On the Decaying Utility of News Recommendation Models

Benjamin Kille
Technische Universität Berlin
Ernst-Reuter-Platz 7
10587 Berlin, Germany
benjamin.kille@tu-berlin.de

Sahin Albayrak
Technische Universität Berlin
Ernst-Reuter-Platz 7
10587 Berlin, Germany
sahin.albayrak@dai-labor.de

ABSTRACT

For how long will a recommendation model provide adequate recommendations? The answer to this question depends on the kind of model, its underlying data, and the domain among other factors. We analyse four types of models in the news domain on how their predictive performances change. Our observations show that replacing or updating models is necessary to maintain high predictive performance. The evaluation suggests that an exponential decay model describes the changing predictive performance accurately.

CCS CONCEPTS

• **Information systems → Data stream mining**; **Recommender systems**;

KEYWORDS

news recommendation, cold-start, model update, time-awareness, decaying utility

ACM Reference format:
Benjamin Kille and Sahin Albayrak. 2017. On the Decaying Utility of News Recommendation Models. In *Proceedings of Workshop on Temporal Reasoning in Recommender Systems, Como, Italy, 31 August, 2017 (TempRec'17),* 6 pages.
https://doi.org/

1 INTRODUCTION

Content providers compete to attract and retain information consumers in what can be described as "attention economy". Therein, consumers trade their attention in exchange for information and entertainment. Brynjolfsson and Oh (2012) stress the difficulty quantifying the value of such exchanges. Their estimate puts the collective annual value for such exchanges in the United States at 100 billion dollars. Ciampaglia *et al.* (2015) emphasise the limited attention span for newly published contents. Publishers employ recommender systems to provide consumers better information access (Billsus and Pazzani 2007). Recommender systems reduce vast collections of items to manageable subsets. In dynamic environments, they seek to maximise the number of interactions thus connecting users and items. The rate at which interactions occur is directly linked to business success. The more users engage with the collection of items the more advertisements they encounter. The more they enjoy the service, the less likely they are to quit using it. As a result, successful recommender systems represent a competitive advantage.

TempRec'17, 31 August, 2017, Como, Italy

Research on recommender systems has produced a myriad of methods. These methods take data related to users, item, or interaction between them. Subsequently, they learn regularities and create a *model* capturing the essential information. The models include global rankings, sets of rules, and latent factor representations among others.

Consequently, businesses continuously contemplate which model to use to generate recommendations. Ideally, they would choose the model maximising users' attention. Although, determining the utility of recommendation models has proven a difficult task. Shani and Gunawardana (2010) point to a variety of properties linked to the performance of recommender systems. These include accuracy, novelty, and diversity. In other words, recommender systems ought to provide relevant, new, and different items.

Frequently, the interaction data is split into disjoint partitions. One partition, the training set, is used to learn a model describing the relation amid users and items. The remaining partitions can be used to (a) optimise parameters, and (b) assess the utility. Cross-validation, a procedure wherein random partitions are permutatively used for training or testing, helps to limit the risk of randomly selecting an unrepresentative sample.

Still, using the described methodology, we merely obtain information about what the best model *would have been* at some point in time. We frame the problem from a slightly different perspective. Suppose we have a set of recommendation models available. Suppose further that we measure utility by models' ability to predict with which items users will interact in the future. We focus on how the utility of a set of recommendation models changes over time. In particular, we posit the hypothesis that the utility change can be modelled in form of an exponential decay function. We use part of the data set released for CLEF NewsREEL 2017 to conduct our evaluation (Lommatzsch *et al.* 2017). The data set comprises logs of various news publishers. News represent a particularly suited domain for our analysis. Publisher publish news articles at high rates. Simultaneously, readers favour novel news. Consequently, we expect models' utility to change rapidly.

This work entails two contributions. First, we formalise the concept of decaying utility of recommender models in the news domain. Second, we conduct experiments for four selected models.

The remainder of this paper commences with Section 2 introducing the notion of *decaying utility*. Section 3 describes the experimental design used to analyse the changes in utility over time. Section 4 presents our observations. Section 5 notes limitations and discusses our findings. Section 6 relates our work to previously published results and ideas. Section 7 summarises our findings and points to directions for future work.

