MÔ HÌNH CHỦ ĐỀ VÀ ỨNG DỤNG TRONG TÌM KIẾM THÔNG TIN

Hà Thị Thanh*, Trịnh Thị Thủy, Ngô Cẩm Tú
Trường Đại học Công nghệ thông tin và Truyền thông – ĐH Thái Nguyên

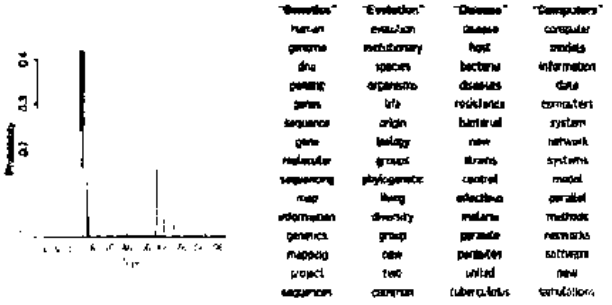
TÓM TẮT

Tất cả thông tin của chúng ta hiện nay được số hóa và lưu trữ dưới nhiều dạng. Tin tức, blog, web pages, bài báo khoa học, sách, hình ảnh, âm thanh, video, mạng xã hội với một lượng lưu trữ lớn. Vì thế tìm kiếm sẽ khó khăn. Ta cần một công cụ tính toán để giúp tổ chức, tìm kiếm, khai phá lượng lớn thông tin đó. Những nhà nghiên cứu về học máy đã phát triển mô hình chủ đề xác suất, một thuật toán thích hợp hướng vào khai phá và giải thích kho dữ liệu văn bản lớn với những thông tin về chủ đề. Thuật toán mô hình chủ đề là phương pháp thống kê mà phân tích các từ của văn bản gốc để khám phá ra các chủ đề của văn bản, cách thức mà các chủ đề liên kết với nhau, cách thức mà các chủ đề đó thay đổi theo thời gian. Mô hình chủ đề cho phép chúng ta tổ chức và tóm tắt kho dữ liệu số. Từ đó giúp việc khai thác và tìm kiếm thông tin nhanh chóng hơn.

Từ khóa. *Tìm kiếm thông tin, học máy thống kê, mô hình chủ đề, suy diễn, mô hình LDA*

GIỚI THIỆU

Chủ đề là tập hợp các từ có mối quan hệ ngữ nghĩa với nhau [3]. Cụ thể: Chủ đề là phân phối qua một tập từ vựng cố định. Mỗi chủ đề khác nhau thì có phân phối khác nhau qua cùng bộ từ vựng. Ví dụ (Hình 1)



Hình 1: *Theo Blei, chọn ra mỗi chủ đề 15 từ với xác suất lớn nhất* [3]

Các chủ đề được đặc tả trước khi dữ liệu được sinh ra. Mặt khác mô hình xác suất giả thiết rằng các chủ đề được sinh ra đầu tiên trước khi có văn bản.

Mô hình chủ đề là một phương pháp phân tích văn bản được rất nhiều học giả quan tâm trong lĩnh vực khoa học xã hội, nhân văn... Mô hình chủ đề cung cấp thuật toán tự động mã hóa nội dung của tập văn bản sang một tập mã có ý nghĩa gọi là "chủ đề". Các nhà nghiên cứu bắt đầu với việc xác định số lượng chủ đề cho thuật toán. Chương trình sẽ định nghĩa số lượng các chủ đề và trả lại xác suất của từ trong mỗi chủ đề.

Mô hình chủ đề đang được quan tâm vì một trong các lý do sau. Thứ nhất, dùng mô hình chủ đề sẽ khai phá ra các chủ đề, sau đó tổ chức lại tập dữ liệu theo chủ đề đã khám phá. Thứ hai, nó được áp dụng cho tập dữ liệu lớn. Thứ ba, mô hình chủ đề được áp dụng cho nhiều loại dữ liệu. Từ dữ liệu có cấu trúc hoặc không có cấu trúc, video, hình ảnh, âm thanh... Đối với dữ liệu văn bản, mô hình chủ đề được khai thác rất nhiều. Đặc biệt trong lĩnh vực tìm kiếm thông tin. Mô hình chủ đề có thể dùng để biểu diễn văn bản qua các chủ đề, phân lớp văn bản, hoặc dùng để xếp hạng văn bản.

