T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET İÇERİK MADENCİLİĞİNDE YAPAY SİNİR AĞLARI VE BİR UYGULAMA

Gülşah AYNEKİN

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA 2006

T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET İÇERİK MADENCİLİĞİNDE YAPAY SİNİR AĞLARI VE BİR UYGULAMA

Gülşah AYNEKİN

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA 2006

T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İNTERNET İÇERİK MADENCİLİĞİNDE YAPAY SİNİR AĞLARI VE BİR UYGULAMA

Gülşah AYNEKİN

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

Bu tez …….... tarihinde aşağıdaki jüri tarafından oybirliği/oy çokluğu ile kabul edilmiştir.

Yrd.Doç.Dr. Seda ÖZMUTLU Yrd.Doç.Dr. Mehmet AKANSEL Prof.Dr. Feray ÇELİKÇAPA (Danışman)

ÖZET

Web, her gün milyonlarca insanın bilgi almak için ulaştıkları çok popüler bir kaynaktır ve sürekli değişime uğramaktadır. Arama motorları ve Web derleyicileri, aradıkları bilgiye Web üzerinden ulaşmak isteyen kullanıcılar için geliştirilmiş araçlardan bir kaçını oluşturmaktadır. Web derleyicileri tarafından yapılan, Web'deki doğru ve ilgili bilginin bulunması görevi çok zor bir iştir. Bu çalışmada kuş gribi ilgili siteleri bulmaya çalışan bir derleyici geliştirilmiştir. Derleyicinin konu ile ilgili siteleri başarılı şekilde bulması sağlamak için yapay sinir ağları tekniği kullanılmaktadır

Özelleştirilmiş derleyici, Web üzerindeki ilgili sayfaları bulmak, indekslemek ve devamında güncellemek için tasarlanmıştır. Geliştirilen sistem ile diğer arama yaklaşımlarında saklı ilgili sayfaları bulmayı amaçlanmaktadır.

Uygulanan Web derleyici, örneklerdeki her kayıt için "ilgili" ve "ilgisiz" şeklinde atamalar yapmaktadır. Derleyici tarafından yapılan atamalar, insanlar tarafından yapılan atamalarla karşılaştırılmış ve derleyicinin performans ölçüleri hesaplanmış ve incelenmiştir.

Anahtar Kelimeler

Veri Madenciliği, Web Madenciliği, Web İçerik Madenciliği, Yapay Sinir Ağları, Web Derleyici, Özelleşmiş Derleyici.

ABSTRACT

The Web is a very popular source for information. Millions of people access the Web everyday to retrieve various types of information from the Web. Search engines, and Web crawlers are some of the search tools that Web users use to reach the required information on the Web. Mining for correct and relevant information in the World Wide Web is a difficult task, handled by Web crawlers. This study outlines the components of a specialized crawler on the topic of Avian Influenza that heavily makes use of artificial neural networks to establish successful focused crawling on the topic of Avian Influenza.

Specialized crawling is designed to find, index and follow the updates of Web pages of interest, and proposes new approaches for reaching relevant pages, which might stay hidden to other crawling approaches.

Using Web crawler “interested” or “uninterested” assignments are made for each record in the sample. The assignments which are made by proposed crawler is compared with the assignments which are made by humans, and performance measures are calculated and analyzed.

Keywords

Data Mining, Web Mining, Web Content Mining, Neural Netwoks, Web Crawler, Focused Crawling

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1 - GİRİŞ

Verilerin dijital ortamda saklanmaya başlanması ile birlikte, yeryüzündeki bilgi miktarının her 20 ayda bir iki katına çıktığı günümüzde veri tabanlarının sayısı da benzer, hatta daha yüksek bir oranda artmaktadır. Birçok veri tabanını insanların kullanımına sunan İnternet aracılığı ile ulaşılabilen zengin bilgi kaynakları, günümüz araştırmacılarına cazip bir veri arama ortamı sunmaktadır. Web üzerindeki erişilebilir bilgi kaynakları, sayılarının çok fazla olması sebebi ile sınıflandırıldıktan sonra kişilere sunulmaktadır. İlk zamanlarda insan aracılığı ile yapılan sınıflandırma, günümüzde doküman sayısının hızlı bir şekilde artması nedeniyle otomatik olarak yapılır hale gelmiştir. Bunun için de bazı ön tanımlı kelimeler ile web sayfaları sınıflandırılmaktadır.

Web ortamından bilgi sağlamakta, arama motorları (search engine) ve web dizinleri kullanılmaktadır. Arama motorları Web'deki çok sayıda dokümanın içeriğinin kelime bazında aranmasını sağlar. Belirtilen anahtar sözcüklerin, arama motorunun veritabanındaki milyonlarca sözcük ile karşılaştırılması olanağının bulunması avantaj sağlamakla birlikte, belirli bir konu ile ilgili doğru anahtar sözcüklerin bulunması önemlidir. Sorgulama sonucu uygun bir sıra ile doküman listesi halinde gösterilir. Fakat, liste genellikle çok uzundur ve kullanıcılar ya isteksizdir ya da araştırmayı daraltacak karmaşık sorgular oluşturmak için yeterli beceriye sahip değildir. Web ortamının oldukça farklı çeşitlilikte materyaller içermesi nedeniyle, kullanıcılar ilgilerini ifade edecek kaynaklar bulmakta zorlanmaktadır ve kelime belirsizliği nedeniyle sürpriz sonuçlar oluşmaktadır.

Diğer bir yöntem olan web dizinleri yani web sayfalarının sınıflandırılıp kullanıcıya sunulmasıdır. Buna göre İnternet kullanıcılarının, İnternet üzerinde aradıkları bilgilere ulaşabilmeleri için konulara göre düzenlenmiş başlıkların yer aldığı İnternet siteleri bulunmaktadır. İnternet kullanıcıları bu başlıkları tarayarak aradıkları bilgilere ulaşabilirler. Web indeksleri, başka sayfalara bağlanmayı sağlayan, çeşitli sözcüklerden oluşan listelerdir. İndekste ana konulardan başlayarak ve giderek alt düzeylere inerek, aranılan konuya doğru ilerlenebilir. Öte yandan, indeks kullanan kullanıcılar, kendileri için en uygun seçeneği indeks üzerinde kendileri bulmak ve

ayrıca, karşılarına çıkacak olan ve çok sayıda sayfadan oluşan sayfa listesinde hangi sayfaların işlerine yarayacağını kendileri belirlemek zorundadırlar. Çoğu kullanıcının böyle bir işlemi tam olarak yerine getirecek kadar zamanı olmamaktadır. Geçmiş yıllarda, İnternet kullanıcılarının aradıkları bilgilere ulaşmak için indeksleri kullanmaları mantıklı olmakla birlikte, günümüzde İnternet'in çok büyük bir bilgi kaynağı haline gelmesiyle, indeksleri kullanarak arama yapmak giderek daha da zorlaşmıştır. Bu yüzden, günümüzde İnternet kullanıcılarının büyük bir çoğunluğu, İnternet'te aradıkları bilgilere ulaşabilmek amacıyla arama motorlarından yararlanmaktadırlar. İlk zamanlarda insan gücü ile yapılan bu çalışma son zamanlarda otomatik olarak yazılan uygulamalar aracılığı ile yapılmaya başlamıştır.

Yahoo'dan bir yetkili "Web bizden daha hızlı gelişmektedir. Yeni web sitelerini geldikleri oranda karşılamak için yeterli personel sağlamak mümkün değil." demektedir. (Akbaba, 2003). Problem gönderilen bir sitenin zaman içerisinde içeriğinin geliştirilebildiği düşünüldüğünde daha da artmaktadır.

Bu çalışmada genel arama motorlarına alternatif olarak, kullanım amacına göre istenilen herhangi bir spesifik konuda ilgili siteleri bulan bir derleyici (crawler) yapısı geliştirilmiştir. Bu tip arama motorlarının geliştirilmesi, güncel Web madenciliği çalışmalarının bir dalıdır ve "focused crawling" olarak adlandırılmaktadır. Bu çalışmanın uygulamasında güncel bir konu olan kuş gribi (Avian Influenza) seçilmiştir.

Veri madenciliği tekniklerinin web verileri üzerinde uygulanmasını konu alan web madenciliğinin bir dalı olan web içerik madenciliği kapsamına giren çalışmada, yapay sinir ağları tekniği kullanılmıştır. Kullanılan yapay sinir tek saklı katmanlı ve ileri beslemeli olup bir denetimli öğrenme tekniği olan geri yayılım algoritmasını kullanmaktadır. Ağın giriş katmanında, anahtar kelime sayısında nöron bulunmaktadır. Giriş verileri anahtar kelimelerin ilgili sitelerde olup olamama durumuna göre kelimelerin ağırlıklarından oluşmaktadır. Çıkış katmanındaki nöron sayısı bir tanedir ve değerlendirilen sitenin durumuna göre "ilgili" veya "ilgisiz" diye iki değerden birisini alır.

Geliştirilen derleyici Web üzerinde arama işlemine, çalışma öncesi verilen konu ile ilgili referans sitelerdeki linklerden başlamaktadır. İlgili bulduğu sitelerin sayfaları

belleğe kaydedilmekte ve bu yeni sitelerde verilen linkler ile aramaya istenildiği kadar devam edebilmektedir.

Karar mekanizmasını oluşturan yapay sinir ağının eğitim yapısına göre iki farklı versiyon derleyici geliştirilmiştir. Birinci statik versiyonda, başlangıçta verilen sitelerle sistem bir kere eğitilmekte ve derleyici çalışırken ağda kullanılacak ağırlıklar tespit edilmektedir. İkinci dinamik versiyonda ise ilk versiyonda olduğu gibi sistem başlangıçta bir kere eğitilmekte fakat daha sonra derleyici çalışırken çok ilgili bir site bulursa bu sitenin eğitimde kullanılıp kullanılmak istendiği kullanıcıya bir arayüz vasıtasıyla sorulmakta ve olumlu cevap alınırsa yeni sitenin eklendiği eğitim kümesine göre yeniden eğitim uygulanmaktadır. Sonuçta oluşan yeni ağırlıklara göre derleyici arama işlemine kaldığı yerden devam etmektedir. Bu iki versiyon ile ilgili çeşitli faktörler belirlenmiş ve bu faktörlerin değişik kombinasyonlarında çeşitli deneyler yapılmıştır. Sonuçlar karşılaştırmalı olarak değerlendirilmiştir.

Bu çalışma aşağıdaki bölümlerden oluşmaktadır:

İkinci kısımda Web yapısı ile arama motorları ve Web içerik madenciliği ile ilgili geçmişteki çalışmalardan bahsedilmiştir. Üçüncü kısımda materyal bölümünde sırasıyla veri madenciliği, Web madenciliği ve yapay sinir ağı yapıları hakkında bilgi verilmiştir. Yöntem kısmında ise yapılan uygulamanın yapısı tanıtılmıştır. Dördüncü kısımda ise geliştirilen uygulama deney sonuçlarına göre değerlendirilmiştir. Beşinci bölüm tartışmada ise çalışmanın genel sonucu ve ilerleme yönü incelenmiştir.

2 – KONU İLE İLGİLİ ÇALIŞMALAR

2.1 – Kuramsal Bilgiler

İnternet, birçok bilgisayar sisteminin birbirine bağlı olduğu, dünya çapında yaygın olan ve sürekli büyüyen bir iletişim ağıdır. İnternet, insanların her geçen gün gittikçe artan "üretilen bilgiyi saklama/paylaşma ve ona kolayca ulaşma" istekleri sonrasında ortaya çıkmış bir teknolojidir. Bu teknoloji yardımıyla pek çok alandaki bilgilere insanlar kolay, ucuz, hızlı ve güvenli bir şekilde erişebilmektedir. İnternet'i bu haliyle bir bilgi denizine, ya da büyükçe bir kütüphaneye benzetebiliriz. İnternet'e, bakış açımıza bağlı olarak farklı tanımlamalar da getirebiliriz : İnternet,

- İnsanların kendi aralarında etkileştiği, bilgi değiş tokuşu yapabildiği ve kendi yazısız kuralları olan büyük bir topluluktur. Bu, İnternet'in sosyal yönüdür.
- Pek çok yararlı bilginin bir tuşa basmak kadar yakın olduğu dev bir kütüphanedir.
- Bir çok bilgisayarın bağlı olduğu çok büyük bir bilgisayar ve iletişim ağıdır.
- Kişilerin değişik konularda fikirlerini serbestçe söyleyebilecekleri ortamlar barındıran bir demokrasi platformudur.
- Evden alışveriş, bankacılık hizmetleri, radyo-televizyon yayınları, günlük gazete servisleri vb gibi uygulamaları ile aslında İnternet aynı zamanda bir hayat kolaylaştırıcıdır.

Sonuç olarak, İnternet, önümüzdeki yıllarda üretilecek bilgilerin dolaşım sistemidir. Ticari boyutunun da ortaya çıkmasıyla yaşamla daha çok iç içe geçmeye başlamıştır. (http://www.po.metu.edu.tr/links/inf/css25/bolum1.html#1)

Günümüzde, İnternet denildiğinde, genelde, İnternet'in sadece belirli bir parçası olan World Wide Web (WWW) veya kısaca Web anlaşılmaktadır. Web sitelerini görmek için kullanılan programlara da Web tarayıcı (Web browser) adı verilmektedir. Web, bilgiye erişim sağlayan kaynakları düzenler. Bunu yaparken de Web sayfaları (Web pages) veya Web belgeleri (Web documents) denilen dosyaları kullanır.

Bir Web sitesi ise içinde bir veya daha fazla sayıda Web sayfası bulundurabilen bir Web alanıdır. Her Web sayfasının kendine has bir evrensel kaynak konumlandırıcısı (Universal Resource Locator – URL) vardır. URL, bir Web dosyasının bulunduğu yeri tam olarak ifade eden Web adresidir (Çavdur, 2005).

Web üzerinde bulunan bazı Web siteleri, İnternet kullanıcılarının aradıkları bilgileri bulmalarına yardım etmek amacıyla tasarlanmışlardır. Bu tür siteler, arama motoru (search engine) adı verilen bir yazılım içerirler. Arama motorları, aranılan bilgiye ulaşılmasına yardımcı olan araçlardır. İnternet kullanıcıları, ihtiyaçları olan bilgiyi aratmak için anahtar sözcükler yazarak arama işlemi yaptırabilirler.

Başlıca arama motorları aşağıdaki gibidir,

- www.altavista.com
- www.google.com
- www.lycos.com
- www.hotbot.com
- www.infoseek.com
- www.askjevees.com
- www.looksmart.com
- www.overture.com
- www.inktomi.com
- www.alltheweb.com
- www.yahoo.com

Daha önce giriş bölümünde de belirtildiği gibi, web de arama işlemi arama motorları veya dizinler ile yapılır. Şekil 2.1'den Google arama motoru Şekil 2.2 'den de aynı arama motorunun dizin arama alternatifi görülmektedir.

Şekil 2.1: Google Arama Motoru

Şekil 2.2: Google Dizin Arama

2.2 – Kaynak Araştırması

Web madenciliği kabaca Web'den faydalı bilginin keşfi olarak tanımlanmaktadır. Detaylı olarak Bölüm 3.1.2 açıklanan Web madenciliği hakkında yapılan araştırmalar, Web kullanım madenciliği ve Web içerik madenciliği konusunda yoğunlaşmıştır. Genel anlamda, otomatik tarama, bilgi alma ve kullanılabilir

kaynakların milyonlarca web sitesi veya online veritabanlarından seçilmesi Web içerik madenciliği konusuna girerken bir veya birçok web sunucu veya online servisten kullanıcı erişim desenlerinin analiz ve keşfi web kullanım madenciliği konusuna girmektedir.

Web kullanım madenciliği çalışması sonucunda bir ziyaretçinin sitede kalma süresi, hizmet stratejileri, etkin kampanyalar ve diğerleri bulunabilir. Bu sayede elektronik ticaret sitesi için en iyi müşteri Web kullanım madenciliği sayesinde tespit edilebilir. Bu nedenle Web madenciliğindeki çalışmaların bir çoğu Web kullanım madenciliği konusunda yoğunlaşmıştır.

Kullanım madenciliği ve içerik madenciliği ile ilgili çalışmalar bu bölümde sıralanabilir:

2.2.1 - Web Kullanım Madenciliği Çalışmaları

Jansen ve Spink (2003) çalışmalarında, arama motoru kullanıcılarının davranışlarının sürekliliğini incelemişlerdir. Sonuçta, kullanıcı davranışlarının incelenmesi amacıyla yapılan çalışmaların çoğunun Amerika merkezli olduğu ve diğer ülkelerdeki kullanıcılar için bu türdeki çalışmaların çok az olduğu belirtilmiştir.

Özmutlu ve arkadaşları (2003) Excite ve FAST arama motorları verilerini incelemişler ve arama motorları kullanıcılarının sorgularında gün içinde değişiklikler olduğunu tespit etmişlerdir. Yazarlar yapılan çalışmada, oturum ve sorguların gelişleri ve süreleri gibi bazı özelliklerin, sabah saatlerinde en yüksek düzeylerinde olup, ilerleyen zamanlarda azaldığını belirlemişlerdir. Bu analizden elde edilen sonuçların ve daha başka veri kümelerinin de bu şekilde incelenmesinin, arama motorlarının arama yapılarını değiştirilmesinde kullanılabileceği üzerinde durulmuştur.

Özmutlu ve arkadaşları (2002), çalışmalarında, büyük Web veri kümelerinin analizi için etkili bir örnekleme stratejisi geliştirmek için Poisson ve sistematik örnekleme tekniklerini değerlendirmişlerdir. Yazarlar çalışmalarında Excite arama motoru verilerini kullanmışlardır. Web arama motorları kayıtlarının çok büyük

boyutlara ulaştığını belirtmişler ve bu boyutlardaki veriler üzerinde, istatistiksel yöntemleri uygulamanın zorluğuna dikkat çekmişlerdir.

Spink ve Özmutlu (2002), soru biçimli Web sorgularının özelliklerini incelemişlerdir. Yazarlar çalışmalarında iki arama motorunu incelemişlerdir. Bunlar, sorguların soru biçiminde girilmesine yönlendiren Ask Jeeves ve bu şekilde bir yönlendirme yapmayan Excite arama motorlarıdır. Çalışma sonucunda, yazarlar, soru biçimli Web sorgularının yapısı için genel kalıplar bulunduğunu ve bu kalıpların daha başka veriler için de test edilmesi gerekliliğini vurgulamışlardır. İleriki araştırmalarda, soru biçimli ve soru biçimli olmayan sorgulardan hangisinin daha fazla sayıda anlamlı terim içerdiğinin tespit edilmesi, soru ve istek biçimli sorgu yapılarının nasıl geliştirilebileceği gibi konular hakkında çalışmalar yapılmasının kapsamı genişleteceği belirtilmiştir.

Spink ve arkadaşları (2004), çalışmasında, cinsellikle ilgili bilgi için Web araması konusunu incelemişlerdir. Yazarlar, cinsellikle ilgili sorguların oranını belirlemeyi ve cinsellikle ilgili ve cinsellikle ilgili olmayan sorguların özelliklerini incelemeyi amaçlamışlardır. Genel olarak, cinsellikle ilgili olan sorgular, diğerlerine göre daha uzun olup, cinsellikle ilgili bir oturum büyük bir olasılıkla 20 sorgudan daha uzun olmaktadır. Ayrıca, yaptıkları çalışmanın tek bir arama motoru verilerine dayalı olmasından dolayı sınırlı olduğunu ve biri Avrupa diğeri de Amerika bazlı olmak üzere iki ayrı arama motoru kayıtlarını, bu şekilde incelemekte olduklarını belirtmişlerdir.

He ve arkadaşları (2002), arama motorları kullanıcı oturumlarındaki konu değişikliklerini tespit etmek için delil birleştirme yaklaşımı uygulamışlardır. Çalışmada, kullanıcılarının arama konularını değiştirmeleri halinde, bunu otomatik olarak tespit edecek bir yaklaşım geliştirilmiştir. Arama motoru kayıtlarından hesaplanan deliller, Dempster-Shafer teorisini kullanarak birleştirmektedir. Yaklaşım, Reuters arama motorunun kayıtlarıyla test edilmiştir. Çalışmada kullanılan veriler, kullanıcıların IP adresleri, arama süreleri ve sorgularından oluşmaktadır. Dempster-Shafer teorisi kullanılarak elde edilen olasılıkları "konu değişikliği yok" veya "konu değişikliği var" şekline dönüştürmek amacıyla bir eşik değeri kullanılmaktadır. Sonuçta, her kayıt için "konu değişikliği yok" veya "konu değişikliği var" şeklinde atamalar yapılmaktadır.

Yaklaşımın performansını değerlendirmek için “duyarlık” ve “anma” performans ölçülerini birlikte dikkate alan bir performans ölçüsü kullanılmıştır.

2.2.2 - Web İçerik Madenciliği Çalışmaları

Chuang ve Chien (2002), bir sorgu-sınıflandırma yaklaşımı geliştirmişlerdir. Bunun için üç arama motorunun (Dreamer, GAIS ve Openfind) kayıtları toplanmış ve bu kayıtlar üzerinde, popüler Web arama konularını gösterecek şekilde, 14 ana ve 100 alt kategoriden oluşan iki aşamalı bir konu sınıflandırması yapılmıştır. Sonuçta, deneysel sonuçların, kullanıcılar tarafından verilen çeşitli yeni terimlerin otomatik olarak sınıflandırılabildiğini göstermiş ve bu yüzden, çalışmalarının, sorgu terimleri kullanarak Web sınıflandırmaları yapmakta veya var olanları geliştirmekte iyi bir başlangıç olabileceği saptanmıştır.

Akbaba (2003), Web sayfalarının otomatik olarak sınıflandırılması üzerine en yakın k komşuluğu tekniği kullanarak bir algoritma geliştirmiştir. Araştırmacı eğitim sayfalarının ve anahtar kelimelerin doğru seçimi ve eşik değerinin doğru tespiti ile beraber otomatik sınıflandırmanın başarılı sonuçlar ürettiği sonucuna varmıştır.

Vaughan (2003), çalışmasında, arama motorlarının değerlendirilmesi için yeni ölçüler önermiştir. Yazar, arama motorları için kullanılan değerlendirme ölçülerinin duyarlık (precision) ve anma (recall) olduğunu belirtmiştir. Bu geleneksel değerlendirme ölçülerine ek olarak bazı yeni ölçülerin de geliştirildiği belirtilmiş ve bu çalışmada yeni ölçülerin test edilmesi amacıyla üç arama motorunun kayıtları kullanılmıştır (Alta Vista, Google ve Teoma). Geleneksel ölçülerin tamamlayıcısı olarak önerilen bu yeni ölçüler, sonuç sıralama kalitesi ve en üstte sıralanan sayfaları getirme yeteneği şeklindedir. Test edilen bu yeni ölçüler ile geleneksel ölçüler arasındaki fark; yeni ölçülerde test belgelerinin sürekli sıralaması (en fazla ilgili olandan en az ilgili olana doğru sıralama) baz alınırken, diğerlerinde kesikli ilgi yargısı (ilgili, kısmen ilgili, ilgisiz gibi) bulunmasıdır. Çalışma sonuçları, yeni ölçülerin, incelenen üç arama motorunun performansını ayırt edebildiğini göstermiş ve bu ölçülere göre en iyi performansı gösteren arama motorunun Google olduğu saptanmıştır Farklı ölçülerin kullanıldığı başka bir çalışmada Google’ın yine daha iyi bir performans gösterdiği ve

aynı zamanda, Google'ın göreceli üstünlüğünün, bu arama motorunun Web üzerindeki en popüler arama motoru olmasından da anlaşılmakta olduğu belirtilmiştir.

Kak ve Shu (1999) çalışmalarında, bir arama motoru geliştirmişlerdir. Bu arama motoru, konu ile ilgili arama sonuçlarını yapay sinir ağı kullanarak sınıflandırmaya çalışmaktadır. Hızlı bir öğrenme tekniği olduğu için yapay sinir ağı kullanılmıştır. Geliştirilen arama motoruna "Anvish" adı verilmiştir. Bu arama motoru ile diğer büyük arama motorlarındaki arama sonuçlarındaki kayıtlardan az ilgili olanların elimine edilmesi amaçlanmıştır.

Aggarwal ve arkadaşları (2000), çalışmalarında özelleştirilmiş bir derleyici geliştirmişlerdir. Derleyici URL adreslerinin yapısına, sitelerdeki linklere ve benzeri tekniklere göre istenen konularda arama yapmaktadır. Burada amaç Yahoo, Alta Vista gibi arama motorlarının verdiği karmaşık ve uzun sonuç listesinden daha kısa ve basit sonuçlar çıkarmaktır.

Chakrabarti (1999), çalışmasında yarı denetimli bir derleyici önermiştir. Bu sayede uzman kişinin bilgisinin de sistemin verimliliğini arttırmak için kullanılması amaçlanmaktadır. Araştırmacı, sistemin verimliliğini arttırmak için anahtar kelime listesinin sürekli güncellenmesini önermiştir.

3 – MATERYAL VE YÖNTEM

3.1 – Materyal

3.1.1 – Veri Madenciliği

Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir (Vahaplar ve İnceoğlu, 2001).

Veri madenciliği ile büyük veri yığınlarından oluşan veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır (Eker, 2004).

Araştırmacıların, geniş, çok hacimli ve dağınık veri setleri üzerinde yapmış oldukları çalışmalar sonucu aşağıdaki sonuçlara varılmıştır (Vahaplar ve İnceoğlu, 2001).

- Veri madenciliği ve bilgi keşfi, özellikle elektronik ticaret, bilim, tıp, iş ve eğitim alanlarındaki uygulamalarda yeni ve temel bir araştırma sahası olarak ortaya çıkmaya başlamıştır. Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini formüle analiz etmeye ve uygulamaya yönelik çalışmaların bütününü içerir. Geniş veri kümelerinden desenleri, değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılır. Bu sayede, web üzerinde filtrelemeler, DNA sıraları içerisinde genlerin tespiti, ekonomideki eğilim ve düzensizliklerin tespiti, elektronik alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edilebilir.

- Sayısal verinin miktarı, son 10 yılda bir patlama yaşayarak tahminlerin dışında bir artış göstermiştir. Buna karşılık, bilim adamlarının, mühendislerin ve

analistlerin sayısı değişmemektedir. Bu orantısızlığı gidermek için yeni araştırma problemlerinin çözümleri birkaç gruba ayrılabilir :

1. Geniş hacimli ve çok boyutlu veri madenciliği için yeni algoritma ve sistemlerin geliştirilmesi,
2. Yeni veri tiplerinin madenciliği için yeni algoritma, teknik ve sistemlerin geliştirilmesi,
3. Dağıtık veri madenciliği için algoritma, protokol ve altyapıların geliştirilmesi,
4. Mevcut veri madenciliği sistemlerinin kullanımının ilerletilip geliştirilmesi,
5. Veri madenciliği için özel gizlilik ve güvenlik modellerinin geliştirilmesi.

- Tüm bu uğraşların başarıya ulaşması ve sonuç verebilmesi için hükümetin ve çok disiplinli ve disiplinler arası çalışan iş sahalarının desteği gereklidir.
- İlgili sistemlerin, ölçülmüş altyapıların ve test ortamlarının oluşturulmasını gerektiren önemli deneysel bileşenlerin gerçekleştirilmesi gerekir.

Veri madenciliğinin gelecek yıllar için üstlenmiş olduğu misyon hakkında fikir sahibi olmak için dünyanın önde araştırma ve danışmanlık firmalarından açıklanan rakamları dikkate aldığımızda, veri madenciliğinin gelecekte oldukça popüler bir konu olacağını görülmektedir. Örneğin, Gartner Group Araştırma Şirketi, gelecek on yıl içinde, hedef pazarlarda Veri madenciliği kullanımının yüzde 80'lere ulaşacağı tahmininde bulunmaktadır.

3.1.1.1 – Genel Bilgiler

Veri madenciliği, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı birçok teknik kullanılmaktadır (Eker, 2004).

Veri Madenciliği, organizasyonların karar aşamaları için yeni bilgiler üreten ya da gelecekle ilgili tahminler ve planlar yapmamızı sağlayan bir dizi teknikler ve anlayışlar bütünü olarak da tanımlanabilir (Karakaş, 2002).

Karar aşamalarında çok kritik bazı bilgiler vardır ki, sonuçların etkileri bu bilgilerin doğruluğuyla orantılıdır. Bir çok durumda cevabı tam olarak verilemeyen sorular doğrultusunda karar verilebilmektedir. Müşterilerin ilgi alanları, firmaya karşı olan bakış açıları, rakip firmalara olan ilgileri, markalarımıza olan bağlılıkları, gelir düzeyleri gibi bilgiler onlara sağladığımız mal ve hizmetlerin kalitesi üzerinde çok net etkiler yapacaktır. Bu tür bilgiler teorik olarak her ne kadar sistemlerde kayıt altında olsa da, kullanılabilir bir şekilde açık ve net cevaplara ulaşılması mevcut kullanımdaki sistemlerle neredeyse imkansız denecek kadar zordur. Çok büyük veri yığınları altında saklı olan bu bilgilere ulaşmak için uzun yıllar boyu yapıla gelen çalışmaların neticesinde bir dizi metodoloji geliştirilmiştir. Veri madenciliği uzun yıllardır özellikle Batı ülkelerinde üzerinde çalışılan bir konu olmasına rağmen, gerçek hayatta, yazılım endüstrisinin son yıllarda üretmiş olduğu ileri teknoloji ürünü yazılımlar ile kullanılmaya başlamıştır.

