AAEW – A recommender system for suggestions in social networking sites

Praggya Pandey  
er.praggyapandey@gmail.com  
Rameshwaran Institute of Technology and Management, Lucknow, Uttar Pradesh

Nancy Vaish  
vaishnancy048@gmail.com  
Rameshwaran Institute of Technology and Management, Lucknow, Uttar Pradesh

ABSTRACT

Online social networks have seen a rapid increment in the number of users across the globe. The giant social networks like Facebook, Twitter, Reddit, Flicker etc do have billions of users and are increasing day by day. A recommender system is an engine which takes a particular user as an input and gives a preferential output which recommends the input user befriend with someone or join a group or get subscribed to an interest or to simply buy a product. To help users every site does have a recommender system. But these recommender systems are mostly based on the fact that the user friends have subscribed to the recommended page/group. In this work, we researched over the Reddit Social Network which is one of the major Social Networking Sites. We engineered a recommender system which recommends the users to follow the more specific Subreddits feeds and is solely based on the current subscribed Subreddits of a user instead of demographic knowledge or Friends knowledge. We used Link Prediction in Social Network as our base and used our novel approach of Weighted Adamic Adar score to find the Subreddits which are closely related to the Subreddits already linked with the user profile. In this way, we predicted the new Subreddits in which the user might have interest. In this work, we addressed the issue of Cold Start for a user who does not have any prior 'like', 'interest' or item purchase history.

Keyword: Social network, Recommender system, Reddit, Subreddit, Betweenness centrality, Cluster, Adamic-adar coefficient.

1. INTRODUCTION

There are millions of Social Networking Site users around the globe. Every day a new social platform takes its root in the arena of social networking. Whichever may be the platform, facebook, twitter, pinster or reddit, the two common essential part of these websites are the Users and the Interests. The whole social network is knitted by these two valuable traits. In facebook, there are Fan Pages and Groups though which users get updates about the topics and items they are interested in. In twitter, the relation becomes Follower and Following and in reddit the relation is called sharing of sub-reddits.

The idea is very basic and so are the social networks. People love to know about the people they like and they more love to know about the different things they are interested in. E.g. programming, baseball, movies, Etc. But people often face problems in finding out the most appropriate source of information sharing and alert gathering. People are left with their own wit and knowledge to find out the correct source of appropriate feeds.

To help users every site does have a recommender system. But these recommender systems are mostly based on the fact that the user friends have subscribed to the recommended page/group. [Reference required]

When a user searches for a particular interest over these sites, the first few results that come up to the screen are the broader interest groups which become over crowded over the time and the user is left joined an interest group which is of lesser interest than the appropriate one

2. LITERATURE REVIEW

Recommender systems (RS) are systems that provide recommendations to customers based on their past purchases, tastes, and preferences (Schafer et al.). Recommendations offered by recommender systems can be as simple as offering a web page (based on
average ratings of web pages) to as complex as providing products in online shopping (by analyzing a customer’s complex click and purchase histories).

In a personalized service, the users are offered customized recommendations based on their stated (explicit) or implied (implicit) needs. MyCDNow can be one of the examples of personalized services in e-commerce sites.

In a non-personalized recommendation service, the same sets of recommendations are offered irrespective of the user who requests the services. Non-personalized recommendations are generated based on what other users have said about the products on an average. Text classification has been one of the key tools that arrange the text information since much time. For more and more subjective information over the net, text classification has a special case known as sentiment classification as the subjective text is becoming very famous in many areas like Data Mining, Information Retrieval etc. It is a complex process which requires more than just text mining techniques. (Qiang Ye et al., 2009)[1] Apply three different machine learning Approaches—Support Vector Machine, Naïve Bayes, and Character-based N-gram model on travel blogs. Experimental results show that standard machine learning techniques perform well on sentiment classification. (ZHU Jian, et al., 2010)[2] also investigate similar task with an individual model (i-model) based on Artificial Neural Network (ANN) to classify movie reviews into positive and negative using sentiment features, feature weight, and prior knowledge base. Supervised machine learning methods are the other related work for sentiment classification that can be found in the paper by (Rui Xia et.al. 2010)[3]

(Minqing Hu and Bing Liu, 2004)[4] Proposed an unsupervised approach to mine opinion features and determine the polarity of opinion sentences. (Qingliang Miao et al., 2009)[5] Proposed a novel ranking mechanism which takes into account the temporal dimension of the opinion. The classification decision at one level in the text influences the decision at another level in the text. The related work is found in the paper by (Yi Hu et al., 2011)[6] based on sentiment classification at document level by modeling and describing topical terms.

