ISSN:1306-3111
e-Journal of New World Sciences Academy
2009, Volume: 4, Number: 3, Article Number: 1A0025

TECHNOLOGICAL APPLIED SCIENCES
Received: October 2008
Accepted: June 2009
Series : 2A
ISSN : 1308-7223
© 2009 www.newwsa.com

Özcan Özyurt
Hacer Özyurt
Besikduzu Vocational School,
Karadeniz Technical University
oozyurt@ktu.edu.tr
Trabzon-Turkey

A STUDY ON LOG BASED INFORMATION RETRIEVAL SYSTEM FOR SOCIAL AND SEMANTIC EXTRACTIONS FROM CHAT MEDIUMS

ABSTRACT

Chat mediums are widely-used as a communication tool nowadays. Contents of chat conversations may be shaped by sex, habit, social behaviors and tendency of the people. In this study, we have presented a log based information retrieval system that is designed to identify the sex of a person in Turkish chat mediums. Here, the sex identification is taken as a base study in the information mining in chat mediums. The proposed sex identification method is compared with the Support Vector Machine and Naïve Bayes methods. The proposed system has achieved accuracy about 90% in the sex identification in the real chat mediums.

Keywords: Information Retrieval, Sex Identification,
Turkish Chat Mediums, Text Mining, Machine Learning

SOHBET ORTAMLARINDAN SOSYAL VE ANLAMSAL ÇIKARIMLAR İÇİN KAYIT TABANLI BİLGİ ÇIKARIM SİSTEMİ ÜZERİNE BİR ÇALIŞMA

ÖZET

Sohbet ortamları günümüzde iletişim aracı olarak yaygın bir biçimde kullanılmaktadır. Sohbet ortamlarındaki konuşmaların içerikleri kişinin cinsiyetine, alışkanlıklarına, sosyal davranış ve eğilimlerine göre şekillenebilmektedir. Bu çalışmada, Türkçe sohbet ortamlarından konuşmacıların cinsiyetlerinin belirlenmesine yönelik kayıt tabanlı bir bilgi çıkarım sistemi ortaya konulmuştur. Burada, cinsiyet belirleme sohbet ortamlarında bilgi madenciliği uygulaması olarak temel alınmıştır. Önerilen cinsiyet belirleme yöntemi veri madenciliği yöntemlerinden SVM ve Naive Bayes sonuçları ile karşılaştırılmıştır. Geliştirilen sistem, gerçek sohbet ortamlarında cinsiyet belirlemede %90'a yakın sonuçlara ulaşmıştır.

Anahtar Kelimeler: Bilgi Çıkarımı, Cinsiyet Belirleme,
Türkçe Sohbet Ortamları, Metin Madenciliği,
Makine Öğrenmesi

1. INTRODUCTION (GİRİŞ)

A chat medium contains a vast amount of information, which is potentially relevant to a society's current interests, habits, social behaviours, crime tendency and other tendencies [1, 2, 3 and 4]. Users may spend a large portion of their time to find out information in chat mediums. An intelligent system may help the users in finding the interested information in the medium [1, 3 and 4]. One of our major targets is to develop a system that automatically determines persons sex identities in chat mediums.

In a conversation, chatters consider the corresponding chatter's sex, and the course and contents of the conversation may be shaped according to the corresponding persons' sexual identity. Therefore, a sample identification system is implemented to determine chatter's sex identity in chat mediums. To do this, many conversations are acquired from a chat medium designed on purpose, and then statistical results are derived from the conversations [1, 2, 3, 4, 5 and 6]. Topics of these conversations are not predetermined and may be about any subjects. These results are used to determine weighting coefficients of the proposed discrimination function.

When we take into consideration of the rapid growth of the internet, one of the most important applications of machine learning techniques is the usefulness of text categorization in real life [2, 7, 8, 9 and 10]. Here, assigning predefined categories to natural language text documents based on their contents is the problem of text categorization [5, 7, 11 and 12]. Manual classifications of natural language texts documents are slow, time consuming and tedious work. Because of this, developing fast and efficient automatic techniques is necessary to classify the bulk amount of texts such as emails, chat conversations and documents in the web mediums [2, 10, 13 and 14].

