TURKMIA'09 Proceedings
VI. Ulusal Tıp Bilişimi Kongresi Bildirileri
ENMI Vol V No 1, 2009

Kulak Burun Boğaz Epikriz Notlarından Birliktelik Kurallarının Çıkartılması

Başak OĞUZ[a, 1], Uğur BİLGE[a], Mehmet Kemal SAMUR[a]

[a] *Biyoistatistik ve Tıp Bilişimi AD, Akdeniz Üniversitesi, Antalya*

Association Rules Extraction from the Otolaryngology Discharge Notes

Abstract. Recently, the number of text mining applications in medical sciences has grown with an increasing rate. Unstructured free-text data, such as patient discharge notes and reports, doctor's notes, clinical trials and studies, research reports, web pages and hospital records are some of the important data sources for physicians. To analyze and access this kind of data by human efforts is difficult and time consuming. Considering the time it takes for decision making, and accesing accurate and required information about patients, this kind of systems have become necessary.

In this study, we developed a software system to transform 600 discharge notes, from the Department of Otolaryngology of Akdeniz University, to a structured form, enabling physicians to access patient information, extracting clinical data from the discharge notes, and codifying them for analysis. First of all, discharge notes which are kept as Microsoft Office Word documents have been transformed into a data table after preprocessing. A query form has also been designed for enabling physicians to access the patient data. To identify the significant content words within each section keyword lists have been used and content words have been converted into a predefined coded structure. Association Rules, that is one of the methods of the traditional data mining, has been applied to the coded data in order to discover the relations between entities/concepts.

Keywords. Otolaryngology; text mining; data mining; association rules

Özet. Metin madenciliği tekniklerinin tıpta kullanımı son birkaç yılda büyük oranda artmıştır. Yapılan klinik çalışmalar, araştırma raporları, hastane kayıtları, doktor notları ve faturalar gibi serbest formatta bulunan metinler tıptaki en önemli veri kaynaklarıdır. Fakat yapılandırılmamış formatta bulunan bu geniş veri yığınlarını insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak hem zordur hem de zaman kaybına yol[1] açmaktadır. Hastayla ilgili karar verme süresinin, doğru verilere erişmenin ve bu verileri kullanarak istenilen bilgilere ulaşmanın zorluğu göz önünde bulundurulduğunda bu tür sistemlerin önemi ön plana çıkmaktadır.

[1] Sorumlu Yazar: Akdeniz Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi AD, basakoguz@akdeniz.edu.tr

Bu çalışmada, Akdeniz Üniversitesi Hastanesi Kulak Burun Boğaz Hastalıkları Anabilim Dalı'ndan alınan ameliyat geçiren hastalara ait 600 adet hasta bilgi formunu yapılandırılmış formata dönüştürmek, hekimlerin hasta ile ilgili ihtiyaç duydukları bilgilere erişimini kolaylaştırmak, hasta bilgi formlarından klinik verileri çıkartmak ve bu verileri analiz etmek amacıyla bir yazılım geliştirilmiştir. Önce Microsoft Office Word belge formatında bulunan hasta bilgi formlarındaki veri alanları ön işlemden geçirilerek veri tablosu haline dönüştürülmüştür. Hazırlanan metin sorgu formuyla birlikte hekimlerin hasta bilgi formlarında aradıkları özellikteki hastalara erişimlerinde kolaylık sağlanmaktadır. Ayrıca her alana özgü oluşturulan anahtar kelime listeleriyle metin içerikleri kodlanabilmekte ve bu veriler üzerinde veri madenciliği teknikleri uygulanabilmektedir. Bu çalışmada, varlıklar/ kavramlar arasındaki ilişkilerin tanımlanabilmesi için veri madenciliğinde kullanılan ilk tekniklerden biri olan Birliktelik Kuralı yöntemi uygulanmıştır.

