Applying Clustering Approach in Blog Recommendation

Zeinab Borhani-Fard [a]
[a] School of Computer Engineering, University of Qom, Qom, Iran

Behrouz Minaei [b]
[b] School of Computer Engineering, Iran University of Science and Technology Tehran, Iran

Hamid Alinejad-Rokny* [c]
[c] School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia
Emails: H.Alinejad@ieee.org and Hamid.AlinejadRokny@UoN.edu.au

***Abstract*—The web has met a significant growth in using weblogs during the recent years. According to the large amount of information in the weblogs, bloggers are facing difficulties to find blogs with similar thoughts and orientations and their popular information. While there is a vast overload of information for blogs, it necessitates having a blog recommender system. Collaborative filtering is a well-known technique in recommender systems. This technique extracts the relations between users and items in according to its neighbor's ratings, and since users have rated just a small part of data, sparsity makes problems for collaborative filtering. This problem leads to an inaccurate comparison among users, and consequently it decreases the accuracy of collaborative filtering algorithms. The use of clustering technique decreases data sparsity and it improves system scalability. We have used clustering to recommend the blog while the blog have reciprocal role, and each blog is both as a user and as an item in the network. In this paper, we use graph clustering based on users' information about social network and we propose blog recommendation framework to get recommendations. Experiments on ParsiBlog [1] data indicated that application of clustering technique with collaborative filtering is better performed that traditional collaborative filtering algorithms, PageRank and etc. A comparison between PageRank algorithm and clustering application showed that graph clustering in recommender system could makes better results in terms of accuracy, quickness and scalability.**

***Index Terms*—Blog networks, Collaborative filtering, Hybrid recommendation system, Graph clustering.**

I. INTRODUCTION

During recent years, blog have changed into a remarkable social media on the internet that enable users to broadcast content on the web consisting thoughts completely personal or Private. Facility of blog contents broadcast likewise willingness for thoughts development is becoming to promote blogs fast and continuously growth. Nowadays there are hundreds of million blogs all over the world that still being increased quickly. A blog is a website consisting data entries (so-called post) having reverse date sequence, and is written and maintained by a blogger who uses a specific tool. Since each blog or blog entry may have links to other blogs and web pages, blog link structure can be considered as a social network.

Recommender systems apply some ideas of users groups to help this individual efficiently to identify their favorite topics amongst vast options. Techniques are divided into three types, content-based recommender system, collaborative filtering recommender system and hybrid recommender systems [1]. Collaborative filtering systems provide the recommendations based on ratings by users set to the active user. Content-based recommender system uses items features (like movie director, actors, etc.) to get recommendations. Hybrid techniques generate recommendations with combining CF methods and content-based recommending methods.

Methods in Collaborative filtering can be divided to memory-based, model-based and hybrid [6]. One of memory-based CF problems is that it must compute similarity between each user (item) with all other users (item) to define their neighbors. This problem is not working in social network or blog recommendation that have equal items and users and numbers of users are very large. To cope with traditional CF technique or memory-based CF weak points, we applied clustering approach to gain more precision, speed and efficiency. Using clustering techniques reduce data sparsity and improve systems scalability because similarity computation is performed only for the users of the same cluster.

doi:10.4304/jetwi.5.3.296-301

Computation of costly and complex clustering is performed off-line. Using clustering methods in the model needs to once more clustering graph and update the model now and then.

Blog recommender system differs from other recommender systems, in several ways. First, the goal of recommending of product, movie, music, news, web page, travel and tourism for all kinds of services, electronic sale and even virtual community is different. It is important to find features of recommendation goals, whereas inappropriate use of recommendation may reflect negative effects. Second, blog recommender system is a provider, and in contrast with meanings, bloggers are dynamic and recommendation changes quickly and blog recommendation mechanism must be more adjustable and flexible than the rest. Blog are human-oriented in other words, blog content are highly subjective and mind-oriented to recommend [4].

This paper is organized as following. Related works for blog recommender system and clustering application in recommender systems are provided in section 2. Section 3 deals with blog recommendation framework in detail that we proposed based on clustering approach. Experiments evaluation and results of applied framework and comparison with other methods are show in section 4. The paper ends up in our conclusion and objectives for future actions.

