PictureSort: Gamification of Image Ranking

Mathias Lux
Klagenfurt University
Klagenfurt, Austria
mlux@itec.aau.at

Mario Guggenberger
Klagenfurt University
Klagenfurt, Austria
mg@itec.aau.at

Michael Riegler
Klagenfurt University
Klagenfurt, Austria
m.a.riegler@tudelft.nl

ABSTRACT

Human computation is a very powerful tool for solving tasks that cannot be solved by computers efficiently. One such problem is ranking images upon their relevance for a semantic query or upon how well they depict a semantic concept. In this paper we investigate a method to leverage human computation in a divide-and-conquer approach to create precise ranking models. We discuss the basic technique, our prototype client, its adoption to a gamification approach, and present the results of a study with the prototype. Results from the study indicate that with our method the ranking aggregated from the user input converges fast to an optimal ranking.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation (e.g., HCI**]: [Miscellaneous]; K.8.0 [**Personal Computing**]: [Games]

Keywords

gamification, image ranking, information retrieval

1. INTRODUCTION

A common approach for search engines is to present a ranked lists of relevant results to users. Relevance is typically judged by a scoring function s, which quantizes the difference of a query representation q to each of the document representations d_i from a corpus $D = \{d_i\}$ with $0 \leq i < |D|$. Especially in multimedia, this relevance function and the document representations are of critical importance for successful retrieval. Different global and local features used for content based image retrieval lead to very different results. Therefore, for each use case or domain, the best combination of scoring function and document representation has to be found. Moreover, relevance is not a crisp concept and many users have problems to express their information need in terms of the query language supported by a search engine.

GamifIR '14 April 13 2014, Amsterdam, Netherlands
ACM 978-1-4503-2892-0/14/04...$15.00.
http://dx.doi.org/10.1145/2594776.2594789.

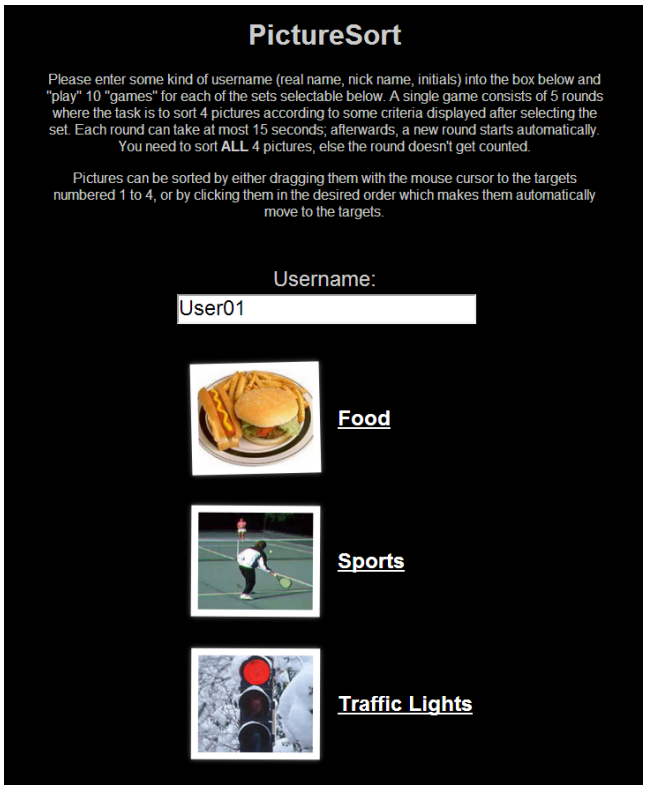

Figure 1: Start screen of the PictureSort game.

