IJICTR
International Journal of Information & Communication Technology Research

Research Note

Volume 10- Number 1 – Winter 2018 (56-61)

Recommending Friends in Social Networks By Users' Profiles And Using Classification Algorithms

Paria Dashtizadeh
Department of Computer Engineering, Ahvaz Branch, Islamic Azad University
Ahvaz, Iran
P_dz65@yahoo.com

Ali Harounabadi*
Department of Computer Engineering, Central Tehran Branch, Islamic Azad University,
Tehran, Iran
a.harounabadi@gmail.com

Received: 22 November, 2017 - Accepted: 18 March, 2018

Abstract—**Nowadays, social networks are becoming more popular, so the number of their users and their information is growing accordingly. Therefore, we need a recommender system that uses all kinds of available information to create highly accurate recommendations. Regarding the general structure of these recommender systems, one criterion is first chosen to calculate the similarity between users and then people who are assumed to have great similarity are proposed to each other as friend. These similar criteria can calculate users' similarity with regard to topological structure and some properties of graph vertices. In this paper, the properties that are required for clustering are extracted from users' profile. Finally, by combining the similarity criteria of mean measure of divergence (MMD), cosine, and Katz, different aspects of the problem including graph topology, frequency of user interaction with each other, and normalization of the same scoring method are considered.**

Keywords: social network, friend recommendation, graph clustering, users' profiles, link prediction

I. Introduction

Services of online social networks such as Facebook have been becoming popular in recent years. Nowadays, millions of people are active in these services and create and share rich online information not previously available in the past [1]. The main reason of their popularity and the difference between social network websites and other websites lie in this fact that they allow the people to virtually have relationship with other people, to send messages and virtual gifts, to use others' shared data and to comment on them [2].

In spite of finding attractive and relevant information of social network users, they face major challenges in identifying information resources such as like-minded users, trusted social friends, and interest groups [3]. Creating friendship takes place through establishing social relationships with others in online social networks via which people can contact their friends in the real world and have access to their favorite information [4]. Nowadays, of the number of people in the social networks is massive, so finding similar people is considered a complex and expensive process. Therefore, a recommender system is required whic can use all available information to create highly precise recommendations and to successfully relate people in social networks [5]. The hypothesis of this recommender system is that people may be very close to desirable social friends, but do not know about them [6].

There are many studies conducted in online social networks in field of friend recommendation systems. The general structure of this recommendation system is such that one criterion is first chosen to calculate the similarity between users and then people who are assumed to have high similarity are proposed to each other as friend. These similar criteria can calculate users' similarity with regard to the topological structure and some properties of graph vertices. In many real applications, both the topological structure of graph and

* Corresponding Author

properties of vertices are important. For example, the properties of vertices in social networks can be values of users' profile (e.g. age, sex, education, employer, place, etc.) and indicate the topological structure of relationship, and interaction between groups of people; both similarity measures between users are important [7].

In this paper, the properties required for clustering are extracted from users' profiles and then sets of data that are available in the social networks are clustered using hierarchal clustering algorithm [8]. Then, the clustered system is trained using a decision tree algorithm. Therefore, when a new member enters a social network, the system puts the specifications of new member in the decision tree and receives appropriate output from the decision tree. Hence, the new member will be placed in the most appropriate cluster. Finally, while combining the similarity criteria of MMD, cosine, and Katz, different aspects of the problem including graph topology, frequency of user interaction with each other, and normalization of the same scoring method are considered. We can use the results of this paper in the electronic commerce. Presenting a suitable configuration for a classifier in order to characterize the users in social networks by the extraction of the effective characteristics from their profiles and combination of similarity measures can be regarded as the study innovations.

This paper is organized as follows: Section 2 presents the works performed in the friend recommender systems in social networks. In Section 3, the suggested system will be explained to predict and recommend friends that have many similarities in common. In Section 4, the suggested system is evaluated. Finally, the conclusion of the paper and future work are suggested in the Section 5.

