Developing a User Recommendation Engine on Twitter Using Estimated Latent Topics

Hiroyuki Koga[1] and Tadahiro Taniguchi[2]

[1] Graduate School of Engineering, Ritsumeikan University, Japan
[2] Department of Human & Computer Intelligence, Ritsumeikan University, Japan
koga@em.ci.ritsumei.ac.jp

Abstract. In recent years, microblogging is popular among people and informal communication becomes important in various communities. Therefore, a number of Web communication tools are developed to facilitate informal communication. In this paper, focusing on microblogging service, Twitter, we develop a user recommendation engine which extracts latent topics of users based on followings, lists, mentions and RTs. This recommendation algorithm is based on Latent Dirichlet Allocation (LDA) and KL divergence between two users' latent topics. This algorithm hypothesizes that the users have latent connection if the distance calculated by KL divergence is short. Additionally, we performed an experiment to evaluate the effectiveness of the algorithm, and this showed that there is correlation between the distance and user's preference obtained through questionnaire.

Keywords: LDA, Twitter, Information Recommendation.

1 Introduction

Informal communication is quite important in various communities because that is not only for fun but also connects a person to other people providing with important information. For example, some researchers insist that communication in a smoking room accelerates participants to share information. Then, many studies have been made on informal communication which is beneficial for their daily works and facilitates information sharing [1,2,3]. These researches don't limit their offline communication on inquiry. To facilitate online informal communication is, rather, a hot research area.

Web communication services become widely used and developed. To the services, like message boards, microblogging and Web chat, everyone can access, read and write in public equally; that is, personal relationships between users had not been included in the services. However, in recent years, services which focus on the personal relationships become more popular; users only accesses, reads and writes other users' pages based on their own private relationships. After the facebook [4] emerges, Web communication tools which treat personal informations and the relationships become popular. Twitter [5] is one of the most popular tools among them. To obtain fruitful information on twitter, a user should find and follow other users who often tweet messages in which he/she is interested. He/she can read only the messages of them

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2011, LNCS 6761, pp. 461–470, 2011.

he/she follows. The messages tweeted by them he/she follows are shown on his/her Time Line (TL). This means that to find users who are suitable for him/her is important on this informal communication architecture. In this paper, we develop a user recommendation algorithm by using Latent Dirichlet Allocation (LDA). LDA is a well-known document clustering method [6]. Usually, LDA is adapted to documents whose elements are words. In contrast, we apply this to a set of followings of a user and other information on Twitter. We also evaluate effectiveness of this recommendation algorithm through an experiment.

2 Background

2.1 Informal Communication Support

Recently, many researchers are studying on informal communication support system. Nakano et al. proposed the Traveling Café to support informal communication in a local community [1]. The system support informal communication between coffee drinkers by prompting person who make a cup of coffee in a coffee room to go to pour another's cup of coffee. A system proposed by Siio et al. helps people interact with each other by assuming a place around a coffee machine as informal communication space [2]. These studies tried to trigger interaction in the real world using ICT.

On the other hand, new online communication services including SNS and microblogging become booming. In these services, users can communicate with the users' friends of the real world naturally on the Web, because users can edit friends and community based on their relationships. In this paper, we focus on informal communication on Twitter. Our goal is to encourage users to enjoy informal communication on Twitter. Therefore, we develop user recommendation engine recommending users who share the similar interest with a target user. In our proposed method, latent topics which represent users' interests are estimated statistically by using LDA.

2.1 Twitter

Twitter is a Web service on which we can talk to each other with short messages, which called “tweets”. The length of a tweet is limited to 140 characters. Thus, the feature of this service is that users can send messages other users more frequently than Web chat or mail. In this service, “following” and “follower” are keywords[1]. If a user follows other users, their tweets are displayed on his/her Time Line (TL). Tweets of users he/she does not follow are not displayed on his/her TL. Therefore, communication is hardly generated based on their tweets. So, if the number of followings is a little, he/she would not use Twitter effectively. In contrast, if he/she follows many users randomly, they may tweet in which he/she is not interested. His/her TL is filled with their uninteresting tweets because their tweets are displayed on his/her TL. So that, communication on Twitter becomes worthless. Therefore it is important for users to find and follow users who post interesting tweets for them. However, on Twitter, we try to find our friends or interesting people who share similar topics with us, keyword search used

[1] We name the users he/she follows as his/her followings and name the users who follow him/her as his/her followers.

frequently on Web is not effective method, because the messages on Twitter is quite shorter than they on Web pages.

