IBIMA Publishing
Communications of the IBIMA
http://www.ibimapublishing.com/journals/CIBIMA/cibima.html
Vol. 2012 (2012), Article ID 775316, 15 pages
DOI: 10.5171/2012.775316

An Association Rule-Based Recommendation Engine for an Online Dating Site

Civan Özseyhan[1], Bertan Badur[2] and Osman N. Darcan[2]

[1]EBI, Istanbul, Turkey

[2]MIS Department, Boğaziçi University, Istanbul, Turkey

Abstract

Being a popular social network type, online dating sites provide a platform for people to find partners for establishing a relationship. In this study, a recommendation engine for one of the prominent online dating sites of Turkey is developed. It works as a support system to suggest potential matches to the site users. As opposed to the traditional systems that match users based on their revealed preferences, the engine is based on a rule set extracted from the past communication data, using association rule mining. A list of best matches based on scoring derived from these rules is presented. The performance of the engine is statistically tested. It is found that the scores of matching couples are found to be significantly higher than the non-matched couples' scores.

Keywords: online dating, recommendation engine, association rule mining.

Introduction

Social networking communities are defined as online communities that focus on the building and verifying of social networks (Romm and Setzekorn, 2009). Social networking communities offer a range of services such as blogging, group creating, support and e-dating (online dating).

Online dating services are examples of a popular type of social network communities where individuals are matched with other individuals based on criteria that the users specify and/or on criteria that the company gleans from information that is provided by the users. These services provide web sites and other tools for the users who have particular demands and expectations from starting a relationship online. In these web sites, users can create personal ads, also known as "profiles", which allow them to submit information about themselves; including their age, gender, location, physical attributes, socio-cultural status and daily habits. This information is collected in a database to enable other users to search the profile base according to their own preferences and contact the ones fitting the search criteria. Members of dating sites can attach photos and videos to their profiles, browse and display other's visual materials, have instant messaging sessions with other individuals, and they can even join real time audio/video conferences. All these features have become present, even indispensable, in almost all online dating services in recent years. However, some services have accomplished to create a difference in advanced matching, having the matching algorithms as one of their essential features.

After the booming of the internet at the end of the century, online dating was first launched in Turkey in the year 2000. One of Turkey's most prominent online dating sites,

 Contact author: Civan Özseyhan E-mail: civan@idefix.com

siberalem.com, referred to as *siberalem* hereinafter was founded in 2000. Having started as a free service for all users online, the site switched to subscription-based membership in 2002. The latest renewal of siberalem.com was in 2007, when the software infrastructure was modified and reinforced. At the end of 2008, statistics indicate that there are approximately 200,000 registered profiles in the database. This group of users is considered "active users".

The simple scenario in siberalem from the user's perspective follows the steps of registration to the site and creating a profile, then searching and browsing others' profiles and finally initiating a communication by sending messages.

The user starts searching other members according to the criteria he/she expects from his/her potential match. The search form includes many fields from physical attributes to religious and political views. Currently, a basic matching system based on filling an additional form requesting the member's preferences, is present. This system can be used by site members to query and list the best matches.

The user can interact with online users via the instant messaging facility built in the website or he/she can send messages to offline users. If the user receives a response, a conversation usually starts and continues up to three or four messages that follow each other, which end up with an exchange of real e-mail addresses, phone numbers, or user names in major instant messaging services. This is the point where users leave the siberalem.com platform and continue the relationship in other media.

The aim of this study is to develop a recommendation engine for the web site working as a support system for the site members, which is capable to output automatically best matches for the user. The recommendation engine is based on a rule set extracted from the sites past communication and matching data, and requires no further expert knowledge or detailed preference input from the user.

Siberalem's messaging logs include specific information such as messaging frequency and message contents, which can be used for tracking couples developing a relationship via the messaging system. In addition to this, the user database has extensive information about certain characteristics of the users from socio-demographical data to physical attributes, habits and cultural preferences. Given a list of couples flagged as "matches" and their available information, it is possible to search for frequent patterns among the features of matching couples to provide an answer to question "who meets who?"

By using association rule mining functionality, the rules can be extracted from the dataset. It is possible to use these rules to build a recommendation engine capable of making automatic suggestions to the newcomers and provide them with a list of potential matches.

The second section of this paper gives a literature survey of research on online dating. The third section introduces the basic concepts of association rule mining. In the fourth section, the development of the engine is explained with emphasis on the extraction of the rules, scoring and testing of the engine. The last section concludes the paper.

