ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG V – MỘT SỐ MÔ HÌNH TRUY XUẤT THÔNG TIN KHÁC

Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR
- MÔ HÌNH LSI
- MÔ HÌNH XÁC SUẤT

NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

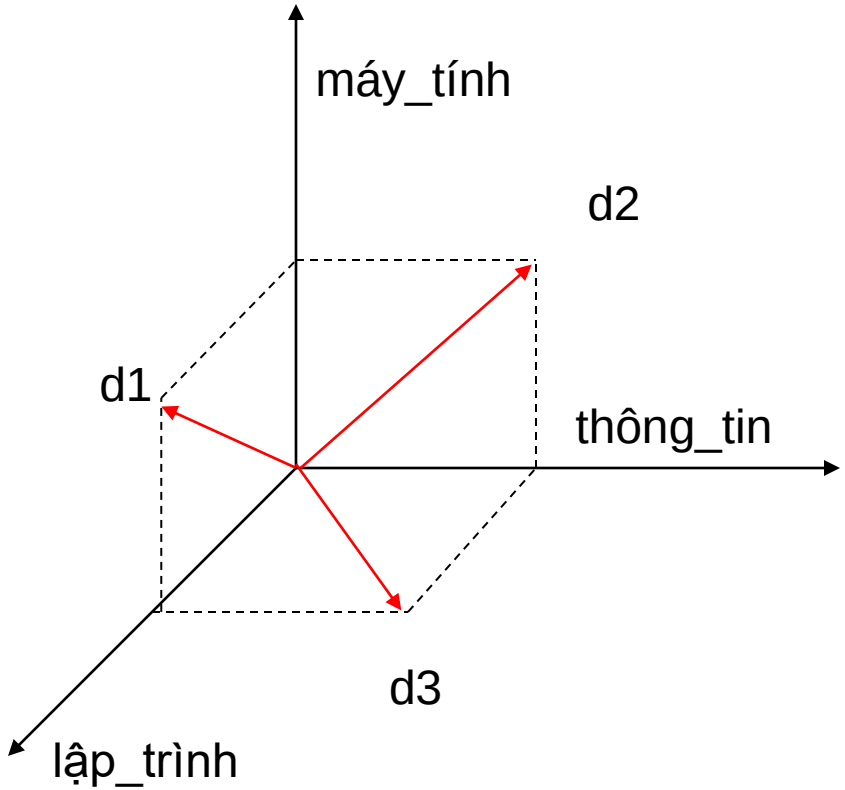
❖NHƯỢC ĐIỂM TRONG BIỂU DIỄN

Giả sử có các tài liệu:

d1: máy_tính lập_trình

d2: máy_tính thông_tin

d3: lập_trình thông tin



NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

❖ NHƯỢC ĐIỂM TRONG BIỂU DIỄN

- Các chiều trong không gian là mỗi từ khóa, không đảm bảo độc lập về ngữ nghĩa.
- Các vector chỉ nằm trong phần dương của không gian.
- Số lượng chiều rất lớn, trong đó có những từ khóa có thể không cần thiết phải biểu diễn (những từ nhiễu)

NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR

❖ NHƯỢC ĐIỂM TRONG SO KHỚP

- So khớp dựa vào từ khóa, nếu tài liệu và truy vấn không có từ khóa chung thì độ tương đồng bằng 0
- Chưa có cơ chế so khớp những từ có nghĩa gần nhau, chẳng hạn: “máy_tính” và “lập_trình”

MÔ HÌNH LSI

❖MỤC TIÊU CỦA MÔ HÌNH LSI

Mô hình LSI (Latent Semantic Index) được đề xuất nhằm:

- Đảm bảo các chiều trong không gian là độc lập.
- Các chiều được chọn mang ý nghĩa của những từ khoá dựa trên sự xuất hiện đồng thời của chúng. Không nhất thiết phải là tập từ khóa sử dụng trong các tài liệu (ngữ nghĩa tiềm ẩn).
- Có thể giảm số chiều trong không gian mà cho kết quả xấp xỉ.

MÔ HÌNH LSI

❖CƠ SỞ TOÁN

Cho ma trận A, vector x được gọi là vector riêng của A nếu tồn tại một số λ, gọi là trị riêng sao cho:

$$Ax = \lambda x$$

→ x là vector riêng của A thì x không thay đổi phương khi nhân với A.