In the literature, most of the studies are concentrated on topic categorizations rather the sex-identification. An offline topic categorization approach for analyzing chat conversation logs related to criminal activities is presented [2]. Here, logs are first pre-processed with stop-word removal and converted into term frequency weighted vectors. Then, categorization techniques including k-NN, Naïve Bayes and linear SVM are employed for topic classification. Another categorization approach for analyzing chat messages from Internet Relay Chat is also adapted to classify chat conversations [10]. In this work, the chat messages are filtered based on time, chat room channel or chat message authors and resulting collections of chat messages are grouped as "sessions" for categorization. Other Independent Component Analysis (ICA) for chat room topic detection is applied for chat message categorization [15]. A similar approach using Complexity Pursuit instead of ICA is also proposed for chat room topic categorization [16].

The rest of this paper is organized as follows. Research significance is given in Section 2. Characteristics of Turkish language are given in Section 3. Collection and analysis of data are given in Section 4. In Section 5, details of log based information retrieval system are given. The implementation and results are discussed in Section 6. The conclusion and future work are given in Section 7.

2. RESEARCH SIGNIFICANCE (ÇALIŞMANIN ÖNEMİ)

In a chat conversation, chatter considers the corresponding chatter's sex, and the course and contents of the conversation may be shaped according to the corresponding persons' sexual identity. The other word, chatter's sex is emphases to conversations' contents mostly. Researchers may spend a large portion of their time to find out information in chat mediums. Manual classifications of large conversations texts are slow and tiring. Hence, an intelligent system may help the researchers in finding the interested information in the mediums. For this purpose, an example information retrieval system is designed to determine chatter's sex identity in a chat medium in this study. To do this, many conversations are acquired from a chat mediums designed on purpose, and then statistical and semantically results are obtained from the conversations. The obtained results are used to determine weighting coefficients of the proposed discrimination function and to evaluate the statistical and semantic analysis.

Many research projects on intelligent systems have been launched to collect and retrieve information from different mediums such as Internet and especially chat, and other communication mediums. These systems posses a common core approaches to learn a probabilistic profile of chatters in the mediums, and then they use the profile to find, or classify the information collected from the mediums. Here, our main goal is to develop a foundation for an intelligent system that collects and uses information obtained from conversations in Turkish chat mediums. In other words, we propose a machine learning system on behalf of users that retrieve information from conversations in chat mediums. This system obtains the information from conversations in a chat medium, so it will assist users collecting information from the medium. Thus, understanding of a natural language by a machine comprises the process of transforming a sentence into an internal representation. So, the system concludes decisions and relationships according to the resulting representations by employing the process of making out the meaning of sentences in natural language.

3. CHARACTERISTICS OF TURKİSH LANGUAGE (TÜRKÇE'NİN KARAKTERİSTİKLERİ)

World Languages may be classified according to their structures and origins. Ural-Altaic languages are one of the most commonly spoken languages in the world [17]. Turkic languages that belong to the group of Ural-Altaic languages form a subfamily of the Altaic languages, and Turkish is an Oghuz oriented member of Turkic languages family. Turkish language has several noticeable characteristics. One of them is that Turkish is an agglutinative language with respect to word structures formed by productive affixations of derivational and inflectional suffixes to root words [17 and 18]. Therefore, many different words can be constructed from roots and phonemes to describe objects and concepts. For example, the word kitaplarımda (in/at my books) is composed of kitap (book), lar = plural suffix, ım = possessive suffix, and "da" = locative suffix.

Another characteristic of the Turkish language is the vowel and consonant harmony. Turkish languages have two kinds of vowel harmony. The vowels in Turkish language are grouped as front and back vowels. The front vowels are produced at the front of the mouth and back vowels are produced at the back of the mouth. Turkish words can contain only one kind of vowel and all suffixes added to the word must conform to the vowel of the syllable preceding them [17]. Turkish language has also consonant harmony as well. If a morpheme follows a word ending in constants k, p, t or ç, the last constant of the word changes to constants g-ğ, b, d or c respectively. Most of the world languages have three grammatical genders (masculine, feminine, and neuter) but Turkish language has only one (neuter). Here, some words in English may expose the sex of the chatter but they don't expose in Turkish. For example, "Ben onun kardeşiyim" in Turkish may be translated into English as "I am his brother/sister". Therefore, the identification of chatters' sex from a conversation in Turkish may be more difficult than in English.

Basically, our identification system takes a sentence from a chat session and analyses all the words morphologically. The sequence of morphemes, appearing in a word, is determined by morphotactics of the language [17 and 18]. Our morphological analyzer recognizes eight types of phonemes; punctuation, possessives, proper names, short forms, quoted items, roots, words and suffixes. Punctuations are symbols such as comma, full stop, question mark, and semi-colon. The roots and words may be nouns, verbs, pronouns, adjectives, adverbs, prepositions, conjunctions, numerals, and interjections. The words are grouped as noun, verbs, adjectives, simple numbers, pronouns and connectives.

