

Human Curation and Convnets: Powering Item-to-Item Recommendations on Pinterest

Dmitry Kislyuk¹, Yuchen Liu^{1,2}, David Liu¹, Eric Tzeng^{1,3}, Yushi Jing¹

¹Visual Discovery & Recommendations, Pinterest, Inc.

²University of California, Los Angeles

³University of California, Berkeley

{dkislyuk, yuchen, dliu, etzeng, jing}@pinterest.com

arXiv:1511.04003v1 [cs.CV] 12 Nov 2015

Abstract—This paper presents Pinterest Related Pins, an item-to-item recommendation system that combines collaborative filtering with content-based ranking. We demonstrate that signals derived from *user curation*, the activity of users organizing content, are highly effective when used in conjunction with content-based ranking. This paper also demonstrates the effectiveness of visual features, such as image or object representations learned from convnets, in improving the user engagement rate of our item-to-item recommendation system.

I. INTRODUCTION

Pinterest is an online catalog used to discover and save ideas. Hundreds of millions of users organize Pins around particular topics by saving them to boards. Each of the more than 50 billion Pins saved on Pinterest has an image, resulting a large-scale, hand-curated collection, with a rich set of metadata.

Most Pins’ images are well annotated: when a person bookmarks an image on to a board, a *Pin* is created around an image and a brief text description supplied by the user. When the same Pin is subsequently saved to a new board by a different user, the original Pin gains additional metadata that the new user provides. Therefore, the data structure surrounding each Pin continues to get richer each time the Pin is re-saved. Furthermore, boards (i.e. collections of Pins) reveal relations *between* Pins: if many users save these two Pins together, there is a high likelihood that another user may find them to be related as well. Such aggregated *image co-occurrence* statistics are found to be useful for related content recommendation.

This work explores how user curation signals can be used in conjunction with content-based features to improve recommendation systems. Specifically we introduce Related Pins, an *item-to-item* content recommendation service triggered when a Pin closeup is shown to the user, and describe in detail our experiments using visual features (such as those obtained from convolutional neural networks), which are of particular interest since this system ultimately recommends visual content. As one of the most popular features on Pinterest¹, Related Pins is a recommendation system that combines collaborative filtering [19] with content-based [6] retrieval. Since May 2015,

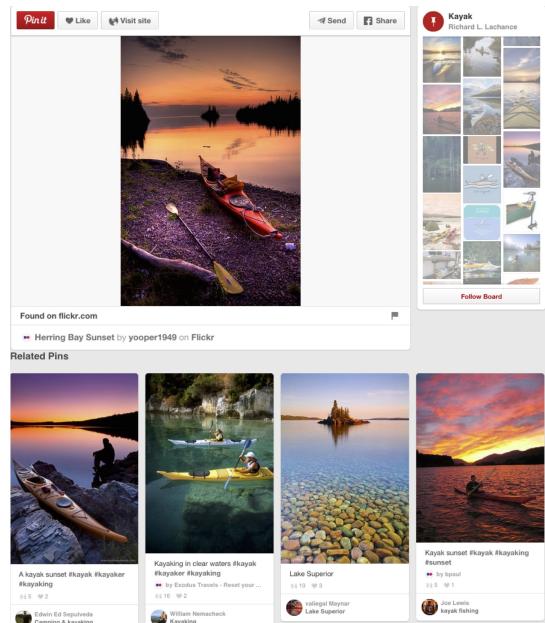


Fig. 1. Related Pins is a product that shows recommendations based on the Pinterest curation graph.

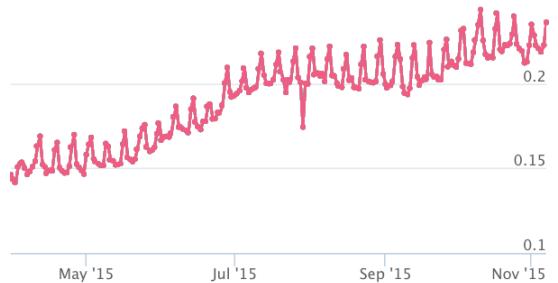


Fig. 2. Since May 2015, the percentage of users engaging with Related Pins recommendations, given they view a Pin closeup, has increased by 50%

the user engagement metric² on Pin recommendations has improved by more than 50%. Note that the improvement is the result of using both visual features and other metadata signals in the learning-to-rank framework—the scope of this paper is limited to the understanding of user curation and visual

¹More than 20% of the page views originate from users clicking on recommended results.

