NewsREEL Multimedia at MediaEval 2018: News Recommendation with Image and Text Content

Andreas Lommatzsch, Benjamin Kille
TU Berlin, Berlin, Germany
{firstname.lastname}@dai-labor.de

Frank Hopfgartner
University of Sheffield
Sheffield, UK
f.hopfgartner@sheffield.ac.uk

Leif Ramming
plista GmbH
Berlin, Germany
leif.ramming@plista.com

ABSTRACT

NewsREEL Multimedia premiers 2018 as part of the MediaEval Benchmarking Initiative. The NewsREEL task combines recommendation algorithms with image and text analysis. Participants must predict the popularity of news items based on text snippets and annotated images. Several major German news portals have supplied data. The algorithms are evaluated in terms of *Precision* and *Average Precision* on unknown data. This paper describes the task and the provided data in detail and explains the applied evaluation approach.

KEYWORDS

Multimedia, News, Recommender Systems

1 INTRODUCTION

Recommender systems help users to find the most interesting items in huge sets of available items [7, 12]. Traditionally, recommender systems focus on Collaborative Filtering (CF), which makes use of users sharing similar tastes [9]. CF-based approaches rely on users being traceable with a user ID and on the possibility to collect enough user feedback or interaction data. If the majority of users browse news anonymously, and if items have short lifecycles and receive few interactions, the resulting "cold start" issue impedes Collaborative Filtering. Empirically, a majority of users browse anonymously. Besides, a majority of items draws attentions for a limited time [5]. As a result, publishers struggle to apply collaborative filtering in their news recommender systems. Content-based recommendation approaches offer an alternative way to address the problem [11]. Usually, news articles come in the form of text accompanied by an image. Both affect readers' perception. NewsREEL Multimedia tasks participants to predict items' popularity based on text snippets and image features.

The remainder of this paper is structured as follows: Section 2 describes the NewsREEL Multimedia task in detail. Section 3 introduces the provided dataset. Section 4 discusses the evaluation procedure. Section 5 outlines the experimental setup. Section 6 presents evaluation results obtained by applying baseline methods to the experiment. Finally, Section 7 concludes the paper.

2 TASK DESCRIPTION

NewsREEL Multimedia tasks participants to predict news items' popularity from texts and images. The task dataset comprises news articles collected by several German publishers over the course of thirteen weeks. The task focuses on non-personalized recommendation. We measure popularity by counting the number of visits for each individual article. In other words, participants must compute which articles receive the most impressions. The data include textual features, such as headline and text snippet, visual features extracted from images, and interaction features, derived from web server logs, such as the number of impressions. Participants receive all item-related data for the entire thirteen weeks. The training set covers item access data of the weeks 0–2 and 6–8. Participants must predict items' popularity for weeks 4, 10, 11, and 12 (evaluation set). The popularity data for the weeks 3, 5, and 9 have been excluded to prevent extrapolation of time series. Information concerning the most recent news trends would allow participants to focus their attention entirely on the impression statistics. Instead, participants should focus on image and text content. The task's goal is to develop methods to estimate the popularity of newly published articles for which previous impressions remain unavailable. Item IDs and features are available for all weeks. Participants must predict the most popular items for the evaluation weeks as well as the number of impressions for the most popular news items.

3 DATA DESCRIPTION

The dataset covers thirteen weeks of four selected publishers, who publish predominantly German articles. We encounter 51 289 images displayed alongside articles during this period. The images distribute unequally with one publisher accounting for 42 003 images. In addition, we provide a total of 1 691 unique labels automatically assigned to images by seven annotators trained on *ImageNet* [4]. The dataset amounts to approximately 8.6 GB in size. We observe a total of about 153 million impressions, 397 million recommendations, and 1.1 million clicks.

The dataset includes the following features for each item:

- item data (ID, URL, image URL, timestamp of publication)
- text features (headline, snippet; in German)
- image features (up to ten labels per image and a weighting, activation weights of a standard deep learning network encoding the image). The images have been annotated by means of different frameworks (*Keras* [2], *TensorFlow* [1] and existing, pre-trained models (VGG16, VGG19 [13]).
- items' popularity data (numbers of visits, clicks, recommendations). The image popularity data cover only the training weeks.

In addition to the provided features, participants may compute further features or integrate data from external sources. Corsini and Larson [3] discuss how to apply image feature extraction for

MediaEval '18, 29-31 October 2018, Sophia Antipolis, France

news recommendation. Kille et al. [8] and Gulla et al. [6] describe additional datasets for news recommendation.

The entire dataset has been collected by plista GmbH. Access to the data is subject to a usage agreement with their providers.

4 EVALUATION AND GROUND TRUTH

News recommender systems determine the most relevant articles. For the NewsREEL Multimedia task, we have computed the number of impressions for the items published on selected news portals. We have split the data into a training and a test set. The test set lacks the number of impressions. Task participants must predict the unknown number of impressions for items in selected weeks. We consider the number of impressions as a proxy for relevance. The quality of the predictions is computed by comparing the predictions with the true number of impressions (observed in the test weeks). In the evaluation we consider different metrics.