4. COLLECTION AND ANALYSIS OF DATA (VERİLERİN TOPLANMASI VE ANALİZİ)

For the evaluation of chat conversations, data were gathered from Specially Designed Chat Medium (SDCM) log files and mIRC (Shareware Microsoft Internet Relay Chat) which are widely used at internet. The gathered data are text files in which conversations are kept. The data existing in these files were subjected to pre-treatment, and were prepared for data mining. In the pre-treatment of the data, basic steps of the data mining were taken into consideration. Total size of the conversations is 4.7 Mb. The conversation which has the shortest duration among the conversations is 1 minute while the longest one is 155 minutes. 154 conversations were gathered from these mediums and 75 of them were used as training sets. The rest 79 conversations were used as test data sets. Main features of the conversations are presented in Table 1.

Table 1. Information of conversations gathered from chat mediums
(Tablo 1. Sohbet ortamlarından elde edilen bilgiler)

Statistical Information	Numerical Values
Total number of conversations	154
Total number of words used in the conversations	24993
Total number of most frequently used twenty words	4195
Words with spelling error	2454
Proportion of number of words with spelling error to the total number of the words	9.8%
The number of Acronyms, short forms and icons	2167
Proportion of the number of most frequently used 20 words to the total number of the words	16.8%
Proportion of the number of most frequently used 50 words to the total number of words	31.4%
Proportion of the number of most frequently used 100 words to the total number of words	44.3%

Statistical information such as total number of words or number of words with spelling errors is also given in Table 1. Besides a list of the total number of words and of most frequently used words in the conversations was made. The proportion of most frequently used 20, 50 and 100 words to total number of words were found as 16.8%, 31.4%, and 44.3% respectively. This demonstrates that specific words are frequently used in conversations.

In accordance with the gathered data, nearly half of the words used in conversations consist of most frequently used 100 words. Considering the total number of the words as 24993, it is seen that almost half of the conversation is made up of the same words. This demonstrates that specific words are frequently used in chat mediums. Another conspicuous situation in chat conversations is proportion of words with spelling errors to the total number of words. As can be read from the table also, the number of words with spelling errors consist 9.8% of the total words. While the number of the words with spelling errors was calculated, exceptions like acronyms, short forms and icon numbers were not taken into considered.

When number of these words is taken into consideration, it is seen that 19% of the words used in conversations were written incorrectly or deficiently. This means that one out of each five words is spelled with errors. When this information was taken into account, it is easily seen that chat conversations bear far much differences from news or normally written texts.

5. DETAILS OF LOG BASED INFORMATION RETRIEVAL SYSTEM (KAYIT TABANLI BİLGİ ÇIKARIM SİSTEMİN AYRINTILARI)

To evaluate the identification system, real information is collected and extracted from chat mediums (mIRC). Also some statistical data collected from the specially designed chat medium (SDCM) is used to evaluate the discrimination function. Weighting coefficients of each word of the function are determined by considering the statistical information. In practice, the usage frequencies of the words may rise dramatically if the number of chatters and duration of the conversation increase. Therefore, a normalization process is applied and the normalized values are used in the identification function. The most frequently used signs obtained from the SDCM, and the mIRC or real Internet medium (RIM) are also used to evaluate the system.

- **Chat Language (Chat Dili):** In real-time and informal environment of IM (Instant Messages) systems, chat messages are very different from conventional text. Therefore, chat language includes acronyms, short forms, polysemes, synonyms and mis-spelling of terms [12]. However, it is possible to find mistakenly written words and irregular short forms apart from formal grammatical rules in the texts. Special expressions and spelling errors which are commonly encountered in chat conversations can be grouped as follows.

  Acronyms are formed by extracting the first letters of a sequence of words. For example, "KIB" is an acronym for "Kendine İyi Bak (Take care)", "SÇS" is an acronym for "Seni Çok Seviyorum (I love you)" and "AEO" is an acronym for "Allah'a Emanet Ol (God be with you)".

Table 2. Acronym and short form examples used in conversations
(Tablo 2. Konuşmalarda geçen sözcüklerin baş harflerinde oluşan sözcük ve kısaltma örnekleri)

Acronyms	Meaning	Short forms	Meaning
KİB	Take care	tmm	OK
AEO	God be with you	tlf	Phone
SÇS	I love you	üniv	University
ARO	God bestow mercy upon you	inş	If God wills
SG	See you later	cvp	Response

Short forms refer to the case in which a lengthy word is replaced with a shorter alternative expression. For example, tmm is a short form for "tamam (okey)", tşk is a short form for "teşekkür ederim (thank you)". Unlike acronyms, it is observed that only some popular short forms have fixed expressions among different chat participants.

