

Associative Lexical Nets

Manfred Wetzler, Reginald Ferber,
Reinhard Rapp, and Bernd Hagen

*Fachbereich 2, Universität –GH– Paderborn,
D-33095 Paderborn, Germany
ferber@psycho2.uni-paderborn.de*

1. Theoretical Framework

Law of Association

"Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also"

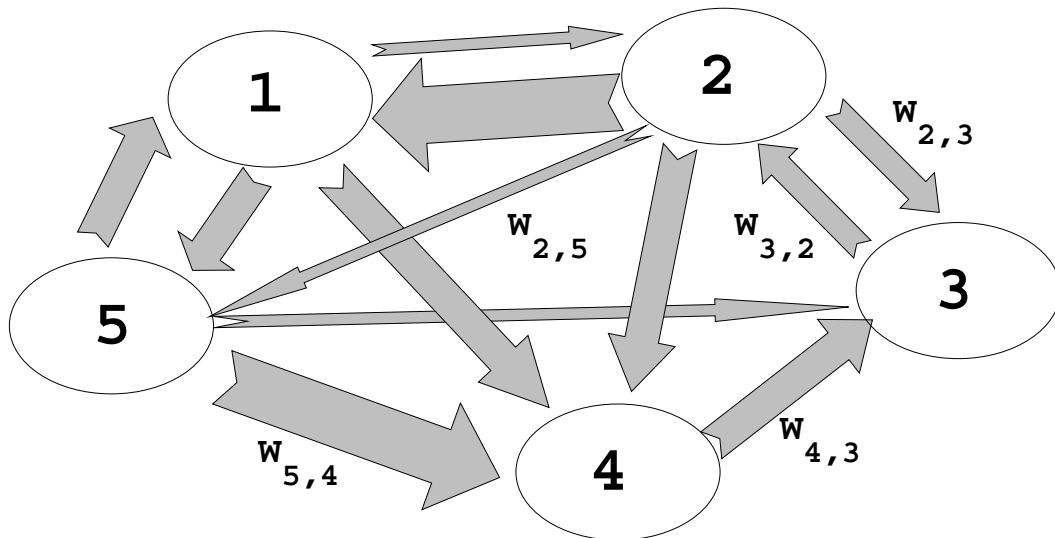
(James, 1890, Vol. 1, p. 561).

Empirical Test

- objects → words
- associations → real numbers
- experienced together → co-occurrence in texts

Associative Lexical Net

- A set of *nodes* representing words denoted by $I = \{1, \dots, n\}$ bearing activities $c(i)$ denoted by a mapping $c : I \rightarrow \mathbb{R}$.
- A set $\{w_{i,j} \in \mathbb{R} \mid i, j \in I\}$ of real valued *weights* representing the associations between the nodes. They can be organized in an $n \times n$ Matrix G .



Operations on a Lexical Net

- **activation of a word i :** set $c(i) = 1$.
- **associative process:** spread the activation in the net by computing new activations for all nodes with the formula

$$c'(i) = \sum_{j=1}^n w_{j,i} c(j) \quad (1)$$

This corresponds to a linear function on \mathbb{R}^n

- **evaluate an associative process:** look for the nodes with the highest activities under c' .

Learning Associations

- $(i \& j)$: word i and j occur together \rightarrow increase association $a_{i,j}$.
- $(i \neg j)$: word i occurs without word j \rightarrow decrease $a_{i,j}$.

learning rule:

$$a_{i,j}(t+1) = \begin{cases} a_{i,j}(t) + \alpha(1 - a_{i,j}(t)) & \text{if } (i \& j) \\ (1 - \alpha)a_{i,j}(t) & \text{if } (i \neg j) \end{cases}$$

with $a_{i,j}(t)$, $\alpha \in [0, 1]$

The following holds:

$$a_{i,j}(t) \xrightarrow[t \rightarrow \infty]{} p(j | i) = \frac{p(i \& j)}{p(i)} =: a_{i,j} \quad (3)$$

