Twitter User's Interest Detection by Using Followee Information Based on LDA Topic Model

Yu Wang and Akira Maeda

Abstract—**Using Latent Dirichlet Allocation model is a very common method to extract topics from documents. We propose a method to extract Twitter user's interests based on their followee information. We use the LDA model to extract user's interested topics by using user's followee information, which means who they are following. We do a pre-processing for our dataset collected from Twitter and then input the followees' tweets into the LDA model. We use the target user's other behavior histories like tweets, retweets, favorites and profile information to do weighting with the topic list extracted by the LDA model. Then, we sort the topics based on their weights, and the topic with the highest weight is our target user's most interested topic. We conducted some interviews to evaluate our method.**

Index Terms—**LDA (Latent Dirichlet Allocation), SNS (Social Networking Service), Twitter, NLP (Natural Language Processing)**

I. Introduction

ONE of the hot topics in NLP (Natural Language Processing) is to analyze text contents on SNS (Social Networking Service) that are posted by the users. This research field aims to understand what SNS users are expressing from the text contents they posted and their behavior history on SNS. If we can infer their interests or trends based on their SNS behavior, better contents can be provided to them through the internet, which leads to a better user experience.

As one of the most popular SNS in the world, Twitter is an ideal social media to research. Not only has Twitter 335 million active users [1], but also provides us with a very convenient API (Application Programming Interface) called Twitter API [2] for researches. We can get almost any data we need easily via this API for researches, such as the user's tweets and retweets, which means what the user has posted on Twitter. Tweets and retweets here are terminology on Twitter. Tweet means what a user has posted onto Twitter. Retweet means that a user posts a tweet by using other's tweets.

An enthusiastic Twitter user tends to open the Twitter app many times in one day. With the rapid development of smart phones, we can see that many youths use SNS apps, such as Twitter, Facebook, and Instagram on their smart phones when they are waiting for a bus, eating food, even in dating. For these users, SNS has become a part of their life. Meanwhile, these heavy Twitter users are our research subject in this paper.

Kwak et al. [3] did a comprehensive research about Twitter in 2010 when Twitter was released less than 3 years ago. They analyzed the Twitter user's relation network, trends on Twitter and the information diffusion by retweet contents.

Takemura et al. [4] classified tweets based on their lifetime duration. Because tweets are a kind of texts which is sensitive to time, it is necessary to tell users which one should be checked at once or which one could be seen later. For example, consider a tweet like *It is raining*. It can provide information users need, but it will lose its value after an hour or two hours. Hence, this kind of tweets with a short lifetime need to be seen immediately. Another example like, *the sky is blue*, has a long lifetime without information loss. This kind of tweets could be pushed to users with no hurry.

Saito et al. [5] use Social Book Mark (SBM) service to extract Twitter user's interests. The tags from SBM are used to build a synonym dictionary to analyze the user's characteristics and interests. Then, the proposed system recommends some contents based on the analysis results.

In this paper, we propose a method to extract target user's interests using the Latent Dirichlet Allocation (LDA) model [6] and combine some weighting methods. The data we mainly use in our research is Twitter user's followee information which means that which users they are following on Twitter. This is information that can represent the target user's interest most, because if a user would like to follow another account on Twitter, it obviously means he/she is interested in the user who he/she will follow. Hence, we use this data as our input into the LDA model. The results from the LDA model will generate several topics with keywords for each of them. Our aim is to use user's behavior on Twitter to weight each topic and find the topic with the highest weight which means the user has interest in.

II. Proposed Method

In this section, we will describe what kind of data we use in our method and how we do pre-process the data first. Then, we will explain the method we propose in detail as shown in Fig. 1. Our method will collect data from Twitter via Twitter API. Then we will do pre-processing for raw Twitter data to get rid of the influence by noise data. Next,

Manuscript received December, 29th, 2018; revised January 20th, 2019.

Yu Wang is with the Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan. (e-mail: gr0320fv@ed.ritsumei.ac.jp)

Akira Maeda is with the College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan. (email: amaeda@is.ritsumei.ac.jp)

data will be inputted into the LDA model that we adjusted parameters in advance. Then we use our weighting structure to calculate the weights for each topic extracted by the LDA model. Finally, we rank the topics, and the topic with the highest weight should be the target user's favorite topic.

