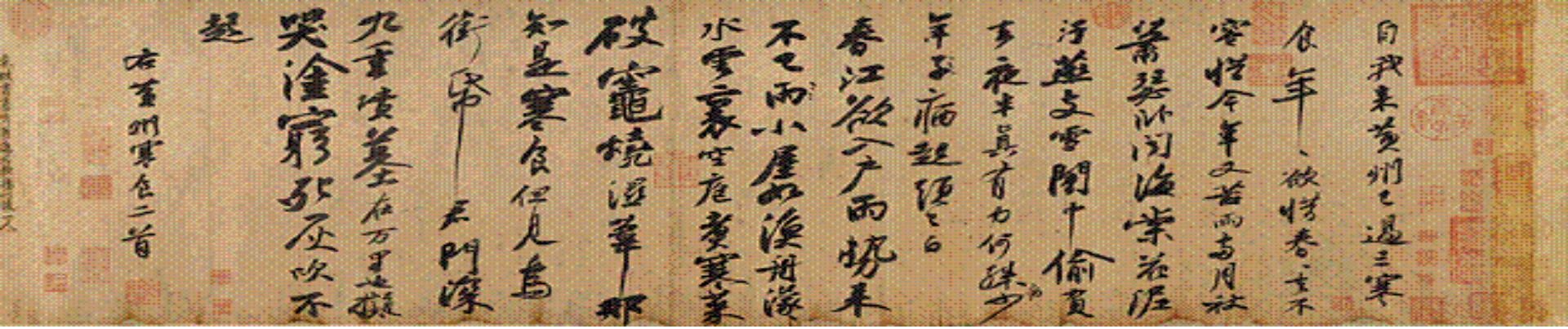

信息检索
Information Retrieval

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：**胡锦毅**

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括放到9#服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

第三章 文本分析及自动标引 (Part 5)

3.5 Thesaurus及term自动关联

Construction of term phrases

- precision – specificity
- phrase generation methods
  statistical methods

$$MI_{kh} = \log_2 \frac{PAIR\_P_{kh}}{TTP_k \times TTP_h} = \log_2 \frac{N \times PAIR\_F_{kh}}{TTF_k \times TTF_h}$$

word 1	*word 2*	*count word 1*	*count word 2*	count of co-occurrences	PMI
puerto	rico	1938	1311	1159	10.0349081703
hong	kong	2438	2694	2205	9.72831972408
los	angeles	3501	2808	2791	9.56067615065
carbon	dioxide	4265	1353	1032	9.09852946116
prize	laureate	5131	1676	1210	8.85870710982
san	francisco	5237	2477	1779	8.83305176711
nobel	prize	4098	5131	2498	8.68948811416
ice	hockey	5607	3002	1933	8.6555759741
star	trek	8264	1594	1489	8.63974676575
car	driver	5578	2749	1384	8.41470768304
it	the	283891	3293296	3347	-1.72037278119
are	of	234458	1761436	1019	-2.09254205335
this	the	199882	3293296	1211	-2.38612756961
is	of	565679	1761436	1562	-2.54614706831
and	of	1375396	1761436	2949	-2.79911817902
a	and	984442	1375396	1457	-2.92239510038
in	and	1187652	1375396	1537	-3.05660070757
to	and	1025659	1375396	1286	-3.08825363041
to	in	1025659	1187652	1066	-3.12911348956
of	and	1761436	1375396	1190	-3.70663100173

3.5 Thesaurus及term自动关联

- Restriction in using term phrases

  only for relatively broad, high-frequency words
  rare terms ➔ over-specific

3.5 Thesaurus及term自动关联

- Automatic indexing process capable of producing high-performance retrieval results:

(1) Terms in the medium-frequency ranges with positive discrimination values are used as index terms directly without further transformation.
(2) The broad high-frequency terms with negative discrimination values are either discarded or incorporated into phrases with low-frequency characteristics.
(3) The narrow low-frequency terms with discrimination values close to zero are broadened by inclusion into thesaurus categories.

