A Month in the Life of a Production News Recommender System

Alan Said[†], Jimmy Lin[‡], Alejandro Bellogín[†], Arjen de Vries[†]
[†]Centrum Wiskunde Informatica, Amsterdam, The Netherlands
[‡]University of Maryland, College Park, Maryland, USA
alan@cwi.nl, jimmylin@umd.edu, alejandro.bellogin@cwi.nl, arjen@acm.org

ABSTRACT

During the last decade, recommender systems have become a ubiquitous feature in the online world. Research on systems and algorithms in this area has flourished, leading to novel techniques for personalization and recommendation. The *evaluation* of recommender systems, however, has not seen similar progress—techniques have changed little since the advent of recommender systems, when evaluation methodologies were "borrowed" from related research areas. As an effort to move evaluation methodology forward, this paper describes a production recommender system infrastructure that allows research systems to be evaluated *in situ*, by real-world metrics such as user clickthrough. We present an analysis of one month of interactions with this infrastructure and share our findings.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Information filtering, Retrieval models

General Terms

Design; Experimentation; Measurement

Keywords

Evaluation; Benchmarking; Live Evaluation; Recommender Systems

1. INTRODUCTION

With the rapid development of dynamic content on the World Wide Web, recommender systems have become a ubiquitous tool to help users generate and consume information online. One unaddressed "big issue" with recommender systems today is how to accurately evaluate them. Traditional evaluation of recommender systems builds on concepts from machine learning, statistics, and information retrieval [2] and is most often *offline*, in the sense that the recommendation algorithms are evaluated without the involvement of users. Instead, datasets containing *recorded* interactions are used as ground truth of "good" recommendations. This is, however, not without problems, as previous research has shown, e.g., [4, 5]. The nature of these problems are related to the concept of "recorded history", e.g., the lack of interaction between a user and an item does not imply that the item would have been a poor recommendation if it had actually been presented. The effect of the missing observations seems more significant than in the IR field (that the methodology has been borrowed from), as there relevant items tend to be rare, and, the results of multiple competing systems are pooled before assessment. Furthermore, it is known that items a user has not interacted with can nevertheless be good recommendations due to aspects such as diversity, novelty, serendipity, etc., but these factors are not considered in traditional evaluation [6].

As an attempt to advance the state of the art in evaluating recommender systems, this paper presents an analysis of the interactions in a production news article recommender system – Plista[1] – over a period of one month (June 2013). The dataset used for this analysis is provided by Plista within the scope of the News Recommender Systems Workshop and Challenge.[2] The Plista system delivers real-time recommendations of news articles to users currently browsing one of the news portals connected to Plista (providers). The most novel aspect of Plista is that it allows external researchers and practitioners to connect their recommendation algorithms to the Plista infrastructure as part of the Plista Contest[3] and deliver recommendations in real time to the system's users [3], thus allowing algorithmic evaluation *in situ*. To our knowledge, this is the first recommender systems setup that allows researchers to test their algorithms in a production environment on real-world users.

2. EVALUATION INFRASTRUCTURE

The recommendation infrastructure used in the Plista Contest allows participants to connect their own recommendation algorithms to Plista's news delivery framework through an HTTP API[4] where messages are sent between the contest server and the participant's client as JSON messages. Fig. 1 shows the flow of messages between the participant's client

LivingLab'13, November 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2412-0/13/10
http://dx.doi.org/10.1145/2513150.2513159 ...$15.00.

[1] http://www.plista.com
[2] https://sites.google.com/site/newsrec2013/
[3] http://contest.plista.com
[4] http://contest.plista.com/wiki/api

Provider	**Type**	**URL**
CFO World	Business	`http://www.cfoworld.de`
CIO	IT News	`http://www.cio.de`
Computerwoche	IT News	`http://www.computerwoche.de`
Gulli	IT & Games	`http://www.gulli.com`
Kölner Stadt-Anzeiger	News	`http://www.ksta.de`
Motor Talk	Automotive	`http://www.motor-talk.de`
Tecchannel	IT	`http://www.tecchannel.de`
Sport 1	Sports	`http://www.sport1.de`
Tagesspiegel	News	`http://www.tagesspiegel.de`
Wohnen und Garten	Home & Garden	`http://www.wohnen-und-garten.de`

Table 1: The 10 news providers, the types of news they deliver, and their URLs.

