www.istatistikciler.org
İstatistikçiler Dergisi 2 (2009) 48-58

İstatistikçiler Dergisi

Veri madenciliği'nde yapısal olmayan verinin analizi: Metin ve web madenciliği

M. Özgür Dolgun
SPSS, Çankaya Mah. Mahmut Yesari Sk. No:8/5
06550-Çankaya, Ankara, Türkiye
odolgun@spss.com.tr

Tülin Güzel Özdemir
SPSS, Zümrütevler Atatürk Cd. Nazmi İlker Sk. No:24
34852-Maltepe, İstanbul, Türkiye

Doruk Oğuz
SPSS, Zümrütevler Atatürk Cd. Nazmi İlker Sk. No:24
34852-Maltepe, İstanbul, Türkiye
doguz@spss.com.tr

Özet

Verinin büyük boyutlara ulaşması ve bilgisayar donanımlarının bu büyük boyuttaki veriyi depolayarak yüksek kapasitede analiz yapabilecek seviyelere gelmeleri ile birlikte analistler karmaşık koşullar ile karşı karşıya kalmaktadırlar. Bu karmaşık koşulların çoğu yapısal olmayan verinin etkin bir şekilde saklanması ve analizi ile ilişkilidir. Merrill Lynch, potansiyel olarak kullanılan bütün verilerin yaklaşık %80'inin yapısal olmayan türde olduğunu ifade etmiştir. Bu büyük ve karmaşık yapıdaki yapısal olmayan veri analistlere yeni fırsatlar açmaktadır. Bu çalışmada, yapısal olmayan verinin metin ve web madenciliği yöntemleri ile yapısal hale dönüştürülmesi sonucu modele dahil edilmesinin, model başarısına yapacağı katkı analiz edilmiştir. Karar ağacı yöntemlerinden C5.0 algoritması kullanılarak elde edilen modeller birbirleri ile karşılaştırılmış ve en iyi model tespit edilmiştir.

Anahtar sözcükler: *Veri madenciliği; Metin madenciliği; Web madenciliği; Model karşılaştırma; Churn analizi.*

Abstract

Unstructured data analysis in data mining: Text and web mining

As data becomes large-scale, as megabytes become cheaper, as CPU speed becomes faster, we as analysts will be faced with more complex requirements. Many of these requirements will depend on the efficient storage and analysis of unstructured data. Merrill Lynch has recently estimated that over 80% of all potentially usable business information exists as unstructured data. The huge amount and complexity of unstructured data opens up many new opportunities for the analyst. In this study, we analyzed the improvement in the model success, which is a result of a extraction process of the useful information from unstructured data, using the text and the web mining methods. All models that are generated by using C5.0 algorithm are compared each other and then discovered which one is the best.

Keywords: *Data mining; Text mining; Web mining; Model comparison; Churn analysis.*

1. Giriş

Son yıllarda bilgi sistemleri ve teknolojinin gelişmesi sonucunda; kamu kurum ve kuruluşları, işletmeler ve diğer kuruluşlar veritabanlarında kuruluşun amacına ve yapısına bağlı olarak çeşitli türlerde veri toplamaktadır. Fakat bu veriler işlenmediği sürece anlamsız bir yığın olarak veritabanlarında depolanmaktadır [2, 3].

Uygun yazılımların gelişimi ve firmaların topladığı veriyi kullanılabilir bilgiye çevirme isteği toplanan bu veriyi işleyerek, verinin içerisindeki kullanılabilir ve ilginç ilişkilerin, birlikteliklerin ve örüntülerin (patterns) ortaya çıkarılmasını gerekli hale getirmiştir. Günümüzde pek çok kurum verilerini müşteri nitelikleri ve müşterilerin satın alma örüntülerine ilişkin yararlı, kullanışlı bilgiler elde edecek yöntemler ile işlemeye başlamamıştır. Ham veri zengini, nitelikli bilgi (knowledge) fakiri durumunda olan kurumların rekabetçi piyasada başarılı olmaları ve başarılarını sürdürmeleri her geçen gün daha da zorlaşmaktadır. Veri toplamanın önemini kavramış olan ve geçmişe yönelik veri tabanı sorgularıyla sadece sorgu bazlı bilginin elde edileceğini, veriden en üst düzeyde fayda sağlayamayacağını görmeye başlayan bütün kurumların en büyük yardımcısı veri madenciliğidir [7].

Veri madenciliği mevcut veriden anlamlı bilgileri, ilişkileri çıkarmada kullanılan tekniklere verilen genel isimdir. Veri madenciliği yapısal veriyi analiz edebilmekte iken; metin ve web madenciliği yapısal olmayan verinin, veri madenciliğinde kullanılmak üzere, yapısal hale dönüştürülmesinde kullanılmaktadır. Farklı birçok alanda kullanılabilen veri madenciliğinin alt alanlarından Metin ve Web Madenciliği bu çalışmada bir uygulama üzerinden incelenecek ve yapısal olmayan verinin metin ve web madenciliği yöntemleri ile yapısal hale dönüştürülmesi sonucu modele dahil edilmesinin model başarısına yapacağı katkı ortaya konulacaktır.

2. Veri, metin ve web madenciliği

Yapısal veri, bir yapı içerisinde organize edilebilen ve bundan dolayı tanımlanabilen veri için kullanılan bir terimdir. En yaygın kullanılan yapısal veri kaynakları SQL (Structured Query Language) ve Access gibi veri kaynaklarıdır. Örneğin SQL, kolon (değişken) ve satır (kayıt) bazlı bilginin seçimine imkan vermektedir. Yapısal veri, içerikteki veri tipine göre organize edilebilen ve arama yapılabilen veridir. Buna karşın yapısal olmayan verinin tanımlanabilir bir yapısı yoktur. En çok bilinen yapısal olmayan veri türleri; resim dosyaları, pdf, word ve text gibi metin dosyaları, web üzerinde tutulan log dosyaları ve e-postalardır. E-postalar veritabanlarında Microsoft Outlook gibi araçlar ile organize edilebilmesine rağmen bu tür veriler herhangi bir yapısal veri türü ile eşleşmediklerinden ham veri olarak düşünülür. Excel gibi hücre yapısına sahip veri türleri yapısal olmasına rağmen halen yapısal olma ve olmama konusundaki yeri tartışılmaktadır .

