Subreddit Recommendation System: A Comparison of ALS and BPR Algorithms

MacLean Hartford, Karolyn Lee, Bill Ma, Isabella Ting

Introduction

Reddit is a social media site on which users can share images, videos, text posts, and links, and discuss these with other members. It is organized into boards known as “subreddits”, which are centered around a single topic ranging from news to hobbies to social discussion and more. Users can subscribe to subreddits to have posts from that board appear on their personalized feed. There are more than 1.2 million subreddits, allowing for a massive variety of content for its 330 million Monthly Active Users.

As Reddit currently offers no in-house recommendation system, our goal is to test two algorithms, Alternating Least Squares and Bayesian Personalized Ranking, in order to create a subreddit recommendation system that can help users discover subreddits they might enjoy but have never interacted with, based on their current interactions with other subreddits.

Data

1. Pushshift API
2. Pushshift Directory
3. Cleaning

1. We originally used calls to the Pushshift API to gather a selection of 5,000 users from the most frequent commenters in the top 100 subreddits in January 2020.

2. After testing this data with our ALS model, we noticed bad results and hypothesized that the data was too sparse to garner meaningful results. This was not enough data to train a well-functioning model, so we augmented and pruned our data with a Pushshift directory of all comments from January 2013. January 2013 was chosen due to its size and computation limits.

3. The final step of preprocessing involved excluding comments from deleted users and known bots. This way, only human users are accounted for and the data better reflects this behaviour. We used the pruned data to track interaction with subreddits for each user, with every comment or post made on that subreddit by a user counting as an interaction. Our final dataset spans 8,181 subreddits and 355,639 users with a sparsity of 99.8%.

Bias in this dataset consists of potentially missed bots and the bias towards frequent commenters found in our original (appended) dataset.

Algorithms

Alternating Least Squares
Using the implicit library’s provided Alternating Least Squares method, we decompose our large matrix of user/subreddit interaction into a user matrix and a subreddit matrix, which we use to recommend subreddits.

Bayesian Personalized Ranking
Bayesian Personalized Ranking aims to tailor recommendations for users by considering the level of interaction of subreddits during training. We feed this into the implicit library’s Bayesian Personalized Ranking method to generate our recommendations.

;

Comparison

We quantitatively and qualitatively compared the two algorithms to determine which method would be best for our recommendation system.

Our chosen evaluation method for quantitative analysis is AUC, presented as the mean AUC across users. We trained our models with 20% of our original non-zero data masked off. For testing we used a binarized version of our data set. In this evaluation setting, false positives are values for which the model incorrectly predicted a positive interaction (the user never interacted with the subreddit), while true positives were correctly predicted (the user did interact with the subreddit). We also offer as comparison the AUC score for suggesting only the 100 most popular subreddits.

Algorithm	AUC Score
Alternating Least Squares	.802
Bayesian Personalized Ranking	.753
Baseline (100 most popular)	.553

As we see from the table above, ALS has a higher AUC score than BPR. This implies that it is a better recommender, as higher AUC means our model is better able to distinguish between relevant and irrelevant subreddits.

As a sanity check, we also wanted to take a look at the outputs of our models. We randomly selected 20 subreddits and based on their ALS and BPR scores and found their top 10 related subreddits.

Subreddit		1	2	3	4
simpleliving	**ALS**	minimalism	Anticonsumption	PhysicGarden	Frugal
	BPR	AskPhysics	ontario	SelfSufficiency	Anticonsumption
animenews	**ALS**	demonssouls	darksouls	Blazblue	anime
	BPR	PERU	MouseReview	AverageMisfires	PSO2

BPR tends to output unrelated subreddits, such as r/AskPhysics for r/simpleliving and r/PERU for r/animenews. ALS seems to be much more consistent with outputs that relate to the topic of the input subreddit. This, combined with its higher AUC score, ultimately led us to choose the ALS matrix factorization for the basis of our recommender.

Recommendation System

To qualitatively analyse our ALS recommendation system, we analysed our data qualitatively by randomly selecting 20 users and creating recommendations for them, determining that the results reflected the performance we expected from AUC scores and model outputs.

User ***fuck_usernames4*** has interacted only with r/AskReddit, a broad/open-ended subreddit so our recommendations are somewhat all over the place. We hypothesize that this is due to how widely-interacted-with r/AskReddit as a general interest and as a default subscription subreddit. This means that every new user automatically subscribes to it. The recommendations fall within our expectations.

fuck_usernames4	
Interacted	**Recommended**
AskReddit	Foofighters
	LetsNotMeet
	shittyadvice
	AMA
	DippingTobacco
	MMFB
	WouldYouRather
	acting
	Assistance
	Swimming

Recommendation System (cont.)

kdmcentire	
Interacted	**Recommended**
Parenting	Mommit
TwoXChromosomes	ABraThatFits
	xxfitness
	femalefashionadvice
	weddingplanning
	Pets
	BabyBumps
	beyondthebump
	breastfeeding
	knitting

User ***kmdcentire***, meanwhile, has interacted with both r/parenting and r/twoxchromosomes, which implies that they are possibly a mother. Our recommender seems to pick up on this subject of interest, suggesting subreddits that make sense for a mother to be subscribed to and falls within our expectations.