Various Studies on Online Dating

Ellison, Heino and Gibbs (2006) investigate self-presentation strategies among online dating participants, exploring how participants manage their online presentation of self in order to accomplish the goal of finding a romantic partner. They found that some of the technical and social aspects of online dating may discourage deceptive communication.

A former research (Hancock et al, 2004) shows that design features of a medium may

affect lying behaviors, and that the use of recorded media will discourage lying.

Sanver's work has enriched the Alvin E. Roth's classical work (Roth & Erev, 1995) about game theoretic analysis of two-sided matching with the assumptions of misrepresentation of preferences (Sanver and Sanver, 2005).

Another area of interest beyond theoretical models is about the characteristics of matching couples. Much empirical evidence shows that female and male partners look alike along a variety of attributes. Couples from same socio-demographical status tend to get together and even their physical attributes and habits show a resemblance (Belot and Francesconi, 2007).

Another concept about the subject is "speed dating", in which, a small group of people meet for matching in a bar café under the control of a moderator. If a match occurs, the moderator gives out the contact information to both parties. This popular game creates a perfect environment for researchers to conduct real life experiments about the dynamics of matching preferences. Many results are acquired from this experiment which reveals information about the importance of first impressions, subconscious preferences, age and height preference in matching (Fisman et al., 2006).

The paper entitled "What Makes You Click: An Empirical Analysis of Online Dating" is about an empirical study conducted in one of the major online dating services of the USA (Hitsch et al, 2005). The study brings a different perspective by using regression analysis to examine the internal dynamics of the online dating. Deriving a popularity index from the number of messages a member receives in a given time period, researchers tried to find out which features of members are important to attract others. By using Gale-Shapley algorithm (Gale and Shapley, 1962) to predict the equilibrium sorting along attributes such as age, income and education, they estimate the most significant factors for forming matches.

A data mining application getting popular recently is making use of recommender systems in e-commerce and social networking sites. Recommender systems provide advice to users about items they might wish to purchase or examine. Recommendations made by such systems can help users navigate through large information spaces of product descriptions, news articles or other items. As on-line information and e-commerce burgeon, recommender systems are an increasingly important tool (Burke, 2000; Jannach et al, 2010; Ricci et al, 2010).

Association Rule Mining

Knowledge Discovery in Databases (KDD) is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases (Han and Kamber, 2006). The phrase *data mining* is used interchangeably with *Knowledge Discovery in Databases,* although data mining is treated as simply a significant step in the process of KDD.

As being a descriptive data mining functionality, *association rule mining* is the discovery of rules for interesting relationships between variables in large transactional databases. It is widely used for market basket analysis to uncover the items that are purchased together. Beyond the market basket analysis, associative mining is used in many applications including *Web mining* and *intrusion detection systems.*

Analyses of the huge transaction database may reveal repeating patterns or rules in the form of $X_1, X_2, \ldots, X_n \rightarrow Y_1, Y_2, \ldots, Y_n$ where X and Y represent the two sides of the rule. X and Y are called the antecedent (left-hand-side or LHS) and the consequent (right-hand-side or RHS) of the rule (Agrawal et al., 1993).

To discover interesting rules efficiently, three main sets of constraints on significance and interest of the rules are defined. These are *support*, *confidence* and *lift* (Han & Kamber, 2006).

According to the original definition of Agrawal, support ($supp(X,Y)$) of a rule is defined as the proportion of transactions including item sets ($X \cup Y$) over all transactions, so

$$supp(X,Y) = p(X \cup Y)$$

In addition to the mentioned "rule support", many modern statistics softwares work with a slightly different definition of support called the "antecedent support". It is defined as the proportion of transactions including item sets (X) over all transactions, so $antecedantsupp(X,Y) = p(X)$
The second measure, the confidence ($conf\ (X,Y)$) is defined as the following:

$$conf\ (X,Y) = \frac{supp(X \cup Y)}{supp\ (X)}$$

Confidence is the proportion of transactions that fulfill the rule completely over the ones having only the left-hand-side of the rule true.

The last measure to be mentioned here is the Lift. Lift measures how much the observed confidence of the rule deviates from the expected confidence.

The lift $lift\ (X,Y)$ of a rule is defined as

$$lift\ (X,Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

Algorithms searching association rules are mostly based on the minimum support and minimum confidence criteria set by the user.