Gartner Grup tarafından yapılan tanımda ise veri madenciliği, istatistik ve matematik tekniklerle birlikte ilişki tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni ilişki ve eğilimlerin keşfedilmesi sürecidir (Eker, 2004).

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analiste, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir (Dunham, 2003).

Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir.

Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği insan merkezlidir ve bazen insan – bilgisayar arayüzü birleştirilir.

Veri madenciliği konusunda bahsi geçen geniş verideki geniş kelimesi, tek bir iş istasyonunun belleğine sığamayacak kadar büyük veri kümelerini ifade etmektedir. Yüksek hacimli veri ise, tek bir iş istasyonundaki ya da bir grup iş istasyonundaki disklere sığamayacak kadar fazla veri anlamındadır. Dağıtık veri ise, farklı coğrafi konumlarda bulunan verileri anlatır.

Veri madenciliği, bilgi keşfetmenin gerekli bir adımı gibi görülmesine rağmen. veri tabanında bilgiyi keşfetmekle eşanlamlı gibi kabul edilir. Bilgi keşfetme işlemi sırasıyla aşağıdaki işlemlerden oluşur (Han, 1999).

1. Veri Temizleme Gürültülü, hatalı, eksik ya da konuyla ilgisi olmayan verilerin ele alınmasıdır.
2. Veri Zenginleştirme: Farklı veri kaynaklarının birleştirilmesidir.
3. Veri Seçimi : Veri tabanında, yapılacak analiz için kullanılacak bağlantılı verilerin düzeltilmesidir.
4. Veri Kodlama: Verilerin özetlenerek madencilik için uygun olan forma dönüştürülmesi ya da birleştirilmesidir.
5. Veri Madenciliği : Veri Modellerinin elde edilmesi için uygun yöntemlerin uygulandığı önemli bir işlemdir.
6. Bilgi Sunma : İşlenmiş bilginin çeşitli tekniklerle kullanıcıya sunulmasıdır.

Sürekli artan verilere rağmen, çok az bilginin elde edildiği bir dünyada yaşanmaktadır. Günümüzde program geliştiriciler, büyük miktarda verilerden bilgi elde etmek için yeni teknikler geliştirmeye çalışmaktadırlar. Dünyadaki verilerin oranı her yıl yaklaşık olarak ikiye katlanmasına rağmen anlamlı bilgilerin miktarı hızla azalmaktadır. Bilgilerdeki bu azalmanın nedeni, anlamlı durumların bulunmasının giderek zorlaşmasıdır.

Bir çok uluslararası şirket çok kısa sürede büyük miktarda veri üretir. Internet gibi evrensel ağlarda, her gün milyonlarca veri dünyaya ayılır. Fakat bu hızla artan gelişmeyi izlemek mümkün değildir. Gelecekte profesyonel bir bilim adamı ve araştırmacı için bilginin mekanik üretimi ve veri yeniden üretimi önemli olacaktır. Bundan dolayı verilerin yorumlanması, seçimi, elenmesi için mekaniksel yöntemler ve stratejiler geliştirmek gerekecektir. Bu şekilde düşünüldüğünde bilgi tek başına bir üretim faktörü olarak kabul edilebilir (Han ve Kamber, 2001).

3.1.1.2 - Veri Madenciliğinin Tarihi

"Veri madenciliği" ve "KDD (Knowledge discovery in databases)" terimlerinin anlamı hakkında karışıklıklar vardır. Bir çok yerde eş anlamlı olarak kabul edilir. 1995 yılında Montreal'de yapılan I.Uluslararası KDD Konferansı'nda, KDD teriminin, veriden bilgi çıkarma işleminin tümünü tanımladığı ileri sürülmüştür. Bu çerçevede veri elemanları arasındaki ilişkiler ve örüntüler bilgi olarak adlandırılır. Bu konferansta, veri madenciliği teriminin KDD işleminin yalnızca aşamalarının keşfinde kullanılması gerektiği belirtilmiştir. KDD, "veriden, kullanışlı ve önceden bilinmeyen, tahmin edilemeyen önemli bilgiyi çıkarma" olarak tanımlanmıştır (Adriaans ve Zantinge 1996). Bu nedenle bilgi yeni olmalı, açık olmamalı ve kullanılabilir olmalıdır. KDD, yeni bir teknik değildir. Araştırmanın birden fazla disiplin kullanılarak yapılmasıdır. Makine öğrenimi, İstatistik, veri tabanı teknolojisi, uzman sistemler ve verilerin görüntülenmesi işlemlerinin hepsinin birlikte kullanıldığı bir tekniktir.

Uzmanlar son yıllardaki gelişmelere bakarak bilgisayarlarda kendi kendine öğrenme kapasiteli programlar uygulanabileceğini düşünebilirler. Yapay öğrenme olmadan yapay zeka ya da makine öğrenimi olmaz. Bu düşünceden yola çıkarak bilgisayar programları, bilgisayar teknolojisinin başladığı 1950'li yıllardan itibaren yapay zekadan yaralanarak geliştirilmiştir. İnsanın zeka yapısına yakın düzeyde bilgisayar yapmanın zor olduğu bilinen bir gerçektir. Günlük gazetelerdeki haberleri yorumlayabilen bir bilgisayar programı yaratmak şimdilerde mümkün olmayan belki başka bir yüzyıl olabilecek bir çalışmadır. 1960'ların sonlarında bilim adamları basit öğrenme kapasiteli bilgisayar programları geliştirebilmişler fakat önemli öğrenme sistemleri yaratamamışlardır. Örneğin Minsky ve Papert şimdi sinir ağları olarak bilinen

“perseptron” ların sadece çok basit kuralları öğrenebildiğini göstermişlerdir (Adriaans ve Zantinge, 1996). Aynı şeyler diğer araştırma programlarında da olmuştur. Bu çalışmalar makine öğreniminin üniversite ve araştırma merkezlerine girmesini sağlamıştır.

1970’li yılların sonunda araştırmacıların bulduğu diğer bir gelişme de makine öğrenimi için uzman sistemler olarak adlandırılan alandır. Jeolojik örneklerin sınıflandırılması ya da hastalıkların teşhisi gibi sınırlı alanlarda uzmanlar basit kurallarla uzman sistemler oluşturmuşlardır. Uzman sistemler binlerce kuraldan oluşur. Basit sistemler için bile çok fazla doğru kural bulmak gerekir. “Akıllı” uzman sistem yaratmak için ilgilenilen alanla ilgili (sigorta poliçeleri ya da vergi konuları gibi) ihtiyaç duyulan her şey kaydedilmelidir. Bu kayıtlar uzman sistemin temeli olarak stoklanır ve uzmanlar tarafından doğru kurallar bulunarak kullanılır. Bir uzman sistemin yapısı için doğru kuralları bulmak oldukça zor bir iştir. Bir uzman sistem için bilgi toplamak özen isteyen pahalı bir süreçtir. Çünkü bilgiler ilgili uzmanlarla görüşülerek toplanır. Bundan dolayı uzman sistem yapıcılar bilgi elde etme konusunda da zorluklarla karşı karşıya kalırlar.

Bütün bunlar göz önünde bulundurulduğunda, öğrenme sistemlerinin bazı avantajlara sahip olduğu ortaya çıkar. Bu avantajlardan biri, makine öğrenimi algoritmalarının bulunması gereken kuralları otomatik olarak genellemesidir. Bundan dolayı da zamandan tasarruf sağlanır. Diğeri ise uzman insanlarla görüşme yerine daha önce edinilen deneyimlerden öğrenme sistemleri kurulabilir. Uzman sistemlerin hatalarından öğrenebilme yoluyla uzman sistemler oluşturmak daha ilginç olabilir. Bütün bu faktörler makine öğreniminde yeni gelişmeler meydana getirmişlerdir.

1980’li yılların başında araştırmacılar makine öğrenimine çok farklı bir gözle bakmaya başlamışlardır. Makine öğrenimi, araştırmacıların yeni keşifler yapmasını sağlamıştır. Bunlar, objelerin karar ağaçlarıyla keyfi olarak sınıflandırılması için yapılan basit algoritmalar, Minsky ve Papert tarafından değerlendirilen “perseptron” lardan daha güçlü sinir ağları için oluşturulan yeni yapılar ve genetik algoritmalar gibi diğer yaklaşımları modellemedir. Aynı zamanda bilgisayar teknolojisinin de ilerlemesinden dolayı daha güçlü hale gelen bilgisayarlarda, yeni algoritmaları gerçek problemlerle

uygulama olanağı bulunmuştur. Basit gibi görünen aslında tam tersi karmaşık olan bir çok problemi çözümlemişlerdir. Üretim programlaması ve zaman tablosu planlaması gibi konuların bilgisayarla çözümlemesi oldukça zordur. Bu konuları tecrübeli planlamacılar daha kolay çözümlerler. Planlamacılar edindikleri tecrübelerle karmaşıklıkları nasıl gidereceklerini öğrenirler. Yapay zekada öğrenme kapasitesinin rolü büyük olduğundan öğrenme algoritmaları önem kazanmıştır (Adriaans ve Zantinge, 1996).

Makine öğrenimi çalışmalarında, felsefi problemlerden, uzman sistemlerden elde edilen muhtemel faydalardan ya da planlama ve programlama sistemlerinden çok daha ilginç bir gelişme olmuştur. Bu gelişme modern toplumda meydan gelen bir veri patlamasıdır. Bu veri patlaması verilerin mekanik üretimi ve verilerin mekanik tüketimi ihtiyacını ortaya çıkarmıştır. Bir çok şirket çok fazla bilgi içeren büyük veri tabanlarına sahiptir. Fakat bu verilerdeki büyük miktardaki artışlar bilgiye ulaşmayı zorlaştıran bir durum ortaya çıkarır. Bu durum, büyük ot yığınları içindeki iğneye benzetilebilir. Ancak otların miktarı her geçen gün daha fazla artar.

Bu bilgiler karşısında “veri madenciliği” yada “Veri tabanında bilgi keşfine (Knowledge discovery in databases, KDD)” ilgi gösterilmeye başlanmıştır. Gerçek bir madencilik işleminde elmas ya da altın gibi değerli bir maden bulmak için sürekli kazmak ve çıkan toprağı elemek gerekmektedir. Bilgisayarla bir benzetme yapıldığında veri tabanındaki milyonlarca veri içinden bilgi bulunur. Burada veriler toprak yığınlarına, bilgi ise elmas ya da altına benzetilebilir.

Veri madenciliğine ilginin artması aşağıdaki faktörlerle açıklanabilir; (Adriaans ve Zantinge, 1996).

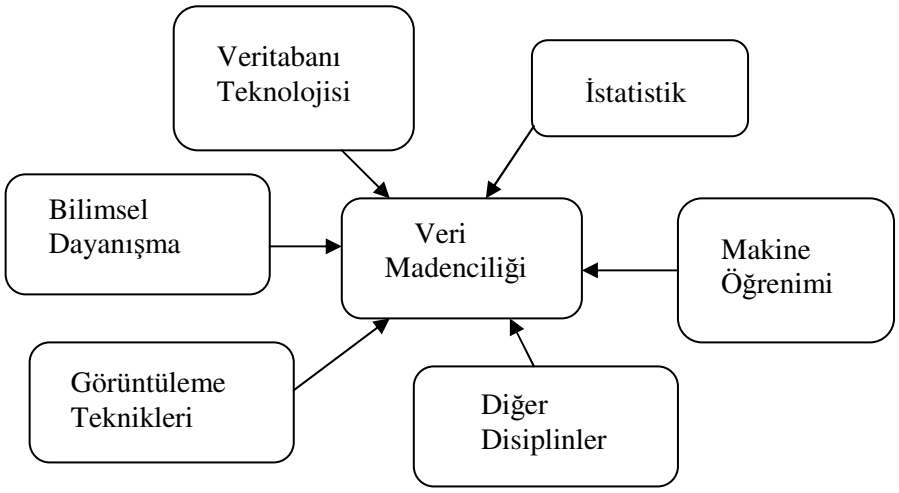
1. 1980’ lerde şirketler, müşterileri, rakipleri, ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanları potansiyel altın madeni gibidir. Sayısı milyonları geçen bu veriler, gizli bilgiler içerirler ve bunlara kolaylıkla SQL (structured query language) veri tabanı sorgulama dili ya da başka yüzeysel sorgulama dilleri kullanılarak ulaşılabilir. SQL sadece bir sorgulama dilidir ve önceden bilinen sınırlamalar altında bilgileri bulmaya yardım eder. Veri madenciliği algoritmaları tipik olarak, veri tabanının alt gruplarında ya da

"uygun" kümelerde belirginleşir. Bir çok durumda tekrarlanabilen SQL sorguları kullanılır ve ortalama sonuçlar elde edilir. Bunu elle yapmak mümkündür fakat oldukça yorucu ve uzun süren bir iştir.

2. Bilgisayarlarda ağ kullanımı gelişmeye devam etmektedir. Bu durumda veri tabanı ile bağlantı kurmak kolaylaşır. Böylece demografik verili dosya ile müşteri dosyası arasında bağlantı kurulabilir ve belirli popülasyon gruplarının kimliklerinin belirlenmesi sağlanabilir.
3. Son birkaç yılda makine öğrenimi teknikleri oldukça gelişmiştir. Sinir ağları, genetik algoritmalar" ve diğer basit uygulanabilir öğrenme teknikleri veri tabanlarıyla ilginç bağlantılar kurmayı kolaylaştırır.
4. Müşteri ile hizmet veren arasındaki ilişki, kişisel bilgileri hizmet verenin masasındaki bilgisayardan merkezi bilgi sistemlerine gönderir. Pazarlamacılar ve sigortacılar da bu yeni kazanılan teknikleri kullanmak isterler.

3.1.1.3 - Veri Madenciliği Yaklaşımı

Veri madenciliği, veri tabanı sistemleri, verilerin depolanması, istatistik, makine öğrenimi, bilgi düzeltme ve yüksek performanslı hesaplamalar gibi alanlardan çıkan genç bir disiplinler arası alandır (Şekil 3.1). Diğer yardımcı alanlar, sinir ağlarını, kimlik belirleme, uzaysal veri çözümlemesini, olasılık grafik teorisini ve tümevarımsal mantık programlamasını içerir. Veri madenciliği birden fazla disiplinden elde edilen yaklaşımların bütünleştirilmesine ihtiyaç duyar (Hand, 1999).



Şekil 3.1 : Veri Madenciliği ve Diğer Disiplinler

Verilen çözümleme metotları istatistikte yıllarca yapılan çalışmalarla geliştirilmiştir. Makine öğrenimi, sınıflama ve tümevarım problemlerinin çözümünde etkilidir. Sinir ağları ise veri madenciliğinde etkinliğini, sınıflama, tahmin ve kümeleme çözümlemesi konularında göstermiştir. Fakat veri madenciliği için bu yöntemler, veri tabanında gittikçe artan büyük miktarda veriler depolandığında, hesaplanabilirliği ve etkinliği açıklamaya yönelir. Etkin veri yapıları, indeksleme ve veri girişi teknikleri veri tabanı araştırmalarında yüksek performansI ı veri madenciliğine yardım etmek amacıyla geliştirilmiştir (Han, 1999).

3.1.1.4 - Veri Madenciliğinin Görevleri

Genel olarak veri madenciliği, betimsel ve kestirim veri madenciliği olmak üzere iki kategoride sınıflandırılır. Betimsel veri madenciliğinde, veri kümesi kısa ve öz bir şekilde tanımlanır ve verilerin ilginç genel özellikleri sunulur. Kestirim veri madenciliğinde ise bir ya da daha fazla modeller kümesi oluşturulup varolan verilerden sonuçlar çıkarılır ve yeni veri kümelerinin davranışları tahmin etmeye çalışılır (Han, 1999)

Veri madenciliğinin başlıca görevleri şunlardır (Chen, 2001),

1. Sınıf Betimlemesi : Veri kümesinin kısa ve öz bir özetini sağlar ve diğerlerinden farklılığını ortaya koyar. Bir veri kümesinin özetlenmesine sınıf nitelendirmesi denilir. İki yada daha fazla veri kümesinin karşılaştırmasına ise sınıf karşılaştırması yada ayrımı denilir. Sınıf betimlemesi sadece özet özellikleri değil merkezi eğilim ölçüleri ve dağılım ölçüleri gibi özellikleri de içerir. Örneğin Avrupa'ya karşı Asya şirketlerinin satışlarını karşılaştırmak, iki sınıfı ayırt edip önemli faktörleri belirlemek ve özet bir bilgi sunmak için sınıf betimlemesi kullanılır.

2. Birliktelik : Birliktelik, modeller kümesi arasındaki, birliktelik ilişkilerini yada korelasyonları keşfetmektir. Birliktelik analizi, pazarlama yönetimi, katalog planı ve diğer işlerde karar verme sürecinde geniş ölçüde kullanılır. Son zamanlarda kurulmuş etkin algoritmalı birliktelik analizinde önemli araştırmalar yapılmıştır. Bunlar, düzey-tarz, önsel araştırma, çoklu düzeyde madencilik, çok boyutlu birliktelikler, sayısal,

kategorik ve sürekli veri, meta-örüntü, yönelimli yada kısıt temelli madencilik ve madencilik korelasyonları için madencilik birliktelikleri içerir.

3. Sınıflama : Sınıflama çalışılan verilerin kümesini çözümler (sınıf ismi, etiketi bilinen objelerin kümesi) ve verilerin içinde özelliklerine göre sınıflar oluşturup her bir sınıf için bir model kurar. Veri tabanındaki her bir sınıfın daha iyi anlaşılması ve daha sonra elde edilen verilerin sınıflanması için sınıflama işlemi bir karar ağacı yada sınıflama kurallarının bir kümesini oluşturur. Örneğin, hastalıkların belirtilerine göre hastalıkları sınıflama, hastalık türlerini tahmin etmeye yardım eder. Makine öğrenimi, istatistik, veri tabanı, sinir ağları, pürüzlü kümeler ve diğer alanlarda geliştirilmiş bir çok sınıflama yöntemi vardır. Sınıflama, müşteri bölümü, iş modelleme ve kredi analizlerinde kullanılır.

4. Tahmin : Bu fonksiyon, bazı eksik verilerin muhtemel değerlerini yada objelerin bir kümesinde kesin özelliklerin değer dağılımını tahmin eder. Aynı zamanda seçilen objelere benzer verilerin kümesini temel alan değer dağılımını tahmin eder. Örneğin, bir işçinin maaşı, çalıştığı yerdeki diğer işçilerin maaş dağılımı temel alınarak tahmin edilebilir.

Genellikle,

- Regresyon analizi
- Genelleştirilmiş doğrusal model
- Korelasyon analizi
- Karar ağaçları
- Genetik algoritmalar
- Sinir ağı modelleri,

nitelikli bir tahmin için kullanılan yöntemlerdir.

5. Kümeleme : Kümeleme çözümlemesi, özellikleri birbirine benzeyen objelerin değerlerinin toplanmasından oluşan bir kümede, verilerin içindeki gizli kümeleri belirlemektir. Benzerlik, uzmanlar yada kullanıcılar tarafından belirlenmiş uzaklık fonksiyonlarıyla tanımlanabilir. İyi bir kümeleme yöntemi, kümeler arası benzerliğin

düşük ve küme içi benzerliğin yüksek olduğu nitelikli kümeler meydana getirir. Örneğin, evler, zemin alanı ve coğrafik yerleşime göre kümelenebilir. Veri madenciliği araştırması, büyük veritabanları ve çok boyutlu veri depoları için yüksek nitelikli ve hesaplanabilir kümeleme yöntemlerine odaklanır.

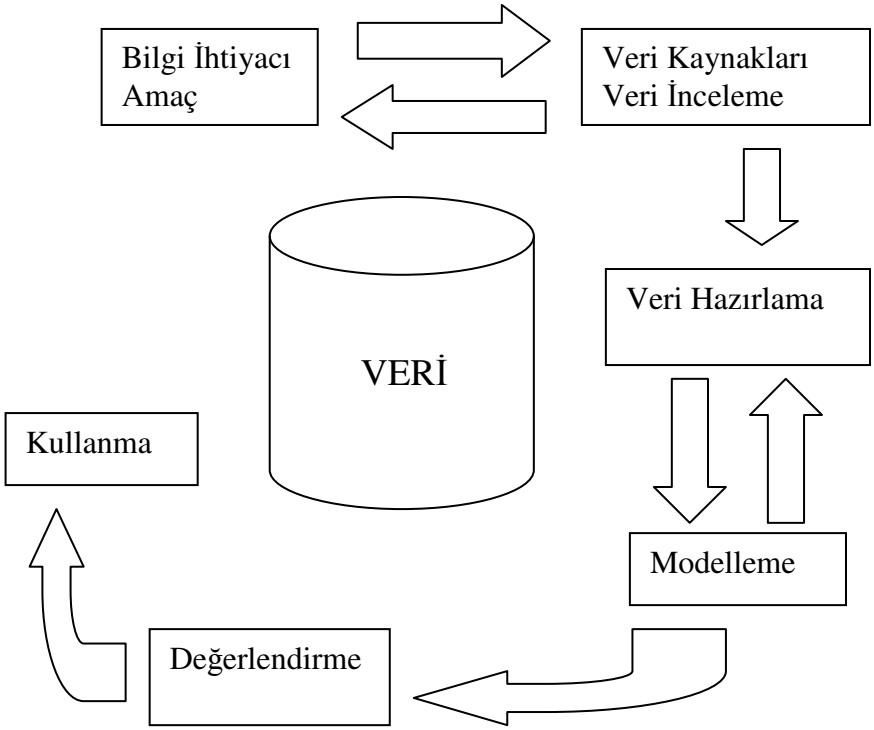
6. Zaman-Serileri analizi : Zaman-serileri çözümlemesi, benzer seriler ve ardışık örüntülerin madenlenmesi, periyodikler, eğilimler ve sapmaların araştırılması, ilginç özellikler ve kesin düzenlerin bulunması için zaman serilerinin geniş kümesini çözümler. Örneğin, bir şirketin geçmişteki stok durumuna, iş durumuna, rekabet performansına ve günlük piyasasına bakarak stok değerlerinin eğilimi tahmin edilebilir.

Aykırı değer çözümlemesi gibi veri madenciliği konuları olduğu gibi yeni veri madenciliği konularından da söz edilebilir (Chen 2001).

3.1.1.5 - Veri Madenciliği Süreci

Ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması mümkün değildir. Bu nedenle yukarıda tanımlanan tüm aşamalardan önce, iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır. Başarılı bir veri madenciliği projelerinde izlenmesi gereken yol Şekil 3.2'deki gibidir (Akpınar, 2000).

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin İzlenmesi



Şekil 3.2 : Veri Madenciliği Süreci

A. *Problemin Tanımlanması* ***:***

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

B. *Verilerin Hazırlanması :*

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının % 50 - % 85'ini harcamasına neden olmaktadır.

Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir. (Akpınar 2000)

1. Toplama **:** Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

2. Değer Biçme **:** Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

3. Birleştirme ve Temizleme **:** Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

Veri Temizliği bir kereye mahsus bir iş olarak görülmemeli ve yaşayan bir süreç olarak ele alınmalıdır. Kullanılan alternatif yaklaşımlar özünde önemli benzerlikler içermektedir: (Kutluay 2004)

- Veri Kalitesi Denetimi – Profilleme : Veri temizliğinde ilk adım olarak yapılması gereken, güncel durumun tespit edilmeye çalışılacağı denetim süreci olmalıdır. Bu aşamada, eldeki verinin durumu ve sorunları (eksiklik, standartlara uygunsuzluk, tutarsızlık gibi) gün ışığına çıkacak ve neyle uğraşıldığı detaylı olarak anlaşılacaktır. Örneğin, bir banka müşteri veri tabanında yapılacak denetim sonucunda müşteriler hakkında

olmazsa olmaz vergi numarası bilgisinin %60 oranında boş; telefon numaralarının da %40 oranında sadece 5 ve daha azı rakamdan oluşan yanlış bilgiden oluştuğu; birden fazla bulunan adres alanındaki bilgilerin de tutarsız doldurulduğu görülebilir. Bu ilk aşama, temizliğe başlamadan önceki durumu tespit etmek ve yol haritasını çizmekte bir temel teşkil etmek amacına hizmet edecektir.

- Hedefler, Kurallar ve İyileştirme Planları : Bu aşamada, profillenen sorunlar ile nasıl baş edileceğinin kuralları konacaktır. Hangi alanlarda ne tür doluluk ve doğruluk hedeflendiği, bunun nasıl sağlanacağı (verinin özelliğine göre dışa dönük -müşteriyi arayarak- veya pasif –müşteri aradığında- yöntemler ile) belirlenmelidir. Örneğin, yukarıda belirlenen vergi numarası, hem telefonu bulunan müşterileri arayarak, hem de müşteri bir işlem yapmak için bankaya başvurduğunda işleme bir ön gereksinim olarak toplanabilir.
- Temizlik (Standardizasyon) : Ancak sağlıklı geçilen ilk iki aşamadan sonra temizlik başlayabilir. Verinin doldurulması, belli standartlara oturtulması bu aşamada gerçekleşecektir.
- Raporlama ve Takip : Yapılan çalışmaların başarısını ölçmek ve yeni çalışmalar planlamak bu aşamanın görevidir. Organizasyonun istediği veri temizliği seviyesine ne kadar ulaşıldığının saptanması önem arz etmektedir. Ulaşılmak istenen seviye kuruma ve veri temizliğin amacına bağlı olarak belirlenir.

4. Seçim **:** Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır. Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır. Verilerin görselleştirilmesine olanak sağlayan grafik araçlar ve bunların sunduğu ilişkiler, bağımsız değişkenlerin seçilmesinde önemli yararlar sağlayabilir. Genellikle

yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin, önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir.

Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüfiliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, tesadüfi olarak örneklenmiş bir veri tabanı parçası üzerinde bir çok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır.

5. Dönüştürme **:** Kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır.

Veri madenciliğinde amaç büyük miktardaki ham veriden değerli bilginin çıkarılmasıdır. Çok miktarda güvenilir (hata ve eksiklerin olmadığı) veri ön şarttır çünkü çözümün, yani çıkarılan kuralların kalitesi öncelikle verinin kalitesine bağlıdır (Alpaydın2000).

C. *Modelin Kurulması ve Değerlendirilmesi :*

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir. Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak

sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır. (Doğruluk Oranı = 1 - Hata Oranı)

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

D. *Modelin Kullanılması :*

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

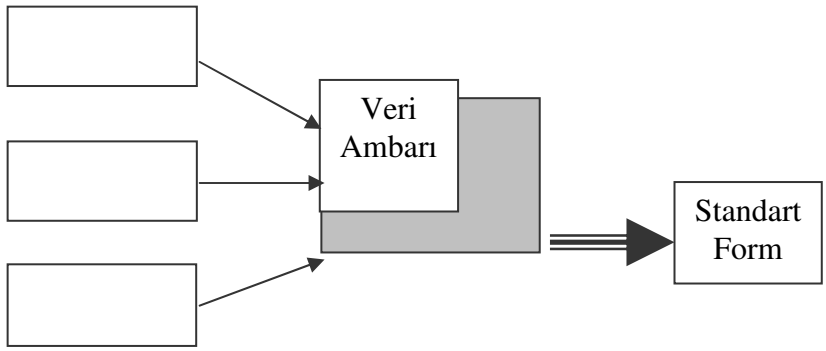
E. *Modelin İzlenmesi :*

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

3.1.1.6 - Veri Ambarları ve Veri Madenciliği

Veri madenciliği büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri

tabanlarının birleştirilmiş ve işlemeye daha uygun bir özetini saklamayı amaçlar (Şekil 3.3) (Alpaydın, 2000).



Şekil 3 3 : Veri Ambarları ve Veri madenciliği

Günlük veri tabanlarından istenen özet bilgi seçilerek ve gerekli ön işlemeden sonra veri ambarında saklanır. Ardından amaç doğrultusunda gerekli veri ambardan alınarak veri madenciliği çalışması için standart bir forma çevrilir. Veri ambarında veri oluşturulduktan sonra bu verinin elle veya gözle analizi yapılabilir. Bunun için OLAP (Online Analytical Processing) programları kullanılır. Bu programlar veriye her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak bakmayı ve incelemeyi sağlar. Böylece boyut bazında gruplama, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik veya rapor olarak sunma olanağı sağlar.

Veri madenciliğinde amaç, kullanıcının bilgi çıkarma sürecinde katkısının olabildiğince az tutulması, işin olabildiğince otomatik olarak yapılabilmesidir. Çünkü OLAP programlarını kullanırken bulunabilecek sonuçlar kullanıcının sormayı düşündüğü sorgularla sınırlıdır. Ama veri içinde kullanıcının hiç aklına gelmeyecek bilgiler de olabilir. Zaten veri madenciliğinde esas amaç bu tip bilgileri bulabilmektir.

