1.GİRİŞ

Bu makalede [1] dokümanların incelenme ve çekici olma ihtimallerinin modelleri kullanılarak, dokümanların tıklanma ihtimallerinin modellenmesi hedeflenmiştir. Çekici olma ihtimali incelenmiş bir dokümanın tıklanmasının ihtimalidir. İnceleme işlemi ise doküman başlık ve özetinin okunmasıdır.

Dokümanların incelenme ihtimalinin, dokümanın bulunduğu sıra ve özellikle de en son tıklanan dokümana uzaklığına göre hesaplayan bir modelin geliştirilmesi hedeflenmiştir. Daha önce dokümanların incelenme ihtimalini buradaki kadar açık bir şekilde modelleyen olmamıştır. Buradakine en yakın çalışmada [2] dokümanların tıklanmaları modellenmiştir. Burada sunulan modeller [2] ile kıyaslanmışlardır.

Elde edilen doküman inceleme modelinin Joachims et al. [3] ile tutarlı sonuç vermesi hedeflenmiştir.

Modellerin anlatılmasında kullanılan değişkenler ve notasyon aşağıda paylaşılmıştır.

q: Arama metnidir.

u: q arama metni sonucu kullanıcıya dönülen dokümandır.

r: Dokümanın pozisyonudur.

d: Son tıklanan doküman ile söz konusu doküman arası mesafedir.

m: Arama metnine göre kullanılan inceleme çeşididir.

e: Dokümanın incelenmesidir.1 ve 0 değerlerini alır.

c: Dokümanın tıklanmasıdır. 1 ve 0 değerlerini alır.

a: İncelenen bir dokümanın tıklanması yani çekici olmasıdır. 1 ve 0 değerlerini alır.

a_{uq}: q arama metni sonucu gelen u dokümanın çekici olma ihtimalidir.

γ_{rd}: r ve d pozisyonunda bulunan dokümanın incelenme ihtimalidir.

γ_{rdm}: m modeline göre incelenen arama metni sonuçlarında, r ve d pozisyonunda bulunan dokümanın incelenme ihtimalidir.

u_{mq}: m inceleme çeşidinin q arama metninde kullanılma ihtimalidir.

Bu makale kapsamında tıklanma ihtimalini tahmin eden 3 farklı model geliştirilmiştir. Bunlar Single Browsing Model, Multiple Browsing Model ve Logistic Model'dir.

Single Browsing Model

İnsanların doküman listesinde yukarıdan aşağı hareket ettiğini, her bir pozisyon için o pozisyondaki dokümanı inceleyip incelemeyeceğine karar verdiğini, incelediği dokümanlar arasında çekici bulduklarına tıkladığını varsayar.

$$P(c = 1|u,q,r,d) = a_{uq}\gamma_{rd} \qquad P(e = 1|r,d) = \gamma_{rd}$$

Dokümanın tıklanma ve incelenme ihtimali yukarıdaki gibi modellenmiştir.

Multiple Browsing Model

İnsanların, arama metnine göre değişen şekilde dokümanları incelediklerini varsayar. Broder [4]'te farklı inceleme çeşitlerine değinilmiştir. Bu tipler ana olarak navigational (bir siteye ulaşma amaçlı) ve informational (bilgi toplamaya yönelik) şeklinde ikiye ayrılmıştır.

$$P(e = 1|r,d,m) = \gamma_{rdm} \qquad P(c = 1|u,q,r,d,m) = \sum_m^M u_{mq}\,\gamma_{rdm}a_{uq}$$

Bir arama metnine göre kullanılan inceleme çeşidi sonucu, dokümanın incelenme ve tıklanma ihtimali yukarıdaki gibi modellenmiştir.

Logistic Model

Tıklanma olasılıklar oranının logaritması modellenmiştir. B_{uq} ve B_{rd} parametreleri kullanılmıştır. B_{uq} dokümanın çekici olması, B_{rd} ise dokümanın bulundu pozisyon ile ilgilidir.

$$\frac{P(c = 1|u,q,r,d)}{1 - P(c = 1|u,q,r,d)} = e^{B_{uq}}\, x\, e^{B_{rd}}$$

Yukarıdaki gibi modellenmiştir. Logistic model, parametrelerinin doğal yorumlanabilmesi, yaygın ve verimli gerçekleştirilmelere sahip olduğu için seçilmiştir.

2.YÖNTEM

Bahsedilen modellerin parametrelerini bulmak için arama motoru kayıtları nadir olaylar filtrelenerek 542,651 adet arama metni ve 336,436,808 oturumdan oluşan bir veri havuzuna dönüştürülmüştür. Modellerin test ve öğrenmeleri için 21 adet veri kümesi hazırlanmıştır. Her bir veri kümesi yukarıda verilen veri havuzunun %1'i kadar olacak şekilde birbirlerinden farklı seçilmiştir. Öğrenme veri setinden yukarıda analtılan model parametreleri bulunmuştur.

Modeller inceleme ve çekici olma ihtimali gibi gözlenemeyen parametrelere sahip oldukları için gözlenebilen tıklanma ve tıklanmama sayılarını tahmin etmeleri istenmiştir. Çapraşıklık (perplexity) yani modelin gözlemler sonucu ne kadar şaşırdığı ise başarı ölçümü olarak kullanılmıştır.

3.BULGULAR

Single Browsing Model

Çapraşıklık (perplexity) değeri diğer modellerden azdır. Model öğrenme ve tahmin alanlarında diğer bahsedilen modelleri geride bırakmıştır. r ve d parametrelerinin artması sonucu doküman inceleme ihtimali genelde azalmıştır ve Joachims et al. [3] ile tutarlı sonuçlar vermiştir.

Multiple Browsing Model

Parametreleri ilk etapta 2 farklı inceleme çeşidinin varlığı kabul edilerek oluşturulmuştur. Single Browsing Model'den daha fazla çapraşıklık (perplexity) değerleri gözlenmiştir. Bu yüzden 2'den fazla inceleme çeşidinin varlığını kabul eden denemeler yapılmamıştır.

Logistic Model

Single Browsing Model'den daha fazla çapraşıklık (perplexity) göstermiştir.

4.SONUÇ

Çalışma sonucu insanların dokümanları incelemeleri ve dokümanların tıklanma ihtimalleri modellenmiştir. Arama motoru kayıtları kullanılarak model parametreleri öğrenilmiş bu parametreler üzerinden modeller test edilmiştir.

Single Browsing Model en başarılı model olmuştur. Ayrıca burada kullanılan doküman inceleme modeli de [3] ile tutarlı sonuçlar vermiştir.

Multiple Browsing Model başarılı olamamıştır.

Dokümanların çekiciliği, editör değerlendirmelerindeki hataları bulmak ve bu değerlendirmeleri tahmin etmek için kullanılabilir. Bu alanların da daha sonra araştırılacağı belirtilmiştir.

5.KAYNAKÇA

[1] Georges E. Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. SIGIR 2008.

[2] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In First ACM International Conference on Web Search and Data Mining WSDM 2008, 2008.

[3] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In Proceedings of ACM SIGIR 2005, pages 154–161, New York, NY, USA, 2005. ACM Press.

[4] A. Broder. A taxonomy of web search. SIGIR Forum, 36(2):3–10, 2002.