YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 1, Sayı 3, Kasım 2014

Sosyal Ağlarda Akan Veri Madenciliği

Sadi Evren SEKER

American University of Middle East, Kuwait, academic@sadievrenseker.com

Özet

Son yıllarda kullanımı hızla artan Facebook veya Twitter gibi sosyal ağlarda her gün çok yüksek miktarda veri akmaktadır. Bu verilerin tamamına yakını çeşitli veri merkezlerinde toplanmakta ve veri madenciliği olarak isimlendirilen yöntemlerle işlenerek anlamlı istatistiksel değerler çıkarılmaktadır. Ayrıca bu ağların çoğunun içeriği internet üzerinden açık olduğu için, veri merkezlerine gerek kalmadan internete bağlı herhangi birisi de benzer yöntemlerle veri toplayabilmekte ve analiz yapabilmektedir. Bu yazıda sosyal ağlardan verinin nasıl toplandığı ve nasıl veri madenciliği çalışmaları yapılabildiğine dair giriş seviyesinde bilgiler verilecektir.

Anahtar Kavramlar: *Veri Madenciliği, Metin Madenciliği, Sosyal Ağlar*

1. Giriş

Siz bu yazıyı okurken, birileri sosyal ağlarda önümüzdeki seçimlere dair, tuttuğunuz futbol takımına veya çalıştığınız şirkete veya okuduğunuz okullara dair görüşlerini paylaşıyorlar.

Bu paylaşılan bilgiler o kadar çok ki, bazı konularda sadece bir saatlik paylaşımı okumak için bir insan ömrü yetmeyebilir.

Örneğin sadece 2012 yılında sosyal ağlarda ve webde üretilen yazılı içerik, yazının icadından sonra 6000 yıl boyunca yazılmış bütün içerikten daha fazla.

2. Veri Madenciliği ile Bilginin İşlenmesi

Bu kadar veri akıp gidiyor, ilgilenenler kendilerine yakın bazı kaynakları vakitleri ölçüsünde okuyor ve yorumluyor, bazı durumlarda da tartışıyor. Peki bu kadar veri, bu kadar emek boşa mı gidiyor?

Hayır, bu verilerin anlamlı birer sonuca dönüşmesi ancak doğru bir işleme ile olabilir. Günümüzde bu kadar büyük ölçekte verinin işlenmesi için **büyük veri** (big data) ismi verilen özel çalışma alanları mevcut (Seker, 2015) . Bu çalışma alanlarından birisi de veri madenciliğidir (Seker, 2015). Yani veri ne kadar büyük olursa olsun özel teknolojiler ile bu verinin işlenmesi ve amaca yönelik olarak sonuçlar elde edilmesi mümkün.

Büyük Veri (Big Data) çalışmaları ile yüksek miktardaki verinin hızlı bir şekilde işlenmesi ve sosyal ağlar gibi yüksek miktarda verinin aktığı ortamların analiz edilmesi mümkündür. Büyük veri çalışmaları, yüksek miktarda işlem gücüne ihtiyaç duyduğu için, günümüzde daha çok bulut bilişim (cloud computing) ile birlikte düşünülmektedir.

Örneğin şu anda yazılan Twitter verilerine bakarak önümüzdeki seçimlere dair bir tahmin yapmak istiyor olalım veya sosyal ağdaki arkadaşlık haritasına bakarak Yunanistan'da bir salgın hastalık olduğunda ilk önce hangi ülkenin etkileneceğini bulmak isteyelim veya internetteki çok sayıdaki web sitesinden hangilerinin terör örgütlerine ait olduğunu veya hangi siyasi görüşün etkisinde olduğunu bulmak isteyelim.