Apriori and *FP-Growth* (Han et al, 2000) are two of the most common associative rule mining algorithms based on the simple definition of Association Rule Mining. Another algorithm implemented by data mining software packages is the *Generalized Rule Induction* (*GRI)* algorithm which offers detailed customization of the model and alternative methods to increase the interestingness level of the rules.

The GRI algorithm is based on the ITRule algorithm and extends it with added functionality (Smyth and Goodman, 1992). It uses a quantitative measure called J to calculate how interesting a rule may be and uses bounds on the possible values this measure may take to constrain the rule search space. Briefly, it maximizes the simplicity/goodness-of-fit trade-off by utilizing information theoretically based on cross-entropy calculation.

The advantage of the GRI algorithm is being able to favor "interesting" rules and eliminate the "uninteresting" ones. This behavior solves the crucial problem of all association algorithms: dealing with huge number of resulting rules most of which are trivial. The GRI algorithm outputs rules in lower quantity and higher quality. The resulting rule set is expected to include significant, meaningful and non-trivial rules (SPSS Inc). These qualities of the rule set will achieve greater importance in the evaluation and deployment process (Aggelis and Christodoulakis, 2004).

Development of the Recommendation Engine

A typical scenario in online dating sites consists of the following steps: registration, filling out the profile, searching and text messaging. During the searching part, which is the third step of this process, the user is expected to inform the system manually about the features he/she is looking for in a potential partner. As a result of numerous queries – including a wide range of basic parameters such as the age and location of the potential partner and many details such as his/her profession and physical qualities – the user is presented with a list of potential matches.

This process requires the user to list his/her expectations precisely, although it does not provide the user with a "support system" to help the user while making his/her preferences.

In fact, the data on siberalem concerning all the required features queried in men and women (that is, the answer to the question "*Which features increase the possibility of a date both for men and women?*") may be used in the construction of a recommendation engine, which will facilitate the searching process of the users. Considering the profile data of the users who were matched through siberalem throughout many years, a model may be built including a rule set concerning which profile types are matched with whichever other profiles. When this model is operated by taking the profile data (of the user who is new to this system) as input, the user may then be presented to a series of potential matches suitable for his/her profile. This result set, prepared without registering the user's own preferences about the possible match, takes its origins from the past experiences of all the matches who met on siberalem.com.

This study also aims at building a recommendation engine specified for the field of online dating. In contrast to the voting/scoring systems that form the basis for the well-known web recommender systems, this model will not make the users rate each other and will be built on the basis of certain rules and patterns (found by scanning the features of already matched users). This model will form the basis for this specified recommendation engine.

Finding Matches

The databases and log files of online dating services keep the record of all user actions including registration, profile updates, recurring visits and messaging. Analysis of the available records can allow us to visualize the inner dynamics of the system. However, a problem arises for identifying a match since observing user interactions outside the system is nearly impossible.

Although it is not possible to pursue what exactly takes place between two people who were matched by siberalem after they meet in real life, an estimator can be used to reveal successful contact in siberalem by using the available data within the system. The proposed method suggests the consideration of user messaging logs in order to find any trace of a successful contact which is probably the starting point of a relationship.

The prerequisite of a successful contact is a two-way communication in the internal messaging system: the user has to send messages and receive ones in response, with the purpose of starting a relationship. Therefore, the elimination of incomplete conversations and determination of real communication throughout the messaging system is the first step of revealing the successful matches.

The conversations including an initial message and a response are likely to continue for 3, or 4 reciprocal messages. After that, another medium of communication is usually preferred by users. Before exchanging phone numbers (which is perceived as an insecure practice), an intermediary step in communication is exchanging instant messenger (IM) addresses in the form of ID, alias, nickname or e-mail address. For any conversation in the internal messaging system of siberalem, the point where IM addresses are exchanged can be accepted as an ending point for the conversation on the mentioned site and a valid starting point for a possible relationship. This indicates that the conversation started in siberalem is transferred to a common IM which provides a more comfortable environment for both sides.

According to the given information about communication habits of site members, a match is defined as follows:

Match: Any conversation which fulfills the following conditions:

1) The minimal number of both sent and received messages must be greater than or equal to 2.

2) In one of the messages other than the initial message, IM messenger IDs or e-mail addresses must be exchanged.

