itüdergisi/***d***
mühendislik
Cilt:3, Sayı:6, 15-24
Aralık 2004

Web kullanıcılarının davranışları için örüntü bulma ve modelleme

Şule GÜNDÜZ*, Eşref ADALI

İTÜ Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34469, Maslak, İstanbul

Özet

İnternetin yaygınlaşması ve her alanda bilgi sağlaması günlük yaşantımıza hızla girmesine neden olmuştur. Haber, ekonomi, kültür, eğitim, sağlık hizmetler ve reklam gibi bir çok alanda bilgi kaynağı olan İnternet ortamında, kullanıcı kendisi için gerekli bilgileri bulmakta çoğu zaman zorlanmaktadır. Bunun nedeni sorgulama araçlarının kısıtlı olması ve bilgilerin fazlalığı olarak görülmektedir. Bu çalışmada kullanıcının bir sonraki istek yapacağı sayfayı öngörerek hızlı ve yüksek oranda doğru öneri yapabilen bir yöntem önerilmiştir. Model tabanlı demetleme yönteminden yaralanarak, kullanıcı oturumları aynı demette bulunan oturumlardaki ortak sayfalarda benzer süreler geçirilmesine göre demetlenmiştir. Ortaya çıkan demetler yeni kullanıcılar için öneri kümesi oluşturmak için kullanılmıştır.
***Anahtar Kelimeler**: Web kullanım madenciliği, kullanıcı örüntüleri, model tabanlı demetleme, Poisson dağılımı.*

Pattern extraction and modelling of the behavior of Web users

Abstract

Making recommendation requires predicting what is of interest to a user at a specific time. Even the same user may have different desires at different times. It is important to extract the aggregate interest of a user from his or her navigational path through the site in a session. In this paper, we present a new model that uses only the visiting time and visiting frequencies of pages without considering the access order of page requests in user sessions. The resulting model has lower run-time computation and memory requirements, while providing predictions that are at least as precise as previous proposals. Our objective in this paper is to assess the effectiveness of non-sequentially ordered pages in predicting navigation patterns. The key idea behind this work is that user sessions can be clustered according to the similar amount of time that is spent on similar pages within a session. We first partition user sessions into clusters such that only sessions which represent similar aggregate interest of users are placed in the same cluster. We employ a model-based clustering approach and partition user sessions according to similar amount of time in similar pages. In particular, we cluster sessions by learning a mixture of Poisson odels using Expectation Maximization algorithm. The resulting clusters are then used to recommend pages to a user that are most likely contain the information which is of interest to that user at that time.
***Keywords:** Web usage mining , usage patterns, model based clustering, Poisson distribution.*

*Yazışmaların yapılacağı yazar: Şule GÜNDÜZ. gunduz@cs.itu.edu.tr; Tel: (212) 285 67 03.
Bu makale, birinci yazar tarafından İTÜ Elektrik-Elektronik Fakültesi'nde tamamlanmış olan "Recommendation models for web users: User interest model and click stream tree" adlı doktora tezinden hazırlanmıştır. Makale metni 02.10.2003 tarihinde dergiye ulaşmış, 14.11.2003 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 31.05.2005 tarihine kadar dergiye gönderilmelidir.

Giriş

Web madenciliği, ilk olarak Etzioni tarafından Web döküman ve servislerinden otomatik olarak bilginin elde edilmesi olarak tanımlanmıştır (Etzioni, 1996). İnternetin hızla yaygınlaşması, Web madenciliği çalışmalarına hız kazandırmıştır. Web madenciliği, *Web yapı madenciliği*, *Web içerik madenciliği* ve *Web kullanım madenciliği* olarak üç sınıfa ayrılabilir (Madria vd., 1999). İnternette erişilebilen veri miktarı zaman içinde hızla artmaktadır. Verilere ulaşım sürecinde kullanıcıyı doğru ve hızlı olarak yönlendirmek hem Web sitesinin etkili kullanımı açısından hem de elektronik ticaret sitelerinin amaçlarına ulaşmaları açısından önemlidir. Bu konu Web kullanım madenciliğinin uygulama alanlarından biri olan öneri modelleriyle çözümlenebilir.

Öneri modelleri için, Web kullanım madenciliğinin en önemli aşamalarından biri Web kullanıcılarının davranışlarını inceleyerek uygun örüntülerin bulunmasıdır. Doğru ve hızlı öneri yapabilmek için bulunan örüntülerin hem kullanıcı davranışlarını etkin biçimde modellemesi hem de önerinin çevrimiçi durumda üretildiği düşünülürse hızlı öneri kümesi üretmeye uygun olması gerekmektedir.

Bu alanda en çok kullanılan yöntemler, Markov modelleri (Sarukkai, 2000, Deshpande vd., 2001), sıralı örüntüler (Pitkow vd., 1999), bağıntı modelleri (Mobasher vd. 2001) ve demetleme (Mobasher vd., 2000, Nasraoui vd., 1999) yöntemleridir. Ancak bu çalışmaların çoğunda kullanıcı oturumları, kullanıcı davranışlarını belirlemek açısından hem yeterli değildir hem de modeller daha sonra öneri kümesi oluşturmak üzere geliştirilmemiştir.

Bir Web kullanıcısının, bir sitede tek bir ziyareti sırasında uğradığı sayfalar, kullanıcı oturumu olarak nitelendirilir. Kullanıcının bir Web sitesini ziyareti sırasında her sayfada geçirdiği süre kullanıcı oturumlarını modellemekte önemli bir rol oynar (Shahabi vd., 1997). Bir kullanıcı bir Web sitesine her bağlanışında farklı konulara (ürünlere) ilgi duyabilir. Genelde, kullanıcılar ilgi duydukları konular ya da ürünler hakkında bilgi vermede istekli değillerdir. Bu durumda kullanıcıların o andaki ilgilerini ziyaret ettikleri sayfalarda geçirdikleri süre ile ölçebiliriz. Eğer bir kullanıcı bir sayfada daha uzun zaman geçiriyorsa, o sayfadaki konulara daha çok ilgi duyuyordur diyebilriz.