MÔ HÌNH CHỦ ĐỀ LDA

LDA là mô hình phát triển từ mô hình pLSI. pLSI là mô hình chủ đề dùng để phân tích ngữ nghĩa ẩn [2].

Ý tưởng cơ bản của LDA là tập các văn bản chứa nhiều chủ đề. LDA là mô hình thống kê của tập văn bản. Nó được mô tả dễ dàng nhất bằng quá trình sinh.

Mô hình LDA là mô hình sinh. Mô hình sinh dùng để giải thích tập dữ liệu quan sát được qua nhóm dữ liệu không quan sát được. Ví dụ, nếu dữ liệu quan sát được là các từ trong văn bản, nó cho rằng mỗi văn bản là hỗn hợp của số lượng nhỏ các chủ đề và việc mỗi từ tạo ra tương ứng nằm trong một chủ đề trong văn bản. LDA cũng là mô hình chủ đề được

* *Tel: 0982 266009, Email: hthanh@ictu.edu.vn*

Blei, Andrew và Michael Jordan đưa ra vào năm 2003.

Một chủ đề là một phân phối qua tập từ vựng cố định. Ví dụ chủ đề "*genetics*" có chứa các từ về di truyền học với xác suất cao và chủ đề *"evolutionary biology"* có các từ về sự tiến hóa của sinh vật học với xác suất cao. Chúng ta giả thiết rằng những chủ đề này được ghi rõ trước khi dữ liệu được sinh ra. Các từ trong văn bản được sinh ra theo quá trình sau: (Hình 2)

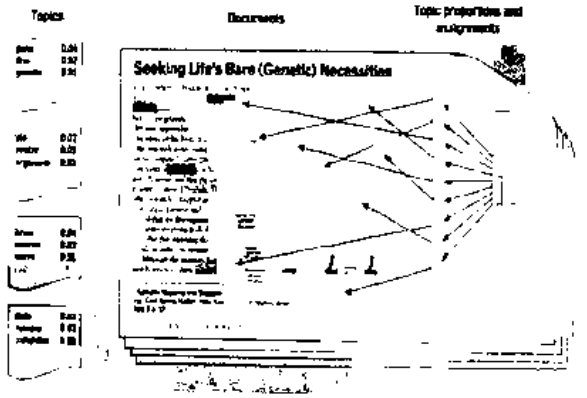
Chọn ngẫu nhiên một phân phối qua các chủ đề.

Với mỗi từ trong văn bản:

+ Chọn ngẫu nhiên một chủ đề từ phân phối qua chủ đề.

+ Chọn ngẫu nhiên một từ từ phân phối tương ứng qua bộ từ vựng

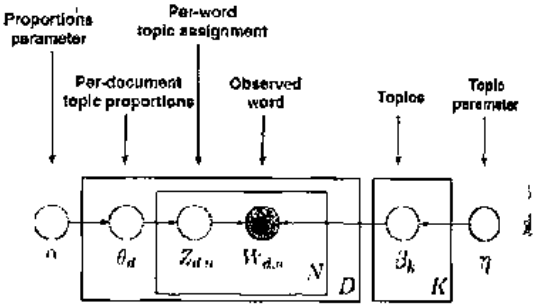
Mô hình thống kê này chỉ ra rằng trong các văn bản có nhiều chủ đề Trong mỗi văn bản có nhiều chủ đề với tỉ lệ khác nhau. Mỗi từ trong mỗi văn bản được rút ra từ một trong các chủ đề. Đây là điểm khác biệt của LDA. Tất cả các văn bản trong tập văn bản dùng chung một tập các chủ đề, nhưng mỗi một văn bản có tỉ lệ chủ đề khác nhau



Hình 2. *Mô hình sinh văn bản từ các chủ đề cho trước* [3]

