VERİ MADENCİLİĞİ

(Web Madenciliği)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Kaynak: M. Ali Akcayol, Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü Ders Notları

İçerik

- İnternet
- World Wide Web
- Web'in Oluşumu
- Web Tarayıcılar
- Web Arama Motorları
- Web Madenciliği
  - Web yapı madenciliği **(Web structure mining)**
  - Web içerik madenciliği **(Web content mining)**
  - Web kullanım madenciliği **(Web usage mining)**

İnternet

- Günümüzde World Wide Web (Kısaca Web) hayatımızın her alanında giderek yaygın bir şekilde kullanılmaktadır.
- **Web, en büyük ve yaygın kullanılan bilgi kaynağı olup arama ve bilgiye erişim hızlı ve kolay bir şekilde yapılabilmektedir.**
- Web üzerinde milyarlarca doküman (Web sayfası) bulunmakta ve milyonlarca kişi sürekli yeni dokümanlar eklemektedir.
- Web, veriye erişimi ve hızlı aramayı sağlamakla birlikte diğer kişilerle bilgi paylaşımını da sağlamaktadır.
- İnternet **diğer kişilerle sesli ve görüntülü görüşme için de kullanılmaktadır.** Bu açıdan **İnternet'in sanal bir topluluk olduğu söylenebilir.**

İnternet

- İnternet günümüzde alışveriş şeklini de değiştirmiştir.
- Mağazaya giderek alışveriş yapmak yerine bilgisayar başında ürünleri almakta ve ödemelerini yapmaktayız.
- Bankacılık, rezervasyon, ödeme başta olmak üzere tüm işlemler elektronik olarak yapılabilmektedir.
- Bu hem maliyet hem de konfor yönünden daha çok tercih edilmektedir.
- **İnternet yaşam kalitesini ve iş yapış şeklimizi de değiştirmiştir.**

World Wide Web

- **Web, kullanıcıların bir bilgisayardan diğer bilgisayarda bulunan veriye ulaşmasını sağlayan İnternet tabanlı bilgisayar ağıdır.**
- Web standart istemci-sunucu (client-server) modelini kullanmaktadır.
- Bu modelde kullanıcılar kendi bilgisayarlarındaki program ile uzaktaki bilgisayar bağlanırlar.
- Web üzerinde gezinti için tarayıcı (browser) denilen programlar kullanılır.
- Browser'lar uzaktaki bilgisayardan istekte bulunurlar ve HTML (HyperText Markup Language) biçiminde gelen bilgiyi yorumlayarak istemci taraftaki kullanıcının ekranında görüntülerler.
- Web üzerinde gezinti yapılırken dokümanlar arasındaki bağlantılar (hyperlink) kullanılır.
- Bu şekilde oluşturulan dokümanlar hypertext olarak adlandırılırlar.

Web'in Oluşumu

- **Web 1989 yılında Tim Berners-Lee tarafından bulunmuştur.** World Wide Web terimini ilk kullanan ve ilk istemci programını yazan kendisidir.
- **Tim Berners-Lee "Information Management: A Proposal" adlı bir öneriyi çalışmakta olduğu CERN laboratuarında 1989 yılında sunmuştur.**
- Bu önerisinde hiyerarşik doküman yapısının avantajlarını ve dezavantajlarını ortaya koymuştur.
- Önerilen doküman yapısıyla bağlantılar (hypertext) aracılığıyla dokümanlar arasında geçiş yapılabilmektedir.
- **Bu öneri dağıtık hypertext sistem olarak adlandırılmıştır ve günümüz Web mimarisinin temelini oluşturmaktadır.**

Web'in Oluşumu

- Başlangıçta destek bulamamış olsa da 1990 yılında Tim-Berners Lee tarafından tekrar önerilmiştir.
- Aynı yıl desteklenen proje ile günümüz Web mimarisi geliştirilmeye başlanmıştır.
- İstemci ve sunucu arasında geliştirilen protokol ile iletişim sağlanmıştır.
- Bu çalışmayla **HyperText Trasfer Protocol (HTTP), HyperText Markup Language (HTML)**ve **Universal Resource Locator (URL)**tanımlanmıştır.

