News Recommendation and Filter Bubble

Jianan Yao
Tsinghua University
Beijing, China, 100084
yaojn15@mails.tsinghua.edu.cn

Alexander G. Hauptmann
Carnegie Mellon University
Pittsburgh, PA, USA, 15289
alex@cs.cmu.edu

Abstract

Recently many literatures have studied the problem of rumor detection on social media and proposed various automatic detection algorithms. In this ongoing work report we exploit the power of the crowd and formulate the *reviewer selection problem*, which aim to find reliable reviews for a possible rumor. Our reviewer selection scheme can be considered complementary to existing methods. We give theoretical analysis and provide a greedy algorithm with approximation guarantee. We conduct experiments on a Twitter dataset about rumors, which validates the effectiveness and efficiency of our algorithm.

1 Introduction

Nowadays people increasingly rely on the Internet to learn what is happening around the world. Among tons of stories and pages available online, news recommendation systems provide users with personalized news articles. However, social media and news platforms, seeking to please users, can shunt information that they guess their users will like hearing, but inadvertently isolate what they know into their own filter bubbles. [Par11] Rumors and fake news often propagate within filter bubbles and some argue that they have affected the outcome of the 2016 U.S. presidential election.

Although many articles have investigated the filter bubble problem, little attention has been paid to recommendation systems themselves. In this paper we examine the role of recommendation algorithms on the formation of filter bubbles. In the following two sections we will analyze if and how content-based and collaborative filtering news recommendation algorithms cause the filter bubble problem. Most current recommendation systems use a hybrid model which takes both news content and user feedback into consideration, but it is more reasonable to study them separately to get an insight of the problem.

2 content-based Methods

Most content-based news recommendation algorithms map users and news into the same feature space and calculate their similarity with certain distance metric. The most common approach is to apply an LDA based topic model to generate news representations.

We train the basic Latent Dirichlet allocation (LDA) model [BNJ03] on a public news dataset[1], which includes 142568 articles from 15 media outlets. The topic number is set to 100. We use the Python topic modeling library gensim[2] for implementation. After training we obtain the topic distribution of each news document.

Fig. 1 shows the topic distribution of news related to President Trump from liberal and conservative websites. To visualize the 100-D data we apply Principal Component Analysis (PCA) for dimensionality reduction. It proves that as for topic distributions, there is minor difference between news articles with left-leaning and right-leaning political stance.

We further extract news on two unrelated issues, *climate change* and *border wall*, and analyze the topic distributions of news from different ideological perspectives. The result is shown in Fig. 2.

In Fig. 2, articles with opposing views on "climate change" issue (green and yellow dots) occupy the same region in the feature space, and articles with different opinions on "border wall" have similar properties, which indicates that LDA model cannot distinguish

[1] https://www.kaggle.com/snapcrack/all-the-news

[2] https://radimrehurek.com/gensim/

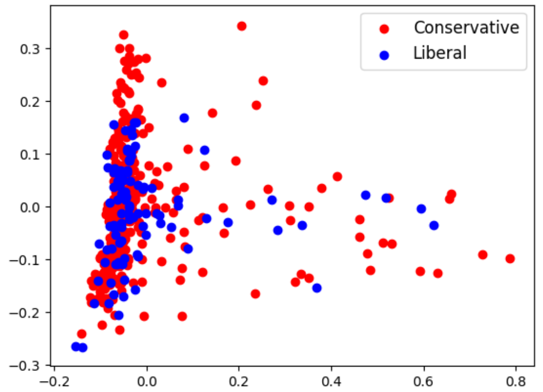

Figure 1: LDA-generated representation of Trump-related news from liberal and conservative media outlets.

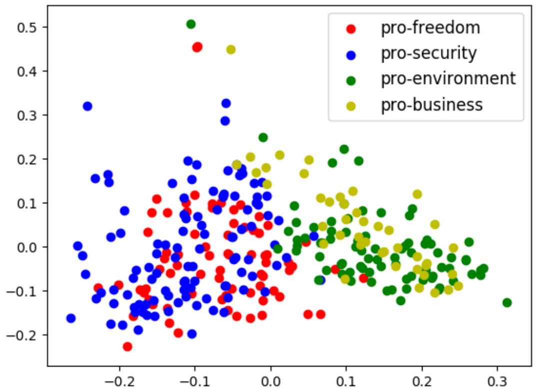

Figure 2: LDA-generated news representations on two issues: climate change and border wall. Two opposing sides of the climate change issue are the pro-environment side which emphasizes the hazard of climate change, and the pro-business side which claims there have already been excessive regulations. Two opposing sides of the border wall issue are the pro-freedom side which fears the wall will sow hatred across the country and pro-security side which prioritizes on stopping illegal immigration.

different opinions on the same issue. A user who has read an article about excessive regulation on business is very likely to get recommendation about climate change. A direct corollary will be pure LDA content-based recommendation systems do not lead to the filter bubble problem.

3 Collaborative Filtering Methods

Collaborative filtering methods proved to be successful in many domains, from movie recommendation to shopping recommendation. Typical collaborative filtering algorithms are built on a user-item-rating matrix, and use two separate feature spaces for users and items. User and item representation can be traditional matrix factorization based vectors [KNK13, Kor08] or neural network based embeddings [WDZE16, HLZ$^+$17].

