

# Posting Topics $\neq$ Reading Topics: On Discovering Posting and Reading Topics in Social Media

Wei Gong<sup>(✉)</sup>, Ee-Peng Lim, and Feida Zhu

School of Information Systems, Singapore Management University,  
Singapore, Singapore

{wei.gong.2011,eplim,fdzhu}@smu.edu.sg

**Abstract.** Social media users make decisions about what content to post and read. As posted content is often visible to others, users are likely to impose self-censorship when deciding what content to post. On the other hand, such a concern may not apply to reading social media content. As a result, the topics of content that a user posted and read can be different and this has major implications to the applications that require personalization. To better determine and profile social media users' topic interests, we conduct a user survey in Twitter. In this survey, participants chose the topics they like to post (posting topics) and the topics they like to read (reading topics). We observe that users' posting topics differ from their reading topics significantly. We find that some topics such as "Religion", "Business" and "Politics" attract much more users to read than to post. With the ground truth data obtained from the survey, we further explore the discovery of users' posting and reading topics separately using features derived from their posted content, received content and social networks.

## 1 Introduction

Social media platforms such as Facebook and Twitter connect millions of users with very large social networks where they create, share and consume content. With regards to content generation and consumption, social media users perform essentially two main types of actions: *posting* and *reading*. Posting is a user action that generates content. For example, tweeting, retweeting and replying are the posting actions in Twitter. Reading, on the other hand, refers to content consumption which often does not generate any public data trace. In social media, some users post often. They are *active users*. Some other users prefer to read content only. When users demonstrate reading as their only actions, they are known as *lurkers* or *silent users* [9, 20, 24].

Users, active or silent, are individuals with topic interests. We call the topics a user likes to post the *posting topics* and the topics a user likes to read the *reading topics*. We postulate that posting topics are not the same as reading topics. This is because, when posting content in social media, users select what

content to post, to whom the content is shared [1,10], and may practise self-censorship when selecting and crafting the content [5,22]. In contrast, reading is typically invisible to others. Users therefore have less worries about how other people perceive them when reading online content. For example, a user interested in politics is likely to read political news and discussions, but may choose not to post political content to avoid unwanted disputes on some controversial issues. In the extreme case, some users may become lurkers who only read but not post.

As discovering user topic interests is important in many applications such as viral marketing, recommendation systems and targeted advertising [6,17,18], a number of studies have focused on predicting users' topic interests [7,19,23,26,28]. While these studies contribute to the discovery of *general* topic interests of users, they do not distinguish between the posting and reading topics. We believe that differentiating user posting and reading topics is important to the above personalization applications. An application (viral marketing, for example) that requires users sharing information (e.g., news, products) with others should focus on the posting topics. A targeted advertising, on the other hand, needs to discover reading topics so as to select the ads that users are likely to pay attention to. To the best of our knowledge, the state of the art research has left out the posting and reading topic consideration, which in turn motivates this work.

**Research Objectives.** Our research aims to answer the following research questions: (a) how different are posting and reading topics? (b) are there topics that are more likely to be reading topics but not posting topics, and vice versa? (c) can we predict posting and reading topics accurately, and finally (d) can we predict lurkers' reading topics accurately?

This paper seeks to answer the above questions by focusing on Twitter platform and formulating two research goals. The first goal is to empirically study the posting and reading topics of Twitter users. In particular, we invest significant efforts in conducting a user survey involving 95 participants who are requested to declare their posting and reading topics. Our analysis of the survey data shows that the topics users like to post can be significantly different from the topics they like to read. We also find that "Politics", "Religion" and "Business" are some topics many users who like to read but not to post.

The second goal of this work seeks to discover user posting and reading topics. This task has two main challenges. First, social media companies may record user browsing history, but often do not make such data available to researchers. The lack of reading behavior data is thus a major challenge for reading topic discovery. Second, lurkers have very little posting behavior, and their reading behavior is also not available. Discovering reading topics for lurkers – who are potential customers and constitute a significant proportion of online users [9] – then becomes another big challenge. To achieve our goal with the limited user behavior data, we make use of users' historical content and following networks so as to develop different ranking strategies to rank user interested topics in posting and reading. We evaluate them using the ground truth data obtained from our survey. We find that predicting user reading topics can be as accurate as predicting user posting topics. We also demonstrate that although predicting

lurkers’ reading topics is harder than that of active users’, we can still predict lurkers’ reading topics with reasonable accuracy.

## 2 Related Work

Posting behavior is a direct way for a user to express herself. Previous studies have shown that social media users *select* what content to post and to whom [1, 5, 10, 22]. For example, Hampton et al. [10] found that people are less willing to discuss a political issue in social media than in person, and people are less likely to express their views online if they believe they have views different from others. Some studies [5, 22] showed that when selecting and crafting the content, users may practise self-censorship. Das and Kramer [5] examined 3.9 million Facebook users and found that 71% of users exercised self-censorship to decide what content to share. Similarly, Sleeper et al. [22] found that Facebook users “thought about sharing but decided not to share”. These studies suggest that users may not disclose their activities, emotions, opinions and topic interests when posting in social media.

