co occurrence

이지현

Intro

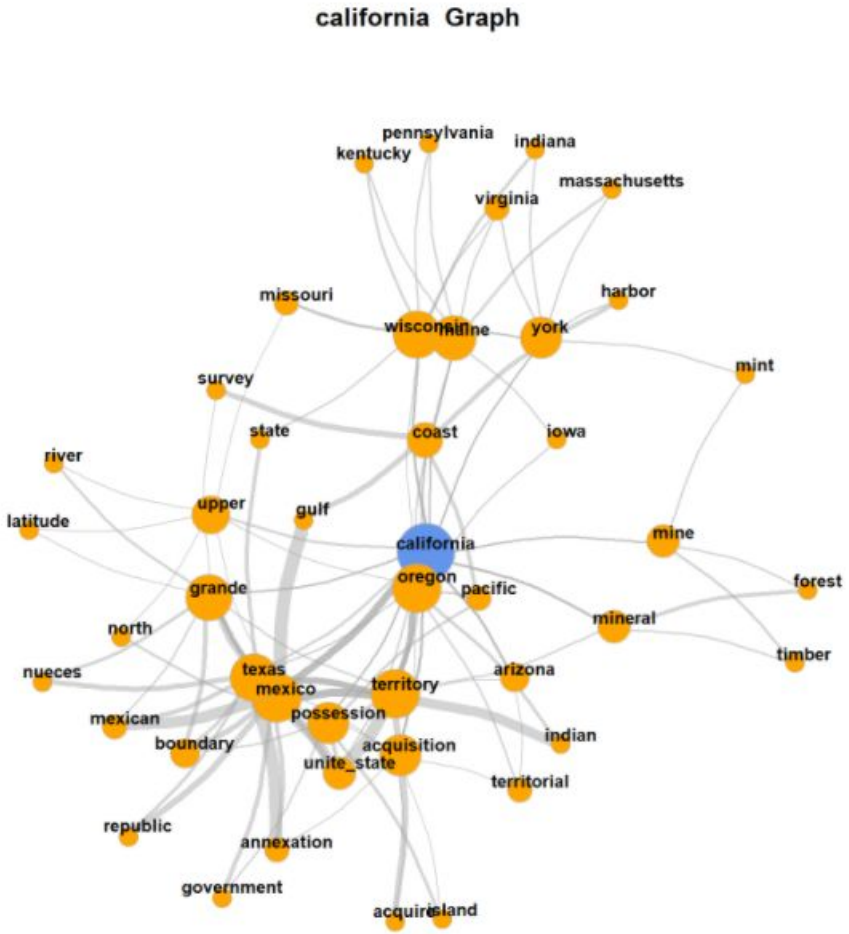
graphical representation of how frequently variables appear together

used in ecology and text mining.

ecology -> how frequently two species ar seen together within sampling site
textminnig -> how frequently two words are present in a single document.

pair들의 동시 출현을 탐색할 수 있다.

Intro


california Graph
pennsylvania
indiana
kentucky
massachusetts
virginia
missouri
harbor
wisconsin
maine
york
mint
survey
state
coast
iowa
river
upper
gulf
latitude
california
mine
grande
oregon
pacific
forest
north
mineral
timber
nueces
texas
mexico
territory
arizona
possession
mexican
indian
boundary
acquisition
unite_state
territorial
republic
annexation
government
acquire
island

word document matrix

한 문서에 단어의 등장 횟수를 행렬로 나타내는 방법.

전제 : 관계가 있는 단어가 같은 문서에서 빈번하게 등장할 것 이다.

$$X = \begin{array}{c} \\ \text{은행} \\ \text{주식} \\ \text{채권} \\ \text{금} \\ \text{원숭이} \\ \text{서울대공원} \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 11 & 0 \\ 6 & 0 \\ 5 & 0 \\ 8 & 0 \\ 0 & 5 \\ 0 & 4 \end{pmatrix} \end{array}$$

1번문서 : 주식관련
2번문서 : 동물원 관련

co occur package

각 pair의 동시출현을 비교한 후, 확률적으로 의미있는 pair만 남긴다.(p value사용)

co occur package

```
Call:
cooccur(mat = finches, spp_names = TRUE)

Of 78 species pair combinations, 14 pairs (17.95 %) were removed from the analysis because expected co-occurrence was < 1 and 6
4 pairs were analyzed

Cooccurrence Table:
```

