Effectiveness of different recommender algorithms in the Mobile Internet: A case study

Kolja Hegelich and Dietmar Jannach
Technische Universität Dortmund
44221 Dortmund, Germany
kolja.hegelich@tu-dortmund.de, dietmar.jannach@tu-dortmund.de

Abstract

Despite the broad use of Recommender Systems (RS) technology in various domains, the number of publicly available reports on the actual business value of such systems is limited.

This paper presents first results of an empirical evaluation of how different recommendation algorithms affect the navigation and buying behavior of a sample of over 155.000 different customers on a commercial Mobile Internet portal for cell phone games. The evaluated RS algorithms include item-based collaborative filtering, SlopeOne, a content-based as well as a hybrid technique, which were compared with naive approaches based on top-selling and top-rated items.

The analysis shows that RS measurably affected the navigation and buying behavior of the portal visitors. The personalized recommendation lists not only attracted more clicks on detailed item descriptions but also lead to an overall sales increase when compared with control groups that received non-personalized recommendations or no recommendations during the evaluation period.

The comparison of different algorithms brought no clear winner that consistently outperformed the others. However, the results indicate that the choice of the recommendation technique should depend on the specific navigational situation in which recommendation lists are presented.

Introduction & previous studies

Although the interest in recommender systems technology has been increasing in the last years both in industry and research and although recommender applications can nowadays be found on many web sites of online retailers, nearly no studies about the actual business value of such systems have been published that are based on real-world transaction data.

In the research community, the performance of a recommender system is mainly measured based on its *accuracy* with respect to predicting whether a user will like a certain item or not[1]. The implicit assumption is that the online user – after establishing trust in the system's recommendations or because of curiosity – will more often buy these recommended items from the shop.

However, a shop owner's key performance indicators related to a personalized web application such as a recommender system are different ones. Establishing a trustful customer relationship, providing extra service to customers by proposing interesting items, maintaining good recommendation accuracy and so on are only a means to an end. While these aspects are undoubtedly important for the long-term success of a business, for an online retailer, the important performance indicators are related (a) to the increase of the *conversion rate*, i.e., how web site visitors can be turned into buyers, and (b) to questions of how to influence the visitors in a way that they buy more or more profitable items.

Unfortunately, only few real-world studies in that context are available because large online retailers do not publish their evaluations of the business value of recommender systems. Only a few exceptions exist. Dias et al. (Dias et al. 2008), for instance, recently presented the results of a 21-month evaluation of their probabilistic item-based recommender system running on a large Swiss e-grocer web portal. Their measures include "shopper penetration", "direct extra revenue" and "indirect extra revenue". Their analysis showed different interesting points. First, a relatively small (when compared to overall sales) extra revenue can be generated directly by the recommender. The fact that direct revenues measurably increased when the probabilistic model went through a periodic update suggests that good recommendation *accuracy* is still important, despite some legitimate criticism of simple accuracy measures (McNee, Riedl, and Konstan 2006). The more important business value, however, comes from *indirect* revenues caused by the recommender systems. Indirect revenues include the money spent on repeated purchases of items initially recommended by the system and on items sold from categories to which the customer was newly introduced to through a recommended item. This in turn also supports the theory that *diversity* in recommendation lists is a valuable property as "unexpected" items in these lists may help to direct users to other, possibly interesting categories.

An earlier evaluation based on real-world data was presented in (Shani, Brafman, and Heckerman 2002), where the authors performed different experiments on an online bookstore. During their experiment, visitors of the web shop re-

[1] See (Herlocker et al. 2004) for an overview on evaluation metrics for recommender systems.

ceived buying proposals either from a “predictive” or a new Markov Decision Process recommender. Thus, they were able to compare the respective *profits* that were generated by different techniques during the observation period. In addition, at least for a period of seven days, the recommendation functionality was fully removed from the web shop. Although this sample is statistically too small, the comparison of sales numbers of two consecutive weeks (one with and one without the recommender) showed a 17% drop in the recommender-less week.

Another initial study on how recommender systems influences the buying behavior of web shop visitors is presented in (Zanker et al. 2006). In this work, it was shown that the recommendations of a virtual advisor for premium cigars can stimulate visitors to buy cigars other than the well-known *Cohibas* and thus increase *sales diversity*, which is interesting from an up-selling and cross-selling perspective and could also create “indirect revenue” as described in (Dias et al. 2008); see also (Fleder and Hosanagar 2007) for a discussion of the role of sales diversity in recommender systems.

In (Zanker et al. 2008), a different study using the same recommendation technology was made in the tourism industry, where it could be observed that the number of accommodation availability enquiries is measurably higher when web site visitors are guided by the virtual advisor. Another evaluation of how different information types and recommendation “sources” influence consumers can be found in (Senecal and Nantel 2004).

Similar to these works, our paper focuses on evaluating the business value of recommender systems in a commercial context. In addition, it aims to answer the question whether certain algorithms perform better than others in a certain environment and application domain in the line of the work of, e.g., (Breese, Heckerman, and Kadie 1998) or (Zanker et al. 2007).

Application and personalization overview

The study presented in this paper was conducted in the context of a Mobile Internet portal of a large telecommunications provider in Germany. Customers access this portal through their mobile devices and are offered a wide range of applications and games, which they can directly purchase and download to their cell phones.

