Discovering compatible users in social networks

Arefeh Kazemi*,
Mohammad Ali Nematbakhsh and
Mohammad Mehdi Keikha

Computer Engineering Department,
University of Isfahan,
Hezarjerib St., Isfahan, Iran
Email: akazemi@eng.ui.ac.ir
Email: nematbakhsh@eng.ui.ac.ir
Email: mehdi.keikha@cs.usb.ac.ir
*Corresponding author

Abstract: One of the most significant features of social networks is the connections made between new compatible individuals. Quite a few research projects have been done on making such connections more convenient and possible. Almost all of these systems interpret compatibility as similarity and consider only exact-matching of users' interests. This research proposes a different approach for finding 'new' compatible friends through social networks. We have proposed three new relations among users' interests. These are, firstly semantic similarity, secondly conceptual complement, and thirdly associative complement. We have used the first two relations in the current system and the third relation is left for future works. We chose 50 members of the LiveJournal and calculated the degree of compatibility between each pair. The results showing an average error of 0.21 which is acceptable in comparison with the previous exact-matching systems. In the latter, the average rate of error was 0.54.

Keywords: social networks; compatible persons; semantic relation; complementary relation.

Reference to this paper should be made as follows: Kazemi, A., Nematbakhsh, M.A. and Keikha, M.M. (2015) 'Discovering compatible users in social networks', *Int. J. Social Network Mining*, Vol. 2, No. 1, pp.1–18.

Biographical notes: Arefeh Kazemi is a PhD student of Computer Engineering in the School of Engineering at the University of Isfahan. She received her BSc and MSc in Computer Engineering from the University of Isfahan in 2008 and 2010, respectively. Her main research interests include social network analysis, natural language processing and machine translation.

Mohammad Ali Nematbakhsh is an Associate Professor of Computer Engineering in the School of Engineering at the University of Isfahan. He received his BSc in Electrical Engineering from Louisiana Tech University in 1981 and his MSc and PhD degrees in Electrical and Computer Engineering from the University of Arizona in 1983 and 1987, respectively. He has published three US patents and co-authored a database book that is widely used in universities. His main research interests include multi-agent systems applications and web semantics in e-commerce and computer networks.

Mohammad Mehdi Keikha is a PhD student of Computer Engineering in Department of Software Engineering at the University of Tehran. He received his BS in Computer Software Engineering from Ferdowsi University of Mashhad, Iran in 2008 and his MS in Artificial Intelligence from the University of Isfahan, Iran in 2010. His research currently focused on ontology matching and semantic web techniques. His interests include semantic web, intelligent agents, e-commerce, data mining, evolutionary algorithms and neural networks.

This paper is a revised and expanded version of a paper entitled 'Finding compatible people on social networking sites, a semantic technology approach' presented at the Intelligent Systems, Modelling and Simulation (ISMS), Kuala Lumpur, 24 January 2011.

1 Introduction

At present, social networks are enjoying a significant increase in popularity, attracting millions of users. Many of these users have integrated these sites into their lifestyles and daily routines. Through social networking, people can share items and information, read relevant news, post comments, etc. However, what makes these sites unique is the further use they have in allowing people to connect to one another through the cyber world. In fact, the main reason for popularity of these sites, making them different from the other cyber spaces, is their 'social' properties. Research shows that as people's connections develop through a social network, there is a significant effect, both on the frequency of use of the network by its members, and on a tendency towards the others in order to become a member of the given network (Moricz et al., 2010). In addition, if these connections occur between compatible people, they lead to fruitful sharing of items and information.

Further, related statistics show that users of social networks are as likely to connect with people that they do not know (considered as 'new' people) as with existing friends. Therefore, enabling searches for known friends and providing opportunities for finding new compatible people to connect with, can be challenging for social networking organisations (DiMicco et al., 2008; Donah et al., 2008).

Many popular social networks such as Facebook and Myspace provide friend-recommendation services in an attempt to overcome the first problem (searching for known friends). These services usually search for users with whom a user might already be acquainted; therefore the users' connections will be as many as their connections in the real world. In fact, these friend-recommendation systems cannot achieve one of the main objectives of the social networks, which is 'increasing' the connections among new users. However, a few friend-recommender systems, aiming to find 'new' compatible friends through social networks, have appeared. In these systems, the compatibility between users is only interpreted as similarity between them. In addition, in order to find similar users, the researchers considered only exact similarity matching among the users' interests, which are believed not to be efficient. These systems use mainly a set of predefined and fixed items, from which users are supposed to specify their interests. Consequently they are not able to describe their interests in their own words.

In order to overcome the mentioned deficiencies, we designed and developed a system to find a degree of compatibility among users of a social network. This system was to be used for finding 'new' compatible friends in friend-recommender systems. In the proposed system, contrary to existing systems, users can specify their interests in their own words; thus, there will be elimination of a major limitation.

The remainder of this article is organised as follows. Firstly, after discussing related works in Section 2, we begin by defining compatibility among individuals and modelling it by defining three relations among the interests of the users in Section 3. We next explain how the relations among the interests can be extracted in Section 4. Then, in Section 5, we describe the system architecture. After that the experimental setup, evaluation and some comparisons with previous approaches are presented, in Section 6 and Section 7. Finally, the conclusions that can be derived from this study are developed in Section 8.

2 Related work

The main purpose of individuals forming social networks is to achieve connection through the internet. Making it more convenient, quite a few researchers have developed viewpoints on friend-recommendation systems. The systems they have described can largely be classified into two groups.