Birçok kurumun verisinin çoğu yapısal olmayan veri olarak veritabanlarında tutulmaktadır. Merrill Lynch, potansiyel olarak kullanılan bütün verilerin yaklaşık %80'inin yapısal olmayan türde olduğunu ifade etmiştir. [4, 10, 11].

Veri madenciliği büyük veri yığınlarında gizli olan örüntüleri ve ilişkileri ortaya çıkarmak için istatistik ve yapay zeka kökenli çok sayıda ileri veri çözümleme yönteminin tercihen görsel bir programlama ara yüzü üzerinden kullanıldığı bir süreçtir. Veri madenciliği algoritmaları; istatistiksel algoritmalar, matematiksel algoritmalar ve yapay zeka algoritmalarını (sinir ağları, karar ağaçları, kohonen ağlar, birliktelik kuralları vb.) bir arada içerir [7].

Veri madenciliği çözümleri ve algoritmaları metin veya web verisindeki kalıpları bulmadan veya model oluşturmadan önce metin veya web verisinin yapısal olması gerekmektedir. Metin ve Web madenciliği işlemleri, veri madenciliğinde kullanılacak yapısal veriye ulaşmak için kullanılan araçlar olarak tanımlanabilir.

Metin ve web madenciliği son yıllarda oldukça fazla çalışılan birbiri ile ilişkili alanlardır. Metin madenciliği, çok büyük belgelerin analizi ve metin tabanlı verinin içerisindeki gizli kalıpların elde edilmesidir. Web madenciliği ise, web içerikleri, sayfa yapıları ve web bağlantı istatistiklerinin de içinde olduğu web ile ilişkili olan verinin analizini içermektedir [10].

2.1. Metin madenciliği

Veri farklı şekillerde bulunabilir. Bazıları otomatik veri analizi için üstesinden gelinebilir ve uygun iken bazıların analizi çok daha zordur. Klasik veri analiz yöntemleri verinin değişken ve kayıt bazlı düzenlendiği varsayımı ile işlem yapmaktadır. Buradaki soru, eğer veri metin formatında yani kayıtların ve değişkenlerin olmadığı bir yapıda ise ne yapmamız gerektiğidir. Metin verisindeki anlamın ortaya çıkarılabilmesi için kullanılan yöntem metin madenciliğidir.

Metin yazımında standart kurallar olmadığından dolayı bilgisayar bunları anlayamamaktadır. Her bir metnin dili ve içerdiği anlam amaca bağlı olarak çeşitlilik göstermektedir. Yapısal olmayan bilgiden içerik çıkarmak için kullanılan geleneksel yöntemler; anahtar kelimeler veya mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan yöntemlerdir.

Bu yöntemler, hem sorgudaki hem de metindeki kelimelerin karakterlerini karşılaştıran bir temele dayanır. Bundan dolayı içeriği açıklayıcı sonuçlar elde edemez. Dili anlamının temeli dilbilimsel yollara dayanır ve bu çoğunlukla Natural Language Processing (NLP) olarak ifade edilir. NLP'yi içeren bir sistemde, karmaşık yapıların bulunduğu ifadeler (örneğin; duştan akan soğuk su ile içilen soğuk su arasındaki fark gibi) akıllı olarak çıkarabilmekte ve terimleri sınıflayarak; ürünler, organizasyonlar veya kişiler gibi sınıflara atamaktadır.

Metin madenciliği doğal dil metinlerinden bilgi ve nitelikli bilgi elde edilmesi sürecidir. İki aşamada gerçekleşir.

- Anahtar içerik/ifadeler metinden elde edilir,
- Elde edilen içerik/ifadeler, yüksek dereceden ilişkili olduğu kategorilere atanır.

Bu aşamaları basit bir örnek üzerinden açıklamak gerekirse;

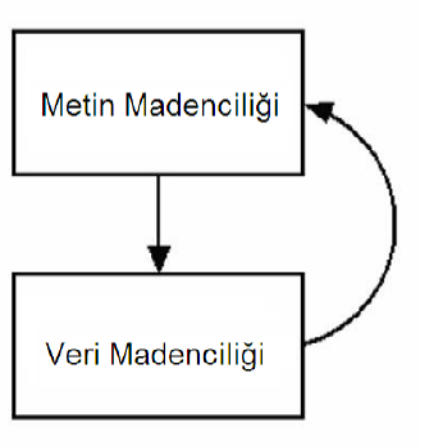
1. Aşama: "CPU" ve "CD-ROM" ifadeleri metinden elde edilir,
2. Aşama: Bu iki ifade, otomatik olarak "Bilgisayar Donanımı" etiketli kategoriye aktarılır.

Metin madenciliği uygulamaları iki ana sınıfta ayrılabilir:

- Metnin anlaşılması/özetlenmesi: Metin madenciliğinin amaçlarından bir tanesi metinden anlamlı nitelikli bilginin çıkarılmasıdır. Böylece metnin içerdiği anahtar içerik anlaşılabilecektir. Örneğin, yavaş tamir veya sipariş gibi sorunlar yüzünden şikayet eden müşterilerin oranını öğrenmek isteyebiliriz.
- Metin ile modelleme: Daha yaygın olarak, metin madenciliği terk etme veya ürün alma gibi müşteri davranışlarının tahmin edildiği bir modelin geliştirilmesi aşamasının bir bölümünü oluşturmaktadır. Metinden elde edilen içerik girdi değişkeni olarak kullanılır ve diğer bilgiler ile beraber öngörüsel model geliştirilir.

Veri madenciliği girdi olarak sadece yapısal veriyi kullandığından dolayı veri madenciliği çözümleri ve algoritmaları kullanılarak metin verisinden kalıplar bulunup, modeller kurulmadan önce metinden elde edilecek bilginin yapısal hale dönüştürülmesi zorunludur. Metin madenciliği sonucunda, kategorilerin oluşturulması ile yapısal olmayan veri yapısal hale dönüşmektedir [5, 9, 12].

Metin ve veri madenciliği arasındaki ilişki Şekil 1'de tanımlanmıştır;



Şekil 1. Süreçler arasındaki ilişki

Şekil 1.'de de görüldüğü gibi, metin ve veri madenciliği arasında interaktif bir ilişki vardır. Metin madenciliği sonucunda elde edilene yapısal veri, veri madenciliği modellerinde kullanılmakta ve elde edilen sonuçlar daha sonra metnin yapısının incelenmesinde kullanılmaktadır.

