Hoche, Susanne, Ingolf Geist, Lourdes Peña Castillo, and Nadine Schulz. 2006. Section 3.5 Databases, Knowledge Discovery, Information Retrieval, and Web Mining, pp. 168-184 of Chapter 3 Methods, Algorithms, and Software, in CIGR Handbook of Agricultural Engineering Volume VI Information Technology. Edited by CIGR-The International Commission of Agricultural Engineering; Volume Editor, Axel Munack. St. Joseph, Michigan, USA: ASABE. Copyright American Society of Agricultural Engineers.

Çevirmenler: Sefa TARHAN, Mehmet Metin ÖZGÜVEN ve Abdullah BEYAZ
Çeviri Editörleri: Sefa TARHAN ve Mehmet Metin ÖZGÜVEN

3.5 Veri Tabanları, Bilgi Keşfi, Bilgi Çıkarma ve Web Madenciliği

Yazarlar: S. Hoche, I. Geist, L. Peña Castillo ve N. Schulz

Çevirmenler: Sefa TARHAN, Mehmet Metin ÖZGÜVEN ve Abdullah BEYAZ

Özet: *Günümüzde dijital bilgi bolluğu bulunmakta ve tarım sektörü gibi alanlarda veri toplamaya büyük yatırımlar yapılmaktadır. Değerli bilgilerin ortaya çıkarılması, kullanışlı, sık veya olağanüstü desenlerin belirlenmesi ya da karmaşık karar işlemlerinin desteklenmesi gibi verinin başarılı kullanımları; veri depolama, verilere ulaşılması ve verilerin analizinde güçlü araçları gerektirir. Veri tabanı yönetim sistemleri (VTYS); verileri depolama ve ulaşmada etkili, entegre ve standart bir platform sunmaktadır. Veri tabanlarında bilgi keşfi (VTBK)'nin amacı; büyük veri kolleksiyonlarının depolandığı veri tabanlarında, kullanışlı bilgilerin yarı otomatik olarak keşfidir. Bilgi çıkarma (BÇ) işlemi; doğal dil metinleri, resimler, ses ve video gibi yapısal olmayan ve yarı bulanık verilerde kullanıcı tanımlı sorgular yapma olarak tanımlanabilir. Web madenciliği kavramı, dünya çapındaki web kaynaklarından kullanışlı bilgilerin çıkarılmasını tanımlamaktadır. Bu bölümde veri tabanı yönetim sistemlerinin mevcut gelişmişlik durumu ve veri tabanlarından bilgi keşfi, bilgi çıkarma, web madenciliğinin gelişen alanları ve tarımsal uygulamalardaki temel metodolojiler hakkında genel bir değerlendirme sunulmaktadır.*

Anahtar Kelimler: *Veri tabanları, Veri madenciliği, Bilgi çıkarma, Web madenciliği.*

3.5.1 Veri Tabanlarına Giriş

Verilerin depolanması amacıyla uygulamalar kendilerine özgü tipik metodlar ve yapılar kullanmakta, uygulama geliştiriciler veri ara yüzleri ve fiziksel depolama gibi konularda detaylara ihtiyaç duymaktadırlar. Bir *veri tabanı (VT)*; veri tabanı yönetim sistemi olarak adlandırılan, uygulamalara fiziksel girişe imkan sağlayan bir bilgi sistemidir. Bu sistemdeki verilerin kullanımı uygulamaya bağlı olarak değişmektedir. Örneğin her uygulama kendi alanına sahiptir ve bu yüzden verilerin belli bölümlerinden belli şekillerde faydalanmaktadır. Veri depolamada, uygulamalarda genel amaçlarla kullanılan dosyalara yapılan giriş sayısına bakıldığında bir veri fazlalığı probleminin olduğu görülmektedir.

Bu veri fazlalığını içeren örnek bir senaryo burada verilmiştir. Bir metin işleme uygulaması; metinler, ürünler ve adresleri yönetir. Bir başka sistem olan kitap yönetimi yazılımında ise hesap, ürün ve adres bilgileri depolanmaktadır. Üçüncü bir sistem olan envanter yönetim uygulamasında; ürün işleme ve adres bilgilerini ele almaktadır. Bu senaryoda adres bilgisi her üç sistemde de aynı anda gereksiz olarak depolanmaktadır.

Bir veritabanı yönetim sistemi bütün verileri tek bir yapısal veritabanında birleştirmekte ve bu verilerin gereksiz fazlalığından kaçınmaya yardımcı olmaktadır. Bu amaçla tanımlayıcı sorgulama dili (TSD) kullanılarak aynı veriye giriş bütün uygulamalar ve kullanıcılar için sağlanabilmektedir. Böyle bir dil, veriye hangi veri yolundan gidileceğini düşünmeden ulaşılmasını sağlamaktadır. Bu durum büyük veri gruplarında veri sorgularının optimizasyonuna olanak sağlamaktadır.

Verilerin tek bir veri tabanına entegre edilmesi, çoklu kullanıcıların ve uygulamaların veri tabanına aynı anda ulaşabilmelerini sağlamaktadır. Veritabanı yönetim sistemi; verileri okuma, yazma ve bunların senkronizasyonu gibi birçok faaliyeti içeren *iş kalemleri* kavramını ortaya koyarak, çoklu kullanıcıların eş girişlerine imkan sağlamıştır.

Her uygulama veri üzerinde kendine has gereklilikleri olduğu için *veri bağımsızlığı* kavramına ihtiyaç vardır. Veri bağımsızlığı, gerçek fiziksel depo şemasından çıkarımlar yapmaya ve standardize edilmiş sorgu arayüzlerinin tanıtılmasına imkan vermektedir. Veri bağımsızlığını gerçekleştiren üç aşamalı mimari aşağıda verilen katmanları kapsamaktadır [1]:

- Verilerin fiziksel depolanmasını tanımlayan iç şema,
- Orta katmanı oluşturan kavramsal şema (bu katman bütün veri tabanı üzerinde mantıksal ve uygulamadan bağımsız bir bakışı tanımlamaktadır),
- Son katmanı oluşturan dış şemalar (bu şemalar farklı uygulamalar için kavramsal şema üzerinde özel bakışlar tanımlamaktadır).

Bunlara ilave olarak veri güvenliği ve koruması için veritabanı yönetim sistemleri destek sağlamalıdır. Öncelikle güçlü geri kazanım ve yedekleme mekanizmaları, hata olması durumunda veri kararlılığını garanti etmelidir. İkinci olarak, detaylı kullanıcı rolleri ve hakları gibi giriş kontrol teknikleri veriye yetkisiz girişleri engellemektedir.

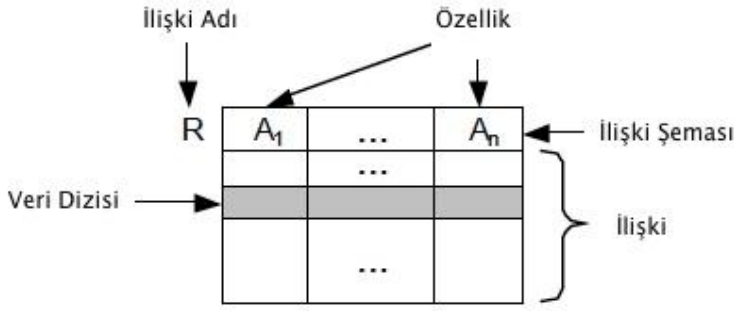
Bir veri tabanını, VTYS tarafından yönetilen yapılandırılmış bir veri kümesi olacak şekilde bütün bu mekanizmalar birleştirilmektedir. Bir veri tabanı yönetim sistemi, veri tabanını ele alan farklı yazılım modüllerini içermektedir. *Veri tabanı sistemi (VTS)*, veritabanı yönetim sistemi ile özel bir veri tabanının birleşimini belirtmektedir.

