Veri Madenciliği

METİN MADENCİLİĞİ

Metin Madenciliği Nedir?

Metin Madenciliği (Text Mining), yapısal olmayan ve düzensiz haldeki elektronik metin yığınlarından; önceden bilinmeyen, potansiyel olarak kullanışlı, yapısal ve düzenli veri elde etme sürecidir.

Elde edilen bilgiyle, analiz edilen metin kaynaklarında açık olarak görülmeyen ilişkiler, hipotezler ve eğilimler tespit edilir.

Metin Madenciliği, veri madenciliğinin bir parçası olarak düşünülmesine rağmen, alışılagelen veri madenciliğinden farklıdır. Ana farklılık, Metin Madenciliğinde örüntülerin olay tabanlı veri tabanlarından daha çok, doğal dil metinlerinden çıkarılmasıdır.

Metin Madenciliği Nedir?

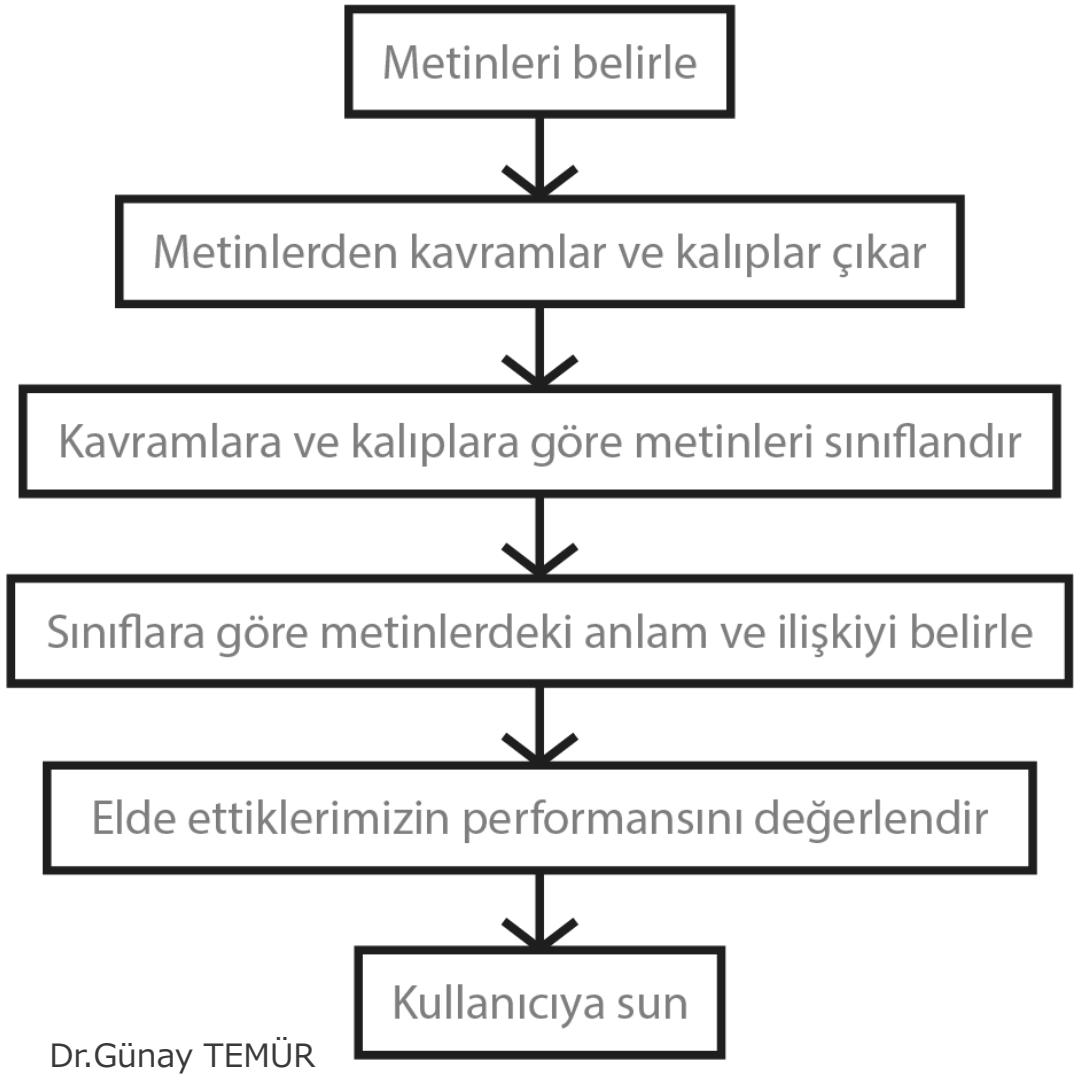
En basit anlamda Metin Madenciliği çalışmaları, metni veri kaynağı olarak kabul eden veri madenciliği çalışmasıdır ve metin üzerinden yapısallaştırılmış veri elde etmeyi amaçlar.