Anahtar Kelimeler. Kulak Burun Boğaz; Metin Madenciliği; Veri Madenciliği; Birliktelik Kuralları

Giriş

İnternet kullanımının hızla artması ve kişisel bilgisayarların yaygınlaşması ile birlikte gittikçe büyüyen hacme sahip doküman yığınları oluşmaktadır. Bu belgeler içinde önemli bilgiler kaybolup giderken, değerli bilgilere ulaşmak için dokümanların içeriğinin belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı kendini hissettirmektedir. Gelenekselleşmiş yöntemleri içeren bilgi erişim (information retrieval) sistemleri belge yığınlarından faydalı ve gerekli bilgileri bulmaya yardımcı olsalar da gerekli detay ve özel bilgilere, bu yöntemler ile ulaşmak zordur. Oysa pek çok açıdan belgeler içindeki bilgilere, ilişkilere ulaşmak son derece önemlidir. Örneğin bir hastalık için ilaç bulmaya çalışan bir araştırmacının, kendisinden önce yapılmış tüm çalışmaları olabildiğince hızlı bir şekilde incelemesi ve bu inceleme sürecinde belgelerin içeriğine, konusuna, içinde geçen kavramlara ve bu kavramların diğer belgelerde geçen farklı kavramlarla ilişkisine ulaşması gerekir [1]. Metin madenciliği metin formatındaki verileri kullanarak içerisindeki bilgileri gün ışığına çıkaran ve özellikle 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alandır [2].

Hekimler karar verme sürecinde veya araştırma yaparken hasta raporları, klinik çalışmalar, araştırma raporları, web sayfaları ve hastane kayıtları gibi serbest metin formatında veya kağıt tabanlı olarak bulunan bu metinleri kullanmaktadır. Yapılandırılmamış formatta bulunan bu geniş veri yığınlarını insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak hem zordur hem de zaman kaybına yol açmaktadır.

Bu çalışmada, KBB (Kulak Burun Boğaz) Hastalıkları Anabilim Dalından alınan ve ameliyat geçiren hastalara ait hasta bilgi formları kullanılmış ve bu metinler üzerinde işlemler yapılmıştır. Çalışmada, yapılandırılmamış formatta bulunan KBB hasta bilgi formlarının yapılandırılmış hale dönüştürülmesi, hekimlerin karar verirken ya da araştırma yaparken hasta ile ilgili ihtiyaç duydukları bilgilere erişimlerinin kolaylaştırılması, hasta bilgi formlarını incelemek için harcadıkları zamanın azaltılması ve veri madenciliği teknikleri ile varlıklar arasındaki gizli ilişkilerin çıkartılması amaçlanmıştır.

1. Veri Madenciliği ve Metin Madenciliği

Veri madenciliği, veri ambarlarında yararlı olma potansiyeline sahip, aralarında beklenmedik/bilinmedik ilişkilerin olduğu verilerin keşfedilerek hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur [3]. Metin madenciliği ise belirli bir formatta olmayan, yazı tipindeki veriler içerisinde gizli olan nitelikli bilginin çıkarılması, düzensiz haldeki verinin formatlanması sürecidir [4]. Veri madenciliği uygulamalarında çoğunlukla yapılandırılmış veriler kullanılmaktadır. Veri madenciliğinde elektronik tablo halinde sunulan veriler kullanılırken metin madenciliği uygulamaları metin formatındaki verileri kullanmaktadır. Metin madenciliğinin ana konularından biri metin verilerin sayısal veri haline dönüştürülüp elektronik tablo şeklinde sunulmasıdır [5]. Böylelikle yapılandırılmamış formatta bulunan metinler veri madenciliği tekniklerinin uygulanabileceği yapılandırılmış formata dönüştürülebilmektedir.

Bu çalışmada hasta bilgi formlarında bulunan varlıklar arasındaki ilişkilerin belirlenmesi için veri madenciliği tekniklerinden **Birliktelik Kuralları** (Association Rules) yöntemi kullanılmıştır. Birliktelik analizi, belirli bir veri kümesinde yüksek sıklıkta birlikte görülen özellik değerlerine ait ilişkisel kuralların keşfidir. Birliktelik kuralı, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır [6]

Birliktelik Kuralında, öğeler arasındaki bağıntı, destek ve güven kriterleri ile hesaplanır. Destek kriteri, veride öğeler arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y öğesinin hangi olasılıkla X öğesi ile beraber olacağını söyler. İki öğenin birlikteliğinin önemli olması için hem destek, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir [7]. Birliktelik kuralı, kullanıcı tarafından minimum değeri belirlenmiş destek ve güven eşik değerlerini sağlayacak biçimde üretilir.