²Defined by the percentage of users clicking off-site or saving recommended content in one day given they see at least one Pin Recommendation.

features in the context of recommendation systems.

This work makes two contributions: first, we demonstrate that “Pinning,” a form of user curation, provides valuable user signals for content recommendation. Specifically we present our use of *image/board co-occurrences*, including the *PinJoin* data structure used to derive this signal. Second, we demonstrate that combining collaborative filtering with content-based retrieval methods, such as applying a learning-to-rank framework [14] to a set of semantic and visual features associated with the candidate Pins, can significantly improve user engagement. In particular, our A/B experiments demonstrate that the use of recently developed visual features (when used in conjunction with other text and graph signals), such as those obtained from VGG [4] and Faster R-CNN [17] yield significant gains in recommendation quality.

II. RELATED WORK

Collaborative filtering [19] using user-generated signals (e.g. co-views, co-clicks) is widely used in commercially deployed recommendation systems such as YouTube related videos [1], [22], [2] and Amazon Related Items [16]. This work investigates the use of *user curation* signals derived from Pins’ image/board co-occurrences, which are unique to Pinterest.

Visual features are widely used in both content-based recommendation systems and image search systems [5], [11], and the learning-to-rank framework used in this paper from [14] has been widely used in industry [8], [9], [12], [18]. To our best knowledge this work contains the first published empirical results on how the latest convolutional neural network (CNN) based visual features (e.g. VGG [4]) and large-scale object detection using Faster R-CNN[17] can improve commercial recommendation systems. Visual features are computed using a distributed process described in our previous work [13].

III. USER CURATION SIGNALS

Content curation is the process of organizing and collecting content relevant to a particular topic of interest. Pinterest is a user-powered content curation service as content is collected and grouped into topic boards, creating a rich set of metadata associated with Pins’ images. For example, during Pin creation, users typically provide a text description of the images as shown in Figure 3. Although any single instance of text description can be noisy, an aggregated collection reveals important annotations relevant to the Pin’s image. Furthermore, when a Pin is saved to a board, one can infer the categorical information of the Pin’s image from the category the user selected for the board.

Formally, we denote the data structures associated with Pins and boards in the following way: each *PinJoin* is a 3-tuple $p = \{u, P, A\}$, where u is the image URL, P is the collection of Pins generated for that image, and A is the aggregation of text annotations or keywords (extracted from Board titles and descriptions). Each *BoardJoin* is represented as a 2-tuple $b = \{t, P\}$, where t is the board title and P is a list of Pins. *PinJoin* is conceptually similar to Visual

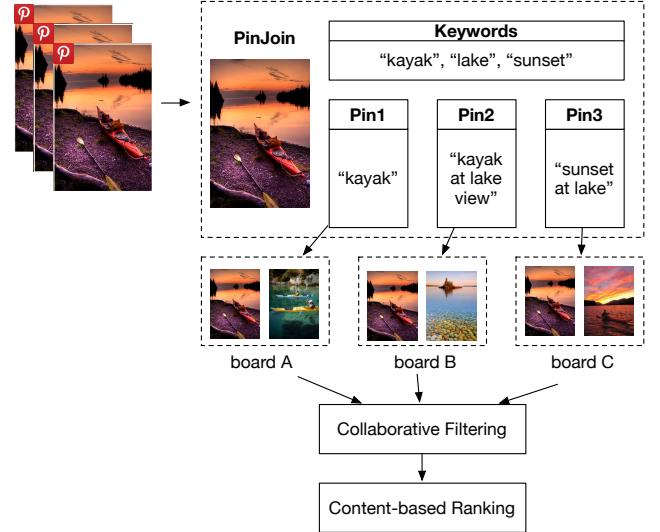


Fig. 3. Pins of the same image are aggregated together to form a *PinJoin*, where rich metadata for the image could be generated based on information from different Pins by different Pinnings.

