Inferring User Interests in the Twitter Social Network

Parantapa Bhattacharya
IIT Kharagpur, India
MPI-SWS, Germany

Muhammad Bilal Zafar
MPI-SWS, Germany

Niloy Ganguly
IIT Kharagpur, India

Saptarshi Ghosh
MPI-SWS, Germany

Krishna P. Gummadi
MPI-SWS, Germany

ABSTRACT

We propose a novel mechanism to infer topics of interest of individual users in the Twitter social network. We observe that in Twitter, a user generally follows *experts* on various topics of her interest in order to acquire information on those topics. We use a methodology based on social annotations (proposed earlier by us) to first deduce the topical expertise of popular Twitter users, and then transitively infer the interests of the users who follow them. This methodology is a sharp departure from the traditional techniques of inferring interests of a user from the tweets that she posts or receives. We show that the topics of interest inferred by the proposed methodology are far superior than the topics extracted by state-of-the-art techniques such as using topic models (Labeled LDA) on tweets. Based upon the proposed methodology, we build a system **Who Likes What**, which can infer the interests of *millions of Twitter users*. To our knowledge, this is the first system that can infer interests for Twitter users at such scale. Hence, this system would be particularly beneficial in developing personalized recommender services over the Twitter platform.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: selection process; H.3.5 [On-line Information Services]: Web-based services

Keywords: Twitter, user interests, Lists, Labeled LDA.

1. INTRODUCTION

The Twitter microblogging site is increasingly being used to discover current information on various topics of one's interest. Many personalized search and recommender systems are being developed to help Twitter users find content that is of interest to them. These systems [3, 6] use a variety of methods ranging from traditional *collaborative filtering* (or variants such as co-factorization machines [5]) to more recent *social recommendations* [2], where the items to be recommended to a user u are drawn from her social network neighborhood. Another approach is *topical recommendations* [7], where a user's topics of interest are first inferred and then leveraged to generate personalized recommendations consisting of items related to those topics. In this paper, we focus on topical recommendations, which have received comparatively less attention in the existing literature.

A key challenge for topical recommendation systems is to accurately infer the topics of interest for an individual user. All prior studies attempted to infer topics of interest from the contents of *tweets*. Most studies inferred interests of user u from the tweets posted by u herself [7, 8], using techniques such as topic models (e.g., Labeled Latent Dirichlet Allocation [8], which is a supervised version of LDA). Some approaches also used the tweets posted by the users whom u follows, i.e., the tweets received by u [2]. However, tweets often contain conversations about mundane daily activities of users [10, 11], making it difficult to identify meaningful topics from tweets.

In this paper, we propose a novel methodology for inferring the topics of interests of a given Twitter user u. Our methodology discovers u's interests from the topical expertise of the users whom u follows (i.e., u's followings). The topical expertise of u's followings are, in turn, inferred from social annotations (collected via the Twitter Lists feature [1]) using a methodology developed in our prior work [4, 10]. We compare the topics of interest inferred for a given user by our methodology with those extracted by a state-of-the-art technique – using supervised topic models (Labeled LDA) on the tweets that are either posted by or received by the given user. We find that topics inferred by our methodology are far superior to the topics inferred from the content of tweets, both in terms of accuracy of the topics, and the completeness with which the interests of a user are identified.

Another important advantage of our methodology is that it can be used to infer interests for any Twitter user who follows a few known topical experts. Using this methodology, we developed and publicly deployed a **Who Likes What** system (`http://twitter-app.mpi-sws.org/who-likes-what/`). We used our system to infer the interests of about *30 million Twitter users*, which is many orders of magnitude more than what prior studies could achieve. Our system – the first one to discover user-interests in Twitter *at scale* – can potentially be a fundamental building block for future designers of content or friend recommendation services over the Twitter platform.

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.

ACM 978-1-4503-2668-1/14/10 ...$15.00.
http://dx.doi.org/10.1145/2645710.2645765.

2. INFERRING USER INTERESTS

In this section, we first propose a novel methodology for inferring the topical interests of users in Twitter and then discuss some alternative state-of-the-art methodologies.