Metin madenciliğinin uygulama alanlarından bazıları;

- Müşteri ilişkileri yönetimi (Customer Relationship Management, CRM): Bütün müşterilerin e-mail, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen metin bilgilerinden nitelikli bilgi çıkarılır. Bu nitelikli bilgi müşterinin terk etme ve çapraz satışlarını tahmin etmek üzere kullanılır.

- Sahtekarlık (Fraud) tespiti: Sağlık, sigorta ve hükümet tarafında toplanan büyük çaptaki metin verilerinde kalıplar ve anormallikler aranarak sahtekarlıklar tespit edilir.

- Bilimsel ve medikal araştırmalar: Hasta raporları, makale başlıkları, yayınlanmış araştırma sonuçları ve diğer yayınlar gibi metin materyallerinden çıkarım yapılır.

- Güvenlik/istihbarat: Organizasyonlar ve bireyler arasındaki kalıplar ve bağlantılar, terörist tehlikeleri ve kriminal davranışları tahmin etmek ve engelleyebilmek için büyük çaptaki metin içerisinde aranır.

- Pazar araştırması: Yayınlanmış belgeler, basın bültenleri ve web sayfaları pazar etkisinin ölçülmesi için aranır ve izlenir. Metin madenciliği kantitatif yöntemler ile açık uçlu anket soruları ve mülakatların değerlendirilmesinde kullanılabilmektedir [5, 12].

2.2. Web madenciliği

Web madenciliği işlemleri kullanılarak yapısal olmayan web verileri yapısal veriye dönüştürülür. Web madenciliği uygulamaları temel olarak üç alt başlık altında toplanabilir;

- Web yapı madenciliği: Web yapı madenciliği ile internetin temel yapısını oluşturan web siteleri, web sayfaları arası ya da web sayfasındaki bağlantılar arasındaki ilişkiler incelenir.

- Web içerik madenciliği: Web içerik madenciliği ile web sayfalarının içerikleri incelenir ve kullanışlı bilgi çıkarımı sağlanır. Web içerik madenciliği kullanarak web sayfalarının başlıkları, içerisinde geçen kelimeler, resimler veya müzik dosyaları incelenir. Bulunan içeriklere göre web siteleri belirli sınıflara veya kümelere ayrılabilir.

- Web kullanım madenciliği: Web kullanım madenciliği ile web sunucularında tutulan kullanıcı erişim kayıtları incelenerek anlamlı ve faydalı kalıplar bulunabilir. Web kullanım madenciliği yöntemleri uygulanarak web sitelerini ziyaret eden kişilerin davranış ve tutumları belirlenebilir.

Web madenciliğinin günümüzde birçok alanda kullanılmasının en önemli sebebi; kişilerin web sayfalarında göstermiş oldukları davranışların, hareketlerin ve yapmış oldukları işlem bilgilerinin var olan iş süreçlerine entegrasyonunu sağlayarak müşterinin en iyi şekilde anlaşılmasını sağlayan müşteri odaklı bir sistem oluşturmasıdır.

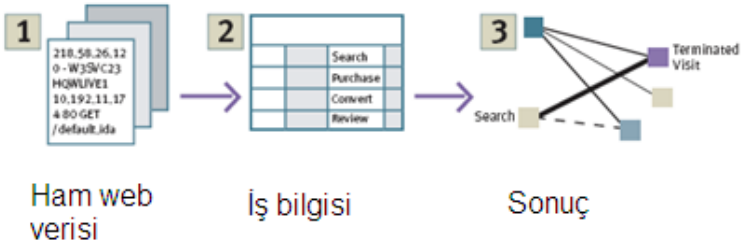
Web madenciliği kullanım alanları aşağıdaki gibidir;

- Web üzerinden ürün satışı gerçekleştiren şirketler web verilerini analiz ederek müşteri profili ve kümeleri oluşturmaktadırlar.

- Google vd. arama motorları web içerik madenciliği uygulayarak aranan anahtar kelimeyi içeren web sitelerini belirlemektedirler.

- Web madenciliği uygulanarak web sitelerinin iyileştirilmesi ve güncel kalması sağlanmaktadır [1, 6].

Web madenciliğindeki süreç Şekil 2'de tanımlanmıştır.



Şekil 2. Web madenciliği süreci

Şekil 2'de görüldüğü gibi, yapısal olmayan web verisi (log dosyaları, vd) iş bilgisi bazlı bir kategori işleminden sonra yapısal hale dönüşmekte ve işlenebilir duruma gelmektedir.

Metin ve web madenciliği hakkındaki genel süreç uygulamanın yer aldığı Üçüncü Bölümde daha detaylı anlatılacaktır.

3. Uygulama

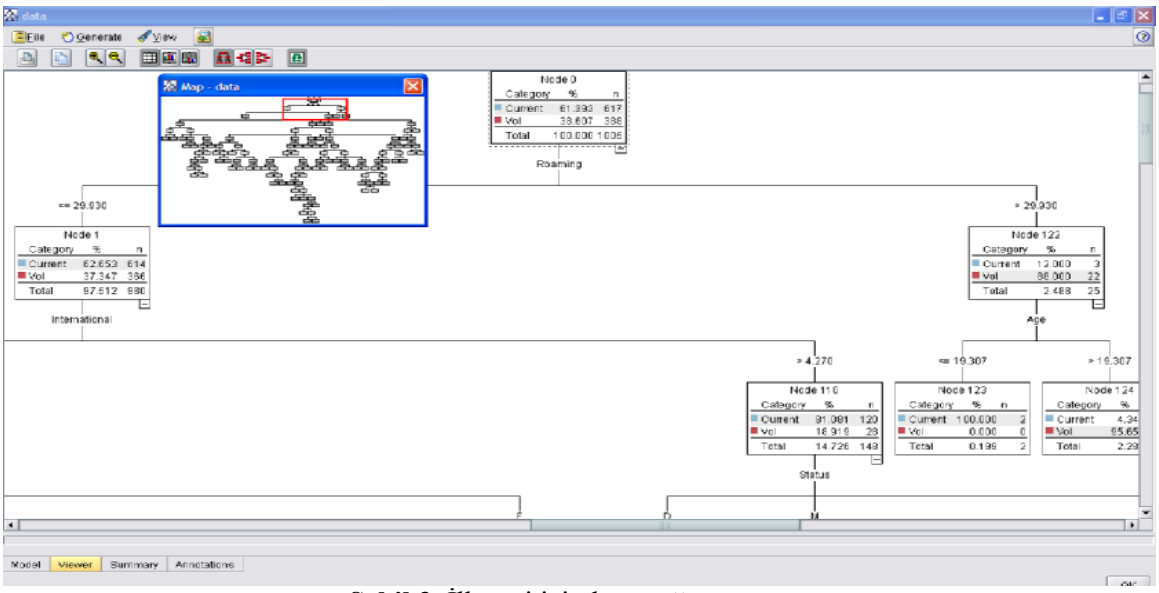
Uygulamada Clementine 12.0 kullanılarak bir telekomünikasyon kurumunun 2070 müşterisine ait 17 değişkenden oluşan şirketi terk etme (churn) yapısal verisi kullanılarak, terk eden müşterilere ait bir profil modeli, karar ağacı algoritmalarından C5.0 kullanılarak elde edilmiştir.