Bu çalışmamızda, kullanıcının ziyaret ettiği sayfalarda geçirdiği süreye göre yeni bir öneri modeli oluşturulmuştur. Kullanıcının ziyaret ettiği sayfaların sırasını modellemek yerine, kullanıcının bir oturumdaki ilgisi için matematiksel bir model oluşturulmuştur. Sayfa sıralarını modellemek çevrimiçi çalışma sırasında öneri oluşturma süresini geciktirdiğinden, daha hızlı ve güvenilir öneri yapabilmek için bu yöntem kullanılmıştır. Kullanıcı oturumlarının oluşmasında iki aşama olduğu varsayılmıştır: 1) Kullanıcı bir Web sitesine bağlandığında oturumu belli bir olasılıkla, demetlerden birine atanır; 2) Kullanıcının sayfaları ziyaret ve sayfalarda kalış süresi, o demete ilişkin Poisson parametreleri kullanılarak oluşturulur. Hem gerçek demet etiketleri hem de demetlere ilişkin Poisson dağılımlarının parametreleri bilinmediğinden, "Expectation Maximization" (EM) algoritması kullanılarak bu bilgiler öğrenilmiştir (Dempster vd., 1977). Oluşturulan demetler için çevrimiçi çalışmada hızlı ve doğru öneri kümesi oluşturmak için sayfa öneri puanları üç farklı yöntem kullanılarak hesaplanmıştır. Yeni bir kullanıcı siteye bağlandığında, oturumu önceden oluşturulan demetlerden birine atanır ve bu demete ilişkin öneri puanları kullanılarak bir öneri kümesi oluşturulur. Bu çalışmanın amacı model doğruluğunu artırmak için uygun veri temizleme tekniklerinin kullanılması, kullanıcı oturumlarının Poisson dağılımı kullanarak modellenebileceğini göstermek ve süre bilgisinin modelin verimini artırdığını doğrulamaktır. Bunun için üç değişik Web sitesi verisi kullanılarak deneyler yapılmış ve iki farklı model ile sonuçlar karşılaştırılmıştır.

Web madenciliği

Giriş bölümünde belirtildiği gibi Web madenciliği üç sınıfa ayrılabilir: 1) Web yapı madenciliği; 2) Web içerik madenciliği; 3) Web kullanım madenciliği. Bu bölümde bu konular hakkında kısaca bilgi verilerek, öneri modellerinin yöntemler dizisi incelenecektir.

Web yapı madenciliğinin amacı, Web sitesi ve Web sayfalarının birbirleri ile bağlantısına bakarak bilgi üretmektir. Web yapı madenciliği Web dökümanları arasındaki bağ yapısını inceleyerek Web sayfaları hakkında bilgi üretir. Web sayfasının yapısını inceler, aynı Web sitesi içindeki Web sayfalarını birbirine bağlayan bağların frekansını, bir Web sitesinden diğer Web sitelerine olan bağların frekansını, özdeş Web sayfalarını belirler.

Web içerik madenciliği Web dökümanlarından yararlı ve gerekli bilgiyi elde etmek için kullanılır. Web üzerindeki metin, işitsel, görsel, yardımcı veri, bağ verisi gibi farklı tiplerde ve yapısız veri, Web içerik madenciliği için uygulanan yaklaşımları karmaşıklaştırır. Bu konudaki yaklaşımlar iki açıdan incelenebilir: 1) Bilgiye erişim, 2) Veri tabanı. Bilgiye erişim yaklaşımları, bilgi bulma yöntemlerini destekleme, geliştirme ve bulunan bilgiyi amaca uygun olarak süzme konularındaki çalışmalarda yoğunlaşırken, veri tabanı yaklaşımları Web üzerinde bulunan verinin modellenmesi, yönetimi ve sorgulanması konularında çalışmalar yapar.

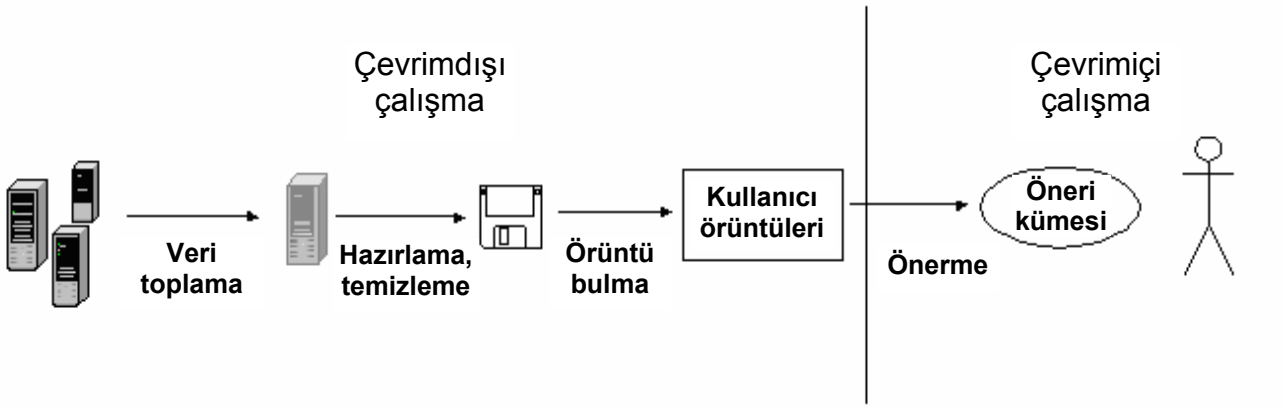
Web kullanım madenciliği veri madenciliği teknikleri kullanılarak, Web kullanıcılarının İnternetteki davranışları için örüntü oluşturmak biçimimde tanımlanabilir. Web kullanım madenciliği, Web yapı madenciliği ve Web içerik madenciliğinden farklı olarak, Web üzerindeki doğrudan erişilebilen veriyi kullanmak yerine, kullanıcıların Web'de dolaşırken hareketlerinden oluşturulan veriden bilgi üretir. Bu konudaki çalışmalar genel site güncelleme sistemleri, sistem iyileştirme ve kişiselleştirme, sahtekarlık ve izinsiz giriş sezisi, kullanıcının bir sonraki faaliyetini öngörme, cep belleğe alma ve önceden getirme başlıkları altında toplanabilir.