Figure 1 shows the entry screen of the games area of the portal. Customers explore the item catalog in the following ways:

- Through *manually-edited or non-personalized lists* such as “New items” or “Top10 items” (top area of screen).
- Through direct text or image links (teasers) to certain items that are shown on the middle area of the start screen.
- Through predefined *standard categories* (lower area) such as “A - Z”, “From 99 Cent”, or “Action & Shooters”.
- In addition, after a purchase, when the payment confirmation is displayed, customers are presented with a list of other, possibly interesting items (post-sales recommendation).

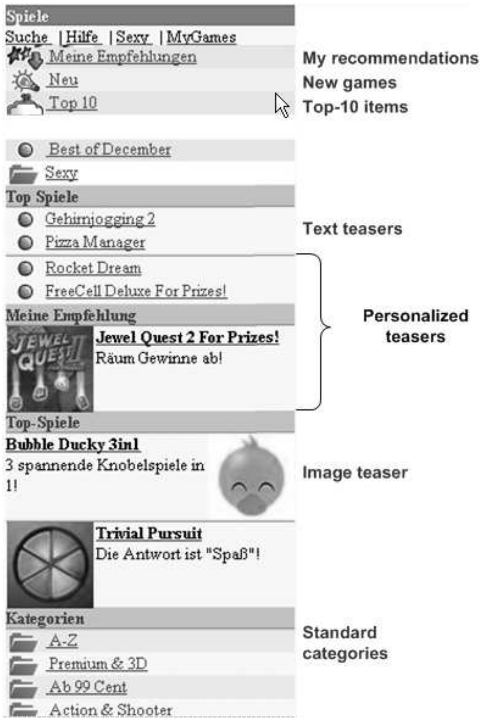

Figure 1: Catalog navigation and categories

Accordingly, the portal was extended with personalized content as follows.

1. A new top-level link “My Recommendations“ was introduced that leads to a personalized recommendation list. (“Meine Empfehlungen” in German).
2. The games presented in the lower two of the four text teasers and the first image teaser on the start page were personalized. Due to existing contracts the first two text links and the two lower image links were manually predefined. The manually edited links remained the same during the whole experiments, which allowed us to analyze the effects of personalizing the other links independently.
3. The lists in the standard categories such as “99 Cent” were personalized except for categories such as “A-Z”, which have a “natural” ordering.
4. The games presented on the post-sales page were also personalized.

During the experiments, different algorithms were used to calculate the personalized recommendations. In order to measure the effect of personalization, members of the control group were shown non-personalized or manually-edited lists which are based on the release date of the game.

Customers can immediately purchase and download games through the portal by choosing items from the presented lists. The relation between their navigation and buying behavior can therefore be easily determined as all portal visitors are always logged-in. Several thousand games (across all categories) are downloaded each day through the platform. The prices for the games range from free evaluation versions (demos) over “99Cent-Games” to a few Euro for premium games and the amounts are directly charged to the customer’s monthly invoice. Note that in contrast to the study in (Dias et al. 2008), where users purchase the same goods repeatedly, customers in our domain only purchase the same item once, i.e., our domain is similar to popular recommender systems application areas such as books and movies.

From the perspective of the application domain, the presented game portal stands in the line of previous works in the area of recommender systems for mobile users. Recent works in the field of mobile recommenders include, e.g., (Miller et al. 2003), (Cho, Kim, and Kim 2004), (van der Heijden, Kotsis, and Kronsteiner 2005), (Ricci and Nguyen 2007), (Li, Wang, and Geng 2008) or (Nguyen and Ricci 2008). Content personalization approaches for the Mobile Internet are presented also in (Pazzani 2002),(Billsus and Pazzani 2007) and (Smyth, Cotter, and Oman 2007). In (Smyth and Cotter 2002), finally, the effects of personalizing the navigational structure on a commercial WAP portal are reported.

It can be expected that this area will attract even more attention in the future because of the rapid developments in the hardware sector and the increasing availability of cheap and fast mobile Internet connections. Note that in contrast to some other approaches, our system does not exploit additionally available information such as the current geographical position or demographic and other customer information known to the service provider. Standard limitations of Mobile Internet applications such as relatively small network capacity and limited display sizes however apply.

Algorithms and ratings

During the four week evaluation period, customers were assigned to seven different groups when they entered the games section of the portal. For each group, the item lists were generated in a different way. For the first four groups, the following recommendation algorithms were used.

- Item-based collaborative filtering (CF) (Sarwar et al. 2001) as also used by Amazon.com (Linden, Smith, and York 2003).
- The recent and comparably simple SlopeOne algorithm (Lemire and Maclachlan 2005).
- A content-based method using a TF-IDF representation of the item descriptions and the cosine similarity measure.
- A “switching” (Burke 2002) hybrid algorithm that uses the content-based method when less than 8 item ratings are available and item-based collaborative filtering otherwise.

Two groups received non-personalized item lists, one based on the average item rating (“TopRating”) and one based on the sales numbers (top sellers). For the final group, the control group, the recommendation lists were manually edited as it was before the personalization features have been introduced. Within most categories, the ordering was based on the release date of the game or chosen based on existing contracts. The top-level link ”My Recommendations” was not available for the control group. During the whole evaluation period, customers remained in their originally assigned group.

From all customers that visited the games portal during the evaluation, a representative sample of over 155.000 was included in the experiment, so each group consisted of around 22.300 customers. Note that only such customers were chosen for which all algorithms were able to produce a recommendation, i.e., users for which a minimum number of ratings already existed. Also only such frequent customers were assigned to the control group and the groups receiving non-personalized recommendations, which guarantees that similar customer segments are compared. The catalog of recommendable items consisted of about 1.000 games.