Reading behavior refers to user actions that consume content. Previous studies on user behavior have showed that social media users spend much more time reading than posting [2, 25]. Despite its prevalence, reading behavior has not been studied extensively like posting behavior [12]. It is partly due to a lack of publicly available data traces of user browsing history. Compared with posting content, users enjoy a higher level of privacy when reading online content. They can read content and choose not to share or discuss about it [20]. Thus, social media users may show different opinions, personal values, personalities and topic interests when come to posting and reading behaviors. However, earlier studies often analyze social media users by considering their posting behavior only [4, 7, 13, 21, 26], which may yield a biased understanding of the users. For these reasons, we analyze and discover social media users’ topics interests by considering both their posting and reading behaviors.

The closest work of ours is probably [15], which studied the difference between user posting topics and the topics of user received content in Twitter. However, as point out in [11], Twitter users typically receive large number of tweets and are not likely to read them all. Thus in our case, we study the difference between user posting topics and reading topics which are the topics that users actually like to read.

## 3 Posting and Reading Topics

To assess the difference between Twitter user posting and reading topics, and obtain the ground truth for evaluating the methods of discovering user posting and reading topics, we conduct a user survey. In this section, we first describe the procedure of this survey. Next, we analyze the survey data and present the findings.

### 3.1 Survey Procedure

Participants in the survey should have used Twitter for some time and have some social connections. We thus require that all participants have their accounts for at least 3 months and each participant at the point of survey follows at least 10 other accounts and is followed by at least 5 other accounts. We sent a recruitment email to all undergraduate students of a Singapore’s public university. We allowed both lurkers and active users to participate in this survey. We finally obtained survey results from 95 Twitter users including 49 protected accounts and 46 public accounts. Each participant received 10 Singapore dollars incentive for completing the survey.

In the survey, participants provided their Twitter account information<sup>1</sup> and activity data including how often they post (i.e., tweet) and how many tweets they read per day. Participants also rated a set of 23 topics (See Table 1 in Sect. 3.3) based on how much they like to post and read these topics. The possible ratings are “Like”, “Somewhat Like”, and “Do Not Like”. We will describe how this set of topics are derived at the end of this section (see Sect. 3.3).

We then crawled all participants’ tweets from March 1st to March 30th, 2015, their followers and followees using Twitter API. For the participants with public accounts, we can crawl their information directly. To collect the protected accounts’ information, we first created a special Twitter account following the protected accounts and then crawled the protected accounts’ tweets and social networks using the special account.

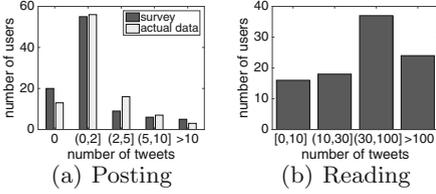
### 3.2 Survey Results and Findings

**Twitter Use.** Figure 1 shows Twitter posting and reading frequency distribution among the participants. In general, these participants read much more than they tweet. To check the reliability of the survey data, we compared the user declared posting frequency with the actual tweet history data from March 1st to March 30th, 2015. Figure 1(a) shows very similar distributions between survey data and tweet history data. It suggests that most of the participants provided information that tally with their actual posting frequencies in Twitter.

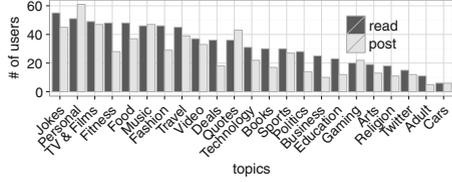
**Difference Between Posting and Reading Topics.** Next, we examine the difference between user posting and reading topics using our survey results. For clarity, we organize this analysis around four questions. The first question is: *What are the popular posting and reading topics?* Fig. 2 plots the posting and reading topics’ popularity among the participants. A posting (reading) topic’s popularity is the number of participants who like to post (read) the topic. We observe that some topics are popular (or unpopular) for both posting and reading. For example, “TV & Films” and “Music” are among the popular topics,

---

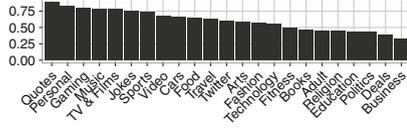
<sup>1</sup> Twitter accounts are considered as personal identifiable information, so we can not use Amazon Mechanical Turk (AMT) for conducting this survey. The restrictions of using AMT: [https://requester.mturk.com/help/faq#restrictions\\_use\\_mturk](https://requester.mturk.com/help/faq#restrictions_use_mturk).



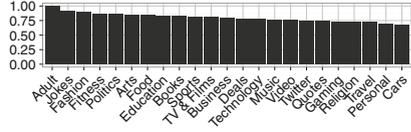
**Fig. 1.** Distribution of posting and reading frequency.



