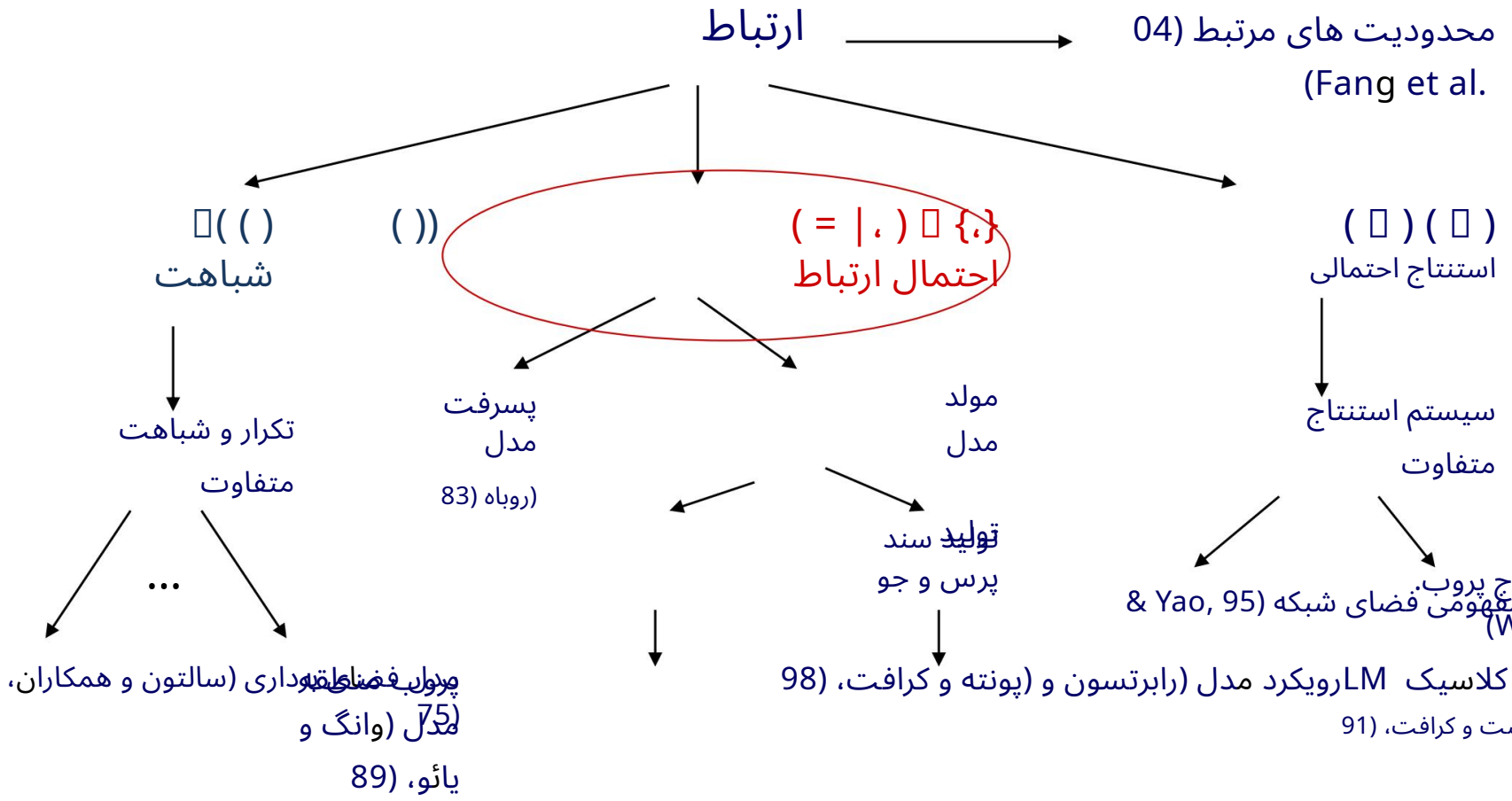
مدل های بازیابی:

احتمالی

بازیابی هوشمند اطلاعات

مفهوم ارتباط



اصل رتبه بندی احتمال

[رابرتسون [77

- برگرداندن یک لیست رتبه بندی شده از اسناد به ترتیب احتمال نزولی که یک سند با پرس و جو مرتبط است، استراتژی بهینه تحت دو فرض زیر است:

1. سودمندی یک سند برای کاربر است مستقل از کاربرد هر سند دیگری.