Birliktelik kuralının en yaygın kullanıldığı örnek market sepet analizidir [8]. Market sepet analizi, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını belirlemeye çalışır. Veriler metin formatında bulunduğunda market sepet analizindeki ürünler yerini kavramlara bırakmaktadır. Kavramların birlikte bulunma durumlarına bakılarak aralarındaki ilişki örüntüleri tespit edilmektedir. Birliktelik Kuralı analizi tıp alanında genel olarak genler arasındaki ilişkilerin belirlenmesinde [9], hastalıkların tahmininde [10], risk faktörlerinin belirlenmesinde vb. uygulamalarda kullanılmaktadır.

2. Gereç ve Yöntem

Bu çalışmada, KBB uzman hekimleriyle yapılan görüşmeler sonucunda ellerinde bulunan ve hastalarla ilgili en kapsamlı veriye erişilebilecek belge olan hasta bilgi formlarının kullanılması kararlaştırılmıştır. Bu sebepten, öncelikli olarak KBB Anabilim Dalı'ndan Microsoft Office Word (doc uzantılı) formatında ve her biri yaklaşık olarak 40 Kb boyutunda, 2002-2007 yılları arasında gelen ameliyat geçiren hastalara ait 600 adet hasta bilgi formu alınmıştır. Daha hızlı işlem yapılabilmesi için bu belgeler Metin Belgesi (txt uzantılı) formatına dönüştürülmüştür. Bu 600 belgeden dokuz tanesinde birçok alan boş bırakıldığı için bu dokümanlar çalışmaya dahil edilmemiştir.

2.1. Yazılım Geliştirme Sırasında Kullanılan Sistemler ve Teknikler

Yazılımın geliştirilmesi sırasında Visual Studio .NET platformu ve Visual C# .NET programlama dili kullanılmıştır. Girdi metinlerinde bulunan yazım hatalarının tespit edilmesi, bu hataların düzeltilebilmesi ve metinlerin yapıtaşı olan kelimelerin gövdelerinin elde edilebilmesi için açık kaynak kodlu bir doğal dil işleme yazılımı olan Zemberek [11] kütüphanesi kullanılmıştır. Türkçede kök sözcüğe yapım ekleri getirilerek türetilen yeni sözcüklere gövde denir. Bir sözcüğün eklenmiş çekim eklerinden arındırılarak gövde veya kökün bulunması işlemine *gövdeleme* adı verilir [12]. Aşağıda, metinlerde uygulanan gövdeleme işlemine örnek verilmiştir.

Örneğin; Dilde=Dil, Kandaki=Kan

Böylelikle program aynı kökten gelen fakat çekim eki almış kelimeleri farklı bir kelime olarak algılamamaktadır. Bu işlemle birlikte, metinlerin boyutu azalmakta, gereksiz yere işlem süresi uzatılmamakta, aynı kelimeler yazılım tarafından farklı kelime olarak algılanmamakta ve daha doğru frekans sonuçları elde edilebilmektedir.

2.2. Oluşturulan Kelime Listeleri

Hasta bilgi formlarından örnek olarak seçilen 100 adet belge incelendiğinde bu belgelerin "Şikayet", "Yaş" gibi alanlardan ve bu alanların içerisinde bulunan hastalara ait bilgilerden oluştuğu gözlemlenmiştir. Yazılım geliştirilirken ilk amaç bu metinler üzerinde işlem kolaylığı sağlayabilmek ve daha kapsamlı analizler yapabilmek için bu formları veri tablosu haline dönüştürmekti. Bu yüzden alan isimleri sütuna ve içerikleri satırlara gelecek şekilde bir tablo oluşturabilmek için alan isimlerinin bulunduğu bir liste oluşturulmuştur. Fakat bu alan isimleri her belgede aynı şekilde yazılmadığı için bunların standart bir forma dönüştürülmesi gerekliliği ortaya çıkmıştır. Bu yüzden alan isimlerinin tüm yazılış şekillerinin ve buna karşılık gelen standart formunun bulunduğu bir düzeltme listesi oluşturulmuştur.

