A Graph-Based Friend Recommendation System Using Genetic Algorithm

Nitai B. Silva, Ing-Ren Tsang, George D.C. Cavalcanti, and Ing-Jyh Tsang

***Abstract*—A social network is composed by communities of individuals or organizations that are connected by a common interest. Online social networking sites like Twitter, Facebook and Orkut are among the most visited sites in the Internet. Presently, there is a great interest in trying to understand the complexities of this type of network from both theoretical and applied point of view. The understanding of these social network graphs is important to improve the current social network systems, and also to develop new applications. Here, we propose a friend recommendation system for social network based on the topology of the network graphs. The topology of network that connects a user to his friends is examined and a local social network called Oro-Aro is used in the experiments. We developed an algorithm that analyses the sub-graph composed by a user and all the others connected people separately by three degree of separation. However, only users separated by two degree of separation are candidates to be suggested as a friend. The algorithm uses the patterns defined by their connections to find those users who have similar behavior as the root user. The recommendation mechanism was developed based on the characterization and analyses of the network formed by the user's friends and friends-of-friends (FOF).**

I. Introduction

In the last few years, social networks have been increasing in both size and services. Social networking services (SNSs) such as Facebook, MySpace, Twitter, Flickr, YouTube and Orkut are growing in popularity and importance and to some extent they are also contributing to a change in human social behavior. Some of these SNSs already provide a service to recommend friends, even though the method used is not disclosed, we believe that an FOF approach is mostly used. The topology of this type of network has been measured and analyzed by different researches [1], [2], [3]. Some interesting structural properties such as power-law, small-world and scale-free network characteristics have been reported [1]. Also the topological patterns of activities and the structure and evolution of online social networks have been studied [3], [7]. Knowledge of the structure and topology of these complex networks combined with quantitative properties such as size, density, average path length or cluster coefficient can be used to develop novel applications such as a recommendation system.

With the increase of the e-commerce, recommendations systems have been of great interest. This is due to the possibility of increase sell obtained from success recommendation. Sites that offer different products such as books, clothes and movies, most often also provides recommendations based on previous brought products. The Netflix prize (http://www.netflixprize.com) is an example of continuous interesting in this field [14]. The problem of product, service, and friend recommendation, or in more global context information, is growing in both commercial and academic research interest.

Here, we proposed a friend recommendation system that suggests new links between user nodes within the network. The central problem can be viewed as a procedure to propose relevant parameters for nodes relationship using the information from the social network topology and statistical properties obtained by using classical metrics of complex networks. Even though, topology based approaches for recommendation systems have already been suggested by other researchers [11], [12], [13], we proposed a different clustering indexes and a novel user calibration procedure using Genetic Algorithm (GA).

The rest of the paper is organized as follows: In Section 2, we briefly describe the Oro-Aro social network used to analyze the proposed recommendation system. In section 3, the recommendation mechanism is explained in details. The process is divided in two phase filtering and ordering, some network measurements are also defined and three important indexes are introduced. In Section 4, we describe the experiments using the proposed system in the Oro-Aro social network. Also, we describe some estimates and comparisons of the obtained samples. Finally, the concluding remarks are presented in Section 6.

II. The Oro-Aro Social Network

A social network is a structured community of individuals or organization composed of nodes that are connected through one or more particular kind of interdependence, like values, ideas, interests, business, friendships, kinship, conflict, and trading [4], [5]. Analysis and measurements of social networks examines the social relations in terms of nodes and connections. Nodes in such network represent individual users of the system, and connections correspond to the relations between the users of the SNSs.

In our experiments, we used the data obtained from the Oro-Aro social network (http://www.oro-aro.com) that was

Manuscript received February 8, 2010. This work was supported in part by FACEPE – Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco.

Nitai B. Silva, Ing-Ren Tsang and George C. D. Cavalcanti are with the Federal University of Pernambuco (UFPE), Center of Informatics (CIn), Av. Prof. Luis Freire, s/n, Cidade Universitária, CEP 50740-540 (phone: +55 81 2126 8430; e-mail: {nbs,tir,gdcc}@ cin.ufpe.br).