Bu çalışmanın konularından biri olan öneri modelleri, Web kullanım madenciliğinin uygulama alanlarından biridir. Öneri modelleri oluşturma çalışması Şekil 1'de gösterildiği gibi dört aşamadan oluşur: 1) Veri toplama; 2) Veri temizleme ve hazırlama; 3) Örüntü bulma ve 4) Önerme.

Veri toplama, hazırlama ve temizleme ve örüntü bulma aşamaları çevrimdışı çalışan aşamalar oldukları için, hızın bu aşamalarda önemi yoktur. Ancak, sitede kullanıcının ziyaret edebileceği sayfalardan veya ürünlerden oluşan bir öneri kümesi oluşturma çevrimiçi bir çalışma olduğundan hızlı olmak zorundadır. Bu durumda kullanıcı örüntülerinin temsil edilebilmesi için basit modellerden oluşturulması gerekmektedir.

Model tabanlı demetleme

Birbirlerine oldukça benzeyen, diğer bir deyişle öznitelik uzayında tanımlanmış bir ölçüte göre birbirlerine yeterince yakın olan nesneleri bir araya getirmeye demetleme denir. Sonuçta ortaya çıkan demetlerin içindeki nesneler verilen ölçüte göre farklı demetlerde bulunan nesnelere göre birbirlerine daha yakınlardır. Demetleme algoritmaları buluşsal yöntemlerden, kurallı yordamlara dayanan istatistiksel yöntemlere kadar geniş bir aralıkta incelenebilir.



Şekil 1. Öneri modelleri için yöntemler dizisi

Model tabanlı demetleme yöntemleri eldeki veri kümesine en iyi uyacak matematiksel bir model bulmaya çalışır. Bu tür yöntemler genelde veri kümesinin belirli olasılık dağılımlarının karışımlarından türediğini varsayar. K nesneden oluşan $\boldsymbol{D}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, ...,\boldsymbol{x}_K\}$ şeklinde bir veri kümesindeki her $\boldsymbol{x}_i$ nesnesi $\boldsymbol{\Theta}$ parametre kümesiyle tanımlanan bir olasılık dağılımından oluşturulur. Olasılık dağılımının $c_j \in \boldsymbol{C}=\{c_1, c_2, ...,c_G\}$ şeklinde G adet bileşeni vardır. Her $\boldsymbol{\Theta}_g$, $g \in [1, ...,G]$ parametre kümesi g bileşeninin olasılık dağılımını belirleyen, $\boldsymbol{\Theta}$ kümesinin ayrışık bir alt kümesidir. Herhangi bir $\boldsymbol{x}_i$ nesnesi öncelikle, $p(c_g|\boldsymbol{\Theta}) = \tau_g$, $\Sigma_G \tau_g=1$ olacak şekilde bileşen katsayısına (ya da bileşenin seçilme olasılığına) göre bir bileşene atanır. Bu bileşen kendi parametrelerine göre $p(\boldsymbol{x}_i|cg; \boldsymbol{\Theta}_g)$ olasılık dağılımına göre $\boldsymbol{x}_i$ değişkenini oluşturur. Böylece bir $\boldsymbol{x}_i$ nesnesinin bu model için olasılığı bütün bileşenlerin olasılıklarının toplamıyla ifade edilebilir:

$$p(x_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} p(c_g \mid \boldsymbol{\Theta}) p(\mathbf{x}_i \mid c_g; \boldsymbol{\Theta}_g) \quad (1)$$

$$p(x_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} \tau_g p(\mathbf{x}_i \mid c_g; \boldsymbol{\Theta}_g) \quad (2)$$

$$p(\boldsymbol{\Theta}_1,...,\boldsymbol{\Theta}_G; \tau_1,...,\tau_G \mid D) = \prod_{i=1}^{K} \sum_{g=1}^{G} \tau_g p(\mathbf{x}_i \mid c_g; \boldsymbol{\Theta}_g) \quad (3)$$

Modelin parametre kümesi, bileşenlerin seçilme olasılığı ve her bileşenin parametre kümesinden oluşur:

$$\boldsymbol{\Theta} = \{\tau_1,...,\tau_G, \boldsymbol{\Theta}_1,...,\boldsymbol{\Theta}_G\} \quad (4)$$

Veri temizleme ve hazırlama

Bu çalışmada modelin değişik Web siteleri üzerindeki başarımını görmek amacıyla üç değişik Web sitesinin sunucu erişim günlükleri kullanılmıştır. Birinci veri kümesi NASA Kennedy Space Center için Temmuz-Ağustos 1995 zaman aralığında tutulan sunucu erişim günlüğüdür. İkinci veri kümesi ClarkNet İnternet servis sağlayıcısının Ağustos-Eylül 1995 zaman aralığındaki sunucu erişim günlüğü kayıtlarıdır. Son olarak Saskatchewan Üniversitesi tarafından Haziran-Aralık 1995 ayları arasında toplanmış sunucu erişim günlükleri kullanılmıştır. Bunlar İnternet üzerinden kolayca erişilebilen, diğer çalışmalarda da kullanılmış sunucu erişim günlükleridir. Veriler genelde buluşsal yöntemler kullanılarak temizlenmiştir. Veri temizleme ve hazırlama süreci genelde sunucu erişim günlüklerinde bulunan ham veriyi temizleyip örüntü bulma tekniklerine uygun hale getirmekten oluşur.