Many short forms are highly subjective to the context of the conversation and chatters. Table 2 shows some example short forms, and some of the most popular acronyms.

Icons are used in conversations, such as :), :))))))), :)) (Laughing), ?, ?-, ????, _?, …!? (Question and asking other meanings?), :P, :PPP, :p (To show tongue), :(, :((, :((((Unhappiness).Some of these icons mean same though their spellings are different. Some icons used in conversations are listed in Table 3.

Table 3. Samples from signs used in conversations
(Tablo 3. Konuşmalarda geçen işaretlere örnekler)

Signs	Meaning
:), :)), :)))), :-)),	Laughing
:D, :d	Laughing loudly
?, ?-, ????, _?, …!?	Question and asking other meanings?
:P, :PPP, :p	To show tongue
;), ;)), ;)))	To blink
:(, :((, :(((	Unhappiness
:-), :=), :-)))	Laughing

Mis-spelling of terms is seen more frequently in chat conversations than formal text documents due to nature of chat document. There are also some cases in which a chat participant purposely mis-spells a word to emphasize its meaning. A common case for mis-spelling is the use of duplicated vowels, such as "evetttttt", "yawwwwww" and "okkkk" instead of "evet (yes)", "yahu (Hey!)" and "okey (okey)" respectively. The number of duplications is not fixed.

- **Classifying Words and Word Groups (Kelimelerin Sınıflanması ve Kelime Grupları):** In a chat medium, many word groups may be defined to identify chatter's sex in a dialogue. In this study, eighth word groups are defined to cover as many sex related concepts and subjects as possible in a chat medium [3 and 4]. Word groups are built by considering the conceptual relations and usage frequency of the words used by female or male chatters. These groups are abbreviations and signs, slang and jargon words, politeness and delicacy words, interjections and shouting, sex and age related words, question words, particle and conjunction words, and other words group. These groups and some important words in the groups are listed in Table 4. The weighting coefficients of each word group are determined considering the usage frequencies and determinative power of words in each group [3, 4 and 6].

Table 4. Some most frequently used words in each word group
(Tablo 4. Her grupta en sık kullanılan kelimelerden bazıları)

No	Abbreviation and signs	Slang and jargon words	Politeness delicacy words
1	Hi (Slm)	My son! (Oglum)	Nice (Güzel)
2	Answer (Cvp)	Man! (Lan)	Thanks (Tşk)
3	What is the news (Nbr)	Uncle! (Dayı!)	Well done (Aferin)
4	You! (u)	Go away! (Defol)	Yes! (Efendim)
5	Thank you (tşk)	Repentance! (Tövbe!)	You (Siz)

No	Interjections Shouting words	Age and sexuality related words	Question words
1	Hey!/Man! (Yaw)	Age (Yaş)	What(for)?(Niye?)
2	Hmm (Hımm)	Sexuality(Cinsiyet)	Why? (Neden?)
3	And, soo (Ee)	My love (Aşkım)	Which? (Hangi?)
4	Oh! (Aa)	My lady (Bayanım)	Where? (Nerde?)
5	Well (İi)	My man/gent. (Erkeğim)	Where are you? (nerdesin?)

No	Particle and conjunction words	Other words	
1	Such/so/that (Öyle)	You (Sen)	
2	If not/otherwise (Yoksa)	I/me (Ben)	
3	In order to (Diye)	If only.. (Olsun)	
4	Another /Other/ (Başka)	You (Seni)	
5	Thus/so/such (Böyle)	Look (Bak)	

Due to nature of chats, contents of conversations vary. Hence, the feature vectors and words in the conversations also chances dynamically. Therefore, new words are added to each word group when a new male or female dominant word is accounted in a conversation. Here, new words are extracted from each conversation under examination. If the identification result is neutral, new words are not taken into consideration. Otherwise new words are taken into consideration according to the identification results. In the first step, these new words are just registered into a transient database. Then, each new word is added to the decision word groups at the lowest dominance level when it is encountered in several other conversations and confirming the same sexual identity. In the long term, the dominance level of a word in a decision word group may be increased if it is more often encountered in conversations relatively.

