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

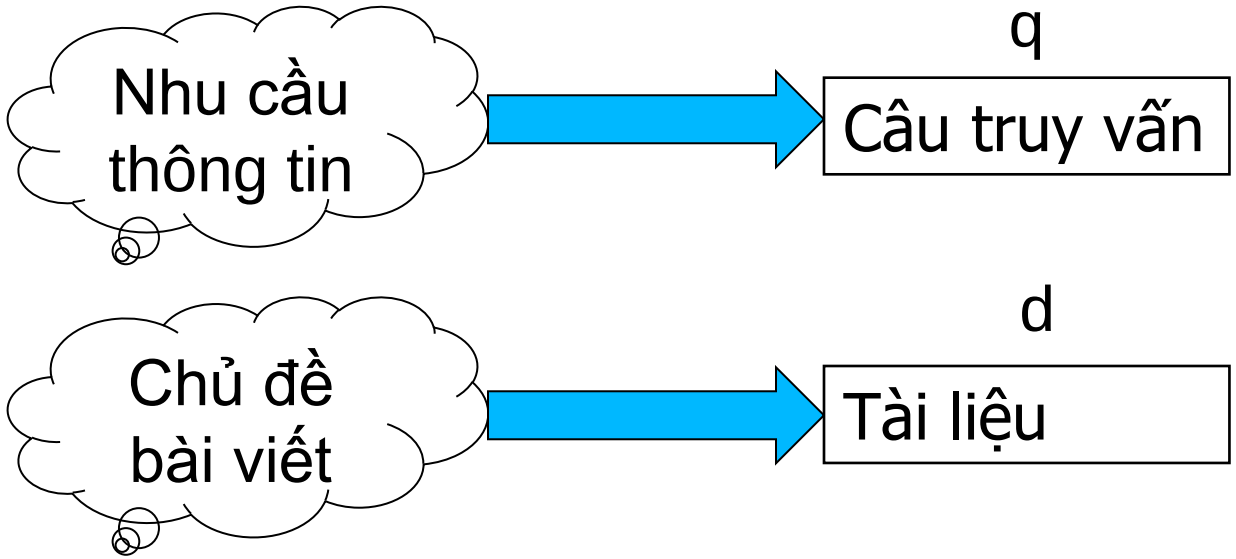
- NHƯỢC ĐIỂM CỦA MÔ HÌNH VECTOR
- MÔ HÌNH LSI
- MÔ HÌNH XÁC SUẤT

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Hàm so khớp của mô hình xác suất được xây dựng trên công thức p(R|d,q). Vấn đề đặt ra là:

- q được tạo ra như thế nào?
- d được tạo ra như thế nào?

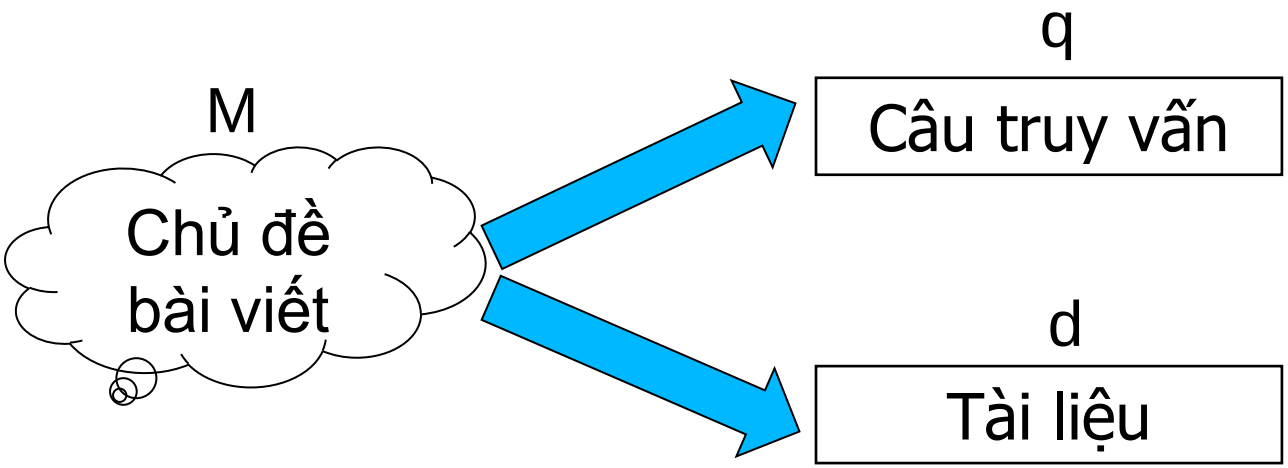


MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Tài liệu thỏa truy vấn nếu nhu cầu thông tin liên quan đến chủ đề. Khi đó, sự liên quan giữa tài liệu và truy vấn có thể được tính theo xác suất sau:

$$rel(d, q) = p(q \mid M_d)$$



MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Cách tiếp cận language model dựa trên giả thiết:

- Việc lựa chọn từ để tạo truy vấn là có chủ đích
- Có nhiễu trong quá trình chọn từ.

→ Giải pháp truy xuất tài liệu theo mô hình ngôn ngữ:

- Xác định mô hình ngôn ngữ M_d cho từ tài liệu d.
- Ước lượng $p(q \mid M_d)$ cho từng tài liệu d.
- Xếp hạng tài liệu giảm dần theo $p(q|M_d)$.