User ***unwind-protect*** seems more representative of a typical user: they interact with both generic (r/AskReddit) and specific (r/science, r/ukbike) subreddits, and our recommender makes a good mix of recommendations that mostly seem relevant. r/worldnews and r/explainlikeimfive are more default-subscription subreddits.

unwind-protect	
Interacted	**Recommended**
AskReddit	worldnews
science	explainlikeimfive
ukbike	europe
unitedkingdom	germany
	Health
	YouShouldKnow
	AskUK
	Israel
	askscience
	britishproblems

Though they don’t seem related to r/AskReddit in topic, it’s possible that many users only interact with the default-subscribed subreddits, which is why they are often suggested when given one as input. Subreddits like r/health, r/askscience, and r/britishproblems seem directly related to this user’s specific interests, and despite the few seemingly out-of-place suggestions, our recommender definitely looks to be working well for this user.

Conclusion

ALS-MF and BPR are fairly common and established techniques in the field of recommendations, but our work examines and compares them to create a well-functioning recommender for subreddits, a topic that has not seen as much work as music or movie recommendations.

In the future, we would like to explore potential improvements to our recommender, such as incorporating the actual text of comments/posts into our recommendations using deep learning or natural language processing; we would also like to update our data to better reflect current Reddit trends, as due to hardware limitations, much of our data comes from 2013.

Citations

"AlternatingLeastSquares." AlternatingLeastSquares - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/als.html.
Huang, Lin. "Stacking Collaborative Filtering for Implicit Feedback." doi:10.14711/thesis-b1106722.
Steffen, Freudenthaler, Christoph, Gantner, and Lars. "BPR: Bayesian Personalized Ranking from Implicit Feedback." ArXiv.org. May 09, 2012. https://arxiv.org/abs/1205.2618.
"BayesianPersonalizedRanking." BayesianPersonalizedRanking - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/bpr.html.

Subreddit Recommendation System: A Comparison of ALS and BPR Algorithms

MacLean Hartford, Karolyn Lee, Bill Ma, Isabella Ting

Introduction

Reddit is a social media site on which users can share images, videos, text posts, and links, and discuss these with other members. It is organized into boards known as “subreddits”, which are centered around a single topic ranging from news to hobbies to social discussion and more. Users can subscribe to subreddits to have posts from that board appear on their personalized feed. There are more than 1.2 million subreddits, allowing for a massive variety of content for its 330 million Monthly Active Users.

As Reddit currently offers no in-house recommendation system, our goal is to test two algorithms, **Alternating Least Squares** (ALS) and **Bayesian Personalized Ranking** (BPR), in order to create a subreddit recommendation system that can help users discover subreddits they might enjoy but have never interacted with, based on their current interactions with other subreddits.

Data

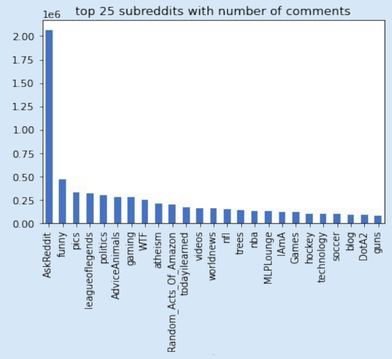
1 Pushshift API → 2 Pushshift Directory → 3 Cleaning

1. We originally used calls to the Pushshift API to gather a selection of 5,000 users from the most frequent commenters in the top 100 subreddits in January 2020.

2. After testing this data with our ALS model, we noticed bad results and hypothesized that the data was too sparse to garner meaningful results. This was not enough data to train a well-functioning model, so we augmented and pruned our data with a Pushshift directory of all comments from January 2013. January 2013 was chosen due to its size and computation limits.

3. The final step of preprocessing involved excluding comments from deleted users and known bots. This way, only human users are accounted for and the data better reflects this behaviour. We used the pruned data to track interaction with subreddits for each user, with every comment or post made on that subreddit by a user counting as an interaction. Our final dataset spans 8,181 subreddits and 355,639 users with a sparsity of 99.8%.

Bias in this dataset consists of potentially missed bots and the bias towards frequent commenters found in our original (appended) dataset.



Comparison

We quantitatively and qualitatively compared the two algorithms to determine which method would be best for our recommendation system.