Daha öncede bahsedildiği gibi Zemberek Kütüphanesi kullanılarak bir yazım denetimi modülü geliştirilmiştir. Yazım denetimi sonucunda Zemberek hatalı bulduğu kelimeler bir öneri listesi sunmaktadır. Tablo 1'de en sık rastlanılan hata türlerinin ve bunlara karşılık sunulan önerilerin bulunduğu bir liste örnek olarak gösterilmiştir. Yazım denetimi modülü, elde bulunan tüm belgeler programa yüklü iken çalıştırılmış, çıktı olarak elde edilen liste yazılım geliştiriciler tarafından incelenmiş ve doğru bulunan sonuçlardan uygun öneriler seçilerek düzeltme listesine eklenmiştir.

Hasta bilgi formları incelendiğinde bazı alanların genellikle hekimler tarafından doldurulduğu bazı alanların ise boş bırakıldığı tespit edilmiştir. Bu sebeple "önemli alanlar" adı altında başka bir liste oluşturularak bu listedeki alanlar dışında kalan tüm alanların ve içeriklerinin yazılım tarafından silinmesi sağlanmıştır. Ayrıca girdi metinlerinin boyutunu ve işlem yapılırken gereksiz kelimeler (acaba, ama, aslında, bazen, vb.) için harcanan zamanı azaltmak için internetten elde edilmiş "Türkçede sık kullanılan kelimeler" listesi [13, 14] kullanılarak girdi metinleri içerisinden bu kelimelerin çıkartılması sağlanmıştır. Bu liste toplamda 275 kelime içermektedir.

Tablo 1. Sık Karşılaşılan Yazım Hatası Türleri

Hatalı Kelime	Hata Türü	Öneriler
Ğüçlüğü	Yanlış yazılan karakter	/ Güçlüğü / güçlüğü
Buruntıkanıklığı	Bitişik yazılan kelimeler	/ burun tıkanıklığı
Boynda	Düşen karakter	/ Boyunda / Boyna
Kulakata	Eklenen karakter	/ Kulakta / Kulaklata / Kulalata
Şişilk	Yer değiştiren komşu karakterler	/ Şişlik / Şişilik / şiş ilk

Yazılım tarafından önemli/anahtar kelimelerin belirlenmesi ve bu kelimelerin metinlerde etiketlenmesi için her alana özgü anahtar kelime listeleri oluşturulmuştur. Bu listeler oluşturulurken metinler yazılım tarafından okunup ham haliyle veri tablosu haline dönüştürüldükten sonra tüm alanlardaki kelimeler tekli, ikili ve üçlü olmak üzere kelime ve kelime gruplarına ayrılmış ve frekansları hesaplanmıştır. İşlemin sonunda elde edilen sonuçlar yazılım tarafından Microsoft Office Excel'e aktarılmış ve elle yapılan incelemeler sonucunda her alana özgü listeler oluşturulmuştur. Kelimelerin tekli, ikili vb. kelime ve kelime gruplarına ya da harf ve harf gruplarına ayrılarak dizilişlerine bakılması ve örüntülerin çıkartılması literatürde "N-gram" yöntemi olarak geçmektedir [15-17]. Bu yöntem kullanılarak tek başına bir anlam ifade etmeyen, ikili ya da üçlü kelime grupları şeklinde anlam ifade eden kelimeler de oluşturulan listelere eklenmiştir. Sadece "Özgeçmiş" alanında kelime listeleri kullanılmamış ve hastanın sigara ve alkol kullanıp kullanmadığını belirlemek için bazı anahtar kelimeleri içerip içermediğine bakılmıştır. Frekanslara bakılarak yapılan inceleme sonucunda sigara kullanan hastaları tanımlamak için genellikle "Paket", "Adet", "Sigara" vb., alkol kullanan hastalar için ise "Alkol", "Kadeh" vb. kelimelerin kullanıldığı gözlemlenmiştir. Eğer hastanın özgeçmiş bilgisinde bu kelimeler geçiyorsa yazılım hastanın özgeçmiş bilgisini "Sigara-Var" veya "Alkol-Var" olarak değiştirmekte, eğer bu kelimelerden herhangi birini içermiyorsa "Özellik-Yok" yazmaktadır.

