SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 14

Yrd. Doç.Dr. Nilüfer YURTAY

14.1 Web Madenciliği

Son dönemde bilişim teknolojilerindeki gelişmeler çeşitli iş alanlarındaki birçok uygulamanın bilgisayar ortamına aktarılmasını sağlamıştır. Veriler birçok organizasyonda kritik kaynak haline gelmiş, verilere etkin bir şekilde ulaşabilmek, verilerin paylaşımı, verilerden kullanılabilir bilgiler (knowledge) çıkarabilmek ve elde edilen bilgilerin etkin kullanımı gibi konular önem kazanmıştır.

Kullanılabilir bilgiler elde edebilmek ve veriler arasındaki ilişkileri ortaya çıkarmak amacıyla Veri Madenciliği yöntemleri kullanılmaktadır. Veri madenciliği, istatistik, veri tabanları ve veri yönetimi, yapay zeka, örüntü tanıma ve konularını kapsamaktadır. Günümüzde "World Wide Web" birçok uygulama için vazgeçilmez bilgi kaynaklarından biri haline gelmiştir. Internette verilerden anlamlı bilgiler elde edebilmek için kullanılan veri madenciliği yöntemi Web Madenciliği'dir.

Web madenciliği kısaca Web sayfaları ve servislerinden otomatik olarak bilgi çekip bunlardaki kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanabilir.

Web madenciliği veri madenciliği tekniklerinin www (World Wide Web) verileri üzerinde uygulanmasını konu alır. Web madenciliğini üç ana başlıkta inceleyebiliriz: Web içerik madenciliği, web yapı madenciliği ve web kullanım madenciliği. Web kullanım madenciliği kullanıcıların web sitelerindeki davranışlarını inceler[1].

Web kullanım madenciliği kullanıcının siteyi kullanırken gerisinde bıraktığı erişim verilerinden bilgi üretmeyi amaçlar. Bu veriler ikinci sınıf verilerdir, yani bir yere girilmiş; bir yerde yazılan, ya da kullanıcının isteğiyle ulaşın veriler değildir. Tamamen kullanıcıdan bağımsız oluşur ve çok ciddi boyutlardadır. Bir e-ticaret sitesinin web kullanım verileri kullanıcı hareketlerini takip etme açısından değerli veriler içerir. Bu sayede site güncelleştirme, sistem iyileştirme ve kullanıcılara kişiselleştirilmiş hizmetler sunmak mümkün olmaktadır.

Bir e-ticaret sitesinin web kullanım verilerinden aşağıdaki analiz ve yorumlara varılabilinir;

Analiz Kısmı:

- Siteyi bugün kaç kişi ziyaret etti?
- Siteye kimler link vermiş?
- En çok hangi sayfadan sonra site terkedilmiş ?
- En çok ziyaret edilen sayfa hangisi?
- Siteyi internette bulabilmek için hangi anahtar kelimeler kullanılmış?
- Kullanım sürelerinin günlere ve saatlere göre dağılımı
- Sayfalara göre istemlerin dağılımı
- Ulaşılmayan sayfalar
- Ulaşılamayan linkler
- İstemlerin statülerine göre dağılımı
- Siteye saldırı var mı yok mu?

[1] http://www.enformatikseminerleri.com/

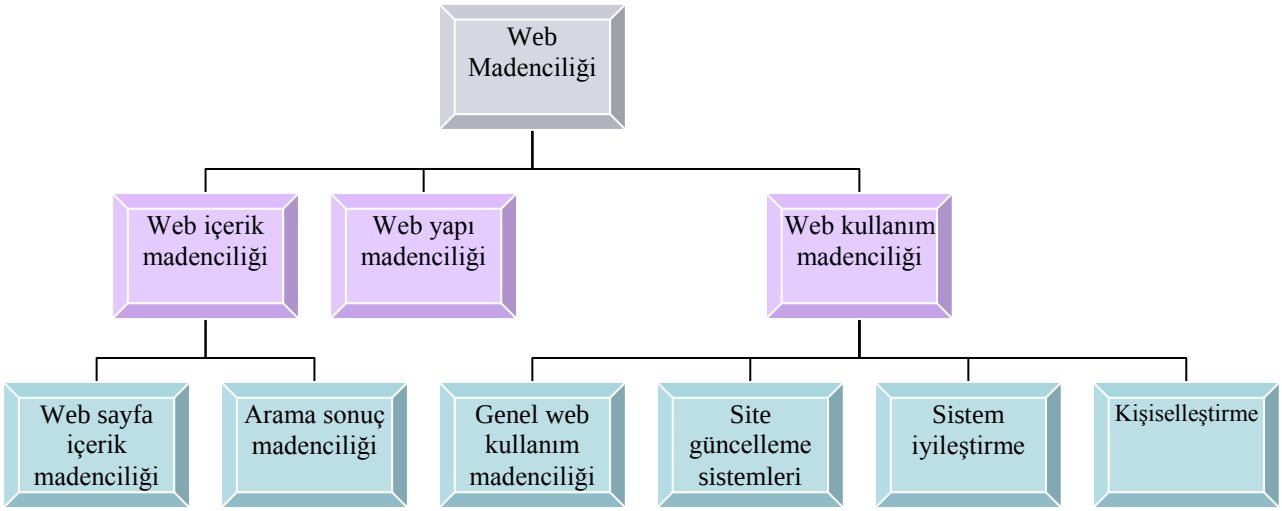
Yorum Kısmı:

