1. Vector (space)model Introduction

$$D \subseteq R^{n+}$$

$$Q \subseteq R^{n}$$

Retrieval functions

$$f : D \times Q \rightarrow R$$

$$\underline{d} = (d_1, d_2, \ldots, d_n)$$

$$\underline{q} = (q_1, q_2, \ldots, q_n)$$

Dot product function

$$\underline{dq}^T = \sum_{i=1}^{n} d_i q_i$$

2. TWO VIEWS OF VECTOR CONCEPT

-- VECTOR (PROCESSING)
"MODEL"

NOTATIONAL OR
DATA STRUCTURAL
ASPECT

-- VECTOR SPACE MODEL

- DOCUMENTS, QUERIES, ETC. ARE ELEMENTS OF A VECTOR SPECE

- ANALYTICAL TOOL

3. THE VECTOR SPACE MODEL

- MATHEMATICAL ASPECTS

- MAPPING OF DATA ELEMENTS TO MODEL CONSTRUCTS

3.1 MATHEMATICAL ASPECTS
3.1.1 BASIC CONCEPTS

- IR OBJECTS (e.g. KEYWORDS DOCUMENS) CONSTITUTE A VECTOR SPACE

- THAT IS, WE HAVE A SYSTEM WITH LINEAR PROPERTIES:

(i) ADDITION OF VECTORS
(ii) MULTIPLICATION BY SCALAR

CLOSURE

- BASIC ALGEBRAIC AXIOMS

e.g. $\underline{x} + \underline{y} = \underline{y} + \underline{x}$

$\underline{x} + \underline{o} = \underline{x}$ i.e. $\underline{o}$ exists

For each $\underline{x}$, $\exists$ $-\underline{x}$

$\alpha (\underline{x} + \underline{y}) = \alpha \underline{x} + \alpha \underline{y}$

.
.
.
etc

LINEAR INDEPENDENCE

A SET OF VECTORS $y_1, y_2 \ldots y_k$ IS <u>LINEARLY</u> <u>INDEPENDENT</u> (L.I.) IF

$\alpha_1 \underline{y}_1 + \alpha_2 \underline{y}_2 + \ldots + \alpha_K \underline{y}_K = \underline{o}$,

WHERE α_i'S ARE SCALARS,

ONLY IF $\alpha_1 = \alpha_2 = \ldots \alpha_K = o$

- BASIS: A <u>GENERATING</u> <u>SET</u> CONSISTING OF L.I. VECTORS
- DIMENSION: $n' \leq n$, where n is the size of the generating set
- $\{ \underline{t}_{i1}, \underline{t}_{i2}, \ldots \underline{t}_{in'} \}$
- ANY subset of L.I. VECTORS of the generating set of size n' FORM A BASIS

(Inner) SCALAR PRODUCT

$\underline{x} \cdot \underline{y} = \|\underline{x}\| \, \|\underline{y}\| \cos\theta$,

WHERE,

θ is the angle between $\underline{x}$ and $\underline{y}$,

$\|\underline{x}\| = \sqrt{\underline{x} \cdot \underline{x}}$

- The above is an instance of a scalar product
- EUCLIDEAN SPACE: A VECTOR SPACE EQUIPPED WITH A SCALAR PRODUCT
- ORTHO GONAL : $\underline{x} \cdot \underline{y} = o$
- NORMALIZING : $\underline{x} / \|\underline{x}\|$
- ORTHONORMAL BASIS

  If underlying basis is orthonormal,

  $$\underline{x} \cdot \underline{y} = \sum_{i=1}^{n} x_i y_i$$

3.1.2 LINEAR INDEPENDENCE VS. ORTHOGONALITY

IF A SET OF NON-ZERO VECTORS
$y_1, y_2 \ldots y_k$ are MUTUALLY ORTHOGONAL ($\underline{x}_i \cdot \underline{y}_j = o$ for all $i \neq j$), then they are LINEARLY INDEPENDENT. But a set of linearly independent vectors is not necessarily mutually orthogonal.

UNDER THE SITUATION OF NON-ORTHOGONAL Generating set, issues of

(i) linear dependence, and
(ii) correlation *

MUST BE CONSIDERED.

