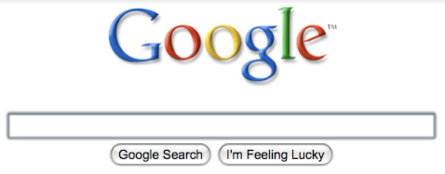
Learning to Rank

srihari@buffalo.edu

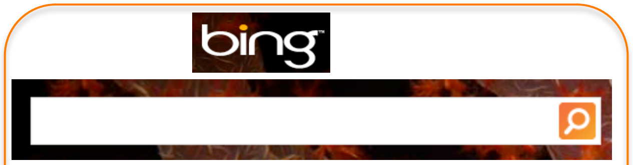
Learning to Rank Problem

- LeToR
- Multiple Inputs
- Target Value
  - t is discrete (eg, 1,$2..6$) in training set but a continuous value in $[1,6]$ is learnt and used to rank objects

Regression with multiple inputs: LeToR

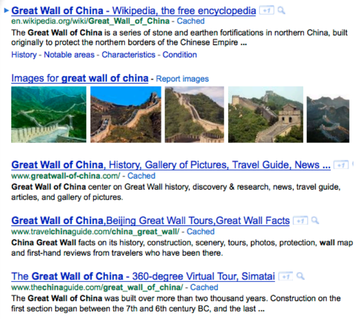


$(d > 200)$



In LETOR 4.0 dataset
46 query-document features
Maximum of 124 URLs/query

Yahoo! data set has $d=700$



Great Wall of China - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Great_Wall_of_China** - Cached
The **Great Wall of China** is a series of stone and earthen fortifications in northern China, built originally to protect the northern borders of the Chinese Empire ...
History - Notable areas - Characteristics - Condition

Images for **great wall of china** - Report images

Great Wall of China, History, Gallery of Pictures, Travel Guide, News ...
www.**greatwall-of-china**.com/ - Cached
Great Wall of China center on Great Wall history, discovery & research, news, travel guide, articles, and gallery of pictures.

Great Wall of China,Beijing Great Wall Tours,Great Wall Facts
www.travel**chinaguide**.com/**china_great_wall**/ - Cached
China Great Wall facts on its history, construction, scenery, tours, photos, protection, **wall** map and first-hand reviews from travelers who have been there.

The **Great Wall of China** - 360-degree Virtual Tour, Simatai
www.the**chinaguide**.com/**great_wall_of_china**/ - Cached
The **Great Wall of China** was built over more than two thousand years. Construction on the first section began between the 7th and 6th century BC, and the last ...

Input (x_i):
(d Features of Query-URL pair)