Other rules are

$$a_{i,j} = \frac{p(i \& j)}{p(i)} \cdot \frac{1}{p(j)} = \frac{p(i \& j)}{p(i) \cdot p(j)}$$

$$a_{i,j} = \frac{p(i \& j)}{p(i)} - p(j)$$

$$a_{i,j} = \frac{p(i \& j)}{p(i) \cdot p(j)} - 1$$

implemented with frequencies

$$w_{i,j} = \frac{\#(i \& j)}{\#(i)}$$

$$w_{i,j} = \frac{\#(i \& j) \cdot q}{\#(i) \cdot \#(j)}$$

$$w_{i,j} = \frac{\#(i \& j)}{\#(i)} - \frac{\#(j)}{q}$$

$$w_{i,j} = \frac{\#(i \& j) \cdot q}{\#(i) \cdot \#(j)} - 1$$

where $\#(i)$ denotes the number of occurrences of the event i and q is the total number of observations.

2. Prediction of Free Word Associations

The Russell & Meseck Association Norms (1959)

- Free associative responses to 100 German stimulus words collected from 331 German subjects in 1957 and 1958.
- *association norms*: a list of all responses and their frequencies for each stimulus word.
- the most frequent response to a word is called the *primary* response.

The Lexical Net for the Simulation

- **corpora**: some 21 million words including newspaper articles and abstracts of psychological articles.
- **node set (vocabulary)**: All 2 012 words from the Russell & Meseck experiment and further 63 344 words occurring at least 10 times in the corpora; altogether 65 356 words.
- **co-occurrence**: occurrence within a distance of 12 words.
- **formula for the weights**:

$$a_{i,j} = \frac{p(i\&j)}{p(i)} * \frac{1}{p(j)^{0.68}} = \frac{1}{p(i)} * \frac{p(i\&j)}{p(j)^{0.68}}$$

implemented as

$$w_{i,j} = \begin{cases} \frac{\#(i\&j)}{\#(j)^{0.68}} & \text{if } \#(j) > 110 \\ \frac{\#(i\&j)}{110} & \text{if } \#(j) \leq 110 \end{cases}$$

- **simulation**: The node of the stimulus word is set to 1, all other nodes to 0 and the activation is spread for one cycle.
- **result**: node with highest activation after the cycle of spreading ($\Leftrightarrow j$ with $w_{i,j} \geq w_{i,k} \forall k \in I$) is the *predicted* response.

Results

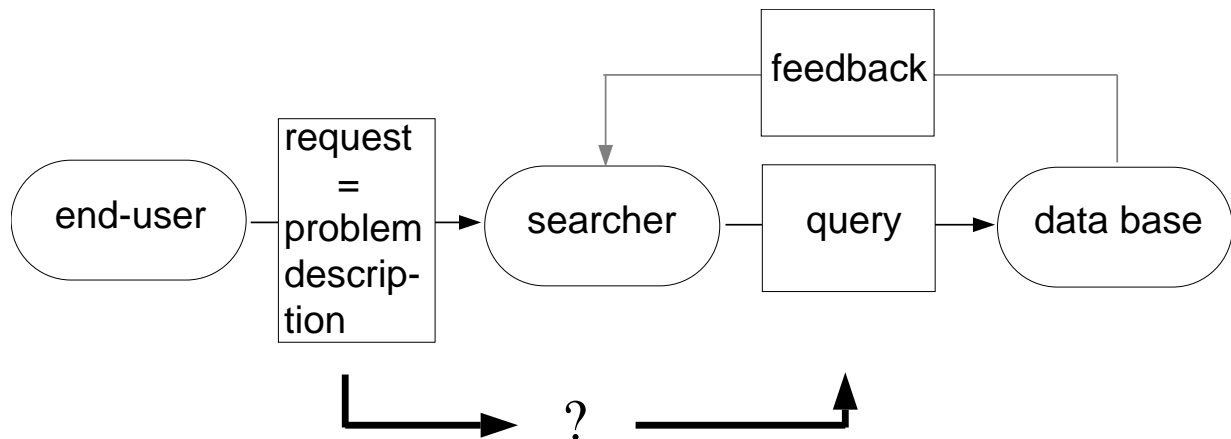
Experiment		Simulation	
subjects response is primary response	22.5	predicted response is primary respons	19
other subjects giving the same response	28.8	subjects giving predicted response	22.9
subjects response was given by no other subject	19.8	predicted response was not given by any subject	57