2. یک کاربر نتایج را به ترتیب مرور می کند

طبق ،PRPتنها چیزی که نیاز داریم این است

"یک تابع اندازه گیری ارتباط"

که راضی می کند

برای , 1 , 2 ,

(2 ،) > (1 ،) (2 ،) < (1 ،)

سوال اساسی

احتمال اینکه THISسند با THIS queryمرتبط باشد چقدر است؟

به طور رسمی...

3متغیر تصادفی: پرس و جو، ارتباط سند {0،1}

با توجه به یک پرس و جو خاص , یک سند خاص ? = = = 1 =
(| ،)

انحراف ...

بررسی مختصر احتمال

مفاهیم اساسی در احتمال

•آزمایش تصادفی: آزمایشی با نتیجه نامشخص
(به عنوان مثال، پرتاب یک سکه، انتخاب یک کلمه از متن)

•فضای نمونه: همه نتایج ممکن، به عنوان مثال، -پرتاب 2سکه منصفانه، ، } =

اگر نتیجه در { } = -(همه سرها) { } = -(همان صورت) اتفاق می افتد

•رویداد:

،به عنوان مثال،

•احتمال رویداد: 1 = ( ) خیلیان(نتیجه همین) در (S

( ) + ( ) = ( ) –اگر = ( ) (مثلاً =صورت یکسان، =چهره متفاوت)

مفاهیم اساسی .Prob(ادامه)

- احتمال شرطی: ( ) /(∩) = (|)
  - (∩) = ( ) (|) = ( ) (|)
  - بنابراین، ( ) /( ) (|) = (|)(قاعده بیز)
  - اگر و مستقل، ( ) = (|)
- احتمال کل: اگر پارتیشنی از (∩) + ∩ + (∩) = ( ) تشکیل دهید
  - ( ) = ( ) (|) + ( ) (|) + + ( ) (|)
  - بنابراین، ( ) /( ) (|) = (|)
  - (|) = ( ) (|) /[( ) (|) + ∩ + ( ) (|)]
  - این به ما امکان می دهد (|)را بر اساس (|)محاسبه کنیم.

تفسیر قانون بیز

{شواهد و مدارک: فضای فرضیه: ... ،1 } =

$$(\mid) = \frac{(\mid) \quad ( )}{( )}$$

می توانیم رها کنیم ( ) ؛ اگر بخواهیم محتمل ترین فرضیه را انتخاب کنیم

احتمال خلفی از

احتمال قبلی از

(|) (|) ( )

احتمال داده ها/شواهد

درست است اگر

انتهای مسیر انحرافی ...

حالا بیایید به بازیابی برگردیم!

مدل‌های بازیابی احتمالی: شهود

فرض کنید تعداد زیادی قضاوت مرتبط داریم (مثلاً تعداد کلیک: = «1»کلیک؛ =
«0»رد شده)

ما می توانیم اسناد را بر اساس Doc(D) Rel(R)؟

پرس و جو (Q)		
Q1		1
Q1	D2	1
Q1	D3	0
Q1	D4	0
Q1	D5	1
...		
Q1	D1	0
Q1	D2	1
Q1	D3	0
Q2	D3	1
Q3	D1	1
Q4	D2	1
Q4	D3	0
...		

$$(,) = \quad D1$$

$$(= 1 | ,) = \frac{(, , = 1)}{(,)}$$

$$(|) \quad = 1\ 1, 1 = ?/2$$

$$(= 1 | 1, 2) = ?/2$$

$$(= 1 | 1, 3) = ?/2$$

چه می شود اگر ما به اندازه کافی نداشته باشیم

ورود به سیستم جستجو؟

اسناد/پرس و جوهای دیده نشده؟

می توانیم تقریبی کنیم (, | 1 =)

احتمال ارتباط

- سه متغیر تصادفی
  - پرس و جو
  - سند
  - ارتباط $\{0,1\}$
- هدف: رتبه بر اساس $P(R=1 \mid d, q)$
  - ارزیابی $P(R=1 \mid d, q)$
  - در واقع، فقط باید $P(R=1 \mid d_1, q)$ را با $P(R=1 \mid d_2, q)$ مقایسه کنید، یعنی اسناد را رتبه بندی کنید
- چندین روش مختلف برای پالایش $P(R=1 \mid d, q)$

پالایش| :1)مدل های شرطی

•ایده اصلی: ارتباط بستگی به این دارد که یک پرس و جو چقدر با یک سند مطابقت دارد

به عنوان مثال، #مطابقات، بالاترین IDFیک ترم منطبق، ،doclenیا حتی (،)با توجه به هر تابع بازیابی دیگری،...

