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

❖CÁCH TIẾP CẬN

- Truy vấn thông tin là một quá trình không chắc chắn
  - Ngữ nghĩa câu truy vấn
  - Tài liệu thỏa truy vấn.
- Lý thuyết xác suất
  - Cở sở của suy luận không chắc chắn
  - Ước lượng khả năng tài liệu liên quan đến truy vấn

MÔ HÌNH XÁC SUẤT

❖CÁCH TIẾP CẬN

- Các mô hình xác suất
    - Các mô hình điển hình: BIM, Two Poisson, BM11, BM25
    - Mô hình ngôn ngữ
    - Mạng Bayes (Bayesian networks)

Các mô hình xác suất là những mô hình đã cũ nhưng có hiệu quả cao.

MÔ HÌNH XÁC SUẤT

❖NGUYÊN LÝ XẾP HẠNG

- Bài toán: cho một truy vấn ***q*** và một tập tài liệu ***D***, xác định thứ tự các tài liệu trong ***D*** theo truy vấn ***q***.
- Sự liên quan giữa tài liệu và truy vấn được thể hiện bằng một biến ngẫu nhiên nhị phân ***R***
  - $R_{d,q}$=1 nếu tài liệu ***d*** liên quan với ***q***
  - $R_{d,q}$=0 nếu tài liệu ***d*** không liên quan với ***q***
- Kết quả xếp hạng được sắp xếp giảm dần theo xác suất của sự liên quan $p(R_{d,q}=1)$ hay $p(R=1|d,q)$

MÔ HÌNH XÁC SUẤT

❖NGUYÊN LÝ XẾP HẠNG

- Các mô hình vấn thông tin theo xác suất cần giải quyết:
  - Xác định các giá trị ước lượng tốt nhất.
  - Phương pháp tính toán xác suất liên quan giữa tài liệu ***d*** và truy vấn ***q***

MÔ HÌNH XÁC SUẤT

❖PHÁT BIỂU BÀI TOÁN

- Tài liệu d = $\{td_1, td_2, \ldots td_m\}$
  - td_i là term thứ i của tài liệu d
  - $p(td_i=$*'từ'*$)$ là xác suất term thứ i của tài liệu d có giá trị là *'từ'*
- Truy vấn q = $\{tq_1, tq_2, \ldots tq_n\}$
  - tq_i là term thứ i của truy vấn q
  - $p(tq_i=$*'từ'*$)$ là xác suất term thứ i của truy vấn q có giá trị là *'từ'*
- Sự liên quan $R \in \{0,1\}$

MÔ HÌNH XÁC SUẤT

❖PHÁT BIỂU BÀI TOÁN

- Xác suất tài liệu ***d*** có liên quan với truy vấn ***q*** là

$$p(R=1|d, q)$$

MÔ HÌNH XÁC SUẤT

❖MỘT SỐ CÔNG THỨC XÁC SUẤT

- Xác suất của hai sự kiện A, B xảy ra đồng thời là p(A,B). Nếu A và B độc lập thì

$$p(A,B) = p(A) * p(B)$$

- Xác suất có điều kiện P(A|B) là xác suất sự kiện A nếu trước đó có sự kiện B xảy ra.
- $p(A,B,C) = p(A) * p(B|A) * p(C|A, B)$
- $p(A) = p(A,B) + p(A,\neg B)$
- $p(A) = p(A,B=b_1) + p(A,B=b_2) + \ldots + p(A,B=b_m)$

MÔ HÌNH XÁC SUẤT

❖MỘT SỐ CÔNG THỨC XÁC SUẤT

- Công thức Bayes:

$$p(A|B) = p(A) * p(B|A) / p(B)$$

$$p(A|B) = p(A) * p(B|A) / [\Sigma_{X\in\{A,\neg A\}} p(B|X)*p(X)]$$

- Tỉ lệ Odds:

$$O(A)= p(A) / p(\neg A) = p(A) / (1-p(A))$$

- log-odds:

$$\log(O(A))= \log(p(A)) - \log(1-p(A))$$

MÔ HÌNH XÁC SUẤT

❖XẾP HẠNG THEO MÔ HÌNH XÁC SUẤT

- Giả thiết: sự liên quan của các tài liệu với một câu truy vấn là độc lập.
- Ước lượng xác suất của sự liên quan p(R=1|d, q) theo:
  - Tần suất của term
  - Tần suất của tài liệu
  - Độ dài của văn bản
- Đặt r là R=1 và ¬r là R=0: các tài liệu được xếp hạng theo thứ tự giá trị p(r|d, q) giảm dần.
- Thay vì xếp hạng theo giá trị p(r|d,q), có thể xếp hạng theo giá trị odds(p(r|d,q)).