**Fig. 2.** Popularity of posting and reading topics.



(a) Proportion of posting participants among reading participants. ( $P_y^p$ )



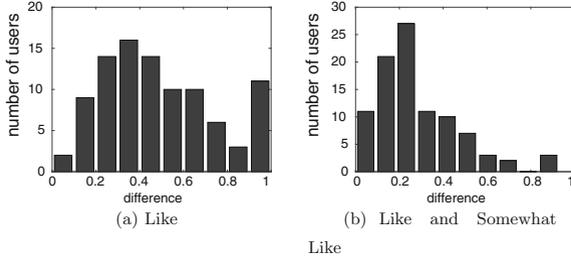
(b) Proportion of reading participants among posting participants. ( $P_y^r$ )

**Fig. 3.** Proportion of users who like to post/read a topic out of those who like to read/post the same topic.

and “Cars” and “Gaming” are among the unpopular topics for both posting and reading. Some topics have significant popularity difference between posting and reading. For example, 48 participants like to read “Fitness” and only 28 participants like to post it. On the other hand, 43 participants like to post “Quotes” and 36 participants like to read it.

The second and third questions are: *Do Twitter users like to post a topic if they like to read it? And do Twitter users like to read a topic if they like to post it?* To answer them, we define the proportion of participants who like to post a topic  $y$  given that they like to read it as  $P_y^p = \frac{|U_y^p \cap U_y^r|}{|U_y^r|}$ , where  $U_y^p$  is the set of participants who like to post topic  $y$ , and  $U_y^r$  is the set of participants who like to read topic  $y$ . Similarly, the proportion of participants who like to read a topic  $y$  given that they like to post it is calculated as  $P_y^r = \frac{|U_y^p \cap U_y^r|}{|U_y^p|}$ . Figure 3(a) and (b) show  $P_y^p$  and  $P_y^r$  respectively for the set of 23 topics.

Figure 3(a) shows that if a user likes to read a topic, on average, she would post it with 0.6 probability as  $avg_y(P_y^p) = 0.6$ . In contrast, the average probability of users liking to read topics which they like to post is significantly higher, with  $avg_y(P_y^r) = 0.8$  (see Fig. 3(b)). In addition,  $P_y^p$  varies largely between topics compared to  $P_y^r$ , as the standard deviations of  $P_y^p$  and  $P_y^r$  are 0.16 and 0.08 respectively. Particularly, only 32% of users who like to read “Business” also like to post it. Similarly, topics such as “Politics” and “Religion” also have low  $P_y^p$  (0.43 and 0.44). Topics such as “Gaming” and “Music” have much higher  $P_y^p$  (0.8 and 0.78). Such topics are more likely to be shared if users like to read them.



**Fig. 4.** Distribution of the differences between user posting and reading topics.

Our fourth question asks: *how different are individual Twitter users’ posting and reading topics?* Suppose a user declares a set of posting topics  $\pi^p$  and a set of reading topics  $\pi^r$ . We compute user posting and reading topic difference as  $d = 1 - \frac{|\pi^p \cap \pi^r|}{|\pi^p \cup \pi^r|}$ , where  $\frac{|\pi^p \cap \pi^r|}{|\pi^p \cup \pi^r|}$  is the Jaccard coefficient of  $\pi^p$  and  $\pi^r$ . Jaccard coefficient is commonly used to measure the similarity of two sets. Hence  $d$  measures the difference between  $\pi^p$  and  $\pi^r$ . Both  $\pi^p$  and  $\pi^r$  can be defined by either topics that are liked with at least the “Like” or “Somewhat Like” rating.

Figure 4(a) shows the distribution of the differences between user’s “Like” posting topics and reading topics. Figure 4(b) shows the distribution of the differences between user’s “Like” and “Somewhat Like” posting topics and reading topics. As the mean differences of 0.5 and 0.28 are significantly larger than 0, we conclude that users have different topic interests in posting and reading.

In summary, we demonstrate that topics are different in attracting users to post and read and that for some topics such as “Business” and “Politics”, only a small proportion of users who like to read them choose to post. We also show that Twitter users’ posting topics are significantly different from their reading topics.

### 3.3 Topics in Tweets

Now, we describe how we obtained the 23 topics to cover all or most of the topics for our participants. We first crawled the tweets generated by a large number of users. We started our crawling process by randomly selecting 434 seed users from Singapore. We then crawled all their followees, who can be based anywhere. In this way, we obtained 93,312 users. Among them, 81,171 users have public accounts. We crawled the latest 200 tweets or whatever available from each public user using Twitter API. Next, we selected the tweets that are posted between Aug 25, 2014 and Nov 25, 2014, discarding tweets that are not written in English, stop words in tweets, and users with less than 10 tweets. Finally, we were left with 50,266 users and their more than 6.2 million tweets.

Next, we adopt T-LDA [28] to learn topics from these tweets. Zhao et al. [28] showed that T-LDA can uncover topics in tweets better than several other