A. Dataset

As we have described before, we consider that the Twitter user's followee contents show the target user's interests. Firstly, we need to get a follow list from our target user who has followed more than 200 accounts by using Twitter API. Then, we traverse all the accounts in the following list, and collect the latest 500 tweets from each account. If some accounts do not have enough tweets which means he/she has less than 500 tweets, we collect all the tweets they have. In total, we should have at least 100,000 tweets as our followee contents. Another part of data we need to collect is the target user's behavior histories, such as tweets, retweets and favorites.

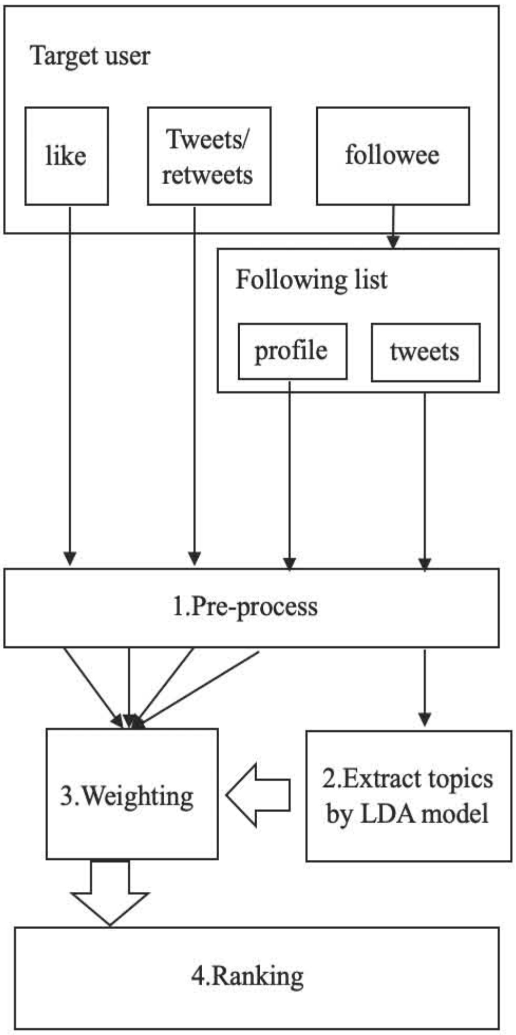
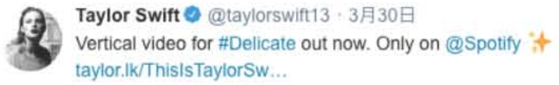


Fig.1 Proposed Method

B. Pre-processing

Pre-processing is a necessary and very important part in the NLP methods. In our research, we aim to analyze both English user's and Japanese user's interests. However, pre-processing will be very different in these two languages.



Taylor Swift @taylorswift13 · 3月30日
Vertical video for #Delicate out now. Only on @Spotify ✨
taylor.lk/ThisIsTaylorSw…

Fig.2 A typical example of an English tweet

a. Tweet pre-processing

Fig.2 is a typical example of a tweet on Twitter. It has all kinds of Twitter elements in this example. We will explain how to analyze these Twitter elements in our research. Although this is an example in English, this process is the same with Japanese.

a) For the trends, which means one word with a hash symbol, we will delete the hash symbol, and get the word behind it.
b) For the direct messages, which means an account with a @, we will delete the @, and collect the word behind it.
c) For the hyperlinks in tweets, since we will delete words whose term frequency is lower than 10% before inputting data into LDA model, all the hyperlinks will be deleted. Because no hyperlinks are the same, we can delete the hyperlinks completely in our dataset.
d) For the emoji in tweets, it is similar with what we do on hyperlinks. Emoji can be transformed to Unicode such as \ue022 or \ue412, so we can delete most of emojis whose term frequency is lower than 10%. For some emojis that are used frequently, we make a dictionary to translate the emoji into words. For example, "⚽: soccer".