Như đã nói trong phần giới thiệu, mô hình chủ đề tự động khai phá các chủ đề từ một tập văn bản (thông qua phần suy diễn các biến ẩn, trình bày ở phần suy diễn). Trong các văn bản quan sát được thì cấu trúc của chủ đề bao gồm các chủ đề, phân phối các chủ đề trên văn bản, chủ đề được gán qua từ trên văn bản được gọi là cấu trúc ẩn. Bài toán tính toán của mô hình hóa chủ đề là sử dụng văn bản được quan sát để suy diễn ra cấu trúc của chủ đề ẩn. Lợi ích của mô hình chủ đề bắt nguồn từ tính chất mà cấu trúc ẩn được suy diễn tương tự với cấu trúc chủ đề của tập dữ liệu. Cấu trúc ẩn dùng để diễn giải cho mỗi văn bản trong tập dữ liệu và những diễn giải này được dùng để giúp cho các bài toán như tìm kiếm thông tin (IR), phân lớp, thăm dò tập dữ liệu (corpus exploration)

Sau đây là mô hình đồ thị của LDA[] do Blei đưa ra vào năm 2003 [1].



Hình 3. *Mô hình đồ thị của LDA*

Trong đó: W là các từ trong từ điển V. Với mỗi văn bản D_i có N_{Di} từ Tức là D=$(W_1, W_2, ..., W_{ND})$ Tập D tập gồm M văn bản, $D=(D_1, D_2, , D_M)$.

Các giả thiết của LDA theo quá trình sinh như sau

- Chọn $\theta \approx Dir(\alpha)$

- Với mỗi từ W_n trong văn bản chọn:

+ Chọn $Z_n \approx multinomial(\theta)$

+ Chọn từ W_n từ xác suất $P(W_n | Z_n, \beta)$ là phân phối xác suất multinomial có điều kiện trên Z_n

Trong đó:

- β là các chủ đề trong kho dữ liệu (là một ma trận cấp $k*V$, trong đó β_{ij} là xác suất từ thứ j trong chủ đề i). β_j là phân phối một chủ đề thứ j qua tập từ vựng V.

- θ_d là tỉ lệ các chủ đề trong văn bản thứ d. θ_{dk} là tỉ lệ chủ đề thứ k trong văn bản d.

- z_d là chủ đề được gán cho văn bản d, $z_{d,n}$ là chủ đề được gán của từ thứ n trong văn bản d.

- W_d là từ trong văn bản thứ d, $W_{d,n}$ là từ thứ n trong văn bản d.

LDA và các mô hình chủ đề khác là một phần trong *probabilistic modeling*. Trong mô hình xác suất nói chung, chúng ta coi như dữ liệu của chúng ta bắt nguồn từ quá trình sinh (bao gồm cả biến ẩn). Quá trình sinh được đinh nghĩa là phân phối xác suất đồng thời qua các biến được quan sát và biến ẩn. Chúng ta thực hiện phân tích dữ liệu bằng cách sử dụng phân phối đồng thời này để tính phân phối có điều kiện của các biến ẩn khi biết các biến được quan sát. Phân phối có điều kiện này gọi là phân phối hậu nghiệm *(posterior distribution)*. Các biến được quan sát là các từ của tập văn bản, biến ẩn là cấu trúc chủ đề. Bài toán tính toán suy diễn cấu trúc chủ đề ẩn từ văn bản là bài toán tính phân phối hậu nghiệm (phân phối có điều kiện của biến ẩn khi biết tập các văn bản).

Quá trình sinh của mô hình LDA tương ứng với công thức phân phối đồng thời của các biến quan sát được và các biến ẩn như sau:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})\right) \quad (1)$$

SUY DIỄN TRONG LDA

Tính phân phối có điều kiện của cấu trúc chủ đề được cho bởi các văn bản được quan sát Ta có posterial [3]:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \quad (2)$$

Tử số của công thức trên là phân phối đồng thời của các biến ngẫu nhiên, mẫu số là xác suất biên của các quan sát (nó là xác suất của các từ trong tập văn bản quan sát được). Về mặt lý thuyết, công thức xác suất này được tính bằng cách lấy tổng của tất cả phân phối đồng thời của tất cả các thể hiện của cấu trúc chủ đề ẩn.

Số lượng cấu trúc chủ đề là rất lớn (theo hàm mũ). Vì vậy nó khó tính toán (vì số lượng từ trong tập văn bản có thể lên đến hàng triệu) Do nhiều mô hình xác suất hiện đại quan trong, chúng ta không tính được posterior bởi vì mẫu số. Mục đích nghiên cứu của mô hình xác suất hiên đại là phát triển các phương pháp hiệu quả để xấp xỉ chúng.