Web sunucuları kendilerine gelen her bir istem için erişim kütüğüne bir kayıt düşerler. Bu kayıtları oluşturan alanlar şunlardır: 1) Kullanıcı IP adresi; 2) Kullanıcı tanımı; 3) İsteğin sunucuya eriştiği zaman; 4) Dosyayı isteme yöntemi ve dosya adı; 5) Protokol; 6) Durum; 7) Dosya büyüklüğü. Bu çalışmanın amacı Web sitesinin içeriğine yönelik Web sayfalarını kullanıcıya önermek olduğu için Web erişim günlüğündeki bazı kayıtlar ve kayıtların bazı alanları örüntü bulmak için gereksizdir. Sunucu erişim günlüğü, sadece "GET…html" şeklinde dosyalar kalacak şekilde temizlenmiştir. Aynı zamanda geçersiz kullanıcı istek kayıtları da temizlenmiştir.

Veri temizleme ve hazırlama adımında yapılması gereken bir diğer işlem kullanıcı oturumlarının belirlenmesidir. Bir oturum herhangi bir işlemin başlangıcından sonuna kadar geçen dönem olarak tanımlanabilir. Web ortamında bir oturum kullanıcının bir Web sitesine girdiği ve ayrılana kadar kaldığı süre içinde yaptığı etkinlikler olarak tanımlanabilir. Ancak hem başka Web sitelerinin sunucu kayıtlarına erişmek mümkün olmadığından, hem de bir kullanıcı bir Web sitesini bir veya daha fazla ziyaret edebileceğinden, kullanıcının bir Web sitesi için oturumunun hangi sayfada başlayıp hangi sayfada bittiğini kestirmek oldukça güçtür. Bu konudaki çalışmalarda genelde, aynı kullanıcıya ait iki istek arasında 30 dakikadan daha fazla süre geçtiğinde o kullanıcı için yeni bir oturum başlatılmasının uygun olduğu belirtilmektedir (Cooley vd., 1999). Bu çalışmada da aynı yaklaşım kullanılmıştır.

Bu çalışmada, kullanıcıların ziyaret ettikleri sayfalarda geçirdikleri süre bilgisinden yararlanmak istediğimizden, bu bilgi veri temizleme ve hazırlama adımında hesaplanır. Sayfa ziyaret

süresi peşpeşe gelen iki Web sayfasının istek zamanlarının farkı olarak hesaplanır. Ancak ağ hızının zaman içinde değişmesi, kullanıcıların alışkanlıkları gibi faktörler nedeniyle bu süreyi çalışmada kullanmak doğru olmaz. Hesaplanan süreyi belli aralıkta normalleştirmek gerekmektedir. Normalleştirilen sayfa ziyaret süresinin en küçük değeri 1 olacak şekilde normalleştirme yapılmıştır. En büyük değer için 2, 3, 5, 10 gibi farklı değerler denenmiştir.

Veriyi temizlemekte en önemli adım farklı dosya isimlerine sahip, ancak kullanıcı tarayıcısında aynı görüntüyü veren sayfaları bulmaktır. Bu, ancak Web sitesi içindeki tüm sayfaları tarayarak gerçeklenebilir. Bu amaç için bir tarayıcı geliştirilmiştir. Bu yazılım ana sayfadan başlayarak Web sitesi içindeki bütün bağların gösterdiği sayfaları tarayarak sayfa numarası, dosya adı, taranılan sayfada kaç tane başka sayfaya bağlantı olduğunu ve o sayfayı tanımlayabilecek 5 anahtar sözcüğü istenen dosyaya yazar. Ancak bu tarayıcıyı sadece NASA Web sitesinde kullanmak mümkün olmuştur, çünkü çalışmada kullanılan diğer veriler güncel Web sitelerine ait değildir. Diğer Web siteleri için basit dizgi eşleme yöntemi kullanılmılştır.

Bu işlemler sonunda her Web sitesi için sayfalar kümesi şu şekilde elde edilir:

$$\mathbf{P} = \{p_1, p_2, ..., p_n\} \tag{5}$$

Her Web sitesi için bulunan oturum sayısı |S| olursa her oturum bir vektör olarak ifade edilebilir. Bu durumda bir kullanıcı oturumu şu şekildedir:

$$S_i = [t_{p_1}, t_{p_2}, ..., t_{p_n}], \quad i \in S_{|S|} \tag{6}$$

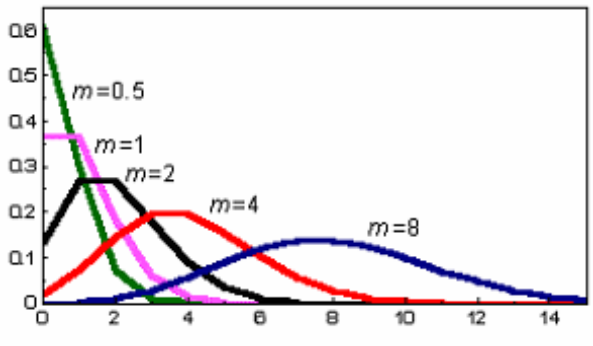
t_{p_j} değeri, eğer S_i oturumunda p_j sayfası ziyaret edilmişse o sayfaya ait normalleştirilmiş sayfa ziyaret süresini, eğer ziyaret edilmemişse sıfır değerini alır.

Kullanıcı oturumlarını demetleme

Bu çalışmada kullanıcı oturumlarının *G* bileşenli karışım bir modelden geldiği varsayılmıştır. Her bileşen bir demete karşılık gelmektedir. Demetleme sorunu, hem hangi kullanıcı oturumunun hangi demete ilişkin olduğunu bulmak hem de bu demetlere ilişkin model parametrelerini bulmaktır. Böyle demetleme sorunları için EM algoritması sıkça kullanılan bir yöntemdir. Başlangıçta model parametreleri için bir başlangıç değeri saptanır. Algoritma iki adımdan oluşur. İlk adımda model parametreleri sabit varsayılarak nesnelerin (kullanıcı oturumlarının) hangi demete ne kadar olasılıkla atandıkları hesaplanır. İkinci adımında (3) ile verilen eşitlik en büyük olacak şekilde model parametreleri güncellenir. Bu iki adım model parametrelerinde değişiklik olmayana kadar ya da değişiklik tanımlanan bir aralıkta kaldığı sürece devam eder.