–ویژگی ها را در ×تعریف کنید

– (= 1 | ,) =

(1 (,), 2 (,), … , (,))

-استفاده از داده های آموزشی (قضاوت های مرتبط شناخته شده) برای پارامتر برآورد

–از مدل برای رتبه بندی اسناد جدید استفاده کنید

پالایش|:2≠) پالایش
مدل های تولیدی

•ایده پایه

-تعریف کردن ،) |)

(, | 1 =) –را با استفاده از قانون بیز محاسبه کنید

$$(\quad = 1 | \ , \quad) = \frac{(= 1 | \ ,)}{(= 0 | \ ,)} = \frac{(\ , \quad | = 1)}{(\ , \quad | = 0)} \times \frac{(= 1)}{(= 0)}$$

برای رتبه Dنادیده گرفته شد

•موارد خاص

(| :"=لسن",)لسن-)

– Query “generation”: (, | ,))=((|)

تولید سند

$$\left(= 1 \mid Q, D \right) = \frac{(= 1 \mid ,)}{(= 0 \mid ,)} = \frac{(, \mid) = 1 (= 1)}{(, \mid) = 0 (= 0)}$$

$$\square \quad \frac{\mid (= 1)}{\mid = 0)} = \frac{(\mid ,) = 1 (\mid = 1)}{(\mid ,) = 0 (\mid = 0)}$$

$$\square \quad \frac{, (\mid = (1 \mid)}{, = 0)}$$

مدل اسناد مربوطه برای Q ←

مدل اسناد غیر مرتبط برای Q ←

تولید سند (ادامه)

- ویژگی های مستقل را فرض کنید ... که در آن { , } مقدار ویژگی است
- اجازه دهید =(به طور مشابه (... = ، ...

(| (|

... ، = 1) ، = 0)

| (= | ، = 1) ، = 0)

(=

=1

= (= 1 | , (= 1 |) ، = 0) (= 0 | , (= 0 |) ، = 0)

=1، =1 =1، =0

تولید سند (ادامه)

$$\ldots = \prod_{=1,\ =1} \frac{(= 1 \mid , (= 1 \mid = 0)}{= 0)} \prod_{=1,\ =0} \frac{(= 0 \mid , (= 0 \mid = 0)}{= 0)}$$

$$\times \prod_{=1,\ =1} \frac{(= 0 \mid , \quad = 1)}{(= 0 \mid , \quad = 0)} \times \prod_{=1,\ =1} \frac{(= 0 \mid , (= 0 \mid , \quad = 0)}{= 1)}$$

$$= \prod_{=1,\ =1} \frac{(= 1 \mid , (= 1 \mid , \quad = 1)}{= 0)} \cdot \frac{(= 0 \mid , \quad = 0)}{(= 0 \mid , \quad = 1)} \boxed{\prod_{=1} \frac{(= 0 \mid , (= 0 \mid , \quad = 1)}{= 0)}}$$

برای رتبه بندی نادیده بگیرید

تولید سند (ادامه)

• فرض کنید 1 = ( ) 0 = (عبارت در پرس و جو ظاهر نمی شود) 1 = | (0 =)

$$\cdots = \frac{(= 1 | , (= 1 | = 1)}{= 0)} \cdot \frac{(= 0 | , \quad = 0)}{(= 0 | , \quad = 1)} \qquad \frac{(= 0 | , (= 0 | = 1)}{= 0)}$$

=1، =1 ... =1

برای رتبه بندی نادیده بگیرید

$$= \frac{(= 1 | , (= 1 | , \; = 1)}{= 0)} \cdot \frac{(= 0 | , (= 0 | , \; = 0)}{= 1)}$$