The first group includes the systems that aim to search for the friends that the user may know in the real world. These researches do not rely on the actual interests of the users and usually only utilise the explicit or implicit links among users, where *explicit* links are the friendship links among users in the social network, and *implicit* ones are built from cooperation among them in the real world, such as collaboration in writing a research paper.

The second group of the researches utilises the users' interest in order to find new friends for them. They estimate similarity degrees among users' interests, and recommend those with interests that correspond, to connect with them. In these researches, the similarity degree between two users is estimated as the number of the 'same' words that they have in common in their content or interests. We refer to this as 'exact similarity matching' and believe that is not efficient.

Chen et al. (2009) has studied four major friend-recommendation algorithms which were designed to help users to find the contacts whom they knew, as well as finding new friends. These four algorithms are: content matching; content plus link; friend of friend; and SONAR. In the first algorithm, similarities of content that users post in a social network is considered as a matching device, and the users associated with similar content will be recommended to make contact. In this algorithm, two contents are considered as being similar if they have many common words with each other. In the second algorithm, the content-matching algorithm is enhanced with social links and information derived from the structure of social networks. The motivation behind this algorithm is that, by disclosing a network path to an unknown person, the recipient of the recommendation will be more likely to accept the recommendation. In the third algorithm however, the friends of the friends of the user are recommended as new friends. Finally in the fourth algorithm, they aggregate social information from different public data sources and discover social relationships among users in the real world in order to find new friends.

Nisgav and Patt-Shamir (2009) proposed an approach to finding similar users in a social network. They assumed that there are *n* users and *m* possible questions. Each user has an answer for each question that represents its interests; however these answers are not known when the algorithm starts. Each user has a vector of preferences, where coordinates correspond to the questions and entries correspond to the answers. Two users are similar if their preference vectors differ in only a few coordinates.

Lo and Lin (2006) proposed an algorithm called WMR which generated a limited, ordered and personalised friend list based on the number of messages communicated among the users in the social network. In this approach, the authors assumed that if there was no interaction between two persons, it would be hard to call them friends, and vice versa. If there are many number of friends communicated between persons A and B, and also B and C, then A and C will be recommend to connect each other as friends.

Chu et al. (2013) have proposed a brand-new friend recommendation approach that utilised location similarity, interest similarity and friendship for recommending new friends to the users. In order to gain the interests similarity, they used pattern matching method.

Some popular social networks such as facebook.com and orkut.com have a friend-recommendation service. Facebook's (2011) friend-recommender system recommends users to connect with one another, based on a 'mutual friends' approach.

In this approach, two users will be recommended to connect each other if they have a number of friends in common. Likewise, Orkut (2008) recommends the friends of the friends of a user to him. However, data on the effectiveness of these approaches is not available at present.

Table 1 Comparison of existing friend recommender systems

Research/system	*Objective*	*Approach*
Chu et al.	Find new friends	Using interests, location and dwell time to recommend a new friend, Pattern matching among the users' interests
Chen et al. Content matching	Find new friends	Exact similarity matching among the used words in the users' content
Chen et al. Content plus link	Find known friends	Enhancing content matching algorithm with social link information
Chen et al. Friend of friends	Find known friends	Friend of friends
Chen et al. SONAR	Find known friends	Use public data sources to discover users relationship in the real world
Nisgav	Find new friends	Exact similarity matching among users' answers to some questions
Lo and Lin	Find known friends	Using the number of messages communicated among users
Facebook	Find known friends	Mutual friends
Orkut	Find known friends	Friend of friend
The proposed system	Find new friends	Using the semantic similarity and conceptual complement relations among users' interests

Table 1 displays a comparison among the mentioned approaches and the proposed system. As it is shown, all of the previous approaches except that of Nisgav, 'content matching' and 'content plus link algorithm', are in the first group of our initial classification, which recommends individuals that a user may know in the real world. In addition, the approach in the second group, that recommends new friends, uses exact similarity matching between users' interests to find new friends.

In this paper, we have designed and implemented a method to find 'new' friends in a social network site. For this purpose, in contrast to the previous systems that use exact string matching, we have used the *semantic similarity* and *conceptual complement* relations among users' interests.

3 Compatibility among users

This research is an attempt to help users find compatible friends on social networking sites. For this purpose, the first task is to define precisely the term 'compatible'. Existing approaches for finding compatible users compute the degree of compatibility between two users as based on the number of interests that they have in common. In other words, in these systems, the term 'compatible' is interpreted as 'similarity'. Thus two users will be considered compatible only if they share similar interests. Focusing on the main motivations which shape friendship between people may seem logical; however the general indications are that this is not always the cause of compatibility. To clarify this point, consider two persons A and B in a social network with the following interests[1]:

- *Person A:* photography, football, fashion designing
- *Person B:* taking pictures, basketball, capitalising in fashion.

We know that if these users meet each other, they may be friends even though they have no common interests. Therefore we argue that exact similarity matching between users' interests is only one factor in deciding whether two users are compatible or not. The authors of this paper believe that other factors that affect compatibility have been disregarded in previous works.

In this research and our previous work (Kazemi and Nematbakhsh, 2011), we define compatibility as being among users with similarity and complementary relations between them. We firstly define compatibility relations between users, afterwards modelling these relations through some new relations among interests. Two individuals are considered similar to one another if they share similar interests, i.e. their interests refer to the same meaning. For example, a user with an interest in 'photography' is similar to a person with an interest in 'taking pictures'. We also consider two users complementary to each other if they can collaborate with each other in order to improve their knowledge of a subject of common interest, or satisfy each other's needs in order to achieve a common objective. For example, a person with an interest in 'football' is a complement to a person who is interested in 'basketball', since they can complete their information regarding 'athletic games'. Likewise, a person with an interest in 'fashion designing' is a complement to someone with an interest in 'capitalising in fashion' because each can satisfy the other's needs to achieve their common objective – 'fashion businesses'.

