



IT4853

Tìm kiếm và trình diễn thông tin

Bài 6. Mô hình ngôn ngữ

IIR.C12. Language models for information retrieval

TS. Nguyễn Bá Ngọc, *Bộ môn Hệ thống thông tin,
Viện CNTT & TT*
ngocnb@soict.hust.edu.vn



Nội dung chính

- Mô hình sinh
- Ước lượng xác suất
- Thử nghiệm

Mô hình sinh dựa trên máy trạng thái hữu hạn

- Đối với mô hình sinh truyền thống, mỗi khi chuyển trạng thái máy trạng thái hữu hạn sẽ sinh một từ. Ở mỗi trạng thái khác nhau máy có thể sinh các từ khác nhau.
- Tập hợp tất cả các văn bản có thể được sinh bởi máy trạng thái hữu hạn gọi là ngôn ngữ của máy trạng thái hữu hạn đó.
- Ví dụ máy trạng thái hữu hạn sau:



- Có thể sinh các văn bản: I wish I wish I wish I wish ...
- Không thể sinh: "wish I wish" hoặc "I wish I".

Thuật ngữ:

Máy trạng thái hữu hạn: finite automaton



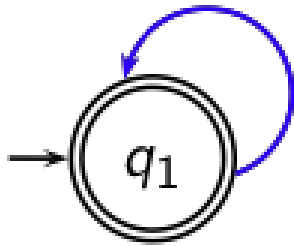
Mô hình ngôn ngữ

- Mô hình ngôn ngữ là một mô hình mở rộng của mô hình sinh truyền thống, bổ xung thêm bảng phân bố xác suất sinh từ thuộc bộ từ vựng cho mỗi trạng thái.
- Cách xếp hạng theo mô hình ngôn ngữ:
 - Thiết lập mô hình ngôn ngữ cho mỗi văn bản;
 - Xếp hạng văn bản theo thứ tự giảm dần xác suất sinh truy vấn của mô hình ngôn ngữ tương ứng.

Ví dụ

mô hình ngôn ngữ dựa trên máy

trạng thái hữu hạn một trạng thái



w	$P(w q_1)$	w	$P(w q_1)$
STOP	0.2	toad	0.01
the	0.2	said	0.03
a	0.1	likes	0.02
frog	0.01	that	0.04
	

string = "frog said that toad likes frog" STOP

Xác suất sinh chuỗi string là:

$$P(\text{string}) = 0.01 \times 0.03 \times 0.04 \times 0.01 \times 0.02 \times 0.01 \times 0.2 = 0.00000000000048$$

Trong đó STOP là trạng thái dừng.

Ví dụ xếp hạng văn bản

Mô hình ngôn ngữ của d_1

w	$P(w .)$	w	$P(w .)$
STOP	.2	toad	.01
the	.2	said	.03
a	.1	likes	.02
frog	.01	that	.04
	

Mô hình ngôn ngữ của d_2

w	$P(w .)$	w	$P(w .)$
STOP	.2	toad	.02
the	.15	said	.03
a	.08	likes	.02
frog	.01	that	.05
	

- string = "frog said that toad likes frog" STOP
 - $P(\text{string}|Md_1) = 0.01 \times 0.03 \times 0.04 \times 0.01 \times 0.02 \times 0.01 \times 0.2 = 0.00000000000048 = 4.8 \times 10^{-12}$
 - $P(\text{string}|Md_2) = 0.01 \times 0.03 \times 0.05 \times 0.02 \times 0.02 \times 0.01 \times 0.2 = 0.00000000000120 = 12 \times 10^{-12}$
- $P(\text{string}|Md_2) > P(\text{string}|Md_1)$
- Thứ tự xếp hạng: d_2 d_1