2 DECAYING UTILITY

Recommender systems provide lists of suggestions upon request. The selection follows a set of rules represented in form of a model. Models are derived from previously recorded data. We define the *utility* of such a model as its ability to correctly predict future interactions amid users and items. Formally, let $U = \{u_m\}_{m=1}^{M}, I = \{i_n\}_{n=1}^{N}$ refer to the sets of users and items. The recommender system monitors interactions amid users and items $r = (u_m, i_n)$. Thereby, the system collects a set of interactions $R_\tau = \{r_\alpha\}_{\alpha=1}^{A}$, where interactions occurred in a closed time interval $\tau = [t_0, T]$, and interactions are chronologically ordered $t_\alpha < t_{\alpha+1}$. A recommendation model M_{R_τ} is a function that takes an interaction r_α and returns a list of suggestions $(i_k, i_{k+1}, \ldots, i_K)$. Let $t = [t_0, T]$ with $t_0 > \tau$. The utility of M_{R_τ} with respect to t refers to the number of interactions $r_\alpha \in R_t$ where u_m previously has been suggested reading i_n. We normalise the utility by dividing through the number of requests. A request refers to each interaction occurring in t. Thereby, we obtain a utility measure which we refer to as *response rate*. In practise, the response rate can be monitored by keeping track of which items have been recommended by the model. We hypothesise that the utility, or more concretely response rate, follows an exponential decay. Similar to radioactive decay, readers perceive an article as particularly interesting close to its publication. As time progresses, the news has spread and the article attract fewer readers. Exponential decays is characterised by the function $f(t) = U \cdot e^{Vt}$, wherein U and V are the parameters. The function describes a decay if $V < 0$. Alternatively, the *half-life* $t_{1/2} = \frac{\ln 2}{-V}$ describes the time it takes to arrive at half the initial quantity.

3 EXPERIMENT

We conducted experiments to measure the change of utility in terms of response rates for a selection of models. We consider the four publishers whose characteristics are shown in Table 1. The data correspond to one week of the NewsREEL 2017 data set. We notice that sessions include few articles. Publisher B observers merely 3.3 articles per session on average. This impedes using models which rely heavily on sufficiently expressive user profiles such as collaborative filtering. For each publisher we consider the time between 1–9 February, 2016. We learn four types of models each with the data of 1 February, 2016. First, the *random* model takes all articles and suggests a random subset. Second, the *freshness* model suggests the articles in chronologically reversed order of publication. Third, the *popularity* model suggests articles proportional to how frequently they had been read. Fourth, the *sequence* model uses the frequency of reading sequences. In other words, given an article i_n, the model suggests another item proportional to the frequency with which it had been read after i_n. We apply the model to all requests in the time 2–8 February, 2016. We determine whether readers subsequently read any of the suggested articles. With this information, we compute the average response rate for each hour. In addition, we monitor newly added articles and derive the coverage of models. The coverage is defined as the proportion of known articles covered by any model. The coverage naturally shrinks as the publishers release more and more articles unknown to the models. We repeat this procedure shifting the period by one day at a time. Thereby, we can compare the differences in response rates for the same day given different models.

4 EVALUATION

We consider the change in response rate as an appropriate proxy for the utility of a recommender model over time. Figure 1 shows the change in response rates over time for all combinations of publishers and models. The response rates are plotted on a logarithmic scale. For all models and publishers, we observe a decreasing trend in response rates. The *sequences* model exhibits the highest response rate for publishers A, B, and C. The *popularity* model exhibits the highest response rate for publisher D. The *random* model performs worst in the initial phase and mostly stagnates at this level. The *popularity* model overtakes the *sequences* model over time. This implies that businesses need to carefully monitor performances.

Figure 2 shows the relation between response rates and coverage for publisher A. We observe that as coverage decreases all models loose predictive accuracy. The effect is most apparent for the *freshness* model.

We analyse how much we could gain by retraining the models on a daily schedule. We focus on the *sequences* model. Figures 3 contrasts the response rate to the number of requests and coverage. The top part of each subfigure shows the number of requests. At times with fewer requests, response rates are based on a smaller set of interactions. We observe this phenomenon particularly at night time. The bottom part of each subfigure shows the coverage. The retrained models are shown in varying colours. Similarly, the centre part shows the response rates in corresponding colour schemes. Initially, models have a relatively high predictive quality.

