Web Madenciliği Teknikleri

Abdullah BAYKAL* ,Cengiz COŞKUN**

* Dicle Üniversitei Fen-Edebiyat Fakültesi Matematik Bölümü, baykal@dicle.edu.tr

** Dicle Üniversitesi Bilgi-İşlem Daire Başkanlığı , ccoskun@dicle.edu.tr

ÖZET

Veri madenciliği, kurumların ileriki kullanımlar için tahminler geliştirmelerine olanak sunan eldeki verilerinde mevcut bulunan gizli ilişkileri ortaya çıkarmayla ilgilidir. Web madenciliği ise veri madenciliği tekniklerinin world wide web verileri üzerinde uygulanmasıdır. Web madenciliği 3 farklı alt bölüm altında analiz edilebilir; Web içerik madenciliği, Web yapı madenciliği ve Web kullanım madenciliği.

ABSTRACT

Data mining is concerned with finding hidden relationships present in business data to allow businesses to make predictions for future use. Web mining is the application of data mining techniques on the world wide web data. Web mining can be analyzed under three different section. Web content mining, Web structure mining and Web usage mining.

Anahtar Kelimeler: Veri Madenciliği,Web Madenciliği, Web Kullanım Madenciliği.

Giriş

Veri Madenciliği , büyük miktardaki birbirinden ilgisiz görünen veriden anlamlı bilginin çıkarılması veya içerisinden gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların bilgisayar programı kullanarak aranmasıdır.

Günümüz bilgisayar sistemleri,gelişen donanım teknolojisi sayesinde çok büyük miktarda verinin saklanmasına izin vermektedir.

Berkeley'de yapılan bir çalışmaya göre bir yılda ortalama bir milyon terabyte veri depolanmakta ve bu sayı her geçen yıl artmaktadır. Bir çığ gibi büyüyen bu veri yığınları son derce dağınık ve düzensiz bir yapıda bulunmaktadır. Veri madenciliğinde amaç, kullanıcının bilgi çıkarma sürecinde katkısının olabildiğince az tutulması, işin olabildiğince otomatik olarak yapılabilmesidir.

Günümüzde birçok işlemin internet üzerinden yürütülmesi sonucu , çok büyük oranda veri yığınları WWW (World Wide Web) ortamında oluşmuş durumdadır.

İnternet üzerinde bir siteye bağlanan herkes bağlantı loglarını tutan sunucularda parmak izi bırakır (IP adresi, browser, cookie'ler, gibi).
Web üzerindeki veri yığınları çok farklı standart ve tiplerde bulunmaktadır. Bu veri yığınları aşağıdaki şekilde sıralanabilir.

- Web sayfaları
- Access Log dosyaları
- Kullanıcı kayıt bilgileri
- Oturum ve hareket bilgileri
- Site yapısı ve içeriği

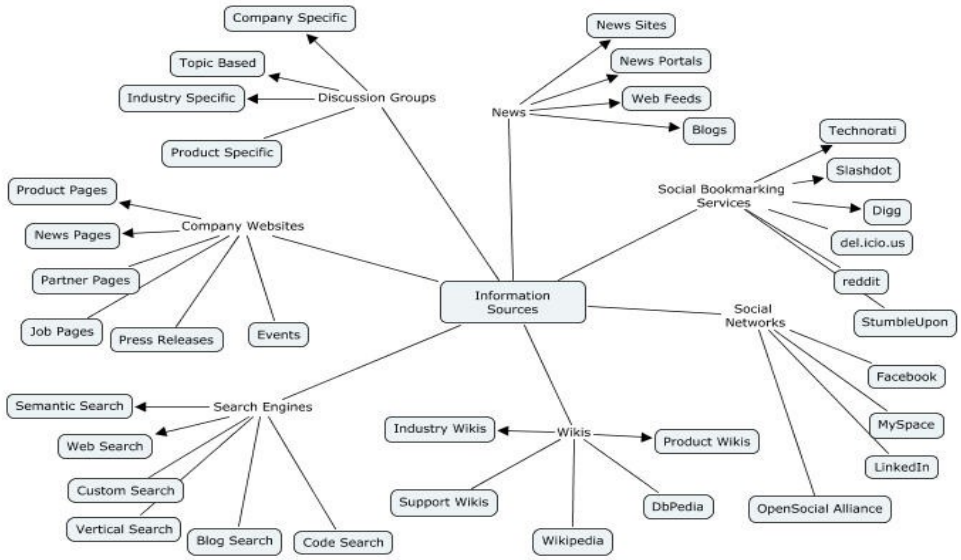
Web madenciliği yukarıda sayılan çeşitli yapıdaki web sayfaları dokümanlarını ve kayıt bilgilerini incelemek, bunlardaki kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak

