SI650: Political Leanings Recommender System

SAM EDDS, SEDDS@UMICH.EDU
TEERTH PATEL, PATELTJ@UMICH.EDU

ACM Reference format:
Sam Edds, sedds@umich.edu **Teerth Patel**, pateltj@umich.edu. 2018. SI650: Political Leanings Recommender System. 1, 1, Article 1 (December 2018), 7 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

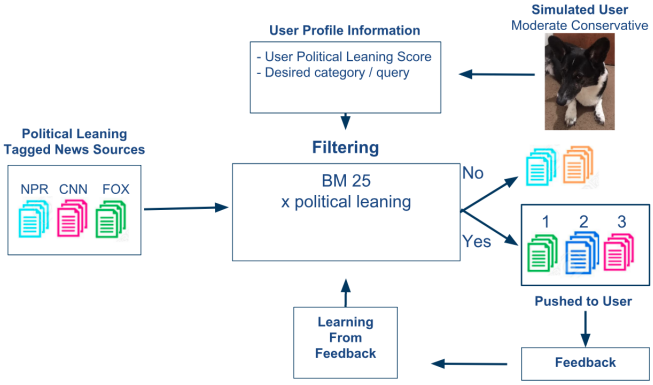
The age of personalized search likely exacerbates political bias, because users are recommended articles based on their past behaviors, which they read, and which cyclically feeds into their next recommendation. We are creating a recommender system for news that incorporates a user's prior known behavior and political leanings, and takes into account a user's feedback. Our system then recommends something the user might not otherwise read. We create simulated users to demonstrate our results, and additionally build an interactive interface to be used in real time, which is attached to our submission (SI_650_Project_Demo.py). Our project can be useful for people to better understand their own political biases.

2 METHODOLOGY

The political leanings recommender system is built using content-based filtering. We first provide a high-level overview of our methods, and each component in our process has a detailed subsection. As shown in Figure 1, we simulate the click logs of 10 users with different political leanings, creating an average political leaning score for each user. That score and their initial broad query are used as inputs for our BM25, which also includes weights for political leanings.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
XXXX-XXXX/2018/12-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Figure 1: Political Recommender System Map



Based on the user's broad query BM25 takes online news articles tagged by category and political leaning, determines the relevant subset by category, and ranks the articles. As mentioned above the political leaning of the source is included as a weight. Results are then filtered to a user based on Precision at k, where $k = 20$.

Our user then provides feedback, marking the articles relevant to them, which is then incorporated as a weight in the next iteration of BM25. Since the user initially provides a very broad category of interest, such as "Business", if they choose mostly "Finance" articles, the next time they ask for "Business" we would rank "Finance" articles higher than other articles.
Finally we asked a few friends to look at our output, with little background on the project, to act as a specific simulated user and have them rank our articles to see if they think our articles recommended are relevant.

3 NEWS SOURCE TAGGING

We scraped 1,019 news articles from 21 news sources [1], assigning each source a political leaning score. For the majority of our sources we utilized the Pew Research Center's Survey on Media Polarization from 2014 [2] in which they categorize respondents into one of the following groups: 'Consistently liberal', 'Mostly liberal', 'Mixed', 'Mostly conservative', and 'Consistently conservative'. Our other sources are ranked based on a different media bias survey.[3] We examine where the average viewer score within the 5 levels (creating three additional sub-levels), and ranked our sources on a scale of -1 to 1, and relative to one another. No two sources were assigned the same score. 1 would be the most liberal possible news source, and -1 would be the most conservative possible news source.

[1] See Appendix A for the political ranking scores. New sources: Al Jazeera, BBC, Bloomberg, Breitbart, CBS News, CNN, Daily Mail, FiveThirtyEight, Fox News, Huffington Post, NPR, NYPost, NYTimes, Page Six, Red State, The Federalist, The Guardian, Town Hall, USAToday, Washington Post, Washington Times, WSJ
[2] http://www.journalism.org/interactives/media-polarization/ The survey responses are based on a scale of 10 political value questions. The overall sample is based on 2,901 respondents, and is considered representative of the 89% of Americans with internet access. While the political landscape has changed in the past few years, we feel it does not change where each news source falls on the spectrum.
[3] http://www.adfontesmedia.com/the-chart-version-3-0-what-exactly-are-we-reading/

4 NEWS SOURCE CATEGORIZATION

We tagged our articles into 5 different categories: Business, Entertainment, News, Political, and Other. The NewsPaper package in Python provided an initial keyword for 535 of our articles[4]. While there is selection bias by source (some sources tagged their articles, others did not), we note it as an augmentation that could be addressed with more time. We use stochastic gradient descent classifier under SVM on 75% our labeled data with 5 fold cross-validation. This is done on the combined title and first paragraph of text.[5] Our accuracy achieved is around 80% and most misclassifications are put into the 'News' category. We then predict on our 584 unlabeled data, which gives us 1,019 total articles.