Mục đích của các mô hình chủ đề là đi xấp xỉ công thức (2) bằng cách chấp nhận một phân phối thay thế qua cấu trúc chủ đề ẩn mà gần với posterior thực sự. Thuật toán mô hình hóa chủ đề nói chung chia làm hai thuật toán: sampling-based algorithms and variational algorithms

Sampling-based algorithms: Cố gắng thu thập mẫu từ posterior tới xấp xỉ nó với một phân phối theo kinh nhiệm. Thông thường sử dụng thuật toán *Gibbs sampling* (chúng ta đi xây dưng chuỗi Markov – biến ngẫu nhiên tuần tự, mỗi biến phụ thuộc vào biến trước, mà phân phối hạn chế của nó là posterior). Chuỗi Markov được định nghĩa trên biến chủ đề ẩn, thuật toán này chạy rất mất thời gian, mẫu thu thập từ phân phối giới hạn, sau đó xấp xỉ phân phối này với mẫu được thu thập (thường thì một mẫu được tạo ra khi một xấp xỉ của cấu trúc chủ đề ẩn với xác suất là lớn nhất).

Variational methods: (phương pháp biến phân)

Phương pháp biến phân thay thế cho thuật toán Sampling-based Phương pháp suy diễn biến phân tốt hơn so với phương pháp lấy mẫu trên. Phương pháp này đặt một họ các phân phối được tham số hóa qua cấu trúc chủ đề ẩn và sau đó tìm thành phần gần nhất với posterior trong họ phân phối đó. Vì thế bài toán suy diễn chuyển sang bài toán tối ưu Thuật toán suy diễn biến phân Coordinate ascent (Blei) trong LDA và thuật toán online (Hoffman) có thể dễ dàng thực hiện được bằng tay với hàng triệu văn bản và phù hợp với tập văn bản streaming của dữ liệu text.

Suy diễn biến phân quay trở về suy diễn hậu nghiệm trong tối ưu hóa. Ý tưởng chính là:

+ Thay thế bằng một phân phối qua các biến ẩn với các tham số tự do (free parameters), gọi là variational parameters.

+ Tối ưu hóa variational parameters để tạo ra một phân phối mà tiến gần đến hậu nghiệm đúng.

- Phương pháp suy diễn biến phân thường nhanh hơn phương pháp lấy mẫu (sampling-based approaches)

Suy diễn biến phân ngẫu nhiên Stochastic variational inference.

+ Đặt điều kiện lên tập dữ liệu lớn và xấp xỉ hậu nghiệm

+ Trong suy diễn biến phân, ta đi tối ưu hóa một họ phân phối để tìm thành viên gần nhất (in KL divergence - đo sự sai khác giữa hai phân phối P và Q) tới hậu nghiệm.

+ Suy diễn biến phân thường đưa về thuật toán như sau:

Phỏng đoán các tham số cục bộ cho mỗi điểm dữ liệu

Dựa vào suy diễn cục bộ này, phỏng đoán lại các tham số toàn cục

Cứ lặp lại như vậy

Cả hai thuật toán trên thực hiện tìm cấu trúc chủ đề. Một tập văn bản được giữ cố định và đảm nhiệm như một hướng dẫn tới nơi tìm kiếm. Phương pháp tiếp cận nào là tốt hơn còn phụ thuộc vào mô hình cụ thể.

ỨNG DỤNG

Sau khi mô hình chủ đề học ra các tham số ẩn (cấu trúc văn bản) thì có rất nhiều ứng dụng. Sau đây là một số ứng dụng trong tìm kiếm thông tin:

- Áp dụng vào bài toán phân cụm các câu truy vấn dựa vào các chủ đề đã tìm được.

- Biểu diễn văn bản mờ mức chủ đề.

