CLEF NewsREEL 2016: Image-based Recommendation

Francesco Corsini[1] and Martha Larson[12]

[1] Delft University of Technology, Netherlands
[2] Radboud University Nijmegen, Netherlands
corsinifrancesco0@gmail.com, m.a.larson@tudelft.nl

Abstract. Our approach to the CLEF NewsREEL 2016 News Recommendation Evaluation Lab investigates the connection between images and users clicking behavior. Our goal is to gain a better understanding of the contribution of visual representations accompanying images (thumbnails) to the success of news recommendation algorithms as measured by standard metrics. We experiment with visual information, namely Face Detection and Saliency Map, extracted from the images that accompany news items to see if they can be used to chose news items that have a higher chance of being clicked by users. Initial results seems to suggest great CTR improvement in the Simulated Environment task, while some decrease in performance has been found in the Living Lab task. The latter result must be further validated in the future.

Keywords: Recommender System, News, Image Analysis, Face Detection, Saliency Map, Evaluation

1 Introduction

The CLEF NewsREEL [5] News Recommendation Evaluation Lab challenges participants to come up with an original and effective solution for providing recommendations for users in the news environment. Our participation is both for Task 1 (Living Lab Evaluation) and Task 2 (Evaluation in Simulated Environment). An overview of this year challenge results can be found at [9].

Typical online news content providers publish images along with their news items. Our work is motivated by the conjecture that these images play a role in the effect of the recommendation, especially whether a user will click on the item. Content providers are well aware of the importance of images and are already taking advantage of them (e.g., both their informative potential, and their potential to act as clickbait). However, the effect of images for automatic recommendations is currently understudied and not well understood. Our research looks for the effect of such images, in order to determine if they can play a crucial role in the definition of a more refined recommendation. Our hypothesis is that people tend to click on news articles because they are curious about the image, as the image catches their eye, and some images depict things clearly making it very easy to see what the article is actually about. Specifically, in this work, we will focus on the usefulness of information about faces appearing and saliency in images. The Open Recommendation Platform (ORP) by plista provided a unique framework to test and benchmark our approach. Given the constraints of the online environment (100ms timeout response time, unpredictable load on the server), new and innovative architectures and algorithms were developed in order to deal with the heavy computational load caused by the image analysis. Our research also investigates whether features extracted from images can be used in a real-time recommendation pipeline.

The rest of the paper is organized as following: in section 2 we discuss the related work on how to trigger interest on images presented, plus the background needed to understand our approach to image classification. Section 3 describes our approach to solve the challenges presented in Task 1 and 2 and here our algorithm is presented. The outcome of our experiments and the results of the evaluations is presentend in section 4. The discussion 5 follows presenting future work and a wrap up for the conclusion.

2 Related Work and Background

2.1 Grabbing Attention

In this section, we discuss factors that trigger our eyes to land on an image. With content-based image retrieval on the rise, there is an increase in the study of cues that could help in ranking the retrieved images. A sound measure that would help to automatically rank is how interesting people find an image. Much research has been devoted to the study of interestingness on the Internet, especially with Flicker images, e.g., [2]. However, this sort of interestingness is different from what we investigate here. Specifically, it implies some sort of community and social behavior that goes beyond the effect of images merely catching the eye. The presence of this kind of behavior cannot be assumed to be present in news recommendation environment, where the images come from the news provider, rather than being contributed by community members. Flickr's interestingness is based on social parameters linked to the behavior, i.e., according to the uploader's score reputation and ratio between views, favorites and comments. As example, images with a positive connotation (smile, bright), tend to always have a higher level of interestingness in social media.