Ayrıca; çağrı merkezlerinden elde edilen müşterilere ait metin dosyası kullanılarak elde edilen yapısal veri var olan yapısal veriye eklenerek ikinci bir veri ve ikinci bir model, müşterilere ait internet üzerinden elde edilen web log dosyası kullanılarak elde edilen yapısal veri ikinci veriye eklenerek üçüncü bir model elde edilmiştir. Bu bölümde, söz edilen üç model karşılaştırılmış ve sonucu açıklanmıştır.

3.1. Veri madenciliği

Kurulan ilk model, 2070 müşteriye ait 17 değişkenden oluşan ve yapısal veri içeren veri dosyası kullanılarak elde edilmiştir. Model, karar ağacı algoritmalarından C5.0 algoritması kullanılarak elde edilmiştir.

Şirketi terk etme değişkeni bağımlı, şehir içi görüşme süresi (saniye), şehirler arası görüşme süresi (saniye), hattın kesilme sayısı, ödeme yöntemi (nakit, kredi kartı, otomatik), tarife bilgisi, kullanıcının cinsiyeti, medeni durumu, yaşı gibi 17 değişken ise bağımsız değişken olarak seçilerek algoritmada kullanılmıştır. Şekil 3'de karar ağacının sonucu verilmiştir.



Şekil 3. İlk veri için karar ağacı sonucu

Şekil 3'deki sonuçlara göre, şirketi terk etmede en önemli değişken yurtdışı dolanım (roaming) olarak bulunmuştur. Karar ağacı modelleri bu çalışmanın ana amacı olmadığından detaylı olarak anlatılmamıştır.

3.2. Metin madenciliği

Metin madenciliği ile ilgili yapılanlar genel hatları ile aşağıdaki şekillerde anlatılmaya çalışılmıştır.

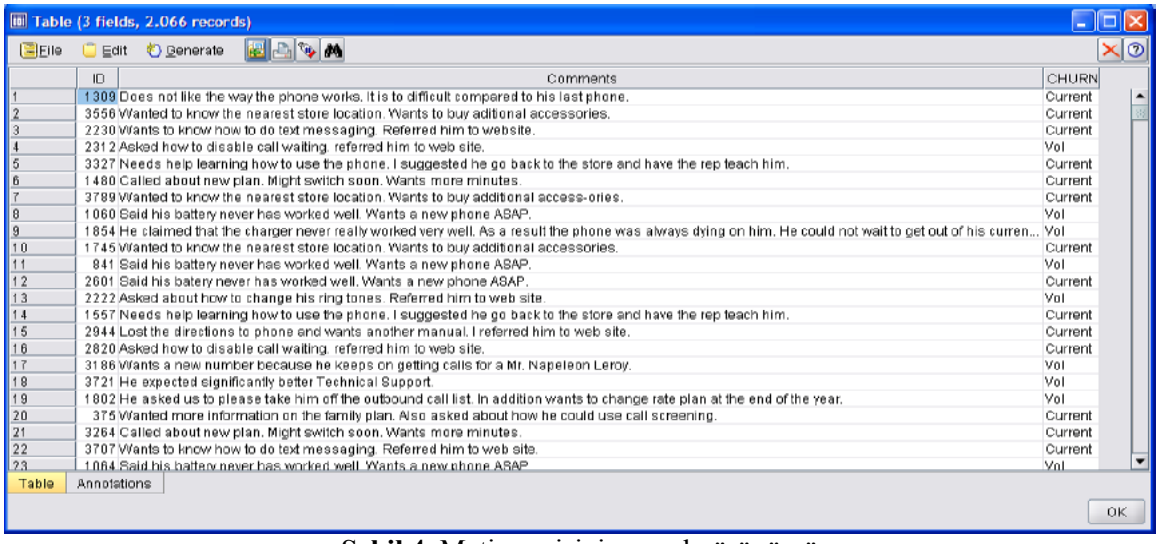


Table (3 fields, 2,066 records)