- **Discirimination Function (Ayrım Fonksiyonu):** A simple discrimination function is designed to identify sex of a person in a chat medium. This function considers each word in conversations separately and collectively. Therefore, statistical information related to each chosen word is collected from the SDCM and Internet chat mediums [3 and 4]. By using the statistical information, a weighting coefficient is determined for each word in each word group. Practically, a normalized (into the interval from 0.0 to 1.0) weighting coefficient of each word is determined. For each conceptually related word group, a sexual identity value is calculated by equation (1).

$$g_i = (\alpha_{i1}w_1 + \alpha_{i2}w_2 +,...,+\alpha_{ik}w_k)/\beta_i. \quad (1)$$

$$\beta_i = \alpha_{i1} + \alpha_{i2} +,..+\alpha_{ik}. \quad (2)$$

where, gi varies from 0.0 to 1.0 and determines the chatters' sexual identities as female or male for ith word group, α_{ij} is the weighting coefficient of jth word in ith word group and varies related to the number of words in the interested text, w_j represent the existing jth words in the interested text (if a word exists in the text, then w_j=1.0 else w_j= 0.0), k is the number of word in ith word group and β_i is normalization divider for the current number of existing words in the ith word group and calculated by equation (2). If a word is female dominant, α varies from 0 to 0.5, but if the word is male dominant, α varies from 0.5 to 1.0.

As explained before, words are also classified in several groups according to their conceptual relations. Thus, the importance of some word groups can be emphasized collectively. So, several word groups are defined by considering words acquired from the conversations in the chat mediums. A weighting coefficient is also determined for each word group. Then, the proposed discrimination function is formed for the sex identification as Equation (3). The equation can be used to determine sex identity of any chatters in a conversation.

$$\gamma = (\lambda_{g1} * g_1 + \lambda_{g2} * g_2 +,...,+\lambda_{gn} * g_n)/\theta. \quad (3)$$

where, γ varies from 0 to 1 and determines the chatters sexual identity as female or male, λ_{gi} is the weighting coefficient for ith female or male word group and θ is normalization divider for the current number of existing groups in a conversation and it is calculated by equation (4).

$$\theta = \lambda_{g1} + \lambda_{g2} +,...+\lambda_{gn}. \quad (4)$$

where, λ_{gi} is the weighting coefficient of ith groups. Hence, the weighting coefficients of each group are determined according to dominant sexual identity of the group. Then, the sex of the chatters may be identified as female when γ is determined between 0.0 and 0.5. On the other hand, chatters may be identified as male when γ is determined between 0.5 and 1.0. Here, the accuracy of the results increases that it shows female or male gender when γ approaches to 0.0 and 1.0 respectfully.

- **Semantic Analysis and Sex Identification (Anlamsal Analiz ve Cinsiyet Belirleme):** Generally, we may ask whether it is possible to further improve the accuracy of the identification system by adding a morphological and semantic analyzer to the system. A morphologic and a semantic analyzer are employed to produce the semantic network of the conversation [3, 19 and 20]. Semantically, some sentence structures in a conversation such as questions, answers and addressed sentences may expose the sex of the chatter [3]. For example "How are you John?", "I am fine

Alice", "This is David", "What is your name? (ismin neydi?) → John", "Name? (ismin?)→ Alan", "Who are you? (sen kimsin?) → David", "U (U) → Buket" and etc. In some conversations, many addressed sentences may also be used such as "Hi John (Merhaba John)", "How are you David (Nasılsın? David)", "What is the news Ahmet (Nbr Ahmnet?)", "I am fine Ali (İyidir Ali)" and etc.

Table 5. A semantic analysis sentences in a conversation
(Tablo 5. Bir konuşma içinde anlamsal analiz cümlesi)

	Chatters	Sentence (Turkish)	English
(1)	GencPrens	:)	Smile.
(2)	GencPrens	yoksa kafana göre birini bulamadın mı?	Couldn't you find someone who is like-minded with you?
(3)	Merix	kafama göre birini bulamadım.	I couldn't find someone like-mined with me.
(4)	GencPrens	O zaman sen evde kalırsın bu gidişle.	At this rate you are not to be able to get married.
(5)	Merix	zaten olmasında.	Anyway, it should not become too.

The analysis of the relation between subjects and personal suffixes may also conclude important clues about the chatting persons in a conversation as presented in Table 5. For example, "At this rate you are not to be able to get married (O zaman sen evde kal-ır-sın bu gidişle)". In the sentence, the suffix "-sın" expresses that other chatter (merix) is female because the sentence, "you are not to be able to get married", is used for female persons.

