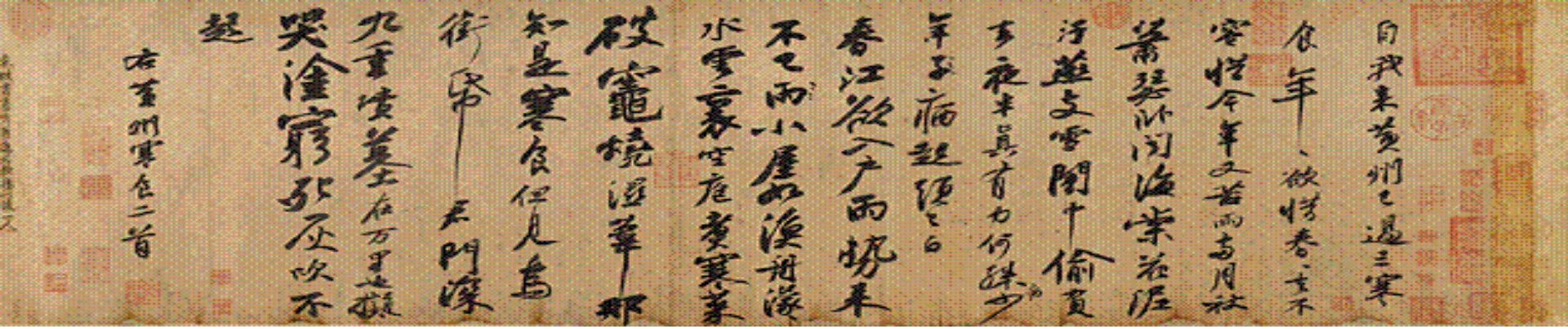


信息检索
Information Retrieval

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》（课号：40240372）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。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。

第四章 基于潜在语义分析(LSI)的信息检索模型

Why Latent Semantic Indexing (LSI)

The goal of LSI:

Eliminate redundant axes

Dimension reduction

Pull together "related" axes – hopefully

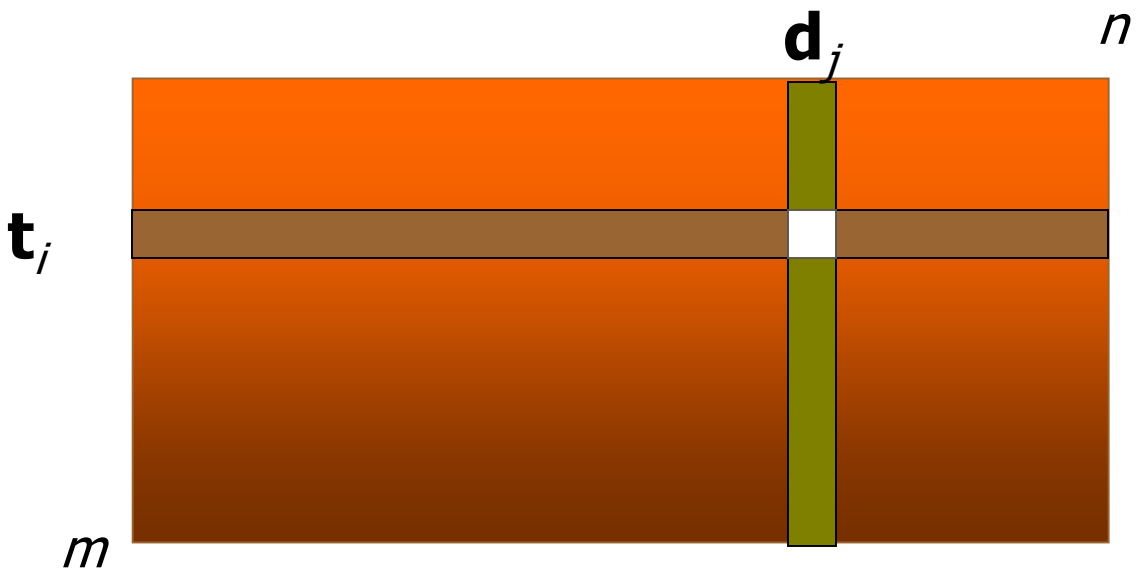
vocabulary mismatch: poor recall

car and ***automobile***

LSI is data-*dependent*

Term-Document Matrix

$w_{i,j}$ = (normalized) weighted count (t_i, $\mathbf{d}_j$)