- Có hai cách tiếp cận để tính độ tương tự giữa cấu truy vấn Q và văn bản d

+ Tính xác suất của truy vấn:

$$P(Q \mid d_i) = \prod_{w_k \in Q} P(w_k \mid d_i) \quad (3)$$

$$P(w_k \mid d_i) = \sum_{j=1}^{|T|} P(w_k \mid z_j) P(z_j \mid d_i) \quad (4)$$

+ So sánh các phân phối chủ đề trong văn bản (sử dụng độ đo sự sai khác: Kullback-Leibler, Jenson-Shannon):

$$D_{KL}(P \| Q) = \sum P(i) \log_2 \frac{P(i)}{Q(i)} \quad (5)$$

$$D_{JS}(P \| Q) = D_{KL}(P \| Q) + D_{KL}(Q \| P) \quad (6)$$

$$D_{JS}(P \| Q) = \frac{1}{2}[D_{KL}(P \| M) + D_{KL}(Q \| M)] \quad (7)$$

- Khả năng phát hiện nghĩa của từ trong văn cảnh (gọi là word sense disambiguation)

- Tìm từ đồng nghĩa. Có hai cách tính độ tương tự giữa các từ.

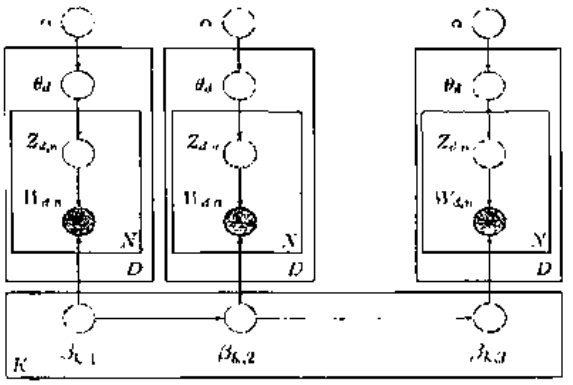
$$P(w_i \mid w_k) = \sum_{j=1}^{|T|} P(w_i \mid z_j) P(z_j \mid w_k) \quad (8)$$

$$P(w_i \mid w_k) = \sum_{j=1}^{|T|} \phi_i^j \theta_j^k \quad (9)$$

Mô hình LDA là mô hình phổ biến nhất và được ứng dụng trong nhiều lĩnh vực. Tùy với đặc điểm của dữ liệu mà mô hình này được mở rộng giả thiết hoặc tích hợp thêm dữ liệu để khai thác thông tin một cách triệt để hơn như sau:

- LDA được nới lỏng và mở rộng giả thiết để khám phá nhiều cấu trúc tinh xảo trong dữ liệu text. Giả thiết thứ nhất cho rằng LDA là "bag of words", mà thứ tự các từ trong văn bản không có thứ tự. Trong khi đó giả thiết này là không thực tế, nó chỉ phù hợp nếu mục đích của ta là khám phá dòng cấu trúc ẩn (sourse semantic structure). Có một số mô hình mở rộng LDA giả thiết rằng các từ là không thay đổi (unexchangeably), ví dụ [4] *đã phát triển mô hình mà nới lỏng túi từ bằng cách giả thiết rằng chủ đề được sinh ra bởi các từ phụ thuộc với từ trước nó*, [5] *phát triển mô hình chủ đề mà chuyển giữa LDA và HMM*. Những mô hình này mở rộng không gian tham số một cách đáng kể nhưng chỉ hiệu suất mô hình ngôn ngữ được cải thiện. Giả thiết thứ hai là *thứ tự của các văn bản* là không quan trọng. Trong công thức (1) là công thức còn lại bất biến để hoán vị thứ tự của văn bản trong tập dữ liệu. Giả thiết này có thể không thực tế khi phân tích tập văn bản dài (long-running) mà nó kéo dài hàng năm hoặc hàng thế kỉ. Với những tập dữ liệu như vậy, chúng ta có thể giả thiết rằng chủ đề thay

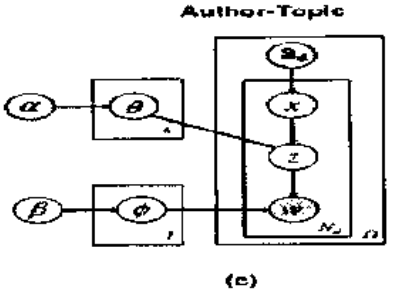
đổi theo thời gian. Một cách tiếp cận cho giả thiết này đó là *mô hình chủ đề động* – dynamic topic model [6]. Mô hình mà nó chú ý tới thứ tự của văn bản và đưa ra cấu trúc chủ đề hậu nghiệm phong phú hơn LDA. Hình 4 chỉ ra một chủ đề động Tốt hơn một phân phối qua các từ, một chủ đề bây giờ *là các phân phối liên tiếp* qua các từ Ta có thể tìm một chủ đề cơ sở của tập văn bản và theo dõi xem bằng cách nào các chủ đề này thay đổi theo thời gian