After processing siberalem database according to our match definition, more than 5 million messages have been reduced to 37476 matches. Features such as demographical, physical, socio-cultural attributes and preferences of both male and female party of every conversation are collected to provide the complete input data for the building of the model.

Features are transformed to categorical variables in the form of multiple binary variables to make them available for the association rule mining. A list of features and their corresponding binary variable examples are given in Table 1.

Selection of the Algorithm

Applications of Apriori and GRI algorithms are tested for a preview to see how different association algorithms impact the results.

Using the same sample set of data, for the same levels of minimum confidence and support, Apriori algorithm produced more trivial rules than the GRI algorithm, which decreased the total quality of the rules as shown in Table 2.

Table 1. Feature List and Their Corresponding Variables

Field name	Possible Values	Variable Examples
Age	Age values are discretized into 5 age groups such as *18-22, 23-27,…, 45+*	Age_Group_*18_22*, Age_Group_*45+*
Location	Five major cities of Turkey, namely, *İstanbul, Ankara, İzmir, Bursa,* and *Eskişehir* are considered separately. People from foreign countries and other small cities are grouped under the titles *Abraod* and *Small Town* respectively	Location_*İstanbul*, Location_*Ankara*, Location_*Abroad* Location_*Small_Town*
Income	Income values are discretized into 5 groups	Income_*1500_2250*
Marital status	Four distinct values are used; these are *single, married, married but separated,* and *divorced.*	MaritalStatus_*Single*, MaritalStatus_*Divorced*
Education	*Primary School, High School, Junior College, Undergraduate, Graduate*	Education_*UnderGraduate*
Smoking / Drinking habits	Smoking values are reclassified into *yes* and *no*. Drinking habit values are *drinker, not drinker* and *social drinker*.	Smoking_*yes*, Drinking_*Social Drinker*
Importance of religion in life	*Very Important, Important, May be Important, Very Little, Not Important*	Religion_*not important*
Having Children	• *I have children, not staying by me* • *I have children, staying by me,* • *No Children*	HasChildern_*i_have children*
Wanting to Have Children	• *I don't know* • *I don't want to have children* • *I want to have children*	WantsChildren_*I don't know*
Eye & Hair Color	Eye Color values are *Hazel, Green, Brown,* and *Blue*. Hair color values are *Red, Gray, Brown, Blond,* and *Black*	EyeColor_*Green* HairColor_*brown*
Weight / Height	Body Mass Index (BMI) is used with values such as, *Underweight, Normal, Overweight* and *Obese*.	BMI_Status_*Normal*
Preferred Relationship	Possible values for relationship are *Friendship, Long-Term, Short-Term,* and *Marriage*	PrefRels_*Friendship*

Table 2. Comparison of Apriori and GRI Algorithms

	Apriori	GRI
Min. Support	2.0 %	2.0 %
Min. Confidence	10.0 %	10.0 %
Discovered number of rules	396424	930

The difference in the numbers of the discovered rules of the two algorithms originates from the facts that similar rules found by the Apriori algorithm are repeated many times in the rule set. The GRI algorithm, on the other hand, output fewer, higher quality and interesting rules by omitting the similar rules which create no information gain. Hence, by considering the test results, the GRI algorithm is selected for training the model.

The GRI algorithm requires that the input variables are given separately as "antecedents" and "consequents". Taking into account the fact that the number of rules grows much larger and complex rules contribute very little to the insights about the data (Borgelt and Kruse, 2002), the discovery of the rules with multiple consequents will also be omitted in the present study.

Considering the two gender types (male and female) in the dataset, two specialized association rule forms will be derived from the general one. The first one shows the existence of which "male features" lead to a potential match with the given "female feature": $Y_1, \ldots, Y_n \rightarrow X_i$. The reverse case shows the inclusion of which "female features" lead to a potential match with the given "male feature": $X_1, \ldots, X_n \rightarrow Y_i$, where X stands for the set of female features and Y for the set of male features.

Obviously, to obtain the rules in both forms, the model should be run twice. In the first run, the "male features" will be given as the antecedents and the "female features" will be the consequents. In the second run, the "female features" will be given as the consequents and the "male features" will be the antecedents.

In addition to antecedent and consequent input variable selections, further settings are available in many of Association Algorithm implementations. These settings mostly cover the minimal cut-off points for the major evaluation criteria of the output rules like support, confidence and lift. The minimum support and confidence level are selected as 2.0 % and 10.0 % respectively. As to the minimum lift value, it is set to 1.0 for the training of the model.

