Recommending without Short Head

Paolo Cremonesi Franca Garzotto Roberto Pagano Massimo Quadrana
Politecnico di Milano - p.zza Leonardo da Vinci 32, Milano - Italy
first_name.last_name@polimi.it

ABSTRACT
We discuss a comprehensive study exploring the impact of recommender systems when recommendations are forced to omit popular items (*short head*) and to use niche products only (*long tail*). This is an interesting issue in domains, such as e-tourism, where product availability is constrained, "best sellers" most popular items are the first ones to be consumed, and the short head may eventually become unavailable for recommendation purposes. Our work provides evidence that the effects resulting from item consumption may increase the utility of personalized recommendations.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Design, Experimentation, Human Factors

Keywords
Recommender Systems, Evaluation, Quality, e-Tourism

1. INTRODUCTION
A wide number of studies indicates that Recommender Systems (RSs) enforce the popularity of already popular items, favoring best-selling or "blockbuster" products, which represent what we refer to as *short head* [2][6]. But what happens if the items in the short-head become unavailable? More specifically, what are the effects of RSs trained on popularity-biased datasets (datasets with a small number of very popular items, and a large number of much less popular items) when recommendations are forced to use niche products in the *long tail* only? This is an interesting issue in domains, such as e-tourism, where products have constrained and variable capacity, i.e., they can be consumed only by a limited number of users. As the first items to be consumed are those most popular, i.e., the short head, recommendation algorithms must face situations in which eventually these items must be considered as "missing". Previous researches considering the role of missing data in RSs focus on missing ratings and how to interpret them (as a negative, positive or neutral user feedback) when training [3] and evaluating [4] a recommender algorithms. These works assume that all of the items are potential candidates for recommendation. In this work our focus is on missing items: we assume that they are not missing at random, as they account for the "short head", and they correspond to products that must be omitted by recommendations.

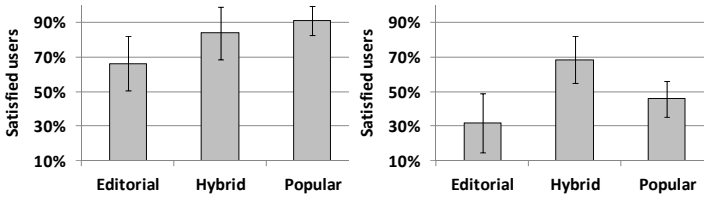
2. EMPIRICAL STUDY
We carried on a vast empirical study in the online hotel booking domain which involved 382 users and focused on the effects of short head removal on RS quality. We implemented a full sized simulation of a hotel booking service called *PoliVenus*, developed a predictive model for short head construction, and measured the quality of recommendations generated using different recommenders and in different conditions of items availability. PoliVenus is a web-based framework that can be configured to perform controlled experiments. It implements the same layout and functionality of the portal of Venere.com (a company of the Expedia group) except payment. Venere.com made us available a catalog of approx. 3,000 hotels and 72,000 related users' reviews which we integrated with reviews extracted from TripAdvisor.

Recommendations quality is defined in terms of *subjective* and *objective variables*. Subjective variables (e.g., *satisfaction* – the perceived quality/value of the reserved hotel) were measured using a web-based questionnaire based on the ResQue model [4]. Objective variables (e.g., *average hotel cost* per night for the reserved hotel, *average task execution time*, *average number of explored hotels*) were assessed using interaction log data. Quality variables are measured under 6 different experimental conditions, defined by the combination of two manipulated variables*: recommendation algorithm* and *hotels availability*. Our study considers three algorithms: (i) *Editorial* is the non-personalized, marketing-based ranking strategy adopted by Venere.com; (ii) *Popular* ranks hotels based on the shrank average rating defined later in the section, and (iii) *Hybrid* interleaves the results from a collaborative-filtering and a content-based algorithm.

For hotel availability we consider two possible values: *low season* (i.e., with full availability of hotels) and *high season* (*without short-head* because the best hotels are fully booked). Experience tells us that hotels which are the best in the users' opinion are the first to be reserved and to become unavailable in high season. Users' opinion is influenced by a number of factors, the predominant being others' opinion, manifested for example in on-line reviews, in terms of amount and average score of ratings. 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 rank hotels according to the *shrank average rating* $r_i = (\mu_i n_i + k n_i)/(n_i + k)$ where μ_i is the average rating of item i and k is a shrink constant that controls the support of the estimation. For $k = 0$ hotels are ranked according to the traditional definition of average rating. For $k \to \infty$ hotels are ranked according to their popularity. By setting a threshold on the shrank rating we can classify the hotels in two classes: the *short head* containing topmost hotels in the ranked list and the *long tail* containing the remaining ones. In our experiments we have simulated an occupancy of 50% (i.e., we assume 50% of hotels are fully booked) and we have used $k = 10$ for the shrank rating.