İlişkisel Veritabanları

Bir kavramsal şema, verinin mantıksal bir görünümünü tanımlar. En popüler kavramsal veri modeli ilişkisel modeldir [2,3]. Bu model; Oracle 10g, Microsoft SQL Server, IBM DB2 gibi birçok ticari veri tabanı sistemi tarafından

kullanılmaktadır. İlişkisel veri tabanı bir tablo grubu içermektedir. Bir tablo yanlız bir ilişki belirtir. Şekil 1, böyle bir ilişki kavramını göstermektedir. Tablo başlığı, tablonun yapısını tanımlar ve şema ilişkisi olarak adlandırılır. Tablonun bir satırı, *tuple* olarak adlandırılır. Bir ilişki, bütün tuplelerin bir kümesi (yani tablonun gövdesi) biraraya gelmesiyle oluşmaktadır. Bir sütun yanlız bir özelliği belirtir. Özelliğin adı tablo başlığında verilir ve değeri sütun girdisi olarak depolanır.

İlişkisel model, basit bir veri modelidir. Ancak ilave *bütünlük sınırlamalarından* faydalanmaktadır. Bütünlük sınırlandırmaları *yerel* ve *küresel* sınırlandırmalar olarak sınıflandırılır. Yerel sınırlamalar, tek bir tablo içerisindeki durumları tanımlamaktadır. Örneğin, bir öznitelik üzerindeki *özgün anahtar durumu* öznitelik değerlerinin her bir tupleyi özgün olarak tanımladığını belirtir. Küresel sınırlamalar, çoklu tabloları etkilemektedir. Buna örnek olarak yabancı anahtar sınırlaması verilebilir. Yabancı anahtar sınırlaması, bir tabloda mevcut olan bir öznitelik değerinin referans edilen ikinci bir tabloda da bulunmasını gerektirir. Hepsi birlikte bir ilişki şeması, yapısal tanımlamaları ve bütünlük sınırlamalarını içermektedir.



Şekil 1.İlişkisel veri modeli.

Daha öncede tanımlandığı üzere, veri tabanı yönetim sistemlerinin avantajlarından biri tanımlayıcı bir sorgu dili kullanmasıdır. İlişkisel modele bağlı olarak farklı sorgulama dilleri önerilmiştir. Bu dillerin en az *ilişkisel cebir* kadar kuvvetli olmaları gereklidir. İlişkisel cebir; küme teorisi işlemlerinin birliği, kesişmesi ve farklılığı yanında projeksiyon, seçim, yeniden isimlendirme ve birleştirme gibi işlemleri kapsamaktadır [4]. *Seçim*; seçim cümlesinde tanımlanan bir durumu tatmin edecek bir ilişkiden tupleleleri çıkartır. Projeksiyon, sütunların seçimine izin verir ve *yeniden isimlendirilme* ise sorgu sonucundaki bir sütuna yeni bir isim verir. *Ekleme*, genel öznitellik ve değerler üzerinden tabloların birleşimini sağlar. En önemli ilişkisel sorgu dili bir sonraki bölümde açıklanacak olan SQL'dir.

SQL: Yapısal Sorgu Dili

Yapısal sorgu dili SQL (Structured Query Language) bir ilişkisel dil olup birçok ticari ilişkisel VTYS tarafından desteklenmektedir. Güncel standard SQL

2003 [5]'dür. SQL 2003; nesne ilişkili veri işleme için özellikler, veri madenciliği için arayüzler, BÇ'yle birlikte, SQL veri tabanlarında XML veri işleme standardının yanında harici veri kaynakları sağlamaktadır. Bu bölümde SQL'in temel sorgulama kavramları hakkında kısa bir değerlendirme verilmektedir. İleri özellikleri yanında SQL'in veri ve bakış tanımlama özellikleri bu bölümde tartışılmamıştır. Konu ile ilgilenen okurların [5] nolu gibi kaynakçada verilen literatürleri incelemeleri önerilmektedir.

Burada, SQL'in temellerini göstermek için bir örnek verilmiştir. Örneklere ait ilişkiler Şekil 2'de sunulmuştur. Milkprod ilişkisi, son yıllardaki her bir inek başına süt üretimiyle ilgili bütün bilgileri tutmaktadır. Hayvanlar hakkındaki bilgileri ikinci özellik olan Cows'da tutulmaktadır. Her iki ilişki birbirleri ile bilindik bir öznitellik numarası ile bağlanmıştır.

Temel bir SQL sorgusu üç kelime içerir: select (seç), from (nereden) ve where (nerede). from kelimesi sorgudaki ilişkileri belirtir. Sonuç kümesinde istenilen öznitelikler sütun yeni isimlendirmeleri ve ek fonksiyonlar select kelimesiyle belirlenir. Bir seçim durumu sorgunun where kelimesiyle tanımlanır. "2004 yılında en az 4000 lt süt veren ve Cow_no ile tanımlanan bütün ineklere ait verileri getir" sorgusu dikkate alınsın. Bu sorgu SQL ifadesi olarak aşağıdaki şekilde tanımlanır:

```
Select No as Cow_no, Milk
FROM Milkprod
WHERE Milk >= 4000 AND Year = 2004
```

Milkprod			Cows		
No	**Süt**	**Yıl**	**No**	**Süt**	**Yıl**
110	3500	2003	110	1000	05.06.1997
111	4000	2003	111	1200	31.05.1999
110	4500	2004	113	800	11.05.2000

Şekil 2. "Milkprod" ve "Cows" örnek tabloları

İkinci örnek ise ekle işlemi ile iki ilişkinin birleşimini göstermektedir. Bu örnekte her bir inek için fiyat, yaş ve süt üretimi arasındaki korelasyonun belirlenmesi istenmektedir. Bu yüzden "No" öznitelliği esas alınarak her iki tablo birleştirilmek zorundadır. SQL'de bir eklemeyi ifade etme yöntemi FROM kelimesi içerisinde natural join anahtar kelimesini kullanmaktır. Bu yüzden Milkprod ve Cows ilişkileri "No" öznitelliği vasıtasıyla birleştirilmiştir. Sonuç olarak istenilen öznitellikler select (seçim) cümlesinde belirtilir:

```
SELECT No, Milk, Price, Birth_date
FROM Milkprod natural join Cows
```

İlişkisel VTYS'lerin yanında diğer kavramsal modelleri kullanan sistemler mevcuttur. Örneğin uzaysal VTS'ler, örneğin sınırlı veri tabanları gibi özel veri modelleriyle haritalar için tipik olarak uzaysal sorgulamaları desteklemektedir [6]. Nesnel ilişkili veri modelleri gibi ilişkisel modellere yapılan eklentiler; veri madenciliği, BÇ sorgulamaları ve XML işleme gibi genişletilmiş veri kavramlarının yanında karmaşık veri tiplerinin veri yönetim sistemlerine dahil edilmesine imkan verir. Veri madenciliği ve multimedya arayüzleri gibi genişletilmiş SQL işlemlerinin yanında SQL sorgulama özelliklerinin ileri kavramları literatür [1,4,8]'de bulunabilir.

3.5.2 Veri Tabanlarında Bilgi Keşfi

Veri tabanlarında bilgi keşfinin (VTBK) amacı, veri tabanlarında genel olarak saklanan büyük veri koleksiyonlarındaki faydalı bilgilerin keşfidir. Verilerin genel tanımlanmasını sağlayan belirli kümeler veya kuralların olağan olmayan veya sıklıkla olan oluş desenlerini veya alt gruplarını belirlemeye veya aynı alanın henüz bilinmeyen nesneler hakkında ileriye dönük tahminleri yapmaya yöneliktir.