Nội dung chính

- Mô hình sinh
- Ước lượng xác suất
- Thử nghiệm



Giả thuyết Unigram

- Xác suất sinh một từ bất kỳ là độc lập với xác suất sinh các từ còn lại:

$$P_{\text{uni}}(t_1 t_2 t_3 t_4) = P(t_1)P(t_2)P(t_3)P(t_4)$$



Giả thuyết phân bố đa thức

- Giả thuyết phân bố đa thức:

$$P(d) = \frac{L_d!}{\text{tf}_{t_1,d}! \text{tf}_{t_2,d}! \cdots \text{tf}_{t_M,d}!} P(t_1)^{\text{tf}_{t_1,d}} P(t_2)^{\text{tf}_{t_2,d}} \cdots P(t_M)^{\text{tf}_{t_M,d}}$$



Xác suất văn bản sinh truy vấn

- Theo luật Bayes

$$P(d|q) = P(q|d)P(d) / P(q)$$

- $P(q)$ là hằng số;
- Giả sử $P(d)$ là đồng nhất đối với tất cả văn bản;
 - Có thể xếp hạng theo $P(q|d)$: Xác suất sinh truy vấn.

Văn bản thường dài hơn so với truy vấn cho nên cũng thuận tiện hơn khi sử dụng để tính các đại lượng xác suất.

Kết hợp giả thuyết unigram và giả thuyết đa thức

$$P(q|M_d) = K_q \prod_{t \in V} P(t|M_d)^{tf_{t,d}}$$

$$K_q = \frac{L_q!}{tf_{t_1,q}! tf_{t_2,q}! \dots tf_{t_M,q}!}$$

Đại lượng cần tính

K_q là hệ số đa thức: là hằng số với một câu truy vấn q xác định, có thể bỏ qua trong xếp hạng.



Đại lượng kết quả tìm kiếm

- Hàm đại lượng kết quả tìm kiếm:

$$RSV(d, q) = \prod_{t \in q} p(t|M_d)$$

$$RSV(d, q) = \prod_{t \text{ duy nhất } \in q} p(t|M_d)^{tf_{t,q}}$$

- Trong đó:

$$p(t|M_d) = \frac{tf_{t,d}}{L_d}$$

Nếu d không chứa một từ truy vấn t thì $RSV(d, q) = 0$
 \Rightarrow Cần làm mịn để tránh giá trị 0.

Thuật ngữ:

Ước lượng khả năng cực đại: Maximum likelihood estimation.



Mô hình ngôn ngữ của bộ dữ liệu

- Tương tự văn bản, xác suất mô hình của bộ dữ liệu sinh từ t được xác định theo công thức sau:

$$p(t|M_C) = \frac{cf_{t,C}}{L_C}$$

- Trong đó:
 - M_C là mô hình ngôn ngữ của bộ dữ liệu C ;
 - $L_C = \sum_{d \in C} L_d$, là số từ trong bộ dữ liệu.



Làm mịn tuyến tính

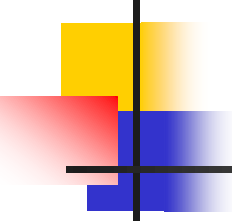
- Kết hợp văn bản và bộ dữ liệu

$$p(t|d) = \lambda p(t|M_d) + (1 - \lambda) p(t|M_c)$$

$$w_{t,d} = \lambda \frac{tf_{t,d}}{L_d} + (1 - \lambda) \frac{cf_{t,c}}{L_c}$$

Thuật ngữ:

Làm mịn tuyến tính: Linear interpolation



Đại lượng kết quả tìm kiếm sau khi làm mịn

$$RSV(q, d) = \prod_{t \in q} \left(\lambda \frac{t f_{t,d}}{L_d} + (1 - \lambda) \frac{c f_{t,c}}{L_c} \right)$$

Các giả thuyết đã sử dụng: Giả thuyết Unigram; phân bố đa thức; làm mịn tuyến tính; khả năng cực đại.