File Edit Generate

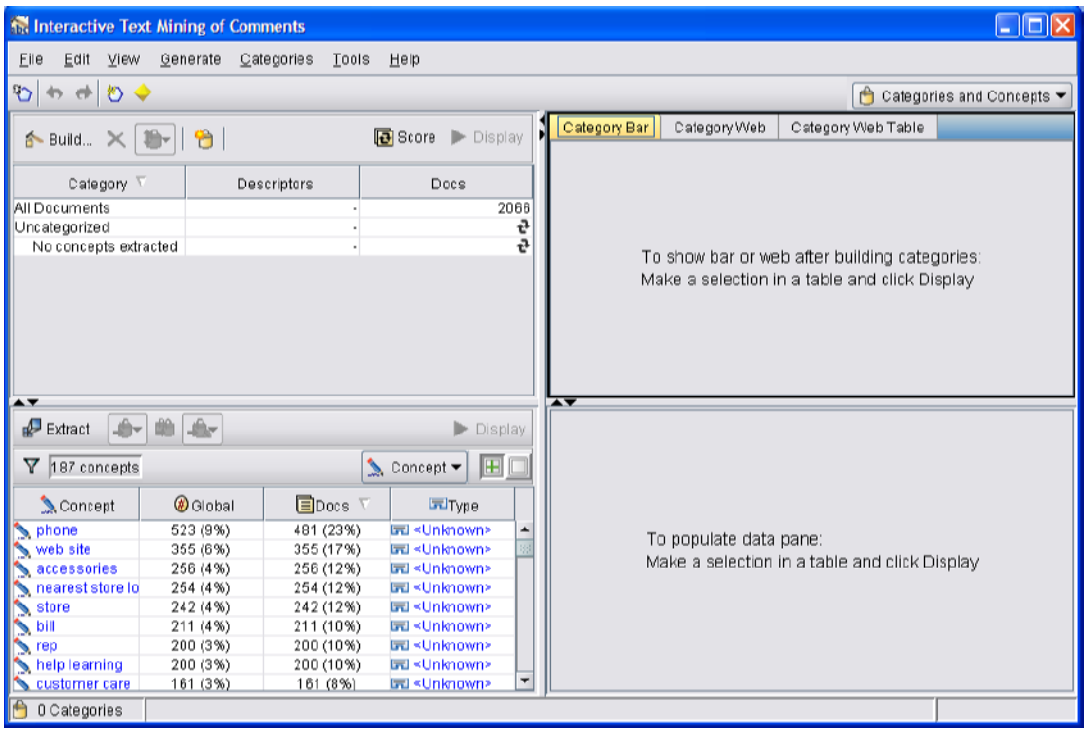
	ID	Comments	CHURN
1	1309	Does not like the way the phone works. It is to difficult compared to his last phone.	Current
2	3556	Wanted to know the nearest store location. Wants to buy aditional accessories.	Current
3	2230	Wants to know how to do text messaging. Referred him to website.	Current
4	2312	Asked how to disable call waiting. referred him to web site.	Vol
5	3327	Needs help learning how to use the phone. I suggested he go back to the store and have the rep teach him.	Current
6	1480	Called about new plan. Might switch soon. Wants more minutes.	Current
7	3789	Wanted to know the nearest store location. Wants to buy additional access-ories.	Current
8	1060	Said his battery never has worked well. Wants a new phone ASAP.	Vol
9	1854	He claimed that the charger never really worked very well. As a result the phone was always dying on him. He could not wait to get out of his curren...	Vol
10	1745	Wanted to know the nearest store location. Wants to buy additional accessories.	Current
11	841	Said his battery never has worked well. Wants a new phone ASAP.	Vol
12	2601	Said his batery never has worked well. Wants a new phone ASAP.	Current
13	2222	Asked about how to change his ring tones. Referred him to web site.	Vol
14	1557	Needs help learning how to use the phone. I suggested he go back to the store and have the rep teach him.	Current
15	2944	Lost the directions to phone and wants another manual. I referred him to web site.	Current
16	2820	Asked how to disable call waiting. referred him to web site.	Current
17	3186	Wants a new number because he keeps on getting calls for a Mr. Napeleon Leroy.	Vol
18	3721	He expected significantly better Technical Support.	Vol
19	1802	He asked us to please take him off the outbound call list. In addition wants to change rate plan at the end of the year.	Vol
20	375	Wanted more information on the family plan. Also asked about how he could use call screening.	Current
21	3264	Called about new plan. Might switch soon. Wants more minutes.	Current
22	3707	Wants to know how to do text messaging. Referred him to web site.	Current
23	1064	Said his battery never has worked well. Wants a new phone ASAP.	Vol

Table Annotations

OK

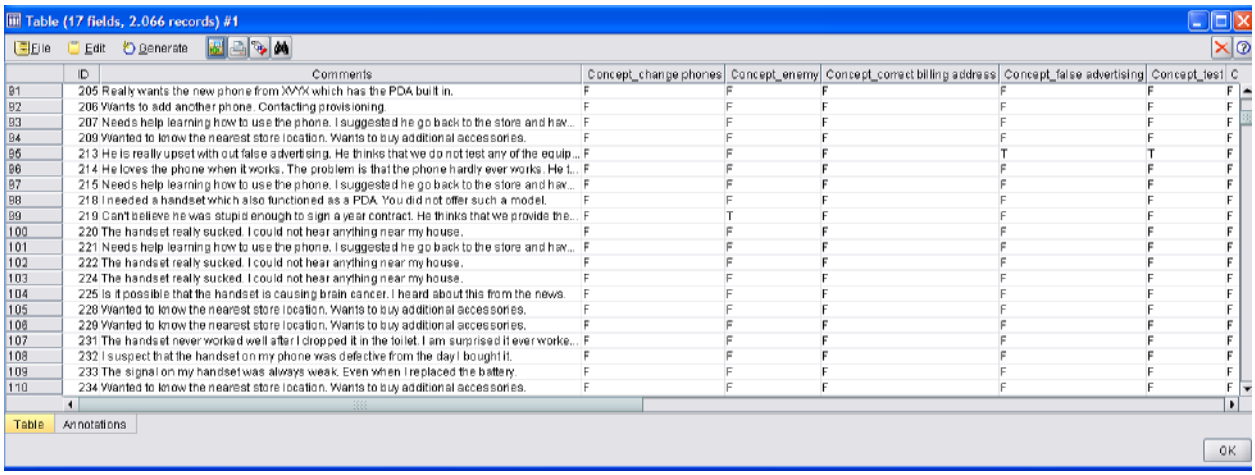
Şekil 4. Metin verisinin genel görünümü

Şekil 4'de metin verisi ile ilgili genel görünüm yer almaktadır. Her bir ID'ye ilişkin bir metin bilgisi (müşterilerin yorumlarını içeren metin alanı) ve şirketi terk etme değişkeni (CHURN) yer almaktadır.



Şekil 5. Metin verisinin analiz aşaması

Şekil 5'de metin verisinin analiz aşaması ile ilgili ekran görüntüsü yer almaktadır. Şekil 5'in sol alt kısmında yer alan görüntüde metinden elde edilen içerikler yer almaktadır.