İşte bütün bu amaçlara yönelik olarak yazacağımız bir kod ile internet üzerinden veri toplamamız, toplanan bu veriyi işlememiz ve istediğimiz sonuçları bulmamız mümkün. Basitçe web madenciliği ismi verilen bu işlemler 5 adımdan oluşur:

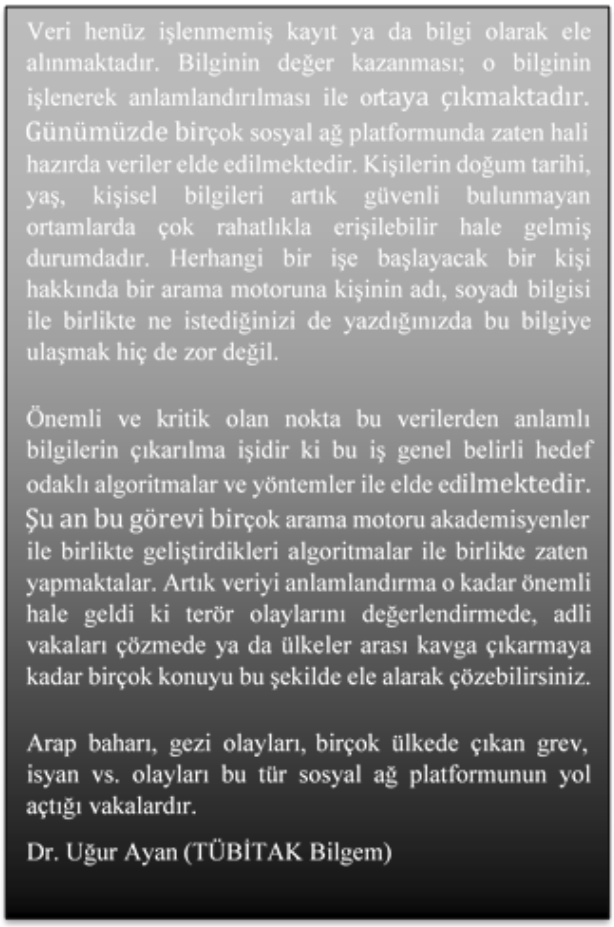
1. Verinin toplanması ve seçilmesi
2. Verinin temizlenmesi
3. Verinin özelliklerinin çıkarılması
4. Makine öğrenmesi algoritmaları kullanılması
5. Anlamlı sonuçların yeniden kullanılabilir hale getirilmesi

İlk adım olan veri toplama için çeşitli yöntemler olmasına karşılık, örneğin internet üzerinde bir örümcek (crawler) dolaşarak web sitelerindeki bağlantıları takip edip hedef bir web sitesinin içeriğini ve bağlantılı web sitelerinin içeriklerini indirmek için kullanılabilir. Elbette bu örümcek internette dolaşırken daha önceden kendisinden istediğimiz özellikte verileri bulmaya çalışacaktır. Verinin özel olarak toplanması için kullanılan bu örümcekler aslında genel anlamıyla bütün arama motorları tarafından kullanılmaktadır. Herhangi bir internet kaynağının ne kadar sıklıkta tekrar ziyaret edileceği ve güncellemelerin yakalanacağı, reklam, bozuk içerikler veya aradığımız bilgi ile ilgisi olmayan içeriğin atlanarak taramaya dahil edilmemesi, taradığımız sitelerin arasında hangi sitelere öncelik vereceğimiz gibi çok sayıdaki problem bu örümceklerin kodlamasında dikkate alınır.

Verinin temizlenmesi aşamasında, veri üzerinde gelen ve bizim amacımıza hizmet etmeyen, örneğin sitenin daha güzel görülmesi için yazılmış kozmetik kodlamalar, reklamlar, bozuk içerikler, eksik ve hatalı olabilecek bilgiler temizlenir. Çünkü raporlama sonucunda genelde çıkan sonucun doğru olması istenir ve sonucu yanlış yönlendirebilecek bu verilerin temizlenmesi hedeflenir. Ancak temizleme işlemi dışında da çok sayıda yöntem mevcuttur.

