Interest based Recommender System for Social Media

Shivangi Garg
Department. of Computer Science & Technology,
Dr. A.P.J Abdul Kalam Technical University,
Lucknow, India
Infosys Limited, Hyderabad, India

ABSTRACT

In great popularity of Social Media, flourishing massive world of online social communities entrust users to formulate themselves over set of the interests. Users love to know about the people they like and the different things they are interested into. Thus, they require most appropriate source of information sharing and alert gathering. Recommender system is a solution for it. Every site does have a recommender system but these recommender systems are mostly based on the subscription choices of user's friends. 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 remains with the choice to join an interest group which is of lesser interest than the appropriate one. Hence, this paper proposes a interest based Recommender System named as 'MAAC' Recommender System, which recommends the users to follow the more specific interest feeds, solely based on the current subscribed interests of a user instead of demographic knowledge or Friends knowledge. Also, it recommends the interests which are closely related to the interests already linked with the user's profile. The proposed recommender system proves to be highly effective in recommending the various interests to the user, thus incorporating the high percent accuracy in terms of recall, precision and F1 measure as compared to other recommender system.

General Terms

Computer Science- Data Mining, Social Networking

Keywords

Social Network, Reddit, Subreddits, Recommender System, Link Prediction, Degree Centrality, Clustering

1. INTRODUCTION

1.1 Basic Concepts

There are millions of Social Networking Site users around the globe. Everyday a new social platform take its root in the arena of social networking. Whichever may be the platform, facebook, twitter, pinster or reddit, the whole social network is knitted upon the two valuable traits i.e. Users and Interests. Facebook have Fan Pages and Groups through 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 [11].

The idea is very basic and so are the social networks. People more likely to know about the other people and about the different things they are interested into. 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 [1].

To help users, every site does have a recommender system, but these recommender systems mostly recommend page/groups subscribed by the user's friends [20]. 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 to join an interest group which is of lesser interest than the appropriate one [2].

1.2 Related Work and Our Contributions

1.2.1 Recommender System

Recommender Systems use a range of algorithms which return a collection of items to users, based on a derived knowledge of their tastes or from previous interactions. Depending on the domain, the recommendations made can be anything like other user-profiles, movies, books or places to visit.

These recommender systems can broadly be classified into three major categories based on their working principles i.e. Collaborative filtering, Content–based filtering, and Hybrid recommendations [4, 5].

Content-based recommender system consists of three problems: New user problem, Limited content analysis problem, Overspecialization problem which provides inaccurate recommendations [6]. For its solution, recommender systems need the user to had given rating to a sufficient number of items or interests [4].

Collaborative recommender systems also suffer from the same New User Problem (Cold Start problem). *New Item Problem* is the additional problem to it which means that if a new item has evolved in the system then it must have been rated by sufficient number of users [17].

Some of the observations are obtained from the previous studies, which further have been adapted in proposed work. Firstly, the social networking websites such as Reddit, Twitter, Facebook etc. have two common things i.e. Users and Interests. Both of them can be linked to form a heterogeneous social network. Secondly, the SNA (Social Network Analysis) techniques like Link Prediction can be used in making recommendations. Thirdly, a framework can be made by combining the first two observations and that can be extended to form a recommender system. Other than that, using the knowledge of hierarchy of a social network can result in good recommendations [7, 8].

1.2.2 Contributions

This section of the paper is the Proposed work which proposed (contributed) the following points in order to overcome the above mentioned observations (problems):

- Construct an *Interest based recommender system* named as *'MAAC' (Modified Adamic Adar Coefficient)* Recommender System, which recommends the users to follow the more specific interest feeds and is solely based on the current interest of a user instead of demographic knowledge or Friends knowledge.

- The recommender system is not confined to be applicable only to a particular social network.
- The system need to be fast as there are millions of users present on social networking sites and equally present are the Feeder pages.

For e.g. If a user is interested in some particular flavour of Linux then there is no need to recommend the user to get subscribed to the more popular linux flavours like Fedora/RedHat etc. Instead the user might be recommended to subscribe to Linux Community which is a combined community for all the Linux users.

The central part of the proposed recommender system 'MAAC', is the proposal of modelling strategy to find the hierarchy and clusters in the social network which after applying the Link prediction gives us the appropriate recommendations.

Below explained are the overall working and explanation of the proposed system 'MAAC'-Interest based Recommender System.