Günümüzde dijital bilginin elde edilmesi kolay ve depolanması pahalı değildir, ayrıca hızlı bir şekilde artan veri toplanabilir. Toplanan veriler ve bir nesneyi tanımlayan özeliklerin sayıları artmaktadır. Büyük miktarlardaki verinin analizi için insan yeteneği, veri toplama ve depolamada kullanılan araçların gerisinde kalmaktadır. Verilerdeki kodlanan değerli bilgiler konvansiyonel metotlarla yapılan analizlerle keşfedilemeyebilir.

Veriden Türetilmiş Bilgiye

Geleneksel yaklaşımlarda veri analizi bazı uzmanlar tarafından oluşturulan hipotezler üzerinde merkezlendirilebilir. Bir hipotez, verinin beklenen bazı özelliklerini tanımlar ve örneğin veri tabanı sistemine yapılan SQL sorgulamalarına dönüştürülür. Cevap olarak VTS, hipotezlenen özellikleri tatmin edecek bütün kayıtları üretir. Veri analiziyle ilgili böyle bir yaklaşım sadece yavaş ve etkisiz olmayıp aynı zamanda oldukça subjektiftir. Eldeki veriler hakkındaki ön yargılar analizi taraflı hale getirir. Daha kötüsü, hipotezlerin herhangi biri tarafından kapsanılmayan perspektifler araştırılmaz ve doğru sorgulama yapılmadığı için önemli desenler belirlenmeden kalabilir.

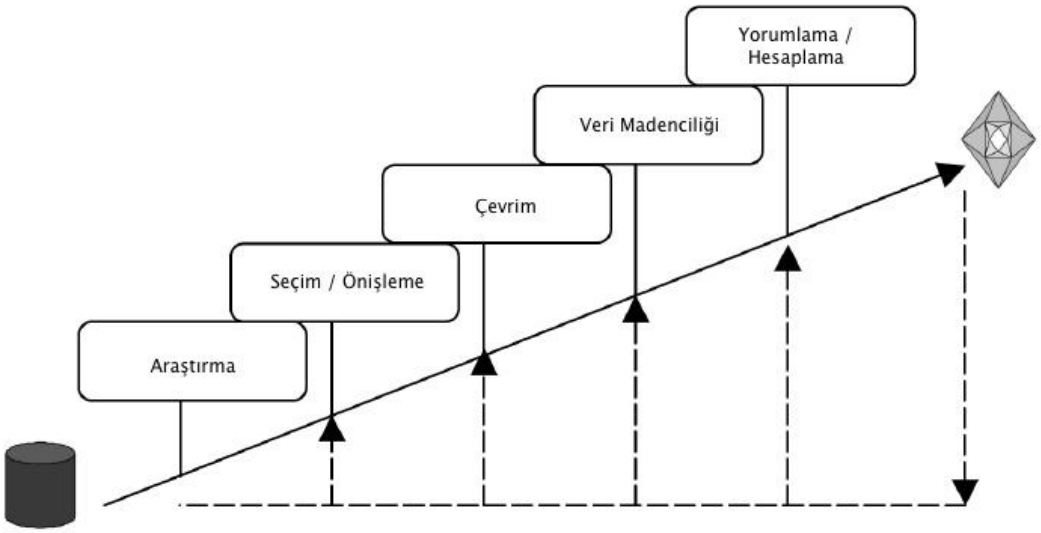
VTBK verilerden türetilmiş bilginin çıkartılması için kullanılan otomatikleştirilmiş bir yaklaşımdır. VTBK makine öğrenmesi, veri tabanları, istatistik ve görselleştirme metotlarını birleştirir [9,10]. VTBK'nın amacı eldeki verilerin sadece kapsamlı anlaşılmasına ait olmayıp daha çok veri içerisinde gizlenmiş değerli bilgi külçelerinin tespitidir [10]. VTBK "verilerdeki geçerli, alışılmamış potansiyel olarak kullanışlı ve kesin olarak anlaşılabilir desenlerin

tanımlanmasının basit olmayan işlemi" olarak tanımlanır [11]. İşlem terimi, VTBK'nın tekrarlayan doğası hakkında ipucu vermektedir (bir sonraki paragrafa bakılabilir). *Basit olmayan (nontrivial)* işlemi; sonuçların apaçık bazı hesaplama adımlarıyla elde edilemediği ancak daha gelişmiş çıkarım yaklaşımlarının uygulanmasıyla elde edilebildiği anlamına gelir. Tanımlanan desenlerin; *doğru, alışılmamış, potansiyel olarak kullanışlı ve anlaşılabilir* olduğu iddiası keşfedilen türetilmiş bilginin eldeki bilgiye yüksek derecelerde uygulanması, yeni iç görülere yol açması ve eldeki görev için faydalı olması ve bütün bunlara ilaveten kapsamlı bir tarzda tanımlanması gerektiği anlamını verir.

VTBK İşlemi

VTBK işlemi Şekil 3'de görüldüğü gibi doğası gereği interaktif ve tekrarlanan bir işlemdir. Kullanıcı ve sistem arasındaki işbirliğinde gerçekleştirilen birçok adımı içermektedir ve bu adımlar tatmin edici sonuçlar alıncaya kadar birçok kez tekrarlanmak zorunda kalabilir [9,11]. İşlem modelleri kişiden kişiye kısmen farketmekle birlikte aynı bileşenleri içerir [kaynak [11], [12]'ye bakılabilir].

- İlk adım *araştırma* adımı olup, uygulama alanının tam olarak anlaşılması geliştirilmek zorundadır ve eldeki VTBK görevinin amacı belirlenmek zorundadır. Örneğin muhtemel hedefler olarak tarımsal uygulamalar ve verim arasındaki korelasyonun belirlenmesi, pazar eğilimleri, tüketici davranışlarındaki değişimler veya potansiyel gelecekteki pazarların belirlenmesi verilebilir.
- *Seçim*, yapılacak analiz için hedef veri kümelerinin oluşturulmasını içermektedir. Bu adım farklı kaynaklardan gelen verilerin birleştirilmesini veya verilen bir verinin belirli bir bölümünü içerebilir. Bir ön işlem adımında gürültü, aykırı değerler ve kayıp verinin ele alınmasına yönelik stratejilerin veriye uygulanması zorunludur.
- Sonrasında, bir çevrim adımında sonuç veri setindeki kayıtlar ve özelliklerin sayısı gerekirse azaltılabilir.
- İlk adımda tanımlanan VTBK işleminin amacı belli bir veri madenciliği metoduna eşleştirilebilir. Verinin yapısı halihazırda özel kullanışlı bir yaklaşımı önerebilir. Model parametreleri belirlenmek zorundadır ve gerçek analiz veya veri madenciliği adımı başlayabilir.



Şekil 3. VTBK işlemi: Veriden bilgiye

- Son olarak belirlenen desenler değerlendirilmeli ve yorumlanmalıdır. Önceki adımların bazıları ya da tümü; başlangıçta tanımlanan amaç karşılanıncaya kadar tekrarlanmak zorunda kalabilir.

Veri Madenciliği

Veri madenciliği bazen VTBK ile aynı anlamda kulanılmaktadır. Bununla birlikte biz genel bir yaklaşım takip edecek ve önceki paragraflarda açıklanan VTBK işleminin bir adımı olarak tanımlayacağız. Söz konusu adımda bazı veri madenciliği algoritmaları esas alınarak veriler analiz edilmektedir.

Geçmiş yıllarda Weka [13], Enterprise Mining [14], Clementine [15] gibi veri madenciliği iş tezgahı sistemleri geniş bir bilimsel alanda (jeofizik [16,17], tıp [17,19]) ve ticari alanlarda (örneğin dolandırıcılık tespiti [17] yatırım [11], risk yönetimi [19] telekominikasyonda [17]) başarıyla uygulanmıştır. Veri madenciliği teknikleri, sistemleri ve uygulamaları hakkında daha geniş bir bakış için ilgili okuyuculara ekte verilen kaynaklar [13,20,21] önerilmektedir.