From the 57 predicted responses that were not given by any subject

- 24 are syntagmatic associations like “Tisch — legen” or “Stuhl — elektrischen”.
- 7 are proper names which were not popular at the time of the experiment (1957–1958) like “Kohl — Helmut” or in the population of subjects like “Adler — Alfred”.
- 3 are new terms like “Baby — Pillen”.
- 4 are specific to psychological literature like “Quadrat — chi”.
- 2 are non-words included in the corpora
- the number of 17 remaining responses given by no subject is close to the number of 19.8 for subject responses.

3. Query Generation in Document Retrieval

Search Process



Selection of terms for the query

- from the written problem description of the end user
- new terms

Problem description

Veränderungen im Selbstkonzept von Partnern nach einer klientenzentrierten Partnertherapie. Changes in the self-concept of partners following a client-centered partner therapy.

Key-words:

self-concept, client-centred therapy, partner-therapy

Database search listing:

```

1.    26 FIND GERM PAARTHERAPIE/PQ
2.   539 FIND CT=SELF CONCEPT
3.     0 FIND 1 AND 2
4.  2314 FIND SELF$/PQ
5.     3 FIND 1 AND 4
6.   152 FIND GERM PARTNER$/PQ
7.   581 FIND GERM EHE$/PQ
8.    82 FIND 4 AND (6 OR 7)
9.    82 FIND 5 OR 8
10.   22 FIND 2 AND (6 OR 7)
11.   25 FIND 5 OR 10
  
```

Example 1: The German and English problem description of an end-user and the corresponding query of a professional searcher.

Material

- 94 search records made independently from this study at the “Zentralstelle für psychologische Information und Dokumentation” at the University of Trier. A record consists of an end-user’s written request in German and English and the corresponding searches in the databases PSYCINFO or PSYINDEX by a professional searcher.
- 246 889 Titles and abstracts of psychological articles from the CD-ROM database PsycLIT.

The Lexical Net

Node set (vocabulary) 2108 words from the records were grouped into 872 *terms* each consisting of

- German words
- their English translations
- different morphological variants and shortened forms with identical roots

occurring in more than 40 documents of PsycLIT.

A EIN EINE EINEM EINEN EINER EINES
AFTER DANACH DARAUf NACH
BEGRIFFS CONCEPT CONCEPTIONS CONCEPTS
CONCEPTUALIZATION KONZEPTS
CENTER CENTERED CENTERS CENTRE ZENTRIEREN
CLIENT CLIENTS KLIENTEN
PARTNER PARTERN PARTNERS PARTNERSCHAFT
PARTNERSCHAFTLICHE
SELBST SELF

Table 1: Some of the 16 terms occurring in example 1

Co-occurrence occurrence of words of the two terms in the same document of PsycLIT (title, abstract or key points).