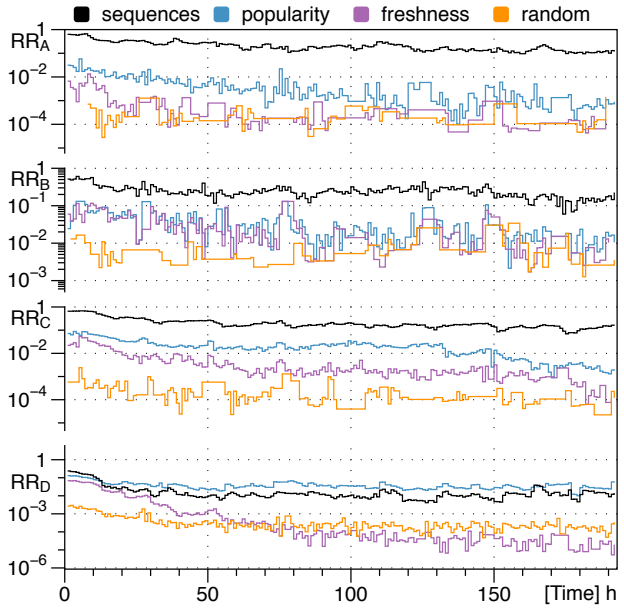


Figure 1: For each publisher, we consider the response rates for four types of models. The response rate is plotted on a logarithmic scale to prevent cluttering.

Table 1: We consider four publishers each referred to by a character label. Content refers to the category of news the publisher offers. The data has been collected during 1–9 February 2016. Sessions refers to the number of unique session cookies observed. Articles refers to the number of unique articles, which users read at least once. Interactions refers to the total number of reads. For each publisher, we present the mean number as well as standard deviation of reads per session. Likewise, we include the mean number and standard deviation of new articles added per hour. Note that besides addition, publishers can change articles to include new information.

Label	Content	Sessions	Articles	Interactions	Interactions per Session	Articles per Hour
A	general news	616 539	74 172	3 860 115	6.2±12.2	19.8±12.6
B	information technology	24 643	2735	82 540	3.3±4.0	1.0±0.2
C	general news	815 260	58 392	5 772 802	7.1±16.9	7.0±4.5
D	sports	1 437 161	12 028	20 227 882	14.1±30.5	7.2±4.5

The predictive performance subsequently decreases and stabilises on a noticeably lower level compared to the initial performance. We observe a noticeable difference in predictive performance amid the retrained models and their predecessors. This effect appears closely linked to the coverage, which shows a similar trend. The observations are consistent on all four publishers and affirm the expectation of a exponential decay phenomenon. Publisher B attracts less visitors and exhibits higher variance compared to the other publishers. Retraining models appears particularly beneficial to publisher D for which the decline in predictive performance quickly renders models useless.

Figure 4 illustrates the loss in predictive performance incurred when using the initial model as opposed to learning a new model on the second day. We observe that the loss is highest on the first day for all publishers. The differences in utility level off over time. For publisher B, we observe that the older model occasionally performs better than the new model.