The training of the model by the GRI algorithm for both genders produced the following summary statistics shown in Table 3.

Table 3. Output Summary of the Model

	Rules for Women	Rules for Men
Discovered number of rules	1239	930
Maximum Support	65,37	61,57
Maximum Confidence	92,19	92,61
Maximum Lift	14,10	14,10
Avarage Support	8,72	6,62
Avarege Confidence	33,08	27,68
Avarege Lift	1,44	1,54

Table 4 and Table 5 present example rules for both men and women. The rules are sorted according to Lift parameter in descending order.

Table 4. Example Rules for Men

Antecedent	Consequent	Supp.	Conf.	Lift
Location_Bursa_M	Location_Bursa_W	4.12	58.38	14.11
Location_Abroad_M	Location_Abroad_W	5.89	31.3	5.52
Education_UnderGraduate_M and MaritalStatus_Single_M and Location_Izmir_M	Location_Izmir_W	3.12	61.06	5.40
Education_UnderGraduate_M and Location_Izmir_M	Location_Izmir_W	4.97	60.27	5.33
Education_UnderGraduate_M and Age_Group_45+_M	Age_Group_45+_W	5.24	48.34	5.23
Location_Izmir_M	Location_Izmir_W	10.04	58.81	5.20
Education_UnderGraduate_M and MaritalStatus_Divorced_M and Age_Group_45+_M	Age_Group_45+_W	3.61	48	5.20
MaritalStatus_Single_M and Location_Izmir_M	Location_Izmir_W	6.05	58.33	5.16
MaritalStatus_Divorced_M and Age_Group_45+_M	Age_Group_45+_W	8.97	47.49	5.14
MaritalStatus_Divorced_M and Location_Istanbul_M and Age_Group_45+_M	Age_Group_45+_W	4.06	46.9	5.08
Location_Istanbul_M and Age_Group_45+_M	Age_Group_45+_W	5.93	45.51	4.93
Education_Graduate_M and Location_Ankara_M	Location_Ankara_W	3.65	56.63	4.87
Age_Group_45+_M	Age_Group_45+_W	14.37	44.42	4.81
MaritalStatus_Divorced_M and Location_Ankara_M	Location_Ankara_W	3.19	55.7	4.79
Education_UnderGraduate_M and Location_Ankara_M	Location_Ankara_W	6.58	55	4.73
Location_Ankara_M	Location_Ankara_W	13.58	54.16	4.66
MaritalStatus_Single_M and Location_Ankara_M	Location_Ankara_W	8.52	53.28	4.58

Table 5. Example Rules for Women

Antecedent	Consequent	Supp.	Conf.	Lift
Location_Bursa_W	Location_Bursa_M	4.14	58.14	14.11
Education_High School_W and MaritalStatus_Single_W and Age_Group_18-22_W	Age_Group_18-22_M	3.3	16.54	7.45
Education_High School_W and Age_Group_18-22_W	Age_Group_18-22_M	3.38	16.28	7.33
MaritalStatus_Single_W and Age_Group_18-22_W	Age_Group_18-22_M	12.51	14.84	6.69
Age_Group_18-22_W	Age_Group_18-22_M	12.88	14.73	6.64
MaritalStatus_Single_W and Location_Small Town_W and Age_Group_18-22_W	Age_Group_18-22_M	5.91	13.65	6.15
Location_Small Town_W and Age_Group_18-22_W	Age_Group_18-22_M	6.08	13.61	6.13
Education_UnderGraduate_W and Location_Izmir_W	Location_Izmir_M	3.5	58.09	5.79
MaritalStatus_Single_W and Location_Izmir_W	Location_Izmir_M	6.68	55.38	5.52
Location_Abroad_W	Location_Abroad_M	5.68	32.48	5.51
Location_Izmir_W	Location_Izmir_M	11.3	52.23	5.20
MaritalStatus_Divorced_W and Location_Izmir_W	Location_Izmir_M	3.27	51.14	5.10
MaritalStatus_Divorced_W and Age_Group_45+_W	Age_Group_45+_M	6.65	71.1	4.95
Location_Istanbul_W and Age_Group_45+_W	Age_Group_45+_M	4.18	70.06	4.88
Age_Group_45+_W	Age_Group_45+_M	9.24	69.08	4.81
MaritalStatus_Single_W and Location_Ankara_W	Location_Ankara_M	7.62	64.79	4.77
Education_UnderGraduate_W and Location_Ankara_W	Location_Ankara_M	3.99	64.08	4.72
Location_Ankara_W	Location_Ankara_M	11.62	63.26	4.66

The Matching Engine

The second phase of the study is to develop a matching engine which utilizes the discovered rules to serve as a recommendation engine for the site members.