2.3. Elde Edilen Verilerin Analiz Edilmesi

Bu adımda, metinler içerisindeki anahtar kelime ve kelime grupları kodlanmış hale dönüştürüldükten sonra elde edilen kodlanmış verilerin veri madenciliği yöntemlerinden Birliktelik Kuralı ile analiz edilmesi işlemi yapılmaktadır. Bu işlem için yazılım içerisinde, yöntemin temel algoritmasını kullanarak kuralların destek ve güven değerlerini hesaplayan özel bir form tasarlanmıştır. Yöntem, sıklıkla girilmiş olan ve analiz için elverişli veriyi içeren "Şikayet", "Yaş", "Cinsiyet", "Özgeçmiş" ve "Ameliyat Öncesi Tanı" alanlarındaki veriler üzerinde uygulanmış ve belirlenen minimum destek ve güven değerlerine göre kurallar elde edilmiştir. Bunun dışında, alanlar içerisinde geçen anahtar kelime ve kelime gruplarının frekansları da hesaplanmıştır.

3. Bulgular

3.1. Sistem Çıktıları

Geliştirilen yazılımda üç adet sekme bulunmaktadır. Bunlardan ilki metinler üzerinde bütün işlemlerin yapılarak metinlerin veri tablosuna dönüştürüldüğü Giriş sekmesi, ikincisi hekimlerin yüklenen metinleri istedikleri hasta özelliklerine göre sorgulayabildikleri sorgulama sekmesi ve sonuncusu da verilere birliktelik kuralı algoritmasının uygulandığı birliktelik kuralı sekmesidir. Her sekmede istenildiği zaman elde edilen sonuçlar Microsoft Office Excel'e gönderilebilmektedir.

Yazılımın 591 adet hasta bilgi formunu tüm işlemleri yaparak veri tablosu formatına dönüştürme süresi, 1,83 GHz çift çekirdek işlemci ve 2 GB RAM'e sahip bir bilgisayarda yaklaşık 2-3 dakikadır.

3.2. Hasta Bilgi Formlarının Analizden Elde Edilen Sonuçlar

Analiz aşamasında hem SPSS paket programı hem de geliştirilen Birliktelik Kuralı modülü kullanılmıştır. SPSS paket programı, analizde kullanılan değişkenlere ait özelliklerin saptanması aşamasında tercih edilmiştir. Her bir hasta bilgi formunda ortalama 410 kelime bulunmaktadır.

Tablo 2. Frekans Tablosu

Şikayet		**Ameliyat Öncesi Tanı**		**Ameliyat Sonu Tanı**	
Kelime	**Frekans**	**Kelime**	**Frekans**	**Kelime**	**Frekans**
Ses Kısıklık	124	Larenks CA	107	Boyun Diseksiyon	145
Boyun Şişlik	75	Kitle	53	Larenjektomi	83
İşitme Azlık	53	Otitis	36	Parotidektomi	44
Yutma Güçlük	39	Dil CA	29	Septumplasti	32
Burun Tıkanıklık	37	Tümör	25	Tümör Rezeksiyon	24
Baş Dönme	34	Septum Deviasyon	19	Hemiglossektomi	18
Kulak Akıntı	31	Hipertrofi	12	Mastoidektomi	16
Kulak Şişlik	28	Vejetasyon	10	Tümör Eksizyon	13
Dil Yara	25	Sinüzit	9	Timpanoplasti	13
Nefes Darlık	24	Dudak CA	7	Biyopsi	13

Tablo 2'de hasta bilgi formlarının analizi sonucunda elde edilen frekanslara örnekler verilmektedir. "Şikayeti" alanında en sık geçen problemler "Ses Kısıklık", "Boyun Şişlik" ve "İşitme Azlık" olarak bulunmuştur. "Ameliyat Öncesi Tanı" alanında "Larenks CA", "Kitle" ve "Otitis", "Ameliyat Sonu Tanı" alanında ise "Boyun Diseksiyon", "Larenjektomi" ve "Parotidektomi" en sık rastlanan kavramlardır.