VTBK işleminin veri tabanı madenciliği adımını başarmak için tasarlanmış sistemler ve temel teknikler türetilmiş bilginin keşfi amacına bağlı olarak farklılık gösterir. Keşif amaçları *tahminleme, sınıflandırma ve tanımlama* olarak kategorize edilebilir [11].

Tahminleme veya sınıflandırma teknikleri; aynı alanın henüz bilinmeyen nesnelerin davranışlarını tahmin etmek için kullanıbilen ampirik verilerden sınıflandırma şemalarının yapılmasını hedeflemektedir. Eldeki verilerin özniteliklerine bağlı olarak önceden tanımlanmış birkaç sınıfın bir tanesine bir veri kalemini haritalayan fonksiyon formatındaki hipotezleri otomatik olarak üretir [22]. Amaç; gelecekteki veri noktalarının mümkün olduğu kadar doğru bir şekilde ve aynı zamanda eldeki verileri kategorize etmektir. Tahminleme teknikleri; karar ağaçları,

destek vektör makineleri, yapay ağlar, kural öğreniciler ve olasılıksal ağlar gibi metotları içermektedir.

Bunların aksine tanımlayıcı sistemler, bölgesel ilgilinin veri bölgelerini tanımlar ve insanların anlayabileceği şekilde keşfedilen desenleri sunar. Popüler tanımlayıcı teknikler; alt grup keşfi, kümeleme, değişen ve sapma belirleme, bağımlılık modellemesi ve özetleme gibi teknikleri içerir.

Tarımda Veri Madenciliği Uygulamaları

Tarımda VTBK son zamanlarda gerekli bir teknoloji olarak ortaya çıkmıştır [23]. Little ve ark. [24], Amerika Birleşik Devletleri Tarım Bakanlığı (USDA, United States Department of Agriculture) ürün sigortası yönetiminde VTBK'nın potansiyel kullanımını gösteren bir proje tanımlamaktadır. Onların yaklaşımı; USDA veri tabanında bir milyondan daha fazla kayıtlardan şüpheli ekim veri tabanı tarımsal uygulamalar, üretim tipi (sulanmışa karşı sulanmamış, daneye karşın silaj) ekilen alan miktarı, hasat edilen alan miktarı, bölgesel karakteristikler ve meteoroloji hakkındaki bilgileri içermektedir. VTBK metotları toplam ekilen ürüne kıyasla elde edilen olağandışı küçük verimlerin dağılımlarını belirlemek için uygulanmıştır. Bu olağandışı verim dağılımları bölgesel meteorolojik olaylarla açıklanamamakta ve USDA ürün sigorta programının yanlış kullanımını işaret etmektedir.

Ayrıca Illinois gıda ve tarım araştırmaları konseyinin (C-FAR) alana özgü tarım projesi için veri madenciliğinin kullanımında [25] VTBK; ürün verimini arttırmak amacıyla hassas tarım ve değişken oranlı uygulamalar alanında kullanılmıştır. Projenin amacı; hava durumu, gübreleme, tohum çeşidi, toprak özellikleri, yetiştiricilik ve yönetim geçmişinin mekansal ve zamansal karakteristikler arasındaki etkileşimlerini esas alarak ürün veriminin mekânsal karakteristiğini tahmin etmektir.

Canteri ve ark. [26]; boyutları ve karmaşıklığı sebebiyle geleneksel metodlarla analiz edilemeyen hassas tarım veri tabanları üzerinde veri madenciliği metodlarının kullanışlılığını keşfetmişlerdir. Bu araştırmacılar ürün verimi ile fiziksel kimyasal toprak özelliklerini ilişkilendirecek başarılı teknikleri araştırmışlardır.

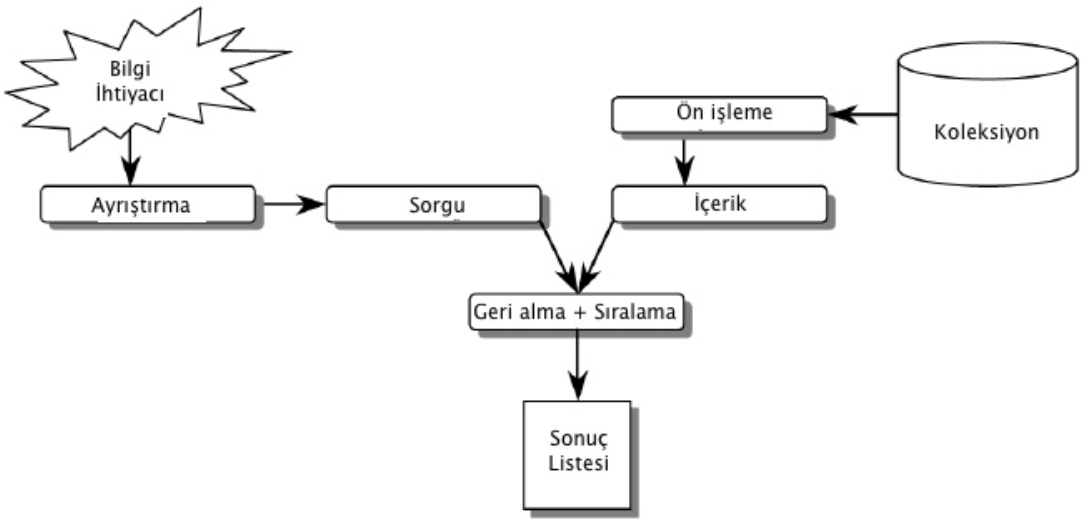
3.5.3 Bilgi Çıkarma

Bilgi çıkarma (BÇ); doğal dil metinleri, görüntüler, ses ve video gibi yapılanmamış ve anlamsal olarak bulanık verilerin koleksiyonlarında bilgi araştırmayla uğraşır. Veri çıkarma sistemleri; örneğin ilişkisel veri tabanında depolanan yapılandırılmış veri ile uğraşırken ve sorguda tanımlanan anahtar kelimeleri içeren bütün kalemleri çıkartırken, BÇ sistemleri verilen bir sorguyu tatmin edecek bir veriden daha çok konu hakkındaki bilgiyi çıkarmayı hedeflemektedir. Aslında, BÇ'deki sorgular katı değildir ve BÇ sistemi kullanıcının bilgi ihtiyacını tanımlayan sorguya uygun bütün bilgileri çıkartmayı hedeflemektedir. Uygunluk (alakalı olma) BÇ'nin ana unsurudur.

İleriki bölümlerde çıkarma performansının değerlendirme ölçülerinin yanında çıkarma modellerinin ana kavramları tanıtılmakta ve klasik yazım bilgisi çıkarımı üzerinde yoğunlaşılmaktadır. BÇ hakkında daha detaylı bilgi Rijsbergen [27], Salton ve McGill [28], Beaza-Yates ve Riderio-Neto [29] tarafından yazılan kitaplarda bulunabilir.

Bilgi Çıkarma Süreci

Şekil 4'de gösterilen BÇ süreci, verilen bir veri koleksiyonunda depolanan belgelerin *ön işlemesini*, onların *çıkarımını* ve uygun dökümanların *derecelenmesini* içermektedir.



Şekil 4. Bilgi çıkarma süreci

Ön işleme fazında belgelerin içeriği anahtar kelimelerle özetlenmektedir. Daha sonra yazının bir *indeksi* yapılır yani bu indeks geniş belge koleksiyonunda hızlı aramayı destekleyen veri yapısıdır. Alışılmış bir indeks yapısı *çevrilmiş dosyadır*. Çevrilmiş dosya koleksiyondaki ayrı bütün anahtar kelimeleri içeren bir vektör ve her bir anahtar kelime için ilgili anahtar kelimenin geçtiği belgelerin listesini içermektedir.