Ing-Jyh Tsang is with Alcatel-Lucent, Bell-Labs, Copernicuslaan 50, 2018 Antwerp, Belgium (e-mail: ing-jyh.tsang@alcatel-lucent.com).

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

developed by the Recife Center for Advanced Studies and Systems (C.E.S.A.R) (http://www.cesar.org.br/). This social network is located in Brazil and it was developed with the intention to facilitate the exchange of experiences of the student of the Center of Informatics and the software engineers at CESAR. The network is composed of a total of 634 nodes and 5076 edges.

Some preprocessing was applied on the data, so that the proposed algorithm could be implemented. The Oro-Aro allows the creation of a one-way relationship, i.e. a user can add another user to his list of contact without the need for the other user to approve the link. This kind of network is similar to twitter, a microblog network. Therefore a filter was used to remove all one-way relationship, i.e. who did not have an inverse relationship. The procedure reduced by 29% the number of edges and 8% in the number of nodes (isolated sub-networks were removed). This procedure was necessary to obtain a network with symmetric connections; also most of the social networks just allow two-ways relationship.

Figure 1 shows a graphical representation of the Oro-Aro social network. Each node represents a user and the edges a two-way relationship between users.

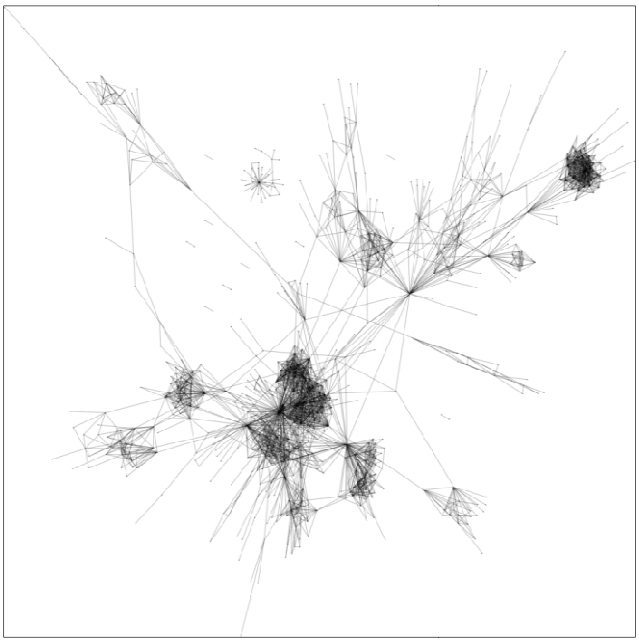

Fig. 1. Visual representation of the Oro-Aro social networks, each node represents a user having the relationships between users show a link.

III. RECOMMENDATION MECHANISM

The proposed friend recommendation system is based on the structural properties of social networks. The topological characteristics, the information and the metrics are derived from the complex network theory. It is observed that these types of networks are defined as being either small-world or scale free [1], [2], [3]. This characteristic can be used to the development of a reliable recommendation mechanism. Figure 2 shows a plot of the degree of the node versus the frequency of occurrence. This plot provides some information on the topological characterization of the Oro-Aro SNS.

The recommendation mechanism procedure utilizes the graph topology of the SNS to filter and order a set of node that have some important properties in relation to a given node v_i. The nodes of the resulting set are recommendations of new edges that should be connected to the node v_i. The creation of these new edges is used to improve the properties of the node v_i, besides providing benefits to the entire network in terms of friendship connection.

The process of recommendation is divided in two steps: a filtering procedure followed by an ordering. Filtering is an important step, because it separates the nodes with higher possibilities to be a recommendation, consequently reducing the total number of nodes to be processed in the network. The ordering step considers some properties to put the most relevant nodes in the top of the resulting list. As expected, the result depends on the user the recommendation is generated for. Therefore, we applied an adaptive solution using Genetic Algorithm.