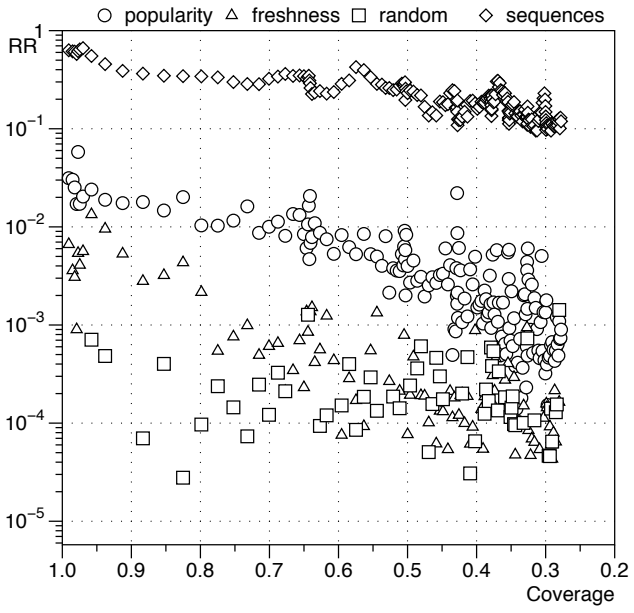
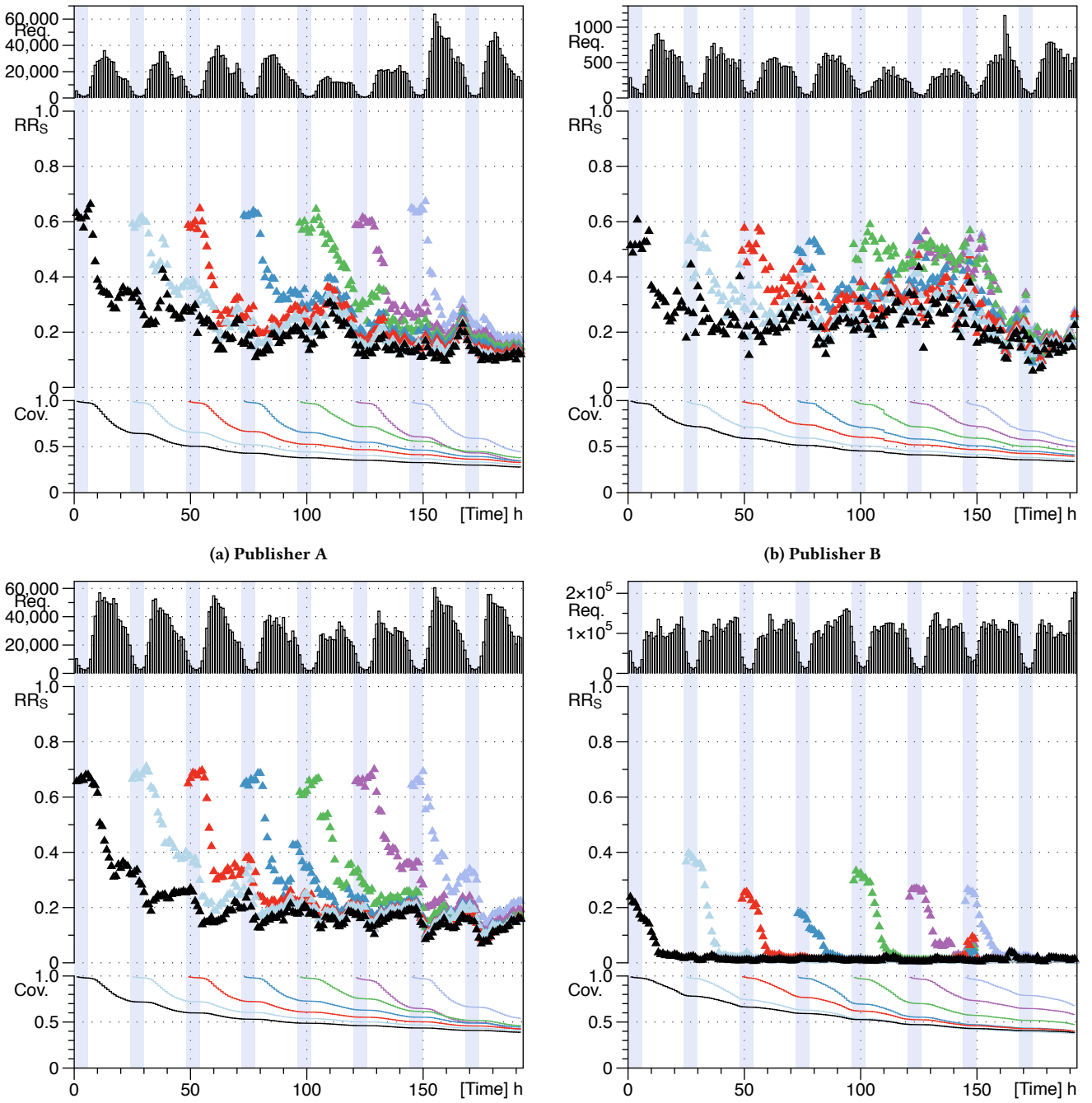


Figure 2: We consider the relation between coverage and response rate for publisher A.

We have fitted an exponential function to our results using the least squares method. Table 2 conveys the exponential fits to the response rates for combinations of publishers and models. We observe that the initial response rates (U) vary considerably. The *random* model has particularly low initial response rates. Conversely, the *sequences* model scores highest with respect to initial response rates. All fits exhibit decay, $V < 0$, with the exception of the random model for publisher B. Recall that publisher B observed less interactions than the other publishers. This could cause higher levels of variance.