II. RELATED WORKS

Because of massive content provided by the blogs, and since most bloggers are non-professional users with difficulties for finding their suitable and favorite blogs, blogs recommending systems have recently attracted researchers' attentions.

In some aspects, meaning of blog ranking is similar to blog recommendation. Abbasi et.al[8] used a personalized PageRank method for blog recommendation . Fujimura and et.al attributed some scores to each blog entry via weighing in based on authority and hub scores on the basis of eigenvector computations [13]. Our study is related blog recommender system and network clustering.

In blog recommending systems domain, different studies were performed both on the basis of blogs content and blogs social network. In Hayes and et.al research, the analysis is performed on the type of suitable recommender strategy for blog, which in their study is applied tags, post subject for blog recommender system [9]. A blog recommendation mechanism is offered in [4] that combines trust model, social relation and semantic analysis. Garc á etal. [3] provide a framework to connect data semantic to web pages links on the basis of special ontology. A blog recommender system that called iTrustU is being offered based on collaborative filtering and multi-facet society [15]. A personalized recommender system is offered in Hart et.al [14] based on tags.

Clustering methods are used in several CF recommender systems to reduce dimensions, data sparsity and to increase scalability. A CF system based on k-means clustering is applied to cope with data sparsity [17]. A CF proposes on the basis of iterative clustering method that extracts internal links of users and items [10]. In this model, users and items are clustered by k-means to solve scalability problem, one part of items are selected by experiencing different clustering algorithms then recommendation are observed separately.

III. BLOG RECOMMENDATION FRAMEWORK

Our proposed blog recommendation framework is explained in this section. We are done the steps of this framework on the ParsiBlog data which is one of the blog hosts in Persian. We produced blog directed graph based on the blogs that each blogger has indicated in his blog roll as favourite blogs. We select favourite blogs in Parsiblog domain. Parsiblog graph has 21305 nods and 257316 Edges. Figure1 shows our blog recommendation framework consisting of two main phases; data preparation and model implementation.

A. Data Preparation

Data preparation contains three stages: data pre-processing, network clustering, data post-processing.

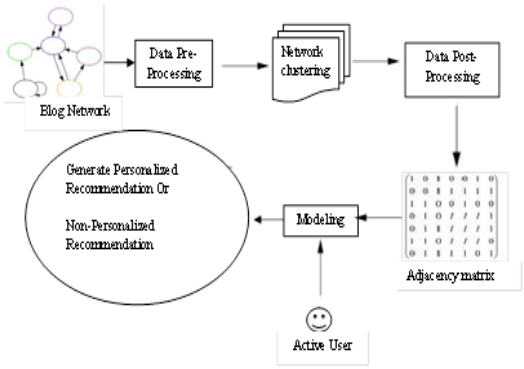

Fig.1. This is the caption for the figure. If the caption is less than one line then it needs to be manually centered.

Pre-processing

We have omitted nodes with no outgoing links in network or in other words with zero out degree, because it's not possible to have any recommendation for these nodes.

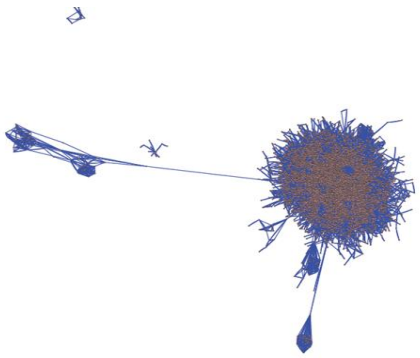

Fig.2. Social Network for Parsiblog

Most networks consisted of strong connected components that have a component with too many nodes. To reduce data sparsity, we can select only strong connected components with bigger nodes numbers. In this work, we selected strong connected components at least with 10 vertices. So, in this stage the derived graph consists of 9065 nodes and 222216 Edges. The biggest strong component has 8933 nodes. Figure 2 shows the blogs social network.

Network Clustering

Clustering is very important stage in our study, because it determines neighbors of active user. All of the next computations depend on clusters. It is important to select suitable clustering algorithm, since using different clustering techniques will have different results, and using a specific clustering algorithm may even decrease recommendations precision.