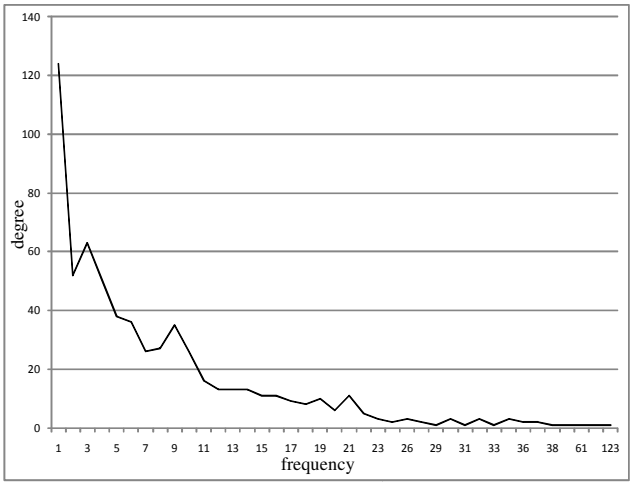

Fig. 2. The graph shows the degree of the node versus the frequency of occurrence. This figure demonstrates the behavior of connected users in the Oro-Aro social network.

A. *Filtering*

The algorithm used to perform the filtering procedure employs the concept of the clustering coefficient, which is characteristic in small world networks [1]. Using the natural idea that "It's more probable that you know a friend of your friend than any other random person" as stated by Mitchell in [4], the filtering step is restricted to select nodes adjacent to each node that is adjacent to the central node of the recommendation process. All nodes that can be reached with two hops are considered.

Figure 3 shows an example of this network for a single user. The node n_i is the central element of the global recommendation process. The nodes within circle "A" are directly connected to the node n_i. The nodes within circle "B" and outside circle "A" are selected by the filtering procedure.

B. Ordering

The ordering procedure consists on the measurement of the coefficients and their normalization by using an adjustment mechanism. It is not different from a regular ordering mechanism, since it also use only one numeric value related to each node to be ordered. The indexing value belonging to each node is a result of a process that measures the interaction strength between that node and the central node of the entire recommendation process.

The measurement of this interaction is chosen as the result of a weighted average among three independent indexes. These indexes measure specific properties of a sub-graph composed by the nodes that are analyzed. Three indicators are analyzed in this present article. However other index can still be used to improve the precision of the weighting values. These metrics were used because of its simplicity and the intuitive ideas of how friend should be connected. In the literature of complex networks several metrics to evaluate the strength of interaction between two nodes were presented [4], [8], [9]. We choose the friend-of-friends (FOF) concept, which is a simple and widely used idea as one of the three measurements. Inspired by the concept of clustering coefficient, we also applied two other metrics. Both of them measure the clustering of a set of nodes that links the two nodes in question. The referred nodes are the node that will receive the recommendation and the possible recommended node. Those measures are chosen based on the simplicity and linkage connection between the user and its possible recommended friend.

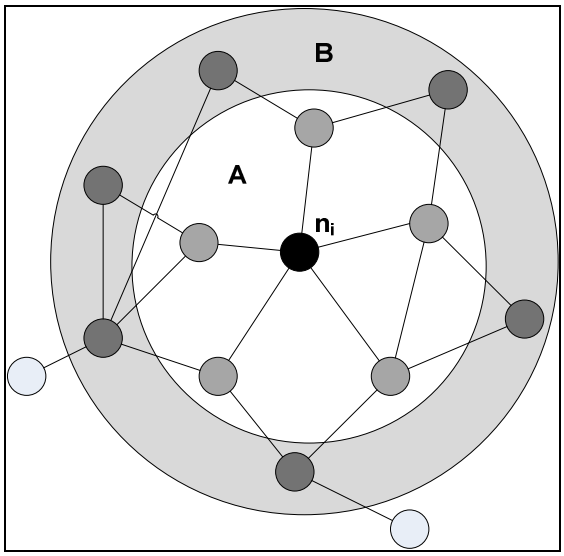

Fig. 3. A visual example of a sub-network showing the links between single users in relation to his connected friends and friend-of-friends in a social network. The network can be spit in two different regions. The region A represents the central user and his friends. Region B represents friends of friends of the main user.

