AN ASSOCIATION RULE-BASED RECOMMENDATION ENGINE FOR ONLINE DATING SITES

YAĞIZ CİVAN ÖZSEYHAN

BOĞAZİÇİ UNIVERSITY

2009

AN ASSOCIATION RULE-BASED RECOMMENDATION ENGINE FOR ONLINE DATING SITES

Thesis submitted to the
Institute for Graduate Studies in the Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Management Information Systems

by
Yağız Civan Özseyhan

Boğaziçi University
2009

Approval Page

Thesis Abstract

Yağız Civan Özseyhan, “An Association Rule-Based Recommendation Engine for Online Dating Sites”

In this study, the database of a Turkish online dating site is analyzed to reveal patterns in the personal features of matching couples.

By using a functionality of data mining, the Association Rule mining, a set of rules is extracted from the available messaging and user data. The rules are used for the development of a recommendation engine which is capable to give list of best potential matches to the site user.

The performance of the recommendation engine is tested with statistical tools to find whether the increase in the quality of results is significant.

Tez Özeti

Yağız Civan Özseyhan, "Arkadaşlık Servisleri İçin Veri Madenciliği Temelli Öneri Sistemi"

Bu çalışmada Türkiye'den bir arkadaşlık sitesinin veritabanını incelenerek, site üzerinden tanışmış çiftlerin sahip olduğu ortak özellikler araştırılmıştır.

Veri Madenciliği araçlarından biri olan "Association Rule Mining" 'ten yararlanılarak, mesajlaşma ve kullanıcı verilerinden bir kural seti türetilmiş; bu kurallar da kullanıcının kendisine en uygun adayları listeleyebileceği bir öneri sisteminin geliştirilmesinde kullanılmıştır.

Öneri sisteminin üretilen sonuçların kalitesinde yaptığı artış istatistik araçlarıyla test edilmiş, verilerin anlamlı olduğu sonucuna varılmıştır.

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION

The rapid growth of the Internet penetration all over the world changed the perception of the computer networks: The Internet is not only a network of computers but it's also the biggest network of people ever created in the human history.

To explain specific features of the Internet community, it's common to refer to conventional societies by mentioning their own rules, dynamics and systems. Interestingly, the Internet – as being an established society – fulfills the needs of its community like many other conventional societies do.

The Internet has its own rules and regulation, creates trade and business activities, and offers services such as education and finance. From sending money to ordering food, the Internet provides online alternatives for the basic daily needs and habits of conventional society. Specialized services called social networking sites, aim to fulfill more complex needs of the Internet community such as communication and socialization. Blogging services, Circle-of-Friends networks, video / image sharing communities and dating sites are examples of social networking sites.

As being a popular type of social networks, the online dating sites provide services like real world's singles bars, providing a platform for people who want to meet others in a romantic context. There exist many global online dating services as successful businesses that have provided this type of services for many years. Also, online dating sites have been popular services in Turkey since its beginning in the year 2000. Finding a partner in a dating site, exchanging e-mail addresses or telephone numbers, meeting in real world and developing a serious relationship has become a common and socially accepted practice.

To study different aspects of this online dating phenomenon, we've worked on the database of siberalem.com, one of the prominent online dating sites of Turkey. The profile information of members and messaging logs made us available for a research.

The aim of the study was to develop a recommendation engine for the web site working as a support system for the site members, which is capable to output best matches of the user automatically. 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.com's messaging logs include certain information such as messaging frequency and message contents, which can be used for tracking couples developing a relationship via the messaging system and probably becoming "matches" outside the site. 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's possible to search for frequent patterns among the features of matching couples to provide an answer to question "who meets who?"

By using a less common functionality of data mining for the given field, the Association Rule Mining can be used to extract rules from the dataset. It's 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 chapter gives background information about social networks, online dating and introduces some basic functionality and methodology of data mining. A literature survey about the subject is also present. The third chapter is about the understanding and preparation of the data. In the fourth chapter, the building of the model and the extraction of the rules are explained, which are used for developing the recommendation engine. The chapter ends with a testing section in which the results of the recommendation engine are tested for statistical significance. The last chapter is the conclusion, where the findings are summarized and suggestions for further research are given.

CHAPTER II: BACKGROUND

This chapter aims to provide fundamental background information about two essential topics of the thesis: social networking and data mining.

After the description of contemporary social networking communities, a special focus is put on online dating: one of the popular social networking tools. Information about the business and operation of online global dating services are followed with a detailed look on the local online dating sites explaining their systems and features.

In the data mining part of the chapter, a brief introduction about the subject will be made by describing the main methodology and functionalities of data mining with a special emphasis on associative rule mining.

Social Networking Communities and Online Dating

Introduction

Since the year 2000, the statistics indicate a surprising boom in the population of the Internet users all over the world: a 336% growth in 8 years. (Internet World Stats, 2008). With this dramatic increase, the global Internet population has reached 1,574,000,000 members, the 26,500,000 of which reside in Turkey. This increase has placed the Internet among the largest communities of people the world has ever witnessed.

However, a question comes to mind with the reference of Internet users as a "social community": Do the Internet users actually embody the very features necessary to mention them under the category of a community or a society (Brym & Lenton, 2001)? The best way to answer this question would be to examine the definition and principal features of a conventional society:

A society is a large, enduring network of social interaction that survives by accomplishing five main tasks: (1) preserving order, (2) producing and distributing goods and services, (3) teaching new members, (4) providing its members with a sense of purpose, and (5) replacing old members (Aberle, 1950).

By creating an analogy between the listed tasks of Aberle and some features of the Internet, Brym and Lenton decide that the Internet society is not different from a conventional society (Brym & Lenton, 2001).

By considering the existing governing structures of the Internet like HTML code convention and domain name regulation; available e-commerce activities and distant education systems, Brym and Lenton state that the Internet society accomplishes many of the same tasks fulfilled by other societies as stated in Aberle's first three assumptions.

The fourth task in Aberle's classification, that is, "providing its members with a sense of purpose" requires more complex relations between the members of society, such as social interaction, communication and sharing. In addition to conventional Internet communication tools like e-mail and instant messaging, Internet users today interact socially by exchanging texts, images, sound and video via social networking communities, which, on the whole create many senses of purpose.

The fifth task of any enduring society is concerned with replacing old members. That is, people ensure the survival of their society by dating, forming long-term relationships, and reproducing. Therefore, it is a necessity for the members of the society to meet each other, form the appropriate matches and reproduce new members for the following generations. This is the point that makes it indispensable to refer to online matching and dating services. Online dating, as a subset of social networking communities, is the newly developing way of the Internet users to get together and fulfill this very task of the society.

Social networking communities are defined as online communities that focus on the building and verifying of social networks for whatever purpose (Romm & Setzekorn, 2009). Social networking communities offer a range of services. Some are merely blog hosting services, some offer the option of joining groups temporarily through chat rooms or for longer periods of time through electronic bulletin boards, newsgroups or online groups. Some social networking communities encourage their members' creativity through sharing of music and video clips, while others enable members to meet, develop relationships and possibly marry.

Many social networking communities are founded and run by Internet companies. The degree of involvement of the company to the activities going on in the virtual community varies from community to community. This property serves as a basis for the following classification of the social networking communities:

1. The Blogging Service: These services provide a space for bloggers to set up a presence on the Web site. Even though bloggers can join groups and can contact individuals that are member of the service, the "company" does not get involved in these interactions in any active way. Facebook (http://facebook.com), MySpace (http://myspace.com), Youtube (http://youtube.com) and Flickr (http://flickr.com) and even the virtual reality site Second Life (http://secondlife.com) fall into this category by creating a space for their users where they can discover their friends, share files, watch videos, listen music etc. by minimal involvement of the company.
2. The groups creating service: These types of services provide a platform on which users can create groups to join or invite other friends. The well-known

example of this category is Yahoo Groups (http://groups.yahoo.com). Members of a group can exchange messages, share files or schedule other activities. Despite the fact that the company provides the infrastructure for the group service, the involvement is again at the minimal level.

3. The support service: In some cases, companies intentionally establish virtual communities to create an added value around their products or services. The purpose ranges from customer service (http://dell.com) to active selling of products (http://amazon.com). In these examples, the company shows a definitely higher level of involvement by developing unique matching algorithms to create the group(s) and match the individuals with the pre-existing group.
4. The e-dating (online dating) service: These are examples of 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. Whereas the major service the company provides is matching the individuals, this model shows the highest level of company involvement.

The study mentions the criterion of the involvement of the company in terms of all online networking communication services (Romm & Setzekorn, 2009). By going one step further, we can also observe different levels of company involvement within the given categories.

For example, in many online dating sevices, the user is free to choose his/her partners throughout the site by searching or browsing the whole available profiles. (match.com, Yahoo! Personals). However, in some cases, the company providing the platform takes an active role in both the initiation and the development of matching, similar to the action of marriage agencies (e-harmony.com).

Alternative perspectives may also be offered while categorizing social networking communities, one of which may be defined as the categorization on the basis of the criterion of anonimity versus real life roles. We may state that the Internet user may either appear to the other users with a masked identity by using a nickname, in the form of anonimity such as match.com or prefer to reflect his/her real life character in Circle-of-Friends networks such as facebook.com. The former approach suits the needs of romantic relationships, whereas the latter extends the use of the service into various domains of daily life as you get news from your friends, make business contacts, share your status, photos and videos with the rest of the world.

As specialized social networking communities, online dating services provide web sites and other tools for the users who have particular demands and expectations from starting a relationship or various forms of relationships online. They provide a platform for this type of activities through the use of “personal ads”, which are posted by the users onto their personal pages with the purpose of finding their potential mates. Personal ads provide the users with the basic means of searching and interacting with the potential mates via messaging. Also known as “profiles”, personal ads help users submit information about them, including their age, gender, location, physical attributes, socio-cultural status and daily habits. This information is collected in database to enable other users to search the profile base according to their own preferences and contact the ones fitting the search criteria.

In addition to simple profile creation and search facilities, many online dating sites enhance their services with advanced features. 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 difference in advanced matching, having the matching algorithms as one of their essential features.

Online Dating Market

Online dating is a large scale business, allowing people from the comfort of their own home to view and read about potential mates all around the world (Bishop, 2009).

American citizens spent over $500 million for online dating services in the year 2005 (Online Publishers Association & comScore Network, 2005). The USA market consists of over 800 enterprises in different sizes; nonetheless, as a direct consequence of a huge networking effect, the market used to tend to create its monopolies. According to their traffic figures, total number of members, and yearly revenues; Yahoo! Personals (http://personals.yahoo.com), match.com (http://match.com) and e-harmony (http://e-harmony.com) have long been the top players that rule the online dating game (Hitwise Inc., 2004).

Nevertheless, recent reports (Hitwise Inc., 2008) show remarkable new entrances to the market. According to a report issued in 2008 respecting the traffic figures, Singlesnet.com (http://www.singlesnet.com) seems to be in the first position followed by PlentyofFish.com (http://www.plentyoffish.com). The success of these newcomer sites mainly depend on technological advances and business models developed by them.

In addition to mentioned mainstream online dating sites, there also exist "niche" businesses which target a specific group of people according to their ethnicity, religious and political views. Specifically oriented sites provide online communication amongst Jewish peoples (http://JDate.com), Christians (http://christianmingle.com) or senior citizens (http://SeniorPeopleMeet.com).

E-bussiness Models and Online Dating

Different business models dominating the Internet domain can be summarized in seven groups: Brokerage, Advertising, Infomediary, Merchant, Manufacturer (Direct), Affiliate, Community, Subscription and Utility (Rappa, 2001). Online services adopt one or more of these business models to generate income.

While the subscription model has always been the pioneering model in the business of online dating, some other companies have also been extremely successful trying out the advertising model in recent years. The two out of Rappa's 7 classified models, the subscription and the advertising models are explained in detail:

1) Models based on subscription: Models based on subscription offer the users a restricted version of the site in order to provide a sound preview of their members to get to know the site better. In this limited version, creating profiles and searching are actually available, whereas sending messages to other members is not applicable. Nevertheless, when the users purchase a premium membership subscription, which, in general, is available in the monthly, quarter or annual form, they acquire unrestricted messaging with other members.

2) Free sites relying on advertisement revenue: These sites are generally directed by either a single individual manager or small companies, and serve their users basic functions of the above-mentioned services. In these services, all the actions of the users are free of any charge including the messaging services, and the financial source of these services is usually based on targeted online advertisement.

   Online dating sites relying on this type of business model typically prefer big online advertisement agencies such as Google's AdSense (http://ads.google.com) to convert the generated traffic into revenue (http://www.plentyoffish.com).

Local Online Dating Market in Turkey

Even though the global attract millions of members around the world, in particular cases they cannot surpass specific solutions and forms possessed by local sites that address to local clients. This fact is an evidence that cultural differences and the factor of localization play a very significant role in this sector.

In year 2000, online dating was first launched in Turkey. During the following decade when the broadband Internet has boomed in the country, online dating turned into a leading industry in the Internet sector.

Unfortunately, the pre-mature market lacks market data and share estimates. A rough estimation from the available sample data shows that, approximately 300000 Turkish people visit any of local dating sites at least once in a month (The number is a calculated estimate based on the 2007 traffic figures of siberalem.com). Whereas many enterprises in different sizes exist, siberalem.com, gayet.net and istanbul.net are the major players in the market.

The foundation of siberalem.com in the year 2000 is followed by the emergence of the other brands in year 2005. By taking the advantage of being the first comer, siberalem.com is the market leader by the means of brand recognition, traffic estimates and sales figures (A.C. Nielsen, 2007).

After their late entrance to the market in 2005, companies like Gayet.net and Istanbul.net have used aggressive advertisement campaigns to establish a member database and to take part in the competition. Gayet.net built its concept around taking quizzes for fun to increase its popularity. Istanbul.net, on the other hand, have emphasized the importance of locality by launching specialized e-dating sites such as izmir.net, ankara.net, adana.net for the major big cities of Turkey.

Case Study: siberalem.com

One of Turkey's most prominent online dating site siberalem.com was founded in 2000, as a brand of the Internet Company, EBI A.S. Having started as a free service for all users online, the site switched to subscription-based membership in 2002. After serious efforts for further development, in 2005, it became the online dating partner of the MSN network. The latest renewal of siberalem.com was in 2007, when the software infrastructure was modified and reinforced.

Statistics

At the end of 2008, statistics indicate that there are approximately 200000 registered profiles in the database. The number of distinct members who log in in a month is around 150000 as a yearly average. This group is users are considered as "active users". When a three-month interval is examined in this respect, the statistics reveal approximately 400000 distinct users login to the site at least once.

One interesting fact is that the site appeals around 10000 new members per day. The total number of messages sent on the site per day is around 180000.

Mechanism

As shown in Figure 1, the simple scenario in siberalem.com 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.