→ Giả sử $x_1, x_2, ..., x_n$ là vector riêng ứng với các trị riêng λ_i khác nhau của A, khi đó, x_i và x_j ($i \neq j$) độc lập tuyến tính.

MÔ HÌNH LSI

❖CƠ SỞ TOÁN

Cho ma trận A, giả sử $\sigma_1^2 \geq \sigma_2^2 \geq .. \geq \sigma_n^2$ là các trị riêng của A^TA tương ứng với các vector riêng $x_1, x_2, .., x_n$.

Đặt: $U = \{x_1, x_2, .., x_n\}$

$y_i = (1/\sigma_i)A\, x_i$ (i = 0..n)

$S = \{y_1, y_2, .., y_n\}$

Σ là matrận đường chéo trong đó các giá trị trên đường chéo là $\sigma_1, \sigma_2, .., \sigma_n$

Khi đó $A = S\, \Sigma\, U^T$

Được gọi là một phép tách ma trận SVD (Singular Value Decomposition)

MÔ HÌNH LSI

❖CƠ SỞ TOÁN

Trong trường hợp chỉ chọn r giá trị riêng đầu tiên (r<<n)

Đặt: $U_r = \{x_1, x_2, .., x_r\}$

$y_i = (1/\sigma_i) A\, x_i \; (i = 0..r)$

$S_r = \{y_1, y_2, .., y_r\}$

Σ là matrận đường chéo trong đó các giá trị trên đường chéo là $\sigma_n, \sigma_n, ..., \sigma_r$

Khi đó $A \approx S_r \Sigma U_r^T$

MÔ HÌNH LSI

❖LATENT SEMANTIC INDEX

Áp dụng phép tách ma trận SVD cho ma trận A là ma trận Term-Document:

$$A = S \Sigma U^T$$

Khi đó, các vector từ khóa K và các vector tài liệu D sẽ được biểu diễn trong cùng không gian với các chiều là các vector riêng qua phép biến đổi:

$$K = S \Sigma$$

$$D = \Sigma U^T$$

MÔ HÌNH LSI

❖LATENT SEMANTIC INDEX

Các vector riêng được xem là những nghĩa tiềm ẩn trong mối liên hệ cùng xuất hiện của các từ khóa

Trong trường hợp muốn giảm số chiều, sẽ chọn r trị riêng và vector riêng đầu tiên, khi đó

$$K = S_r \Sigma_r$$

$$D = \Sigma_r U_r^T$$

MÔ HÌNH LSI

❖LATENT SEMANTIC INDEX

Ví dụ:

	d_1	d_2	d_3
máy_tính	1	1	0
lập_trình	1	0	1
thông_tin	0	1	1

$$S = \begin{pmatrix} 0.57 & -0.4 & 0.71 \\ 0.57 & -0.4 & -0.71 \\ 0.57 & 0.82 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0.57 & -0.81 & 0 \\ 0.57 & 0.41 & 0.71 \\ 0.57 & 0.41 & -0.71 \end{pmatrix}$$

MÔ HÌNH LSI

❖LATENT SEMANTIC INDEX

$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Công cụ tách ma trận theo SVD trực tuyến:

http://users.telenet.be/paul.larmuseau/SVD.htm

MÔ HÌNH LSI

❖TÍNH ĐỘ TƯƠNG ĐỒNG GIỮA TÀI LIỆU VÀ TRUY VẤN

- Chuyển truy vấn thành vector dựa trên các vector từ khóa.
- Tính độ đo cosine (hoặc độ đo nào khác tùy chọn) dựa trên vector truy vấn và vector tài liệu.

MÔ HÌNH LSI

❖TÍNH ĐỘ TƯƠNG ĐỒNG GIỮA TÀI LIỆU VÀ TRUY VẤN

Ví dụ:

d_1 Romeo and Juliet

d_2 Juliet: Oh happy dagger

d_3 Romeo died by dagger

d_4 “live free or die” is from New-Hampshire

d_5 he is from New-Hampshire

Xếp hạng tài liệu theo truy vấn q: "die dagger"