Hình 4. *Mô hình chủ đề với giả thiết chủ đề thay đổi theo thời gian*

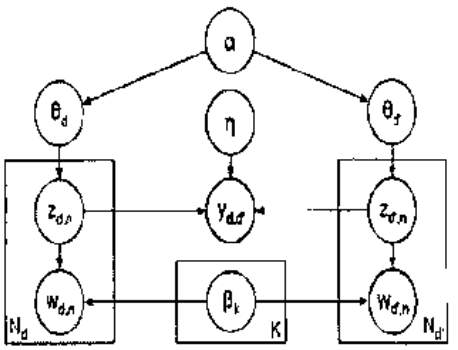
Giả thiết thứ 3 là số lượng chủ đề là cố định Trong mô hình chủ đề phi tham số [7] cung cấp một giải pháp tìm số lượng chủ đề qua tập dữ liệu trong quá trình suy diễn Posterior, và hơn nữa những văn bản mới có thể bộc lộ những chủ đề không biết trước. Mô hình chủ đề phi tham số Bayes được mở rộng cho mô hình phân cấp, mô hình chủ đề phân cấp là loại mà ta có thể tìm được một cây chủ đề, cấu trúc chủ đề đặc biệt của nó được suy diễn từ dữ liệu [8] Còn những mở rộng khác của LDA mà nới lỏng giả thiết của các biến. Mô hình chủ đề tương quan [9] và Pachinko allocation Machine [10] cho biết số lần xuất hiện của chủ đề để bộc lộ mối tương quan giữa các chủ đề (ví dụ như chủ đề về địa chất sẽ có tương quan gần với chủ đề hóa học hơn là chủ đề thể thao). Spherical topic model [11] cho biết các từ không nằm trong chủ đề. Mô hình chủ đề thưa [12] làm mô hình có phân phối chủ đề mạnh và "bursty" mô hình chủ đề cung cấp nhiều mô hình hiện thực về số lượng từ [13]

- Tích hợp thêm dữ liệu vào mô hình LDA, các văn bản chứa thêm thông tin như tác giả, tiêu đề, vị trí địa lý, liên kết, .. Tùy vào từng loại ứng dụng và loại dữ liệu mà những thông tin này có thể thích hợp với mô hình chủ đề Có một lực lượng nghiên cứu mạnh về tích hợp mô hình chủ đề trong metadata. The author-topic model là một nghiên cứu rất sớm về hướng này Tỉ lệ chủ đề được gắn vào author. Các bài báo với nhiều tác giả được giả thiết để gắn với mỗi từ tới một tác giả. Mô hình chủ đề - tác giả cho phép suy diễn về các tác giả tốt như với các văn bản



Hình 5. *Mô hình chủ đề tác giả*

Rất nhiều tập văn bản được liên kết (ví dụ như các bài báo khoa học được liên kết bằng cách trích dẫn hoặc các trang web được liên kết bởi các hyperlink). Một vài mô hình chủ đề được phát triển để giải thích cho những liên kết này khi ước lượng các chủ đề Mô hình chủ đề quan hệ (relational topic Model) giả thiết rằng mỗi văn bản được mô hình hóa như trong LDA và các liên kết giữa các văn bản phụ thuộc vào khoảng cách giữa các tỉ lệ chủ đề của chúng. Mô hình này vừa là mô hình chủ đề mới, vừa là mô hình mạng mới. Không giống với mô hình thống kê của mạng truyền thống, mô hình chủ đề quan hệ đưa vào các thuộc tính nút tính toán trong mô hình hóa các liên kết



Hình 6. *Mô hình chủ đề quan hệ*

Hướng khác mà tích hợp metadata trong mô hình chủ đề bao gồm:

+ Các mô hình của cấu trúc liên quan đến ngôn ngữ.

