

Database_한국어1분반 Team 6

Phrase 1

22000758 최윤성

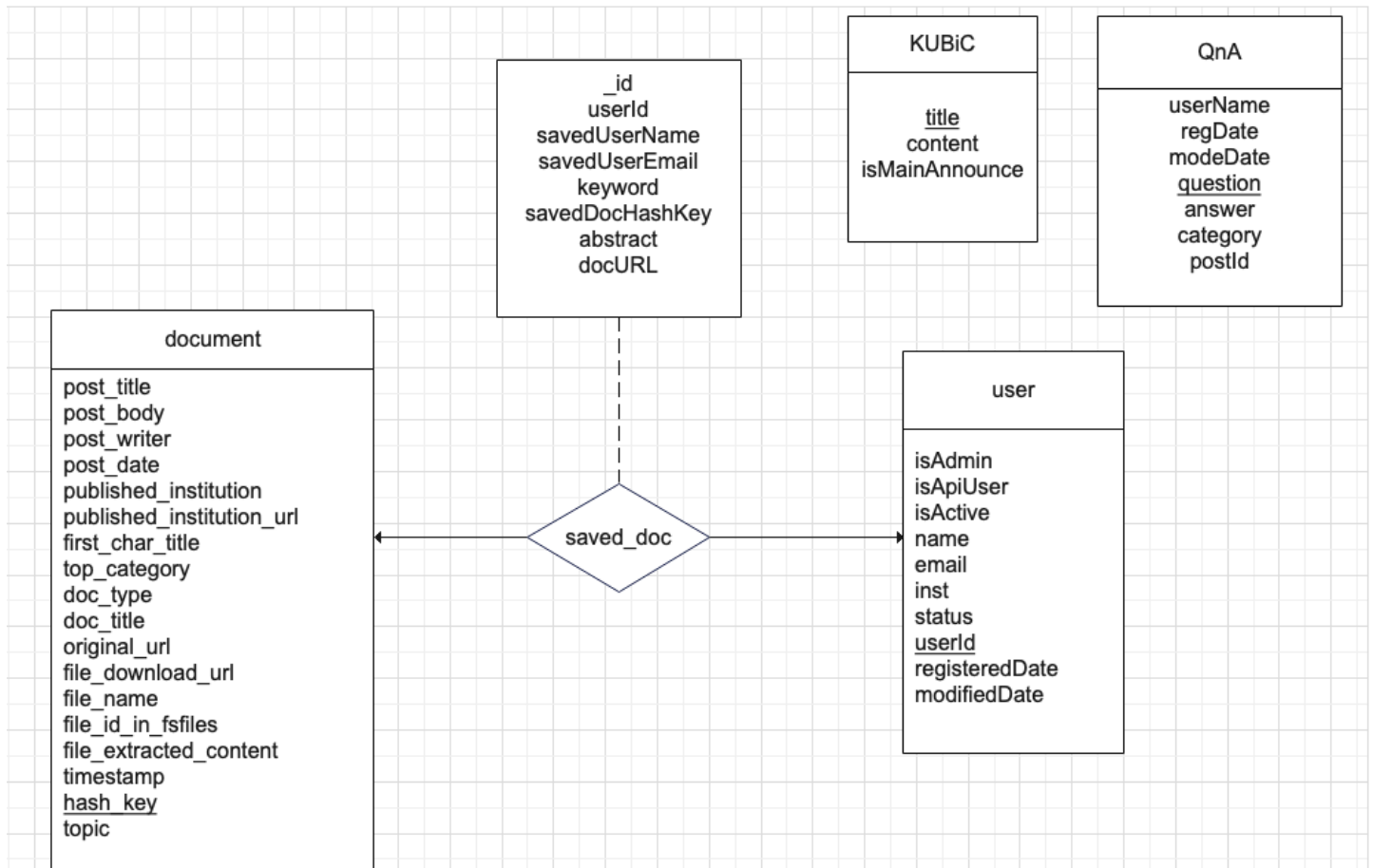
22100511 이선환

22200527 이성현

a) 문제를 해결한 방법에 대한 자세한 설명

대표되는 값들과 1대1 대응되는 attribute들을 count함수를 사용하여 찾아 분류하고, record에서 중복성을 띠는 attribute들을 normalize하였다.