tanımlanabilir. Bir web sitesinin kullanımının artırılması için , bu siteyi kullanan kullanıcıların ilişki ve hareket eğilimleri ortaya çıkarılarak çeşitli ipucları elde edilebilir. Bir online alış veriş sitesindeki kayıtlarından, alış veriş yapan kullanıcıların hareketleri incelenerek, elde edilen sonuçlar satışların artırılması için kullanılabilir.

Web Madenciliği Teknikleri

Web madenciliği , yukarıda da bahsedildiği gibi web de bulunan veriden faydalı bilgilere ulaşmaktır. Bir çok kaynağa göre web madenciliği terimini ilk kez Etzioni 1996 yılında kullanmıştır [1].

Web Madenciliğinde kullanılan veriler , web üzerinde çok geniş bir alandan toplanmaktadır Şekil 1.



Şekil 1. Web madenciliği veri kaynakları [2]

Web madenciliğinde kullanılacak verilerin yapısına göre 3 gruba ayrılır [5].

1. Web İçerik madenciliği
2. Web yapı madenciliği
3. Web kullanım madenciliği

1.Web İçerik Madenciliği

Web içerik madenciliği , yapay zeka ve akıllı yazılım programları, bilgi tarama teknikleri kullanarak web kaynaklarının içeriklerinden yararlı bilgiyi elde etmek olarak tanımlanabilir. Son zamanlarda XML dili de bu konuda kullanılmaya başlanmıştır[3].

Web üzerindeki çok farklı yapıdaki veri (metin, görsel, link , resim,..) web içerik madenciliği için yapılacak uygulamaları zorlaştırır.

Web sitelerinin dokümanlarındaki link ve hyperlinkler bularak sayfanın ve web sitesinin yapısal raporu çıkarılmaya çalışılır.

Kullanıcıların istedikleri bilgiyi bulup bulmadıkları, sayfanın yapısının çok geniş veya derin olup olmadığı, içerik elemanlarının doğru yere yerleştirilip yerleştirilmediği, az ziyaret edilen bölümlerin sayfa düzeni ile

yerleştirildiği yerin ilişkisi hakkında bilgiler ediniriz.

Web içerik madenciliğinde, araştırmamızın amacına bağlı olarak üç çeşit rapor elde edebiliriz.

- Web sayfasının hyperlinklere bağlı olarak sınıflandırılması
- Web sitesi yapısını gösterir rapor
- Belirli bir alan adının (domain) in web sitesindeki yapısal hiyerarşisi ve hyperlink ağının raporu.

Elde edilen bilgiler kullanıcılara bilgi aramada kullanmaları için görsel olarak grafik sunumlara dönüştürülür.

2. Web Yapı Madenciliği

Web yapı madenciliğin amacı web sayfaları arasındaki linkleri takip ederek bilgi üretmektir. Bir web sitesinin bağlantı yapısının analiz edilmesinde Graph Teorisi kullanılır. Web yapı madenciliği , yapısal veri tiplerine göre 2 ye ayrılır.