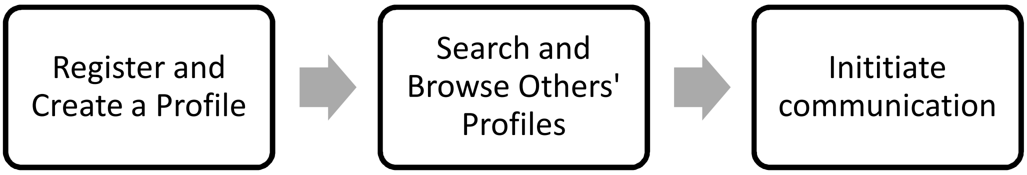

Register and Create a Profile
Search and Browse Others' Profiles
Inititiate communication

Figure 1. Simple communication scenario in siberalem.com

A new comer to the site initially creates a membership to the site by filling in a registration form. After the registration form is submitted, the user is redirected to pages where he/she creates his/her detailed profile. This process also includes the step in which the user uploads his/her profile photo. After logging-in, the user comes face to face with a detailed control panel shown if Figure 2. By using it, the user can search other members, get in touch with them (by sending messages, chat requests, ice-breakers, gifts, and questionnaires), manage account settings and update profile information.

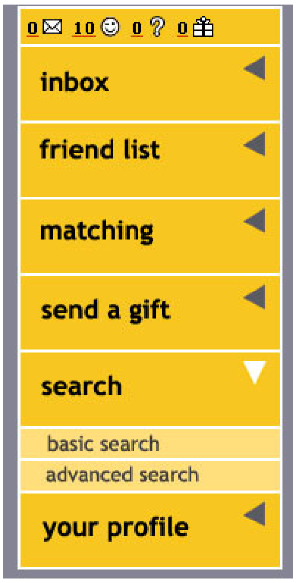
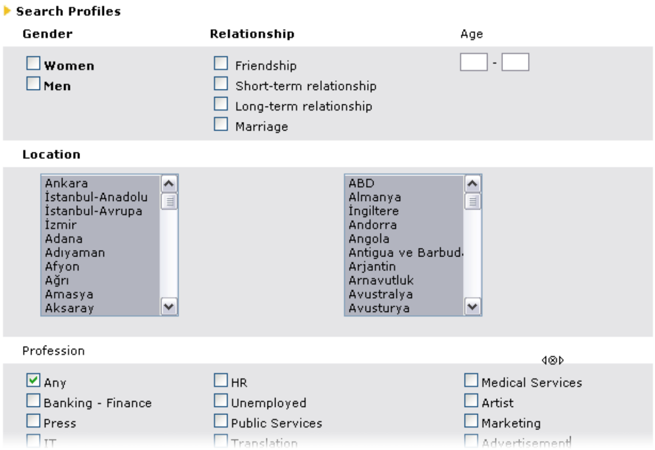


Figure 2. Control panel of siberalem.com

By using either simple or detailed search forms, the user starts searching other members' according to the criteria he/she expects from his/her potential match. The basic search form includes the age and location fields; whereas the advanced search form shown in Figure 3 includes many fields from physical attributes to religious and political views.



Search Profiles

Gender

□ Women
□ Men

Relationship

□ Friendship
□ Short-term relationship
□ Long-term relationship
□ Marriage

Age

□ - □

Location

Ankara
İstanbul-Anadolu
İstanbul-Avrupa
İzmir
Adana
Adıyaman
Afyon
Ağrı
Amasya
Aksaray

ABD
Almanya
Ingiltere
Andorra
Angola
Antigua ve Barbuda
Arjantin
Arnavutluk
Avustralya
Avusturya

Profession

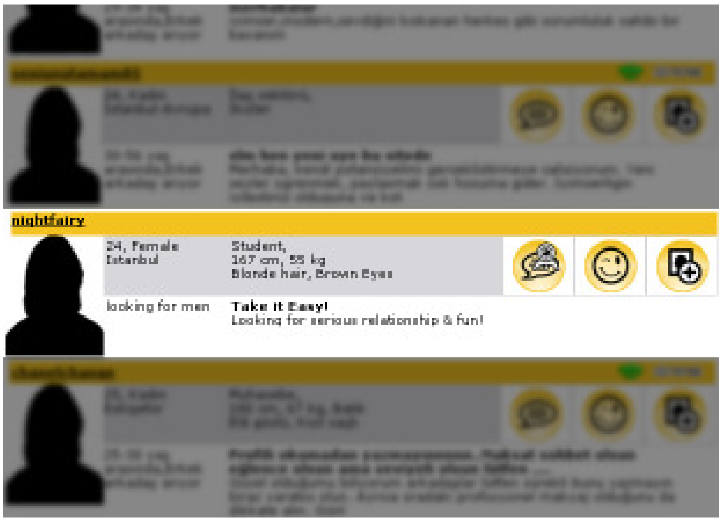
☑ Any
□ Banking - Finance
□ Press
□ IT

□ HR
□ Unemployed
□ Public Services
□ Translation

□ Medical Services
□ Artist
□ Marketing
□ Advertisement

Figure 3. Detailed search form

The search results as shown in Figure 4 are ordered according to the date of last activity of the owners, so online users which are currently using the service are listed on the top.



nightfairy

24, Female
Istanbul

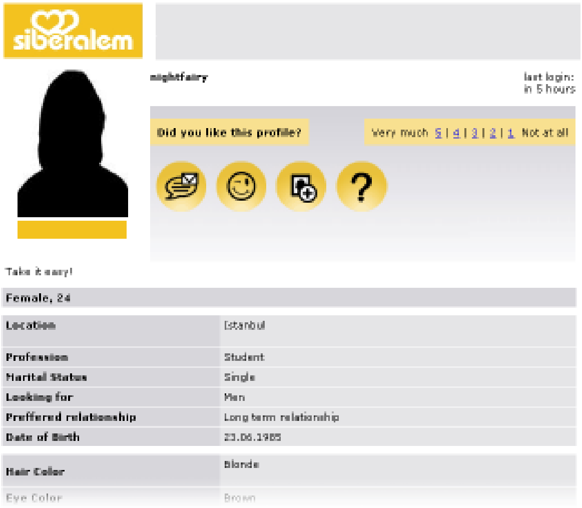
Student,
167 cm, 55 kg
Blonde hair, Brown Eyes

looking for men

Take it Easy!
Looking for serious relationship & fun!

Figure 4. Search results following a detailed search

After scanning the profile summaries in search results, the user can open profiles for a detailed view as presented in Figure 5.



nightfairy

last login:
in 5 hours

Did you like this profile? Very much 5 | 4 | 3 | 2 | 1 Not at all

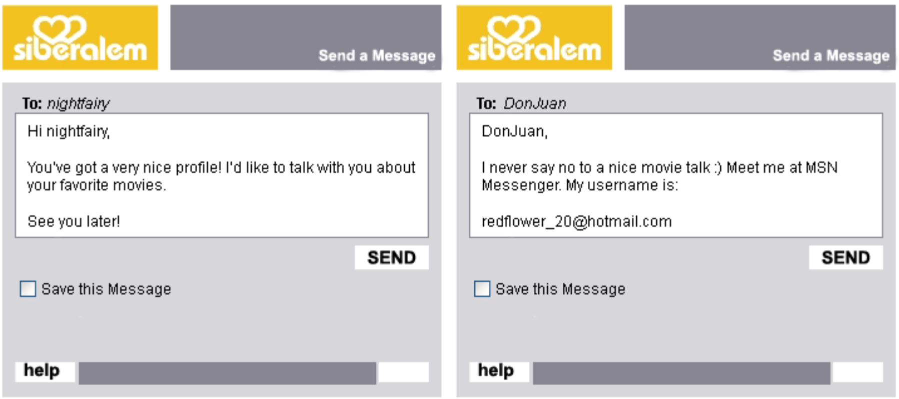
Take it easy!

Female, 24

Location	Istanbul
Profession	Student
Marital Status	Single
Looking for	Men
Preffered relationship	Long term relationship
Date of Birth	23.06.1985
Hair Color	Blonde
Eye Color	Brown

Figure 5: Detailed profile page

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 big instant messaging services like MSN Messenger or G-Talk. This is the point where users leave the siberalem.com platform and continue the relationship in other media. Figure 6 shows an example of communication via internal messaging system.



To: *nightfairy*

Hi nightfairy,

You've got a very nice profile! I'd like to talk with you about your favorite movies.

See you later!

□ Save this Message

To: *DonJuan*

DonJuan,

I never say no to a nice movie talk :) Meet me at MSN Messenger. My username is:

redflower_20@hotmail.com

□ Save this Message

Figure 6. Communication via messaging

Features

In the global market of online dating business, companies try to enrich their services by adding new products and solutions to the main line of standard functions like search and messaging. In order to be a part of the competition, these companies launch new tools and features ranging from audio-video chat to advanced matching systems.

To compete with local rivals, and also to follow the major global players, siberalem.com tries to maintain numerous features. The following list summarizes a total list of functions available at siberalem.com.

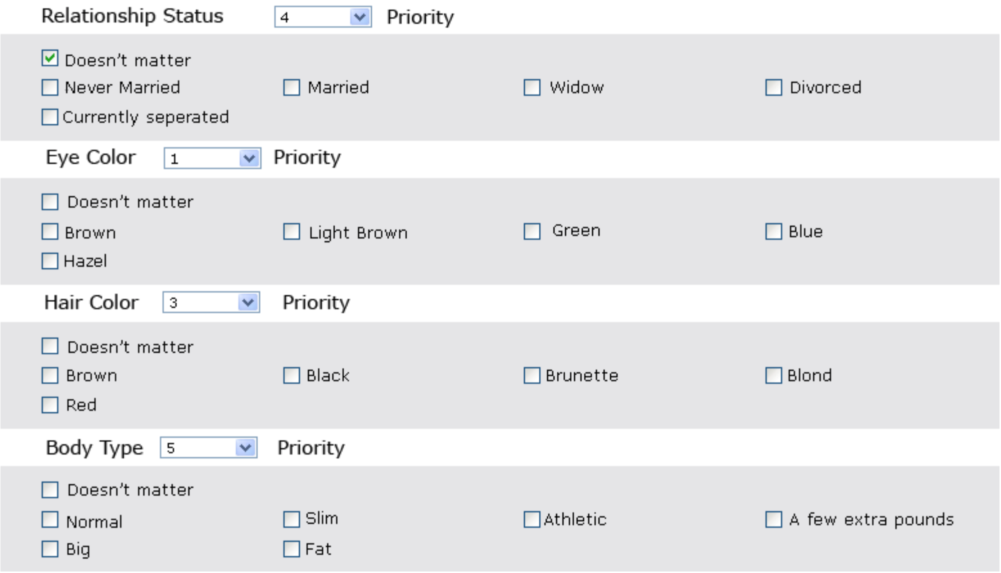
- Online text, audio & video chat
- Offline messaging
- Photo & Video upload to profile
- Advanced search, including keyword search
- Forward, backward and mutual matching
- Icebreakers (Blink)
- Sending gifts
- Template questions, questionnaires
- Friend list
- Desktop Messenger Application

Matching

A basic matching system is presented in siberalem.com which can be used by site members to query and list the best matches.

The system requires filling of an additional form asking the member his / her preferences about the potential partner. Preferences about features of height, weight, profession, marital status, eye and hair color, body form, education, spoken foreign languages, ethnicity, religion, smoking and children can be selected thorough the settings form of the matching system.

Whereas the form seems to be similar with the detailed search form, the matching form includes an importance measure for every available feature which is non-existent in the detailed search form shown in Figure 7. The user can adjust how much he/she cares about a specific feature, by selecting the level of importance from a Likert-like scale presented for every feature (from "not important at all" to "very important"). By using this form, the member reveals his preferences by selecting the features and assigning weighted values of importance for each.



Relationship Status 4 Priority

☑ Doesn't matter
□ Never Married □ Married □ Widow □ Divorced
□ Currently seperated

Eye Color 1 Priority

□ Doesn't matter
□ Brown □ Light Brown □ Green □ Blue
□ Hazel

Hair Color 3 Priority

□ Doesn't matter
□ Brown □ Black □ Brunette □ Blond
□ Red

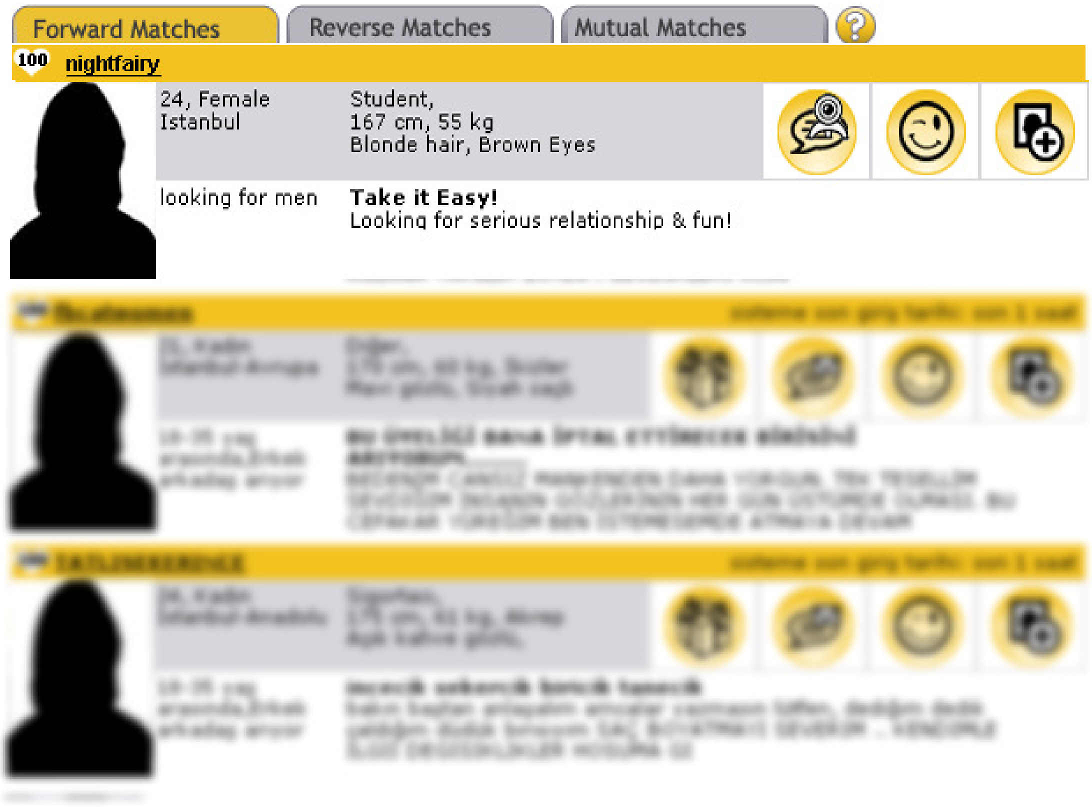
Body Type 5 Priority

□ Doesn't matter
□ Normal □ Slim □ Athletic □ A few extra pounds
□ Big □ Fat

Figure 7: Settings for matching

After the matching settings form is submitted, three types of queries can be run as shown in Figure 8:

- Forward matching: The users who are fulfilling the matching criteria of the member are listed and ordered by a matching score calculated according the weighted importance preferences of the member.
- Reverse matching: The users for whom the member fulfills their matching criteria are listed and ordered by the matching score.
- Mutual matching: The users, the sums of whose forward matching and reverse matching scores are highest, are listed.



nightfairy

24, Female
Istanbul

Student,
167 cm, 55 kg
Blonde hair, Brown Eyes

looking for men

Take it Easy!
Looking for serious relationship & fun!

