Mining Potential Domain Expertise in Pinterest

Ana-Maria Popescu[1], Krishna Y. Kamath[2] and James Caverlee[2]

[1] Research Consulting
anamariapopescug@gmail.com
[2] Texas A&M University, College Station TX 77840, USA,
krishna.kamath@gmail.com, caverlee@cse.tamu.edu

Abstract. This paper describes a first investigation of *potential domain expertise* in Pinterest. We introduce measures for characterizing the volume and coherence of Pinterest users' pinning activity in a given category, their perceived and declared category-specific expertise and the response from the social network. We use such signals in the context of a supervised ML framework and report encouraging preliminary results on the task of mining *potential experts* for 4 popular content categories.

Keywords: experts, social network, interest network

1 Introduction

Pinterest is an image-based social platform which has seen rapid growth [16] in 2012. The site allows users to curate image collections (*boards*) as well as interact with other users and their content. Pinterest employs a set of >30 content categories to help in curation, search and discovery; when a board is created, the user can select a category label (e.g., "Home Decor"). The site also showcases category-specific time-sensitive feeds which expose users to the newest content and encourage pinning. Given a category, some pinners have more relevant real-life experience or sustained, deep interest in it than others; their collections can be used for high-quality recommendations and search results.
This paper describes a preliminary investigation of *mining potential experts* for Pinterest categories. First, we describe a set of signals used to capture potential expertise. Second, we report on encouraging initial experiments for identifying highly knowledgeable (*potential expert*) users for 4 popular Pinterest categories. Finally, we outline our ongoing work on the topic.

2 Related Work

Pinterest is starting to attract the attention of the research community: recent studies have focused on generic site analysis [4], gender roles and behaviors [14] and initial content quality measures [9]. Our focus is on identifying top users for given Pinterest categories. Extensive previous work has been done on identifying *global* and *topic-sensitive* authorities in other social networks or QA communities, using a variety of approaches (link analysis, text-based methods, etc.) [2, 1, 5, 15, 12, 7, 10, 11, 3]. We leverage insights from previous research for mining potential experts in a new network with a blend of interest-driven and socially motivated activities.

3 User Features

Given a category c (e.g., Design), we characterize a user u's pinning activity, interest, declared and perceived expertise for c using the features in the top 5 rows of Table 1. To start, we rely on the user-supplied category labels to find category-specific boards. Each final feature reflects a signal of potential expertise (see Table 2). In the following, we describe these features in more detail.

Table 1. User-level feature space for category c and user u. $f_{domExpert}$ relies on two automatically acquired lexicons, LexGenExpert (3.1) and LexCat(c) (3.2).

Final features: f(u,c)	**Description**
$f_{domExpert}$	$f_{catRel} * (f_{genExpert} * w_{gen} + f_{selfProm} * w_{prom})$
f_{vol}	$f_{\%boardsCat} * (f_{numBoards} * w_{ct} + f_{boardSize} * w_s)$
f_{coh}	$f_{\%boardsCat} * (f_{semCoh} * w_d + f_{linkCoh} * w_l)$
$f_{socDirect}$	$f_{\%boardsCat} * (f_{repins} * w_r + f_{cumEFR} * w_{efr})$
f_{socNet}	1 if u is "authority" in repin graph for c; 0 otherwise