* (term, term) relationship

3.1.3 REPRESENTATION IN IR

KEYWORDS:

$t_1, t_2, t_3 \ldots t_n$

VECTORS:

$\underbrace{\underline{t}_1, \underline{t}_2, \underline{t}_3 \ldots \underline{t}_n}$

Generating set

$\underline{d}_\alpha = (a_{1\alpha}, a_{2\alpha}, \ldots a_{n\alpha})$

OR

$$\underline{d}_\alpha = \sum_{i=1}^{n} a_{i\alpha} \underline{t}_i$$

3.1.4 IMPORTANT RELATIONSHIPS

ASSUME:

$$n' = n = p$$

$$\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_n$$

$$\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_n$$

Basis can be either

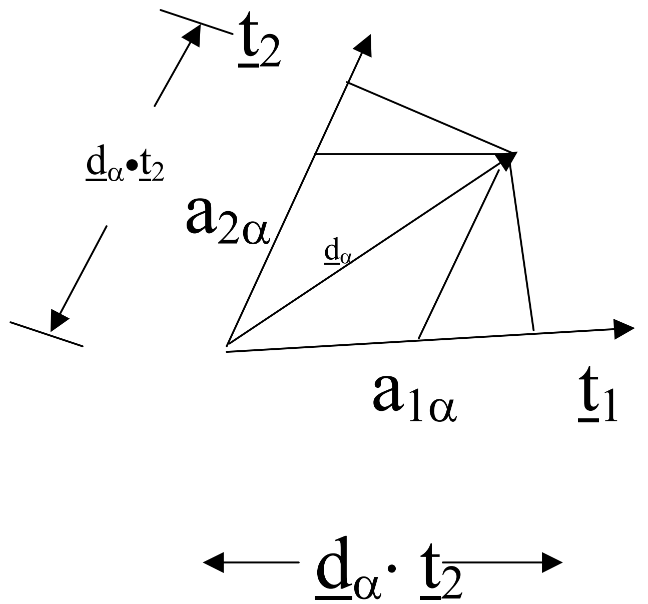
$$\|\underline{t}_i\| = 1, I = 1, 2, \ldots n$$

THUS,

$$\underline{d}_\alpha = \sum_{i=1}^{n} a_{i\alpha} \underline{t}_i \quad \ldots (1)$$

OR

$$\underline{t}_i = \sum_{\alpha=1}^{n} b_{\alpha i} \underline{d}_\alpha \quad \ldots (2)$$



$n=2$
assume $\underline{t}_1$ and $\underline{t}_2$ are normalized

Projection and component are NOT the same, when the basis vectors are non-orthogonal

3.1.5 PROJECTION VS. COMPONENTS

FOR VECTORS, $\underline{x}$, $\underline{y}$

$(\underline{x} / \|\underline{x}\|) \cdot \underline{y}$ IS THE

PROJECTION OF $\underline{Y}$ ONTO $\underline{X}$.

3.1.4 (Contd.)

By MULTIPLYING equ. (1) by $\underline{t}_j$ ON BOTH SIDES,

$$\underline{t}_j \cdot \underline{d}_\alpha = \sum_{i=1}^{n} a_{i\alpha} t_j \cdot t_i \, ,$$

$$1 \le \alpha, j \le n \ldots (3)$$

If $\underline{t}$'s ARE NORMALIZED, THE LEFT HAND SIDE IS THE PROJECTION OF $\underline{d}_\alpha$ ONTO $\underline{t}_j$

