Investigating Topic Models for Social Media User Recommendation

Marco Pennacchiotti
Yahoo! Labs
Sunnyvale, CA, 94089
pennac@yahoo-inc.com

Siva Gurumurthy
Yahoo! Labs
Sunnyvale, CA, 94089
shiiva@yahoo-inc.com

ABSTRACT

This paper presents a user recommendation system that recommends to a user new friends having similar interests. We automatically discover users' interests using Latent Dirichlet Allocation (LDA), a linguistic topic model that represents users as mixtures of topics. Our system is able to recommend friends for 4 million users with high recall, outperforming existing strategies based on graph analysis.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language models*

General Terms

Algorithms

Keywords

social media, user recommendation, topic models, LDA

1. LDA FOR USER RECOMMENDATION

With the advent of social-media, the internet consumer base is switching from traditional Blogging and Email communication to a short text based communication. Unlike a pure social networking, which primarily involves interacting with friends, social-media is more of a broadcast network, where the user wants to spread his message as wide and swift as possible. Hence, users need diverse audience than their friends from the social network. In this work, we present a user recommendation system that represents users as mixtures of topics, and given a target user recommends new friends that have similar mixtures of topics – i.e. shared interests. Topics and user interests are automatically inferred by using Latent Dirichlet Allocation (LDA), as follows.

User-level LDA model. Our model is an adaptation of the original LDA proposed by Blei et al. [1] where documents are replaced by users' streams. In practice, while Blei at al. use documents represented by the bag of words they contain, we use social-media users' streams represented by the words that they emit in the social media (i.e. the words in their tweets). The generative model works as follows (see Figure 1). Given a number U of users and a number K of topics, each user u is represented by a multinomial distribution θ_u over topics, which is drawn from a Dirichlet prior with parameter α. Also a topic is represented by a multinomial distribution β_k drawn from another Dirichlet prior with parameter η. The generative model states that each word position n in a user stream is assigned a topic $z_{u,n}$ drawn from θ_u, and that the word in that position $w_{u,n}$ is drawn from the distribution $\beta_{z_{u,n}}$.

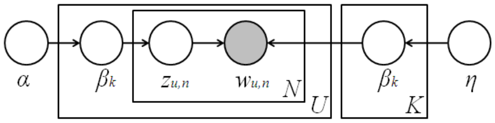

Figure 1: Plate representation of our user-level LDA model.

As the number of users are in the order of millions and their generated tokens are much higher than a typical document, we adopt collapsed Gibbs sampling to learn the distributions. This allows for a compact representation of the model whenever the parameters are large. In particular, we adopt the high performance large-scale LDA parallel implementation presented in [3], which operates on a Hadoop cloud computing architecture of about 1000 machines. This allows us to build a large-scale LDA model, learn topics from the Twitter stream, and find the topic distribution of each user across the topics –i.e. each user is modeled by a topic-vector, where each dimension is the probability to emit the topic.

Our recommendation system uses the LDA output as follows. Given a target user u_t and its topic distribution θ^{u_1}, it recommends to u_t those users that have a distribution highly similar to θ^{u_1}. In detail, given two user distributions θ^{u_1} and θ^{u_2}, similarity is measured by applying either symmetric Kullback Leibler (KL) divergence,or cosine similarity between the users by assuming the topic probabilities as the weights of the vector. Apart from automatic recommendation, our system can be easily adapted to support users to search for other users based on a query.

Pre-processing. We build the LDA model from a repository of about 1.3 billion tweets from April 2010. We discard all users with less than 5 tweets, 5 friends and 5 followers. This step reduces the number of users of more than 70%. We then apply a dictionary-based *spam-filter* to discard spam users, and a dictionary-based *language filter*, discarding users that post most of their tweets not in English, i.e. more than 50% of the words are out of the English dictionary. This further reduces the number of users of about 50%, thus obtaining a final set of 4,050,230 users. Finally, for each user we discard all words that appear in a stop-word dictionary, e.g. most frequent English words, particles, etc. For computational reasons, we max the number of words per user to 20,000. Random word sampling is applied to users exceeding this threshold. Finally, LDA is applied to the set of 4M users, with a dictionary per user capped to 20K, amounting to a total input size of 100GB.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

topic 27	league, arsenal, chelsea, inter, football, liverpool, barcelona, ...
topic 75	tht, wen, sum, luv, knw, nite, hav, gud, rite, lol, wht, jst, ...
topic 96	android, mobile, flash, mac, apps, windows, ipad, iphone, ipod, ...

Table 1: Examples of topics extracted by our LDA model.

2. EXPERIMENTAL RESULTS

Friends re-ranking. In this experiment, we automatically evaluate the LDA model over the existing Twitter graph, similarly to [2]. We create a gold standard by randomly sampling 100 target users; for each target user, we select as positive set 10 random friends, and as negative set 10 random non-friend users. We evaluate the system on the task of re-ranking positive and negative examples for each user: ideally, higher similarity score should be assigned to positive examples with respect to negatives. We use area under the ROC curve (AUC) as the evaluation metric. We experiment different configurations of LDA, by varying the number of topics and iterations (denoted with t and i in Table 2), and by using as similarity measure either cosine similarity or KL-divergence (denoted as cs and kl). We compare to a baseline system that recommends users by computing cosine similarity between the raw tf-idf vectors of the users. We repeat the experiment over three buckets (head, torso and tail) created according to the number of followers of each user, in order to test the systems' behavior.