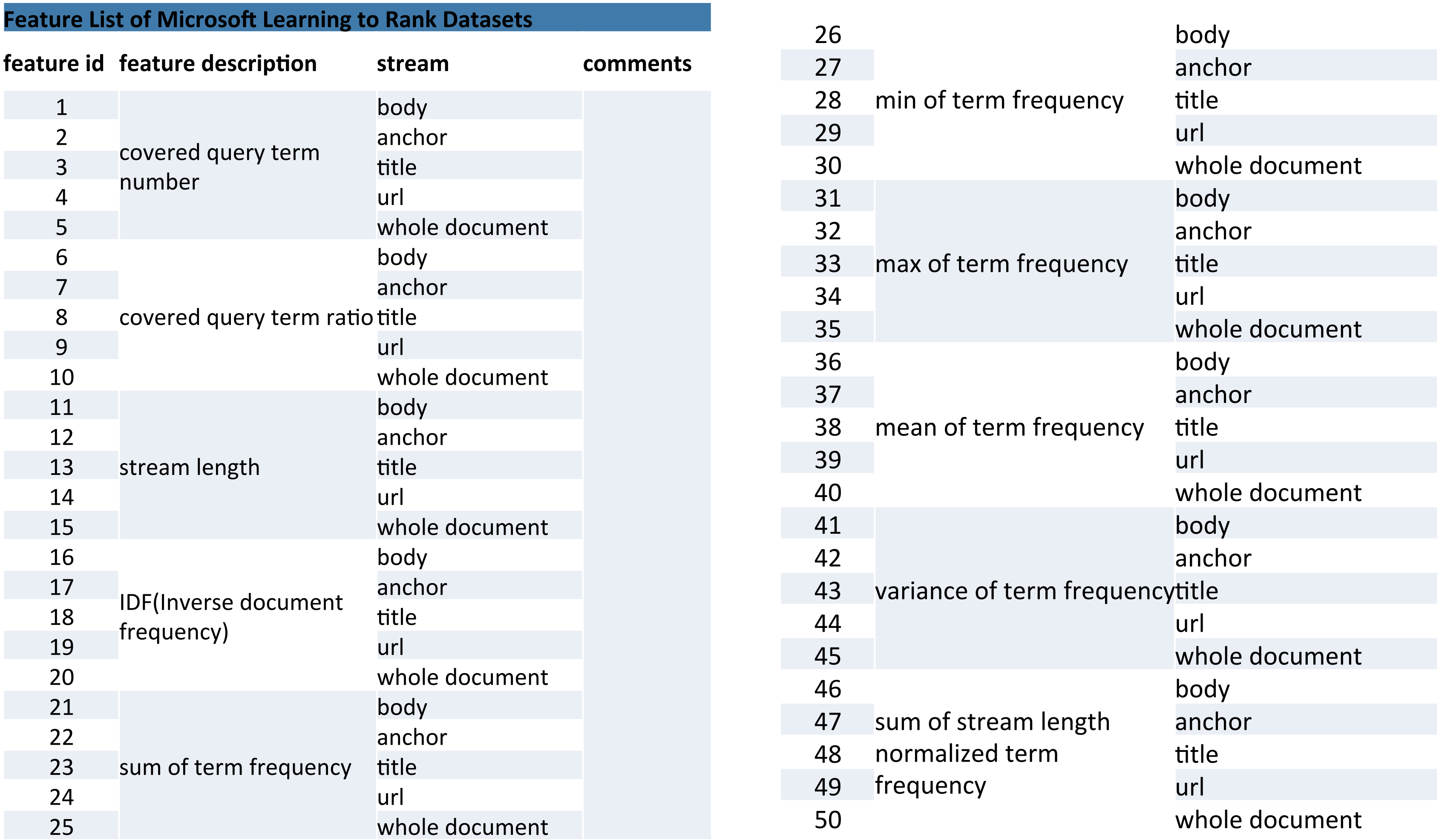
- Log frequency of query in anchor text
- Query word in color on page
- # of images on page
- # of (out) links on page
- PageRank of page
- URL length
- URL contains "~"
- Page length
- TF/IDF

Target (t):
Relevance Value
(0,1,2):higher value is better match

Regression returns
continuous value, y
Allows fine-grained ranking of URLs

Query-URL Features

See http://research.microsoft.com/en-us/projects/mslr/feature.aspx



Feature List of Microsoft Learning to Rank Datasets

feature id	feature description	stream	comments
1	covered query term number	body	
2		anchor	
3		title	
4		url	
5		whole document	
6	covered query term ratio	body	
7		anchor	
8		title	
9		url	
10		whole document	
11	stream length	body	
12		anchor	
13		title	
14		url	
15		whole document	
16	IDF(Inverse document frequency)	body	
17		anchor	
18		title	
19		url	
20		whole document	
21	sum of term frequency	body	
22		anchor	
23		title	
24		url	
25		whole document	
26	min of term frequency	body	
27		anchor	
28		title	
29		url	
30		whole document	
31	max of term frequency	body	
32		anchor	
33		title	
34		url	
35		whole document	
36	mean of term frequency	body	
37		anchor	
38		title	
39		url	
40		whole document	
41	variance of term frequency	body	
42		anchor	
43		title	
44		url	
45		whole document	
46	sum of stream length normalized term frequency	body	
47		anchor	
48		title	
49		url	
50		whole document	

raw frequency: $tf(t,d) = f_{t,d}$, log-normalized: $1 + \log(f_{t,d})$

$idf(t,D) = \frac{\log N}{|\{d \in D: t \in d\}|}$ N=|D|. no. of docs in corpus, $|\{d \in D: t \in d\}|$: no. of docs where term t appears

Feature Statistics

- Most of 46 features are normalized as continuous values from 0 to 1, exception some features are all 0s'.

Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1
Mean	0.254	0.1598	0.1392	0.2158	0.1322	0.1614	0	0	0	0	0	0.2841	0.1382	0.2109	0.1218

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2879	0.1533	0.2258	0.3057	0.3332	0.1534	0.5473	0.5592	0.5453	0.5622	0.1675	0.1377	0.1249	0.126	0.2109	0.1705

32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
0.1694	0.1603	0.1275	0.0762	0.0762	0.0728	0.5479	0.5574	0.5502	0.5673	0.4321	0.3361	0	0.1065	0.1211

Returning to LeToR Problem

- Try:
- Several Basis Functions
- Quadratic Regularization
- Express results as E_{RMS}
  - rather than as squared error $E(\mathrm{w}^*)$ or as Error Rate with thresholded results

$$E_{RMS} = \sqrt{2E(\mathrm{w}^*)/N}$$