Örneğin; metinlerin sınıflandırılması (classfication), bölütlenmesi (clustering), metinlerden konu çıkarılması (entity extraction), sınıf taneciklerinin üretilmesi (production of granular taxonomy), duygusal analiz (sentimental analysis), metin özetleme (document summarization), ve varlık ilişki modellemesi (entity relationship modelling) gibi çalışmaları hedefler.

Çalışma Alanları

Belirtilen hedeflere ulaşılması için Metin Madenciliği çalışmaları kapsamında; bilgi getirimi (information retrieval), hece analizi (lexical analysis), kelime frekans dağılımı (word frequency distribution), örüntü tanıma (pattern recognition), etiketleme (tagging), bilgi çıkarımı (information extraction), veri madenciliği (data mining) ve görselleştirme (visualization) gibi yöntemler kullanılmaktadır.

Metin madenciliğinin aşamaları şu şekilde özetleyebiliriz:



- Analizimize uygun olan metinleri belirlemek.
- Metinlere istatistiksel, yapısal ve dilsel teknikler uygulayarak kavramlar ve kalıplar çıkarmak.
- Statik, makine öğrenimi, ve kalıp eşleştirme teknikleri ile kavramlara ve kalıplara göre metinleri sınıflandırmak.
- Sınıflara göre metinlerdeki anlam ve ilişkiyi belirlemek.
- Elde ettiklerimizin performansını; doğruluğunu, tutarlılığını, alakasını, hassasiyetini kontrol ederek değerlendirmek.
- Kullanıcıya sunmak.

Metin madenciliğinin aşamaları

Enformasyon Getirimi (Information Retrieval): Bu aşama ilgilenilen külliyet (derlem, corpus) hakkında ön bilginin toplandığı aşamadır. Örneğin metin madenciliği web üzerindeki veri kaynakları üzerinde yapılacaksa web sayfaları, adresleri veya dosya sistemi üzerindeyse dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin bilgileri gibi bilgilerin toplandığı aşamadır.

Doğal dil işleme aşaması (natural language processing): Bu aşama bütün metin madenciliği aşamalarında kullanılmasa bile genelde özellik çıkarımı ve metinden bazı anlamsal bilgilerin elde edilmesinde sıklıkla başvurulan aşamadır. Örneğin, konuşma parçalarının etiketlenmesi (part of speech tagging) veya cümle bilimsel parçalama (syntactic parsing) veya diğer dilbilimsel işlemler doğal dil işleme aşamasında yapılır.

Metin madenciliğinin aşamaları

Adlandırılmış varlık tanıma (named entity recognition): Genellikle metin işleme aşamasında istatistiksel bazı özelliklerin çıkarılması için kullanılır. Örneğin, metnin içerisindeki kişi isimleri, yer isimleri, semboller, kısaltmalar v.s. bu yöntemle bulunur. Metin madenciliği çalışmalarının her zaman temiz metinlerde yapılmadığını hatırlatmakta yarar vardır.

