Collaborative filtering recommender system for Reddit

Arnab Borah
aborah@ucsc.edu

Nehal Bengre
nbengre@ucsc.edu

Abstract

We shall be evaluating the performance of Probabilistic Relational Learning,a Statistical Relational Learning Language(SRL) over other SRL Languages and existing proven methods in context to a recommendation system

1 Introduction

Reddit is a large discussion forum where people post content and news.According to current statistics,,around 250 million people are registered users on Reddit. Reddit users can follow,upvote,downvote and post comments.A part of reddit which is one of its most important feature is the SubReddit which is a dedicated subforum on a particular topic.Reddit users currently follow subreddits according to their interest.It would help the users if they get recommended subreddits similar to the ones they already follow.

Statistical Relational Learning is an area of machine learning where relational links between the various objects can be predicted. Statistical Relational Learning methods can be applied to recommend SubReddits to the users based on the their comments on similar SubReddits. We intend to apply PSL(Probabilistic Soft Logic),a first order logic language for our user-based collaborative filtering recommendation system to predict similar sub-reddits of interest to users.We shall be comparing the results thus obtained,to the other established recommendation system methods like Matrix Factorization and other Statistical Relational Learning languages like Markov Logic Network and evaluating as to how PSL performs compared to those,in or context.

2 DataSet

We shall be using Reddits comments consolidated dataset which consist of the fields of author,comments,subredditid,upvotes,downvotes amongst others.The dataset consist of comments ranging from 2007-2016.The dataset was readily available.We shall be using different attributes of the dataset to define the various rules to be incorporated in the model for recommendation. The dataset consist of over 1.6 billion JSON objects pulled from Reddit's API.

3 Approach

Before we describe the approach, we have to represent the data in the form of a graphical model for applying PSL. Hence, Our objective is to convert the data into a directed graph of where subreddits and users are the nodes. The nodes are connected by weighted edges between the nodes. Every user node will have an edge to each subreddit and every subreddit will have a node between each user node. Hence, it is a bipartite graph. The edges will be of type subreddit-user, user-subreddit, user-user, subreddit-subreddit. Our approach is to use four of the attributes :ie:Upvotes,Downvotes,Comments and SubredditIds to define each of the rules.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The first attribute included rule for User-to User Recommendation,where in the idea is that each user would be grouped to a subreddit if he has activity above a threshold(comments,upvotes,downvotes) which indicates that he is active in that subreddit.Common users active in a particular subreddit would be grouped and each user would be suggested subreddits on the basis of the subreddits of the common users.

As a part of the second attribute,we are using the Upvotes feature which indicates that if any user has his comment upvoted beyond a certain threshold,this means that the user is highly positively active in the subreddit.Hence the recommendations of his other subreddits are highly desirable to the other similar users.As a part of the third attribute,we are using the Downvotes feature which indicates that if any user has his comment downvoted beyond a certain threshold,this means that the user is negatively active in the subreddit and the recommendations of his other subreddits should not be recommended to the other similar users. We shall also be considering the sentiment of the comments posted by the user.If the average of the comments posted by the user in any subreddit has high subjectivity and positive polarity,this means that the user has contributed his opinions positively to the subreddit.

4 Model

4.1 Data

The data ,mostly refined,though needed some minor forms of pre-processing. For calculating the similarity of the users,we calculated user-to user similarity using set of similarity functions like Euclidean distance and Cosine distances.

4.2 Rules

We have defined PSL rules of this form:

User-User Recommendation

$$10 : SimilarUsersSimilarity(u1; u2) \wedge UserSubReddit(u1; i) \Rightarrow UserSubReddit(u2, i) \quad (1)$$

This rule captures the intuition that similar users tend to have common interests and hence can be active on similar SubReddits.The predicate UserSubReddit(u1; i) takes a value in the interval [0; 1] and represents the normalized value and UsersSimilarity(u1; u2) is binary, with value 1 if u1 is one of the common related user to u2. The similarities are calculated with similarity measures like cosine similarity and euclidean similarity. UsersSimilarity indicates the user to user mapping based on similarity measures.UsersSubReddit indicates observed mapping between users and subreddits and corresponding target mappings. Currently ,we have given a weight of 10 to the rule.

