Evaluating Top-N Recommendations "When the Best are Gone"

Paolo Cremonesi, Franca Garzotto, Massimo Quadrana
Politecnico di Milano
Milano – Italy
{paolo.cremonesi, franca.garzotto, massimo.quadrana}@polimi.it

ABSTRACT

In a number of domains of interest for recommender systems, items are characterized by constrained and variable "capacity": the same product or service can be consumed by a limited number of users and the possibility of item consumption depends on contextual circumstances (e.g., time). Our work explores recommenders in the context of these "bounded" domains. We consider online hotel booking as a case study, and investigates if and how "missing" items (hotels that eventually becomes unavailable for users' consumption) affect the quality of recommendations. The paper proposes a technique for defining "missing" items as "best items", and presents an articulated empirical research in which recommendations for hotel online booking are evaluated in different experimental conditions with a user centric approach involving 142 participants.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Experimentation, Human Factors

Keywords

Top-N recommendation task, User-Centric Evaluation, Perceived Quality, Perceived Accuracy, E-tourism, Missing Items

1. INTRODUCTION

An implicit assumption of most existing studies on Recommender System (RS) evaluation is that all items are always available, regardless the number of users who have "consumed" (bought, used) them. In the video-on-demand or e-book business, for example, the electronic format of products allows a potentially infinite number of customers to consume them at any time. Still, in many domains, hereinafter referred to as *bounded domains,* items are characterized by *constrained* and *variable "capacity"*: the same product or service can be consumed by a limited number of users and the possibility for a customer to use or buy it depends on *contextual circumstances,* e.g., varies over time. Examples of items of this type are events and travel services, or tangible products such as clothing.

The research interest of bounded domains originates from the observation that, whenever the capacity of an item is emptied, the recommender system should consider that item as "*missing*". Still, *"missing" items of this kind are not missing at random*: as the number of available items decreases, their quality tends to decrease, because *the items that "are gone" are typically the "best ones"*. In e-tourism, for instance, the hotels that are optimal in terms of price/comfort rate are normally booked first and may become unavailable during high season or when the booking time is close to the desired time of usage; what remains available in these circumstances are typically the most expensive accommodations, or the ones that offer low quality services or are in the less attractive locations. Our work investigates how the time-varying capacity of items in bounded domains plays in the recommendation process. Specifically, this paper explores if and how the quality of Top-N recommendations is affected in situations where the "missing items" are those fully consumed and potentially the best ones.

Some prior studies focused on the "missing ratings" problem - a seemingly similar yet very different issue. The missing rating problem is related to the sparsity of user-rating matrixes and if/when missing ratings should be considered as a negative, positive or neutral user feedback when training [6] and evaluating [8] a recommender algorithm. Research in "unbounded" domains – where all items are always available - shows that high-rated items introduce biases in the design and evaluation of recommenders. These are due to the so called popularity effect and positivity effect [8]: the majority of high ratings are condensed in the small fraction of the most popular items, or short-head. The findings of a wide off-line study [3] pinpoint that when the user rating is not taken into account, the accuracy of non-personalized algorithms is comparable to the performance of sophisticated personalized algorithms. When the most popular and widely ranked items in the short head are removed, personalized algorithms are ranked higher than the non-personalized algorithm.

To explore the "missing items" problem arising in bounded domains, our work considers the case-study of hotel on-line booking. We performed a series of experiments that extend a *preliminary* research presented in [4] and are organized in two steps. The first step explores the process through which, in hotel online booking, users select the items they choose to "consume" first, and defines a technique to identify missing items. The second step compares the quality of personalized and non-personalized recommendations in two scenarios – full and limited availability of hotels – using on-line evaluation techniques (142 users). In both steps we exploit a relatively large dataset made available by Venere.com, a subsidiary company of the Expedia group.

