset, so making a comparison meaningless. On the contrary, the maximum value of the normalized index is 1 independently of the number of categories in each dataset. The normalized Shannon's diversity index (E) is computed as:

$$E = -\sum_i p_i \log_k p_i$$

where p_i is the proportion of recommended items in the ith category and k is the number of categories. In this case the normalization factor is the base of the logarithm which is set equal to k, i.e. the number of categories. Using the normalized Shannon's index allows us to compare the diversity of the same CARS in different datasets because its maximum value is always equal to 1. The Tidemann & Hall's diversity index (TH)[1] is measured as:

$$TH = 1 - \frac{1}{(2\sum_i rp_i) - 1}$$

where r is the rank of the ith category (ranked with 1 as the largest category). In order to provide each dataset with a ranking of categories, we used the number of distinct items contained in each category as defined by the relative website. Therefore, the category with the highest number of distinct items is ranked with 1. In the case of TH there is no need to normalize the index because it always tends to 1 when the number of items increases, and therefore 1 is always the maximum value that TH can take.

It is useful to underline the difference between the measurement of "recommendation diversity" described in the last paragraph and that of "customers' behavior heterogeneity" described in Sect. 3.1. The latter measures the Shannon's entropy in each user's vector in the known $User \times Item$ matrix and represents the heterogeneity in the customers' behavior in each application. The former measures the diversity in the vector of recommended items generated by the CARS for each user and represents the ability of a CARS to generate diverse recommendations.

3.3 Experimental settings

We conducted our experiments across the following three main settings. First, we analyze the CARS' accuracy and diversity in the two most popular recommendation tasks, "finding all good items" (Find-all) and "recommending the top-k items"

[1] The Tidemann-Hall index is very similar to the Rosenbluth index, the only difference between them is that they rank categories differently. The categories are ranked in ascending order in the Rosenbluth index and in descending order in the Tidemann-Hall index. We decided to use the Tidemann-Hall index since our aim was to compare datasets with different categories (Meilak 2008) and the Tidemann-Hall is better suited for this task than the Rosenbluth index because it is more sensitive to the absolute number of categories (the largest category receives a weight equal to one in the formula).

(Top-k). In the "find all good items" approach, the recommender system suggests all the "recommendable" items, i.e., the items having the rating value above a certain threshold. In our experiments, we selected the threshold value for the rating being 2 because it is a popular practice in similar recommendation and other settings, such as Information Retrieval (i.e., if an item is selected more than once, it is "good enough" for the recommendation purposes). Furthermore, since all the recommended items are ordered based on the value of their ratings, the choice of a particular threshold value for the "find-all-good-items" strategy is not very crucial because the "long tail" of the recommendation list will, most likely, not be seen by the user in any case since the user will mainly focus only on the first couple of pages of recommended items. In the "recommend top-k items" approach, only the "top-k" items having the k highest ratings for a particular user are recommended to that user. In our study, we varied the number of top-k recommended items from 1 to 4 in order to focus on the recommendations of only the very best items, as is often done in recommendation applications. However we present the results only for $k = 4$ because they do not change significantly when k is varied between 1 and 4 and therefore additional numbers for other values of k do not add any significant insights to the paper.

Second, we analyze the performance of our methods at the following two levels of contextual granularity. In two out of three datasets context is represented by a 2-level hierarchy (see Fig. 1). At the first level (C_1) the granularity of the contextual information is coarser, at the second level (C_2) the granularity is finer. In the three datasets context represents the "period of the year", the "intent of a purchase", the "store" where items are bought, respectively (additional details are presented in Sect. 3.1).

Third, we analyze accuracy and diversity of the CARS approaches varying the type of data used by the recommender systems to generate recommendations. The three datasets are characterized by different structures of the User-Item-Context matrix. We considered two main features to characterize the matrix, the data sparsity and the heterogeneity of customers' behavior. The data sparsity is measured as the number of empty cells in the *User* $\times$ *Item* matrix divided by the total number of cells. As it was mentioned above, the heterogeneity of customers' behavior is measured by looking at how many items customers had purchased in each product category, that is by computing the average entropy of each customers' vector of known ratings. High entropy means that the behavior is heterogeneous, while low entropy means that the behavior is homogeneous. The combination of *User* $\times$ *Item* $\times$ *Context* matrix's entropy and sparsity may describe the type of data used by the recommender system and it may affect recommendations performance. In fact, it was shown that both these parameters affect recommender systems' performance (Herlocker et al. 2004).

In summary, we used 3 different data sets, 7 contextual approaches (one pre-filtering, two post-filtering and four contextual modeling), 6 different performance measures, 2 recommendation strategies, 4 values of k for the "top-k" strategy, and 3 different contextual variables in our experiments. As a result, we have conducted 3,780 individual experiments in total. In the next section, we present the results of our experiments described in this section.

4 Results

In this section we present the results of our empirical study described in Sect. 3. In particular, we examine the effects of the three main factors considered in our study and described in Sect. 3 (i.e., recommendation task, context granularity and the type of data) on the performance of different CARS methods (pre-filtering, post-filtering and CM methods) in terms of the accuracy and diversity of recommendations that these methods provide. We start our presentation by summarizing the results of a *marginal* analysis in Sect. 4.1. This analysis examines how each of the three main factors *separately* affects the performance of the CARS methods. Although it is important to know how each of the factors separately affects the performance of the CARS methods by doing the marginal analysis, it is the *regional* analysis that constitutes the determining factor in comparison of various CARS approaches. Unlike the marginal analysis, the regional analysis determines how each region in the 3-dimensional factor space, defined by the combination of the recommendation task, context granularity and the data type, affects the performance of various CARS methods in terms of their accuracy and diversity measures. Therefore, this regional analysis constitutes the core of this section because it answers the main research question of which of the CARS approaches dominates the others and in which circumstances (i.e., regions of the factor space), whereas the marginal analysis provides additional evidence for answering the main research question. For this reason, and due to the space limitation, Sect. 4.1 presents the summary of the results of the marginal analysis. The specific results of the marginal analysis are reported in much greater detail in the working paper (Panniello et al. 2012).

4.1 Marginal analysis of accuracy and diversity of various CARS methods

We first analyzed the effect of the two recommendation tasks, "finding all good items" (Find-all) vs. "recommending the top-k items" (Top-k), on recommendations accuracy and diversity. We computed the average value of each accuracy and diversity metric of each CARS across all the experimental settings excluding the recommendation task. Both accuracy and diversity change when the recommendation task changes, but no clear trend emerges. The Precision increases when moving from "Find-all" to "Top-k". The Recall decreases when moving from "Find-all" to "Top-k". As a combination of these results, the F-measure of CARS is slightly higher in the "Find-all" task and lower in the "Top-k" task. When moving from "Find-all" to "Top-k," the Simpson's D slightly increases, the Shannon's E strongly increases, while the TH index decreases. The results are shown in Fig. 2 a and b, for the sake of brevity only the F-measure and the Shannon's E, respectively, are reported.