**Table 1.** Topics and some related keywords.

Topics	Some related keywords	Topics	Some related keywords
Arts	Art, Artwork, @fineartamerica	Adult	Adult, Porn, Sex, Pornography
Books	Journal, Book, Poet, Writer, Author	Business	Business, Economy, Finance, Oil
Cars	F1, Formula, Driver, BMW, Car	Deals	Chance, Deals, Contest, Cashback
Education	Education, Library, Publish	Food	Food, Cook, Recipe, Restaurant
Fitness, Health	Fitness, Health, Workout, Gym, Muscle, Weight, Training, Treatment	Fashion	Fashion, #ootd, #nyfw, Carpet, Dress, Collection, Beauty, Style
Gaming	Game, Xbox, PS4, Gaming, Dota, League	Jokes, Funny	Funny, Joke, Humor, LOL, Humour, Fun
Music	Music, #mtvstars, Concert, kpop	Quotes	Quote, Happiness, Positive
Personal activity	Eating, Super, God, Hell, Moment, Feeling, Asleep, Weather	Politics	Politics, Obama, War, Immigration, Election, Congress, Minister, Military
Religion	Religion, Lord, Buddhism, Christain	Sports	Sports, Basketball, NBA, Football, Goal
Technology, Science	Technology, Tech, Google, Apple, Mobile, NASA, Science, Solar, Comet, Earth	Twitter	Twitter, Followers, Unfollowers, Fustunfollow, Unfollowed
TV & Films	TV, Movie, Trailer, Plot, IMDb	Travel	Travel, Tour, Vacation, Hotel, Island
Video	Video, Youtuber, Youbube, Viewer		

LDA based methods. We call the topics generated by T-LDA the L-topics. In T-LDA, each L-topic is represented as a word distribution. We manually read the word distribution and then assigned a topic name to it. For example, a L-topic with top words: *collection*, *fashion*, *dress*, *wearing*, and *makeup* was assigned the topic name “Fashion”. We manually checked all the L-topics generated with the number of L-topics  $K' = 20, 30, 40, 50$  and  $60$ . Note that multiple L-topics may be assigned with the same topic name and L-topics without clear topic may not be assigned with topic names. We finally obtained the 23 topics used in our survey, i.e.,  $Y = \{y_1, y_2, \dots, y_T\}$  where  $T = 23$ . For each topic  $y_t \in Y$ , we manually selected a set of keywords  $\gamma_{y_t}$  from the top words in each of the L-topics that are assigned as  $y_t$ . Table 1 shows the 23 topics and some related keywords.

## 4 Posting and Reading Topic Discovery

Another goal of this work is to discover user posting and reading topics. We consider this problem as a form of ranking problem: we use ranking strategies to rank topics and aim to give user interested topics higher ranks and uninterested topics lower ranks. A ranking strategy takes certain information (e.g., content

and following networks) of a user as input and outputs a topic ranking for her. We define some notations first for easy reading. Let  $Y = \{y_1, y_2, \dots, y_T\}$  be a set of topics to be ranked. A ranking  $\sigma$  is a bijection from  $\{y_1, y_2, \dots, y_T\}$  to itself. We use  $\sigma(y_t)$  to denote the rank or position given to topic  $y_t$ ,  $\sigma^{-1}(k)$  to denote the topic at the  $k$ -th position,  $\sigma^{-1}(1..k)$  to denote the set of topics in the top  $k$  positions, and  $\pi$  to represent a set of ground truth posting or reading topics according to which type of topic interests we want to discover.

To evaluate a ranking strategy on a set of testing users  $U_{test}$ , we use measurer *mean average precision at position  $n$*  (MAP@ $n$ ) which is a common way to measure rankings. In our case,  $n$  represents the number of top ranking topics chosen as the predicted topics. For example, if  $n = 5$ , then we will use  $\sigma^{-1}(1..5)$  as the predicted topics. To calculate MAP@ $n$  for  $U_{test}$ , we first calculate *average precision at position  $n$*  (ap@ $n$ ) for each user in  $U_{test}$ :  $ap@n = \frac{\sum_{k=1}^n P(k)}{n}$  where  $P(k)$  represents the precision at the cut-off  $k$  topics in the ranking, i.e.,  $P(k) = \frac{|\sigma^{-1}(1..k) \cap \pi|}{k}$  if  $\sigma^{-1}(k) \in \pi$ , otherwise,  $P(k) = 0$ . The MAP@ $n$  for  $U_{test}$  is the average of the average precision of each user, i.e.,  $MAP@n = \frac{\sum_{u \in U_{test}} ap_u@n}{|U_{test}|}$ .

The rest of this section is organized as follows. First, we present three different ranking strategies: Popularity, Content, and Followee-Expertise. Each ranking strategy takes different information of a user for posting or reading topic discovery. Next, we propose a model that learns to combine rankings determined from different strategies. Finally, we show the performance of discovering user posting and reading topics.

## 4.1 Ranking Strategies

**Popularity.** Popularity strategy ranks posting and reading topics according to their popularity. We call the Popularity strategy *Post-Popularity (Read-Popularity)* if we aim to discover posting topics (reading topics). The intuition of Popularity strategy is that users are likely to be interested in popular topics. The popularity of each posting or reading topic is obtained from a set of training users  $U_{train}$ . Let  $\pi^{(u)}$  be the set of ground truth posting or reading topics for user  $u$ . For each topic  $y \in Y$ , we obtain its popularity measured by the number of training users interested in  $y$ , i.e.,  $|\{u | y \in \pi^{(u)}, u \in U_{train}\}|$ . We then rank the topics by their popularity.