One way to involve the user to find the best matching document representation and scoring function is to use test data sets with predefined topics, like it is common for MediaEval [6] or TRECVID [2]. Based on the topics, search engine developers can try and select the best performing methods, i.e., those that lead to the highest precision, recall, mean average precision or any other appropriate retrieval performance measure. However, creating test data sets and topics for each and every domain and use case can be a very tedious task, and that is what our submission focuses on. Instead of asking experts to create topics that include an exemplary query and a ranked list of results, we employ human computation and gamification to create ranked lists of results. In our proposed interface we give a query and ask people to rank a small subset of images, i.e., four images, instead of the whole list at once. While the idea is simple and is based on a divide and conquer scheme for sorting things, we can show in an evaluation that the overall ranking of images

stabilizes fast, i.e., the minimum number of subset rankings is surprisingly low. The query itself can be keywords, a concept, a single image or even a set of images. This implies that our approach can be adopted to many use cases. Moreover, we discuss the gamification of this approach to employ it in a community of experts. The therefore developed game can be seen in Figure 1.

2. RELATED WORK

Image ranking in information retrieval is a challenging task that has been explored in many different ways. For instance, Siddiquie et al. [7] use a method based on multi attribute queries and Hardoon and Pasupa [4] try to solve the problem by using implicit feedback from the eye movements of the users. The most relevant work, however, is from Janssens [5]. In his work, Janssens uses human computation to rank images based on semantic attributes, so people are asked how well images represent a semantic concept. The conclusion of Janssens' work is basically that human computation can help to solve the ranking problem. More particularly, people can reach a very high consensus in the way how they rank images, i.e., people agree on the ranking of a set of images according to a semantic concept easily. This makes ranking by human computation a promising candidate for our gamification approach. To the best of our knowledge there is not other work which uses gamification in context of image ranking.

3. METHODOLOGY

Our approach, PictureSort, is based on the idea that people are able to rank a small set of images more easily than a large one. Figure 2 shows the interface of our HTML 5 based client application. In each turn people are shown four different images at a time and have to put them with drag and drop mouse gestures into the slots corresponding to their rank. The actual criterion for ranking the images is shown before each round, which consists of five turns. While the task gets simpler by reducing the number of images shown, and related work has already indicated that with such a method agreement over different people can be reached, each of the images has to be shown multiple times, i.e., in several turns and rounds. The actual ranking of the whole set is then aggregated from the rankings of the small sets in a way that every image that has been ranked at least once gets a score assigned, based on the ranking it received in the turns it showed up. If an image has been ranked first, it is scored one point, if it is ranked fourth, it is awarded four points. The sum of the points an image received in all turns is summed up and normalized by the number of turns. It then gives an overall score $s(I_i)$ for the global ranking of the image I_i

$$s(I_i) = |T(I_i)|^{-1} \sum_{t \in T(I_i)} r(I_i, t)$$

where $r(I_i, t)$ is the rank of image I_i in turn t and $T(I_i)$ is the set of turns t the image I_i has been ranked.

In our contribution we focus on the question how fast people can reach an agreement, in other words if the ranking converges fast to a stable state, or not. Fast, in this case, is defined by a low number of rounds or turns necessary to reach the agreement. From a numerical point of view we investigate Spearman's Footrule Distance over consecutive rounds. This measure is defined as the sum of displacements between two rankings R_1 and R_2 of a set of elements $X = \{x_i\}$, where x_i^j is the ranking of the i-th element in R_j. Spearman's Footrule Distance d_s is expressed as

$$d_s(R_1, R_2) = \sum_{x_i \in X} |x_i^1 - x_i^2|$$

Our research question for the experiments with PictureSort is then easily defined with R_n being the ranking after round n: How large (or small) is n, if n rounds have to be played until $d_s(R_k, R_{k+1}) < m$ for all $k > n$, whereas m is very small minimum distance, or 0? In other words, how many rounds have to be played, so that the rankings between consecutive rounds do not change significantly or not at all for all following rankings?