To facilitate the modelling of complement and similar users, we define the following three new relations among users' interests that effect compatibility between them:

1 semantic similarity

2 conceptual complement

3 associative complement.

The first relation is applied in modelling similar users; while the other two are applied in modelling complementary ones.

Defining semantic similarity between two interests is straightforward, while defining the conceptual and associative complement needs further investigation. In order to obtain some basic ideas for defining complementary relations, we investigate the definitions of complementary relations in other scientific fields. It seems that the definition of complementary sets in Mathematics is closely related to our work.

In Mathematics, a complement of a set A refers to things that not in A. The relative complement of A with respect to set B is the set of elements in B but not in A. Meanwhile, all sets under consideration are said to be subsets of a given set U.

From the definition of complement sets, it is clear that two sets can be complements of each other only if both of them belong to the same common set. This common set is named the *universal set* in the Mathematics. In other words, this universal set is the key point that leads to a complementary relation between the two sets. These two sets complement each other to make a whole universal set. This simple and major point guides us to an interesting idea, which leads us to define complementary relation between interests: "Two interests are complements of each other if they are similar on a higher level in an IS-A or PART-OF hierarchy". With this idea, notable because these are the first definitions that can be applied in modelling the complementary relations among the users, we define the relations among interests as follows.

1 *Definition 1: semantic similarity*

   Two interests are in semantic similarity relation if both refer to the same meaning. In other words, two interests are semantically similar if they are synonyms.

   'Photography' and 'taking pictures are in semantic similarity relation since they are synonyms, according to the above definition.

2 *Definition 2: conceptual complement*

   Two interests are in conceptual complementary relation if both belong to the same concept. In other words, two interests are conceptual complements if they are in an IS-A relation with the same concept.

   'Football' and 'basketball' are in conceptual complementary relation since both of them are in IS-A relation with athletic games concept, according to the above definition.

3 *Definition 3: associative complement*

Two interests are in associative complementary relation if both are constituent parts of a same concept, in other words, they are in PART-OF relation with the same concept.

'Fashion designing' and 'capitalising in fashion' are in associative complementary relation because are in the PART-OF relation with 'Fashion business' concept.

It is worth mentioning that in definition 1 and 2, the 'same concept' is the key reason that causes the complementary relationship between two users, and can be considered as the universal set in the definition of complement sets in Mathematics. In fact, this common interest concept is an illustration of the cause of complementary relation between these users.

After defining relations between interests, it is the next phase, that of extracting these relations among the users' interests, that is explained in the next section.

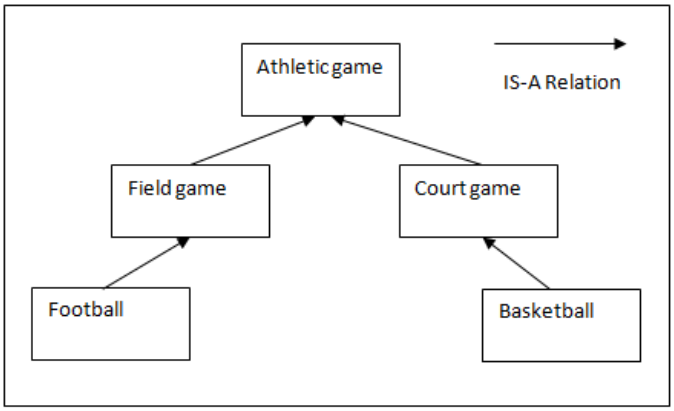
4 Extracting complementary relations among the users' interests

As mentioned before, two interests are called conceptual- or associative-complement if they are in IS-A or PART-OF relation with a common concept. We used WordNet (Rita Wordnet, 2013) for extracting these relations.

WordNet is a large comprehensive database of English words that provides the system with some relations between words. Three relations that are defined in WordNet are through the terms 'synonym', 'hypernym' and 'holonym' (Wordnet, 2013). X and Y are synonym if both of them denote the same concept. For example car and automobile are synonymous. Y is a hypernym of X if every X is a (kind of) Y. For example, *canine* is a hypernym of *dog*, because every dog is a member of the larger category of canines. Moreover, Y is a homonym of X if X is a part of Y. For instance, *building* is a holonym of *window*. In fact, WordNet arranges words in an IS-A hierarchy. Obviously the hypernym and holonym relations specify the IS-A and PART-OF relations respectively; therefore, they can be used in finding conceptual and associative complementary relations among the users' preferences. Two interests are conceptual- or associative-complement, if they have the same hypernym or holonym in an acceptable short distance in IS-A or PART-OF hierarchy. Moreover, two words are in semantic-similarity relation if they are both synonyms in WordNet. In our initial work, semantic similarity and conceptual complementary relation were extracted while associative complementary relation was left for future study.

Rita WordNet distance function was used in measuring the distance between interests in IS-A hierarchy. If the distance between two interests is zero, they are synonyms. If the distance is not zero but is less than a predefined threshold, then the interests are in conceptual complementary relation with each other. This threshold should be computed by training on the train dataset – explained in Section 7.1.