WRITING EQN. (3) IN A MATRIX FORM, WE HAVE

$$P = G_t A \ldots (4)$$

WHERE

$(P)_{j\alpha} = \underline{t}_j \cdot \underline{d}_\alpha$

$(G_t)_{ji} = \underline{t}_j \cdot \underline{t}_i$

$(A)_{i\alpha} = a_{i\alpha}$

RESPECTIVELY,

PROJECTIONS,

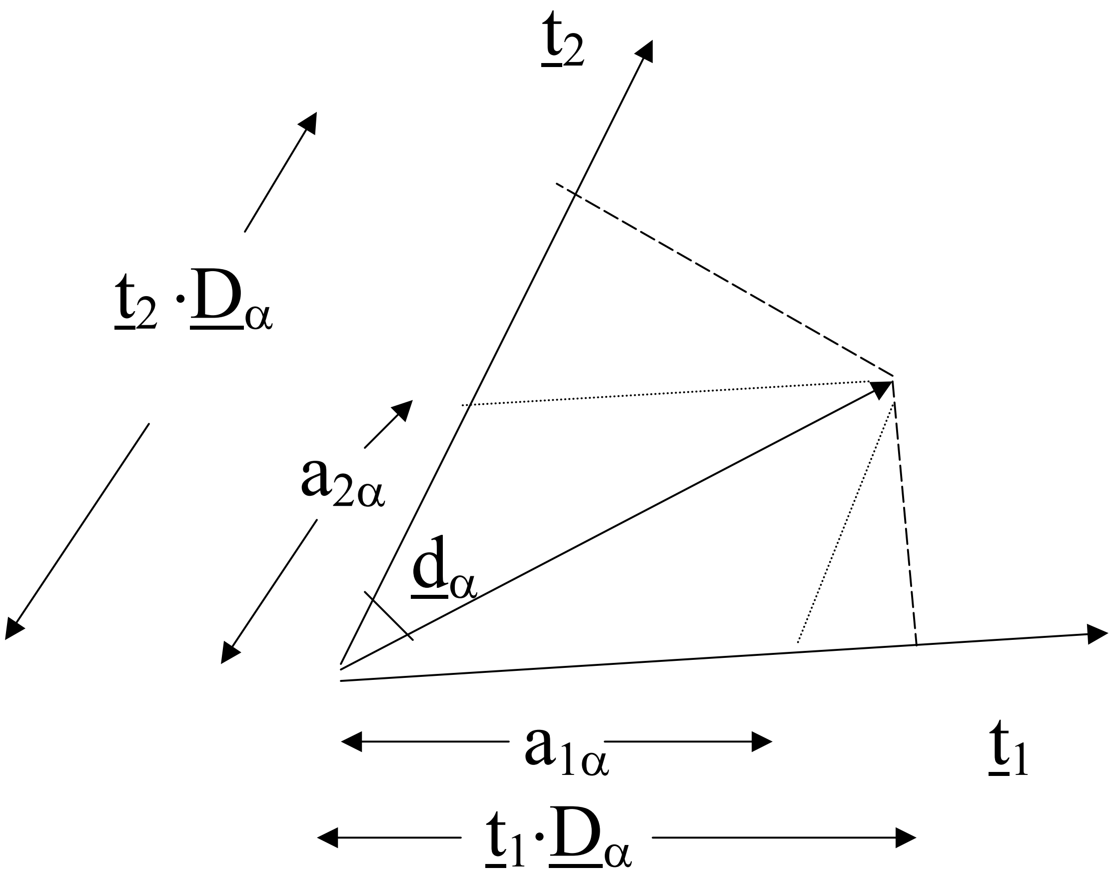
TERM CORRELATIONS

&

COMPONENTS OF $\underline{d}$'s

<u>EXAMPLE</u> 1

n=2



$\underline{d}_\alpha = a_{1\alpha}\underline{t}_1 + a_{2\alpha}\underline{t}_2$ … (5)

LET $\underline{d}_1, \underline{d}_2$ BE A BASIS (L.I.)

THEN,

$$G_t A = \begin{bmatrix} \underline{t}_1\cdot\underline{t}_1 & \underline{t}_1\cdot\underline{t}_2 \\ \underline{t}_2\cdot\underline{t}_1 & \underline{t}_2\cdot\underline{t}_2 \end{bmatrix} \begin{bmatrix} \underline{a}_{11} & \underline{a}_{12} \\ \underline{a}_{21} & \underline{a}_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \underline{t}_1 \cdot (a_{11}\,\underline{t}_1 + a_{21}\underline{t}_2) & \underline{t}_1 \cdot (a_{12}\,\underline{t}_1 + a_{22}\underline{t}_2) \\ \underline{t}_2 \cdot (a_{11}\,\underline{t}_1 + a_{21}\underline{t}_2) & \underline{t}_2 \cdot (a_{12}\,\underline{t}_1 + a_{22}\underline{t}_2) \end{bmatrix}$$

USING EQN. (5), WE HAVE

$$= \begin{bmatrix} \underline{t}_1 \cdot \underline{d}_1 & \underline{t}_1 \cdot \underline{d}_2 \\ \underline{t}_2 \cdot \underline{d}_1 & \underline{t}_2 \cdot \underline{d}_2 \end{bmatrix}$$

$$= P$$

SIMILARLY,

STARTING FROM EQN. (2) AND MULTIPLYING BOTH SIDES BY $\underline{d}_\beta$, AND WRITING IN MATRIX FORM.

$$P^T = G_d B \; \ldots \; (6)$$

WHERE

$$(G_d)_{\beta\alpha} = \underline{d}_\beta \cdot \underline{d}_\alpha$$
$$(B)_{\alpha i} = b_{\alpha i}$$