Some variable and concept are defined to derive the indexes used for the friend recommendation system. First we define C_i as the set of the nodes adjacent to node v_i and D_C are defined as the adjacency density among the node contained in the C set. This measure is inspired in the clustering coefficient definition, whereas D_{C_i} is the clustering coefficient of node v_i. The density is calculated as the real size of edges where the origin and destiny nodes are in C divided by the quantity of edges of the clique (complete graph) formed by all the nodes contained in C.

$$D_C = \frac{\sum_{i \in C}(\sum_{j \in C}(M_{ij}))}{(|C| * (|C| - 1))/2}$$

where M_{ij} is the element *ij* of the adjacency matrix.

1) First Index

The first index is defined as the number of adjacent nodes, that are linked at the same time to node i and node j, where i is the center node of the analysis and j is the node that is being ordered for the recommendation system. So, we have:

$$I_{1_{ij}} = |C_i \cap C_j|$$

In a social networks this index measures the quantity of common friends between person i and person j.

2) Second Index

The second index refers to the density of the result measured by the first index:

$$I_{2_{ij}} = D_{C_i \cap C_j}$$

This index measures the cohesion level inside the "small" group formed by the common friends of person i and person j. If this index has a small value then the people inside this group are not well-related. Figure 4 shows a sub-graph representing common friends between two users, n_i and n_j.

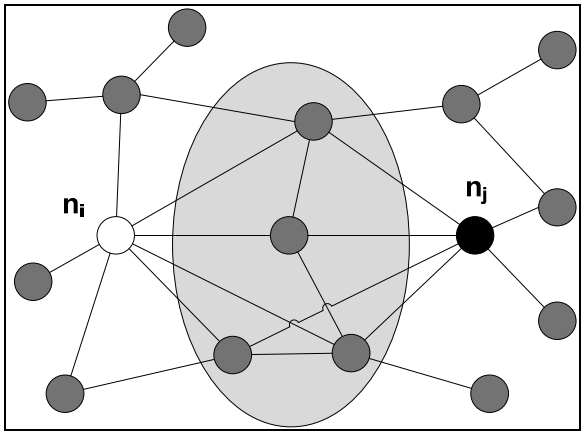

Fig. 4. Sub-graph of the relationship from white and black vertex. The common adjacent vertexes important for both users are shown in the circled region.

3) Third Index

The third index is a variation of the second. This index measures the density of the group formed by the adjacent vertices of node n_i and node n_j. Instead of an intersection,

there is a union of the two adjacencies set as shown in Figure 5 and represented by:

$$I_{3_{ij}} = D_{C_i \cup C_j}$$

This index measures the cohesion between the "big" set formed by the friends of i and friends of j. A good index 2 does not necessarily means a good index 3. In other words, a person that belongs to my work environment, for example, does not necessarily have friends that is relate in any way with my friends in general. So the second index between me and a work friend can appear strong, but as their friends have no relationship with my friends, the third index will appear weak.

C. Index Weight Calibration Using Genetic Algorithm

The calibration step is the procedure that combines the multiple indexes, three in this case, into a single value. This value is used to obtain the final set of ordered results for the recommendation mechanism. A simple procedure to obtain this value is to use a weighted average among the indexes. However, an Genetic Algorithm is used to optimize the indexes value so to obtain the best ordered results.

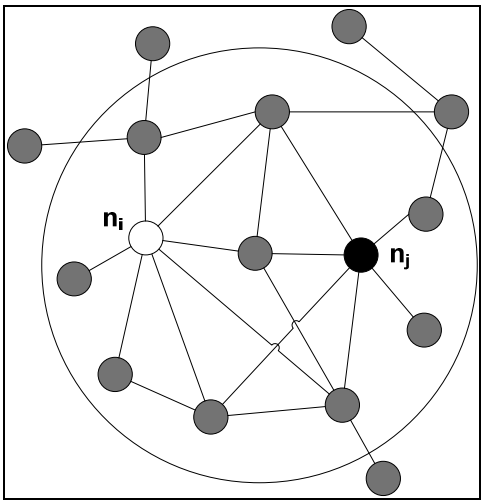

Fig. 5. The rounded vertex, minus the black and white vertex, composes the analyzed set on third index.