- Kullanıcıların profilleri
- Kullanıcıların zaman içindeki değişimleri
- Sitede beğenilen sayfalar,beğenilmeyen sayfalar
- Kullanıcıların gezinti şekli/hızı
- Sitenin içeriği nasıl olmalıdır?

İnternet üzerindeki veri yığınlarını web sayfaları, Access Log dosyaları, Kullanıcı kayıt bilgileri, Oturum ve hareket bilgileri ve Site yapısı ve içeriği olarak sıralayabilriz.

Yukarıda sayılan çeşitli yapıdaki web sayfaları dokümanlarını ve kayıt bilgilerini incelemek, bunlardaki kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanabilir.

Web madenciliğindeki veri kaynakları; web İçerik madenciliği, web yapı madenciliği, web kullanım madenciliği olarak üç sınıfta incelenebilir(Şekil 14.1).



Şekil 14.1 Web madenciliği[2]

Web içerik madenciliği , web kaynaklarının içeriklerinden yararlı bilgiyi elde etmek olarak tanımlanabilir. Son zamanlarda XML dili de bu konuda kullanılmaya başlanmıştır.

Web yapı madenciliğinin amacı web sayfaları arasındaki linkleri takip ederek bilgi üretmektir. 2 ayrı veri tipine göre yapılmaktadır: Hyperlink web sayfası verisi ve HTML yada XML sayfa verisi.

Web kullanım Madenciliğinde kullanılan veriler, web üzerindeki çeşitli sunucularda tutulan kullanıcı erişim hareketlerinin yer aldığı çeşitli log dosyalarından elde edilir. Web kullanım madenciliği bir veya birçok web sunucudan kullanıcı erişim desenlerinin otomatik keşfinin ve analizin yapıldığı bir tip veri

[2] http://www.teknoturk.org/docking/yazilar/tt000119-yazi.htm

madenciliği etkinliğidir. Birçok organizasyon Pazar analizleri için geliştirdikleri stratejileri ziyaretçi bilgilerine dayanarak yerine getirir .

Web kullanım madenciliği yapılırken kullanılan bilgilerin çoğu günlük dosyalarındaki bilgilerdir. Günlük dosyalarına web sitesinden istenen her sayfa bir kayıt olarak eklenir.

Web kullanım madenciliği;

- İlk işlem
- Desen keşif
- Desen analiz

aşamalarından oluşmaktadır.

İlk işlem aşamasında kullanım verisi ilişkisiz sahalardan temizlenir. Veri temizleme, kullanıcı tanıma ve oturum tanıma işleri yerine getirilir.

Desen keşfi aşamasında ilk işlemden geçirilen veriye istatistik, uyum kuralları, kümeleme, sınıflama ve sıralı desenler gibi teknikler uygulanarak desen bulunmaya çalışılır.

Desen analiz aşamasında ise bir önceki aşamada bulunan desenler üzerinde derinlemesine analizler yapılır. Analiz aracı olarak SQL benzeri diller ile OLAP işlemleri kullanılabilir [3]

Veri madenciliğinde yapılan işlerin bir kısmı anormal durumların tespiti ile ilgilidir. Veri madenciliği tekniklerinden birisi ise istisna saptanmasıdır. Kredi kartı yolsuzluklarını tespit için kullanılan bu yöntem saldırı tespitine yakın bir konudur.

Veri madenciliği ile saldırı tespiti yapılmasının en önemli bir tane nedeni vardır o da daha önceden meydana gelmemiş bir saldırıyı tanımadır. Veri madenciliği kullandığı kümeleme tekniği ile ilk olarak meydana gelen bir durumu tanıyabilmektedir. Kümelemede kullanıcılar genel özelliklerine dayalı olarak gruplara ayrılmaktadırlar.

14.2 İstatistiksel Sınıflandırma Modelleri

Sınıflandırma işlemi için kullanılan diğer tekniklerden biri de istatistiksel tekniklerdir. Bayes teoremi de bu teknikler içerisinde sıkça başvurulan bir teoremdir. Buna bağlı olarak bayes ağları kullanılır.