b. English pre-processing

Doing pre-processing for English papers or news articles will be a pretty easy task because English words are split by spaces. The only thing needed to do is to delete the stop words that are defined in a stop word dictionary, such as "you", "me", etc., in the texts. However, it is a very difficult task to do pre-processing for English Twitter texts. Since there are too many internet slang words such as "lol", which means laugh out loudly, "xD" which is an expression of laughing. To solve this stop word problem, we build a stop word dictionary for Twitter, which is called Twitter-stop-words-dictionary. In order to recognize stop words on Twitter as far as possible, we are still expanding the dictionary. Until now, we have already recorded 362 words in our Twitter-stop-words-dictionary.

c. Japanese pre-processing

Due to the fact that Japanese sentences are not split by space, doing pre-processing for Japanese is quite a different task from English. In order to decide which parts we should use from a Japanese sentence, we need to perform morphological analysis first. Morphological analysis is a process to divide sentences into smallest units. For example, a sentence will be divided into a noun, an adjective or a verb.

For our morphological analysis, we use Janome [7] to divide Japanese Twitter texts. As the result of morphological analysis, we can get a set of {morpheme, part-of-speech} pairs such as {私, 名詞}, which means "私

(I, my, me, etc.)" is a noun. We can use this method to divide Japanese Twitter texts into the smallest units with their part of speech.

In our research, we only keep nouns which may show user's interests directly and adjectives which may show user's interests indirectly. We do not use verbs because it has a very low influence on user's interests.

After pre-processing English and Japanese Twitter texts, we can get a BoW (Bag of Words) model as our dataset. The BoW model means a model with a word list just like a bag of words. Every word in a BoW model can be described as a one-hot vector which could be a vector like (0, 2, 0, 0, 0). As the result, we use these vectors as the input into the LDA model.

C. LDA model

After we obtain the data from pre-processing, we can input them into the LDA model. LDA is a topic model that is used to extract topics from documents. Since one tweet can only have less than 140 words, it will have a bad performance if we treat one tweet as one document. In our case, we treat 500 tweets from one following account as one document which is a method known as Author-topic-model in NLP field [8]. We think that the tweets from one account can represent what this account is about, so we treat them as one document. That is, if our target user is following 250 accounts on Twitter, we will have 250 documents as our input data. Also, this way can let us avoid generating sparse matrixes from the BoW model.

The result of the LDA model will generate *n* topics with *m* keywords for each. *n* and *m* are two parameters we need to set in advance. We will explain the parameters setting part in detail in section 4.

D. Weighting

We consider 3 factors to do topic weighting in our research. The first one is the target user's favorites contents on Twitter. Target's favorites contents are determined by the target user's clicks of 'like' button when they see other accounts' tweets if they like. In this case, we give a weighting factor 3, in our formula of favorite calculation.

$$W_{T_m} = \sum_{n=1} k \cdot Similarity\,(T_m,\ L_n) \qquad (1)$$

where T_m means topic *m*, and meanwhile L_n means *n*-th favorite contents in the target's like list. The result W_{Tm} means the weight of topic *m*. *k* is the weight we give to different factors. We calculate the sum of cosine similarity between topic *m* and like *n* with weight 3.

$$Similarity\,(T_m,\ L_n) = \frac{T_m L_n}{\sqrt{\sum_{m=1} T_m^2}\sqrt{\sum_{n=1} L_n^2}} \qquad (2)$$

Formula (2) shows how we calculate the similarity between liking list and topic list. Basically, we calculate cosine similarity between each pair of two vectors.

Similarly, we also use the target user's tweet contents to do weighting. Besides, we treat retweets as a part of tweet contents to do this process. For tweets/retweets weighting, we set a weighting factor as 2, since there are many users who post their daily life or mood rather than posting anything about their interests on Twitter. Hence, we give a lower weighting factor than favorite contents.

The third factor we will use is the profile information from the followees. Technically, this part should be the most important one in our weighting process. Because the profile is supposed to describe what this account is about. However, there are many accounts which trend to write aphorism that has no direct relation with their interested topics. Considering this, we set the profile part's weighting factors as 1 eventually.

E. Ranking

The last part of our proposed method is to rank the topics with weights. The higher weights that topic has, the more interest our target has in this topic. Hence, we can infer the target's interests via our weighting results. The topic with the highest weight should be the most possible interests that the target user could have.

III. EXPERIMENT

In this part, we will explain how we determine two parameters for our LDA model and show the result calculated by our proposed method using a Twitter dataset.