Cluster is a collection of data object that the members of the same cluster are similar and differ from other cluster members. Clustering methods are divided into three groups: partitioning method, density-based methods and hierarchical methods [7]. Term of cluster in graph is also called community in some papers. Today one of the main network subjects being mostly noticed and studied is communities' structure in network, and its goal is collecting vertices in to groups; so these groups will have larger density of edges inside the groups among the others. There are different algorithms to find such communities. During recent years, new algorithms are proposed. Newman and Girvan proposed an algorithm using Edge Betweenness as a metric to identify communities' boundaries [5]. The algorithm complexity is O(m3) on sparse graphs, while regarding available hardware. It limits the algorithm application to the networks having at least thousands of nodes.

We have used FastGreedy algorithm proposed by Claust et.al [13]. Algorithm complexity in the worst case is O(mdlog n) that d indicates depth of dendrogram, and m shows number of edges and n show number of vertices in the network. But algorithm complexity is O(nlog2n) for sparse graphs. Since most blog networks are sparse graphs so algorithm will be performed in linear time.

$$\mathbf{Q} = \sum_{\mathbf{i=1}}^{\mathbf{q}} (\mathbf{e_{ii}} - \mathbf{a_i^2}) \qquad (1)$$

$$\mathbf{a_i} = \sum_{\mathbf{j=1}}^{\mathbf{q}} \mathbf{e_{ij}} \qquad (2)$$

With $\mathbf{e_{ij}}$ consists of edges that connect vertices of community i to vertices of community j and q show number of clusters. $\mathbf{e_{ii}}$ consists of edges that connect nodes of cluster i to each other. This algorithm is a greedy implementation of hierarchical clustering algorithm. Algorithm consist of finding changes in q magnitude which is obtained merging each couple of communities and selecting the biggest one and at last doing the related mergence. Empirically, modularity more than 0.3 is a good index for suitable community structure in a network [2].

We have used igraph package [11] in R open source software for clustering implementation. We identified 192 clusters in Parsiblog graph and modularity amount was 0.372.

Post processing

This stage consists of clusters refinement to increase model accuracy. In this stage we have identified clusters with very few members as an outlier and omitted them.

We selected clusters at least with 50 members. Having done such operations on the clusters, clusters number declined in to 6 and there were 8435 nodes in the network. Table 1 shows the general information about features of primary blog graph, blog graph after pre-processing and after post- processing to be compared. With comparing network features, we can see that number of graph edges in each stage have not any remarkable reduction, but density, clustering coefficient, graph degree are increased. Figure 3 shows the distribution of clusters size (magnitude) and distribution of strong components size.

TABLE1.

THIS IS THE CAPTION FOR THE TABLE. IF THE CAPTION IS LESS THAN ONE LINE THEN IT IS CENTERED. LONG CAPTIONS ARE JUSTIFIED TO THE TABLE WIDTH MANUALLY.

	Vertices#	***Edge #***	***Degree Avg.***	***Density***	**Clustering Coefficient**
Initial Network	21305	257316	24.1554	0.0005669	0.31747
Pre-Processing Network	9065	222216	49.0272	0.0027042	0.37995
Post-Processing Network	8435	218207	51.7384	0.0030669	0.37663

