Bilgi Erişim Sistemlerinde Olasılıksal Yöntemler

Ahmet Kaşif 501731005
Mehmet Burak Koca 501631004
Kerim Can Kalıpcıoğlu 501731001

Erişimde Olasılıksal yaklaşımlar

Kullanıcı sorgularının ve dökümanların bulunduğu bir sistemin, sorguyu sağlayan dökümanları döndürmesini bekliyoruz.

Bilgi erişimde, mantıksal veya vektör uzay modellerinde, sorgu-döküman eşleştirmeleri biçimsel olarak tanımlanır fakat anlamsal olarak bir şey ifade etmezler.

Dolayısıyla, bilgi erişim sistemi kullanıcı sorguları hakkında bilgi değildir, dökümanın sorguyu sağlayıp sağlamadığı hakkında kesin olmayan tahminler yapar.

Erişimde Olasılıksal yaklaşımlar

Olasılık teorisi, bilgi erişim sistemlerindeki kesinlik içermeyen bu ortamda, sisteme muhakeme yeteneği kazandırmak için bazı ilkeleri temel olarak kullanır.

Olasılıksal modeller, bir dökümanın sorguyla ne kadar ilgili olduğunu, bu temellere dayandırarak tahmin etmeye çalışırlar.

Olasılıksal Modellerin Gelişim Süreci

Olasılıksal erişim modellerinin kullanılması ile alakalı ilk çalışma, Maron ve Kuhns tarafından 1960 yılında yapılmıştır.[1]

Daha sonra Miller tarafından 1971 yılında yine olasılıksal bilgi erişim üzerine temel oluşturan bir çalışma yapılmıştır.

Sonraki çalışmalar, çeşitli olasılıksal yöntemler kullanlamaları bakımından farklılaşmıştır.

Olasılık Teorisi

A ve B olaylarımız olduğunu varsayalım.
P(A, B) : A ve B' nin birlikte olma olasılığı
P(A|B) : B olayının olduğu durumda, A'nın olma olasılığını veren şartlı olasılığı ifade eder.

Olasılık Teorisi

Zincir kuralını kullanarak, şartlı olasılık ve birleşik olasılık arasındaki ilişkiyi gözlemleyelim:

$P(A, B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

Benzer şekilde, B olayı gerçekleştiği durumda A olayının gerçekleşmeme olasılığı:

$P(\bar{A}, B) = P(\bar{A}|B)P(\bar{A})$

Bölünebilirlik Kuralı : Eğer B, birden fazla alt olaya bölünebiliyor ise, B olayının gerçekleşme olasılığı, tüm alt olayların gerçekleşme olasılığının toplamıdır.

$P(B) = P(A,B) + P(\bar{A},B)$

Bayes Teoremi

Bayes teoremine göre, A olayına bağlı B olayının gerçekleşme olasılığı, B olayına bağlı A olayının gerçekleşme olasılığı ile A olayının tek başına gerçekleşme olasılığının çarpımının, B olayının gerçekleşme olasılığına bölünmesiyle bulunabilir. Matematiksel gösterimi şu şekildedir:

$P(A|B) = P(B|A)P(A) / P(B)$

Doküman Derecelendirme Problemi

- Bir doküman koleksiyonuna yapılan sorgu sonucu, sorgu ile ilişkili dokümanların ilişki derecesi ile sıralanıp döndürülmesi problemidir
- Örnek olarak $R_{d,q}$ 'yu binary ilişki skoru kabul edersek;
  - $R_{d,q}$ = 1 eğer doküman *d, q sorgusu ile ilişkili ise*
  - $R_{d,q}$ = 0 diğer durumlarda
- Olasılıksal derecelendirme, dokümanları, sorgu ile tahmini ilişki derecelerine göre sıralayarak döndürmektedir $P(R = 1 | d, q)$

Probability Ranking Principle (PRP)

- Eğer bir bilgi erişim sistemi gelen tüm sorgulara, dokümanları sistemdeki verileri kullanarak gerçeğe en yakın olasılıklar ile sıralanmış olarak(en ilişkili olandan en düşük ilişkili olana doğru) cevap verebiliyorsa, sistemin performansı en etkin düzeydedir.