Synsets [21] except in this case, all the images within a visual synset are exact duplicates of each other and the collection is curated manually. In practice, both of these structures contain additional metadata, such as category and topic information, which is used for deriving other features during re-ranking.

User curation reveals relations *among* images: we observed that images of Pins on the same board are semantically (and to some extent visually) related to each other. Therefore, if enough users save these two Pins together, there is a high likelihood that a new user may also find them to be related. This is helped by the fact that users on Pinterest actively curate content—our engaged users have an average of 24 boards. An example of image/board co-occurrences is shown in Figure 4.

IV. RELATED PIN RECOMMENDATIONS

This section presents the architecture that powers the Pinterest Related Pins Recommendation system. The first step described relies on collaborative filtering over user curation signals (image co-occurrences on boards) to generate a candidate sets. The second step uses content-based ranking approach to rank the candidates based on content signals such as visual features, textual signals, and categories signals derived from *PinJoins*.

A. Candidate Generation from User Curation

The first step of the pipeline is to generate a set of image candidates for each query image, which will serve as candidate sets for content-based re-ranking. We adopt a classic collaborative filtering approach to exploit the Pin/board co-occurrences as described in Section 3. For each Pin, we select up to 10,000 Pins with the highest number of shared boards. In practice, the candidate generation process is accomplished through a MapReduce job, which takes $BoardJoin B = b_1, b_2, \dots, b_n$ as input. The mapping stage outputs image pairs (p_i, p_j) for all image pairs in each board b , and in the reduce stage all



Fig. 4. Examples of candidates with low, medium and high board co-occurrences with the query image. The relevance of the candidate gradually increases with higher co-occurrences with the query image. *Top example:* low: travel destinations, medium: Yosemite viewpoints, high: Half Dome. *Bottom example:* low: animals, medium: dogs, high: golden retrievers.

the related images are grouped by the same query image. For computational efficiency, we sample images based on the quality/popularity of the images. For Pins that do not generate enough candidates through board co-occurrence, we rely on a content-based retrieval service described in our previous paper [13].

B. Content-based Ranking

After generating a set of candidates, we re-rank with a set of both content pair-features (defined between a query and a candidate) and query-independent features. In addition to standard features such as text annotation match and topic vector similarity, we were particularly interested in the effectiveness of visual similarity features for our re-ranking step. Examples of visual features include the *fc6* and *fc8* activations of intermediate layers of deep convolutional neural networks (CNNs) [7] based on AlexNet [15] and VGG [20]. These features are binarized (*fc6*) and sparsified (*fc8*) for representation efficiency and compared using Hamming distance (*fc6*) and cosine similarity (*fc8*), respectively. We use the open-source Caffe [10] framework to perform training and inference of our CNNs on multi-GPU machines. In this work, we also trained an object detection module using Faster R-CNN [17], initially fine-tuned on a dataset containing the most common objects found on Pinterest, including home decor and fashion categories such as various furniture types, shoes, dresses, glasses, bags and more.

To learn the weight vector for our linear model, we adopted the learning-to-rank approach from Joachims [14]. Given training data in the form of relative ranking triplets $(q^{(k)}, d_+^{(k)}, d_-^{(k)})$, where document $d_+^{(k)}$ is considered to be more relevant to query $q^{(k)}$ than document $d_-^{(k)}$, the RankSVM algorithm described in [14] approximates a weight vector \vec{w} which maximizes the number of training examples satisfying $\vec{w}\Phi(q, d_+) > \vec{w}\Phi(q, d_-)$, where $\Phi(q, d) \in \mathbb{R}^m$ gives m features of the document d in the context of query q .