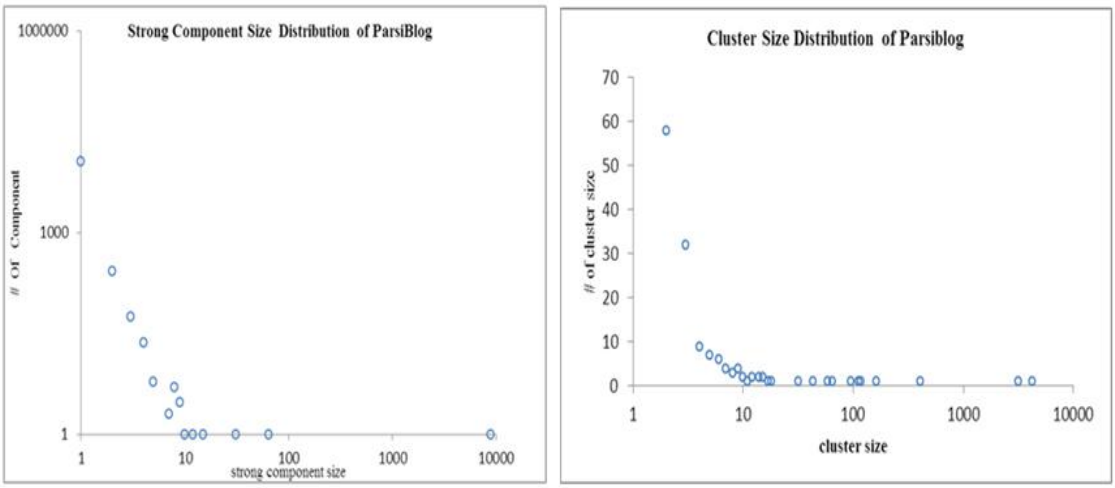

Fig.3. Strong component and cluster size distribution in Parsiblog

B. Model Implementation

Model implementation consists of two steps; model construction and generate recommendations. To implementation the model, one must convert data set to an adjacency matrix based on directed graph of blogs relations.

$$\boldsymbol{A[u,v]} = \begin{cases} \mathbf{1}, & \boldsymbol{if\ u\ links\ to\ v} \\ \mathbf{0}, & \boldsymbol{else} \end{cases} \qquad (3)$$

We used blogs link (blog roll) as the bloggers' favorite items rating, in this paper. So item-blog matrix is a asymmetric, square and binary one in which number of users and items are equal, and each blogger is an item and also a user. Each blog external links list shows the items preferred by blogger. Adjacency matrix in Parsiblog network has 8435 rows and 8435 columns.

Model construction

Development and design of models such as machine learning and data mining algorithms provides the system with learning opportunity to identify complex patterns based on train data, and then create intelligence recommendation for test data or real world data which is based on learner's models. In model construction, we predict a class to which each blogger belongs.

To construct the model, we divided adjacency matrix or data set into 70% train data obtained in the graph to classify 6 main clusters. We used C5 algorithm in Clementine software for classification. Accuracy of train data set was 81.03%. Test data set will be used for efficiency evaluation and accuracy of recommender system. Mean accuracy of test data was 79.60% after repeated practices.

Generate recommends

Generating recommend actions can be done in personalized and non-personalized format.

In non-personalized recommendations, some cases are recommended to the blog regardless of his characteristics that the most famous method is on the basis of ranking. Generally, there are three suitable approaches (input degree, HITS, PageRank) to rank nodes on the network. For each blogger regarding the cluster to which he/she belongs, we recommend the blog the k-Top highest rank node (blog) in that cluster.

In personalized, it's better to use personalized information to recommend the user. One of most well-known personalized recommending methods is collaborative filtering. We recommend each blog regarding the cluster to which he/she belongs an N-Top recommendation by collaborative filtering. The advantage of this method to traditional CF is that there is no need to compute active user similarity with the whole network users, and computing the users' similarity of the same cluster is enough.

We used PageRank algorithm to generate non-personalized recommendation and we also used collaborative filtering algorithm on the basis of a memory-based collaborative filtering method. In this method, we use cosine similarity standard to compute similarity. At the end of this section, Page Rank algorithm and collaborative filtering method based on neighborhood as our experiments basics is introduced. We named get of non-personalized recommendation as clustPR and get of personalized recommendation as clustCF.

PageRank algorithm: This algorithm is the most well-known link analysis algorithm offered in 1998 and it was

applied in Google search engine [12]. Assigning weight to each page, the algorithm sorts out search results based on the weight. Suppose that a random walker is searching through the created graph by Internet pages. Entering each site, the walker selects each of outgoing links with equal probability.

So, different pages with different weight would be seen. The main and valid page in PageRank is the one to which other valid and important pages offered link. This criterion indicates the popularity of each page through the whole graph, and it can be defined recursively as follows:

$$\boldsymbol{pagerank(u) = \frac{1-p}{N} + p \sum_{v \in S_u} \frac{pagerank(v)}{outdegree(v)}} \quad (4)$$

P is damping factor that in most cases it equals 0.85. $\boldsymbol{S_u}$is a set of all pages linked to page u and outdegree(v) shows the whole output pages of v.