Matrix Factorizations: SVD

定理： 设 $A \in C_r^{m\times n}\ (r>0)$,则有 $A^T A$ 的特征值：

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$$

并称 $\sigma_i = \sqrt{\lambda_i}\ (i = 1,2,\ldots,n)$ 为 A 的奇异值。

定理： 设 $A \in C_r^{m\times n}\ (r>0)$,则存在 m 阶酉矩阵 U 和 n 阶酉矩阵 V，使得

$$U^T A V = \begin{bmatrix} \Sigma & O \\ O & O \end{bmatrix}$$

其中 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, 而 $\sigma_i\ (i = 1,2,\ldots,r)$ 为 A 的非零奇异值。

A 的奇异值分解： $A = U \begin{bmatrix} \Sigma & O \\ O & O \end{bmatrix} V^T$

Matrix Factorizations: SVD

定义： 设 $A \in C^{n\times n}$，若 A 满足

$$A^T A = I \qquad \text{或} \qquad A^{-1} = A^T$$

则称 A 为酉矩阵。

定理（酉矩阵的性质）：

设 $A, B \in C^{n\times n}$，

(1) 若 A 是酉矩阵，则 A^{-1} 也是酉矩阵；

(2) 若 A，B 是酉矩阵，则 AB 也是酉矩阵；

(3) 若 A 是酉矩阵，则|detA|=1；

(4) A 是酉矩阵的充分必要条件是，它的 n 个列向量两两正交的单位向量.

Matrix Factorizations: SVD

$$A_{m\times n} = U_{m\times m}\begin{bmatrix} \Sigma_{r\times r} & O_{r\times(n-r)} \\ O_{(m-r)\times r} & O_{(m-r)\times(n-r)} \end{bmatrix} V_{n\times n}^{T}$$

$$= \begin{bmatrix} U_{m\times r} & U_{m\times(m-r)} \end{bmatrix} \begin{bmatrix} \Sigma_{r\times r} & O_{r\times(n-r)} \\ O_{(m-r)\times r} & O_{(m-r)\times(n-r)} \end{bmatrix} V_{n\times n}^{T}$$

$$= \begin{bmatrix} U_{m\times r}\Sigma_{r\times r} + U_{m\times(m-r)}O_{(m-r)\times r} & U_{m\times r}O_{r\times(n-r)} + U_{m\times(m-r)}O_{(m-r)\times(n-r)} \end{bmatrix} V_{n\times n}^{T}$$

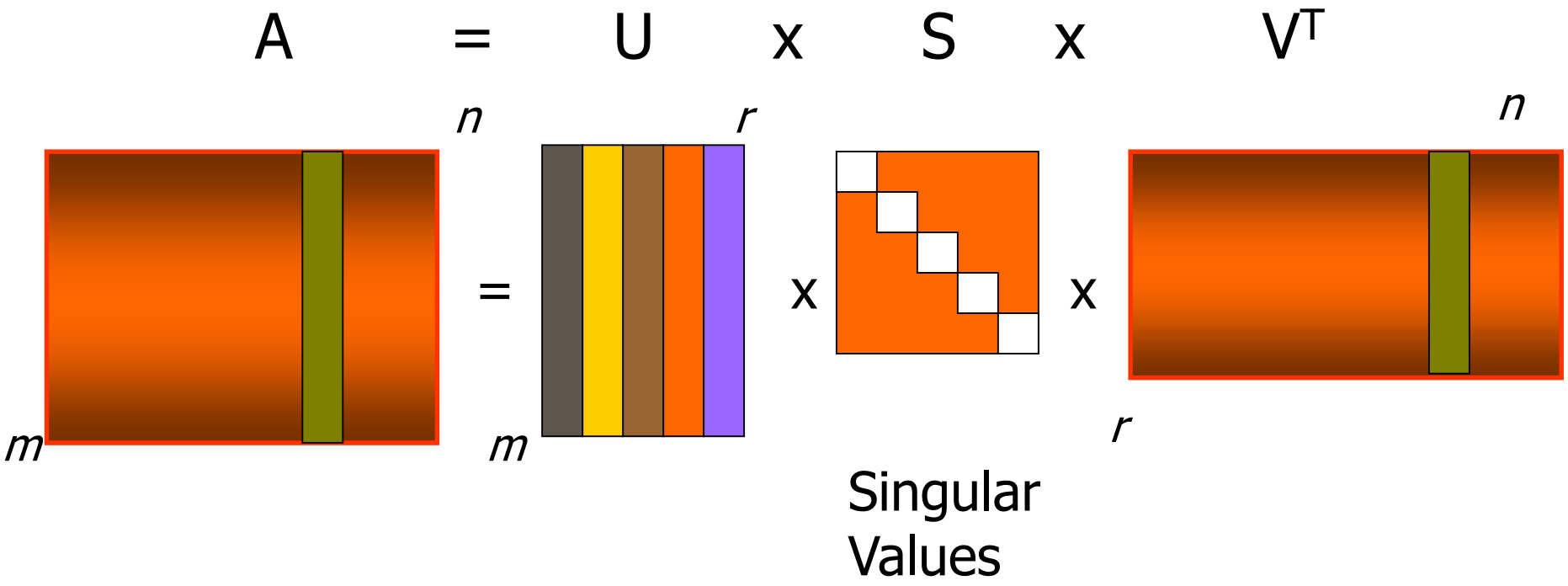
$$= \begin{bmatrix} U_{m\times r}\Sigma_{r\times r} & O_{m\times(n-r)} \end{bmatrix} V_{n\times n}^{T}$$