A five-point rating scale from -2 to $+2$ was used in the experiments. Since the number of explicit item ratings was very low and only about two percent of the customers have issued at least one rating, also implicit ratings have been taken into account: both clicks on item details as well as actual purchases were interpreted as implicit ratings. When no explicit rating was given, a view on item details was interpreted as a rating of 0 (medium); several clicks on the same item were not counted. An actual purchase was interpreted as a rating of 1 (good) for the item. Explicit ratings override these implicit ratings.

In order to achieve the best possible recommendation accuracy, the item similarities and the average differences for the collaborative filtering and the SlopeOne techniques were computed using the full customer base and not only the 155.000 customer sub-sample.

Evaluation

The following hypotheses are in the center of our evaluation.

- H1: Personalized recommendations attract more customers to detailed product information pages (item view conversion rate).
- H2: Personalized recommendations help to turn more visitors into buyers (sales conversion rate).
- H3: Personalized recommendations stimulate individual customers to view more items.
- H4: Personalized recommendations stimulate individual customers to buy more items.

The detailed evaluation will show that depending on the navigational situation of the portal visitor different phenomena with respect to the effectiveness of recommendation algorithms can be observed. Before considering the overall effect of the use of recommendation technology on the portal, we will discuss the individual results obtained for these different situations.

Measurement 1: My Recommendations

The following results are related to the personalized recommendation list that is presented when the customer clicks on the “My Recommendations” link as shown in the top area of Figure 1. Throughout the evaluation, we will use different colors to highlight data rows in the charts that are significantly different ($p < 0.01$) from each other.

The conversion rate measurements (hypotheses H1 and H2) are given in Figure 2, which depicts the *item view conversion rate* for visitors of the "My Recommendations" list, and Figure 3 that shows how many of the users that visited the "My Recommendations" section actually purchased an item[2].

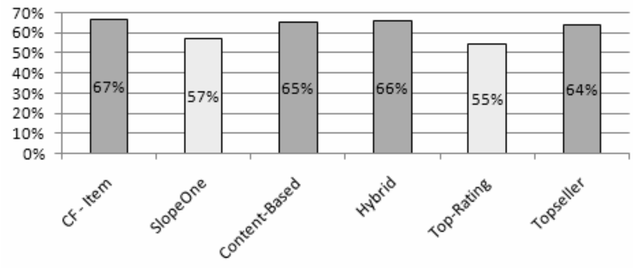

Figure 2: Conversion rate: Item views to "My Recommendations" visits

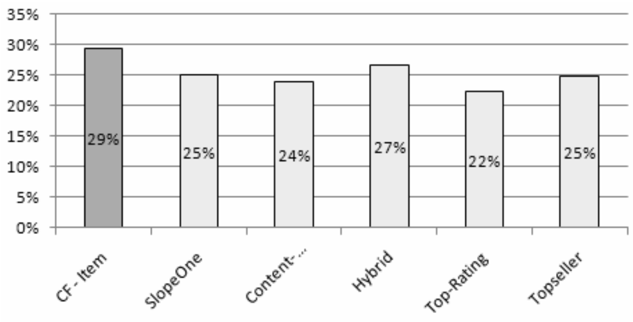

Figure 3: Conversion rate: Buyers to My Recommendations visits

In Figure 2 we see that the different algorithms fall into two groups: One, where about two thirds of of the customers actually click on at least one of the presented items and one, where only 55% are interested in the recommended items. Considering the actual numbers, the differences between the two groups are significant ($p < 0.01$).

From the personalized methods, only the SlopeOne algorithm did not attract significantly more visitors than the non-personalized list of top rated items. Interestingly, the non-personalized top seller list also has a good item view conversion rate, i.e., placing generally-liked, top-selling items in a recommendation list seems to work quite well in the domain.

When the sales conversion rate is considered, we see in Figure 3 that only the CF method helps to turn more visitors into buyers (Hypothesis H2).

[2]In Figures 2 to 5, the control group is not depicted, because the "My Recommendations" section, which was newly introduced for measuring the impact of personalization, was not available for them.

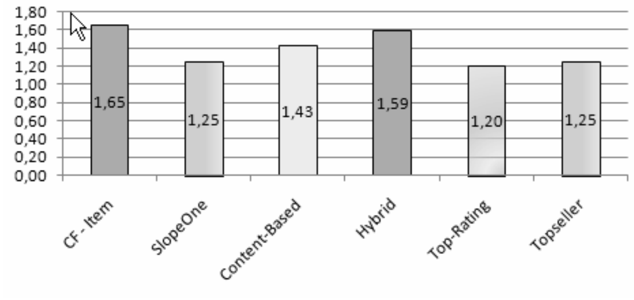

Figure 4: Item views per "My Recommendations" visits

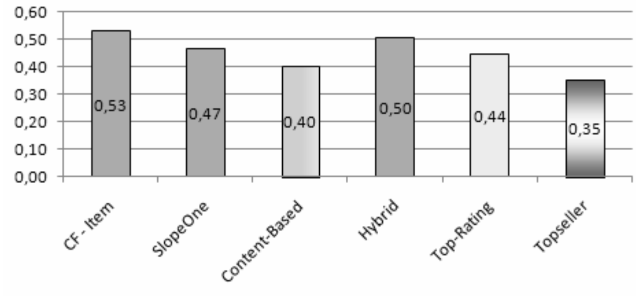

Figure 5: Item purchases to "My Recommendations" visits

The evidence for our hypotheses H3 (more item views per customer) and H4 (more purchases per customer) in the context of the "My Recommendations" section can be seen in Figures 4 and 5. Figure 4 shows that all recommendation algorithms (except for SlopeOne) stimulate users to click on more items. Compared with the findings with respect to the conversion rates, this can be interpreted as follows: while top seller lists help to stimulate one or the other customer to click on an item detail, personalized lists seem to contain more items that are interesting to a customer.