Şekil 6. Yapısal veri-değişken ve kayıt bazlı gösterim

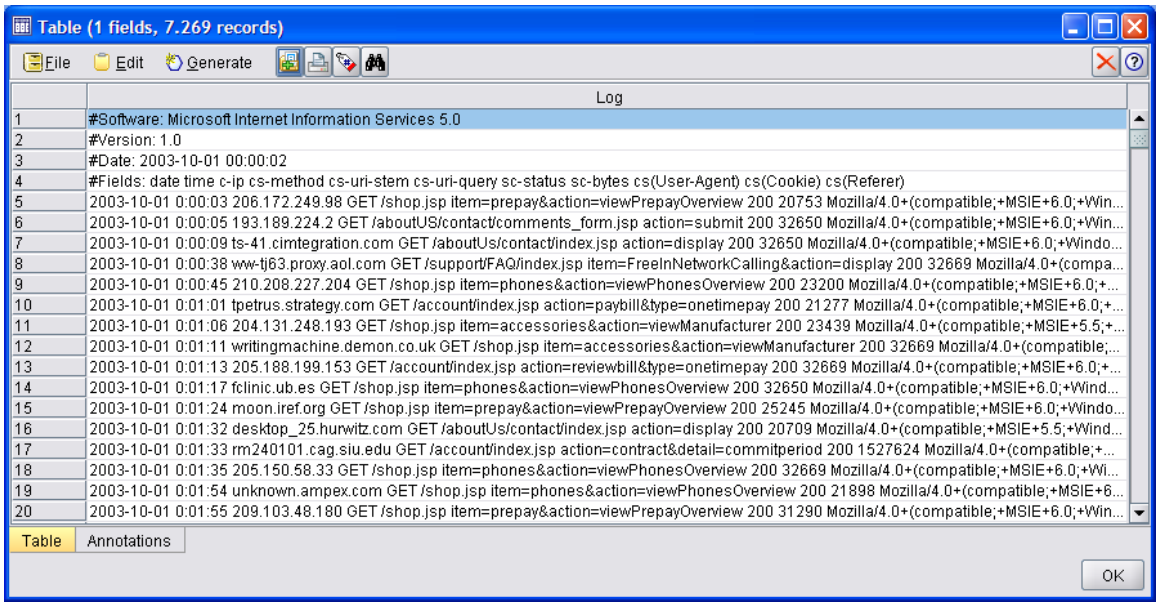
Şekil 6'da ise her bir ID'ye karşılık gelen metin dosyalarının, yapısal şekle nasıl dönüştüğü görülmektedir. Görüldüğü gibi her bir metnin (yapısal olamayan şekil) yanında, O metnin hangi kategoriye atandığı bilgisi (yapısal şekil) yer almaktadır. Her bir metnin hangi kategoriye atandığı bilgisi, ilgili kategorideki "T" harfi, hangi kategoride yer almadığı bilgisi ilgili kategorideki "F" harfi ile kodlanmıştır.

Metin verisinin metin madenciliği işlemi sonucunda yapısal şekle dönüştürülmesi ile elde edilen verinin, birinci veri ile birleştirilmesi ile ikinci veri elde edilmiştir. İkinci model, ikinci veri kullanılarak elde edilmiştir. Bu işlemin ve aslında bu makalenin asıl amacı, yapısal olmayan veri içerisindeki bilginin modele eklenmesi durumunda model başarısının arttığının gösterilmesidir.

Bir sonraki aşamada, web madenciliğinden gelen yapısal veride ikinci veriye eklenecek ve elde edilen yeni veriden yeni bir model oluşturulacaktır.

3.3. Web madenciliği

Web madenciliği ile ilgili yapılanlar genel hatları Şekil 7-9 ile anlatılmaya çalışılmıştır.



Şekil 7. Log dosyasının genel görünümü

Şekil 7'deki log dosyalarının yapısal olmadığı görülmektedir. Dosyada sırasıyla, hangi tarihte web sayfasına erişim sağlandığı, kullanıcının IP adresi, istek tipi (GET veya POST), hangi web sayfasına erişim sağlandığı, statü (200 veya 300), boyut (gönderilmiş olan dosyanın byte cinsinden boyutu) ve hangi web tarayıcısının (Mozilla, Explorer, vd) kullanıldığı gibi bilgiler yer almaktadır.

Log dosyalarının incelenerek web sitesinin yapısının ortaya konduğu tanımlama dosyası, olay dosyası (event definition)'dır. Web sunucularından elde edilen yapısal olmayan log dosyaları olay dosyasında yapılan tanımlamalardan yola çıkılarak yapısal bir hale getirilir. Web madenciliği işlemcisinin çalışması için mecburi bir dosyadır.

Standart bir olay dosyası 4 temel alandan oluşur:

1. Olay kategorisi (event category),
2. Olay ismi (event name),
3. Olay tanımı (event definition),
4. Olay nitelikleri (event attributes).

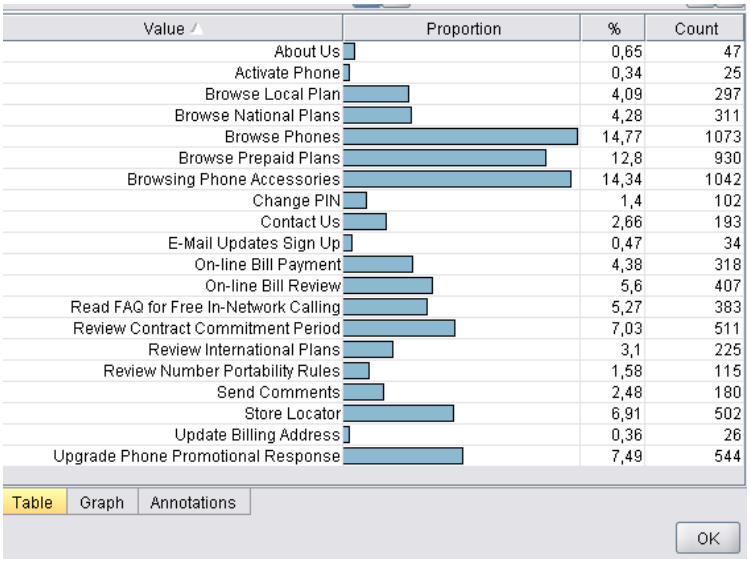
Olay kategorisi: Olayları anlamlı gruplar altında toplamak için kullanılır. İstenilen bir ifade tanımlanabilir.

Olay ismi: Olayı açıklayan bölümdür. İstenilen bir ifade tanımlanabilir. Ancak olay isimleri tekil olmalıdır, olay dosyasında her olay ismi bir kere kullanılmalıdır.

Olay tanımı: Web madenciliği işlemcisinin log dosyalarında bulduğu sayfalar ile tanımlanan olay dosyası arasında eşleştirme yapması için kullanılacak alandır.

Olay nitelikleri: İlgili olayla hangi özel parametrenin kullanıldığını gösteren bilgidir. Tek bir olay için birden fazla nitelik tanımlanabilir [1, 6, 13].