The weight calibration of each single index must be adjusted to obtain an optimized result. In this case, optimization means classifying the most important users in the beginning of the list. The importance of a user on the social network depends in the context. This context can change depending on the user who is searching for the recommendation. Therefore, the optimization function must consider the existing relationship structure of the user.

To create the optimization function for the weights, a modification has been proposed in the filtering step. The filtering step needs also to include the nodes directly related to the central node of the global recommendation process. The fitness function used the classification of these nodes as a measure of rightness. Since the ordering procedure defines positions to each node, the mean positions of these nodes that are already related to the central node, is the fitness function value. The smaller this value is the better is the weighting set being considered.

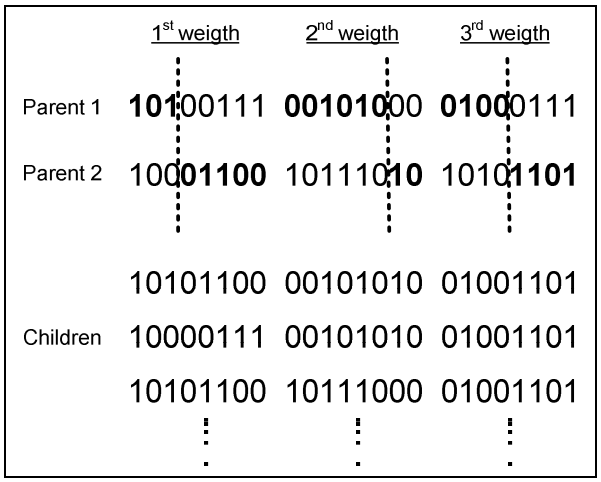

Fig. **6**. Generation of several children strings obtained from two parents. The crossover of the gene from the parents is randomly obtained.

A self-adjustment mechanism should be more practical for the purpose of the recommendation system. This mechanism should choose different indexes for different users. Since each user has his particular relationships pattern, preserving these patterns is the best choice for a recommendation mechanism for it to be more effective. A self-adjustment mechanism needs to be able to calibrate the weights in relation to the optimization function, which is given by:

$$M(n, w) = I_1(n_c, n).w_1 + I_2(n_c, n).w_2 + I_3(n_c, n).w_3$$

The calibration step can be defined as an optimization problem. In this case, we use a genetic algorithm to solve this optimization. The fitness function is defined in the above equation, where I_i, represents the index and w_i the weights that we wish to optimize in order to find the best solution. Each individual of the population is composed of 3 bytes. Each byte, ranging from 0 to 255 represents the weight of each index. The evolutionary process is done with a binary genetic algorithm with uniform state replacement [10]. Each new generation replaces the worst individuals by children of the best individuals. The crossover is done in a single random cutoff point as shown in Figure 6. Each pair of parents can produce many children as long as there are no duplicates in generation. After a generation of children every bit of the new string is passed through the process of mutation. A high mutation rate was applied, 4% for each bit, enough to prevent the loss of diversity in the population, which could lead to premature convergence. The initial population contained 200 individuals randomly generated. The algorithm runs until no improvement occurs in the fitness of the best individual for 4 consecutive generations.

TABLE I
EVOLUTION OF 3 GENERATIONS IN GENETIC ALGORITHM

Iteration	1st Weight	2nd Weight	3rd Weight	Optimization function
	71	69	134	27,341
1st	247	58	104	35,089
	104	7	32	43,764
	75	12	147	18,706
2nd	87	10	104	21,142
	104	7	200	25,997
	70	10	155	10,975
3rd	70	11	142	11,129
	71	11	141	11,439
nth	…	…	…	…

Example of the weights obtained with three iterations using genetic algorithm. The goal is to minimize the value of the optimization function.