Web Tarayıcılar

Mosaic ve Netscape Browser'lar

- **Web'in önemli gelişmelerinden birisi de 1993 yılında mosaic tarayıcının geliştirilmesidir.**
- Mosaic grafik arayüze sahiptir ve Unix işletim sistemi için geliştirilmiştir. Kısa süre sonra mosaic tarayıcının Windows ve Macintosh versiyonları geliştirilmiştir.
- **1994 yılının ortalarında Netscape tarayıcı geliştirilmiştir.**
- Microsoft tarafından geliştirilen **Internet Explorer tarayıcı 1995 yılında geliştirilmiştir.**
- **Web'in popüler ve başarılı olmasında en önemli aşamalardan birisi Mosaic tarayıcının geliştirilmesidir.**

Web Arama Motorları

Internet

- **Internet, Web'in iletişim ağını sağlar.**
- **Internet'e ilişkin çalışmalar ARPA (Advanced Research Projects Agency) tarafından desteklenmiştir.**
- **İlk ARPANET bağlantısı 4 node ile 1969 yılında yapılmıştır.** 1972 yılında ise 40 node ile bağlantı yapılmıştır.
- **1973 yılında Vinton Cerf ve Bob Kahn** tarafından **TCP/IP (Transmission Control Protocol / Internet Protocol)** protokolünün ilk versiyonu geliştirilmiştir.
- Geliştirilen TCP/IP protokol yığını ile birbirinden uzakta farklı ağlar içinde yer alan bilgisayarlar birbirine bağlanmıştır.
- **1982 yılında TCP/IP protokolünü kullanan Internet doğmuştur.**

Web Arama Motorları

Search Engines

- Bilginin Dünya üzerinde dağıtık ve çok büyük boyutlarda bulunmasından dolayı bilgiyi bulmak ve erişmek daha önemli hale gelmeye başladı.
- **Çok büyük bir alanda ve dağıtık bulunan bilginin bulunması için arama motorları geliştirilmeye başlanmıştır.**
- **Excite arama motoru 1993 yılında** 6 Stanford Üniversitesi öğrencisi tarafından **geliştirilmiştir.**
- 1994 yılında EINET Galaxy geliştirilmiştir ve **1994 yılında Yahoo! geliştirilmiştir.**
- Yahoo! diğer alternatiflerine göre favoriler listesi ve öneriler dizini sunmaktaydı.
- Ardından Lycos, Infoseek, Alta Vista, Inktomi, Ask Jeeves, Northernlight gibi arama motorları geliştirilmiştir.

Web Madenciliği

- **Son on yılda Web'in gelişimi sonucunda Dünya'nın en büyük veri kaynağı ortaya çıkmıştır.**
- **Web kendine özgü çok sayıda karakteristik özelliğe sahiptir ve çok büyük veri üzerinde veri madenciliği önemli ve zor bir iş haline gelmiştir.**
- Web üzerindeki veri miktarı çok büyüktür ve gün geçtikçe hızla artmaktadır. Aranan her türlü bilgi Web üzerinde bulunabilmektedir.
- **Web üzerinde yapılandırılmış tablolar, yapılandırılmış Web sayfaları, düz metinler** ve **multimedia dosyaları gibi çok farklı dosyalar bulunmaktadır.**
- **Web üzerindeki veri heterojendir.**

Web Madenciliği

- **Aynı bilgiye sahip Web sayfaları çok farklı biçimlerde ve içeriğe sahip** şekilde Web üzerinde bulunabilmektedir.
- **Bu farklılık Web sayfalarındaki bilgilerin entegrasyonunu çok zor hale getirmektedir.**
- **Web üzerindeki bilginin çok önemli bir kısmı bağlantılara**sahiptir.
- **Hyperlink'ler aynı site üzerindeki Web sayfaları arasında** veya **çok farklı sitelerdeki Web sayfaları arasında olabilmektedir.**
- Hyperlink'ler Web sayfaları için çok önemlidir.
- **Çok sayıda** Web sayfası tarafından **link verilen sayfalar otorite sayfalar**

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi gürültüye sahiptir. Gürültü iki farklı kaynaktan dolayı oluşmaktadır.**
- **Bunlardan birincisi,** Web sayfası **gezinti linkleri**, **reklamlar**, **copyright bilgileri**, **privacy bilgileri**, v.b. gibi çok farklı türde veriye sahiptir.
- İyi bir **Web**bilgisi **analizi için gürültüleri ortadan kaldırmak gereklidir**.
- **İkincisi**, **Web üzerindeki bilginin kalite kontrolü bulunmamaktadır** ve herhangi birisi istediği bilgiyi bir link üzerindeki Web sayfasına yazabilir.
- **Web üzerindeki verinin büyük bir kısmı düşük kalitede, hatalı ve eksiktir.**
- Web üzerinde ticari uygulamalar bulunmaktadır ve insanlar çok sayıda