A. Number of topics

Firstly, we need to determine how many topics we want to extract from our Twitter dataset. Technically, the LDA model will generate 10 topics from documents based on the default setting. However, we need to judge if it is an appropriate setting for our dataset. However, it is very difficult to judge whether a topic model is correct or not, because we need to judge them by human. Fortunately, there is a method called Word Intrusion proposed by Chang [8], which can be used to evaluate our results. It is a task to measure how well the inferred topics match the human concepts.

$$MP_k^m = \sum_s 1\,(\,i_{k,s}^m = w_k^m\,)\,/S \qquad (3)$$

Here w_k^m is the index of word generated from *k*-th model by model m. $i_{k,s}^m$ is the word selected by subject *S* on the set of *k*-th topic inferred by model m and Subject *S*. Fig.3 shows the result of word intrusion by adjusting topic numbers.

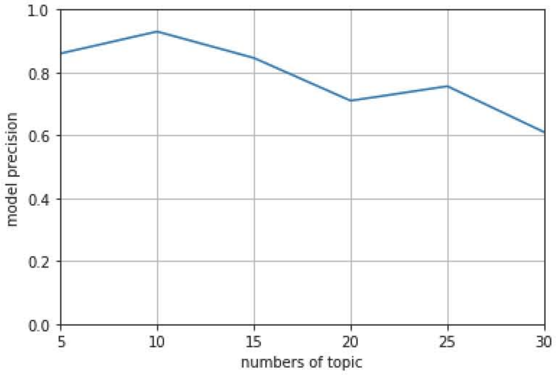
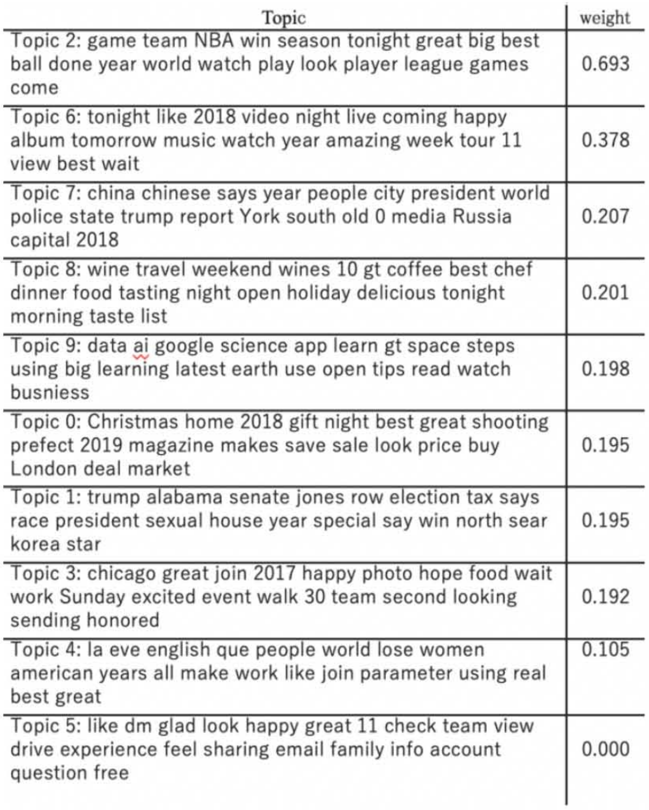


Fig.3 Rating topic numbers

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table.1 Example results in English



Topic	weight
Topic 2: game team NBA win season tonight great big best ball done year world watch play look player league games come	0.693
Topic 6: tonight like 2018 video night live coming happy album tomorrow music watch year amazing week tour 11 view best wait	0.378
Topic 7: china chinese says year people city president world police state trump report York south old 0 media Russia capital 2018	0.207
Topic 8: wine travel weekend wines 10 gt coffee best chef dinner food tasting night open holiday delicious tonight morning taste list	0.201
Topic 9: data ai google science app learn gt space steps using big learning latest earth use open tips read watch busniess	0.198
Topic 0: Christmas home 2018 gift night best great shooting prefect 2019 magazine makes save sale look price buy London deal market	0.195
Topic 1: trump alabama senate jones row election tax says race president sexual house year special say win north sear korea star	0.195
Topic 3: chicago great join 2017 happy photo hope food wait work Sunday excited event walk 30 team second looking sending honored	0.192
Topic 4: la eve english que people world lose women american years all make work like join parameter using real best great	0.105
Topic 5: like dm glad look happy great 11 check team view drive experience feel sharing email family info account question free	0.000

With the increment of topic numbers, the precision of the LDA model decreases in Fig.3. When we extracted 10 topics, model will have the highest precision with 92%. At this point, we can prove that 10 is a suitable topic number parameter we should give to the LDA model.

