Enhancing Session-Based Recommendations through Sequential Modeling

Stéphane Martin
EPFL, LIA
Lausanne, Switzerland
stephane.martin@epfl.ch

Boi Faltings
EPFL, LIA
Lausanne, Switzerland
boi.faltings@epfl.ch

Vincent Schickel
Prediggo SA
Lausanne, Switzerland
schickel@prediggo.com

ABSTRACT

Recommender systems typically determine the items they should recommend by learning models of user-preferences. Most often, those preferences are modeled as static and independent of context. In real life however, users consider items in sequence: TV series are watched episode by episode and accessories are chosen after the main appliance. Unfortunately, since sequences are more complex to model, they are often not taken into account.

We developed an efficient sequence-modeling approach based on *Bayesian Variable-order Markov Models* and combined it with an existing content-based system, the *Ontology Filtering*. We tested this approach through live evaluations on two e-commerce sites. It dramatically increased performance, more than doubling the CTR and strongly increasing recommendation-mediated sales. These tests also confirm that the technique works efficiently and reliably in a production setting.

CCS CONCEPTS

• **Information systems → Online shopping**; **Recommender systems**;

KEYWORDS

Recommmender Systems, Sequence-modeling, E-commerce, Context-tree, Variable-order Markov Model

ACM Reference Format:

Stéphane Martin, Boi Faltings, and Vincent Schickel. 2018. Enhancing Session-Based Recommendations through Sequential Modeling. In *UMAP '18: 26th Conference on User Modeling, Adaptation and Personalization, July 8–11, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3209219.3209259

1 ONTOLOGY FILTERING AND BAYESIAN VARIABLE-ORDER MARKOV MODELS

Ontology Filtering (OF, [5]) is a fast recommendation technique that has been used commercially for years with great success by Prediggo SA[1]. On the other hand, the *Bayesian Variable-order Markov Models* (BVMM, [2]) have already proven quite effective for news recommendation [3] but can be slower. Our implementation combines them by first generating a candidate set of M items with OF. The final set of $N < M$ recommendations is then selected from the candidate set according to a model of sequential page-views embodied in the BVMM.

OF proceeds by organizing a catalogue of products as a hierarchical ontology in order to infer ratings. As most businesses lack a predefined ontology for their products, the ontology is constructed automatically out of the attributes of the products and their descriptions in natural language. Ratings are then inferred based on the user's behaviour. The method has been shown to provide excellent results [6], especially on long-tail items, and has been exploited commercially for years by Prediggo SA. It has proven to work well with large numbers of users and items (more than 1,000,000 products).

A BVMM is a variant of Markov-chain models in which the probability of the next item in a sequence is computed by averaging the probabilities predicted by a series of *experts* associated with subsequences of varying length. In its applications to online recommendation [3], the sequences model the succession of page-views by the visitors of a website. In each step, the aim is then to predict the N pages with the highest probability of being visited next.

The algorithm organizes those sequences into a context tree [1], i.e. a suffix-tree in which the page most recently visited is a direct child of the root. Each path from the root to a node n_s matches a sequence s, observed at least once, called the *context* of the node. Each node is associated with a weight w_s and an *expert* object e_s, that keeps statistics on the items visited immediately after the node's sequence. The function of e_s is to estimate the probability for each item to be visited next by a user having just gone though s.

A user currently browsing the last item of a sequence s has also traversed all the suffixes of s. The probabilities emitted by the experts of all those suffixes, including the root, are thus aggregated into a weighted average that forms the predicted probability of the next item. The computation of this weighted probability can be done in time linear to the height D of the tree [2]. The weights are also adjusted in $O(D)$ in accordance with the observations, following a Bayesian approach — this is the peculiarity of the BVMM.

Although the BVMM computes the probability of each item in $O(D)$, it has to do so over the whole set of candidates to determine the best ones. When the size K of the catalogue becomes too big, this approach in $O(KD)$ does not scale well. We thus use the learned

UMAP '18, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5589-6/18/07…$15.00
https://doi.org/10.1145/3209219.3209259

[1]Prediggo SA — http://www.prediggo.com/ — is a Swiss-based start-up specializing in recommendation and search solutions for e-commerce. The OF technology is protected by a pending US patent and is available commercially through Prediggo SA.

product-ontology to preselect a subset of $M \ll K$ candidates by extracting from the ontology the M products closest to the user's current product. Besides lowering the computational complexity to $O(MD)$, this policy of *ontological preselection* ensures the semantic relevance of the candidates by adding a content-based aspect to the collaborative filtering nature of the BVMM.

2 LIVE-USER EXPERIMENTS

We tested this method by organizing live experiments on the websites of two of Prediggo's clients. The first one, a retailer of furniture, home appliances and electronic equipment, operates a high-traffic e-commerce website in Switzerland. The second one manages a medium-sized e-boutique specializing in youth fashion. The tests were run from the 21st of November to the 30th of December 2017, on the same servers where Prediggo's engine operates normally.