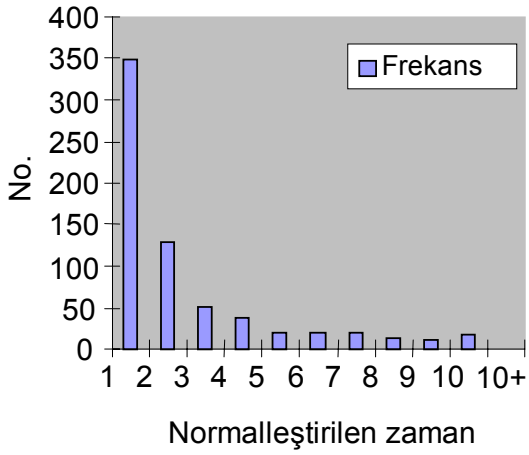
Algoritmayı başlatmak için, demet sayısı ve model başlangıç parametreleri gibi başlangıç koşullarını seçmemiz, buna ek olarak verinin hangi dağılımdan geldiğini belirtmemiz gerekmektedir. Değişik demet sayıları denenerek algoritma çalıştırılmıştır. Model başlangıç parametreleri ise, ilk başta sadece bir demet olacak şekilde hesaplanmış, daha sonra hesaplanan bu değerlere rastgele değerler eklenip çıkartılarak demet sayısı kadar parametre kümesi elde edilmiştir. Burada en önemli adım, verinin hangi dağılımdan geldiğini kestirmektir. Web sitesi içinde çok sayıda sayfa olabileceği göz önünde bulundurulursa, çok boyutlu bir veri uzayı ile ilgilenildiği açıktır. Çok boyutlu verilerin olasılık dağılımını bulmak oldukça güçtür. Olasılık uygulamalarında kullanılan bir varsayımla, her boyutun birbirinden bağımsız olduğu varsayılabilir. Bu durumda bir nesnenin olasılığı, nesneyi oluşturan boyutların olasılıklarının çarpımı şeklinde yazılabilir. Böyle bir varsayım yaptığımızda problem her boyut için bir olasılık dağılımı bulmaya indirgenmiş olur. Bu çalışmada Poisson dağılımının tanımından da yararlanılarak her boyuttaki verinin bir Poisson dağılımdan geldiği varsayılmıştır. $m>0$ parametreli bir Poisson dağılımına göre herhangi bir X rastgele değişkenin k değerini alma olasılığı şu şekilde yazılabilir:

$$p(X = k) = \frac{m^k e^{-m}}{k!} \quad k=0,1,... \tag{7}$$



Şekil 2. m'nin seçilmiş değerleri için Poisson dağılımı

Bir kullanıcı oturumdaki herhangi bir boyutun Poisson dağılımından gelip gelmediğini anlamak için, her boyutun aldığı değerlere ait histogramlar çizilmiştir. Ortaya çıkan histogramların şekli bizim yaklaşımımızın doğruluğunu kanıtlamıştır. Şekil 2 bazı *m* parametre değerleri için Poisson dağılımının şeklini göstermektedir. Şekil 3 ise NASA verisinden elde edilmiş bir kullanıcı oturumunda bir sayfa için çizilmiş histogramı göstermektedir. Normalleştirme 1 ile 10 değerleri arasında yapılmıştır. Şekil 3'te görüldüğü gibi histogram küçük parametreli bir Poisson dağılımı biçimindedir.



Şekil 3. NASA verisinden bir sayfaya ilişkin histogram

Algoritmanın çalışması sonucunda her demet için demetin seçilme olasılığı, τ_g, ve her demete ilişkin parametre kümesi, $\boldsymbol{\Theta}_g$ öğrenilir. Her demet için parametre kümesi şu şekilde gösterilebilir:

$$\mathbf{p}_{c_g} = \{\tau_g, \boldsymbol{\Theta}_g\} \tag{8}$$

$\boldsymbol{\Theta}_g$ site içindeki her sayfa için EM algoritması ile öğrenilen Poisson dağılımının parametrelerinden oluşur. *n* tane sayfa bulunan bir site için $\boldsymbol{\Theta}_g$ parametresi şöyle yazılabilir:

$$\boldsymbol{\Theta}_g = \{\theta_{g1}, \theta_{g2}, ..., \theta_{gn}\} \tag{9}$$

Böylece S_i, $i \in |S|$, oturumunun G demetten oluşan Poisson modelden gelme olasılığı aşağıda belirtildiği gibidir::

$$p(S_i \mid \Theta) = \sum_{g=1}^{G} \prod_{j=1}^{n} \tau_g \frac{\theta_{gp_j} e^{-\theta_{gp_j}}}{t_{p_j}!} \tag{10}$$

Demet özellikleri

Öneri kümesi üretme çevrimiçi çalışma sırasında olduğundan, bu kümeyi hızlı üretmek önemlidir. Öneri kümesi kullanıcı oturumunda o ana kadar henüz ziyaret etmediği, ancak modelin, kullanıcının bir sonraki adımda ziyaret edebileceğini öngördüğü sayfalardan oluşur. Aynı zamanda yapılan önerilerin doğruluk başarımının da yüksek olması gerekmektedir. Hızlı ve doğru öneri kümesi üretmek için, her demette her sayfa için öneri puanı hesaplamak gerekmektedir. Böylece her demet için bir öneri puanı kümesi olan, $\mathbf{RS}_g = \{rs_{g1}, rs_{g2}, ..., rs_{gn}\}$, oluşturulur. Her demet için parametre kümesi şu şekilde yazılabilir:

$$\mathbf{p}_{c_g} = \{\tau_g, \boldsymbol{\Theta}_g, \mathbf{RS}_g\} \tag{11}$$

Modelin öneri yapmak için bellekte saklaması gereken bütün parametreler bunlardır. Bu çalışmada bellekte saklanması gereken parametre sayısı model büyüklüğü olarak tanımlanmıştır. Model büyüklüğü ne kadar küçük olursa çevrimiçi çalışmanın o kadar hızlı olacağı açıktır.