The aim of the engine is to recommend potential partners having higher matching probability with the member. In contrast to conventional matching systems, the member does not have to reveal his/her preferences about the opposite gender. The member's own features are sufficient to get results from the engine.

To start generating recommendations, the engine needs to be fed with profile data of members actually using the system. In the production environment, the profile data of 150000 "active members" will serve as the source of this input data.

To get the best matches for a user in the system, a two step process is used. First, the system finds all the rules for which the antecedent conditions are completely satisfied by the user. These rules form a group of features which is a subset of the user's own features. In the second step, the system iterates through the members of the opposite gender to score them according to these rules satisfied for the user.

To normalize the scores of members satisfying a multiple number of rules, the score based only on its confidence is divided by the number of rules for each member. Sorting of the aggregated scores of the members determines the best matches for the user.

The following algorithm shown in Figure 1 is used to return the best matches for a user in the form of a list. The following pseudo code shows how the program outputs a list of female users with their calculated match scores for a given single male user:

- R_m is the list of all rules containing rules in the form Y->X. Every rule r has antecedants, a consequent, support, confidence and lift.
- R_{temp} is the empty list to be filled with rules rt which of their antecedent conditions are completely satisfied by the user.
- W is the list of all female members loaded into the matching engine. Every member w has a MemberID and features.
- M is the list of all male members loaded into the matching engine. Every member has a MemberID and features.
- P is the UserID entered to the User Interface of the program

```
for each r ∈ R_m{
    if  (M_p.features ⊂ r.antecedants){
        add r to R_temp;
    }
    for each w ∈ W{
        score=0;
        for each rt ∈ R_temp{
            if (rt.consequent ⊂ w.features){
                score+=rt.confidence;
            }
        }
    print w.UserID, score/R_temp.length
}
```

Figure 1. Pseudo Algorithm of Matching Engine

The matching algorithm is run for a selected random user; as if the user visits the production site and requests a query to get his/her matches. The test is repeated for both random male and female users. The user interface to control the matching program is shown in Figure 2.

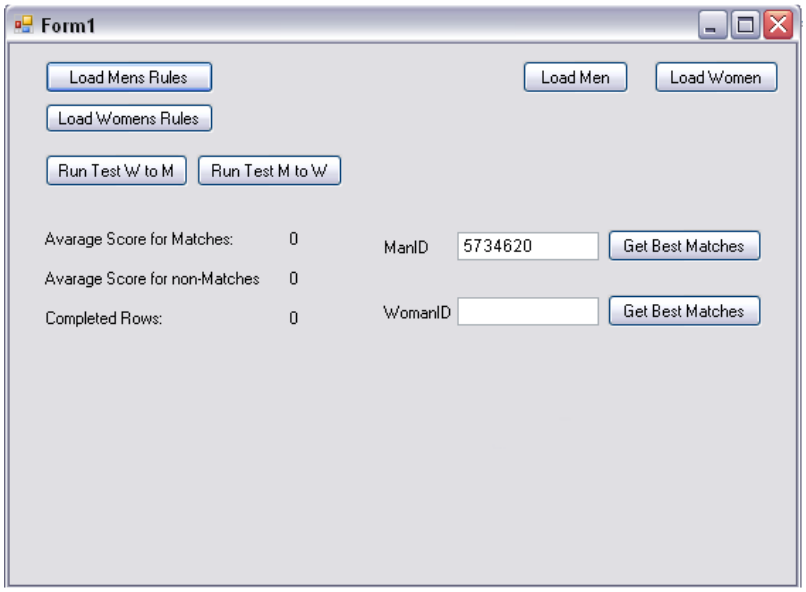

Figure 2. User Interface of the Matching Program

The selected male user with UserID 5734620 has the following features shown in Figure 3.

MaritalStatus_Divorced, Location_Istanbul, Income_1500-2250, Age_Group_33-37, BMI_Status_Normal, PrefRel_LongTerm, PrefRel_ShortTerm, PrefRels_Friendship, HairColor_brown, EyeColor_Green

Figure 3. Features of the Test User

By running the match engine to get the best matches of the user, we obtain top ten best matches as seen in Table 6.