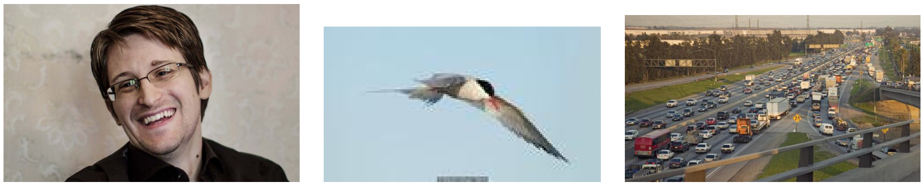
Other related research comes from the area of advertising. An accurate prediction of the probability that users click on ads is crucial for the online advertisement business. Even if with different methods, both our work and ads business share the same goal: predict (and increase) how many clicks an image(or an ad) receives. State-of-the-art click through rate prediction algorithms rely heavily on historical information collected for advertisers, users and publishers. However, recent work has seen the integration of multimedia features extracted from display ads into the click prediction models [1] [3]. The features related to an increase in CTR are numerous. In particular, Cheng et al. [1] present an extensive list of image features and their correlation with CTR. In this study, we focus on key features from [1], chosen because of their promise and their feasibility in being deployed in an online environment. From a study of the literature, we found two of most interesting and investigation-worthy features: the presence of a person [13] [2] [3] [1] [12], especially when having a face clearly visible facing the camera, and the analysis of the saliency map to detect aesthetics and simplicity [1] [3] [4] [11]. However, due to unexpected technical issues during the implementation of these features, only the presence of a person (face detection) was fully developed at the start of the Task 1 challenge. For this reason, it was the only one adopted for consistency throughout all the Task 1 evaluation window. However both features have been tested together in the Task 2 part of the challenge.

2.2 Image Classification

Our approach is based on a straightforward binary image classifier, which classifies the image of the target item (thumbnail) as either “interesting” or “not interesting”. The motivation behind this choice of binary classifiers is the lack of time resources and easy management of the results; a better and more refined approach to the classification (e.g. degrees of interestingness) is planned in future work 5.3. The classification process can be summarized simply as follows: According to our research an image is interesting if it either has:

- The presence of a person: A single central person (portrait) is preferred over multiple people all over the image
- A single cluster in the middle of the image with a flat background. A single object is preferred over multiple objects

As for example, the Fig. 1a and 1b are considered “interesting”, 1a for the presence of a face and 1b for satisfying the single object in the center. While 1c does not satisfy either of the two requirements.



(a) Face (b) Salient (c) Not interesting

Fig. 1

3 Approach

Our approach was designed to validate our hypothesis that images impact user clicks on recommendations rather than to reach the maximum possible CTR.

The Living Lab Evaluation [6] (Task 1) was executed on the ORP, where part of plista's traffic is redirected. The ORP makes it possible to deploy and test algorithms in a real environment. The platform uses HTTP protocol supporting JSON format for data. Communication is handled by four types of messages: Recommendation requests, Impressions, Item Updates, Error Messages. The timeout for the waiting for the response is 100ms: if the system does not answer within this timeframe, the request is considered as an "error"

The Evaluation in Simulated Environment [10] (Task 2) officially makes use of a set of data provided by the NewsREEL organizers. The set includes item updates and event notification [8]. However, this official dataset did not have a crucial field which was required by our image-based algorithm: the img_url. Although the field itself is present, the official data set was collected in June 2013 and the most part of the links have disappeared since the images are hosted by the publishers themselves. Domains tend to remove the items (especially images) after some time of inactivity, by cleaning their databases of old dated articles, as they take much space and do not generate any kind of traffic. Our participation in CLEFNewsREEL using the "official" dataset is, for this reason, compromised. However, this fact did not prevent us from testing our algorithms on another offline dataset. The data used are daily dumps from the plista ORP platform, just like the original dataset with a much more recent date (May 2016).

The algorithms developed and tested are the following:

- Task 1: Baseline1
- Task 1: Baseline1 + Faces
- Task 2: Baseline1
- Task 2: Baseline1 + Faces
- Task 2: Baseline1 + Faces + Salience
- Task 2: Baseline2
- Task 2: Baseline2 + Faces + Salience

Baseline1 is a Popularity with a freshness windows of 100 items, while Baseline2 is Random with the same freshness windows. For the remaining part of the paper these two algorithms will be called Pop100 and Rand100. By looking at the difference between the image enhanced algorithm and the relative baseline we can understand the effectiveness of image-based recommendation in the news environment.