Bir veri ambarının temel yapısını oluşturan bazı özel kurallar vardır. Bunlar ;

- *Zamana bağlı olmalıdır* : Zaman içinde toplanan bilgileri kapsamalı ve ambardaki bilgi ile ambara giren bilgi arasında bağ olmalıdır. Bu, bilgiyi belli bir periyoda göre arama açısından önemlidir.
- *Dağınık olmamalı* : Bir veri ambarındaki veriler eski tarihli olmamalı ve yalnız arama amaçlı olmalıdır. Çünkü yalnızca en son bilgiler güncelleştirilebilir, silinebilir ya da değiştirilebilir ve böyle veriler diğer veri tabanlarına da yüklenebilir. Örnek olarak kullanılmaya hazır veri tabanı verilebilir.
- *Konuya yönelimli olmalıdır* : Bu kullanılmaya hazır yerinin, var olan bütün uygulama alanlarına göre oluşturulması demektir. Kullanılmaya hazır veri tabanındaki bütün bilgiler veri ambarı için yararlı olmayabilir. Çünkü veri ambarı karar vermeyi desteklemek için özel olarak planlanırken, kullanılmaya hazır veri tabanı günlük kullanım için gerekli bilgiyi içerir.
- *Bütünlük* : Kuruluşlar için gerekli mesleki bilgileri yansıtmak demektir. Bir kullanılmaya hazır veri tabanında uygulamaya yönelik çeşitli bilgiler vardır ve bazı uygulamalarda aynı şeyin farklı isimleri için kullanılır. Fakat veri ambarındaki bu bilgileri bütünleştirmek ve tutarlı hale getirmek gerekir. Bir isim altında her şey tanımlanmalıdır (Adriaans ve Zantinge 1996).

Sonuç olarak, kullanıcının gerekli olan son bilgiye ulaşması için bilgiler uygun olarak düzenlenmelidir. Bir veri ambarı yalnızca karar destekleme için planlanmıştır ve yalnızca karar için gerekli olan veriler kullanılmaya hazır veri tabanından çıkarılır ve depolanır.

Bir veri ambarı planlamak için özel bilgiye ihtiyaç vardır. Çünkü kullanıcı ihtiyaç duyduğu veri modeline hızlı bir şekilde ulaşmak ister. Bundan dolayı kullanılmaya hazır veri tabanındaki veriler, veri ambarındaki verilerden farklı planlanmalıdır. Veri ambarı için uygun bir model oluşturulduktan sonra özel bir veri ulaşım çevresi planlanmalıdır. Kullanılmaya hazır verilen destekleyen çok sayıda veri tabanı varsa, bunların veri ambarı için kopyalanması gerekir. Bu sayede kontrol altında tutulabilir.

3.1.1.7 - Veri Madenciliğini Etkileyen Eğilimler

Temel olarak veri madenciliğini 5 ana harici eğilim etkiler (Vahaplar ve İnceoğlu, 2001):

a) Veri : Veri madenciliğinin bu kadar gelişmesindeki en önemli etkendir. Son yirmi yılda sayısal verinin hızla artması, veri madenciliğindeki gelişmeleri hızlandırmıştır. Bu kadar fazla veriye bilgisayar ağları üzerinden erişilmektedir. Diğer yanda bu verilerle uğraşan bilim adamları, mühendisler ve istatistikçilerin sayısı hala aynıdır. O yüzden, verileri analiz etme yöntemleri ve teknikleri geliştirilmektedir.
b) *Donanım* : Veri madenciliği, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir. Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmayı mümkün hale getirmiştir.
c) *Bilgisayar Ağları* : Yeni nesil internet, yaklaşık 155 Mbits/sn lik hatta belki de daha da üzerinde hızları kullanmamızı sağlayacak. Bu da günümüzde kullanılan bilgisayar ağlarındaki hızın 100 katından daha fazla bir sürat ve taşıma kapasitesi demektir. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır. Bundan 10 yıl önceki bilgisayar ağları teknolojisinde hayal edilemeyenler artık kullanılabilmektedir. Buna bağlı olarak, veri madenciliğine uygun ağların tasarımı da yapılmaktadır.
d) *Bilimsel Hesaplamalar* : Günümüz bilim adamları ve mühendisleri, simülasyonu teori ve deneyden sonra bilimin üçüncü yolu olarak görmektedirler. Veri madenciliği ve bilgi keşfi, bu 3 metodu birbirine bağlamada önemli rol almaktadır.
e) *Ticari Eğilimler* : Günümüzde ticaret ve işler çok karlı olmalı, daha hızlı ilerlemeli ve daha yüksek kalitede servis ve hizmet verme yönünde olmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği

sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

3.1.1.8 - Veri Madenciliğinin Kullanım Alanları

Günümüzde veri madenciliğinin başlıca ilgi alanları olarak aşağıdakiler sayılabilir; (Eker, 2004).

1. Pazarlama

- Müşteri segmentasyonu,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulması,
- Çeşitli pazarlama kampanyaları,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulması,
- Pazar sepeti analizi,
- Çapraz satış analizleri,
- Müşteri değerlemesi,
- Müşteri ilişkileri yönetimi,
- Çeşitli müşteri analizleri,
- Satış tahminleri,

2. Bankacılık

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Müşteri segmentasyonu,
- Kredi taleplerinin değerlendirilmesi,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,

3. Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,

- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri tipinin belirlenmesi.

4. Perakendecilik

- Satış noktası veri analizleri,
- Alış-veriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu,

5. Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin optimizasyonu,

6. Telekomünikasyon

- Kalite ve iyileştirme analizleri,
- Hisse tespitleri,
- Hatların yoğunluk tahminleri,

7. Sağlık ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis,
- Tedavi sürecinin belirlenmesi,

8. Endüstri

- Kalite kontrol analizleri,
- Lojistik,
- Üretim süreçlerinin optimizasyonu,

9. Bilim ve Mühendislik

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi.

Yeni Kullanımlar

Veri madenciliği disiplini, bugünkü teknoloji ile tam olarak desteklenemeyen yeni yeteneklere sahip uygulamalara ihtiyaç doğurmuştur. Bu uygulamalar, genel olarak 4 ana kategoride toplanmaktadır (Vahaplar ve İnceoğlu 2001).

a. İş ve Elektronik Ticaret Verileri : Arka ofis, ön ofis ve ağ uygulamaları iş süreçleri sırasında geniş çaplarda veri üretirler. Bu veriyi karar verme mekanizmalarında efektif olarak kullanmak, ilgili ticari kuruluşun temel yapı taşlarından olmalıdır.

b. Bilimsel, Mühendislik ve Sağlık Bakım Verileri : Günümüzde bilimsel veriler, iş sahası verilerinden daha da karmaşık hale gelmişlerdir.

c. Web Verileri : Internet ve web üzerindeki veriler hem hacim hem de karmaşıklık olarak hızla artmaktadır. Sadece düz metin ve resimden başka akan ve nümerik veriler de web verileri arasında yer almaktadır.

3.1.1.9 - Veri Madenciliğinin Maliyeti

Bir kuruluşta, KDD uygulanması, veri depolama, yeniden yönetme süreci ve veri madenciliği için fiyat çıkarımı gibi işlemleri içerir. Bu da kısmi olarak bu disiplinlerin fiyatlandırılmasına bağlıdır. Veri madenciliği için bir maliyetten söz edildiğinde, temelde, eldeki verilerden örüntü bulmak için makine öğrenimi teknikleri fiyatlandırılır. Bu fiyatlandırma, aynı konuda bir uzmanın performansının maliyeti ile karşılaştırılır. İnsan problem çözmede, genelde bilgisayardan daha iyidir. İnsanların çoğu yaratıcılık yeteneğine sahip olduğundan bazı problemlerin çözümünde yeni yaklaşımlar ortaya koyabilirler. Bilgisayarlar bunu yapma kapasitesine sahip değildirler. Fakat bazı alanlarda bilgisayarlar insanları yapabilirlik konusunda geçmiştir (Adriaan and Zantinge, 1996).

- Hız : Bilgisayar kısa sürede milyonlarca kaydı kolaylıkla okuyabildiğinden çok sayıdaki veriyi ele almada insandan daha donanımlıdır.

- Karmaşıklık : Bilgisayarlar özellikle hesaplama alanlarında insanlardan öndedir. Örneğin, makine öğrenimi konusunda insanların yapamayacağı çeşitli türdeki hesaplamaları yapar.
- Yenileme : Bilgisayarlar işlemleri tekrar etmekten yorulmaz. Örneğin, bilgisayarda makine öğrenme algoritmalarının yönetilmesi büyük bir avantajdır. Çünkü tekrar etme işlemi ne kadar istenirse o kadar yapılır. Yenilemenin bu unsuru, maliyet çıkarma içinde büyük kolaylıktır. Veri madenciliği çevresi oluşturularak, belirli bir zamanda satış sonuçlarını kestirmeye çalışmak oldukça zordur. Fakat bu çevre bir kez oluşturulduğunda istenilen sıklıkta ve sürede kullanılabilir.

Bu üç unsur, veri madenciliği çevresi oluşturmakta önemli ölçüde güç gerektirmesine rağmen, gerçekten faydalı olduğu zamanla anlaşılır. Bir kuruluş için müşteri davranış kestirim modeline sahip olunduğunda, sürekli kullanılabilir ve pazarlama hareketleri buna göre düzeltilebilir.

Makine öğrenme tekniklerinin maliyetinin çıkarılmasında başka bir yolda, uzman sistemlerdir. Uzman sistemler basit olarak uzmanların bilgilerini içeren sistemlerdir. Bratko, makine öğrenimi konusunda uzmandır ve bu sistemlerin kullanılmasıyla makine öğrenme algoritmalarının etkisini analiz etmiştir. Uzman sistemlerin planlanmasında temel zorluk, uzman bilginin elde edilmesidir. Uzman bilgi, uzmanlarla görüşerek elde edildiğinden uzun zaman almaktadır. Makine öğrenimi araçlı uzman sistemi ve makine öğrenimi araçsız uzman sisteminin planlanması maliyet açısından karşılaştırıldığında, makine öğrenimi tekniklerinin kullanımının, bütün işlemin fiyatlandırılmasında daha etkili sonuçlara sahip olduğu görülür. İlk uzman sistemlerden biri olan Mycin, 400 kural içerir ve bu kuralları oluşturmak yıllar alır. Yeni bir uzman sistem olan Gasoil ise yaklaşık 2800 kuraldan oluşur. Makine öğrenimi kullanıldığından beri, kurallar çok daha kısa sürede belirlenmiştir. Ayrıca makine öğrenimi, bilgiyi sistemli bir şekilde depoladığından, bilginin korunma maliyeti önemli ölçüde azalır. Bütün bunlar veri madenciliği için düşünüldüğünde, uzmanların makine öğrenimi algoritmaları yardımıyla, büyük veri kümelerini analizinin maliyeti oldukça düşük olacaktır. Öğrenme algoritmalarının, bilgiyi ortaya çıkarmada sağladığı tasarruf pazarlama alanında daha belirgindir (Adriaan and Zantinge, 1996).

3.1.1.10 - Veri Madenciliğinin Uygulanabilirliliği

Veri madenciliği pratikte Amerika ve Avrupa'da uygulanmaktadır. American Express ve AT8 gibi büyük kuruluşlar, müşteri dosyalarını analiz etmekte KDD'yi kullanmaktadırlar. UK ve BBC, sunulan programların analizinde veri madenciliği tekniklerini kullanmaktadır. Avrupa ülkelerinin çoğunda çok sayıda banka ve sigorta şirketleri KDD ile ilk tecrübelerini edinmişlerdir. KDD'nin çok problem içerdiği ortaya çıkmıştır. KDD'nin %80'i veri hazırlama olup geriye kalan %20'si ise madenciliktir. Elde bulunan verilerin, normal veri tabanında rutin olarak yapılan temizleme ve kodlamada kullanılması KDD işleminin önemli bir bölümüdür. Bu, model oluşturmadan daha önemlidir. Çünkü doğru veri olmadan doğru bilgiye ulaşılamaz.

Veri madenciliği çalışmalarının olabildiğince otomatik olması istenilse de uzmanların yardımı ve desteği olmadan başarılı olmak söz konusu değildir. Uzmanlar amacı tanımlar. Uygulama ile ilgili sonuca yararlı olabilecek her tür bilginin sisteme verilmesi gerekir ve bunları da ancak uzmanlar bilir. Ayrıca çalışma ile alınan sonuçların yorumlanması ve geçerlenmesi uzmanlar tarafından yapılır.Veri madenciliği tek aşamalı bir çalışma değildir, tekrarlıdır. Sistem ayarlanana dek birçok deneme gerektirebilir. Çalışma uzun olabilir. Buna çalışan ekibin ve yönetimin hazırlıklı olması, kısa vadede çok büyük beklentilere sahip olunmaması gerekir (Alpaydın2000). Veri madenciliğini uygulamak zor değildir. Küçük bir veri kümesinde, veri madenciliği yapılmak istenildiğinde, uzman tarafından sağlanan bir kontrol kümesiyle, kümeleme algoritmasının çalıştırılması gerekir. Bunun için gerekli standart araçlar hali hazırda piyasadan sağlanabilir. Diğer taraftan büyük bir veri tabanında veri madenciliği yapmak daha zordur. Problemlerle karşılaşılabilir ve bundan dolayı projeler başarısızlığa uğrar. Bu problemler,

1. İleri görüşün olmaması: Bir dosyanın, gelecekte nelerle karşılaşılacağı ve ne isteneceği düşünülmeden hazırlanması
2. Hazırlanan bütün dosyaların güncel olmaması: Verilerin eksik ya da yanlış olması
3. Kurumlar arasında işbirliği olmaması: Bazı kurumların verilerini vermek istememesi

4. Bilgisayar ortamında veri işlemleri yapan birimler arasında iletişim eksikliğinin olması
5. Resmi ve gizli sınırlamalar: Bazı verilerin gizli olmasından dolayı kullanılamaması
6. Dosyalarla teknik nedenlerden dolayı bağlantı kurulamaması: Veri modellerinin günümüze uygun olmaması yada hiyerarşik ve ilişkisel veri tabanı arasında farklılık olması
7. Zamanlama sorunu: Dosyaların merkezi olarak toplanması sırasında zaman kaybedilmesi
8. Yorumlama problemi: Veri tabanında keşfedilen ilişkiler yorumlanamadığı taktirde kullanılamaz.
9. Sınırlı bilgi: Veri tabanlarının genel olarak veri madenciliği dışındaki amaçlar için tasarlanması ve bu yüzden, öğrenme görevini kolaylaştıracak bazı özelliklerin bulunmaması (Vahaplar ve İnceoğlu, 2001).
10. Gürültü ve Eksik Değerler: Veri özellikleri yada sınıflarındaki hatalara gürültü adı verilir. Veri tabanlarındaki eksik bilgi ve bu yanlışlardan dolayı veri madenciliğinin amacına tam olarak ulaşmaması.
11. Güncellemeler ve konu dışı sahalar : Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar. Öğrenme sisteminin, kimi verilerin zamanla değişmesine ve keşif sisteminin verinin zamansızlığına karşın zaman duyarlı olmamasıdır.

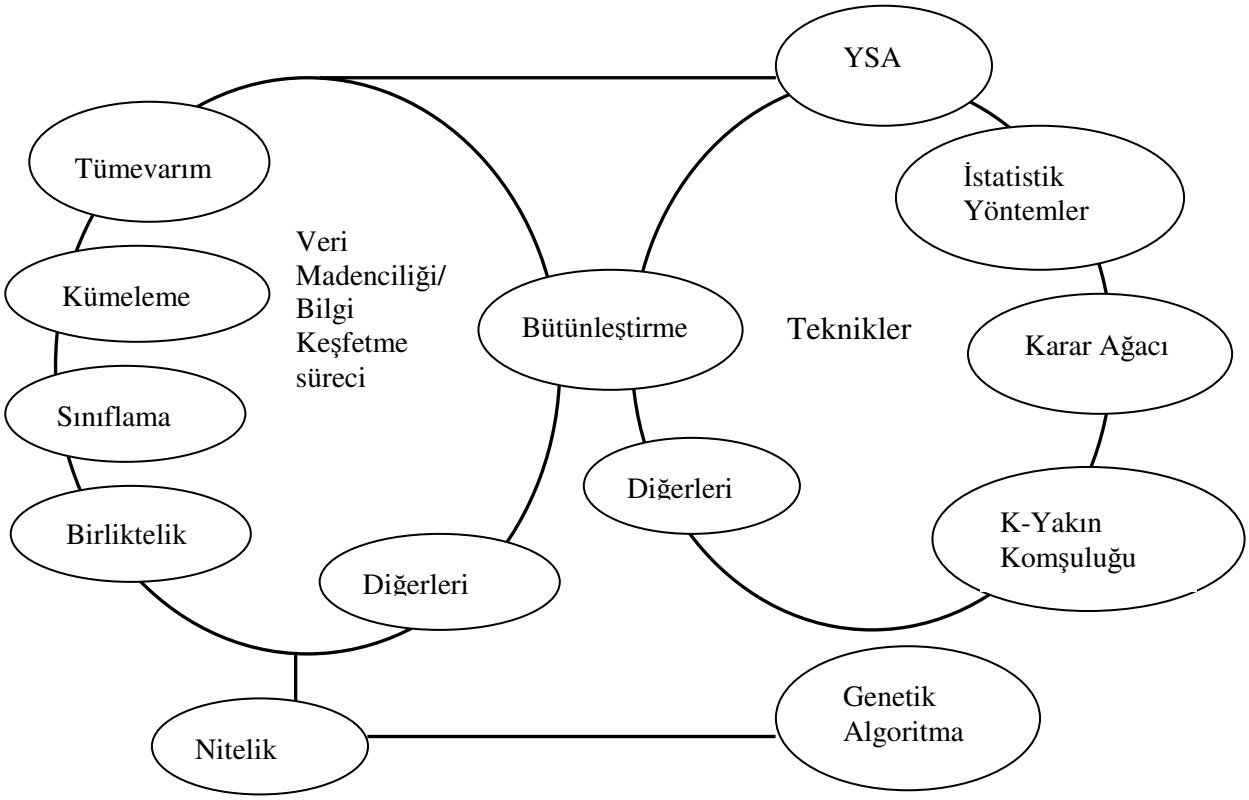
Yukarıda bahsedilen problemler sağlıklı ve esnek bir yapıya sahip olan organizasyonlarda çözümsüz değildir (Adriaans ve Zantinge, 1996).

3.1.1.11 - Veri Madenciliğinde Kullanılan Yöntemler

Veri madenciliği, veri kümelerinden gizli kalmış bilgilerin çıkarıldığı tek başına bir teknik değildir. Çok farklı tekniklerin kullanıldığı bir işlemdir. Şekil 3.4'de veri madenciliğinde kullanılan tekniklerden bazıları verilmiştir.

- Sorgulama Araçları
- Görüntüleme

- İstatistik Teknikler
- On Line Analitik İşlemler (OLAP)
- K-Yakın Komşuluğu
- Karar Ağaçları
- İlişki Kuralları
- Yapay Sinir ağları
- Genetik Algoritmalar



Şekil 3.4 : Veri madenciliği ve Kullanılan Teknikler

1. *Sorgulama Araçları ile Veri Kümesinin Ön Analizi:*

Veri madenciliğinde, ilk aşama geleneksel sorgulama araçlarının kullanılarak veri kümesinin yüzeysel bir analizini yapmaktır. Veri kümesine, basit yapılandırılmış sorgulama dilinin (SQL) uygulanmasıyla, sağlıklı bilgiler elde edilir. Fakat ileri örüntü analizi algoritmalarını uygulamadan önce veri kümesinin yapısının bilinmesi gerekir. SQL ile veri kümesinden kolaylıkla elde edilebilen yüzeysel bilgiler ortaya çıkarılır. Geriye kalan gizli kalmış bilgiler ise daha ileri teknikler uygulanarak elde edilir

(Moxon, 1996). Önemli pazarlama araştırmaları yapan büyük kuruluşlar için de bu gizli kalmış bilgiler çok önemlidir.

2. *Görüntüleme Teknikleri:*

Görüntüleme teknikleri, veri madenciliği işleminin başlangıcında, veri kümelerinde bulunan örüntülerin kalitesini belirlemek için kullanılır (Westphal ve Blaxton, 1998). Görüntüleme teknikleri, ileri grafik teknikleri ile kayıtları üç boyutlu ele alarak ilginç olasılıkları ortaya çıkarır. Böylece kullanıcıların, üç boyutlu yapılardaki karşılıklı etkileşimleri araştırmasını sağlar. Fakat bir çok kullanıcı bu ileri düzey teknikleri uygulamakta zorluk çekebilir. Bu yüzden basit grafiksel gösterim teknikleri de kullanılabilir ve sağlıklı bilgiler edinilebilir. Bu yöntemlerden biri olan, serpme diyagramı başta uygulanır. İki özellikteki bilginin bir kartezyen uzayında gösterilmesidir. Veri madenciliği işleminde, serpme diyagramları, veri kümelerinin ilginç alt kümelerini belirlemede kullanılır (Adriaans ve Zantigne, 1996).

3. *İstatistiksel Teknikler:*

Veri madenciliği çalışması esas olarak bir istatistik uygulamasıdır. İstatistik literatüründe son elli yılda bu amaç için değişik teknikler önerilmiştir. Bu teknikler istatistik literatüründe çokboyutlu analiz başlığı altında toplanır ve genelde verinin parametrik bir modelden (çoğunlukla çok boyutlu bir Gauss dağılımından) geldiğini varsayar. Bu varsayım altında sınıflandırma, regresyon, kümeleme, boyut azaltma, hipotez testi, varyans analizi, bağıntı kurma için teknikler istatistikte uzun yıllardır kullanılmaktadır (Alpaydın 2000).

Çok boyutlu bir veri uzayında kayıtlar nokta olarak kabul edilir. Böylece iki kayıt arasındaki uzaklık daha kolay belirlenir. Veri madenciliğin de, iki kayıt arasındaki uzaklık benzetme ile belirlenir. Birbirine yakın olan kayıtlar oldukça benzerdir, birbirinde uzak olanlar ise benzerliği az olan kayıtlardır (Westphal ve Blaxton, 1998). Bütün özellikler aynı büyüklük sırasında ölçülürse, farklı kayıtlar arasındaki uzaklık ölçümü güvenilir olur.

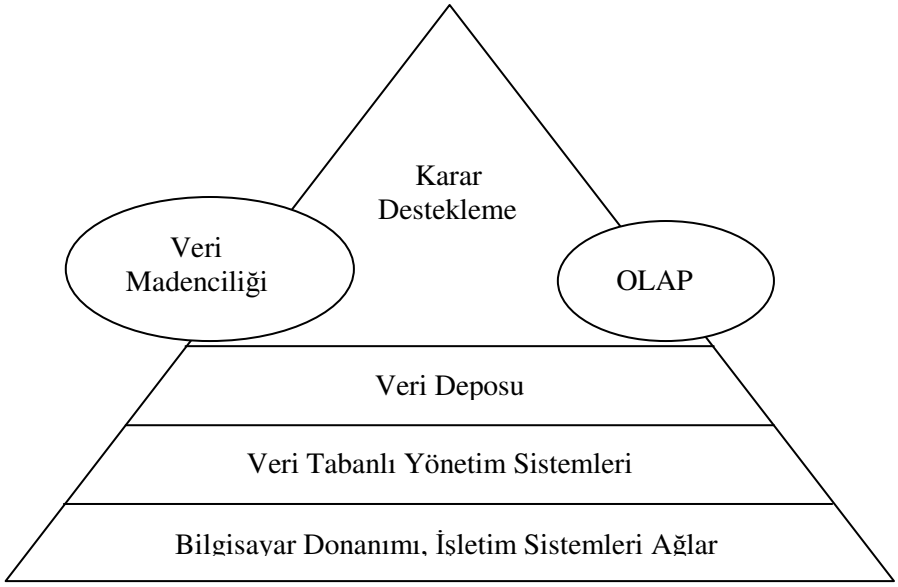
Az boyutlu veri uzayında veri kümelerini belirlemek kolaydır. Dolayısıyla bu kümelerin arasından ilginç olanlarını seçmek bir çok durumda mümkün olduğu gibi bazen de ileri araştırma programlarına gerek duyulabilir.

4. *On Line Analitik İşlemler (OLAP) :*

Bağımsız p tane özellik, p boyutlu uzay olarak görülebilir. Buna göre, veri kümesinde çok boyutlu veri kümesinde çok boyutlu analizler ilginç yapılar ortaya çıkarabilir. Kuruluşların cevaplarını aradıkları sorular çok boyutlu analizler gerektirir. İlişkisel veri tabanları, bu şekilde raporlara izin vermezler, fakat raporlama araçlarının yetenekleri ile, belirli bir noktaya kadar kabul edilebilir sonuçlar oluşmaktadır. Fakat daha karmaşık analizler işin içine girdiğinde, bir OLAP yapısı kurmadan bu raporları almak imkansız hale gelebilir.

Çok boyutlu ilişkiler, çok çözüm yolu gerektirir. Boyut sayısı azaldığında, çözüm yolu da sınırlanır. Fakat kuruluşların, verilerden elde etmek istedikleri bilgilerin sonu yoktur. Dolayısıyla geniş veri kümelerinin online kullanılması tercih edilir. OLAP araçları bu tür problemleri, çözmek için geliştirilmiştir. (Adriaans ve Zantigne, 1996).

OLAP, veri madenciliğinde önemli bir evre olmasına rağmen aralarında önemli bir fark vardır. OLAP araçları öğrenemez, yeni bilgiler yaratamaz, yeni çözümler için araştırma yapamaz. Bundan dolayı, bir veri tabanından, veri madenciliği yolu ile çıkarılan, çok boyutlu bilgi ve bu bilginin türü arasında önemli farklılıklar vardır. Veri madenciliği OLAP'tan daha güçlüdür. Diğer bir avantajı da, veri madenciliği algoritmaları çok özel depolamaya ihtiyaç duymaz. Bağıntılı bir veri tabanında depolanmış verilerle direkt olarak çalışılabilir (Chen 2001).



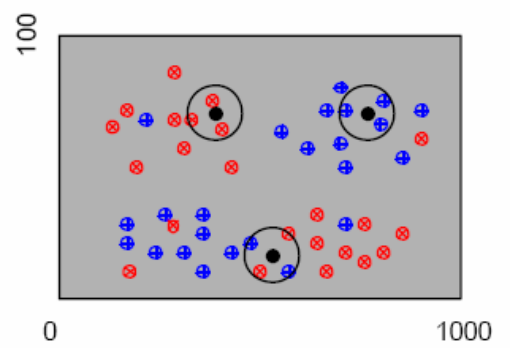
Şekil 3.5 : Veri madenciliği ve OLAP

5. *K-Yakın Komşuluğu :*

Kayıtlar, bir veri uzayındaki noktalar olarak düşünülürse, birbirine yakın olan kayıtlar, birbirinin civarında (yakın komşusu) olur.K-yakın komşuluğunda temel düşünce "komşunun yaptığı gibi yap"tır. Eğer belirli bir kişinin davranışı tahmin edilmek isteniyorsa, veri uzayında o kişiye yakın, örneğin on kişinin davranışlarına bakılır. Bu on kişinin davranışlarının ortalaması hesaplanır ve bu ortalama belirlenen kişi için tahmin olur. K-yakın komşuluğunda, K harfi araştırılan komşuların sayısıdır. 5-yakın komşuluğunda, 5 kişiye ve 1-yakın komşuluğunda 1 kişiye bakılır (Han ve Kamber, 2001).

K-yakın komşuluğu bir öğrenme tekniği değildir. Daha çok bir araştırma yöntemidir. K-yakın komşuluğu, veri kümesini daha iyi anlamaya yardımcı olur. K-yakın komşuluğu tekniği ile n tane kayıttan oluşan bir veri kümesinde, her bir kayıt için tahmin yapılmak istendiğinde, her kayıt, diğer kayıtlarla karşılaştırılmak zorundadır. Bu da büyük veri kümelerinde karesel karmaşıklığa yol açar. Eğer, bir milyon kayıtlı veri tabanında basit bir K-yakın komşuluğu incelemesi yapılacaksa, bir milyar karşılaştırma yapılması gerekir. Bu, araştırmada sorunlara neden olur. Genelde veri madenciliği algoritmaları n kayıt sayısı kadar karmaşıklığa sahip olmalıdır. Bu nedenle K-yakın

komşuluğu tekniği alt örneklemlerle ya da sınırlı sayıda veri kümesinde kullanılmalıdır. Şekil 3.6'da K-yakın komşuluğu yapısı genel anlamda gösterilmiştir.



Şekil 3.6 : K-Yakın Komşuluğu Yapısı

6. *Karar Ağaçları:*

Verileri sınıflandırmak ve bu verilerle kestrim yapmak, birbirleriyle çok yakın bağlantılı işlemlerdir. Bu karar ağaçlarıyla daha iyi görülebilir. Belirli bir müşterinin, belirli bir davranışı gösterip göstermeyeceği kestrilmek istenebilir. Bu durumda müşterinin, belirli bir müşteri sınıfında yer aldığı ve bundan dolayı da belirli bir davranışı göstereceği açıktır.