Figure 8. The list of matches

Some of the other local dating sites also feature different matching systems. Istanbul.net has a non-weighted matching engine similar to siberalem.com’s existing solution.

Data Mining

Introduction

Data Mining refers to extracting or "mining" knowledge from large amounts of data (Han & Kamber, 2006). In recent years, overwhelming accumulation of scientific, industrial and commercial data has led to a serious problem: converting the raw data into meaningful information. Today, the problem is gradually worsening, since researches estimate that the total amount of data doubles every three years (Varian & Varian, 2003).

Data Mining is mainly referred to as a computer aided field. In fact, it is highly possible to carry out knowledge extraction with manual methods, which are also historically practiced. As the amount of data grows and complexity increases, computer automation becomes a must in Data Mining. Significant advances in both computer hardware and development of efficient software algorithms have made the Data Mining a more practical and popular tool in recent years.

Data Mining exhibits a wide range of application in different areas, from Business to Science, from Medicine to Military, Engineering, Genetics, Education and so on. There have been efforts to define standards for Data Mining activities among different fields that led to the emergence of various examples.

Data Mining Functionalities

Data mining functionalities can be classified into four main classes of Tasks: Classification, Clustering, Regression and Association Rule Mining.

Classification

Classification is learning a function that maps (classifies) a data item into one of several predefined classes (Weiss & Kulikowski, 1991). Examples of common classification algorithm include *Naive Bayes Classifier*, *k-nearest Neighbor*, *Decision Trees*, *Neural Networks* and *Bayesian Networks*. A graphical example for classification is shown in Figure 9.

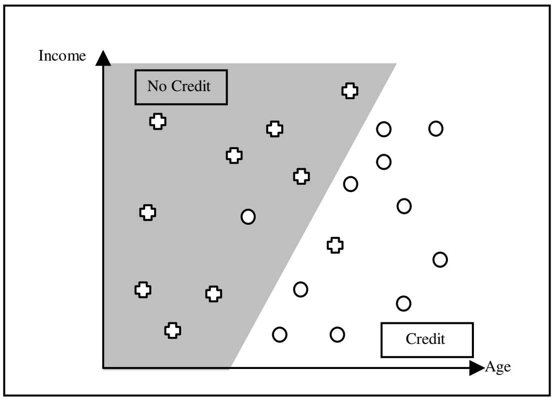


Figure 9. Separation by using classification

The classification process can be transparent similar to that in Decision Trees, or it can be opaque like the case in Neural Networks.

The example illustrated in Figure 10 is a typical decision tree classifying the customers of a financial institute.

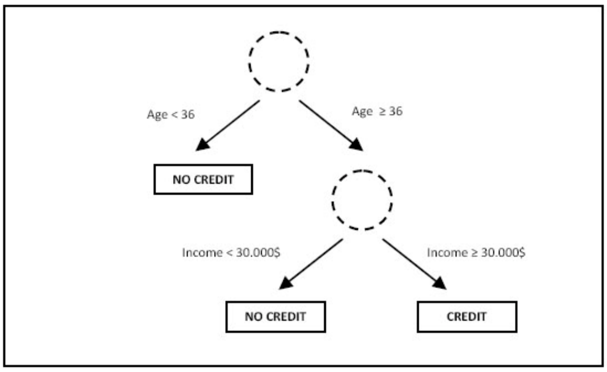


Figure 10. Typical decision tree

According to “Age” and “Income” attributes of customers, they are separated into two different classes: Credit and No Credit. Credit classes imply that they are "suitable for loaning Credits" and No Credit classes are "not suitable for loaning Credits". According to the tree, a customer below the age of 36 never gets a credit from bank. If the customer is above 36, there are two options. He/she gets the credit if his/her income is equal or greater than 30.000$. If the income is below 30.000$, the customer again fails to get the credit.

Regression

Regression is a statistical tool where a "dependent" variable is modeled as a function of the "independent" variables. The parameters of the "regression equation" are selected to maximize the "fit" of the data. "Least Squares Method" is one of the mostly used and most appropriate algorithms (Fisher, 1922).

Regression is widely used in different fields - especially in economics - to construct models, observe trends or make forecasts and predictions. Figure 11 illustrates application of a simple regression to the given dataset.

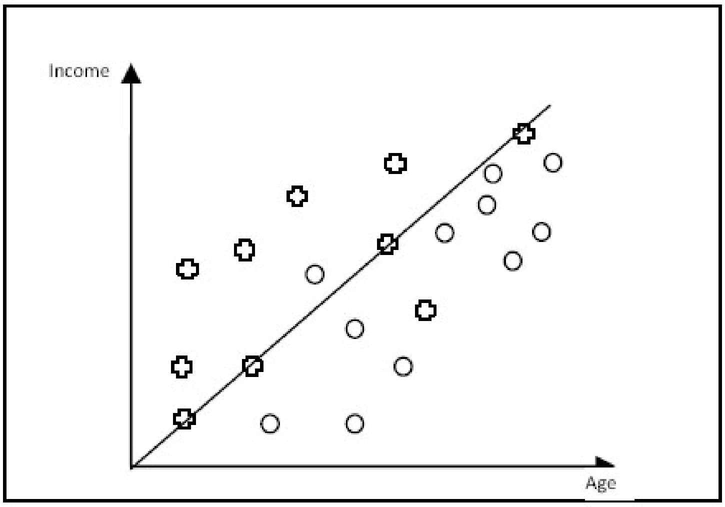


Figure 11. Simple regression

Clustering

Clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data (Jain & Dubes, 1988).

The assignment of the objects to clusters assures that objects from the same cluster are more similar to each other when compared to objects from different clusters. *K-means* is one of the most common algorithms.

Clustering applications range from simple basket analysis to web mining, image processing and automatic document classification.

Figure 12 illustrates a clustering algorithm discovering three clusters in the given dataset.

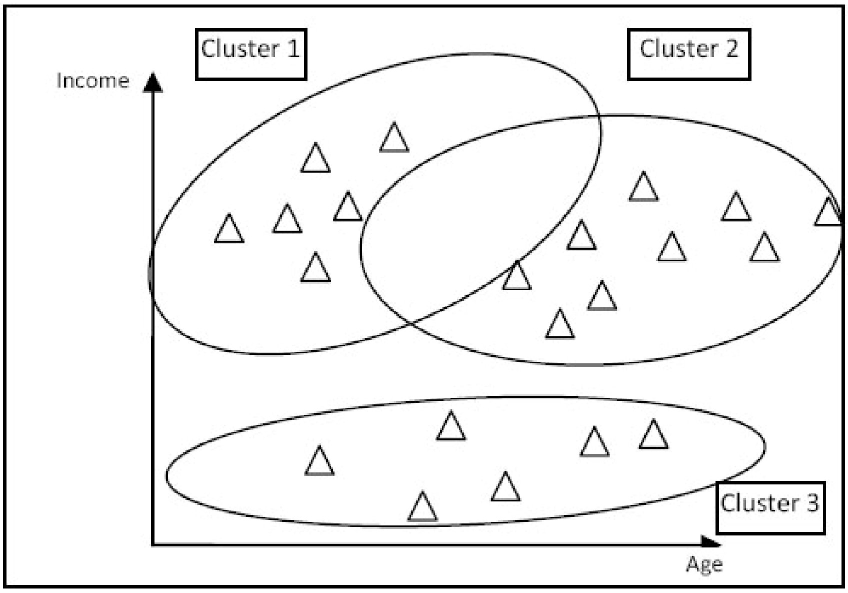


Figure 12. Clustering example

Associative Rule Mining

Associative Rule Mining is a common name for methods and algorithms that help discover interesting relationships between variables in large databases. Beyond the market basket analysis associative mining is used in many applications including *Web mining* and *intrusion detection systems.*

A typical and classical example of Associative Rule Mining is the rule discovery process from the Point of Sale transaction logs of supermarkets. Analyses of the huge transaction database of the supermarket may reveal repeating patterns or rules in the form of $X_{1,}X_2, \dots , X_n \rightarrow Y_{1,}Y_2, \dots , Y_n$ where the X and Y represent the two sides of the rule.

The self-explanatory rule $\{\text{ Tomatoes, Onions }\} \rightarrow \{ \mathit{Olive\ Oil} \}$ indicates that customers who buy tomatoes and onions together are likely to buy olive oil. Customers who plan to make a salad with tomatoes and onions need olive oil for dressing. The supermarket may increase its sales by putting the olive oil bottles near the vegetable section.

Agrawal's definition of Associative mining helps us go deeper into the concept and define the interestingness of rules (Agrawal, Imielinski, & Swami, 1993):

Let $I = \{i_{1,}i_2, \dots , i_n\}$ be a set of n binary attributes called items. Let $D = \{t_{1,}t_2, \dots , t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

Many trivial, uninteresting rules may be discovered in large scale databases, which usually hide the interesting ones from our sight. An associative rule of mining algorithm aims to discover interesting rules efficiently.

The following real-life example illustrates this on a minimal set of data given in Table 1:

Table 1. Transaction Data for Associative Rule Learning

transaction ID	tomato	onion	olive oil	Corn
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Table 1 consists of 5 transactions including the item set $I = \{tomato, onion, olive\ oil, corn\}$. The binary number on the table represents the presence or absence of items in the transaction (1 and 0 respectively). It's possible to extract many rules fulfilling Agrewal's definition, but if our objective is selecting the interesting ones, we have to set constraints on significance and interest of rule.

The two main threshold measures are *support* and *confidence*, which follow directly from the definition of the Association Mining. Also, in addition to these parameters we will also provide information about *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 mentioned "rule support", many modern statistics software 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 $antecedantsupp(X, Y) = p(X)$

This is also the definition of support which will be used in the following chapters.

In our example, the item set $\{tomato, onion\}$ has a support of $p(X) = \frac{2}{5} = 0{,}4$ since it occurs in 2 of 5 transactions.

Rules with higher support values are always preferred because of the increasing significance of the rule with the growing level of representation of the left-hand-side of the rule. To eliminate the rules below a desired support value, tools of Association Mining algorithms let users set a minimum support value before starting the mining process. By selecting a threshold value for support, following factors should be taken into consideration to achieve desired results and match the goals of the research:

If the minimum support level is set too high, the output will lack many rules with lower support but high interestingness factor. The occurrences of rule may be low, but the information it brings can be significant and interesting for the research. On the other hand, if the support level is set too low, the output grows very large in size. In such a case the elimination of the trivial rules and filtering out the interesting ones becomes harder.

Due to calculation method of the association algorithm, setting a lower minimum support level may decrease the performance of the algorithm dramatically. Especially for the mining of large datasets, the relation between minimum support level and the time required for the training of the model gains a greater importance.

The second measure, the Confidence $(Conf\ (X,Y))$ is defined as the following:

$$\mathrm{C}onf\ (X,Y) = \frac{supp(X \cup Y)}{supp\ (X)}$$

In other words, confidence is the proportion of transactions that fulfill the rule completely over the ones that have only the left-hand-side of the rule is true.

In our example the confidence of the rule $\{\ \mathrm{Tomatos, Onions}\ \}\ =>$ $\{\ Olive\ Oil\ \}$ is $\frac{0{,}2}{0{,}4} = 0{,}5$. That means, only in one case Olive Oil is bought with both Tomatoes and Onions together.

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 o rule is defined as $lift\ (X,Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$

The rule $\{\text{ Tomatoes, Onions }\} => \{\text{ Olive Oil }\}$ has a lift of $\frac{0,2}{0,4*0,4}$ = 1,25 which indicates that the observed confidence is higher than the expected confidence.

Algorithms searching association rules are mostly based on the minimum support and minimum confidence criteria set by the user. In the first part of the two-step-process all frequent item-sets are found according to given minimum support. In the second phase,

the frequent item-set below the minimum confidence threshold is eliminated.

Apriori and *FP-Growth* (Han, Pei, & Yin, 2000) are two of the most common associative rule mining algorithms based on the simple definition of Association Rule Mining. However, statistics and data mining software packages also include advanced algorithms like *The continuous association rule mining algorithm (Carma)* and *Generalized Rule Induction* (*GRI)*, which offer detailed customization of the model and alternative methods to increase the interestingness level of the rules.

One of these algorithms, the GRI algorithm is based on the ITRule algorithm and extends it with added functionality (Smyth & Goodman, 1992).

GRI 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 an information theoretic based cross-entropy calculation.

The quantitative interestingness measure J is defined as

$$J(X|Y) = p(Y)(p(X|Y)\log\frac{p(X|Y)}{p(X)} + \left(1 - p(X|Y)\right)log\frac{(1 - p(X|Y))}{1 - p(X)})$$

In which $p(Y)$ is the probability of the rule's antecedent matching an example from the dataset, $p(X)$ is the probability of the rule's consequent matching an example from the dataset, and $p(X|Y)$is the conditional probability of the rule's consequent conditioned on the antecedent.

A rule in GRI takes the form

If $Y = y$ then $X = x$ with probability p where X and Y are two fields (attributes) and x and y are values for those fields. The consequent is restricted to being a single value assignment expression while the antecedent may be a conjunction of such expressions, for example

If $X = x$ and $Z = z$ then $Y = y$ with probability p.

The complexity of a rule is defined as the number of conjuncts appearing in the rule's antecedent.

GRI generates rules through the following steps (SPSS Inc):

- Process each output field Y_i in turn. GRI derives all rules for the current output field before moving on to the next. In other words, GRI uses depth-first search to generate the rule set. For each output field, select each possible output value y_k. Again, processing is depth-first, so all rules predicting the current output field value are generated before the next output field value is considered.
- For each output value, select each input field X_m.
- For each input field, select each possible condition x_q.
- For the rule, compute the J statistic.
- If the value of J is greater than the highest J for any rule in the table predicting the same outcome ($Y_i = y_k$), or if the number of rules in the table is less than the maximum number of rules in the table, and if the minimum support and confidence criteria are met, insert the rule in the table (replacing

the lower-*J* rule if necessary) . Otherwise, proceed to the next input field value.