We then examined the effect of context granularity on accuracy and diversity. The results are somehow different with respect to the previous case, because while accuracy changes significantly when context granularity changes, diversity does not. All the accuracy measures decrease when context becomes "Fine". Figure 3a reports this result in terms of F-measure. This behavior is quite expectable in RS because when context becomes finer, the quantity of information available in each context decreases

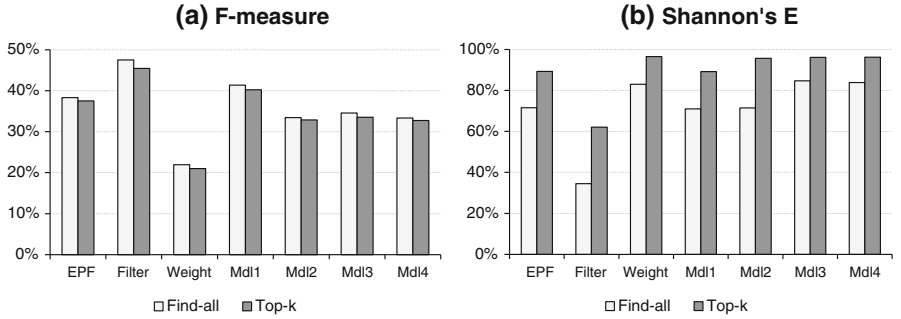

Fig. 2 Effect of the two recommendation tasks (Find-all vs. Top-k) on recommendations accuracy (**a**) and diversity (**b**) of various CARS methods

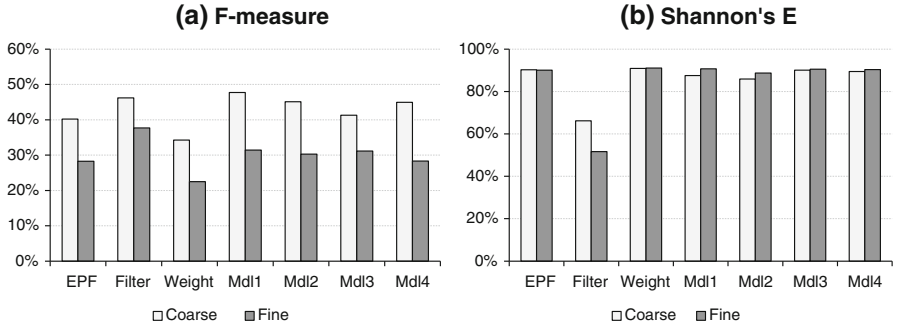

Fig. 3 Effect of context granularity (coarse vs. fine) on recommendations accuracy (**a**) and diversity (**b**) of various CARS methods

thus making the prediction problem harder. On the contrary, diversity changes very little across the two levels of context granularity for all the CARS approaches. Figure 3b reports this result in terms of Shannon's E. The results for each CARS method are discussed further in the working paper (Panniello et al. 2012).

Finally, we examined the effects of the type of data on accuracy and diversity of CARS. In general, accuracy increases when moving from DSet 1 to DSet 3, except that for Weight PoF. On the contrary, the diversity generated by all the CARS decreases when moving from DSet 1 to Dset 3. The results are shown in Fig. 4a and b where the F-measure and the Shannon's E, respectively, are reported. The reason for this is that the heterogeneity of customers' behavior decreases from DSet 1 to DSet 3, i.e., DSet1 exhibits the most diverse behavior of customers and DSet3 the least one. The heterogeneity of customers' behavior across contexts is beneficial when the goal is to generate and deliver diverse recommendations, while it is detrimental for accuracy because it decreases the ability of any recommender system to correctly predict the preferences of a user.

In summary, we examined how recommendation task, context granularity and the type of data, *individually* affect accuracy and diversity of recommendations across different CARS methods. We found that the recommendation task affects both accuracy and diversity in a way which depends on the specific measure considered. Context

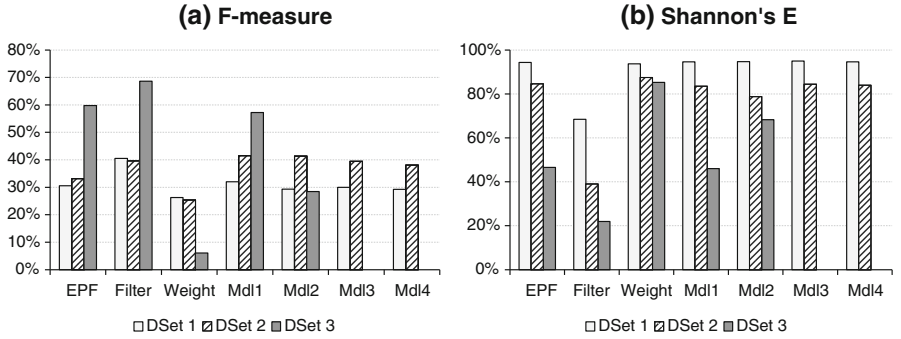

Fig. 4 Effect of type of data (DSet1 to DSet3) on recommendations accuracy (**a**) and diversity (**b**) of various CARS methods

granularity only affect accuracy while not diversity: when context becomes finer, the accuracy of CARS decreases. The type of data also affects both accuracy and diversity showing an interesting trade-off: if the heterogeneity of customers' behavior increases, the accuracy of a CARS decreases while its diversity increases. The regional analysis in the next section will focus on the direct comparison of the CARS methods across the regions of the factor space.

4.2 Regional analysis of accuracy and diversity of various CARS methods

In the previous section we summarized the results of a "marginal" analysis of our experiments in which we analyzed the effects of each factor on CARS performance at a time. In this section, we analyze the effects of all the three factors on the performance of CARS methods *simultaneously*, this three factors being (a) the recommendation task (defined with values Find-all vs. Top-k recommendations), (b) context granularity (coarse vs. fine granularity of contextual variables) and (c) type of data, i.e., datasets DSet 1, DSet 2 and DSet 3 characterized by the combination of different levels of data sparsity and heterogeneity of the users. Collectively, these factors form the 3-dimensional *factor space* consisting of various *regions* (hence the name "regional" analysis). Since in one of the three datasets (DSet 3, characterized by high sparsity and low heterogeneity) the context hierarchy has only one level (see Fig. 1c), the number of the regions in the overall factor space is only ten (and not 12, as it should have been in the completely orthogonal case). Note that this analysis presents only the high-level "picture" capturing the aggregate results of our experiments. As pointed out in Sect. 3.3, we used 3 different data sets, 7 contextual approaches, 6 different performance measures, 2 recommendation strategies, 4 values of k for the "top-k" strategy, and 3 different contextual variables in our experiments, thus generating 3,780 experiments in total. The figures and tables from this section constitute the summaries of these experiments.

Our regional analysis is structured in three parts. First, we identify which CARS method(s) dominates the others in terms of recommendation accuracy in a statistically significant manner and provide an explanation of these results. Second, we identify

which CARS method significantly dominates the others in terms of the diversity of recommendations and provide an explanation of this behavior. Third, we combine the accuracy and the diversity measures to identify which CARS approach(es) provides the best performance for a combination of these two measures. In this study, we combine accuracy and diversity by (a) averaging the standardized measures, (b) combining the ordinal ranking among the approaches and (c) analyzing the Pareto frontier of the two measures in each region. The first two options are alternative methods to evaluate the combination of accuracy and diversity quantitatively (Miller and Salkind 2003; Bracalente et al. 2009). The Pareto frontier is a very old concept in economics representing the optimal combination of two factors (Pareto 1896). Database researchers introduced a new concept of "skyline queries" recently that resembles the Pareto frontier to a large extent (Sharifzadeh and Shahabi 2006).