Örneğin facebook, twitter mesajları, telefonlardan yollanan SMS mesajları gibi mesajların çoğunda yazım hataları hatta kısaltmalar kullanılmaktadır. Metin madenciliği bu ihtimallerin de göz önünde tutulması gereken çalışmalardır. Örneğin ‘’osman bey’’ kelimesi, istanbulda bir semt ismi olabileceği gibi bir kişi ismi de olabilir. Adlandırılmış varlık tanıma çalışmalarında, hedeflenen kelime gruplarının metin içerisinden çıkarılması, sayılması, yoğunluğunun bulunması, etiketlenmesi gibi işlemler yapılabilir.

Metin madenciliğinin aşamaları

Örüntüsü tanımlı varlıkların bulunması (pattern identified entities): Bazı durumlarda, metnin içerisinden özel bazı bilgilerin metin madenciliğine konu olması mümkündür. Örneğin e-posta adresleri, telefon numaraları, adresler, tarihler gibi bazı bilgileri özel olarak almak isteyebiliriz. Genelde bu durumlarda düzenli ifadeler (regular expressions) veya içerik bağımsız gramerler (context free grammers) tanımlanarak metin üzerinde çalıştırılır.

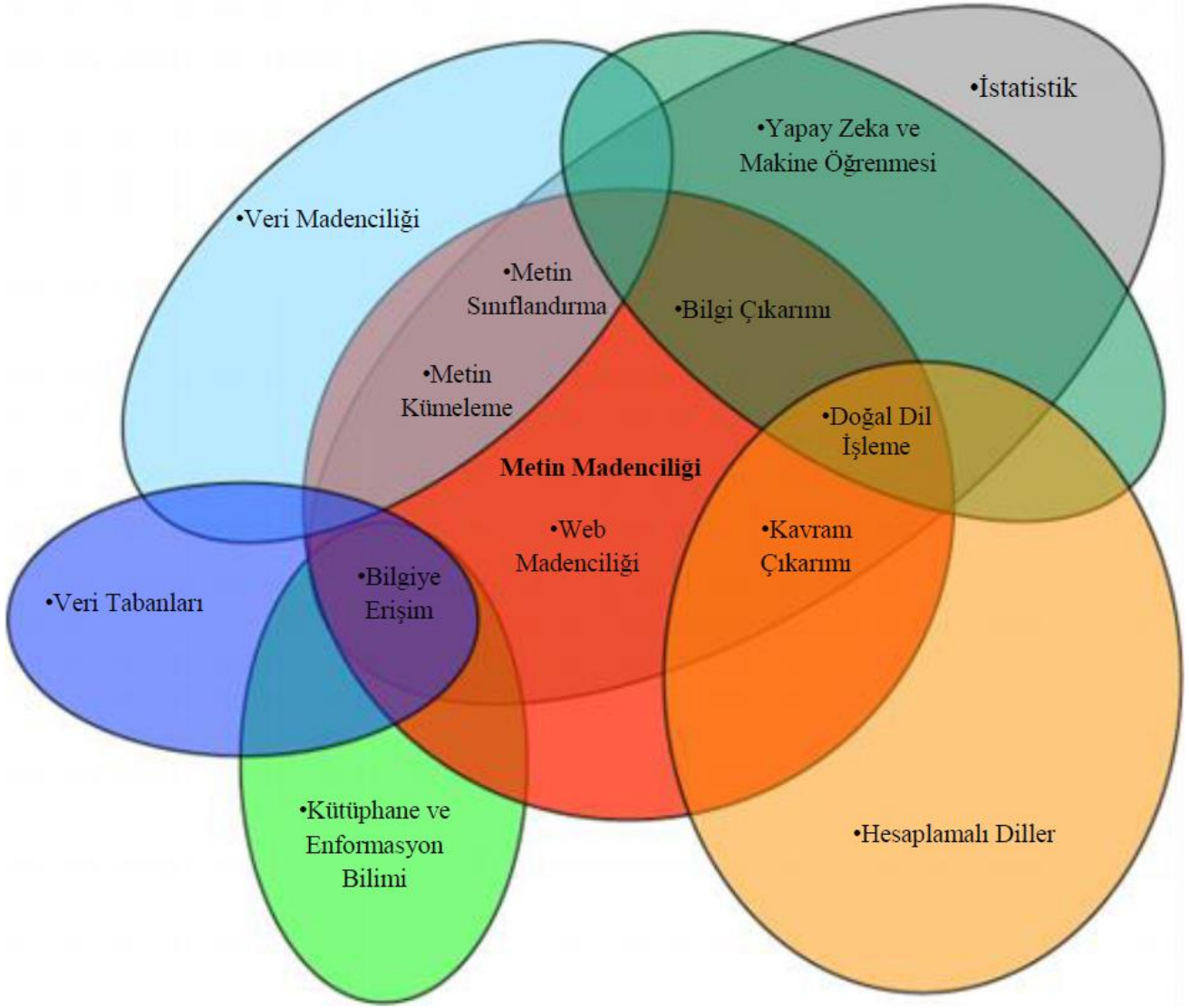
Eş Atıf (Coreference): Bir varlığa işaret eden (atıf eden) isim kelime gruplarını ve diğer terimlerin bulunması/ayrılmasını hedefler.