Tümevarım algoritmaları, büyük veri kümelerini çok iyi analiz eder ve tümevarım algoritmaları ile kurulan karar ağacı ile karar süreci kolaylaştırılır. Yapılan kestrimler de daha iyi sonuçlar verir (Komorowski ve Zytkow, 1997).

İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veriden oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (IF-THEN rules) yazılabilir. Bu şekilde kural çıkarma, veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda analiste bilgi verir ve daha sonraki analizler için yol gösterici olabilir.

Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğinde, kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahiptir (Akpınar, 2000).

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalar,

- Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktardaki değişken ve veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması**,**
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikliye dönüştürülmesidir.

Karar ağacı temelli tipik uygulamalar ise,

- Hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi,
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi*,*
- Geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi,
- Üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesidir.

Gerçek dünyanın sosyal ve ekonomik olaylarını daha güvenilir bir şekilde gösterebilmek için standart istatistik tekniklerin dışında yeni analiz tekniklerinin geliştirilmesi ile ilgilenen Morgan ve Sonquist tarafından University of Michigan'da

1970'li yılların başlarında kullanıma alınan Automatic Interaction Detector – AID karar ağacı temelli ilk algoritma ve yazılımdır. AID tekniği en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasındaki mümkün bütün ilişkilerin incelenmesine dayanmaktadır. Karar ağacı tekniğinin sağladığı kuruluş ve yorumlama kolaylıkları, AID yazılımının başlangıçta istatistikçi ve veri analistleri tarafından büyük coşku ile karşılanmasına neden olmuştur (Akpınar 2000).

Ancak AID'in bağımlı ve bağımsız değişkenler arasındaki ilişkilerin tanımlanmasında aşırı saldırgan davrandığı ve bunun sonucunda anlamlı ve anlamsız ilişkileri ayırt edemediği yönünde Einhorn başta olmak üzere bir çok araştırmacı tarafından yayınlar yapılmıştır.

İlk temelleri AID yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür. Geliştirilen bu algoritmalar içerisinde CHAID (Chi-Squared Automatic Interaction Detector; G.V. Kass; 1980), C&RT (Classification and Regression Trees; Breiman, Friedman, Olshen ve Stone; 1984), ID3 (Quinlan; 1986), Exhaustive CHAID (Biggs, de Ville ve Suen; 1991), C4.5 (Quinlan; 1993), MARS (Multivariate Adaptive Regression Splines; Friedman), QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997), C5.0 (Quinlan), SLIQ (Supervised Learning in Quest; Mehta, Agarwal ve Rissanen), SPRINT (Scalable Parallelizable Induction of Decision Trees; Shafer, Agrawal ve Mehta) başlıcalarıdır

7. *Birliktelik Kuralları:*

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi adı altında veri madenciliğinde yaygın olarak kullanılmaktadır (Akpınar 2000).

Pazarlama araştırmalarındaki hedef, müşteri profillerini açık olarak belirlemektir. Örneğin, spor bir arabaya sahip olan ve evinde köpek besleyen kadınların %90'ı Chanel marka parfüm kullanırlar. Veri madenciliği teknikleriyle bu tür ilişkiler

bulunabilir. Bu ilişkiler veri madenciliğinde birliktelik kuralı olarak adlandırılır (Westphal ve Blaxton, 1998). Bu örnek için müşterilerin cinsiyeti, sahip oldukları araba türü ve hayvan besleyip beslemedikleri ve satın almaktan hoşlandıkları ürünlerden oluşan bir veri tabanında oluşturulacak kural: cinsiyeti bayan, arabası spor ve hayvan besleyen kayıtların %90'ı Chanel marka parfüm kullanabilir.

Bir birliktelik algoritması oluşturmadan önce kurallar belirlenmelidir. Büyük veri tabanında ilişkileri bulacak algoritmalar geliştirmek çok zor değildir. Fakat, geliştirilen algoritmalar önemli ilişkileri ortaya çıkaracağı gibi önemsiz bir çok ilişkiyi de ortaya çıkarır. Bu yüzden, büyük veri tabanlarında küçük alt kümeler bulunmalıdır. Ayrıca, büyük veri tabanlarında çok sayıda ilişki bulunabileceğinden, birliktelik kuralları sayısı da sınırsız olabilir. Dolayısıyla ilginç ilişkilerle önemsiz ilişkilerin ayrılması gerekir (Zhong ve Zhon 1999).

8. *Yapay Sinir Ağları:*

1980'lerden sonra yaygınlaşan yapay sinir ağlarında amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir, ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez.

Bu çalışmada yapay sinir yağları kullanılması nedeniyle, bölüm 3.1.3'de daha ayrıntılı bir şekilde konu ile ilgili bilgi verilmektedir.

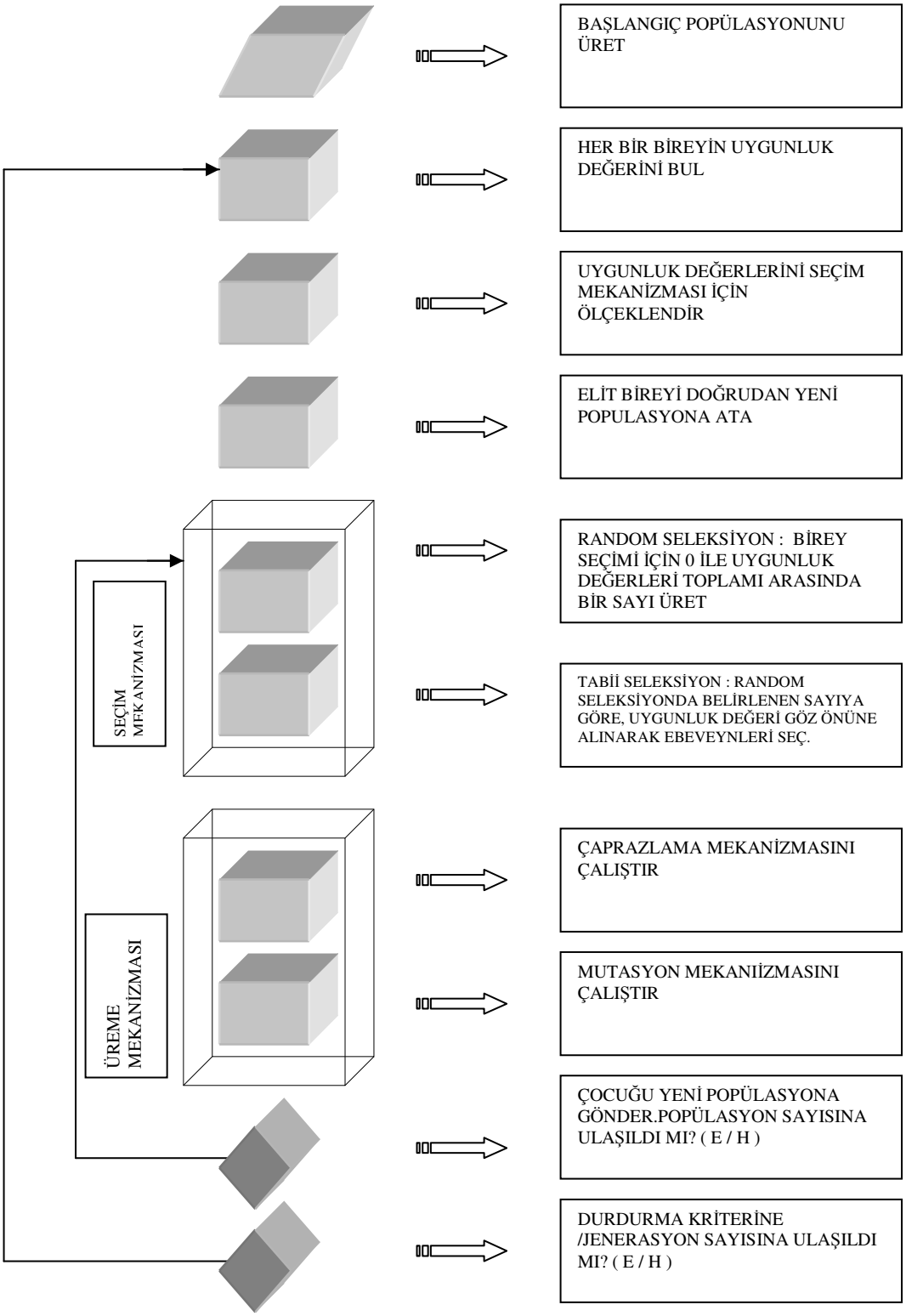
9. *Genetik Algoritma:*

Genetik algoritmalar yapay zekanın gittikçe genişleyen bir kolu olan evrimsel hesaplama tekniğinin bir parçasını oluşturmaktadır. Adından da anlaşıldığı üzere, evrimsel hesaplama tekniğinin bir parçası olan genetik algoritma Darwin'in evrim teorisinden esinlenerek oluşturulmuştur. Herhangi bir problemin genetik algoritma ile çözümü, problemi sanal olarak evrimden geçirmek suretiyle yapılmaktadır. (Semetay ve Kurt, 2001).

Genetik algoritmalar; doğal seleksiyon prensibinden yola çıkılarak geliştirilmiş arama algoritmalarıdır. Algoritma, belirli bir uzunluğa sahip dizilerden oluşmuş bir veri yığınına sahiptir. Yığın içindeki her bir dizi, çözüm uzayında bir noktayı temsil eder. Bu diziler aynı zamanda üreme yolu ile varlığını sürdürmeye aday olan birer bireydir. Algoritmanın temel işleyişi , çözüme uygun olmayan bireyleri elemek, çözüme daha uygun bireyleri seçmek ve seçilen bireylerden yeni bireyler üretmek doğrultusundadır. Algoritmanın işleyişi aşamalı olarak düşünüldüğünde temel prensibinin eleme olduğu anlaşılmaktadır. İkinci prensip ise, elemeyi aşan bireylerden yararlanılarak olası yeni çözümler elde etmektir. Bu da bireyler arasındaki bilgi alışverişi ile sağlanır. Bireyler arası rassal bilgi alışverişi, arama işleminin, çözüm uzayının daha uygun noktalarında devam etmesini sağlar. (Oğuz ve Akbaş, 1999).

Genetik algoritma(GA) stokastik bir arama yöntemidir. Biyolojik sistemlerin gelişim sürecini modelleyen GA, ilk olarak Holland (1975) tarafından önerilmiştir. Sezgisel bir yöntem olan GA, problem için en optimum sonucu bulamayabilir, ancak bilinen metotlarla çözülemeyen veya çözüm zamanı çok büyük olan problemlerde optimuma çok yakın çözümler vermektedir (Goldberg, 1989).

Doğal sistemler oldukça sağlam kararlı yapıdadırlar. Uyum yeteneğinin nasıl gelişip işlediğini öğrenmenin en iyi yolu, biyolojik sistemler üzerindeki çalışmalardır. GA'ların karmaşık arama uzaylarına uygulandığında kararlı çözümlere ulaştıkları, kuramsal ve deneysel çalışmalar ile kanıtlanmıştır. Etkili ve verimli bir yöntem olduğu kanıtlandıktan sonra çeşitli bilim dallarının yanı sıra iş dünyası ve mühendislikte geniş bir uygulama alanı bulmasının en önemli nedeni, kuşkusuz hesaplamalarda sağladığı basitlik ve arama işlemlerindeki iyileştirme gücüdür. Standart genetik algoritma akış diyagramı Şekil 3.7'deki gibidir.



Şekil 3.7 : Standart Genetik Algoritma Akış Diyagramı

3.1.2 – Web Madenciliği

Veri madenciliği ve Web son zamanların geçerli iki araştırma sahasıdır. Bu iki sahanın doğal kombinasyonu Web madenciliği olarak adlandırılır. Veri madenciliği uygulamalarından biri olan Web madenciliği, Web verileri üzerinde veri madenciliği fonksiyonlarını yerine getirir (Özakar ve Püskülcü, 2002).

Birçok yazara göre web madenciliği terimi ilk kez Etzioni tarfından 1996'da ortaya atılmıştır. Bu bildiride Etzioni Web madenciliğinin veri madenciliği tekniklerini kullanarak World Wide Web'de bulunan dosya ve servislerden otomatik olarak paternler bulmak ve öngörülmeyen bilgiye ulaşmak olduğunu iddia etmektedir (Etzioni, 1996). Araştırmacıların çoğu çalışmalarında bu tanımlamayı esas almışlardır

Burada bu işlemlerden bazılarının rahatlıkla arama motorları tarafından yapılabileceği akla gelebilir. Bu durumda Web Madenciliğine ihtiyaç duyulmasının iki sebebi vardır. Bunlar:

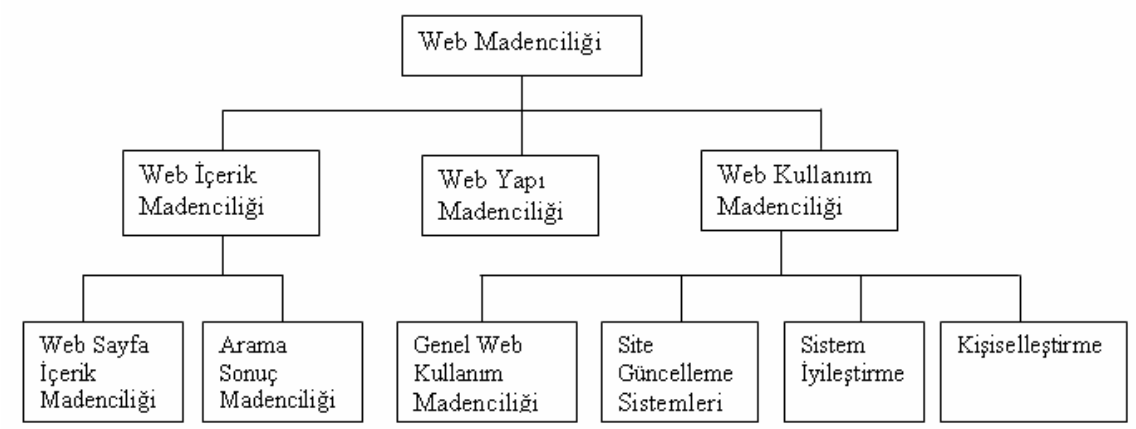
1. Google, Yahoo gibi arama motorlarını kullanıldığında genelde iki çeşit sorunla karşılaşılır: "Veri madenciliği" ile ilgili dokümanlar araştırılırken sonuç olarak çok fazla doküman listelenebilir ama bunların birçoğu araştırılan konuyla yeteri kadar ilgili değildir. Ayrıca dokümanlar sıralanırken araştırılan konuyla en çok ilgili olandan en az ilgili olana doğru sıralanmış değildir. Ancak dokümanlar incelendikten sonra istenilen sırada konuyla ilgili siteler bulunabilmektedir.

2. Arama motorlarında yine "veri madenciliği" konusunun araştırıldığı varsayılırsa, bu konu ile yakından ilgili olan *makine öğrenmesi , bilgi keşfi ile ilgili dokümanlar* içerisinde "veri madenciliği " kelimeleri geçmediği için sonuç olarak listelenmeyecektir. Bu sebeple son zamanlarda araştırmacılar veri madenciliği kavramını Web'e uyarlamışlardır (Şakiroğlu ve ark., 2003).

Web madenciliği kabaca Web'ten faydalı bilginin keşfi olarak da tanımlanabilir. Bu tanım içinde otomatik tarama, bilgi alma ve kullanılabilir kaynakların milyonlarca web sitesi veya online veritabanlarından seçilmesi web içerik madenciliği konusuna girerken bir veya birçok web sunucu veya online servisten kullanıcı erişim desenlerinin

analiz ve keşfi Web kullanım madenciliği konusuna girmektedir (Takcı ve Soğukpınar, 2002).

Web üç tip veri bulundurur; içerik, Web log dosyaları ve Web yapı verisi. Şekil 3.8 'de madencilik yapılabilecek verinin sınıflandırması incelenmektedir. Bunlar Web içerik madenciliği, Web yapı madenciliği ve Web kullanım madenciliğidir.

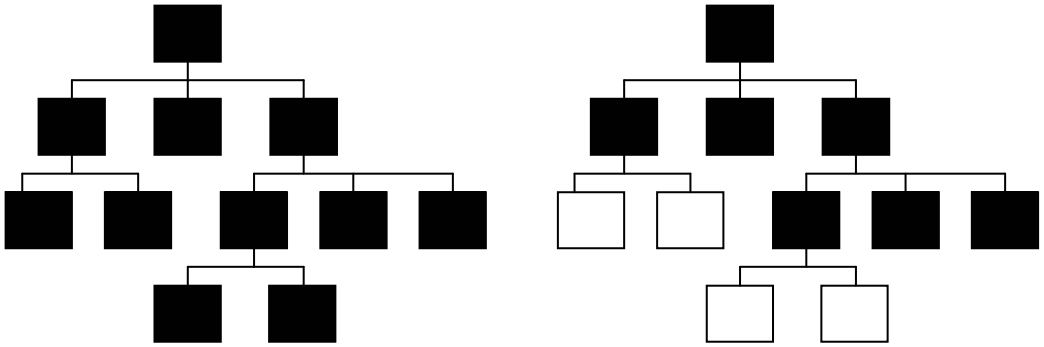


Şekil 3.8 : Web Madenciliği Sınıflandırması

1. *Web İçerik Madenciliği:* Web içerik madenciliği temel olarak İnternet de saklı bilgiyi bulma üzerine yoğunlaşmıştır (arama motorları, vs.). Kısaca konusu, site içeriğidir. Adından da anlaşılacağı gibi web dokümanlarının içeriklerini yorumlamak ile ilgilenir. Web içerik madenciliği akıllı yazılım ajanları (web robotları, web örümcekler vs.) daha doğrusu makine öğrenimi veya yapay zeka ile ilgilidir. Son zamanlarda dokümanlardan bilgi çıkarma için XML de kullanılmaya başlanmıştır. Burada; saniyede binlerce web sayfasını inceleyen geniş ölçekli programlara "derleyici" (Crawler) denilmektedir (Belen ve arkadaşları, 2003). Web içerik verilerinin çoğu belli bir düzene sahip olmayan düz metinlerdir. Lycos, Alta Vista, Web Crawler gibi bilinen çeşitli arama motorları bu tekniklerden faydalanırlar.

Web içerik madenciliği, arama motorlarındaki yapının genişletilmiş hali olarak düşünülebilir. İnternetde arama yapılırken bir çok teknik kullanılmaktadır.Bu tekniklerden, klasik arama motorlarında en çok kullanılan kelime tabanlı arama yaklaşımıdır. Bunun dışında, içerik hiyerarşisi, kullanıcı davranışları ve sayfalar arası link ilişkileri de kullanılan en temel yaklaşımlardandır.

Derleyicide, çekirdek URL adres setine bakarak değerlendirme başlamakta ve çekirdek URL adreslerindeki linkler kaydedilip arama bu linklerden devam etmektedir. Web'deki muazzam büyük yapı, özelleşmiş derleyici yapılarının geliştirilmesine neden olmuştur. Şekil 3.9'da genel derleyici ve özelleşmiş derleyicilerdeki arama mantığı görülmektedir. Şekildeki siyah gölgeli kısımlar derleyicinin değerlendirmeye aldığı sayfaları temsil etmektedir. Buna göre özelleşmiş derleyici bir sayfayı ilgili bulduysa sayfanın linklerini değerlendirmeye almakta, aksi halde diğer sayfaları değerlendirmeye geçmekte bir alt seviyeye inmemektedir (Dunham, 2003).



Şekil 3.9 : a) Genel Derleyici b) Özelleşmiş Derleyici

2. *Web Yapı Madenciliği:* Web yapı madenciliği sitenin yapısal dizaynını iyileştirmek için kullanılır. Web sayfaları arasındaki bağlantılarını (hyperlink) ilişkilerini keşfetmekle ilgilenir. Yani HTML kodlarındaki <a href> etiketleri arasında yer alan veriyi yorumlar. Web içerik madenciliği web sayfasının içeriği ile ilgilenirken, web yapı madenciliği doğrudan web sayfaları arasındaki bağlantılar ile ilgilenir (Şakiroğlu ve arkadaşları, 2003).

3. *Web Kullanım Madenciliği:* Web kullanım madenciliği; bir veya birçok web sunucudan kullanıcı erişim desenlerinin otomatik keşfinin ve analizin yapıldığı bir tip veri madenciliği etkinliğidir. Birçok organizasyon pazar analizleri için geliştirdikleri stratejileri ziyaretçi bilgilerine dayanarak yerine getirir. Organizasyonlar günlük operasyonlarla her gün yüzlerce MB veri toplamaktadır. Bu bilgilerin çoğu web sunucuların otomatik olarak tuttuğu günlük dosyalarından elde

edilir. Günlük dosyalarında, istemcinden sunucuya gönderilen her bir istek bir kayıt olarak tutulur (Takcı ve Soğukpınar, 2002).

Web verilerinin analizi sonucunda bir ziyaretçinin sitede kalma süresi, hizmet stratejileri, etkin kampanyalar ve diğerleri bulunabilir. Ayrıca siteye bağlanan bir kullanıcının hangi amaçla siteye bağlandığı, kötü niyetli bir kullanıcı olup olmadığı da bulunabilmektedir. Bir elektronik ticaret sitesi için en iyi müşteri veri madenciliği sayesinde bulunabildiği gibi bir "hacker" da aynı yöntemlerle bulunabilir.

Web kullanım madenciliği başlıca üç fazdan oluşmaktadır: (Belen ve arkadaşları, 2003)

1. *Ön Işleme :* Ön işleme veri kaynağından alınan verinin desen bulmaya hazır hale getirilmesi adımıdır. Belki de web kullanım madenciliğinin en önemli aşamasıdır. Çünkü etkili bir şekilde yapıldığından zaman ve kaynak tasarrufu sağlayacaktır. Bu adımda esas olarak veri gürültüden temizlenir.

2. *Desen Bulma:* Veri madenciliğinde desen bulmak için kullanılan bir çok yöntem ve algoritma vardır ve bunların çoğu web kullanım madenciliğinde de kullanılmaktadır.

3. *Desen Analizi:* Desen analizi web kullanım madenciliğinin son adımıdır. Desen analizinin amacı bulunan desenlerden ilginç olmayan desenleri elemektir. Desen analizinin en çok karşılaşılan şekli SQL gibi bilgi sorgulama dilleri ile yapılan uygulamalardır. Bir başka yöntem ise verilerin veri küplerine yüklenerek OLAP işlemlerinin yapılmasıdır.

Web içerik madenciliği dokümanların içinden bilgi çıkarırken web kullanım madenciliği kullanıcıların erişimlerinden bilgi çıkarmaktadır. Erişimlere dayalı bilgilerle kullanıcı davranışları bulunabilmekte ve kişiye özel hizmet olanağı sağlanabilmektedir.

3.1.3 – Yapay Sinir Ağları

Beynin üstün özellikleri, bilim adamlarını üzerinde çalışmaya zorlamış ve beynin nörofiziksel yapısından esinlenerek matematiksel modeli çıkarılmaya çalışılmıştır. Beynin bütün davranışlarını tam olarak modelleyebilmek için fiziksel bileşenlerinin doğru olarak modellenmesi gerektiği düşüncesi ile çeşitli yapay hücre ve ağ modelleri geliştirilmiştir. Böylece Yapay Sinir Ağları denen yeni ve günümüz bilgisayarlarının algoritmik hesaplama yönteminden farklı bir bilim alanı ortaya çıkmıştır. Yapay sinir ağları; yapısı, bilgi işleme yöntemindeki farklılık ve uygulama alanları nedeniyle çeşitli bilim dallarının da kapsam alanına girmektedir (Elmas, 2003).

Genel anlamda YSA, beynin bir işlevi yerine getirme yöntemini modellemek için tasarlanan bir sistem olarak tanımlanabilir. YSA, yapay sinir hücrelerinin birbirleri ile çeşitli şekillerde bağlanmasından oluşur ve genellikle katmanlar şeklinde düzenlenir. Donanım olarak elektronik devrelerle ya da bilgisayarlarda yazılım olarak gerçeklenebilir. Beynin bilgi işleme yöntemine uygun olarak YSA, bir öğrenme sürecinden sonra bilgiyi toplama, hücreler arasındaki bağlantı ağırlıkları ile bu bilgiyi saklama ve genelleme yeteneğine sahip paralel dağılmış bir işlemcidir. Öğrenme süreci, arzu edilen amaca ulaşmak için YSA ağırlıklarının yenilenmesini sağlayan öğrenme algoritmalarını içerir (Haykin, 1994).

3.1.3.1 - Yapay Sinir Ağlarının Özellikleri

YSA'nın hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir. Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA' nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA' nın karmaşık problemleri çözebilme yeteneğini gösterir. Günümüzde birçok bilim alanında YSA, aşağıdaki özellikleri nedeniyle etkin olmuş ve uygulama yeri bulmuştur (Elmas, 2003).

1. *Doğrusal Olmama* : YSA' nın temel işlem elemanı olan hücre doğrusal değildir. Dolayısıyla hücrelerin birleşmesinden meydana gelen YSA da doğrusal değildir ve bu özellik bütün ağa yayılmış durumdadır. Bu özelliği ile YSA, doğrusal olmayan karmaşık problemlerin çözümünde en önemli araç olmuştur.

2. *Öğrenme* : YSA' nın arzu edilen davranışı gösterebilmesi için amaca uygun olarak ayarlanması gerekir. Bu, hücreler arasında doğru bağlantıların yapılması ve bağlantıların uygun ağırlıklara sahip olması gerektiğini ifade eder. YSA' nın karmaşık yapısı nedeniyle bağlantılar ve ağırlıklar önceden ayarlı olarak verilemez yada tasarlanamaz. Bu nedenle YSA, istenen davranışı gösterecek şekilde ilgilendiği problemden aldığı eğitim örneklerini kullanarak problemi öğrenmelidir.

3. *Genelleme* : YSA, ilgilendiği problemi öğrendikten sonra eğitim sırasında karşılaşmadığı test örnekleri için de arzu edilen tepkiyi üretebilir. Örneğin, karakter tanıma amacıyla eğitilmiş bir YSA, bozuk karakter girişlerinde de doğru karakterleri verebilir yada bir sistemin eğitilmiş YSA modeli, eğitim sürecinde verilmeyen giriş sinyalleri için de sistemle aynı davranışı gösterebilir.

4. *Uyarlanabilirlik* : YSA, ilgilendiği problemdeki değişikliklere göre ağırlıklarını ayarlar. Yani, belirli bir problemi çözmek amacıyla eğitilen YSA, problemdeki değişimlere göre tekrar eğitilebilir, değişimler devamlı ise gerçek zamanda da eğitime devam edilebilir. Bu özelliği ile YSA, uyarlamalı örnek tanıma, sinyal işleme, sistem tanılama ve denetim gibi alanlarda etkin olarak kullanılır.

5. *Hata Toleransı* : YSA, çok sayıda hücrenin çeşitli şekillerde bağlanmasından oluştuğundan paralel dağılmış bir yapıya sahiptir ve ağın sahip olduğu bilgi, ağdaki bütün bağlantılar üzerine dağılmış durumdadır. Bu nedenle, eğitilmiş bir YSA'nın bazı bağlantılarının hatta bazı hücrelerinin etkisiz hale gelmesi, ağın doğru bilgi üretmesini önemli ölçüde etkilemez. Bu nedenle, geleneksel yöntemlere göre hatayı tolere etme yetenekleri son derece yüksektir.

6. *Donanım ve Hız* : YSA, paralel yapısı nedeniyle büyük ölçekli entegre devre (VLSI) teknolojisi ile gerçeklenebilir. Bu özellik, YSA'nın hızlı bilgi işleme yeteneğini artırır ve gerçek zamanlı uygulamalarda arzu edilir.

7. *Analiz ve Tasarım Kolaylığı* : YSA' nın temel işlem elemanı olan hücrenin yapısı ve modeli, bütün YSA yapılarında yaklaşık aynıdır. Dolayısıyla, YSA'nın farklı uygulama alanlarındaki yapıları da standart yapıdaki bu hücrelerden oluşacaktır. Bu nedenle, farklı uygulama alanlarında kullanılan YSA'ları benzer öğrenme algoritmalarını ve teorilerini

paylaşabilirler. Bu özellik, problemlerin YSA ile çözümünde önemli bir kolaylık getirecektir.

3.1.3.2 - Yapay Sinir Ağlarının Tarihçesi

Yapay Sinir Ağları (YSA), beynin fizyolojisinden yararlanılarak oluşturulan bilgi işleme modelleridir. Bazı bilim adamları, beynimizin güçlü düşünme, hatırlama ve problem çözme yeteneklerini bilgisayara aktarmaya çalışmışlardır. Bazı araştırmacılar ise, beynin fonksiyonlarını kısmen yerine getiren bir çok modelleri oluşturmaya çalışmışlardır (Efe ve Kaynak, 2000).