3.1 Algorithm

Although the algorithms deployed in the Living Lab Evaluation (Task 1) differed from the one deployed in the Evaluation in Simulated Environment (Task 2), the logic behind them is quite similar and can be summarized as follows:

A recency windows for each combination of category/domain is created, each window encompassing 100 items. Every time a new update comes in, it is processed by taking the url_img field and scraping the corresponding image from the website. Features for the image are computed with our image processing algorithms, namely Viola-Jones [14] for face detection and spectral residuals [7] for the saliency map. The saliency map involves the extraction of several subfeatures (e.g., number of objects and their positions, background to foreground ratio) which are then used to detect if the image satiefies the requirement of being a single cluster in the middle of the image. This newly processed item is then added to the possible recommendations list, while the oldest item in the list is discarded (if full).

For the Pop100 algorithm: These items are sorted by a popularity score, which is an aggregation of how many impressions the item has received plus how many clicks it received in previous recommendations. Whenever a recommendation request arrives, the top N items are selected and only picked if they individually satisfy the "visual requirements" (see 2.2). If not enough items have been gathered before the top C elements have been considered, then standard popularity is used instead, without taking into consideration the "visual requirements" in order to fill the remaining spots. For the Rand100 algorithm, the logic is the same, however the ranking step is replaced by a random picking of items. The first C random times the item will be picked only if it satisfies the "visual requirements", after C times this restriction decays.

The constant C has been determined from empirical testing, and it can be interpreted as a tradeoff between "being interesting" and "following the baseline". In case of Pop100, the smaller C the most the items will be popular and less "visually interesting". As our intention here is to test if the visual component has an effect, C has been intentionally exaggerated in order to make the effect more notable.

4 Results

4.1 Living Lab Evaluation

The Online results show the data obtained from the scoreboard in the ORP during the evaluation window. Although the evaluation itself ran for around 40 days, not all the days have been taken in consideration due to issues which resulted in the recommender receiving a low volume of requests. As a result, only 24 days have been considered for the results. In order to answer our research question we decided to benchmark our image enhanced algorithm against its own baseline without image information. As for the Online, Pop100 is the baseline.

As can be seen from the Fig. 2: although the image enhanced recommender had more overall clicks, the baseline performed better in CTR value over long period of time. The Pop100+Faces sees a 28% decrease in CTR over the baseline Pop100. Our conclusion is that the lower result is actually due to a mixture

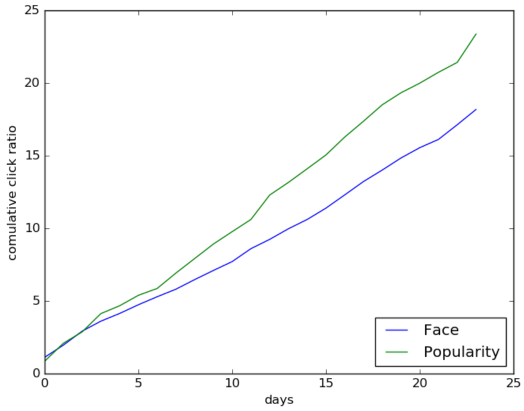


Fig. 2: Online Pop100 vs Pop100+FacesDetection: Cumulative CTR

of technical problems that most likely undermined the performance of the algorithm. A rundown of the problems can be found in the discussion in section 5.1