$$= \begin{bmatrix} U_{m\times r}\Sigma_{r\times r} & O_{m\times(n-r)} \end{bmatrix} \begin{bmatrix} V_{n\times r} & V_{n\times(n-r)} \end{bmatrix}^{T}$$

$$= U_{m\times r}\Sigma_{r\times r}V_{n\times r}^{T} + O_{m\times(n-r)}V_{n\times(n-r)}^{T}$$

$$= U_{m\times r}\Sigma_{r\times r}V_{n\times r}^{T}$$

Matrix Factorizations: SVD

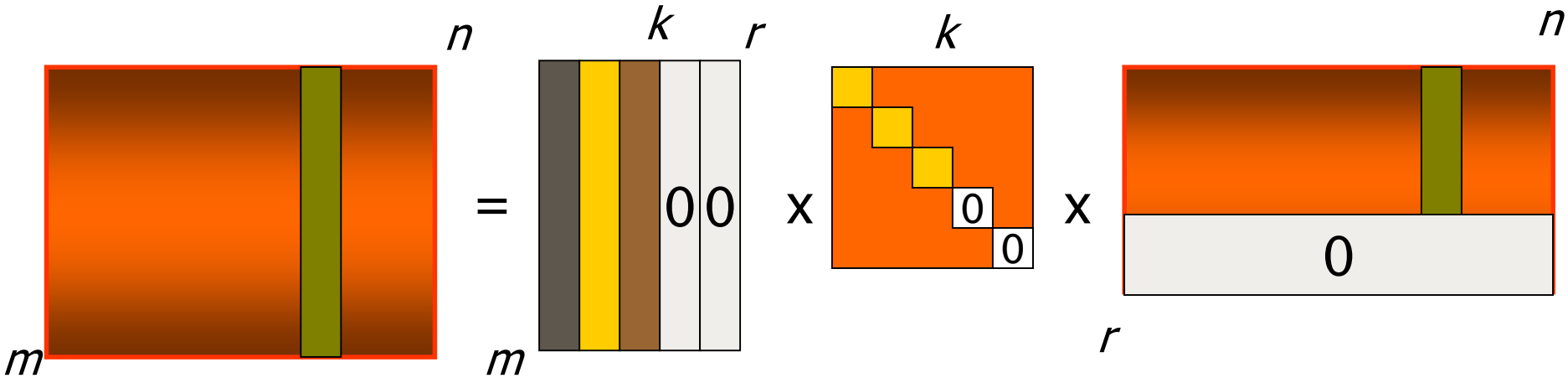


U, V orthonormal; S diagonal

A: a term by document matrix; U: a term by dimension matrix;
S: a singular value matrix; V: a document by dimension matrix