Figure 1 'Football' and 'basketball' in the WordNet noun taxonomy



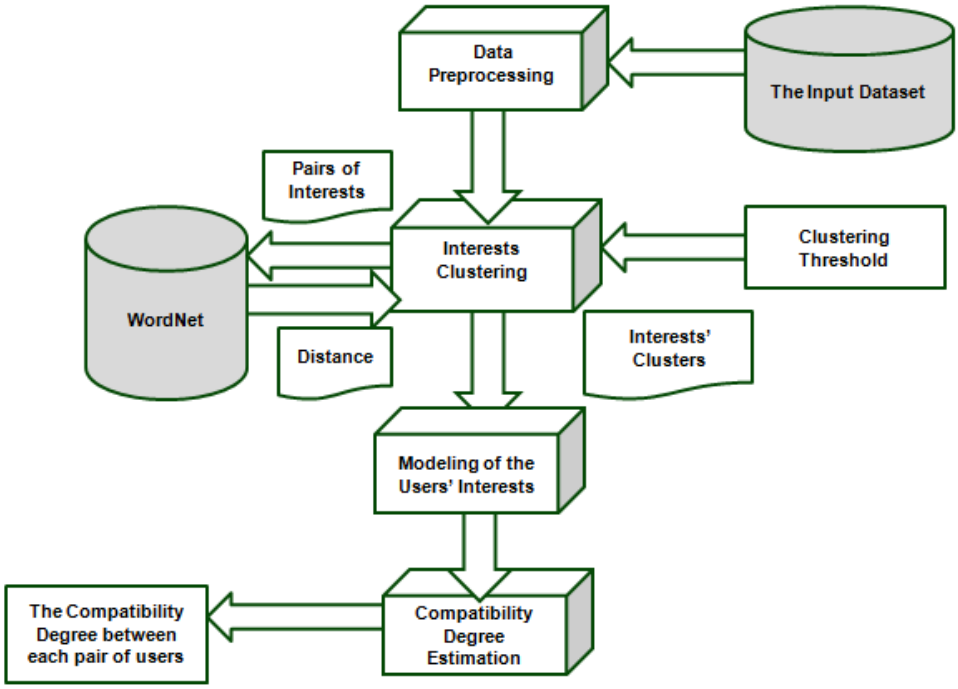
To clarify this point, consider our previous example of football and basketball interests. These two interests are not synonymous, so they are not semantically similar. In order to find complementary relation, the distance between them in IS-A hierarchy must be estimated. For simplicity, suppose that the distance between two words is the number of links between them. These words in IS-A hierarchy are shown in Figure 1. Here, the common concept that connects these two words is *athletic games* and the distance between them is 4. Suppose that the trained threshold distance is 5. If so they will be considered as complement interests. The result is desirable because two users with these interests can collude with each other in getting information about the athletic games as a common interest.

5 System architecture

The input of this system is a dataset consisting of a number of social network users, with their interests. The objective here is to estimate the compatibility degree between each pair within the group. To begin with, this dataset was preprocessed to remove irrelevant and meaningless words. Then, the compatibility relations among interests were extracted, as illustrated in Section 4. Finally, the compatibility degree among users was estimated, based on the number of compatibility relations between their interests. To reduce the system response time, a clustering algorithm was also used.

The architecture of the system is shown in Figure 2. Here, we observe that the system consists of four main units: data preprocessing, interests clustering, users' interests modelling, and compatibility degree estimation unit. The contribution and responsibility of each unit is described in detail in the next section.

Figure 2 Architecture of the proposed system (see online version for colours)



5.1 Data preprocessing

The input dataset consist of the interests that each user states about him or her self in a personal profile. As a result it may contain meaningless and irrelevant words. Any interest not registered in the WordNet dictionary is considered meaningless. In the data preprocessing unit, all of the meaningless and irrelevant words were removed from the input dataset. The ratio of the meaningless interests to the users' interests is only 3.6% which shows that the most of the interests that users specify in their own words are registered in the Wordnet.

5.2 Interests clustering

In order to estimate the compatibility degree between each pair of persons in the input dataset, the compatibility relations between each pair of their interests should be extracted. Suppose that there are n users in the input dataset and each user has m interests. The system's response time to estimate compatibility degree among each pair will thus be computed by the following equation:

$$\binom{n}{2} \times m^2 \times t_{WordNet} + \binom{n}{2} \times m^2 \tag{1}$$

The first expression of the equation (1) represents the needed time to extract compatibility relations between each pair of interests, using Wordnet – taking $t_{WordNet}$ time. After extracting these relations, the next step is to estimate the compatibility degree between each pair of users by counting the number of compatibility relations between their interests – taking $\binom{n}{2} \times m^2$ time. Since the $t_{WordNet}$ is long, the second part of the equation can be ignored and the order of the equation (1) is n^2m^2. Given that the number of users in a social network is potentially great, the response time of the system is extremely high. To decrease the system time, a clustering method was used in order to reduce the coefficient of $t_{WordNet}$ in the equation (1).

Using the clustering method, those interests that are close and compatible to each other in IS-A hierarchy in WordNet are classified in a same cluster. After clustering, each interest will be compared to only the clusters' centres, instead of the whole interests; hence, a significant decrease in system time. The system time in this case is as equation (2) where C indicates the number of clusters. The first and second expressions of the equation (2) represent the time necessary to cluster the users' interests and the time for calculating the compatibility degree between each pair of users respectively. Since the $t_{WordNet}$ is long, the second part can be ignored, and the order of equation (2) is nm.

$$n \times m \times C \times t_{WordNet} + \binom{n}{2} \times m^2 \tag{2}$$

The clustering algorithm that was used is shown in Figure 3. In this algorithm, if the distance between two interests is less than a predefined threshold, they are compatible with each other and will be posed in the same cluster.