THAT IS,

DOCUMENT CORRELATIONS

AND

COMPONENTS OF $\underline{t}$'s ALONG DOCUMENTS

CAN further SHOW, $PB = G_t \ldots (7)$

$P^T A = G_d \ldots (8)$

3.1.6 DOCUMENT RANKING

$$\underline{q} = \sum_{i=1}^{n} q_i t_i$$

$$\underline{d}_\alpha \cdot \underline{q} = \left(\sum_{i=1}^{n} a_{i\alpha} \cdot t_i\right) \cdot \left(\sum_{j=1}^{n} q_j t_j\right)$$

$$= \sum_{i,j=1}^{n} a_{i\alpha} q_j t_i t_j \quad (9)$$

EXAMPLE 2

$n=2$ $\qquad$ $A^T G_t q^T$

$$\underline{q} = q_1 \underline{t}_1 + q_2 \underline{t}_2$$

$$\underline{d}_\alpha = a_{1\alpha} \underline{t}_1 + a_{2\alpha} \underline{t}_2$$

$$\underline{d}_\alpha \underline{q} = a_{1\alpha} q_1 \underline{t}_1 \cdot \underline{t}_1$$

$$+ a_{2\alpha} q_2 \underline{t}_2 \cdot \underline{t}_2$$

$$+ a_{1\alpha} q_2 \underline{t}_2 \cdot \underline{t}_1$$

$$+ a_{2\alpha} q_1 \underline{t}_1 \cdot \underline{t}_2$$

3.2 MAPPING OF DATA ELEMENTS TO MODEL CONSTRUCTS

Term Frequency Data

term

$$\mathbf{W} = \text{document} \begin{bmatrix} & & \\ & w_{\alpha i} & \\ & & \end{bmatrix}$$

May be interpreted as

A^T or B or P^T

But, this alone is NOT enough

*By interpretation we mean how data obtained from real-world documents are mapped to model constructs such as, A, B and G_t.

Text Analysis

- Controlled vs. Free vocabulary
- Single term Indexing
    a. Extract words
    b. Stop list
    c. Stemming
    d. Term weight assignment

$$\text{RSV}(q, d_\alpha) = \sum_i \frac{\left(0.5 + 0.5\frac{f_{\alpha i}}{\max\limits_j \left(f_{\alpha j}\right)}\right)\log\left(\frac{N}{n_i}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(0.5 + 0.5\frac{f_{\alpha i}}{\max\limits_j \left(f_{\alpha j}\right)}\right)^2 \left(\log\left(\frac{N}{n_i}\right)\right)^2}}$$

- More general descriptions

a. phrases
b. thesaurus entries

3.2.1 TWO WAYS OF MAPPING W TO THE MODEL

Method I. Mapping W^T to A

$A \equiv W^T$

$RSV_{\underline{q}} = (\underline{d}_1 \cdot \underline{q}, \ \underline{d}_2 \cdot \underline{q}, \ \ldots$
$\ldots \ \underline{d}_p \cdot \underline{q})$

$\underline{q} = (q_1, \ q, \ \ldots \ q_n)$

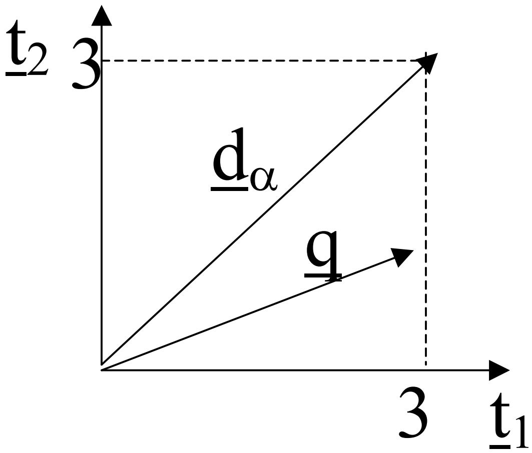
q_i – is the component of $\underline{q}$ along t_i

$RSV_{\underline{q}}^T = WG_t\underline{q}^T$
$= P^T\underline{q}^T$, since

$P^T = A^TG_t \equiv WG_t$, then

$P = G_tA \quad RSV_{\underline{q}}^T = W\underline{q}^T$

n=2



$$
\begin{array}{c c}
 & \begin{array}{cc} \underline{t}_1 & \underline{t}_2 \end{array} \\
\begin{array}{c} \underline{t}_1 \\ \underline{t}_2 \end{array} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\end{array} = G_t
$$

$$\underline{d}_\alpha = \overset{a_{1\alpha}}{(\,3,} \; \overset{a_{2\alpha}}{3\,)} \qquad \underline{q} = \overset{a_1}{(\,3,} \; \overset{q_2}{1\,)}$$

$$\underline{d}_\alpha = 3\,\underline{t}_1 + 3\,\underline{t}_2$$

$$\underline{q} = 3\,\underline{t}_1 + \underline{t}_2$$

$$\underline{d}_\alpha \cdot \underline{q} = |d_\alpha|\,|q| \cos\theta$$

$$= \sum_{i=1}^{2} a_{\alpha i} \underline{q}_i$$

$$(3\,\underline{t}_1 + \underline{t}_2) \cdot (3\,\underline{t}_1 + 3\,\underline{t}_2)$$