Xiwang Yang (Xiwang Yang, 2014)[7] illustrates how recommender system utilizes social network information thus shown some real benefits to users.

Julia Hoxha (Julia Hoxha, 2014)[8] proposed the investigation how to model Web browsing behavior of the people at various multiple websites and based on the investigated model, it predicts future preferences so as to produce relevant recommendations across different domains.

Yoonki Song (Yoonki Song, 2014)[9] Proposed Recommendation systems for software engineering (RSSEs) which are emerging to assist developers in various tasks. Robillard and colleagues defined an RSSE (Robillard et al., 2010)[10]: “a software application program that yields information items approximated to be invaluable for a software engineering task within the given context.” Yue shi (Yue shi, 2013)[11] proposed his recommender system as Unified Recommendation model based on the foundation of the work carried out on over Collaborative filtering. Abhiroop Gupta (Abhiroop Gupta, 2012) [12] proposed Mobile Recommender Systems. Mobile devices, such as cell phones and PDAs, are growing as a major source of information (Nurmi and Floréen, 2004)[13]. Jian Wang (Jian Wang, 2013)[14] proposed the session-aware recommender system in the e-commerce domain. Olsan and Neal (Olsan and Neal, 2015)[15] proposed a method in which they create a network map of the primary interest topics in a social network, for a solution of navigating the social world for finding the specific topic interest. Murata and Moriyasu (Murata and Moriyasu, 2007)[16] gives the link prediction method in a social network which has been based on weighted proximity measures.

3. PROPOSED WORK

In this work, we created a Recommender System for datasets in which there is minimal information available as training data. Also, we proved our approach to be best over the variety of other domains such as Shopping sites and Social Networks like Reddit and Twitter where there is no social structure present in terms of User to User relation. We have used the link prediction technique as the engine in our recommender system. As there are many link prediction techniques but most of them have limitations such that they cannot be applied in real-world recommender systems. Our work is closely related to Adamic-Adar coefficient.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of users</td>
<td>876,961</td>
</tr>
<tr>
<td>Total # of subreddits</td>
<td>15,122</td>
</tr>
<tr>
<td>Average # of subreddits per user</td>
<td>9.69</td>
</tr>
<tr>
<td>Minimum # of subreddits per user</td>
<td>1</td>
</tr>
<tr>
<td>Maximum # of subreddits per user</td>
<td>112</td>
</tr>
<tr>
<td>Average # of users per subreddit</td>
<td>561.8</td>
</tr>
<tr>
<td>Minimum # of users per subreddit</td>
<td>1</td>
</tr>
<tr>
<td>Maximum # of users per subreddit</td>
<td>523.025</td>
</tr>
</tbody>
</table>
Steps to conduct the proposed work

1) Create a graph of the user-subreddit mapping

2) Create subreddit-subreddit mapping by applying criteria similar to the above-mentioned example.

3) Create subreddit clusters using newman-girvan clustering algorithm

4) Find the main actor in each cluster using betweenness centrality.

5) Find users who are creating subreddits for each cluster

6) Find a criterion on which the users will be recommended to post in a subreddit which is a part of the cluster which is already linked with the user

7) Find a criterion on which the users will be recommended to post in a subreddit which is a part of the cluster which is NOT already linked with the user

4. RESULT AND DISCUSSIONS

4.1 Dataset used

The dataset which is used contains data of around 850,000 anonymous reddit user’s interest areas. This dataset is quite unique as it is the largest available dataset of Reddit users.

The data is a collection of anonymous users which are numbered to give them an identity. The numbers are unique in the dataset. Each numbered user is followed by the list of his Sub-reddit interests which are separated by a comma.

In this section, we have shown results and comparative analysis of our approach to previous approaches.