2. INTEREST BASED RECOMMENDER SYSTEM

The goal of this proposed research is to formulate a recommendation method that takes a general way to make effective recommendations in a broader range of applications. However, there is no comparison in between hierarchical link prediction and collaborative filtering [10].

Proposed work used the advantages given by both the hierarchical link prediction and collaborative filtering. Communities/Clusters of the interests in the dataset are formed, that is based on the hierarchical model and from those clusters which contain the prior interests of the user, the proposed recommender system 'MAAC', recommends few items based on the novel approach of link prediction.

2.1 Reddit Dataset

Proposed system took a case study of *Reddit social network* as explained in [3]. As mentioned earlier, main aim is to develop a hybrid recommender system that has an ability to achieve higher precision than ordinary single component system. In this study, *Figshare user posting behaviour have been examined* which comes with the CSV (Comma Separated Values) [18]. The referred dataset contains data of 850,000 anonymized reddit user's interests. This is a real world social network and largest dataset available for researchers to study. While Reddit is among the largest online social networks but it has been relatively understudied [9].

Table 1. Statistics of Reddit Dataset

STATISTICS	**VALUE**
Total # of users	876,961
Total # of subreddits	15,122
Average # of subreddits per user	9.69
Minimum # of subreddits per user	1
Maximum # of subreddits per user	112
Average # of users per subreddit	561.8
Minimum # of users per subreddit	1
Maximum # of users per subreddit	523,025

The statistics of reddit dataset shown in Table 1, is a collection of anonymous users which are numbered (unique) to give them an identity, followed by the list of his sub-reddit interests (*603, pics, gaming, funny, tattoos, Askreddit, Minecraft*).

Similarly, more users from 604 to 612 are there, a graph created out of these users and sub-reddits can be visualized in Fig.1 (visualized by Gephi Software- https://gephi.org/) [19].

For the proposed system, data for 5000 users has been taken among the whole dataset, and experiment is performed over this 5000 user's data set. A *Social network G = < V, E>* is represented, in which each edge *e* = <*u, v*>, E represents an association between *u* and *v* that existed at a particular time *t(e)*. The representation of the graph for 5,000 users makes **9127 Nodes and 99,666 Edges** out of which 5000 nodes are Reddit User nodes and remaining 2967 Nodes are Sub-Reddit Nodes. The density of the graph is quite high as the number of edges between nodes is high.

3. METHODOLOGIES

Proposed recommender system 'MAAC', has been implemented by using following methodologies:

3.1 Interest to Interest Linkage

A bipartite network named X is created, where Xij = 1 if user i posts actively in subreddit j, otherwise 0. This following network is proposed as a weighted unipartite network named Y as XX0, where Yij denotes the number of users that post in both the subreddits i and j.

Two subreddits are linked if the number of users who post in both the subreddits is above a local threshold value. Similar idea can be applied in case of other social networking sites to form relations between entities. This methodology is implemented using Algorithm 1:

Algorithm 1. Interest To Interest Linkage

```
For every user pair $U_{ij}$      {
Find pair of common subreddit $R_{nm}$ which exist in interest list of both i and j
If found   {
If ( there exists an edge E between subreddits n and m having weight =x)
x=x+1
else
create an edge between both subreddits n and m
weight=1 }
Else
Continue }
```

The graph formed by this algorithm is a weighted undirected graph in which each node corresponds to a subreddit and every edge has weight as its attribute.

3.2 Refining Data

Refining data is to remove the noise present in data before interest to interest linkage methodology and for that; noise has been defined as "The user who posted in less than 10 Subreddits in last 8 months.

It has been found during the research study that out of the selected 5000 users 1115 users has posted in less than 10 subreddits and hence considered as noise. This results in 3885 active users to get shortlisted for the further experiments. From these active users, **7967 Nodes and 92,887 Edges** are obtained**.** 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 [13,14]. Algorithm 2 is used for refining the graph.

Algorithm 2. Refining The Data

```
1.  For every edge e in the graph G     {
2.  If e.weight < .001*G.number_of_nodes
3.  G.remove_edge(e)                }
```

In this proposed system, G.number_of_nodes=9127 which when multiplies with .001 gives 9.127. So, every edge in the SubReddit interest graph whose weight is lesser than 9 will be removed. The final graph obtained by applying all filtering methods contains; *Number of nodes: 1785 Number of edges: 41926.*

3.3 Clustering

Clustering form various groups in social networks depending on various similarity metrics and bond (strong) of connections between the vertices nodes. Clustering is basically applied for identifying the community structures among the network. It lies into two vast categories, agglomerative and divisive, depending on the removal or addition of edges to the network or from the network [11].