- Hyperlink bir web sayfasını faklı bir lokasyona yönlendiren yapısal eleman olduğu için webdeki hyperlinklerin modelinin çıkarılmasıdır,
- Web sayfası dokümanlarındaki HTML ya da XML etiketleri analiz ve tanımlarında ağaç (tree) benzeri yapıların kullanılmasıdır.

3. Web Kullanım Madenciliği

Web Kullanım Madenciliğinde kullanılan veriler, web üzerindeki çeşitli sunucularda tutulan kullanıcı erişim hareketlerinin yer aldığı çeşitli log dosyalarından elde edilir.

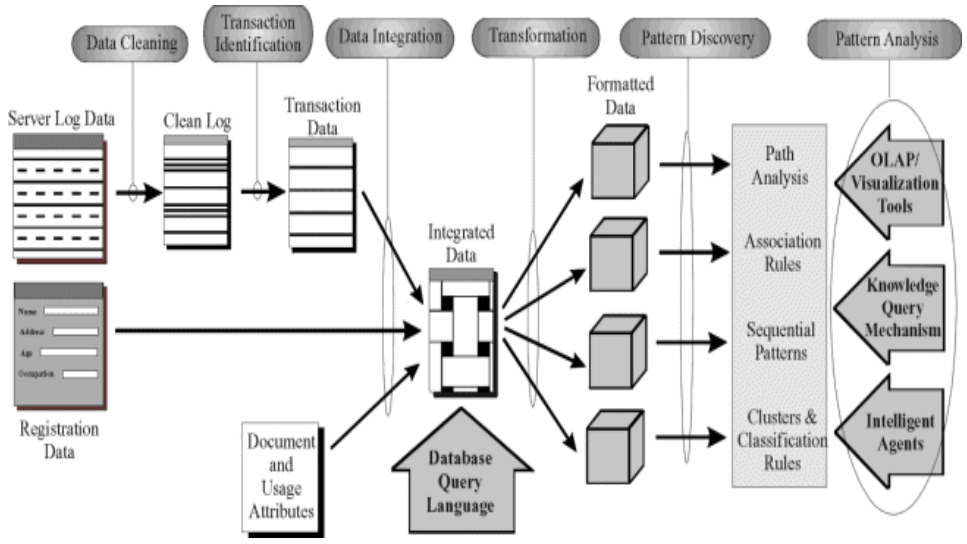
İstemcilerden gelen her istek , bir kayıt olarak , metin tabanlı log dosyalarına eklenir. Bu log dosyalarındaki kayıt desenindeki veriler kullanıcı hakkında , ayrıntılı bilgiler içerir. Log dosyasındaki kayıt formatı verilen servis çeşidine ve kullanılan işletim sistemine göre faklılıklar gösterir. Bu log dosyalarından bazıları, access log (erişim), mail log , error log, referrer log, ftp log şeklindedir. Bunun dışında , sunucu üzerinde verilen farklı servislerde isteye baglı log dosyaları tutulmaktadır. Özellikle web sunucularının (Apache, Microsoft IIS) access log dosyaları, tutukları veriler nedeniyle , web madenciliğinde önemli bir veri kaynağı olmaktadır.

Web Kullanım madenciliği işlemleri 2 ana bölümden oluşur Şekil 2. İlk bölüm, web verilerini uygun yapılara dönüştüren web sitesi bağımlı işlemleri kapsar.

Bu bölüm, ön işleme, işlem tanımlanma ve veri birleştirme işlemlerinden oluşur. İkinci bölüm ise web sitesinden bağımsız olarak, veri madenciliği temelini oluşturan uygulamaları içerir.

Ön işleme kısmında yer alan veri temizleme web kullanım madenciliğinde gerçekleştirilen ilk basamaktır. Bu basamakta, web loglarında yer alan ilgisiz öğeler temizlenir. Geçersiz web adresleri kontrol edilir ve ayıklanır.Log kayıtlarında bulunan ,jpg, JPEG, jpeg, gif, GIF gibi benzer dosya uzantıları düzeltilir. Proxy sunucular üzerinden gelen kullanıcıların, log dosyalarındaki aynı kayıt gibi görünmesi giderilir. Çoklu log kayıtlarını birleştirme, bütünleştirme gibi alt seviye veri bütünleştirme işlemleri bu basamakta yapılır.