- Examine the inserted rule to determine whether there is any potential benefit to specialize the rule, namely adding more conditions to the antecedent of the rule. If there is an information gain based on *J* calculation, specialize the rule.
- Repeat until all input field values, input fields, output field values, and output fields have been considered.

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 which of them are mostly 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 (Vasilis Aggelis, 2004).

Methodology of Knowledge Discovery in Databases

KDD (Knowledge Discovery in Databases) is the generic name of process involving data mining methods and algorithms.

KDD basically covers three important phases of the process: data preparation, mining and interpreting. (Fayyad, Piatetsky-Shapiro, & Smyth, 1997). There exist numerous extensions of the basic concept, which try to set an appropriate standard for the data mining processes.

CRoss Industry Standard for Data Mining (CRISP-DM) is one of the best known methodologies, which is explained in detail.

CRISP-DM

CRISP-DM is a freely available standard process model developed by a consortium of industry data mining pioneers (CRoss Industry Standard Process).

According to CRISP-DM model, the lifecycle of a Data Mining project consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. As shown in Figure 13, the methodology utilizes a life cycle rather than a *waterfall* type approach.

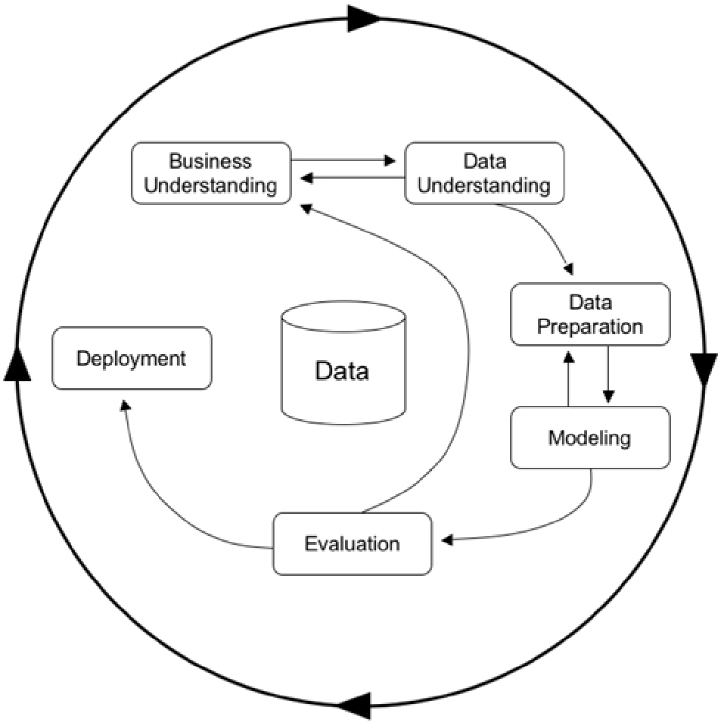


Figure 13. Life cycle in CRISP-DM model

The outer cycle in Figure 13 illustrates that the data mining project itself is a never ending process. After the deployment of the solution, a new data mining process starts with new business questions in mind and lessons learned from the previous case.

Each step is explained briefly as the following:

1) Business understanding

The main goal of the initial phase of the data mining project is to focus on business requirements and objectives onto the formation of a problem definition. A roadmap for achieving the project's goals is also necessary.

2) Data understanding

After the collection of data, a first look into the data is necessary to become familiar with the available data, identify possible problems and form hypotheses regarding hidden information.

3) Data preparation

The raw data is not suitable for data mining processing in many cases. The data has to be selected, cleaned and transformed. Data preparation tasks are likely to be performed multiple times.

4) Modeling

The modeling phase covers selection of the right data mining model, applying it to data and optimizing the model's parameters. If the data is not suitable for the model's needs, it's possible to go back and repeat the previous data preparation step.

5) Evaluation

In this phase, the models built in the Fourth Phase are evaluated considering their modeling quality and their fit into the business objectives.

6) Deployment

The creation and evaluation of models itself increase the knowledge of the data, but the presentation and organization requires a further step.

The deployment of a model can be as simple as filling a report of findings, but in many cases the deployment requires a live integration into the organizations existing systems like on personalized web pages of online retailers.

Various Studies on Online Dating and Data Mining

Online dating services provide an interesting new area for academic research and many researchers from different disciplines are attracted to the topic.

The huge data which is accumulating on databases of the online dating service provider makes the online dating domain interesting for academic research, especially for the data mining field. On the other hand, the fact that social communities and online dating sites are defining a new type of human interaction, many social scientists dealing with communications are working on the subject.

Descriptive research over online dating exists but it's not as much as it is expected. The fact that online dating is a relatively new phenomenon may explain the lack of research about this subject.

The article "Love Online: A report on Digital Dating in Canada" is an extensive analysis of the Canadian online dating market. It first tries to explain the global social environment which led people to online dating and then focuses on the local online dating statistical figures of Canada (Brym & Lenton, 2001).

According to the report, there are four social forces that drive the rapid growth of online dating:

1) A growing population is composed of singles

2) Career and time pressures are increasing

3) Single people are more mobile

4) Workplace romance is in decline.

The conclusion is that the demand for dates is on the rise, but the above listed social circumstances make it difficult for people to find good dating partners. Online dating services provide a convenient way for people looking for dates but can't create enough time in the modern world.

By comparing the survey data of Canadian online dating users and non-online dating users, researchers have found differences between these two populations: Online daters are more likely to be male, single, divorced, employed, and urban. They are also more likely to enjoy higher income.

Another article investigates 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 (Ellison, Heino, & Gibbs, 2006).

Data collected by phone interviews with users of a large dating site revealed the following results: The majority of online dating participants claim they are truthful and research suggests that some of the technical and social aspects of online dating may discourage deceptive communication. For instance, anticipation of face-to-face communication influences self-representation choices and self disclosures because individuals will more closely monitor their disclosures as the perceived probability of future face-to-face interaction increases and will engage in more intentional or deliberate self-disclosure.

A former research (Hancock, Thom-Santelli, & Ritchie, 2004) showed that design features of a medium may affect lying behaviors, and that the use of recorded media will discourage lying. This is the case in online dating profiles which are recorded medium for user's self declaration and preferences.

These findings are very important for the future researchers of the field because of the fact that it provides a solid basis for the reality and the reliability of the available data in online dating sites.

There has always been a need for human beings to meet new people with the aim of establishing romantic relationships with the opposite sex. Therefore, online dating can be seen as one of the alternative ways of finding a good “match” among the tens of other conventional alternatives.

Historically, marriage is seen as the ultimate form of “matching” and it attracted the curiosity of researchers. Among social disciplines such as sociology, anthropology and psychology, which focus on the cultural and humanistic side of this phenomenon, economics handled marriage as market equilibrium of couples.

A more generic approach is the matching topic applied to social sciences which tries to create two sided matching models applicable to different situations like “employers looking for employees” or “men looking for women for marriage”. Alvin E. Roth’s classical works about game theoretic analysis of two-sided matching is extended by further publications about the topic (Roth & Erev, 1995). Sanver’s work enriched the models with the assumptions of misrepresentation of preferences. (Sanver & Sanver, 2005).

Another area of interest beyond numerical models is about the characteristics of matching couples. A significant number of researches are done with dating partners and married couples to answer a single simple question: “Who is likely to get matched with whom?” The question is about the self characteristics of couples and their preferences. “Do similarities among couples attract each other, or , is the common myth which states exactly the opposite true?”

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 & Francesconi).

Another concept about the subject is “speed dating”. In this hype of modern ages, a small group of people meet in a bar café under the control of a moderator. The participants are rotated to meet each other for small sessions of 5 to 10 minutes and in the end of the rotation, every man and woman reports his selection to the moderator in secret. 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, and even the role of pheromones in matching (Fisman, Iyengar, Kamenica, & Simonson, 2006). In this study researchers made the following conclusions about the selectivity:

“Women put greater weight on the intelligence and the race of partner, while men respond more to physical attractiveness. Moreover, men do not value women's intelligence or ambition when it exceeds their own. Also, we find that women exhibit a preference for men who grew up in affluent neighborhoods. Finally, male selectivity is invariant to group size, while female selectivity is strongly increasing in group size.”

As mentioned before, although online dating sites stand as interesting data sources, there aren’t sufficient data mining applications in the field. In this section, one of the existing studies will be examined, and next, the potential applications where data mining and online dating meet will be explained.

Another paper titled “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. As an introduction to the paper, the authors focus on the major characteristics of online dating services by using the available data from their case study.

According to a survey on registration, site members are motivated to become a member of the site due to different reasons. 39% of the users state that they are hoping to start a long-term relationship, 26% state that they are “just looking/curious”, and 9% declare that they are looking for a casual relationship. Perhaps not surprisingly, men seem to be more eager for a short term/casual relationship (14%) than women (4%) (Hitsch, Hortaçsu, & Ariely, 2005).

In the same manner, the study also presents statistics for sexual preferences, demographic socioeconomic characteristics and reported physical characteristics of members.

In the remaining parts of the study; Hitsch, Hortaçsu and Ariely bring a different perspective by using statistical and data mining tools 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.

In addition to available data from profiles, they let a group of paid students to rate the member photos according to their physical attractiveness. This effort brings an additional parameter into the game which is otherwise not available by default.

As the methodology, regression is used. Different variables including socio-demographic and physical features are regressed to number of received messages to find out which attributes cause an increase the demand for the member. By using Gale-Shapley algorithm (Gale & Shapley, 1962) to predict the equilibrium sorting along attributes such as age, income, and education they estimate the most significant factors for forming matches.

The paper claims that the results confirm many findings obtained in psychology, anthropology, and sociology studies, which are based on stated preference data. As a given example stronger emphasis on a partner's income is found among women rather than men.

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).

Typically, a recommender system compares the user's profile to some reference characteristics, and seeks to predict the 'rating' that a user would give to an item they had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).

When building the user's profile a distinction is made between explicit and implicit forms of data collection (Adomavicius & Tuzhilin, 2005).

Explicit forms of data collection can be listed as following:

- Asking a user to rate an item on a sliding scale
- Asking a user to rank a collection of items from favorite to least favorite
- Presenting two items to a user and asking him/her to choose the best one
- Asking a user to create a list of items that he/she likes

These forms of data collection require the active participation of the user for the accumulation of the ratings.

The list of implicit forms of data collection includes:

- Observing the items that a user views in an online store
- Keeping a record of the items that a user purchases online
- Obtaining a list of items that a user has listened to or watched on his/her computer
- Analyzing the user's social network and discovering similar likes and dislikes

Online websites making use of both implicit and explicit forms of data collection for their recommendation systems achieved good results by creating a different type of user experience. Systems like Amazon.com’s (http: //amazon.com) product recommendations or last.fm’s (http: //last.fm) music service enable users discover new types of products, songs, movies or other digital media, without which they would not be able to search or browse.

Online dating sites can also make use of these new types of technology to let users discover new people according to their own ratings or browsing habits. Both explicit and implicit data collection types may be utilized for a recommender system in an online dating site; and an application for the latter will be introduced in the remaining part of the thesis.

CHAPTER III: DEFINITION OF THE PROBLEM AND PREPARATION OF DATA

Business Understanding: Definition of the Problem

Even though online dating seems to be based on technology at first sight, in fact it consists of a series of phenomena, such as relationships between men and women and the concept of marriage. These phenomena are related with the essential problems of humanity as a whole. For this reason, the mostly asked questions and possible answers have a direct relation with social life. Some examples of these questions are; “In general, what type of women are regarded as attractive women in general?”, “How does a man begin to show his interest to a woman?”, or “What are the features of particular men who date with particular women, and vice versa?”

When social networks on the Internet and online dating sites are accepted as virtual reflections of complex social relationships, they turn into laboratories where researchers may easily observe human relationships. When one correctly examines the data containing personal and physical properties of millions of people, their friends and their communicational records, he/she will clearly see that the existing data stands as a potential resource to answer the questions we have been asking.

The research carried out by Hitsch, Hortaçsu and Ariely on an online dating site also aims to find answers to such questions (Hitsch, Hortaçsu, & Ariely, 2005). The article explains in detail how a wide range of factors, including socio-demographic features or physical attractiveness affect popularity and attraction.

The findings in this study naturally lead us to a new question: On online dating sites, what types of individuals prefer to be matched with one another? Can we predict whether a relationship will begin or not by examining the personal features of a man or a woman? When we examine the couples who meet on an online dating site, can we find repeating patterns or rules in the features of both sides? All in all, the basic question is; "Who meets who?"

The possible answers given to this question will not only enlighten a social phenomenon that has been intriguing researchers for a long time, but will also provide new horizons for online dating sites in terms of technological tools and the variety of services.

As it has already been described in previous chapters, the typical scenario on 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 many 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 reaches a list of potential matches he/she desires to find.

This process expects the user to list his/her expectations in an exact manner; 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.com 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.com throughout many years, a model may be built including a rule set concerning what types of profiles are matched with what types of 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 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.

All the "recommender systems" which have recently appeared on the web with algorithms such as "collaborative filtering", have become recognized in many different fields ranging from e-commerce to entertainment sector. 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.

Preparation and Understanding of Data

Finding Matches

The databases and log files of online dating services keep the record of any user action including registration, profile updates, recurring visits and messaging. Analysis of the available records allows us to visualize the inner dynamics of the system; it also can provide exact information about the key indicators of the site, including member statistics, member profile characteristics and volumes of user interaction.

A problem arises by matching: observing the user interactions outside the system is nearly impossible. The further information from the users who initiated a contact in siberalem.com is not available. It's not possible to know if they had a further relationship including meeting in the physical environment. The possibilities vary from a short conversation on the phone to marriage.

Although it is not possible to pursue what exactly takes place between two matches who met on siberalem.com after they meet in real life, an estimator can be used to reveal successful contact in siberalem.com by using the available data within the system. The suggested 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 on the internal messaging system: The user has to send messages and also receive ones in response, with the purpose of starting a relationship.

As a direct consequence of the imbalanced men/women ratio in siberalem.com, women tend to select their mates as opposed to being selected by men. This results in many unsuccessful conversation attempts initiated by men. A first message is sent by a man but the conversation doesn’t start when the woman doesn’t respond. The elimination of the incomplete conversations and finding of real communication throughout the messaging system is the first step of revealing the matching users.

The conversations including an initial message and a response are likely to continue for 3-4 reciprocal messages. For these types of conversations, there are, 7.2 total sent and received messages on average. After that, another medium of communication is preferred by users.

Due to limitations of the environment on the web, users do not prefer to continue the communication on online dating sites after they initiated a first contact. This argument is confirmed by user testimonials and focus group meetings (Kalita Araştırma, 2009).