2.1 Proposed methodology

For a given Twitter user u (whose interests are to be inferred), our methodology consists of the following two steps. First, we check which other users u is following, i.e., users from whom u is interested in receiving information. Second, we identify the topics of expertise of those users (whom u is following) to infer u's interests, i.e., the *topics* on which u is interested in receiving information.

Inferring topical expertise using Twitter Lists: Lists are an organizational feature, by which users can group experts on topics that interest them [1]. To create a List, a user specifies a name and an optional description, and then adds other users as members of the List; for instance, a user can create a List named "Music and musicians", and add accounts such as Lady Gaga, Katy Perry, Yahoo Music.

In our prior work [4, 10], we proposed a methodology for utilizing List names and descriptions to discover the topical expertise of popular users in Twitter. To identify topics of expertise of a user v (whom u has subscribed to), we collect the Lists which have v as a member, and extract the most common terms that appear in the names and descriptions of the Lists. We identify v as an expert on a topic t if and only if v has been listed on t at least 10 times, i.e., if the term t appears at least 10 times in the names or descriptions of the Lists containing v.[1] Similar to [4, 10], we considered as topics only unigrams (single words such as 'politics', 'music') and bigrams (two words which frequently occur together, e.g., 'social media', 'video game', 'bay area') which are identified as *nouns* or *adjectives* by a standard part-of-speech tagger. Our prior work [4, 10] has shown that this methodology accurately infers the topics of expertise of millions of popular users in Twitter. For instance, some topics of expertise of the user-account @BarackObama, as inferred by the above methodology, are 'politics', 'celebs', 'government', and so on. Similarly, some topics inferred for the user account @linuxfoundation are 'tech', 'linux', 'software', 'computer', and 'developer' (see [4, 10] for more examples).

Inferring user interests: For the given user u (whose interests are to be inferred), we use the above List-based methodology to identify the topics of expertise of those to whom u has subscribed. Intuitively, if a user subscribes to tweets from several experts on a certain topic, then the user is likely to be interested in that topic. We considered u to be interested in topic t if and only if u subscribes to at least 3 experts on topic t. Thus, we obtain an *interest vector* for u, which is a ranked list of topics, ranked on the basis of the number of experts on a topic whom u subscribes to.

2.2 Alternative methodologies

As stated earlier, prior attempts to infer interests of a given user u rely on the tweets posted by either u herself [7, 8], or by the users whom u follows [2]. Hence, for a given user u, we collected two sets of tweets – (i) **self-tweets:** up to 3,200 most recent tweets *posted by* u herself[2], and (ii) **received-tweets:** the most recent tweets *received by* u, i.e., up to 3,200 most recent tweets posted by the users whom u follows. We pre-process the tweets by removing a standard set of stopwords, URLs and @user mentions. Thereafter, we apply Labeled Latent Dirichlet Allocation (L-LDA) to infer latent topics from the two tweet-sets.

L-LDA [9] is a supervised version of the popular LDA topic model, and has recently been used by several studies to infer topics from tweets [8]. L-LDA requires each document (tweet) to be tagged with zero or more 'labels'. Similar to LDA, each topic inferred by L-LDA is a distribution over the set of distinct terms in the corpus, and each document is a distribution over the topics. However, unlike LDA, each topic discovered by L-LDA maps to one of the labels specified in the input [9].

Following the methodology in [8], we used as the input labels for a given set of tweets, the K most frequent hashtags contained in that tweet-set. For instance, for the self-tweets of user u, the labels are the top K hashtags that are most frequently tweeted by u. Similarly, for the received-tweets of u, the labels are the top K hashtags that were tweeted most frequently by the users whom u follows. Then we use L-LDA to infer K topics for each of the tweet-sets.[3]

3. EVALUATION

We now evaluate the comparative performance of the above three methodologies. Each methodology potentially infers several hundreds of topics for a given user; since it is difficult to evaluate so many topics, we decided to focus on the top 20 topics of interest inferred for a given user.