II. Related Works

Since social networks are becoming increasingly popular, the number of their users and their shared information are progressively growing. Since, there are many varieties of users in social networks, a recommender system is required that uses all kinds of available information to create high-precision recommendation and to successfully relate people in social networks. Many studies have been conducted in the field of friend recommender systems in online social networks. Many friend-finding systems recommend friends to users based on the similarity in users' profiles, the number of shared neighbors, geographical position of users, and prediction of edge by means of available nodes [9]. This section briefly introduces some studies conducted in recent years by different techniques. Symeonidis et al. [10] presented a multi-way spectral clustering to predict the communication in social networks. They used little information obtained from few vectors, eigenvalues, and normal Laplace matrix and calculated multi-stage partitions of data. First, kth first eigenvector and the corresponding eigenvalues (k is the number of clusters) were calculated; then the nodes were clustered while using the k-means clustering algorithm and the eigenvectors obtained. The center of each cluster was calculated and then the distance between each node from the center of cluster was calculated. Then, the similarity of each node in each specified cluster was separately calculated in relation to the nodes in the same cluster and nodes in other clusters and was stored in one vector. Finally, n- users with a high score in similarity were proposed to the intended user.

Hamid et al [11] suggested a friend recommendation system based on cohesion. Mainly, cohesion is defined as all factors that attract all people into one special group. In this method, first, the sub-network of social networks with random number of people supposed to be introduced to each other as the friend was extracted. Then, the number of properties common in the intended users was chosen and the strength of communication between two users was measured with regard to the chosen properties. In the next stage, the network could be completed by adding communications between the users which can be created in future but are not available in the network now. Then, Louvain method was used to identify the potential communities in a social network graph. This method can analyze large networks within the shortest time and do this task in hierarchical clustering. Finally, people who are in the society and who are not friend with each other are proposed to each other.

In [6], Papadimitriou et al. presented a friend recommendation system called FriendLink to obviate the problems identified in the prediction of communication. It operates by traversing all paths of limited length based on the algorithmic small-world hypothesis, where it does not use all paths with different lengths in the network and considers the maximum length l€ [2 and 6] between the user and nominated friends. To do this, they defined a new similarity criterion between nodes. It uses a combination of local and global properties. Using a new criterion, it calculates the similarity score of each user with all users with whom they have no relationship. Finally, the users who have high score similarities are proposed to the intended users. Shalforoushan et al [12] introduced a new method for link prediction in social networks using Bayesian networks. Bayesian network is a reliable model to understanding the relationship between variables. Their solution has two phases. The first phase is related to new users who register in the social networks and do not find their favorite friends. In this condition, the properties of the network and properties of common friends cannot be used; the only information that can be used in this phase includes primary and personal properties registered by users in the registration time. The second phase is related to the time during which the user is a member in social networks and finds some friends. Therefore, we can use some properties such as common friends in this phase. Modeling at this stage is done based on adding the characteristics of common friends. Bayesian inference in the first stage suggests that the property of an area as the most personal effective factor is required for creation of the friendship between users. This means that the most probable friendship in the social network

belongs to users that have a common living region. On the other hand, having the same gender is not a suitable factor for friendships. For the recommendation, a list with k friends is proposed to each user.

Tian et al [13] presented a friend- recommendation system with the shortest path approach in social network, suggesting all potentially common friendships between two users with indirect connection. In their approach, first graph adjacency matrix is created. Then, k shortest path between two vertices is calculated by Floyd- Warshall algorithm and graph adjacency matrix. In the next stage, the search results are optimized using pruning method. In order to recommend friend to users, the largest common subcategory between the two vertices is calculated from the paths obtained from the previous stage and is stored in an array. The elements of this array represent all potentially common friends between two specific users who are connected indirectly and they can be proposed to the users.

III. The Proposed Classification Method

This section presents the main strategy of the suggested method. In order to simplify the expression of the material, first the most important symbols and definitions corresponding to them used later in this article are stated.

A. Primary Definitions in the Graph

A graph G= (V, E) include a set of nodes V and set of edges E; with each edge connecting two nodes to each other. In this paper, the graph G is always a graph with no direction and weight. Therefore, (v_i, v_j) and (v_j, v_i) show an equal edge in the graph. Furthermore, we assume that graph G has no multiple edge. Therefore, the two nodes of v_i and v_j are connected by an edge in E and no other edge connects them in E. Finally, we assume that graph G contains no loop (that is, a graph cannot attach to itself). This graph represents the friendship between users in online social networks. The total number of edges connected to the vertex is called the degree of that vertex and is shown by deg(v_i). Adjacency matrix A related to graph G is a square matrix with labeled rows and columns along with graph vertices; whether two users vi and vj are friends or not, values 0 and 1 might be put in their relevant cells (vi, vj). Adjacency matrix is symmetric for graphs with no direction.