Before exchanging phone numbers (which is perceived as an insecure practice), an intermediary step in communication is exchanging instant messenger (IM) addresses in form of ID, alias, nickname or e-mail address. Here, it should be pointed out that in the year 2008, over 28 million MSN Live Messenger accounts are owned by Turkish people (Microsoft, 2008). The MSN Live Messenger's user IDs are in the form of e-mail addresses.

For any conversation in the internal messaging system of siberalem.com, 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: It indicates that the conversation started in siberalem.com 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, we have made the following definitions:

Message: A text message sent from one user to another, represented as a single row in the database.

Conversation: Series of text messages between two users. The conversations which contain any "auto-reply" messages are not valid because of the unintentional user action.

Match: Any conversation which fulfills the following conditions:

1) The minimal number of both sent and received messages must be equal or greater than 2.
2) In one of the messages other than the initial message, IM messenger IDs or e-mail addresses must be exchanged.

If the exchange happens in the initial message, the conversation doesn't count as a match, because in many cases some impatient men send their e-mail addresses in the first message although a real contact does not exist.

The message table in siberalem.com's database has the following data structure presented in Table 2 with two rows of example data in:

Table 2: Simplified View of Message Table in the Database

MessageID	FromUserID	ToUserID	MessageBody	Date
1	112323	34589	Hi, how are you?	01/01/2007
2	34589	112323	Thank you. I'm fine.	01/01/2007

The message table is a snapshot of the production environments database taken in the year 2008. The dataset that will be used in the remaining part of the analysis consists of the 700000 rows of this table covering the messages sent between February 2006 and January 2008

To extract the "matches" from the database according to the rules, a small program is written in C# language. The program scans every row in the message table and finds all the conversations containing more than one sent and more than one received message. The messages in the conversation are searched for IM IDs by using regular expressions. If a string in the form of e-mail addresses is found, the conversation is flagged as a match.

Table 3 gives the summary statistics:

Table 3: Conversations Found by Program

	Total	Percent
Total Number of Conversations	5056841	100,00%
Conversations with Received Messages >=1 and Sent Messages >=1	196624	3,89%
Conversations with Received Messages >=2 and Sent Messages >=2	37476	0,74%

The results further indicate that 24408 (71%) of 37476 conversations contain messages in which e-mail addresses are exchanged, so these are flagged as matches. 14691 (61%) of the "matching" conversations are initiated by men, in 9501 (38%) the first message is sent by a woman. In 216 conversations (1%), one of the sides of the conversation does not exist in the main member tables of the database any more (due to clean-up or resignation of member), so these conversations are excluded from further analysis.

To test the validity of assumptions, a simple empirical study is done on the results: Samples from conversations, marked as "matches" are examined in terms of their content while keeping the user information anonymous. The majority of messaging in conversation was similar to the following samples shown in Figure 14, which prove the connection between communication on siberalem.com and the starting of a relationship:

User A: Vallahi burada bir daha seni görür müyüm bilmiyorum. Bir merhaba diyeyim dedim! (I don't know if we'll ever meet here again, but I wanted to say Hello to you!)

User B: Selam! MSN adresim xxx@xxx.com! (Hello! My MSN address is xxx@xxx.com !)

User A: Ekledim seni MSN listeme ama online görünmüyorsun. (I've added you to my friend list, but you don't seem to be online.)

User B: Bekle bekle. Yarım saate oradayım… (Wait for me! I will be there in 30 minutes.)

User C: Profilin başlığın ilginçmiş. "Yalnızlığa alışmak" yazmışsın ya, hoşuma gitti! (What an interesting profile title. I liked the phrase "getting used to loneliness"!)

User D: Şansım dönüyor mu ne? Sanırım artık yalnız kalmayacağım. MSN kullanıyor musun? (I think my tide is turning, no more loneliness for me! :) Do you have an MSN address?)

User C: Evet, xxxx@xxxx.com. Adım da xxx bu arada. (Yes, xxxx@xxxx.com. My name is xxxx, indeed.)

User D: Tamam xxxx, benim adresim de xxxx@xxxx.com! (Ok xxxx. My address is xxxx@xxxx.com.)

Figure 14. Conversation examples

In the next step, demographical, physical, socio-cultural attributes and preferences of both man and woman of every conversation will be collected to provide the complete input data for the building of the model. The following section will explain the phases of selection of these features and transformation of data.

Selection of Data

For every man and woman who appear as sides of the 24192 available matches, we have to select features available from the siberalem.com's user data. The features will be used as input variables for the model.

Registration to siberalem.com requires filling a web form to obtain specific user information. The form consists of multiple pages and many form elements of different types to let the user reveal information about him/herself and his/her preferences.

The registration form includes text boxes for free text input, radio buttons and drop-down boxes for single selections of and checkboxes for multiple selections from the defined sets. The user doesn’t have to fill out every single field except for the ones which are explicitly marked as “required”. After filling the form and logging in to the site, the user may go back to the form and fill the missing values any time he/she wants.

Table 4 presents an example of profile fields at siberalem.com with additional information of data types and rate of missing values. Table 4 shows a selection of available values.

Table 4: siberalem.com Profile Field Examples

Field	Data Type	is Required?	Answered in % of male profiles	Answered in % of female profiles
Gender	Categorical	yes	100,00%	100,00%
Location	Categorical	yes	100,00%	100,00%
Age	Continuous	yes	100,00%	100,00%
Relationship you looking for	Categorical, Multiple Choice	yes	100,00%	100,00%
Preferred Gender	Categorical	yes	100,00%	100,00%
Preferred Age	Categorical	yes	100,00%	100,00%
Preferred Location	Categorical	yes	100,00%	100,00%
Profession	Categorical	yes	99,30%	99,54%
Marital status	Categorical	yes	98,98%	99,18%
Profile title	Free text	yes	100,00%	100,00%
About yourself	Free text	yes	100,00%	100,00%
Eye color	Categorical	no	64,92%	68,30%
Hair color	Categorical	no	63,65%	67,57%
Weight	Continuous	no	64,47%	63,33%
Height	Continuous	no	64,40%	63,52%
Outfit	Categorical	no	59,43%	62,23%
Sports you do	Free text	no	44,30%	41,76%
Football Club	Free text	no	36,96%	38,07%
Food you cook well	Free text	no	59,75%	64,49%
Name three items in your bedroom	Free text	no	50,39%	55,90%
Education	Categorical	no	42,58%	48,01%
Languages you speak	Categorical, Multiple Choice	no	39,27%	40,55%
Monthly income	Categorical	no	41,53%	34,47%
Ethnicity	Categorical, Multiple Choice	no	45,91%	48,79%
Religion	Categorical, Multiple Choice	no	41,52%	42,78%
Importance of religion in your life	Categorical	no	41,66%	46,59%
Smoking habits	Categorical	no	45,18%	50,02%
Drinking habits	Categorical	no	45,78%	50,03%
Eating habits	Categorical	no	44,50%	48,49%
Do you have children?	Categorical	no	39,40%	43,31%
Do you want children?	Categorical	no	37,09%	41,49%

Four types of fields are eliminated from the input variables list of our model because some of their characteristics avoid them from being an input parameter for the model.

1) Fields in free text form: Parsing required information and passing it to the model is not feasible.
2) Fields about detailed character features and lifestyle: Some profile questions on siberalem.com aim to let users tell more about the details of their lives and enrich their profiles. Questions like “food you cook well” or “name three items in your bedroom” are examples of this type which are not included in the search queries on site and which are found irrelevant for any matching algorithm.
3) Fields having high number of (>50) categorical values on which grouping is not possible (like Profession field).
4) Fields asking user preferences about the potential partner.

After the elimination of the features according to the above characteristics, the following feature list is formed:

- Age
- Location
- Income
- Marital status
- Education
- Smoking / Drinking habits
- Importance of religion in life
- Having Children
- Wanting to Have Children
- Eye & Hair Color
- Weight / Height
- Preferred Relationship

Data Transformation

The following steps explain how selected variables are handled one by one to apply necessary transformations like reclassification and discretization. Sections also include descriptive information and frequency distributions of resulting variables.

Age

Age is a calculated field which is based on the birth date information filled by the users. The continuous variable which is kept as an integer in the database is discretized by using the following mapping shown in Table 5:

Table 5: Distribution of Age

Old Values	Age Groups	% of Women	% of Men
Age>=18 and Age<=22	18-22	12.98%	2.24%
Age>22 and Age<=27	23-27	22.55%	17.38%
Age>27 and Age<=32	28-32	20.10%	26.16%
Age>32 and Age<=38	33-38	20.05%	20.83%
Age>38 and Age<=45	38-45	15.79%	18.21%
Age>45	45+	9.31%	14.48%

This distribution has indicated two significant facts. Firstly, the ages of members cluster around 27-32, unlike the other social networking sites where most members are younger than 25 years-old. Secondly, male members are older than female members in general.

Location

The members using siberalem.com are generally located in the three big cities of Turkey; namely, Ankara, Istanbul and Izmir. These cities are followed by Bursa and Adana.

When the total number of the members living in rural areas of the country is calculated, it also constitutes an important sum, shown under the title “small town”. Location information is stored in database by using integer IDs assigned to cities in Turkey. For foreign countries, the ID value is -1.

Using the following mapping in Table 6, the city variable is reclassified into another one containing significantly less values than the original:

Except the 5 big cities of Turkey according to their population and the foreign countries, other small cities are grouped under the title “Small Town”.

Table 6: Distribution of Location

Original Value	Location	% of Women	% of Men
1	Ankara	11.72%	13.69%
2,3	Istanbul	35.13%	36.45%
4	Izmir	11.39%	10.12%
24	Bursa	4.17%	4.15%
5	Adana	2.39%	2.27%
N/A	Small Town	30.28%	28.19%
-1	Abroad	5.72%	5.94%

Income

When the distribution of income is analyzed, it may be stated that men receive higher income when compared to women.

No transformations are applied to the income variable; the same classification is used as it’s stored in the database. Men and women of match couples show the following distribution of income levels as shown in Table 7:

Table 7: Distribution of Income

Income	% of Women	% of Men
less than 750 TL / month	9.38%	1.82%
750-1500 TL/month	20.23%	14.10%
1500-2250 TL/month	9.88%	20.82%
2250-3750 TL/month	4.35%	16.39%
more than 3750 TL/month	4.27%	26.88%
N/A	52.68%	20.79%

Marital Status

The distributions indicated that men and women are equal in the distribution of their marital status.

In addition to “single”, “married” and “married but separated” the status of being “divorced” and “widowed” are combined in to the “divorced” as the fourth value. The following distributions on Table 8 are obtained:

Table 8: Distribution of Marital Status

Marital Status	% of Women	% of Men
Single	65.90%	62.06%
Married but separate	3.46%	6.55%
Married	5.37%	6.19%
Divorced	25.85%	25.89%

Education

The education values are stored as integer IDs on the database. After converting the IDs to labels of education levels and combining the Primary School values (same label is used for 5 years and 8 years of Primary Education) the following distributions on Table 9 are obtained:

Table 9: Distribution of Education

Education	% of Women	% of Men
Graduate	8.49%	21.79%
Undergraduate	28.42%	43.78%
Junior College	12.25%	9.06%
High School	18.06%	9.37%
Primary School	1.82%	0.81%
N/A	31.76%	15.99%

Drinking and Smoking Habits

Smoking values are reclassified into “smoker” and “not smoker”, also the Drinking values are transformed into “drinker”, “not drinker” and “social drinker”. Table 10 and Table 11 show the frequency distributions of these values.

Table 10: Distribution of Drinking Habits

Drinking Habits	% of Women	% of Men
Drinker	29.21%	12,51%
Social Drinker	35.43%	49.48%
Non-Drinker	9.48%	24.80%
N/A	25.88%	13.37%

Table 11: Distribution of Smoking Habits

Smoking Habits	% of Women	% of Men
Non-Smoker	33.14%	35.01%
Smoker	39.75%	52.64%
N/A	27.92%	12.35%

Importance of religion in life

No transformation is applied on data. The distribution for the variable is given in Table 12:

Table 12: Distribution of Importance of Religion

Importance of religion in life	% of Women	% of Men
Very Important	%18,04	%20,6
Important	%22,33	%20,6
May be Important	%14,36	%9,5
Very Little	%7,15	%10,35
Not Important	%5,55	%13,42
N/A	%32,55	%25,42

Having Children

No transformation is applied on data. The distribution for the variable is given in Table 13:

Table 13: Distribution of Having Children

Having Children	% of Women	% of Men
I have children, not staying by me	%3,0	%11,63
I have children, staying by me	%14,4	%9,77
No Children	%46,5	%54,30
N/A	%35,9	%24,29

Wanting to Have Children

No transformation is applied on data. The distribution for the variable is given in Table 14:

Table 14: Distribution of Wanting to Have Children

Wanting to Have Children	% of Women	% of Men
I don't know	%12,70	%17,95
I don't want to have children	%19,59	%17,34
I want to have children	%28,67	%34,18
N/A	%39,02	%30,5

Eye & Hair Color

The following figures show the distribution of other physical attributes among men and women: the eye and hair color. The distributions are given in Table 15 and Table 16:

Table 15: Distribution of Eye Color

Eye Color	% of Women	% of Men
Hazel	18.29%	23.43%
Green	10.36%	11.84%
Brown	52.02%	54.10%
Blue	3.88%	4.05%

Table 16: Distribution of Hair Color

Hair Color	% of Women	% of Men
Red	9.23%	0.18%
Gray	0.41%	10.09%
Brown	40.80%	29.20%
Blond	20.61%	3.23%
Black	13.08%	49.50%

Height and Weight

Both height and weight are stored as continuous variables on database. Because of the inter-dependence of these variables by defining body shape in people, a new combined value is created to obtain a better estimator representing the body form.

Body Mass Index (BMI) is a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults. It is defined as the weight in kilograms divided by the square of the height in meters (World Health Organization, 1995).

After the calculation of BMI score, the values can be classified as shown on Table 17.

Table 17: Distribution of BMI

Original BMI Score	BMI	% of Women	% of Men
BMI<18.5	Underweight	12.43%	0.19%
BMI>=18.5 and BMI<24.9	Normal	57.40%	65.42%
BMI>=24.9 and BMI<29.9	Overweight	5.92%	25.38%
BMI>=29.9	Obese	1.52%	1.55%

Preferred Relationship Type

When the types of relationship preferred by male and female members are considered, it may be observed that friendship, which is a socially accepted type of relationship, is preferred the most. Short term relationships, which may be described as a more marginal type, is preferred less in general, and is more preferred by men more than it is by women. It may also be concluded that men tend to prefer marriage more than women do.

The values for multiple-choice question of preferred relationship types are stored in a single field of database by making the use of bitwise notation.

The necessary bitwise operations are done to extract the available information. The Table 18 summarizes the selections made by both women and men.