**Content.** A user related content can be tweets posted by herself or received from her followees. The posted tweets are the content she likes to share. The received tweets include the content she likes to read. We therefore have two ranking strategies based on posted content and received content to predict user posting and reading topics respectively. They are *Posted-Content* and *Received-Content* strategies respectively. User posted and received content is commonly used for topic discovery [23, 26, 27]. The intuition of the Content strategies is that users are likely to be interested in the topics that their posted and received content is associated with.

The Content strategies rank topics as follows. We first obtain tweets from a set of users including the users whose topic interests we aim to infer and their followees. We then use T-LDA to generate all users’ L-topic distributions from their content. Recall that to differentiate the topics learned by T-LDA from the topics to be ranked ( $Y$ ), we call the former the L-topics  $X = \{x_1, x_2, \dots, x_K\}$ .

Next, we map L-topics in  $X$  to topics in  $Y$ . For each topic  $y_t \in Y$ , we have defined a set of related keywords, i.e.,  $\gamma_{y_t}$ . Each L-topic  $x_k \in X$  is represented as a word distribution. We empirically use the top 30 words in the distribution as  $x_k$ ’s keywords, i.e.,  $\gamma_{x_k}$ . We then find a topic  $y_{t_k}$  for  $x_k$  such that they share the most common keywords, i.e.,  $y_{t_k} = \arg \max_{y_t} |\gamma_{x_k} \cap \gamma_{y_t}|$ . In this way, we can map every L-topic  $x_k$  to a topic  $y_{t_k}$ . It is possible to have multiple L-topics mapped to one topic in  $Y$ .

Finally, with the mapping from  $X$  to  $Y$ , we determine user topic distribution as follows. From T-LDA, each user is assigned a L-topic distribution, i.e.,  $\langle l_1, l_2, \dots, l_K \rangle$  where  $l_k$  represents how likely the user is interested in  $x_k$ . For each  $y_t \in Y$ , we obtain the likelihood that the user is interested in  $y_t$  by summing up  $l_k$  for  $x_k$ ’s that are mapped to  $y_t$ , i.e.,  $z_t = \sum_{t_k=t} l_k$ . Thus we obtain a topic distribution  $\langle z_1, z_2, \dots, z_T \rangle$  for this user. The Content ranking strategy returns the topics according to their topic ordering in  $\langle z_1, z_2, \dots, z_T \rangle$ .

**Followee-Expertise.** A user’s choice of following other users can reveal her reading topic interests. We particularly focus on followees who are well known to be associated with topics. These users are known as *topic experts* [8]. For example, if a Twitter account is well known to post content related to sports events, then this account is an expert in topic “Sports”. The topic a user is well known to be associated with is her *topic expertise* or *expertise*. When a user has an expertise, it is likely to be followed by other users interested in that expertise. For example, if a user likes sports, she may follow sports news accounts or stars whose expertise is “Sports”. Thus, the intuition behind *Followee-Expertise* strategy is that a user is likely to be interested in reading a topic if many of her followees have expertise in that topic [3].

We adopt a method proposed in [8] to obtain followees with expertise. This method exploits the *List* feature of Twitter. In Twitter, users can create lists to organize their followees. Each list has a name given by the user who created this list. Some list names do not carry any meaning (e.g., “list #2”). Some list names show the social relationships of the members (e.g., “family”). There are also many list names that reveal the members’ expertise (e.g., “music”).

We therefore make use of list names to obtain followees with expertise. First, we crawled the number of lists each followee is member of and the names of the lists. The users who are member of only very few lists are usually not well known and these lists are usually for social purpose. We therefore only included those followees who appear in at least 10 lists. For our survey participants, we obtained 15,395 followees. 8,601 of them are public users. Among the 8,601 followees, 43 percent of them appear in at least 10 lists. As Twitter API has rate limits, we collected at most 1000 lists per followee. Next, for each followee, we removed the stop words from the names of the lists she is member of and chose at most 20

top frequent words that appear in the names. We use  $\beta^{(f)}$  to denote the chosen words for followee  $f$ . Finally, to know  $f$ 's expertise, we again utilize the keywords from each topic in  $Y$ :  $f$ 's expertise is  $y^{(f)} \in Y$  if  $\beta^{(f)}$  and  $y^{(f)}$ 's related keyword set  $\gamma_{y^{(f)}}$  share the most number of words, i.e.,  $y^{(f)} = \arg \max_y |\beta^{(f)} \cap \gamma_y|$ . For example, for account @latimesports, we obtained  $\beta^{(\text{@latimesports})} = \{\text{sports, news, media, lakers, nfl, baseball, ...}\}$ , and the topic expertise is "Sports".

For a user whose reading topics are to be predicted, we use the above way to derive a set of her followees with expertise, i.e.,  $F^e$ . Each followee  $f \in F^e$  has an expertise  $y^{(f)}$ . Followee-Expertise ranks topic  $y \in Y$  in higher position than  $y' \in Y$ , if the number of followees with expertise  $y$  is larger than the number of followees with expertise  $y'$ , i.e.,  $|\{f | y^{(f)} = y, f \in F^e\}| > |\{f | y^{(f)} = y', f \in F^e\}|$ . For example, if a user follows 8 accounts with expertise "Sports", 4 accounts with "Politics" and 10 accounts with "Music", then the user's reading topic ranking is "Music", "Sports" and "Politics".