3.5 Thesaurus及term自动关联

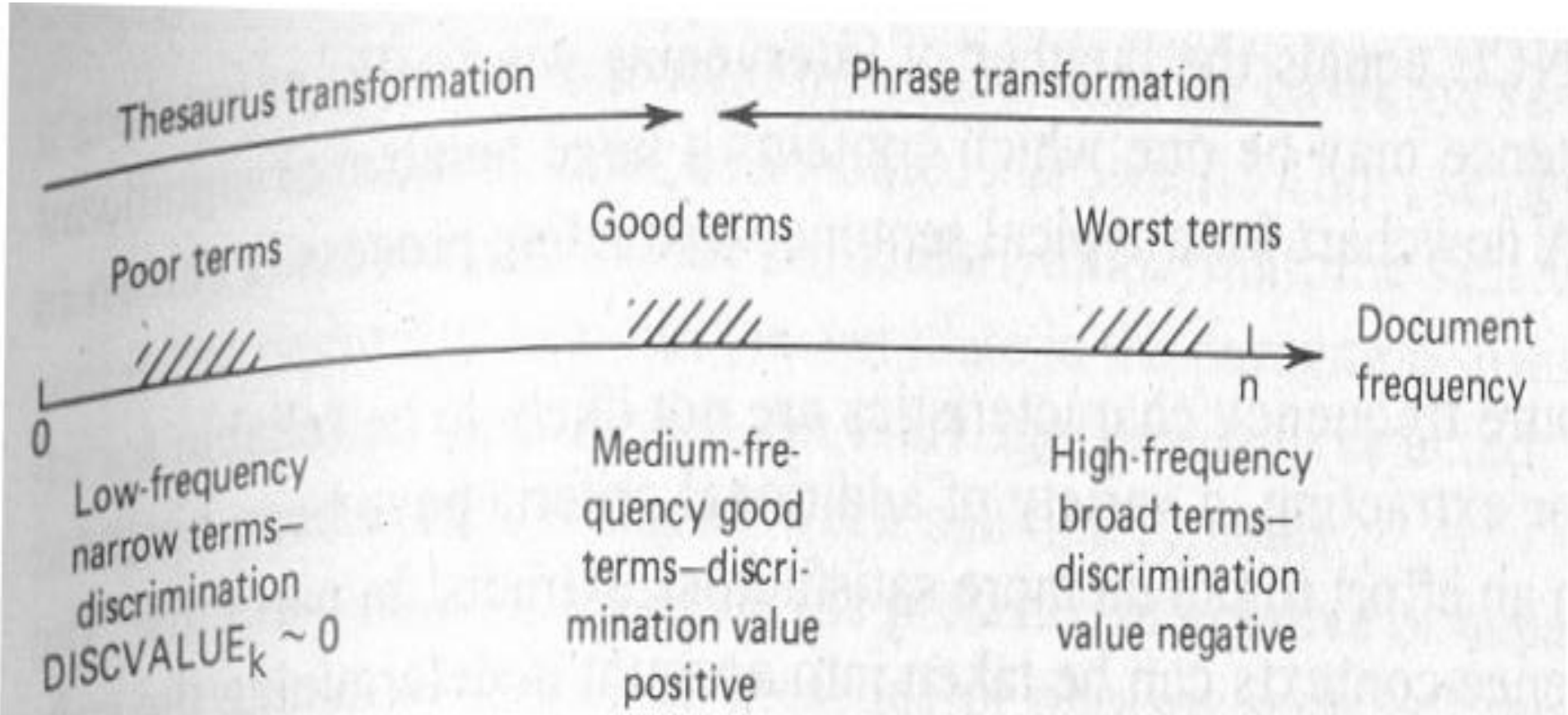

Figure 3-7 Term characterization in frequency spectrum.