Sometimes many implication sentences may also be accounted in these conversations. Here, sex of the chatters is identified indirectly through the implying sentences. For example, "I am not Murat, I am his older brother (Ben murat değilim abisiyim)", "No, I am a house girl (yok ev kızıyım)" and etc. In the semantic analysis, idiomatic expressions, and the relations between subjects and suffixes are specially analyzed for the identification. Here, the personal suffixes used in the semantic analysis are listed in Table 6.

Table 6. Personal suffixes used in semantic analysis
(Tablo 6. Anlamsal analizde kullanılan kişi ekleri)

Suffixes	Persons
-m	Singular first person
-(y)Im (im,ım,um,üm,yim,yım,yum,yüm)	Singular first person
-n	Singular second person
-sIn (sin, sın, sun, sün)	Singular second person

In Turkish, the personal suffixes determine the acting person. For example, the verb "gitti" (went) takes suffix for first person "-m" and "-n" for second person. Here, "-m" and "-n" suffixes determine subject as explained in Table 7.

Table 7. Example of using personal suffixes in semantic analysis
(Tablo 7. Anlamsal analizde kişi eklerinin kullanım örneği)

Verb	Example sentences	Morphological analysis
gitmek (to go)	Gittim (I went)	git-ti-m >> git: verb; -ti: simple past tense suffix; -m: singular first person suffix.
	Gittin (You went)	git-ti-n >> git: verb; -ti: simple past tense suffix; -n: singular second person suffix.

Some sentences, phrases and expressions can also be used to determine the persons' sexual identity. Here, the personal suffixes used in the key phrases and expressions are determined to define the sexual identity of the subject. For example, "Yakışıklıyım (I am handsome)" can morphologically be analyzed as yakışıklı-(y)ım [18]. Here, "handsome" determines the dominant sex and the suffix "-(y)ım" determines the singular first person. Then, the chatter can be identified as male.

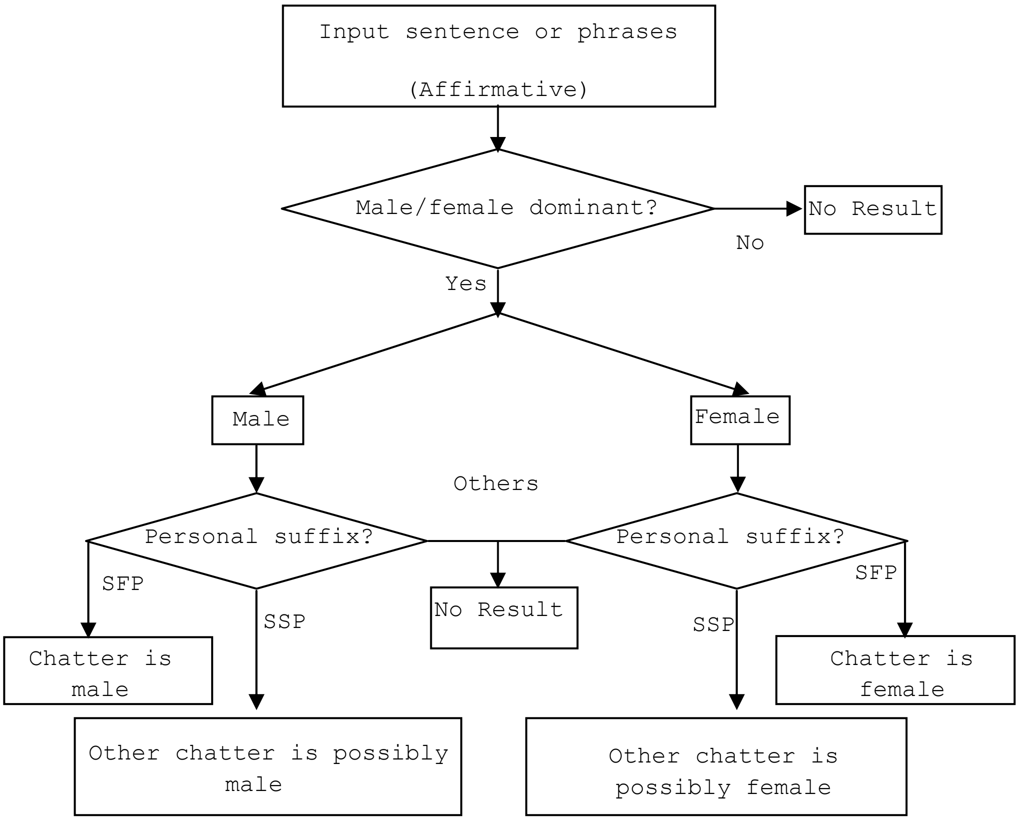

Figure 1. A flowchart of the semantic analysis approach
(Şekil 1. Anlamsal analize ait akış diyagramı)

A flowchart of this simple semantic analysis approach is given in Figure 1. Here, the SFP and SSP represent the singular first and second person respectively.