We start with comparing different CARS methods in terms of predictive accuracy as determined by the F-, Precision and Recall measures. The results of the comparison in terms of the F-measure and Precision are presented in Tables 2 and 3 and show that the best-performing CARS methods are shared by the Filter PoF and one of the Contextual Modeling approaches. The results are concordant except for three regions where the most precise approach is not the one with the highest F-measure (see the first column and the last two in Tables 2, 3). In contrast, there is no clear winner emerging from such comparison in terms of the Recall measure (Table 4). In particular, the Weight PoF method provides the highest Recall in three regions, EPF in two regions, Filter PoF in one region, and one of the Contextual Modeling approaches in the remaining ones.

Figure 5 is a graphical representation of the same results, as in Table 3, where we only present the CARS approach which dominates the others in each region in terms of the F-measure. The three axes in Fig. 5 represent the three factors used in our experiments: recommendation task, context granularity, and type of data. Furthermore, we have divided this 3-dimensional factor space into 10 regions, each region being the combination of the three factors (2 recommendation tasks, 2 context granularities and 3 types of data; note that, since context granularity for one of the factors is defined only by a single level, this means that we have only 10 regions in total and not 12, as the full combination of these three factors would produce). For instance, the region at the axes' origin corresponds to the combination of "Find-all" task, "Coarse" context granularity, and a type of data "DSet 1". Each region of the factor space in Fig. 5 has its cube that specifies the best performing CARS method for that region in terms of the F-measure (note that the "Mdl" label in Fig. 5 indicates the "best-of-breed" method among the four Contextual Modeling approaches). For instance, in the region at the axes' origin the best performing methods are "Filter" and "Mdl", where Mdl refers to the best-of-breed method among the CM approaches. We use the F-measure here because it represents the harmonic mean of the Precision and Recall measures, and because using the F-measure in such cases is a common practice in the data mining and the recommender systems communities. We also checked the statistical significance of the difference between the average accuracy of the dominant and the second-best approaches using the t test. The cubes with diagonal stripes in Fig. 5 represent the cases in which the difference of the means between the dominant and the second best approaches is not statistically significant ($p > 0.05$). All the other

Table 2 F-measure of CARS methods for the ten regions of the factor space

Regions: CARS methods	DSet 1 Find-all Coarse (%)	DSet 1 Find-all Fine (%)	DSet 1 Top-k Coarse (%)	DSet 1 Top-k Fine (%)	DSet 2 Find-all Coarse (%)	DSet 2 Find-all Fine (%)	DSet 2 Top-k Coarse (%)	DSet 2 Top-k Fine (%)	DSet 3 Find-all Coarse (%)	DSet 3 Top-k Coarse (%)
EPF	42.26	26.17	42.23	24.93	33.56	30.81	34.39	30.77	41.46	40.15
Filter PoF	47.47	36.82	40.74	35.15	39.49	33.54	39.47	33.54	54.10	54.06
Weight PoF	39.92	21.91	35.95	21.38	28.14	22.91	27.40	22.58	6.41	7.36
Mdl1	47.92	26.27	44.97	25.41	39.32	34.40	44.14	36.05	40.99	40.99
Mdl2	40.80	25.33	38.99	24.52	45.36	30.90	45.45	30.90	50.29	50.49
Mdl3	42.27	26.27	42.53	25.17	38.00	36.34	37.93	36.14		
Mdl4	44.08	25.12	41.69	24.96	42.64	35.95	42.71	35.87		

Table 3 Precision of CARS methods for the ten regions of the factor space

Regions: CARS methods	DSet 1 Find-all Coarse (%)	DSet 1 Find-all Fine (%)	DSet 1 Top-k Coarse (%)	DSet 1 Top-k Fine (%)	DSet 2 Find-all Coarse (%)	DSet 2 Find-all Fine (%)	DSet 2 Top-k Coarse (%)	DSet 2 Top-k Fine (%)	DSet 3 Find-all Coarse %	DSet 3 Top-k Coarse %
EPF	30.98	18.55	47.41	21.79	28.53	27.59	29.43	27.59	33.70	34.47
Filter PoF	38.68	34.09	45.88	34.81	39.53	33.00	39.53	33.00	51.73	51.87
Weight PoF	29.14	14.08	40.54	18.75	20.39	15.91	21.13	16.51	3.45	5.96
Mdl1	30.85	20.20	50.44	22.51	42.84	36.04	51.51	39.41	32.93	32.93
Mdl2	29.08	17.83	43.62	21.50	52.13	27.62	52.14	27.62	53.32	53.71
Mdl3	30.85	20.20	47.59	22.49	36.49	39.63	36.48	39.61		
Mdl4	34.58	17.49	46.84	21.90	44.47	35.11	44.47	35.10		

Table 4 Recalls of the CARS methods for the ten regions of the factor space

Regions: CARS methods	DSet 1 Find-all Coarse (%)	DSet 1 Find-all Fine (%)	DSet 1 Top-k Coarse (%)	DSet 1 Top-k Fine (%)	DSet 2 Find-all Coarse (%)	DSet 2 Find-all Fine (%)	DSet 2 Top-k Coarse (%)	DSet 2 Top-k Fine (%)	DSet 3 Find-all Coarse (%)	DSet 3 Top-k Coarse (%)
EPF	74.54	55.45	40.57	34.96	51.71	41.78	51.03	41.51	75.71	57.31
Filter PoF	68.80	49.64	38.98	42.32	48.60	40.62	48.53	40.62	65.96	65.41
Weight PoF	71.90	66.16	34.45	29.84	57.90	51.06	46.89	42.25	74.33	13.06
Mdl1	74.24	50.37	43.30	35.26	43.31	40.23	43.95	37.34	67.63	67.63
Mdl2	75.96	56.72	37.64	34.40	45.89	41.03	45.86	41.03	60.62	58.63
Mdl3	74.24	50.37	40.95	34.91	47.65	38.19	47.37	37.30		
Mdl4	69.53	59.92	40.07	34.99	49.81	43.83	49.75	43.77		