In Table 1 three iterations of the evolutionary process is shown. The interaction occurs until the convergence criteria are satisfied. Only the top 3 combinations of weights in each generation are shown. In the last column the numbers represent the value of the function optimization for each weight combination. Figure 6 shows a graphical representation of the optimization function of the best weights combination in each generation. The gray color represents the classification of the nodes already connected to the central node. The black color represents the average classification of all nodes in gray, in other words the fitness function value. And the white color represents the nodes that are not connected to the central node, i.e. nodes that will be recommended.

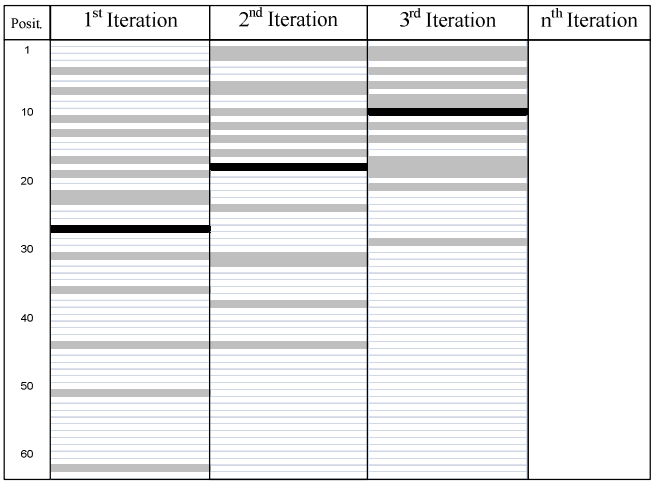

Fig. 7. Graphical representation of the values of the optimization functions for the best weights in each iteration of the example in the table.

This step was developed using the TSQL and SQL Server 2008 database. As the network is considered small, it takes no more than 20 seconds, however the time can increase depending on the network size. Below, we describe part of the algorithm focusing on genetic algorithm procedure.

Below it is described the recommendation process for only one user. The experiment is executed for a set of user, so the recommendation process is performed for each one of this group.

In the first step a set of candidate is chosen. This set is composed by users with higher chances to be recommended in relation to the central user. In the second step, the index measures between each candidate and the central user is evaluated once, then the genetic algorithm is used to select the best weight combination to balance the indexes for each candidate. Below the summary of the main code used for the genetic algorithm procedure is presented.

The main procedure that return the best weight combination for recommend friends for a specific user:

- *Generate an initial population with random weight values.*
- *Until the fitness function value of the best individual of the population do not improve for four generations, do:*
  - *Evaluate the fitness function for each individual in the population.*
  - *Exclude the worst individuals in the population according the fitness function value.*
  - *Generate new individuals applying crossover in remaining individuals.*
  - *Apply mutation operations on children.*
  - *Merge children and parents eliminating duplicates.*
- *Return the weight combination of the best individual*

The function to generate population sons from two parents is summarized as:

- *Crossover is applied over each pair from the Cartesian product of individual of the elite population.*
- *Crossover is applied on same chromosomes type of two parents. Two chromosomes generate two new chromosomes, and the combination of three chromosomes from each parent six different new individual can be generated. Crossover is performed in a single random cutoff point*

The fitness function, considering the candidate set of users as the real candidates plus all friends already connected to the central user. Also all three indexes between each candidate and the central user are evaluated, this procedure is summarized:

- *The weight combination is received as parameter and normalized.*
- *For each candidate, the weighted average is evaluated using indexes and a set of weight.*
- *The result set is sorted in descending order over the weighted average measure.*
- *A new set is created selecting the position of all the candidates that are friends of the central user i.e. that are already connected to it.*
- *The fitness function measure is defined as the average of all entries in the previous set*

The weight combination solution is obtained by applying the genetic algorithm. Using this combination the initial set of candidate is ordered to produce the recommendation list. The filtering step can reduce the set of candidate based on the first index, avoiding measuring the second and third index for less relevant nodes. The indexes evaluated from nodes n_1 to n_2 are the same as n_2 to n_1, so reducing by half the time needed for the solution for the entire network. This does not mean that if n_2 is recommended for n_1, n_1 will be recommended for n_2, since each user has his own context with singular weight combination. In a applied situation, the recommendation can be previously generated, and the weight combination could be preserved for the user for a period of time and them recalculated, depending on increase of the friend network list of the user.