A path p (v0, vx) from source vertex to destination vertex is the sequence of the edges in the form of (v0, v1), (v1, v2), …, (v(x-1), vx), where ei= (vi, v(i+1) Є E (0<i<x). For two vertices of vi and vj, the shortest path between them is the path with the minimum number of edges.

B. Architecture of the Suggested System

In this paper, a new architecture is presented for recommending friends to users of social networks. Figure 1 demonstrates all stages of this architecture. As observed in the figure, the suggested architecture consists of a number of stages. In the first section, the properties required for clustering based on users' profiles are extracted. The datasets that are available in the social network are clustered using hierarchal clustering algorithm. In the second section, the clustered system is taught using decision tree algorithm. Therefore, when a new member enters the social network, the system gives the characteristics of the new member to the decision tree and receives the most appropriate cluster as the output from the decision tree. Therefore, the new member will be placed in the most appropriate cluster. Finally, while combining the similarity criteria of mean measure of divergence (MMD), cosine, and Katz, different aspects of the problem including graph topology, frequency of user interaction with each other, and normalization of the same scoring method are considered. Then, each stage will be thoroughly explained.

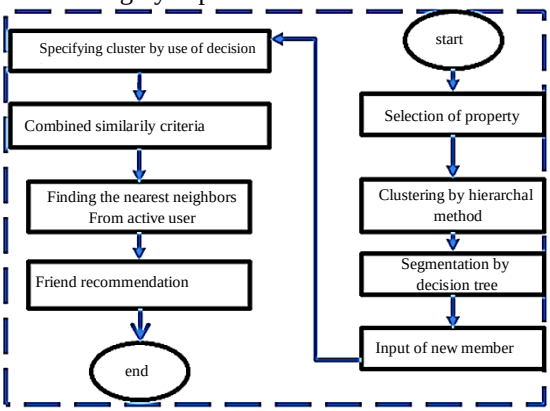

Fig. 1. Flowchart of the proposed method

Data Collection:

In this stage, information of the profile as well as the information of users' communication is collected. Since, the users in different social networks have profiles with different properties, this stage is dependent on the properties of the utilized social networks and the required application.

Clustering by Complete Linkage Hierarchical Algorithm:

This method that is like single linkage method is considered as exclusive and a hierarchical clustering method. This clustering method is called the farthest neighbor. In this method, in order to calculate the similarity between two clusters of A and B, Relation (1) is used as follows: (1)

$$d_{AB} = \max_{i \in A, j \in B} d_{ij}$$

Where, i is the data sample belonging to cluster A and j is data sample belonging to cluster B. Indeed, in this method, the similarity between two clusters is the greatest distance between a member of a cluster and the other member of another cluster.

Combined Similarity Criterion:

The similarity in the suggested method is captured by sum of values obtained from mean measure of divergence (MMD), cosine and Katz methods.

Similarity Criterion of Mean Measure of Divergence (MDD):

This criterion is the most commonly used criterion in recommendation systems and calculates the biological

distance based on non-dimensional traits. Traditional similarity criterion does not calculate the personal habits of people to state people's preference. Every person has personal habits which state their preferences. Some users want to have higher or lower reliability as compared to other people. This prejudice of ranking influences the relationship between users. As mentioned earlier, the traditional similarity of this criterion is ignored. In the similarity criterion of MMD, the personal habit of people is calculated in their preferences.

The similarity between two users u and u' [sim (u, u')] is represented by Relation (2) [14]:

$$sim(u,u')^{MMD}=\frac{1}{1+(\frac{1}{r}\sum_{i=1}^{r}\{(\theta_u-\theta_{u'})^2-\frac{1}{|I_u|}-\frac{1}{|I_{u'}|}\})} \quad (2)$$

Where, Θu is the vector of user u ranking;

|Iu| represents the number of total ranking made by user u; and

r indicates the number of co-rated items between two users u and u'. Here, the value of r is based on the vector of property. The cold-start problem is solved by this approach as it is taken from users' profile and their similarities.