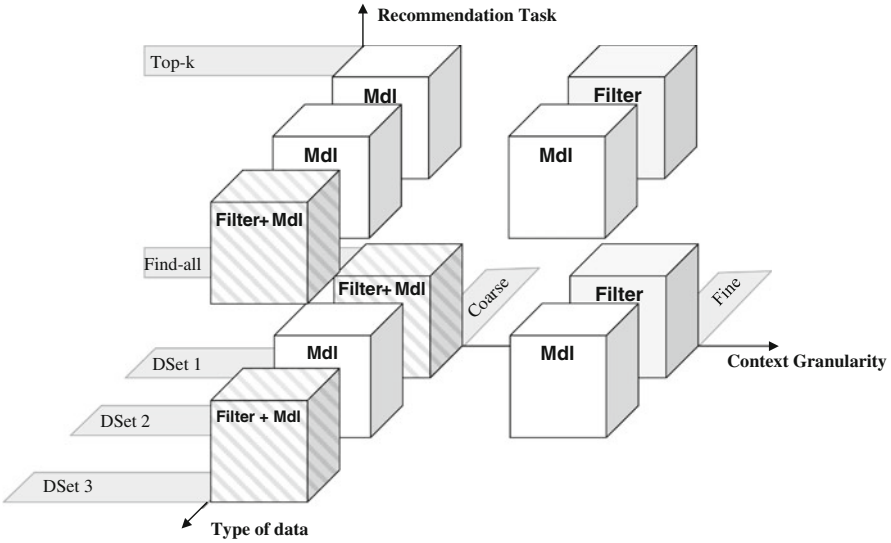

Fig. 5 Which CARS approach dominates the others in terms of the recommendations accuracy (as defined by the F-measure)

cases are statistically significant with $p < 0.001$. In the regions where the t tests are not significant, the differences of the means between the second best approach and each one of the remaining CARS methods *are* statistically significant. For instance, in the region at the axes' origin, the difference of the means between "Filter" and "Mdl" is not statistically significant. However, the difference between these two approaches and the remaining ones is statistically significant. Therefore we can state that both Filter and Mdl dominate the others, but not that Filter and Mdl are different.

As commented above, in all the regions the most accurate CARS approach is either Filter PoF or one of the Contextual Modeling approaches. Figure 5 shows that the Filter PoF approach dominates the others in terms of accuracy when context is "Fine" and the type of data is characterized by low sparsity and high heterogeneity (DSet 1), regardless of the recommendation task. The Filter PoF approach also dominates (together with Mdl) when the type of data is characterized by high sparsity and low heterogeneity (Dset 3). The reason of this results is that in the regions where context is finer and in those where the type of data is characterized by high sparsity (DSet 3) the prediction problem is made harder due to the lack of information.

This is consistent with the marginal analysis (see Sect. 4.1 and Panniello et al. 2012). The Filter PoF exploits all the information available by generating the recommendations via the uncontextual *User* $\times$ *Item* matrix and its accuracy may turn out to be higher than that of other CARS in these regions. When the type of data is characterized by medium levels of sparsity and heterogeneity (DSet 2) the best Contextual Modeling (CM) approach always dominates. However, the best CM approach is not always the same. As shown in Table 2 there may be differences among the four Mdl approaches. Mdl1 is the most accurate CM approach in the region corresponding to DSet 1, "Find-all", "Coarse" (where it shares the dominant position with Filter PoF)

and in the region corresponding to DSet 1, "Top-k", "Coarse". Mdl2 is the most accurate CM approach in the regions defined by DSet 2, "Find-all", "Coarse" and DSet 2, "Top-k", "Coarse". Mdl3 is the most accurate CM approach in the regions defined by DSet 2, "Find-all", "Fine" and DSet 2, "Top-k", "Fine". Finally, Mdl4 is never the most accurate approach among those in the CM category.

Table 5 reports similar results for the diversity measure. We computed the average diversity in each region for each of the CARS methods across all the users and the three measures of diversity.

Figure 6 is the graphical representation of these results from Table 5, where we only report the CARS approach which dominates the others in each region in terms of diversity. As in Fig. 5, the three axes represent the three factors used in our experiments and each cube represents the best performing CARS methods when the specific combination of factors is considered. The cubes with diagonal stripes represent the cases when the t test between the means of the dominant approach and that of the second best is not significant. All other cases in Fig. 6 identify the methods exhibiting statistically significant differences between the best- and the second-best-performing methods (with $p < 0.001$). In all the three non-significant cases (cubes with stripes in Fig. 6), the difference of the means between the second best approach and each one of the *remaining* CARS methods is significant. This means that although we cannot state that one approach significantly dominates the others in those three regions, we can still state that the two best approaches indeed *statistically* dominate the remaining ones.

Figure 6 clearly shows that the Weight PoF approach is the one generating the most diverse recommendations (where diversity is measured as an average among the three metrics presented in Sect. 3.2) in all the regions defined by DSet 2 and DSet 3. In the regions defined by datasets DSet 1, i.e. when the type of data is characterized by low sparsity and high heterogeneity, Weight PoF dominates only when the recommendation task is "Top-k" and the context is "Coarse" (together with a Contextual Modeling approach). In the remaining regions Contextual Modeling and/or EPF dominate. The reason is that when customers' behavior is heterogeneous (i.e., in DSet 1) and the quantity of information is high (sparsity is low in DSet 1) all the CARS are able to generate diverse recommendations except Filter PoF which only exploits context to filter out recommendations. When heterogeneity decreases and sparsity increases (i.e., moving to DSet 3) increasing diversity becomes a harder problem, and the best performing CARS is Weight PoF which exploits all the information available to generate recommendations (via the uncontextual $User \times Item$ matrix) but does not filter out those irrelevant to the context, rather places them at the bottom of the list. This interpretation is confirmed by the marginal analysis (see Sect. 4.1). In particular, in the regions defined by DSet 1, "Find-all", "Coarse" Mdl4 is the Contextual Modeling approach providing the highest diversity, as well as in the region defined by DSet 1, "Top-k", "Coarse". In the region defined by DSet 1, "Find-all", "Fine" Mdl1 is the Contextual Modeling approach providing the highest diversity.

After identifying the regions of the 3-dimensional space in which each approach dominates the others for accuracy and diversity individually, it is important to combine accuracy and diversity measures and to compare the CARS methods in terms of a single combined performance measure for each region. The problem of combining

Table 5 Average diversity of the CARS methods for the ten regions of the factor space

Regions: CARS methods	DSet 1 Find-all Coarse (%)	DSet 1 Find-all Fine (%)	DSet 1 Top-k Coarse (%)	DSet 1 Top-k Fine (%)	DSet 2 Find-all Coarse (%)	DSet 2 Find-all Fine (%)	DSet 2 Top-k Coarse (%)	DSet 2 Top-k Fine (%)	DSet 3 Find-all Coarse (%)	DSet 3 Top-k Coarse (%)
EPF	96.07	95.86	94.24	94.98	83.98	81.87	91.74	89.92	69.64	85.23
Filter PoF	87.88	69.22	94.12	76.85	52.49	45.05	65.58	55.20	47.73	64.83
Weight PoF	95.58	95.25	94.72	86.45	85.20	85.42	93.45	93.79	90.11	96.59
Mdl1	95.03	96.62	92.85	87.96	81.52	81.86	88.84	89.77	64.67	83.65
Mdl2	95.28	96.32	94.32	87.74	76.33	76.34	89.34	89.99	74.51	92.82
Mdl3	96.11	96.53	94.73	87.93	84.11	81.84	91.11	89.70		
Mdl4	96.12	96.18	94.83	87.37	82.81	81.67	90.37	91.51		