$$= 9\,\underline{t}_1 \cdot \underline{t}_1 + 9\,\underline{t}_1 \cdot \underline{t}_2 + 3\,\underline{t}_2\,\underline{t}_1 + 3\,\underline{t}_2\,\underline{t}_2$$

$$= 12$$

Method II. B $\underline{=}$ W   USE SAME W as Method I

$$RSV_q^T = P^T \underline{q}^T$$

$$\downarrow$$

$$G_d B$$

$$RSV_{\underline{q}}^T = G_d B \underline{q}^T$$
$$= G_d w \underline{q}^T$$

- Columns of W are used as components of term vectors along document vectors
- Elements of $\underline{q}$ are components of $\underline{q}$ along term vectors

3.2.2 USING THE MODEL COMPARISON TO EARLIER WORK

I. THE STANDARD SPECIAL CASE

- TERMS FORM AN ORTHONORMAL BASIS, $G_t=I$
- HERE, P=A (FROM(4))
- W IS INTERPRETED AS

A^T ($=P^T$) $\sum_{i=1}^{n} a_{1\alpha}\cdot q_i$ when $G_t=I$

In this case

$$\underline{d}_\alpha\cdot\underline{q} = \sum_{i=1}^{n} a_{i\alpha}\cdot q_i$$

$$= \sum_{i=1}^{n} w_{\alpha i}\cdot q_i$$

II. WHILE THE ABOVE RESTRICTIONS APPEAR COMPATIBLE, ONE OF THE PRACTICES DEFINES TERM VECTOR $\underline{t}_i$ as follow:

$$\underline{t}_i = (w_{1i},\ w_{2i},\ \ldots\ w_{ni})$$

This suggests,

$$A^t = B$$

But, according to the vector space model,

$$P = G_t A$$

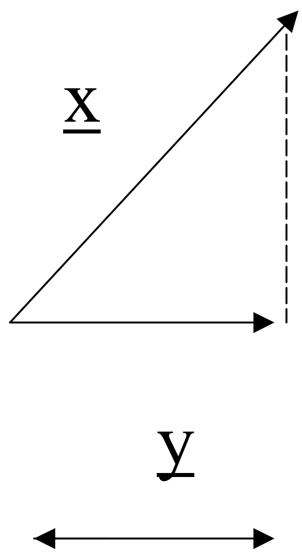
and

$$PB = G_t$$

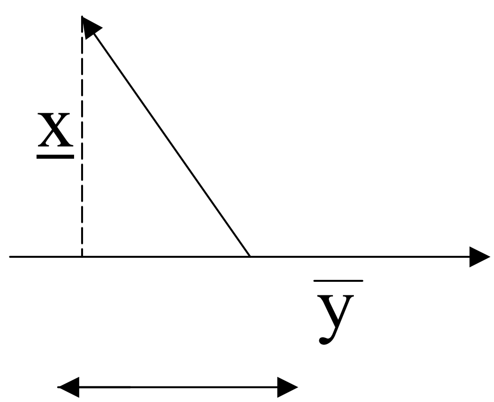
Thus, $A^{-1} = B$

IF EACH ROW OF W REPRESENTS DOCUMENTS, THEN EACH COLUMN DOES **NOT** REPRESENT TERM VECTOR, THUS, WHAT IS KNOWN TO BE COMMON PRATICE IS CONTRADICTIRY TO WHAT WE SHOW TO BE THE RELATIONSHIP BETWEEN **A** AND **B** MATRICES.

Can Projection be negative?



Projection of
x on y is +



Projection of
x on y is -

3.2.3 Other commonly used retrived functions

Measures of vector similarity

Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Dice coefficient	$2\frac{\lvert X \cap Y\rvert}{\lvert X\rvert + \lvert Y\rvert}$	$\frac{2\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y\rvert}{\lvert X\rvert^{1/2} \cdot \lvert Y\rvert^{1/2}}$ $\quad \frac{\sum x_i y_i}{\lVert X\rVert \cdot \lVert Y\rVert}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y\rvert}{\lvert X\rvert + \lvert Y\rvert - \lvert X \cap Y\rvert}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$

$X = \{t_i\}$
$Y = \{t_j\}$

$X = (x_1, x_2, \ldots x_t)$
$|X|$ = number of terms in X
$|X \cap Y|$ = number of terms appearing jointly in X and Y