- Dokümanın sorgu ile ilişkili olma olasılığı

P (ilişkili olma durumu | doküman, sorgu)

İkili Bağımsız Model (BIM)

Doğru çalışan bir olasılıksal bilgi erişim sistemi tasarlamak için ilişkilendirmeye katkıda bulunan parametrelerin belirlenmesi gerekmektedir

- Dokümanın ilişki durumuna etki eden istatiksel değerler hesaplanır (terim frekansı, dokümanın frekansı, doküman uzunluğu)
- Bu veriler birleştirilerek ilgili dökümanın ilişki derecesi belirlenir
- Dokümanlar ilişki derecesine göre çoktan aza doğru sıralanır $P(R|d, q)$

İkili Bağımsız Model (BIM)

Yaklaşım 1

➢ Her bir dokümanın ilişkili olma olasılığı diğer dokümanlardan bağımsızdır.

- Dokümanlar ve sorgular binary vektörler olarak ifade edilir.
  - Örneğin Doküman d, $\vec{x} = (x_1, \ldots, x_M)$ olarak ifade edilir
    - *t* terimi *d dokümanında varsa* $x_t = 1$*, yoksa* $x_t = 0$
  - Farklı dokümanlar aynı vektör gösterimine sahip olabilirler.

11

İkili Bağımsız Model (BIM)

$$rank = \frac{P(R = 1 \mid D)}{P(R = 0 \mid D)} = \frac{P(D \mid R = 1) * P(R = 1)}{P(D \mid R = 0) * P(R = 0)}$$

D= Doküman R={0,1} ilişkili olup olmama durumu

- P(R=1) ve P(R=0) 'ın ranka etkisi yoktur.
- Tüm sözcükler için ilişkili olma durumu aynıdır.

Dokümanın ilişki derecesi diğer dokümanların ilişki derecelerinden bağımsız olarak kabul edilerek hesaplanır

İkili Bağımsız Model (BIM)

Yaklaşım 2

- Her bir sözcüğün ilişkili olma durumu diğer sözcüklerden bağımsızdır.

$$rank = P(R = 1 \mid D) = \frac{P(D \mid R = 1)}{P(D \mid R = 0)} = \frac{\prod_w P(D_w \mid R = 1)}{\prod_w P(D_w \mid R = 0)}$$

İlişkili kelimelerin bulunup bulunmaması rankı etkilemez.

İkili Bağımsız Model (BIM)

Yaklaşım 3

$$rank = P(R = 1 \mid D) = \frac{\prod_{w \in D}(P_w)}{\prod_{w \in D}(q_w)} * \frac{\prod_{w \notin D}(1 - P_w)}{\prod_{w \notin D}(1 - q_w)}$$

$$\text{rank} = \frac{\prod_{w \in D} (P_w) * (1 - q_w)}{\prod_{w \in D} (q_w) * (1 - P_w)}$$

$$P_w = P(D_w = 1 \mid R = 1)\ q_w = P(D_w = 1 \mid R = 0)$$

İlişkiler ile tahminde bulunma

- İlişkili olarak belirlenen N_1 ve ilişkili olmayan N_0 dokümanlarımız olduğunu varsayalım;
- W kelimesi $N_1(w)$ ve $N_0(w)$ olarak gözlemlensin

$$P_w = \frac{N_1(w) + 0.5}{N_1 + 1} \qquad q_w = \frac{N_0(w) + 0.5}{N_0 + 1}$$

P_w= En az 1 kez ilgili kelime geçen doküman sayısının tüm dokümanlara oranı

Sorguda olan veya olmayan kelimeler ilişkili ve ilişkili olmayan dökümanlarda eşit sayıda bulunabilir.