YSA'ların öğrenme özelliği, araştırmacıların dikkatini çeken en önemli özelliklerden birisidir. Çünkü herhangi bir olay hakkında girdi ve çıktılar arasındaki ilişkiyi, doğrusal olsun veya olmasın, elde bulunan mevcut örneklerden öğrenerek daha önce hiç görülmemiş olayları, önceki örneklerden çağrışım yaparak ilgili olaya çözümler üretebilme özelliği YSA'lardaki zeki davranışın da temelini teşkil eder

1943 yılında bir nörobiyolojist olan Warren McCulloch ve bir istatistikçi olan Walter Pitts, "Sinir Aktivitesindeki Düşüncelere Ait Bir Mantıksal Hesap" başlıklı bir makale ile ilk dijital bilgisayarlara ışık tutmuştur. John Von Neumann bu makaleyi, "elektronik beyinler" için bir kopya olarak görmüştür. Yapay zeka alanındaki araştırmacılar içerisinde istisnai bir yeri olan Marvin Minsky, bu makaleden aldığı ilhamla makroskobik zeka fikrini ortaya atmış ve uzman sistemlerin doğmasına neden olmuştur. Bronx Yüksek Bilim Okulu'ndan Frank Rosenblatt, gözün hesaplamaları ile ilgilenmiştir. Bu bilim adamları, öğrenmenin ve zekanın herhangi bir özelliğinin simülasyonunda bilgisayarların aktif olarak nasıl kullanılabileceğini, 1956 yılında düzenlemiş oldukları ilk yapay zeka konferansında tartışmışlardır.

1959'da, Stanford üniversitesinden Bernard Widrow, basit nöron benzeri elemanlara dayanan ve "adaline" (Adaptive Linear Neuron) olarak adlandırılan bir adaptif lineer elemanı geliştirmiştir. Adaline ve iki tabakalı biçimi olan "madaline" (Multiple Adaline); ses tanıma, karakter tanıma, hava tahmini ve adaptif kontrol gibi çok çeşitli uygulamalar için kullanılmıştır. Daha sonraları adaline, ayrık bir çıkış yerine sürekli bir çıkış üretmek için geliştirilmiştir. Widrow, telefon hatları üzerindeki ekoları

elimine etmeye yarayan adaptif filtreleri geliştirmede, adaptif lineer eleman algoritmasını kullanmıştır. Bununla ilk defa YSA'lar gerçek bir probleme uygulanmıştır

Helsinki Teknik Üniversitesi'nden Teuvo Kohonen, 1970'lerin ilk yıllarında adaptif öğrenme ve birleşik hafızalar üzerine temel çalışmalar yapmış ve bu olduğu çalışmaları ile danışmansız öğrenme metotlarının gelişmesine ışık tutmuştur

Minsky ve Papert'in perseptron isimli kitaplarında, YSA'nın temel olarak ilgi çekici konular olmadığını belirtmeleri bir çok araştırmacının bu alanda çalışmaktan vazgeçmelerine sebebiyet vermiştir. YSA konusunda çalışmaya devam eden Grossberg , YSA modellerini yapılandırmak için nörolojik verinin kullanılması, algı ve hafıza için YSA tabanlı mekanizmaların önerilmesi, belirgin eşitliklerle bütünleşen bir sinaptik model için bir ilişkilendirici kural üzerinde çalışmıştır.

1982 yılında ilgi çeken bir başka gelişme, moleküller biyolojiden beyin kuramcılığına geçiş yapan bir model Caltech fizikçisi Hopfield tarafından sunulmuştur. Kendi adıyla anılan bir ağ yapısı mevcuttur ve bir çok alana uygulanmıştır.

1987 yılında yapılan ilk yapay sinir ağları sempozyumundan sonra YSA uygulamaları yaygınlaşmıştır. Günümüzde, YSA'larla ilgili araştırmalar yapan çok sayıda bilim adamı ve araştırma grupları vardır. Farklı bilim ve ilgi alanlarında çalışan birçok araştırmacı, birçok yeni gelişmeleri sunmaya devam edeceklerdir.

3.1.3.3 - Yapay Sinir Ağının Yapısı

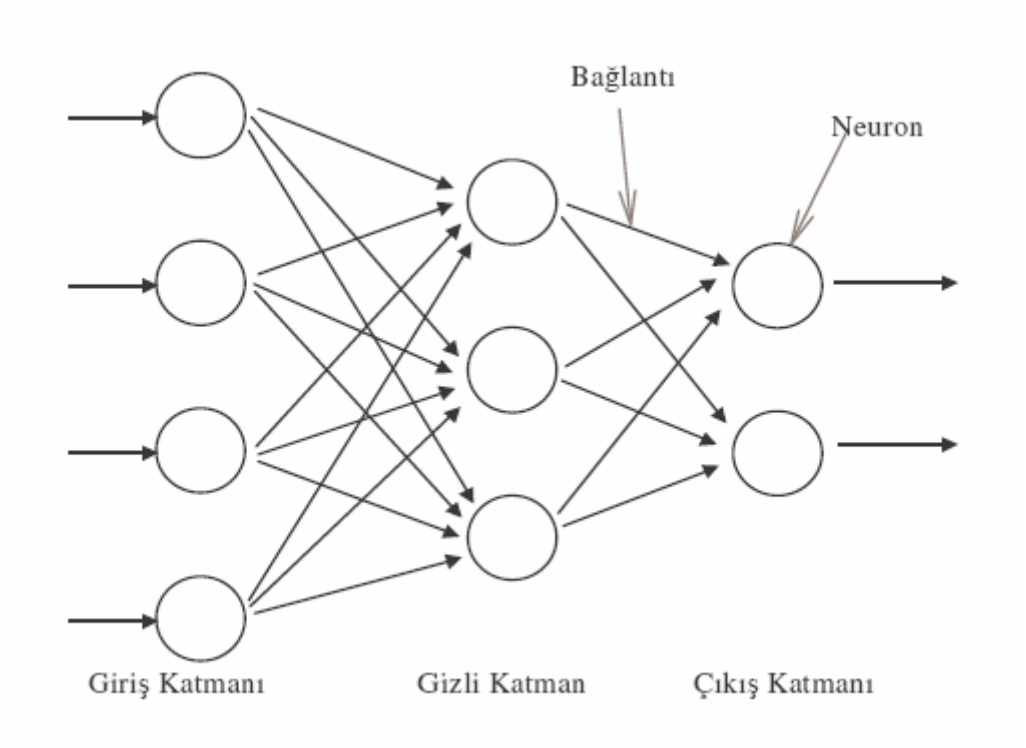
Sinir hücreleri bir grup halinde işlev gördüklerinde ağ olarak adlandırılırlar ve böyle bir grupta binlerce nöron bulunur. Yapay nöronların birbirleriyle bağlantılar aracılığıyla bir araya gelmeleri yapay sinir ağını oluşturmaktadır. Yapay sinir ağıyla aslında biyolojik sinir ağının bir modeli oluşturulmak istenmektedir. Nöronların aynı doğrultu üzerinde bir araya gelmeleriyle katmanlar oluşmaktadır (Haykin, 1994).

Katmanların değişik şekilde bir birleriyle bağlanmaları değişik ağ mimarilerini doğurur. YSA'ları üç katmadan oluşur. Bu katmanlar sırasıyla;

a) *Girdi Katmanı* : Bu katmandaki proses elemanları dış dünyadan bilgileri alarak ara katmanlara transfer ederler. Bazı ağlarda girdi katmanında herhangi bir bilgi işleme olmaz.

b) *Ara Katman (Gizli Katman)* : Girdi katmanından gelen bilgiler işlenerek çıktı katmanına gönderilirler. Bu bilgilerin işlenmesi ara katmanlarda gerçekleştirilir. Bir ağ içinde birden fazla ara katman olabilir.

c) *Çıktı Katmanı* : Bu katmandaki proses elemanları ara katmandan gelen bilgileri işleyerek ağın girdi katmanından sunulan girdi seti için üretmesi gereken çıktıyı üretirler. Üretilen çıktı dış dünyaya gönderilir.



Şekil 3.10 : Yapay Sinir Ağı Modeli

3.1.3.4 - Yapay Sinir Ağı Nöron Modeli

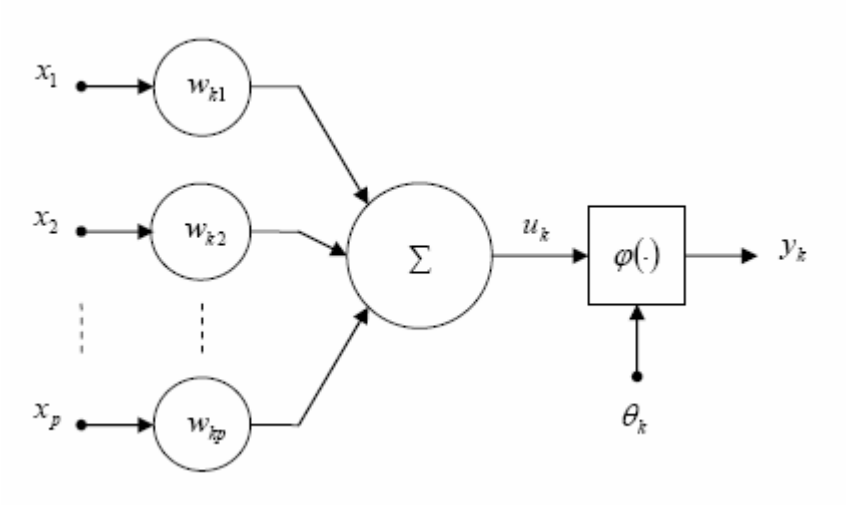
YSA'lar, insan beyninin çalışma prensibi örnek alınarak geliştirilmeye çalışılmıştır ve aralarında yapısal olarak bazı benzerlikler vardır. Bu benzerlikler Çizelge 3.1'de verilmiştir.

Çizelge 3.1 : Sinir Sistemi ile YSA'nın benzerlikleri

SİNİR SİSTEMİ	**YSA SİSTEMİ**
Neuron	İşlem elemanı
Dendrit	Toplama fonksiyonu
Hücre gövdesi	Transfer fonksiyonu
Aksonlar	Eleman çıkışı
Sinapslar	Ağırlıklar

Bir YSA modelinin temel birimi, Şekil 3.11'de gösterilen işlem elemanıdır. Bu elemen yapay nöron olarak adlandırılır. Nöron modelinin üç temel elemanı vardır (Çavdur, 2005):

- Her biri kendine ait bir ağırlıkla karakterize edilen, sinapsisler (synapses veya connecting links) kümesi. Özel olarak, j sinapsisindeki k nöronuna bağlı olan x_j sinyali, w_{kj} ağırlığıyla çarpılır. w_{kj} ağırlığında indislerinin yazım şekli önemlidir. İlk indis incelenen nörona, ikinci indis de sinapsisin giriş tarafındaki nörona karşılık gelmektedir. Literatürde bu gösterimin tersi de kullanılmaktadır. İlgili sinapsis yükseltici ise w_{kj} ağırlığı pozitif, indirgeyici ise negatiftir.
- Sırasıyla nöronun sinapsisleri tarafından ağırlıklandırılmış giriş sinyallerini toplamak için bir toplayıcı, burada açıklanan işlemler bir doğrusal birleştirici oluşturmaktadır.
- Nöron çıkışının büyümesini sınırlandırmak için bir aktivasyon fonksiyonu (activation veya squashing function). Tipik olarak, normalize edilmiş bir nöron çıkışı [0,1] veya $[-1,+1]$ aralıklarındadır.



Şekil 3.11 : Yapay Sinir Ağı Nöron Modeli

k nöronu,

$$U_k = \sum_{j=1}^{p} w_{kj} x_j \tag{3.1}$$

ve

$$y_k = \varphi(u_k - \theta_k) \tag{3.2}$$

denklemleriyle tanımlanabilir. Burada, $x_1, x_2, \ldots, x_p$ giriş sinyalleri, $w_{k1}, w_{k2}, \ldots, w_{kp}$ sinaptik ağırlıklar, u_k doğrusal birleştirici çıktısı, θ_k eşik, $\varphi(.)$ aktivasyon fonksiyonu, ve y_k çıkış sinyalidir.

θ_k eşiğinin uygulanmasıyla, doğrusal birleştiricinin u_k çıkışında,

$$v_k = u_k - \theta_k \tag{3.3}$$

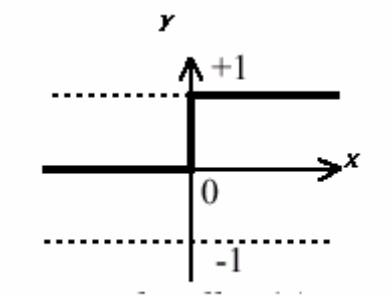
şeklinde bir dönüşüm olmaktadır.

3.1.3.5 - Aktivasyon Fonksiyonları

Transfer fonksiyonu olarak da geçen aktivasyon fonksiyonu, birleştirme fonksiyonundan elde edilen net girdiyi bir işlemden geçirerek hücre çıktısını belirleyen

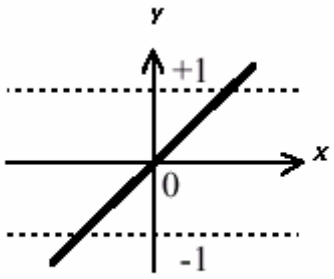
ve genellikle doğrusal olmayan bir fonksiyondur ve yapay nöron modelinde $\varphi(.)$ olarak ifade edilmektedir. Nöron modellerinde, hücrenin gerçekleştireceği işleve göre çeşitli tipte aktivasyon fonksiyonları kullanılabilir (Elmas, 2003). Bunlar;

a) Eşik aktivasyon fonksiyonu : Eğer net değer sıfırdan küçükse sıfır, sıfırdan daha büyük bir değer ise net çıkışında +1 değeri verir. Eşik aktivasyon fonksiyonunun –1 ile +1 arasında değişeni ise signum aktivasyon fonksiyonu olarak adlandırılır. Signum aktivasyon fonksiyonu, net giriş değeri sıfırdan büyükse +1, sıfırdan küçükse –1, sıfıra eşitse sıfır değerini verir. Şekil 3.12'de eşik aktivasyon fonksiyonunun grafiği görülmektedir.



Şekil 3.12 : Eşik Aktivasyon Fonksiyonu.

b) Lineer aktivasyon fonksiyonu : Bu fonksiyonun çıkışı girişine eşittir. Sürekli çıkışlar gerektiği zaman çıkış katmanındaki aktivasyon fonksiyonunun lineer aktivasyon fonksiyonu olabildiğine dikkat edilmelidir. Şekil 3.13'da doğrusal aktivasyon fonksiyonu görülmektedir.

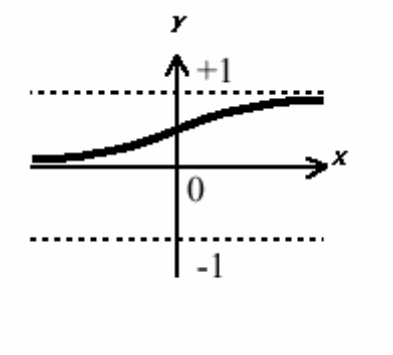


Şekil 3.13 : Doğrusal Aktivasyon Fonksiyonu

$$f(x) = x \tag{3.4}$$

şeklinde ifade edilir.

c) Lojistik fonksiyon : Bu fonksiyonunun lineer olmamasından dolayı türevi alınabilmektedir Böylece daha sonraki bölümlerde görülecek olan geri yayınımlı ağlarda kullanmak mümkün olabilmektedir. Şekil 3.14'da logaritma sigmoid transfer fonksiyonu görülmektedir.



Şekil 3.14 : Logaritma Sigmoid Aktivasyon Fonksiyonu.

$$\text{Lojistik fonksiyonu,} \quad f(x) = \text{lojistik}(x) = \frac{1}{1+\exp(-\beta x)} \tag{3.5}$$

şeklinde ifade edilir.

Buradaki β eğim sabiti olup genelde bir olarak seçilmektedir.

d) Hiperbolik tanjant aktivasyon fonksiyonu : Bu fonksiyonda lineer olmayan türevi alınabilir bir fonksiyondur. +1 ile –1 arasında çıkış değerleri üreten bu fonksiyon lojistik fonksiyona benzemektedir.Denklemi aşağıda görüldüğü gibidir.

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.6}$$

Bu aktivasyon fonksiyonlarından başka fonksiyonlar da vardır. Yapay sinir ağında hangi aktivasyon fonksiyonunun kullanılacağı probleme bağlı olarak değişmektedir. Yukarıda verilen fonksiyonlar en genel aktivasyon fonksiyonlarıdır.

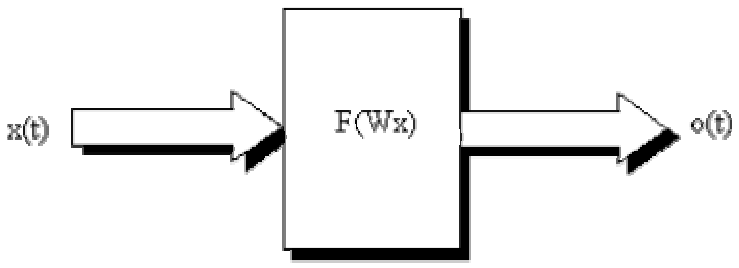
3.1.3.6 - Yapay Sinir Ağlarının Sınıflandırılması

YSA'lar, genel olarak birbirleri ile bağlantılı işlemci birimlerden veya diğer bir ifade ile işlemci elemanlardan (nöron) oluşurlar. Her bir sinir hücresi arasındaki bağlantıların yapısı ağın yapısını belirler. İstenilen hedefe ulaşmak için bağlantıların nasıl değiştirileceği öğrenme algoritması tarafından belirlenir. YSA'lar yapılarına ve öğrenme algoritmalarına göre sınıflandırılırlar (Haykin, 1994).

1. *YSA'ların Yapılarına Göre Sınıflandırılması :*

Yapay sinir ağları, yapılarına göre, ileri beslemeli ve geri beslemeli ağlar olmak üzere iki şekilde sınıflandırılırlar.

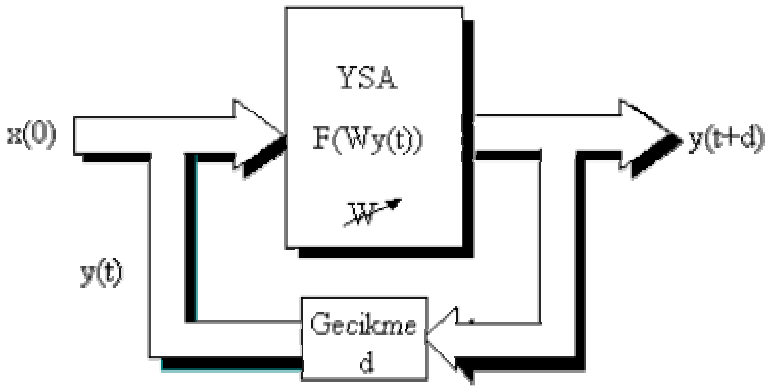
a) İleri Beslemeli Ağlar : İleri beslemeli bir ağda işlemci elemanlar (İE) genellikle katmanlara ayrılmışlardır. İşaretler, giriş katmanından çıkış katmanına doğru tek yönlü bağlantılarla iletilir İE'ler bir katmandan diğer bir katmana bağlantı kurarlarken, aynı katman içerisinde bağlantıları bulunmaz. Şekil 3.15'de ileri beslemeli ağ için blok diyagram gösterilmiştir. İleri beslemeli ağlara örnek olarak çok katmanlı perseptron (Multi Layer Perseptron-MLP) ve LVQ (Learning Vector Quantization) ağları verilebilir.



Şekil 3.15 : İleri Beslemeli Ağ İçin Blok Diyagram

b) Geri Beslemeli Ağlar : Bir geri beslemeli sinir ağı, çıkış ve ara katlardaki çıkışların, giriş birimlerine veya önceki ara katmanlara geri beslendiği bir ağ yapısıdır. Böylece,

girişler hem ileri yönde hem de geri yönde aktarılmış olur. Şekil 3.16'de bir geri beslemeli ağ görülmektedir. Bu çeşit sinir ağlarının dinamik hafızaları vardır ve bir andaki çıkış hem o andaki hem de önceki girişleri yansıtır. Bundan dolayı, özellikle önceden tahmin uygulamaları için uygundurlar. Bu ağlar çeşitli tipteki zaman-serilerinin tahmininde oldukça başarı sağlamışlardır. Bu ağlara örnek olarak Hopfield, SOM (Self Organizing Map), Elman ve Jordan ağları verilebilir.

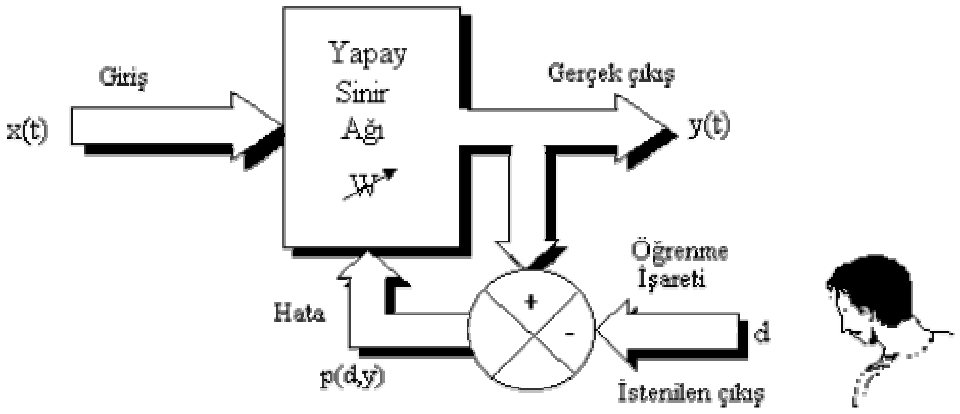


Şekil 3.16: Geri Beslemeli Ağ İçin Blok Diyagram

2. *YSA'ların Öğrenme Algoritmalarına Göre Sınıflandırılması :*

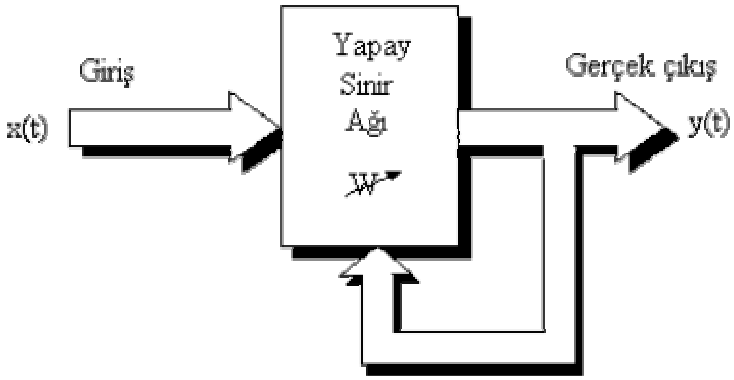
Öğrenme; gözlem, eğitim ve hareketin doğal yapıda meydana getirdiği davranış değişikliği olarak tanımlanmaktadır. O halde, birtakım metot ve kurallar, gözlem ve eğitime göre ağdaki ağırlıkların değiştirilmesi sağlanmalıdır. Bunun için genel olarak üç öğrenme metodundan ve bunların uygulandığı değişik öğrenme kurallarından söz edilebilir. Bu öğrenme kuralları aşağıda açıklanmaktadır.

a) Danışmanlı Öğrenme : Bu tip öğrenmede, YSA'ya örnek olarak bir doğru çıkış verilir. İstenilen ve gerçek çıktı arasındaki farka göre İE'ler arası bağlantıların ağırlığını en uygun çıkışı elde etmek için sonradan düzenlenebilir. Bu sebeple danışmanlı öğrenme algoritmasının bir "öğretmene" veya "danışmana" ihtiyacı vardır. Şekil 3.17'de danışmanlı öğrenme yapısı gösterilmiştir. Widrow-Hoff tarafından geliştirilen delta kuralı ve Rumelhart ve McClelland tarafından geliştirilen genelleştirilmiş delta kuralı veya geri yayılım algoritması danışmanlı öğrenme algoritmalarına örnek olarak verilebilir.



Şekil 3.17 : Danışmanlı Öğrenme Yapısı.

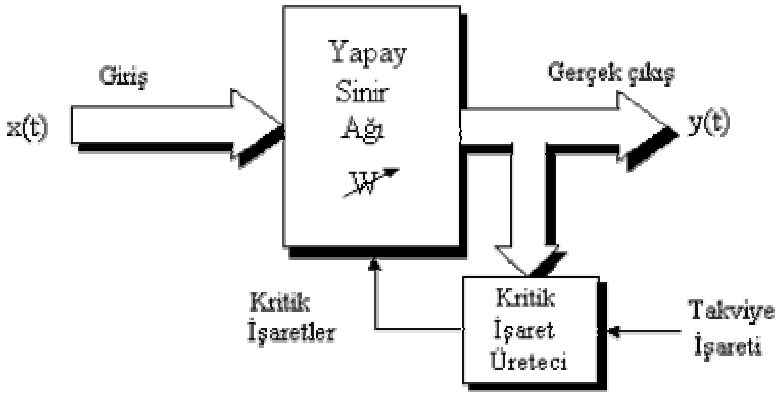
b) Danışmansız Öğrenme : Girişe verilen örnekten elde edilen çıkış bilgisine göre ağ sınıflandırma kurallarını kendi kendine geliştirmektedir. Bu öğrenme algoritmalarında, istenilen çıkış değerinin bilinmesine gerek yoktur. Öğrenme süresince sadece giriş bilgileri verilir. Ağ daha sonra bağlantı ağırlıklarını aynı özellikleri gösteren desenler oluşturmak üzere ayarlar. Şekil 3.18'de danışmansız öğrenme yapısı gösterilmiştir. Grossberg tarafından geliştirilen ART (Adaptive Resonance Theory) veya Kohonen tarafından geliştirilen SOM (Self Organizing Map) öğrenme kuralı danışmansız öğrenmeye örnek olarak verilebilir.



Şekil 3.18 : Danışmansız Öğrenme Yapısı.

c) Takviyeli öğrenme : Bu öğrenme kuralı danışmanlı öğrenmeye yakın bir metoddur. Denetimsiz öğrenme algoritması, istenilen çıkışın bilinmesine gerek duymaz. Hedef çıktıyı vermek için bir "öğretmen" yerine, burada YSA'ya bir çıkış verilmemekte fakat

elde edilen çıkışın verilen girişe karşılık iyiliğini değerlendiren bir kriter kullanılmaktadır. Şekil 3.19'da takviyeli öğrenme yapısı gösterilmiştir. Optimizasyon problemlerini çözmek için Hinton ve Sejnowski'nin geliştirdiği Boltzmann kuralı veya GA takviyeli öğrenmeye örnek olarak verilebilirler.



Şekil 3.19 : Takviyeli Öğrenme Yapısı

3.1.3.7 - Yapay Sinir Ağlarının Eğitilmesi

Yapay sinir ağlarının öğrenme sürecinde de, tıpkı dış ortamdan gözle veya vücudun diğer organlarıyla uyarıların alınması gibi dış ortamdan girişler alınır, bu girişlerin beyin merkezine iletilerek burada değerlendirilip tepki verilmesi gibi yapay sinir ağında da aktivasyon fonksiyonundan geçirilerek bir tepki çıkışı üretilir. Bu çıkış yine tecrübeyle verilen çıkışla karşılaştırılarak hata bulunur. Çeşitli öğrenme algoritmalarıyla hata azaltılıp gerçek çıkışa yaklaşılmaya çalışılır. Bu çalışma süresince yenilenen yapay sinir ağının ağırlıklarıdır. Ağırlıklar her bir çevrimde yenilenerek amaca ulaşılmaya çalışılır. Amaca ulaşmanın veya yaklaşmanın ölçüsü de yine dışarıdan verilen bir değerdir. Eğer yapay sinir ağı verilen giriş-çıkış çiftleriyle amaca ulaşmış ise ağırlık değerleri saklanır. Ağırlıkların sürekli yenilenip istenilen sonuca ulaşılana kadar geçen zamana öğrenme adı verilir (Efe ve Kaynak, 2003)

Yapay sinir ağı öğrendikten sonra daha önce verilmeyen girişler verilip, sinir ağı çıkışıyla gerçek çıkışı yaklaşımı incelenir. Eğer yeni verilen örneklere de doğru yaklaşıyorsa sinir ağı işi öğrenmiş demektir. Sinir ağına verilen örnek sayısı optimum değerden fazla ise sinir ağı işi öğrenmemiş ezberlemiştir. Genelde eldeki örneklerin

yüzde sekseni ağa verilip ağ eğitilir, daha sonra geri kalan yüzde yirmilik kısım verilip ağın davranışı incelenir diğer bir deyişle ağ böylece test edilir.