RecSys '13, October 12 - 16 2013, Hong Kong, China
Copyright 2013 ACM 978-1-4503-2409-0/13/09…$15.00.
http://dx.doi.org/10.1145/2507157.2507225

Table 1. Dataset used in the study

	Total Venere+TripAdvisor	Venere	TripAdvisor (crawled)
Hotels	3,100	3,100	3,100
Users (reviewers)	210,000	72,000	138,000
Reviews and ratings	246,000	81,000	165,000
Hotel features	481	481	–

2. INSTRUMENTS

For the purpose of our study, we developed *PoliVenus*, a web-based testing framework that can be easily configured to facilitate the execution of controlled empirical studies in on-line booking services. PoliVenus implements the same layout as Venere.com online portal and simulates all Venere.com functionality except payment functions. Users can explore and filter the catalogue of items according to their characteristics (e.g., budget range, stars, accommodation type, city area of hotels). PoliVenus is based on a modular architecture, can operate *with* and *without* recommendations, and can be easily customized to different datasets and 20 different recommendation algorithms. In the baseline configuration (i.e., without recommendations) hotels are presented to the users ordered according to the editor's ranking criteria (e.g., best sellers, marketing strategies). With recommendations, hotels are ordered according to the recommender system ranking.

The user profile required by personalized algorithms to generate recommendations is based on the user's current interaction with the system (*implicit elicitation*). Whenever a user interacts with an object on the interface (e.g., link, button, map, picture, etc.), the system assigns a score to the hotel related to that object. With all these *signals*, PoliVenus builds the user profile for the current user session. The user profile contains implicit hotel ratings, where each rating is computed as a linear combination of all the signals generated for that hotel. The user profile is continuously updated by effect of every new signal and the list of recommended hotels is updated accordingly.

Venere.com has made us available with a catalog of more than 3,000 hotels and 72,000 related users' reviews. Each accommodation is provided with a set of 481 features concerning, among the others: accommodation type (e.g., residence, hotel, hostel, B&B) and service level (number of stars), location (country, region, city, and city area), booking methods, average single-room price, amenities (e.g., spa), and added values (e.g., in-room dining). Users' reviews associated to each accommodation consist of numeric ratings and free-text. We have enriched the original Venere.com dataset with additional reviews extracted from the TripAdvisor.com web site using a web crawling tool. Table 1 reports the detailed statistics of the dataset used in our experiments. The dataset is available by contacting the authors.

3. FIRST STEP: DEFINING "THE BEST"

In the first step of our research, we focused on how to create scenarios of low availability of items that simulate a realistic user experience. The not obvious problem is how to build the missing items set, identifying the items to be removed from the service provider's catalogue during the recommendation process in situations of scarce availability of items. In the hotel online booking domain, experience tells us that the number of hotels that can be reserved decreases during high season, and the best hotels are the first one to be booked and consequently the first ones to become unavailable. Hence we will refer to the missing items situations as the *"the best are gone"* scenario. For our simulation in the empirical study (step 2), we defined a technique to identify *which* are "best" hotels and *how many* of them should be included in the missing items set of the "best are gone" scenario.

To define "best" hotels, we can use two different metrics. According to common sense, the best hotels are the ones with the largest *average rating* $\bar{r}_i$, defined as

$$\bar{r}_i = \frac{\sum_u r_{ui}}{n_i}$$

where r_{ui} is the rating from user *u* to item *i*, and n_i is the number of users who rated item *i*.

As average ratings computed over a larger support n_i are considered more reliable by the user, *popularity* n_i can be used as an alternative metric to define "best" hotels. The two metrics are not necessarily correlated, as low popularity may come along with a high hotel rating and vice versa. To overcome this ambiguity, we adopted the definition of *shrunk average rating* $\bar{\bar{r}}_i$ introduced in [1]

$$\bar{\bar{r}}_i = \frac{\sum_u r_{ui}}{n_i + k}$$

where k is a shrinkage constant that controls the support of the estimate. For $k = 0$ hotels are ranked according to the traditional definition of average rating. For $k \rightarrow \infty$ hotels are ranked according to their popularity. In our experiments we set $k = 10$.

Once hotels are sorted based on their shrunk ratings $\bar{\bar{r}}_i$, the topmost hotels that capture *66%* of ratings are included in the missing items set. In our dataset, the missing set of items comprises 1500 "best" hotels, almost 50% of the 3100 hotels in the complete catalogue of PoliVenus. This value is close to the percentage of fully booked hotel during high season periods in a top level tourism destination in Italy like Rome as reported by *Venere.com*.

It is worth noting that the high season scenario removes from the dataset the short-head items as defined in [3], the only difference being in the notion of popularity (shrunk rating vs. number of ratings). Consequently, the hotels available in the high season scenario are those corresponding to the long-tail, as defined in [3].