In agglomerative method, similarities are computed between nodes pairs and edges. Starting with the nodes pairs having highest similarity, it then added to the empty graph. The same procedure continues and can be halted at any count, and the final components left in the network are the communities.
Agglomerative method has the limitation that it does not take edge removal into consideration therefore, *the proposed approach lies in the second category called divisive method* which starts with the interest network and removes the edges of the least similar connected pairs of vertices. Doing this repeatedly, network is divided into smaller and smaller groups (presented in the form of dendrogram), which is halted at any position and the components involved at the last stage will be the network communities.

The divisive approach works on concept called *"betweenness*" i.e. instead of finding least connected vertex pairs, approach will search for the edges in the network that are connecting and by passing between many pairs of nodes. Such edges will be the strongest in terms of similarity.

Thus, the Algorithm for community structure finding works as-

i. Compute edges betweenness scores in the network.

ii. Find and Remove edge with the highest score from the network.

iii. Recompute betweenness for all other remaining edges.

iv. Repeat step 2.

The clusters are formed by *'Newman algorithm'* [12]. The nodes are coloured on the basis of their cluster numbers. *7 clusters are obtained* from the graph shown in Fig.2, after all filtering and refining as discussed above. The most subreddits corresponds to the general interests like pics, videos, science, technologies, etc, hence form largest clusters as shown in Fig.2.

The nodes having Red, Black and Yellow colour corresponds to such subreddits. It is obvious that the largest clusters are formed by such subreddits as a common man would like to get updates regarding such topics in his day to day life only. All the remaining clusters have nodes with a clear similarity between them. The Blue sub-reddits contains the nodes who are specifically for the Mac/Apple users. The SRS community do have all those nodes that are specifically related to SRS.

Cold start problem is overcome by using clustering and weighted degree. Consider the case, in which a person has only two interests 'baseball' and 'Iphone'. Iphone lies in cluster 3 and baseball lies in cluster 6. So, the proposed recommender system will also recommend the top subreddits from each cluster 3 and 6 who have maximum degree centrality. By using this technique the suggested subreddits are more popular than the one he already likes and also it will keep maintaining the user's existing interests.

3.4 Link Prediction

Link prediction technique has been used as an engine in proposed recommender system. There exist many link prediction techniques like common neighbours, preferential attachment, adamic-adar coefficient, Jaccard's coefficient etc [15]. But most of them have limitations such that they cannot be applied in real world recommender systems. This proposed research work is closely related to *Adamic-Adar coefficient* [16].

Adamic-Adar measures the similarity between the node in question and the non-neighbour nodes present in the graph [16]. However, it does not take weight of the edges into the count. But, the edge weight is an important state for this proposed work which tells how strongly two nodes are connected with each other. Therefore, it can't be ignored, as it needs to be used for predicting the future links more efficiently. So, proposed work need to modify it. Hence, *new modified coefficient has been called as Modified Adamic-Adar coefficient (MAAC),* which also takes edge weights into consideration. E.g. if there are 3 nodes in network x, y, z; x to y has edge weight 1, and y to z is also 1, then for node z, MAAC is defined as 1+1/2=1(sum of the weights of edges that lie between Z to X) / number of Hops).

3.5 Degree Centrality

Degree Centrality is used for recommending the interest to the new user who does not have any input data. It means that no node is connected to the user node. In such cases all the recommender systems which are based on the link prediction technique fails to deliver any result.

Hence, for a new user who joins Reddit and doesn't have any prior interest, proposed recommender system will provide recommendations based on the top elements of different clusters formed during clustering. For every cluster formed, proposed system selects the top 3 nodes found based on the Degree Centrality. In the proposed research, algorithm found

7 clusters in graph G, hence 21 recommendations. This method is the boundary testing for cold-start problem.

4. RESULTS

After applying all the above methodologies and using proposed *modified Adamic Adar coefficient*, proposed recommendation system 'MAAC', recommend other subreddits to the user based on his prior interest subreddits list. For each user, *50% training* data has been given out of whole interest set of the user input to the algorithm and receive out the recommendations linked to those interests.