Tablo 3. Analizde Kullanılan Hasta Verilerinin Frekans ve Yüzdeleri

		N	%
Ameliyat Öncesi Tanı	Larenks-CA	96	41,7
	Septum Deviasyon	18	7,8
	Tümör	57	24,8
	Otitis	34	14,8
	Dil-CA	25	10,9
Yaş	0-25	15	6,5
	26-40	29	12,6
	41-60	114	49,6
	61 ve üstü	72	31,3
Sigara	Var	65	28,3
	Yok	165	71,7
Şikayet	Ses Kısıklık	94	40,9
	Boyun Şişlik	27	11,7
	Kulak Akıntı	23	10,0
	İşitme Azlık	11	4,8
	Kulak Şişlik	17	7,4
	Burun Tıkanıklık	19	8,3
	Nefes Darlık	7	3,0
	Kitle	11	4,8
	Dil Yara	21	9,1
Cinsiyet	Kadın	65	28,3
	Erkek	165	71,7
Toplam		230	100

Metinlerde bulunan varlıklar arasındaki ilişkilerin belirlenebilmesi için yapılan birliktelik analizinde genel olarak tüm hastalar için girilmiş olan "Yaşı", "Özgeçmiş", "Cinsiyet", "Şikayeti" ve "Ameliyat Öncesi Tanı" alanlarındaki veriler kullanılmıştır. Toplamda 230 hastaya ait veri analize dahil edilmiştir. Hastaların yaşları da 0-25, 26-40, 41-60 ve 61 ve üstü olmak üzere dört kategoriye ayrılmıştır. Hasta verilerine ait frekans ve yüzdeler Tablo 3'te gösterilmiştir.

Birliktelik Kuralı analizi ile yüksek sıklıkta birlikte görülen kelime ve kelime grupları bulunmakta ve belirlenen minimum destek ve güven değerine göre kurallar üretilmektedir. Tablo 4'te, % 4 minimum destek ve % 50 minimum güven değerine göre "Yaşı", "Özgeçmiş", "Cinsiyet", "Şikayeti" ve "Ameliyat Öncesi Tanı" alanları kullanılarak toplamda 230 hasta için yapılan analizden çıkartılan 26 kuraldan bazıları örnek olarak gösterilmiştir.

Tablo 4. Analiz Sonucunda Elde Edilen Kurallar

Kural No	Birliktelik Kuralları (A=>B)	Destek (%)	Güven (%)
1	41-60, Burun Tıkanıklık=>Septum Deviasyon	4,3	100
2	Burun Tıkanıklık, Erkek=>Septum Deviasyon	5,2	100
3	Burun Tıkanıklık=>Septum Deviasyon	7,8	94,74
4	Dil Yara, Kadın=> Dil CA	4,8	91,67
5	Dil Yara=> Dil CA	7,8	85,71
6	Kadın, Kulak Akıntı=> Otitis	5,2	100
7	İşitme Azlık=> Otitis	4,8	100
8	Kulak Akıntı=> Otitis	10	100
9	Erkek, Kulak Akıntı=> Otitis	4,8	100
10	41-60, Kulak Akıntı=> Otitis	4,8	100
11	Kulak Şişlik=> Tümör	7,4	100
12	41-60, Boyun Şişlik=> Tümör	6,1	87,5
13	Boyun Şişlik=> Tümör	9,1	77,78
14	41-60, Erkek, Ses Kısıklık=> Larenks CA	19,1	95,65
15	41-60, Ses Kısıklık=> Larenks CA	20	93,88

Analiz sonucunda elde edilen kuralların sol tarafında bulunan ve "Yaşı", "Özgeçmiş", "Cinsiyet" ve "Şikayeti" alanlarına ait verileri içeren kısım "kuralın gövdesi" olarak adlandırılmaktadır. Sağ tarafta bulunan ve "Ameliyat Öncesi Tanı" alanına ait verileri içeren kısım ise "kuralın başı" olarak adlandırılmaktadır [18]. Kurallar yorumlanırken hem destek değerinin hem de güven değerinin yüksek olmasına dikkat edilmelidir. Kuralların nasıl yorumlandığına dair aşağıda üç örnek verilmiştir.

- **Kural 1:** Yaşı 41-60 arasında olan ve Burun Tıkanıklığı şikayeti olan hastaların tamamına Septum Deviasyonu tanısı konulmuştur (Güven=%100).
- **Kural 12:** Yaşı 41-60 arasında olan ve Boyun Şişlik şikayeti olan hastaların %87, 5'ine Tümör tanısı konulmuştur (Güven=%87, 5).
- **Kural 15:** Yaşı 41-60 arasında olan ve Ses Kısıklığı şikayeti olan hastaların %93, 88'ine Larenks CA tanısı konulmuştur (Güven=%93, 88).