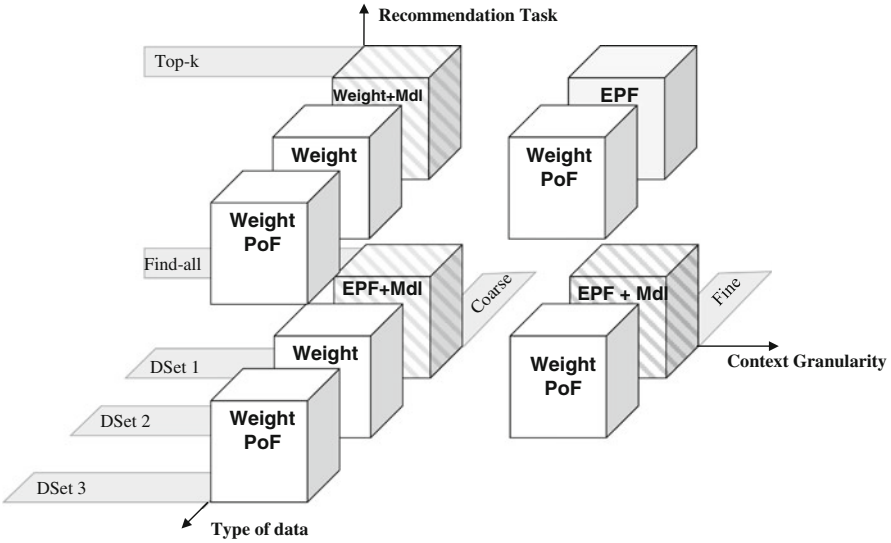

Fig. 6 Which CARS approach dominates the others in terms of the recommendations diversity (specified as the average of Simpson, Shannon and Tidemann & Hall diversity indexes)

the two measures is not straightforward, however. In fact, accuracy is measured by the F-measure while diversity by the average of three different measures D, E and TH. Although both measures are calculated as percentage values ranging from 0 to 1, they cannot be simply averaged because they represent very different performance metrics and they have very different scales. Finally, the relative importance of accuracy and diversity depends on several factors, including the domain, the business application and the specific goals of the company using the recommender system. Therefore, the problem is not only in combining accuracy and diversity into a single concise index and studying which CARS dominates the others but also studying which CARS achieves the best balance between accuracy and diversity in certain conditions. In order to investigate this problem, and according to the literature, we adopt three strategies. The first is to consider the two metrics as numerical variables expressed in an equal interval ratio scale (Miller and Salkind 2003; Bracalente et al. 2009). Since accuracy and diversity are percentages, they qualify for this type of measure. In this case, the only method needed to combine the variables is to make the scales homogeneous by standardizing the metrics and computing the average (since they are not standardized in their initial format it would not be possible to simply average the different measures). The results of combining the two measures according to this method are plotted in Fig. 7.

As Fig. 7 shows, the dominant CARS approach is the CM. In fact, it is the dominant approach in 9 out of 10 regions, while the EPF outperforms all the other approaches only in the region identified by DSet 1, "Fine" and "Top-k". In particular, the Mdl2 is the best CM approach when the type of data is DSet 3 regardless of the recommendation task used. The Mdl1 is the best contextual modeling approach in all the regions corresponding to DSet 1, "Find-all" regardless of the context granularity. In

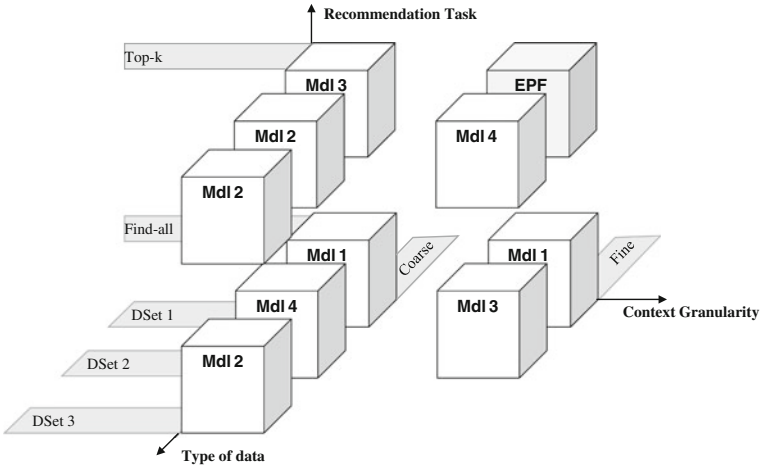

Fig. 7 Which CARS approach dominates the others in terms of average of standardized accuracy and diversity

one region (DSet 1, "Find-all", "Fine") the EPF approach provides the best combined performance, while in the regions defined by DSet 2, the best performing approaches are Mdl2, Mdl3 and Mdl4.

The second strategy is based on the hypothesis that the two measures reflect two different properties which cannot be combined in a single index. In this case we can still use the ordinal rankings among the CARS approaches (Miller and Salkind 2003; Bracalente et al. 2009), the first based on the comparison of accuracy and the second on diversity. The rankings can be simply combined by calculating which approach is placed in the best position in both the rankings. Moreover, the Goodman and Kruskal's Gamma index (Bracalente et al. 2009) can be computed to compare the ranking. In general, a Gamma index close to 1 means that the two rankings are very similar, while a value close to -1 means the rankings are opposite one another. Figure 8 reports these results.

In most regions, the Gamma index has a negative value. This confirms the fact that the most accurate CARS approach tends to be one of the worst in terms of diversity. Therefore, maximizing both accuracy and diversity is normally impossible, while it is possible to identify a good compromise between the two performance measures. This observation will be confirmed by the analysis of the Pareto frontier. Again, the best balance is provided by the CM approaches, although there are differences among them.

The third strategy consists of analyzing the Pareto frontier in each region, therefore identifying the dominating approaches (those on the frontiers) and excluding the others. We plotted the CARS approaches in the graphs presented in Figs. 9, 10 and 11 where the accuracy measure is plotted on the x-axis and the diversity on the y-axis for each one of the 10 regions of the 3-dimensional factor space.

Figure 9 reports the plots for the four graphs for the plan identified by DSet 1 dataset, where sparsity is low and heterogeneity high. In this plan the Pareto frontier

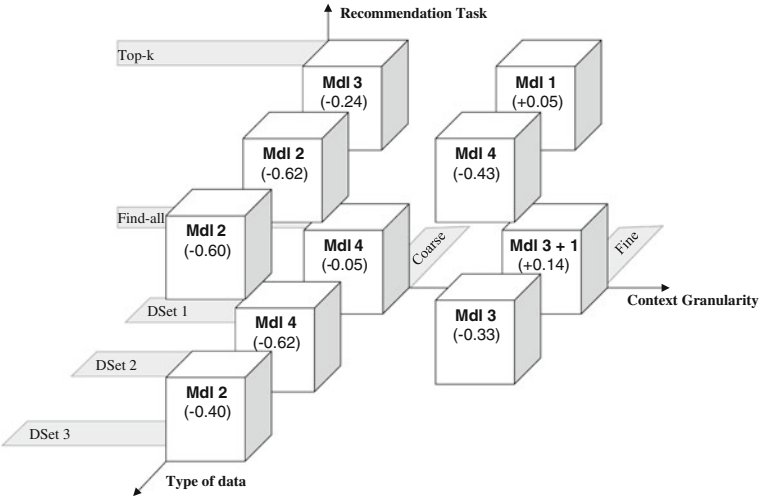

Fig. 8 Which CARS approach dominates the others in terms of combined ordinal ranking of accuracy and diversity (numbers in brackets are the Goodman and Kruskal's Gamma indexes)