Due to our computational constraint, we selected 5,000 users from the original Reddit dataset. Tail users that are subscribed to less than 10 sub-Reddits are filtered out in the training and evaluation data, which follows the similar pre-processing in related work (Wang and Zhang, 2011) [17] and (Gang and Mong, 2012)[18]. Our final refined dataset contains data of 3,884 users who each has subscribed to a variable number of sub-Reddits.

From this dataset, we created Sub-Reddit to Sub-Reddit mapping which is a weighted undirected graph ‘G’. As per our algorithm, we will add one user each time to the graph G and will link that user as per the training data available with that user. For each user, we clearly partition the sub-Reddit list for that user into two halves. The 50% of the list is used for the training purpose and the remaining 50% is used for the testing purpose.

Each time when we run our recommender system we make the user join the graph G with the first 50% of its listed subreddits. Then on this graph, our recommender algorithm predicts the future links for the user. These recommended future links are the recommendations made for that particular user. The output is in the form of a list of recommendation whose length varies for every user. The number of recommendations made is between 10 to 21.

Chart-1: Data plot of Recall, Precision and F1 measure for AAEW. Data sorted on the value of Recall
4.2 Comparative Analysis

To test our recommender system we compared it with these two coefficients using the similar approach we used for our own system (Jing et al., 2014)[19]. We used the refined data same as we used to collect our values for Recall, Precision, and F1 measure.

Table -2: Comparison of our own recommender system AAEW with the other most efficient coefficient measures

<table>
<thead>
<tr>
<th>#</th>
<th>AAEW</th>
<th>Adamic Adar</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users having Recall value =1</td>
<td>24</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Users having Recall value =0</td>
<td>0</td>
<td>454</td>
<td>3075</td>
</tr>
<tr>
<td>Users having Recall value &gt;=.5</td>
<td>1145</td>
<td>808</td>
<td>9</td>
</tr>
<tr>
<td>Users having F1 measure value &gt;.5</td>
<td>70</td>
<td>351</td>
<td>5</td>
</tr>
<tr>
<td>Users having Precision value =1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table -3: Averaged comparison of our own recommender system AAEW with the other most efficient coefficient measures

<table>
<thead>
<tr>
<th>#</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAEW</td>
<td>0.391756</td>
<td>0.212546</td>
<td>0.2622</td>
</tr>
<tr>
<td>Adamic Adar</td>
<td>0.31892</td>
<td>0.21601</td>
<td>0.236722</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.024266</td>
<td>0.032055</td>
<td>0.025376</td>
</tr>
</tbody>
</table>

Following chart-2 shows the values present in the above table.

Chart-2: The graph showing average values for the collected measure for all three link prediction methods

The overall result for AAEW recommender system is better than Adamic Adar in terms of Recall and F1 measure

5. CONCLUSION AND FUTURE WORK

This research expounds an entirely different approach to solve the problem of discovering people who have particular interests or expertise and furnishing suitable recommendation to a specific user by retrieving his desired information based on his previous interest. Using Python networkX library and Social Network Analysis techniques we engineered our system named AAEW (Adamic-Adar with Edge Weight Coefficient) Recommender System. This system puts forwards a constructing algorithm of Interest Group by uncovering shared interest relationships between people, based on their previous likings or in case of online shopping, their purchase history. The approach we took to construct our recommender system overcomes the problem of Cold Start. Although our algorithm can be applied to many domains for our experimental purposes, we took Reddit user dataset. From that dataset, we chose 5000 users based on which our recommender system forms a base graph. The practice result proves that this algorithm has the characteristics of highly effective group arranging and providing relevant recommendation and further it is easy to be extensible.
5.1. FUTURE WORK

The focus of this section is to outline the direction for future research and the advancements in the proposed framework and techniques. The following subsections present possible further investigations that could be performed and potentially result in improving the techniques so that they contribute to higher retrieval accuracy.

1) Taking the Demographic Information into account

2) Optimizing the Interest-to-Interest linkage algorithm

3) Create own module to collect the live data from sites other than reddit

6. ACKNOWLEDGEMENT

This research paper is made possible through the help and support from everyone, including my parents, family and my best friend. First and foremost I would like to thank God for his unconditional guidance and wisdom as I make my research. Second, I would like to thank my paper guide Er. Nancy Vaish for her full support and encouragement.

7. REFERENCES