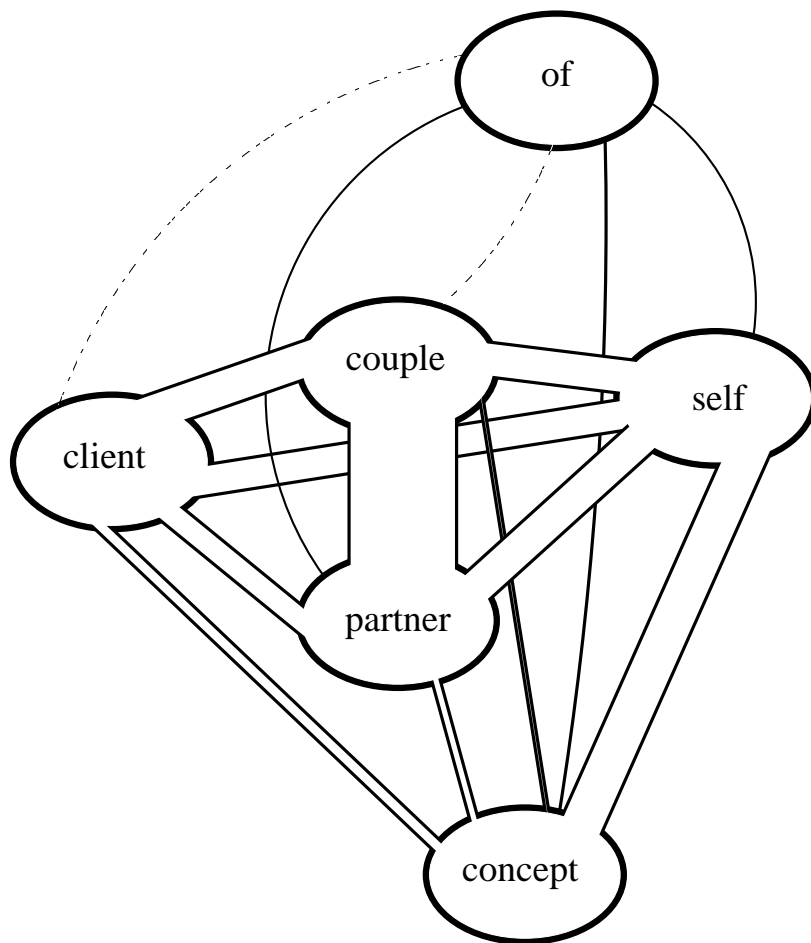
Formula for the weights

$$a_{i,j} = \frac{p(i\&j)}{p(i) \cdot p(j)} - 1$$

implemented as

$$w_{i,j} = \frac{\#(i\&j) \cdot 246\,889}{\#(i) \cdot \#(j)} - 1$$

- nonlinear monotonic transformation onto $[-0.1, 1]$
- division by the spectral radius of the resulting matrix.