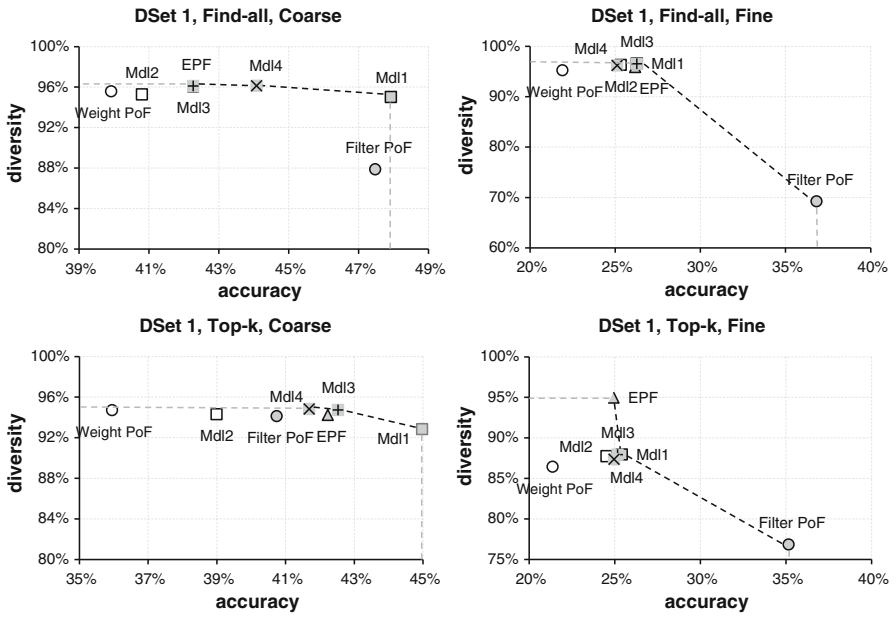

Fig. 9 Pareto frontiers in the four regions of the plan defined by DSet 1

always includes at least three CM approaches. The Filter PoF is on the frontier in two regions, as well as the EPF, while the Weight PoF is never on the frontier. The Filter PoF approach is placed in the right-bottom side of the diagram, meaning that in a multi-criteria decision-making problem it would the best approach if the weight of

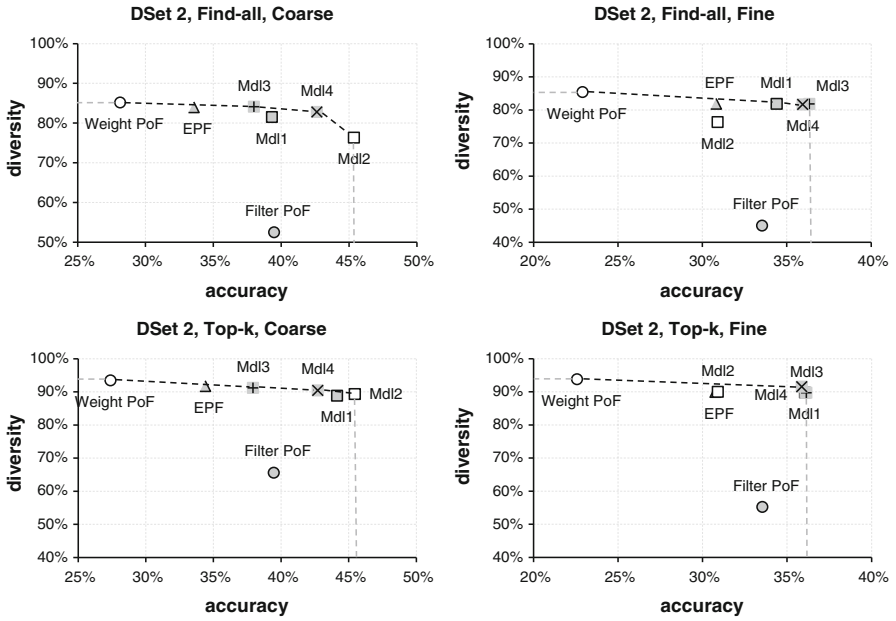

Fig. 10 Pareto frontiers in the four regions of the plan defined by DSet 2

accuracy is much higher than that of diversity. Among the CM approaches, Mdl2 is the one which is not on the frontier in three regions.

Figure 10 presents the plots for the four regions in the vertical plan defined by DSet 2. Also in this case at least three CM approaches are on the frontier. The Filter PoF is never on the frontier, while the EPF is very close to the frontier. Weight PoF is always on the frontier, in the upper-left part of the diagram, meaning that it should be used if the weight of diversity is much higher than that of accuracy.

Figure 11 presents the plots for the two regions in the vertical plan defined by DSet 3. In this case Mdl2 is on the frontier, while Mdl1 and EPF are not. Filter PoF and Weight PoF are on the frontier, at the extreme of it. Again, Filter PoF would be considered the best if the weight of accuracy is much higher than that of diversity in a multi-criteria decision-making problem. Vice-versa for Weight PoF.

As the graphs in Fig. 9 through Fig. 11 show, the Contextual Modeling (CM) approaches are the only ones appearing in each one of the ten regions. This is consistent with the results depicted in Figs. 7 and 8 which show that in almost all the regions the CM approaches are those providing the best combination of accuracy and diversity, considering both the average of standardized measures and only the ordinal ranking. The only region showing a little inconsistency is that defined by DSet 1, "Top-k", "Fine", in which EPF would prevail over the CM if the average between standardized accuracy and diversity is used. The reason is that this is the only region in which EPF provides the highest diversity. For this reason, we can state that, *in general, the CM approaches are those which provide the best balance between accuracy and diversity*. However, as the plots of the Pareto frontiers show, there may be differences among the