5 SIMULATED USERS

We simulated click logs for 10 users varying along the political spectrum. Our sources are broken into 8 categories from left to right: Mostly Liberal Left, Mostly Liberal Center, Mostly Liberal Right, Mixed Far Left, Mixed Left, Mixed Middle, Mixed Right, and Mostly Conservative Right. Each click log is 30 articles long, randomly sampled as below. We then create a political leaning score for each user by calculating the mean score across all articles.

(1) Very Conservative User (Only Mostly Conservative Right)
(2) Very Liberal User (Only Mostly liberal Left)
(3) Moderate Liberal User (Mostly Liberal Left/Mid/Right)
(4) Moderate Conservative User (Mostly Conservative Right, Mixed Right)
(5) Middle Ground User (Only Mixed Middle)
(6) Left-Mid Mix User (Mostly Liberal Middle/Right, Mixed Middle)
(7) Right-Mid Mix User (Mixed Right, Mixed Middle)
(8) Left and Right Mix User (Mixed Left, Mixed Right)
(9) Random User (Any)
(10) Reads Everything Equally User (All)

6 BM 25 VARIANT

In order to rank the articles based on relevance to our queries, we implement a BM25 ranking. BM25 gives us a value for each document (article) based on the query term that appears in each document. This function is given by:

$$\text{score}D,Q = \sum_{i=1}^{n} \log \frac{N - nq_i + .5}{nq_i + .5} \frac{fq_i,Dk_1 + 1}{fq_i,D + k_1 1 - b + b\frac{|D|}{\text{avg}|D|}}$$

where fq_i,D is the term frequency of q_i in document D. Our BM25 takes both the term frequency and the inverse document frequency and gives a value for each individual document.

In order to accommodate the political leanings of each individual user and the political rating of each source, we implement the following modification:

[4]Specifically we used section, template top, or keyword in that order depending on availability
[5]Our model includes text feature extraction of cleaned title and first paragraph of text both for counts and TF-IDF. We then tune our model parameters, grid searching to optimize over the number of n-grams, use of IDF, and alpha.

$$w_p = \frac{1}{|p_u - p_d|}$$

where w_p gives the weight assigned to each document based on the user's political leaning, p_u gives the user's political leaning and p_d gives the document's political rating based on source. This modifier allows us to see which documents have the closest values for political leaning to the political rating. As the difference between these values becomes smaller, the values for the modifier becoming increasingly large.

7 FEEDBACK

For this recommender system, feedback is a key component that we utilize to improve the quality of our document rankings, particularly over large iterations. Our system's intuition is that the more the system is utilized, the higher quality our results become. This is because we account for previous information based on each individual user's preferences so we improve with each iteration.

8 EVALUATION

We evaluate our recommender system by Precision at k, where k is the number of documents retrieved. This tells us the proportion of relevant documents retrieved for the first k documents retrieved based on our query. We use Precision at k because it will determine the change in our metrics over several iterations. Since we are using feedback information in our model for our recommender system, we are able to see whether the feedback that we provided through each iteration caused the value of our metric for evaluation to change.

9 INTERACTIVE CODE

In addition to our simulated users code we create an interactive version of our code to be used in real-time. In our simulation we have a set history (SI_650_Project_Simulations.py), which adds to the query list. In our interactive mode (SI_650_Project_Demo.py), a user starts with no political leaning, and just inputs a category/query. After every iteration the model updates results based on previous queries and updated political leaning. These agile results allow a user to better see how our system works and learn from it.

10 RESULTS

For our interactive code, a user can judge their own results. To evaluate how well we did with our simulated results we chose two simulated users and evaluated them on a chosen category. Our users are "Very Conservative User" and "Very Liberal User", with the category being Politics. We asked three people while providing very little project detail to evaluate if they would find this article relevant (if they were a given user, and interested in political articles). They were given output based on the articles seen in Appendix A, Figures 2 and 3, using Precision at k. Table 1 shows the results from our 3 evaluators, each of whom had a different response. This ad hoc method still illuminates that our process initially struggles to evaluate relevant articles. This is expected to some extent, as we have very broad categories. With more time we would

expect to better specify what a user is interested in reading because the user provides feedback. Additionally, because of the time constraints, the "Potential Augmentations" section details issues that could have exacerbated our results, and what could be done to alleviate them.

Table 1: Simulated User Evaluation

Evaluator	Very Conservative	Very Liberal
1	.7	.65
2	.15	.4
3	.1	.15

11 POTENTIAL AUGMENTATIONS

11.1 Selection Bias

Given the time constraints, there are a number of areas which we would augment with more time. One major issue is selection bias from our sources, which greatly impacts our results. As we decided to scrape our data, the sources which label their data versus do not have some inherent differences which means that we will predict well on our labeled data in a train/test scenario, but will have bias in our unlabeled data. If we incorrectly classify categories we will be struggle to recommend relevant documents. With more time we would try to create labels beforehand for our labeled data and learn more the differences in sources to better account for selection bias.