The **Precision** measures how precisely participants identify the most relevant items. We consider two cut-off points. First, we compute the **Precision@n** to check whether participants manage to identify the most popular items. We analyze $n = 10$ and $n = 10\%$ of the number of items in the bin. Second, we compute the **Average Precision@n** (AP). We define the AP as the mean of the top n precision scores: $\text{AP} = 1/M \sum_{n=1}^{M} \text{Precision@}n$, where M describes the number of elements in the test set. For computing Precision@n we assume the top n items to be the target. In other words, task participants succeed if they manage to identify the most relevant items. We compute the precision metrics for each publisher separately.

Baseline strategies and the observed evaluation results are discussed in the subsequent section. Baseline strategies and their evaluation results are discussed in [10].

5 RUN DESCRIPTION

The data cover thirteen weeks indexed from 0 to 12. Participants receive the content-related features for all items. The interaction-related features, such as impressions and clicks, remain unavailable for the weeks 3 to 5 as well as 9 to 12. Participants must create a predictor using the data from weeks 0 to 2 and estimate the number of impressions for items in week 4. Likewise, they must use the data from weeks 6 to 8 to predict impressions in weeks 10 to 12. We obtain a prediction for each combination of item and week in the specified periods. For each of the weeks we compute three metrics: Precision@10, Precision@Top10%, and AP@Top10%. We average those measurements over the weeks to determine the submission's overall score.

6 EVALUATION

We have implemented three baseline algorithms:

(1) The random recommender shuffles the itemIDs randomly and assigns each item the average number of impressions for an item at that rank as the prediction.

(2) The text similarity-based recommender computes the similarity of each item in the test set with all items in the training set. We employ the cosine similarity on a bag-of-word representation of terms in the articles' text. Subsequently, we compute the weighted average of the impression count of items identified as similar to the target article. We use the similarity score as weight.

(3) The image label-based recommender determines similar items based on the overlap of image labels. Therein, we consider only labels with confidence above thirty percent. We obtain the estimated number of impression through the average number of impressions of similar items weighted by their similarity scores.

Table 1: Baseline Evaluation Results (portal 13554). P@ and AP@ refer to precision and average precision. We cut off the lists at ten items, or ten percent of items.

	Random			Text-based			Image-based		
Week	P@ 10	P@ 10%	AP@ 10%	P@ 10	P@ 10%	AP@ 10%	P@ 10	P@ 10%	AP@ 10%
04	0.00	0.11	0.07	0.40	0.19	0.23	0.40	0.21	0.18
10	0.00	0.10	0.06	0.40	0.19	0.22	0.30	0.24	0.18
11	0.00	0.09	0.04	0.40	0.18	0.21	0.30	0.22	0.16
12	0.00	0.10	0.04	0.40	0.19	0.21	0.30	0.22	0.17
avg.	0.00	0.10	0.05	0.40	0.18	0.21	0.33	0.22	0.17

Table 2: Baseline Evaluation Results (portal 39234). P@ and AP@ refer to precision and average precision. We cut off the lists at ten items, or ten percent of items.

	Random			Text-based			Image-based		
Week	P@ 10	P@ 10%	AP@ 10%	P@ 10	P@ 10%	AP@ 10%	P@ 10	P@ 10%	AP@ 10%
04	0.00	0.08	0.04	0.30	0.13	0.09	0.10	0.13	0.09
10	0.00	0.12	0.07	0.30	0.14	0.09	0.10	0.08	0.05
11	0.00	0.10	0.04	0.30	0.12	0.05	0.00	0.10	0.06
12	0.00	0.11	0.06	0.40	0.11	0.06	0.00	0.10	0.04
avg.	0.00	0.10	0.05	0.33	0.13	0.07	0.05	0.10	0.06

Tables 1 and 2 list evaluation results for two of the publishers. The random recommender achieves a very low precision. The image label-based recommender shows a slightly better precision. The text-based recommender outperforms the image label-based recommender. This indicates that text provides more information than the image label when computing the popularity of news items. Moreover, we observe noticeable differences between the portals. This indicates that the importance of images depends on the specific news portal. In addition, different weeks show a significant variance. This suggests that user behavior shifts between weeks.

7 CONCLUSION

NewsREEL Multimedia is a challenging task combining recommendation with text and image analysis. The task provides a real-world dataset collected by several major German news portals. The evaluation centers on anticipating the most popular articles by their contents. We gauge methods' ability to predict items' popularity in terms of precision and average precision. Three baselines have been introduced allowing participants to evaluate their methods' performance. Details on the developed methods and the obtained results are reported in the workshop working notes of the MediaEval workshop.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[2] F. Chollet et al. Keras. https://keras.io, 2015.

[3] F. Corsini and M. Larson. CLEF NewsREEL 2016: Image-based Recommendation. CLEF 2016: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, 2016.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[5] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 169–176. ACM, 2014.

[6] J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, and X. Su. The adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, pages 1042–1048. ACM, 2017.

[7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[8] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz. The plista dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, pages 16–23. ACM, 2013.

[9] Y. Koren and R. Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

[10] Lommatzsch, Andreas and Kille, Benjamin. Baseline Algorithms for Predicting the Interest in News based on Multimedia-Data. In *Proceedings of the MediaEval Workshop*, 2018.

[11] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[12] F. Ricci, L. Rokach, and B. Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.

[13] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, 2015.