Toplanan veriler ne yazık ki genelde bilgisayarlar tarafından doğrudan işlenemeyen verilerdir. Örneğin şu anda Türkçe olarak yazdığım bu yazıyı insanlar okuyunca bir anlam verebilir, kendi hayatları ve geçmiş tecrübeleri ile ilişkilendirebilir, bu yazıdaki bazı konulara dayanarak, hayal kurabilir, tahmin yapabilir, ancak bu özelliklerin hiçbiri olmayan, yani hayal kuramayan, hayatı ve geçmiş tecrübeleri olmayan, tahmin yeteneği olmayan bilgisayarlar, insanlar gibi davranmazlar. *Sonuçta bilgisayarların işleme şekli, sayılar ve bu sayılar üzerinde tanımlı*

dört işlem gibi basit işlemlerdir. İşte, bilgisayarlardan istediğimiz ve bizim yerimize bir seçimin sonucunun tahmini veya tuttuğumuz futbol takımı ile ilgili görüşlerin analizi veya yazılan ekonomi haberlerinden borsa tahmini gibi işlemler bilgisayarlar için oldukça karmaşık işlemlerdir ve bu çalışmalar genel olarak **yapay zeka** çatısı altında toplanırlar. Bu çalışmalar kapsamında, toplanan verinin, bilgisayarların işleyebileceği seviyeye indirilmesi yani sayılarla ifade edilmesi gerekir. Bu sayılarla ifade işlemi ise özellik çıkarımıdır.


Veri henüz işlenmemiş kayıt ya da bilgi olarak ele alınmaktadır. Bilginin değer kazanması; o bilginin işlenerek anlamlandırılması ile ortaya çıkmaktadır. Günümüzde birçok sosyal ağ platformunda zaten hali hazırda veriler elde edilmektedir. Kişilerin doğum tarihi, yaş, kişisel bilgileri artık güvenli bulunmayan ortamlarda çok rahatlıkla erişilebilir hale gelmiş durumdadır. Herhangi bir işe başlayacak bir kişi hakkında bir arama motoruna kişinin adı, soyadı bilgisi ile birlikte ne istediğinizi de yazdığınızda bu bilgiye ulaşmak hiç de zor değil.

Önemli ve kritik olan nokta bu verilerden anlamlı bilgilerin çıkarılma işidir ki bu iş genel belirli hedef odaklı algoritmalar ve yöntemler ile elde edilmektedir. Şu an bu görevi birçok arama motoru akademisyenler ile birlikte geliştirdikleri algoritmalar ile birlikte zaten yapmaktalar. Artık veriyi anlamlandırma o kadar önemli hale geldi ki terör olaylarını değerlendirmede, adli vakaları çözmede ya da ülkeler arası kavga çıkarmaya kadar birçok konuyu bu şekilde ele alarak çözebilirsiniz.

Arap baharı, gezi olayları, birçok ülkede çıkan grev, isyan vs. olayları bu tür sosyal ağ platformunun yol açtığı vakalardır.

Dr. Uğur Ayan (TÜBİTAK Bilgem)

Veri üzerinden özellik çıkarılması aşamasında, işlenen verinin yapısına bağlı olarak çeşitli yöntemler kullanılabilir. Örneğin web sitesinden alınan resimler üzerinden bir işlem yapmak isteniyorsa, resim işleme yöntemleri ile resimlerin özelliklerinin çıkarılması, resimdeki kişilerin, sembollerin, yer işaretlerinin vs. tespit edilmesi mümkündür. Benzer şekilde yazılmış bir *twitter veya facebook mesajının da işlenerek bu mesajda geçen anlamın anlaşılması, mesajın olumlu veya olumsuz anlama göre sınıflandırılması veya mesajın ilgili olduğu kişi, kurum ve ifadelerin ilişkilendirilmesi mümkündür.*

Şayet kaynak olarak kullanılan veri, yazılı metinse, bu durumda doğal dil işleme veya metin madenciliği yöntemleri kullanılarak bilgisayarların yazılı metni anlaması hedeflenir (Seker, 2015).

Bu aşamaya kadar veriyi toplayıp, veriyi işlenebilir şekilde temizledik ve veri üzerinden bilgisayarların işleyebileceği sayısal özellikleri çıkarttık.