In both experiments, the traffic was divided into two buckets each covering 50% of all sessions. Sessions in the first bucket received recommendations selected as usual by the OF, while the sessions of the second were assigned to the BVMM with ontological preselection. The trees were built at intervals of 30 minutes out of the 500,000 most recent page-views and were allowed to grow until a limit of 100,000 nodes. The purchases of the visitors and their clicks on the recommendations were logged for analysis.

During the first experiment (E1), the system received ≈ 5,700,000 calls for recommendations to be chosen among 10,951 products. Tab. 1 (top) shows the number of requests for recommendations, the number of clicks performed by the users, the CTR (clicks / requests ratio), the mean time to generate one recommendation and the values of sales following a recommendation within 24 hours[2]. There is little doubt that the BVMM is more effective at triggering the interest of the users. Taking the equivalence of the recommenders as the 0-hypothesis, Fisher's exact test yields a p-value over the clicks so small that our statistical library rounded it to zero.

Table 1: CTR, time for 1 reco. (ms), post-recom. revenue.

	Clicks	Total Calls	CTR	Time	Revenue
E1, OF	18,148	2,782,676	0.652%	1.912	100 %
E1, BVMM	66,631	2,886,756	2.308%	3.487	298 %
E2, OF	1371	129,431	1.059%	1.523	100 %
E2, BVMM	2827	133,804	2.113%	0.755	124 %

The second experiment (E2) covered ≈ 263,000 calls for recommendations for 2219 products. Its results confirm the conclusions of the first one, although the differences were not as striking (Tab. 1, bottom). This can be because a smaller site lacks sufficient traffic to train the sequence model. The BVMM achieves a CTR clearly higher than OF, 1.059% against 2.113%. The p-value on clicks, also computed with Fisher's exact test, is just as tiny as it was in the first experiment (≈ 0). The difference between the post-recommendation revenues is also in favour of the BVMM, although the volume of purchase was smaller. The improvements in accuracy do not entail a significant increase in computational complexity. Both methods can run on a standard server.

[2]For reasons of confidentiality, the amounts are shown here in percent of OF.

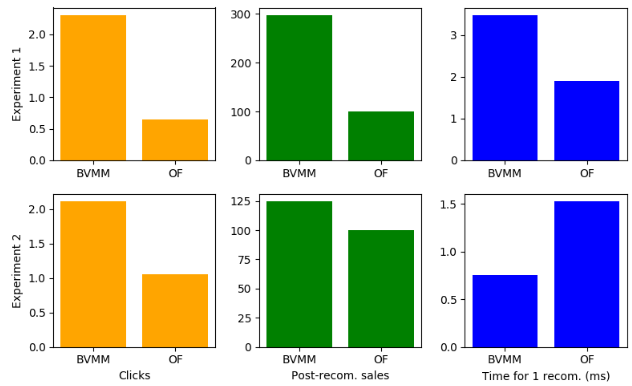

Figure 1: CTR, post-recom. revenue and time.

3 CONCLUSIONS

Our experiments confirm that taking into account the *sequence* of preferences strongly enhances the predictive power of user-behaviour modeling [4]. Markov-chain models such as the BVMM are ideally suited to model the sequential nature of user-preferences, and they can be efficiently implemented using context-trees.

Our implementation relies on an initial step of candidate preselection. Since the time taken to collect the candidates does not depend on the total number of items, and since the ontologies are created automatically through an offline procedure, the efficiency of the method does not depend on the size of the catalogue.

Our evaluations were carried out through live A/B tests on operational e-commerce sites. They are thus free of the biases that often distort the results obtained using static datasets. Moreover, they were realized with the normal settings and infrastructure that Prediggo, or similar-sized companies, would normally use to serve their customers. The results presented above can hence be taken as representative of the normal operation of an online business, and demonstrate that BVMM models are suitable to be applied in practical recommendation systems.

ACKNOWLEDGMENTS

This work was supported by a grant from the Swiss Commission for Technology and Innovation CTI/KTI under contract 18384.1 PFES-ES.

REFERENCES

[1] Ron Begleiter, Ran El-Yaniv, and Golan Yona. 2004. On Prediction Using Variable Order Markov Models. *J. Artif. Int. Res.* 22, 1 (Dec. 2004), 385–421. http://dl.acm.org/citation.cfm?id=1622487.1622499
[2] Christos Dimitrakakis. 2010. Bayesian variable order Markov models. In *JMLR*.
[3] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized News Recommendation with Context Trees. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/2507157.2507166
[4] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *CoRR* abs/1802.08452 (2018). arXiv:1802.08452 http://arxiv.org/abs/1802.08452
[5] Vincent Schickel-Zuber. 2007. *Ontology Filtering*. PhD. Thesis No 3934. Swiss Federal Institute of Technology (EPFL), Lausanne (Switzerland).
[6] Vincent Schickel-Zuber and Boi Faltings. 2006. Inferring User's Preferences using Ontologies. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*. AAAI press, 1413–1416.