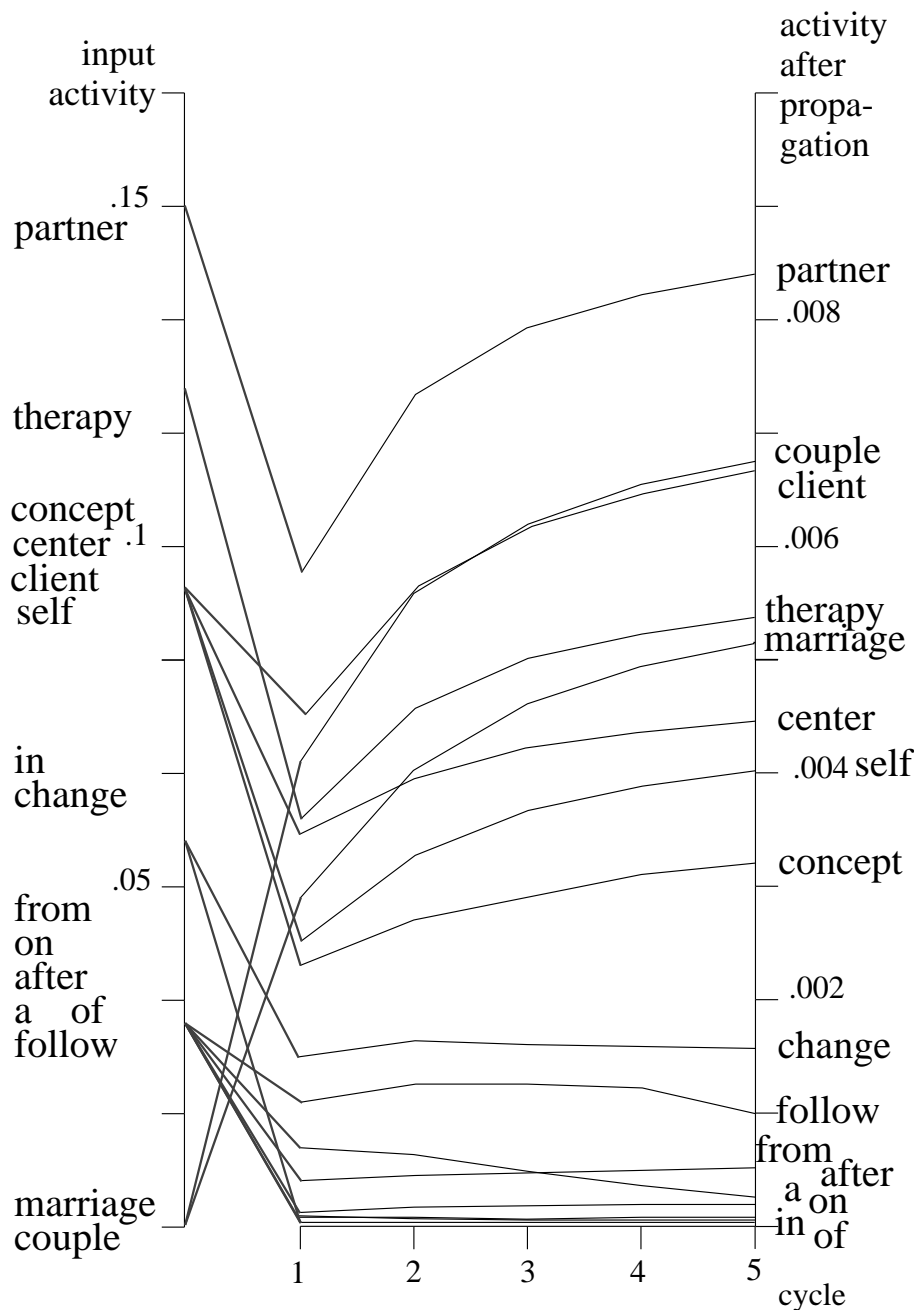


	concept	couple	client	partner	self
of	25	-1	-11	8	7
concept		32	446	77	4173
couple			3244	26020	2377
client				1388	2208
partner					2863

*The associations between some of the terms in example 1.
Values in the table are the associations multiplied by 10^5 .*

Simulation of a record

- activation of the terms occurring in the end users problem description
- spreading of activation
- calculation of the activity ranks of the terms of the record
- addition of activation to the terms of the problem description
- spreading of activation
- ...



The evolution of the activity of the terms in example 1.

Evaluation of a simulation

The terms of each record were grouped into four classes:

- *P&Q-terms* appearing in the problem description and in the query.
- *P&¬Q-terms* appearing in the problem description but not in the query.
- *¬P&Q-terms* not appearing in the problem description but in the query
- *¬P&¬Q-terms*: all remaining terms appearing neither in the problem description nor in the query

P&Q-terms: SELF, CONCEPT, PARTNER, THERAPY

P&¬Q-terms: AFTER, FROM, CHANGE, IN, OF, ON,
FOLLOW, A, CLIENT, CENTER

¬P&Q-terms: COUPLE, MARRIAGE.

¬P&¬Q-terms: All other terms

Table 2: Classification of the 16 terms from example 1 into four classes for evaluation.

Mean activity ranks are calculated for the classes.

Good results are achieved if

	activity	rank
P&Q-terms	high	low
P&¬Q-terms	low	high
¬P&Q-terms	high	low
¬P&¬Q-terms	low	high