5,000 users from the original Reddit dataset has been selected. Tail users that are subscribed to less than 10 sub-Reddits are filtered out in the training and evaluation data. And final refined dataset contains data of 3,884 users, each of them has subscribed to variable number of sub-Reddits.

From this dataset Sub-Reddit to Sub-Reddit mapping has been created, which is a weighted undirected graph *'G'*. As per proposed algorithm, one user will be added each time to the graph G and that user will be linked as per the training data available with that user. For each user, the sub-Reddit list has been partitioned into two halves for that user. The 50% of the list is used for the training purpose and the remaining 50% is used for the testing purpose.

Each time whenever recommender system is executed, the user joins the graph G with the first 50% of its listed subreddits. Then on this graph, proposed 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 and 21.

In testing phase, the actual recommendations made are compared with the remaining 50% of the subreddit list. For each user, the data is collected on the following proven measures.

- **Precision** measures the degree of accuracy of recommendations produced by the system.

Precision = Found/ (Found + False positives)

- **Recall** measures the degree of relevant recommendations to the total number of recommendations.

Recall = Found/ (Found + False negatives)

- **F1 measure** evaluates the top most relevant items to a user which is a harmonic mean of precision and recall used in the field of information retrieval i.e.

F1 measure = 2 x Precision x Recall/ (Precision + Recall)

Fig. 3 shows the sample output for the user input '4789'.

The output in terms of Recall, Precision and F1 Measure are collected for each user and Fig.4 shows the graph for it.

The value of Recall, Precision and F1 measure lies between 0 and 1 where higher means better. For user, if recall value is 1 then it means, that all the items present in the testing set for that user have been recommended by our system. Precision get his maximum value of 1 when every recommended item lies in the testing list of the user.

If the user does not have any prior interest, proposed recommendation system will also recommend for that user also. This method is known as boundary testing for cold-start problem, the whole recommendation process is as discussed in section 3.5.

5. COMPARATIVE ANALYSIS

To test proposed 'MAAC' recommender system, it is compared with the two coefficients, Adamic Adar and Jaccard coefficient (explained in section 3.4). And for comparison, same approach and same data has been used which was used for the original proposed system [19]. Table 2 shows the average comparison between the three.
Hence, it is clearly visible by table 2, that MAAC recommender system is giving the best result for recommendations with high accuracy for the above statistics.
Fig.5 shows the graph of table 2. From the graph, it is clearly visible that the overall result for MAAC recommender system is better than Adamic Adar in terms of Recall and F1 measure.

Table 2. Averaged Comparison Of Our Own Recommender System 'Maac' With The Other Most Efficient Coefficient Measures

#	Recall	Precision	F1 Measure
MAAC	0.391756	0.212546	0.2622
Adamic Adar	0.31892	0.21601	0.236722
Jaccard	0.024266	0.032055	0.025376

6. CONCLUSION & FUTURE DIRECTIONS

This paper presents the research for the evaluation and implementation of the MAAC (Modified Adamic Adar Coefficient Recommender system) designed to overcome some of the challenges associated with conventional recommender systems. The key feature of this research system is to provide good recommendations, without needing much data and on the basis of the user's previous interest or no interest. Proposed approach is extensible and does not stick to the domain of Social Networks only.

This new proposed approach overcomes the problem associated with conventional recommender systems, which are based on collaborative filtering or content based filtering. The proposed research proved the results for the cold start problem which is a prevalent problem in recommender systems. It will also prove its applicability on real-life recommender systems which does not have ample data, as used dataset only takes 50% of the user data for the training purpose. Algorithm used in the research work is not processor extensive, i.e. every time, the system need not to load the backend graph again and again. Once the graph is formed at the backend, the new user will connect itself to it with the desired nodes and the recommendations will be made on that basis. Every time the system need not to load the backend graph which is a great advantage, as the system will perform very fast.

However in Future, Demographic Information can be taken into account, Interest-to-Interest linkage algorithm can be optimized and a module can be created to collect the live data from sites other than reddit.

