Jornada del Proyecto del Corpus Lingüístico UAM, U. de Tokio y TUFS

Universidad de Estudios Extranjeros de Tokio. 2016/5/12.

Reglas de relación entre rasgos lingüísticos y factores espacio-temporales

Hiroto Ueda

1. Reglas de asociación de variables

D1	A	B	C	D
i1	1	1	0	0
i2	0	0	1	0
i3	0	1	0	0
i4	0	0	1	1
i5	1	1	1	0

Support(A, B) = 2 / 5 = .400

Confidence(A, B) = P(B|A) = 2 / 2 = 1.000

Confidence(B, A) = P(B|A) = 2 / 3 = 0.667

A > B = Confidence(A, B) = P(B|A)

Lift(A, B) = Confidence(A, B) / P(B) = (2 / 2) / (3 / 5) = 1 / 0.6 = 1.667

Lift(A, B) = [c(A, B) / t(A)] / [t(B) / s] = [c(A, B) * s] / [t(A) * t(B)]
Lift(B, A) = [c(A, B) / t(B)] / [t(A) / s] = [c(A, B) * s] / [t(B) * t(A)]

Lift(A, B) = Lift(B, A) [c: coocurrencias, t: total-columna, s: suma]

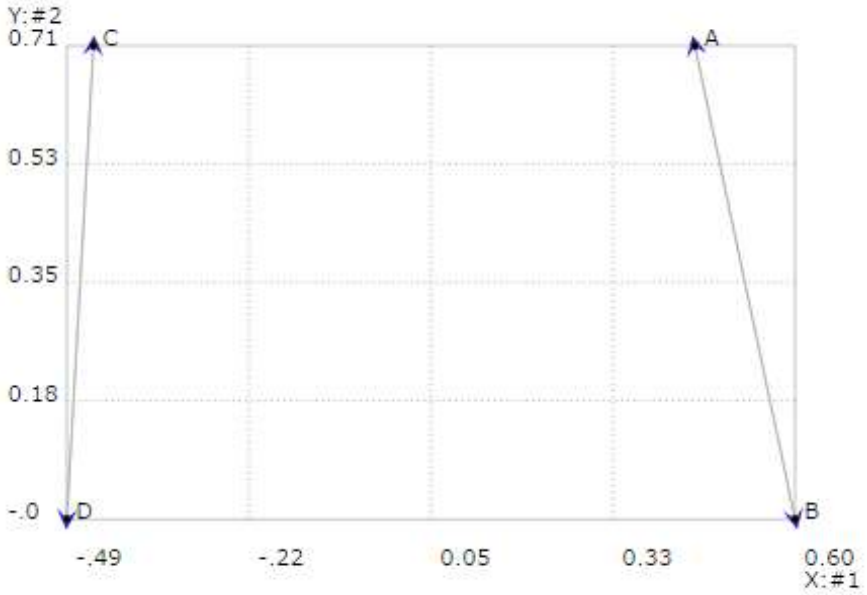
Lift(A, B) = Confidence(A, B) / P(B) = (2 / 2) / (3 / 5) = (2*5 / 2*3) = 1.667
Lift(B, A) = Confidence(B, A) / P(A) = (2 / 3) / (2 / 5) = (2*5 / 3*2) = 1.667

«Synthesis» = $(\text{Support} * \text{Lift})^{1/2}$

«Tabla de asociación» (Association table: AT):

AT	L	R	Ln	>	Rn	Ls	Rs	Lp	Rp	Co.	Sup.	Conf.	«Lift»	Synt.
1	1	2	A	>	B	2.	3.	.400	.600	2.000	.400	1.000	1.667	.816
2	2	1	B	>	A	3.	2.	.600	.400	2.000	.400	.667	1.667	.816
3	4	3	D	>	C	1.	3.	.200	.600	1.000	.200	1.000	1.667	.577
4	3	4	C	>	D	3.	1.	.600	.200	1.000	.200	.333	1.667	.577
5	1	3	A	>	C	2.	3.	.400	.600	1.000	.200	.500	.833	.408
6	3	1	C	>	A	3.	2.	.600	.400	1.000	.200	.333	.833	.408
7	2	3	B	>	C	3.	3.	.600	.600	1.000	.200	.333	.556	.333
8	3	2	C	>	B	3.	3.	.600	.600	1.000	.200	.333	.556	.333

