

# 情報検索システム特論

## Advanced Information Retrieval Systems

### 第3回 Lecture #3の訂正

2023-05-01

湯川 高志 Takashi Yukawa

# Documents

- ▶  $d_1$ :  
Information retrieval (IR) in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources.
- ▶  $d_2$ :  
The World Wide Web (WWW), commonly known as the Web, is the world's dominant software platform.[1] It is an information space where documents and other web resources can be accessed using a web browser and (more recently) web-based applications.
- ▶  $d_3$ :  
The quick brown fox jumps over the lazy dog.

# Document -> Set of Index Terms

- ▶  $d_1$ :  
information retrieval compute information science  
process obtain information system resource relevant  
information need collection resource
- ▶  $d_2$ :  
world wide web common know web world dominant  
software platform information space document web  
resource access use web browser recent web  
application
- ▶  $d_3$ :  
quick brown fox jump lazy dog

# Put ID Numbers to the Index Terms

Information:1  
retrieval:2  
compute:3  
science:4  
process:5  
obtain:6  
system:7  
resource:8  
relevant:9  
need:10  
collection:11  
world:12

wide:13  
web:14  
common:15  
know:16  
dominant:17  
software:18  
platform:19  
space:20  
document:21  
access:22  
use:23  
browser:24

recent:25  
application:26  
quick:27  
brown:28  
fox:29  
jump:30  
lazy:31  
dog:32

# Document x ID Numbers of Index Terms

- ▶  $d_1$ :  
information=1 retrieval=2 compute=3 information=1 science=4  
process=5 obtain=6 information=1 system=7 resource=8  
relevant=9 information=1 need=10 collection=11 resource=8  
{1,2,3,1,4,5,6,1,7,8,9,1,10,11,8}
- ▶  $d_2$ :  
world=12 wide=13 web=14 common=15 know=16 web=14  
world=12 dominant=17 software=18 platform=19 information=1  
space=20 document=21 web=14 resource=8 access=22 use=23  
web=14 browser=24 recent=25 web=14 application=26  
{12,13,14,15,16,14,12,17,18,19,1,20,21,14,8,22,23,14,24,25,14,26}
- ▶  $d_3$ :  
quick=27 brown=28 fox=29 jump=30 lazy=31 dog=32  
{27,28,29,30,31,32}

Redundant tokens are included in the previous material

# Document x Term Frequency

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$d_1$	4	1	1	1	1	1	1	2	1	1	1	0	0	0	0	0
$d_2$	1	0	0	0	0	0	0	1	0	0	0	2	1	5	1	1
$d_3$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$d_1$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$d_2$	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
$d_3$	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1

# Document x TF Value

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$d_1$	1	0.25	0.25	0.25	0.25	0.25	0.25	0.5	0.25	0.25	0.25	0	0	0	0	0
$d_2$	0.2 2	0	0	0	0	0	0	0.2 2	0	0	0	0.4 4	0.2 2	1	0.2 2	0.2 2
$d_3$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$d_1$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$d_2$	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0.2 2	0	0	0	0	0	0
$d_3$	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1

# IDF Values

$$N = 3$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$n_i$	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
$\frac{N}{n_i}$	1.5	3	3	3	3	3	3	1.5	3	3	3	3	3	3	3	3
idf	0. 18	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 18	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$n_i$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\frac{N}{n_i}$	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
idf	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48	0. 48



# Document x TF-IDF Value

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$d_1$	0.18	0.12	0.12	0.12	0.12	0.12	0.12	0.09	0.12	0.12	0.12	0	0	0	0	0
$d_2$	0.04	0	0	0	0	0	0	0.04	0	0	0	0.19	0.1	0.48	0.1	0.1
$d_3$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

  

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$d_1$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$d_2$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0	0	0	0	0
$d_3$	0	0	0	0	0	0	0	0	0	0	0.48	0.48	0.48	0.48	0.48	0.48

$$|\vec{d_1}| = 0.41$$

$$|\vec{d_2}| = 0.63$$

$$|\vec{d_3}| = 1.18$$