İlişki, kural, olay çıkarımları: Çeşitli amaçlarla metnin içerisinden bazı bilgilerin çıkarılması istenebilir. Örneğin bir doktora çalışmasında, verilen bir metnin içerisindeki olayları çıkararak sıralamak (event ordering) üzerine çalışmış, Türkçedeki fiil yapılarını, olay belirten kelime gruplarını, zaman kalıplarını ve bütün bu kelime grupları arasındaki olası ilişkileri gösteren özel bir matematiksel formül tasarlanmıştır.

Metin madenciliğinin aşamaları

Duygu analizi (sentimental Analysis): Metinlerde geçen duygusal ifadelerin çıkarılmasını amaçlar. En sık kullanılanı duygusal kutupsallıktır (sentimental polarity). Buna göre bir konu hakkında geçen mesajların veya yazıların olumlu veya olumsuz olmasına göre iki sınıfa ayrılması hedeflenir. Ancak duygu analizi bunun dışında, metinlerdeki ruh hali, kanaat ve daha karmaşık duyguların çıkarılması üzerinde de çalışmaktadır.

Metin Madenciliğinin ilişkili olduğu disiplinler ve yöntemler



Metin Verilerini Sayısallaştırma

Çok sayıda küçük doküman vs. az sayıda büyük doküman: eğer çok büyük olan az sayıda dokümandan "kavram" (concept) çıkarma niyetindeyseniz, o zaman değişken sayısı (çıkarılmış kelime) çok fazla iken durum sayınızın (doküman) çok az olmasından dolayı genel olarak istatistiksel analizler daha az güçlü olurlar.

Belirli karakter, kısa kelime, sayı vb. çıkarmak: harflerin belirli sayısından daha uzun ya da daha kısa olan kelime, karakter sırası, ya da belirli karakterleri çıkarmak, girdi dokümanlarını indekslemeye başlamadan önce yapılabilir. Ayrıca işlenen dokümanda küçük bir yüzdede görülen "seyrek (rare) kelimeleri" de çıkarabilirsiniz.

Metin Verilerini Sayısallaştırma

Listeye alma, listeden çıkarma (durdurma kelimeleri - stop words): sıralanacak kelimelerin belirli bir listesi tanımlanabilir, bu durum belirli kelimelere ulaşmak istediğinizde ve bu kelimelerin görülmesi ile frekansına dayalı girdi dokümanlarını sınıflandırmada kullanışlıdır. Ayrıca "durdurma kelimeleri" sıralamadan çıkarılacak olan terimleri tanımlar. İngilizcede bu kelimeler "the", "a", "of" 'dir.