5 DISCUSSION AND LIMITATIONS

The evaluation indicates that exponential decay models represent a suited first attempt to mathematically describe how the utility of recommendation models changes over time. The parameters vary among publishers and models. Still, Figure 3 shows similar trends for the *sequence* models across all publishers and despite which day we picked. The coverage appears highly related to the decaying response rates. Figure 3 and Figure 2 illustrate this relation. As time passes, publishers add new articles to their collections. Unless we update the models used to provide recommendations, they cover a lesser proportion of articles. The distribution of requests over the course of the day affects the response rates. Figure 3 illustrates the differences in requests for all four publishers. We observe a periodic pattern with more requests during the day and fewer requests at night. In addition, we observe that as the coverage arrives at 50 % the response rates level off for the *sequence* models and all four publishers. Figure 4 shows that switching to a retrained model is most beneficial on the first day. This suggests that publishers should replace or update their models at least once a day. Additional experimentation is necessary to analyse how the choice of data used to create the model affects its utility. We have kept the training data set to the length of one day in our experiments. Using more data and/or different types of models represents the direction to further explore. Our experiments used recorded data and inferred the utility rather than observing actual interactions resulting from recommendations generated by our models. Joachims *et al.* (2017) discuss how counterfactual reasoning facilitates using logged information more effectively. Unfortunately, we lack the required information on internal parameters of the recommender systems



(a) Publisher A

(b) Publisher B

(c) Publisher C

(d) Publisher D

Figure 3: Evaluation Results Overview. Each subfigure refers to a single publisher. Each subfigure contains three parts: at the top, the frequency of requests, at the centre the response rate (RR) referring to the sequence model, and at the bottom the coverage. For response rate and coverage, a colour scheme refers to the day at which the model has been created. Night times are highlighted in light blue.

to apply their method. Our experiments are based on part of the NewsREEL data set. In order to verify our findings, we have to conduct experiments with the feedback of actual readers. This will

Table 2: Exponential fit to the response rates observed for combinations of publishers and models.

Publisher	random	freshness	popularity	sequences
Publisher A	$0.0000 \cdot e^{-1.8058t}$	$0.0069 \cdot e^{-0.0712t}$	$0.0292 \cdot e^{-0.0376}$	$0.4406 \cdot e^{-0.0088t}$
Publisher B	$0.0018 \cdot e^{0.0028t}$	$0.0778 \cdot e^{-0.0206t}$	$0.0655 \cdot e^{-0.0106t}$	$0.3498 \cdot e^{-0.0039t}$
Publisher C	$0.0005 \cdot e^{-0.0176t}$	$0.0400 \cdot e^{-0.0505t}$	$0.0590 \cdot e^{-0.0140t}$	$0.4853 \cdot e^{-0.0107t}$
Publisher D	$0.0027 \cdot e^{-0.0531t}$	$0.0877 \cdot e^{-0.0875t}$	$0.0646 \cdot e^{-0.0057t}$	$0.2887 \cdot e^{-0.1079t}$

confirm whether the selection of publishers or the time period may have biased the findings.

6 RELATED WORK

The decreasing predictive performance of models has been discussed by Jambor *et al.* (2012) for the domain of movies. They employed methods from Control Theory to devise an optimised updating strategy. Movies exhibit different characteristics than news. In particular, people tend to revisit movies much more frequently than news thus impeding comparisons to our work. Koren (2009) focused on collaborative filtering. He introduced a latent factor model which captures the temporal development of preferences. Thereby, he could more accurately predict how users rate movies. Collaborative filtering requires expressive user profiles with sufficiently clearly stated preferences. News consumption happens anonymously disallowing creating such profiles. As Table 1 illustrates, publishers generally get to know readers' preferences for few articles. News recommender systems have to work in conditions

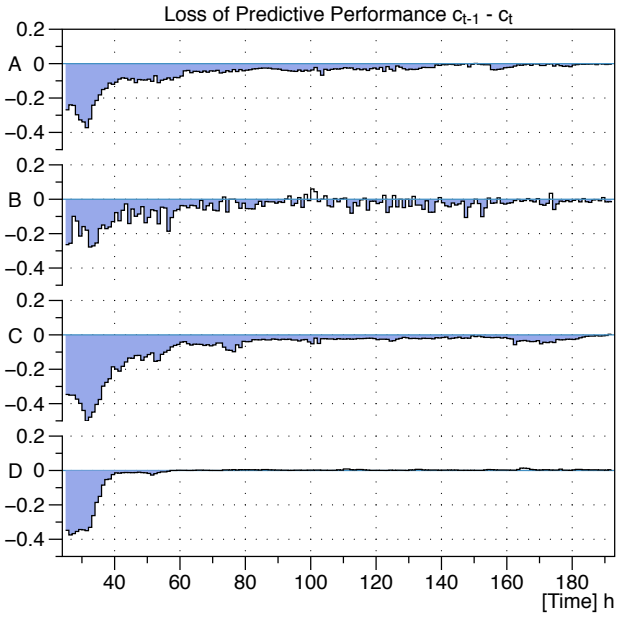


Figure 4: Comparison of response rates for the sequence models learnt on 1 February ($t-1$) and 2 February (t) in the period 2–9 February, 2016. The highlighted areas show the loss in predictive performance by using the older model.