Table 18: Distribution of Wanting to Have Children

Preffered Relationship	% of Women	% of Men
Friendship	94.20%	91.08%
Long-Term	42.81%	77.68%
Short-Term	15.71%	65.47%
Marriage	33.16%	43.55%

Binary Encoding of Data

The last step in the data preparation phase is bringing the categorical variables in the form of multiple binary variables to make them available for the association rule mining. The following example variable shown in Table 19 explains how the encoding is conducted:

Table 19: Encoding Example of a Variable

Categorical Form	Binary Form
MaritalStatus	RC_MaritalStatus_Single RC_MaritalStatus_Married RC_MaritalStatus_Divorced RC_MaritalStatus_Married but seperate

One single categorical variable is encoded to 4 binary variables having 0 or 1 as values. The encoded variables have the prefix “RC_”.

To distinguish the feature variables of two genders of a couple, certain suffixes are used. The female features become the suffix _W (like RC_MaritalStatus_Single_W) and the male features become the _M (like RC_MaritalStatus_Single_M). An example of the two variables is given in Table 20.

Table 20: Example List of Input Variables

Man’s Features	Woman’s Features
RC_Education_Primary School_M	RC_Education_Primary School_W
RC_Education_High School_M	RC_Education_High School_W
RC_Education_Junior College_M	RC_Education_Junior College_W
RC_Education_UnderGraduate_M	RC_Education_UnderGraduate_W
RC_Education_Graduate_M	RC_Education_Graduate_W
RC_MaritalStatus_Single_M	RC_MaritalStatus_Single_W
…	…

CHAPTER IV: BUILDING OF THE MODEL AND THE MATCHING ENGINE

This chapter explains the development of the recommendation engine in two phases. In the first phase, a model is trained by using the prepared data to get a set of rules defining patterns found in features of matching couples. In the second phase, a computer program is developed to generate recommendations for site users full of members with high matching potential, by making use of the discovered rule set.

The last section of the chapter describes the test environment created to measure the performance of the matching engine and summarizes the evaluation of the results.

Building of the Model

The Prepared Dataset

In Chapter III, by identifying the matching couples, querying their features from the database and making the necessary data transformations, we have achieved the final form of the dataset.

As presented in the end of the last chapter, the table consists of WomanID and ManID columns of matching couples in addition to the profile feature attributes of man and woman in binary form. There are 12 features for every man and women which count up to 60 feature attributes in binary form.

The table has 15265 rows, representing the couples marked as "matches". The couples are formed from a group of 9159 distinct men and 1160 distinct women.

Selecting of the Data Mining Functionality

The information we have to extract from the data is supposed to consist of common patterns or so-called rules found repeatedly among the features of matching couples. As it may obviously be seen in the data set, in the training data, we lack a sample of non-matching couples, and therefore we are devoid of a target variable which classifies a couple as a match or non-match. Consequently, all the supervised methods such as classification or regression which require a target variable to be trained are not suitable for the solution of the problem (Han & Kamber, 2006). In such a case, when we take unsupervised methods into consideration, the only functionality which can perform rule discovery is the association algorithm.

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 21.

Table 21. 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 big difference in numbers of the discovered rules of the two algorithms originates from the similar rules found by the Apriori algorithm repeated many times in the ruleset. An example is shown in Table 22.

Table 22. Redundant Rules of Apriori Algorithm

Antecedant	Consequent
RC_Age_Group_W_18-22 and RC_HairColor_W_brown and RC_BMI_Status_W_Normal and RC_PrefRel_W_Friendship	Age_Group_M_18-22
RC_Age_Group_W_18-22 and RC_HairColor_W_brown and BMI_Status_W_Normal	Age_Group_M_18-22
RC_Age_Group_W_18-22 and RC_Smoking_W_no and RC_HairColor_W_brown and RC_PrefRel_W_Friendship	Age_Group_M_18-22

The GRI algorithm, on the other hand, output fewer rules in higher quality and interestingness by omitting the similar rules which create no information gain. By considering the test results, the GRI algorithm is selected as the tool to train the model.

Variables

The GRI algorithm requires that the input variables will be given separately as "antecedents" and "consequents".

The output rules will be in the form of $X_1, \dots, X_n \rightarrow Y_1$, where X_i represent antecedent variables and Y_1 the consequent one, respectively. The number of antecedents in a rule can be any number greater than one, if no other constraint is set explicitly in the model settings. On the other hand, every rule has a single consequent on the right hand side.

The necessity of rules in the form of $X_{1,}, \dots, X_n \rightarrow Y_1, Y_2$ with multiple consequents is questionable (Christian Borgelt, 2002). This phenomenon is explained clearly in the documents of an Apriori algorithm implementation developed by Christian Borgelt and Rudolf Kruse. The developers state that due to the following reasons they excluded the option of multiple consequents from their software:

1- The number of rules grows much bigger

2- Complex rules contribute very little to the insights about the data.

According to both arguments, addition of multiple consequents to the right hand side of the rule, let alone being useful, turns out to diminish the quality of the rule set. Therefore, the discovery of the rules with multiple consequents will also be omitted in our study.

Considering the two gender types (man and woman) in the dataset, two specialized association rule form will be derived from the general one. The first one shows, the existence of which "man's features" lead to a potential match with the given "woman's feature": $Y_1, \ldots, Y_n \rightarrow X_i$. The reverse case shows, the inclusion of which "woman's features" lead to a potential match with the given "man's 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 "man's features" will be given as the antecedents and the "woman's features" will be the consequents. In the second run, the "woman's features" will be given as the consequents and the "man's features" will be the antecedents.

Model Settings

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. Other parameters include detailed settings such as maximum number of antecedents and maximum number of rules.

Minimum Support

Setting of minimum support variable for the model limits the output rule set with the rules having support values greater than the minimum support.

Our data originating from an e-dating service represents the preferences of site members about the opposite sex, which shows a great variety among members. To discover the different patterns from minor groups, the minimum antecedent support has to be set below a certain limit. By taking this fact into consideration, the minimum support is selected as 2.0.

Minimum Confidence

For our mining process we tried to set the minimum confidence level as high as possible. The only trade-off between the minimum confidence level and the quality of the results is the following:

At higher minimum confidence levels, the distinct number of consequents in total rule set decreases (some of the features do not appear in the right-hand-side of any rule).

The minimum confidence level is set at 10%. As the low level of minimum confidence may seem problematic, the addition of a minimum Lift parameter guarantees the validity and the quality of the rules.

Minimum Lift

A minimum lift value of 1.0 is set for the training of the model.

Building of the Model

The training of the model twice for both genders produced the following summary results in the GRI implementation shown in Table 23

Table 23. 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

Conversion of the Rules

The rules in the form of variable names as strings is not easy to store and use in the deployment process. To simplify the notation of the rules, the string representation of the rules are parsed and converted to integer IDs according to following mapping shown in Table 24:

Table 24. The Conversion Mapping for the Rules

Feature	FeatureIndex
RC_Education_Primary School	1
RC_Education_High School	2
RC_Education_Junior College	3
RC_Education_UnderGraduate	4
RC_Education_Graduate	5
RC_MaritalStatus_Single	6
RC_MaritalStatus_Married	7
RC_MaritalStatus_Divorced	8
RC_MaritalStatus_Married but seperate	9
…	…

Table 25 and Table 26 present example rules for both men and women. The rules are sorted according to Lift parameter in descending order. The last two columns shown in Table (Antecedent IDs and Consequent IDs) represent the integer IDs of the rule antecedents and consequent after conversion.

Table 25. Example Rules for Men

Antecedent	Consequent	Supp.	Conf.	Lift	Ante. IDs	Cons. IDs
RC_Location_Bursa_M	RC_Location_Bursa_W	4.12	58.38	14.11	15	15
RC_Location_Abroad_M	RC_Location_Abroad_W	5.89	31.3	5.52	10	10
RC_Education_UnderGraduate_M and RC_MaritalStatus_Single_M and RC_Location_Izmir_M	RC_Location_Izmir_W	3.12	61.06	5.40	4;6;14	14
RC_Education_UnderGraduate_M and RC_Location_Izmir_M	RC_Location_Izmir_W	4.97	60.27	5.33	4;14	14
RC_Education_UnderGraduate_M and RC_Age_Group_45+_M	RC_Age_Group_45+_W	5.24	48.34	5.23	4;27	27
RC_Location_Izmir_M	RC_Location_Izmir_W	10.04	58.81	5.20	14	14
RC_Education_UnderGraduate_M and RC_MaritalStatus_Divorced_M and RC_Age_Group_45+_M	RC_Age_Group_45+_W	3.61	48	5.20	4;8;27	27
RC_MaritalStatus_Single_M and RC_Location_Izmir_M	RC_Location_Izmir_W	6.05	58.33	5.16	6;14	14
RC_MaritalStatus_Divorced_M and RC_Age_Group_45+_M	RC_Age_Group_45+_W	8.97	47.49	5.14	8;27	27
RC_MaritalStatus_Divorced_M and RC_Location_Istanbul_M and RC_Age_Group_45+_M	RC_Age_Group_45+_W	4.06	46.9	5.08	8;13;27	27
RC_Location_Istanbul_M and RC_Age_Group_45+_M	RC_Age_Group_45+_W	5.93	45.51	4.93	13;27	27
RC_Education_Graduate_M and RC_Location_Ankara_M	RC_Location_Ankara_W	3.65	56.63	4.87	5;12	12
RC_Age_Group_45+_M	RC_Age_Group_45+_W	14.37	44.42	4.81	27	27
RC_MaritalStatus_Divorced_M and RC_Location_Ankara_M	RC_Location_Ankara_W	3.19	55.7	4.79	8;12	12
RC_Education_UnderGraduate_M and RC_Location_Ankara_M	RC_Location_Ankara_W	6.58	55	4.73	4;12	12
RC_Location_Ankara_M	RC_Location_Ankara_W	13.58	54.16	4.66	12	12
RC_MaritalStatus_Single_M and RC_Location_Ankara_M	RC_Location_Ankara_W	8.52	53.28	4.58	6;12	12

Table 26. Example Rules for Women

Antecedent	Consequent	Supp.	Conf.	Lift	Ante. IDs	Cons. ID
RC_Location_Bursa_W	RC_Location_Bursa_M	4.14	58.14	14.11	15	15
RC_Education_High School_W and RC_MaritalStatus_Single_W and RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	3.3	16.54	7.45	2;6;22	22
RC_Education_High School_W and RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	3.38	16.28	7.33	2;22	22
RC_MaritalStatus_Single_W and RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	12.51	14.84	6.69	6;22	22
RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	12.88	14.73	6.64	22	22
RC_MaritalStatus_Single_W and RC_Location_Small Town_W and RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	5.91	13.65	6.15	6;11;22	22
RC_Location_Small Town_W and RC_Age_Group_18-22_W	RC_Age_Group_18-22_M	6.08	13.61	6.13	11;22	22
RC_Education_UnderGraduate_W and RC_Location_Izmir_W	RC_Location_Izmir_M	3.5	58.09	5.79	4;14	14
RC_MaritalStatus_Single_W and RC_Location_Izmir_W	RC_Location_Izmir_M	6.68	55.38	5.52	6;14	14
RC_Location_Abroad_W	RC_Location_Abroad_M	5.68	32.48	5.51	10	10
RC_Location_Izmir_W	RC_Location_Izmir_M	11.3	52.23	5.20	14	14
RC_MaritalStatus_Divorced_W and RC_Location_Izmir_W	RC_Location_Izmir_M	3.27	51.14	5.10	8;14	14
RC_MaritalStatus_Divorced_W and RC_Age_Group_45+_W	RC_Age_Group_45+_M	6.65	71.1	4.95	8;27	27
RC_Location_Istanbul_W and RC_Age_Group_45+_W	RC_Age_Group_45+_M	4.18	70.06	4.88	13;27	27
RC_Age_Group_45+_W	RC_Age_Group_45+_M	9.24	69.08	4.81	27	27
RC_MaritalStatus_Single_W and RC_Location_Ankara_W	RC_Location_Ankara_M	7.62	64.79	4.77	6;12	12
RC_Education_UnderGraduate_W and RC_Location_Ankara_W	RC_Location_Ankara_M	3.99	64.08	4.72	4;12	12
RC_Location_Ankara_W	RC_Location_Ankara_M	11.62	63.26	4.66	12	12

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.

As stated in the thesis goals before, 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.

Preparation of User Input Data

To start generating recommendations, the engine needs to be fed with profile data of members actually using the system. On the production environment, the profile data of 150000 "active members" will serve as the source of this input data. For the testing purposes of the system a sample of 145692 users are imported to the database. Before the insertion, all the member features are transformed and converted to integer arrays to match the notation used by the rule set.

The General Principles of the Matching Engine

After the registration, every member may request recommendations from matching engine for him / herself. The results are listed in a separate listing page of the site. The number of the returning results is limited with a certain value. If there are more matches than the upper limit, the top results will be listed.

To get the best matches of a user in the system, a two step process is used. Firstly, the system finds all the rules which of their 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 loops through the members of the opposite gender to score them according to the found rules satisfied for the user.

To normalize the scores of members satisfying different number of rules, the calculated score is divided by the number of rules. Sorting of the aggregated scores of the members determines the best matches for the user.

Scoring

Among all the available members for matching, the ones satisfying a user rule's "consequent" criteria get a cumulative score as high as the confidence of the corresponding rule. No other rule parameters are used for scoring purposes.

The Algorithm

A computer program is developed according to the above stated principles.

On initializing, the program establishes the proper database connections and loads the rules and user data to the computer memory as arrays of special "structs". This process makes the in-memory calculations available, decreases the number of slow database operations and drastically increases the performance of the software.