Basic features: f(u, c)	**Description**
f_{catRel}	$\alpha_0 * \frac{\vert T(u)\vert \cap LexCat(c)\vert}{\vert T(u)} + \alpha_1 * \frac{\vert T(u)\vert \cap LexCat(c)\vert}{\vert LexCat(c)\vert}$ $T(u)$ = tokens in u's profile description
$f_{genExpert}$	$\alpha_0 * \frac{\vert T(u)\vert \cap LexGenExpert\vert}{T(u)\vert} + \alpha_1 * \frac{\vert T(u) \cap LexGenExpert\vert}{\vert LexGenExpert\vert}$
$f_{selfProm}$	$f_{url} * w_{url} + f_{accts} * w_{ac} + f_{desc} * w_d + f_{urlProm} * w_{up}$
$f_{url}, f_{accts}, f_{desc}$	binary features indicating the presence/absence of a url, Twitter or FB account, populated description field
$f_{urlProm}$	binary feat.: 1 if user pins from URL in profile; 0 otherwise
$f_{\%boardsCat}$	$\vert B_{u,c}\vert / \vert B_u\vert$, B_u = set of u's boards; $B_{u,c}$ =u's boards in cat. c
$f_{numBoards}$	$1 - e^{(-(\alpha * \vert B_{u,c}\vert^2))}$ ($B_{u,c}$ = set of u's boards in cat. c)
$f_{boardSize}$	$1 - e^{(-(\beta * meanBoardSize^2))}$, $meanBoardSize$ = mean. num. pins for all $b \in B_{u,c}$
f_{semCoh}	mean semantic coherence for $B_{u,c}$ board set (based on [9])
$f_{linkCoh}$	$\frac{\sum_{b \in B_{u_c}} linkCoh(b)}{\vert B_{u,c}\vert}$, where $linkCoh(b) = 1 - \frac{\vert uniqueOriginUrls(pins(b))\vert}{\vert pins(b)\vert}$
$f_{repins}; f_{cumEFR}$	$1 - e^{(-(\gamma * f^2_{catRepins}))}$; $1 - e^{(-(\delta * f^2_{cumCatEFR}))}$
$f_{catRepins}$	$\frac{\sum_{b \in B_c} repinsStat(b)}{\vert B_c\vert}$, $repinsStat$ = avg. num. of repins for b's pins
$f_{cumCatEFR}$	$f_{\%boardsCat} * \sum_{b \in B_c} efr(b)$, where $efr(b) = 1 - \frac{followers(b)}{followers(u)}$

Profile Expertise Clues: Users may claim expertise in a category c by using generic expertise terms (e.g., "expert", "maker", "author") in the context of c (e.g., for $c = Food$: "nutrition expert", "cookbook author", etc.). Some include links to their Facebook/Twitter/Instagram accounts, blogs, shops on Etsy and more. Users may also pin from their websites (linked in the profile) in order to

Table 2. Signals for potential expertise in *Design* category (Data: section 4)

Volume f_{vol}	**Coherence** f_{coh}	**Social(direct)** $f_{socDirect}$	**Social(network)** f_{socNet}	**DomExpert** $f_{domExpert}$	**Potential Expert** **Model(M)**
karyna	rodhunt	karyna	zsazsabellagio	111creative	itscloudcuckoo
dilekarisoy	vtloc1989	plentyofcolour	vickiah	itscloudcuckoo	satsukishibuya
1000pin	woodbridgebuild	2dstudio	tristan50	brittanysharp22	luxe
vtloc1989	tinycastlecrtv	designlovefest	stacier	vbroussard	plentyofcolour
beverbal	HighPointMarket	psimadethis	shellytgregory	nyclq	howaboutorange

better "self-promote" [14]. The $f_{domExpert}$ feature seeks to capture these factors and identify users whose profile suggests potential category-specific expertise, knowledge or experience (e.g., for the Design category, the top users are *111creative*, *itscloudcuckoo* or *brittanysharp22*, professional graphic designers with their own design studios).

Volume: Users with significant category pin or board volume are of interest, especially if the relative volume for the target category with respect to others is large. f_{vol} takes these factors into account: for Design, top users based on volume include *karyna* and *dilekarisoy*, active users who focus on design content.

Coherence: We hypothesize that users with significant knowledge of a given category c tend to have better organized, more coherent content than beginners or users with a passing interest in c. f_{coh} reflects the combined semantic and URL-based coherence of the boards in c. *Semantic board coherence* uses the topic diversity measure in [9] while *URL-based coherence* checks if category boards feature content from a focused set of urls. Users with coherent category-specific content are more likely to be professional (as can be seen from the profiles of example users in Table 2).

Direct Social Feedback: The direct response to a user's boards in the form of *repins*, *likes*, *comments* or board-specific *followers* is a good indicator of audience interest and can help find high quality users (see Table 2). We found that *repins* are the most common form of social feedback and correlate with *likes* (comments are sparse). We leverage repins together with board followers who are not followers of the target user (this suggests a strong interest in the particular board).

Global Social Authority: We also make use of global authority measures for a user and a category-specific repin graph by checking if u is among the top $k = 250$ authorities/hubs (we used the NetworkX package [8]).