For the proposed List-based methodology, we consider the top 20 topics in the interest vector for u, i.e., the 20 topics on which u follows most number of topical experts. For the L-LDA-based methods, we use $K = 20$, i.e., we use the top 20 hashtags in a tweet-set as the input labels to L-LDA, so that L-LDA infers 20 topics. Each topic inferred by L-LDA is a probability distribution over the set of distinct terms in the corpus, and the terms within a topic are ranked according to a probability score. We select from each topic the term with the highest probability score assigned by L-LDA, so that we have 20 terms representing the 20 topics inferred for a user.

Thus, for a given user, we have three sets of 20 topics each: (i) **List-topics:** inferred using the proposed methodology, (ii) **self-llda:** inferred using L-LDA on the tweets posted by the user herself, and (iii) **received-llda:** inferred using L-LDA on the set of tweets received by the user.

To evaluate the quality of these topic-sets, one needs to compare the inferred topics with some *ground truth*, i.e., known interests for some specific Twitter users. Since such ground truth is difficult to obtain for random Twitter users, we adopt two strategies for our evaluation, as described in Section 3.1 and Section 3.2 below.

3.1 Using declared interests of users

Methodology: We focus on some well-known Twitter users who have indicated some topics of their interest in their 'bio' (short autobiography written by a user in her account

[1]The threshold 10 is selected based on the observations in our prior studies [4, 10].

[2]One can obtain at most the 3,200 most recent tweets of a user through the Twitter API.

[3]Similar to [8], we set the parameters of L-LDA as $\alpha = 0.167$ and $\beta = 0.001$. The number of topics K was set to 20.

User, with extracts from bio	Top topics of interest, inferred by different methodologies		
	List-topics (proposed)	self-llda	received-llda
Michelle Zhou: into interior design, love shopping & food ...	interior design, decor, fashion, shopping, lifestyle, travel, drinks, hotel	design, system, night, check, gatos, food	#theskenstheory, #fashionstar, win, gardening, daily, love
Erin Marshall: sharing my life in fashion, fun ... Love red lipstick, high heels	fashion, fun, style, fashion designers, beauty	wedding, seatle, cute, today, thanks, love	seattle, obama, #fashion, wedding, thanks, wine
Jesse Millar: Computer sc. student, addicted to programming, game design	developers, game dev, technology, games	awesome, love, last, photo, time, week	games, time, app, thanks, years
Mattia Pontacolone: loves social media, mobile, web apps and start-ups	technology, social media, marketing, startup,	google, facebook, iphone, mayors, love, days	#ted, obama, facebook, thanks, books, new
Cege Smith: Obsessed with The Vampire Diaries, Cult, and Being Human. Author ...	authors, publishing, writing, vampirediaries, writing resources	#ghost, #halloween, #paranormal, #vampires, writers, #amreading	#fantasy, #amwriting, #mystery, #books, #wlcauthor, #horror

Table 1: Examples of top topics of interest inferred by the three methodologies, for some well-known Twitter users who declare their interests in their bio. All topics are case-folded to lower case.

profile). Specifically, we looked for well-known users (researchers, writers, politicians) whose bio contains a phrase such as "like <topic>" or "love <topic>", or some similar phrases. We then check whether the top topics inferred by the various methodologies for these users include the interests declared by the users themselves.[4]

Results: For almost all the users we studied, the topics inferred by the proposed methodology matched the interests declared in the account bio. Table 1 shows the declared interests for some well-known users (as given in their bio) and the top topics of interest inferred by the three methodologies. It is evident that the List-topics (inferred using the proposed methodology) capture a large fraction of the topics of interest stated by the users themselves. For instance, *Michelle Zhou*, one of the General Chairs of the RecSys 2014 conference, mentions in her Twitter bio – "into interior design, love shopping & food". We find that her List-topics include 'interior design' and 'shopping' as well as topics such as 'drinks' and 'hotel' which are closely related to food. Even in the cases where the List-topics do not include the specific interests mentioned by the user (e.g., 'love red lipstick, high heels for user *Erin Marshall*), the List-topics are very relevant broader topics such as 'fashion' and 'style'.