IV. Experiments

Usually experiments to validate solutions in machine learning uses well established data sets. This data are normally divided in training, validation and test set which are prepared specifically for the purpose of evaluating the performance of the algorithm used to solve the problem. The procedures enable the comparison of different solutions obtained from different method applied in the same data set. However, the problem of relationship recommendation in social networks does not have a straightforward method to validate the results since the objective is to generate a list of recommendation that is dependable on the user.

Complex networks have been shown to be a promising area for research in recent years. Several problems can be modeled in network structures. Despite all the progress, recommendation in complex networks has been a problem poorly explored. A validation technique to assess the quality of the proposed algorithm would be necessary to compare the recommended list of relationships, to a certain user, with the desired objective. For privacy reasons social networks normally do not provide or give access to information necessary to perform this experiments. Most of the times, each SNSs develops their own methods and algorithms for recommendation but the limited access to this method make it difficult a cross-validation. A direct consequence of this fact is that there is no public data set prepared for experiments on the recommendation of relationships in social networks.

To solve the problem of lack of a common public data, we obtained through C.E.S.A.R the data of their social network, Oro-Aro to test the proposed method. The experiment consisted of analyzing the algorithm being used by a selected group of users. For each user we present a list of recommendations of new relationships. Then, we evaluated the acceptance of the recommendations. For comparison, we used the algorithm FOF (friend-of-friend) for a subset of the users. Of the 655 users of the Oro-Aro 70 were selected. They satisfy the condition of having at least 13 relationships and have accessed the network in the last 45 days. The first condition is necessary because the algorithm uses the topology, making sure that the subnet of each user with a minimum size is relevant and that it display a pattern. The second condition focuses on the users most likely to respond to the experiment. For each of the 70 selected user, it was generated a set of 10 recommendations, among which 20% using the FOF algorithm. However, before sending the recommended list of friends to the recipients, we still have to replace those that were already part of their friends list before the pre-processing network, i.e., the one-way relationships. This step conveys the first result. We call it the correct set of one-way relationships. Our proposed solution presents a rate of 20.83% while the FOF algorithm achieved 13.33%. Note that the low value for this rate does not imply a bad results since a person can link to another person without having a common interest. These are common cases of celebrity user such as movie actors. In the twitter site the best connect person does not necessary is connected to all the users that are connected to him. After this procedure, the recommendations are sent in the form of links to the profile of users recommended and instructions to accept or reject the recommendation. After a period of 10 days, 31 of 70 users used the recommendations. It was counted only the acceptance of recommendation without the need of the recommended user to have done the reverse action. Using our algorithm 77.69% of the recommendations were accepted while the FOF algorithm achieved 72.22% success.

TABLE II
Experimental Results

	GBFRGA	FOF
correct set of one-way relationships	20,83%	13,33%
Acceptance of the recommendation	77,69%	72,22%

V. Conclusion and Future Works

We have presented a friend recommendation system based on the topology of a social network. The knowledge of the structure and topology of these SNSs combined with quantitative properties of the graph are used to develop the recommendation system. The social network used, the Oro-Aro, is smaller than the most of the popular social networks, such as Facebook, Orkut or Myspace. Besides being smaller in size, it is also less accessed, hence the rate of 44% of user response to the experiment. As expected, our algorithm was better than the FOF, another solution that is also based on network topology. Despite the small difference between the performances, we believe that the size and dynamics of the network plays a major role. The Oro-Aro network has only 634 users, and the average of the friend per users is only 8.1. Thus, we can assume that the FOF algorithm performs well in small networks; however the resulting list is probably small. In larger networks, like Facebook or Orkut in which the average friend size exceeds 200, the FOF could not distinguish the best recommendations when there are small differences in FOF criteria. The reason why our solution can perform better in larger networks is because of its hybrid nature of taking in consideration three different indexes. Since the FOF is used as part of our solution which implies