Bilgisayarların işlemesi için hazırlanmış olan verinin üzerinde artık istediğimiz sonuca yönelik olarak çeşitli **makine öğrenme algoritmalarını** kullanabiliriz. Örneğin elimizde yazılı bir metin var ve bu metni hangi yazarın yazdığını bulmak istiyor olalım veya yazılmış yazının ekonomi, spor, magazin veya politika konularından hangisi ile ilgili olduğunu bulmak istiyor olalım, işte bu tip problemlere sınıflandırma problemleri ismi verilir ve genel olarak sınıflandırma algoritmaları ile çözülür. Örneğin bilgisayar kendisini daha önceden toplanmış verilere göre eğiterek bu sınıfları öğrenir ve daha sonra yeni gelen bir metin için hangi sınıfa ait olduğunu tahmin etmeye çalışır.

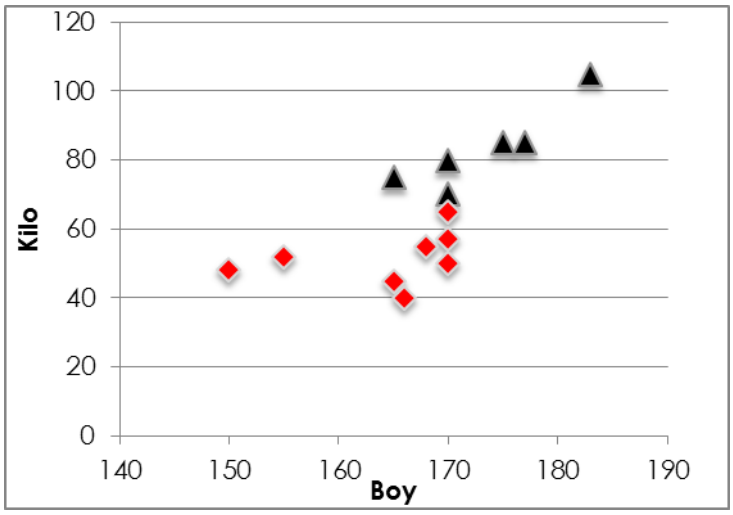
Benzer bir problem elimizdeki iki bilgi arasında da kurulabilir. Örneğin milyarlarca yazıdan birbirine en yakın iki yazıyı bulmak istiyor olalım veya yazıların sahibi yazarlardan birbirine en yakın üslupla yazan yazarları bulmak istiyor olalım. Bu problem tipi için ilişkilendirme algoritmaları kullanılabilir.

Çok sayıda farklı problem çeşitleri ve kullanılabilecek algoritmalar bulunmaktadır. Bu problemler arasında geleceğe dair tahmin problemleri, bir sistemde oluşan bir anormalliğin tespiti problemleri, verilen verilerden matematiksel bir gösterim çıkarmaya yönelik problemler veya verilen verinin daha kısıtlı bir alanda özetlenmesine dair problemler gibi problemler sayılabilir.

Diyelim ki bir suçlunun yakalanması için bir bilgisayar yazılımı hazırlamamız isteniyor. Örneğin bir makine öğrenmesi algoritması ile suçlunun boyu ve kilosunu biliyorsak, cinsiyetini tahmin etmeye çalışalım.

Örnek olarak sıkça kullanılan KNN algoritmasını ve nasıl çalıştığını burada açıklayalım.

Basitçe elimizdeki şimdiye kadar toplanmış boy ve kilo bilgilerini iki boyutlu uzaya dağıtarak işe başlayalım:

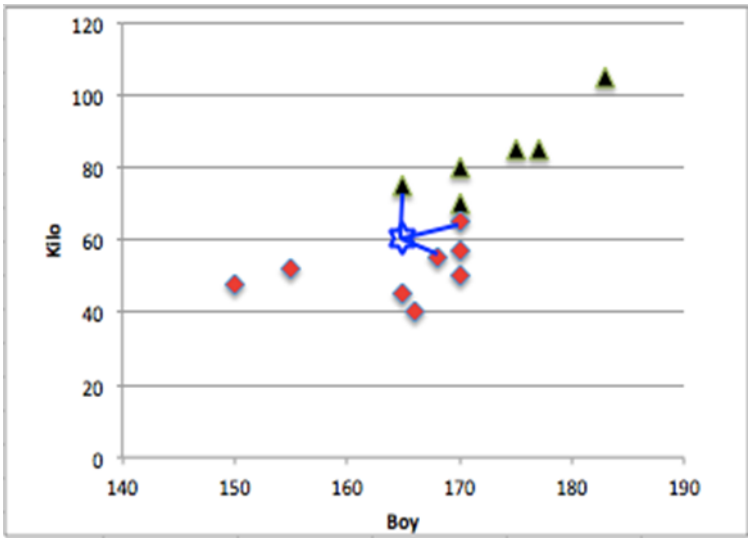


Şekil 1 *Kilo ve boy bilgilerinin grafiği*

Bu grafikte, boy bilgisi yatay, kilo bilgisi düşey eksende gösterilmiştir ve kadın bireyler kırmızı, erkek bireyler ise siyah renkle işaretlenmiştir.

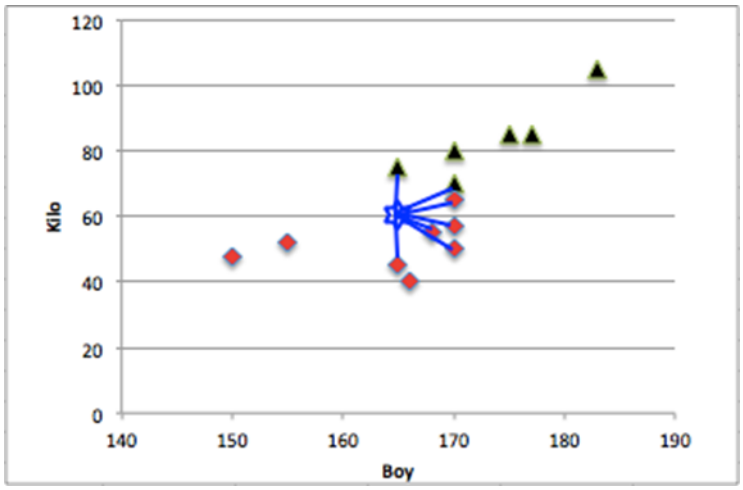
Algoritmamıza göre elimizde boyu ve kilosu olan bir bireyin cinsiyetini tahmin etmek için sayısal olarak en yakın 3 bireyin cinsiyetlerini kontrol edeceğiz ve buna göre bir tahminde bulunacağız.

Örneğin cinsiyetini tahmin etmek istediğimiz bireyin boyunun 165 ve kilosunun 60 olduğunu kabul edelim. Bu durumda şekildeki bireyin konumu ve en yakınındaki 3 birey Şekil 2'deki gibi olacaktır.



Şekil 2 *En yakın 3 komşuya bakma*

Tahminde bulunmak istediğimiz kişinin verilerine, elimizdeki en yakın 3 kişiden 2 tanesi kadın ve 1 tanesi erkek olduğu için bu kişinin %66 olasılıkla kadın olduğunu söyleyebiliriz. KNN algoritmasında, sizin de fark edeceğiniz üzere taranacak komşu sayısı her zaman tek sayı olarak tutulur. Bunun sebebi çift sayıdaki komşularda eşitlik olması ihtimalidir. Daha detaylı bir bilgi için 3 yerine 7 komşuya bakalım.



Şekil 3 *En yakın 7 komşuya bakma*

7 komşudan 5 tanesinin kadın ve 2 tanesinin erkek olduğunu bulduk. Demek ki daha detaylı bir aramayla 5/7 = %72 oranında bu bireyin kadın olduğunu söyleyebiliriz.

Kaç komşuya bakılacağı, kararın kesinliğini arttırırken işlem süresini uzatmaktadır. İşlem hızı ise sosyal ağlar gibi büyük verilerle çalışılırken dikkate alınması gereken parametrelerdendir.