## 4.2 Learning to Combine Rankings

The above three ranking strategies utilize different information to infer users' posting or reading topic rankings. It is possible that different ranking strategies can complement each other so as to achieve better performance [16]. We therefore propose a model that learns to combine rankings generated from different ranking strategies.

We are given a set of training users  $U_{train}$  that we wish to uncover their posting or reading topics. For each user, we have a collection of rankings which are generated by different ranking strategies. We use  $\sigma_i^{(u)}$  to represent the  $i$ -th ranking for user  $u$ . Remember that we use  $\pi^{(u)}$  to denote the set of ground truth topics for user  $u$ . We have Posted-Content and Post-Popularity strategies for predicting posting topics, and Received-Content, Read-Popularity, and Followee-Expertise strategies for predicting reading topics.

For the  $i$ -th ranking strategy, we define a set of parameters  $w_i = \{w_{i1}, w_{i2}, \dots, w_{iT}\}$  where  $w_{it}$  represents how important the topic at position  $t$  is in the  $i$ -th ranking strategy and  $0 < w_{it} < 1$ . We then combine user  $u$ 's rankings as follows: for each topic  $y \in Y$ , we obtain its overall (or combined) importance by summing up the topic  $y$ 's importance in all ranking strategies, i.e.,  $\sum_i w_i \sigma_i^{(u)}(y)$  where  $\sigma_i^{(u)}(y)$  represents the rank assigned to  $y$  by the  $i$ -th ranking for user  $u$ . We then can re-rank all the topics based on their overall importance, and get a combined ranking  $\phi^{(u)}$  for user  $u$ .

A good combined ranking  $\phi^{(u)}$  should rank the topics from ground truth topics  $\pi^{(u)}$  in front positions. Thus the topics in  $\pi^{(u)}$  should be much more important than the other topics. This means we need  $\frac{\sum_{y \in \pi^{(u)}} \sum_i w_i \sigma_i^{(u)}(y)}{\sum_{y \in Y} \sum_i w_i \sigma_i^{(u)}(y)}$  to be close to 1.

In other words, we want the total importance of the user interested topics (the numerator) to be close to the total importance of all topics (the denominator).

We then can write our model as follows. We minimize the following function:

$$F(w) = \frac{1}{2|U_{train}|} \sum_{u \in U_{train}} \left(1 - \frac{\sum_{y \in \pi^{(u)}} \sum_i w_{i\sigma_i^{(u)}(y)}}{\sum_{y \in Y} \sum_i w_{i\sigma_i^{(u)}(y)}}\right)^2 \quad (1)$$

To simplify the representation, we can rewrite  $F(w)$  as  $F(w) = \frac{1}{2|U_{train}|} \sum_{u \in U_{train}} \left(1 - \frac{\sum_i \sum_t a_{it}^{(u)} w_{it}}{\sum_i \sum_t w_{it}}\right)^2$  where  $a_{it}^{(u)}$  equals to 1 if there exists a topic  $y \in \pi^{(u)}$  such that  $\sigma_i^{(u)}(y) = t$ . Otherwise,  $a_{it}^{(u)}$  equals to 0.

In order to ensure  $w_{it}$  falls within  $(0, 1)$ , we transform it using logistic function:  $w_{it} = \frac{1}{1+e^{-\theta_{it}}}$ . Thus, instead of learning  $w$ , we learn  $\theta$ . To avoid overfitting, we add a regularization term to our objective function.

$$F(\theta) = \frac{1}{2|U_{train}|} \sum_{u \in U_{train}} \left(1 - \frac{\sum_i \sum_t a_{it}^{(u)} w_{it}}{\sum_i \sum_t w_{it}}\right)^2 + \frac{\lambda}{2|U_{train}|} \sum_i \sum_t \theta_{it}^2 \quad (2)$$

where  $w_{it} = \frac{1}{1+e^{-\theta_{it}}}$  and  $\lambda$  is a control of the fitting parameters  $\theta$ . As  $F$  is not convex, in order to improve the chances of finding a global minimum, a common strategy is to use gradient descent with random restart, which performs gradient descent many times (e.g., 100 times) with randomly chosen initial points, and selects the locally optimized point with the lowest  $F$  value. We write the derivative of  $F$  of  $\theta_{jv}$ :

$$\begin{aligned} \frac{\partial}{\partial \theta_{jv}} F(\theta) = & - \frac{1}{|U_{train}|} \sum_{u \in U_{train}} \left( \left(1 - \frac{\sum_i \sum_t a_{it}^{(u)} w_{it}}{\sum_i \sum_t w_{it}}\right) \right. \\ & \left. \frac{a_{jv}^{(u)} \sum_i \sum_t w_{it} - \sum_i \sum_t a_{it}^{(u)} w_{it}}{(\sum_i \sum_t w_{it})^2} \frac{e^{-\theta_{jv}}}{(1+e^{-\theta_{jv}})^2} \right) + \frac{\lambda}{|U_{train}|} \theta_{jv} \end{aligned} \quad (3)$$

The update rule is  $\theta_{jv} := \theta_{jv} - \alpha \frac{\partial}{\partial \theta_{jv}} F(\theta)$ , where  $\alpha$  is the learning rate. After we learn  $\theta$  and then obtain parameter  $w_i$  for each ranking strategy  $i$ , we can get the combined ranking for user  $u$  by computing the overall importance for each topic  $y$  using  $\sum_i w_{i\sigma_i^{(u)}(y)}$ .