Çizelge 3.2 : Öğrenme Algoritmaları ve Uygulandıkları Alanlar

Uygulama Tipi	Yapay Sinir Ağı
Öngörü Tanıma	* Geri Yayılım * Delta Bar Delta * Geliştirilmiş Delta Bar Delta * Yönlendirilmiş Rastsal Tanıma * Geri Yayılım içinde Self Organizing Map * Higher Order Neural Networks
Sınıflandırma	* Learning Vektor Quantization * Counter-Propagation * Olasılıklı Yapay Sinir Ağları
Veri İlişkilendirme (Data Association)	* Hopfield * Boltman Makinesi * Bidirectional Associative Memory * Spantion - Temproal Pattern Recogniation
Veri Kavramlaştırma (Data Conceptualization)	* Adaptive Resonance Network * Self Organizing

Birçok öğrenme algoritmasının bulunmasından dolayı bu kısımda sadece en popüler öğrenme algoritması olan Geri Yayınım Algoritması anlatılacaktır.

Geri Yayılım Algoritması

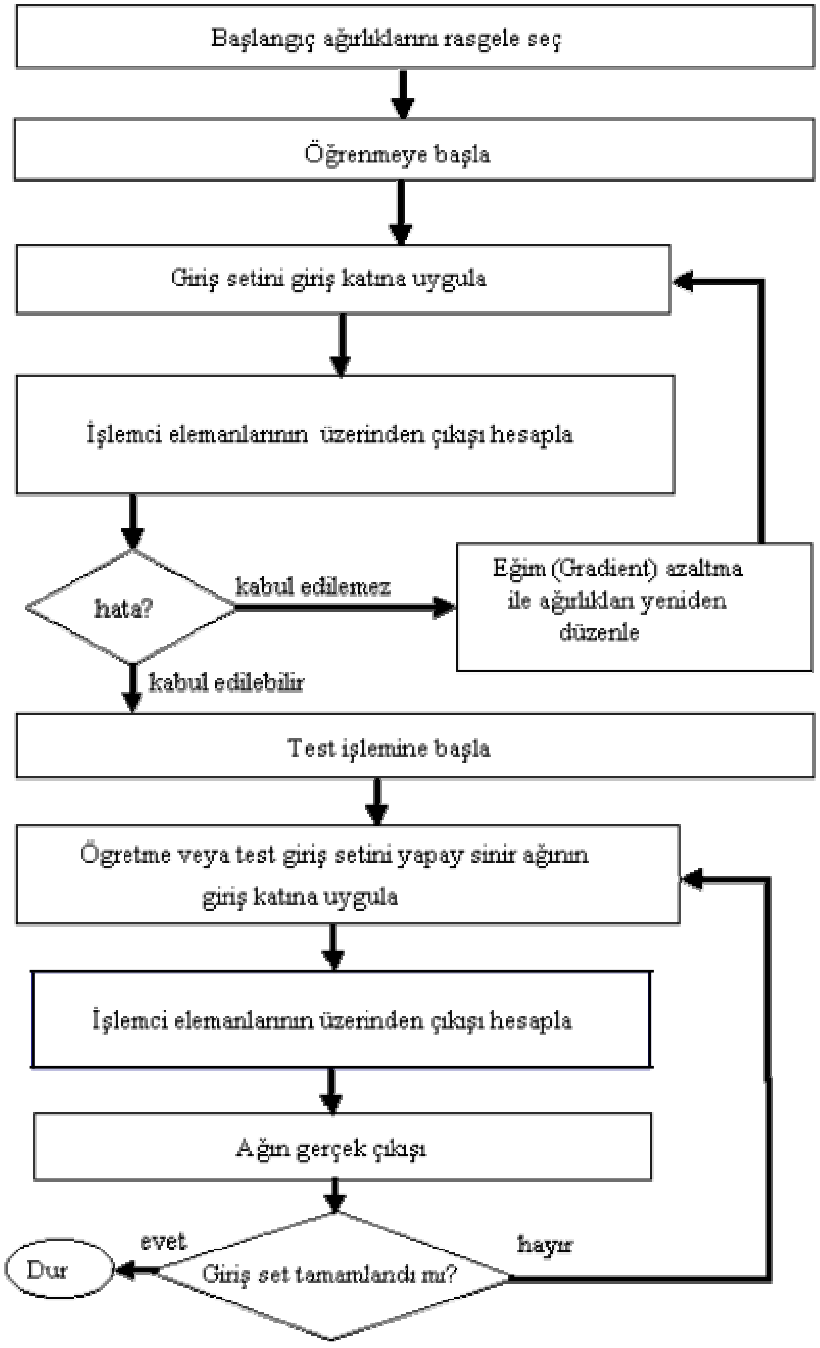
Yapay sinir ağları belki de en çok tahmin amacıyla kullanılmaktadır. Tahmin için kullanılan yapay sinir ağları içinde de en yaygın olarak kullanılanı geri yayılım algoritmasıdır. Geri yayılım algoritması ileri beslemeli ve çok katmanlı bir ağ mimarisini gerektirmektedir. Geri yayılım Algoritması, YSA'lar açısından en önemli tarihsel gelişmelerden biridir.Geri yayılım algoritması veya bir başka adıyla Genelleştirilmiş Delta Algoritması uygulamada en çok kullanılan öğretme algoritmasıdır. Hata, ağdaki ağırlıkların bir fonksiyonu olarak görülür ve hataların kareleri ortalaması dereceli azaltma yöntemi kullanılarak, minimize edilmeye çalışılır.

Bu algoritma, hataları çıkıştan girişe geriye doğru azaltmaya çalışmasından dolayı geri yayılım ismini almıştır (Haykin, 1994).

Geri yayılım algoritması, çok katmanlı ağları eğitmede en çok kullanılan temel bir öğrenme algoritmasıdır. Eğitme işlemi ve eğitimden sonraki test işlemi bu akışa göre gerçekleştirilir. Geri yayılım algoritması, danışmanlı öğrenme yöntemini kullanılır. Örnekler ağa öğretilir ve ağa hedef değeri verilir. Öğrenme, her örnek için ağın çıktı değeri ile hedef değerinin ile karşılaştırılır. Hata değeri, ağa tekrar geri besleme şeklinde verilir. Örnek setindeki hata kareleri toplamını azaltmak için nöronlar arasındaki bağlantı ağırlıkları değiştirilir.

Burada, gizli katman sayısını tespit etmek zor bir işlemdir. Gizli katmanlar, doğrusal olmayan yada değişkenler arasında etkileşim var ise kullanılır. Bu etkileşim ne kadar karmaşıksa, o kadar çok gizli katmana ihtiyaç duyulur. Eğer az sayıda gizli katman kullanılırsa ağ öğrenmeyi başarmaz. Gereğinden fazla sayıda gizli katman bulunması halinde ise, ağ ezberler. Bu da ağın, yeni örnekler için genelleme yeteneğini azaltır. Amaç, ağın genelleme yeteneğini optimum yapabilecek en az sayıda katmanı kullanmaktır. Ağı doğru zamanda durdurmak, ağın ezberlemesini önler. Gizili katman sayısını belirlemenin hızlı bir yolu yoktur. Bunu belirlemek için benzer veri seti kullanılmış tecrübelerden faydalanılabilir. Geri yayılım algoritmasının bir dezavantajı, yakınsama hızının yavaş olması ve yerel en iyi çözümlerde durabilmedir.

Tipik çok katlı geri yayılım ağı, daima; bir giriş tabakası, bir çıkış tabakası ve en az bir gizli tabakaya sahiptir. Gizli tabakaların sayısında teorik olarak bir sınırlama yoktur. Fakat genel olarak bir veya iki tane bulunur. Bu algoritmanın akış şeması Şekil 3.20'de verilmiştir.



Şekil 3.20 : Geri Yayılım Algoritmasının Akış Şeması

Öğrenme algoritması olarak geri yayılım algoritması seçildiğinde öğrenme katsayısı önem kazanmaktadır. η: Öğrenme katsayısıdır. Öğrenme katsayısı, ağırlıkların bir sonraki düzeltmede hangi oranda değiştirileceğini göstermektedir. Küçük öğrenme katsayıları, ağın sonuca ulaşmasını yavaşlatır. Büyük öğrenme katsayıları, ağın sonuca daha kısa sürede ulaşmasını sağlar. Bununla birlikte çok yüksek oranlar ağın

hesaplamalarında büyük salınımlara neden olur ve ağın dip noktayı bulmasını engelleyebilir. Öğrenme katsayısı için tipik değerler 0,01 ile 0,9 arasında değişir. Karmaşık ve zor çalışmalar için küçük öğrenme katsayıları seçilmesi önerilir.

Geri yayılım algoritmasının matematiksel gösteriminde kullanılan notasyonlar aşağıdaki gibidir (Haykin, 1994):

- *i, j* ve *k* , ağ boyunca soldan sağa doğru yayılan farklı nöronları göstermektedir. *j* nöronu *i* nöronunun sağındaki katmanda ve *k* nöronu j nöronunun sağındaki katmanda bulunmaktadır.
- *n* iterasyonu, ağa verilen *n*. eğitim örneğini göstermektedir.
- $\varepsilon(n)$, *n*. iterasyondaki hata kareleri toplamını göstermektedir. Tüm *n* değerleri için $\varepsilon(n)$ ortalaması, ε_{av} ortalama karesel hatasını göstermektedir.
- $e_j(n)$, *n*. iterasyonda, *j* nöronu çıkışındaki hata sinyalini göstermektedir.
- $d_j(n)$, *n*. iterasyonda, *j* nöronu çıkışındaki istenen cevabı göstermektedir.
- $y_j(n)$, *n*. iterasyonda, *j* nöronu çıkışındaki fonksiyon sinyalini göstermektedir.
- $w_{ji}(n)$, *n.* iterasyonda, *i* nöronu çıkışını, *j* nöronu girişine bağlayan sinaptik ağırlığı göstermektedir. Bu ağırlığa *n*. iterasyonda uygulanan düzeltme $\Delta W_{ji}(n)$ ile gösterilmektedir.
- $v_j(n)$, *n.* iterasyonda, *j* nöronunun net içsel aktivite düzeyini göstermektedir. Bu değer, *j* nöronundaki doğrusal olmayan yapıya uygulanan sinyali göstermektedir.
- $\varphi_j(.)$, *j* nöronunun giriş çıkış arasındaki doğrusal olmayan yapının fonksiyonel ilişkisini tanımlayan aktivasyon fonksiyonunu göstermektedir.
- θ_j , *j* nöronuna uygulanan eşik değerini göstermektedir. Eşik değerinin etkisi, -1 değerine eşit olan bir $w_{j0} = \theta_j$ ağırlık sinapsi ile gösterilmektedir.
- $x_i(n)$, giriş vektörünün , *n.* iterasyondaki, *i.* elemanını göstermektedir.
- $o_k(n)$, çıkış vektörünün , *n.* iterasyondaki, *k*. elemanını göstermektedir.
- η, öğrenme katsayısını göstermektedir.

n. iterasyonda, j nöronu çıkışındaki hata sinyali,

$$e_j(n) = d_j(n) - y_j(n), \text{ j nöronu bir çıkış düğümü} \tag{3.7}$$

şeklinde tanımlanmaktadır. Buna göre, ağın hata kareleri toplamı,

$$\varepsilon(n) = \frac{1}{2}\sum_{j\in c} e_j^2(n) \tag{3.8}$$

olarak yazılmaktadır. Verilen bir eğitim kümesi için ε_{av}, eğitim kümesi öğrenme performansının ölçüsü olarak maliyet fönksiyonunu göstermektedir.Öğrenme sürecinin amacı, ε_{av} değerini en küçükleyecek şekilde ağ parametrelerinin ayarlanmasıdır.

Solundaki nöron katmanı tarafından üretilen fonksiyon sinyalleri kümesini giriş olarak alan bir *j* nöronu ele alınsın, *j* nöronunun net içsel aktivite düzeyi,

$$v_j(n) = \sum_{i=0}^{p} w_{ji}(n) y_i(n) \tag{3.9}$$

olarak tanımlanmaktadır.Buna göre, *n.* iterasyonda , *j* nöronuna uygulanan fonksiyon sinyali,

$$y_j(n) = \varphi_j(v_j(n)) \tag{3.10}$$

ve $w_{ji}(n)$ sinaptik ağırlığına uygulanan, $\Delta W_{ji}(n)$ düzeltme değeri Denklem 3.11'deki gibidir.

$$\Delta W_{ji}(n) = \eta \delta_j(n) y_i(n) \tag{3.11}$$

elde edilir. Burada yerel eğim,

$$\delta_j(n) = e_j(n)\varphi_j'(v_j(n)) \tag{3.12}$$

şeklinde tanımlanmaktadır. Denkleme göre, j nöronu için yerel eğim $e_j(n)$ hata sinyali ile $\varphi_j^t(v_j(n))$ aktivasyon fonksiyonunun türevinin çarpımıdır. $\delta_j(n)$'nun değeri hesaplanırken, j nöronunun konumuna bağlı olarak iki farklı durum oluşmaktadır. Bunlar;

- j nöronunun çıkış düğümü olması durumunda, kendisine ait bir istenen cevabı bulunmaktadır. Bu nedenle, $e_j(n)$ hata sinyali ve $\delta_j(n)$ doğrudan hesaplanabilmektedir.
- j nöronunun ara düğüm olması durumunda ise hata sinyali, j nöronuna, doğrudan bağlı olan, sonraki gizli katman veya çıkış katmanındaki nöronlara ait δ değerlerinin ağırlıklarının çarpılmasıyla

$$\delta_j(n) = \varphi_j^t(v_j(n)) \sum_k \delta_k(n) w_{kj}(n) \tag{3.13}$$

denklemi elde edilmektedir.

Denklem 3.14'deki gibi ağırlıklar revize edilir. Bu çevrim ε_{av}'nin değeri istenen minimum seviyeye gelinceye kadar devam eder.

$$w_{ji}(Yeni) = w_{ji}(Eski) + \Delta w_{ji}(n) \tag{3.14}$$

3.2 – Yöntem

3.2.1 – Genel Bilgiler

Uygulama çalışması, veri madenciliğinin İnternet üzerindeki uygulama alanı Web madenciliğinin bir dalı olan, Web içerik madenciliğinin kapsamına girmektedir. Web içerik madenciliği temel olarak İnternet de saklı bilgiyi bulma üzerine yoğunlaşmıştır (arama motorları, vs.). Web içerik madenciliği akıllı yazılım ajanları (Web robotları, Web örümcekleri vs.) ile ilgilidir.

Genel anlamda Web madenciliği, veri madenciliği tekniklerini kullanarak Web'de bulunan dosya ve servislerden otomatik olarak öngörülmeyen bilgiye ulaşmaktır.

Günümüzde bu işlemlerden birçoğu rahatlıkla arama motorları veya dizin arama siteleri tarafından yapılabilmektedir. Buna rağmen web madenciliğine ihtiyaç duyulmasının ana nedenleri ;

- Arama motorlarında arama yapıldığında, listelenen sonuçlar içerisinde araştırılan konuyla yeteri kadar ilgisi olmayan bilgiler ile karşılaşılmakta ve sonuçlar sıralanırken en çok ilgiliden en az ilgili olana doğru ilerleyen bir sıra bulunmamaktadır. Bu durum kullanıcıları isteksiz yapmakta ve karmaşık sorgular sonucu istenen bilgiye ulaşılması gerekliliği kullanıcıların uzman bilgisine sahip olmasını sorunlu hale getirmektedir.
- Dizin arama yönteminde, ana konulardan başlayarak ve giderek alt düzeylere inerek, aranılan konuya doğru ilerlenebilmektedir. Bu durumda kullanıcıların, kendileri için en uygun seçeneği kendilerinin bulması ve karşılarına çıkacak olan çok sayıda sayfadan oluşan sayfa listesinde hangi sayfaların işlerine yarayacağını kendilerinin belirlemesi gerekmektedir. Çoğu kullanıcının böyle bir işlemi tam olarak yerine getirecek kadar zamanı olmamaktadır.

Bu sorunlar nedeniyle kullanıcılara farklı bir alternatif olarak, belirlenen spesifik bir konuda İnternet üzerinde arama yapan bir derleyici geliştirilmiştir. Konu bu çalışmada kuş gribi (Avian Influenza) olarak seçilmiştir. Kuş gribinin seçilmesinin nedeni bu günlerde popüler bir konu olmasındandır. İleride de anlatılacağı gibi önerilen derleyici ve ilgililik değerlendirici yapısının esnek olması sebebiyle, geliştirilen derleyici, istenilen herhangi bir konuda çalışabilmektedir. İstenilen konuya ait belirtilen sayıda anahtar kelimenin ve başlangıç noktası olacak referans sitelerin uzman kişi tarafından belirlenmesi yeterli olmaktadır.

Derleyici taradığı sitelerin ilgili olup olmadıklarına bir veri madenciliği tekniği olan yapay sinir ağı kullanarak karar vermektedir. Burada, Özmutlu ve arkadaşlarının ASIS&T (The Information Society for the Information Age) 2004'de sundukları "Güvenlik konusunda geliştirilmiş özel bir web derleyicisi" konulu çalışma referans alınmıştır. Bu çalışmada önerilen derleyici tasarımı kısmen hayata geçirilmiştir. Bu çalışma aşamalar şeklinde gerçekleştirilmiştir. Daha önceki çalışmalarda sayfaların ilgililik seçimi için anahtar kelime sıklıklarını sayma tekniği kullanılmıştır ve 2005

yılında Özmutlu ve arkadaşları tarafından Bursa Emniyet Müdürlüğü ile bilişim suçlarını belirlemek amacıyla ortak bir proje kapsamında uygulamaya geçirilmiştir. Bu çalışmanın önceki çalışmalardan farklılığı ise sayfaların ilgililiğini yapay sinir ağları kullanılarak değerlendirmesi ve geliştirilen dinamik versiyon ile uzman kişi bilgisini karar mekanizmasına katmasıdır.

Sayfaların ilgililik değerlendirmesi için web sayfalarının metin içeriğini kullanılmaktadır. Yaklaşımın doğru değerlendirme yapabilmesi için çok kaliteli anahtar kelimeler kullanılması gerekir. Anahtar kelime seçiminde kuş gribi ile ilgili sayfalarda çok sıklıkla bulunan kelimelerin tespit edilmesi önemlidir. Bu değerlendirmede doküman yapısı, imajlar gibi diğer ilgili özellikleri kullanmamaktadır. İmaj gibi objelerin değerlendirilmesi imaj tanıma gibi çok kapsamlı bir çalışmayı da gerektireceğinden bu tez kapsamında ele alınmamıştır.

Derleyicinin çeşitli Web sayfalarına ulaşma yöntemi olarak, bu çalışmada temelde link tabanlı yaklaşım kullanılmaktadır. Buna göre bir dokümanı referans gösteren bir web sayfasının gerçekte onun içeriği hakkında yeterli izleri içermesi gerektiğine dayanan bir değerlendirme mantığı kullanmaktadır (Chakrabarti, 1999).

Geliştirilen derleyici, istenilen herhangi bir konuda çalışabilmektedir. İstenilen konuya ait belirtilen sayıda anahtar kelimenin ve başlangıç noktası olacak referans sitelerin uzman kişi tarafından belirlenmesi yeterli olmaktadır.

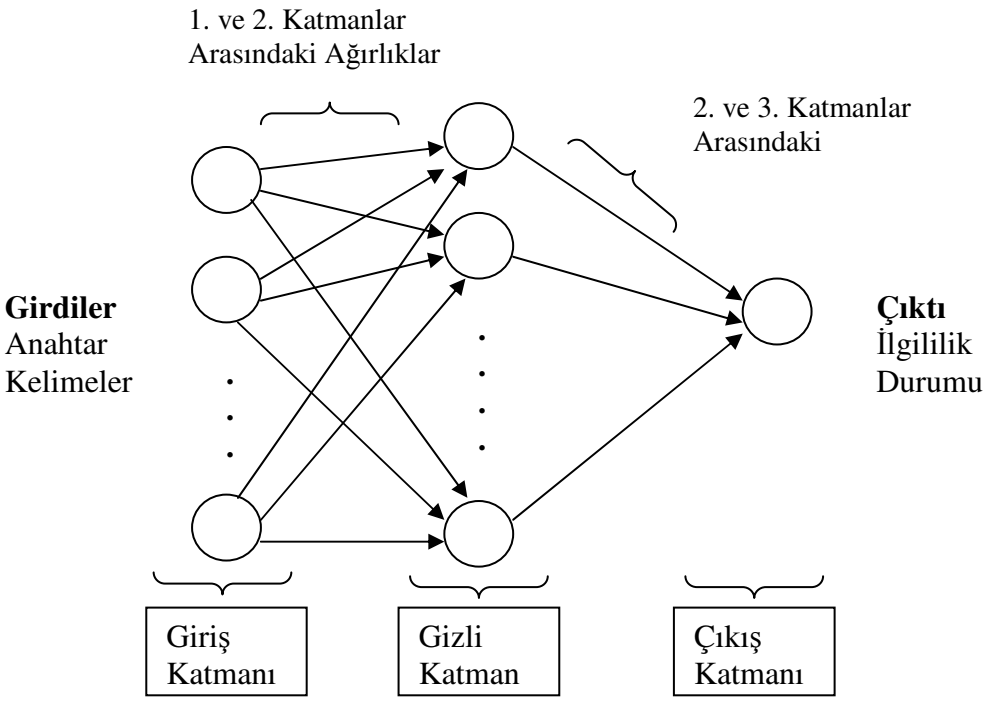
3.2.2 - Sistem Yapısı

Geliştirilen özelleştirilmiş derleyici, kuş gribi konusuyla ilgili Web'deki siteleri bulmayı amaçlamaktadır. Fakat daha önce de belirtildiği gibi, herhangi bir konuya ait anahtar kelimelerin ve başlangıç noktası olacak referans sitelerin uzman kişi tarafından belirlenmesi ile her konuya uygulanabilir.

Derleyici, başlangıçtaki verilen sitelerin linklerini belirmekte ve sırasıyla bütün linkleri taramakta ilgili bulduklarını kaydetmekte, ilgisiz sitelerinde adresi bilgi olarak tutulmaktadır. İkinci çevrimde ilgili bulunan sitelerle büyüyen referans sitelerin linkleri

tekrar belirlenmekte ve derleyici tekrar çalışarak yeni ilgili siteler belirlenmektedir. Bu döngü sistemi istenildiği kadar çalışmaya devam ettirilebilir.

Bu çalışmada sayfaların ilgililik değerlendiricisi olarak yapay sinir ağları kullanılmaktadır. İleri beslemeli, çok katmanlı ve bir danışmanlı öğrenme tekniği olan geri yayılım algoritmasını kullanan yapay sinir ağı kullanılmıştır. Ağdaki gizli katman sayısı birdir. Öğrenme katsayısı (η) 0.05 olarak alınmıştır. Kullanılan ağ yapısı Şekil 3. 21'de gösterilmektedir.



Şekil 3.21 : Kullanılan Yapay Sinir Ağı

- *Giriş Katmanı :* Ağın bu katmanında anahtar kelime sayısında nöron bulunmaktadır. Anahtar kelimenin ilgili sitede olup olamama durumuna göre girdi değer almaktadır. Anahtar kelime sitede yok ise değer sıfır olmakta, aksi halde ilgili kelimenin ağırlığı girdi olarak ağa verilmektedir. Yapılan deneylerde 10 ve 15 anahtar kelimeli yani giriş katmanı nöron sayılı YSA'ları kullanılmıştır.
- *Gizli Katman :* Bu katmandaki nöron sayısı deneylerde 15 veya 30 alınmıştır

- *Çıkış Katmanı :* Bu katmandaki nöron sayısı birdir. Değerlendirilen sitenin ilgili olup olmamasına göre bir sonuç üretir.

Derleyicinin karar verme mekanizmasını oluşturan YSA'nda geri yayılım algoritması kullanılmıştır. Bu eğitim algoritmasının adımları Bölüm 3.1.3.7.1'de detaylı olarak verilmiştir. Derleyicide amaçlanan doğru tahmin yapmaktır. Tahmin problemlerinde de çoğunlukla geri yayılım eğitim algoritması tercih edilmektedir.

YSA'da başlangıçta ağırlıklar rasgele atanmaktadır. Bu çalışmada rastgele ağırlıklar [0,1] aralığında verilmiştir. Veritabanında eğitimin kaç çevrimde tamamlandığı ve her bir çevrimin hata değeri tutulmaktadır.

Yapılan iki versiyonda Visual Basic 6.0 programlama dilinde yazılmıştır. Başlangıçtaki referans sitelerin ve daha sonra ilgili bulunan sitelerin dokümanı sabit diskte C:\ep\veriler\sayfalar adresinde tutulmaktadır. Ayrıca eğitimde kullanılan sitelerin metinleri de C:\ep\veriler\Egitim_Sayfalari adresinde kayıtlıdır. Programda kullanılan veriler vt isimli bir Access veri tabanı dosyasında tutulmaktadır. Bu dosyada kullanılan tablolar ve görevleri aşağıdaki gibidir ;

- *Agirlik_12* : Kullanılan yapay sinir ağının birinci ve ikinci katmanı arasındaki ağırlıklar Şekil 3.22'de gösterilen tabloda tutulmaktadır. Ağırlıkların aktif oldukları alan Şekil 3.21'de görülmektedir.

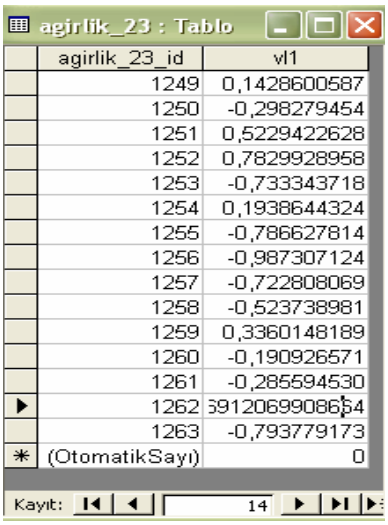
agirlik_12 : Tablo

agirlik_12_id	wj1	wj2	wj3	wj4	wj5	
833	0,439342627	-0,055556204	0,3293254770	-0,268562801	-0,577696681	0,6
834	0,975838487	0,7190299501	-0,853279251	1,0336668453	-0,366649029	0,1
835	0,697626358	0,6203990738	0,2687138926	1,1106084221	0,7847945879	-0,
836	-0,84341167	-0,893279403	0,6823299827	-0,402298204	-0,992429139	-0,
837	-0,18675455	-0,235480975	126601607393	-0,295522930	0,2169620418	-0,
838	0,886904094	0,2433268465	0,3231694525	-0,089272381	-0,908442621	0,1
839	0,032228212	-0,140910600	-0,227476329	-0,113094761	-0,565169052	-0,
840	0,280978056	-0,100352613	-0,48056738	1,1850161653	0,0840816186	0,1
841	-0,8936332	-0,893599811	-0,273325226	-0,695377430	-1,132517181	0,
842	0,433580696	0,839943002	0,1439955504	-0,77700112	0,4625085144	-0,
(OtomatikSayı)	0	0	0	0	0	

Kayıt: 5 / 10

Şekil 3.22 : 1. ve 2. Katmanlar Arasındaki Ağırlıkların Tablosu

- *Agirlik_23* **:** Kullanılan yapay sinir ağının birinci ve ikinci katmanı arasındaki ağırlıklar Şekil 3.23'de gösterilen tabloda tutulmaktadır. Ağırlıkların aktif oldukları alan Şekil 3.21'de görülmektedir.



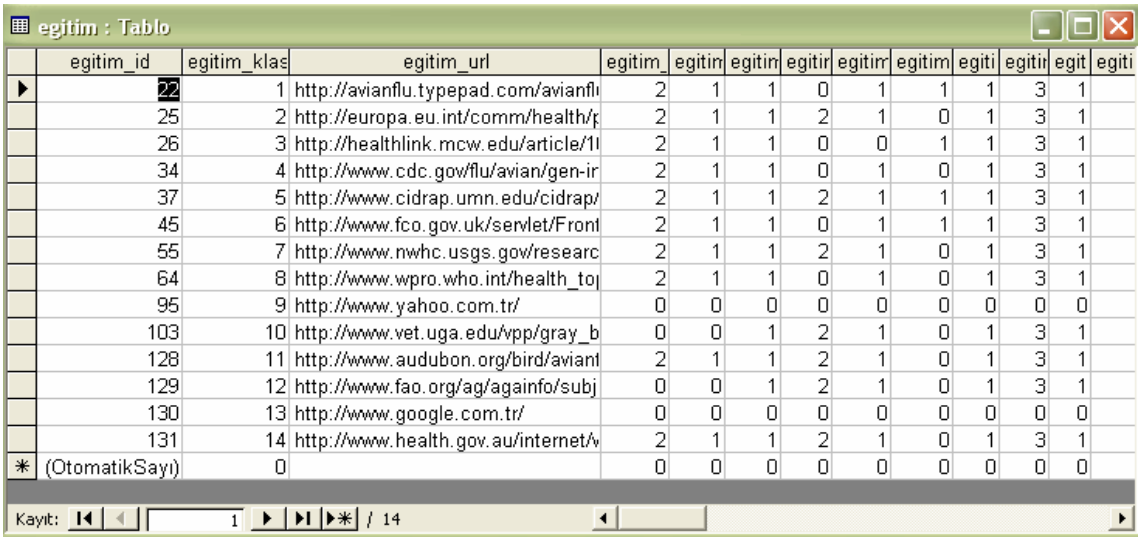
agirlik_23 : Tablo

agirlik_23_id	vl1
1249	0,1428600587
1250	-0,298279454
1251	0,5229422628
1252	0,7829928958
1253	-0,733343718
1254	0,1938644324
1255	-0,786627814
1256	-0,987307124
1257	-0,722808069
1258	-0,523738981
1259	0,3360148189
1260	-0,190926571
1261	-0,285594530
1262	391206990865 4
1263	-0,793779173
(OtomatikSayı)	0

Kayıt: 14

Şekil 3.23 : 2. ve 3. Katmanlar Arasındaki Ağırlıkların Tablosu

- *Egitim :* Eğitimde kullanılan siteler, anahtar kelimenin ilgili sitelerde olup olamama durumuna göre girdi olarak kullanılacak ağırlıklar ve çıktı olarak kullanılacak ilgililik durumu Şekil 3.24'de gösterilen tabloda bulunmaktadır. Eğitimde ilgisizlik durumunu ağa öğretmek için sitelerin yaklaşık %15'ini ilgisiz sitelerden seçilmiştir.