+ Các mô hình mà giải thích cho khoảng cách giữa tập văn bản

+ Các mô hình của các thực thể được đặt tên.

HƯỚNG PHÁT TRIỂN

Qua khảo sát thì mô hình chủ đề hoạt động hiệu quả khi học trên tập dữ liệu lớn và trong một số trường hợp thì nó cũng cải thiện hiệu xuất tìm kiếm. Trong mấy mấy năm gần đây hiệu quả của mô hình học sâu đang được áp dụng vào nhiều lĩnh vực Trong thời gian tới tác giả sẽ kết hợp mô hình chủ đề với học sâu để khai thác về đô tương tự ngữ nghĩa trong văn bản và môt số bài toán của tìm kiếm thông tin.

KẾT LUẬN

Qua bài này chúng tôi muốn giới thiệu về mô hình chủ đề LDA, suy diễn trong mô hình chủ đề, các nghiên cứu gần đây về các mô hình chủ đề và ứng dụng của mô hình chủ đề trong tìm kiếm thông tin.

TÀI LIỆU THAM KHẢO

1. David M. Blei, Andrew Y. Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), pp 993-1022

2 Thomas Hofmann. 1999 Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '99). ACM, New York, NY, USA, pp.50-57

3 David M. Blei. 2012 Probabilistic topic models. *Commun. ACM* 55, 4 (April 2012), pp. 77-84.

4 Hanna M Wallach 2006. Topic modeling. beyond bag-of-words In *Proceedings of the 23rd international conference on Machine learning* (ICML '06) ACM, New York, NY, USA, 977-984.

5. Griffiths, T , Steyvers, M., Blei, D , Tenenbaum (2005), J integrating topics and syntax. *Advances in Neural Information Processing Systems* 17 1 K saul, y weiss, and l bottou, eds. mit Press, Cambridge, ma, 537–544

6. Blei, D , Lafferty, J Dynamic topic models *in International Conference on Machine Learning* (2006), acm, new york, ny, usa, 113–120.

7. Teh, Y , Jordan, M , beal, M , Blei, D. hierarchical Dirichlet processes *J. Am Stat. Assoc* 101, 476 (2006), 1566–1581.

8. Blei, D , Griffiths, T., Jordan, M (2010) "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies" *J. ACM 57, 2 (2010)*, pp 1–30

9. Blei, D , Lafferty, J (2007) "A correlated topic model of science". *Ann. Appl. Stat.*, vol. 1, pp 17–35.

10 Li, W., Mccallum, A Pachinko allocation. Dagstructured mixture models of topic correlations *in International Conference on Machine Learning* (2006), pp. 577–584.

11 Reisinger, J , Waters, B., Silverthorn, B., mooney, r spherical topic models. *in International Conference on Machine Learning* (2010).

12. Wang, C., Blei, D. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process *Advances in Neural Information Processing Systems 22. y. bengio, D. schuurmans, J Lafferty, C. K. i. Williams, and a. culotta*, eds 2009, pp. 1982–1989

13. Doyle, G., Elkan,C., accounting for burstiness in topic models. *in International Conference on Machine Learning* 2009, acm, pp 281–288.

SUMMARY

TOPIC MODELS AND INFORMATION RETRIEVAL

Ha Thi Thanh*, Trinh Thi Thuy, Ngo Cam Tu

University of Information and Communication Technology - TNU

All our information is stored in the form of news, blogs, web pages, scientific articles, books, photos, audio, video, social networking . with a large information So it is too difficult to search information. We need a tool to help organizations, search, explore large amounts of this information The machine learning researchers have developed probabilistic topic model, an algorithm to explore and explain the large text data repository with information on topics Topic models are statistical methods that analyze the of the text data to discover the topic of the text, the way in which the interlinked themes, the way in which that topic instead change over time Topic model allows us to organize and summarize data warehouse number. Thus helping to exploit and find information more quickly

Keywords: *infomation retrieval, machine learning, topic model, LDA model*

Ngày nhận bài:23/10/2016; Ngày phản biện:04/11/2016; Ngày duyệt đăng: 31/5/2017

* *Tel. 0982 266009, Email: htthanh@ictu.edu.vn*