Figure 3 The clustering algorithm

1. *Initial the number of clusters to 1 and set the center of cluster*$_1$ *as interest*$_1$ *in the data set.*

2. *For each interest*$_i$ *of each person*$_j$ *in the data set, do the following two steps:*

   a. Compute the WordNet distance between interest$_i$ *and each of the existing cluster centers CC. Let m be the cluster index such that the distance between interest*$_i$ *and* CC_m *is the smallest. Name this distance as mindistance.*

   b. If the mindistance≤Threshold, then put the interest$_i$ *incluster*$_m$*. If mindistance>Threshold, then add a new cluster to the existing clusters and set interest*$_i$ *as its center.*

5.3 Modelling of the users' interests

After clustering the users' interests, each user was modelled with a preference vector, where coordinates correspond to clusters' numbers and the entries correspond to the

number of user's interests that are in this cluster. Equation (3) shows the preference vector initialisation for person *A*.

$$PreferVector\ A[i] = the\ number\ of\ interests\ of\ person\ A\ that\ are\ in\ cluster\ i \quad (3)$$

After constructing the preference vector for every user in the input dataset, these vectors were used to estimate the compatibility degree between each pair of users in the input dataset.

5.4 Compatibility degree estimation

As a final point, the compatibility degree between each pair of users was estimated, based on the total number of compatibility relations between their interests as equation (4).

$$personsComDeg(Person\ A, Person\ B) = \overrightarrow{PreferVector\ A} \,.\, \overrightarrow{PreferVector\ B} \quad (4)$$

In equation (4), the inner product between two users' preference vectors is equal to the number of complementary relations between their interests. Through the outcome of this equation, we can determine the compatible individuals in the social network and recommend them to connect with one another.

6 Experimental setup

To evaluate our system and compare it with the latest achievement in this field, a reference dataset that contains users' interests and the real compatibility degree between each pair is required. Satisfying this requirement, the relevant sources were searched and no such sets were found.

For assessing the quality of a method that estimates the compatibility degree between users, i.e., its accuracy, a common method is to compare its output against in human judgments. The more the means was similar to human judgment, the more accurate it was. To evaluate the proposed method, it was necessary to have a reference dataset which would consist of the following: some users along with their interests; also the human ratings for compatibility degree between each.

We employed the dataset crawled on July, 2010 from *LiveJournal* (2013) which is available in social computing data repository at ASU (Zafarani and Liu, 2009). This dataset contains some information about users like their preferences. We randomly sampled 50 users from this dataset, each of whom had 30 words indicating his or her preferences. Since neither in the dataset nor on the web was there reference to degree of compatibility among users, in order to evaluate our system we made the reference dataset manually. We first discovered complementary and semantic similarity relations between each pair of preferences manually. Then we calculated the compatibility degree among the users (namely real compatibility degree) by counting the number of compatibility relations among their interests.

We used this file as a reference dataset to evaluate our system, estimating the compatibility degree within each pair of users in the dataset with our proposed system. An example of estimated compatibility degree among four users in the input set is described in Appendix.

7 Evaluation

We applied K-fold cross validation to calculate the average system error for finding the optimal values for clustering threshold and number of users' interests. Since there are 50 users in our dataset we chose $K = 5$ and divided the data into five equal parts. For each part $k = 1, 2,\ldots, 5$, we fitted the model with parameter λ to the other K-1 parts as equation (5) and equation (6) and computed its error on the k^{th} part based on the equation (7). We trained the system twice. In the first round, we considered the number of interests constant and tried to find the optimal clustering threshold. In the second round, we used the optimal clustering threshold to find the optimal number of users' interests. In each round, λ is considered as the parameter which should be optimised.

$$Error_{i,j}(\lambda) = \frac{abs\begin{pmatrix} real\ compatibilityDegree_{i,j} \\ -system\ compatibilityDegree_{i,j}(\lambda) \end{pmatrix}}{\max\begin{pmatrix} real\ compatibilityDegree_{i,j}, \\ system\ compatibilityDegree_{i,j}(\lambda) \end{pmatrix}} \tag{5}$$

$$\hat{\beta}^{-k}(\lambda) = \arg\min_{\lambda} \sum_{i,j \in other K-1 parts} Error_{i,j}(\lambda) \tag{6}$$

$$systemError_k(\lambda) = \frac{\sum_{i,j \in k'th\ part} Error_{i,j}\left(\hat{\beta}^{-k}(\lambda)\right)}{num\ of\ pairs} \tag{7}$$

$$num\ of\ pairs = \begin{pmatrix} \left(\frac{N}{K}\right) \\ 2 \end{pmatrix}$$

In equation (5) $system\ compatibilityDegree_{ij}(\lambda)$ is the compatibility degree between persons i and j that our system had estimated based on the parameter λ. Obviously, equation (5) represents the proportion of the real compatibility degree that the proposed system had estimated it incorrectly; therefore, it is a good representation for the system's fault. For example, if the estimated compatibility degree between two users via our system is 30, and the reference compatibility degree is 50, the system's error for this case is: (50–30) / 50 = 0.4.