Matrix Factorizations: SVD

A_k = U x S_k x V^T



$$A_{m\times n} \approx U_{m\times k}\Sigma_{k\times k}V_{n\times k}^{\ T}$$

Matrix Factorizations: SVD

定义：矩阵A的Frobenius范数（简称F-范数）

$$\| A \|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} | a |_{ij}^2}$$

F-范数的性质：

对任意 $A \in C^{m\times n}$，A 的 F-范数满足：

（1）非负性；

（2）齐次性：对任意 $\lambda \in C, \| \lambda A \| = | \lambda | \| A \|$；

（3）三角不等式：对任意 $A, B \in C^{m\times n}$，都有 $\| A + B \| \le \| A \| + \| B \|$；

（4）相容性：对任意 $A, B \in C^{m\times n}$，都有 $\| AB \| \le \| A \| \| B \|$；

（5）酉不变性.

对比：向量 2-范数

Matrix Factorizations: SVD

Relative distances are (approximately) preserved by projection:

Of all $m \times n$ rank k matrices, A_k is the best approximation to A according to F-norm

.

定理： $\| A - A_k \|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + ... + \sigma_r^2}$

存储单元：m*k+k+k*n=k(m+n+1)
原始：m*n

复杂度： $(m+n)^2 \times k^3$

Query Matching

for query q:

$$\cos\theta_j = \frac{(\mathrm{A}_k \mathrm{e}_j)^{\mathrm{T}} q}{\|\mathrm{A}_k \mathrm{e}_i\|_2 \|q\|_2} = \frac{(U_{m\times k}\Sigma_{k\times k}{V_{n\times k}}^{T}\mathrm{e}_j)^{\mathrm{T}} q}{\|U_{m\times k}\Sigma_{k\times k}{V_{n\times k}}^{T}\mathrm{e}_j\|_2 \|q\|_2} = \frac{{\mathrm{e}_j}^{\mathrm{T}} V_{n\times k}\Sigma_{k\times k}(U_{m\times k}{}^{\mathrm{T}} q)}{\|\Sigma_{k\times k}{V_{n\times k}}^{T}\mathrm{e}_j\|_2 \|q\|_2}$$

let:

$$s_j = \Sigma_{k\times k}{V_{n\times k}}^{T}\mathrm{e}_j$$

(the scaled document vector)

then:

$$\cos\theta_j = \frac{{\mathrm{s}_j}^{\mathrm{T}}(U_{m\times k}{}^{\mathrm{T}} q)}{|\mathrm{s}_j\|_2 \|q\|_2}, j = 1,2\ldots,n$$

Term-Term & Doc-Doc Matrix

Computing two terms:

$$A_k A_k^T = U_{m\times k}\Sigma_{k\times k}V_{n\times k}{}^T(U_{m\times k}\Sigma_{k\times k}V_{n\times k}{}^T)^T = U_{m\times k}\Sigma_{k\times k}V_{n\times k}{}^T V_{n\times k}\Sigma_{k\times k}{}^T U_{m\times k}{}^T$$
$$= U_{m\times k}\Sigma_{k\times k}{}^2 U_{m\times k}{}^T = (U_{m\times k}\Sigma_{k\times k})(U_{m\times k}\Sigma_{k\times k})^T$$

rows of $U_{m\times k}\Sigma_{k\times k}$ as coordinates for terms

Computing two documents:

$$A_k^T A_k = V_{n\times k}\Sigma_{k\times k}{}^2 V_{n\times k}{}^T = (V_{n\times k}\Sigma_{k\times k})(V_{n\times k}\Sigma_{k\times k})^T$$

rows of $V_{n\times k}\Sigma_{k\times k}$ as coordinates for documents

Term-Doc Matrix

$$A_k = U_{m\times k}\Sigma_{k\times k}V_{n\times k}{}^T = (U_{m\times k}\Sigma_{k\times k}{}^{1/2})(V_{n\times k}\Sigma_{k\times k}{}^{1/2})^T$$

Example

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Query: human interaction with computers
C3 and C5

Example

"human vs. user"

