Türk Kütüphaneciliği 28, 3 (2014), 358-361

Eleştiri / *Research Critique*

Identifying Web search session patterns using cluster analysis: A comparison of three search environments

Wolfram, D., Wang, P. ve Zhang, J. (2009) Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 60(5): 896-910. http://www3.interscience.wiley.com/cgi-bin/fulltext/ 121675939/PDFSTART

Öz

Kullanıcıların daha etkin kullanabileceği erişim sistemleri tasarlamak için bilgi arama modellerinin incelenmesi önemlidir. Bu amaçla bilgi erişim sistemleri işlem kayıtları üzerinde oturum bazlı kümeleme çalışmaları yapılmış fakat farklı türdeki ortamlarda birbirine uyumlu grupların oluşup oluşmadığı ile ilgili karşılaştırma yapılmamıştır. Bu çalışmada üç farklı türdeki[1] Web tabanlı bilgi erişim sistemini temsil eden işlem kayıtları üzerinde kümeleme tekniği kullanılarak arama oturum modellerini incelemişlerdir. Sonuçlar arama davranışlarının oturum karakteristiklerine dayanan belirgin gruplar halinde kümelenebildiğini ve farklı sistemler olsa da benzerlik gösterdiğini ortaya çıkarmıştır. Oturum bazlı analizler kullanıcı arama davranışlarının anlaşılması için önemlidir, sistem tasarımcılarının çeşitli kullanıcı gruplarının ihtiyaçlarını daha iyi karşılayabilecek sistemler geliştirmesine yardımcı olabilir.

***Anahtar Sözcükler**: Bilgi erişim; kümeleme; oturum bazlı analizler; oturum karakteristikleri.*

Abstract

Information seeking models are important to design more efficient information retrieval systems. For this purpose, information retrieval systems transaction log studies were performed based on the Session, but there has never been a comparison whether different types of session groups are compatible with each other or not. In this study, three different types[2] of Web-based information retrieval systems are studied the search session models using clustering methods. Results have shown that searching behaviors are clustered into distinct groups by characteristic of sessions and revealed although being different groups, they show similarities. Session-based analysis is important for the understanding of user search behavior; this can help to system designers to develop systems to meet the needs of various user groups in a better way.

***Keywords:** Information retrieval; clustering; session-based analysis; characteristic of sessions.*

Makale Eleştirisi

Araştırma Problemi

Çalışmada farklı kullanıcı gruplarına sahip üç ayrı Web tabanlı arama ortamının işlem kayıtlarındaki belirli özelliklerin (oturum uzunluğu, sorgu başına düşen terim sayısı, ortalama sorgu intervali vb.) niceliksel analizi yapılarak şu sorulara yanıt aranmıştır:

- Farklı türdeki Web tabanlı arama ortamlarında benzer oturum karakteristiklerine göre oluşturulan gruplar farklılık gösteriyor mu?
- Arama oturum modelleri zamanla değişir mi?

[1] Akademik Web sitesi (University of Tennessee-Knoxville), Genel Arama Motoru (Excite) ve Tüketici Sağlık Bilgisi Portalı (HealthLink).

[2] Academic Web Site (University of Tennessee-Knoxville), Public Search Engine (Excite) and Consumer Health Information portal (HealthLink).

Önceki çalışmalarda oturum karakteristikleri incelenmiş fakat belirlenen grupların zamanla değişip değişmediği ve farklı türdeki bilgi erişim sistemi ortamlarında birbirine uyumlu grupların oluşup oluşmadığı incelenmemiştir. Bu açıdan çalışmada yeni bilgi üretileceği söylenebilir.

Literatür Değerlendirme

Web arama aktivitelerinin analizi yaklaşık 15 yılı aşkın süredir yapılmaktadır. Fakat literatür değerlendirmesi kısmında daha çok son 10 yıllık süreçteki veri madenciliğine dayanan, matematiksel modelleme ya da açıklayıcı istatistiksel modelleri kullanan çalışmalar incelenmiştir. Kullanıcıların, aramaların ve oturumların karakteristiklerini raporlayan ilk çalışmalara değinilmemiştir. Bütünleştirme açısından sorunun genel olarak hangi bağlamda yer işgal ettiğinin görülmesi için açıklayıcı türdeki çalışmalar hakkında da bilgi verilebilir. Kümeleme teknikleri kullanılarak oturum seviyesinde işlem kayıtları analizi yapan çalışmalar ayrıntılı biçimde incelenmiştir.

Veri toplama

Çalışmada kullanılan veri setleri, üç farklı tür Web arama ortamının işlem kayıtlarıdır. Veri setlerinin özellikleri şu şekildedir:

- *Akademik Web sitesi (University of Tennessee-Knoxville):* UTK arama motoruna girilen sorguları içermektedir. 3.9 M büyüklüğündedir ve 2 yılı (2003-2004 kapsamaktadır).
- *Genel Arama Motoru (Excite):* 1999 ve 2001 yılında toplanmış iki ayrı veri kümesidir. Büyüklükleri sırasıyla 622K ve 587K'dır.
- *Tüketici Sağlık Bilgisi Portalı (HealthLink):* Büyüklüğü 377K'dır. 2005 yılına aittir.