On the other hand, most of the topics inferred using L-LDA on tweets (self-llda and received-llda) are much less relevant to the interests of the users. In fact, several of the terms extracted from the tweets are about recent events (e.g., 'wedding', 'halloween') or are globally popular topics (e.g., 'obama' and 'facebook') which are likely to be posted by the masses.

3.2 Using human feedback

Methodology: Since the interests of a certain user are best known to that user herself, one of the best ways of evaluating the quality of inferred interests is through direct human feedback. We selected 10 volunteers to give us feedback about the quality of the interests inferred for them. The volunteers, all of whom actively use Twitter, are researchers at the home institutions of the authors of this paper (but none of them is an author of this paper herself). Each evaluator was shown the top 20 topics inferred for him / her by the three methodologies. To prevent bias in judgment, the topic-sets were anonymized, i.e., the evaluator was not told which topic-set is from which methodology. Then the evaluator was asked to *rank the three topic-sets* with respect to (i) accuracy, and (ii) completeness of the inferred topics of interest. Note that both *accuracy* (how correct the inferred interests were, analogous to *precision*) and *completeness* (whether a large fraction of the user's interests could be inferred, analogous to *recall*) are important aspects which determine the quality of a set of inferred topics of interest.

[4]Note that only 22% of Twitter users have a bio on their profile, with only 3% having phrases "like <topic>" or "love <topic>". Thus, while this method is good for validating interests, it is not suitable for inferring user interests at scale.

Eval. id	list-topics		self-llda		received-llda	
	Acc	Com	Acc	Com	Acc	Com
1	1	1	2	2	3	3
2	1	1	3	3	2	2
3	1	1	3	3	2	2
4	2	1	1	2	3	3
5	1	1	3	3	2	2
6	1	1	3	3	2	2
7	1	2	2	3	2	1
8	1	1	3	3	2	2
9	1	1	2	2	3	2
10	1	1	2	2	3	3

Table 2: Rankings given by 10 evaluators to the topics inferred for their own Twitter accounts by three methodologies. Rankings are based on two aspects: accuracy (Acc) and completeness (Com) of the inferred topics.

Results: Table 2 summarizes the rankings given by the 10 evaluators to the topic-sets inferred by the three methodologies. *All except one of the evaluators ranked* **list-topics** *as the most accurate.* Between the two LLDA-based methodologies, the *received-llda* set (obtained using L-LDA on the tweets received by a user) was ranked better than self-llda by a majority of the evaluators. The results for completeness were very similar – all except one evaluator ranked **list-topics** as the best, and received-llda was ranked higher than self-llda by most evaluators.

These results clearly indicate that the topics of interest inferred by the proposed methodology are far superior than topics inferred from the contents of tweets. This is because tweets primarily contain day-to-day conversations [10, 11], which makes it difficult to identify meaningful topics even when using state-of-the-art techniques such as L-LDA.

4. THE WHO-LIKES-WHAT SYSTEM: INFERRING USER INTERESTS AT SCALE

Having established the accuracy of the proposed methodology, we used the methodology to develop a system for large-scale discovery of user-interests in Twitter.

Large-scale inference of user-interests: Given the restrictions on using the Twitter API, it is infeasible to crawl data of all 600 million users presently using Twitter. Hence, we started crawling user-accounts in the chronological order of their account creation date, and were able to gather data for about 38.4 million users. Using the proposed methodology, we could infer as many as 36,481 distinct topics of interest, for as many as 29.7 million of these users (i.e., 77.25% of all users whose data we could gather).[5] These numbers show that apart from being accurate, the proposed methodology also has very high *coverage*, i.e., it can be used to infer topics of interest for a large fraction of active Twitter users, which is several orders of magnitude larger than what any of the prior studies of inferring user-interests could achieve.

We have developed a novel Web-based system called **Who Likes What** (deployed at `http://twitter-app.mpi-sws.org/who-likes-what/`), where one can enter the name of any Twitter user, and see word clouds containing the top interests inferred for the given user. We invite readers to use the system for themselves. To our knowledge, this is the first system which can infer user-interests in Twitter at the scale of millions of users.