When it comes to actual purchases (game downloads), Figure 5 shows that most personalized methods and even the simple SlopeOne algorithm outperform the non-personalized approaches.

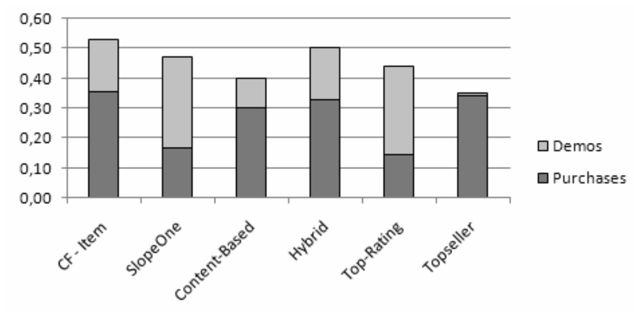

Figure 6: Game purchases and demo downloads in "My Recommendations"

Note that for some of the games provided on the mobile portal, free evaluation versions (demos) are available. If not mentioned otherwise, all numbers given with respect to conversion rates and sales figures are related to *all item downloads*, i.e. free demos plus actual game purchases. Figure 6 now repeats the numbers of Figure 5, but also shows the fraction of demo downloads and purchased games. Due to the nature of the algorithms and the particularities of the application (see more details in Measurement 4) , the recommendation lists produced by the TopRating and SlopeOne methods contain a relatively high portion of demo games. Given the high number of actual downloads, these demo recommendations seem to be well-accepted, but unfortunately, these two techniques perform particularly poor when the games are not free. The item-based, content-based and hybrid technique, on the other hand, not only help to sell as many items as a simple top-seller promotion but also make users curios about demo games. The TopRating method only raises interest in demo versions. The list of top selling items is generally dominated by non-free, mainstream games, which explains the fact nearly no demo games are chosen by the users.

Measurement 2: Post-sales recommendations

The next navigational situation in which product recommendations are made is when a customer has purchased an item and the payment receipt is displayed. About 90.000 customers who actually bought at least one item during the evaluation period have been involved in the experiment. Overall, the evaluation sample contains more than 230.000 views of the post-sales 5-item recommendation lists, meaning that on average, customers bought more than one item.

The experimental setup is nearly identical with Measurement 1 and customers received their recommendations based on different recommendation algorithms. The recommendation list of the control group was manually edited and ordered by game release date. Items that the current customer has already purchased before were removed from these lists.

The same hypotheses were tested in this experiment, i.e., to what extent recommender systems stimulate customers to view and buy more items. The results are shown in Figures 7 to 10.

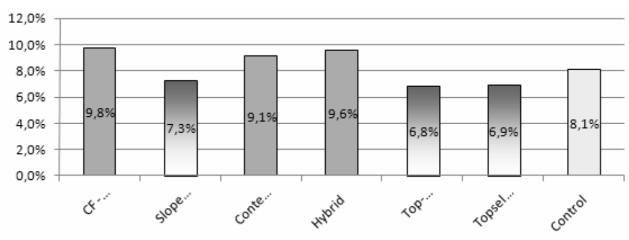

Figure 7: Conversion rate: Item views to post-sales list views

With respect to the conversion rates, the following observations can be made. First, the manually edited list of recent items (viewed by the control group) worked quite well and has raised more customer interest than the non-personalized techniques and even the SlopeOne algorithm

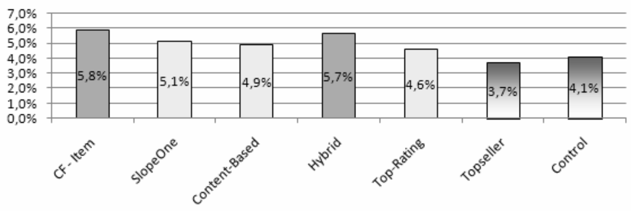

Figure 8: Conversion rate: Buyers to post-sales list views

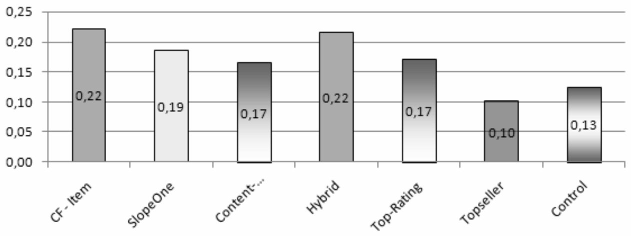

Figure 9: Item visits per post-sales list views

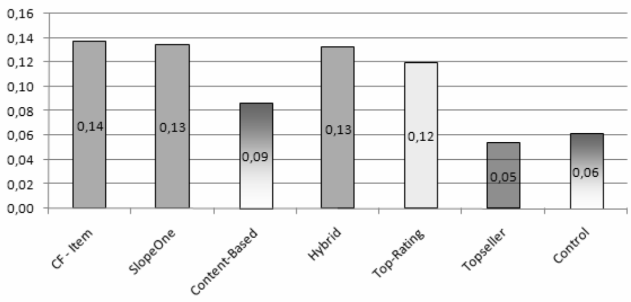

Figure 10: Item purchases to post-sales list visits

(Figure 7). When it comes to actual purchases (Figure 8), however, the manually-edited list did not help well to turn more visitors into buyers. Interestingly, the relative improvement caused by personalized recommendations with respect to this conversion rate is higher on the post-sales recommendation page than in the "My Recommendations" sections. Again, the CF algorithm worked best; in absolute numbers, the differences between the various techniques are significant, $p < 0.01$. With respect to the number of item visits and purchases per customer (Figures 9 and 10), it can again be observed that the different recommendation techniques not only stimulated visitors to view more items but actually also helped to increase sales. It can also be seen that displaying a list of top-selling items after a purchase leads to a particularly poor effect with respect to the overall number of downloads.