Here, some input may generate more precise and some others less precise outputs. As can be seen from the table, the reply or replies must confirm the previous sentence. For example, if the first sentence is "Güzelmiyim? (Am I beautiful?) > Güzel+mi+(y)im > {ADJ N+Question Suffix+(y)Im}", the reply may be "Evet (Yes you are)" and it confirms the first sentence.

A more complex semantic analysis method analyzes input with dominant male/female phrases or words. The first simple approach analyzes only the affirmative sentences but a more generalized method should analyze negative, question and negative question sentences as well. This generalized method is summarized in Table 8.

Table 8. A generalized semantic analysis
(Tablo 8. Anlamsal analizin genellenmesi)

Input sentences	Dominant	Suffixes	Result
Affirmative sentences	Male	SFP	Chatter is Male
		SSP	Other Chatter is possibly Male (Check the next response for more precise decision)
	Female	SFP	Chatter is Female
		SSP	Other Chatter is possibly Female (Check the next response for more precise decision)
Negative sentences	Male	SFP	No Decision (Check the next response for more precise decision)
		SSP	Other Chatter is possibly Male (Check the next response for more precise decision)
	Female	SFP	No Decision (Check the next response for more precise decision)
		SSP	Other Chatter is possibly Female (Check the next response for more precise decision)
Questions	Male	SFP	Chatter is possibly Male
		SSP	Other Chatter is possibly Male (Check the next response for more precise decision)
	Female	SFP	Chatter is possibly Female
		SSP	Other Chatter is possible Female (Check the next response for more precise decision)
Negative Questions	Male	SFP	Chatter is Male
		SSP	No Decision (Check the next response for more precise decision)
	Female	SFP	Chatter is Female
		SSP	No Decision (Check the next response for more precise decision)

Hence, semantic relations may contribute to the final decision and strengthen the accuracy of the identification system. Equation (5) and (6) combines the statistic and semantic identification outputs and produces a single identification output.

$$\lambda = (\lambda_{sta} * \gamma_{sta} + \lambda_{sem} * \gamma_{sem}) / \eta. \qquad (5)$$

$$\eta = \lambda_{sta} + \lambda_{sem}. \qquad (6)$$

Where λ is the final result that identifies the sex of the chatter, λ_{sta} and λ_{sem} are the statistic and semantic weigh coefficients respectively, γ_{sta} and γ_{sem} are statistic and semantic identifications respectively, and η is the normalization divider.

6. RESULTS (SONUÇLAR)

In this paper, we have presented a full-scale implementation of a chat system to collect information from conversations and a method to identify chatters profiles. This approach describes how to use a discrimination function and semantic analysis method for sex identification in the medium. Of the 154 conversations gathered from internet medium, 75 were used as training set while the rest 79 were used as test dataset. Tests results for the same data sets are also obtained on WEKA's Support Vector Machine (SVM) and Naive Bayes (NB) implementations as given in the following tables [21]. Experimental results indicate that the proposed discrimination function has sufficient discriminative power for the sex identification in the chat mediums. We also find that the system can quite accurately predict the chatter's sex in the mediums.

Table 9 presents the sex classification results for the conversations between chatters in the mediums. Here, general identification results including female and male chatter are presented to show the performances of the Naïve Bayes, Support Vector Machine, and the Proposed Methods. The decision accuracy of our system reaches to 89.9%.

Table 9. The general results of sex identification for the data of SDCM and mIRC
(Tablo 9. Özel tasarlanmış sohbet ortamı ve mIRC'den alınan veriler için cinsiyet belirleme sonuçları)

	NB	SVM	Our Method
Number of chatters	79	79	79
Number of correct decision	68	70	71
Number of wrong decisions	11	9	8
Percentage of correct decision	86.0%	88.6%	89.9%
Percentage of wrong decisions	14.0%	13.4%	10.1%

7. CONCLUSION AND FUTURE WORKS (TARTIŞMA VE GELECEK ÇALIŞMALAR)

Nowadays chat mediums are becoming an important part of human life and provide quite useful information about people in a society. In this paper, we have presented a log based information retrieval system that determinate to chatters' sex identities as an information-mining problem in chat mediums. A simple discrimination function is defined and evaluated for the sex identification in these mediums. Our simple and computationally less expensive sex identification system with semantic analysis method provides better performance if it is compared with other methods. These results show that the proposed identification function is quite useful for binary classification (male-female) such as sex identification. This identification system with the discrimination function achieves accuracy about 90% in the sex identification in the mediums.