Currently, users usually use alternative methods. The methods include browsing other users' followings, followers, RT, lists and mentions. RT is a user's tweet by using another user's tweet (Fig.1). List is a grouping function of a user's followings. He/she can group his/her followings by using this function. Mention is his/her tweet to a specific user (Fig.2). These functions give hyperlinks for him/her to find referential users. We presuppose that the user and the user's followings, users who retweets messages which generate interests and users who mention for the user have common interest. Though, there are problems for followings and lists. If we can find a complete list which follows entire users we want to follow, only we have to do is simply to follow the list. However, a list is always customized just for one user who made the list. A user always customizes a list for his/her self. So, we always cannot find a complete list for us. List is usually edited from one user's personal point of view for personal use. In the same way, the users cannot make a complete set of their followings.

In this paper, we extract statistically latent topics that generate followings and lists by using probabilistic model. We assume that users who share same latent topics with a user are worth following for him/her. We also assume that topics are represented by clusters of human relations and categories by users on Twitter.

Fig. 1. Example of RT: kogame5 tweets by using tweet of tryal

Fig. 2. Example of mention: tanichu tweets to kogame5

3 Latent Topics Extraction

In this section, we introduce Latent Dirichlet Allocation (LDA) [6] to extract latent topics which generate followings and lists in a sense of generative model. It was developed to cluster for document, originally. It generates clusters of words and documents based on Bayesian statistical model. It is a generative model that assumes a document is multinomial distribution of topic z which generates word w. Our algorithm clusters users on Twitter using LDA by considering a user as a document. A user's followings' names, names of lists that include the user, names of target users whom the user mentioned, and labels of RT are considered as words contained in a document. The labels of RT are user's name who tweets the original message and number which was put in time-series when the original message was tweeted. Fig.3 represents LDA's graphical

model. α and β is hyperparameter. We use four kinds of Twitter user information to make corpus for using LDA. These are described briefly in the next section.

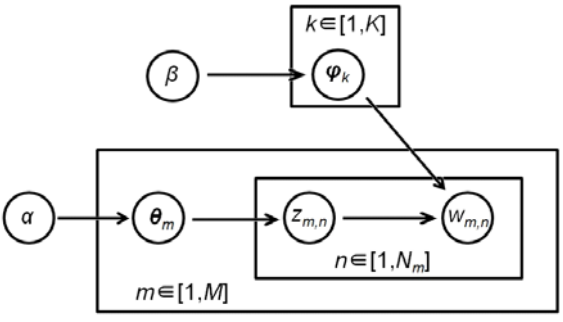

Fig. 3. Graphical model of LDA

3.1 Types of Corpora

We made four types of corpora using four kinds of user information; following, RT, mention and list. Our algorithm extracts latent topics by applying LDA to this corpus.

1. following corpus
2. RT corpus
3. mention corpus
4. list corpus

Following corpus is a set of names of users whom the user follows. It also contains his/her names. *RT corpus* is a set of labels which consist of a user's name who tweets the original message and number which was put in time-series when the original message was tweeted; that is, we do not use text of RT[2]. *Mention corpus* is a set of names of users to whom a user tweets addressing. *List corpus* is a set of names of list which contains the target user and it was made by the other users. We made these four kinds of corpora. By combining some of them, a training data set for a user is prepared. Additionally, we appended a meaningless string to each users' dataset to avoid each user's dataset to be empty. In our experiment, we used Twitter API to get Twitter users' data [7].

3.2 Latent Dirichlet Allocation (LDA) [6]

LDA is a method to estimate latent topics outputting each words by hypothesizing that document is constructed by words which are generated based on *k* topic. Griffiths proposed the method [8] which sample from a mixture model with multiple multinomial distributions based on Gibbs Sampling [9,11,12] with LDA. Thus, using this method, we can sample around global optimal solution without local solution. Distribution (1) can be obtained by a probabilistic argument [8,12].