Another observation is that the items that are recommended by the SlopeOne technique and the TopRating method are also downloaded very often (see 10), presumably because the recommendation lists again contain many free demos. Figure 11 therefore shows the ratio of demo downloads to game purchases, which is quite similar to the one from the “My Recommendations” section, i.e., recommending top-selling or newly released items does not stimulate additional interest in free evaluation versions (demo games). The trend toward interest in demo versions seems to be a bit more amplified than in the “My Recommendations” section, which indicates that after a purchase transaction, customers first have a look on another, but free, game.

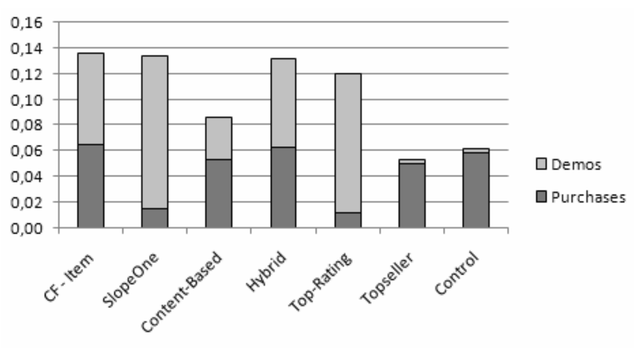

Figure 11: Game purchases and demo downloads on post-sales page

Finally, in this navigational context, the content-based method could raise some initial customer interest (Figure 9), perhaps because games are recommended that are quite similar to previously downloaded ones. However, while customers viewed some of the items, they had no strong tendency of purchasing them, probably because the games were – according to the general tendency of content-based methods – too similar to games they already know. The list of top selling items again contained mostly of non-free games, which explains the small fraction of demo games here; the same holds for the control group.

Measurement 3: Start-page recommendations

This measurement analyzes the effect of the personalized recommendations on the start page as shown in Figure 1. Remember that some elements in these lists are edited manually but were static during the experiment. Thus, we did not include item visits or purchases from these links (that could have been other banner advertisements as well) in the evaluation.

During the experiment, the personalized elements of the list, i.e., the last two text teasers and the first image teaser, were determined based on the top-3 list of the individual recommendation algorithms or based on the non-personalized lists of top-selling and top-rated items. Customers assigned to the control group received manually-determined recommendations which were ranked by release date.

For this experiment, we will only show the conversion rate figures for the different teaser elements on the start page.

Figure 12 shows the percentage of portal visitors that followed one of the personalized product links on the start page. On average, the image teaser was clicked on by around 6% of the users. Although the image only represents the third-ranked item of the recommendation algorithms and is also positioned after the text links, its conversion rate is significantly higher than for the text links. Since this also holds for the non-personalized methods, the attractiveness of the third link can be attributed to its visual representation. Interestingly, however, the image teaser leads to a good conversion rate with respect to actual sales (Figure 13). With respect to these conversion rates, both the CF method and the content-based method lead to a significant increase of item detail clicks and purchases. It can be also observed that the conversion rates of the first text teaser can even be better than the image teaser, when the text links are personalized. Thus, personalization can partially even outweigh the disadvantages of the unflashy representation.

Another particularity of this measurement on the start page is that the manually-selected items used for the control group lead to comparably good conversion rates, especially with respect to item visits. A possible explanation could be that customers have no special expectations with respect to the offers on the start page. The fact that the manually selected items are newly released ones might further contribute to the good acceptance.

Although recommending items based on their average customer rating (as done by the SlopeOne and the TopRating technique) worked well in the first two experiments, this approach does not work particularly well on the start page, i.e., customers seem to prefer either new items or items that are somehow related to their previous buying history.

Finally, when it comes to the number of purchases induced by the recommendation lists, the personalized techniques clearly outperformed the manually defined lists, at least for the first two teaser elements, see Figure 14.

Note that we also compared the item click and sales numbers of the other four and statically defined image and text teasers with the personalized ones. It could be seen that although the personalized items are partially placed lower on the screen and are thus harder to select, the received significantly more clicks and lead to more sales than the non-personalized links.

Measurement 4: Overall effect on demo downloads

In Measurement 1 and 2 we have seen that SlopeOne and the non-personalized technique based on item ratings lead to significantly more views and downloads of demo games. In this measurement, the goal was to analyze whether this trend also exists when the entire platform is considered, including, e.g., all other personalized and non-personalized navigation possibilities.

Note that no explicit category in the navigation tree for “free demos” exists. Games for which free evaluation versions exist, can however appear in all other personalized and non-personalized item listings in the portal. In addition, customers are pointed to demos in two additional ways: a) through direct-access links that are sent to them in sales promotions b) through pointers to other demo games that are displayed after a demo has been downloaded.