Recommendation based on Upvotes

$$15 : SimilarUsersUpVotes(u1; u2) \wedge UserSubRedditVotes(u1; i) \Rightarrow UserSubRedditVotes(u2, i) \quad (2)$$

We are still building the rule but,we are going to give the rule a higher weight compared to the initial rule.We shall be giving the rule an initial weight of 15.

Recommendation based on DownVotes We are still building the rule but,we are going to give the rule a higher weight compared to the initial rule.We shall be giving the rule an initial weight of 5.

Recommendation based on Sentiments of Comments

We are still working on defining the model and the initial weight for this rule based on subjectivity and sentiment of the comments.

5 Evaluated model

We ran our model on the sample subset of 2008 year on one subset.It consisted of around 84500 comments.We have currently initialise initial weights and inferred from the model.Following is the evaluation and the observation of our model.

Table 1: Evaluation of the current implemented model

Target user recommendation		
User-reddit recommendation observed	Correctly identified	Accuracy (in percentages)
1514	991	65.455

We created all the possible combinations of 5046 subreddits and 3673 users.We had around 5046 observed user-subreddit mappings.We had around 79433 target user-mappings.We observed a 70,30 split between the training set and the target set of the observed user-subreddit mappings and evaluated our observation based on this In the below image,USERSSUBREDDITS is the mapping between users and potential subreddits

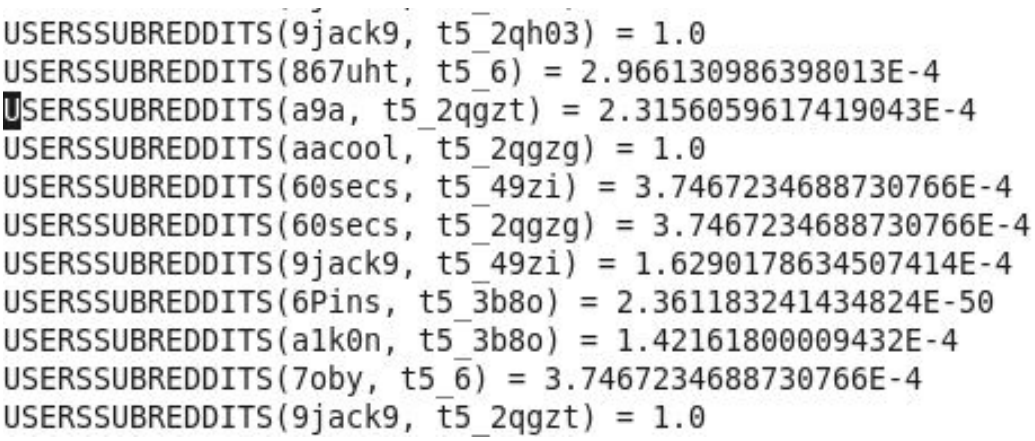

Figure 1: A snippet of our results

6 Timeline

- Week 7:Evaluate based on weight learning.Complete and test the remainder of the 3 rules.ie:Upvotes recommendation,downvotes recommendation and comments sentiment recommendation on the current dataset.Evaluate the results against the results of the standard Matrix Factorization methods.
- Week 8:Scale the model to a larger dataset.Get basic model running on Alchemy,MLN framework and compare the results thus obtained from the PSL Model to that of the MLN.
- Week9:If time permits,we shall work on item based collaborative filtering for our model and compare the results,thus obtained using user based collaborative filtering and item based collaborative filtering.Start work on posters and final report.
- Week 10:Prepare for the presentation and continue work on final report.

We ran our model on the sample subset of 2008 year on one subset.It consisted of 89000 comments.Following is the evaluation and the observation of our model.

7 Potential Issues

Scalability issues :It is possible that PSL will not be able to handle huge data. Hence, at first we will try to create the recommendation system on a small instance of the whole dataset after trimming the dataset small enough that it is a representative of the entire dataset.

References

[1] Pigi Kouki, Shobeir Fakhraei.,James Foulds,, Magdalini Eirinaki &Lise Getoor (2015) HyPER: A Flexible and Extensible Probabilistic Framework for Hybrid Recommender Systems. *9th ACM Conference on Recommender Systems (RecSys)*

[2]Vishnu Sundaresan, Irving Hsu & Daryl Chang. (2014)Subreddit Recommendations within Reddit Communities