Bu çalışmada öneri puanı hesaplamak için üç değişik yöntem kullanılmıştır. Daha sonra öneri puanları en yüksek puan 1 olacak şekilde normalleştirilmiştir. Büyükten küçüğe doğru sıralanan öneri puanları çevrimiçi çalışmada öneri kümesi oluşturma zamanını daha da kısaltır.

Yöntem 1: İlk yöntem için sadece Poisson dağılımının parametreleri kullanılarak öneri kümesi oluşturulur. Bu durumda her sayfa için öneri puanı şu şekilde yazılabilir:

$$rs_{gi} = \theta_{gi}, \qquad g \in [1,...,G], i \in [1,...,n] \qquad (12)$$

Bundan sonraki yöntemler için her kullanıcı oturumunun her demet için olasılığı hesaplanır ve olasılığı en yüksek demete atanır. Her demet içinde her sayfaya yapılan istek sayısı, $R_{p_i}, i \in n$, bütün sayfalara yapılan toplam istek sayısı, R_p hesaplanır. Her sayfa için popülerlik katsayısı diye adlandırılan bir terim tanımlanır ve her sayfanın popülerliği:

$$f_{gi} = \frac{R_{p_i}}{R_p} \qquad (13)$$

olarak hesaplanır.

Yöntem 2: İkinci yöntemde öneri puanları popülerlik katsayıları kullanılarak hesaplanır.

$$rs_{gi} = f_{gi} \qquad (14)$$

Bu yöntemde popülerlik katsayısı kullanılmasının nedeni ise bir demet içinde birçok kullanıcı tarafından ziyaret edilen sayfaları önermektir.

Yöntem 3: Son yöntemde popülerlik katsayıları Poisson parametreleri ile çarpılarak öneri puanları hesaplanır:

$$rs_{gi} = \theta_{gi} * f_{gi} \qquad (15)$$

Öneri kümesi oluşturma

Modelin çevrimiçi çalışan bölümü öneri kümesi oluşturan bölümdür. Yeni bir kullanıcı geldiğinde ilk önce oturumu belli bir olasılıkla bir demete atanır. Daha sonra o demetin öneri puanları kullanılarak, kullanıcının bir sonraki adımı model tarafından öngörülür ve bir öneri kümesi oluşturulur. Öneri kümesi en yüksek öneri puanına sahip *N* sayfa olacağı gibi belli bir eşik değerinden daha yüksek öneri puanına sahip sayfalar da olabilir. Hangi yaklaşımın kullanılacağı bir sonraki bölümde açıklanacak olan değerlendirme ölçütlerine göre değişir.

Deney sonuçları

Bu çalışma için üç farklı sunucu erişim günlükleri kullanılmıştır (Tablo 1). Veriler temizlendikten ve kullanıcı oturumu olarak düzenlendikten sonra yaklaşık olarak %30'u deneme kümesi, %70'i ise öğrenme kümesi olarak ayrılmıştır. Model parametreleri öğrenme kümesi kullanılarak öğrenilmiş, sonuçlar ise deneme kümesi üzerinde değerlendirilmiştir. Modeli değerlendirmek üzere iki ölçüt kullanılmıştır: 1) Doğruluk; 2) Başarı oranı.

Doğruluk (DK): Deneme kümesindeki her kullanıcı oturumunun ilk *w* sayfası seçilir. Bu bölüm kullanılarak oturum bir demete atanır. Model bu demette öneri puanı ζ değerinden daha büyük sayfalardan bir öneri kümesi oluşturur. Kullanıcı oturumunun kalan ikinci bölümü kullanılarak öneri kümesindeki kaç sayfanın bu bölümdeki sayfalara uyduğu hesaplanır. Bu sayının öneri kümesindeki toplam sayfa sayısına oranı ise doğruluğu verir.

Başarı oranı (BO): Deneme kümesindeki her kullanıcı oturumunun her sayfasından sonra oturumun hangi demete ilişkin olduğu hesaplanır ve demete ilişkin en yüksek öneri puanına sahip *N* sayfa kullanıcıya önerilir. Eğer bu *N* sayfadan biri kullanıcının bir sonraki isteği ile aynı ise başarı olarak tanımlanır. Başarı oranı ise toplam başarı sayısının toplam yapılan öneri sayısına bölünmesi ile bulunur.

Deneyler 5 ile 30 arasında değişen farklı demet sayıları, 0.1 ile 0.9 arasında değişen ζ değerleri için tekrar edilmiştir. *w* için 1 ve 2 değerleri denenmiştir. Yapılan deneyler sonucunda *w* için en uygun değerin 2, ζ içinse en uygun değerin 0.5 olduğu gözlenmiştir. *w* değeri 1 seçilirse kullanıcı oturumları ilk sayfadan sonra bir demete atanır. Sadece ilk sayfaya bakarak kullanıcı oturumlarının hangi demette olduğunu saptamak doğru değildir. Eğer ζ değeri büyük seçilirse model çok az sayıda sayfa önerir, bu durumda kullanıcının bir sonraki isteğini tahmin etmek güçleşir. Ancak ζ değeri çok küçük

olursa, bu durumda da önerilen sayfa sayısı çok fazla olur, bu da kullanıcının seçim yapmasını güçleştirir. Bir öneri modelinin amacı da kullanıcıya yol göstererek amacına ulaşmasını kolaylaştırmak olduğu için bu değerler *w* için 2, ζ için 0.5 olarak seçilmiştir.