in which little information is available about user preferences. Baltrunas and Amatriain (2009) extended the time-aware collaborative filtering to implicit feedback. Implicit feedback can be derived from log files such as the ones used in our experiment. Still, they apply their method to movies, which again exhibit characteristics different to news. Campos *et al.* (2014) discussed time-aware evaluation protocols. They introduce a scheme to categorise evaluation protocols focussing on rating prediction. Their scheme assigns our work the time-dependent cross-validation category. Much of the work on time-aware evaluation of recommender systems has focused on movies and rating prediction.

Das *et al.* (2007) present the news personalisation systems used for Google's news aggregator. Their system employs covisitation counts similar to our *sequence* model. In addition, they use probabilistic latent semantic indexing and MinHash clustering to improve their response rates. The news aggregator has access to much more comprehensive user profiles for the subset of users reading news while logged in with their Google accounts. Li *et al.* (2010) represent news recommendation as contextual-bandit problem. Therein, the system has a set of choices modelled figuratively as arm of bandit found in casinos. The system learns how to choose depending on the context. Garcin *et al.* (2013) introduce the notion of context trees to news recommendation. Context trees capture particularities of situation and use them to select a better set of article to be recommended.

7 CONCLUSION AND FUTURE WORK

We have introduced the notion of *utility decay* for news recommender systems. The utility decay refers to a model's decreasing ability to correctly anticipate future interactions amid users and items. Experiments with data from four publishers have confirmed that exponential decay functions can be used to describe the changes of response rates over time. We observed a similar pattern for the coverage, the proportion of articles a model can potentially suggest. We conjecture that there is a strong relation between the two quantities. The relation depends on factors including the publisher and the type of model. Further evaluation is necessary to improve the understanding of utility decay in news recommendation. First, we will consider varying the time span used to learn a model. This will show whether reducing or increasing the amount of data describes the changes of response rates more accurately. Second, we will consider additional types of models. With little information concerning users, we plan to evaluate an item-based latent factor model. We intend to participate in the next edition of NewsREEL to verify our findings with the feedback of actual news readers. Finally, we will evaluate additional time periods to verify that the observed pattern is not due to choosing a particular time.

REFERENCES

Linas Baltrunas and Xavier Amatriain. 2009. Towards Time-dependant Recommendation based on Implicit Feedback. *Workshop on Context-aware Recommender Systems* (2009).

Daniel Billsus and Michael J Pazzani. 2007. Adaptive News Access. *The Adaptive Web* (2007), 550–570.

Erik Brynjolfsson and JooHee Oh. 2012. The Attention Economy - Measuring the Value of Free Digital Services on the Internet. *ICIS* (2012).

Pedro G Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware Recommender Systems: a Comprehensive Survey and Analysis of Existing Evaluation Protocols. *User Modeling and User-Adapted Interaction* 24, 1-2 (2014), 67–119.

Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2015. The production of information in the attention economy. *Scientific reports* 5, 1 (May 2015).

Abhinandan Das, Mayur Datar, Ashutosh Garg, and Shyamsundar Rajaram. 2007. Google news personalization - scalable online collaborative filtering.. In *WWW*. ACM, New York, New York, USA, 271–280.

Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized news recommendation with context trees.. In *RecSys*. ACM Press, New York, New York, USA, 105–112.

Tamas Jambor, Jun Wang, and Neal Lathia. 2012. Using Control Theory for Stable and Efficient Recommender Systems.. In *WWW*. ACM, New York, New York, USA, 11–20.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *the Tenth ACM International Conference*. ACM Press, New York, New York, USA, 781–789.

Yehuda Koren. 2009. Collaborative Filtering with Temporal Dynamics. *KDD* (2009), 447.

Lihong Li, Robert E Schapire, Wei Chu, John Langford, and John Langford. 2010. A contextual-bandit approach to personalized news article recommendation. In *the 19th international conference*. ACM Press, New York, New York, USA, 661–670.

Andreas Lommatzsch, Benjamin Kille, Frank Hopfgartner, Martha Larson, Torben Brodt, Jonas Seiler, and Özlem Özgöbek. 2017. *CLEF 2017 NewsREEL Overview: A Stream-based Evaluation Task for Evaluation and Education*. Springer.

Guy Shani and Asela Gunawardana. 2010. Evaluating Recommendation Systems.