The relevance triplets we use in training are generated through user clicks and impression logs, normalized by position and device to account for position bias, using a clicks over expected clicks [3] (COEC) model. For each query $q^{(k)}$ in our training set, given the set of observed results $D^{(k)}$, we

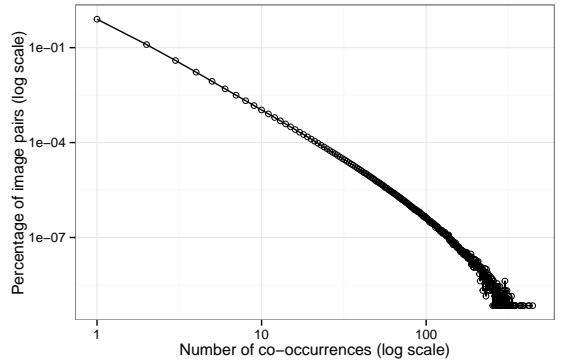


Fig. 5. Percentage of image pairs on Pinterest having different number of board co-occurrences: majority of pairs (80%) co-occur on boards once.

generate the training triplet:

$$q^{(k)}, \text{argmax}_{d \in D^{(k)}} \text{COEC}(d), \text{argmin}_{d \in D^{(k)}} \text{COEC}(d)$$

corresponding to the query, best engaged document, and worst engaged document. We also generate random negative examples:

$$q^{(k)}, \text{argmin}_{d \in D^{(k)}} \text{COEC}(d), d_{\text{RANDOM}}$$

on the intuition that even poorly engaged candidates generated through our board co-occurrence signal should still be more relevant than a random document from our corpus of Pins.

V. EXPERIMENTS

A. Analysis of User Curation Signals

In this subsection we present a qualitative analysis of using user curation signals to generate candidates for Related Pins. As we described previously, board co-occurrences for Pins is a strong signal of relevance and has been the foundation of candidate generation for our system. Figure 4 illustrates that the relevance of the candidates grows gradually when the number of co-occurrences with the query Pin increases.

We also show the percentage of image pairs having different number of board co-occurrences in Figure 5. Note that the majority of the image pairs (around 80%) only co-occur once on the same board, which suggests that ranking based on

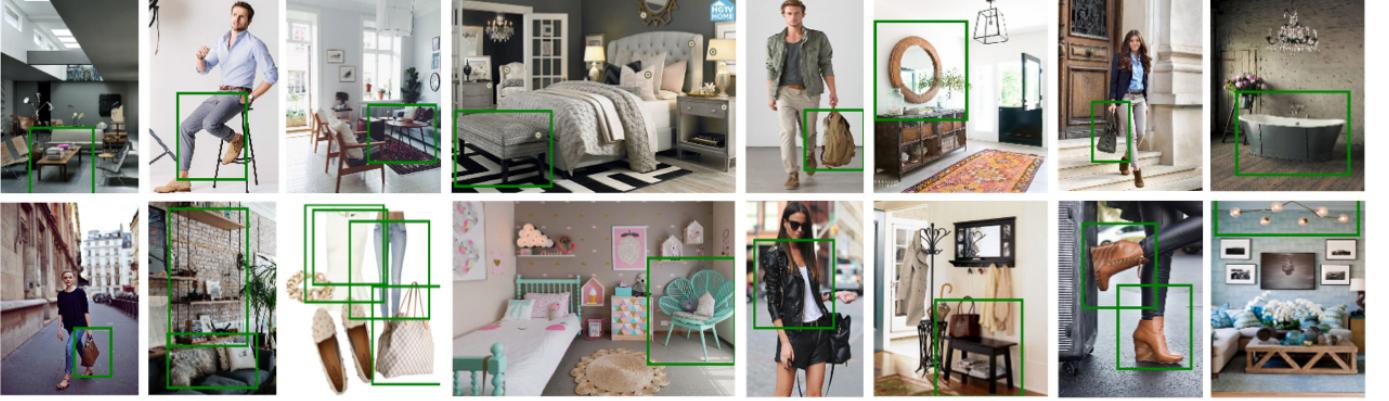


Fig. 6. Examples of detected visual objects from Pinterest’s object detection pipeline. Detection of objects allows for more targeted visual features to be computed for Related Pins.

content features as the next step is important for finding high-quality recommendations. On the other hand, there are only a handful of image pairs which co-occur many times on the same boards.