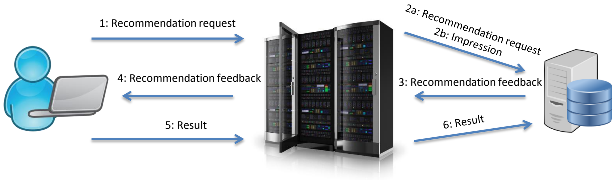

1: Recommendation request
4: Recommendation feedback
5: Result
2a: Recommendation request
2b: Impression
3: Recommendation feedback
6: Result

Figure 1: The flow of messages sent between a user reading an article, the Plista server, and the participant's recommendation algorithm.

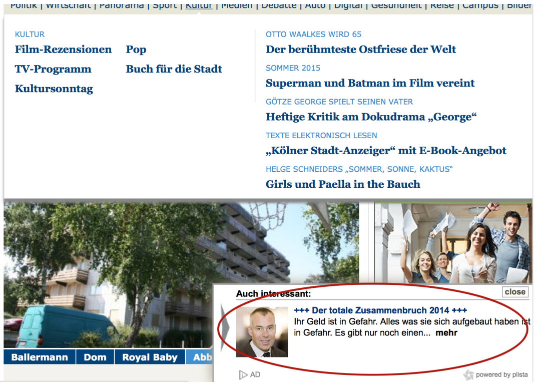
running a recommendation algorithm and the user reading a news article. The recommendation request (1) is sent as the user starts reading an article from one of the providers. Plista's servers forward the request to one of the participants (2a), while other participants are sent the impression (the user's id and information on the news article being read) without a recommendation request (2b). This step ensures all participants have access to information on all user-news article interactions. Whenever a participant receives a recommendation request (3), his or her system must provide a response within 200 ms, otherwise the recommendation will be served by Plista. Messages (5) and (6) are only sent when the reader clicks on the recommended news article. Fig. 2 show an example of a news recommendation being delivered to a reader.

Recommended items must be from the same provider as where the recommendation request was issued, and must still be "recommendable". Whether or not an item is recommendable is decided by Plista and is communicated to the participants in the impression messages, i.e., message (2) in Fig. 1. Due to the ephemerality of news, some articles loose significance over time and should thus not be recommended.

3. ANALYSIS

3.1 Basic Statistics

We present an analysis of the complete Plista logs over a period of 30 days (June 1st 2013 – June 30th 2013). The analyzed dataset contains the interactions, recommendations, and limited content (article title, URL, etc.) for a set of 14 news article providers. In total, we observed 335 thousand clicks and 34 million impressions from approximately 15 million users. Although there are a total of 14 providers, only 10 of them have clicks resulting from recommendations.


Politik | Wirtschaft | Panorama | Sport | Kultur | Medien | Debatte | Auto | Digital | Gesundheit | Reise | Campus | Bilder
KULTUR
Film-Rezensionen
Pop
TV-Programm
Buch für die Stadt
Kultursonntag
Ballermann
Dom
Royal Baby
Abb
close
AD
powered by plista

OTTO WAALKES WIRD 65

Der berühmteste Ostfriese der Welt

SOMMER 2015

Superman und Batman im Film vereint

GÖTZE GEORGE SPIELT SEINEN VATER

Heftige Kritik am Dokudrama „George“

TEXTE ELEKTRONISCH LESEN

„Kölner Stadt-Anzeiger“ mit E-Book-Angebot

HELGE SCHNEIDERS „SOMMER, SONNE, KAKTUS“

Girls und Paella in the Bauch

Auch interessant:

+++ Der totale Zusammenbruch 2014 +++

Ihr Geld ist in Gefahr. Alles was sie sich aufgebaut haben ist in Gefahr. Es gibt nur noch einen... mehr

Figure 2: An example of a recommendation being presented to the reader on Kölner Stadt-Anzeiger – `www.ksta.de`.