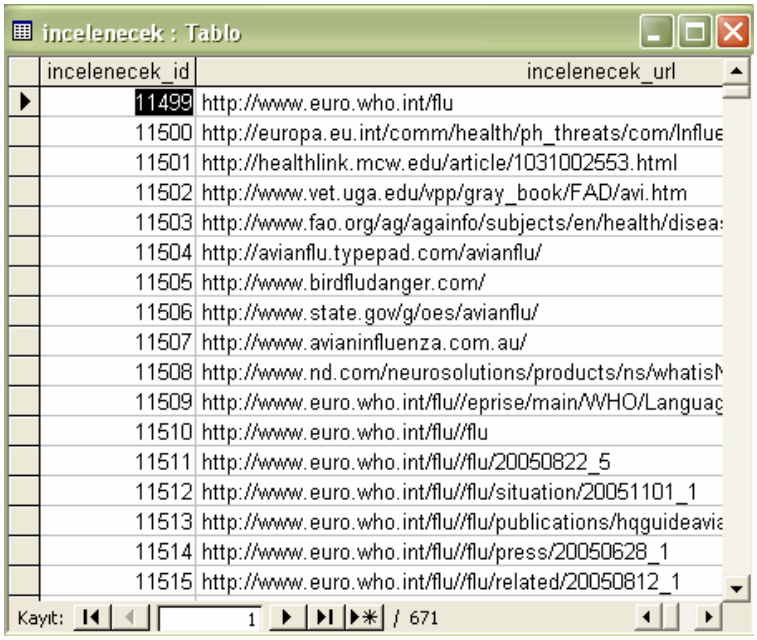
egitim : Tablo

egitim_id	egitim_klas	egitim_url	egitim_	egitim	egitim	egitim	egitim	egitim	egiti	egitim	egit	egiti
22	1	http://avianflu.typepad.com/avianfl	2	1	1	0	1	1	1	3	1	
25	2	http://europa.eu.int/comm/health/p	2	1	1	2	1	0	1	3	1	
26	3	http://healthlink.mcw.edu/article/1l	2	1	1	0	0	1	1	3	1	
34	4	http://www.cdc.gov/flu/avian/gen-ir	2	1	1	0	1	0	1	3	1	
37	5	http://www.cidrap.umn.edu/cidrap/	2	1	1	2	1	1	1	3	1	
45	6	http://www.fco.gov.uk/servlet/Front	2	1	1	0	1	1	1	3	1	
55	7	http://www.nwhc.usgs.gov/researc	2	1	1	2	1	0	1	3	1	
64	8	http://www.wpro.who.int/health_to	2	1	1	0	1	0	1	3	1	
95	9	http://www.yahoo.com.tr/	0	0	0	0	0	0	0	0	0	
103	10	http://www.vet.uga.edu/vpp/gray_b	0	0	1	2	1	0	1	3	1	
128	11	http://www.audubon.org/bird/aviant	2	1	1	2	1	0	1	3	1	
129	12	http://www.fao.org/ag/againfo/subj	0	0	1	2	1	0	1	3	1	
130	13	http://www.google.com.tr/	0	0	0	0	0	0	0	0	0	
131	14	http://www.health.gov.au/internet/v	2	1	1	2	1	0	1	3	1	
(OtomatikSayı)	0		0	0	0	0	0	0	0	0	0	

Kayıt: 1 / 14

Şekil 3.24 : Eğitim Tablosu

- *İncelenecek :* Derleyici çalışırken tarayacağı sitelerin adreslerini Şekil3.25'de gösterilen tablodan almaktadır.



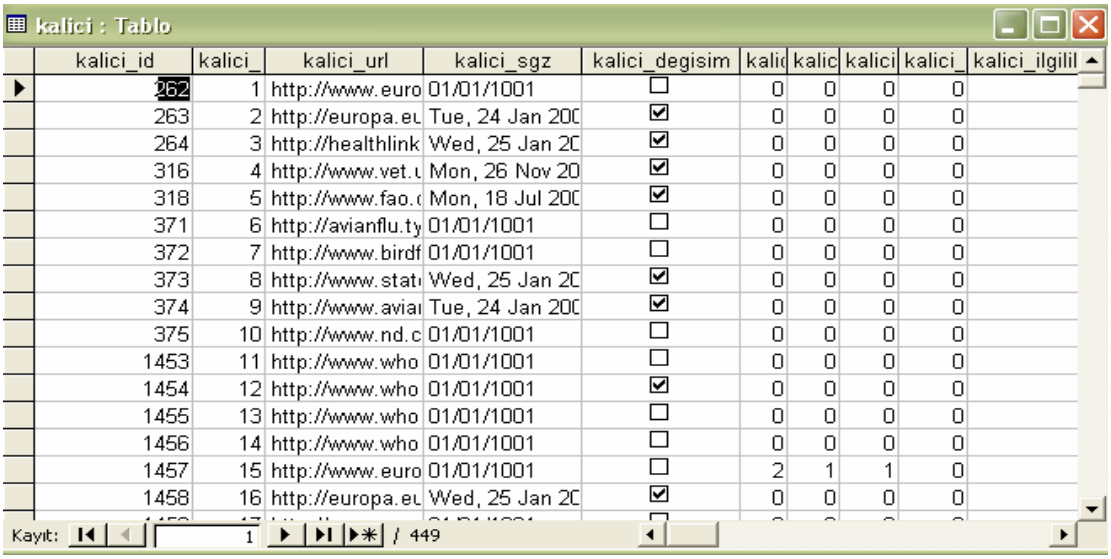
incelenecek : Tablo

incelenecek_id	incelenecek_url
11499	http://www.euro.who.int/flu
11500	http://europa.eu.int/comm/health/ph_threats/com/Influe
11501	http://healthlink.mcw.edu/article/1031002553.html
11502	http://www.vet.uga.edu/vpp/gray_book/FAD/avi.htm
11503	http://www.fao.org/ag/againfo/subjects/en/health/disea
11504	http://avianflu.typepad.com/avianflu/
11505	http://www.birdfludanger.com/
11506	http://www.state.gov/g/oes/avianflu/
11507	http://www.avianinfluenza.com.au/
11508	http://www.nd.com/neurosolutions/products/ns/whatisN
11509	http://www.euro.who.int/flu//eprise/main/WHO/Languag
11510	http://www.euro.who.int/flu//flu
11511	http://www.euro.who.int/flu//flu/20050822_5
11512	http://www.euro.who.int/flu//flu/situation/20051101_1
11513	http://www.euro.who.int/flu//flu/publications/hqguideavia
11514	http://www.euro.who.int/flu//flu/press/20050628_1
11515	http://www.euro.who.int/flu//flu/related/20050812_1

Kayıt: 1 / 671

Şekil 3.25 : İncelenecek Tablosu

- *Kalici :* Başlangıçtaki ilgili referans siteler ve daha sonra bulunan ilgili siteler Şekil 3.26'da gösterilen tabloya kaydedilmektedir.



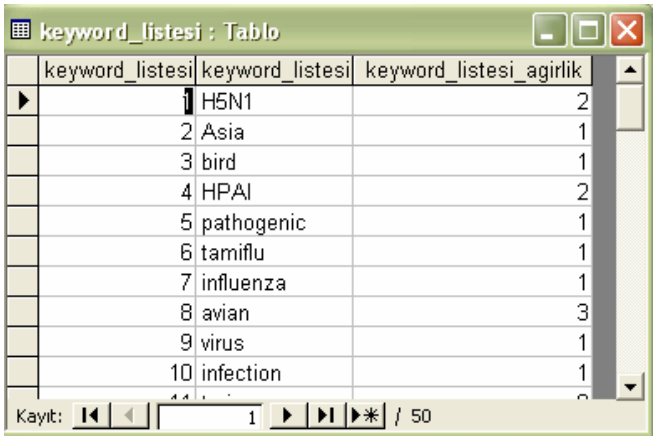
kalici : Tablo

kalici_id	kalici_	kalici_url	kalici_sgz	kalici_degisim	kalic	kalic	kalici	kalici_	kalici_ilgilil
262	1	http://www.euro	01/01/1001	☐	0	0	0	0	
263	2	http://europa.eu	Tue, 24 Jan 200	☑	0	0	0	0	
264	3	http://healthlink	Wed, 25 Jan 20	☑	0	0	0	0	
316	4	http://www.vet.u	Mon, 26 Nov 20	☑	0	0	0	0	
318	5	http://www.fao.(	Mon, 18 Jul 200	☑	0	0	0	0	
371	6	http://avianflu.ty	01/01/1001	☐	0	0	0	0	
372	7	http://www.birdf	01/01/1001	☐	0	0	0	0	
373	8	http://www.stat	Wed, 25 Jan 20	☑	0	0	0	0	
374	9	http://www.aviar	Tue, 24 Jan 200	☑	0	0	0	0	
375	10	http://www.nd.c	01/01/1001	☐	0	0	0	0	
1453	11	http://www.who	01/01/1001	☐	0	0	0	0	
1454	12	http://www.who	01/01/1001	☑	0	0	0	0	
1455	13	http://www.who	01/01/1001	☐	0	0	0	0	
1456	14	http://www.who	01/01/1001	☐	0	0	0	0	
1457	15	http://www.euro	01/01/1001	☐	2	1	1	0	
1458	16	http://europa.eu	Wed, 25 Jan 20	☑	0	0	0	0	

Kayıt: 1 / 449

Şekil 3.26 : Kalıcı Tablosu

- *Keyword_listesi :* Anahtar kelimeler ve ağırlıkları Şekil 3.27'de gösterilen tabloda tutulmaktadır.



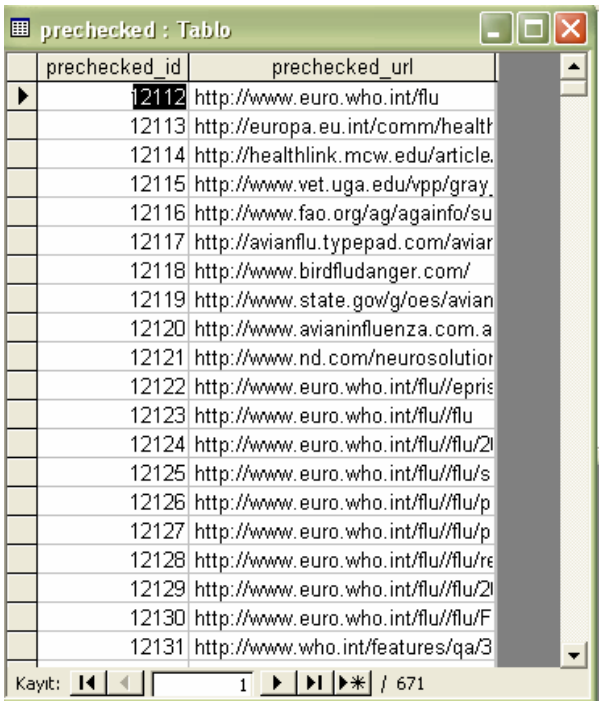
keyword_listesi : Tablo

keyword_listesi	keyword_listesi	keyword_listesi_agirlik
1	H5N1	2
2	Asia	1
3	bird	1
4	HPAI	2
5	pathogenic	1
6	tamiflu	1
7	influenza	1
8	avian	3
9	virus	1
10	infection	1

Kayıt: 1 / 50

Şekil 3.27 : Anahtar Kelime Tablosu

- Prechecked : İncelenen bütün sitelerin adresleri Şekil 3.28'de gösterilen tabloda kayıt altına alınmaktadır.



prechecked : Tablo

prechecked_id	prechecked_url
12112	http://www.euro.who.int/flu
12113	http://europa.eu.int/comm/healt
12114	http://healthlink.mcw.edu/article.
12115	http://www.vet.uga.edu/vpp/gray_
12116	http://www.fao.org/ag/againfo/su
12117	http://avianflu.typepad.com/aviar
12118	http://www.birdfludanger.com/
12119	http://www.state.gov/g/oes/avian
12120	http://www.avianinfluenza.com.a
12121	http://www.nd.com/neurosolutior
12122	http://www.euro.who.int/flu//epris
12123	http://www.euro.who.int/flu//flu
12124	http://www.euro.who.int/flu//flu/2
12125	http://www.euro.who.int/flu//flu/s
12126	http://www.euro.who.int/flu//flu/p
12127	http://www.euro.who.int/flu//flu/p
12128	http://www.euro.who.int/flu//flu/re
12129	http://www.euro.who.int/flu//flu/2
12130	http://www.euro.who.int/flu//flu/F
12131	http://www.who.int/features/qa/3

Kayıt: 1 / 671

Şekil 3.28 : Prechecked Tablosu

- *Ysa_degerlendirme :* Değerlendirilen bütün sitelerin adresleri ve YSA değerlendirme sonuçları Şekil 3.29'da gösterilen tabloya kaydedilmektedir.

ysa_degerlendirme : Tablo

ysa_degerlendi	ysa_degerlendi	ysa_degerlendi	ysa_degerlendi
242976	http://www.euro	1	0,997818435
242977	http://europa.eu	1	0,9988359341
242978	http://healthlink	1	0,9969729229
242979	http://www.vet.u	1	0,9999242413
242980	http://www.fao.(	1	0,9999242413
242981	http://avianflu.ty	1	0,9992301563
242982	http://www.birdf	1	0,9981275099
242983	http://www.stat	1	0,9982033658
242984	http://www.aviar	1	0,9991462611
242985	http://www.nd.c	0	0
242986	http://www.euro	0	0
242987	http://www.euro	0	0
242988	http://www.euro	0	0
242989	http://www.euro	0	0
242990	http://www.euro	0	0
242991	http://www.euro	0	0
242992	http://www.euro	0	0

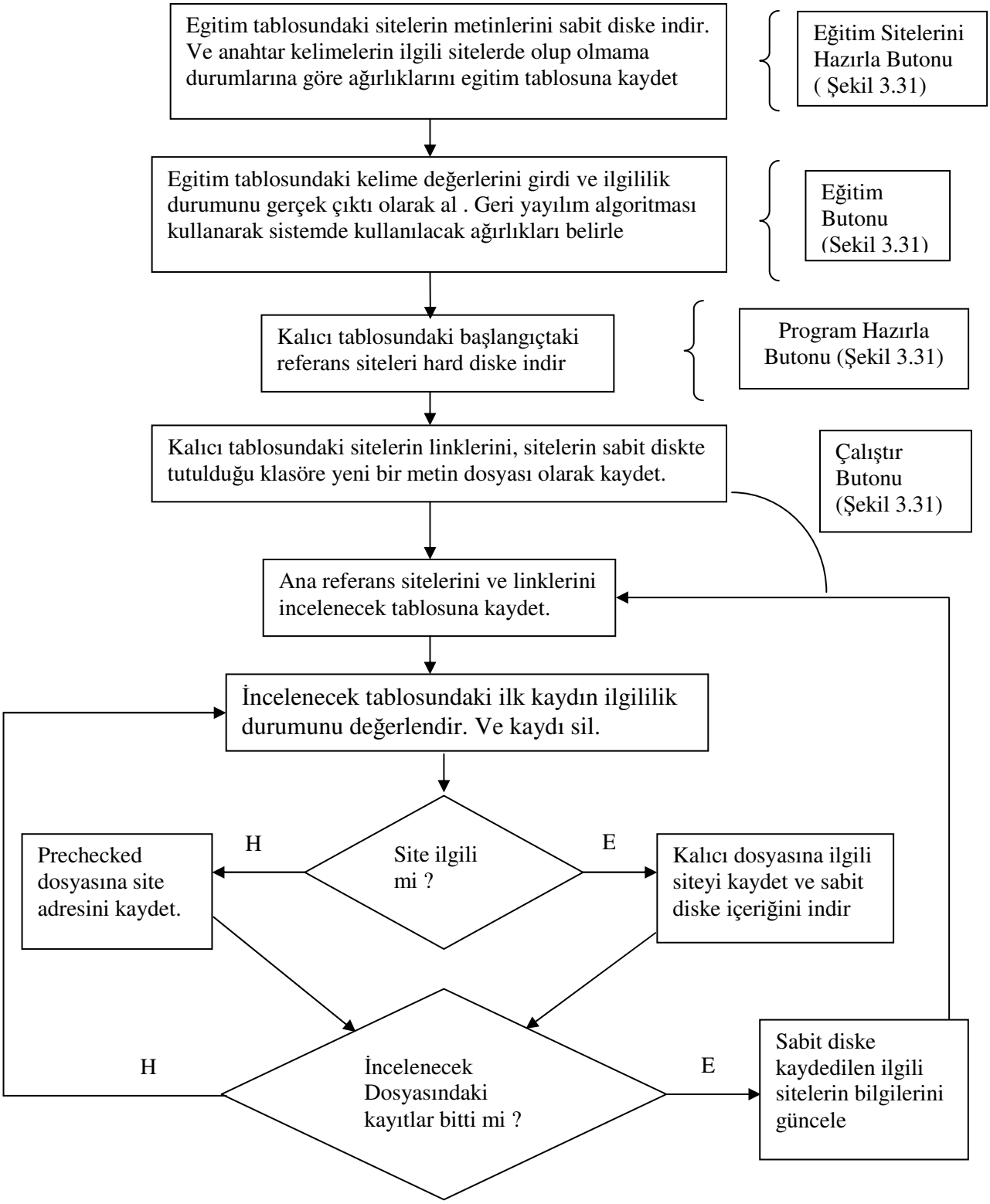
Kayıt: 15 / 671

Şekil 3.29 : Yapay Sinir Ağı Değerlendirme Tablosu

Sistemde derleyicinin incelediği her sitenin ilgililik durumu yapay sinir ağı sonucuna göre değerlendirilmektedir. Bu noktada yapay sinir ağının kararını belirleyen ağırlıklar yani ağın eğitimi önem arz etmektedir. Bu nedenle iki farkı versiyon derleyici geliştirilmiştir. İkinci dinamik versiyonda Chakrabarti (1999) tarafından önerilen yarı denetimli öğrenme tekniği uygulanmaktadır. Bu sistemde yapay sinir ağının eğitiminde kullanılan eğitim sitelerinin dinamik olarak uzman kişi tarafından güncellenmesi önerilmektedir. Ayrıca Chakrabarti (1999) tarafından anahtar kelime listesinin de uzman kişi tarafından güncellenmesi önerilmektedir. Bu çalışmada sadece ilk öneri olan derleyici çok ilgili bir site bulduğunda aramanın durması ve kullanıcıya arayüz vasıtasıyla ilgili sitenin eğitim sitesi olarak kullanılması ve istenen herhangi bir sitenin eğitim setinden çıkarılması yönündeki metot uygulanmıştır. Sırasıyla bu iki versiyon incelenirse:

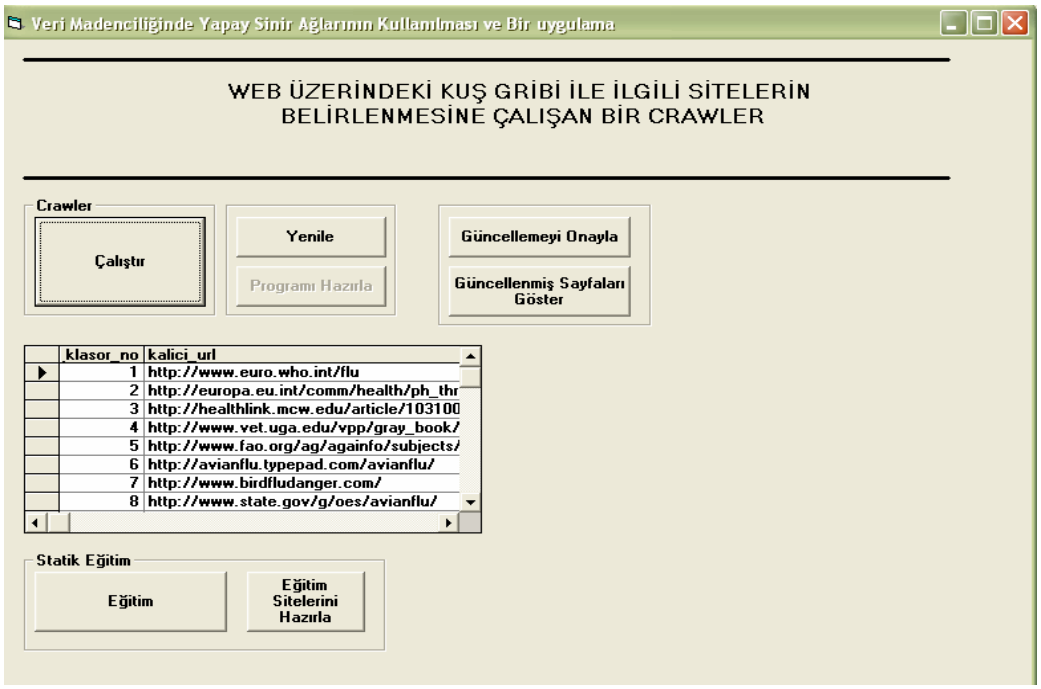
1. *Statik Versiyon :* Temel olarak, bu versiyonda eğitim sitelerine ve anahtar kelimelere göre sinir ağı program ilk çalıştırıldığında eğitilmektedir. İstenilirse daha sonra kullanıcı tarafından eğitim tekrarlanabilmektedir. Burada temel mantık eğitim

sayfaları setinin sabit kalmasıdır. İlgili bulunan siteler arayüzden görülebilmektedir.Statik versiyonun akışı Şekil 3.30'de görülmektedir.



Şekil 3. 30 : Statik Versiyon Akış Diyagramı

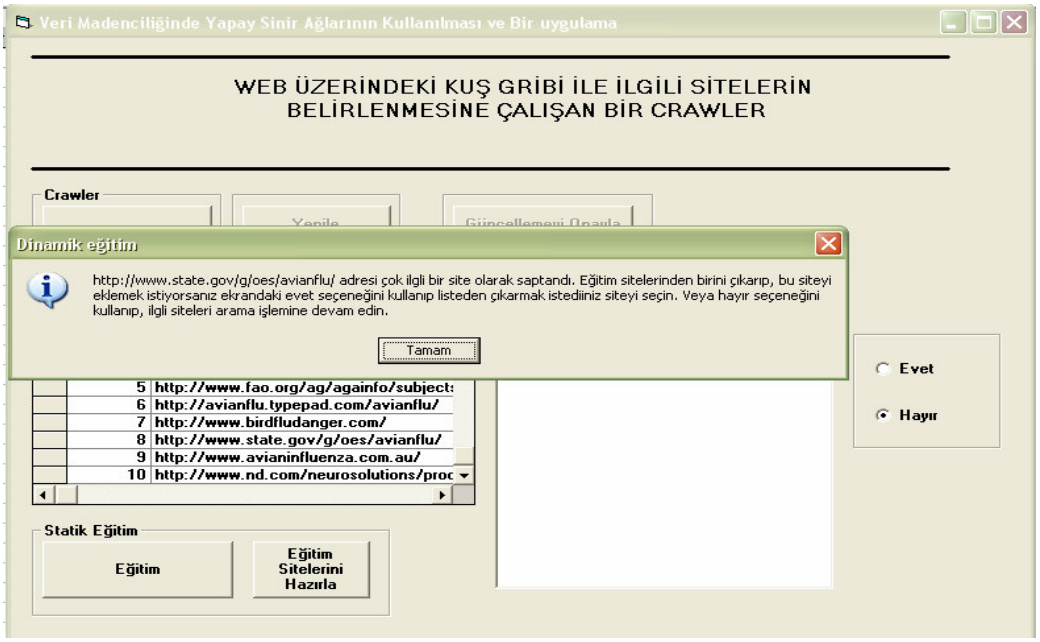
Bu versiyonda kullanılan arayüz Şekil 3.31'de gösterilmiştir.



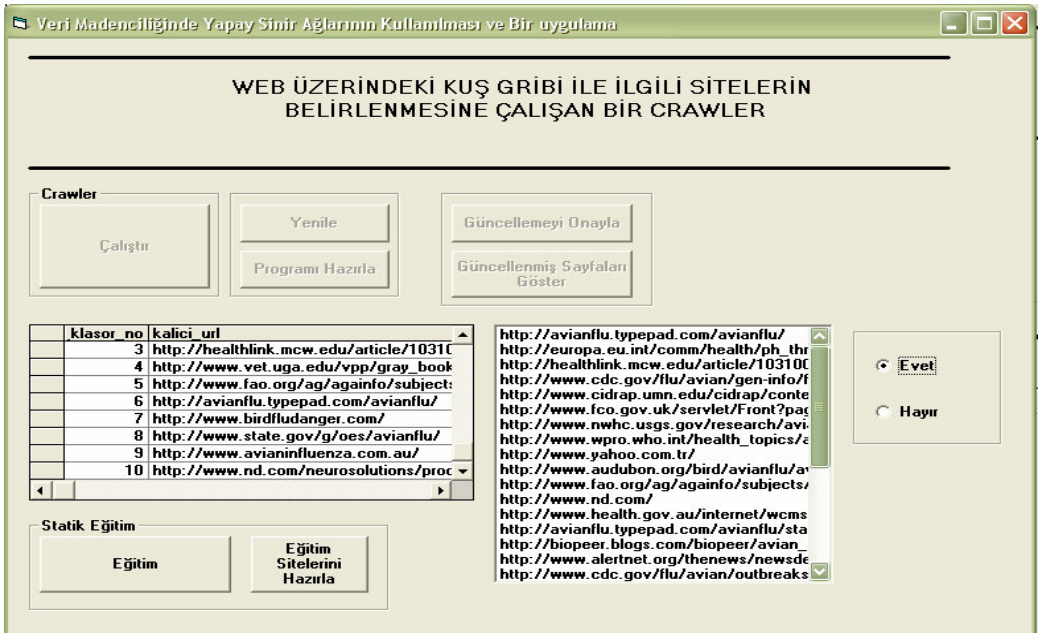
Şekil 3.31 . Statik Versiyon Kullanıcı Arayüzü

2. *Dinamik Versiyon :* Bu ikinci versiyonda eğitim dinamik şekilde yapılmaktadır. Burada temel mantık, sayfa değerlendirici çalışırken çok ilgili bir sayfa bulunursa, bu sayfanın eğitim sayfaları setine katılabilmesidir, yani eğitim sayfaları seti sabit değildir. Dinamik versiyon ile bir kere eğitim yerine uzman kullanıcı bilgisinin eklenmesi yoluyla eğitim listesinin sürekli güncellenmesi ve eğitim sayısını arttırıp yapay sinir ağının sonuçlarının doğruluğunu arttırmak amaçlanmıştır.

Statik versiyondaki gibi ilk program çalıştığında ağ eğitilmektedir. Çalıştır butonuna basılıp derleyici çalıştığında çok ilgili bir site bulursa aşağıdaki arayüz açılmakta ve Şekil 3.32'de görüldüğü gibi bulunan sitenin eğitim sitelerinden biri ile değiştirilmek istenip istenmediği sorulmaktadır. Eğer kullanıcı evet seçeneğini işaretlerse Şekil 3.33'de görülen arayüze eğitimde kullanılan sitelerin listesi gelmektedir Ve buradan çıkarılmak istenen eğitim sitesi seçilmektedir. Yeni eğitim sitelerine göre eğitim yapılmakta ve derleyici sinir ağındaki yeni ağırlıklara göre incelenecek dosyasında kaldığı yerdeki kayıtdan çalışmaya devam etmektedir.



Şekil 3.32 : Dinamik Versiyon Kullanıcı 1. Arayüzü



Şekil 3.33 : Dinamik Versiyon Kullanıcı 2. Arayüzü

3.2.3 - Uygulama Çalışması

Geliştirilen özelleşmiş derleyicinin başarılı ve kullanılabilir bir çalışma olup olmadığına karar vermek için deneyler yapılmıştır. Sonuçları karşılaştırabilmek için

derleyicinin değerlendireceği siteler sabitlenmiş ve linklerle beraber toplam 570 ile sınırlandırılmıştır. Sonuçları karşılaştırabilmek için öncelikle bu sitelerin hepsinin gerçekteki ilgililik durumu belirlenmiştir.