Equation (6) selects the best value of the parameter λ that fits to the K-1 parts of the dataset. In equation (7) we made an estimation of the overall error of the system on the k^{th} part of the data which was to be the average of calculated errors for each pair of persons in this part. Since we calculated compatibility degree among each pair once and there are $\frac{N}{K}$ persons in each part, the $\begin{pmatrix} \frac{N}{K} \\ 2 \end{pmatrix}$ values were calculated as the compatibility degree errors among each pair of users. After calculating the system error on each part of the data, the average system error was estimated based on equation (8). As it mentioned before, this error could be used for finding the impact of some parameters λ such as clustering algorithm thresholds, as well as the number of interests for each user in our system

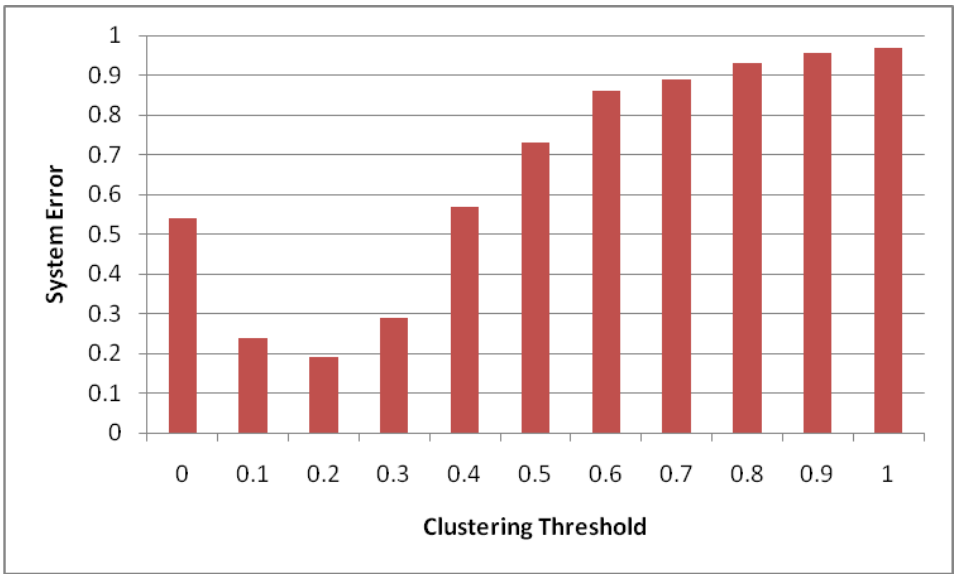
$$AverageSystemError(\lambda) = \frac{1}{K}\sum_{k=1}^{k=K} systemError_k(\lambda) \quad (8)$$

7.1 Finding the optimal clustering threshold

As mentioned before, we should determine an optimal threshold for the clustering algorithm. For this purpose, we split up the reference dataset in five sets for applying 5-fold cross validation. In the training phase, the clustering threshold was altered, with 0.1 intervals between zero and one and the system error for each step was calculated based on the equation (5). Then in the equation (6), the best clustering threshold for this iteration, which had the lowest system error value, was selected for evaluating the system on the test dataset. After calculating the best threshold for each part of data, we selected the final optimal threshold as the average of these thresholds. The optimal cluster threshold in the dataset is 0.18. We therefore chose 0.18 as a clustering threshold, in order to evaluate our system. The average system error with this clustering threshold is 0.21, explained in Section 7.3.

The impact of the clustering threshold on system error for all of the iterations in 5-fold cross validation was identical. Figure 4 shows the diagram of changing system error based on different clustering thresholds in the first iteration. Since the clustering threshold determines the maximum distance between the interests in a cluster, the number of clusters becomes lower and each cluster should contain more interests if the threshold is high. In addition, if two interests had a same cluster, it means that they are two compatible interests. Higher clustering threshold therefore indicates that the number of interests that have compatibility relation with each other is also high.

Figure 4 The impact of the clustering threshold on average system error in the first iteration (see online version for colours)



For example, when clustering threshold is one, all interests take place in one cluster so consequently, all of them are considered compatible with each other. On the other hand, if the clustering threshold is zero, each interest has to take place in a separate cluster. This means that, only two interests that are represented by the same string will take place in the same cluster; therefore, the noted condition here is exactly similar to the strategy of the previous recommender systems (which use only string matching strategy for finding compatible users).

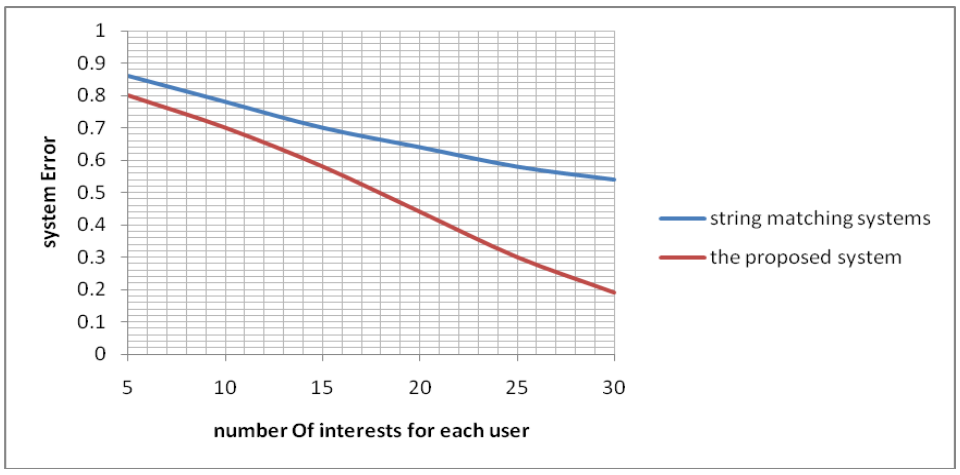
According to Figure 4, the optimal cluster threshold in the first iteration is 0.2.While the system error of this point is 0.19. The system Error decreases when the cluster threshold changes between zero and 0.2 increasing the cluster threshold results in more compatible interests that pose correctly in the same cluster. Increasing the cluster threshold from 0.2 increases the system error. This is because the number of interests in the same cluster increases, and consequently the number of interest that have been identified incorrectly as compatible interests with each other will be increased. By increasing the clustering threshold, the number of clusters(C) will decrease, and consequently, according to equation (2), the system's speed will increase.