Memory-based collaborative filtering: Memory-based CF algorithms use all or a sample of user-item data to create a prediction. Each user is a part of a group of individual with similar interests. Priorities predications in new items are produced for the blogger by determining what a new user's neighbor is nominated [6]. Neighborhood-based CF algorithm [16] is a memory-based CF algorithm, containing below stages:

Computing similarity or weight $\mathbf{w}_{i,j}$ between active user/item i and active user/item j. Neighborhood formation: selecting K item/user having most similarities with active item/user. Offering N-Top recommendation by weighed-in average neighbors' item/user is obtained. There are different methods for similarity or weight computation between users and items such as Pearson Correlation, cosine similarity, etc.

IV. EXPERIMENTAL EVALUATION

To evaluate clustCF and clustPR, we compare them with PageRank algorithm and traditional collaborative filtering algorithm.

A. Data Set

Evaluation is applied on data set of Parsiblog graph. Construct of Parsiblog graph described in pre-processing in section 3. Pre-processing data are used for PageRank algorithm and Traditional CF algorithm, Post-processing data is used for clustCF method and clustPR method that we propose in generate recommends in section 3. ClustCF is for personalized recommends and ClustPR is for non-personalized recommends.

B. Evaluation Metrics

To evaluate our offered framework; we applied recall and precision metric which were defined in information retrieval. These metrics are defined in blogs as follows:

$$\boldsymbol{Precision = \frac{|Favorite\ blogs \cap Recommended\ blogs|}{|Recommended\ blogs|}} \quad (5)$$

$$\boldsymbol{Recall = \frac{|Favorite\ blogs \cap Recommended\ blogs|}{|Favorites\ blogs|}} \quad (6)$$

C. Experimental Results

To perform such an evaluation, we selected 1000 nodes randomly in PR methods and traditional collaborative filtering (CF), and we computed recall and precision average. We obtained 20-Top recommendations for each user of this candidate set in CF algorithm, and we computed average for recall and precision.

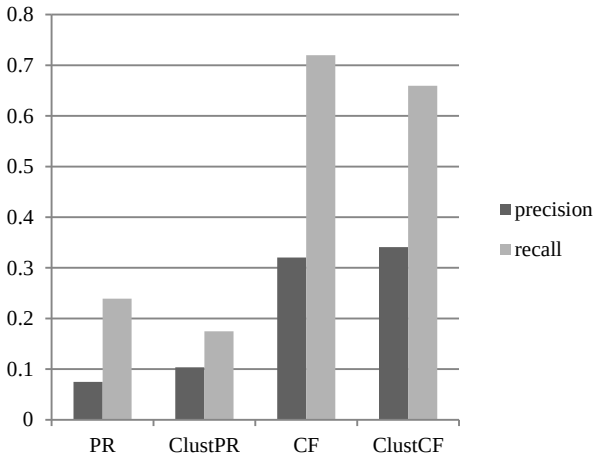

Fig.4. Comparison of algorithms (precision and recall)

We computed recall and precision average for 10 test data set in clustPR and clustCF methods, and then we computed total average for recall and precision. Figure 4 shows average of recall and precision for four algorithms.

Results indicate that clustPR as a non-personalized recommendation increases precision but recall decrease because of network clustering. ClustPRcan increase precision but this amount is less than personalized recommendation methods.

In clustCF method, amounts of precision are larger than traditional CF method but its amount is not big and recall is smaller because of network clustering.

V. CONCLUSION AND FUTURE WORKS

In this paper, we offered a blog recommendation framework that makes use of clustering approach to generate recommendation. A complex clustering network was used on blogs social network to find similar users group. Then we used neighborhood-based CF algorithm to generate recommendation in each cluster. We tried our experiments on the real world data set. We also did it for non-personalized recommendations to demonstrate that our framework with clustering approach increases accuracy for recommendations.

In future works, we intend to recommend a framework that can assign bloggers into several clusters (overlapping cluster). Overlapping clusters can depict real world conditions that bloggers participate in different communities. To do that, we intend to assign bloggers into several clusters with combining blog social network data and content-based recommender systems. In this study, we also used blogs links (blog roll) as an item-user rating matrix that is a binary matrix. In future studies, we

are going to consider the other links such as post-to-post, comment-to-post, and then combining them and obtaining the strength of blog relationship, we will compare the results by using non-binary matrix.

REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6):734–749.

[2] Clauset, A., Newman, M. E. J., and Moore, C. 2004. Finding community structure in very large networks. Phys. Rev. E, 70(6):66111

[3] Garcı́a-Crespo Á., Colomo-Palacios R., Gómez-Berbı́s J.M.l, Garcı́a-Sánchez F.2010.SOLAR: Social Link Advanced Recommendation System , Future Generation Computer Systems 26 (3): 374_380.

[4] Li Y.M., Ching-Wen C. 2009.A synthetical approach for blog recommendation: Combining trust, social relation and semantic analysis, Expert Systems with Applications 36 (3): 6536–6547.

[5] Newman M. E. J., Girvan M.2004. Finding and evaluating community structure in networks, Phys. Rev. E, 69(2): 26113

[6] Su X., Khoshgoftaar T.M. 2009. A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, 2009 (January 2009) Hindawi Publishing Corp. New York, NY, United States.

[7] Han J., KamberM.2001. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, Calif, USA.

[8] Abbasi, Z. and Mirrokni, V.S. 2009. A Recommender System Based on Local Random Walks and Spectral Methods. LNCS 5439, pp. 139–153.

[9] Hayes C., Avesani P., Bojars U.2007. An Analysis of Bloggers, Topics and Tags for a Blog Recommender System, Lecture Notes in Computer Science4737, 1-20.

[10] Jiang, X., Song, W., and Feng, W. 2006. Optimizing collaborative filtering by interpolating the individual and group behaviors. In APWeb. Lecture Notes in Computer Science3841,2006, 568-578

[11] Nepusz T., Csardi G. 2007. igraph Reference Manual, Technical Report, CRAN repository.

[12] Page L., Brin S., Motwani R., Wingord T. 1999. The PageRank Citation Ranking: Bringing Order to the Web, Technical Report. Stanford University.

[13] Fujimura, K., Inoue, T., Sugisaki, M. 2005. The Eigen Rumor Algorithm for Ranking Blogs, In proceeding of WWW 2005, May 10--14, 2005, Chiba, Japan.

[14] Hart M., Johnson R., Stent A.2009., iTag: a personalized blog tagger, In Proceedings of the third ACM conference on Recommender systems.

[15] Peng T .C, T. Chou S.c. 2009. iTrustU: a blog recommender system based on multi-faceted trust and collaborative filtering, In Proceedings of the 2009 ACM symposium on Applied Computing (New York, USA).

[16] Sarwar B.M., Karypis G., Konstan J.A, Riedl J. 2001. Item based collaborative filtering recommendation algorithms, in Proceedings of the 10th International Conference on World Wide Web (WWW ’01), Pp. 285–295.

[17] Xue, G., Lin, C., and Yang, Q. 2005. Scalable collaborative filtering using cluster-based smoothing, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.

Hamid Alinejad-Rokny is a member of *School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW, Australia*. He is the author/co-author of more than 65 publications in technical journals and conferences. He served on the program committees of several national and international conferences. He was *Guest Editor-Chief* for special issue at *IJFIPM*. Also He is *Deputy Editor-Chief* at *International Journal of Software Engineering And Computing* and he is editorial board member at *IJSEI, IJFIPM, JETWI, IJSCIP, IJCSCS, IJCNT and IJEIS.* His research interests are in the areas of Data Mining, Bioinformatics, Artificial Intelligence and Biological Computing.

Behrouz Minaei-Bidgoli obtained his Ph.D. degree from Michigan State University, East Lansing, Michigan, USA, in the field of Data Mining and Web-Based Educational Systems in Computer Science and Engineering Department. He is working as an assistant professor in Computer Engineering Department of Iran University of Science & Technology, Tehran, Iran. He is also leading at a Data and Text Mining research group in Computer Research Center of Islamic Sciences, NOOR co. developing large scale NLP and Text Mining projects for Persian and Arabic languages. He is the author/co-author of more than 60 publications in technical journals and conferences**.**