İndeks yapıldıktan sonra çıkarma süreci kullanıcı tarafından tanımlanan sorguyu işleyerek başlatılabilir. Sorgu kullanıcının bilgi ihtiyacının sistem temsilidir. Başlangıçta oluşturulan indeks yapısı hızlı sorgulama işlemeye imkan sağlar.

Verilen sorguya uygunluklarına bağlı olarak çıkartılan belgeler derecelendirilir ve sıralı sonuç listesi kullanıcıya sunulur. Bir sonraki bölümde BÇ sürecinin her bir adımını daha detaylı tanımlayacağız.

Ön İşleme ve İçerikleme

Klasik metin bilgisi çıkarmada, her döküman *indeks terimleri* olarak ifade edilen anahtar kelimeler kümesiyle tanımlanır. İndeks oluşturma, aynı içeriğe sahip

belgelerin aynı anahtar kelimelerle tanımlanmasını gerektirir. Bu anahtar kelimeler otomatik olarak çıkarılabilir veya bir uzman tarafından belirlenebilir.

Bir belgenin içeriğini tanımlamadaki indeks teriminin uygunluğu araştırılan bütün belgelerdeki bu terimin bulunma sayısına bağlıdır. Bir indeks teriminin birçok belgede olması çok da önemli değildir. Ancak incelenen belge koleksiyonunun sadece küçük bir kısmında bulunan bir terim uygun belge sayısını daraltır ve böylece daha anlamlı olur. Bu sebeple bir belgenin içeriğini tanımlamada önemini ortaya koyan ifade bir indeks terimi olarak atanır.

Belgelerin ön işlemi aşağıda belirtilen işlemleri içerir:

- *Sözcüksel analiz (Lexical analysis)*: Bu analizin amacı yazı içerisindeki kelimeleri tanımlanmak ve tireleri, dijitleri, noktalama işaretlerini ve büyük ve küçük harfleri ele almaktır.
- *Durdurma kelimelerinin çıkarılması*: bu aşamada belgelerin sınıflandırılmasında önemi olmayan kelimeleri filtrelemek için (örneğin *"the"* ve *"to"*) durdurma kelimeleri çıkartılır.
- *Kelimelerin köklerine indirilmesi*: bu aşamada kelimeler, kendilerinin kök formlarına indirgenir ve sorgulama terimlerinin söz dizimi farklarını içeren belgelerin çıkarılmasına müsade edilir.
- *İndeks terimlerinin seçimi*: bu aşamada indeks terimi olarak kullanılacak kelimeler belirlenir.
- *Terim kategorizasyon yapısının oluşturulması*: bu aşamada ilgili terimlerle sorgulamayı genişletmek için eş anlamlılar sözlüğü gibi yapılar oluşturulur.

Ön işleme basamağından sonra, indekslenmiş belgeler sorgulanabilir. Verilen sorgular, temel çıkarma modeli esas alınarak değerlendirilir. Takip eden bölümlerde birçok BÇ modelleri tartışılacaktır.

Bilgi Çıkarma Modelleri

Boolean, vektör ve olasılık modelleri olmak üzere 3 farklı klasik BÇ modeli bulunmaktadır. Daha detaylı bilgi için ilgili okuyuculara kaynak [29] önerilmektedir.

Boolean modeli klasik küme teorisi ve boolean matematiğine dayanan basit bir BÇ modelidir. Bir sorgu, açık tanımlı (crisp semantics) boolean ifadesiyle belirlenmektedir ve çıkarma stratejisi herhangi bir derecelendirme skalası olmadan boolean mantığına dayanmaktadır. Bu yüzden boolean modeli, BÇ'den daha çok veri çıkarmaya yöneliktir. Eğer indeks terimi dökümanda bulunuyor ise *doğru*, yoksa *yanlış* sonucunu vermektedir. Bir indeks teriminin değeri iki farklı sonuçtan (uygundur veya uygun değildir) birisidir ve yani her belge ya uygundur ya da değildir.

Vektör modeli [30] çok boyutlu vektörlerin benzerliğine dayanır. Vektör modeli, araştırılan belgeler ve sorgudaki indeks terimlerinin bulunmalarını yansıtır. Daha detaylı bakılırsa bir belgenin bir sorguya olan uygunluğu; belge ve sorgu

indeks terim vektörleri arasındaki açının kosinüsüne göre değerlendirilir. Çıkarılan belgeler, uygunluklarına göre derecelendirilir ve sadece kısmen sorguya uyan belgelerde çıkartılabilir.

Olasılık modeli [31], bir belgede bulunan bir terimin görünme olasılığını veya bir belgenin bilgi ihtiyacını tatmin etme ve verilen sorgu için uygunluk olasılığını tahmin eder. Çıkartılan belgeler olasılık değerlerine göre sıralanır.

Genel olarak kısmi eşleştirmeler yapamadığı için boolean modeli en zayıf klasik model olarak değerlendirilmektedir [28]. Vektörel modelin olasılık modelden üstün olup olmadığı tartışması devam etmektedir.

Geçmiş yıllarda, her bir klasik model için alternatifler oluşturulmuştur. Boolean modeli, *genişletilmiş Boolean modeli* ve *bulanık veri modeli* olarak iyileştirilmiştir [29]. *Genelleştirilmiş vektör modeli* [32] ve *gizli anlamsal indeksleme modeli* [33] klasik vektör modeline dayanmaktadır. Sonuç çıkarım ağı ve inanç ağı, klasik olasılık modelinin gelişmiş halleridir [29].

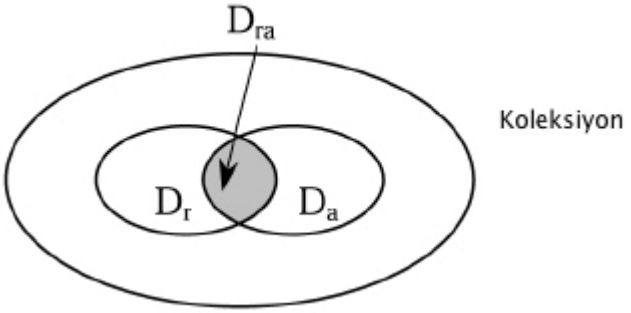
Derecelendirme

Çıkarma işleminden sonra uygun belgelerin sıralı listesi kullanıcıya sunulur. Vektör ve olasılık modellerinde belgelerin derecelendirme sırası, verilen sorguya olan uygunluğuna bağlıdır. Boolean modelinde, derecelendirme mümkün değildir. Bir belge ya sonuç kümesindedir veya değildir ve böylece bütün çıkartılan belgeler eşit uygunluğa sahiptir.

Böyle bir durumda kullanıcı belgeleri interaktif olarak uygun ya da uygun değil (alakalı ve alakasız) olarak işaretleyerek düzenleyebilir. Kullanıcının yargısına bağlı olarak sorgu, vektör modelinde indeks terimlerinin yeni bir vektörü olarak temsil edilir yani yeniden formüle edilir. BÇ modeli, yeniden düzenlenen sorguya uygun olan belgeleri tekrar çıkartır. Bu tekrarlı sorgu iyileştirme süreci *uygunluk geri beslemesi* olarak adlandırılır [34].

Çıkarım Performans Değerlendirmesi

BÇ sistemlerinin uygun olmayan (alakasız) belgeleri çıkarması ya da uygun (alakalı) belgeleri çıkarmaması olasıdır. Bu sebeple BÇ sistemlerinin performansını ölçme teknikleri vardır. BÇ performansının standart ölçümleri, *geri çağırma* ve *kesinliktir*. Bir sorguyu q ve verilen bir koleksiyonda q'ya uygun olan bir belge kümesini D_r olarak isimlendirdiğimizde, D_r kümesi içerisindeki belge sayısı $|D_r|$ ile ifade edilmektedir. q sorgusuna gore sistem D_a cevap kümesini çıkarmaktadır. $|D_a|$ cevap kümesindeki belge sayısıdır. $|D_{ra}|$ q sorgusuna uygun olan ve D_a cevap kümesinde bulunan belgelerin sayısıdır. $|D_{ra}|$; D_a ve D_r kesim kümesinin eleman sayısıdır. Şekil 5' de görüldüğü gibi, $D_{ra} = D_a \cap D_r$ şeklinde ifade edilebilir.