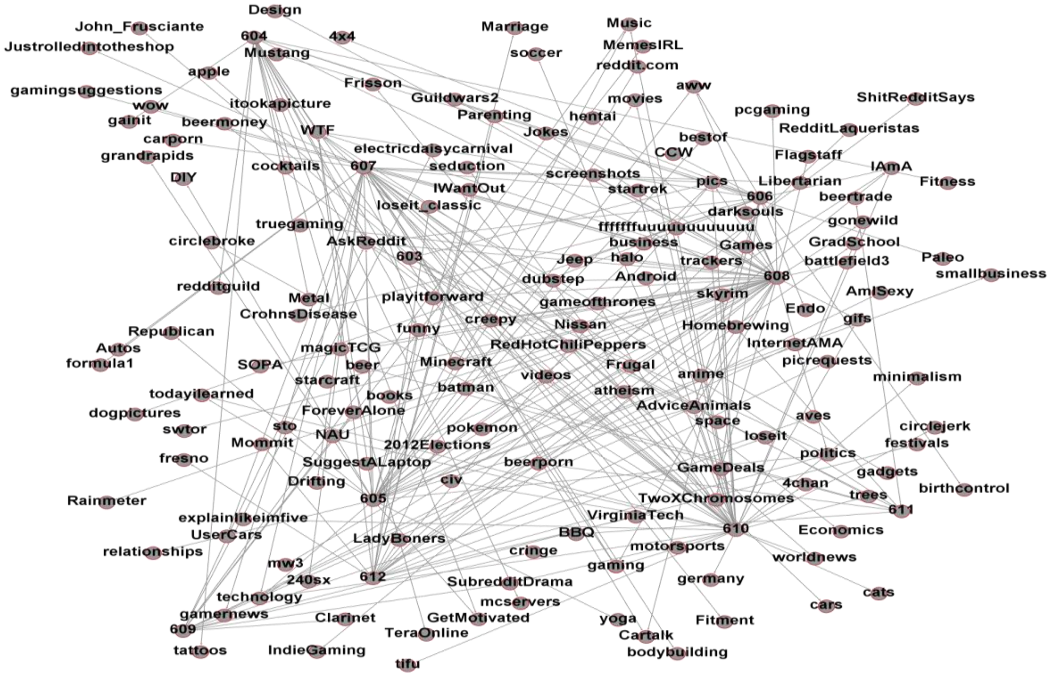


Fig 1: The network of the 10 Reddit users and their subscribed Sub-Reddit

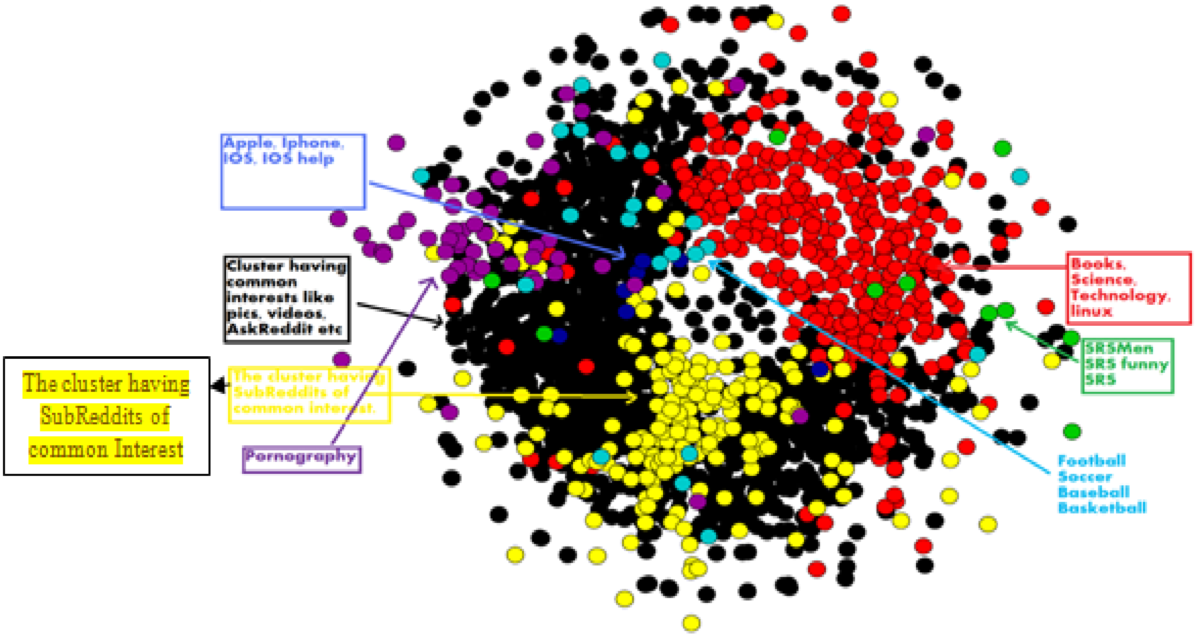
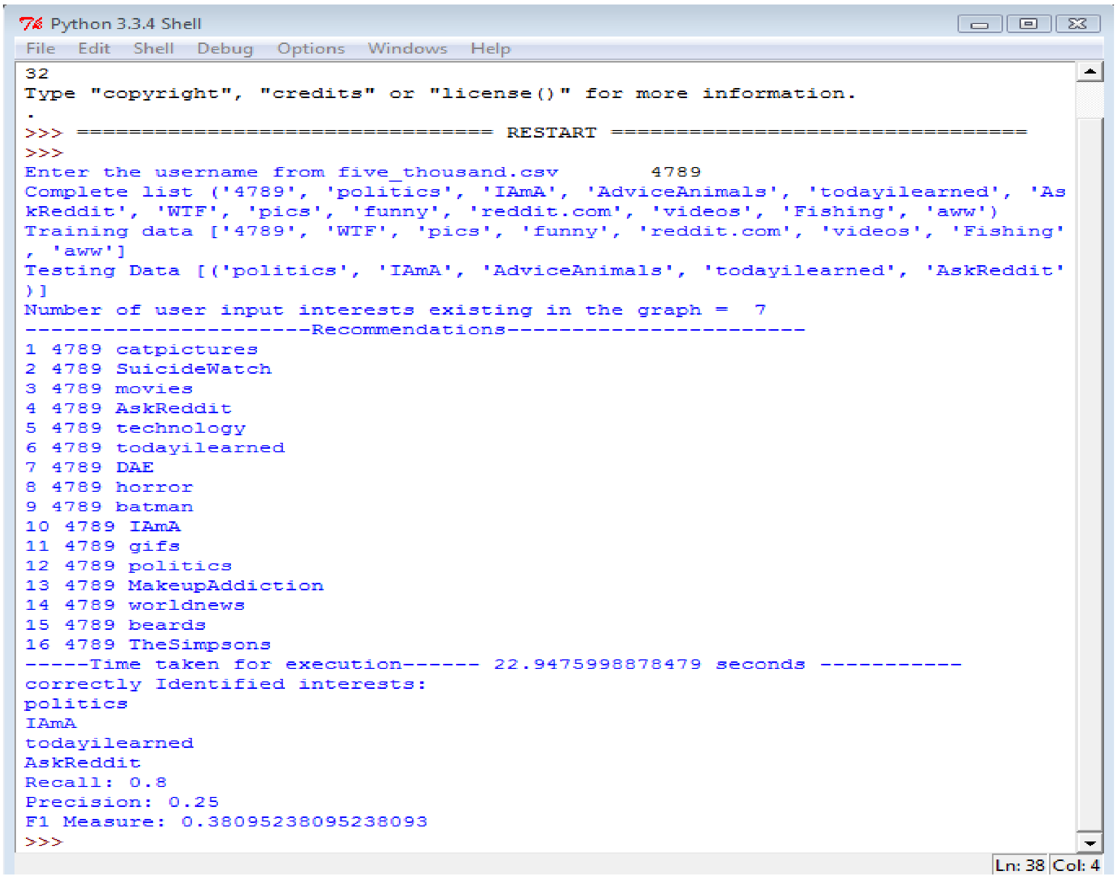


Fig 2: The groups of nodes after clustering. The clusters are shown in different colours.



```
32
Type "copyright", "credits" or "license()" for more information.
.
>>> ================================ RESTART ================================
>>>
Enter the username from five_thousand.csv        4789
Complete list ('4789', 'politics', 'IAmA', 'AdviceAnimals', 'todayilearned', 'As
kReddit', 'WTF', 'pics', 'funny', 'reddit.com', 'videos', 'Fishing', 'aww')
Training data ['4789', 'WTF', 'pics', 'funny', 'reddit.com', 'videos', 'Fishing'
, 'aww']
Testing Data [('politics', 'IAmA', 'AdviceAnimals', 'todayilearned', 'AskReddit'
)]
Number of user input interests existing in the graph =  7
----------------------Recommendations-----------------------
1 4789 catpictures
2 4789 SuicideWatch
3 4789 movies
4 4789 AskReddit
5 4789 technology
6 4789 todayilearned
7 4789 DAE
8 4789 horror
9 4789 batman
10 4789 IAmA
11 4789 gifs
12 4789 politics
13 4789 MakeupAddiction
14 4789 worldnews
15 4789 beards
16 4789 TheSimpsons
-----Time taken for execution------ 22.9475998878479 seconds -----------
correctly Identified interests:
politics
IAmA
todayilearned
AskReddit
Recall: 0.8
Precision: 0.25
F1 Measure: 0.38095238095238093
>>>
```