Similarity Criterion of Cosine:

In order to find common data items, the similarity should be measured. In the perspective of data item based recommendation, the similarity of cosine is defined as the standard criterion. This similarity criterion measures the angle between two n-dimensional vectors. This method is usually used in the field of information retrieval and text mining in order to compare textual document presented as vectors of terms.

According to the similarity criterion of cosine, the similarity between two data items a and b (as ranking vectors corresponding to $\vec{a}$ and $\vec{b}$) is calculated by means of Relation (3):

$$sim(\vec{a},\vec{b})=\frac{\vec{a}.\vec{b}}{|\vec{a}|*|\vec{b}|} \quad (3)$$

Sign . is the internal multiplication of vector and $|\vec{a}|$ is Euclidean length of vector defined as the root square of the internal multiplication of the vector by itself. The similarity values are between 0 and 1, where a value that is close to 1 shows high similarity. In the field of data item recommendation, these criteria can be used for calculating user similarities while the user u is considered as vector u which is $x_u \in R^{|I|}$. If the user u ranks the data item i, $x_{ui}=r_{ui}$, else 0. Therefore, the similarity between users u and v (cv) is calculated by Relation (4) [14].

$$CV(u,v)=\cos(x_u,x_v)=\frac{\sum_{i\in I_{uv}} r_{ui}r_{vi}}{\sqrt{\sum_{i\in I_u} r_{ui}^2 \sum_{j\in I_v} r_{vj}^2}} \quad (4)$$

Where, I_{uv} is the data item ranked by the users u and v.

Katz's Similarity Criterion:

First, graph adjacency matrix is developed. Graph adjacency matrix represents the user with A. The global index of Katz investigates all paths available between both of them in order to the capture the similarity between two nodes and has different lengths L, and is based on the shortest path. This means that it will damped with the length of path incrementally and finally the shortest path gains the highest weight. The formulation of Katz is in the form of Relation (5):

$$katz(x,y)=\sum_{l=1}^{\infty}\beta^l|paths^l_{v_x,v_y}| \quad (5)$$

Where, $|paths^l_{vx.vy}|$ indicates the number of paths with length L between two nodes x and y; and β is the damping coefficient which is a free parameter. The controller of paths' weight is obtained by Relation 6:

$$\beta=\frac{1}{1+K} \quad (6)$$

K in the above relation is equal to the maximum sum of rows or columns of adjacency matrix.

The similarity matrix between the users is obtained by Relation (7) [15]:

$$katz(A;\beta)=\beta A+\beta^2A^2+\beta^3A^3+\ldots=(I-\beta A)^{-1}-I \quad (7)$$

In the above relation, I is the unit matrix.

Combination of Three Criteria for Friend Recommendation:

After calculating the similarity between every pair of users by means of the three mentioned methods, the similarity in the suggested method is obtained by sum of their values and Relation (8).

$$sim(u,u')^{PM}=sim(u,u')^{cos}+sim(u,u')^{Katz}+sim(u,u')^{MMD} \quad (8)$$

After calculating similarities between the users, the next step is to find users with high K value and to obviously discover the highest similarity to the active user. These users are neighbors of active users. Relation (9) is used in order to predict the ranking from the active user for without-ranking properties, [14]:

$$pred_{u,i}=\bar{r}_u+\frac{\sum_{u'\in N(u)} sim(u,u')\times(r_{u',i}-\bar{r}_{u'})}{\sum_{u'\in N(u)} sim(u,u')} \quad (9)$$

Sim (u, u'): combined similarity calculated by Relation (8).

$\bar{r}_u$: the average rating of user u;

N(u): the number of neighbors of user u;

$r_{u,i}$: rating given by user u to the property i;

IV. The Results of Experiments

A. Datasets

In this section, the method that is presented in this paper is empirically compared with two available friend recommendation systems [12]. The suggested method which is called PM will be described in detail. The database used in this study is the popular dataset of

Slovakian social network. This dataset known as POKEC is composed of two general files. The first file shows the profile of 1632803 users. On the other hand, the second file shows the friends and people with whom each user has relationship [16].