The distribution of views and downloads of demo games during the four-week evaluation period for the different recommendation groups is shown in Figure 15. Overall, about

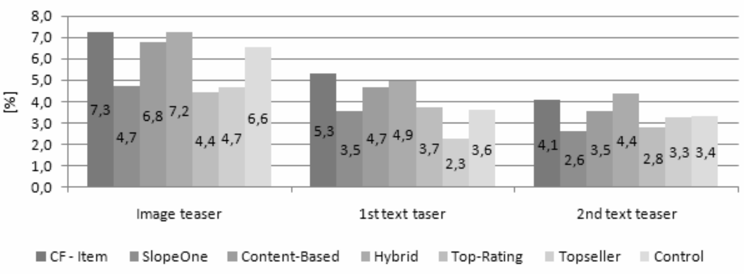

Figure 12: Conversion rate: Item views to start page visits

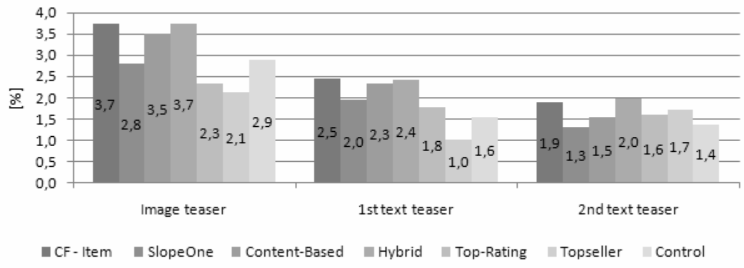

Figure 13: Conversion rate: Purchases from start page visits

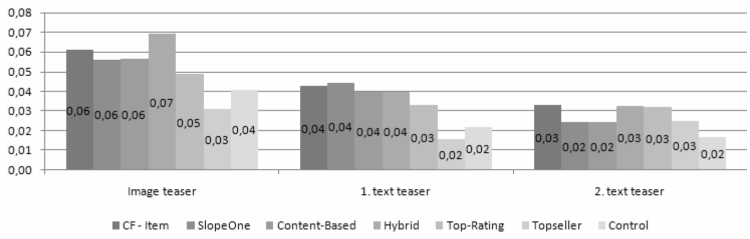

Figure 14: Purchases per start page visits

38.000 downloads have been observed for the selected subsets of customers. When considering the actual downloads, we see that the ranking of the algorithms remains the same; the differences are even amplified.

As already quickly mentioned in previous sections, this result can be explained by different facts that are related to the particular application setting and the nature of SlopeOne and the top-rating algorithm, which both tend to rank demo games highly in the different categories described above for the following reasons. First, as demo games can be downloaded at no cost and user ratings are only possible on the platform after a download, more explicit ratings are available for these games. Next, explicit ratings tend to be above-average also in this domain. Note that a similar phenomenon can also be observed in other datasets such as the MovieLens rating database. Finally, as customers receive a non-personalized pointer to another demos after downloading a free game, a reinforcement of the effect occurs.

An in-depth analysis, whether the downloads that were stimulated by the different algorithms lead to significantly *different* demo-download/purchase conversion rates is beyond the scope of the current study. What could, however, be observed in a first analysis is that the demo/purchase conversion rate is significantly higher when the demo was promoted by a recommendation list (as opposed to a banner advertisement).

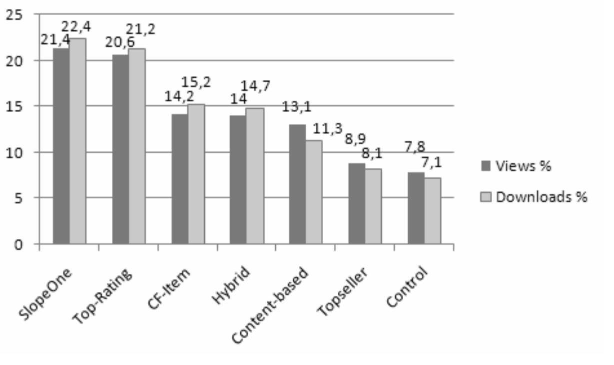

Figure 15: Distribution of demo game item views and downloads

Measurement 5: Overall effects

In this final measurement reported in this paper, the overall effect of the personalized recommendations (as an add-on to the other navigational options) situations was evaluated. Again, the interesting figures are related to item view and sales conversion rates (H1 and H2) as well as to the question whether more items are viewed and sold by individual customers (H3 and H4).

With respect to the conversion rates (hypotheses H1 and H2), no significant differences between the personalized and non-personalized variant could be observed, when the platform is considered as a whole. On average, about 80% of all observed customers have viewed at least one item and around 57% have bought at least one game, independent of the recommendation algorithm group they were assigned to. These figures are nearly identical for all seven test groups. For the item view conversion rate, for instance, the numbers only range from 79.6% to 80.3%. Thus, although slight improvements could be observed in individual (personalized) situations as described above, the influence on the overall conversion rate is too small and thus, the percentage of portal visitors that view or purchase items could not significantly be increased *by the additional use* of personalized recommendation lists.

There could be different reasons for this non-effect. First, remember that beside the described personalized lists, there are various other ways in which customers can access the item catalogs. Many customers for instance use the built-in search functionality of the portal; the ranking of search results is not personalized. The list of *new items* (see Figure 1) is also one of the most popular ways of browsing the catalog and used by significantly more people than, for instance, the new "My Recommendations" section. Our analysis shows that personalizing this particular list does not improve the conversion rates as customers always prefer to see the latest releases on top of such a list. Second, remember that only customers have been considered in the evaluation, for which a minimum number of rating already existed, i.e. users who are in generally interested in games. An evaluation of whether more *new users* can be tempted to purchase items was not in the focus of the evaluation.