Fig 3: Output from our MAAC recommender system for user '4789'

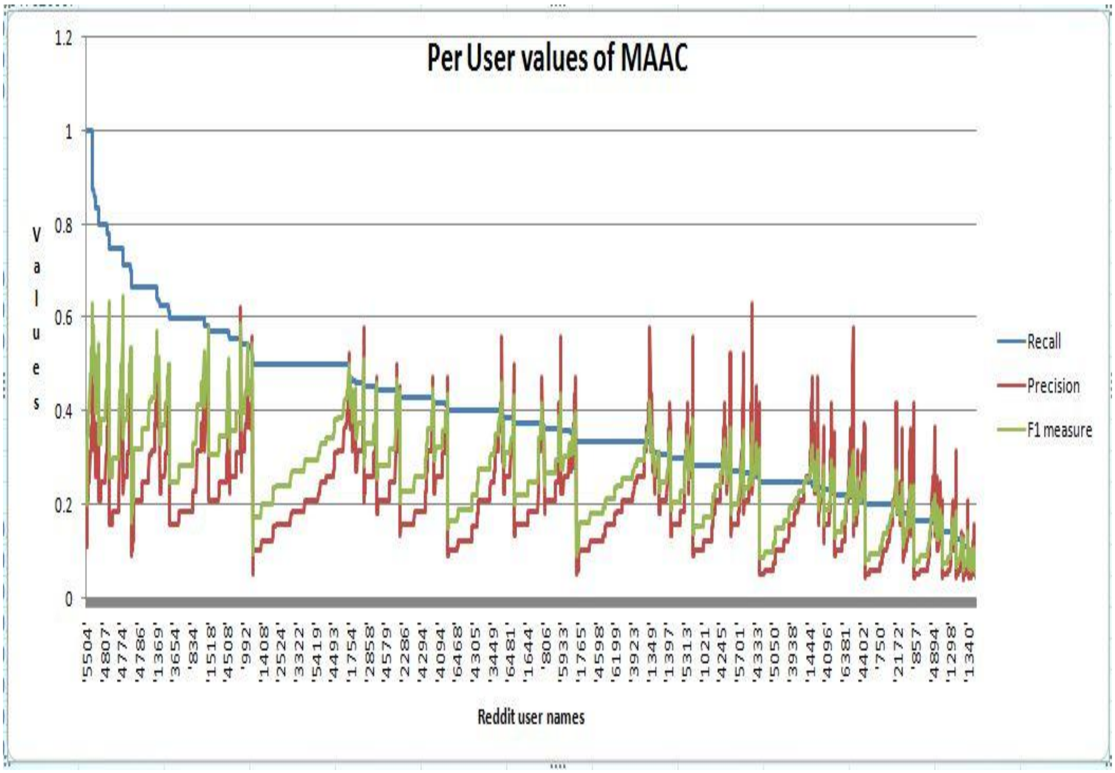


Fig 4: Data plot of Recall, Precision and F1 measure for MAAC. Data sorted on value of Recall