14.2.1 Koşullu Olasılık

O_1 ile O_2, S örnek uzayında iki olay olsun. $P(O_2) > 0$ olmak üzere; O_2 olayının gerçekleşmiş olması halinde O_1 olayının olasılığına, O_1 olayının O_2 olayına bağlı koşullu olasılığı veya kısaca O_1 in O_2 koşullu olasılığı denir ve $P(O_1 / O_2)$ şeklinde gösterilir. Burada O_2 olayı ilgili koşullu olasılığın örnek uzayı olmaktadır.

$P(O_1/O_2) = \frac{P(O_1 \cap O_2)}{P(O_2)}$ ile hesaplanır.

- [3] Veri Madenciliği ,Gökhan Silahtaroğlu 06'2008

14.1 Örnek

Bilgisayar Mühendisliği Bölümünde Yüksek Lisans yapan öğrencilerin lisans ve medeni hal durumlarını gösteren veriler aşağıdaki tablodadır.

	Evli	Bekar	Toplam
Bilgisayar Müh.	3	5	8
Elektrik-Elektronik Müh.	12	4	16
Toplam	15	9	24

O1: Evli öğrenciler
O2:Bekar öğrenciler
O3:Bilgisayar Lisanslı öğrenciler
O4: Elektrik-Elektronik lisanslı öğrenciler

$$P(O_1)=\frac{15}{24}\quad P(O_2)=\frac{9}{24}\quad P(O_3)=\frac{8}{24}\quad P(O_4)=\frac{16}{24}$$
$$P(O_1\cap O_3)=\frac{3}{24}\quad P(O_1\cap O_4)=\frac{12}{24}$$
$$P(O_2\cap O_3)=\frac{5}{24}\quad P(O_2\cap O_4)=\frac{4}{24}$$

Olasılıkları mevcuttur. Bilgisayar lisanslı olduğu bilinen bir öğrencinin bekar çıkma olasılığını bulalım.

$P(O_2/O_3)=\frac{P(O_2\cap O_3)}{P(O_3)}=\frac{\frac{5}{24}}{\frac{8}{24}}=5/8$ elde edilir.

14.2.2 Bayes Teoremi

$O_1,O_2,\ldots O_n$ aynı örnek uzaydaki karşılıklı ayrık ve bütüne tamamlayan olaylar olmak üzere, F aynı örnek uzaydaki bir başka olay olsun. Bu durumda

$P(O_k/F)=\frac{P(O_k)P(F/O_k)}{\sum_{i=1}^{n}P(O_i)P(F/O_i)}$ dir. Burada $P(F/O_i)$ önceki olasılıklar, $P(O_k/F)$

değerlerine de sonraki olasılıklar adı verilir. $\sum P(O_i)=1$ dir.

14.2 Örnek

Bir öğretmenin A dersindeki 11 öğrenci başarılı 4 öğrenci başarısız, B dersinden 8 öğrenci başarılı 7 öğrenci başarısız ve C dersinden de 5 öğrenci başarılı 10 öğrenci de başarısızdır. Bütün öğrencilerin aynı ortamda olduğu bilindiğine göre, bu öğretmenin başarılı bir öğrencisinin B dersini alma olasılığını bulalım.

$$P(B/\text{Başarılı}) = \frac{P(B).P(\text{Başarılı}/B)}{P(A).P(\text{Başarılı}/A) + P(B).P(\text{Başarılı}/B) + P(C).P(\text{Başarılı}/C)}$$

$$P(B/\text{Başarılı}) = \frac{p.8/15}{p.\frac{11}{15} + p.\frac{8}{15} + p.\frac{5}{15}} = 1/3$$ elde edilir.

14.2.3 Bayes Sınıflama

X kümesi sınıf üyeliği bilinmeyen bir veri kümesi olsun. H ise bu X veri örneğinin C sınıfına ait olduğu iddaa edilen hipotez olsun. Bu durumda H'ın C sınıfına ait olduğu kabul edildiği için P(H|X) olasılığı hesaplanabilir. (P(H|X)=sonrasal olasılık)

$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$ bayes bağıntısı olarak yazılabilir.

