Baseline Algorithms for Predicting the Interest in News based on Multimedia Data

Andreas Lommatzsch
DAI-Labor, TU Berlin, Berlin, Germany
andreas.lommatzsch@dai-labor.de

Benjamin Kille
DAI-Labor, TU Berlin, Berlin, Germany
benjamin.kille@dai-labor.de

ABSTRACT

The analysis of images in the context of recommender systems is a challenging research topic. NewsREEL Multimedia enables researchers to study new algorithms with a large dataset. The dataset comprises news items and the number of impressions as a proxy for interestingness. Each news article comes with textual and image features. This paper presents data characteristics and baseline prediction models. We discuss the performance of these predictors and explain the detected patterns.

KEYWORDS

Multimedia, News, Recommender Systems, Image Analysis

1 INTRODUCTION

The NewsREEL Multimedia tasks supplies participants with different kinds of data. These include low-level features, image labels, and texts. Thus, participants may apply a broad spectrum of machine learning approaches. There is little existing work as NewsREEL Multimedia represents the first task of its kind. The tasks' overview paper [3] presents an outline and detailed description.

In this paper, we study ways to predict the popularity of news items relying on multimedia data. We analyze differences among publishers, especially, how they affect the quality of predictions.

The remainder of this paper is structured as follows: Section 2 analyzes the dataset. Subsequently, we introduce different predictors (Section 3). Section 4 discusses the baseline results. Finally, Section 5 concludes and suggests directions for future research.

2 DATA DESCRIPTION

The dataset covers thirteen weeks of four selected publishers. Three publishers—17614, 13554, and 39234—make up most of the impressions. Fig. 1 illustrates how the number of impressions is distributed. We recognize the downward trend on the log-log plots. This indicates power law distributed quantities. In other words, few articles collect most attention whereas a majority of articles receives little attention. As a result, the predictors must accurately pick the best articles to perform well. The automatic annotators have assigned a frequent subset of labels to articles. For publisher 17614, these include 'stage,' 'suit,' and 'wig.' The dataset provides the labels computed using six different labeler configurations. All annotators rely on ImageNet, which had been trained on publicly available images. The annotators differ with respect to the used frameworks (TensorFlow, Keras) and the applied pre-trained network (VGG16, VGG19, InceptionV3, ResNet50). The task incentivizes participants to find the relation between configuration and performance.

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

3 BASELINES

NewsREEL Multimedia tasks the participants to find the news items which users will read most frequently. The participating teams must predict the number of impressions for each item listed in the test weeks. We introduce three baseline strategies for predicting the number of impressions: random, document-based, and feature-based.

3.1 Random

The *random* baseline assigns each item a random non-negative integer as number of impressions. This random guessing should be the lower bound for all prediction strategies.

3.2 Document-based Approach

The *document-based* approach centers on the notion of document similarity. The algorithm employs the basic concept of the k nearest neighbor classifier [1, Chapter 4.4]. First, we represent each news items as a bag of words. We obtain the words either from the articles' texts or image annotations. Next, we determine the ten most similar news items by means of cosine distances amid their term vectors. The computation exhibits linear complexity in the number of news items. With the NewsREEL Multimedia dataset, the computation took several minutes. Finally, we estimate the number of impressions as the sum of the ten neighbors' impressions.

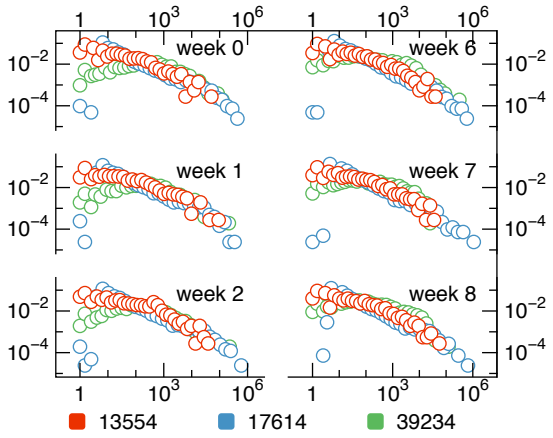

Figure 1: Distribution of Impressions for three publishers in the training set. Publishers have been color-coded according to the legend. The x-axis shows the number of impressions. The y-axis refer to the proportion of articles. Both axes are plotted logarithmically.

3.3 Feature-based Approach

The *feature-based* approach considers features rather than documents. We derive features as terms occurring in the news article as well as labels assigned to images. For each term and label, we compute the average number of impressions of all articles related to them. We estimate the number of impression for a given article by averaging the expected impressions of all its features.

The NewsREEL Multimedia dataset contains further information facilitating variations of this approach. Image labels carry a reference to their annotator's configuration. Thus, the baseline can focus on particular annotators' input or combinations thereof. In addition, each label entails a confidence score. The score indicates how confident the annotator is that the label applies to the image. We can modify the baseline to consider these scores as weights.