According the Figure 4 and equation (2), there is a tradeoff between the system's speed and its accuracy. By increasing the clustering threshold from 0.2 to 1, the system's speed could be increased. This would be, however, at the expense of a decrease in the system's accuracy.

7.2 The impact of the number of users' interests on the average system error

In the input dataset, each user has 30 interests, describing himself or herself in 30 words. Changing the number of interests for each user affects both the response time and accuracy of the system. We can find an optimal value for the needed number of the interests offered by each user, in order to find accurate compatible persons for them within reasonable time. The impact of the number of user's interests on system error for all of the iterations in 5-fold cross validation was similar. The impact of the number of interests on the system error in the first iteration of cross validation in comparison with previous string matching systems, are shown in Figure 5.

Figure 5 The impact of the number of interests for each person – in the proposed system and in previous string matching systems (see online version for colours)



As can be seen in Figure 5, the systems' error decreases in accordance with a logarithmic function of the number of interests for each user. Likewise, the gradient in the proposed system is more than the gradient in the previous systems. We see here that the number of users' interests has a higher impact on the decrease of the proposed system's error, compared with the previous systems. Note that if a person describes himself or herself as having a greater number of interests, the system can gather more information about users' characteristics and consequently compatible friends are found with more accuracy.

7.3 Our system vs. previous systems

Based on Figure 6, the performance of our system has performed about 33% better than previous systems, which used string matching approaches to find similar interests.

The proposed approach to estimation of the compatibility degree among individuals can be used in a friend-recommendation system, allowing us to evaluate the performance of our system from this viewpoint, as well.

In a friend-recommender system, if the estimated compatibility degree between two users exceeds a predefined threshold, they are recommended to be friends. We can estimate the average compatibility degree of each user with the others, and then compare this estimated degree with the real one. If the differences in these values are high, the recommended users, to this user, are not good recommendations, so the proposed approach is not an efficient one for such purposes.

The average of the estimated compatibility degrees in our system, together with exact string matching systems, is shown in Figure 7. This figure shows that the proposed system works better than a string matching system for applying in a friend recommendation system.

Figure 6 The performance of our system vs. the performance of previous systems (see online version for colours)

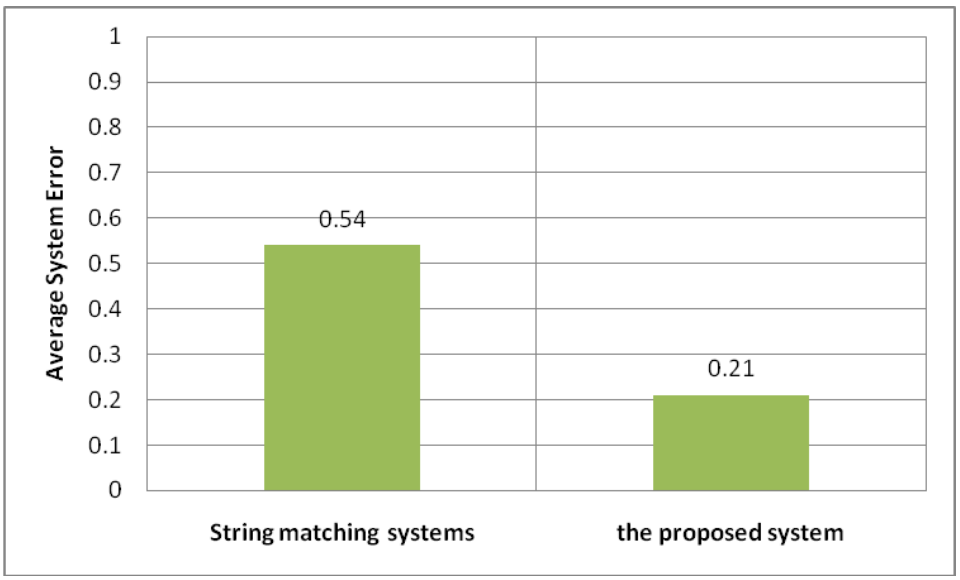
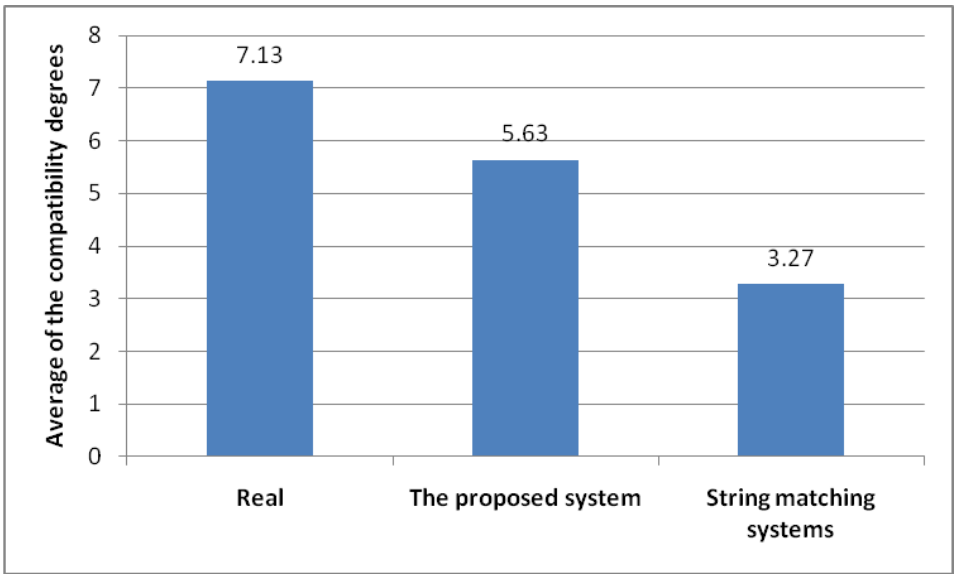