[2] There are two kinds of RT that are official and non-official. We used non-official RT.

$$P(z_i = k \mid z_{-i}, w) \propto \frac{n_{k,-i}^{(t)} + \beta}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + V\beta} \frac{n_{m,-i}^{(k)} + \alpha}{\sum_{k=1}^{K} n_m^{(k)} + K\alpha} \tag{1}$$

where, z_i is the i th topic; w is a set of words; $n_{k,-i}^{(t)}$ is the number of t th word appeared in the k th topic excluding i th topic; $n_{m,-i}^{(k)}$ is the number of k th topic appeared in the m th document excluding i th document; α is the hyperparameter for θ_m which is a parameter of a distribution which generates probability of the m th document; β is the hyperparameter for φ_k which is a parameter of a distribution which generates probability of the k th topic; V is number of words; K is number of topics; M is number of documents. First right-hand member is probability of i th word in k th topic, second right-hand member is probability of k th topic in m th document.

Gibbs sampling algorithm for latent Dirichlet allocation
while not finished do
 for all documents $m \in [1, M]$ do
  for all words $n \in [1, N_m]$ in document m do
   for the current assignment of k to a term t for word $w_{m,n}$:
  decrement counts and sums: $n_m^{(k)} - 1$; $n_m - 1$; $n_k - 1$;
   multinomial sampling acc. to Eq.(1) (decrements from previous step):
  sample topic index $\tilde{k} \sim p(z_i \mid z_{-i}, w)$
   use the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$ to:
  increment counts and sums: $n_m^{(\tilde{k})} + 1$; $n_m + 1$; $n_{\tilde{k}}^{(t)} + 1$; $n_{\tilde{k}} + 1$
  end for
 end for
end while

Each step of the Gibbs sampling procedure involves replacing the value of one of the variables by a value drawn from the distribution of that variable conditioned on the values of the remaining variables. That is, z_i is replaced by the extracted value from distribution $f(z_i \mid z_{-i}, w)$. This procedure is repeated either by cycling through the variables in some particular order or by choosing the variable to be updated at each step at random from some distribution. Here, M is the number of data, T is the number of step.

3.3 Recommendation Method

Our method calculates distance between two users by comparing the users' probabilistic latent topics. It measures the distance between users' multinomial distributions representing topics they are interested in with KL divergence. It selects potential users recommended to a target user if the KL divergence is small enough. KL divergence from the target user P to a recommended user Q is defined as bellow.

$$D_{KL}(P \parallel Q) = \sum_{x=0}^{T} p(x)\{\log(p(x)) - \log(q(x))\} \tag{2}$$

where, x is the topic index, T is the number of topics, $p(x)$ is a generation probability of the x th topic of a user P. The x th topic is generated from a user P. $q(x)$ is a generation probability of the x th topic of a user Q. A user Q whose distance $D_{KL}(P \parallel Q) < 1.5$ [3] and whom user P doesn't follow is recommended to user P.

4 Experiment

4.1 Experimental Condition

We extracted RT data and mention data from 50 users' TLs data from 2010/6/28 to 2010/7/1, following data and list data from profiles of 50 users selected for an experiment at 2010/7/1. We prepared corpora by mixing following corpus, list corpus, RT corpus and mention corpus. We didn't use RT corpus alone because the amount of RT data is not enough to construct multinomial distributions. Similarity, we didn't use mention corpus alone, either. By using these data, the LDA constructed multinomial distributions of topics for each users and multinomial distribution of words for each topic.Fig.4 indicates an example multinomial distribution of one user.

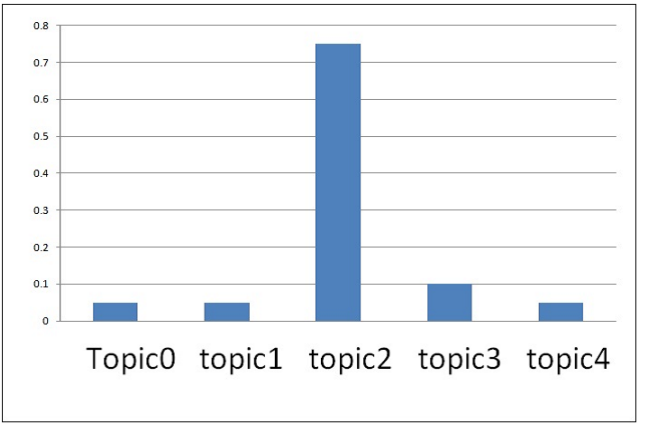

Fig. 4. Example of multinomial distribution of user's topic

[3] Here, 1.5 is decided with heuristic.