Articles	70,353
Users	14,897,978
Impressions	34,346,816
Clicks	334,865

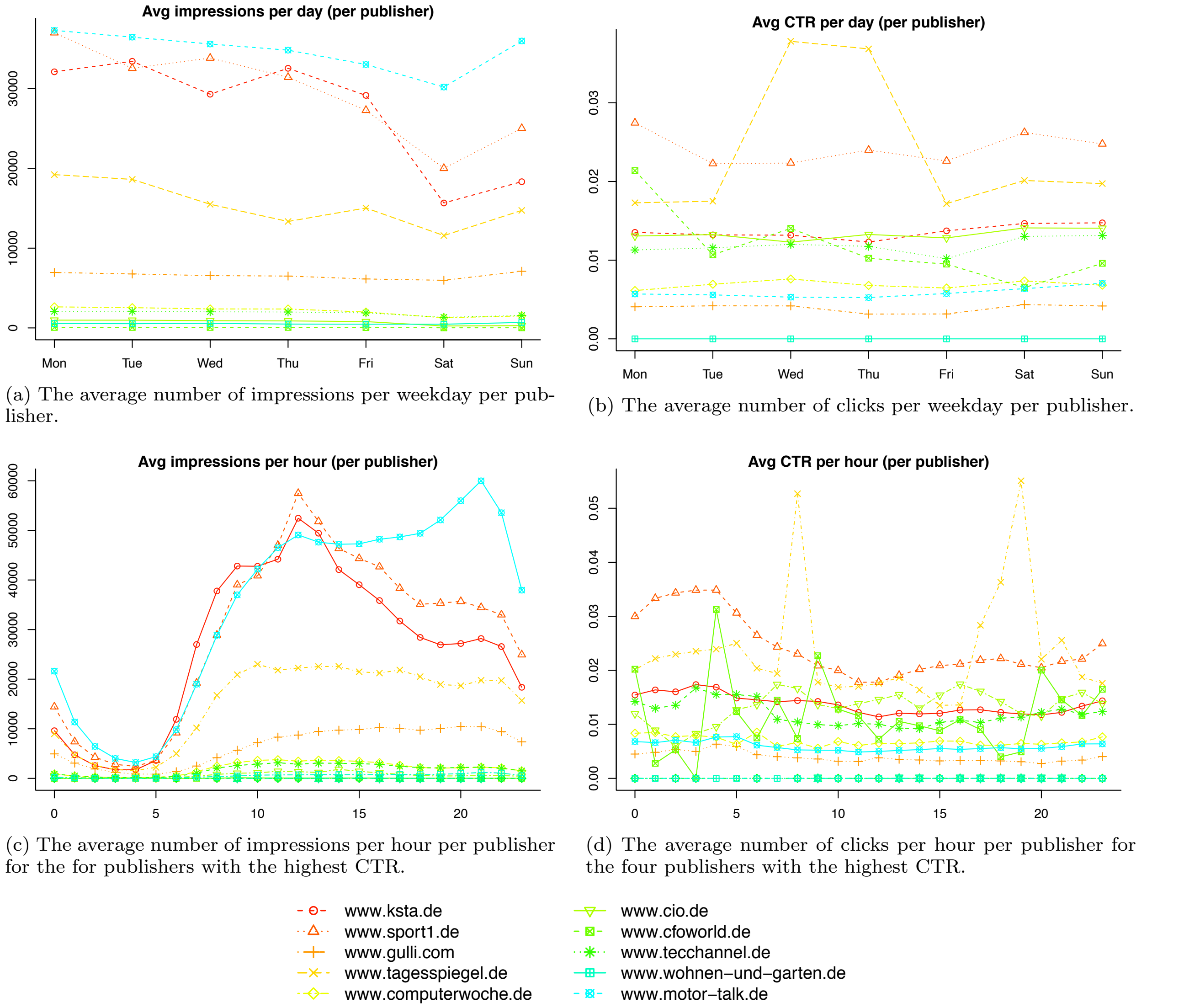
Table 2: The size of the analyzed data.

The impressions from the four providers without valid recommendations have been omitted in the data analysis.

