Board Coherence in Pinterest: Non-visual Aspects of a Visual Site

Krishna Y. Kamath
Texas A&M University
College Station, TX 77843
kykamath@cse.tamu.edu

Ana-Maria Popescu
Research Consulting
anamariapopescug@gmail.com

James Caverlee
Texas A&M University
College Station, TX 77843
caverlee@cse.tamu.edu

ABSTRACT

Pinterest is a fast-growing interest network with significant user engagement and monetization potential. This paper explores quality signals for Pinterest boards, in particular the notion of *board coherence*. We find that coherence can be assessed with promising results and we explore its relation to quality signals based on social interaction.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

social media; pinterest; quality analysis

1. INTRODUCTION

The *board-based organization* of a user's pin collection is a core Pinterest feature. Understanding the factors that contribute to board quality can help with board search and recommendation, domain expert mining and more. Such factors include the creation intent and theme of the board. Pinterest boards are assembled with various intents: tracking projects (`wedding planning`), product classes (`Red dresses`), aspirational intent (`House of my dreams`) and more. Although many boards have a central theme, this is not always the case. Fig. 1 shows a board with items of interest to the user, but no unique use case or object type. This paper investigates a quality metric called *coherence* to measure the degree to which a user's board has a central theme. More specifically, we are interested in the following questions (i) How can we discover a board's core topics? (§ 2) (ii) How can we measure a board's coherence? (§ 2) (iii) How can we evaluate the performance of our coherence metrics? (§ 3)

2. BOARD COHERENCE

We consider a Pinterest board *coherent* if it has a small set of discernible themes (topics) and its pins predominantly adhere to these themes. We now delve into two aspects of board coherence: board topic discovery and computing board coherence based on these topics.

Board Topics Discovery: Pinterest offers users a choice of more than 30 category labels – however, only a single category can be assigned to a board (e.g., Weddings). Moreover, users frequently skip the labeling step. In our experience with > 150,000 crawled boards, 35-56% of the boards lacked a category label (depending on whether highly active users or all users were considered). Hence, we explore

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

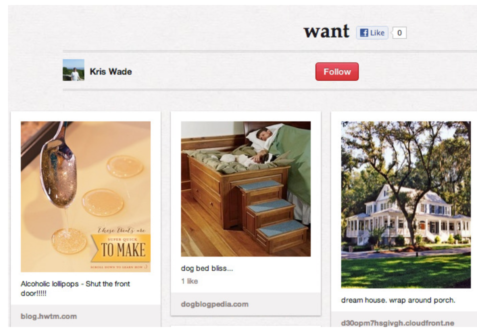


Figure 1: Low-coherence board sample

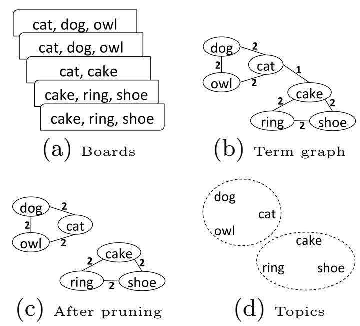


Figure 2: Term graph-based topic discovery

two other means of identifying board topics: (i) Term-graph based topic discovery; and, (ii) LDA-based topic discovery. The *term graph* method first constructs a graph whose nodes are terms in pin descriptions and whose edges correspond to board-level term co-occurrence. Edge weights indicate the number of boards in which the terms co-occur. Fisher's exact test and Pearson's χ^2 test of independence eliminate spurious edges caused by stop words and chance term pairs. The resulting term graph was clustered using affinity propagation [1] to discover topics (see Fig. 2).

Inspired by past work [2], we also experiment with a *LDA-based topic discovery* method [3]. The approach generates one document per board by concatenating the board description, title, and pin descriptions. Topics are then learned from a training set of 25,000 boards (>9 pins each), and the learned model is used to label test boards. Fig. 3 shows the top words for learned LDA topics capturing *dessert*, *drinks*, *lunch/dinner* and *breakfast* terms which can refine the official Food & Drink category. LDA topics can also uncover new user interests suggesting new categories: e.g., multiple *religious* and *spiritual* topics were discovered, though none appear in Pinterest's 30 official category labels.

Figure 3: Topics related to Food & Drink category

Measuring Board Coherence: We now describe our board coherence estimation approach. Let $\mathcal{T} = \{T_1, T_2, \ldots T_{|\mathcal{T}|}\}$ be the set of topics, where T_i is a topic (term cluster). Fig. 2 shows example topics $T_1 = \{\text{cat}, \text{dog}, \text{owl}\}$ and $T_2 = \{\text{cake}, \text{ring}, \text{shoe}\}$. Let P be a pin represented by the set of terms in its descriptions, and B a board represented by the set of terms in all of its pin descriptions, i.e, $B = \bigcup_j P_j$. Given board B and topic set $\mathcal{T}$, we use an *entropy-based* measure to compute the *topical diversity* of a board, which reflects the number of relevant topics and how closely the pins in B adhere to them. A *coherent* board will have low topical diversity, while an *incoherent* one will have high diversity. Let P_i^B be the probability that B has pins from topic T_i:

$$P_i^B = \frac{\{r \in T_i | \forall r \in B\}}{|B|}$$

Let $\mathcal{D}^B_{\text{Graph}}$ and $\mathcal{D}^B_{\text{LDA}}$ denote the topical diversity of a board estimated based on term graph-based topics ($\mathcal{T}_{\text{Graph}}$) and, respectively, inferred LDA topics for a board ($\mathcal{T}^B_{\text{LDA}}$):

$$\mathcal{D}^B_{\text{Graph}} = -\sum_{i=0}^{|\mathcal{T}_{\text{Graph}}|} P_i^B \log_2 P_i^B; \quad \mathcal{D}^B_{\text{LDA}} = -\sum_{i=0}^{|\mathcal{T}^B_{\text{LDA}}|} P_i^B \log_2 P_i^B$$

A 0 value for topic diversity indicates pins from a single topic (e.g., board 1 in Fig. 2 a)); higher values for topic diversity indicate a less coherent board (e.g., board 3 in Fig. 2 a)).

3. EXPERIMENTAL EVALUATION

We start with an initial evaluation of our coherence estimation methods and continue with a larger scale analysis of board coherence.

Coherence Estimation: To evaluate our coherence estimation methods, we use a gold standard set of 401 boards with quality images (spam was removed from a larger starting set of randomly sampled boards). We adopt a broad notion of coherence. A board was labeled as *coherent* if: 1) it had a user assigned category label (64% of the cases) which fit most (i.e. >90%) of its pins; 2) no such label had been assigned (36%), but one was found by the annotators; 3) the board fell outside Pinterest's categorization scheme, but one of Wikipedia's main topic categories (e.g., Religion) fit most of the pins instead. 313 boards (78%) were marked as *coherent* and 88 as *incoherent*. We compared 5 topical diversity estimators: a baseline which assumed all boards are coherent, two term graph-based methods and two LDA methods with different learned topics sets (we exclude worse-performing estimators). Our evaluation measures included:

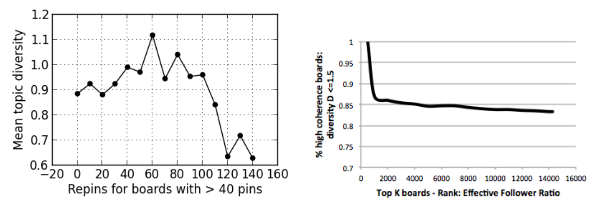
- *Mean diversity difference* ($\mathcal{D}_{\text{diff}}$) defined as the difference between the mean values assigned to incoherent and coherent boards.
- *AUC (area under the ROC curve)* values used to compare SVM classifiers (one per metric) employing topical diversity values for binary board classification.

The results of our evaluation are shown in Table 1. We see that the *LDA-200* metric performs best. Hence, we select this for further analysis.

Coherence Analysis: We further explored coherence using a random sample of 18,998 boards and the *LDA-200* method. We found that *LDA-200* finds core topics (prob. > 0.08) for 14,543 (72%) of the boards. The rest were too sparse (61% of rest had at most 5 pins) or too incoherent; in some cases, topics outside of the learned set were required (e.g., a WWE board). For the 14,543 labeled boards, the median number of topics was 2; about 40% had a single topic while about 90% had at most 3 topics. Overall, we found that the majority of boards were of reasonable coherence.

Estimation Method	$\mathcal{D}_{\text{diff}}$	AUC
Baseline: no topical diversity	0.00	0.50
Term graph - Fisher's exact test	0.20	0.79
Term graph - χ^2 test	0.11	0.69
LDA - 100 topics	0.34	0.71
LDA - 200 topics	**0.44**	**0.81**

Table 1: Topical diversity estimation methods



(a) Topical diversity vs. Repins

(b) Topical diversity vs Effective Follower Ratio

Figure 4: Strongly coherent boards attract higher user interest

Based on mean coherence values, Animals, Travel and Health Fitness were the most coherent categories, while Holidays & Events, Kids, Food & Drink, Weddings and DIY & Crafts were the least coherent.

User Interaction: We also examined the relation between coherence and board-level social actions: *repinning* and *following*. First, we bucketed 1,674 boards with higher pin counts (> 40) based on total repins (bucket size: 10) and computed mean bucket coherence (see Fig. 4(a)). We found that boards with high repin counts had lower topical diversity (i.e., were more coherent). A similar result was obtained for *likes* instead of *repins*. We then defined a follower-based quality signal:

$$\text{Effective follower ratio} = 1 - \frac{\#\text{ users following the board's owner}}{\#\text{ users following the board}}$$

The measure distinguishes between users who follow a board by default (a Pinterest choice when one follows a user) and those who follow that board, but not others from the same user. We found that boards with high effective follower ratios were more likely to be strongly coherent.

Conclusions: We presented an initial investigation of Pinterest *board coherence*; we found that it can be assessed with promising results and that it is related to (but not identical with) board quality signals rooted in social interaction. Our ongoing work is combining these diverse signals for Pinterest board- and user-level analysis.

4. REFERENCES

[1] B. J. J. Frey and D. Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, Feb. 2007.

[2] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[3] X.-H. Phan. GibbsLDA++, 2007.