co occur package

Field name	Field definition
sp1	Numeric label giving the identity of species 1, assigned based on the order in the input matrix
sp2	Numeric label for species 2
sp1_inc	Number of sites (or samples) that have species 1
sp2_inc	Number of sites that have species 2
obs_cooccur	Observed number of sites having both species
prob_cooccur	Probability that both species occur at a site
exp_cooccur	Expected number of sites having both species
p_lt	Probability that the two species would co-occur at a frequency less than the observed number of co-occurrence sites if the two species were distributed randomly (independently) of one another
p_gt	Probability of co-occurrence at a frequency greater than the observed frequency
sp1_name	If species names were specified in the community data matrix this field will contain the supplied name of sp1
sp2_name	The supplied name of sp2

co occur package (정확히 j 번 나타날 확률)

$$P_j = \frac{\binom{N_1}{j} \times \binom{N-N_1}{N_2-j}}{\binom{N}{N_2}}$$

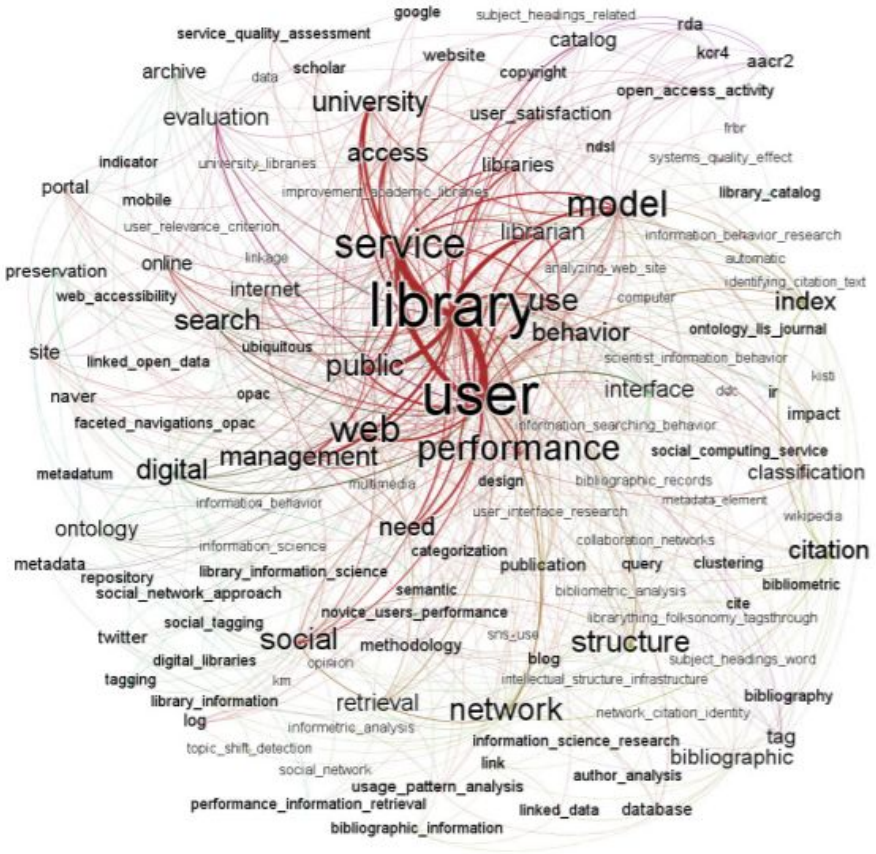
N1 = spc1이 출현한 장소(sample, document)수
N2 = spc2 ‘’’
N = 조사된 총 장소

N1 c j = N1의 장소중 J개를 고르는 경우의 수
(N - N1) c (N2-j) = N1이 출현하지 않은 곳에서 N2 - j 개의 장소를 고르는 수.
N c N2 = 전체에서 N2개를 고르는 경우의수

예제

활용 방법

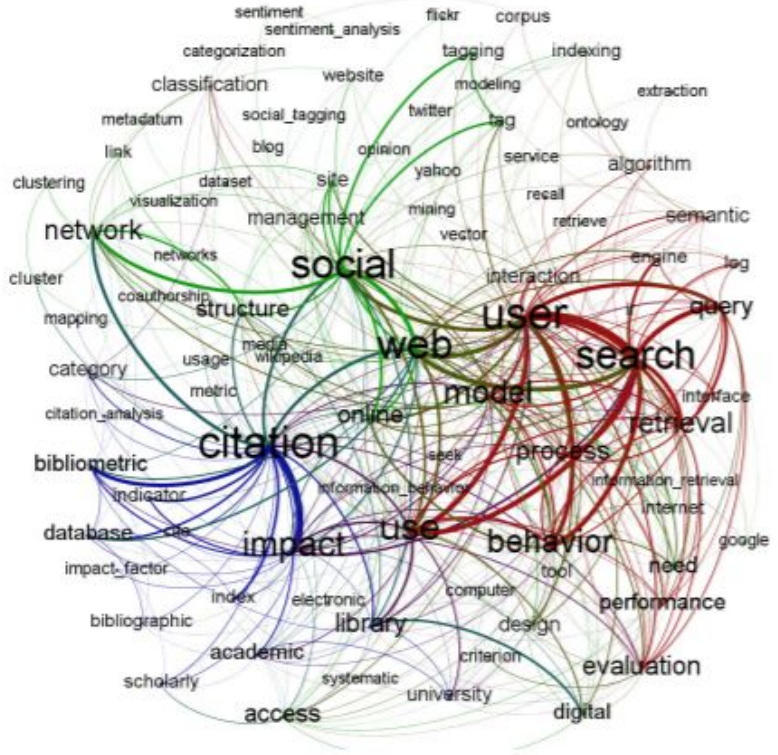
연구동향 분석
(정보관리학회지)