Table 6. Top 10 Best Matches for User 5734620

Rank	UserID	Score
1	7735529	24.5182
2	4521193	24.4041
3	10821729	24.3015
4	6788407	24.3013
5	8474802	24.1162
6	3274864	24.1059
7	9775325	23.8650
8	8878239	23.8626
9	10745655	23.8626
10	8449754	23.8626

The best matches for the test user with UserID 7735529 has the following features shown in Figure 4.

Education_UnderGraduate, MaritalStatus_Married but separate, Location_Istanbul, Income_1500-2250, Age_Group_33-37, BMI_Status_Normal, WantsChildren_I_don't_know, HasChildren_I have children, staying by me, Drinking_Social Drinker, Religion_not important, Smoking_yes, PrefRel_Marriage, PrefRel_LongTerm, PrefRel_ShortTerm, PrefRels_Friendship, HairColor_brown, EyeColor_Hazel

Figure 4. Features of the Best Match of the Test User

A subjective comparison of the two members concludes that the selected male and female users are similar in socio-demographical features and there is no serious incapability observed to prevent a potential relationship.

To test if the higher scored members are more likely to be matched, another query is run for user 7735529. There are 136 records in the "couples" table for the user 7735529, which means that she has communicated with 136 different members. According to "couples" table, 128 of these conversations are "non-matches" and 8 of them are flagged as "matches".

The matching engine calculates the matching scores for these two groups separately and obtain the following results shown in Table 7.

Table 7. The Comparison of the Average Scores of Matches and Non-Matches

	Number of Partners	Avarage Matching Score
Non-Matches	128	8,66
Matches	8	12,37

The average score for the matches is significantly greater than the score for non-matches. For user 7735529, the matching engine produces higher scores for the partners who are real matches of the user 7735529.

Testing of the Matching Engine

The scoring algorithm calculates matching scores for a selected sample of couples whose matching status are already known. The expectation is to observe higher matching scores from the algorithm for the already matched couples in comparison to the scores of couples labeled as non-matches. The test is performed on a test sample which is different from the dataset used for building the model. The sampling period of this test dataset ranges from January 2009 to April 2009. First, matching scores for the 8000 couples in the sample dataset are computed. 4000 couples labeled as "matches" and another 4000 labeled as "non-matches" are randomly selected.

As presented in the previous section, the scoring algorithm produces different scores for the "*Male user looking for female user*" by using male user's rules and female user's features and for the "*Female user looking for male user*" by using female user's rules and male user's features. So, for any given couple two different scores are generated.

Hence, two consecutive tests are performed, one for calculating Man → Woman and one for Woman → Man scores respectively.

For both types of rules, the means of scores for different groups of couples (matching and non-matching) are hypothesized to be different. The mean of matching scores are expected to be greater than that of the non-matching ones. Student's t-test is used to compare the means of the independent samples of matching and non-matching groups. Table 8 and Table 9 show the descriptive statistics of the Man → Woman and the Woman → Man scores for the sample dataset.

Table 8. Statistics for the Man → Woman Scores

	Match Status	N	Mean	Std. Deviation	Std. Error Mean
Score	Non-Match	4000	9.24	3.8916294	.0615321
	Match	4000	11,92	4.3950297	.0694915

Table 9. Statistics for the Woman → Man Scores

	Match Status	N	Mean	Std. Deviation	Std. Error Mean
Score	Non-Match	4000	19,62	8.4867193	.1341868
	Match	4000	20,56	9.0009027	.1423168

The null hypothesis is stated as follows: $H_0 = \mu_{match} = \mu_{non-match}$

For the Man → Woman scores, the t-statistic under the assumption of equal variances is -28,811 (df: 7998, p-value: 0.00). The null hypothesis of equal average scores is rejected at a 1% significance level. It can be concluded that the mean of the score for the matching couples is significantly greater than that of the means of the non-matching couples for the Man → Woman scores.

Similarly, for the Woman → Man scores, the t-statistic under the assumption of equal variances is -4,711 (df: 7998, p-value: 0.00), the null hypothesis of equal average scores is rejected at a 1% significance level. It can be concluded that the mean of the score for the matching couples is significantly greater than that of the means of the non-matching couples for the Woman → Man scores.