3.1 LexGenExpert: Category-independent Expertise Terms

The $f_{domExpert}$ feature checks if the generic expertise terms in a mined lexicon (LexGenExpert - see Table 3) are used in profiles in the context of the target category (e.g. "nutrition expert", "founder of a design firm"). Terms are mined as described in the following. For a sample $k = 10$ of Pinterest categories, users are ranked with respect to how consistently their category-specific content is *repinned* (we use $f_{catRepins}$ in Table 1). The top 150 users per category are retained and their profile descriptions are tokenized. Resulting tokens t are scored using $s(t, c)) = \frac{freq(c,t)}{totalFreq(t)}$, a measure of term frequency in the top 150 user descriptions for c versus total frequency in description corpus for all Pinterest users in our dataset. Terms directly reflecting category names (e.g., designer) are automatically removed and added to a category-specific lexicon (3.2). For each category c, the top $n = 50$ terms according to $s(t, c)$ are retained. A final score $exp(t)$ is computed for all $t : exp(t) = \sum_{0<i<=k} topN(c_k, t)$, where $topN(c_k, t)$ is 1 if t is in the top n terms for category c_k and 0 otherwise. The top $m = 50$ terms based on $exp(t)$ are retained as *potentially* indicating category-agnostic expertise.

Table 3. Examples of automatically mined *generic potential expertise* terms.

lover	founder	blogspot	blogger	owner	graphic	vintage
write	market	enthusiast	addict	content	entertain	southern
author	writer	official	lifestyle	director	shop	brand

3.2 LexCat(c): Category-specific Interest Terms

Given a category c, we look for profile terms indicating interest in or expertise for c. We start with the terms ranked based on $s(t, c)$ in 3.1 above and remove LexGenExpert entries and terms with $freq(c, t) = 1$. We also leverage *twellow.com*, a public directory where users ”list” themselves under category labels and which was helpful in other user modeling research [13]. For the $k = 10$ Pinterest categories, we obtain the 200 top users (w.r. to follower count) for related Twellow category labels and tokenize their profile descriptions. Terms are ranked using $s(t, c)$ limited to the available Twitter user profiles; the top 50 terms are retained for each c (e.g., DIY& Crafts: “beading”, “crochet”; Food & Drink: “vegetarian”, “produce”).

4 Experiments: Identifying Potential Experts

In the following, we describe preliminary results for identifying *potential experts* in Pinterest categories.

Dataset: Our dataset contained 12,543 Pinterest users whose boards and pins were crawled in Dec 2012. The median number of boards per user is 16 and the median user follower count is 82. The >12,000 users are a fully-crawled sample of a larger set of Pinterest accounts mined from the feeds “pinterest.com/popular”, ”pinterest.com/everything” and pinterest.com/source” (in combination with example domains - e.g., “tumblr.com”).[3]

Potential Experts: Given a subsample of 400 users and 4 *popular* Pinterest categories (Food/Drink, DIY/Crafts, Home Decor and Design), users were annotated with respect to their experience in and knowledge of each category after inspecting their pins and boards, social media accounts, website, and all other publicly available information. Users with *relevant experience* (e.g., chefs for Food/Drink, interior designers for Home Decor) were considered potential

[3] We also used BFS-style crawling to augment the user set. As a note, April 2013 experiments found Pinterest has become difficult to crawl due to site changes.

experts. Creators of *category-relevant content* recognized publicly (e.g, by means of awards, etc.) were also labeled as potential experts. In addition to this *strict* expertise definition, we used a more *relaxed* criterion: users with experience or recognized contributions in *closely related* categories were also labeled potential experts (e.g., for DIY/Crafts, a graphic designer whose DIY/Crafts boards cover invitation design, etc.). Remaining users were not considered potential experts[4]. Table 4 summarizes the labeling results for the *relaxed* potential expertise annotation - a larger-scale study and more in-depth guidelines are needed before providing general % numbers. As a note, ongoing work shows that other domains (e.g., Travel) have a drastically lower percentage of potential experts.