We assume that this distribution corresponds to the target user's interests. This figure indicates generation probability of user's topics. For example, this figure shows that topic 2 is generated with the probability of 75%. Our algorithm recommends users based on the KL divergence between two users, we conducted experiments to evaluate effectivity of our proposed method. We selected eight participants from fifty users, we showed TL of recommended candidates to participants. The candidates include *near users*, *distant users* and *half distant users*. The near users have distance of $D_{KL}(P \| Q) < 1.5$. The distant users are up to 2 users of maximum of $D_{KL}(P \| Q)$. The half distant users are up to 3 users of half maximum of $D_{KL}(P \| Q)$ around. We got eight participants to answer three questionnaires after looking through the TLs, followings and lists of the candidates.

Table.1 shows kinds of questionnaires.

Table 1. The three kinds of questionnaires for participants

	Questionnaire	Answer
Questionnaire 1	Do you want to follow this user?	1. I want to follow this user. 2. I may follow this user. 3. I do not want to follow this user.
Questionnaire 2	Are you acquaintance with this user?	1. I am acquaintance with this user. 2. I am not acquaintance with this user.
Questionnaire 3	Had you followed this user?	1. Yes, I had. 2. No, I had not.

We got participants to answer questionnaire 1 “Do you want to follow this user” on a scale of 1 to 3 that are (1)I want to follow this user (2)I may follow this user (3)I don't want to follow this user. We got participants to answer questionnaire 2 “Are you acquaintance with this user” on a scale of 1 to 2 that are (1)I am acquaintance with this user (2)I am not acquaintance with this user. Questionnaire 1 and 2 are asked to evaluate efficiency of this recommendation algorithm. Besides, Twitter users often remove their followings. If a target user had followed and removed some recommended users once, he/she would not want to follow them again. However, we could not get the history, the result contains unrelated factor which we want to take out. Therefore, we prepared questionnaire 3 to know the recommended users are followed or not followed. In this experiment, participants looked twenty four users in average. We prepared four kinds of corpora. Table.2 shows four types of corpora. We conducted an experiment and questionnaire each experimental conditions. In this experiment, we used LDA program “LDA implementation in C++ using Gibbs Sampling” developed and provided by Phan [10].

Table 2. The four kinds of corpora in experiment

Condition names	Kind of corpus
C1	Following
C2	List
C3	Following + list
C4	Following+list+RT+mention

4.2 Experiment a Result

In this section, we show results of the experiments. Fig.5 shows the results of questionnaire 1, this indicates the average of KL divergence with error bars representing standard deviations each experimental conditions. Fig.5 shows distance between multinomial distributions correlate to averages of questionnaire 1's answer. More in detail, participants have tend to answer (1)"I want to follow this user" or (2)"I may follow this user" if the distance between participant and a recommended user is small enough. Fig.6 shows the number of the target users' answer of questionnaire 1 to the candidates in each condition. On $C1$, $C3$ and $C4$, we took similar results. Answer rate of (1)"I want to follow this user" is slightly higher on $C3$ than $C1$ (46% on $C3$, 40% on $C1$). The number of candidates tends to be decreasing on $C1$ because participants

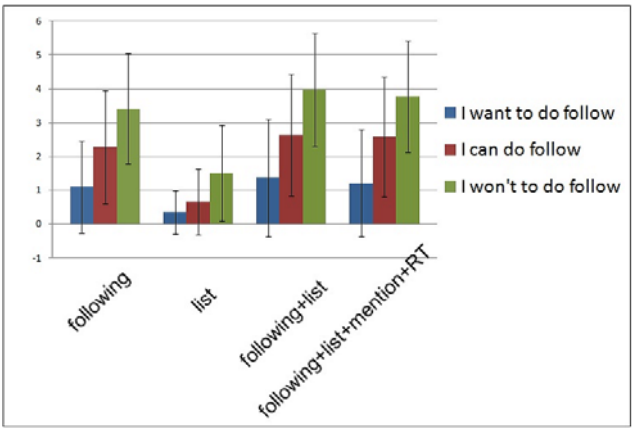

Fig. 5. Relation between result of questionnaire and KL divergence each experimental conditions