Giá trị tham số

- Sử dụng λ lớn có xu hướng trả về văn bản chứa tất cả từ truy vấn
 - Hiệu ứng sử dụng điều kiện AND
- Giá trị λ nhỏ thích hợp cho xử lý truy vấn dài
 - Hiệu ứng sử dụng điều kiện OR
- Cần tùy chỉnh λ để đạt được chất lượng cao.



Giả thuyết mô hình ngôn ngữ

- Người dùng có những hình dung nhất định về văn bản cần tìm. Chính mô hình văn bản trong tưởng tượng đó đã làm nảy sinh câu truy vấn.
- Xác suất $p(q|d)$ thể hiện khả năng văn bản d chính là văn bản trong tưởng tượng của người dùng.



Nội dung chính

- Mô hình sinh
- Ước lượng xác suất
- Thử nghiệm

Thử nghiệm của Ponte và Croft

Rec.	precision			significant?
	tf-idf	LM	%chg	
0.0	0.7439	0.7590	+2.0	
0.1	0.4521	0.4910	+8.6	
0.2	0.3514	0.4045	+15.1	*
0.4	0.2093	0.2572	+22.9	*
0.6	0.1024	0.1405	+37.1	*
0.8	0.0160	0.0432	+169.6	*
1.0	0.0028	0.0050	+76.9	
11-point average	0.1868	0.2233	+19.6	*

- Mô hình ngôn ngữ trả về kết quả tốt hơn so với VSM trong thử nghiệm này...
- ...Tuy nhiên chưa hoàn toàn thay thế được VSM trong thực tế



Ví dụ 1

- Bộ dữ liệu: d_1 và d_2
- d_1 : Jackson was one of the most talented entertainers of all time
- d_2 : Michael Jackson anointed himself King of Pop
- Truy vấn q : Michael Jackson
- Xếp hạng văn bản theo mô hình ngôn ngữ, sử dụng làm mịn tuyến tính với $\lambda = 0.5$



Ví dụ 1

- $RSV(q, d_1) = [(0/11 + 1/18)/2] \times [(1/11 + 2/18)/2] \approx 0.003$
- $RSV(q, d_2) = [(1/7 + 1/18)/2] \times [(1/7 + 2/18)/2] \approx 0.013$

d_2 được xếp hạng cao hơn d_1



Ví dụ 2

- Bộ dữ liệu: d_1 và d_2
- d_1 : Xerox reports a profit but revenue is down
- d_2 : Lucene narrows quarter loss but decreases further
- Truy vấn q: revenue down
- Xếp hạng văn bản theo mô hình ngôn ngữ, sử dụng làm mịn tuyến tính với $\lambda = 0.5$



Ví dụ 2

- $P(q|d_1) = [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2]$
- $P(q|d_1) = 1/8 \times 3/32 = 3/256$
- $P(q|d_2) = [(1/8 + 2/16)/2] \times [(0/8 + 1/16)/2] =$
- $P(q|d_2) = 1/8 \times 1/32 = 1/256$
- Xếp hạng d_2 cao hơn d_1



Bài tập 6.1

Hãy xây dựng mô hình ngôn ngữ cho văn bản sau sử dụng giả thuyết Unigram và MLE:

“The martian has landed on the latin pop sensation Ricky Martin”



Bài tập 6.2

Hãy viết cho mỗi đại lượng sau một câu mô tả về ảnh hưởng của những đại lượng này đến xếp hạng văn bản theo mô hình ngôn ngữ:

- a) Tần suất từ trong văn bản;
- b) Tần suất bộ dữ liệu của từ;
- c) Tần suất văn bản của từ;
- d) Chuẩn hóa theo độ dài;



Bài tập 6.3

Cho bộ dữ liệu văn bản:

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

Cho hệ số làm mịn = 0.5, hãy điền vào bảng phân bố xác suất sau:

Query	Doc 1	Doc 2	Doc 3	Doc 4
click				
shears				
click shears				

Xếp hạng các văn bản cho câu truy vấn: click shears