4.2 Evaluation in Simulated Environment

The Task 2 evaluation was done by using the dataset from ORP daily dumps. Three non-consecutive days have been used as a test set. We consider three days to be the minimum-sized data set large enough to provide a reliable comparison. Each day has an average of 68.000 requests. Since the algorithm running in the Task 1 environment accumulated a total of 175.000 requests over a month, we needed three days to reach approximately the same number of requests to have a comparable size for the dataset. Further testing is planned over a larger dataset in the future. The evaluation metric works as following: A recommendation is a successful hit if the user lands on the recommended page within 10 minutes of navigating the website. In this testing we conducted tests over two different baselines: Rand100 and Pop100. Rand100 was introduced in order to "weaken" the strength of the baseline algorithm in order to better show the effect of the Image features. The results can be seen in the Table 1

Introducing Image-based recommendation leads to a clicks increase of 51% with respect to the baseline Rand100, while the increase is 36% with respect to the Pop100 when considering only faces, 22% with both features.

Table 1: Task 2 Results

Algorithm	Clicks	Requests	CTR
Rand100	258	204456	0.13%
Rand100+Face+Salience	390	202254	0.19%
Pop100	630	204120	0.31%
Pop100+Face	857	203893	0.42%
Pop100+Face+Salience	771	203979	0.38%

5 Discussion

The results from the Task 1 and Task 2 evaluation differ: we think that this may be due to the inherent difference between the testing environments. We discuss this with more details in this section.

5.1 Living Lab Evaluation

The results gathered during the evaluation window of a month suggest that the baseline (Pop100) performs better than the image-based algorithm. This can be partially attributed to the technical problems which the image-based algorithm faced when running online.

One of the problem encountered was to make the algorithm fast enough to keep up with the ORP rate of updates. While the requests sent by the platform do follow the performance of the algorithm (if the algorithm is struggling less requests are sent), this does not apply to the updates; therefore all the updates are sent at anytime. Updates are the "computationally intensive" part in our algorithm, as each update usually comes with an image that needs to be downloaded and analyzed. Updates tend to come in groups of 10 or more, making it necessary to queue them. Even when trying to solve the matter with various strategies, it sometimes happened that the next batch of updates came before the queue was all processed, making the queue longer and the processing time even longer, thus making the problem worse: if repeated enough times the server would crash and get rebooted, therefore going through a new cold start period. Longer queue and longer processing time meant longer delay to answer recommendation requests as well, thus failing due to the timeout time. The time resources available for this research were necessary limited and not all solutions to this problem have been explored.

5.2 Evaluation in Simulated Environment

The Evaluation method used in this task does make the CTR quite worse than the one obtained in Task 1, as there is no actual user answering directly to the recommendation shown. Therefore no direct CTR comparison can be made.

However the difference between the baseline and the baseline+visual information can be used to infer the effect of such features.

For both baselines Rand100 and Pop100 we can see a significant improvement of the CTR when we make use of the Image information. As expected the increase is bigger in the "weaker" baseline, Rand100. However the most striking difference is the improved performance over the Pop100, especially when compared with the results of the similar experiment conducted Task 1. This strengthens our idea that the Task 1 implementations results were jeopardized by the poor technical performance rather than the Image-based recommendation model.

5.3 Future Work

The algorithm and the approach developed during this challenge was intended to be an exploratory task. Much is still needed to indeed prove the real effect of images on the recommendation.

Both Task 1 and Task 2 testing needs to be continued on all the possible combinations of baselines and features used in this paper, in order to test both the single effect of the features independently and their strength against different baselines. This is especially needed in order to investigate further the difference between Task 1 and Task 2, especially in light of the results obtained in this paper. A larger dataset (including images) needs to be used for testing in Task 2. This is our aim in the forthcoming future. Improvement in efficiency and running times are needed in order to allow the algorithm to properly work in an Living Lab environment. The current implementation has many flaws that likely resulted in many delays and worse CTR. A possible approach could be to not compute images until they reach a minimum level of popularity: this would filter out many “socially uninteresting” images.

Although this paper has focused its attention on the exploitation of high level visual clues (people, saliency map), a more in depth analysis of other feature classes may reveal useful insights. Notable global features include colorfulness, brightness and saturation. Another interesting approach could be the inclusion of visual information of how and where the recommendation is displayed (website related features). All of this on top of a more refined approach to the classification, by introducing different degrees of interestingness in the process.