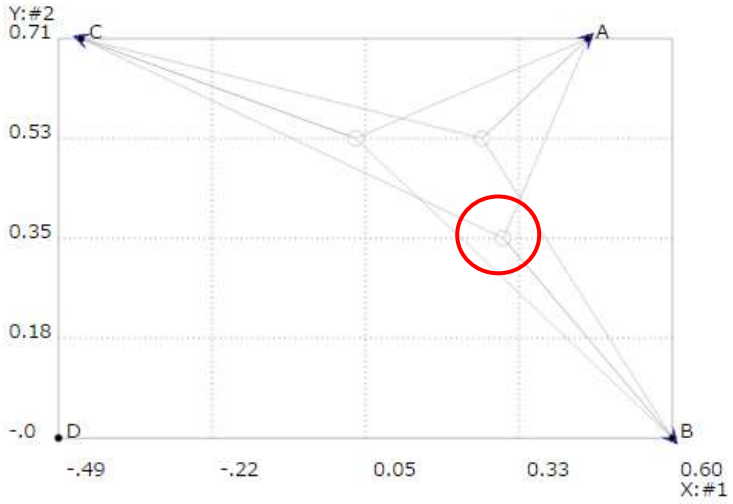
[E.vector]	#1	#2
A	.449	.707
B	.599	-.000
C	-.449	.707
D	-.489	.000



● Variable conjuntiva

encontrado los tres casos de asociación, B:C > A, A:C > B, A:B > C:

AT	L.	R.	Ln	>	Rn	Ls	Rs	Lp	Rp	Cooc.	Sup.	Cnf.	«Lift»	Synt.
1	2:3	1	B:C	>	A	1	2	.200	.400	1	.200	1.000	2.500	.707
2	1:3	2	A:C	>	B	1	3	.200	.600	1	.200	1.000	1.667	.577
3	1:2	3	A:B	>	C	2	3	.400	.600	1	.200	.500	.833	.408



● Reglas de asociación aplicadas a la matriz no negativa (variables cuantitativas)

D2	A	B	C	D	E
i1	10	19	14	7	12
i2	11	7	10	0	1
i3	0	0	1	12	1
i4	0	1	2	3	3

$$Cooc(A, B) = \sum_{[i = 1, N]} \min(A_i, B_i)$$

$$= \min(10, 19) + \min(11, 7) + \min(0, 0) + \min(0, 1)$$

$$= 10 + 7 + 0 + 0 = 17$$

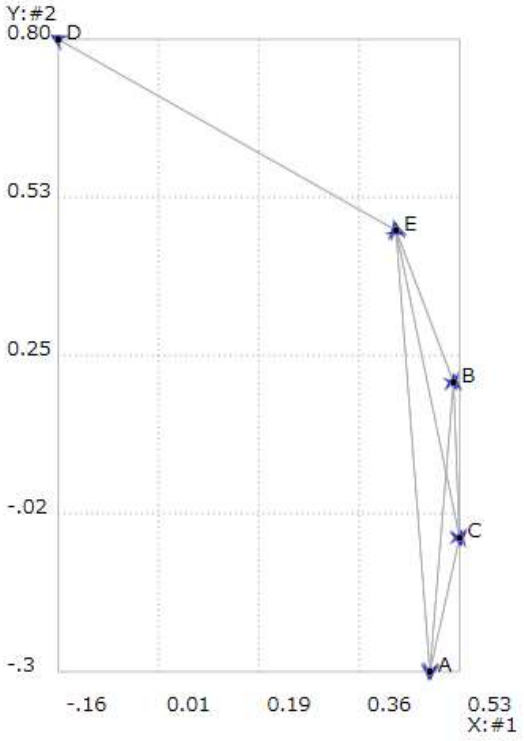
Ls = 21, Rs = 27

L.p.= P(A) = 21 / 76 = .276, R.p. = P(B) = 27 / 76 = .355

Sup = 17 / 76 = .224, Conf = 17 / 21 = .810, «Lift» = (17 / 21) / (27 / 76) = 2.279.