Şekil 5. Bir doküman koleksiyonunda kesinlik ve geri çağırma

Geri çağırma **R**; bir koleksiyondan çıkartılan uygun belgelerin sayısını ifade eder ve çıkartılan uygun belgelerin sayısının, koleksiyondaki uygun belgelerin toplam sayısına oranı olarak tanımlanır. Yani $\mathbf{R}=|D_{ra}|/|D_r|$ şeklinde ifade edilir.

Kesinlik **P;** çıkartılan belgelerden uygun olanlarının sayısını tanımlar ve çıkartılan uygun belgelerin sayısının, çıkartılan toplam belge sayısına oranı olarak ifade edilir. Yani $\mathbf{P}=|D_{ra}|/|D_a|$ şeklinde ifade edilir.

İdeal olarak, kesinlik ve geri çağırmanın %100 olması gerekir. Ancak, her iki ölçü birbirine bağlıdır ve geri çağırmadaki artış genellikle kesinlikte azalmaya neden olur. Böylece, BÇ sistemleri aynı anda kesinlik ve geri çağırmayı maksimize etmeye çalışır.

Tarımda Bilgi Çıkarma Uygulamaları

Tarım sektöründe BÇ; tarımsal uygulamalar, pazar eğilimleri, tarım ve orman ürünlerinin fiyatları, uluslararası ticaret kanunları, resmi dökümanlar vb. hakkında bilgi ve türetilmiş bilginin yayılmasında önemli bir rol oynamaktadır.

Otuka [35], danışmanlar ve çiftçiler arasında türetilmiş bilgi ile tecrübenin paylaşılması ve tekrarlı kullanımı için BÇ tekniklerini uygulamıştır. Başarılı ve başarısız tarımsal durumları tanımlayan yazılı belgeler, web tabanlı bir sistemde saklanmaktadır ve özel bir probleme uygun bilgiler için bu belgeler sorgulanabilmektedir. Verilen bir problem için uygun tarım uzmanlarını çıkarma imkanı BÇ sistemine verilerek, sistemin problem çözme yeteneği iyileştirilmiştir. Bu yaklaşımda tarım uzmanları, yayınlarında ilgili terimlerin bulunmasına göre karakterize edilmiştir [36].

Hoa [37], Vietnamda tarımsal gelişim süreci için altyapının oluşturulması amacıyla bilimsel, teknolojik ve ekonomik bilgilerin kullanımını tanımlamıştır. Bu çalışmada Vietnam Tarım ve Kırsal Kalkınma Bilgi Merkezinde BÇ'nin kullanımını göstermiştir. Bu merkez; bitki türleri, gübreler ve tarımsal kimyasallar, yerli/yabancı tarımsal üretim ve ana pazarlardaki pirinç, buğday, mısır, kahve, kauçuk ve gübrenin uluslarası fiyatlarıyla ilgili bir çok veri tabanına sahiptir. Çıkarılan bilgiler; politika üretimi, stratejik planlama ve karar süreçlerinin desteklenmesinde, tarımsal ürünler

için ihracat pazarlarının araştırılmasında ve yabancı ülkeler arasında yatırım işbirliği ve ortak girişimleri arttırmada uygulanmıştır.

3.5.4 Web Madenciliği

Küresel bilgi paylaşımı için internet tabanlı hipermedya girişimi olarak 1989'daki icadından bu yana WWW çok hızlı bir gelişim göstermiştir. Web kullanıcılarının sayısı tam olarak bilinmemekle beraber tahminen üssel oranlarda artmıştır. WWW'da bulunan bilgi kaynakları çoğalmıştır ve web siteleri, programlar ve teknoloji sürekli olarak değişmiştir. Bütün bu yaşananlar kullanıcıya istenen bilgiyi bulmasında ve ağın yapısının ve kullanımının analizinde yardımcı olmak için gerekli otomatik araç ve gereçlere ihtiyacı ortaya çıkarmıştır. Bu yüzden, *web madenciliği* ya da WWW'da veri madenciliği tekniklerinin uygulaması, geçmiş yıllardaki birçok araştırma projesinin odak noktasını oluşturmuştur. Scime [38], web madenciliğinde kullanılan pratik uygulamalar ve hali hazırdaki araştırmalar konusunda kayıtlar tutmuştur. Web madenciliğinin genel tanımı "WWW'den kullanışlı bilgilerin keşfi ve analizi"dir [39]. Daha açık tanımı ise "WWW ile ilgili faaliyet ve yapay olgulardan, ilginç ve potansiyel olarak kullanışlı desenlerin ve üstü kapalı bilginin çıkarılmasıdır [40].

Web madenciliği araştırmaları web içeriği madenciğiliği, web yapısı madenciliği ve web kullanımı madenciliği olarak sınıflandırılabilir [41]. Web içeriği madenciliği, web sitelerinin içeriğinden kullanışlı bilgilerin keşfiyle ilgilenmektedir. Web yapısı madenciliği, web'in yapısı ve web siteleri arasındaki linkler konusuna yoğunlaşmaktadır. Web kullanımı madenciliği ise web kullanıcılarının giriş desenlerini çalışmaktadır.

Web İçeriği Madenciliği

Web içeriği madenciliği; metinler, resimler, ses ve video içeren web belgeleri içeriklerinden veya böyle belgelerin tanımından türetilmiş BÇ sürecidir. Online bilgi kaynaklarının otomatik olarak aranmasına yoğunlaşmaktadır. Web içeriği madenciliği teknikleri web de arama motorlarının ilk versiyonunda çoğunlukla kullanılan anahtar kelime çıkarmanın ötesine geçmiştir.

Web içeriği madenciliği kapsamında sistemler; kendi madencilik stratejilerine göre, belgelerin içeriklerinden doğrudan bilgi çıkaran sistemler, arama motorları ve web örümcekleri gibi diğer araçların arama sonuçlarını iyileştiren sistemler şeklinde sınıflandırılır [40]. İlave olarak web içeriği madenciliğinde, ajan temelli ve veritabanı yaklaşımları arasında farklılıklar oluşturulabilir [39].

Web'deki bilgiye ulaşmak ve analiz etmek için WebLog [42] ve ARANEUS [43] gibi veritabanı yaklaşımları; standart veritabanı sorgulama mekanizmaları ve veri madenciliği tekniklerini birleştirir. Bu yaklaşımlar; web'deki heterojen ve yarı yapılandırılmış verilerin, ilişkisel veri tabanları gibi daha çok yapılandırılmış veri kolleksiyonlarına birleştirilmesi üzerine yoğunlaşmaktadır.

Ajan temelli sistemler otomatik olarak kullanıcı adına web de uygun bilgileri araştırır ve organize eder. Bu sistemler üç kategoriye ayrılabilir. OCCAM [44], FAQ-Finder [45], ShopBot [46] gibi akıllı araştırma ajanları; özel bir alanı ve muhtemelen bir kullanıcı profilinin karakteristiklerini kullanarak uygun bilgileri araştırır. HyPersuit [47] gibi veri filtreleme ajanları; web belgelerini otomatik olarak çıkarmak, filtrelemek ve kategorize etmek için bilgi çıkartma tekniklerini kullanır. WebWatcher [48] gibi kişiselleştirilmiş web ajanları kullanıcı tercihlerini öğrenmek ve uygun web belgelerini keşfetmeyi hedefler.