Şekil 7'de görülen log dosyaları Web Mining for Clementine 12.0 ile analiz edilmiş ve yapısal olmayan log dosyası Şekil 8 ve 9'da görüldüğü gibi kullanılabilir olan yapısal şekle dönüştürülmüştür.



Value	Proportion	%	Count
About Us		0,65	47
Activate Phone		0,34	25
Browse Local Plan		4,09	297
Browse National Plans		4,28	311
Browse Phones		14,77	1073
Browse Prepaid Plans		12,8	930
Browsing Phone Accessories		14,34	1042
Change PIN		1,4	102
Contact Us		2,66	193
E-Mail Updates Sign Up		0,47	34
On-line Bill Payment		4,38	318
On-line Bill Review		5,6	407
Read FAQ for Free In-Network Calling		5,27	383
Review Contract Commitment Period		7,03	511
Review International Plans		3,1	225
Review Number Portability Rules		1,58	115
Send Comments		2,48	180
Store Locator		6,91	502
Update Billing Address		0,36	26
Upgrade Phone Promotional Response		7,49	544

Table Graph Annotations

OK

Şekil 8. Yapısal veri-grafiksel gösterim

Şekil 8'de görüldüğü gibi log dosyasında yer alan veriler, olay dosyası baz alınarak çeşitli kategorilere dönüştürülmüştür. Örneğin, analiz edilen bu log dosyası içerindeki kayıtların %0,65'inin "About Us" sayfasına giriş yapan müşterilerden oluştuğu artık bilinmektedir.

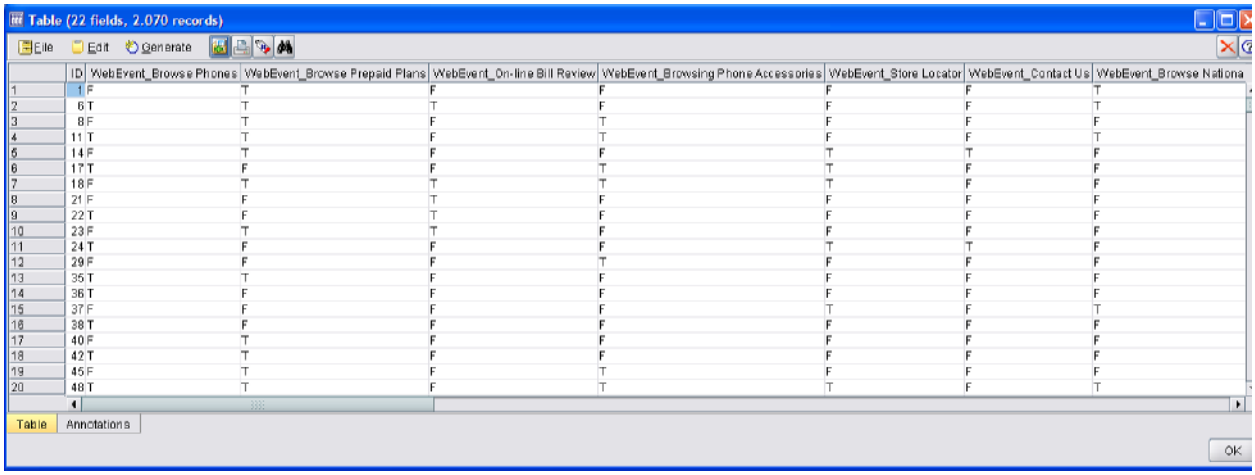


Table (22 fields, 2.070 records)

File Edit Generate

	ID	WebEvent_Browse Phones	WebEvent_Browse Prepaid Plans	WebEvent_On-line Bill Review	WebEvent_Browsing Phone Accessories	WebEvent_Store Locator	WebEvent_Contact Us	WebEvent_Browse Nationa
1	1	F	T	F	F	F	F	T
2	6	T	T	F	F	F	F	T
3	8	F	T	F	T	F	F	F
4	11	T	F	F	T	F	F	T
5	14	F	T	F	F	T	T	F
6	17	T	F	F	T	T	F	F
7	18	F	T	T	T	T	F	F
8	21	F	F	T	F	F	F	F
9	22	T	F	T	F	F	F	F
10	23	F	T	T	F	F	F	F
11	24	T	F	F	F	T	T	F
12	29	F	F	F	T	F	F	F
13	35	T	T	F	F	F	F	F
14	36	T	F	F	F	F	F	F
15	37	F	F	F	F	T	F	T
16	38	T	F	F	F	F	F	F
17	40	F	T	F	F	F	F	F
18	42	T	T	F	F	F	F	F
19	45	F	T	F	T	F	F	F
20	48	T	T	F	T	T	F	T

Table Annotations

OK

Şekil 9. Yapısal veri-değişken ve kayıt bazlı gösterim

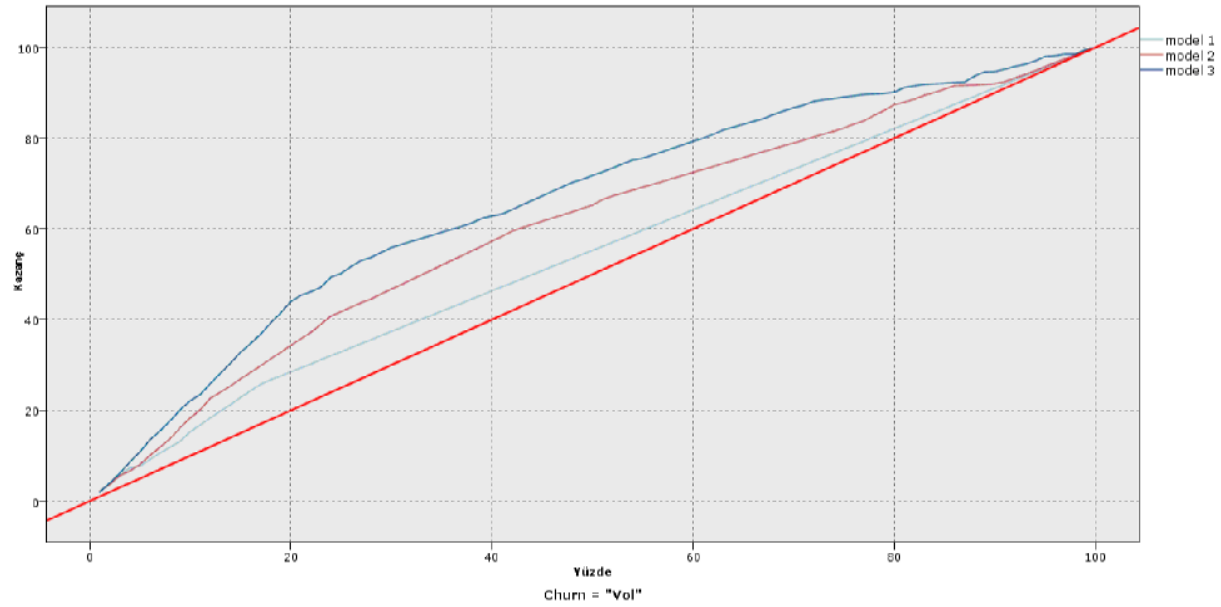
Şekil 9'da ise her bir ID'ye karşılık gelen log dosyalarının, web madenciliği işleminden sonra yapısal şekle nasıl dönüştüğü görülmektedir. Log dosyaları tanımlanan ilgili olay isimleri kategorilerine atandığı ("T" ve "F" harfleri ile) görülmektedir.