Web Üzerindeki Verilerin Özellikleri

- **Web üzerindeki bilgi dinamiktir ve sürekli değişmektedir.**
- **Değişiklikleri anlık izlemek bazı uygulamalar için çok önemlidir.**
- **Web sanal bir topluluktur.** Web sadece insanlar arasında veri iletişimini değil **insanlar arasındaki etkileşimi de sağlamaktadır.**
- Yukarıdaki özelliklerin hepsi Web üzerindeki bilginin elde edilmesi için kullanılacak yöntemler için hem fırsatları hem de zorlukları beraberinde getirmektedir.
- **Web madenciliği, veri madenciliğinde kullanılan tüm tekniklerin uygulanmasını içermez.**
- Çok zengin ve farklı özelliklere sahip veriyi bulundurmasından dolayı **Web madenciliği kendine özgü algoritmalara sahiptir.**

Web Madenciliği

- **Web madenciliği kullanılabilir bilgiyi** Web **bağlantılarından**, **sayfa içeriklerinden** ve **kullanılan veriden** elde eder.
- Web madenciliği çok sayıda veri madenciliği tekniğini kullanır ancak **sahip olduğu verinin heterojen olması**, **yarı yapılandırılmış** veya **yapılandırılmamış** olmasından dolayı **sadece veri madenciliği uygulaması olarak görmek doğru değildir.**
- Çok sayıda veri madenciliği yöntemi son on yılda geliştirilmiştir.
- Web mining üç türde ele alınmaktadır. Bunlar;
  - **Web yapı madenciliği**
  - **Web içerik madenciliği**
  - **Web kullanım madenciliği**

  yöntemleridir.

Web yapı madenciliği

- Web yapısı madenciliği **faydalı ve kullanılabilir bilgiyi** Web sayfalarında bulunan **bağlantılardan çıkarır.**
- **Bağlantılar kullanılarak hangi sayfanın daha önemli olduğu gibi bilgiler elde edilebilir.**
- **Ayrıca aynı ortak ilgilere sahip olan benzer kullanıcıları belirleyebiliriz.**
- Klasik **veri madenciliğinde bu tür bilgiler bulunmaz.**

Web içerik madenciliği

- **Web içerik madenciliğinde** faydalı ve kullanılabilir bilgiler **Web sayfalarının içeriğinden elde edilir.**
- Örneğin **Web sayfaları içeriklerine göre sınıflandırılabilir.**
- **Bu özellikler klasik veri madenciliğinde de kullanılmaktadır.**
- Web sayfalarında **kullanıcıların forum bilgilerine müşteri görüşlerine dayanarak çıkarımlar yapılabilmektedir.**

Web kullanım madenciliği

- **Web kullanım madenciliği,**kullanıcıların **Web sayfalarına erişim bilgilerini kullanır.**
- **Kullanıcıların tıklama bilgileri**, **sayfalarda gezinme bilgileri**, **sayfalar üzerindeki etkileşim bilgileri** gibi veriler kullanılır.
- Yukarıdaki işlerin yanı sıra Web üzerindeki verilerin zengin ve çok çeşitli oluşu Web madenciliğinde çok farklı uygulama alanları oluşturmaktadır.
- **Web madenciliği süreci ile veri madenciliği süreci birbirine benzemektedir.**Sadece **veri toplama aşaması farklıdır.**
- **Klasik veri madenciliğinde veriler bir veri ambarında tutulur.**
- **Web madenciliğinde** ise veriler **dağıtık bulunan Web üzerinde bulunur** ve **toplanması çok önemli ve zor bir iştir.**
- Veriler elde edildikten sonra **ön işleme, Web madenciliği ve post-processing** işlemleri gerçekleştirilir.

Web madenciliğinin kullanım alanları

- Web madenciliğinin günümüzde birçok alanda kullanılmasının en önemli sebebi, kişilerin web sayfalarında göstermiş oldukları davranışların, hareketlerin ve yapmış oldukları işlem bilgilerinin var olan iş süreçlerine entegrasyonunu sağlayarak müşterinin en iyi şekilde anlaşılmasını sağlayan müşteri odaklı bir sistem oluşturmasıdır.
- Web madenciliği kullanım alanları aşağıdaki gibidir.
- Web üzerinden ürün satışı gerçekleştiren şirketler web verilerini analiz ederek müşteri profili ve kümeleri oluşturmaktadırlar.
- Google vd. arama motorları web içerik madenciliği uygulayarak aranan anahtar kelimeyi içeren web sitelerini belirlemektedirler.
- Web madenciliği uygulanarak web sitelerinin iyileştirilmesi ve güncel kalması sağlanmaktadır.