Table 4. Potential category expertise (*relaxed* version). Dataset: 400 users

Category	**Potential expert** (example)	**Not expert** (example)	**Potential expert** (%)	**Not expert** (%)
DIY & Crafts	vintagerevivals	boulderlocavore	**20.5%**	79.5%
Home Decor	dbohemia	elle_tea	**21.5%**	78.5%
Food & Drink	30aeats	nellieo	**9%**	91%
Design	980ds	1059alexandra	**22.75%**	77.25 %

Table 5. Results: Category-specific models (balanced GS). Notation: $M \setminus F$ = model using all features but F, B_F = baseline using only F

Cat.	**Method**	**Avg. P**	**Avg. R**	**Avg. F1**	**Cat.**	**Method**	**Avg. P**	**Avg. R**	**Avg F_1**
Food	$B_{domExpert}$	96.7	70.2	79.02	**Design**	M	89.5	80.3	83.95
Food	M	87.7	68.8	74.9	**Design**	$M \setminus f_{vol}$	89.3	76.8	81.6
Food	$M \setminus f_{vol}$	89.2	69	74.5	**Design**	$M \setminus f_{socNet}$	88	75.2	80.5
Food	$B_{socDirect}$	97.5	65.5	74.1	**Design**	$M \setminus f_{domExpert}$	86.5	72.5	77.5
Home Decor	$B_{domExpert}$	90.6	64.3	74.4	**DIY/Crafts**	M	71.7	70	68.9
Home Decor	M	81	67.7	72.1	**DIY/Crafts**	$M \setminus f_{vol}$	74.8	62.4	66.1
Home Decor	$M \setminus f_{coh}$	72.4	71.6	71.2	**DIY/Crafts**	$M \setminus f_{socNet}$	76.7	55.9	63.7
Home Decor	$M \setminus f_{vol}$	82.2	62.9	69.4	**DIY/Crafts**	$M \setminus f_{coh}$	75.9	57.7	62.5

4.1 Results

Category-specific models We used the manually labeled data to test if potential experts can be automatically identified. First, each category was targeted separately, with the relevant labeled data subset used for gold standard creation. We experimented with both a *balanced* gold standard set and an *unbalanced* set (entire labeled set). We used Generalized Additive Models (GAM) [6] as our ML framework in a 10-fold cross-validation setting. The full model (M) was compared to models using feature set subsets, including single-feature baselines. We focused on *precision*, *recall* and F_1 values for the *potentialExpert* class (averaged over 10 folds). Tables 5 and 6 show the best performing model versions ranked

[4] Cohen's kappa coefficient for 2 annotators was 0.59, in the "fair-to-good" range. We are devising more specific guidelines for ongoing work

by average F_1. We find that : a) potential experts can be identified with encouraging results; b) *potential expertise* profile clues and *direct social feedback* are particularly useful.

Generic Models We then recast the *potential expert* mining task as follows: given $e(u, c)$, where u is a user and c a category, can $e(u, c)$ can be automatically labeled as “potential expert” or not? A balanced gold standard (482 examples) was derived by combining the category-level balanced gold standards. The unbalanced gold standard corresponded to the entire labeled dataset (1600 examples). Table 7 summarizes the relevant results. The full model M and the model using all but the *volume* information perform best. Among the baselines, the *profile-based expertise clues*, the *coherence* and the *direct social feedback* ones perform best.

Table 6. Results: Category-specific models (unbalanced GS). Notation: $M \backslash F$ = model using all features but F, B_F = baseline using only F

Cat.	Method	Avg P	Avg R	Avg. F_1	Cat.	Method	Avg P	Avg R	Avg. F_1
Food	M	54.2	35.2	39.8	Design	$M \setminus f_{coh}$	64.3	48.04	54.3
Food	$M \setminus f_{socDirect}$	70	29.2	37.7	Design	$M \setminus f_{vol}$	73.1	45.08	53.5
Food	$M \setminus f_{socNet}$	77.7	34.8	37.3	Design	$M \setminus f_{socDirect}$	73.3	44.6	52.9
Food	$M \setminus f_{domExpert}$	66.7	29.3	36.2	Design	$M \setminus f_{socNet}$	75.2	40.8	50.4
Food	$B_{socDirect}$	96.7	22.5	29	Design	M	71.1	39.7	50.2
Home Decor	M	65.4	27.9	32.8	DIY/Crafts	$M \setminus f_{coh}$	60	34.7	38.4
Home Decor	$M \setminus f_{socNet}$	62.3	24.8	32.5	DIY/Crafts	$M \setminus f_{socNet}$	52.2	22.4	30.5
Home Decor	$M \setminus f_{vol}$	53.5	23.9	31.6	DIY/Crafts	M	55	22.3	30
Home Decor	$M \setminus f_{socDirect}$	62.1	24.8	31.5	DIY/Crafts	$M \setminus f_{vol}$	75	23.9	28.1
Home Decor	$M \setminus f_{coh}$	67.7	19	27.6	DIY/Crafts	$M \setminus f_{domExpert}$	54.2	18.8	25.7