Our chosen evaluation method for quantitative analysis is AUC, presented as the mean AUC across users. We trained our models with 20% of our original non-zero data masked off. For testing we used a binarized version of our data set. In this evaluation setting, false positives are values for which the model incorrectly predicted a positive interaction (the user never interacted with the subreddit), while true positives were correctly predicted (the user did interact with the subreddit). We offer as comparison the AUC score for suggesting only the 100 most popular subreddits.

Algorithm	AUC Score
Alternating Least Squares	.802
Bayesian Personalized Ranking	.753
Baseline (100 most popular)	.550

As we see from the table above, ALS has a higher AUC score than BPR. This implies that it is a better recommender, as higher AUC means our model is better able to distinguish between relevant and irrelevant subreddits.

As a sanity check, we also wanted to take a look at the outputs of our models. We randomly selected 20 subreddits and based on their ALS and BPR scores and found their top 10 related subreddits.

Subreddit		1	2	3	4
simpleliving	**ALS**	minimalism	Anticonsumption	PhysicGarden	Frugal
	BPR	AskPhysics	ontario	SelfSufficiency	Anticonsumption
animenews	**ALS**	demonssouls	darksouls	Blazblue	anime
	BPR	PERU	MouseReview	AverageMisfires	PSO2

BPR tends to output unrelated subreddits, such as r/AskPhysics for r/simpleliving and r/PERU for r/animenews. ALS seems to be much more consistent with outputs that relate to the topic of the input subreddit. This, combined with its higher AUC score, ultimately led us to choose the ALS matrix factorization for the basis of our recommender.

Recommendation System

To qualitatively analyse our ALS recommendation system, we analysed our data qualitatively by randomly selecting 20 users and creating recommendations for them, determining that the results reflected the performance we expected from AUC scores and model outputs.

User ***fuck_usernames4*** has interacted only with r/AskReddit, a broad/open-ended subreddit so our recommendations are somewhat all over the place. We hypothesize that this is due to how widely interacted with r/AskReddit is as a default subscription subreddit and as the most popular subreddit in our dataset by a huge margin as suggested by our data graph. The recommendations fall within our expectations.

fuck_usernames4

Interacted	Recommended
AskReddit	Foofighters
	LetsNotMeet
	shittyadvice
	AMA
	DippingTobacco
	MMFB
	WouldYouRather
	acting
	Assistance
	Swimming

Recommendation System (cont.)

kdmcentire

Interacted	Recommended
Parenting	Mommit
TwoXChromosomes	ABraThatFits
	xxfitness
	femalefashionadvice
	weddingplanning
	Pets
	BabyBumps
	beyondthebump
	breastfeeding
	knitting

User ***kmdcentire***, meanwhile, has interacted with both r/parenting and r/twoxchromosomes, which implies that they are possibly a mother. Our recommender seems to pick up on this subject of interest, suggesting subreddits that make sense for a mother to be subscribed to and falls within our expectations.

User ***unwind-protect*** seems more representative of a typical user: they interact with both generic (r/AskReddit) and specific (r/science, r/ukbike) subreddits, and our recommender makes a good mix of recommendations that mostly seem relevant. r/worldnews and r/explainlikeimfive are more default-subscription subreddits.

unwind-protect

Interacted	Recommended
AskReddit	worldnews
science	explainlikeimfive
ukbike	europe
unitedkingdom	germany
	Health
	YouShouldKnow
	AskUK
	Israel
	askscience
	britishproblems

Though they don’t seem related to r/AskReddit in topic, it’s possible that many users only interact with the default-subscribed subreddits, which is why they are often suggested when given one as input. Subreddits like r/health, r/askscience, and r/britishproblems seem directly related to this user’s specific interests, and despite the few seemingly out-of-place suggestions, our recommender definitely looks to be working well for this user.

Conclusion

ALS-MF and BPR are fairly common and established techniques in the field of recommendations, but our work examines and compares them to create a well-functioning recommender for subreddits, a topic that has not seen as much work as music or movie recommendations.

In the future, we would like to explore potential improvements to our recommender, such as incorporating the actual text of comments/posts into our recommendations using deep learning or natural language processing; we would also like to update our data to better reflect current Reddit trends, as due to hardware limitations, much of our data comes from 2013.

Citations

"AlternatingLeastSquares." AlternatingLeastSquares - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/als.html.

Huang, Lin. "Stacking Collaborative Filtering for Implicit Feedback." doi:10.14711/thesis-b1106722.

Steffen, Freudenthaler, Christoph, Gantner, and Lars. "BPR: Bayesian Personalized Ranking from Implicit Feedback." ArXiv.org. May 09, 2012. https://arxiv.org/abs/1205.2618.

"BayesianPersonalizedRanking." BayesianPersonalizedRanking - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/bpr.html.