11.2 Location Data

In future iterations of the project we would also include location data. This includes adding a location weight to BM25 based on distance from a user's IP address, weighting articles that take place closer to the user higher.

11.3 Live Draw For Sources

Finally, we would also update our data daily or every few hours so we could recommend articles for a user in close to real-time.

12 APPENDIX A

Table 1: Appendix A New Source Political Leaning Score

Source	Political Leaning	Score
The Guardian	Mostly liberal left	1.000
Al Jazeera	Mostly liberal left	0.950
NPR	Mostly liberal middle	0.900
NYTimes	Mostly liberal right	0.850
BBC	Mostly liberal right	0.750
Huffington Post	Mostly liberal right	0.700
Washington Post	Mostly liberal right	0.650
CNN	Mixed far left	0.400
CBS News	Mixed left	0.250
USAToday	Mixed left	0.125
WSJ	Mixed middle	0.100
Fox News	Mixed right	-0.400
Breitbart	Mostly conservative right	-0.600
FiveThirtyEight	Mixed middle	0.150
Daily Mail	Mostly conservative right	-0.700
NYPost	Mostly conservative right	-0.800
Page Six	Mostly conservative right	-0.900
Red State	Mostly conservative right	-0.950
Washington Times	Mixed right	-0.550
The Federalist	Mostly conservative right	-0.650
Town Hall	Mostly conservative right	-0.850

Figure 2: Very Liberal User Output

	Index	Title	Source
296	1	Jamal Khashoggi case: All the latest updates	aljazeera
601	2	How the 'rugby rape trial' divided Ireland	theguardian
613	3	George W Bush delivers eulogy at his father's funeral – full text	theguardian
607	4	May to raise Khashoggi killing with Saudi ruler at G20	theguardian
606	5	Fifa examining claims of sexual and physical abuse on Afghanistan women's team	theguardian
553	6	In this high-stakes game of Brexit, how much of a gambler are you? | Jonathan Freedla	theguardian
573	7	Facebook discussed cashing in on user data, emails suggest	theguardian
564	8	Washington mourns George HW Bush as Trump gives cold shoulder to Clintons	theguardian
592	9	Missing Emirati princess 'planned escape for seven years'	theguardian
608	10	G20 summit: can world leaders find unity – or is it simply showboating?	theguardian
570	11	May tries to woo Brexit MPs with Irish backstop 'parliamentary lock'	theguardian
617	12	Oscars host Kevin Hart's homophobia is no laughing matter | Benjamin Lee	theguardian
629	13	Facebook documents published by UK – the key takeaways	theguardian
600	14	Hey, that's our stuff: Maasai tribespeople tackle Oxford's Pitt Rivers Museum	theguardian
571	15	Unite leader warns Labour against backing second EU referendum	theguardian
584	16	Full Brexit legal advice to be published after government loses vote	theguardian
568	17	'Brutal news': global carbon emissions jump to all-time high in 2018	theguardian
575	18	Macron calls on community leaders to help end protests	theguardian
612	19	No deal or no Brexit if MPs vote down May plan, says Tusk	theguardian
554	20	Congress is finally pushing the US to withdraw from Yemen. It's about time | Mark Wei	theguardian

(1)

Figure 3: Very Conservative User Output

	Index	Title	Source
1255	1	George H.W. Bush makes his final journey home as Air Force One arrives in Texas	dailymail
1206	2	George H.W. Bush funeral: Final salute for 41st President	dailymail
1246	3	Trump at the center of Mueller's Russia probe after former attorney Cohen admits to LY	dailymail
1218	4	Politicians and dignitaries arrive for George HW Bush's State Funeral	dailymail
1184	5	Putin and Bin Salman share a high five before leader's summit at G20	dailymail
1262	6	Does equality mean women are dying younger?	dailymail
1193	7	New Jersey businessman pleads not guilty to murdering brother and his family	dailymail
1107	8	BBC Radio host killed herself after walking off show	nypost
1134	9	Meet the trainer who made Archie a hunk on 'Riverdale'	nypost
1261	10	Missionary killed on a remote island may have been trying to bring about the apocalyps	dailymail
1209	11	Hillary Clinton ignores Donald Trump as he and Melania arrive for front row seats for Bu	dailymail
1236	12	Kate Middleton looks very festive in tartan at Christmas party	dailymail
1239	13	Meghan Markle attended Michelle Obama's talk in London	dailymail
1232	14	Jeffrey Epstein trial is called off after reaching last-minute settlement	dailymail
1124	15	Americans are into pretty quirky holiday traditions	nypost
1478	16	How Charles Dickens Put A Holly Branch Through The Heart Of The Worst Economics Ev	townhall
1373	17	NeverTrump Clings To Russia Collusion Theory Despite Lack Of Evidence	thefederalist
1190	18	Video of Chris Watts learning he failed lie detector test and then confessing to murderin	dailymail
1125	19	This drug is so dangerous, even dark web dealers refuse to sell it	nypost
1130	20	People drink twice as much alcohol over the holidays	nypost

(2)