4 EVALUATION

We have evaluated the implemented algorithms paying attention to the configurations used to annotate the images. Table 1 shows that the results differ strongly in between domains. The random baseline performs at ≈ 10% for all three publishers. In contrast, the text-based method achieves 34.7% for publisher 13554, 19.2% for publisher 17614, and 22.5% for publisher 39234. The image-based method exhibits noticeable differences as well. While it scores 19.0% for publisher 13554 with configuration 7, it barely exceeds the random baseline for publisher 17614 and 39234. The good performance of image based recommenders for domain 13554 ("cars") compared with the other domains ("world and local news") could be explained by the fact, that articles on 13554 are have a longer lifecycle and are less influenced by breaking news.

Comparing the text-based predictors with the image-based predictors, we find that text feature-based methods on average show a better performance. The approach focusing on selected text features performs significantly better than the text terms based document similarity method. The document similarity method which uses images obtains similar results to the image-based feature methods. For publisher 13554, they score 23.2% with configuration 7, whereas they remain on the random baseline level for the remaining publishers. Specific terms appear to affect items' popularity more than assigned images do. A suitable weighting scheme is of major importance. Comparing word features with image features, the results indicate that the words are more suitable for forecasting the popularity of items than the computed images labels. An analysis of the correlation between image labels and text terms should be conducted. The use of different languages—English for image labels and German for news texts—introduces an additional difficulty.

We analyze the differences between the feature-based and the document-based approaches. On average, the feature-based methods outperform the document-based approaches. This could be explained by considering more robust data (when using features) instead of merely considering the documents most similar to the current news item. Top text terms in domain 13554 (domain cars) are *middle-class*, *unique*, *mar* and *grand*; the top image labels are *snake* (referring to cables), *roof*, and *folding chair*.

Comparing the influence of the image labeler configuration, we find that the labeler 4 based on the InceptionV3 [4] performs worse than the predictors using the VGG [2] component. Analyzing the labels computed by the algorithms, we found, that the labels typically describe selected objects in the image, but are not optimized for interestingness prediction. An additional challenge is raised by example ("stock") images used by the publishers with news items for that no recent photos exist.

Table 1: Prec@10% for the baseline algorithms We will add the missing numbers.

recommender name	labeler config.	domain 13554	17614	39234
doc. similarity using images	2	0.207	0.103	0.110
doc. similarity using images	3	0.223	0.109	0.104
doc. similarity using images	4	0.200	0.114	0.104
doc. similarity using images	5	0.224	0.112	0.104
doc. similarity using images	6	0.227	0.109	0.121
doc. similarity using images	7	0.232	0.109	0.091
doc. similarity using text	-	0.186	0.100	0.137
image feature-based	2	0.159	0.097	0.123
image feature-based	3	0.137	0.099	0.127
image feature-based	4	0.091	0.108	0.113
image feature-based	5	0.108	0.104	0.110
image feature-based	6	0.129	0.109	0.110
image feature-based	7	0.124	0.106	0.096
text feature-based	-	0.347	0.192	0.225
random	-	0.101	0.102	0.102

Overall, the evaluation results between the configurations and domains. The underlying rules should be researched in detail to improve the prediction algorithms and to optimize the parameter configurations.

5 CONCLUSION

In this paper, we have presented several ways to estimate how popular news items will become based on multimedia data. The results suggest that performance strongly depends on the individual publisher. We have observed that text-based features perform better than image-based features. This could be due to terms being more closely linked to the events reported by the articles.

While text-based methods have outperformed the random baseline consistently, image-based approaches merely overcome the random baseline for some publishers. This indicates that news articles' popularity may be disconnected from images for some publishers. Furthermore, we have seen that the quality of image-based recommendations depends on the annotator used to create the labels.

Future Work. We see several ways to extend this research:
(1) Our work has focused exclusively on "high-level" features such as image labels. Low-level features deserve further attention.
(2) In our experiments, we have analyzed annotators' configurations and the token-based methods separately. A weighted combination of both might yield an performance boost for some publishers. For a live recommender the context of the item should be considered as well.
(3) Our feature-based approach linearly combines features. More complex methods—such as neuronal networks or SVMs—should be tested. They could capture the underlying distributions more accurately.

REFERENCES

[1] R. O. Duda, P. E. Hart, D. G. Stork, et al. Pattern classification. 2nd. *Edition. New York*, 55, 2001.

[2] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

[3] A. Lommatzsch, B. Kille, F. Hopfgartner, and L. Ramming. NewsREEL Multimedia at MediaEval 2018: News Recommendation with Image and Text Content. In *Procs. of the MediaEval*, 2018.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.