4. GAMIFICATION

As already introduced, the main task for the user in the game is to sort four images in the correct order. Since the picture sorting part of the game can obviously be boring after a while, long term and replay motivation needed to be considered. Like in common games with a purpose, implementing simple game mechanics helps [8]. In our case we chose a sophisticated scoring and awarding system. During the game the user does not get information about the scores awarded. Scores are only shown after the game is finished. Players get awards in three different main categories:

- **Time**. The players get points based on the time they need for each round and for the whole game (five rounds in a row). The faster a player finishes, the higher the scores he gets.
- **Precision**. If a player agrees with the majority on the ranking, points are awarded.
- **Distance**. Moreover, the spatial distance from the original image position and the rank it ended up is considered (c.p., Figure 2). So if a user has to move a picture farther to the right slot, more points are awarded compared to users moving the image to the right slot just beside.

The precision part is based on how precise the user did the ranking compared to the majority of rankings. This, however, leads to a cold start problem for the first user. Our approach in this case is that an *honest user* plays the first round and therefore initiates a possible global ranking. For each consecutive player two images out of the four pictures shown in each turn are chosen from the set of already ranked images. Scores are awarded only if the relative ranking of the two already ranked images is correct. As the players do not know which of the four images are used for computing the score, players have to be honest. This is the same principle as it is employed in reCAPTCHA [9], the popular automated Turing test.

All in all a player maximizes the score if s/he gives fast and precise answers. In addition to that it is partially a game of chance if the random placement of the images is optimal, i.e., the known images are far away from their actual position.

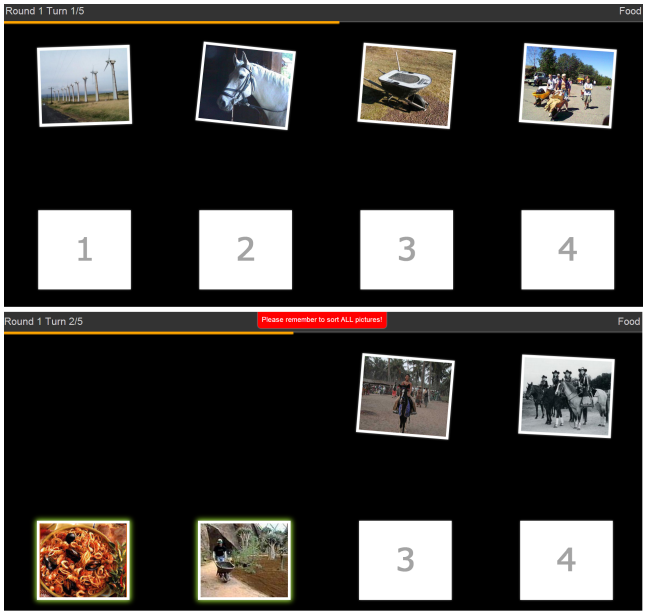

Figure 2: The ranking part of the PictureSort game. The upper picture shows how the start of each round looks. The lower image shows how the user can rank the images.

5. EXPERIMENTS

For our experiments we had a group of ten participants, ranging from students to professors. Each of the users played at least 50 rounds per category. The categories were *food*, *sports* and *traffic lights*. The images of the categories are taken from the Caltech-256 dataset [3]. Each category consists of a set of 50 pictures divided into 25 negatives and 25 positives, which were chosen randomly.

To measure the performance of the users and the system, we used a ground truth (the optimal ranking based on expert votes) for each category. It should be stated that the final system will not use a ground truth. Table 1 shows the results of the experiment for the food category, Table 2 for sports and Table 3 for the traffic lights. Each row shows the precision, precision@10 and the Spearman's Footrule Distance d_s after n games for all users of the current category. The Spearman's Footrule Distance in this case shows the distance of the ranks from the current rank to the next one.

The experiment shows that the stabilization of the ranks is reached very fast. In all three tables a very high precision of 0.80 is reached after at most the fourth run. All of them get a precision above 0.90 after the 13th run. After this peak no important further improvement is reached. This is also indicated by the values of the d_s. At the beginning the distance is very high, but it gets lower very fast, although it never reaches $d_s = 0$. This is somehow clear because at some point the users have, for example, to decide between three images from the same category. This is a very subjective decision which differs between all the users and also differs in our ground truth. For two pictures of a hamburger, it is, e.g., hard to say which one is "more" hamburger than the other one. Because of this, the precision will never reach a value of 1 and the Spearman's Footrule Distancee will never be 0. However, considering precision@10 one can see that the first 10 images are always in the correct categories. The precision@10 for all three categories is always maximized after at most the third run, which is a further indication that the ranking converges fast to near optimal fast.