İlişkiler ile tahminde bulunma (Örnek)

İlişkili Dokümanlar D_1= "a b c b d" D_2= "a b e f b"

İlişkili olmayan Dok. D_3= "b g c d" D_4= "b d e" D_5= "a b e g "

Kelime	:	a	b	c	d	e	f	g	h	
$N_1(w)$	:	2	2	1	1	1	1	0	0	N_1=2
$N_2(w)$	:	1	3	1	2	2	0	2	0	N_2=3
P_w	:	$\frac{2+0.5}{2+1}$	$\frac{2+0.5}{2+1}$	$\frac{1+0.5}{2+1}$	$\frac{1+0.5}{2+1}$	$\frac{1+0.5}{2+1}$	$\frac{1+0.5}{2+1}$	$\frac{0+0.5}{2+1}$	$\frac{0+0.5}{2+1}$	
q_w	:	$\frac{1+0.5}{3+1}$	$\frac{3+0.5}{3+1}$	$\frac{1+0.5}{3+1}$	$\frac{2+0.5}{3+1}$	$\frac{2+0.5}{3+1}$	$\frac{0+0.5}{3+1}$	$\frac{2+0.5}{3+1}$	$\frac{0+0.5}{3+1}$	

Yeni doküman : D_6= "b g h" :

$$rank = \frac{\prod_{w \in D} (P_w) * (1 - q_w)}{\prod_{w \in D} (q_w) * (1 - P_w)} = \frac{\left(\frac{2.5}{3} * \left(1 - \frac{3.5}{4}\right)\right) * \left(\frac{0.5}{3} * \left(1 - \frac{2.5}{4}\right)\right) * \left(\frac{0.5}{3} * \left(1 - \frac{0.5}{4}\right)\right)}{\left(\frac{3.5}{4} * \left(1 - \frac{2.5}{3}\right)\right) * \left(\frac{2.5}{4} * \left(1 - \frac{0.5}{3}\right)\right) * \left(\frac{0.5}{4} * \left(1 - \frac{0.5}{3}\right)\right)} = \frac{1.64}{13.67}$$

İlişkiler bilinmeksizin tahminde bulunma

Yaklaşım 4

$$\text{rank} = \prod_{w \in D \cap Q} \frac{N - \mathrm{N_w} + 0.5}{\mathrm{N_w} + 0.5}$$

N = Doküman Sayısı

$\mathrm{N_w}$ = Kelimenin dokümanlarda bulunma sayısı

İlişkiler bilinmeksizin tahminde bulunma(örnek)

Dokümanlar D_1= "a b c b d" D_2= "a b e f b"

D_3 = "b g c d" D_4= "b d e" D_5= "a b e g" D_6= "b g h"

Kelime	:	a	b	c	d	e	f	g	h	
N(w)	:	2	6	2	3	3	1	3	1	N=6
$N - N_w/N_w$	:	$\frac{4+0.5}{2+0.5}$	$\frac{0+0.5}{6+0.5}$	$\frac{4+0.5}{2+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{5+0.5}{1+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{5+0.5}{1+0.5}$	

Sorgu = "a c h" :

$$\mathbf{rank_{D_1}} = \prod_{w \in \mathbf{D_1} \cap Q} \frac{N - \mathrm{N_w} + 0.5}{\mathrm{N_w} + 0.5} = \frac{4.5}{2.5} * \frac{4.5}{2.5} = 3.24$$

$$\mathbf{rank_{D_2}} = 1$$

$$\mathbf{rank_{D_3}} = \frac{4.5}{2.5} = 1.8$$

$$\mathbf{rank_{D_4}} = 1$$

$$\mathbf{rank_{D_5}} = \frac{4.5}{2.5} = 1.8$$

$$\mathbf{rank_{D_6}} = \frac{5.5}{1.5} = 3.66$$

Ranklar :

D_6

D_1

D_3

D_5

D_2

D_4

BIM Varsayımları

- Maron & Kuhns, 1960: Bir bilgi erişim sistemi dokümanların ilişki seviyesini belirleyemeyecek durumda ise ilişkili olma olasıklarına göre değerlendirilmelidir.
- BIM ‘de kabul edilebilir olasılıksal tahminlerde bulunabilmek için :
  - Dökümanların, sorgunun ve ilişkilerin boolean olarak ifade edilmesi
  - Terimlerin birbirinden bağımsız olması
  - Sorguda bulunmayan terimlerin sorgu sonucuna etkisizliği
  - Doküman ilişki durumunun diğer dokümanlardan bağımsız olması varsayımlarında bulunmak gerekmektedir.