Kayıtlar, MS Access ve SQL sunucu veri tabanlarında depolanmıştır.

Veri analizi

Farklı sistemlerden alınan veri setlerinin yapıları farklıdır. Dolayısıyla benzer değişken sayıları az olduğu için sabit küme belirlemede etkili olan değişken sayısı 4 ila 6'dır. Fakat göz ardı edilen bu değişkenlerin oturum modellerinde belirleyici özellikleri olabilir (Örneğin Boole işleçleri kullanılan sorgu sayısı).

Küme belirlemede terim sayısının eşik belirlemede bu kadar etkili olması geçerlik sorunları yaratabilir. Çünkü sayılar bilgi ihtiyacı ve sistemin türüne göre (çok özel (spesifik) ya da çok genel) farklılık gösterebilir. Mevcut ölçüme göre HealthLink'te "*pulmonary thromboembolism*" sorgusunun C2 kümesine, genel arama motorunda "*embolism*" sorgusunun C1 kümesine girme olasılığı yüksektir (genelde tek terim). Öte yandan daha düşük bir eşik noktası daha az sorgu ile daha fazla küme sonucu doğururdu.

İşlem kayıtlarının zaman aralıklarının farklı olması da diğer bir sorundur.[3] Üç veri kümesi kendi içlerinde analiz edilmiştir. Fakat araştırmanın amacı bunların karşılaştırılmasıdır. Zaman aralığı olayı bu amaca uymamaktadır.

Çalışmada iki aşamalı kümeleme (Two-step Cluster) yöntemi kullanılmıştır. Küme sayısını PASW yazılımının belirlemesine izin vermiş ise her veri setinde de üç küme çıkması düşük bir ihtimal (Gaskin, 2012). Araştırmacılar kendileri belirlemişlerse de geçerlik sorunları olabilir.

Sorgu başına düşen görüntülenen sayfa sayısı bilgisi sadece Excite verilerinde bulunmaktadır. Dolayısıyla Excite veri setinde küme oluştururken kullanılan değişken sayısı diğerlerinden farklı olarak 6'dır (HealthLink ve UTK için 5). Karşılaştırma yapılmayacağı için sorgu başına düşen görüntülenen sayfa sayısı bilgisini değişkenler arasına almak çok gerekli olmayabilir.

[3] University of Tennessee-Knoxville: 2003-2004. Excite: 1999 ve 2001, HealthLink: 2005

Oturumların birden fazla günde devam etmeyeceğine karar verilmiştir[4]. Benzer şekilde eğer uzun süre aktivite olmazsa konu aynı bile olsa yeni bir oturum olarak sayılmıştır. İdeal olarak sorgular benzeşiyorsa iki sorgu aynı oturuma atanmalıdır.

Orjinal makaledeki Tablo 2'de boole işleçleri kullanılarak yapılan arama sayılarından bahsedilmiştir. Fakat mevcut veri tabanı yapısında böyle bir alan bulunmamaktadır. Ayrıca araştırma sonuçlarında da bahsedilmemiştir. Aynı şekilde tabloda tanınmayan ya da standart olmayan sözcükler yer almaktadır. Standart sözlük ile karşılaştırılarak belirlenmiştir ancak sözlük hakkında bilgi verilmemiştir (Örneğin HealthLink çok özel (spesifik), Excite ise çok geneldir bunların ikisini de kapsayacak nitelikte bir sözlük varsa belirtilmesi gerekir).

Kümelerin geçerliklerini sınamak için 200 örnek oturum çekilmiştir. Sonrasında bir insan hakem her oturumu küme karakteristiklerine dayanarak belirlenmiş üç kümeden birine ataması konusunda görevlendirilmiştir. Sonuçta otomatik kümeleme yöntemi ile insan hakem arasında %66 ve %70 arasında mantıklı bir uyuşmayı gözlenmiştir. Fakat buradaki sorun insan hakemin terim popülerliğini değerlendirmesi ile ilgilidir. İnsan hakemin ortalama terim popülerliğini değerlendirmesi pek güvenilir değildir. Çünkü her veri kümesinde binlerce sorgu terimi bulunmaktadır ve bu değerlendirmenin oturum örneklerindeki yüzlerce terimin her biri için yapılması gerekmektedir.

Zorluklar

Web arama davranış modellerini belirlemek için tek tek sorguları incelemek yeterli değildir. Sorgu gruplarının oturum bazlı incelenmesi gerekmektedir. Buradaki zorluk tek bir etkileşim oturumunun sınırlarını belirlemektir. İşlem kayıtları genellikle kullanıcı tarafında belli bir bilgisayar ya da IP'ye atanan çerezler ile belirlenmiştir. Teknik olarak belli bir bilgisayar ya da bir IP adresini paylaşan ve tek olmayan bir grup bilgisayarı temsil etmektedirler. Araştırmacının bu durumda oturumun sınırlarını netleştirmesi zordur. Ayrıca verilerin sağlandığı bilgi erişim sistemlerinde oturum açma zorunluluğu da bulunmamaktadır.