All news providers in the dataset are from German news outlets with topics ranging from traditional news to specialized blogs (e.g., motorcycles, technology, business, etc.); a list of news types and provider URLs is available in Table 1. The recommendations in the dataset were provided by participants in the Plista contest. In the scope of the analyzed dataset, an *interaction* is the event where a user (a reader) views an article. The clickthrough rate (CTR) is the percentage of recommended articles that are clicked by the readers. The dataset details are shown in Table 2.

3.2 Interaction Analysis

The impressions and clicks performed during the analyzed time span are shown in Fig. 3. Specifically, Fig. 3a shows the average number of impressions per provider per weekday. The three largest providers (in terms of impressions) show how readership varies throughout workdays and weekends. The *traditional* news website KSTA appears to follow a similar impression pattern as the sports-related website Sport 1. It appears most readers visit these websites during workdays without any large fluctuations between different



(a) The average number of impressions per weekday per publisher.

(b) The average number of clicks per weekday per publisher.

(c) The average number of impressions per hour per publisher for the for publishers with the highest CTR.

(d) The average number of clicks per hour per publisher for the four publishers with the highest CTR.

Figure 3: The number of impressions and clicks per publisher averaged by the day of the week (Figs. 3a and 3b) and averaged over hours of the day (Figs. 3c and 3d).

workdays. There is, however, a significantly lower amount of articles read during the weekend (∼30%). The car-related website Motor Talk displays a different interaction behavior. Here, the amount of readers stays fairly constant over weekdays and the weekend, with a slight drop on Saturdays. Looking at the clicks on recommended articles over the same period, as shown in Fig. 3b, we see that a high number of impressions does not necessarily translate into a high number of clicks (CTR) on recommended articles. Instead, it would appear that traditional news (KSTA, Tagesspiegel) and sports-related websites tend to have a higher number of users clicking on recommended articles than users reading articles on more topic-centered websites, e.g., Computerwoche, Motor Talk, etc.

The (traditional) news website Tagesspiegel seems to have a large increase in clicked recommendations on Wednesdays and Thursdays specifically. However, this is caused by a significant drop in the number of impressions during 42 consecutive hours during the first week of June on the Tagesspiegel website, as shown in Fig. 4. This drop consequently affected the average number of impressions per weekday for the same provider in Fig. 3a where a slight downwards slope is visible for Tagesspiegel between Wednesday and Thursday. An analogous effect is seen in the two peaks in the hourly CTR shown in Fig. 3d at around 07:00 and 19:00. During these 42 hours, the average number of impressions per hour dropped to between 5 and 10, while the CTR during the same time span reached 20%, compared to the typical 1.5%. The cause of the lower amount of interactions was service disruptions at the provider.[5]

Looking at the hourly averages in Figs. 3c and 3d, specifically in terms of impressions (Fig. 3c) the majority of news-related impressions appear during the first half of the day, peaking around lunch. For other types of news, e.g., Motor Talk, Gulli, the trend is almost reversed: the peaks appear much later during the day. This is perhaps expected, as personal interests and hobbies are often dealt with after working

[5] This was confirmed in email communication between the authors and the News Recommendation Challenge organizers.