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

T_m =

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

S_m =

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

Example

$D_m =$

0.20	-0.06	0.11	-0.95	0.05	-0.08	0.18	-0.01	-0.06
0.61	0.17	-0.50	-0.03	-0.21	-0.26	-0.43	0.05	0.24
0.46	-0.13	0.21	0.04	0.38	0.72	-0.24	0.01	0.02
0.54	-0.23	0.57	0.27	-0.21	-0.37	0.26	-0.02	-0.08
0.28	0.11	-0.51	0.15	0.33	0.03	0.67	-0.06	-0.26
0.00	0.19	0.10	0.02	0.39	-0.30	-0.34	0.45	-0.62
0.01	0.44	0.19	0.02	0.35	-0.21	-0.15	-0.76	0.02
0.02	0.62	0.25	0.01	0.15	0.00	0.25	0.45	0.52
0.08	0.53	0.08	-0.03	-0.60	0.36	0.04	-0.07	-0.45

Example

$\hat{X} =$

T

0.22	-0.11
0.20	-0.07
0.24	0.04
0.40	0.06
0.64	-0.17
0.27	0.11
0.27	0.11
0.30	-0.14
0.21	0.27
0.01	0.49
0.04	0.62
0.03	0.45

S

3.34	
	2.54

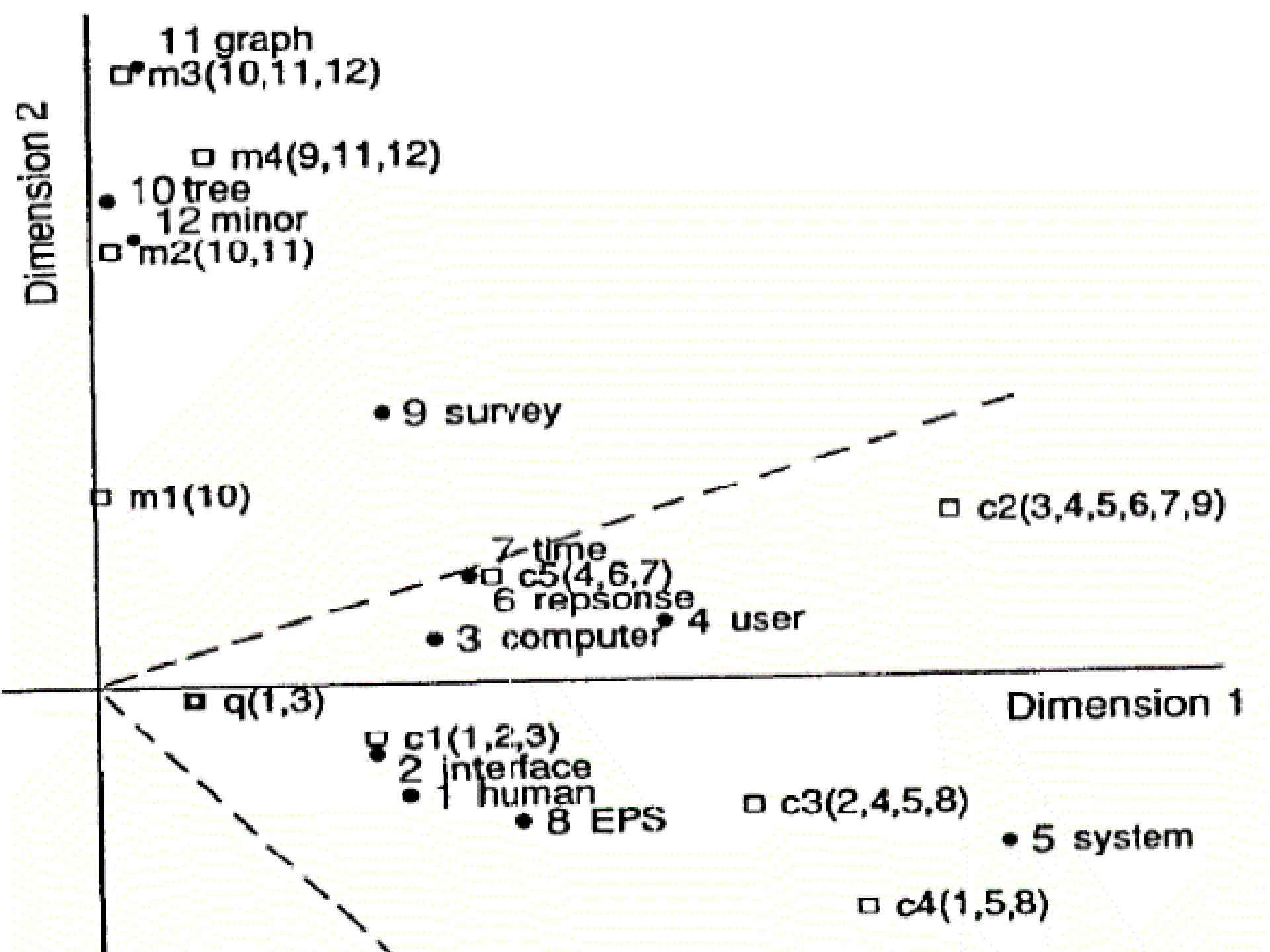
D′

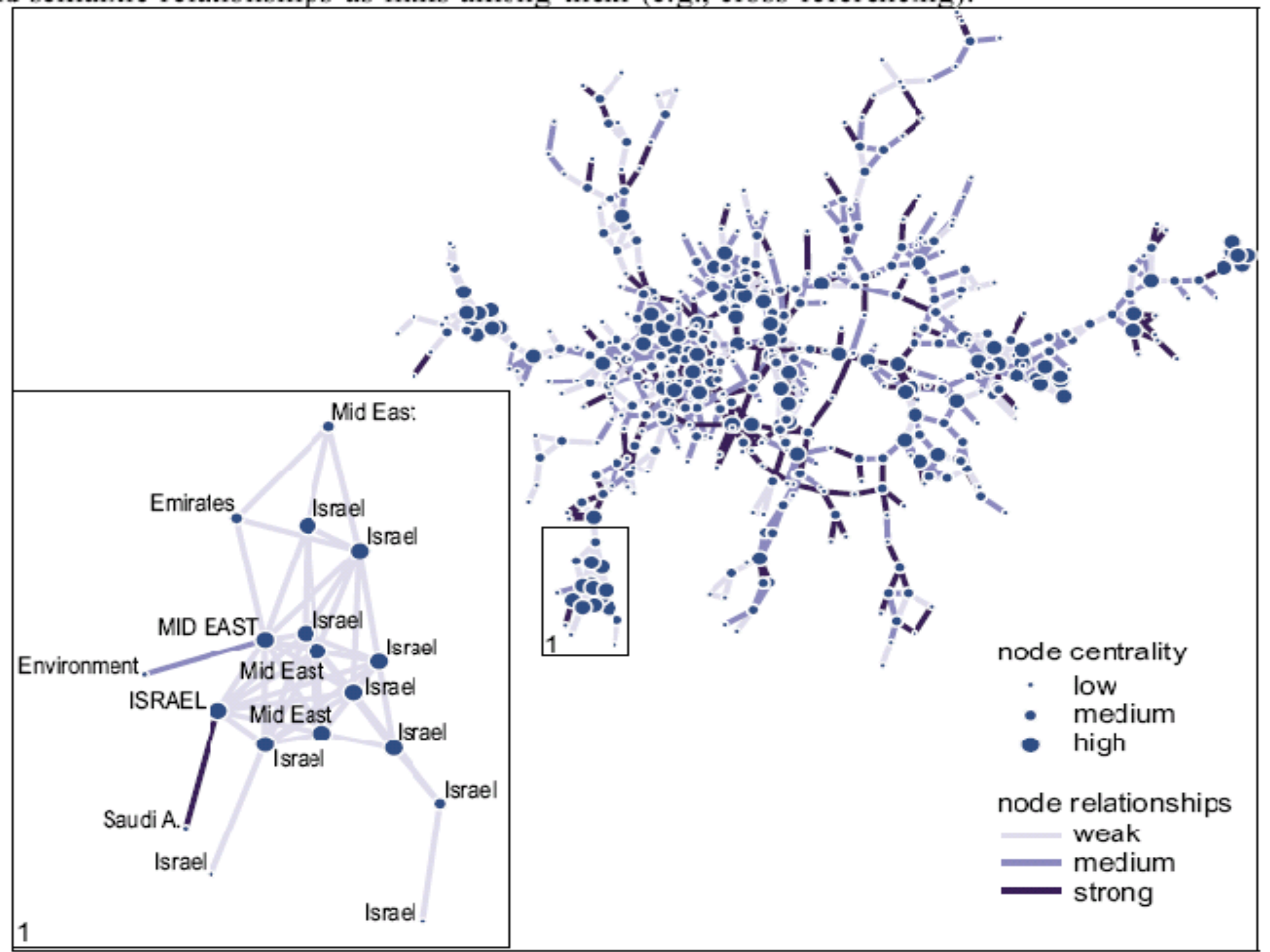
0.20	0.61	0.46	0.54	0.28	0.00	0.02	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53