B. *Number of keywords for each topic*

As well as the number of topics has been decided, we need to determine another parameter, which is the number of keywords for each topic. The default setting is 10 as the same with number of topics. We have tried from 5 to 25 keywords. As the result, 20 keywords for each topic have the best distribution performance in our Twitter dataset.

C. *Experimental Data*

Before we made a questionnaire to evaluate our method, we signed up a new Twitter account to test our proposed method whether it can infer Twitter user's interests correctly. The reason we made a new account is that we want to control what a user will do on Twitter, so that we can know that if our method infers correct interests or not. The new account follows 200 hundred accounts randomly from Twitter recommended. Next, we clicked like buttons on some tweets.

As shown in Table 1, topic 2 on the top of the table is a topic about basketball. Apparently, its weight is much higher than other topics. Our method will infer that basketball is a specific topic that this Twitter user is interested in. Also, we can see that some topics at the bottom have very low weights, which means that the target user has low interests in them.

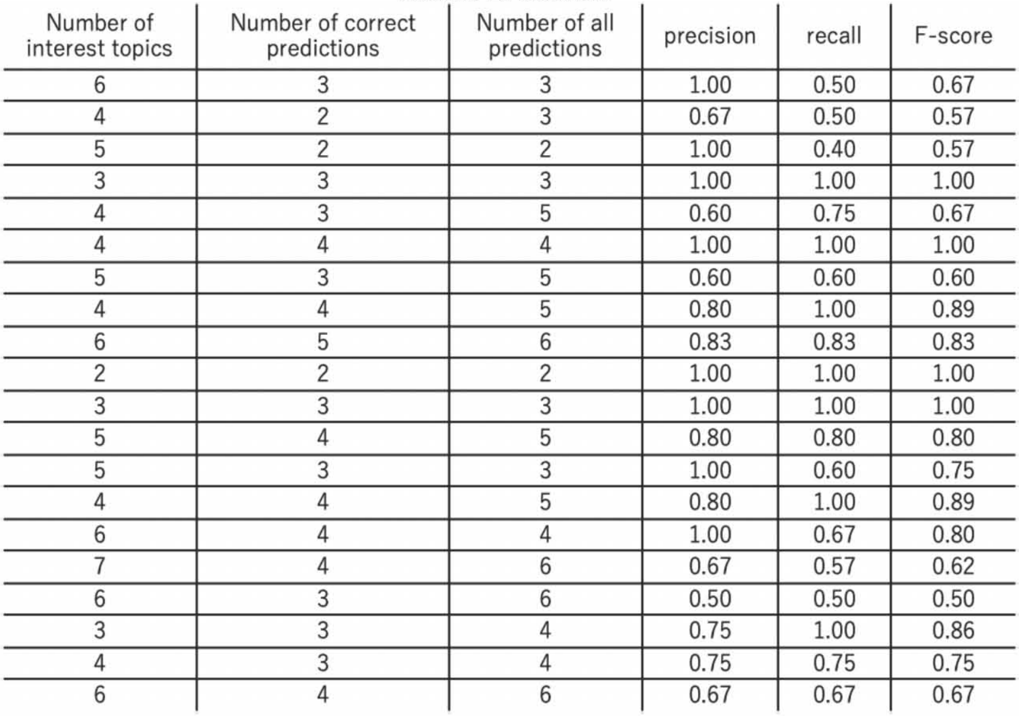
To certify that our proposed method gives us a correct result, we checked the account we used in this experiment. This user is following 17 accounts related to basketball which is less than 23 accounts about music. However, this account has 10 favorite tweets about basketball in total 15 tweets.