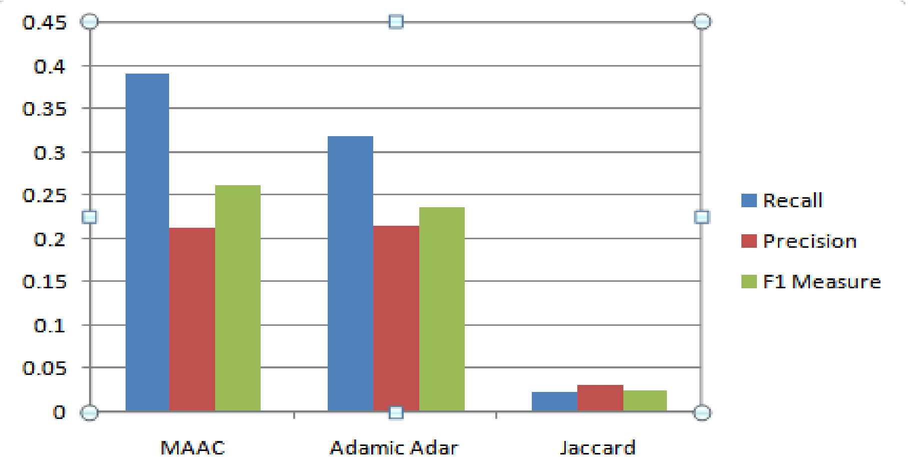


Fig 5: The graph showing average values for the collected measure for all three link prediction methods

7. REFERENCES

[1] Albert R, Jeong H, Barab´asi AL. 1999. Internet: diameter of the world-wide web. Nature401:130–131 DOI 10.1038/43601.

[2] Barabasi AL, Albert R, Jeong H. 2000. Scale-free characteristics of random networks: the topologyof the world-wide web. Physica A 281:69–77 DOI 10.1016/S0378-4371(00)00018-2.

[3] Sanderson B, Rigby M. 2013. We've reddit, have you? What librarians can learn from a site full of memes. College & Research Libraries News 74:518-521.

[4] Pazzani, Michael J. and Daniel Billsus: Content-Based Recommendation Systems. In Brusilovsky, Peter, Alfred Kobsa and Wolfgang Nejdl (editors): The Adaptive Web, volume 4321 of lecture Notes in Computer Science, chapter 10, pages 325-341. Springer-verlag Berlin, Germany, May 2007.

[5] Schafer, J.Ben, Dan Frankowski, Jon Herlocker and Shilad Sen: Collaborative Filtering recommender Systems. In Brusilovsky, Peter, Alfred Kobsa And Wolfgang Nejdl (Editors): The Adaptive Web, Volume 4321 of Lecture Notes in Computer Science, Chapter 9, Pages 291-324. Springer-Verlag, Berlin, Germany, May 2007.

[6] Adomavicius and Tuzhilin: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEETKDE: IEEE Transactions on Knowledge and Data Engineering*, 17, 2005.

[7] Herbert A. Simon. *Administrative Behavior*, 4th Edition, pages 155, 184, 216. Free Press, 4 sub edition, March 1997.

[8] Herbert A. Simon. *The sciences of the artificial* (3rd ed.), pages 5, 216. MIT Press, Cambridge, MA, USA, 1996.

[9] Olson and Neal (2015), Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Comput. Sci. 1:e4*; DOI 10.7717/peerj-cs.4.

[10] Huang Zan, Hsinchun Chen, and Daniel D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):116{142, January 2004.

[11] Scott J., Social Network Analysis: A Handbook. *Sage Publications, London, 2nd edition* (2000).

[12] Newman, M.E.J. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E 69*, 26113(2004).

[13] Wang Jian and Zhang Yi. Utilizing marginal net utility for recommendation in e-commerce. In Proceedings of the *34th ACM SIGIR*'11, pages 1003–1012. *ACM*, 2011.

[14] Gang Zhao, Mong Li Lee, Wynne Hsu, and Wei Chen. Increasing temporal diversity with a. purchase intervals.

a. In Proceedings of the 35th *ACM SIGIR*, pages 165–174, New York,

b. NY, USA, 2012. *ACM.*

[15] Murata Tsuyoshi and Moriyasu Sakiko (2007), Link Prediction of Social Networks Based on Weighted Proximity Measures. *IEEE/WIC/ACM International Conference on Web Intelligence*; DOI 10.1109/WI.2007.52

[16] Adamic, L. A., Adar, E., Friends and Neighbors on the Web, Social Networks, Vol.25, No.3, pp.211-230, 2003.

[17] Ekstrand M. D., J. T. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2):81{173, 2011.

[18] FigShare:http://figshare.com/articles/reddituserposting behavior/874101.

[19] Jing Li, Lingling Zhang, Fan Meng, Fenhua Li (2014). Recommendation Algorithm Based On Link Prediction And Domain Knowledge In Retail Transactions. *International Conference on Information Technology and Quantitative Management, ITQM*; pp 875 – 881.

[20] Ma H., D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender Systems with Social Regularization. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.