b) 구현된 database의 E-R diagram



c) table의 모든 schema

document

	Field	Type	Null	Key
1	post_title	varchar(255)	YES	
2	post_body	mediumtext	YES	
3	post_writer	varchar(255)	YES	
4	post_date	varchar(15)	YES	
5	published_institution	varchar(20)	YES	
6	published_institution_url	varchar(512)	YES	
7	first_char_title	varchar(10)	YES	
8	top_category	varchar(50)	YES	
9	doc_type	varchar(10)	YES	
10	doc_title	mediumtext	YES	
11	original_url	varchar(512)	YES	
12	file_download_url	mediumtext	YES	
13	file_name	varchar(512)	YES	
14	file_id_in_fsfiles	varchar(30)	YES	
15	file_extracted_content	mediumtext	YES	
16	timestamp	varchar(30)	YES	
17	hash_key	varchar(30)	NO	PRI
18	topic	mediumtext	YES	

saved_doc

	Field	Type	Null	Key
1	_id	mediumtext	YES	
2	userId	varchar(30)	YES	MUL
3	savedUserName	varchar(255)	YES	
4	savedUserEmail	varchar(255)	YES	
5	keyword	varchar(255)	YES	
6	savedDocHashKey	varchar(128)	YES	MUL
7	abstract	varchar(200)	YES	
8	docURL	mediumtext	YES	

user

	Field	Type	Null	Key
1	isAdmin	tinyint	YES	
2	isApiUser	tinyint	YES	
3	isActive	tinyint	YES	
4	name	mediumtext	YES	
5	email	mediumtext	YES	
6	inst	mediumtext	YES	
7	status	mediumtext	YES	
8	userId	varchar(30)	NO	PRI
9	registeredDate	datetime	YES	
10	modifiedDate	datetime	YES	

KUBiC

	Field	Type	Null	Key
1	title	varchar(128)	NO	PRI
2	content	text	YES	
3	isMainAnnounce	tinyint	YES	

QnA

	Field	Type	Null	Key
1	userName	varchar(30)	YES	
2	regDate	datetime	YES	
3	modeDate	datetime	YES	
4	question	varchar(128)	NO	PRI
5	answer	text	YES	
6	category	varchar(10)	YES	
7	postId	int	YES	

d) view에 대한 설명

1

DDL Query : CREATE VIEW userCount AS (

SELECT COUNT(DISTINCT userId) AS COUNT FROM user

);

Size of the resulting table : 1

Table header and records :

userCount	
1	383

2

DDL Query : CREATE VIEW docCount AS (

SELECT COUNT(DISTINCT hash_key) FROM document

);

Size of the resulting table : 1

Table header and records :

count	
1	30317

3

DDL Query : CREATE VIEW instPubCount3 AS (

WITH instDocCount(published_institution, CNT) AS (

SELECT published_institution, COUNT(DISTINCT hash_key)

FROM document GROUP BY published_institution

),

instDocRank(published_institution, CNT, docRank) AS (

SELECT published_institution, CNT,

RANK() OVER (ORDER BY CNT DESC)

FROM instDocCount

)





SELECT published_institution, CNT

FROM instDocRank WHERE docRank = 3

);

Size of the resulting table : 1

Table header and records :

	 published_institution		 CNT	
1	통일부		3383	

4

DDL Query : CREATE VIEW yearPubCount AS (

SELECT SUBSTRING(post_date, 1, 4) AS post_year, COUNT(*) AS cnt

FROM document

GROUP BY

post_year

ORDER BY cnt DESC);

Size of the resulting table : 56

Table header and records :

	post_year	cnt
1	2009	4306
2	2008	2596
3	2007	2430
4	2006	2062
5	2017	1815

5

DDL Query : CREATE VIEW catSummary AS (

SELECT top_category, category_count,

RANK() OVER (ORDER BY category_count DESC) AS category_rank

FROM (

SELECT top_category, COUNT(DISTINCT hash_key) AS category_count

FROM document

WHERE top_category IS NOT NULL

GROUP BY top_category

) AS categoryCnt

);

Size of the resulting table : 139

	post_year	cnt
1	2009	4306
2	2008	2596
3	2007	2430
4	2006	2062
5	2017	1815

Table header and records :

6

DDL Query : CREATE VIEW crawlHist AS (

SELECT post_title, post_writer, published_institution, post_date, timestamp

FROM document

WHERE timestamp IS NOT NULL

ORDER BY timestamp

);

Size of the resulting table : 30316

Table header and records :

	post_title	post_writer	published_institution	post_date	timestamp
1	[2017년 3월호] 평화누리통일누리 163호(2017. 2. 28)	평통사	평화와통일을여는사람들	2017-04-03	2022-04-09 01:0...
2	[2017. 8월] 평화누리통일누리 통권 167호	평통사	평화와통일을여는사람들	2017-08-09	2022-04-09 01:0...
3	[2018년 7-8월 합본호] 평화누리통일누리 177호(2018. 8. 14)	평통사	평화와통일을여는사람들	2018-08-14	2022-04-09 01:0...
4	[2018년 6월호] 평화누리통일누리 176호(2018. 6. 27)	평통사	평화와통일을여는사람들	2018-06-26	2022-04-09 01:0...
5	[2018년 특별호] 평화누리통일누리 174호(2018. 3. 26)	평통사	평화와통일을여는사람들	2018-03-27	2022-04-09 01:0...