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Giả sử truy vấn q = {tq_1, tq_2, … tq_m} và mô hình ngôn ngữ M_d, khi đó:

$p(q|M_d) = p(tq_1|M_d) * p(tq_2|M_d,tq_1) * p(tq_3|M_d,tq_1,tq_2) * \ldots * p(tq_m|M_d,tq_1,tq_2,\ldots,tq_{m-1})$

Với M_d là mô hình ngôn ngữ Unigram:

$p(q|M_d) = p(tq_1|M_d) * p(tq_2|M_d) * p(tq_3|M_d) * \ldots * p(tq_m|M_d)$

Với M_d là mô hình ngôn ngữ Bigram:

$p(q|M_d) = p(tq_1|M_d) * p(tq_2|M_d,tq_1) * p(tq_3|M_d,tq_2) * \ldots * p(tq_m|M_d,tq_{m-1})$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Ví dụ:

M_1:

w	I	love	this	fun	film	…
p(w)	0.1	0.1	0.01	0.05	0.1	…

M_2:

w	I	love	this	fun	film	…
p(w)	0.2	0.001	0.01	0.005	0.1	…

q: I love this fun film

$p(q|M_1) = 0.1 * 0.1 * 0.01 * 0.05 * 0.1$

$p(q|M_2) = 0.2 * 0.001 * 0.01 * 0.005 * 0.1$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Phương pháp ước lượng mô hình ngôn ngữ:

Giả sử:

- Tập từ vựng V = $\{w_1, w_2, \ldots, w_n\}$
- Ngữ liệu D
- Ước lượng các xác suất $p(w_{ti}|w_{ti-k+1}\ldots w_{ti-1})$, với k tương ứng với mô hình k-gram theo phương pháp maximum likelihood:

$$p(w) \approx c(w)/N$$

$$p(w_{ti}|w_{ti-k+1}\ldots w_{ti-1}) \approx c(w_{ti-k+1}\ldots w_{ti})/c(w_{ti-k+1}\ldots w_{ti-1})$$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Các trường hợp chưa xuất hiện trong ngữ liệu huấn luyện → xác suất là 0, không so sánh được.

Kỹ thuật smoothing:

- Tránh trường hợp không tính được xác suất do:
  - Có những gram không xuất hiện khi ước lượng.
  - Xuất hiện từ mới → thêm từ đại diện vào tập V.
- Tính toán lại kết quả ước lượng.

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Smoothing với công thức Laplace:

- Giả sử tất cả các gram đều xuất hiện → tần số thấp nhất là 1
- Tính toán lại xác suất của các gram khi huấn luyện:

$$p(A) = (c(A)+1)/(N+|V|)$$

$$p(A|B) = (c(A,B)+1)/(c(B)+|V|)$$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Ví dụ: Cho các tài liệu sau

d_1 chinese beijing chinese

d_2 chinese chinese shanghai

d_3 chinese macao

d_4 tokyo japan chinese

Cho truy vấn:

q chinese chinese chinese tokyo japan

Xử lý truy vấn q với tập tài liệu đã cho theo mô hình ngôn ngữ Uni-gram

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Smoothing với công thức Backoff

Tính toán từ kết quả huấn luyện n-gram theo nguyên tắc:

- $p(w_{ti}|w_{ti-k+1}\dots w_{ti-1})$ nếu $p(w_{ti}|w_{ti-k+1}\dots w_{ti-1})$ tính toán được, ngược lại:
- $p(w_{ti}|w_{ti-k+2}\dots w_{ti-1})$ nếu $p(w_{ti}|w_{ti-k+2}\dots w_{ti-1})$ tính toán được, ngược lại:
- …
- $p(w_{ti})$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Smoothing với công thức nội suy (interpolation):

Tính toán từ kết quả huấn luyện n-gram theo nguyên tắc tổng hợp các xác suất của các k-gram:

- Trường hợp đơn giản

$$p(w_{ti}|w_{ti-k+1}\dots w_{ti-1}) = \lambda_1 * p(w_{ti}|w_{ti-k+1}\dots w_{ti-1}) + \lambda_2 * p(w_{ti}|w_{ti-k+2}\dots w_{ti-1}) + \dots + \lambda_n * p(w_{ti})$$

Trong đó, $\Sigma\lambda_i = 1$.

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Smoothing với công thức nội suy (interpolation):

- Trường hợp sử dụng ngữ cảnh:

$$p(w_{ti}|w_{ti-k+1}\dots w_{ti-1}) = \lambda_{1(c)}*p(w_{ti}|w_{ti-k+1}\dots w_{ti-1}) + \lambda_{2(c)}*p(w_{ti}|w_{ti-k+2}\dots w_{ti-1}) + \dots + \lambda_{n(c)}*p(w_{ti})$$

Trong đó, $\Sigma\lambda_{i(c)} = 1$.

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Smoothing với công thức nội suy (interpolation):

Ước lượng các giá trị λ_l bằng cách:

- Chia bộ dữ liệu huấn luyện thành 2 phần: huấn luyện và thử
- Xác định giá trị λ_l sao cho giá trị của hàm sau là lớn nhất trong bộ dữ liệu thử:

$$p(w_{t1},w_{t2}, \ldots, w_{tn}|\lambda_1,\lambda_2,\ldots,\lambda_k)$$

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Ví dụ: Áp dụng mô hình ngôn ngữ Bi-gram, xác định giá trị λ_1 và λ_2. Cho ngữ liệu huấn luyện là

football football team hockey

Dữ liệu thử là:

football hockey team

MÔ HÌNH XÁC SUẤT

❖LANGUAGE MODEL

Bài tập: Cho các tài liệu sau

d_1 chinese beijing chinese

d_2 chinese chinese shanghai

d_3 chinese macao

d_4 tokyo japan chinese

Cho truy vấn: *q chinese chinese chinese tokyo japan*

Áp dụng mô hình ngôn ngữ Bi-gram, sử dụng công thức smoothing backoff.

1) Lập chỉ mục cho các tài liệu đã cho.
2) Xử lý truy vấn q và xếp hạng các tài liệu.