Veri madenciliği çalışmamızın son aşamasında da çıkan sonuçları yeniden kullanılabilir hale getirmeye çalışıyoruz. Örneğin yukarıdaki bir örnek için veri topladık, toplanan veriler üzerinden testler yaptık ve bilgisayarımız bize çeşitli tahminlerde bulundu. Bu tahminler ne kadar doğru? Hata oranlarına bakarak veri toplama şeklimizi, verilerden özellik çıkarma şeklimizi veya algoritmalarımızı gözden geçirmemiz gerekir.

Yukarıdaki ufak örneğimizde kişilerin sadece boy ve kilosuna bakarak bir suçlunun cinsiyetini bulan yazılımın nasıl geliştirildiğine baktık. Öncelikle 13 farklı bireyin verilerini topladık. Bu verilerden boy ve kiloyu kullanmaya karar verdik ve bu iki veriyi birbiri ile ilişkilendirdik. Ardından elimizdeki bireyin özelliklerine göre KNN algoritması kullandık.

Hiçbir veri madenciliği yöntemi tam başarı garanti etmez. Mutlaka bir hata payı vardır. Bizim amacımız hatayı en aza indirirken, çalışma hızımızı makul seviyede tutmaktır.

3. Akan Veri Madenciliği

Günümüzde hızla gelişen konulardan birisi de İnternet kullanımı ve İnternet gibi değişken (dinamik) bir ortamdaki verinin analiz edilmesi sorunudur. Hem veri çok büyük boyutlara ulaşmakta hem de problem karmaşıklaşmaktadır.

Yine bir örnek üzerinden gidelim. Mesela, bu sefer elimizde kimin yazdığı belli olmayan ve bir suçu aydınlatacak, bir yazı olsun ve biz de bu yazının sahibini arıyor olalım. Diyelim ki şüphelendiğimiz 3 farklı kişi var.

İşe bu 3 kişinin daha önceki yazılarını toplayarak başlıyoruz. Her yazarın ne kadar çok yazısını bulursak o kadar işimiz kolaylaşır.

Ardından her yazarın kullandığı kelime sıklıklarını sayıyoruz. Problemi basit tutmak adına çok fazla kelime kullanmadıklarını kabul ediyorum (gerçek uygulamalarda bu farklı kelime sayıları milyonlar mertebesinde oluyor) ve şekil 4'teki gibi bir tablo hazırlıyoruz:

	Yazar 1	Yazar 2	Yazar 3	Toplam
Kelime1	-	10	3	13
Kelime2	10	1	-	11
Kelime3	5	5	5	15
Kelime4	8	3	2	13
Kelime5	2	1	-	3
Toplam	25	20	10	55

Şekil 4

Bu aşamadan sonra suçu aydınlatacağını düşündüğümüz yazıyı ele alabiliriz. Örneğin yazı da aşağıdaki gibi 3 kelimeli bir yazı olsun:

Yazı: Kelime1, Kelime2, Kelime2, Kelime3

Yani yazımızda 2 kere ikinci kelime ve birer kere birinci ve üçüncü kelimeler kullanılmış.

Şimdi artık bu yazıdaki kullanım sıklıklarına ve yazarların kelime kullanma alışkanlıklarına bakarak yazarı tahmin edebiliriz.

Normalde kelime frekansları (kullanım sıklıkları) normalleştirilmekte (normalization) ve farklı istatistiksel modeller kullanılmakta ancak ben problemi ve çözümü basit tutmak adına aşağıdaki şekilde her kelime için ayrı ayrı hangi yazara ait olma ihtimali olduğunu inceliyorum.

Kelime1: Yazar 2 10/25

Kelime2: Yazar 1 10/25

Kelime2: Yazar 1 10/15

Kelime3: Bütün yazarlar için eşit

Buna göre 3. Kelimenin bizim yazarı tahminimiz için bir katkısı yok ancak diğer kelimelere göre yazar 1 olma ihtimali, yazar 2 olma ihtimaline göre daha kuvvetli görülüyor. Demek ki bu yazıyı, örneklerini topladığımız 1. Yazarın yazmış olması daha yüksek ihtimal.

Buradaki örnek aslında klasik bir veri madenciliği çalışmasıdır. Bu örnek ne zaman akan veri madenciliği çalışması olur sorusuna ise aşağıdaki gibi sosyal ağlarda sıkça karşılaştığımız üç durumu açıklayarak cevap vermek gerekir.