After the initializing process, the following algorithm is used to return the best matches of the user in form of a list. The following pseudo code shows how the program outputs a list of Women with their calculated match scores for a given single men:

- 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

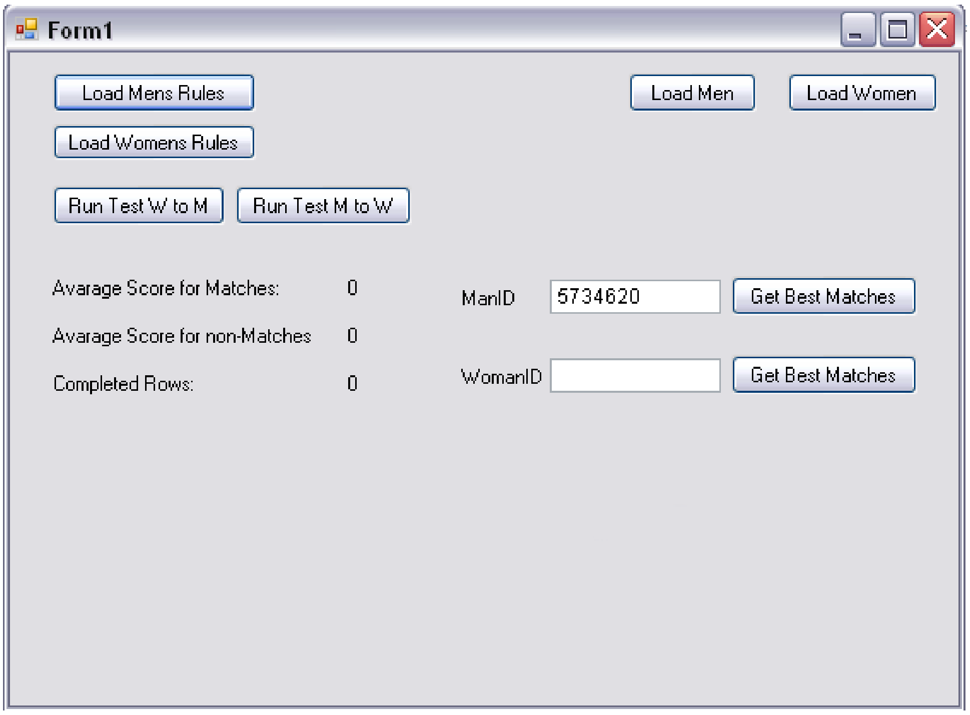
```
for each r ∈ $R_m${
if   ($M_p$.features ⊂ r.antecedants){
add r to $R_{temp}$;
}
for each w  ∈ W{
score=0;
        for each rt ∈ $R_{temp}${
                if (rt.consequent ⊂ w.features){
                score+=rt.confidence;
                }
}
print w.UserID, score/$R_{temp}$.length
}
```

Figure 15. Pseudo Algorithm of Matching Engine

The full source-code of the program is available on the Appendix A section.

Running of The Algorithm

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 own matches. The test is repeated for both random male and female user. The user interface to control the matching program is shown in Figure 16.



Avarage Score for Matches: 0

Avarage Score for non-Matches 0

Completed Rows: 0

ManID 5734620

WomanID

Figure 16. User interface of the matching program

The selected male user with UserID 5734620 has the following features shown in Figure 17:

RC_MaritalStatus_Divorced, RC_Location_Istanbul,RC_Income_1500-2250, Age_Group_33-37,

BMI_Status_Normal, RC_PrefRel_LongTerm, RC_PrefRel_ShortTerm, RC_PrefRels_Friendship,

RC_HairColor_brown,RC_EyeColor_Green

Figure 17. Features of the test user

By running the match engine to get the best matches of the user, we obtain the following top list of best matches:

Table 27: 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

As seen in Table 27, the scores for the 10 best matches are mostly different from each other, so they create no tie condition for the rankings except the last three positions.

The best match with UserID 7735529 has the following features shown in Figure 18:

RC_Education_UnderGraduate, RC_MaritalStatus_Married but separate, RC_Location_Istanbul, RC_Income_1500-2250, Age_Group_33-37, BMI_Status_Normal, RC_WantsChildren_I dont know, RC_HasChildren_i have children, staying by me, RC_Drinking_Social Drinker, RC_Religion_not important, RC_Smoking_yes, RC_PrefRel_Marriage, RC_PrefRel_LongTerm, RC_PrefRel_ShortTerm, RC_PrefRels_Friendship, RC_HairColor_brown, RC_EyeColor_Hazel

Figure 18. Features of the best match of the test user

A subjective comparison of the two members concludes that the selected man and women 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 matches, another query is run for user 7735529. There are 136 records in the “couples” table for user 7735529, which means that 7735529 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”.

We let the matching engine to calculate the matching scores for these two groups separately and obtain the following results shown in Table 28:

Table 28. 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 which of them are real matches of the user 7735529.

In the next section this test will be applied to a sample set of multiple users to measure the significance of the matching engine’s results.

Testing of the Matching Engine

Testing Methodology

In our first design, 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 flagged as non-matches. The test will be run for different samples taken both from the dataset used for the building of the model and a new dataset derived from the 2009 data.

Significance of the Scoring Algorithm

Dataset

Two sets of sample data are prepared for the test. For the preparation of the first dataset, 15000 couples flagged as “matches” and 15000 couples flagged as “non-matches” are randomly selected from the list of couples used for the building of the model.

For the second dataset, the program used to extract matches from the messaging table is re-run on messaging data from January 2009 to April 2009. A new set of 30000 couples are constructed from the matching and non-matching couples by following the very same procedure used for the preparation of the first dataset.

If we examine the usage patterns of site members, we find that the average lifetime (from registration to leaving the site forever) is three months on average. There are doubtlessly frequent users of siberalem.com that stay as members for many years, as well as members who stay for a few days and leave. This phenomenon creates a problem for the evaluation of non-matching couples. The system marks a couple as non-matching whenever a reply message is not present. If the receiver of the message has already left the site permanently and gave up to check his/her message inbox, the sender may not receive a response, however much their matching scores may be. For the solution of the problem, for every member, two dates are queried form the data base: a starting date for the first appearance in the messaging table, and an end date for the last message they sent in the system. If the start and end dates of two non-matching members are not overlapping, we eliminate these non-matching couples from the sample. Also “non-matching” couples, one side of whom is a “non-responsive” member (he/she never sends a message), are eliminated during the above-described procedure. This control makes the test more related with real life conditions.

Calculating of the Scores

A new testing function is written and added to the matching program. When the test is started, the function loops through the 30000 couples in the dataset and computes the corresponding matching scores for every single couple.

As presented in the previous sectiont, the scoring algorithm produces different scores for the “Men looking for women” by using man’s rules and woman’s features and for the “Women looking for men” by using woman’s rules and man’s features. So, for any given couple two different scores are possible by considering the order of gender types.

The test is run twice, once for calculating Man → Woman scores and once for Woman → Man scores. To different outputs are produced in the form of the Table 29:

Table 29. Output of the test program

CoupleID	Match Status	Score
1	Match	12,23
2	Non-Match	8,2
...	...	...

Interpretation

The score variables for different groups of couples (matching and non-matching) are expected to differ in their means by favoring the mean of the matching couples. The descriptive statistical analyses of the two different groups approves the assumption by showing the mean of the score for the matching couples is 11% higher than the mean of the non-matching couples when Man → Woman scores are calculated as shown in Table 30.

Table 30. Statistics for the Man to Woman Scores using the 1. Dataset

	Match Status	N	Mean	Std. Deviation	Std. Error Mean
Score	Non-match	15000	12,499	4,729	0,0386155
	Match	15000	13,396	4,853	0,0396312

Further statistical testing is needed to tell exactly that the difference between the means is significant and it's not by chance. Student's t-test is used to compare the means of the independent samples of matching and non-matching groups.

We test the null hypothesis:

$$H_0 = \mu_{match} = \mu_{non-match}$$

We calculate the t-statistic for the two groups by using statistics software. The software produces the output as shown in Table 31.

Table 31: t-test Results for the Man to Woman Scores using the 1. Dataset

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Differenc	Lower	Upper
Score	Equal variances assumed	11.15	.001	-26.486	29998	.000	-1.4655805	.0553334	-1.5740363	-1.3571247
	Equal variances not assumed			-26.486	29988	.000	-1.4655805	.0553334	-1.5740363	-1.3571247

We reject the null hypothesis if $t < -t(^{\propto}/_{2}, v)$ or $t > t(^{\propto}/_{2}, v)$.

The t-statistic under the assumption of equal variances is -26,486. We reject the Null Hypothesis and conclude that the means of the two groups are not equal.

Similar results are obtained by testing the Woman → Man scores. The Table 32 and the Table 33 show the descriptive statistics and t-test results respectively.

Table 32. Statistics for the Woman to Man Scores using the 1**.** Dataset

	Match Status	N	Mean	Std. Deviation	Std. Error Mean
Score	Non-Match	15000	7.794	5.521	0,04502
	Match	15000	8.501	5.394	0,04402

Table 33. t-test Results for the Woman to Man Scores using the 1. Dataset

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Differenc	Lower	Upper
Score	Equal variances assumed	6.118	.013	-11.209	29998	.000	-0.706	.063	-0.829	-0.582
	Equal variances not assumed			-11.209	29988	.000	-0.706	.063	-0.829	-0.582

The significant difference between the means of the two groups implies the positive impact of matching engine recommendation on the potential matches.

Given a site member, following the recommendation of the engine and sending messages to recommended users increases the chance of a potential match in a significant way.

Until this point, the tests conducted with the Dataset 1 reflected the results obtained from the data used for the building of the model. The same tests are repeated on Dataset 2, to measure the impact of the matching engine on new members used the messaging system in the year 2009.

Table 34 shows the descriptive statistics of the Man → Woman scores for the second dataset. We observe again that the mean of the scores for matched couples are greater than the non-matches.

Table 34: Statistics for the Man to Woman Scores using the 2**.** Dataset

	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

To prove the statistical significance of the observation we test the null hypothesis:

$$H_0 = \mu_{match} = \mu_{non-match}$$

Table 35: t-test Results for the Man to Woman Scores using the 2. Dataset

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
score	Equal variances assumed	57.750	.000	-28.811	7998	.000	-2.6741464	.0928185	-2.8560948	-2.4921981
	Equal variances not assumed			-28.811	7.882E3	.000	-2.6741464	.0928185	-2.8560952	-2.4921976

We reject the null hypothesis if $t < -t(^{\propto}/_{2}, v)$ or $t > t(^{\propto}/_{2}, v)$.

According to Table 35, the t-statistic under the assumption of equal variances is -28,811. We reject the Null Hypothesis and conclude that the means of the two groups are not equal.

Similar results are obtained by testing the Woman → Man scores. The Table 36 and the Table 37 show the descriptive statistics and t-test results respectively.

Table 36. Statistics for the Woman to Man Scores using the 2**.** Dataset

	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

Table 37. t-test Results for the Woman to Man Scores using the 2. Dataset

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
score	Equal variances assumed	18.708	.000	-4.781	7998	.000	-.9351352	.1956021	-1.318566	-.5517042
	Equal variances not assumed			-4.781	7.970E3	.000	-.9351352	.1956021	-1.318566	-.5517040

As seen on all of the t-test results, the null hypothesis is rejected and the difference between the means of the scores of matching and non-matching couples is found significant. We can conclude that the recommendation engine produces higher scores for "matching" couples.

From the site user's perspective, this result can also be interpreted as following: "If a user follows the suggestions of the recommendation engine and send messages to that men or women on the list, the chance of a match is increases by 10 to 15%". The success rate of the engine may increase if input features, rule settings are well adjusted and feedback data is incorporated to the system.

CHAPTER V: CONCLUSION

In our study we were able to develop a new type of recommendation engine for online dating sites giving lists of men or women for the site user with higher potential to start a relationship. The rules used in the engine are extracted from the past data kept on the sites database and no additional expert knowledge is used. The overall performance of the engine is tested for the statistical significance and it's 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.com users) available for developing solutions for individuals' needs.

We made certain assumptions 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" which were able to start relationship according to definition. This classification methodology of "matching" and "non-matching" couples was the first finding of the study.

In the "data understanding" phase, the frequency distributions of specific features for the male and female sides of couples, flagged as "matches" are given. Commenting about the distribution tables revealed differences between male and female users, and also they gave a picture of the average man and woman who has been successful on the dating site.

We've introduced a less practiced way of extracting rules from data by using the Association Rule Mining functionality of the data mining. Experiments with different algorithms and settings have provided us an acceptable rule set in the sense of quality and quantity of the rules.

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.

Associative Rule Mining produced satisfying results on the siberalem.com’s database, which is promising also for studies on other social networks like circle-of-friend networks. Any other available data for the users feature sets such as personality traits, social contacts, preferences of digital content etc. may produce different rule sets including interesting 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 user. The recommendation engine should also be expanded to provide suggestion list for users of the same kind of gender. These issues can also be handled in further research.

APPENDICES

A. Source Code of the Matching Engine

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Data.OleDb;
using System.Collections;
using System.IO;
using System.Data.SqlClient;

namespace rule_engine
{
    class EngineManager
    {

        SqlConnection sqlconn = new SqlConnection("Data
Source=xxx.xxx.xxx.xxx;Initial Catalog=EBI;Integrated
Security=SSPI");
        const string connectionString =
"Provider=Microsoft.Jet.OLEDB.4.0;Data Source= c:\\engine1.mdb";
        OleDbConnection conn = new
OleDbConnection(connectionString);

        rules[] srulesm;
        rules[] srulesw;
        public double matchavg = 0;
        public double nonmatchavg = 0;
        public int totaltestcount = 0;
        public struct rules
        {
            public double confidence;
            public double lift;
            public double support;
            public int[] ante;
            public int cons;
        }

        Hashtable manfeatures = new Hashtable();
        Hashtable womanfeatures = new Hashtable();

        public void FindSharedLives()
        {
            string SQL = "select ManID,WomanID from Couples where
SharedLifeTime is null";
            OleDbCommand cmd2 = new OleDbCommand(SQL);
            cmd2.Connection = conn;
            conn.Open();
            OleDbDataReader reader2 = cmd2.ExecuteReader();
            while (reader2.Read())
            {
                int result = 0;
                double ManID = reader2.GetDouble(0);
                double WomanID = reader2.GetDouble(1);

                    string sqltest="select max (MessageID), min
(MessageID) from (select * from Message_Main where FromUyeID=" +
ManID.ToString() + ") a";
```

```
                    SqlCommand sqlcmd = new SqlCommand(sqltest);
                    sqlcmd.Connection = sqlconn;
                    sqlconn.Open();


                        SqlDataReader rdr = sqlcmd.ExecuteReader();
                    rdr.Read();
                    int MaxMessageM = 0;
                    int MinMessageM = 0;
                    try
                    {
                        MaxMessageM = rdr.GetInt32(0);
                        MinMessageM = rdr.GetInt32(1);
                    }
                    catch
                    {
                        MaxMessageM = 0;
                        MinMessageM = 0;
                    }
                    rdr.Close();
                    SqlCommand sqlcmd2 = new SqlCommand("select max
(MessageID), min (MessageID) from (select * from Message_Main where
FromUyeID=" + WomanID.ToString() + ") a");
                    sqlcmd2.Connection = sqlconn;
                    SqlDataReader rdr2 = sqlcmd2.ExecuteReader();

                    rdr2.Read();
                    int MaxMessageW = 0;
                    int MinMessageW = 0;
                    try
                    {
                        MaxMessageW = rdr2.GetInt32(0);
                        MinMessageW = rdr2.GetInt32(1);
                    }
                    catch
                    {

                        MaxMessageW = 0;
                        MinMessageW = 0;
                    }
                    rdr2.Close();
                    sqlconn.Close();


                    if (MinMessageM > 0 && MinMessageW > 0)
                    {
                        if (MinMessageM <= MinMessageW)
                        {
                            if (MinMessageW <= MaxMessageM)
                            {
                                result = 1;
                            }
                        }
                        if (MinMessageW <= MinMessageM)
                            if (MinMessageM <= MaxMessageW)
                            {
                                result = 1;
                            }
                    }
```

