Merhaba Arkad

ⁿ

aşlar bugün sizlere A User Browsing Model to Predict Search Engine Click Data from Past Observations adlı makalenin özetini anlatacağım. Bu makale SIGIR 2020 de Honourable Mentionsa layık görülmüştür.

----------------------------------------------------------------------------------------------------------------

Bir arama metnine göre listelenen dokümanlardan, arama metnine uygun olanlarının yukarı sırada gözükmesi istenilen bir durumdur. Bu uygunluğun belirlenmesinde dokümanların tıklanma sayıları önemli fakat yetersiz bir ölçüttür.

Çünkü insanlar genelde listenin altındaki dokümanlara bakmazlar ve dolayısıyla tıklamazlar. Fakat bu dokümanların alakasız olduğu kanıtlamaz. Sadece tıklanma sayısına bakılarak dokümanlar sıralanırsa uygun olan dokümanların hak ettiklerinden alt sırada kalma ihtimali oluşur.

Bu makalede anlatılan çalışmada insanların dokümanları inceleme ihtimalini modelleyen ve bu modeli kullanarak dokümanların tıklanma ihtimalini hesaplayan modeller geliştirilmiştir. Bulunan sonuçlar başka çalışmalar ile kıyaslanmıştır.

----------------------------------------------------------------------------------------------------------------

Modellerin anlatılmasına geçmeden önce makalede ve bu sunumda kullanılan değişkenler ve notasyonları anlatmak istiyorum.

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 metodudur.

e: Dokümanın incelenmesidir.binary değerler alır.

c: Dokümanın tıklanmasıdır.binary değerler alır.

𝑎: İncelenen bir dokümanın tıklanması yani çekici olmasıdır. Binary değerler 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 metninde r ve d pozisyonunda bulunan dokümanın incelenme ihtimalidir.

𝑢\_𝑚𝑞: m inceleme modelinin q arama metninde kullanılma ihtimalidir.

Burada doküman inceleme ile bahsedilen durum, dokümanın özeti ve başlığının kullanıcı tarafından incelenmesidir.

----------------------------------------------------------------------------------------------------------------

Bu makalede anlatılan çalışmada 3 farklı model geliştirilmiştir.

Bunlar Single ve multi browsing model ve logistic modeldir.

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.

Model parametrelerine bakarsak bir dokümanın tıklanması yani c nin1 olma ihtimali dokümanın çekici olması ve incelenme ihtimalinin çarpımına yani 2 bağımsız koşulunda gerçekleşmesine eşittir.Doküman inceleme ihtimalininin d ye bağlılığının gösterilmesi çalışmanın önemli motivasyonları arasındadır.

Multiple Browsing model

İnsanların, arama metnine göre değişen şekilde dokümanları incelediklerini varsayar. Farklı inceleme tipleri mevcuttur fakat çalışma kapsamında 2 ana tip ele alınmıştır. Bunlar navigational yani bir siteye ulaşma amaçlı ve informational yani bilgi toplamaya yönelik olanlar şeklindedir.Modele bakılırsa bir dokümanın incelenme ihtimali pozisyonuna ek olarak kullanılan inceleme şekline de bağlıdır.bu inceleme şekli de bir arama metninde umq ihtimali ile kullanılır.bir dokümanın tıklanma ihtimali ise çekici ve incelenme ihtimalinin çarpımına eşittir. Fakat burada incelenme ihtimalinin modellenmesi single browsinge göre biraz farklıdır. Her bir umq x γrdm ikilisi tek bir browsing model için inceleme ihtimali verir. Büyük M kadar farklı inceleme metodu olduğu durumda ise umq x γrdm ikililerinin toplamı incelenme ihtimalini verir.

Logistic Model

Tıklanma olasılıklar oranının logaritması modellenmiştir. Buq ve Brd parametreleri kullanılmıştır. Buq dokümanın çekici olması ile Brd ise dokümanın bulundu pozisyon ile ilgilidir.

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

----------------------------------------------------------------------------------------------------------------

Bahsedilen modellerin parametrelerini bulmak için arama motoru kayıtları nadir olaylar filtrelenerek yaklaşık 500bin adet arama metni ve 336 milyon oturumdan oluşan bir veri havuzuna dönüştürülmüştür. Modellerin test ve öğrenmeleri için 21 adet veri kümesi hazırlanmış. Her bir veri kümesi yukarıda verilen veri havuzunun %1’i kadar olacak şekilde birbirlerinden farklı seçilmiştir.Öğrenme veri setinden model parametreleri bulunmuştur.

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

----------------------------------------------------------------------------------------------------------------

Single Browsing Model

Çapraşıklık (perplexity) değeri diğer modellerden azdır. Model öğrenme ve tahmin alanlarında diğer bahsedilen modellerden daha iyidir. r ve d parametrelerinin artması sonucu doküman inceleme ihtimali genelde azalmıştır ve bu sonuçlar başlangıçta hedeflendiği gibi eye tracking deneyleri ile tutarlıdır.

Multi Browsing Model

Parametreleri ilk etapta 2 çeşit inceleme modelinin varlığı kabul edilerek oluşturulmuş. Single Browsing Model’den daha fazla çapraşıklık (perplexity) değerleri gözlemlenmiştir. Bu yüzden 2’den fazla inceleme modelinin varlığını kabul eden denemeler yapılmamıştır.

Logistic Model

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

Bu durumda Single Browsing Model en iyi model olarak seçilmiştir.

----------------------------------------------------------------------------------------------------

dokümanların çekiciliği editör değerlendirmelerindeki hataları bulmak ve editör değerlendirmelerini tahmin etmek için kullanılabilir. Bu alanların da daha sonra araştırılacağı belirtilmiştir.