IV. Evaluation

In order to verify how well our method will perform in real situations, we invited 20 heavy Twitter users who almost use Twitter every day. In this experiment, we use precision,

$$precision = \frac{|\{relevent\ docs\} \cap \{retrieved\ docs\}|}{|\{retrieved\ docs\}|} \quad (4)$$

Table.2 Evaluation



Number of interest topics	Number of correct predictions	Number of all predictions	precision	recall	F-score
6	3	3	1.00	0.50	0.67
4	2	3	0.67	0.50	0.57
5	2	2	1.00	0.40	0.57
3	3	3	1.00	1.00	1.00
4	3	5	0.60	0.75	0.67
4	4	4	1.00	1.00	1.00
5	3	5	0.60	0.60	0.60
4	4	5	0.80	1.00	0.89
6	5	6	0.83	0.83	0.83
2	2	2	1.00	1.00	1.00
3	3	3	1.00	1.00	1.00
5	4	5	0.80	0.80	0.80
5	3	3	1.00	0.60	0.75
4	4	5	0.80	1.00	0.89
6	4	4	1.00	0.67	0.80
7	4	6	0.67	0.57	0.62
6	3	6	0.50	0.50	0.50
3	3	4	0.75	1.00	0.86
4	3	4	0.75	0.75	0.75
6	4	6	0.67	0.67	0.67

recall,

$$recall = \frac{|\{relevent\ docs\} \cap \{retrieved\ docs\}|}{|\{relevent\ docs\}|} \tag{5}$$

and F-measure,

$$F - measure = 2\,\frac{precision\ \cdot recall}{precision + recall} \tag{6}$$

to evaluate our method. Where precision is the number of correct results returned by model divided by the number of all the results returned by model. Recall is the number of correct results returned by model divided by the number of all the results that should have been returned by model. These three measures are very common evaluation method in information retrieval field.

We asked 20 users to write down the number of their interested topics. Our proposed method analyzed their Twitter one by one. We set a threshold number to 1.0 for deciding if the topic we get is a user's interest. If the final weights for a topic is greater than 1.0, we assume that this is a topic that user has interested in. As shown in Table 2, basically, our proposed method has a high precision but with a low recall. Also, we found that when users have little numbers of interested topics, our proposed method can get a high precision, recall and F-measure. However, if the users have more than 5 interested topics, our method's performance will become lower.

V. CONCLUSION

In this paper, we proposed a method to infer Twitter user's interests by using the LDA model based on their followee information. We use the Twitter user's followee information as the main factor in our research which is the factor often ignored in existing researches. Although our method can infer the target user's interests in most cases, there are still many aspects needed to be improved.

Analyzing Twitter texts and doing pre-processing is very difficult because there are spelling mistakes, slang words or some meaningless words. We realized that we still need to do more in our pre-processing part, such as stemming which aims to eliminate the differences between singular word and plural word and different tense of verbs needs to be unified as well. On the other hand, emoji should not be simply deleted, since it could express the user's emotion towards one topic. If emoji can be used effectively, the precision of our proposed method will be improved. These tasks will be our next goal to achieve in the future.

REFERENCES

[1] Most popular social networks worldwide ranked by number of active users, https://www.statista.com/statistics/272014/globalsocial-networks-ranked-by-number-of-users/, 2018.10

[2] Twitter API, https://developer.twitter.com, 2018.10

[3] H. Kwak, C. Lee, H. Park, S. Moon. "What is Twitter, a social network or a news media?," in *Proc. of the 19th international conference on World Wide Web*. ACM, pp. 591-600, 2010

[4] H. Takamura, K. Tajima, "Tweet Classification Based on Their Lifetime Duration," in *Proc. of ACM CIKM*, pp.2367-2370, 2012

[5] J. Saito, T. Yukawa, "A method of extracting and recommanding Twitter user's interest based on Social Book Mark," The Special Interest Group Technical Report of IPSJ, pp.1-8, 2011(in Japanese)

[6] D. M. Blei, Y. N. Andrew, M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp.993-1022, 2003

[7] Janome, http://mocobeta.github.io/janome/, 2018.10

[8] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.306-315, 2004

[9] J. Chang, J. Bpyd-Graber, S. Gerrish, C. Wang, D. M. Blei. "Reading tea leaves: How humans interpret topic models," in *Proc. of Advances in neural information processing systems (NIPS2009)*, pp.288-296, 2009