B. Ranking with Visual Features

We found that one of the most important features for re-ranking the Pin candidates generated through board co-occurrence is visual similarity. We validated this by setting up a series of A/B experiments, where we selected five million popular Pins on Pinterest as queries, and re-ranked their recommendations using different sets of features. The control group re-ranked Related Pins using a linear model with a set of standard features: text annotation similarity, topic vector similarity, category vector similarity, as well as query-independent features. The treatment group re-ranked using fine-tuned VGG $fc6$ and $fc8$ visual similarity features along with indicator variables (in addition to the features used in control).

Across the 5M query Pins, the treatment saw a 3.2% increase in save/clickthrough rate³. After expanding the treatment to 100M query Pins, we observed a net gain of 4.0% in propensity to engage with Related Pins, and subsequently launched this model into production. Similar experiments with a fine-tuned AlexNet model yielded worse results (only 0.8% engagement gain).

When broken down by category, we noted that the engagement gain was stronger in predominantly visual categories, such as art (8.8%), tattoos (8.0%), illustrations (7.9%), and design (7.7%), and lower in categories which primarily rely on text, such as quotes (2.0%) and fitness planning (0.2%). Given the difference in performance among categories, we performed a follow-up experiment where we introduced a cross feature between the category vector of the query and the scalar $fc6$ visual similarity feature (between the query and candidate). This introduces 32 new features to the model, one for each of our site-wide categories (these features are sparse, since the Pinterest category vector thresholds most values to zero). The

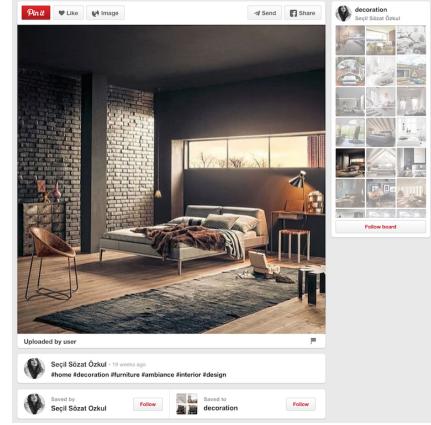


Fig. 7. Instead of generating recommendations for this entire interior design Pin, we may also want recommendations for the individual objects seen in the image, such as the chair on the left.

result from this was a further 1.2% engagement increase in addition to the gains from the initial visual re-ranking model. Further work into component and cross product features is of interest to us, as they are essentially free to compute at rank-time, since the raw feature data is already stored.

C. Ranking with Detected Objects

Users are sometimes interested in the *objects* in the Pin’s image, instead of the full image (as shown in Figure 7). We therefore speculate that object detection, when feasible, should improve relevance targeting. After applying non-maximum suppression (NMS) to the proposals generated by our fine-tuned Faster R-CNN module mentioned in section IV, we considered query Pins where the largest proposal occupies at least 25% of the Pin’s image, or if the proposal is smaller, it passes a confidence threshold of 0.9 in Faster R-CNN. We categorize these images as containing a dominant visual object, and using the best-performing fine-tuned VGG re-ranking variant from the previous section as our control, we experimented with the following treatments:

- *Variant A*: if a dominant visual object is detected in the query Pin, we compute visual features (VGG) on just

³This metric was measured across a 14 day period in Sep. 2015.

that object.

- *Variant B*: same as *variant A*, but we also hand-tune the ranking model by increasing the weight given to visual similarity by a factor of 5. The intuition behind this variant is that when a dominant visual object is present, visual similarity becomes more important for recommendation quality.
- *Variant C*: if a dominant visual object is detected in the query Pin, we still use the features from the entire image (as the control does), but increase the weight given to visual similarity by a factor of 5, as in *variant B*. In this variant, we assume that the presence of detected visual objects such as bags or shoes indicates that visual similarity is more important for this query.