Figure 7 The average of compatibility degrees (see online version for colours)



8 Conclusions

In this study, we proposed a method for matching compatible users in a social networking website. Our system uses a semantic and complementary approach for finding the degree of compatibility among individuals. In order to model compatible users, we defined three new relations among users' interests: *semantic similarity*, *informative complement* and *associative complement*. We used WordNet to elicit the required information for extracting *semantic similarity* and *informative complement* relations. The *associative complement* is left for the future work. We chose 50 members from *LiveJournal* social network as our experimental cases in our study. Then, we calculated compatibility degree between each pair of them. The results show the superiority of our approach because the average error rate in our approach was 0.21 – quite satisfying compared to existing systems', which use only string similarity matching algorithms.

References

Chen, J. et al. (2009) 'Make new friends, but keep the old: recommending people on social networking sites', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pp.201–210, ACM, New York, NY, USA.

Chu, C. et al. (2013) 'Friend recommendation for location-based mobile social networks', *Proceeding of Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pp.365–370.

DiMicco, J. et al. (2008) 'Motivations for social networking at work', *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*, pp.711–720, ACM, New York, NY, USA.

Donah, M. et al. (2008) 'Social network sites: definitions, history, and scholarship', *Journal of Computer-Mediated Communication*, pp.210–230.

FaceBook (2013) [online] http://www.Facebook.com (accessed 20 September 2013).

Kazemi, A. and Nematbakhsh, M. (2011) 'Finding compatible people on social networking sites, a semantic technology approach', *Proceedings of the Intelligent Systems, Modelling and Simulation (ISMS)*, pp.306–309, Kuala Lumpur.

Livejournal (2013) [online] http://www.Livejournal.com (accessed 2013).

Lo, S. and Lin, C. (2006) 'WMR – a graph-based algorithm for friend recommendation', *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)*, pp.121–128, IEEE Computer Society, Washington, DC, USA.

Moricz, M. et al. (2010) 'PYMK: friend recommendation at myspace', *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10)*, pp.999–1002, ACM, New York, NY, USA.

Nisgav, A. and Patt-Shamir, B. (2009) 'Finding similar users in social networks: extended abstract', *Proceedings of the Twenty-First Annual Symposium on Parallelism in Algorithms and Architectures (SPAA '09)*, pp.169–177, ACM, New York, NY, USA.

Orkut now suggestion new friends (2008) [online] http://devilsworkshop.org/orkut-now-suggesting-new-friends-new-feature (accessed 12 January 2008).

Rita Wordnet (2013) [online] http://www.rednoise.org/rita/index.html (accessed 12 September 2013).

Wordnet (2013) [online] http://en.wikipedia.org/wiki/WordNet (accessed 14 February 2013).

Zafarani, R. and Liu, H. (2009) *Social Computing Data Repository at ASU*, Arizona State University, School of Computing, Informatics and Decision Systems Engineering, Tempe, AZ [online] http://socialcomputing.asu.edu (accessed 14 March 2009).

Notes

1 In common, the users in the social network sites express their preferences with single or compound words in their profiles. In this article, we refer to these words as interests.

Appendix

Estimated compatibility degree among four users in the input dataset

user: fireflytrance

Preferences:

absinthe, action, aim, anime, art, astronomy, California, cats, chocolate, clubs, college, comedy, computers, cosmology, dancing, disturbed, dreams, drinking, dvds, fantasy, feminism, flying, friends, guys, html, imperfection, Ireland, Japan, karaoke, London.

user: heathbar224

Preferences:

Ankara, apples, art, berries, Boston, California, Casablanca, Chippendale, clerks, clouds, coffee, Colorado, couscous, crepes, dispatch, earmuffs, eggs, Europe, existentialism, fog, French, goldfish, Italian, kisses, leaves, lemurs, licorice, lightening, London, love

user: hoshinokokoro

Preferences:

animals, animation, anime, art, astrology, astronomy, batman, books, capsule, cartoons, cats, collecting, comics, computers, decent, disney, drawing, ecology, figures, firefly, flying, folklore, food, gaming, gems, genetics, goosebumps, halo, history, internet

user: i_anagram_i

Preferences:

alias, bags, bands, beaches, Boston, Britain, cars, cats, cooking, dancing, democrats, equestrian, food, Germany, icons, ipod, Ireland, lamb, lanyards, makeup, movies, music, politics, psychology, royals, Russia, sailing, shopping, spies, sports

user: kiiitty,

Preferences:

acting, alcohol, anime, art, astrology, astronomy, backs, ballet, black, bondage, boobs, books, bowling, California, campfires, camping, candles, cats, cds, cheese, Christmas, clouds, cloves, comedy, creativity, dancing, dikes, Disney, dogs, drawing

Persons	*Compatibilty degree*
fireflytrance and heathbar224	4.0
fireflytrance and hoshinokokoro	17.0
fireflytrance and i_anagram_i	8.0
fireflytrance and kiiitty	15.0
hoshinokokoro and i_anagram_i	10.0
i_anagram_i and kiiitty	8.0
heathbar224 and hoshinokokoro	3.0
heathbar224 and i_anagram_i	2.0
heathbar224 and kiiitty	5.0
kiiitty and hoshinokokoro	19.0