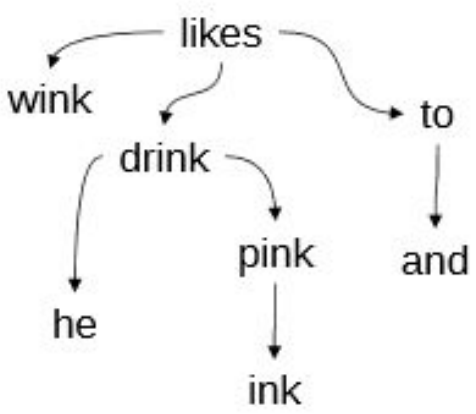
Çözüm ?

Kelime bağımlı modeller

Kelime Bağımlı Modeller

- Geleneksel modeller (BIM gibi) tüm kelimeleri bağımsız kabul eder.
  - Neredeyse tüm bilgi erişim sistemlerinin ortak eksiği ve üzerinde düşünülen varsayımıdır

- Kelime bağımlı modeller sorgudaki kelimelerin olasılığa ayrı ayrı etki etmesi yerine bağımlı olarak etki etmelerini sağlamaktadır.
  - Bağımlılıklar en küçük kapsar ağaç olara
  - Her bir kelime ebeveynine bağımlıdır.



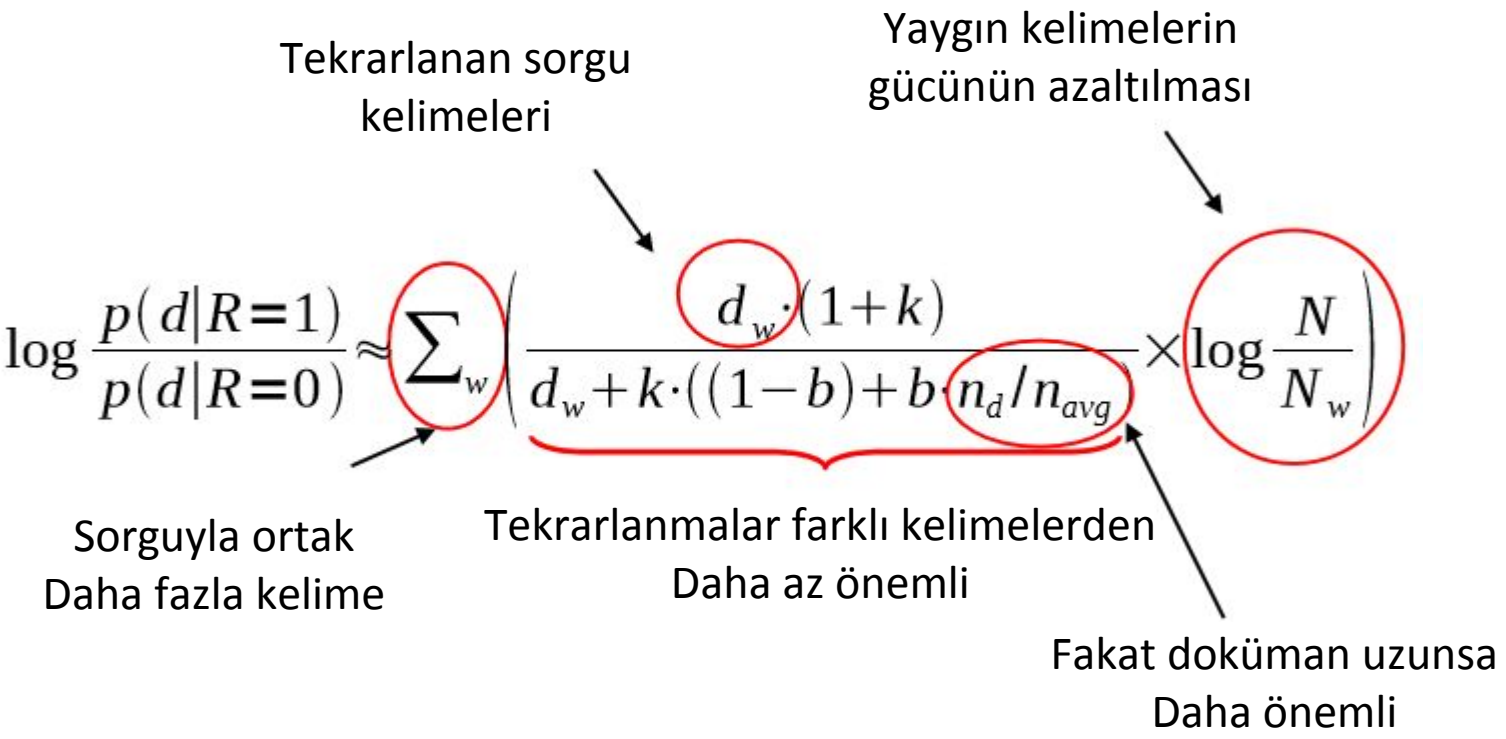
P("he likes to wink and drink pink ink")