=1، = =1

مدل رابرتسون-اسپارک جونز

(رابرتسون و اسپارک جونز 76)

مدل (RSJ

(1-)

ورود به سیستم

(1-)

(= 1 | ,) ورود به سیستم

=1، = =1

دو پارامتر برای هر عبارت: | = (= | = (=

مشکل آن عبارت در یک سند مرتبط =): prob. این=): اصطلاح در یک سند ،

غیر مرتبط وجود دارد ،

چگونه پارامترها را تخمین بزنیم؟

فرض کنید قضاوت های مرتبطی داریم،

= #(·) + 0.5 / #(·) + 1

=

مدل :RSJبدون اطلاعات مرتبط

(کرافت و هارپر (79

(مدل RSJ)

ورود به سیستم (, | 1 =)

=1، = =1

چگونه پارامترها را تخمین بزنیم؟

فرض کنید قضاوت مربوطه نداریم، -ما ثابت فرض می کنیم -با فرض غیر مرتبط بودن همه اسناد، تخمین بزنید.

ورود به سیستم (, | 1 =)

=1، = =1

#اسنادی که عبارت در آنها وجود دارد :#اسناد در مجموعه :

مدل RSJ: خلاصه

- مهم ترین پروب کلاسیک. مدل IR • فقط از عبارت حضور/غیبت استفاده کنید، بنابراین به عنوان مدل استقلال باینری نیز شناخته می شود
- اساسا بیز ساده برای رتبه‌بندی اسناد • طبیعی‌ترین بازخورد مرتبط/شبه • وقتی بدون قضاوت مرتبط بودن، پارامترهای مدل باید به‌صورت موقتی تخمین زده شوند.

- عملکرد به خوبی مدل VS تنظیم شده نیست

بهبود :RSJاضافه کردن TF

سند پایه مدل نسل:

تعداد دفعات ترم کجاست ، ... بگذار =

2- مدل مخلوط پواسون

بسیاری از پارامترهای بیشتر برای تخمین زدن!

تقریبی BM25/Okapi

(رابرتسون و همکاران 94)

• Idea: تقریبی $P(R=1|q,d)$ با یک تابع ساده تر که دارای ویژگی های مشابه است

• مشاهدات:

$\log P(R=1|q,d)$ – مجموع وزن های عبارت است، $0 = P(R=1|q,d) = 0$ – اگر $0 =$

-یکنواخت با افزایش می یابد -دارای حد مجانبی است

• تابع ساده $=$ است $\frac{(k_1+1)\,tf}{k_1+tf}$ $\frac{(K+1)\,n}{K+n}$ ورود به سیستم

افزودن Doc. طول و پرس و جو TF

- گنجاندن طول سند
  - "با دقت" سند طولانی را جریمه کنید
- ترکیب پرس و جو TF
  - یک تبدیل TF مشابه
- فرمول نهایی BM25 نامیده می شود که به بالا می رسد عملکرد TREC

فرمول BM25

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \quad (1)$$

Okapi TF/BM25 TF

where

Q is a query, containing terms T

$w^{(1)}$ is the Robertson/Sparck Jones weight [5] of T in Q

$$\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \quad (2)$$

N is the number of items (documents) in the collection

n is the number of documents containing the term

R is the number of documents known to be relevant to a specific topic

r is the number of relevant documents containing the term

K is $k_1((1-b)+b.dl/avdl)$

k_1, b and k_3 are parameters which depend on the on the nature of the queries and possibly on the database; k_1 and b default to 1.2 and 0.75 respectively, but smaller values of b are sometimes advantageous; in long queries k_3 is often set to 7 or 1000 (effectively infinite)

tf is the frequency of occurrence of the term within a specific document

qtf is the frequency of the term within the topic from which Q was derived

dl and $avdl$ are respectively the document length and average document length measured in some suitable unit.

پسوندهای
مدل های "نسل سند".

- وابستگی اصطلاحی را ضبط کنید (Rijsbergen & Harper 78)
- روش های جایگزین برای ترکیب TF (Croft 83، Kalt96)
- انتخاب ویژگی/مدت برای بازخورد (TREC Okapi گزارش ها)
- برآورد مدل ربط بر اساس بازخورد شبه [Lavrenko & Croft 01]

سوالات؟