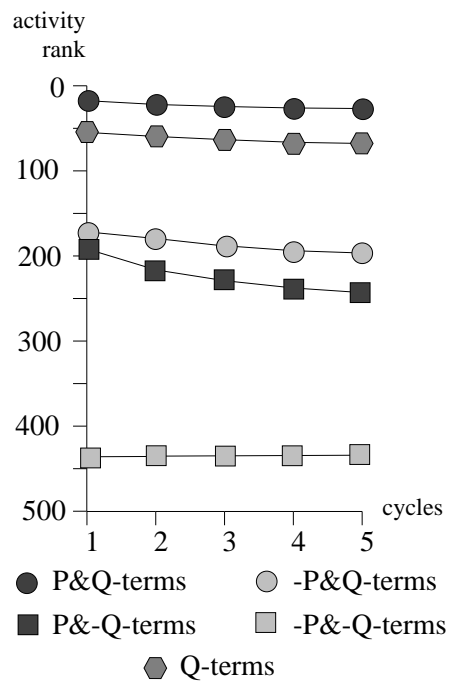
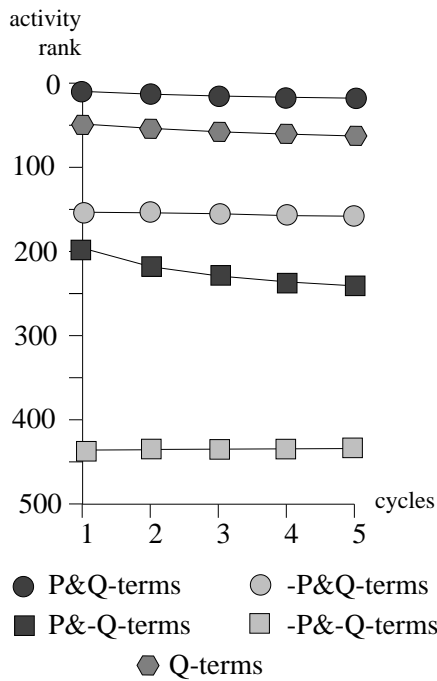
RANK	CLASS	ACTIVITY	TERM
1.	P&Q	0.007922	PARTNER PARTNERN...
2.	-P&Q	0.006178	COUPLES PAAR COUPLE
3.	P&-Q	0.006142	CLIENT KLIENTEN CLIENTS
4.	- -	0.005975	DISCLOSURE
5.	P&Q	0.005	THERAP THERAPEUTEN...
6.	- -	0.004995	INTERCOURSE INTERCOURSES
7.	- -	0.004981	ORGASM
8.	- -	0.004967	SPOUSES SPOUSE
9.	- -	0.00492	MASTURBATION
10.	- -	0.004902	BIBLIOTHERAPY
11.	- -	0.004609	IMPOTENCE
12.	-P&Q	0.004603	EHE MARRIAGE MARITAL
13.	- -	0.00456	LIEBES LOVE LOVING
14.	- -	0.004169	EMPATH EMPATHIE EMPATHY
15.	P&-Q	0.004152	CENTER CENTERED...
16.	- -	0.004102	ESTEEM
17.	- -	0.004077	COUNSELING COUNSELOR
18.	- -	0.004035	RESTRUCTURING...
19.	- -	0.003994	PSYCHOTH PSYCHOTHERAP...
20.	- -	0.003772	PSYCHOSEXUAL
22.	P&Q	0.003667	SELBST SELF
50.	P&Q	0.002974	CONCEPT CONCEPTIONS...
145.	P&-Q	0.001589	CHANGE CHANGING...
201.	P&-Q	0.001246	FOLLOW FOLLOWING...
391.	P&-Q	0.000489	DANACH AFTER NACH
408.	P&-Q	0.000446	VON FROM VOM
499.	P&-Q	0.000175	EIN EINE EINEM A ONE
528.	P&-Q	0.000071	AUF ON
534.	P&-Q	0.000058	IN INS IM
549.	P&-Q	0.000007	OF AUS

Table 3: Ranks and activities of the terms in example 1 in the third cycle of the simulation.

The mean ranks of the four classes are: P&Q-terms, 16.1; \neg P&Q-terms, 7.0; P& \neg Q-terms, 249.3; and \neg P& \neg Q-terms: 440.7. Overlap is 0.6.

Overall Results

- 47 calibration records to estimate parameters
- 47 test records to compare the results with those of the calibration sample.



Calibration sample:

Cy	P&Q	-P&Q	P&-Q
1:	18.5	155.6	194.8
2:	24.5	156.0	216.3
3:	29.1	158.6	227.8
4:	32.2	161.2	235.0
5:	34.4	162.9	239.5

Test sample:

Cy	P&Q	-P&Q	P&-Q
1:	18.9	172.3	203.9
2:	22.9	180.4	227.9
3:	25.7	188.7	241.3
4:	27.7	194.7	249.9
5:	29.0	198.2	255.5

References

Ferber, R. Vorhersage der Suchwortwahl von professionellen Recherchereuren in Literaturdatenbanken durch assoziative Wortnetze. In *Mensch und Maschine – Informationelle Schnittstellen der Kommunikation. (Proceedings ISI '92)* (1992), H. H. Zimmermann, H.-D. Luckhardt, and A. Schulz, Eds., Universitätsverlag Konstanz, pp. 208–218.

James, W. *The Principles of Psychology*. New York: Holt, 1890. Reprinted New York: Dover Publications, 1950.