= P(likes) * P(to|likes) * P(wink|likes)
* P(and|to) * P(drink|likes) * P(he|drink)
* P(pink|drink) * P(ink|pink)

Okapi BM25

- BIM kısa kataloglar için tasarlanmıştı ve bu koşullar altında oldukça iyi sonuç vermektedir
- Modern metin arama işlemlerinde , model terimlerin frekanslarına ve dokümanın uzunluğuna da önem vermek zorundadır
- BestMatch25 (BM25 veya Okapi) bu parametrelere hassasiyet göstermektedir
- BM25 1994 ten bugüne kadar en çok kullanılan ve hala popülaritesini koruyan bilgi erişim modelidir.

Okapi BM25


Tekrarlanan sorgu kelimeleri
Yaygın kelimelerin gücünün azaltılması
$\log \frac{p(d|R=1)}{p(d|R=0)} \approx \sum_w \left(\frac{d_w \cdot (1+k)}{d_w + k \cdot ((1-b) + b \cdot n_d / n_{avg})} \times \log \frac{N}{N_w} \right)$
Sorguyla ortak
Daha fazla kelime
Tekrarlanmalar farklı kelimelerden
Daha az önemli
Fakat doküman uzunsa
Daha önemli

Okapi BM25 (örnek)

Dokümanlar D_1= "a b c b d" D_2= "a b e f b"

D_3 = "b g c d" D_4= "b d e" D_5= "a b e g " D_6= "b g h h"

Sorgu = "a c h" , k=1 ve b=0.5

Kelime	:	a	b	c	d	e	f	g	h	
N(w)	:	2	6	2	3	3	1	3	1	N=6
$N - N_w/N_w$	:	$\frac{4+0.5}{2+0.5}$	$\frac{0+0.5}{6+0.5}$	$\frac{4+0.5}{2+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{5+0.5}{1+0.5}$	$\frac{3+0.5}{3+0.5}$	$\frac{5+0.5}{1+0.5}$	

$$\mathbf{rank_{D_1}} = \log \frac{P(D_1|R=1)}{P(D_1|R=0)} = 2 * \left(\frac{1*(1+1)}{1+1*\left(0.5+\left(0.5*\frac{5}{4}\right)\right)} * \log \frac{6+1}{2+0.5} \right)$$

$$\mathbf{rank_{D_6}} = \log \frac{P(D_6|R=1)}{P(D_6|R=0)} = 1 * \left(\frac{2*(1+1)}{2+1*\left(0.5+\left(0.5*\frac{4}{4}\right)\right)} * \log \frac{6+1}{1+0.5} \right)$$

Saf Bayesian yaklaşımlar ve Dil Modelleri

- Döküman d'yi ve sorgu q'yu sınıflandırmak istiyoruz.
  - Coğrafi sınıflandırma veya her döküman bir sınıf

- Sorgu ve dökümanın üretken bir model (generative model) tarafından üretildiğini düşünürsek

Saf Bayesian yaklaşımlar ve Dil Modelleri

- “üretken model”: Sözel bir girdi için dilin tüm elemanlarını ve sadece dilin elemanlarını üreten sonlu kurallar yaklaşımını simgeler. (Google)

- Hangi sınıfların belirlenen dökümanı ve sorguyu üretme şansı vardır?