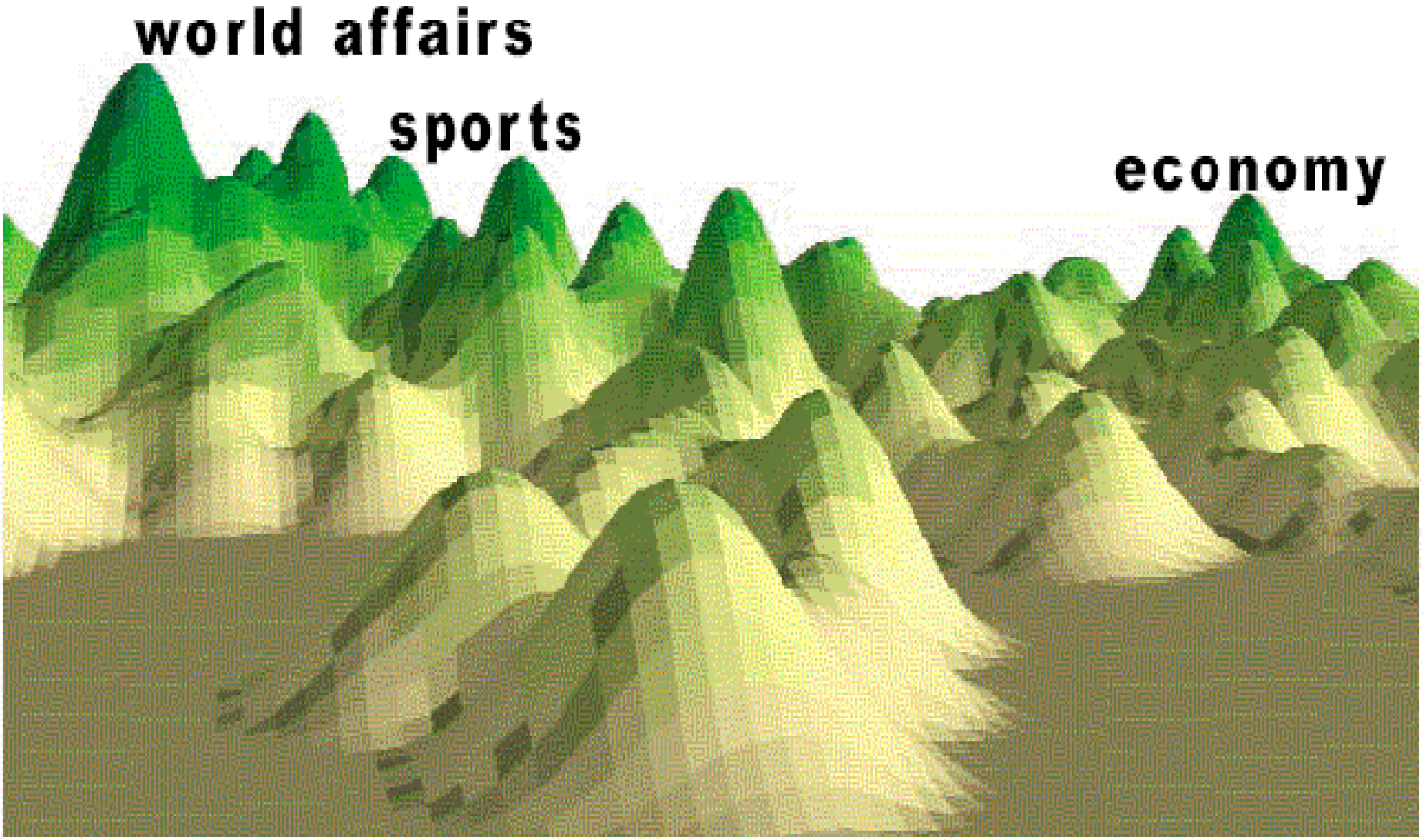
Example

$X =$

0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62


11 graph
m3(10,11,12)
m4(9,11,12)
10 tree
12 minor
m2(10,11)
9 survey
m1(10)
c2(3,4,5,6,7,9)
7 time
c5(4,6,7)
6 repsonse
4 user
3 computer
q(1,3)
c1(1,2,3)
2 interface
1 human
8 EPS
c3(2,4,5,8)
5 system
c4(1,5,8)
Dimension 2
Dimension 1


Mid Eas:
Emirates
Israel
Israel
MID EAST
Israel
Environment
Israel
Mid East
ISRAEL
Israel
Mid East
Israel
Israel
Israel
Saudi A.
Israel
Israel
1
1
node centrality
low
medium
high
node relationships
weak
medium
strong



Topic Density Surface

Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure

George W. Furnas (Bellcore)
Scott Deerwester (University of Chicago)
Susan T. Dumais (Bellcore) *Thomas K. Landauer (Bellcore)*
Richard A. Harshman (University of Western Ontario)
Lynn A. Streeter (Bellcore) *Karen E. Lochbaum (Bellcore)*

ABSTRACT

In a new method for automatic indexing and retrieval, implicit higher-order structure in the association of terms with documents is modeled to improve estimates of term-document association, and therefore the detection of relevant documents on the basis of terms found in queries. Singular-value decomposition is used to decompose a large term by document matrix into 50 to 150 orthogonal factors from which the original matrix can be approximated by linear combination; both documents and terms are represented as vectors in a 50- to 150-dimensioal space. Queries are represented as pseudo-documents vectors formed from weighted combinations of terms, and documents are ordered by their similarity to the query. Initial tests find this automatic method very promising.

A Mathematical View of Latent Semantic Indexing: Tracing Term Co-occurrences

April Kontostathis
Lehigh University
CSE Department