〈그림 1〉 정보관리학회지 2010년-2013년 (Nodes: 127 Edges: 681 Graph Density: 0.085)

활용 방법

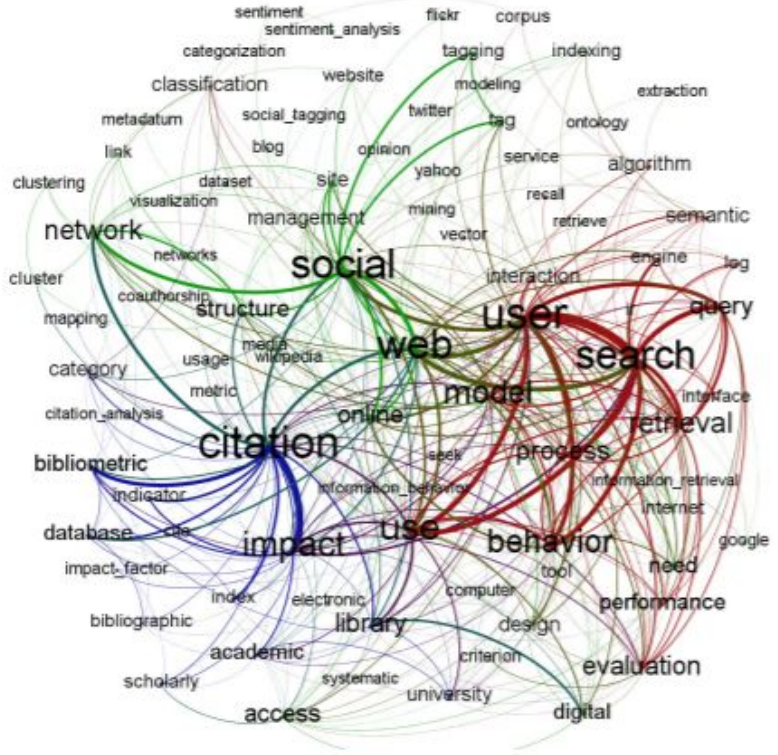
연구동향 분석(JASIST)



〈그림 2〉 JASIST 2010년-2013년 (Nodes: 89 Edges: 439 Graph Density: 0.112)

활용 방법

연구동향 분석(JASIST)



〈그림 2〉 JASIST 2010년-2013년 (Nodes: 89 Edges: 439 Graph Density: 0.112)

활용 방법

스팸 분류

예를 들어, Fig. 8.의 문장 (예: I am a student)가 spam인지 ham인지 판단하려면 6가 지 동시 단어 (예: I-am, I-a, I-student, am-a, am-student, a-student)의 각 조합들의 비율이 spam성향이 강한지 ham성향이 강한지 식 (1)을 바탕으로 계산하게 된다. 즉, 새로 고려하는 한 문장 의 단어 대 단어 조합의 경향이 spam경향이 크면 spam으로 ham경향이 크면 ham으로 판별하게 된다.

활용 방법

스팸 분류

예를 들어, Fig. 8.의 문장 (예: I am a student)가 spam인지 ham인지 판단하려면 6가 지 동시 단어 (예: I-am, I-a, I-student, am-a, am-student, a-student)의 각 조합들의 비율이 spam성향이 강한지 ham성향이 강한지 식 (1)을 바탕으로 계산하게 된다. 즉, 새로 고려하는 한 문장 의 단어 대 단어 조합의 경향이 spam경향이 크면 spam으로 ham경향이 크면 ham으로 판별하게 된다.

활용 방법

스팸 분류

예를 들어, Fig. 8.의 문장 (예: I am a student)가 spam인지 ham인지 판단하려면 6가 지 동시 단어 (예: I-am, I-a, I-student, am-a, am-student, a-student)의 각 조합들의 비율이 spam성향이 강한지 ham성향이 강한지 식 (1)을 바탕으로 계산하게 된다. 즉, 새로 고려하는 한 문장 의 단어 대 단어 조합의 경향이 spam경향이 크면 spam으로 ham경향이 크면 ham으로 판별하게 된다.