B. Evaluated Metrics

In order to evaluate the suggested method, two different standard criteria called recall, precision, and F1 are used. The precision is equal to division of real positive cases into the total sum of real positive cases and false positive. It involves the ratio of the correct recommendation chosen by the test to the total recommendation that shows the predicted friend.

$$precision = \frac{TP}{TP + FP} \quad (10)$$

Recall is equal to division of real positive cases into the sum of real positive cases and false negative cases. More specifically, recall shows the ratio of the correct recommendation chosen by the test to the total recommendation that is real friends.

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

The criterion F1 is the harmonic mean of precision and recall.

$$F=\frac{1}{\alpha\frac{1}{per}+(1-\alpha)\frac{1}{Rec}}=\frac{(\beta^2+1)Per.Rec}{\beta^2 Per+Rec} \quad (12)$$

Where β=1 and α= 0.5, consequently

$$F1=\frac{2Per.Rec}{Per+Rec} \quad (13)$$

C. Results and Discussion

First, the database under consideration was clustered via hierarchical clustering. In the clustering, six properties were used. These six properties included geographical region, age, music interest, gender, film interest, and relationship with children. Since the number of cluster was not specified, hence the trial-and-error method was used to determine the number of clusters. To do this, the accuracy of the classifier was used. Clustering was done through hierarchal clustering for clusters whose numbers were 2, 4, 6, 8, and 10, for which the classifier was then used. For each of them, classification was separately used via the decision tree. As observed in Table 1, in the state in which six classifiers were used, the accuracy of classification has had the highest value equal to 0.98. The following results were obtained based on the six clusters.

TABLE I. the effect of the number of clusters on the accuracy of classification

The number of clusters	The accuracy of classification (percent)
2	57
4	83
6	98
8	72
10	64

The dendrogram chart displays that kind of hierarchal clustering in which six clusters have been used. Then, the decision tree-based model was achieved based on the performed clustering, and new users will be placed in the cluster with which they have the greatest matching. The third property had the highest entropy and was selected as the root. In the pruning process, only the third property remains in the tree. This property shows the age. The decision tree was used only for new users. When the new users entered the network, first their clusters were determined via the decision tree. Then according to the above mentioned process, the user received friendship recommendation in the same cluster.

D. Comparing the Suggested Method with Other Methods

Here, the suggested method abbreviated as PM is compared with other methods [12]. After clustering and using the decision tree to offer the recommendation, the suggested combination method is applied. Since the database under consideration is social network, so it does not suffice to only pay attention to such parameters as precision. Thus, values such as precision and recall are used. The precision shows what ratio of the positives is really positive. Here, the positive is considered as friend recommendation while the negative is considered as wrong recommendation. The values of precision in the suggested method decrease with increase in the number of the recommendation.

On the other hand, the feedback shows what ratio of the real positives is recognized as appropriate positive. Here, the suggested method acts more weakly with increase in the number of the recommendation.

Criterion F1 is the harmonic mean of precision and recall parameters. The harmonic mean of the suggested method declines with increase in the number of friend recommendation.

TABLE II. comparing the suggested method with other methods

Method	Precision	Recall	F_1
PM	0.51	0.38	0.46
Bayesian	0.41	0.25	0.31
FOF	0.33	0.195	0.25

V. Conclusions and Future Works

Nowadays, online social networks are becoming increasingly popular, since social media platforms allow users to form ties or connections among themselves in the process of sharing images, texts, videos, and other digital artifacts [17]. In this paper, a framework was presented to recommend a friend in social networks. In this method, the properties required for clustering from users' profile were extracted. Then, while using hierarchal clustering algorithm, the dataset available in the social network was clustered. After

that, the clustered system was trained using the decision tree algorithm. Finally, when a new member was registered in the social networks, the system gave the characteristics of the new member to the decision tree and received the most appropriate cluster as the output from decision tree. Therefore, the new member would be placed in the most appropriate place. While combining the similarity criteria of MMD, cosine, and Katz, different aspects of the problem including graph topology, frequency of user interaction with each other, and normalization of the same scoring method were considered. The results of this paper can be used in the electronic commerce. The results obtained from the evaluation based on the precision and recall suggest that the method proposed in the current paper, used for recommending appropriate friends to a user, is more successful than other similar methods and can offer appropriate recommendations for friendship with regard to similarities in characteristics, interest, and their interaction with each other. In future studies, daily interaction of users such as comment about people's post, common image labeling, similar product ratings, and so on can be used to define similarity criteria in order to create a precise recommendation system. The precision and recall quantities in the suggested method will operate more weakly by increasing the number of suggestions. Making use of other similarity measures can be studied in order to examine the problem improvement in system results.