Sistemin verimliliğini etkileyen ana parametreleri bulmak ve kullanılan yapay sinir ağının doğru nöron sayılarını saptamak için deneylerde aşağıdaki faktörler kullanılmıştır:

A. *Kullanılan versiyonlar:* Bu sayede dinamik ve statik versiyonların sonuçlarının karşılaştırılması amaçlanmaktadır (S / D)
B. *Anahtar kelime sayısı:* 10 ve 15 kelimeli iki alternatif oluşturulmuştur. Anahtar kelime sayısındaki değişim yapay sinir ağının giriş katmanındaki nöron sayısını değiştirmektedir.
C. *Eğitimde kullanılan site sayısı:* 14 ve 30 adetli iki alternatif eğitim seti seçilmiştir. Dinamik versiyonda kullanılan siteler değişmektedir. Fakat kullanıcı tarafından bir site eklenirken bir site de listeden çıkarılmaktadır, dolayısıyla toplam site sayısı sabit kalmaktadır.
D. *Yapay sinir ağının gizli katman nöron sayısı:* 15 ve 30 olarak iki alternatif değerlendirilmiştir.
E. *Yapay sinir ağını değerlendirmekte kullanılan eşik değeri:* 0,75 ve 0,50 değerleri kullanılmıştır.

Faktörler deney sürelerinin uzunluğundan dolayı 2 ile sınırlandırılmıştır. Faktör sayısının 5 ve her bir faktörün seviyesinin 2 olması nedeniyle toplam 32 (2^5) deney yapılmıştır.

Çalışmada kullanılan gösterimler aşağıdaki gibidir (Ozmutlu ve Cavdur, 2005):

- N_{ilgili} : Derleyici tarafından konu ile ilgili bulunan site sayısı.
- $N_{ilgisiz}$: Derleyici tarafından konu ile ilgisiz bulunan site sayısı.
- N_{gercek_ilgili} : İnsanlar tarafından konu ile ilgili bulunan site sayısı.
- $N_{gercek\_ilgisiz}$: İnsanlar tarafından konu ile ilgisiz bulunan site sayısı
- N_{dogru_ilgili} : İnsanlar ve derleyici tarafından konu ile ilgili bulunan site sayısı.

- $N_{dogru\_ilgisiz}$: İnsanlar ve derleyici tarafından konu ile ilgisiz bulunan site sayısı.
- A Tipi Hata : Derleyici tarafından gerçekte konu ile ilgili olup, ilgisiz bulunan site sayısı.
- B Tipi Hata : Derleyici tarafından gerçekte konu ile ilgisiz olup, ilgili bulunan site sayısı.

Yukarıdaki ifadelerle ilgili en çok kullanılan eşitlikler aşağıdaki gibidir;

$$N_{gercek\_ilgili} = N_{dogru\_ilgili} + \text{A Tipi Hata} \quad (3.15)$$

$$N_{gercek\_ilgisiz} = N_{dogru\_ilgisiz} + \text{B Tipi Hata} \quad (3.16)$$

$$N_{ilgili} = N_{dogru\_ilgili} + \text{B Tipi Hata} \quad (3.17)$$

$$N_{ilgisiz} = N_{dogru\_ilgisiz} + \text{A Tipi Hata} \quad (3.18)$$

Bu çalışmada geliştirilen derleyicinin performansını belirlemek için temelde iki performans ölçütü kullanılmaktadır. Bunlar, duyarlık (precision) *P* ve anma (recall) *R* performans ölçütleridir. Kullanılan bütün performans ölçütleri aşağıdaki gibidir (Ozmutlu ve Cavdur, 2005):

- $P_{ilgili} = \dfrac{N_{dogru\_ilgili}}{N_{ilgili}}$ (3.19)
- $P_{ilgisiz} = \dfrac{N_{dogru\_ilgisiz}}{N_{ilgisiz}}$ (3.20)
- $R_{ilgili} = \dfrac{N_{dogru\_ilgili}}{N_{gercek\_ilgili}}$ (3.21)
- $R_{ilgisiz} = \dfrac{N_{dogru\_ilgisiz}}{N_{gercek\_ilgisiz}}$ (3.22)
- A tipi hata
- B tipi hata

4 – ARAŞTIRMA SONUÇLARI

Çizelge 4.1'de deney sonucundaki ilgililik verileri ve buna göre performans ölçütlerinin değerleri görülmektedir.

Çizelgedeki deney sonuçlarına göre geliştirilen derleyici başarılı sonuçlar vermiştir. Belirtilen performans ölçütlerinden P_{ilgili} ve R_{ilgili}'nin değerlerine göre sistem ilgili siteleri bulma konusunda %89 ile %97 arasında başarı göstermiştir. Aynı şekilde $P_{ilgisiz}$ ve $R_{ilgisiz}$'nin değerlerine göre sistem ilgisiz siteleri bulma konusunda da %44 ile %88 arasında doğru sonuçlar bulmuştur. A tipi ve B tipi hata sayıları da, değerlendirilen site sayısının 570 olduğu düşünüldüğünde kabul edilebilir seviyededir.

İlgisiz sitelerin tespitindeki düşük başarı oranında, yapay sinir ağının değerlendirmede kullandığı anahtar kelimelerin 10'dan 15'e çıkarılmasının önemli bir etkisi olmuştur. Özellikle $R_{ilgisiz}$ oranındaki düşüş ve B tipi hatadaki artış, ilgisiz sitelerin belirlenmesinde nispeten başarısız sonuçlar oluştuğunu göstermektedir. Bu duruma göre az fakat özellikli kelime ile sistemin daha verimli çalıştığı sonucuna varılabilir.

Hangi faktörlerin sistem üzerinde daha fazla etkili olduğu ve değişime neden olduğunu anlamak için geliştirilen hipotezlere ortalamaların karşılaştırılması için t-testi uygulanmaktadır (Özmutlu ve Cavdur, 2005).

Bu istatistiksel test yönteminde,

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} \tag{3.23}$$

$\bar{d}$: İki örnek grubu arasındaki ortalama farkları

s_d : İki örnek grubu arasındaki standart sapma farkı

n : Örnek büyüklüğü

Bu t testi, bir örneğin ortalamalarının ayrık olup olmadığını saptamak için, eşli iki örnekli gruba uygulanır. Her iki popülasyon varyanslarının eşit olduğu varsayılmaz. Bir çiftli sınama, örneklerdeki gözlemlerde doğal bir eşleme olduğunda, örneğin, bir örnek grup (deneyden önce ve sonra) iki kere sınandığında, kullanılabilir.

Bu çalışmada 5 faktör ve 6 performans ölçütü için toplam 30 tane ortalamaların karşılaştırılması için t-testi yapılmıştır. Ortalamaların farklılığı ve diğer faktörler sabit iken ilgili faktörün sistemde değişiklik oluşturup oluşturmadığının tespiti amaçlanmaktadır.

%95 güvenilirlikte t değeri [-2,1314 , 2,1314] aralığında olduğu zaman H0 hipotezi kabul edilir. Aksi durumda H0 hipotezi reddedilir. t testi sonuçları Çizelge 4.2'de görülmektedir.

Deney No	A	B	C	D	E	Ngercek_ilgili	Ngercek_ilgisiz	Nilgili	Nilgisiz	Ndogru_ilgili	Ndogru_ilgisiz	A Tipi Hata	B Tipi Hata	Pilgili	Pilgisiz	Rilgili	Rilgisiz
1	S	10	14	15	0,75	480	90	456	114	446	80	34	10	0,978	0,702	0,929	0,889
2	S	10	14	30	0,75	480	90	469	101	456	77	24	13	0,972	0,762	0,950	0,856
3	S	10	30	15	0,75	480	90	457	113	444	77	36	13	0,972	0,681	0,925	0,856
4	S	10	30	30	0,75	480	90	467	103	453	76	27	14	0,970	0,738	0,944	0,844
5	S	15	14	15	0,75	480	90	466	104	431	55	49	35	0,925	0,529	0,898	0,611
6	S	15	14	30	0,75	480	90	486	84	460	64	20	26	0,947	0,762	0,958	0,711
7	S	15	30	15	0,75	480	90	486	84	447	51	33	39	0,920	0,607	0,931	0,567
8	S	15	30	30	0,75	480	90	489	81	460	61	20	29	0,941	0,753	0,958	0,678
9	S	10	14	15	0,50	480	90	461	109	449	78	31	12	0,974	0,716	0,935	0,867
10	S	10	14	30	0,50	480	90	504	66	468	54	12	36	0,929	0,818	0,975	0,600
11	S	10	30	15	0,50	480	90	461	109	446	75	34	15	0,967	0,688	0,929	0,833
12	S	10	30	30	0,50	480	90	501	69	465	54	15	36	0,928	0,783	0,969	0,600
13	S	15	14	15	0,50	480	90	485	85	448	53	32	37	0,924	0,624	0,933	0,589
14	S	15	14	30	0,50	480	90	503	67	465	52	15	38	0,924	0,776	0,969	0,578
15	S	15	30	15	0,50	480	90	498	72	457	49	23	41	0,918	0,681	0,952	0,544
16	S	15	30	30	0,50	480	90	506	64	465	49	15	41	0,919	0,766	0,969	0,544
17	D	10	14	15	0,75	480	90	456	114	441	75	39	15	0,967	0,658	0,919	0,833
18	D	10	14	30	0,75	480	90	459	111	449	80	31	10	0,978	0,721	0,935	0,889
19	D	10	30	15	0,75	480	90	446	124	431	75	49	15	0,966	0,605	0,898	0,833
20	D	10	30	30	0,75	480	90	406	164	393	77	87	13	0,968	0,470	0,819	0,856
21	D	15	14	15	0,75	480	90	429	141	402	63	78	27	0,937	0,447	0,838	0,700
22	D	15	14	30	0,75	480	90	484	86	457	63	23	27	0,944	0,733	0,952	0,700
23	D	15	30	15	0,75	480	90	480	90	446	56	34	34	0,929	0,622	0,929	0,622
24	D	15	30	30	0,75	480	90	484	86	457	63	23	27	0,944	0,733	0,952	0,700
25	D	10	14	15	0,50	480	90	461	109	444	73	36	17	0,963	0,670	0,925	0,811
26	D	10	14	30	0,50	480	90	475	95	453	68	27	22	0,954	0,716	0,944	0,756
27	D	10	30	15	0,50	480	90	448	122	432	74	48	16	0,964	0,607	0,900	0,822
28	D	10	30	30	0,50	480	90	437	133	424	77	56	13	0,970	0,579	0,883	0,856
29	D	15	14	15	0,50	480	90	477	93	450	63	30	27	0,943	0,677	0,938	0,700
30	D	15	14	30	0,50	480	90	501	69	462	51	18	39	0,922	0,739	0,963	0,567
31	D	15	30	15	0,50	480	90	489	81	454	55	26	35	0,928	0,679	0,946	0,611
32	D	15	30	30	0,50	480	90	499	71	462	53	18	37	0,926	0,746	0,963	0,589

Çizelge 4.1 : Deney Sonuçları ve Performans Ölçütlerinin Değerleri

Çizelge 4.2 : t Testi Sonuçları

Faktör	Hipotez Testi	İstatistik	Sonuç
A	$H0 : D_{Ahata,\ ststik,dinamik} = 0$ $H1 : D_{Ahata,\ ststik,dinamik} \neq 0$	t= -3,0025	İstatistiksel olarak H0 hipotezi geçersizdir..
	$H0 : D_{Bhata,\ ststik,dinamik} = 0$ $H1 : D_{Bhata,\ ststik,dinamik} \neq 0$	t= 2,0824	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Pilgili,\ ststik,dinamik} = 0$ $H1 : D_{Pilgili,\ ststik,dinamik} \neq 0$	t= -1,7704	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Pilgisiz,\ ststik,dinamik} = 0$ $H1 : D_{Pilgisiz,\ ststik,dinamik} \neq 0$	t= 3,1001	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgili,\ ststik,dinamik} = 0$ $H1 : D_{Rilgili,\ ststik,dinamik} \neq 0$	t= 3,0025	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgisiz,\ ststik,dinamik} = 0$ $H1 : D_{Rilgisiz,\ ststik,dinamik} \neq 0$	t= -2,0824	İstatistiksel olarak H0 hipotezi geçerlidir.
B	$H0 : D_{Ahata,\ 10kelime,15kelime} = 0$ $H1 : D_{Ahata,\ 10kelime,15kelime} \neq 0$	t= 1,4578	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Bhata,\ 10kelime,15kelime} = 0$ $H1 : D_{Bhata,\ 10kelime,15kelime} \neq 0$	t= -9,1043	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Pilgili,\ 10kelime,15kelime} = 0$ $H1 : D_{Pilgili,\ 10kelime,15kelime} \neq 0$	t= 9,0189	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Pilgisiz,\ 10kelime,15kelime} = 0$ $H1 : D_{Pilgisiz,\ 10kelime,15kelime} \neq 0$	t= 0,0853	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Rilgili,\ 10kelime,15kelime} = 0$ $H1 : D_{Rilgili,\ 10kelime,15kelime} \neq 0$	t= -1,4578	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Rilgisiz,\ 10kelime,15kelime} = 0$ $H1 : D_{Rilgisiz,\ 10kelime,15kelime} \neq 0$	t= 9,1043	İstatistiksel olarak H0 hipotezi geçersizdir.
C	$H0 : D_{Ahata,\ 14egitim,30egitim} = 0$ $H1 : D_{Ahata,\ 14egitim,30egitim} \neq 0$	t= -0,5440	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Bhata,\ 14egitim,30egitim} = 0$ $H1 : D_{Bhata,\ 14egitim,30egitim} \neq 0$	t= -1,7184	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Pilgili,\ 14egitim,30egitim} = 0$ $H1 : D_{Pilgili,\ 14egitim,30egitim} \neq 0$	t= 1,8031	İstatistiksel olarak H0 hipotezi geçerlidir.

Çizelge 4.2 : t Testi Sonuçları (devamı)

	$H0 : D_{Pilgisiz,\ 14egitim,30egitim} = 0$ $H1 : D_{Pilgisiz,\ 14egitim,30egitim} \neq 0$	t= 0,8507	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Rilgili,\ 14egitim,30egitim} = 0$ $H1 : D_{Rilgili,\ 14egitim,30egitim} \neq 0$	t= 0,5440	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Rilgisiz,\ 14egitim,30egitim} = 0$ $H1 : D_{Rilgisiz\ 14egitim,30egitim} \neq 0$	t= 1,7184	İstatistiksel olarak H0 hipotezi geçerlidir.
D	$H0 : D_{Ahata,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Ahata,\ 15aranöron,30aranöron} \neq 0$	t= 2,4244	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Bhata,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Bhata,\ 15aranöron,30aranöron} \neq 0$	t= -0,8537	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Pilgili,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Pilgili,\ 15aranöron,30aranöron} \neq 0$	t= 0,5227	İstatistiksel olarak H0 hipotezi geçerlidir.
	$H0 : D_{Pilgisiz,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Pilgisiz,\ 15aranöron,30aranöron} \neq 0$	t= -3,6686	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgili,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Rilgili,\ 15aranöron,30aranöron} \neq 0$	t= -2,4244	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgisiz,\ 15aranöron,30aranöron} = 0$ $H1 : D_{Rilgisiz,\ 15aranöron,30aranöron} \neq 0$	t= 0,8537	İstatistiksel olarak H0 hipotezi geçerlidir.
E	$H0 : D_{Ahata,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Ahata,\ 0,75\ eşik,0,50eşik} \neq 0$	t= 3,4393	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Bhata,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Bhata,\ 0,75\ eşik,0,50eşik} \neq 0$	t= -3,7441	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Pilgili,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Pilgili,\ 0,75\ eşik,0,50eşik} \neq 0$	t= 3,3132	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Pilgisiz,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Pilgisiz,\ 0,75\ eşik,0,50eşik} \neq 0$	t= -3,0780	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgili,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Rilgili,\ 0,75\ eşik,0,50eşik} \neq 0$	t= -3,4393	İstatistiksel olarak H0 hipotezi geçersizdir.
	$H0 : D_{Rilgisiz,\ 0,75\ eşik,0,50eşik} = 0$ $H1 : D_{Rilgisiz,\ 0,75\ eşik,0,50eşik} \neq 0$	t= 3,7441	İstatistiksel olarak H0 hipotezi geçersizdir.

5 – TARTIŞMA

Bu tez çalışmasında, genel arama motorlarına alternatif olarak, kullanım amacına göre istenilen herhangi bir spesifik konuda ilgili siteleri bulan bir derleyici yapısı geliştirilmiştir. Bu tip arama motorlarının geliştirilmesi, veri madenciliği tekniklerinin Web verileri üzerinde uygulanmasını konu alan Web madenciliğinin bir dalı olan Web içerik madenciliği kapsamına girmektedir. ve "focused crawling" olarak adlandırılmaktadır.

Derleyici taradığı sitelerin ilgili olup olmadıklarına bir veri madenciliği tekniği olan yapay sinir ağı kullanarak karar vermektedir. Burada, Özmutlu ve arkadaşlarının ASIS&T (The Information Society for the Information Age) 2004'de sundukları "Güvenlik konusunda geliştirilmiş özel bir web derleyicisi" konulu çalışma referans alınmıştır.

Özmutlu ve arkadaşlarının(2004) çalışmasında derleyicinin Web'de dolaşırken hangi sayfaları değerlendireceği noktasında farklı teknikler önerilmektedir. Bunlar link tabanlı yaklaşım, IP tabanlı yaklaşım, anahtar deyim tabanlı yaklaşım ve ilgi alanı (domain) tabanlı yaklaşımdır. Bu çalışmada link tabanlı yaklaşım kullanılmıştır. İlerideki çalışmalarda diğer tekniklere göre de Web'de dolaşan derleyiciler geliştirilip, sonuçları karşılaştırılabilir. Özmutlu ve arkadaşları(2004), derleyicinin dört farklı bilgisayarda ve dört farklı mantıkta da Web'de dolaşması ve kaynaklardan alınan bütün adreslerin tek bir veri tabanında birleştirilip değerlendirilmesine yönelik bir sistem önermektedirler. Bu tez çalışması bu kapsamlı sistemin kısmi bir parçasıdır.

Yapay sinir ağının öğrenme mekanizmasına göre iki farkı versiyon derleyici geliştirilmiştir. Bunlar statik ve dinamik versiyonlardır. Dinamik versiyonda Chakrabarti (1999) tarafından önerilen yarı denetimli öğrenme tekniği uygulanmaktadır. Bu sistemde yapay sinir ağının eğitiminde kullanılan eğitim sitelerinin dinamik olarak uzman kişi tarafından güncellenmesi önerilmektedir. Ayrıca Chakrabarti (1999) tarafından anahtar kelime listesinin de uzman kişi tarafından güncellenmesi önerilmektedir. Buna göre derleyici çok ilgili bir site bulduğunda anahtar kelimeler içerinde olmayan fakat yeni sitede çok kullanılan bir kelimeyi kullanıcıya anahtar kelime listesine eklemesi için önermektedir. Bu sayede anahtar kelime listesinin

dinamikliği sağlanarak eğitimden daha iyi sonuçlar alınması amaçlanmaktadır. İlerideki çalışmalarda Chakrabarti (1999) tarafından önerilen anahtar kelimelerin dinamikliğine dayanan sistem geliştirilebilir.

Geliştirilen derleyicinin sonuçlarını değerlendirmek için beş faktör ve her bir faktörün iki alternatif değerine göre 32 deney yapılmıştır. Deney sonuçları altı tane performans ölçütüne göre değerlendirilmiştir. Değerlendirilen altı faktörün seviyeleri arasındaki değişimde sistemin etkilenip etkilenmediğini anlamak için, ortalamalarının farkı paired t-testi ile değerlendirilmiştir. Beş faktör ve altı performans ölçütü için 30 tane paired t-testi yapılmıştır.

Testlerin sonucunda, sistem üzerinde en fazla değişiklik yapan faktörün anahtar kelime sayısı olduğu belirlenmiştir. Deneylere göre kelime sayısı 10'dan 15'e çıkarıldığında B tipi hata artmış ve $R_{ilgisiz}$ oranında büyük oranda düşüş görülmüştür. Sonuçta sistemin verimliliğini önemli oranda gerilemiştir. Bu durumdan yola çıkılarak anahtar kelimelerin sayısının gürültüyü azaltmak için kısıtlı miktarda tutulmasının gerektiği sonucuna varılabilir. Az fakat özellikli kelime ile sistem daha iyi sonuçlar vermektedir.

Diğer bir faktör olan eşik değerinin sonuçları değerlendirme noktasında kilit rol oynadığı ve değerinde yapılan değişikliğin sistemi önemli oranda etkilediği tespit edilmiştir.

Yapay sinir ağının gizli katmanında nöron sayısının 15 ve 30'lu iki alternatifinin ve eğitim deney setinin 14 ve 30'lu iki alternatifinin deney sonuçlarındaki performans ölçütlerinin ortalamalarında önemli bir değişikliğe neden olmadığı sonucuna varılmıştır.

Chakrabarti (1999), tarafından önerilen yarı denetimli öğrenme tekniği uygulanan dinamik versiyon sayesinde, statik versiyona göre büyük oranda bir gelişme kaydedilememiştir. Bu durumda kullanıcı tarafından eğitim listesine eklenen sitelerin çok doğru tercihler olamama durumu etkili olmuş olabilir.

KAYNAKLAR

Adriaans, P. ve D, Zantinge. 1996. Data Mining. Addison Wesley Longman, England, 158 P.

Akbaba, O. 2003. Web Sayfalarının Otomatik Olarak Sınıflandırılması Üzerine Yaklaşımlar ve Örnek Simülasyon Uygulaması. Gebze Yüksek Teknoloji Enstitüsü, Yüksek Lisans Tezi (yayınlanmamış), 55 S.

Akpınar, H. 2000. Veri tabanlarında bilgi keşfi ve veri madenciliği. İ.Ü.İşletme Fakültesi Dergisi. Nisan-2000, 1-22.

Alpaydın, E. 2000. Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Bilişim 2000 Veri Madenciliği Eğitim Semineri., İstanbul.

Belen, E. Özgür, Ç. ve Özakar, B., 2003. WALA : Web Erişim Kütük Araştırmacısı. Türkiye Bilişim Derneği 3. Bilişim Haftası, İstanbul. http://kurultay.tbd.org.tr/kurultay20/Bildiriler/Belgin_Ozakar/bildiri.pdf

Chakrabarti, S., Van den Berg, M. Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery, *Computer Networks,* 31 (1999) 1623–1640.

Chen, Z. 2001. Data Mining and Uncertain Reasoning. A Wiley-Interscience Publication, USA, 392 P.

Chuang, S.-L., Chien, L.-F. 2002. Enriching Web taxonomies through subject categorization of query terms from search engine logs. Decision Support Systems, 35 (2003), 113-127.

Çavdur, F. 2005. Arama Motorlari Kullanici Oturumlarindaki Konu Değişikliklerinin Tespit ve Tahmin Yöntemleri. Uludağ Üniversitesi, Yüksek Lisans Tezi. (yayınlanmamış), 153 S.

Dunham, M.H. 2003. Data Mining Introductory and Advanced Topics. Prentice Hall, New Jersey, 5-19 P, 195-220 P.

Efe, Ö., O, Kaynak. 2000. Yapay Sinir Ağları ve Uygulamaları. Boğaziçi Üniversitesi Yayını, İstanbul, 141 S.

Eker, H. Veri Madenciliği Veya Bilgi Keşfi. 2004. http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538

Elmas, Ç. 2003. Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama), Seçkin Yayıncılık, Ankara. 192 S.

Etzioni, O. 1996. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 65-68. http://www.cs.washington.edu/homes/etzioni/papers/cacm96.pdf

Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, USA, 1-10 P.

Han, J. 1999. What is data mining?. Kluwer Academic Publishers, USA, 170P.

Han, J. ve M, Kamer. 2001. Data Mining: Concept and Techniques. Morgan Kaufmann Publishers, San Francisco, 703 P.

Hand, D.J. 1999. Statistics and data mining: intersection disciplinse. Knowledge Discovery & Data Mining. ACM. 16-19.

Haykin, S. 1994. Neural Networks. Macmillan College Publishing Company, USA, 696P.

He, D., Göker, A., Harper, D. J. 2002. Combining evidence for automatic Web session identification. Information Processing and Management, 38 (2002), 727-742.

Jansen, J. B., Spink, A. 2003. An analysis of Web searching by European AlltheWeb.com users. Information Processing and Management, (baskıda).

Karakaş, M. 2002. Veri Madenciliği Üzerine.
http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=132

Komorowski, J. ve Zytkow, J. 1997. Principles of Data Mining and Knowledge Discovery. Springer, Germany, 510 P.

Kutluay, K. 2004. Veri Kalitesi Madencilik Elele. CRMPro Dergisi, 2004-Ağustos sayısı.

Moxon, B. 1996. Defining data mining. DBMS Data Warehouse Supplement. http://www.dbmsmag.com/9608d53.html.

Oğuz, M. ve Akbaş S., 1999. Genetik Algoritmalar ve Bir Üretim Problemine Uygulanması. YTÜ Doktora Tezi (yayınlanmamış), İstanbul, 105 S.

Ozmutlu, S. and Cavdur, F., Neural Network Applications for Automatic Topic Identification, Online Information Review, 29 (1): 34-53, (2005).

Ozmutlu, H.C. and Cavdur, F., Application of automatic topic identification on excite web search engine data logs, Information Processing and Management, 41 (5): 1243-1262 (2005).

Ozmutlu, H.C. and Ozmutlu, S., 2004. An Architecture For SCS: A specialized Web crawler on the topic of security. ASIST 2004: 67th Annual Meeting of the American Society for Information Science and Technology, Providence, RI, Nov. 2004.

Ozmutlu, H. C., Spink, A., & Ozmultu, S. 2002. Analysis of large data logs: an application of Poisson sampling on Excite web queries. *Information Processing and Management,* 38(4), 473-490.

Özakar, B. ve Püskülcü, H., 2002. Web içerik ve web kullanım madenciliği tekniklerinin entegrasyonu ile oluşmuş bir veri tabanından nasıl yararlanılabilir?. Türkiye'de İnternet Konferansları-VIII. http://inet-tr.org.tr/inetconf8/bildiri/119.doc

Semetay, C. ve Kurt, M. Genetik Algoritma ve Uygulama Alanları. TMMO Makine ve Mühendis Dergisi, Ekim-2001 sayısı.

Spink, A., Ozmutlu, H. C., Lorence, P. D. 2002. Web searching for sexual information: an exploratory study. Information Processing and Management, 40 (2004), 113-123.

Şakiroğlu, M. Tuğ, E. ve Bulun, M., 2003. Web Log Dosyalarından Genetik Algoritma Yöntemiyle Sıralı Erişimlerin Tespit Edilmesi.Türkiye Bilişim Derneği, 3. Bilişim Haftası, İstanbul.
http://kurultay.tbd.org.tr/kurultay20/Bildiriler/Merve_Sakiroglu/bildiri.pdf

Takcı, H. ve Soğukpınar, İ., 2002. Erişim Desenleriyle Saldırı Tespiti. Bilgi Teknolojileri Kongresi, Pamukkale Üniversitesi, 6-8 Mayıs 2002, Denizli.

Vahaplar, A. ve İnceoğlu, M.M. 2001. Veri Madenciliği ve Elektronik Ticaret Türkiye'de İnternet Konferansları-VII. Elektronik Poster Bildiri. http://inet-tr.org.tr/inetconf7/eposter/inceoglu.doc

Vaughan, L. 2003. New measuraments for search engine evaluation proposed and tested. Information Processing and Management, 40 (2004), 677-691.

Westphal, C. ve Blaxton, T. 1998. Data Mining Solutions. Wiley Computer Publishing, USA, 5-23.

Zhong, N. ve Zhou, L. 1999. Methodologies for Knowledge Discovery and Data Mining. Springer, Germany, 533 P.

http://www.po.metu.edu.tr/links/inf/css25/bolum1.html#1

http://www.google.com.tr

TEŞEKKÜR

Bu çalışma ve yüksek lisans eğitimim süresince bana yardımcı olan bölüm başkanım sayın Prof. Dr. Erdal Emel'e, yüksek lisans tezini birlikte yaptığım danışmanım sayın Yrd. Doç. Dr. Seda Özmutlu'ya ve eşi sayın Yrd. Doç. Dr. H. Cenk Özmutlu'ya teşekkürü bir borç bilirim. Ayrıca eğitimim boyunca bana her zaman destek olan aileme teşekkür ederim.

ÖZGEÇMİŞ

Gülşah Aynekin 01.09.1980 tarihinde Kayseri'de doğmuştur. İlk ve orta öğrenimini 50. Yıl Dedeman İlköğretim Okulu'nda, lise öğrenimini Sümer Lisesi'nde tamamlamıştır. 2002 yılında Erciyes Üniversitesi Endüstri Mühendisliği Bölümün'den mezun olan Gülşah Aynekin, 2003 yılında Uludağ Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda yüksek lisans eğitimine başlamıştır. Halen özel bir tekstil kuruluşunda Planlama Mühendisi olarak çalışmaktadır.