Hem metin hem de web madenciliği işlemlerindeki amaç daha öncede açıklandığı gibi, yapısal olmayan verinin yapısal şekle dönüştürülmesidir. Şekil 6 ve Şekil 9'da anlatılmak istenen bu yapı açıkça görülmektedir.

Web verisinin web madenciliği işlemi sonucunda yapısal şekle dönüştürülmesi ile elde edilen verinin, ikinci veri ile birleştirilmesi ile üçüncü veri elde edilmiştir. Üçüncü model, üçüncü veri kullanılarak elde edilmiştir.

Başlangıçta var olan yapısal verinin kullanıldığı Model 1, metin madenciliği ile elde edilen yapısal verinin var olan yapısal veri ile birleştirilmesinden elde edilen verinin kullanıldığı Model 2 ve web madenciliğinden elde edilen verinin de eklenmesiyle elde edilen verinin kullanıldığı Model 3'ün karşılaştırması Şekil 10'da verilmiştir.

Şekil 10'daki grafik, elde edilen üç karar ağacı modelini kazanç yüzdesi (gain) ölçütü ile karşılaştıran kazanç grafiğini (gain chart) göstermektedir. Grafikte; Y ekseni kazanç, X ekseni ise erişilebilecek kayıtları (bu uygulamada müşteriler) göstermektedir.



Şekil 10. Modellerin karşılaştırılması

Karar ağaçları ile beraber her bir adıma ilişkin kazanç (%) değerleri elde edilir. Elde edilen bu değerlerde kazanç grafiği üzerinde yer alır. Kazanç, ilgili adımdaki hedef kategori sayısının geneldeki hedef kategori sayısına oranıdır. Köşegendeki doğru (kırmızı grafik), hiçbir modelin kullanılmadığı durumda tüm örneklem için beklenen olumlu cevapları temsil eder.

Uygulamada bu tür bir grafik için beklenen, ilk %20'lik dilimde (X ekseni), model kazanç değerlerinin yaklaşık %50 ve üzerinde olmasıdır. Yani, mevcut verinin %20'sini kullanarak, model kazancının yüksek olması beklenmektedir [5, 8, 13].

Model 1 için bu grafik yorumlandığında; mevcut kayıtların %20'sine ulaşıldığında model kazancının yaklaşık %30 olması beklenmektedir. Buna göre, Model 3'ün kazancının yaklaşık %45 ile diğer modellerden fazla olduğu açıkça görülmektedir.

4. Sonuç ve öneriler

Yapısal veri kullanılarak elde edilen model ile yapısal olmayan verinin metin ve web madenciliği yöntemleri kullanılarak yapısal hale getirilen ve buradan elde edilen model karşılaştırılmıştır. Metin ve web madenciliği yöntemleri kullanılarak elde edilen modelin sonuçta daha başarılı olduğu görülmüştür (Şekil 10). Yapısal olmayan verideki nitelikli bilginin çıkarılıp modele entegre edilebilmesi ile en son modelin daha başarılı olduğu sonucu beklenmeyen bir olgu değildir.

Öngörüsel diğer model algoritmaları (CHAID, C&RTree, Lojistik Regresyon, vd.) kullanılarak yeniden modelleme yapılması ve algoritmalar arasında hangisinin daha başarılı olduğu sonucunun tespit edilmesi diğer bir çalışmaya bırakılmıştır.

Dünya üzerindeki potansiyel olarak kullanılan bütün verilerin yaklaşık %80'inin yapısal olmayan türde olduğu düşünüldüğünde, bu verilerin kullanılması kesinlikle araştırmalara katma değer katacaktır.

Kaynaklar

[1] Chakrabarti, S. (2003), Mining the Web: Discovering Knowledge from Hypertext Data, *Morgan Kaufmann Publishers*, San Francisco.

[2] Dolgun, M.Ö. (2006), Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, *Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü*, Ankara.

[3] Han, J., Kamber, M. (2001), Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, San Francisco.

[4] Hearst, M. (2009), What is text mining, *http://www.sims.berkeley.edu/~hearst/textmining.html*.

[5] Introduction to Text Mining (2008), *SPSS Inc*.

[6] Liu, B. (2007), Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, *Springer*.

[7] Özdemir Güzel, T., Dolgun, M.Ö., Şatır, U., Deliloğlu, S., Korkmaz, H.E. (2007), 2005 Yılı Öğrenci Seçme Sınavı (ÖSS) Verileri Kullanılarak Öğrenci Profilinin Belirlenmesi, *5. İstatistik Kongresi*, Antalya.

[8] Shapiro-Piatetsky, G., Steingold, S. (2000), Measuring Lift Quality in Database Marketing, *ACM SIGKDD Explorations Newsletter,* 2(2), 76-80.

[9] Sholom M.W., Indurkhya N., Zhang T., Damerau F. (2004), Text Mining: Predictive Methods for Analyzing Unstructured Information, *Springer.*

[10] Tan, A.H., Yu, P.S. (2004), Guest Editorial: Text and Web Mining, *Applied Intelligence 18, 239-241, Kluwer Academic Publisher*.

[11] Unstructured data (2009), *http://en.wikipedia.org/wiki/Unstructured_data*.

[12] W. Fan, L. Wallace, S. Rich, Z. Zhang. (2006), Tapping into the power of text mining, *Communications of ACM*, 49(9), 76-82.

[13] Web Mining for Clementine 12.0 User's Guide (2007), *SPSS Inc*.