### 4.3 Results of Posting and Reading Topic Discovery

We use the ground truth topics obtained from our survey to evaluate the ranking strategies. All the following results are the average MAP by repeating 5-fold cross-validation 10 times. We empirically set  $\lambda = 0.1$  and  $\alpha = 20$ .

**Posting Topic Discovery.** We use 69 participants who posted no less than 5 tweets from March 1st to March 30th, 2015 for this part of evaluation, and the remaining users are considered as lurkers who mainly focus on reading. We apply Posted-Content and Post-Popularity to predict user posting topics. Table 2 shows the performance of these two ranking strategies and the performance of the combined rankings. To determine the significance of results, we use the randomly shuffled topics (i.e., the Random predictor) as baseline. In the Table,  $n$  represents

**Table 2.** Performance (MAP@n) of posting topic discovery.

		Random	Posted-content	Post-popularity	Combined
Like	n = 3	0.22	0.38	0.55	<b>0.58</b>
	n = 5	0.2	0.31	0.48	<b>0.51</b>
	n = 7	0.18	0.31	0.50	<b>0.52</b>
Like and somewhat like	n = 3	0.49	0.65	0.85	<b>0.86</b>
	n = 5	0.46	0.57	0.79	<b>0.80</b>
	n = 7	0.44	0.55	<b>0.76</b>	<b>0.76</b>

the number of topics that are chosen as the predicted topics. “Like” means that we use the topics that a user likes to post as the ground truth topics, and “Like and Somewhat Like” means that we use the topics that a user rates “Like” or “Somewhat Like” as the ground truth topics.

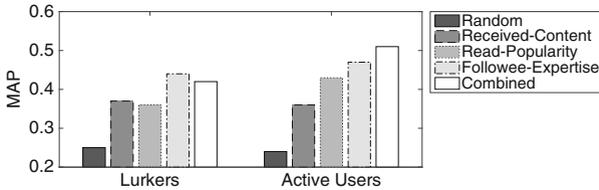
We observe firstly that all our ranking strategies yield performance significantly better than Random. Secondly, Post-Popularity performs much better than Posted-Content. One possible reason is that inferring topics from tweets is still a challenging problem as tweets are short and people use many informal and idiosyncratic words in tweets [14]. The performance of Post-Popularity shows that there are some “universal” posting topics such as “TV & Films” and “Music”. Finally, the combined ranking method achieves the best performance.

**Reading Topic Discovery.** We use all the survey participants in reading topic discovery evaluation. Table 3 shows the performance of Received-Content, Read-Popularity and Followee-Expertise and their combined rankings. We summarize our observations as follows. First, all our ranking strategies perform significantly better than Random. Secondly, compared with Read-Popularity and Followee-Expertise, Received-Content does not predict user reading topics well. One possible reason is the difficulty of inferring topics in tweets. Another possible reason is that Twitter users are only interested in a subset of tweets they received. Thirdly, Followee-Expertise, an unsupervised method, mostly performs better than Read-Popularity. Fourthly, again, the combined ranking can achieve the best performance. Lastly, comparing Tables 2 and 3, we notice that reading topic discovery can achieve comparable performance as posting topic discovery, which suggests that although we do not have user reading behavior data traces, we can still predict user reading topics with reasonable accuracy.

**Reading Topic Discovery for Lurkers.** In order to see how well we can predict lurkers’ reading topics, we divide the testing users into lurker group and active user group. The lurker group consists of the users who post less than 5 tweets from March 1st to March 30th, 2015 and the remaining users belong to the active user group. Figure 5 shows the performance of predicting reading topics for lurkers and active users. We set  $n = 5$  and the ground truth topics are the “Like” topics. Other settings have consistent findings. We first observe that all our methods perform much better than Random for both lurker and active user

**Table 3.** Performance (MAP@n) of reading topic discovery.

		Random	Received-content	Read-popularity	Followee-expertise	Combined
Like	n = 3	0.29	0.39	0.50	0.52	<b>0.56</b>
	n = 5	0.25	0.34	0.41	0.43	<b>0.48</b>
	n = 7	0.22	0.33	0.41	0.42	<b>0.46</b>
Like and somewhat like	n = 3	0.61	0.71	0.82	0.86	<b>0.87</b>
	n = 5	0.60	0.66	0.78	0.79	<b>0.80</b>
	n = 7	0.57	0.65	0.78	0.77	<b>0.80</b>

**Fig. 5.** Performance of predicting lurkers and active users' reading topics.

groups. Secondly, overall, predicting active users' reading topics is easier than predicting lurkers'. Thirdly, Read-Popularity does not perform well for lurkers. It shows that compared with active users, lurkers are less likely to pay attention to the popular reading topics. Lastly, we find that Followee-Expertise performs best for the lurker group. Thus, using only this unsupervised method, we can achieve promising prediction results for lurkers.