4. Tartışma ve Sonuç

Bu çalışmada, yapılandırılmamış formatta bulunan KBB hasta bilgi formlarının yapılandırılmış hale dönüştürme, hekimlerin hasta ile ilgili ihtiyaç duydukları bilgilere erişimini kolaylaştırma, karar verirken ya da araştırma yaparken hasta bilgi formlarını incelemek için harcadıkları zamanı azaltma, veri madenciliği teknikleri ile varlıklar arasındaki gizli ilişkileri çıkartma amaçlanmıştır.

Metinlerde bulunan varlıklar arasındaki ilişkilerin saptanması için veri madenciliği yöntemlerinden Birliktelik kuralı kullanılmıştır. Bu yöntemle kavramlar/varlıkların birlikte bulunma durumlarına bakılarak aralarındaki yüksek sıklıkta görülen ilişki örüntüleri tespit edilmekte ve kurallar çıkartılmaktadır [19]. Ayrıca Birliktelik Kuralları

hem analistler hem de normal kullanıcılar tarafından kolayca anlaşılıp yorumlanabildiği için çalışmalarda sıklıkla tercih edilmektedir [20]. Analiz sürecinde yaşanan en büyük problem, doktorların birçok formda tüm alanları doldurmamış olması ve bu yüzden analize dahil edilen özellik ve hasta sayısında düşüş yaşanmasıdır. Çalışmada, 591 hasta bilgi formundan elde edilen veriler incelendikten sonra analiz için elverişli alanlar (yaş, cinsiyet, şikayet, özgeçmiş, ameliyat öncesi tanı) belirlenmiş ve bu alanların tümüne sahip olan 230 hasta verisi Birliktelik kuralı analizinde kullanılmıştır. Analiz sonucunda minimum destek ve güven değerine göre 26 kural elde edilmiştir. Bu tür kuralların, gelecekte geliştirilecek olan karar destek sistemlerine veya klinik rehberlere fayda sağlayacağı düşünülmektedir. Şerban et al. [21] çalışmasında kanserli hastalara ait tıbbi veriler üzerinde Birliktelik kuralı analizi uygulanmış ve elde edilen kurallar kullanılarak hastaların semptomlarına göre kanserli olup olmadıklarını tahmin eden küçük ölçekli bir sistem geliştirilmiştir. Ordonez'in çalışmasında [10] hastanın kronik hastalıkları, yaşı, cinsiyeti, sigara içip içmediği, kan basıncı vb. verileri içeren ve kalp rahatsızlığı olan hastalara ait 655 adet hasta kaydı kullanılarak Birliktelik Kuralı analizi yapılmış ve risk faktörlerine göre kalp hastalığının olup olmadığı tahmin edilmeye çalışılmıştır. Mahgoub et al. [20] çalışmasında kuş gribi ile ilişkili birçok kaynaktan (Reuters, BBC, Medical News Today, Yahoo vb.) toplanan örnek 100 adet internet sayfası XML formatına dönüştürülmüş ve anahtar kelimeler arasındaki ilişkilere bakılarak hastalıkla ilişkili özellikler (hastanın durumu, lokasyon vb.) EART adlı metin madenciliği sistemi ile çıkartılmıştır. Literatürde KBB bölümüne özgü yapılan benzer bir çalışma bulunmamaktadır.

Hekimlerin karar verirken veya araştırma yaparken hasta ile ilgili bilgilere kolaylıkla erişebilmelerini sağlamak ve hasta bilgi formlarını incelemek için harcadıkları zamanı azaltmak için hekimlerin istedikleri özellikteki hasta bilgilerine erişimlerini sağlayacak bir sorgu formu tasarlanmıştır. Hekimlerin ihtiyaç duydukları hasta bilgilerine erişebilmek için elde bulunan tüm belgeleri inceledikleri göz önünde bulundurulduğunda, geliştirilen yazılımın bu süreci kolaylaştırarak harcanan zamanı büyük bir oranda azaltacağı söylenebilir.