MÔ HÌNH XÁC SUẤT

❖XẾP HẠNG THEO MÔ HÌNH XÁC SUẤT

- Thay vì xếp hạng theo giá trị p(r|d,q), có thể xếp hạng theo giá trị odds(p(r|d,q)).

→ Xếp hạng theo giá trị

$$p(d|r,q)/p(d|\neg r,q)$$

Giá trị p(d|r,q) và p(d|¬r,q) được ước lượng tùy theo mô hình.

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Tần số của mỗi term là 0 (không xuất hiện) và 1 (có xuất hiện)

1) Giả thiết độc lập:

- Các term trong một tài liệu và một truy vấn độc lập với nhau
- Xác suất của một term xuất hiện trong các tài liệu liên quan không ảnh hưởng đến xác suất của các term khác trong các tài liệu liên quan

MÔ HÌNH XÁC SUẤT

❖BOOLEAN INDEPENDENCE MODEL (BIM)

1) Giả thiết độc lập:

- Giả thiết này đơn giản hóa việc tính toán và khá hiệu quả.

$$p(d|q,r) = \prod_{i\in[1,m]} p(td_i|q,r)$$

$$p(d|q,\neg r) = \prod_{i\in[1,m]} p(td_i|q,\neg r)$$

MÔ HÌNH XÁC SUẤT

❖BOOLEAN INDEPENDENCE MODEL (BIM)

2) Các term của câu truy vấn là yếu tố duy nhất xác định sự liên quan giữa tài liệu và truy vấn:

- Với term $td_i \notin q$ thì $p(td_i|q,r)$ không phụ thuộc vào r:

$$p(td_i|q,r) = p(td_i|q,\neg r)$$

→ Chỉ cần tính xác suất các term trong truy vấn q

$$p(d|q,r) = \prod_{td \in q} p(td|q,r)$$

$$p(d|q,\neg r) = \prod_{td \in q} p(td|q,\neg r)$$

MÔ HÌNH XÁC SUẤT

❖BOOLEAN INDEPENDENCE MODEL (BIM)

Dựa trên hai giả thiết trên, sự liên quan được thể hiện qua giá trị:

$$\prod_{td \in q} p(td|r,q)/p(td|\neg r,q)$$

Việc ước lượng giá trị p(td|r,q) và p(td|¬r,q) được thực hiện theo hai trường hợp:

- Trường hợp không có ngữ liệu mẫu
- Trường hợp có ngữ liệu mẫu

MÔ HÌNH XÁC SUẤT

- **BOOLEAN INDEPENDENCE MODEL (BIM)**
- Trường hợp không có ngữ liệu mẫu

Sử dụng hai giả thiết:

- Term của truy vấn xuất hiện hay không xuất hiện trong tài liệu liên quan là như nhau:

$$p(tq_i|r,q)=0.5$$

- Xác suất term xuất hiện trong tài liệu không liên quan (N_{td}: số tài liệu chứa td, N: tổng số tài liệu)

$$p(td|\neg r,q)=N_{td}/N$$

MÔ HÌNH XÁC SUẤT

❖BOOLEAN INDEPENDENCE MODEL (BIM)

❖ Trường hợp không có ngữ liệu mẫu

- Độ liên quan giữa tài liệu và truy vấn:

$$rel(d,q) = \prod_{td \in q} p(td|r,q)/p(td|\neg r,q)$$

$$= \prod_{td \in q} 0.5*N/N_{td}$$

Sử dụng độ đo log-odds:

$$rel(d,q) = \Sigma_{td \in q} \log(0.5*N/N_{td})$$

→ Trọng số của mỗi term là $w_{td}=\log(0.5*N/N_{td})$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Ví dụ: Tính độ liên quan giữa truy vấn q và các tài liệu sau theo mô hình BIM

d_1 Romeo and Juliet

d_2 Juliet: Oh happy dagger

d_3 Romeo died by dagger

q: die dagger

MÔ HÌNH XÁC SUẤT

❖BOOLEAN INDEPENDENCE MODEL (BIM)