Tablo 1. Deney sonuçları karşılaştırması

Veri Kümesi	Poisson Modeli	Model 1	Model 2
NASA	52.0	4	47.84
ClarkNet	49.6	15	49.30
Saskatchewan Uni.	50.8	5	44.59

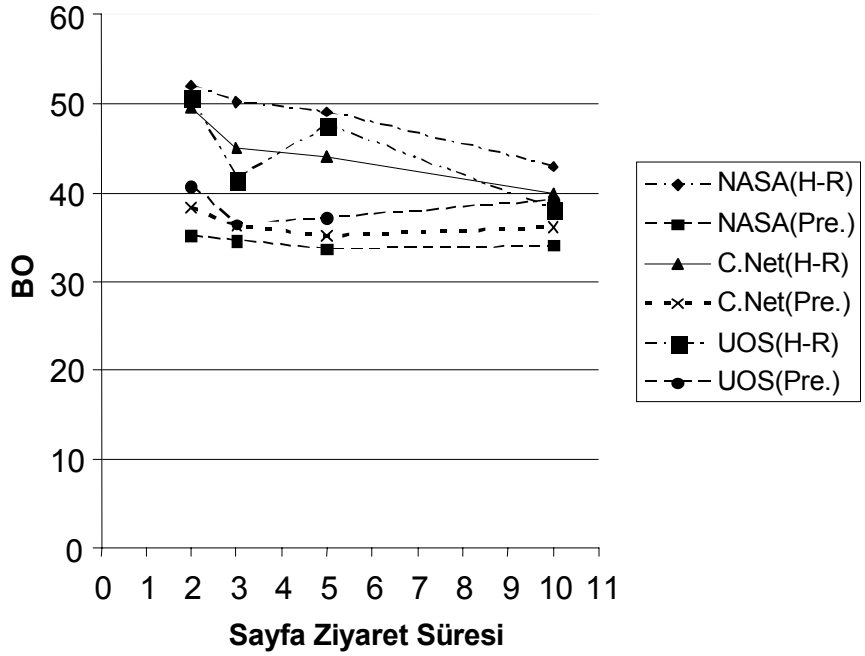
Yapılan deneyler, sayfa ziyaret süresinin 1ve 2 arasında normalleştirilmesinin modelin başarımını artırdığını göstermiştir. Tablo 2 bu sonuçları modelin en iyi sonuç verdiği demet sayıları için göstermiştir. Sayfa ziyaret süresini normalleştirme aralığı genişledikçe sonuçlar kötüleşmiştir. Bu da bize süreyi uygun şekilde normalleştirdiğimizde modelin başarımının arttığını göstermektedir. Şekil 4 model başarımının sayfa süresini normalleştirme değerlerine göre nasıl değiştiğini göstermektedir. Şekil 5 ise model başarımı ile kullanılan demet sayısı arasındaki ilişkiyi göstermektedir. Kullanıcı oturumlarının Poisson dağılımı ile modellenebileceğini göstermek için deneyler başka bir dağılım kullanarak tekrarlanmıştır. Bunun için ilk olarak Multinomial dağılım seçilmiştir. Sonuçların kötü çıkması nedeniyle aynı deneyler Binomial dağılım kullanılarak tekrarlanmıştır. Binomial dağılım sayfa ziyaret sürelerini kullanmadan, kullanıcının bir oturumda herhangi bir sayfayı ziyaret olasılığını kullanarak oturumları modeller. Bu dağılım bize kullanıcı oturumlarını modellerken sayfa ziyaret süresini kullanmanın yararı hakkında da bilgi verecektir. Tablo 3 binomial dağılım kullanarak yapılan deney sonuçlarını vermektedir. Kullanılan demet sayılarının Tablo 2'deki demet sayılarından farklı olmasının nedeni, binomial dağılımın Tablo 3'deki demet sayıları ile en iyi sonucu vermesidir. Tablo 2 ve Tablo 3 karşılaştırıldığında Poisson dağılım kullanılarak oluşturulan modelin daha iyi sonuç verdiğini görmekteyiz.

Tablo 2'de görüldüğü gibi başarı oranı sonuçları daha iyi çıkmıştır. Bunun nedeni ise başarı oranı ile ölçülen deneylerde kullanıcın istek yaptığı her sayfadan sonra kullanıcı oturumunun hangi demete atanacağı hesaplanmaktadır. Bu da sonucun daha iyi çıkmasına neden olmuştur. Yine Tablo 2'de görüldüğü gibi NASA veri kümesine ait sonuçlar daha iyi çıkmıştır. Bunun nedeni NASA veri kümesinin daha iyi temizlenmesi olabilir.

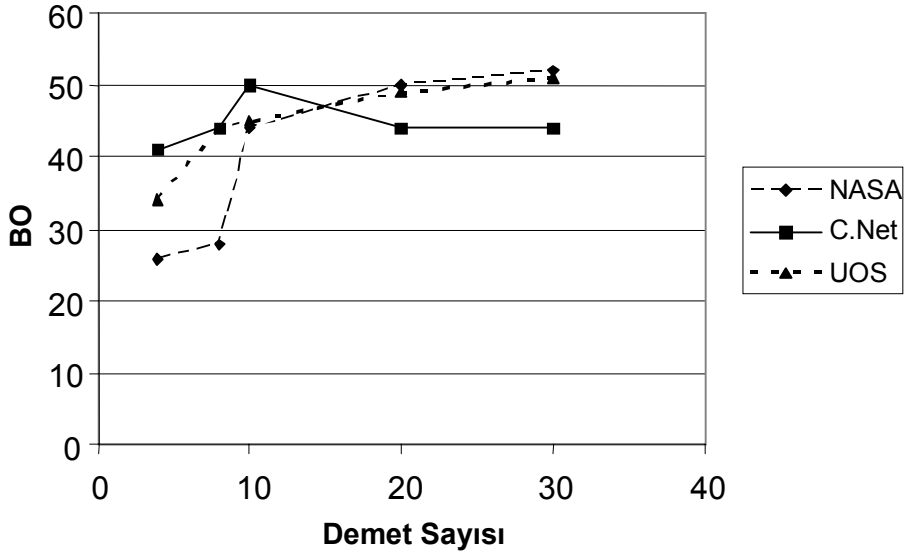
Son olarak modelin başarımı farklı iki model kullanılarak karşılaştırılmıştır. Tablo 3 bu sonuçları göstermektedir. Birinci model k-means demetleme yaklaşımı kullanarak kullanıcı oturumlarını demetlemiş, daha sonra doğruluk ölçümü kullanarak modelin başarımı ölçülmüştür (Mobasher vd., 2000). Alınan sonuçlar Poisson dağılımı kullanarak oluşturulan kullanıcı ilgisi modelinin çok daha iyi sonuç verdiğini göstermiştir. Karşılaştırılan ikinci model kullanıcı oturumları arasında bağıntı örüntülerini bulmaktadır (Mobasher vd., 2001). Bu modeli seçmemizin nedeni hem modelin zaman bilgisini kullanmaması hem de demetleme yaklaşımımını kullanmamasıdır. Böylece farklı bir yaklaşım kullanarak kullanıcı ilgisi modeli karşılaştırılabilir. Ancak ikinci modelin model büyüklüğünün kullanıcı ilgisi modelinden oldukça büyük olduğu açıktır. Bu durumda da kullanıcı ilgisi modeli daha iyi sonuç vermiştir.