As the difference between the means of the scores of matching and non-matching couples are found to be significant, it can be concluded that the recommendation engine produces higher scores for "matching" couples. From the site user's perspective, this result can also be interpreted as follows: If a user follows the suggestions of the recommendation engine and sends messages to those users on the recommendation list, the chance of a potential match increases compared to a match from a message sent by his (her) own judgments.

Conclusion

In this study, the researchers are able to develop a new type of recommendation engine for online dating sites, giving lists of men or women for the site users with higher potential of starting a relationship. The rules used in the engine are extracted from past data kept on the site's database and no additional expert knowledge is used. The overall performance of the engine is tested for statistical significance and it is found that it may create a real benefit for the users of the site. In this sense, the study is an example of making use of collaborative information (in this case the past experience of siberalem users) available for developing solutions for individuals' needs.

Certain assumptions were made about the messaging routines and message contents of online dating site users to define a notion of "matches". This definition helped to differentiate between successful and unsuccessful couples and build a model by using the members called "matches" who were able to start a relationship according to our definition. This classification methodology of "matching" and "non-matching" couples is the first finding of the study.

Rules are extracted from data by using the GRI algorithm of Association Rule Mining functionality of data mining. The developing of the recommendation engine turned the list of rules to a functional support system for the site members. For this development, a scoring system is proposed by making use of the "confidence" attribute of the rules.

The implementation of our recommendation engine to a working online dating site environment is possible, by solving a list of issues. By experimenting with different input features, the success rate of the recommendation engine may be increased to afford even more benefit for site users. The recommendation engine should also be expanded to provide a suggestion list for users of the same kind of gender. These issues can also be handled in further research.

References

Aggelis, V. & Christodoulakis, D. (2004). "e-Trans Association Rules for e-Banking Transactions," *IV. International Conference on Decision Support for Telecommunications and Information Warsaw*, Poland.

Agrawal, R., Imielinski, T. & Swami, A. (1993). "Mining Association Rules between Sets of Items in Large Databases," *SIGMOD Conference*, 207-216.

Belot, M. & Francesconi, M. (2007). Can Anyone Be "The" One? Field Evidence on Dating, The Institute for Social and Economic Research, University of Essex, Working Paper no 17.

Borgelt, C. & Kruse, R. (2002). Induction of Association Rules: Apriori Implementation, 15th Conference on Computational Statistics. Berlin.

Burke, R. (2000). "Knowledge-based Recommender Systems," In Encyclopedia of Library and Information Science, 1-23.

Ellison, N., Heino, R. & Gibbs, J. (2006). "Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment," *Journal of Computer-Mediated Communication* , 11(2), article 2.

Fisman, R., Iyengar, S. S., Kamenica, E. & Simonson, I. (2006). "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," *The Quarterly Journal of Economics*, 673-697.

Gale, D. & Shapley, L. S. (1962). "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69(1), 9-15.

Hancock, T. J., Thom-Santelli, J. & Ritchie, T. (2004). "Deception and Design: The Impact of Communication Technology on Lying Behavior," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, 129-134.

Han, J. & Kamber, M. (2006). "Data Mining: Concepts and Techniques 2nd Ed," Academic Press.

Han, J., Pei, J. & Yin, Y. (2000). "Mining Frequent Patterns without Candidate Generation," ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00), Dallas, 1-12.

Hitsch, G. J., Hortaçsu, A. & Ariely, D. (2005). "What Makes You Click: An Empirical Analysis of Online Dating," 2005 Meeting Papers, University of Chicago.

Jannach, D., Zanker, M., Felfernig, A. & Friedrich, G. (2010). Recommender Systems: An Introduction*, Cambridge University Press*.

Kantor, P. B., Ricci, F., Rokach,L. & Shapira, B. (2010). "Recommender Systems Handbook," *Springer.*

Özkal-Sanver, İ. & Remzi Sanver, M. (2005)."Implementing Matching Rules by Type Pretension Mechanisms," *Mathematical Social Science*, 50(3), 304-317.

Romm-Livermore, C. & Setzekorn, K. (2009). Social Networking Communities and E-Dating Services, *IGI Global*, London, UK.

Roth, A. E. & Erev, I. (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Journal of Economic Literature* , 164-212.

Smyth, P. & Goodman, R. M. (1992). "An Information Theoretic Approach to Rule Induction from Databases," *IEEE Transactions on Knowledge and Data Engineering*, 301-312.

SPSS Inc. (2009). Clementine 12.0 User Manual.