Kaynak Gösterme Yanlışları

897. sayfa: Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. Computer Magazine, 35(3), 107–109. Bu kaynağın tarihi metin içi atıfta tarih 2003 olarak belirtilmiştir.

900.sayfa: Metin içi atıf (He & Göker, 2000) değil, (Göker & He, 2000) olmalıdır.

907. sayfa: Wang'dan alıntı, 2003 yılındaki değil 2007'deki çalışmasından metin içi atıf verimemiş kaynak: Novak, J.D. (1998). Learning, creating, and using knowledge. Mahwah, NJ: Lawrence Erlbaum.

Sonuç

Sabit küme değişkenleri çeşitliliğini artırmak için benzer yapıdaki sistemler karşılaştırılabilir. Kullanıcı girişi yapılan sistemlerin oturum kayıtlarını tercih etmek daha sağlıklı olur.

Oturum sınırlarının belirlenmesi veri kümesinin karakteristiğine bağlıdır. Sınırlandırmanın sorguların konu analizine göre yapılmasını öneren çalışmalar da yapılmıştır (Huang, Peng, An, ve Schuurmans, 2004; He, Göker ve Harper, 2002). Belki bu çalışma için aynı yöntem denenebilir. Bunun dışında konu analizi, makine öğrenme teknikleri ya da istatistik dil modeli ile de yapılabilir.

Farklı bilgisayarlar aynı IP adresini paylaşabilir veya kütüphane gibi ortak kullanım alanlarındaki çok kullanıcılı bilgisayarlar farklı kullanıcıların oturumlarını kaydetmiş olabilirler. Bunun çözümü için oturum sınırları belirlenirken belli zaman aralığındaki ardışık sorguları belirleyen bir çalışma yapılabilir ya da farklı zamanlarda yapılmış fakat benzer özellikler gösteren sorguların interval değerleri çok yakın ise aynı oturuma atanabilir (iki çok yakın

[4] Aynı IP adreslerinden gece yarısından 5 dakika önce ve 5 dakika sonra gelen sorgu sayısı çok az olduğu belirtilmiş fakat sayı verilmemiştir.

sorgu interval değeri, eşik değerinden daha az ise aynı oturuma aittir). Nitekim bunu yaparak kullanıcıları tanıyan çalışmalar bulunmaktadır (Hu, Zeng, Li, Niu ve Chen, 2007).

Çalışma nitel yöntem de kullanılarak daha zengin hale getirilebilir. Örneğin C1 kümesi için (kısa oturumlar genelde tek kelime) ilgili sonuçlar hızlı bulunuyor. Tartışma kısmında kullanıcıların sorgularını bile değiştirmeden oturumu terk ettiği belirtilmiştir. Fakat bunların yüksek verimde mi yoksa verimsiz oturumları mı temsil ettiğini bilmiyoruz. Bu durumda kullanıcının bilgi ihtiyacını karşılayıp karşılamadığı ancak nitel yöntemlerle ortaya çıkar.

Teşekkür

Bu makale eleştirisi Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümünde 2013 Bahar döneminde verilen BBY 606 Araştırma Yöntemleri dersi kapsamında hazırlanmıştır. Metni okuyarak önerilerde bulunan hocam Prof. Dr. Yaşar Tonta'ya çok teşekkür ederim.

Kaynakça

Gaskin, J. [James Gaskin]. (19 Mart 2012). *Two-step Cluster Analysis in SPSS* [Video dosyası]. 12 Nisan 2013 tarihinde http://www.youtube.com/watch?v=DpucueFsigA adresinden erişildi.

He, D., Göker, A. ve Harper, D.J. (2002). *Combining evidence for automatic Web session identification*. Information Processing&Management, 38(5), 727–742. 12 Nisan 2013 tarihinde http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.88.1441 adresinden erişildi.

Hu, J., Zeng, H.J., Li, H., Niu, C. Ve Chen, Z. (2007). *Demographic pre- diction based on user's browsing behavior*. WWW '07 Proceedings of the 16th International Conference onWorld Wide Web (ss. 151–160). New York: ACM. 20 Nisan 2013 tarihinde http://wwwconference.org/www2007/papers/paper686.pdf adresinden erişildi.

Huang, X., Peng, F., An, A. ve Schuurmans, D. (2004). *Dynamic Web log session identification with statistical language models*. Journal of the American Society for Information Science and Technology. 55 (14),

1290–1303. 12 Nisan 2013 tarihinde http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.6602 adresinden erişildi.

Müge Akbulut
mugeakbulut@gmail.com