Tablo 2. Poisson dağılım kullanarak elde edilen deney sonuçları (%)

Veri	Demet	Yöntem 1		Yöntem 2		Yöntem 3	
Kümesi	Sayısı	DK	BO	DK	BO	DK	BO
NASA	30	34.4	51.1	34.7	51.3	35	52
ClarkNet	10	37.9	48.7	37.6	49.2	38.2	49.6
Saskatchewan Uni.	30	40.6	50.8	40.7	50.6	40.7	50.8



Şekil 4. Model başarımının sayfa ziyaret süresine göre değişimi



Şekil 5. Model başarımının demet sayısına göre değişimi

Tablo 3. Binomial dağılım kullanarak elde edilen deney sonuçları (%)

Veri	Demet	Yöntem 1		Yöntem 2		Yöntem 3	
Kümesi	Sayısı	DK	BO	DK	BO	DK	BO
NASA	23	33.33	39.97	33.60	36.78	35.92	48.84
ClarkNet	10	34.78	34.17	34.04	34.19	38.72	34.16
Saskatchewan Uni.	30	34.78	41.22	34.88	41.22	38.65	41.27

Sonuçlar ve tartışma

Bu çalışmanın amacı bir Web sitesini ziyaret eden yeni bir kullanıcıya, kullanıcının o anki ilgisine göre Web sitesinde ziyaret edebileceği sayfaları öneren bir model oluşturmaktır. Bu modelin doğru öneri yapabilmesi kadar bunu hızlı yapabilmesi de önemlidir. Çünkü öneri kümesi çevrimiçi çalışmada oluşturulmaktadır. Bunun için kullanıcı oturumları Poisson dağılımı kullanarak modellenmiştir. Ayrıca bu çalışmada kullanıcı oturumlarının modellenmesi ve sayfa ziyaret süresinin kullanılan modelde etkisi incelenmiştir. Deneyler kullanıcı oturumlarının Poisson dağılım kullanarak modellenebileceğini ve sayfa ziyaret süresinin dar bir aralıkta normal-leştirilmesinin modelin başarımını artıracağını göstermiştir.

Bu çalışmada yeni bir kullanıcıya öneri kümesi oluşturabilmek için sayfa öneri puanları üç değişik yöntem kullanılarak hesaplanmıştır. Deney sonuçları kullanılan üçüncü yöntemin en iyi sonucu verdiğini göstermektedir. Bu yöntemde öneri puanları hesaplanırken hem Poisson dağılımının parametrelerinden hem de demet içinde sayfanın ne kadar sık ziyaret edildiği bilgisinden yararlanılmıştır.

Yapılan çalışma sonunda örüntü bulmadan önce eldeki veri kümesinin amaca uygun olarak temizlenmesinin deney başarımını artırdığı görülmüştür. Veri madenciliği uygulamalarında model geliştirme kadar verinin temizlenmesi de çok önemlidir.

Geliştirilen model bir Web kullanıcısının bir Web sitesini ziyareti sırasındaki ilgisini modellemek açısından yeni bir bakış açısı kazandırmıştır.

Kaynaklar

Cooley, R., Mobasher, B., Srivastava, J., (1999). Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, **1**, 1, 5-32.

Dempster, A. P., Laird, N. M., Rubin, D. B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of Royal Statistical Society*, **39**, 1, 1-38.

Deshpande, M., Karypis, G., (2001). Selective Markov Models for Predicting Web-Page Accesses. In Proceedings of 1st SIAM International Conference on Data Mining (SDM'2001).

Etzioni, O., (1996). The World Wide Web: Quagmire or Gold Mine, *Communications of the ACM*, **39**, 11, 65-68.

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.

Mobasher, B., Dai, H., Luo, T., Nakagawa M., (2000). Discovery of Aggregate Usage Profiles for Web Personalization, In Proceedings of International {WEBKDD} Workshop – Web Mining for E-Commerce: Challenges and Opportunities.

Mobasher, B., Dai, H., Luo, T., Nakagawa, M., (2001). Effective Personalization Based on Association Rule Discovery from Web Usage Data. In Proceedings of 3rd ACM Workhop on Web Information and Data Management.

Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R., (1999). Mining Web Access Logs Using a Fuzzy Relational Clustering Algorithm Based on a Robust Estimator, In Proceedings of the Eighth International Fuzzy Systems Association Congress.

Pitkow, J., Pirolli, P., (1999). Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. In Proceedings of USENIX Symposium on Internet Technologies and Systems (USITS'99)

Sarukkai, R. R., (2000). Link Prediction and Path Analysis Using Markov Chains. In Proceedings of 9th International World Wide Web Conference, 377-386.

Shahabi, C., Zarkesh, A., Adibi, J., Shah, V., (1997). Knowledge Discovery from Users Web-Page Navigation, In Proceedings of 7th Int. Workshop on Research Issues in Data Engineering.