Overall, all LDA systems outperform the tf-idf baseline with statistical significance, proving our claim that topic models are a good representations of user-level interests. In detail, the best configurations are with LDA and cosine similarity (rows 1-7), gaining +0.20 AUC on the tf-idf baseline, and +0.11 AUC on the LDA configurations with KL, suggesting that cosine is a better measure than KL for comparing topic vectors. Increasing the number of topics does not seem to produce any improvement in performance, while the number of iterations slightly improves AUC up to +0.02, indicating that a small model with 100 topics and 100 iterations may suffice to obtain good accuracy. Results across different buckets indicate that LDA performs better for head and torso than for the tail, suggesting that better topic-models correspond to more active users –i.e. LDA performs well when data sparseness is low.

We indirectly compare our LDA model with the L-LDA model proposed by Ramage et al. [2], where users are recommended using a tweet-level LDA model –i.e. a LDA document corresponds to single tweets instead of the full set of tweets of a user as in our case. On our same task, the authors report that L-LDA does not outperform tf-idf, while we do by +0.20 AUC. We can then conclude that for user-oriented applications, it is much better to adopt topic models at the user-level than topic models at the tweet-level. Using user-level LDA implies a huge reduction in the final LDA space (of a factor corresponding to the average number of tweets per user); but, at the same time, implies the need of more computational resources, in order to keep in memory very large document vectors. As a qualitative analysis, we report in Table 1 some of the topics from our LDA system: some topics are good at defining personal interests (topic 27,96), some at defining users' vocabulary (topic 75). All these aspects play a key role in recommendation.

System	Head	Torso	Tail	*Overall*
lda-100t-100i-cs	0.834 ±0.007	0.832 ±0.012	0.769 ±0.002	*0.811*
lda-100t-200i-cs	0.842 ±0.002	0.839 ±0.009	0.780 ±0.001	*0.820*
lda-100t-500i-cs	0.854 ±0.005	0.837 ±0.003	0.795 ±0.001	*0.829*
lda-200t-100i-cs	0.832 ±0.001	0.825 ±0.003	0.771 ±0.002	*0.809*
lda-200t-200i-cs	0.832 ±0.002	0.845 ±0.011	0.790 ±0.001	*0.822*
lda-400t-100i-cs	0.811 ±0.000	0.834 ±0.000	0.752 ±0.002	*0.799*
lda-400t-200i-cs	0.812 ±0.003	0.843 ±0.000	0.767 ±0.002	*0.807*
lda-100t-100i-kl	0.704 ±0.007	0.664 ±0.001	0.645 ±0.003	*0.671*
lda-100t-200i-kl	0.702 ±0.001	0.674 ±0.002	0.686 ±0.008	*0.687*
lda-100t-500i-kl	0.670 ±0.017	0.652 ±0.000	0.669 ±0.000	*0.664*
lda-200t-100i-kl	0.724 ±0.002	0.741 ±0.000	0.723 ±0.000	*0.729*
lda-200t-200i-kl	0.715 ±0.005	0.721 ±0.018	0.721 ±0.001	*0.719*
lda-400t-100i-kl	0.765 ±0.000	0.768 ±0.000	0.742 ±0.001	*0.758*
lda-400t-200i-kl	0.765 ±0.000	0.797 ±0.000	0.726 ±0.002	*0.763*
tf-idf	0.599 ±0.014	0.638 ±0.005	0.682 ±0.005	*0.612*

Table 2: AUC results of re-ranking experiment.

System	Head		Torso		Tail		*Overall*	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
graph-follower	0.92	0.48	1.00	0.21	1.00	0.21	*0.97*	*0.30*
graph-friend	1.00	0.48	0.81	0.54	0.93	0.50	*0.92*	*0.51*

Table 3: Coverage of graph-based methods with respect to positives of lda-100topic-500iteration-cos.

Comparison with graph-based models. We here provide a preliminary comparison of our LDA best model (*lda-100topic-500iteration*) to methods based on the social graph, largely adopted in existing systems (e.g. Google Follower Finder). Intuitively, graph-based methods have high precision, since graph information are known to be reliable estimators of social influence, but also low recall, due to possible low connectivity. The goal of this experiment is to evaluate how much gain in recall topic-models give with respect to graph-based models. Our intuition is that topic-models are able to recommend users that have similar interests as the target user, but that are not in its network, and thus missed by graph-based models. We compare LDA with two systems: *graph-friend*, recommending users that have many friends in common with the target users, and; *graph-follower*, recommending users that have many followers in common with the target user. The experimental setup is as follows: We select 30 random users (10 for head, torso and tail), and for each of them we sample 3 random recommendations among the top-50 returned by the LDA model. Then, we manually label the recommendations as good or bad, thus creating a gold standard of 90 recommendations, of which 77 are good and 13 are bad (this meaning that our LDA system has an accuracy of 0.86 on the top-50 suggestions). We evaluate both the precision and the recall of graph-based systems over this gold standard.

Results reported in Table 3 indicate, as expected, that graph-based models are highly precise, but miss a big part of the good recommendations that are captured by LDA. These results allow us to draw four main conclusions. (1) Graph-based methods are highly precise but their very low recall indicates that they should be integrated with general purpose topic models: this is our future work. (2) Graph-based models are not effective for target users with few or no connections. Unfortunately, these are the users that usually need recommendations, as they want to establish new friends. (3) Despite graph-based models, topic-models are effective also for users with few connections, as they are only dependent on the linguistic profile of the user. (4) Topic-models can be adopted also for new users that have not yet issued tweets, by asking the user to enter keywords representing his general interests (e.g. 'soccer, Italy, rock'), and then mapping the keywords in the topic-model; this being also future work.

3. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, (3):993–1022, 2002.

[2] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proc. of ICWSM*, 2010.

[3] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. In *Proc. of VLDB*, 2010.