7

DDL Query : CREATE VIEW fileSummary AS (

SELECT timestamp, file_id_in_fsfiles, file_name, file_download_url

FROM document

WHERE file_id_in_fsfiles IS NOT NULL

);

Size of the resulting table : 12815

Table header and records :

	timestamp	file_id_in_fsfiles	file_name	file_download_url
1	2022-04-27 12...	62680dbdcddd8b369f71d5fa	sf_20090526.hwp	http://knsi.org/knsi/admin/work/works/mosf_20090526.hwp
2	2022-05-04 03...	62721c809861967f4338...	외교통일위원회 소위원회 위원 명단('21.01.07.현재).pdf	https://uft.na.go.kr:444/uft/reference/reference02.do?mode=downl...
3	2022-04-30 01...	626c149ef9d529df5543...	[주요 경제지표]로동 생산 능력의 장성	http://www.kinu.or.kr/com/file/filedown?_ci=92008_&ck=pmfmvthDNfo...
4	2022-04-29 11...	626bfa46fc8e4b822302...	To Build a National Community through the Korean Commonwealth : ...	http://www.kinu.or.kr/com/file/filedown?_ci=14388_&ck=967a2d52cd0...
5	2022-05-17 03...	62833a5d801db7deff24...	1082643494_MOPH_UNICEF_Medicine_Delivery_Plan_2003_(final).pdf	http://www.nkhealth.net/INC/download.php?code=sub_0302&number=28...

8

DDL Query : CREATE VIEW userInstCount2 AS (

SELECT institute, max_status, max_status_count, COUNT(*) AS data_count

FROM (SELECT inst AS institute,

status AS max_status,

COUNT(status) OVER (PARTITION BY status) AS max_status_count

FROM user

WHERE inst IS NOT NULL) AS instStatus

GROUP BY institute, max_status, max_status_count

ORDER BY data_count DESC

LIMIT 1 OFFSET 1

);

Size of the resulting table : 1

Table header and records :

	institute	max_status	max_status_count	data_count
1	한동대	대학생	350	23

9

DDL Query : CREATE VIEW docTopicQ9 AS (

WITH tfidf2021(word, topicCnt) AS (

SELECT tfidfWord, COUNT(DISTINCT docId)

FROM frequency

WHERE score > 0 and docId IN (

SELECT hash_key FROM document WHERE LEFT(post_date, 4) = '2021'

)

GROUP BY tfidfWord

)

SELECT word, topicCnt

FROM (

SELECT word, topicCnt,

RANK() OVER (ORDER BY topicCnt DESC) AS topicRank

FROM tfidf2021

) AS topicRank2021

WHERE topicRank = 4

);

Size of the resulting table : 1

Table header and records :

	word	topicCnt
1	미국	163

10

DDL Query : CREATE VIEW docTopicQ10 AS (

WITH tfidfInstCnt3(word, topicCnt) AS (

SELECT tfidfWord, COUNT(DISTINCT docId)

FROM frequency

WHERE score > 0 and docId IN (

SELECT hash_key FROM document

WHERE published_institution = (SELECT published_institution FROM instPubCount3)

)

GROUP BY tfidfWord

)

SELECT word, topicCnt

FROM (

SELECT word, topicCnt,

RANK() OVER (ORDER BY topicCnt DESC) AS topicRank

FROM tfidfInstCnt3

) AS topicRankInstCnt3

WHERE topicRank = 1

);

Size of the resulting table : 1

Table header and records :

	word	topicCnt
1	문헌	234

e) database의 크기와 **table**의 크기를 **KB** 단위로 요약

Size of the Database :

DatabaseName	Size(KB)
db_proj_06	3291312.0

Size of the tables :

TABLE_SCH EMA	TABLE_SIZE	data(KB)	idx(KB)
db_proj_06	KUBIC	16.0	0.0
db_proj_06	QnA	16.0	0.0
db_proj_06	document	28224.0	0.0
db_proj_06	frequency	3259392.0	0.0
db_proj_06	saved_doc	3600.0	0.0
db_proj_06	user	64.0	0.0