	Romeo	Juliet	happy	dagger	die
$d(td\|r,q)$	0.5	0.5	0.5	0.5	0.5
$d(td\|\neg r,q)$	2/3	2/3	1/3	2/3	1/3
w_{td}	-0.41	-0.41	0.58	-0.41	0.58

$rel(d_1,q)=?$

$rel(d_2,q)=-0.41$

$rel(d_3,q)=0.17$

MÔ HÌNH XÁC SUẤT

- **BOOLEAN INDEPENDENCE MODEL (BIM)**
- Trường hợp có ngữ liệu mẫu

- r_{td} là số tài liệu liên quan chứa term td
- N_R là tổng số tài liệu liên quan
- Ước lượng các xác suất

$$p(td|r,q) = r_{td}/N_R$$

$$p(td|\neg r,q) = (N_{td}-r_{td})/(N-N_R)$$

MÔ HÌNH XÁC SUẤT

- **BOOLEAN INDEPENDENCE MODEL (BIM)**
- Trường hợp có ngữ liệu mẫu

- Để tránh trường hợp r_{td}=0 và $r_{td}=N_{td}$, thực hiện smoothing:

$$p(td|r,q) = (r_{td}+0.5)/(N_R+1)$$

$$p(td|\neg r,q) = (N_{td}-r_{td}+0.5)/(N-N_R+1)$$

- Độ liên quan:

rel(d,q)

$$= \Sigma_{td\in q}\log([(r_{td}+0.5)*(N-N_R+1)]/[(N_{td}-r_{td}+0.5)*(N_R+1)])$$

MÔ HÌNH XÁC SUẤT

- **BOOLEAN INDEPENDENCE MODEL (BIM)**
- Trường hợp có ngữ liệu mẫu

- Trọng số của mỗi term:

$$w_{td}=\log([(r_{td}+0.5)*(N-N_R+1)]/[(N_{td}-r_{td}+0.5)*(N_R+1)])$$

MÔ HÌNH XÁC SUẤT

❖ BOOLEAN INDEPENDENCE MODEL (BIM)

Ví dụ: Tính độ liên quan giữa truy vấn q và các tài liệu sau theo mô hình BIM. Biết N=30, N_R=6, r_{die}=3, r_{dagger}=4, N_{die}=15, N_{dagger}=16.

d_1 Romeo and Juliet

d_2 Juliet: Oh happy dagger

d_3 Romeo died by dagger

q: die dagger

MÔ HÌNH XÁC SUẤT

❖TWO POISSON MODEL

Tần số của term là số lần xuất hiện của term trong tài liệu

→ Xác suất term td xuất hiện k lần trong tài liệu d là p(td=k|d)

Để ước lượng xác suất p(td=f|d,r), giả thiết td tuân theo quy luật phân phối Poisson, khi đó:

$$p(td=k|d) = \lambda^{k} * e^{-\lambda}/k!,\ k=0,1,2,..$$

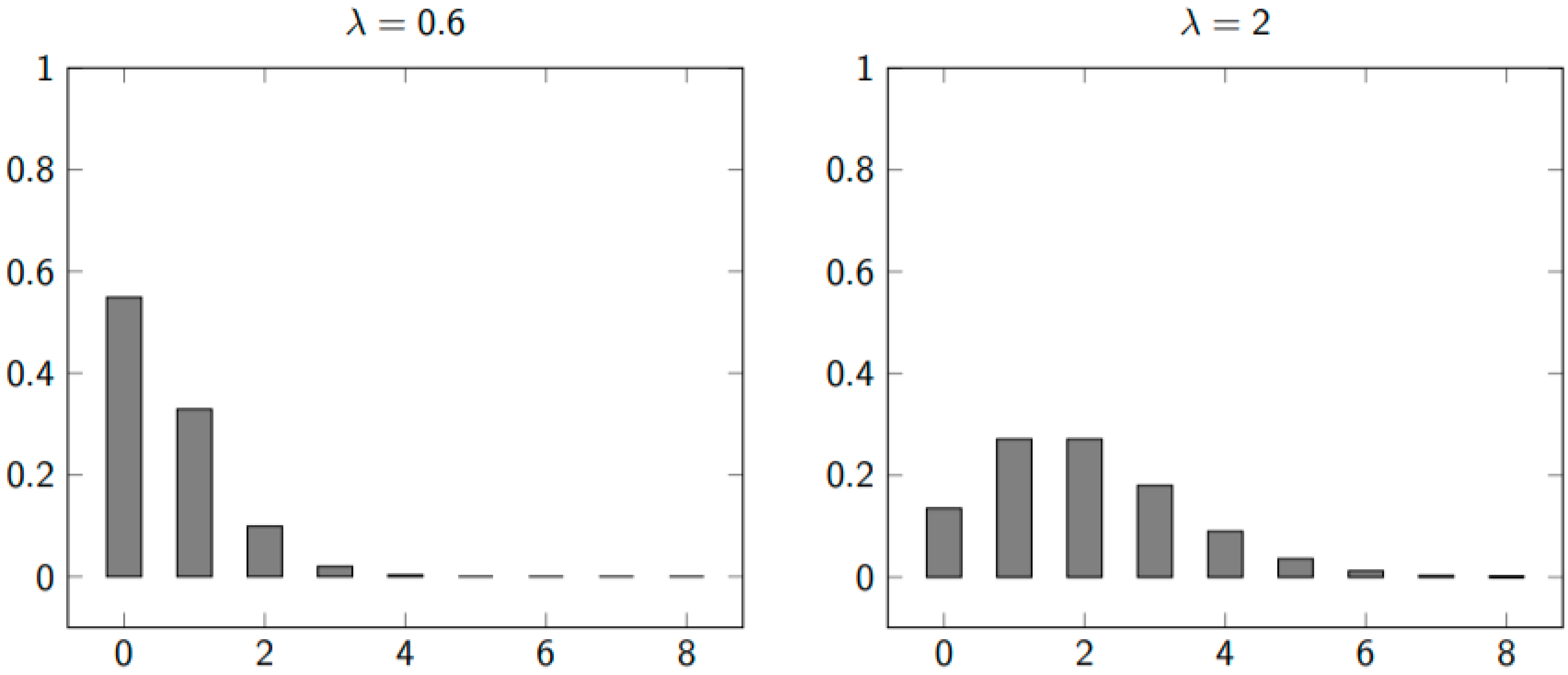
Giá trị λ được ước lượng như sau

$$\lambda=count(td)/N_{term}$$

MÔ HÌNH XÁC SUẤT

❖TWO POISSON MODEL

Ví dụ phân phối xác suất Poisson



MÔ HÌNH XÁC SUẤT

❖TWO POISSON MODEL

Tuy nhiên, một quy luật Poisson không thể hiện đúng thực tế nên sử dụng hai quy luật phân phối Poisson, gọi là Elite (E) và non-Elite($\neg$E):

$p(td=k|r,d)=u*\lambda^{k}*e^{-\lambda}/k! + (1-u)*\mu^{k}*e^{-\mu}/k!, k=0,1,2,..$

$p(td=k|\neg r,d)=v*\lambda^{k}*e^{-\lambda}/k! + (1-v)*\mu^{k}*e^{-\mu}/k!, k=0,1,2,..$

Trong đó:

- u là xác suất tài liệu là Elite và có liên quan
- v là xác suất tài liệu là Elite và không liên quan

MÔ HÌNH XÁC SUẤT

❖TWO POISSON MODEL

Vì vấn đề ước lượng các tham số u, v, λ và μ, Robertson và Walker đề xuất một cách xấp xỉ theo hình dạng của hàm tính trọng số term sao cho:

- $w_{td} = 0$ nếu $k=0$
- w_{td} đồng biến với k
- w_{td} có dạng log-odds

Công thức đề nghị: $w'_{td} = [k/(k_1+k)]*w_{td}$

Với w_{td} là trọng số được tính như mô hình BIM