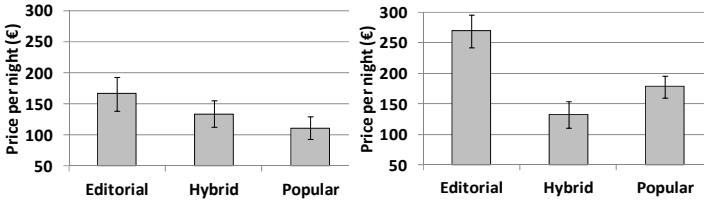
Study participants aged between 20 and 40, had some familiarity with the use of the Web and had never used Venere.com before the study. They were randomly assigned to one of the six experimental conditions and were asked to make a hotel reservation for 2 nights in a specific period.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577286



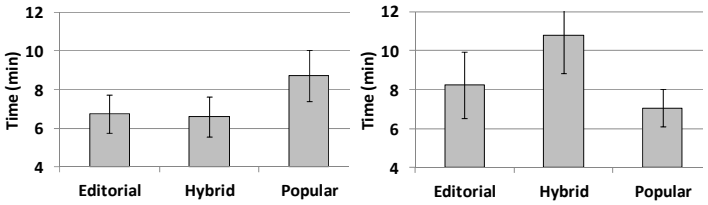
(a) with short head (low season) (b) without short head (high season)

Figure 1. Satisfaction



(a) with short head (low season) (b) without short head (high season)

Figure 2. Average cost per night



(a) with short head (low season) (b) without short head (high season)

Figure 3. Average task execution time

3. RESULTS

Analysis of variance suggests that algorithm and hotel availability have a significant impact on all quality variables. Concerning satisfaction, users receiving *Popular* non-personalized recommendations are the most satisfied in the low season condition (Figure 1.a): 90% of positive answers to satisfaction questions, in line with other studies showing that popularity-based recommenders do better or just as well as personalized ones. In all 3 high season conditions (Figure 1.b), users are overall less satisfied than in low season, being obviously disappointed by the limited offer of products. While satisfaction of users receiving *Editorial* and *Popular* recommendations decreases of about 50% in high season, it remains stable in users receiving personalized recommendations, who are now the most satisfied (70%). There is a statistically significant negative correlation between satisfaction and average price of reserved hotels. In low season (Figure 2.a) users with *Popular* recommendations are more satisfied and spend significantly less (100 € vs. 150 € in average per night) than users with *Editorial* recommendations. In high season, when most hotels are fully booked and costs are higher, the average price of booked hotels increases by more than 50% with the *Editorial* and *Popular* recommendations (Figure 2.b), while it does not significantly differ for users' with personalized recommendations. Data concerning *task execution time* confirms the intuition that searching for hotels in low season takes less time (Figure 3) than in the high season period when the decision making process is more complex. Less intuitively, users who invested more time on the decision process in both conditions of hotel availability are the most satisfied. Data concerning another measure of effort – *number of explored hotels* – confirm this phenomenon: users who explore the largest number of hotels are those with *Popular* recommendations in low season and with personalized recommendations in high season.

4. DISCUSSION AND CONCLUSIONS

The low season condition when all items are available is comparable to situations of potentially unlimited capacity that characterize most domains considered by RS research (e.g., movies). Hence our results on satisfaction in low season are in line with prior findings (e.g.[1]) pinpointing that the perceived quality of non personalized algorithms is comparable to the one of personalized algorithms. When a large amount of items potentially satisfy the specified characteristics, the opinion of the crowd is strongly influential on the decision making process. Hence popularity-biased algorithms are more appreciated than algorithms that rely on other persuasion criteria like personalization and suggest products less in line with the most popular ones. In contrast, scarcity of resources improves the relative quality of personalized recommendations: they are more appreciated than popularity-based recommendations. When the most popular solutions are gone (missing short head/high season condition), available items are in the long tail, have few user ratings and tend to be "below threshold" and indistinguishable from one another with respect to popularity. Other product qualities become important such as the match between product characteristics and personal needs, hence personalized algorithms, which are unbiased by popularity, increase their persuasion strength. When using personalized recommendations it is also interesting to notice that the average cost of reserved hotels is not affected by the scarcity of items, remaining stable in the two availability conditions: missing the most popular items in the short head, users are forced to spend more effort in search (exploring more items and digging more deeply into product features) and seem to become more conscious of alternative offers, and more able to discover hotels at reasonable prices.

In summary, our study shows that product consumption and unavailability of short head items weakens the performance of popularity based recommenders while enforcing the benefits of personalized recommendations. These results, although deserving further validation studies, may suggest reflections for the design of future recommender systems.

5. REFERENCES

[1] Cremonesi, P., Garzotto, F., and Turrin, R.,. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: an Empirical Study, *ACM Transactions on Interactive Intelligent Systems*, 2(2), 2012, 1–41.

[2] Lee. K. My head is your tail: applying link analysis on long-tailed music listening behavior for music recommendation. *Proc. RecSys 2011, ACM* pp. 213-220 2011.

[3] Marlin, B.M. Zemel, R.S., Collaborative prediction and ranking with non-random missing data. *Proc. RecSys 2009,* ACM pp. 5–12, 2009.

[4] Pradel, B., Usunier, N. and Gallinari, P., Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. *Proc. RecSys 2012*. ACM, pp. 147–154, 2012.

[5] Pu, P., Chen, L., and Hu, R., A user-centric evaluation framework for recommender systems. *Proc. RecSys 2011* ACM pp. 157–164, 2011

[6] Traverso, S., Huguenin, K., Trestian, I., Erramilli, V., Laoutaris, N., and Papagiannaki, K., TailGate: handling long-tail content with a little help from friends. *Proc. WWW 2012.* ACM, pp. 151–160, 2012.