AT	L	R	L.n.	>	R.n.	L.s.	R.s.	L.p.	R.p.	Co.	Sup.	Conf.	«Lift»	Synt.
1	1	3	A	>	C	21.	27.	.375	.482	20.	.357	.952	1.975	.840
2	3	1	C	>	A	27.	21.	.482	.375	20.	.357	.741	1.975	.840
3	2	3	B	>	C	27.	27.	.355	.355	22.	.289	.815	2.294	.815
4	3	2	C	>	B	27.	27.	.355	.355	22.	.289	.815	2.294	.815
5	5	3	E	>	C	17.	27.	.304	.482	16.	.286	.941	1.952	.747
6	3	5	C	>	E	27.	17.	.482	.304	16.	.286	.593	1.952	.747
7	1	2	A	>	B	21.	27.	.276	.355	17.	.224	.810	2.279	.714
8	2	1	B	>	A	27.	21.	.355	.276	17.	.224	.630	2.279	.714

9	5	2	E	>	B	17.	27.	.224	.355	14.	.184	.824	2.318	.653
10	2	5	B	>	E	27.	17.	.355	.224	14.	.184	.519	2.318	.653
11	5	1	E	>	A	17.	21.	.354	.438	11.	.229	.647	1.479	.582
12	1	5	A	>	E	21.	17.	.438	.354	11.	.229	.524	1.479	.582
13	5	4	E	>	D	17.	22.	.354	.458	11.	.229	.647	1.412	.569
14	4	5	D	>	E	22.	17.	.458	.354	11.	.229	.500	1.412	.569



2. Reglas de relación

D1	A	B	C	D	smH
i1	1	1	0	0	2
i2	0	0	1	0	1
i3	0	1	0	0	1
i4	0	0	1	1	2
i5	1	1	1	0	3
smV	2	3	3	1	9

«Support» = Xnp / smA

«Confidence» = Xnp / smH

«Lift» = «Confidence» / R.p. = (Xnp / smH) / [smV / smA]
= (1 / 2) / (2 / 9) = .500 / .222 = 2.250

.

D1	A	B	C	D	smH
i1	1	1	0	0	2
i2	0	0	1	0	1
i3	0	1	0	0	1
i4	0	0	1	1	2
i5	1	1	1	0	3
smV	2	3	3	1	9

«Synthesis» = («Support» * «Lift»)$^{1/2}$

Por lo tanto;

= [(Xnp / smA) * (Xnp / smH) / (smV / smA)]$^{1/2}$ ←Definición
= [(Xnp / smA) * (Xnp / smH) * (smA / smV)]$^{1/2}$
←Arreglar el numerador
= [(Xnp) * (Xnp / smH) / smV]$^{1/2}$ ←borrar smA
= [Xnp / (smH * smV)]$^{1/2}$ ←Arreglar el numerador

El «Synthesis» de [i1 > A] es 1 * [(2 * 2)]$^{1/2}$ = .500.

RT	L	R	Ln	>	Rn	Ls	Rs	Lp	Rp	Co.	Sup.	Conf.	«Lift»	Synt.
1	4	4	i.4	>	D	2.	1.	.222	.111	1.	.111	.500	4.500	.707
2	2	3	i.2	>	C	1.	3.	.111	.333	1.	.111	1.	3.	.577
3	3	2	i.3	>	B	1.	3.	.111	.333	1.	.111	1.	3.	.577
#4	1	1	i.1	>	A	2.	2.	.222	.222	1.	.111	.500	2.250	.500
5	1	2	i.1	>	B	2.	3.	.222	.333	1.	.111	.500	1.500	.408
6	4	3	i.4	>	C	2.	3.	.222	.333	1.	.111	.500	1.500	.408
7	5	1	i.5	>	A	3.	2.	.333	.222	1.	.111	.333	1.500	.408
8	5	2	i.5	>	B	3.	3.	.333	.333	1.	.111	.333	1.	.333
9	5	3	i.5	>	C	3.	3.	.333	.333	1.	.111	.333	1.	.333