Differentiating between general and niche topics: During our evaluation using human feedback (see Section 3.2), some evaluators commented that though the top inferred topics accurately captured their broad interests, the topics were sometimes too general. They preferred to see more specific interests, such as 'machine learning' or 'big data' instead of 'science' or 'technology'. We observed that the more specific interests are indeed inferred by the proposed methodology; however, these specific interests were not getting included in the top 20 topics (ranked according to the number of topical experts that a user follows) that were initially shown to the evaluators.

To take this feedback into account, we classified topics into two categories based on their generality, which we estimate by the global number of users who are interested in a topic. Out of the 36 thousand distinct topics inferred (as stated above), we consider the top five percentile of topics as 'general' topics (on which there are thousands of interested users), and the rest of the less popular topics as 'niche' topics. Table 3 shows the general and niche interests for some volunteers who participated in the evaluation.

For a given user, **Who Likes What** shows three word clouds – one showing the top topics considering all inferred interests of the user, the second one showing only the top general topics, and the third showing only the top niche topics of interest. The evaluators were later shown the three word clouds inferred for their accounts, and the *niche* word cloud was unanimously voted as the best by all evaluators.

Eval. id	Top general interests	Top niche interests
1	technology, movies, geek	star trek, programmers, pythonistas, vim
2	technology, politics, science	machine learning, big data, networks
3	football, sports, technology	barcelona, soccer, premier league, la liga
6	football, sports, world, motor sport	man utd, epl, formula1, premier league
10	developers, technology, programmers	haskell, scala, functional programming

Table 3: Examples of general and niche topics of interests for some of the evaluators.

5. CONCLUSION

We developed a novel methodology to infer topics of interests of users in the Twitter social network. Comparison with topics extracted by content-based techniques reveals interesting insights – besides observing that the proposed technique is much superior, we also find that the tweets which a user *receives* are a better indicator of her interests, than the tweets she herself posts. Our findings imply that (i) using social signals can lead to better discovery of user-interests than content-based methods, and (ii) interests are passive traits of users, and inferring them from the user's activity (tweets posted) may be misleading. We also developed a complete system which can infer interests of millions of Twitter users. The potential of such a system is enormous – several topical search / recommender applications can be built using *Who Likes What* as a foundation structure, and our future plan is to develop such systems over the Twitter platform.

Acknowledgment: This research was supported in part by a grant from the Indo-German Max Planck Centre for Computer Science (IMPECS). P. Bhattacharya was supported by a fellowship grant from Tata Consultancy Services.

[5] Of the rest 22.75% users, whose interests we could not infer, 95% follow less than 2 users (with 83% following zero users), and have posted less than 10 tweets in their entire lifetime. Thus, almost all those users for whom our methodology fails to infer interests are *inactive* users.

6. REFERENCES

[1] Twitter help center: how to use Twitter lists. `http://bit.ly/how-to-use-twitter-lists`.

[2] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. H. Chi. Short and tweet: experiments on recommending content from information streams. In *ACM SIGCHI*, 2010.

[3] E. Diaz-Aviles et al. What is Happening Right Now ... That Interests Me?: Online Topic Discovery and Recommendation in Twitter. In *ACM CIKM*, 2012.

[4] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *ACM SIGIR*, 2012.

[5] L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization Machines: Modeling User Interests and Predicting Individual Decisions in Twitter. In *ACM WSDM*, 2013.

[6] Y. Kim and K. Shim. TWITOBI: A Recommendation System for Twitter Using Probabilistic Modeling. In *IEEE ICDM*, 2011.

[7] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on Twitter: a first look. In *ACM Workshop on Analytics for Noisy Unstructured Text Data*, 2010.

[8] R. Ottoni et al. Of Pins and Tweets: Investigating how users behave across image- and text-based social networks. In *AAAI ICWSM*, 2014.

[9] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *EMNLP*, 2009.

[10] N. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi. Inferring Who-is-Who in the Twitter Social Network. In *Workshop on Online Social Networks*, 2012.

[11] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier. It's not in their tweets: modeling topical expertise of Twitter users. In *IEEE SocialCom*, 2012.