In this paper, we use the classic LFM (Latent Factor Model) as a representative for collaborative filtering algorithms. [KBV09, MS08] LFM models the preference $\hat{y}_{ui}$ as the dot product of latent factor vectors p_u and q_i, representing the user and the item, respectively.

$$\hat{y}_{ui} = p_u^\top q_i$$

We collect data from Twitter using Twitter standard API.[3] We select 11 popular U.S. news media with different political leanings listed in Table 1. You can refer to Wikipedia pages[4] to learn about the liberal-conservative divide in U.S. politics. For each of them we retrieve the most recent 200 tweets and query for their retweeters. After removing users with less than 3 retweets and news with no retweets, we get a dataset of 7119 users, 2065 news and 55746 retweets (viewed as positive ratings). Since no negative feedback can be retrieved on Twitter, we randomly choose user-item pairs as negative samples. Finally we use Matrix Factorization to generate latent factor vectors for users and news. Here we set number of factors to 5. We also try other latent dimension numbers and they show a similar pattern. LFM is implemented in Python using Surprise[5].

Table 1: List of liberal and conservative news outlets in our dataset.

liberal	conservative
CNN	Fox News
New York Times	Breitbart
The Economist	The Blaze
Politico	National Review
Washington Post	New York Post
	Rush Limbaugh

Fig. 3 shows the topic distribution of news related to President Trump from liberal and conservative websites. There is obvious difference between liberal and conservative news representations, especially when compared with Fig. 1.

[3] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

[4] https://en.wikipedia.org/wiki/Social_liberalism, https://en.wikipedia.org/wiki/Social_conservatism

[5] http://surpriselib.com/

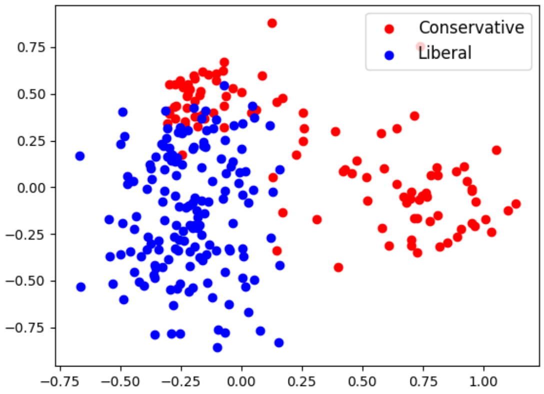

Figure 3: LFM-generated representation of Trump-related news from liberal and conservative media outlets.

To further investigate the situation, we choose two unrelated categories, *international politics* and *California wildfire*, and inspect the news representations from liberal and conservative media. The result is shown in Fig. 4.

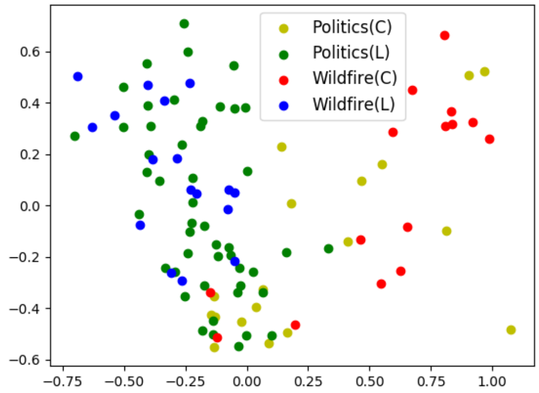

Figure 4: LFM-generated news representations on two categories: international politics and California wildfire. Red and blue dots stand for news about California wildfire from conservative and liberal media outlets respectively. Yellow and green dots stand for news on international politics from conservative and liberal media outlets respectively.

Fig. 4. reveals an interesting phenomenon. From the figure it seems that blue and green come from the same category while red and yellow constitute another category, but that is exactly the opposite. The news represented by blue dots and green dots both originate from liberal news outlets, but they tell completely different stories. So are red and yellow dots. By manual inspection, we find that for California wildfire, nearly all news websites tell exactly the same story and there is no ideological perspective on this event, but news from outlets with different political leaning are still mapped to different regions in the feature space. Articles on California wildfire from liberal media are mapped close to articles on international politics from liberal media. So is conservative media.

Up to now we have found an explanation for filter bubbles. Users tend to read (or retweet for Twitter) news with similar political leaning with their own, and when an article is published it will first be read by like-minded people, which finally leads to a dead lock. News articles sharing common audience will be mapped to the same region in the feature space, even if the articles are about different topics, and then users will be recommended with what like-minded people tends to read, and the filter bubble will be reinforced. Finally we have strong filter bubbles and leave users isolated from different ideological perspectives.

4 Conclusion

In this paper we analyze the role of recommendation systems in filter bubbles. We analyze topic distributions of news under LDA on different issues and from different sources, and discover that typical content-based news recommendation algorithms cannot distinguish different opinions on the same topic. We analyze news representations under Latent Factor Model and indicate that collaborative filtering algorithms tend to map news from the same outlets or with similar political leaning into contiguous regions, thus leaving users in filter bubbles.

5 Future Work

For content-based methods, we only test on original LDA. In our future work we will try topic modeling combined with sentiment analysis and opinion mining.

After figuring out where the filter bubbles comes from, we should consider how to overcome this problem. We need to strike a balance between breaking the filter bubbles and still making users enjoy what they see. A reinforcement learning framework, which learn users' open-mindedness or tolerance on different topics and adjust recommendation policy accordingly, could be a reasonable consideration.

References

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[HLZ+17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[KNK13] Santosh Kabbur, Xia Ning, and George Karypis. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 659–667. ACM, 2013.

[Kor08] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[MS08] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[Par11] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[WDZE16] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 153–162. ACM, 2016.