With respect to the hypotheses H3 and H4 (increased number of item views and sales per customer), the following observations can be made. Regarding the average number of item views per customer (H3), we see that all personalized algorithms outperform the non-personalized topseller list and the control group. Similar to the effect of Measurement 4, SlopeOne and the simple ranking based on average customer rating raised the most attention. Thus, H3 could only partially be validated at the global scale as also the non-personalized top-rating technique was successful.

The observations made with respect to the number of purchased/downloaded items per customer (H4) are shown in Figure 16.

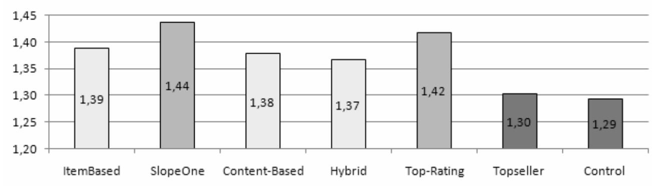

Figure 16: Average number of purchases including free downloads per customer on entire platform

The figure shows that the additional attention raised by SlopeOne and the "Top Rating" algorithm also leads to an measurably increased number of items purchased and downloaded per customer. Figure 17 shows the number of downloaded items (including the demos) for the different algorithms. If we, finally, look at the actual sales numbers for non-free games only (Figure 18), we can see that although the Top-Rating list raised attention for free demos, it did not lead to increased sales for non-free items. Overall, all personalized techniques were more successful than the non-personalized one. On the global scale, the difference was however – a bit surprisingly – only significant for the content-based method, which indicates that customers tend to spend money on items that are similar to those they liked in the past. In fact, a closer look on the performance of the algorithms in specific sub-categories shows that the content-based method often slightly outperforms the other methods with respect to non-free games. While the differences were not significant in the individual situations – which is why we did not include these figures here – these slightly higher sales numbers sum up to a significant difference on the global scale. Examples of categories in which the content-based method worked slightly better with respect to non-free games are the "new games", "half-price", or "erotic games" section of the download portal.

Overall, the increase in actual sales that are directly stimulated by the recommender system is between 3.2% when compared to the Top-Rating technique and around 3.6% when no personalized recommendation is available.

In general, these last observations suggest that in situations where the user has no strong expectations on a certain genre (such as the "MyRecommendations" section), collaborative methods – which also recommend items of categories that the user has not seen before – work particularly well. In many other situations, however, users tend to prefer recommendations of game sub-categories that they already know. One exception is the post-sales situation, where users are, non-surprisingly, not interested in purchasing games which are very similar to the one he or she has bought a moment ago.

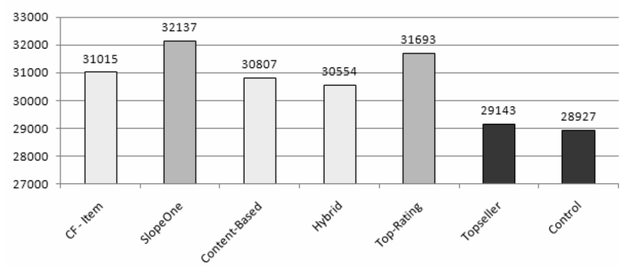

Figure 17: Total number of purchases and downloads

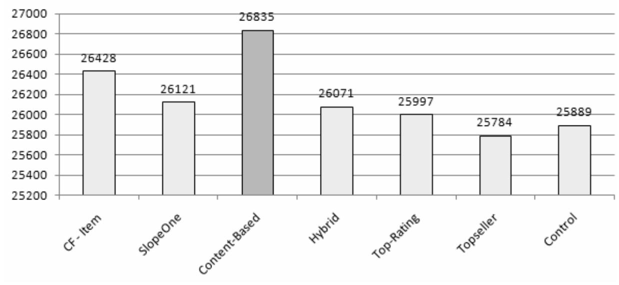

Figure 18: Total number of purchases (without demos)

Summary

In this work, the effects of personalized item recommendation in various navigational contexts on a Mobile Internet game portal were analyzed. Different standard recommendation techniques have been implemented on the portal and deployed in parallel in a real-world setting for a period of four weeks. In addition, non-personalized techniques based on top-selling or top-rated items have been used for comparison purposes.

The findings can be summarized as follows.

Ratings in the Mobile Internet

The amount of explicit item ratings was very low on the considered Mobile Internet portal and only about 2% of the users issued explicit ratings. While we are not aware of any studies that compare the willingness of customers to rate items in different settings, we suspect that the relatively high effort for submitting an item vote using a mobile device compared to a web browser discourages users to participate in this community process. When using only explicit ratings (in combination with a minimum number of neighbors as to decrease the Mean Absolute Error MAE), the *coverage* and applicability of individual algorithms quickly degrades. An analysis for the item-to-item algorithm for instance showed that the *item coverage* degrades to less than 50% when the minimum number of neighbors is set to the value of 20. In our experiments we therefore also took two types of implicit ratings into account: item views and item purchases, as the offline analysis showed that this combination leads to a good coverage (between 90% and 95%) for the item-based recommender. Note however that the MAE was also not very helpful to predict the accuracy of different algorithms in advance when including these implicit ratings. An offline evaluation showed that due to the high number of similar ratings very similar and low (below 0.2 points on the five-point scale) MAE values were achieved. An analysis of whether or not using different values for implicit ratings can help to further improve the recommendation accuracy, was not part of the current study.

Recommending in navigational context

In this study, the effects of personalized recommendations have been measured in different navigational situations such as the start page of the portal or the post-sales situation. In addition, we differentiated between the interest that was raised by the recommendations and the actual effect on the buying behavior of the customers.