Sonuç olarak, metin madenciliği ve kullanılan tekniklerle ile ilgili olarak yapılan uzun soluklu bir araştırma sonucunda elde edilen bilgiler ve belirlenen hedefler doğrultusunda KBB hasta bilgi formlarının analizi için bir yazılım geliştirilmiş ve süreç içerisinde metinlerle çalışmanın getirdiği birçok deneyim kazanılmıştır. Metin madenciliği, Türkiye'de son birkaç yılda ilgi gören ve bilgi eksikliği, deneyimsizlik ve altyapı yetersizliğinden dolayı özellikle sağlık alanında yeterli sayıda çalışmanın yapılamadığı bir alandır. Çalışmanın en büyük hedeflerinden biri de bu alanda araştırma yapmak isteyen kişilere yol gösterici olmak ve konu ile ilgili temel bilgileri vererek süreç içerisinde ne tür problemlerle karşılaşıldığını ve bu problemlerin nasıl çözümlendiğini göstermektir.

Kaynakça

[1] Güven A. Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği. Doktora Tezi, 2007, İstanbul.
[2] Konchady M. Text Mining Application Programming. 1st ed. Charles River Media, Boston, 2006.
[3] Dolgun MO, Özdemir TG, Deliloğlu S. Öğrenci Seçme Sınavında (ÖSS) Öğrencilerin Tercih Profillerinin Veri Madenciliği Yöntemleriyle Tespiti. Bilişim'07 Kongresi, Ankara, 2007.
[4] Sehgal AK, Text Mining: The Search for Novelty in Text, http://www.cs.uiowa.edu/~sehgal/Papers/comp04.pdf , Last accessed: 12.02.2007.

[5] Weiss SM, Indurkhya N, Zhang T, Damerau FJ. Text Mining; Predictive Methods for Analyzing Unstructured Information. Springer Science+Business Media, Newyork, 2005.
[6] Özçakır FC, Çamurcu AY. Birliktelik Kuralı Yöntemi için Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi 2007: 6(12): 21-37.
[7] Karabatak M, İnce MC, Apriori Algoritması ile Öğrenci Başarısı Analizi, http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf, Last accessed: 29.01.2009.
[8] Kononenko I, Kukar M, Machine Learning and Data Mining; Introduction to Principles and Algorithms, Horwood Publishing, 2007, Chichester.
[9] Nam H, Lee K, Lee D. Identification of Temporal Association Rules from Time-Series Microarray Data Sets. BMC Bioinformatics 2009: 10(3): 6.
[10] Ordonez C. Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction. IEEE Transactions On Information Technology In Biomedicine 2006: 10(2).
[11] Zemberek; Project Home Page, https://zemberek.dev.java.net/, Last accessed: 15.12.2007.
[12] Tülek M. Türkçe İçin Metin Özetleme. Yüksek Lisans Tezi, 2007, İstanbul.
[13] Turkish Stopwords, http://www.ranks.nl/stopwords/turkish.html, Last accessed: 06.01.2007.
[14] Can F, Kocberber S, Balcik E, Kaynak C, Ocalan HC, Vursavas OM. Information Retrieval on Turkish Texts. Journal Of The Amerıcan Socıety For Informatıon Scıence and Technology 2008: 59(3): 407–421.
[15] Bruijn LM, Hasman A, Arends JW. Supporting the classification of pathology reports: comparing two information retrieval methods. Computer Methods and Programs in Biomedicine 2000: 62: 109–113.
[16] Heja G, Surjan G. Using n-gram method in the decomposition of compound medical diagnoses. International Journal of Medical Informatics 2003: 70: 229-236.
[17] Çebi Y, Dalkılıç G. Turkish Word N-gram Analyzing Algorithms for a Large Scale Turkish Corpus – TurCo. Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).
[18] DB2 Universal Database, http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.im.model.doc/c_lift_in_an_association_rule.html, Last accessed: 03.03.2009.
[19] Feldman R, Sanger J. The Text Mining Handbook; Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Newyork, 2007.
[20] Mahgoub H, Rösner D, Ismail N, Torkey F. A Text Mining Technique Using Association Rules Extraction. International Journal of Computational Intelligence 2008: 4: 1.
[21] Şerban G, Czibula IG, Campan A. A Programming Interface For Medical Diagnosis Prediction. Studia Unıv. Babes Bolyai Informatica 2006: 1(1): 21-30.