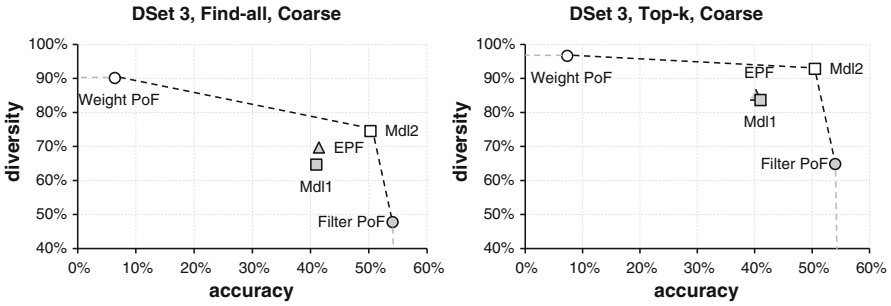

Fig. 11 Pareto frontier in the two regions of the plan defined by DSet 3

four CM approaches. Looking at the plots, Mdl1 should be preferred to any other CM approach when the type of data is similar to the cases of DSet 1 and DSet 2. Except one case, Mdl1 provides the highest accuracy while the difference in diversity is minor. When the type of data is similar to DSet 3, i.e., sparsity is around 98 % but the users' behavior is quite homogeneous, Mdl2 dominates Mdl1. The EPF approach is also always very close to the frontier, except in the two regions of DSet 3. Considering the EPF is probably the less complex CARS methods, from a practical viewpoint using this approach when the type of data is similar to DSet 1 and DSet 2 is reasonable. The result is confirmed by the fact that EPF is the only non-CM approach appearing in Fig. 7. A different comment has to be done for the post-filtering approaches. They should *not* be used if the goal is achieving a good balance between accuracy and diversity because they always are at the extreme of the Pareto frontier. Filter PoF is often the most accurate approach but its diversity is significantly (in a statistical way) lower than that of Weight PoF. On the contrary, Weight PoF provides high diversity but poor accuracy. This analysis is confirmed by Figs. 7 and 8.

5 Conclusions

In this paper, we compared the performance of various pre-filtering, post-filtering and contextual modeling CARS methods in terms of their predictive performance and diversity measures across various experimental conditions to determine which method dominates the others and under which circumstances. We have identified three key factors affecting performance of the CARS methods, including the type of the recommendation task (Find-All vs. Top-k), context granularity (coarse vs. fine granularity of the contextual information) and the type of the data set (DSet 1 characterized by low sparsity and high heterogeneity, DSet 3 characterized by high sparsity and low heterogeneity, DSet 2 with medium levels of sparsity and heterogeneity). Then we have compared the performance of different CARS methods using the marginal and regional analysis techniques. Using the marginal analysis, we have examined how each of the three factors *separately* affects the performance of the CARS methods and concluded that, for our experimental settings, the recommendation task affected both accuracy and diversity in a way which depends on the specific performance measure considered. Context granularity only affected accuracy in our experiments: when

context became finer, the accuracy of CARS decreased. The type of data used in our experiments affected accuracy and diversity showing a trade-off: if the heterogeneity of customers' behavior increased, the accuracy of CARS decreased while diversity increased. Most of the performance differences in our experiments were statistically significant, as explained in the paper.

Using the regional analysis, we have examined which of the CARS methods dominated the others in each of the regions of the 3-dimensional factor space defined by the recommendation task, context granularity and the data type. It turned out that none of the CARS methods uniformly dominated the others in all the regions for both the recommendation accuracy and diversity measures in our experiments. However, the Mdl and the Filter PoF methods statistically outperformed other CARS alternatives in terms of the accuracy measure across all of the factor space. Similarly, Weight PoF and the EPF methods statistically outperformed the other CARS methods in terms of the diversity measure across most of the 10 regions in our experiments. Finally, the Mdl class of methods outperformed the rest of the CARS methods in terms of the combination of the accuracy and the diversity measures.

Based on this analysis, the Mdl-oriented contextual modeling methods constitute a reliable "best bet" when choosing a sound CARS approach because these methods provide a nice performance balance in terms of accuracy *and* diversity measures. However, even such good CARS methods as Mdl do not dominate all other techniques across all the experimental settings, and other methods, such as Filter PoF and even Weight PoF, also constitute viable alternatives for certain regions of the factor space, certain experimental settings and specific performance measures.

This work constitutes the first step towards an ambitious goal of understanding the differences among the three main approaches to CARS. Therefore, much more work is required in order to achieve this goal. For example, as a future work, it would be interesting to study how "hybrid" combinations of pre-, post-filtering and contextual modeling approaches work and compare with the "pure" non-combined methods. It is also important to understand *why* some specific individual methods (such as the Mdl1 in the contextual modeling approach) outperform the others (such as Mdl2, Mdl3 and Mdl4) using theoretical analysis in conjunction with the empirical studies described in this paper. In addition, a comparison of content-based CARS approaches also needs to be done, as opposed to the collaborative-filtering-based methods presented in this paper. Another interesting point for future research could be to investigate how to use the changes in customer behavior based on the long and short term preferences in order to improve recommendations. Finally, it would also be interesting to compare different CARS approaches for other types of measurements, such as serendipity and novelty.

References

Adomavicius, G., Tuzhilin, A.: Extending recommender systems: a multidimensional approach, pp. 4–6. IJCAI workshop on intelligent techniques for web personalization, Seattle (2001)

Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans. Inf. Syst. **23**, 103–145 (2005)

Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems'. ACM RecSys Tutorial, pp. 335–336 (2008)

Adomavicius, G., Kwon, Y.: Toward more diverse recommendations: Item re-ranking methods for recommender systems. 19th Workshop on information technologies and systems (WITS), Phoenix

Adomavicius, G., Mobasher, B., Ricci, F., Tuzhilin, A.: Context-aware recommender systems. AI Mag. **32**(3), 67–80 (2011)

Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Handbook on Recommender Systems, pp. 217–253. Springer (2011)

Adomavicius, G., Kwon, Y.: Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans. Knowl. Data Eng. **24**(5), 896–911 (2012)

Bader, R., Neufeld, E., Woerndl, W., V. Prinz: Context-aware POI recommendations in an automotive scenario using multi-criteria decision making methods. Workshop on context-awareness in retrieval and recommendation, pp. 23–30. Chicago (2011)

Baltrunas, L., Amatriain, X.: Towards time-dependant recommendation based on implicit feedback. First workshop on context-aware recommender systems, pp. 1–5. New York (2009)

Baltrunas, L., Ludwig, B., Ricci, F.: Matrix factorization techniques for context aware recommendation. Fifth ACM conference on recommender systems, pp. 301–304. Chicago (2011)

Baltrunas, L., Ricci, F.: Experimental evaluation of context-dependent collaborative filtering using item splitting. User Modeling and User-Adapted Interaction, Special issue on Context-Aware Recommender Systems (this issue) (2014)

Billsus, D., Pazzani, M.: User modeling for adaptive news access. User Model. User Adapt. Interact. **10**(2–3), 147–180 (2000)

Bracalente, B., Cossignani, M., Mulas, A.: Statistica aziendale. McGraw-Hill Italia, Milano (2009)

Campos, P.G., F. Díez and I. Cantador, : Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. User Modeling and User-Adapted Interaction, Special issue on Context-Aware Recommender Systems (this issue) (2014)

Cena, F., Console, L., Gena, C., Goy, A., Levi, G., Modeo, S., Torre, I.: Integrating heterogeneous adaptation techniques to build a flexible and usable mobile tourist guide. AI Commun. **19**(4), 369–384 (2006)

Cremonesi, P., Garza, P., Quintarelli, E., Turrin, R.: Top-N Recommendations on Unpopular Items with Contextual Knowledge, 2011 Workshop on Context-aware Recommender Systems. Chicago (2011)

De, A., Desarkar, M.S., Ganguly, N., Mitra, P.: Local learning of item dissimilarity using content and link structure. Sixth ACM conference on recommender systems, pp. 221–224. Dublin (2012)