4. SECOND STEP: ON-LINE STUDY

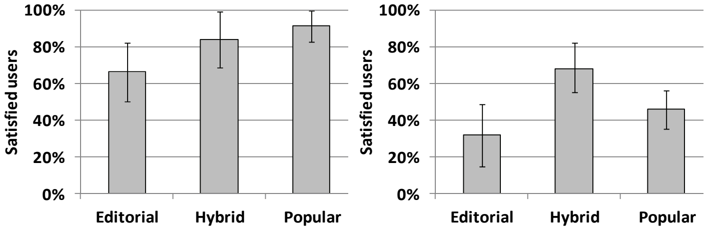
We model the quality of recommendations using *subjective* and *objective* variables, among which:

- *choice satisfaction* (subjective): the subjective evaluation of the reserved hotel in terms of quality/value for the user;
- *hotel price* (objective): the cost of one night for the reserved hotel;
- *extent of hotel search* (objective): an indicator of user's effort in terms of number of hotels that have been explored and for which detailed information has been acquired.

Subjective variables are measured using a web based questionnaire (containing 10 questions). Objective variables are assessed by analyzing interaction log data.

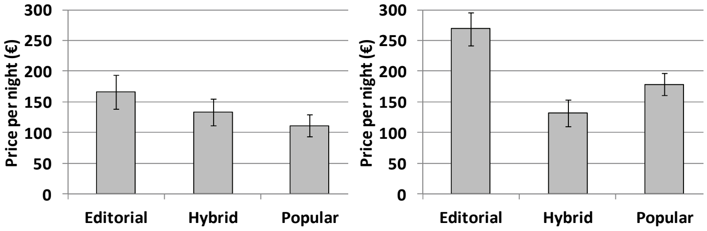
The effects of recommendations are explored under different experimental conditions, defined by the combination of two manipulated variables*: hotels availability* and *recommendation algorithm*. Hotels availability can assume two possible values: *low season* (full availability) and *high season* ("when the best are gone"). For algorithms, our study considers three recommendation techniques: *Editorial*, *Hybrid* and *Popular*.

- *Editorial*. We assume as *baseline* "algorithm" the most common approach of online booking systems, which ranks



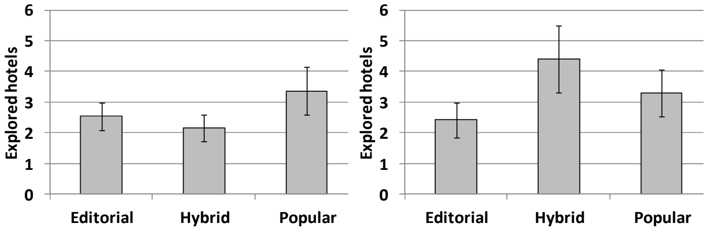
(a) low season (b) high season

Figure 1. Percentage of satisfied users



(a) low season (b) high season

Figure 2. Average cost per night



(a) low season (b) high season

Figure 3. Number of explored hotels

hotels based on some marketing strategy. We adopted the default ranking of Venere.com, which is (mainly) based on the number of users who booked a hotel.

- *Hybrid*. This technique generates a list of recommended hotels interleaving the results from *PureSVD* and *DirectContent* algorithms. Interleaving has been proposed in [2] with the name of "mixed hybridization". *PureSVD* is a collaborative algorithm based on matrix-factorization [3]. *DirectContent* recommends hotels whose content is similar to the content of hotels the user has rated [5]. Content analysis takes into account 481 features (e.g., category, price-range, facilities), the free text of the hotel description, and the free text of the hotel reviews. *DirectContent* is a simplified version of the LSA algorithm described in [1].
- *Popular*. This technique generates a ranked list of items based on the shrunk average rating $\bar{\bar{r}}_i$.

Our main research audience is represented by users aged between 20 and 40 who have some familiarity with the use of the web and had never used Venere.com before the study (to control for the potentially confounding factor of biases or misconceptions derived from previous uses of the system). The total number of recruited subjects who completed the task and filled the questionnaire was *142*. They were equally distributed in the 6 experimental conditions. We recruited participants from current students, alumni, and administrative personnel at the School of Engineering and the School of Industrial Design at our university. They were contacted by e-mail, using university mailing lists, and were not informed aware of the true goal of the experiment. To encourage participation and realistic behavior, we used a lottery incentive [7]: a randomly selected person who completed the assigned task and filled the final questionnaire by a given deadline would win a 100€ coupon to be used to stay in the hotel fictitiously reserved using PoliVenus.