19 Memorial Drive West
Bethlehem, PA 18015
apk5@lehigh.edu

William M. Pottenger, Ph.D.
Lehigh University
CSE Department
19 Memorial Drive West
Bethlehem, PA 18015
billp@cse.lehigh.edu

ABSTRACT

Current research in Latent Semantic Indexing (LSI) shows improvements in performance for a wide variety of information retrieval systems. We propose the development of a theoretical foundation for understanding the values produced in the reduced form of the term-term matrix. We assert that LSI's use of higher orders of co-occurrence is a critical component of this study. In this work we present experiments that precisely determine the degree of co-occurrence used in LSI. We empirically demonstrate that LSI uses up to fifth order term co-occurrence. We also prove mathematically that a connectivity path exists for every nonzero element in the truncated term-term matrix computed by LSI. A complete understanding of this term transitivity is key to understanding LSI.

Table 4: Order of Co-occurrence Summary Data

k =100 for all Collections

	Number of pairs with order of co-occurrence					
Collection	First	Second	Third	Fourth	Fifth	Sixth
MED Truncated	1,110,485	15,867,200	17,819	-	-	
MED Original	1,110,491	15,869,045	17,829	-	-	
CRAN Truncated	2,420,520	18,817,356	500	-	-	
CRAN Original	2,428,588	18,836,832	512	-	-	
CISI Truncated	2,327,918	29,083,372	17,682	-	-	
CISI Original	2,328,026	29,109,528	17,718	-	-	
LISA Words Truncated	5,380,788	308,556,728	23,504,606	-	-	
LISA Words Original	5,399,343	310,196,402	24,032,296			
LISA Noun Phrase Truncated	51,350	10,976,417	65,098,694	1,089,673	3	
LISA Noun Phrase Original	51,474	11,026,553	68,070,600	2,139,117	15,755	34

Theorem: if the ij_{th} element of the truncated term by term matrix is nonzero, then there is a transitivity path between term i and term j.