X={x1,x2,…,xn} sınıf üyeliği bilinmeyen bir veri kümesi olsun. Bu küme için m adet sınıf olduğunu varsayalım.c1,c2,…,cm de sınıf değerleri olsun. Sınıfı belirleyecek bir örneğe ilişkin

$P(Ci/X) = \frac{P(X/Ci)P(Ci)}{P(X)}$ olasılıkları hesaplanabilir

İşlem yükünü azalmak amacı ile

$P(X/Ci) = \prod_{k=1}^{n} P(Xk/Ci)$ yazılabilir. (örneğe ait xi değerlerinin biribirinden bağımsız olduğu kabul edilirse). Bilinmeyen örnek X'i sınıflandırmak için

$P(Ci/X) = \frac{P(X/Ci)P(Ci)}{P(X)}$ de sadece pay değerlerinin karşılaştırılması yeterli olacaktır. Bu değerler içinden en büyük olanı seçilerek bilinmeyen örneğin bu sınıfa dahil olduğu belirlenebilir. Sonuç olarak bayes sınıflandırıcısı

$C_{MAP} = \arg\max_{ci} \prod_{k=1}^{n} P(Xk/Ci)$ ile hesaplanabilir.

14.3 Örnek

Aşağıdaki verileri kullanarak x={E,172,65} olarak verien yeni bir verinin sınıfının ne olduğunu Bayes kuralına göre bulmaya çalışalım[4]:

Cinsiyet	Kilo	Boy	Beden
K	48	170	ORTA
K	49	151	KÜÇÜK
K	52	158	ORTA
K	56	165	ORTA
E	59	160	KÜÇÜK
K	61	159	ORTA
E	62	162	KÜÇÜK
E	63	174	ORTA
K	68	168	ORTA
K	69	177	BÜYÜK
E	72	170	ORTA
E	74	165	KÜÇÜK
E	85	175	ORTA
E	85	190	BÜYÜK
E	98	190	BÜYÜK

1.grup	150-160
2.grup	161-170
3.grup	171-180
4.grup	181-190
5.grup	191 ve üstü

1.grup	45-55
2.grup	56-65
3.grup	66-75
4.grup	76-85
5.grup	86-95
6.grup	96 ve üstü

Bu grupları tek tabloda tekrar düzenlersek;

- [4] Veri Madenciliği ,Gökhan Silahtaroğlu 06'2008, syf:61-62
-

Cinsiyet	Kilo	Boy	Beden
K	1	2	ORTA
K	1	1	KÜÇÜK
K	1	1	ORTA
K	2	2	ORTA
E	2	1	KÜÇÜK
K	2	1	ORTA
E	2	2	KÜÇÜK
E	2	3	ORTA
K	3	2	ORTA
K	3	3	BÜYÜK
E	3	2	ORTA
E	3	2	KÜÇÜK
E	4	3	ORTA
E	4	4	BÜYÜK
E	5	4	BÜYÜK

Elde edilir.

x={E,172,65} yani x={E,3,2} nin sınıfı?

P(CJ) değerlerinin hesaplayalım:

P(küçük)=P(C1)= 4/15=0,267
P(orta)=P(C2)= 8/15=0,534
P(büyük)=P(C3)= 3/15=0,2

P(x/Cj) değerlerini bulalım:

P(x/küçük)=1/4=0,25
P(x/orta)=1/8=0,125
P(x/büyük)=0/3=0,3

$$P(xi) = \sum_{i=1}^{m} P(xi / Cj) P(Cj)$$

$$P(xi) = (0{,}267x0{,}25) + (0{,}534x0{,}125) + (0{,}2x0) = 0{,}1335$$

$$P(kucuk / x) = \frac{0,267x0,25}{0,1335} = 0,5$$

$$P(orta / x) = \frac{0,534x0,125}{0,1335} = 0,5$$

$$P(büyük / x) = \frac{0,2x0}{0,1335} = 0$$

Bulunan değerler küçük, orta ve büyük sınıfına dahil olma olasılıklarını göstermektedir.

KAYNAKLAR

- Wavecluster: A multi-resolution clustering approach for very large spatial databases, Sheikholeslami, Gholamhosein and Chatterjee, Surojit and Zhang, Aidong (1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. In Proceedings of the 24th VLDB conference .
- Veri Madenciliği Yöntemleri, Yalçın Özkan 06'2008
- Veri Madenciliği ,Gökhan Silahtaroğlu 06'2008
- İstanbul Ticaret Üniversitesi Derğisi Veri Madenciliği Modelleri Ve Uygulama Alanları (Serhat ÖZEKES)
- www.bilmuh.gyte.edu.tr/~htakci/vm/verimadenciligi.doc
- http://www.gurunlu.com/?tag=/k-means
- Ayhan Adsız, Metin Madenciliği, Ahmet Yesevi Üniversitesi, Bilişim Sistemleri ve Mühendislik Fakültesi,Dönem Projesi,sayfa:46,2006.
- http://tr.wikipedia.org/wiki/Genetik_algoritma
- http://www.yapay-zeka.org/modules/wiwimod/index.php?page=GA