All participants were given the following instructions: "*Imagine that you are planning a vacation in Rome and are looking for an accommodation during Easter season; choose a hotel and make a reservation for at most two 2 nights; dates and accommodation characteristics (stars, room type, services, and location) are at your discretion. After confirming the reservation (simulated), please complete the final questionnaire*".

After accessing PoliVenus and agreeing on the study conditions (lottery participation and privacy rules), each participant was automatically redirected to the homepage of the PoliVenus reservation system and randomly assigned to one of the six experimental conditions. After committing the reservation, the user was directed to the questionnaire page.

5. RESULTS

We polished the collected data by removing the ones referring to subjects who showed apparent evidences of gaming with the testing system (e.g., those who interacted with the system for less than 2 minutes) or left too many questions unanswered. In the end, we considered the data referring to *125* participants. They were almost equally distributed in the six experimental conditions, each one involving a number of subjects between *20* and *24*. This section presents the main results, focusing on the comparison between high and low season conditions, i.e., full availability of items vs. when "the best are gone".

Analysis of variance suggests that the six different experimental conditions (type of algorithm and hotel availability) have a significant impact ($p<0.05$) on all variables. We ran multiple pair-wise comparison post-hoc tests using Tukey's method on the mean value of the dependent variables. The results are shown in Figures 1–3, together with the 95% confidence interval.

Figure 1 describes *user satisfaction* in the 6 experimental conditions. In the *low season* scenario (Figure 1.a) more than 90% of the users are happy with top popular recommendations (percentage of "yes" answers to the question about choice satisfaction). In the *high season* scenario users are overall less satisfied than in the low season scenario for all the recommendation strategies. This is not surprising as users may interpret the scarcity of resources in a given period as a weakness of the catalogue of services and ascribe the phenomenon to the service provider rather than an objective contingent situation. For users receiving editorial and top popular recommendations the percentage of satisfied users reduces to half of its value: satisfied users drop from 60% down to 30% in the editorial case, and from 90% down to 45% in the top popular case. In contrast, *users receiving personalized recommendations in the high season situation are now the most satisfied (70%) and their number does not significantly differ from the low season scenario.*

It is interesting to compare these results against an objective variable, the average cost per night of the hotels reserved by users (Figure 2). There is a statistically significant negative correlation between hotel price and satisfaction. In the *low season* scenario (Figure 2.a) users who receive popular recommendations are more satisfied and spend significantly less than users receiving editorial recommendations. In high season, when most hotels are fully booked and the average cost per room is higher, the average price of booked hotels increases by more than 70% in the editorial condition and by approximately 50% in presence of top popular recommendations (Figure 2.b). Still, hotel price it is not significantly different between low and high season scenarios for the users receiving personalized recommendations.

Figure 3 plots the effort measured as *average number of hotels explored* by the users. The diagram confirms the intuition that searching for hotels in the low season period intuitively requires exploring fewer hotels than in the high season period. Less intuitively, in both low and high season conditions, the most satisfied users explore more hotels than the least satisfied users.

6. DISCUSSION

The analysis of three recommendation strategies in the two experimental conditions – low and high season – suggests a number of observations.

In the low season scenario the quality of non-personalized algorithm is comparable to – or even better than – the quality of personalized algorithms. A possible interpretation of this phenomenon is that, when the offer of products abounds, the opinion of the crowd has a stronger persuasion effect than a good match with individual preferences. When a large amount of products potentially satisfy the characteristics that are explicitly specified by the user (e.g., stars, services, location), popularity tends to be the most important attribute of items that drive a user's decision process. Hence in low season conditions the algorithms that are biased by popularity are more appreciated that personalized algorithms where popularity counts less. By their very nature, personalized algorithms rely more on a second order persuasion criterion (personalization), offering recommendations that do not necessarily match the opinion of the crowd (which is evident to the user from the popularity values and ratings of suggested items). In the high season scenario, when available items are in the long tail and have no or few user ratings, all recommended items appear to be "below threshold". They are somehow indistinguishable from one another with respect to persuasive attributes such as average rating and popularity, which become neutral in the decision process; therefore popularity-based algorithms reduce their effectiveness. At this point, other qualities of items become important, such as an acceptable match between item characteristics and personal needs; hence personalized algorithms which are unbiased by popularity increase their persuasion strength as they provide better support to users in discovering novel yet satisfactory alternative solutions.