Dourish, P.: What we talk about when we talk about context. Pers. Ubiquitous Comput. **8**, 19–30 (2004)

Fagà, R., Furtado, B.C., Maximino, F., Cattelan, R.G. da Pimentel, M.G.C.: Context information exchange and sharing in a peer-to-peer community: a video annotation scenario. 27th ACM international conference on design of communication SIGDOC2009, pp. 265–272. Bloomington (2009)

Feng, Q., Liu, L., Dai, Y.: Vulnerabilities and countermeasures in context-aware social rating services. Trans. Internet Technol. **11**(3), 1–27 (2012)

Ge, Y., Liu, Q., Xiong, H., Tuzhilin, A., Chen, J.: Cost-aware travel tour recommendation. 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 983–991. San Diego (2011)

Gorgoglione, M., Panniello, U.: Including context in a transactional recommender system using a pre-filtering approach: two real E-commerce applications. The 23rd IEEE international conference on advanced information networking and applications (AINA-09), pp. 667–672. Bradford (2009)

Gorgoglione, M., Panniello, U., Tuzhilin, A.: The effect of context-aware recommendations on customer purchasing behavior and trust. Fifth ACM conference on recommender systems, pp. 85–92. Chicago (2011)

Hariri, N., Mobasher, B., Burke, R.: Context-aware music recommendation based on latenttopic sequential patterns. Sixth ACM conference on recommender systems, pp. 131–138. Dublin (2012)

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)

Hu, R., Pu, P.: Helping users perceive recommendation diversity. Workshop on novelty and diversity in recommender systems, DiveRS. Chicago (2011)

Hussein, T., Linder, T., Gaulke, W., Ziegler, J.: Hybreed: a software framework for developing context-aware hybrid recommender systems. User Modeling and User-Adapted Interaction, Special issue on Context-Aware Recommender Systems (this issue) (2014)

Kaminskas, M., Ricci, F.: Location-adapted music recommendation using tags. 19th International conference on user modeling, adaption, and personalization, pp. 183–194. Girona (2011)

Levi, A., Mokryn, O., Diot, C., Taft, N.: Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. Sixth ACM conference on recommender systems, pp. 115–122. Dublin (2012)
Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)
Mahmood, T., Ricci, F., Venturini, A.: Improving recommendation effectiveness by adapting the dialogue strategy in online travel planning. Int. J. Inf. Technol. Tour. **11**(4), 285–302 (2010)
McDonald, D., Dimmick, J.: The conceptualization and measurement of diversity. Commun. Res. **30**(1), 60–79 (2003)
Mcginty, L., Smyth, B.: On the role of diversity in conversational recommender systems. Fifth international conference on case-based reasoning, pp. 276–290
Meilak, C.: Measuring export concentration: the implications for small states. Bank Valletta Rev. **37**, 35–48 (2008)
Miller, D.C., Salkind, N.J.: Handbook of Research Design and Social Measurement. Sage Publications, London (2003)
Oku, K., Nakajima, S., Miyazaki, J., Uemura, S.: Context-aware SVM for context-dependent information recommendation. 7th International conference on mobile data management, p. 109 (2006)
Palmisano, C., Tuzhilin, A., Gorgoglione, M.: Using context to improve predictive modeling of customers in personalization applications. IEEE Trans. Knowl. Data Eng. **20**(11), 1535–1549 (2008)
Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano, C., Pedone, A.: Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems. Third ACM conference on recommender systems, pp. 265–268. New York (2009)
Panniello, U., Gorgoglione, M.: Incorporating context into recommender systems: an empirical comparison of context-based approaches. Electron. Commer. Res. **12**(1), 1–30 (2012)
Panniello, U., Tuzhilin, A., Gorgoglione, M.: Comparing context-aware recommender systems in terms of accuracy and diversity: which contextual modeling, pre-filtering and post-filtering methods perform the best. Working paper CBA-12-01. Stern School of Business, NYU (2012)
Pareto V.: Cours d'économie politique professé l'université de Lausanne, 3 volumes. Lousanne (1896)
Reddy, S., Mascia, J.: Lifetrak: music in tune with your life. 1st ACM international workshop on human-centered multimedia, pp. 25–34. Santa Barbara (2006)
Ricci, F.: Mobile recommender systems. Int. J. Inf. Technol. Tour. **12**(3), 205–231 (2011)
Sae-Ueng, S., Pinyapong, S., Ogino, A., Kato, T.: Personalized shopping assistance service at ubiquitous shop space. 22nd International conference on advanced information networking and applications, pp. 838–843. Los Alamitos (2008)
Said, A., Berkovsky, S., De Luca, E. W.: Group recommendation in context. 2nd Challenge on context-aware movie recommendation, pp. 2–4. Chicago (2011)
Sharifzadeh, M., Shahabi, C.: The spatial skyline queries. VLDB '06 Proceedings of the 32nd international conference on very large data bases, pp. 751–762. Seoul (2006)
Wang, S.L., Wu, C.T.: Application of context-aware and personalized recommendation to implement an adaptive ubiquitous learning system. Expert Syst. Appl. **38**(9), 10831–10838 (2011)
Zhang, M., Hurley, N.: Avoiding monotony: improving the diversity of recommendation lists. Second ACM conference on recommender systems. Lausanne (2008)
Ziegler, C.N., McNee, S., Konstan, J. Lausen, G. Improving recommendation lists through topic diversification. 14th International conference on world wide web. Chiba (2005)

Author Biographies

Umberto Panniello is a post-doc and lecturer in Management Engineering at the Politecnico di Bari (Polytechnic University of Bari, Italy). He completed his Ph.D. degrees in 2011 at the Politecnico di Bari in the area of marketing and data mining. His primary interests lie in the areas of recommender systems and data mining techniques in business applications. His contribution is based on experiences gained during his Ph.D. work as well as his current research projects.

Alexander Tuzhilin is a Professor of Information Systems and the NEC Faculty Fellow at the Stern School of Business, NYU. He has received Ph.D. in Computer Science from the Courant Institute of Mathematical Sciences, NYU. His current research interests include data mining, recommender systems and

personalization. Dr. Tuzhilin has produced over 100 research publications on these and other topics. He has served on the organizing and program committees of numerous CS and IS conferences and as the Chair of the Steering Committee of the RecSys Conference. He has also served on the Editorial Boards of the IEEE Transactions on Knowledge and Data Engineering, the ACM Transactions on Management Information Systems, the INFORMS Journal on Computing (as an Area Editor), the Data Mining and Knowledge Discovery Journal, the Electronic Commerce Research Journal and the Journal of the Association of Information Systems.

Michele Gorgoglione is Associate Professor of Marketing and e-Business Models at the Politecnico di Bari (Polytechnic University of Bari, Italy). Dr. Gorgoglione received his Laurea degree in Electronic Engineering from the Politecnico di Bari and Ph.D. degree in Management from the Università of Tor Vergata - Roma (Italy). Dr. Gorgoglione has worked in the areas of artificial intelligence, knowledge management and personalization technologies and business models. He is author of over fifty international scientific publications.