활용방법

관광학 분야

Table 2. Co-word occurrence matrix

	K1	*K2*	*K3*	*K4*	*K5*	*K6*	*K7*	*K8*	*K9*	*K10*	*K11*	*K12*	*K13*	*K14*	*K15*	...
K1	30	2	0	0	0	0	0	1	0	0	0	0	0	0	0	...
K2	0	29	0	0	0	0	0	0	0	0	0	0	0	1	0	...
K3	13	8	23	0	4	6	6	0	4	0	0	0	2	0	2	...
K4	1	5	0	12	5	1	3	2	0	0	2	0	1	0	0	...
K5	4	4	0	0	12	0	2	2	0	0	0	0	0	0	0	...
K6	7	2	0	0	2	9	2	0	1	0	0	0	0	0	0	...
K7	4	4	0	0	0	0	10	1	1	0	0	0	0	0	0	...
K8	0	2	0	0	0	0	0	5	0	0	0	0	0	0	0	...
K9	4	1	0	0	0	0	0	0	5	0	0	0	0	0	0	...
K10	0	4	0	0	0	0	0	0	0	5	0	2	0	0	0	...
K11	1	2	0	0	0	0	1	2	0	0	4	0	0	0	0	...
K12	1	4	0	0	0	0	0	0	0	0	0	4	0	0	0	...
K13	2	2	0	0	0	0	0	0	0	0	0	0	4	0	0	...
K14	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	...
K15	1	2	0	0	0	0	0	0	0	0	0	0	0	0	3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

활용방법

관광학 분야

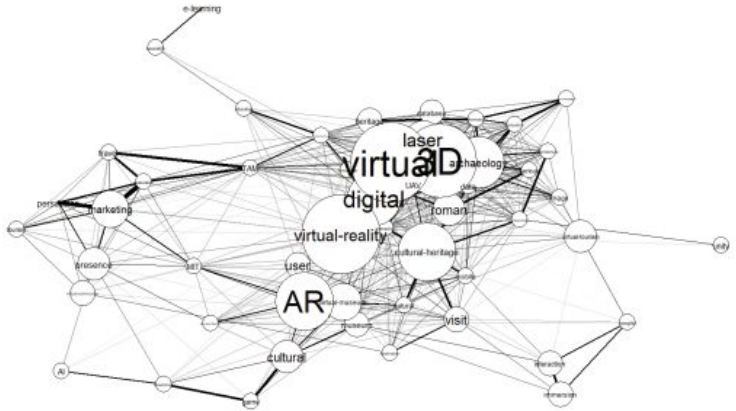


Figure 2. Network visualization of virtual reality-related keywords

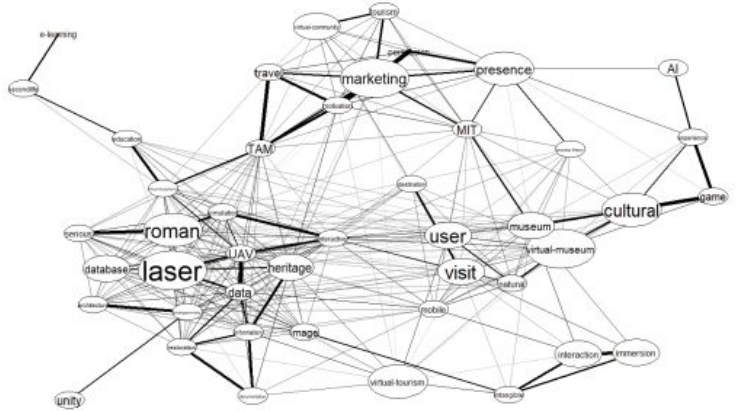


Figure 3. Network visualization of virtual reality-related keywords (excluded the most frequent keywords)

활용방법

관광학 분야

주요결과를 살펴보면, '가상의(virtual) — 3차원(3D) — 디지털(digital) — 고고학(archaeology)', '문화유산 (cultural-heritage) —방문(visit) —모바일 (mobile) —사 용자(user) —자연스러운(natural)', '증강현실(AR) —가 상박물관(virtual-museum) —박물관(museum)', '문화적 인(cultural) —게임 (game) —경험(experience) —인공지 능(AI)', '마케팅(marketing) —설득 (persuasion) —여행 (travel) —동기부여(motivation) —기술수용모형(TAM)' , '이러닝(e-learning) — 세컨드라이프(secondlife) —교 육(education)' 등의 키워드가 서로 매우 높은 연관성을 보인다는 것을 알 수 있다. 또한, 'virtual', '3D', 'digital', 'AR', 'virtual-reality', 'laser' 등의 키워드는 빈도가 상대 적으로 높을 뿐만 아니라 다른 키워드와의 연관성도 높 은 것으로 나타났다.