References

[1] Boyd, D., Ellison, N. B, "Social network sites: Definition, history, and scholarship", Journal of Computer-Mediated Communication, 13(1), pp. 210-230, 2007.

[2] Kazemi, A., Nemath, M, "Finding compatible people on social networking sites, a semantic technology approach", Second International Conference on, Intelligent System, Modelling and Simulation (ISMS), pp. 306-309, 2011.

[3] Moricz, M., Dosbayev, Y., Berlyant, M, "PYMK: friend recommendation at myspace", In Proceedings of the ACM SIGMOD International Conference on Management of data, pp. 999-1002, 2010.

[4] D'cunha, A., Patil, V, "Friend recommendation techniques in social network", International Conference on, Communication, Information & Computing Technologhy (ICCICT), pp. 1-4, 2015.

[5] Rai, P., Singh, S, "A Survey of Clustering Techniques", International Journal of Computer Applications, VOL. 7, No. 12, pp. 1-5, 2010.

[6] Papadimitriou, A., Symeonidis, P., Manolopoulos,Y, "Fast link prediction in social networking systems", ELSEVIER, The Journal of Systems and Software,VOL. 85, No. 9, pp. 2119-2132, 2012.

[7] Zhou, Y., Cheng, H., Yu, J. X, "Graph clustering based on structural/attribute similarities", Proceedings of the VLDB Endowment, 2(1), pp. 718-729, 2009.

[8] Alsaleh, S., Nayak, R., Xu, Y., Chen, L, "Improving matching process in social network using implicit and explicit user information", the Asia-Pacific Web Conference Lecture Notes in Computer Science, VOL. 6612, pp. 313-320, 2011.

[9] Yigit, M., Bilgin, B., Karahoca, A, " Extended Topology Based Recommendation System For Unidirectional Social Networks", Expert Systems with Applications, Vol. 42, Issue. 7, PP. 3653-3661, 2015.

[10] Symeonidis, P., Lakovidou, N., Mantas, N., Manolopoulos, Y, "From biological to social networks: Link prediction based on multi-way spectral clustering", The Journal of China Universities of Posts and Telecommunications, VOL. 7, No. 4, pp. 226-242, 2013.

[11] Hamid, N., Naser, A., Hasan, k., Mahmoud, H, "A Cohesion-Based Friend Recommendation System", Journal of Social Network Analysis and Mining, Vol. 4, Issue. 1, pp. 1-11, 2014.

[12] SHalforoushan, H., Jalali, M, " Link Prediction in Social Networks Using Bayesian Networks", Iternational Symposium on Artificial Intelligence and Signal Processing, DOI: 10.1109/AISP.2015.7123483, pp. 246-250, 2015.

[13] Tian, X., Song, Y., Wang, X., Gong, X, "shortest path based potential common friend recommendation in social networks", Second International Conference on, cloud and Green Computing (CGC), pp. 541-548, 2012.

[14] Suryakant., Mahara, T, " A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment", ELSEVIER, Twelfth International Multi-Conference on Information Processing (IMCIP), DOI:10.1016/j.procs.2016.06.099, pp. 450-456, 2016.

[15] Symeonidis, p., Perentis, c, "Link Prediction in Multi-model Social Networks", In European Conference, ECML PKDD 2014, Nancy, France, pp. 147-162, 2014.

[16] "POKEC", Available at: http://snap.stanford.edu/data/soc-pokec.html. Access Time 06, Dec, 2016.

[17] Espina, C., Himelboim, I., Rainie, L., Shneiderman, B, "Classifying Twitter Topic-Networks Using Social Network Analysis", Journal of Social Media+ Society, pp. 1-13, 2017.

AUTHOR BIOGRAPHIES

Paria Dashtizadeh is graduated in computer engineering(software) with M.S. degree from Ahvaz Azad university. Her field of interest is mainly focused on classification algorithms.
E-mail: p_dz65@yahoo.com

Ali Harounabadi is an assistant professor of computer engineering at central Tehran branch, Azad University. His research is focused on Recommender Systems, Web mining and methodologies in software Engineering.
E-mail: a.harounabadi@gmail.com