3.5 Thesaurus及term自动关联

利用Automatic term association的其他途径

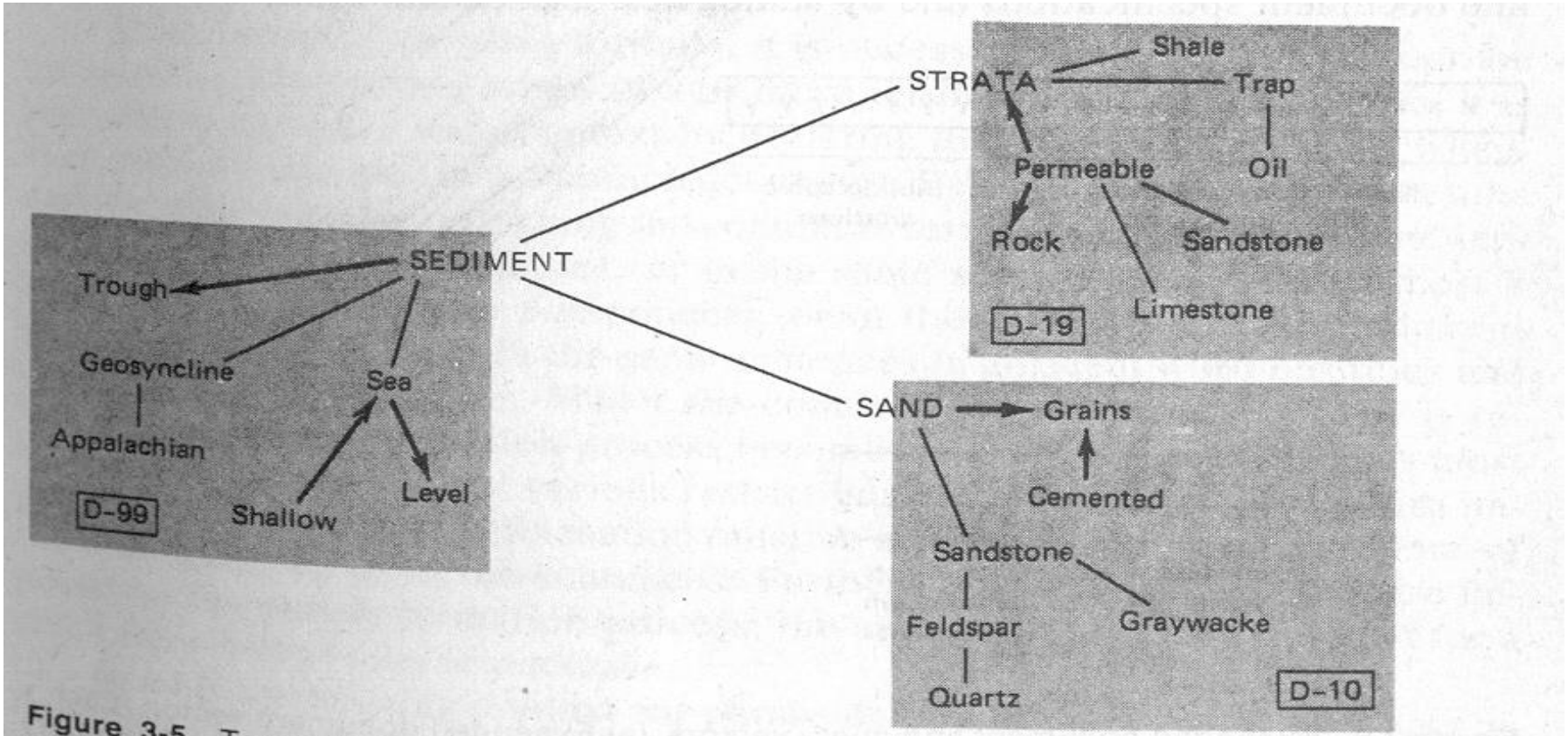

Figure 3-5 Term association map. (Different types of connecting lines denote different strengths of associations between terms). (*Adapted from reference 35.*)

3.5 Thesaurus及term自动关联

	A	B	C	D	E
A	1	1	0	0	0
B	1	1	0	1	0
C	0	0	1	0	1
D	0	1	0	1	1
E	0	0	1	1	1

(a)

Original term	Associated terms
A	B
B	A, D
C	E
D	B, E
E	C, D

(b)

$$q = \begin{pmatrix} A = 4 \\ B = 2 \\ C = 1 \\ D = 1 \\ E = 0 \end{pmatrix} \quad \begin{matrix} \text{add } B = 2 \\ \text{add } A = 1, D = 1 \\ \text{add } E = \tfrac{1}{2} \\ \text{add } B = \tfrac{1}{2}, E = \tfrac{1}{2} \\ \text{add nothing} \end{matrix} \quad q' = \begin{pmatrix} A = 5 \\ B = 4\tfrac{1}{2} \\ C = 1 \\ D = 2 \\ E = 1 \end{pmatrix}$$

(c)

3.5 Thesaurus及term自动关联

Term ambiguity may introduce irrelevant statistically correlated terms.

"Apple computer" → "Apple red fruit computer"

Only expand query with terms that are similar to *all* terms in the query.

"fruit" not added to "Apple computer" since it is far from "computer."

"fruit" added to "apple pie" since "fruit" close to both "apple" and "pie."