Matriz no negativa:

D2	A	B	C	D	E	smH
i1	10	19	14	7	12	62
i2	11	7	10	0	1	29
i3	0	0	1	12	1	14
i4	0	1	2	3	3	9
smV	21	27	27	22	17	114

RT	L	R	Ln	>	Rn	Ls	Rs	Lp	Rp	Co.	Sup.	Conf.	«Lift»	Synt.
#1	3	4	i.3	>	D	14.	22.	.123	.193	12.	.105	.857	4.442	.684
#2	1	2	i.1	>	B	62.	27.	.544	.237	19.	.167	.306	1.294	.464
3	2	1	i.2	>	A	29.	21.	.254	.184	11.	.096	.379	2.059	.446
4	1	5	i.1	>	E	62.	17.	.544	.149	12.	.105	.194	1.298	.370
5	2	3	i.2	>	C	29.	27.	.254	.237	10.	.088	.345	1.456	.357
6	1	3	i.1	>	C	62.	27.	.544	.237	14.	.123	.226	.953	.342
7	1	1	i.1	>	A	62.	21.	.544	.184	10.	.088	.161	.876	.277
8	2	2	i.2	>	B	29.	27.	.254	.237	7.	.061	.241	1.019	.250
9	4	5	i.4	>	E	9.	17.	.079	.149	3.	.026	.333	2.235	.243
10	4	4	i.4	>	D	9.	22.	.079	.193	3.	.026	.333	1.727	.213
11	1	4	i.1	>	D	62.	22.	.544	.193	7.	.061	.113	.585	.190
12	4	3	i.4	>	C	9.	27.	.079	.237	2.	.018	.222	.938	.128
13	3	5	i.3	>	E	14.	17.	.123	.149	1.	.009	.071	.479	.065
14	4	2	i.4	>	B	9.	27.	.079	.237	1.	.009	.111	.469	.064
15	3	3	i.3	>	C	14.	27.	.123	.237	1.	.009	.071	.302	.051
16	2	5	i.2	>	E	29.	17.	.254	.149	1.	.009	.034	.231	.045

3. 1. Ejemplo de aplicación

Variación léxica en el español de América

Cahuzac (1980)[1]:

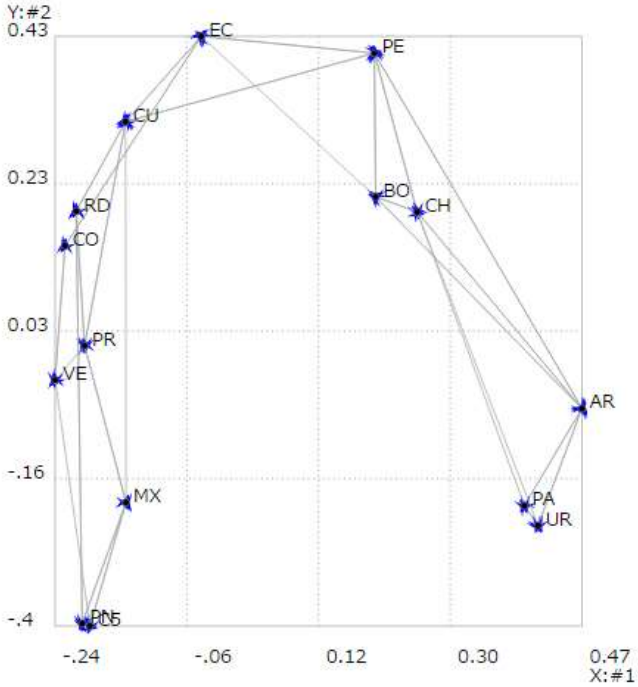
Cahuzac (1980)	MX	CU	RD	PR	C5	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
01 cacahuero	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
02 cafetalista	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
03 camilucho	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
04 campero	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