5.4 Conclusion

Task 1 and Task 2 results seems to contradict each other at the first look. Task 2 shows an increase of the recommender performance while Task 1 shows a decrease. We can partially explain the difference by the fact that early Task 1 implementation ran in technical difficulties typical of the online environment, which partially jeopardized the final outcome.

By looking at the Task 2 results we can clearly see an improvement of the CTR when introducing image-based recommendations. This initial result seems to suggest a great improvement even when combined with already strong baselines (Popularity/Recency). More experiments with different baseline combinations and settings are required in the future to definitively prove the effectiveness of image-based recommendation in the news environment. We think that the results shown in this paper provide a good initial confirmation of its potential.

References

1. Haibin Cheng, Roelof Van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. Multimedia Features for Click Prediction of New Ads in Display Advertising. In *18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–785, 2012.
2. Sagnik Dhar, Tamara L. Berg, Stony Brook, Vicente Ordonez, and Tamara L. Berg. High level describable attributes for predicting aesthetics and interestingness. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1657–1664, 2011.
3. Xiaoli Fern. The Impact of Visual Appearance on User Response in Online Display Advertising. *Proceedings of the 21st international conference companion on World Wide Web*, pages 457–458, 2012.
4. M Gygli, H Grabner, H Riemenschneider, F Nater, and L Van Gool. The Interestingness of Images. *Computer Vision (ICCV), 2013 IEEE International Conference on*, (iii):1633–1640, 2013.
5. Frank Hopfgartner, Torben Brodt, Jonas Seiler, Benjamin Kille, Andreas Lommatzsch, Martha Larson, Roberto Turrin, and András Serény. Benchmarking news recommendations: The CLEF NewsREEL use case. *SIGIR Forum*, 49(2):129–136, January 2016.
6. Frank Hopfgartner, Benjamin Kille, Andreas Lommatzsch, Till Plumbaum, Torben Brodt, and Tobias Heintz. *Benchmarking News Recommendations in a Living Lab*, pages 250–267. Springer International Publishing, 2014.
7. Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (800):1–8, 2007.
8. Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista Dataset. In *2013 International News Recommender Systems Workshop and Challenge*, pages 16–23, 2013.
9. Benjamin Kille, Andreas Lommatzsch, Gebrekirstos Gebremeskel, Frank Hopfgartner, Martha Larson, Jonas Seiler, Davide Malagoli, Andras Sereny, Torben Brodt, and Arjen de Vries. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In Norbert Fuhr, Paulo Quaresma, Birger Larsen, Teresa Goncalves, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Evora, Portugal, September 5-8, 2016.* Springer, 2016.
10. Benjamin Kille, Andreas Lommatzsch, Roberto Turrin, András Serény, Martha Larson, Torben Brodt, Jonas Seiler, and Frank Hopfgartner. *Stream-Based Recommendations: Online and Offline Evaluation as a Service*, pages 497–517. Springer International Publishing, 2015.

11. Judith A . Redi and Isabel Povoa. The Role of Visual Attention in the Aesthetic Appeal of Comsumer Images: a Preliminary Study. In *Visual Communications and Image Processing (VCIP)*. Intelligent Systems, Delft University of Technology, The Netherlands, 2013.
12. Paola Ricciardelli, Cristina Iani, Luisa Lugli, Antonello Pellicano, and Roberto Nicoletti. Gaze direction and facial expressions exert combined but different effects on attentional resources. *Cognition and Emotion*, 26(6):1134–1142, 2012.
13. Andreas E. Savakis, Stephen P. Etz, and Alexander C. P. Loui. Evaluation of image appeal in consumer photography. *Proc. SPIE 3959*, 3959:111–120, 2000.
14. P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR)*, 1:I—-511—-I—-518, 2001.