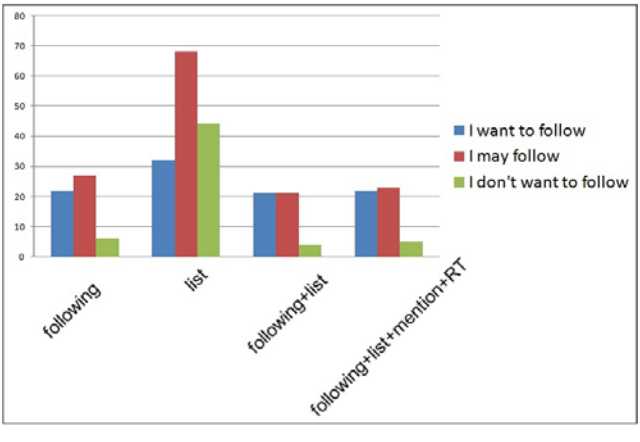

Fig. 6. Number of answer of questionnaire 1 each experimental condition

already followed the candidates. We found that it is not necessarily that recommendation result is interior division blending a number of corpora. This result indicates that a method to blend some kinds of data is important. Skipping detailed result, rate of answer (1) on $C1$ (27%) and on $C2$ (11%) vary greatly. This result indicates that there is qualitative difference between recommended users on each condition. On the other hand, lists are made by a user editing his/her followings. Therefore, the lists are not simple subsets. We skip explaining about questionnaire 3 because there is not discriminative result.

5 Conclusion

In a document clustering with LDA, Blei focuses on a relation between a document and words in the document. In this paper, we replaced the relation with a relation between a user and his/her followings or lists. We extracted multinomial distributions of users' topics by replacing relations between words and documents in LDA with between a user and the user's followings and lists. We found out similarities of users by comparing distance of multinomial distributions of users' topics. However, users' RTs and mentions data that are active information of users are not valuable. The reason why RTs and mentions are not valuable is that an experiment term is short. We obtained good result by recommending users who have a multinomial distribution that is similar to a multinomial distribution of a target user. We treated user's information which has quantitative difference between the user's followings and lists as same quality things. On another front, it is important to combine effectively these informations by considering difference of quality of these informations. To combine effectively these informations, we refer to Multimodal LDA [13] which is proposed by Nakamura et al. Multimodal LDA that treats a number of information as same quantitative information, for concept acquisition.

References

1. Nakano, T., Kamewada, K., Sugito, J., Nagaoka, Y., Ogura, K., Nishimoto, K.: The Traveling Cafe: A Communication Encouraging System for Partitioned Offices. In: Conf. on Human Factors in Computing Systems (CHI 2006) (April 2006)
2. Siio, I., Mima, N.: Meeting Pot: Coffee Aroma Transmitter. In: UbiComp 2001: International Conference on Ubiquitous Computing. ACM, Atlanta (2001)
3. Matsubara, T., Sugiyama, K., Nishimoto, K.: Raison D'etre Object: A Cyber-Hearth That Catalyzes Face-to-face Informal Communication. In: Han, Y., Tai, S., Wikarski, D. (eds.) EDCIS 2002. LNCS, vol. 2480, pp. 537–546. Springer, Heidelberg (2002)
4. facebook, `http://www.ja-jp.facebook.com/`
5. Twitter, `http://www.twitter.com/`
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation The Journal of Machine Learning Research (2003)
7. Twitter API Wiki, `http://www.apiwiki.twitter.com/w/page/22554648/FrontPagetwitterAPI20.txt`

8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101 (suppl. 1), 5228–5235 (2004)
9. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. Chapman & Hall CRC, Boca Raton (2003)
10. Phan, X.H., Nguyen, C.T.: LDA implementation in C++ using Gibbs Sampling, http://gibbslda.sourceforge.net/
11. Bishop, C.M.: Pattern Recognition And Machine Learning, pp. 542–546. Springer Science+Business Media, LLC, Heidelberg (2006)
12. Heinrich, G.: Parameter estimation for text analysis. Technical report (2004)
13. Nakamura, T., Nagai, T., Iwahashi, N.: Grounding of word meanings in multimodal concepts using LDA. In: Proc. IEEE/RSJ International Conference on Robotics and Automation 2009, pp. 3943–3948 (October 2009)