Features	Queries	Engagement
FT-VGG (control)	-	-
FT-VGG + category cross features	5M	+1.2%
FT-VGG + object detection <i>variant A</i>	315k fashion	+0.5%
FT-VGG + object detection <i>variant B</i>	315k fashion	+0.9%
FT-VGG + object detection <i>variant C</i>	315k fashion	+4.9%

TABLE I
RESULTS WHEN USING CROSS FEATURES AND OBJECT DETECTION,
MEASURED OVER A 7 DAY PERIOD IN OCT. 2015

Results for these variants are listed in Table I. Variants A and B of the object detection experiments suggest that the tight bounding boxes from our object detection module do not provide enough context for our CNN models, but Variant C, which results in an additional 4.9% engagement gain over the VGG similarity feature control, demonstrates that the presence of visual objects indicates that visual similarity should be weighed more heavily. Based on these results, our future focus is scaling up the number of object categories we can detect, and tuning the weight given to visual similarity in Variant B and C.

VI. CONCLUSION AND FUTURE WORKS

The Related Pins system described in this work has improved user engagement metric and traffic on Pin recommendations by more than 50% from May 2015 to November 2015. This demonstrates that signals derived from user curation and the activity of users organizing content contain rich information about the images and are very effective when used in conjunction with collaborative filtering. We also demonstrate that visual features such as representations learned from CNNs or presence of detected visual objects can be used in the learning-to-rank framework to improve item-to-item recommendation systems. One important component not discussed in this work is our use of user signal in the form of *Navboost*, which also uses a model based on COEC (extended to actions beyond clicks) to re-rank content based on user engagement. Our future work includes exploring a richer set of features (e.g. sparse features, dense features, cross-product features, more object categories) and real-time recommendations (enabling re-ranking based on locale, current search query, and other forms of personalization).

VII. ACKNOWLEDGMENTS

We would like to thank our colleagues on the Visual Discovery and Recommendations teams at Pinterest, in particular Dmitry Chechik, Yunsong Guo, and many others. We'd also like to acknowledge Jeff Donahue and Trevor Darrell from Berkeley Vision and Learning Center (BVLC) for their collaboration with Pinterest and their work on Caffe.

REFERENCES

- [1] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: Taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 895–904, New York, NY, USA, 2008. ACM.
- [2] M. Bendersky, L. Garcia-Pueyo, J. Harmsen, V. Josifovski, and D. Lepikhin. Up next: Retrieval methods for large scale related video suggestion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 1769–1778, New York, NY, USA, 2014. ACM.
- [3] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 1–10, New York, NY, USA, 2009. ACM.
- [4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.
- [5] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey*, 40(2):5:1–5:60, May 2008.
- [6] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: Approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR '05*, pages 253–262, New York, NY, USA, 2005. ACM.
- [7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.
- [8] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdoch, E. Nyberg, J. M. Prager, N. Schlaefke, and C. A. Welty. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.
- [9] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, pages 1–8. IEEE, 2007.
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [11] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 30(11):1877–1890, 2008.
- [12] Y. Jing, M. Covell, D. Tsai, and J. M. Rehg. Learning query-specific distance functions for large-scale web image search. *IEEE Transactions on Multimedia*, 15(8):2022–2034, 2013.
- [13] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel. Visual search at pinterest. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, pages 1889–1898, New York, NY, USA, 2015. ACM.
- [14] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02*, pages 133–142, New York, NY, USA, 2002. ACM.
- [15] A. Krizhevsky, S. Ilya, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. 2012.
- [16] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan. 2003.
- [17] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.
- [18] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: Machine learning for static ranking. In *Proceedings of the 15th International Conference on World Wide Web, WWW '06*, pages 707–715, New York, NY, USA, 2006. ACM.

- [19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.
- [20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [21] D. Tsai, Y. Jing, Y. Liu, H. A.Rowley, S. Ioffe, and J. M.Rehg. Large-scale image annotation using visual synset. *ICCV*, 2011.
- [22] J. Weston, R. Weiss, and H. Yee. Affinity weighted embedding. In *Proc. International Conference on Machine Learning (ICML)*, pages 1215–1223, Beijing, China, June 2014.