```
                string SQL3 = "Update Couples set 
SharedLifeTime="+result.ToString()+" where 
WomanID="+WomanID.ToString()+" and ManID="+ManID.ToString();
                OleDbCommand cmd3 = new OleDbCommand(SQL3);
                cmd3.Connection = conn;
                cmd3.ExecuteNonQuery();


            }


        }
        public int LoadRulesM()
        {

            string SQL = "select * from Rules where 
AntecedantGenderID=2";
            string SQL2 = "select count(1) from Rules where 
AntecedantGenderID=2";
            int i = 0;
            OleDbCommand cmd = new OleDbCommand(SQL2);
            cmd.Connection = conn;
            OleDbCommand cmd2 = new OleDbCommand(SQL);
            cmd2.Connection = conn;

            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();
            reader.Read();
            int rulecount = reader.GetInt32(0);
            srulesm = new rules[rulecount];

            OleDbDataReader reader2 = cmd2.ExecuteReader();
            while (reader2.Read())
            {

                string antestring = reader2.GetString(1);
                string[] antearray = antestring.Split(';');
                int[] intValues = new int[antearray.Length];

                for (int n = 0; n < antearray.Length; n++)
                {

                    intValues[n] = Convert.ToInt32(antearray[n]);

                }
                srulesm[i].ante = intValues;
                string consstring = reader2.GetString(2);
                srulesm[i].cons = Convert.ToInt32(consstring);
                srulesm[i].support = reader2.GetDouble(3);
                srulesm[i].confidence = reader2.GetDouble(4);
                srulesm[i].lift = reader2.GetDouble(5);
                i++;
            }

            conn.Close();
            return rulecount;
        }



        public int LoadRulesW()
        {

            string SQL = "select * from Rules where 
AntecedantGenderID=1";
            string SQL2 = "select count(1) from Rules where 
AntecedantGenderID=1";
            int i = 0;
            OleDbCommand cmd = new OleDbCommand(SQL2);
            cmd.Connection = conn;
            OleDbCommand cmd2 = new OleDbCommand(SQL);
            cmd2.Connection = conn;

            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();
            reader.Read();
            int rulecount = reader.GetInt32(0);
            srulesw = new rules[rulecount];

            OleDbDataReader reader2 = cmd2.ExecuteReader();
            while (reader2.Read())
            {

                string antestring = reader2.GetString(1);
                string[] antearray = antestring.Split(';');
                int[] intValues = new int[antearray.Length];

                for (int n = 0; n < antearray.Length; n++)
                {

                    intValues[n] = Convert.ToInt32(antearray[n]);

                }
                srulesw[i].ante = intValues;
                string consstring = reader2.GetString(2);
                srulesw[i].cons = Convert.ToInt32(consstring);
                srulesw[i].support = reader2.GetDouble(3);
                srulesw[i].confidence = reader2.GetDouble(4);
                srulesw[i].lift = reader2.GetDouble(5);
                i++;
            }

            conn.Close();
            return rulecount;
        }
        public int LoadMen()
        {

            string SQL = "select  UserID,Features from Features 
where GenderID=2";
            string SQL2 = "select  count(1) from Features where 
GenderID=2";

            OleDbCommand cmd = new OleDbCommand(SQL2);
            cmd.Connection = conn;
            OleDbCommand cmd2 = new OleDbCommand(SQL);
            cmd2.Connection = conn;

            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();
            reader.Read();
            int mancount = reader.GetInt32(0);
            OleDbDataReader reader2 = cmd2.ExecuteReader();
            while (reader2.Read())
            {
                int ID = reader2.GetInt32(0);
                string features = reader2.GetString(1);
                manfeatures.Add(ID, features);
            }

            conn.Close();
            return mancount;
        }

        public int LoadWomen()
        {

            string SQL = "select  UserID,Features from Features
where GenderID=1";

            string SQL2 = "select count(1) from Features where
GenderID=1";


            OleDbCommand cmd = new OleDbCommand(SQL2);
            cmd.Connection = conn;
            OleDbCommand cmd2 = new OleDbCommand(SQL);
            cmd2.Connection = conn;

            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();
            reader.Read();
            int womancount = reader.GetInt32(0);
            OleDbDataReader reader2 = cmd2.ExecuteReader();
            while (reader2.Read())
            {
                int ID = reader2.GetInt32(0);
                string features = reader2.GetString(1);
                womanfeatures.Add(ID, features);
            }

            conn.Close();
            return womancount;
        }

        public  double[] GetMansBestMatches(int ID)
        {

            rules[] manrules = getMansRules(ID);

            double[] scorelist=new double[womanfeatures.Count];
            int[] womanx=new int[womanfeatures.Count];
            int n=0;
            foreach (int WomenID in womanfeatures.Keys)
              {
            double score=calculateMtoWscore(ID, WomenID,manrules);

            scorelist[n] = score;
                );
            n++;
              }


           Array.Sort(scorelist);
           Array.Reverse(scorelist);
           return scorelist;

        }

        public bool isContacted (int ManID, int WomanID)
        {
            string SQL2 = "select  from couples where FromUyeID=" +
ManID.ToString() + " and ToUyeID=" + WomanID.ToString();


            OleDbCommand cmd = new OleDbCommand(SQL2);
            cmd.Connection = conn;


            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();
            reader.Read();
            int womancount = reader.GetInt32(0);

            return true;
        }

        public double[] GetWomansBestMatches(int ID)
        {

            rules[] womanrules = getWomansRules(ID);

            double[] scorelist = new double[manfeatures.Count];
            int n = 0;
            foreach (int ManID in manfeatures.Keys)
            {
                double score = calculateWtoMscore(ManID, ID,
womanrules);
                scorelist[n] = score;
                n++;
            }

            Array.Sort(scorelist);
            Array.Reverse(scorelist);
            return scorelist;

        }




        public rules[] getMansRules(int ID)
        {


            string rulestring = manfeatures[ID].ToString();
            string[] rulesarray = rulestring.Split(';');
            int[] features = new int[rulesarray.Length];

            for (int n = 0; n < rulesarray.Length; n++)
            {
                if (rulesarray[n] != "")
                {
                    features[n] = Convert.ToInt32(rulesarray[n]);
                }

            }

            ArrayList matchingrules = new ArrayList();
            for (int n = 0; n < srulesm.Length; n++)
            {

                bool isSubset =
!srulesm[n].ante.Except(features).Any();

                if (isSubset)
                {
                    matchingrules.Add(srulesm[n]);
                }

            }
            rules[] amatchingrules=new rules[matchingrules.Count];

            int m = 0;
            foreach (rules matchingrule in matchingrules)
            {
                amatchingrules[m] = matchingrule;
                m++;
            }
            return amatchingrules;


        }

        public rules[] getWomansRules(int ID)
        {


            string rulestring = womanfeatures[ID].ToString();
            string[] rulesarray = rulestring.Split(';');
            int[] features = new int[rulesarray.Length];

            for (int n = 0; n < rulesarray.Length; n++)
            {
                if (rulesarray[n] != "")
                {
                    features[n] = Convert.ToInt32(rulesarray[n]);
                }

            }

            ArrayList matchingrules = new ArrayList();
            for (int n = 0; n < srulesm.Length; n++)
            {

                bool isSubset =
!srulesw[n].ante.Except(features).Any();

                if (isSubset)
                {
                    matchingrules.Add(srulesw[n]);
                }

            }
            rules[] amatchingrules = new rules[matchingrules.Count];

            int m = 0;
            foreach (rules matchingrule in matchingrules)
            {
                amatchingrules[m] = matchingrule;
                m++;
            }
            return amatchingrules;


        }

        public int[] getWomanFeatures(int ID)
        {

            string rulestring = womanfeatures[ID].ToString();
            string[] rulesarray = rulestring.Split(';');
            int[] features = new int[rulesarray.Length];


            for (int n = 0; n < rulesarray.Length; n++)
            {
                if (rulesarray[n] != "")
                {
                    features[n] = Convert.ToInt32(rulesarray[n]);
                }

            }
            return features;

        }


        public int[] getManFeatures(int ID)
        {

            string rulestring = manfeatures[ID].ToString();
            string[] rulesarray = rulestring.Split(';');
            int[] features = new int[rulesarray.Length];


            for (int n = 0; n < rulesarray.Length; n++)
            {
                if (rulesarray[n] != "")
                {
                    features[n] = Convert.ToInt32(rulesarray[n]);
                }

            }
            return features;

        }


        public double calculateMtoWscore(int ManID, int
WomanID,rules[] manrules)
        {
            double score=0;

            int[] wfeatures = getWomanFeatures(WomanID);

            for (int n = 0; n < manrules.Length; n++)
            {

                if (wfeatures.Contains(manrules[n].cons))
                {
                    score = score + manrules[n].confidence;
                }

            }
            double finalscore = score/manrules.Length;
            return finalscore;
        }

        public double calculateWtoMscore(int ManID, int WomanID,
rules[] womanrules)
        {
            double score = 0;

            int[] wfeatures = getManFeatures(ManID);

            for (int n = 0; n < womanrules.Length; n++)
            {

                if (wfeatures.Contains(womanrules[n].cons))
                {
                    score = score + womanrules[n].confidence;
                }

            }
            double finalscore = score / womanrules.Length;
            return finalscore;
        }



        public void MassCalculate(int param)
        {

            string SQL = "select
ID,ManID,WomanID,ReceivedMessages,SentMessages,MailExchange,SharedLi
feTime from couples";

            OleDbCommand cmd = new OleDbCommand(SQL);
            cmd.Connection = conn;
            conn.Open();

            OleDbDataReader reader = cmd.ExecuteReader();

            int matchcount = 0;
            int nonmatchcount = 0;

            totaltestcount = 0;
            while (reader.Read())
            {
                int ID = reader.GetInt32(0);
                double dManID = reader.GetDouble(1);
                double dWomanID = reader.GetDouble(2);
                double dReceivedMessages = reader.GetDouble(3);
                double dSentMessages = reader.GetDouble(4);
                double dMailExchange = reader.GetDouble(5);
                int SharedLifeTime = reader.GetInt16(6);
                int WomanID = Convert.ToInt32(dWomanID);
                int ManID = Convert.ToInt32(dManID);

                int ReceivedMessages =
Convert.ToInt32(dReceivedMessages);
                int SentMessages = Convert.ToInt32(dSentMessages);
                int MailExchange = Convert.ToInt32(dMailExchange);


                double score = 0;

                if (param == 1)
                {
                    rules[] womanrules = getWomansRules(WomanID);
                     score = calculateWtoMscore(ManID, WomanID,
womanrules);
                }
                else if (param == 2)
                {

                    rules[] manrules = getMansRules(ManID);
                     score=calculateMtoWscore(ManID,
WomanID,manrules);
                }




                if (ReceivedMessages > 1 && SentMessages > 1 &&
MailExchange == 1)
                {

                    double matchtotal = matchavg * matchcount;
                    matchtotal = matchtotal + score;
                    matchcount++;
                    matchavg = matchtotal / matchcount;


                }
                if (ReceivedMessages <= 1 && SentMessages <= 1 &&
MailExchange == 0 && SharedLifeTime==1)

                {
                    double nonmatchtotal = nonmatchavg *
nonmatchcount;

                    nonmatchtotal = nonmatchtotal + score;
                    nonmatchcount++;
                    nonmatchavg = nonmatchtotal / nonmatchcount;
                }

                totaltestcount++;

            }

            conn.Close();

        }


    }
}
```

B. Contents of the CD

dataset.mdb	MS Access 2007 database of the ruleset, matching / non-matching couples and their features. Both 2007 and 2009 data for the couples is included.
/Matching Engine	.NET Project folder of the matching engine.
/Match Finder	.NET Project folder of the match finder program.

REFERENCES

A.C. Nielsen. (2007). Siberalem.com Marka Bilinirlik Anketi.

Aberle, D. F. (1950). The Functional Prerequisites of a Society. *Ethics 60* , 100-111.

Adomavicius, G., & Tuzhilin, A. (2005, June). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering 17* , pp. 734–749.

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

Belot, M., & Francesconi, M. (2007). *Can Anyone Be "The" One? Field Evidence on Dating.* The Institute for Social and Economic Research.

Bishop, J. (2009). *Understanding and Faciliatating the Development of Social Networks in Online Dating Communities: A Case Study and Model.* Retrieved March 2009, from http://www.jonathanbishop.com.

Brym, R. R., & Lenton, R. L. (2001). *Love Online: A Report on Digital Dating in Canada.* Toronto : msn.ca.

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

Christian Borgelt, R. K. (2002). Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics* . Berlin.

CRoss Industry Standard Process. (n.d.). Retrieved 2008, from http://www.crisp-dm.org/

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

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1997). From Data Mining to Knowledge Discovery in Databases. *AI Magazine* , pp. 37-54.

Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Royal Statist. Soc 85* , pp. 597–612.

Fisman, R., Iyengar, 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. (1962, January). College Admissions and the Stability of Marriage. *The American Mathematical Monthly, Vol. 69, No. 1* , pp. 9-15.

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

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, (pp. 1-12). Dallas.

Hancock, T. J., Thom-Santelli, J., & Ritchie, T. (2004). Deception and design: the impact of communication technology on lying behavior. *Conference on Human Factors in Computing Systems*, (pp. 129-134). New York.

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

Hitwise Inc. (2004). *Hitwise Internet Dating Rankings.*

Hitwise Inc. (2008). *Hitwise Internet Dating Ratings.*

Internet World Stats. (2008). *Internet World Stats*. Retrieved from http://www.internetworldstats.com/stats.htm

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. In *Prentice-Hall Advanced Reference Series.*

Kalita Araştırma. (2009). *Üye Memnuniyeti ve Beklenti Araştırması.* İstanbul.

Microsoft. (2008). *Microsoft Advertising - Turkey.*

Online Publishers Association & comScore Network. (2005). *Paid Online Content U.S. Market Spending Report.*

Rappa, M. (2001). Business Models on the Web.

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

Roth, A. E., & Erev, I. (1995). 8,) Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Journal of Economic Literature* , pp. 164-212.

Sanver, İ., & Sanver, R. (2005). Implementing matching rules by type pretension mechanisms. *Mathematical Social Science* .

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

SPSS Inc. (n.d.). Clementine 12.0 User Manual.

Varian, P., & Varian, H. R. (2003). How Much Information.

Vasilis Aggelis, D. C. (2004). e-Trans Association Rules for e-Banking Transactions . IV International Conference on Decision Support for Telecommunications and Information .

Weiss, S., & Kulikowski, C. (1991). *Computer systems that learn: classification and prediction methods from statistics.* Morgan Kaufmann Publishers.

World Health Organization. (1995). *Physical Status: The Use and Interpretation of Antrophometry.* Geneva.