All in all the experimental results are very promising. They confirmed the outcome of the paper from Janssens [5] that contributors of human computation approaches can indeed reach a very high consensus in their decisions. But more important is the fact that this consensus is reached after a very small number of games in a short period of time. This is an indicator of the strength and possibilities of this information retrieval method, because even a small number of games from a small group of players with a considerably small number of inputs can help to build a ranking model.

The gamification of the task is necessary. Since the task is not entertaining by itself, gamification helps to motivate the users to do it over and over again. The approach can be considered as a crowdsourcing task, and gamification of the task typically leads to a better commitment of workers. We assume that such a game, if it is for example released in Google Play Store, can be a source for a continuous stream of ranked images.

Table 1: Evaluation results for the food category. d_s denotes Spearman's Footrule Distance between the consecutive ranks.

Round	Precision	d_s	Precision@10
1	0.44	436	0.8
2	0.52	370	1
3	0.8	184	1
4	0.8	194	1
5	0.8	120	1
10	0.88	68	1
15	0.96	68	1
30	0.96	28	1
50	0.96	34	1
70	1	16	1

Table 2: Evaluation results for the sports category.

Round	Precision	d_s	Precision@10
1	0.64	250	0.9
2	0.76	152	1
3	0.8	264	1
4	0.88	78	1
5	0.88	62	1
10	0.96	58	1
15	0.96	52	1
30	0.96	32	1
50	0.96	14	1
70	0.96	18	1

6. CONCLUSION

In this paper we presented a gamification approach for letting people build an optimal ranking of images according to a semantic concept. Our experiments indicated that human based ranking of images leads to a very high consensus in a very short time and small amount of played games.

In the future we want to investigate how the rankings created by users can be used in search engines to tune relevance functions and feature fusion approaches, or to select

Table 3: Evaluation results for the traffic lights category.

Round	Precision	d_s	Precision@10
1	0.6	250	0.8
2	0.68	266	0.9
3	0.72	186	1
4	0.8	118	1
5	0.84	168	1
10	0.84	50	1
15	0.88	42	1
30	0.92	24	1
50	0.92	26	1
70	0.96	8	1

content based features for retrieval. Our immediate plan is to include the approach in a medical information system, where only a small number of experts, i.e., highly trained and specialized surgeons, would use such a system to indicate their understanding of relevance on a semantic level. Hence, the high consensus and the fast approach of a near optimal ranking are crucial characteristics for our work. In case of a medical information system the gamification has to be more subtle and, therefore, we plan to integrate scoring badges (vertical achievements) as a more subtle approach for awarding the experts contributing. Still, we believe that even in a professional environment gamification increases the users incentives like described in [1].

7. ACKNOWLEDGEMENTS

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/25557/37319.

8. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, pages 95–106. International World Wide Web Conferences Steering Committee, 2013.

[2] N. Ballas, B. Labbé, A. Shabou, H. Le Borgne, P.-H. Gosselin, M. Redi, B. Merialdo, H. Jégou, J. Delhumeau, R. Vieux, et al. Irim at trecvid 2012: Semantic indexing and instance search. In *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, 2012.

[3] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[4] D. R. Hardoon and K. Pasupa. Image ranking with implicit feedback from eye movements. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 291–298. ACM, 2010.

[5] J. H. Janssens. Ranking images on semantic attributes using human computation. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, 2010.

[6] B. Loni, A. Bozzon, M. Larson, and L. Gottlieb. Crowdsourcing for Multimedia at MediaEval 2013: Challenges, data set, and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[7] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 801–808. IEEE, 2011.

[8] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, Aug. 2008.

[9] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.