Burada bazı web içeriği madenciliği projeleri kısaca açıklanmıştır:

- *Occam* [44] bir bilgi toplama motorudur. Kullanıcı istenilen bilgiyi, bir veri tabanı sorgusu olarak belirleyebilir. Occam istenen bilgiyi elde etmek için bir faaliyet planı oluşturmak amacıyla farklı siteler hakkında onunla ilgili türetilmiş bilgiyi kullanmaya çalışır.
- FAQ Finder [45] web'deki sıklıkla sorulan (FAQs, Frequently Asked Questions) sorular dosyasını kullanan otomatik soru cevap sistemidir. Kullanıcı sisteme herhangi bir konu hakkında bir soru yöneltir ve FAQ Finder bir cevap üretmesi yüksek muhtemel olan FAQ dosyasını bulur, bu dosya içerisinde benzer soruları araştırır ve kullanıcıya verilen cevapları getirir.
- ShopBot [46] karşılaştırmalı bir alışveriş ajanıdır. Online mağazaların ana sayfalarından girdiler alır, bu sitelerden nasıl alışveriş yapıldığını öğrenir, bu siteleri ziyaret eder, ürün bilgilerini alır ve sonuçları kullanıcı için özetler.
- WebWatcher [48] kullanıcıların web de sörf yapmasına yardım eden bir tur rehberi ajanıdır. Kullanıcının istediği bilgiye bağlı olarak kullanıcıya sayfadan sayfaya eşlik eder, uygun olduğuna inanılan linklerin altını çizer ve tecrübelerden öğrenir.
- CiteSeer [49] NEC Araştırma Enstitüsü tarafından yapılan bir araştırma olup dijital kütüphaneleri kullanmak için algoritmalar teknikler ve yazılım sağlamaktadır. Bütün bunlar NEC araştırma indeksi (http://citeseer.ist.psu.ed/) içerisinde yer verilmiştir. Bu araç web de gezinir ve bilimsel makalelerin yerini tespit eder; atıflar, atıf içeriği, makale başlığı, yazarlar vb. bilgileri çıkartır ve tam metin indekslemesi ve otomatik atıf indekslemesini gerçekleştirir.

Web Yapısı Madenciliği

Web yapısı madenciliği web'in link yapısından web sayfaları hakkında bilgi çıkarma işlemidir. Uygunluk ve kalite için web sayfaları arasındaki linkler, genellikle göstergeler olarak yorumlanır [50]. Bu yaklaşımdaki mantık; "sıklıkla başvurulan bir web sayfası, nadiren başvurulan bir web sayfasından daha önemlidir" esasına dayanır. Böylece bir belgenin başvurduğu web sayfası sayısı bu belgenin zenginliğini ve konu çeşitliliğini gösterebilir. Nihayetinde çok sayıda linke sahip bir belge muhtemelen daha iyi bir bilgi kaynağıdır.

Web yapısı madenciliği; araştırma, tarama ve trafik tahmininde uygulamara sahiptir. Örneğin, Pagerank [51] web link yapısındaki konumlarını esas alarak bütün web sayfalarının küresel bir derecelemesidir ve popüler araştırma motoru olan google'a (http://google.com/) temel oluşturur.

HITS (Hyperlink-Induced Topic Search) algoritması [50]; *otorite sayfaları* (sıklıkla başvurulan web belgeleri) ve *merkez sayfaları* (otorite sayfalarına link veren belgeler) bulmak için web sayfaları koleksiyonlarını araştırır ve ayrıca birçok web uygulamasında kullanılmıştır.

Yapısal web madenciliği projesi Clever [52]; link yapısı analizini esas alarak bir web kaynağını sınıflandıran birçok algoritmayı birleştirmektedir. SALSA algoritması [53], web belgeleri arasındaki link yapısından türetilen grafikler üzerinde rastgele yürüyüşleri esas alan bir stokastik web yapısı madenciliği yaklaşımıdır.

Web Kullanım Madenciliği

Her web sunucusu aldığı erişime ait kütük kaydı oluşturur ve saklar. Kütük dosyaları çok geniş yapılandırılmış bilgi koleksiyonlarını oluşturur. Web kullanım madenciliği aynı zamanda web kütük madenciliği olarak bilinir ve web sunucular tarafından tutulan kütüklerden ilginç kullanıcı desenlerinin otomatik keşif işlemidir [40]. Web kullanım madenciliğinden elde edilen türetilmiş bilgi, web sunucuları tarafından sağlanan hizmetlerin kalitesinin arttırmak için kullanılabilir. Örneğin web siteleri kullanıcıların ihtiyacına göre düzenlenebilir; web sitelerinin tasarımı ve navigasyonu iyileştirilebilir; elektronik ticaret için hedef kulalnıcılar belirlenebilir ve pazarlama kararlarına destek olunabilir. WebSIFT [52], WebLogMiner [40], Web Utilization Miner [54] gibi web kullanım madenciliğiyle ilgili birçok araştırma ve ticari proje bulunmaktadır. Web kullanım madenciliği genellikle üç basamaktan oluşmaktadır:

1. *Ön işleme*: bu aşama uygun olmayan ve gürültülü verinin çıkarılması ve verilerin kullanışlı veri özetleri şeklinde gruplanması gibi veri temizlemeyi içermektedir;
2. *Desen keşfi*: bu aşama veriden türetilmiş bilginin çıkartılması için uygun metotlar ve algoritmaların uygulanmasını içermektedir. İstatistiksel analiz, birlik kuralları, kümeleme ve sınıflandırma gibi metotlar kullanılmaktadır.
3. *Desen analizi*: bu aşama ikinci aşamada bulunan desenlerin veya kuralların anlaşılması içermektedir ve alakasız desenler filtrelenmektedir.

Tarımda Web Madenciliği Uygulamaları

Web madenciliği teknikleri; tarım sektöründe kullanılacak bilgi ve türetilmiş bilgi yayım sistemleri ve karar destek sitemleri için hayati bir temel oluşturur. Pan [55], tarım sektöründeki kullanıcılar arasında bilginin yayılmasının arttırılması amacıyla, web'den bilgi kaynaklarının toplanması ve yönetilmesine ait metotlar ve teknikleri değerlendirmiştir.

Gandhi [56] online kaynaklardan otomatik olarak bilgi toplayan ve bunu işleyerek çiftçilere, parekende satıcılarına, toptan satıcılara ve devlet görevlilerine ileten özel bir market bilgi sistemini tanımlamaktadır.

Kurlavicius ve Kurlavicius [57], web tabanlı bir tarım sektörü karar destek sistemini tanıtmışlardır. Bu sistem, verilen bir tarımsal problem için optimum çözümü araştırarak tarım sektöründe stratejik planlama ve işlevsel karar almaya yardım etmektedir. Bu yaklaşım hem web madenciliği hemde bilgi çıkarma tekniklerini birlikte kullanmaktadır. Uygun belgeler için bilgi ajanları WWW'yu otomatik olarak araştırmakta ve bu belgeleri kategorize etmektedir. Alan bilgi ajanları ise toplanan bilgilerden alana özel türetilmiş bilgileri çıkarmaktadır.