1. Sosyal ağlarda, eğitim kümemizi (yazarların, yazı örnekleri) toplarken, sabit sayıda yazarımız yoktur, sürekli yeni yazarlar sisteme dahil olabilmektedir
2. Sosyal ağlarda uzun süre çalıştığımızı düşünürsek bir süre sonra bazı yazarların yazı yazmayı bıraktığını biliyoruz. Bu yazarların sistemden silinmesi gerekir çünkü kısıtlı bilgisayar imkanları ile çalışılmaktadır ve tarihte yazı yazan her yazarın yazı karakteristiklerini toplamak şimdiki bilgisayarların kapasitelerinin çok üzerindedir.
3. Silinmiş bazı yazarlar uzun süre sonra tekrar yazı yazmaya başlayabilir, bu yazarların eski karakteristik özelliklerinin geri yüklenmesi gerekebilir.

Yukarıda sadece üç örneği verilen durumlar için akan veri madenciliği çalışmaları yürütülmektedir. Genelde mevcut yazarların karakteristiğinin çok dışında farklı karakteristik özellikler gösteren yazıların, farklı bir yazardan gelmesi ihtimali üzerinde durulur.

Örneğin Yazı olarak aşağıdaki şekilde bir yazı gelmiş olsun:

Yazı: Kelime6, Kelime7, Kelime6, Kelime1

Bu durumda yazının %80'lik kısmı yeni kelimelerden oluşuyor ve daha önceki yazarlarımızdan tanımadığımız kelimelerin sayısı beklenenden çok. Bu durumda bu yeni yazının, yeni bir yazar tarafından yazılmış olma ihtimali üzerinde durulabilir.

Henüz akademide geliştirilen ve yeni yeni uygulamaları olmaya başlayan akan veri madenciliği çalışmaları ile, değişmekte olan ortamlarda verinin daha sağlıklı analizi ve değişimin yakalanması mümkün olacaktır.

Veri madenciliği uygulamalarına ihtiyaç her geçen gün hızla artıyor ve veri madenciliği ile ilgili çok sayıda yeni iş alanları da çıkıyor. Belki de siz bu yazıyı okurken yeni bir yöntem geliştirilmiş olabilir.

Kaynakça

Şadi Evren ŞEKER, "İş Zekası ve Veri Madenciliği (WEKA ile)", İstanbul, Cinius Yayınları, ISBN 978-605-127-671-7, 2013

Ethem ALPAYDIN, 'Yapay Öğrenme', İstanbul Boğaziçi Üniversitesi Yayınevi, ISBN: 9786054238491, 2. Basım, 2013

Sadi Evren SEKER, Khaled Al-NAAMI, Latifur KHAN, " Author Attribution on Streaming Data",Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on , IEEE IRI pp. 497 – 503, Aug. 2013

Sadi Evren SEKER, Cihan MERT, Khaled Al-Naami, Ugur AYAN, Nuri OZALP, "Ensemble classification over stock market time series and economy news", Intelligence and Security Informatics (ISI), Proceeding of 2013 IEEE International Conference, pp 272 – 273, ISBN 978-1-4673-6214-6

Seker, S. E. (2014), "Büyük Veri Yaşam Döngüsü (Big Data Life Cycle)", YBS Ansiklopedi, v. 2, is. 3, pp. 10 – 17

Seker, S. E. (2015), "Metin Madenciliği (Text Mining)", YBS Ansiklopedi, v. 2, is. 3, pp. 30 – 32

Seker, S. E. "Weka ile Veri Madenciliği", draft2digital, Bilgisayar Kavramları Yayınları, İstanbul, 2015, ISBN: 9781524255350

I. Ocak, S. E. SEKER (2013), Calculation of surface settlements caused by EPBM tunneling using artificial neural network, SVM, and Gaussian processes, Environmental Earth Sciences, Springer-Verlag, Vol. 70, Is. 3, pp. 1263-1276, DOI: 10.1007/s12665-012-2214-x, Oct. 2013