## 5 Discussion and Conclusion

One of the main contributions of this work is to show that social media users' posting topics are different from their reading topics. We also observe that topics are different in attracting people to post and to read. For example, users seem to have less concerns when posting topics such as "TV & Films" and "Music". However, for topics such as "Adult", "Religion", "Politics" and "Business", many users who are interested in reading them do not share them in Twitter. Our findings imply that to measure the popularity of a tweet or an event, we need to consider its topic. For example, if a tweet is about "Politics", then the number of users sharing it could possibly underestimate its popularity or influence.

Our work also contributes to the prediction of users' posting and reading topics. We have evaluated the prediction performance using different ranking strategies. We demonstrated that predicting reading topics can achieve similar performance as predicting posting topics, although the reading content is not observed. We also showed that we can predict lurkers' reading topics using the

topic experts among the lurkers' followees. Posting and reading topics can be useful in different practical scenarios. For example, posting topics can be used to predict if users will share an event or product. Users' reading topics can be used to predict if they will pay attention to an advertisement.

In the future work, we could examine and compare the differences between posting and reading topics for a much larger user community and in other social media platforms such as Facebook. Another future direction is to study users' views and opinions when they are interested in certain topics but do not share them, and the context which encourages people to speak up.

**Acknowledgments.** The authors gratefully thank Ingmar Weber for the inspiring discussion. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

## References

1. Balasubramanian, S., Mahajan, V.: The economic leverage of the virtual community. *Int. J. Electron. Commer.* **5**(3), 103–138 (2001)
2. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. In: *IMC* (2009)
3. Bhattacharya, P., Zafar, M.B., Ganguly, N., Ghosh, S., Gummadi, K.P.: Inferring user interests in the twitter social network. In: *RecSys* (2014)
4. Chen, J., Hsieh, G., Mahmud, J.U., Nichols, J.: Understanding individuals' personal values from social media word use. In: *CSCW* (2014)
5. Das, S., Kramer, A.: Self-censorship on Facebook. In: *ICWSM* (2013)
6. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D.: The youtube video recommendation system. In: *RecSys* (2010)
7. Diao, Q., Jiang, J.: A unified model for topics, events and users on Twitter. In: *EMNLP* (2013)
8. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: crowd-sourcing search for topic experts in microblogs. In: *SIGIR* (2012)
9. Gong, W., Lim, E.P., Zhu, F.: Characterizing silent users in social media communities. In: *ICWSM* (2015)
10. Hampton, K., Rainie, L., Lu, W., Dwyer, M., Shin, I., Purcell, K.: Social media and the "spiral of silence". Pew Research Internet Project (2014). <http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/>
11. Hong, L., Bekkerman, R., Adler, J., Davison, B.D.: Learning to rank social update streams. In: *SIGIR* (2012)
12. Hsieh, G., Chen, J., Mahmud, J.U., Nichols, J.: You read what you value: understanding personal values and reading interests. In: *CHI* (2014)
13. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. *PNAS* **110**(15), 5802–5850 (2013)
14. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG! In: *ICWSM* (2011)
15. Kulshrestha, J., Zafar, M.B., Noboa, L.E., Gummadi, K.P., Ghosh, S.: Characterizing information diets of social media users. In: *ICWSM* (2015)

16. Lebanon, G., Lafferty, J.D.: Cranking: combining rankings using conditional probability models on permutations. In: ICML (2002)
17. Li, X., Guo, L., Zhao, Y.E.: Tag-based social interest discovery. In: WWW (2008)
18. Linden, G., Smith, B., York, J.: Amazon. com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* **7**(1), 76–80 (2003)
19. Michelson, M., Macskassy, S.A.: Discovering users’ topics of interest on twitter: a first look. In: AND (2010)
20. Nonnecke, B., Preece, J.: Lurker demographics: counting the silent. In: CHI (2000)
21. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC (2010)
22. Sleeper, M., Balebako, R., Das, S., McConahy, A.L., Wiese, J., Cranor, L.F.: The post that wasn’t: exploring self-censorship on facebook. In: CSCW (2013)
23. Spasojevic, N., Yan, J., Rao, A., Bhattacharyya, P.: Lasta: large scale topic assignment on multiple social networks. In: KDD (2014)
24. Tagarelli, A., Interdonato, R.: “Who’s out there?”: identifying and ranking lurkers in social networks. In: ASONAM (2013)
25. Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B.Y.: You are how you click: clickstream analysis for Sybil detection. In: SEC (2013)
26. Xu, Z., Ru, L., Xiang, L., Yang, Q.: Discovering user interest on twitter with a modified author-topic model. In: WI-IAT (2011)
27. Yang, Z., Xu, J., Li, X.: Data selection for user topic model in twitter-like service. In: ICPADS (2011)
28. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011)



<http://www.springer.com/978-3-319-28360-9>

Advances in Network Science

12th International Conference and School, NetSci-X

2016, Wroclaw, Poland, January 11-13, 2016,

Proceedings

Wierzbicki, A.; Brandes, U.; Schweitzer, F.; Pedreschi, D.

(Eds.)

2016, XII, 213 p. 61 illus. in color., Softcover

ISBN: 978-3-319-28360-9