Metin Verilerini Sayısallaştırma

Eş anlamlılar ve deyimler ve kelime kökleri: "ekmek" gibi eş anlamlılar ve belirli bir anlam ifade eden deyimler sıralama (indexing) için birleştirilebilir. Örneğin "Microsoft Office" bir bilgisayar paket programı şeklinde deyim olarak tanımlanabilir, veya analizde kullanılan doküman veri setinde listesi çıkarılmış kelimeler tekil veya çoğul olabilirler. Analizi yapılacak kelime sayısını azaltmada kelime köklerini belirleme ve bu kelimelerin çoğul olduğu durumları kökü belirlenmiş olan kelimeye atfetme ile durum karmaşasından kurtulunabilinir. Örneğin "olmak", ile "oldu" kelimeleri metin madenciliği programı tarafından aynı kelime olarak tanımlanır. Burada amaç kelimeleri köklerine indirgemektir.

Metin Madenciliği İçin Özellik Seçimi

Son zamanlarda internette bulunan dokümanların sayısında muazzam bir büyüme vardır. Yapılandırılmamış veriler ile bu dokümanlar çevrimiçi olarak depolanmış baskın veriler haline gelir.

Metinsel verilere etkin bir biçimde uygulanabilecek olan birçok özellik seçimi yaklaşımı mevcuttur. Bunların çoğu terimlerin bir puanlama tablosuna dayalıdır. Özelliklerin puanı, doküman veri setindeki ifadelerin kalitesini temsil eder.

Özellik seçme süreci veri madenciliğinde veri hazırlamayı takiben çok önemli bir stratejidir. Veri madenciliğinin en önemli problemi birçok potansiyel tahminleyici ile büyük veri setlerindeki boyutluluk lanetidir (curse of dimensionality).

Metin Madenciliği İçin Özellik Seçimi

Özellik seçimi modeldeki değişkenleri azaltmayı amaçlar, bu sebeple alakasız veya gereksiz değişkenler veya gürültülü veriyi kaldırarak lanetin (curse) etkisini azaltır. Analiz için aşağıdaki pozitif etkileri vardır.

-Algoritma sürecini hızlandırır

-Veri kalitesini arttırır

-Algoritmanın tahminleme gücünü arttırır

-Sonuçları daha anlaşılır yapar

Örnek Metin Madenciliği uygulaması

Örneğin elimizde 100 adet yazı olsun. Bu yazıları yazan yazarları biliyor olalım (diyelim ki 5 farklı yazarın 20'şer adet yazısı olsun). Yeni gelen 101. Yazının bu 5 yazardan hangisine ait olduğunu bulmak, klasik bir metin madenciliği uygulamasıdır ve literatürde yazar tanıma (author recognition) olarak da geçer.

Burada örnek olarak metinlerdeki kelime kullanma sıklıklarını özellik çıkarımı için kullanmak isteyelim. Yani yazarlarımızı kullandıkları kelime sıklıklarından tanıyabileceğimizi düşünüyoruz (author attribution). Her metinde ve dolayısıyla her yazar için hangi kelimeyi ne sıklıkla kullandığı bilgisi bizim özellik çıkarımı aşamamız oluyor.

Örnek Metin Madenciliği uygulaması

Ardından kullanılan kelime sıklıklarını örnek olarak makine öğrenme algoritması olan KNN algoritmasına veriyoruz ve diyelim ki yazarını tanımak istediğimiz 101. Yazı için her kelime için en çok kullanan yazarları listeliyoruz. Neticede bize bir olası yazarlar listesi çıkıyor ve biz de en yüksek ihtimalle hangi yazarın bu yazıyı yazmış olabileceğini söylüyoruz. Bu çıkan sonuç aslında 101. Yazı için anlamlı ve yapılandırılmış bir sonuç olarak kabul edilebilir.

ÖRNEK BİLİMSEL ÇALIŞMALAR

[PDF] **Metin madenciliği** ile soru cevaplama sistemi [PDF] emo.org.tr

S İlhan, N Duru, Ş Karagöz, M Sağır - Elektronik ve Bilgisayar ..., 2008 - emo.org.tr

... Veri **madenciliğinin** alt dalı olarak ele alınan **metin madenciliği** ise yazılmış farklı dokümanlardan ... **Metin** madenciliğini veri **madenciliğinden** ayıran en büyük fark **metin** madenciliğinde ...

☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 28 İlgili makaleler 5 sürümün hepsi »

Metin madenciliği ile e-ticaret sitelerinin belirlenmesi [PDF] dergipark.org.tr

T KAŞIKÇI, H Gökçen - Bilişim Teknolojileri Dergisi, 2013 - dergipark.org.tr

... için çözüm **metin madenciliği** ve **metin** sınıflandırmadır. **Metin madenciliği**, doğal dil işleme ile veri **madenciliğinin** bir arada ... Veri **madenciliği** büyük veri yığınları içerisinden gelecekle ...

☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 31 İlgili makaleler 4 sürümün hepsi »

KNN algoritması ve r dili ile **metin madenciliği** kullanılarak bilimsel makale tasnifi [PDF] dergipark.org.tr

D KILINÇ, E BORANDAĞ, F YÜCALAR... - Marmara Fen Bilimleri ..., 2016 - dergipark.org.tr

... **Metin** tabanlı veri setleri üzerinde analiz işlemi gerçekleştirebilmek için Veri **Madenciliğinin** alt alanı olan **Metin Madenciliği** ... Bu çalışmada, akademik yayınlar üzerinde **metin madenciliği** ...

☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 45 İlgili makaleler 6 sürümün hepsi »

ÖRNEK BİLİMSEL ÇALIŞMALAR

Metin madenciliği ile **metin** sınıflandırma [PDF] yildiz.edu.tr
İF Pilavcılar - 2007 - dspace.yildiz.edu.tr
... **Metin** halindeki verilerin bulunduğu veritabanlarındanbilgiyi kolayca elde etmek için **metin** ... yardımıyla **metin** halindeki verilersınıflandırılabilmektedir.Tezde, amaç doğrultusunda, ...
☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 24 İlgili makaleler »

[PDF] **Metin madenciliği** ile benzer haber tespiti [PDF] researchgate.net
A Karadağ, H Takçı - Akademik Bilişim, 2010 - researchgate.net
... analizi için veri **madenciliğinin** metinler üzerinde çalışan, daha farklı özelliklere sahip bir uyarlaması olan **metin madenciliği** kavramı tanımlanmıştır. **Metin madenciliği**; yapısal olmayan ...
☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 16 İlgili makaleler 3 sürümün hepsi »

Reklam İçerikli Epostaların **Metin Madenciliği** Yöntemleri ile Otomatik Tespiti [PDF] dergipark.org.tr
K ÇALIŞ, O GAZDAĞI, O YILDIZ - Bilişim Teknolojileri Dergisi, 2013 - dergipark.org.tr
... Bu çalışmada **metin madenciliği** yöntemleri kullanılarak Türkçe içerikli reklam epostalarının tespiti gerçekleştirilmiştir. Bu amaçla Destek Vektör Makinesi, k En Yakın Komşu ve Naive ...
☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 35 İlgili makaleler 4 sürümün hepsi »

Muhasebede Analiz Yöntemi Olarak **Metin Madenciliği** [PDF] dergipark.org.tr
Ş AĞDENİZ, B YILDIZ - Muhasebe Bilim Dünyası Dergisi, 2018 - dergipark.org.tr
... Büyük verinin yaklaşık %80-90'ı **metin**, ses, görüntü gibi yapısal olmayan ... **Metin madenciliği**, literatürde sıklıkla veri **madenciliğinin** bir alt dalı olarak anılsa da veri **madenciliğinden** farklı ...
☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 9 İlgili makaleler 6 sürümün hepsi »

Metin madenciliği ile Shakespeare külliyatının incelenmesi [PDF] dergipark.org.tr
S ÇELİK - MANAS Sosyal Araştırmalar Dergisi, 2020 - dergipark.org.tr
... yapılandırılmamış formatta saklanan doğal dil **metni** ile ilgilenmektedir (Weiss vd., 2010). **Metin** veri **madenciliği** olarak da bilinen **metin madenciliği**, veri **madenciliği**, makine öğrenimi, ...
☆ Kaydet 99 Alıntı yap Alıntılanma sayısı: 5 İlgili makaleler 5 sürümün hepsi »