When the most popular and widely ranked items are removed (i.e., "the best are gone") *the quality of the recommendations for non-personalized algorithms decreases.* In contrast, the effect of missing items is *negligible on the personalized hybrid algorithm*. The robustness of the hybrid algorithm with respect to missing popular items can be explained considering the partial content-based nature of the algorithm itself.

Finally, it is interesting to consider the results of the "price of booked hotels" in the different experimental conditions. It is intuitive that higher levels of choice satisfaction are related to lower price of the chosen hotel. More surprising is that the average cost of reserved hotels in the condition of personalized recommendation is not affected by the scarcity of offer, remaining stable in low and high season. By effect of personalized recommendations, users tend to explore more items, become more conscious of alternative offers, and seem to be more able to discover hotels at reasonable prices.

7. CONCLUSIONS

Our work focuses on business sectors that we call *bounded* domains and are characterized by two features: the *capacity of items*, i.e., the number of users that can consume them, is *limited* and *variable*, i.e., depends on circumstantial factors. We explore the consequences of these characteristics in an exemplar case of bounded domain, hotel online booking. We hypothesize (on the base of common sense and marketing data) that the variable reduced availability of items does not creates "missing data" at random, but it mainly involves "best" items, i.e., those most and highest rated. After defining a technique to calculate missing items and simulate low availability, "best are gone" scenarios, we carried on a series of studies to compare the effects on recommendation quality in two conditions: full availability of items (the implicit assumption underlying most existing research, which works with unbounded domains) and scarcity of items, in which "the best" items are removed from the search space used by the recommendation process.

Our work provides a number of contributions. We pinpoint how bounded domains are a challenge arena for recommender systems research. In addition, we provide empirical evidence that, in the considered case study, the *subtractive effect* resulting from item consumption and removal of missing items may strongly *weaken the performance of non-personalized popularity based recommendations*. In contrast, *personalized recommendations do not exhibit such a negative behavior,* suggesting that their relative quality is stronger in "the best are gone" condition. Finally, we propose a robust technique to define the "missing items" concept, which is needed to realistically simulate "best are gone" scenarios. This method, and the articulated design of our empirical study, can be adopted by other researchers in order to replicate our experiments in different case studies, to test our results using different algorithms, or to explore new research questions in bounded domains.

8. REFERENCES

1. Bell, R., Koren, Y., Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights. In *Proc. ICDM '07.* IEEE, pages 43–52, 2009.
2. Burke, R., Hybrid web recommender systems. In *The adaptive web*, Brusilovsky P., Kobsa A, Nejdl W. (Eds.). LNCS Vol. 4321. Springer-Verlag, pages 377–408, 2007.
3. Cremonesi, P., Koren, Y. and Turrin, R., Performance of recommender algorithms on top-n recommendation tasks. In *Proc. RecSys '10.* ACM, pages 39–46, 2010.
4. Cremonesi P., Garzotto F., Smoothly Extending e-Tourism Services with Personalized Recommendations: A Case Study. In *Proc. EC-Web 2013*. Springer, 2013 (to appear)
5. Lops, P., De Gemmis, M., and Semeraro, G., 2011. Content-centric recommender systems: State of the art and trends. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, P. B. Kantor (Eds). Springer, pages 73–105, 2011.
6. Marlin, B.M. and Zemel, R.S., Collaborative prediction and ranking with non-random missing data. In *Proc. RecSys '09.* ACM, New York, NY, USA, 2009, 5–12.
7. Porter, S. R., and Whitcomb, M. E., The Impact of Lottery Incentives on Survey Response Rates. *Research in Higher Education, 44*(4), 389–407, 2003.
8. Pradel, B., Usunier, N. and Gallinari, P., Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. In *Proc. RecSys '12*. ACM, pages 147–154, 2012.