Table 7. Results: Generic models. M, $M \setminus f_{vol}$, $M \setminus f_{coh}$ outperform top baselines on Avg. F1 (stat. significance: $*$:$p < 0.05$; $**$:$p < 0.1$).

Dataset	Method	Avg. P	Avg. R	Avg. F_1	Dataset	Method	Avg. P	Avg. R	Avg. F_1
Balanced	M	79.5	65.08	71.04^{*}	Unbalanced	$M \setminus f_{vol}$	70.1	23	34.1^{**}
Balanced	$M \setminus f_{coh}$	80.1	61.3	69^{*}	Unbalanced	M	61.6	23.4	33.7^{**}
Balanced	$B_{domExpert}$	82.7	48.3	60.2	Unbalanced	$B_{domExpert}$	59.5	16.4	25.3
Balanced	B_{coh}	67.6	48.4	55.6	Unbalanced	B_{coh}	54	10.5	17
Balanced	$B_{socDirect}$	84.6	38.4	52.5	Unbalanced	$B_{socDirect}$	72.3	9.3	15.9

4.2 Conclusions and Ongoing Work

This paper presented preliminary results showing that potential experts can be identified for specific Pinterest categories. Our ongoing work is focused on testing on larger gold standard sets and additional categories (both necessary for robust conclusions), improve our models and finally, integrate potential experts and their content in other Pinterest-related tasks.

References

1. Aral, S., Walker, D. Identifying influential and susceptible members of social networks . In: Science, Vol.337 no.6092 pp.337-341, 2012
2. Bakshy, E., Mason, W., Hofman, J., Watts, D.: Everyone's an Influencer: Quantifying Influence on Twitter. In: Proceedings of WSDM 2011
3. Cataldi, M., Mittal, N., Aufaure, M-A.: Estimating Domain-based User Influence in Social Networks. In: Proceedings of SAC, 2013
4. Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: I Need to Try This: A Statistical Overview of Pinterest. In: Proceedings of CHI-2013.
5. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Crowdsourcing search for topic experts in microblogs. In: Proceedings of SIGIR 2012.
6. Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models. In: Chapman & Hall/CRC, 1990
7. Jurczyk, P., Agichtein, E.: Discovering authorities in QA communities by using link analysis. In: Proceedings of CIKM 2007.
8. Hagberg A., Schult D., Swart P.: Exploring network structure, dynamics and function using NetworkX. In: Proceedings of SciPy2008, 2008.
9. Kamath, K., Popescu, A., Caverlee, J. : Board Coherence: Non-visual Aspects of a Visual Site. In: Proceedings of WWW 2013 (Companion Volume).
10. Liu, Lu., Tang, J., Han, J., Jiang, M., Yang, S.: Mining Topic-level influence in heterogeneous networks. In: Proceedings of CIKM 2010.
11. Luiten, M., Kosters, W., Takes, F. Topical influence on Twitter: A feature construction approach. 2012
12. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: Proceedings of WSDM 2011
13. Pennacchiotti, M., Popescu, A.: Democrats, Republicans and Starbucks Afficionados. In: Proceedings of KDD 2011.
14. Ottoni, R., Pesce, J.P., Las Casas, D., Franciscani, G., Kumaruguru, P., Almeida, V.: Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. In: Proceedings of ICWSM 2013
15. Sharma, N.: Discovering topical experts in Twitter social network. In: Master's Thesis, IIT - Kharagpur, 2012.
16. Wikipedia: http://en.wikipedia.org/wiki/Pinterest