that in the worst case the algorithm have the same performance, and that the combined clustering indexes with the adaptive feature of the Genetic Algorithm is responsible for the improved performance. The experiments results showed to be very promising; however it is still necessary to apply this algorithm in networks much larger in size and activity. Several papers [1], [2], [3], [4] and [7] have investigated the topological structure of SNSs and others complex networks. However, we have showed that these analyses can be used to develop some practical applications in the field of recommendation system.

For future work, it is important to test the proposed mechanism more intensively in a larger network using several test groups. In addition, this approach based on network topology can be also used for other type of recommendation networks system apart from SNSs. An example is the innumerous e-commerce systems that constitute bipartite networks composed by users and products and could be mapped as single entity graphs.

ACKNOWLEDGMENT

We would like to acknowledge the Recife Center for Advanced Studies and Systems (C.E.S.A.R) for providing the Oro-Aro database. In particular we would like to thanks Ricardo Araújo Costa for assistance in managing the database information. Also, FACEPE – Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, for financial support.

REFERENCES

[1] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacherjee, B. Measurement and Analysis of Online Social Networks. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. (San Diego, California, USA. October 24-26, 2007). IMC'07. ACM Press, New York, NY, 29 - 42.

[2] Ahn,Y.Y., Han,S., Kwak, H., Moon, S., and Jeong, H. Analysis of Topological Characteristics of Huge Online Social Networking Services. Proceeding of the 16th International World Wide Web Conference, (Banff, Alberta, Canada, May 8-12, 2007).WWW '07. ACM Press, New York, NY, 835-844, 2007.

[3] Wilson, M., and Nicholas, C. Topological Analysis of an Online Social Network for Older Adults. Proceeding of the 2008 ACM workshop on Search in Social Media. Napa Valley, California, USA, October 30, 2008). SSM'08. ACM Press, New York, NY, 51-58, 2008.

[4] Mitchell, M. Complex Systems: Network Thinking. Artificial Intelligence, 170(18), pp. 1194-1212, 2006.

[5] Berkowitz, S. D. An Introduction to Structural Analysis: The Network Approach to Social Research. Toronto: Butterwort, 1982

[6] Chau, D. H., Pandit, S., Wang, S., and Faloutsos, C., Parallel Crawling for Online Social Networks**.** Proceedings of the 16th international conference on World Wide Web. (Banff, Alberta, Canada, May 8-12, 2007).WWW '07. ACM Press, New York, NY, 1283 – 1284, 2007.

[7] Kumar, R., Novak, J., and Tomkins, A. Structure and Evolution of Online Social Networks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. (Philadelphia, PA, USA, August 20-23, 2006). KDD'06. ACM Press, New York, NY, 611 – 617. 2006.

[8] Karrer, B., Levina, E., and Newman, M.E.J. 2008. Robustness of community structure in networks. Phys Rev. E 77 046119, 2008.

[9] R. Albert and A.-L. Barabási. 200. Topology of complex networks: Local events and universality. Phys. Rev. Let. 85, 5234-5237, 2000.

[10] Goldberg, David E. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison Wesley (1989).

[11] Guy, I., Ronen I., and Wilcox E. Do you know? Recommending people to invite into your social network. Proc. IUI pp. 77-86. 2009.

[12] Chen, J., Geyer, W., Dugan, C., and Guy, I. Make new friends, but keep the old: recommending people on social networking sites. Proc. CHI, pp. 201-210. 2009.

[13] Liben-Nowell, D., and Kleinberg, J. The link prediction problem for social networks. Proceedings of the twelfth international conference on Information and knowledge management, pp 556-559. 2003.

[14] R. Bell, Y. Koren, C. Volinsky (2008). "The BellKor solution to the Netflix Prize 2008". http://www.netflixprize.com/assets/ProgressPrize2008_BellKor.pdf