Kaynaklar

1. Date, C. J. 2000. *An Introduction to Database Systems.* 7th ed. Reading, MA: Addison-Wesley Publishing Company.
2. Codd, E. 1970. A relational model for large shared data banks. Communication of the ACM 13(6): 377-387.
3. Codd, E. 1982. Relational database: A practical foundation for productivity. *Communication of the ACM* 25(2): 109-117.
4. Elmasri, R., and S. B. Navathe. 2000. *Fundamentals of Database Systems*. 3rd ed. Reading, MA: Addison-Wesley.
5. Melton, J., ed. 2003. Database Languages—SQL. ISO/IEC 9075-1:2003.
6. Kuper, G., L. Libkin, and J. Paredaens. 2000. *Constraint Databases.* Heidelberg, Germany: Springer-Verlag.
7. Stonebraker, M., P. Brown, and D. Moore. 1999. *Object-Relational DBMSs - Tracking the Next Great Wave.* 2nd ed. San Francisco, CA: Morgan Kaufman Publishers.
8. Ullman, J. D., and J. Widom. 1997. *A First Course in Database Systems.* Upper Saddle River, NJ: Prentice-Hall, Inc.
9. Mannila, H. 1996. Data mining: Machine learning, statistics, and databases. *Proc. of the 8th International Conference on Scientific and Statistical Database Management*, 1-6.
10. Keim, D., and H. Kriegel. 1996. Visualization techniques for mining large data-bases: A comparison. *IEEE Transactions on Knowledge and Data Engineering, Special Issue on Data Mining* 8(6): 923-938.
11. Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 1996: 37-54.
12. Mannila, H. 1997. Methods and problems in data mining. *Proc. of International Conference on Database Theory*, 41-55.
13. Witten, I.H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* 2nd ed. San Francisco, CA: Morgan Kaufmann.
14. http://www.sas.com/technologies/analytics/datamining/miner/
15. http://www.spss.com/spssbi/clementine/
16. Fayyad, U., D. Haussler, and P. Stolorz. 1996. VTBK for Science data analysis: Issues and examples. *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining* (VTBK-96), 50-56.
17. HGM, J., R. B. Altmad, V. Kumar, H. Mannila, and D. Pregibon. 2002. Emerging scientific appligations in data mining. *Communication of the ACM* 45(8): 50-58.

18. Lloyd-Williams, M. 1997. Discovering the hidden secrets in your data-The data mining approach to information. SCGISA & RRL.net Workshop on Health and Crime Data Analysis.
19. Apte, C., B. Liu, E. P. D. Pednault, ann P. Smyth. 2002. Business applications of data mining. *Communication of the ACM* 45(8): 49-53.
20. Dunham, M. H. 2002. *Data Mining: Introductory and Advanced Topics:* Englewood Clifts, NJ: Prentice Hall, Inc.
21. Larose, D. T. 2000. *Discovering Knowledgein Data: An Introduction to Data Mining.* New York, NY: John Wiley.
22. Weiss, S. I., and C. Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems.* San Francisco, CA: Morgan Kaufmann.
23. Persidis, A. 2000. Data mining in biotechnology. *Nature Biotechnology* 18: 237-238.
24. Little, B., W. Johnston, A. Lovell, S. Steed, V. O'Conner, G. Westmorland, and D. Stonecypher. 2001. Data mining U.S. corn fields. *Proc. of the First SIAM International Conference on Data Mining* 1: 99-104.
25. http://www.gis.uiuc.edu/cfardatamining/default.htm
26. Canteri, M. G, B. C. Ávila, E. L. dos Santos, M. K. Sanches, D. Kovalechyn, J. P. Molin, and Gimenez. 2002. Application of data mining in automatic description of yield behavior in agricultural areas. *Proc. of the World Congress of Computers in griculture and Natural Resources,* 183-189.
27. van Rijsbergen, C. J. 1975. *Information Retrieval*. London, UK: Butterworths.
28. Salton, G., and M. J. Gill. 1983. *Introduction to Modern Information Retrieval.* New York, NY: McGraw-Hill Book Co.
29. Baeza-Yates, R., and B. Ribeiro-Neto. 1999. *Modern Information Retrieval.* Harlow, UK: Addison-Wesley.
30. Salton, G. 1971. The SMART *Retrieval System–Experiments in Automatic Document Processing.* Englewood Cliffs, NJ: Prentice-Hall Inc.
31. Robertson, S. E., and K. S. Jones. 1976. Relevance weighting of search terms. J. *American Society for Information Science* 27: 129-146.
32. Wong, S. K. M., W. Ziarko, and P. C. N. Wong. 1985. Generalized vector space model in information retrieval. *Proc. of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*
33. Dumais, S. T. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers* 23: 229-236.
34. Rocchio, J. J. 1971. Relevance feedback in information retrieval. *The SMART Retrieval System–Experiments in Automatic Document Processing,* ed. G. Salton, 313-323. Englewood Cliffs, NJ: Prentice-Hall Inc.
35. Otuka, A. 1999. Case retrieval and management system for agricultural case base. *Proc. of the 2nd Conference of the European Federation for Information Technology in Agriculture, Food and the Environment.*
36. Otuka, A. 2000. Retrieval of specialists and their works for agriculture case base. *Proc. of the 2nd Asian Conference for IT in Agriculture.*
37. Hoa, T. T. T. 1998. Database for Agriculture in Information Centre for agriculture and rural development of Vietnam. *Agricultural Information Technology in Asia and Oceania,* 25-28.
38. Scime A. 2004. Web Mining. Idea Group Inc. (IGI).
39. Cooley, R., J. Srivastava, and B. Mobasher. 1997. Web mining: Information and pattern discovery on the World Wide Web. *Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence.*

40. Zaïane, O. R. 1999. Resource and knowledge discovery from the internet and multimedia repositories. PhD thesis. B.C., Canada: Simon Fraser University.
41. Kosala, R., and H. Blockeel. 2000. Web mining research: A survey. *ACM SIGKDD Explorations* 2(1): 1-15.
42. Lakshmanan, L., F. Sadri, and I. N. Subramanian. 1996. A declarative language for querying and restructuring the web. Proc. of the 6th International Workshop on *Research Issues in Data Engineering: Interoperability of Nontraditional Database Systems (RIDE-NDS'96).*
43. Merialdo P., P. Atzeni, and G. Mecca. 1997. Semistructured and structured data in the web: Going back and forth. *Proc. of the Workshop on the Management of Semistructured Data.*
44. Kwok, C., and D. Weld. 1996. Planning to gather information. *Proc. of the 14th National Conference on Artificial Intelligence.*
45. Hammond, K., R. Burke, C. Martin, and S. Lytinen. 1995. Faq-finder: A casebased approach to knowledge navigation. *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments.* Menlo Park, CA: AAAI Press.
46. Doorenbos, R. B., O. Etzioni, and D. S. Weld. 1997. A scalable comparisonshopping agent for the World-Wide Web. *Proc. of the 1st International Conference on Autonomous Agents.*
47. Weiss, R., B. Velez, M. A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. K. Gifford. 1996. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. *Proc. of the 7th ACM Conf. on Hypertext.*
48. Joachims, T., D. Freitag, and T. Mitchell. 1997. WebWatcher: A tour guide for the World Wide Web. *Proc. of the 11th International Joint Conference on Artificial Intelligence.*
49. Lawrence, S., K. Bollacker, and C. L. Giles. 1999. Indexing and retrieval of scientific literature. *Proc. of the 8th International Conference on Information and Knowledge Management.*
50. Kleinberg, J. 1999. Authorative sources in a hyperlinked environment. *Journal of the ACM* 46(5): 604-632.
51. Page, L., S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank Citation Ranking: Bringing order to the web. Technical Report, Stanford Digital Library Technologies Project. 184 Chapter 3 Methods, Algorithms, and Software
52. Chakrabarti, S., B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. 1999. Mining the web's link structure. *IEEE Computer* 32(8): 60-67.
53. Lempel, R., and S. Moran. 2001. SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems* 19: 131-160.
54. Spiliopoulou, M. 1999. The laborious way from data mining to web mining. International *J. Computer Systems Science and Engineering, Special Issue on "Semantics of the Web"* 14: 113-126.
55. Pan, S. 2004. Approaches on collecting and integrating of agricultural network information resources. *AFITA/WCCA Joint Conference on IT in Agriculture.*
56. Gandhi, V.P. 2004. An IT based market information system for improving marketing efficiency of fruits and vegetables in India. *AFITA/WCCA Joint Conference on IT in Agriculture.*
57. Kurlavicius, A., and G. Kurlavicius. 2004. Agricultural information and farm management decision support by internet. *AFITA/WCCA Joint Conference on IT in Agriculture.*