WEB MADENCİLİĞİ

Web madenciliği, veri madenciliğinin bir alt dalı olup web üzerindeki bilgileri işleyerek analiz etmeyi amaçlar. Temel olarak 3 grup altında incelenebilir:

Web kullanım madenciliği

Web içerik madenciliği

Web yapı madenciliği

Web kullanım madenciliği (web usage mining)

çalışmalarında, kullanıcıların web sayfaları ile olan ilişkileri incelenmektedir. Örneğin kullanıcıların tıklama alışkanlıkları ve sıklıkları, dolaştıkları siteler, hangi sayfaya hangi sayfadan sonra girdikleri, en çok hangi reklamlara tıklandığı, resim içerikli mi yazı içerikli mi yoksa video içerikli mi sitelere daha çok tıkladıkları gibi sorulara cevap aranır. Bu soruların cevapları karşılıklı olarak ilişkilendirilmeye çalışılır. Örneğin "video içerikli sayfaları dolaşan kişiler mi resim yoğun siteleri dolaşan kişiler mi daha fazla internetten alışveriş yapmaktadır?" şeklinde karşılaştırmalar yapılması mümkündür.

Web kullanım madenciliği, genelde sunucu kayıtlarını (server logs), kullanıcıların bilgisayarlarına yüklenen ufak çerezleri (cookies) temel alarak istatistiksel sonuçlar üretmeye çalışır. Ayrıca günümüzde çeşitli kaynakların dağıttığı ve internet gezginine (browser) eklenerek kullanıcı hakkında istatistiksel bilgi toplayan araç çubukları (toolbars) da bulunmaktadır. Örneğin google toolbar, alexa toolbar, yahoo toolbar gibi araç çubukları kullanıcı davranışlarını istatistiksel amaçla toplamaktadır

Web yapı madenciliği (web structure mining)

Aslında bir çizge kuramı (graph theory) çalışması olarak düşünülebilir. Bu gruptaki çalışmalar, web'te bulunan kaynakları kullanarak birer çizge çıkarmayı (graphic) ve bu çizge üzerinde analizler yapmayı hedeflerler. Örneğin hangi sitelerin, hangi sitelere bağlantı (link) verdiği bilgisi bir grafik şeklinde çizilebilir. Buradan en çok bağlantı alan veya en çok bağlantı veren siteleri analiz etmek mümkündür. Benzer şekilde site içeriklerinde kullanılan bilgilerin de çizgeye dökülmesi ve analiz edilmesi mümkündür. Bir sitenin kendi içindeki bağlantı yoğunluğu veya resim yoğunluğu veya kullanıcı ile iletişimi sağlayan formların yoğunluğu site yöneticilerine veya site tasarımcılarına faydalı bilgiler sunabilir. Bu tip sitenin içeriğine yönelik analizler de yine web yapı madenciliğinin bir alanı olarak düşünülebilir.

Web içerik madenciliği (web content mining)

Çalışmaları ise web sitelerinin içeriğine yoğunlaşır. Örneğin sitenin içerisindeki sayfaların dillerini tespit etmek, kullanılan kelimelerin yoğunluğunu bulmak, otomatik olarak anahtar kelime (keyword) çıkarımı yapmak veya sitelerin kategorize edilmesi (sohbet, oyun, haber, spor vs.) bu tip çalışmalara birer örnektir. Bu çalışma grubunda içerik analizi yapılması sırasında doğal dil işleme (natural language processing) veya resim işleme (image processing) gibi konulardan istifade edilmektedir.

Yukarıdaki her üç yöntem için de genel olan bir durum ise çeşitli istatistiksel yöntemlerden yararlanıldığıdır. İstatistiksel modellerin çıkarılması ve çeşitli amaçlara yönelik olarak bu modellerin kullanılması analizlerin bir parçası olmuştur.

BİTTİ ☺