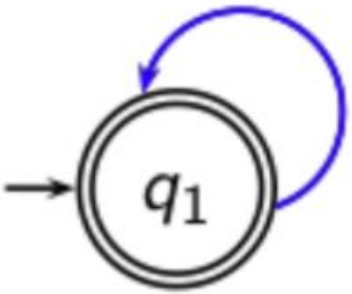
Dil Modelleri (DM)

- Dökümanı sorguyu üreten bir üretken model olarak düşünürsek
- Yapmamız gerekenler:
  - Üretken modeli tanımlanır.
  - Parametreler belirlenir.
  - Sıfır noktaları düzenlenir.
  - Sorguya uygulanarak o sorguyu üretmesi en muhtemel döküman seçilir.

Dil Modelleri (DM)

- Sonlu durum makinalarını (Finite-state automata) deterministik bir dil olarak kabul edebiliriz.
- Bizim modelimizde her döküman bunun gibi farklı sonlu durum otomatları tarafından üretilmiştir. Ancak bunlar olasılıksal dönüşüm fonksiyonuna sahiptirler.

Olasılıksal bir dil modeli



w	$P(w\|q_1)$	w	$P(w\|q_1)$
STOP	0.2	toad	0.01
the	0.2	said	0.03
a	0.1	likes	0.02
frog	0.01	that	0.04
		...	...

string = “frog likes toad”
P(string) = 0.01 * 0.02 * 0.01

Olsalılıksal dil modellerinin BE sistemlerinde kullanımı

- Sorgu hangi dökümanın olasılıksal dil modelinde daha olasıysa o döküman sorguyla daha benzerdir.
- Dökümanlar $P(d \mid q) = P(q \mid d) * P(d) / P(q)$ değerleri için sıralanır.
  - P(q) bütün dökümanlar için aynı
  - P(d) çoğunlukla ihmal edilir.
  - $P(d \mid q) =\sim P(q \mid d)$

Olsalılıksal dil modellerinin BE sistemlerinde kullanımı

- DM yaklaşımı için üretim modeli tanımlanmalıdır.
- Dökümanlar sorgunun dökümanların modelinden rassal bir çıktı olarak elde edilme olasılığına göre sıralanmalı

$$P(q \mid M_d) = \prod_{herterim} P(t \mid M_d)^{tf_{t,q}}$$

$$\hat{P}(t \mid M_d) = \frac{tf_{t,d}}{|d|}$$

Yumuşatma ve karışık model

$$P(q|d) \propto \prod_{1 \leq k \leq |q|} (\lambda P(t_k|M_d) + (1-\lambda)P(t_k|M_c))$$

- Terimlerin koleksiyon frekansını da katarak yumuşatır. (Sıfır değer çıktısını engeller.)
- Eşitlik dökümanın kullanıcının istediği döküman olma olasılığını verir.

Örnek

- d1 : “Jackson was one of the most talented entertainers of all time”
- d2: “Michael Jackson anointed himself King of Pop”
- Sorgu q: “Micheal Jackson” ve $\lambda = ½$ için
  - $P(q \mid d1) = [(0/11 + 1/18)/2] \cdot [(1/11 + 2/18)/2] \approx 0.003$
  - $P(q \mid d2) = [(1/7 + 1/18)/2] \cdot [(1/7 + 2/18)/2] \approx 0.013$

Kaynaklar

[1]http://160592857366.free.fr/joe/ebooks/ShareData/Modern%20Information%20Retrieval%20-%20A%20Brief%20Overview.pdf