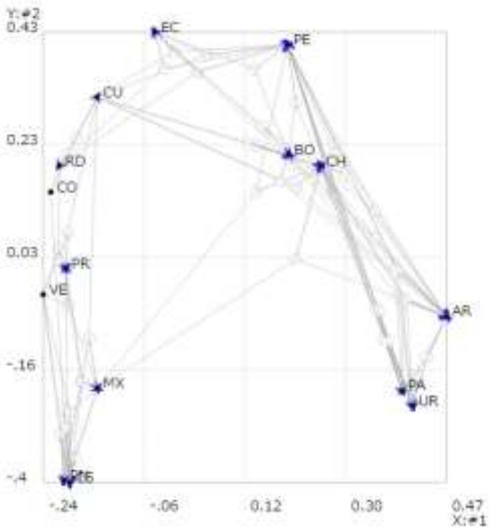
Reglas de asociación entre los países hispanohablantes:

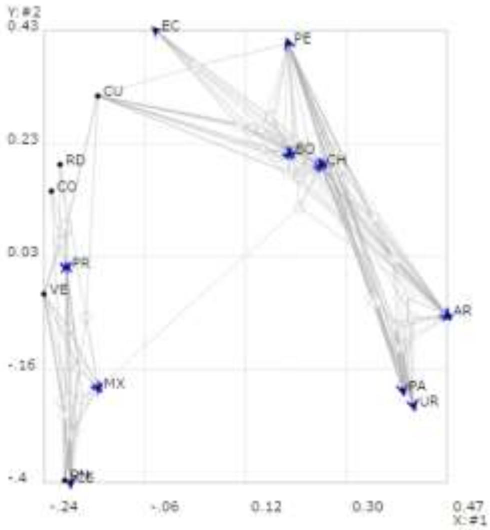
AT	L.	R.	Ln	>	Rn	Ls	Rs	Lp	Rp	Cooc.	Sup.	Cnf.	«Lift»	Synt.
1	13	14	PA	>	UR	10	12	.213	.255	10	.213	1.000	3.917	.941
2	14	15	UR	>	AR	12	16	.255	.340	12	.255	1.000	2.938	.909
3	14	13	UR	>	PA	12	10	.255	.213	10	.213	.833	3.917	.886

[1] Cahuzac, Philippe. (1980) "La División del español de América en zonas dialectales: Solución etnolingüística o semántico-dialectal." *Lingüística Española Actual*, 10, pp. 385-461. Los cinco países centroamericanos (GU, HO, EL, NI, CR) no presentan variaciones entre sí, de modo que los agrupamos en nombre de C5. Las filas destacadas, 24 *guaso* y 36 *montubio*, serán mencionadas posteriormente.

4	13	15	PA	>	AR	10	16	.213	.340	10	.213	1.000	2.938	.855
5	7	8	VE	>	CO	9	11	.191	.234	8	.170	.889	3.798	.831
6	15	14	AR	>	UR	16	12	.340	.255	12	.255	.750	2.938	.825
7	12	10	CH	>	PE	5	10	.106	.213	5	.106	1.000	4.700	.794
8	8	7	CO	>	VE	11	9	.234	.191	8	.170	.727	3.798	.778
9	5	6	C5	>	PN	7	10	.149	.213	6	.128	.857	4.029	.761
…														







Association: Variable conjuntiva doble / Association: Variable conjuntiva triple

Referencia:

Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. http://www.almaden.ibm.com/cs/quest/papers/sigmod93.pdf [2016/4/25]

尾崎隆『ビジネスに活かすデータマイニング』技術評論社(2014) pp.163-184.

豊田秀樹『データマイニング入門』東京書籍(2008) pp.149-183.

REDES-web:

http://lecture.ecc.u-tokyo.ac.jp/~cueda/numeros/index-j.php