With respect to the navigational context, customers seem to react slightly differently to recommendations, probably because of different expectations. In the dedicated "My Recommendations" section of the portal, classical collaborative filtering and the hybrid technique are particularly good at raising customer interest as customers view many of the recommended items. While customers are also easily stimulated to download free games by the comparably simple SlopeOne and TopRating method, these techniques do not lead to a significant increase in non-free games. A similar effect can be observed in the post-sales situation; the trend toward free demo downloads is even amplified in this situation. Thus, the item-based, content-based and hybrid technique which lead to a good number of purchases but also raise additional interest in demos, seems to be a good choice here.

On the portal entry page, the recommendation of top-rated items (or topsellers) has a particularly poor effect and the personalized methods lead to significantly better results. A listing of newly released items on the start page works however also quite well.

In certain navigational situations, we observed that personalization worsens the conversion rates and sales numbers. In the section on new items, which contains games of the last three weeks, the strict chronological order with the newest items on top works best. Most probably, the visitors of the "New" category enter this section regularly and only check the first few lines for new arrivals.

Finally, when measuring the number of game downloads including the demos on the entire platform, it shows that naive approaches such as TopRating and the comparably simple SlopeOne technique work sufficiently well to raise the users' interest in individual games. The important result, however, is that with respect to actual sales, the content-based and the item-based methods were clearly better than all others. Overall, it could be demonstrated that recommender systems are capable to stimulate a *measurable increase in overall sales* by over 3 percent on the entire platform.

Acknowledgement

The work was supported by Twistbox Games Ltd. & Co. KG, Germany. The described portal is based on Twistbox's Nitro-CDP™ content-download platform.

References

Billsus, D., and Pazzani, M. J. 2007. Adaptive news access. In Brusilovsky, P.; Kobsa, A.; and Nejdl, W., eds., *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, 550–570. Springer.

Breese, J. S.; Heckerman, D.; and Kadie, C. M. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Cooper, G. F., and Moral, S., eds., *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 43–52.

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction* 12(4):331–370.

Cho, Y. H.; Kim, C. Y.; and Kim, D.-H. 2004. Personalized image recommendation in the Mobile Internet. In *Procedings of 8th Pacific Rim International Conference on Artificial Intelligence*, volume 3157 of *Lecture Notes in Computer Science*, 963–964. Auckland, New Zealand: Springer.

Dias, M. B.; Locher, D.; Li, M.; El-Deredy, W.; and Lisboa, P. J. 2008. The value of personalised recommender systems to e-business: a case study. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, 291–294.

Fleder, D. M., and Hosanagar, K. 2007. Recommender systems and their impact on sales diversity. In *EC '07: Proceedings of the 8th ACM conference on Electronic commerce*, 192–199.

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

Lemire, D., and Maclachlan, A. 2005. Slope one predictors for online rating-based collaborative filtering. In *Proceedings of SIAM Data Mining (SDM'05)*, 471–480.

Li, Q.; Wang, C.; and Geng, G. 2008. Improving personalized services in mobile commerce by a novel multicriteria rating approach. In *WWW'08: Proceeding of the 17th international conference on World Wide Web*, 1235–1236.

Linden, G.; Smith, B.; and York, J. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE* 7(1):76–80.

McNee, S. M.; Riedl, J.; and Konstan, J. 2006. Accurate is not always good: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*.

Miller, B. N.; Albert, I.; Lam, S. K.; Konstan, J. A.; and Riedl, J. 2003. Movielens unplugged: experiences with an occasionally connected recommender system. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, 263–266.

Nguyen, Q. N., and Ricci, F. 2008. Long-term and session-specific user preferences in a mobile recommender system. In *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*, 381–384.

Pazzani, M. J. 2002. Commercial applications of machine learning for personalized wireless portals. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence*, 1–5.

Ricci, F., and Nguyen, Q. N. 2007. Acquiring and revising preferences in a critique-based mobile recommender system. *IEEE Intelligent Systems* 22(3):22–29.

Sarwar, B.; Karypis, G.; Konstan, J.; and Reidl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, 285–295.

Senecal, S., and Nantel, J. 2004. The influence of online product recommendations on consumers' online choices. *Journal of Retailing* 80(2):159 – 169.

Shani, G.; Brafman, R. I.; and Heckerman, D. 2002. An MDP-based recommender system. *Journal of Machine Learning Research* 6:453–460.

Smyth, B., and Cotter, P. 2002. Personalized adaptive navigation for mobile portals. In *Proceedings of the 15th European Conference on Artificial Intelligence*, 608–612.

Smyth, B.; Cotter, P.; and Oman, S. 2007. Enabling intelligent content discovery on the mobile internet. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, 1744–1751.

van der Heijden, H.; Kotsis, G.; and Kronsteiner, R. 2005. Mobile recommendation systems for decision making "on the go". In *ICMB '05: Proceedings of the International Conference on Mobile Business*, 137–143.

Zanker, M.; Bricman, M.; Gordea, S.; Jannach, D.; and Jessenitschnig, M. 2006. Persuasive online-selling in quality and taste domains. In *Proceedings of 7th Intl. Conference E-Commerce and Web Technologies*, 51–60.

Zanker, M.; Jessenitschnig, M.; Jannach, D.; and Gordea, S. 2007. Comparing recommendation strategies in a commercial context. *IEEE Intelligent Systems* 22(3):69–73.

Zanker, M.; Fuchs, M.; Höpken, W.; Tuta, M.; and Müller, N. 2008. Evaluating recommender systems in tourism - a case study from Austria. In *Proceedings Information and Communication Technologies in Tourism, ENTER 2008*, 24–34.