Veri temizliğinden sonra ,log kayıtları bir veya birçok işlem tanımlama uygulamaları kullanılarak mantıksal gruplara ayrılır. Temizlenmiş sunucu logları iki şekilde düşünülebilir; birçok sayfa referansının tek bir işlemi yada her biri tekbir sayfa referansından oluşan bir işlemler kümesi şeklinde.



Şekil 2. Web Kullanım Madenciliği Genel Mimarisi [4]

İşlem tanımlama (Transaction identification) 'nın amacı her kullanıcı için anlamlı referans grupları yaratmaktır. Bu yüzden, işlem tanımlama işi, büyük bir işlemi çok sayıda küçük işleme bölmek, yada küçük işlemleri birleştirerek daha az sayıda büyük işlem oluşturma işidir. Bu işlem, verilen veri madenciliği işine uygun işlemler yaratmak için birkaç kez -bütünleştir yada böl- basamaklarına kadar genişletilebilir.

Web madenciliği işlemi için sadece Erişim log kayıtları kaynak olmayabilir. Örneğin, herkese açık olmayan, kayıtlı kullanıcıların kullanabildiği uygulamaları çalıştıran sunuculardaki kullanıcı kayıt bilgileri erişim log kayıtlarıyla birleştirilebilir.

Referans sayfalarına, sayfa tipleri, sınıflandırma, kullanım sıklığı, sayfa meta bilgileri ve link yapıları dahil edilebilir.
Alan adına bağımlı verilerin birleştirme aşaması tamamlandığın da , ortaya çıkan işlem verileri ilgili veri madenciliği kurallarına uygun bir şekilde formatlanmalıdır. Örneğin, birleştirme kurallarının belirlenmesi formatı, Sıralı Model (Sequential Patterns) incelerken gereken formattan farklı olabilir.
Model Keşfi (Pattern Discovery) için sınıflandırma , istatistiksel analiz, kümeleme,ilişkilendirme kuralları gibi birçok yöntem kullanılabilir.

Model Analizi(Pattern Analyis) web madenciliğinin son aşamasıdır. Bu aşamada, OLAP, SQL veri tabanı uygulamaları ve görselleştirme araçları kullanılarak amaca uygun çeşitli filtreler uygulanır [6].

SONUÇ

Web madenciliği günümüzde internetin yoğun bir şekilde kullanımının artması nedeniyle üzerinde önemli ölçüde araştırma yapılan bir alan haline gelmiştir.

Web madenciliği , kullanıcıların web sitesindeki davranışlarını çeşitli kayıtlar üzerinden inceleyerek web sitelerinin yeniden tasarım yada geliştirilmesi konusunda ipucları sunar.

Her gün web de sunucular üzerinde kullanıcılara ait, web madenciliği araştırması yapılabilecek milyarlarca kayıt oluşmaktadır.

Referanslar

[1] Oren Etzioni, The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65

[2] dorai.wordpress.com, Erişim tarihi Aralık/2008

[3] Hidayet Takıcı, “Kütüphane Kullanıcıların Erişim Desenlerinin Keşfi”

[4] http://maya.cs.depaul.edu/webminer/survey/img13.gif Erişim tarihi Aralık/2008

[5] Madria, S. K., Bhowmick, S. S., Ng, W. K., Lim, E. P., (1999). Research Issues in Web Data Mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, 303-312.

[6] Resul Daş,İbrahim Türkoğlu ve Mustafa POYRAZ, “Web Kayıt Dosyalarından İlginç Örüntülerin Keşfedilmesi”, Fırat Üniv.Fen ve Müh. Bil. Dergisi 19 (4), 493-503,2007