Although some satisfactory results are obtained, the system is still needed to be improved and tested on larger data sets. Our experiments and results show that the methods proposed for sex identification may also be applied to the other concepts and subjects.

In this application, misleading questions and answers are not taken into account. This may effects the identification results negatively. In the future implementation of the system, the problem will be considered to minimize or eliminate these misleading sentences.

REFERENCES (KAYNAKLAR)

1. Khan, F.M., Fisher, T.A., Shuler, L.A., Tianhao, W., and Pottenger, W.M., (2002). Mining Chat-room Conversations for Social and Semantic Interactions. Lehigh University Technical Report, LU-CSE-02-011.
2. Elnahrawy, E., (2002). Log-Based Chat Room Monitoring Using Text Categorization: A Comparative Study. The International Conference on Information and Knowledge Sharing, US Virgin Islands.
3. Kose, C., and Ozyurt, O., (2006). A Target Oriented Agent to Collect Specific Information in a Chat Medium. Lecture Notes in Computer Science, Springer Verlag, 4263, pp:697-706.
4. Kose, C., Nabiyev, V., and Özyurt, O., (2006). A statistical approach for sex identification in chat mediums, The international scientific conference on Problems of Cybernetic and Informatics (PCI), Azerbaijan.
5. Yang, Y., (1999). An evaluation of statistical approaches to text categorization .In Information Retrieval Journal, 1(2), pp:69–90.
6. Ozyurt, O. and Kose, C., (2006). Information extraction in the chat mediums: statistical and semantic approaches for sex identification, ELECO 2006, Electrical- Electronics-Computer Engineering Workshop.
7. Bing, L., Xiaoli, L., Wee, S.L., and Philip, S.Y., (2004). Text Classification by Labeling Words. Nineteenth National Conference on Artificial Intelligence, pp:425–430.
8. Koppel, M., Argamon, S., and Shimoni, A.R., (2003). Automatically Categorizing Written Texts by Author Gender. Oxford Journals, Humanities, Literary and Lingustic Computing, 17, pp:401-412.
9. Amasyalı, M.F., and Diri, B., (2006). Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. Lecture Notes in Computer Science, Springer Verlag, 3999, pp:221-226.
10. Bengel, J., Gauch, S., Mittur, E., and Vijayaraghavan, R., (2004). Chattrack: chat room topic detection using classification. Lecture Notes in Computer Science, Springer Verlag, 3073, pp:266-277.
11. Yan, X. and Yan, L., (2006). Gender Classification of Weblog Authors. The Twenty-First National Conference on Artificial Intelligence, pp:228-230.
12. Haichao, D., Siu, C.H., and Yulan, H., (2006). Structural analysis of chat messages for topic detection, Online Information Review, 30(5), pp:496-516.
13. Joachims, T., (1998). Text categorization with support vector machines: learning with many relevant features. Lecture Notes in Computer Science, Springer Verlag, 1398, pp:137-142.
14. George, H.J. and Pat, L., (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, pp:338-345.
15. Kolenda, T., Hansen, L.K., and Larsen, J., (2001). Signal detection using ICA: application to chat room topic spotting. Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Source Separation, ICA'2001, San Diego, USA, pp:540-545.
16. Bingham E., Kab A., and Girolami M., (2003). Topic identification in dynamical text by complexity pursuit, Neural Processing Letters, (17), pp:69-83.
17. Hengirmen, M.,(2002). Türkçe Dilbilgisi. Engin Yayınevi, Ankara.

18. Oflazer, K., (1994). Two-level Description of Turkish Morphology. Literary and Linguistic Computing, 9, pp: 137- 148.
19. Kanagaluru, C.S., and Janaki, R.D., (2002). The dynamics of language understanding. Language Engineering Conference, Hyderabad, India, pp:197-199.
20. Baumgartner, R., Eiter, T., Gottlob, G., Herzog, M., and Koch, C., (2005). Information extraction for the semantic. Lecture Notes in Computer Science - Reasoning Web, 3564, pp:275-289
21. www.cs.waikato.ac.nz/ml/weka/