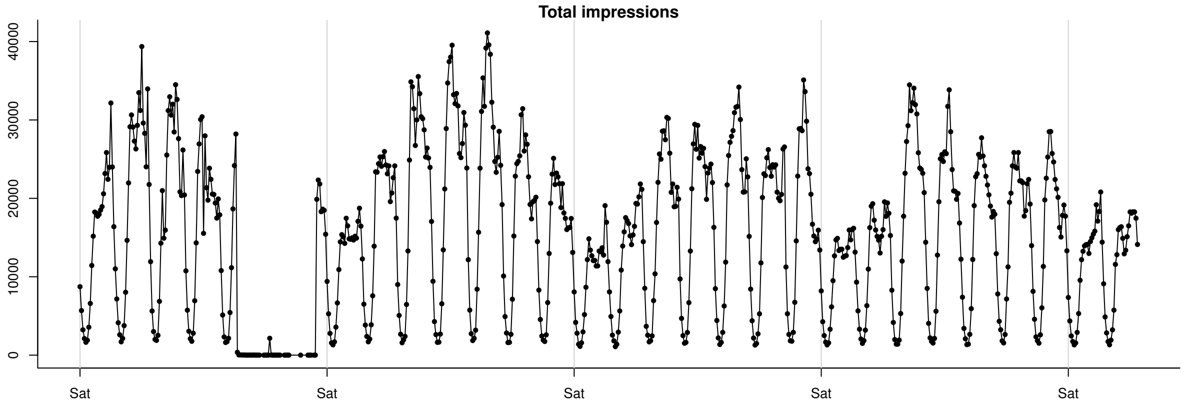


Figure 4: Impressions per hour for Tagesspiegel during June 2013. There is a significant drop in impressions at the end of the first week, which explains the spike in CTR in Figs. 3b and 3d.

hours. The hourly CTR trend, shown in Fig. 3d, appears to be similar to that of the weekly CTR trend, i.e., traditional news sources have higher CTR, whereas more topic-focused sources remain at low CTRs throughout the day. With the exception of CFO World and Sport 1, there are no significant changes in CTR over the day. The spikes in CFO World are likely related to a lower number of new articles produced during certain periods, lowering the average CTR.

Our analysis points to two distinguishable trends: First, traditional and sports-related news sources like Tagesspiegel, KSTA, and Sport 1 generally receive the bulk of impressions during the first half of the day, and, second, traditional news sources generally have higher CTRs than more topic-focused sources, e.g., compare KSTA and Motor Talk.

4. DISCUSSION AND CONCLUSIONS

In this paper, we have analyzed and visualized the weekly and hourly impressions and clickthrough rates in the Plista news recommendation system. By analyzing one month worth of interaction data we have identified a few trends in news recommendation and shown that in situ evaluation is sensitive to factors not related to the recommendation itself. We have isolated an unnaturally high CTR for one of the news providers in the dataset, the cause of which is service disruptions. Not knowing about disruptions like this, a recommendation algorithm could potentially be tuned incorrectly based on service malfunction instead of recommendation quality. This points to the importance of transparency in online evaluation frameworks and infrastructures, which can be problematic for the service providers from privacy and business insight perspectives. The identified trends point to conceptual differences in news across different domains, e.g., traditional news sources are mainly consumed during the first half of the day whereas topic-focused news receive the bulk of their interaction latter in the day. Readers of traditional news are more likely to interact with recommendations than readers of topic-focused news. We believe this can be an effect of traditional news being read for generally informative reasons, whereas readers of news related to certain personal interests are more likely to seek out information interesting to them – without the aid of a recommender system.

Traditional evaluation of recommender systems mainly focuses on the predictive qualities of algorithms, but in a live evaluation framework like the one presented here, other factors need to be taken into consideration. For example, the speed of a recommendation algorithm is rarely examined in a research context, but in a live system it is imperative that the recommendations are delivered with low latency. Indeed, Plista enforces a strict time budget of 200 ms, after which the participant loses the opportunity to make a recommendation. Another example of the importance of this aspect is the Netflix Prize,[6] where the winning algorithm was awarded the one million dollar prize, but was never implemented in the production system due to too high complexity and running time [1].

In situ recommender evaluation frameworks such as the Plista infrastructure described here may provide a solution to the thorny problem of recommender systems evaluation. This also represents a nice example of mutually-beneficial collaborations between academia and industry, and opens many avenues for future research. We are currently investigating whether results from one domain can be translated into another thus alleviating problems related to sparsity.

5. ACKNOWLEDGMENTS

Parts of this work were carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°246016.

6. REFERENCES

[1] X. Amatriain. Building industrial-scale real-world recommender systems. In *RecSys*, 2012.

[2] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM TOIS*, 22(1):5–53, Jan. 2004.

[3] B. Kille, F. Hopfgartner, T. Brodt, and T. Heinzt. The plista dataset. In *NRS*, 2013.

[4] R. Kohavi. Online controlled experiments: introduction, learnings, and humbling statistics. In *RecSys*, 2012.

[5] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI EA*, 2006.

[6] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *RecSys*, 2011.

[6] http://www.netflixprize.com