بسمه تعالی

پژوهشکده آماد، فناوری دفاعی و پدافند غیرعامل

گروه علوم و فناوری‌های دفاعی

میز تخصصی رصد و پایش علم، فناوری و نوآوری دفاعی و امنیتی

پیش‌بینی پیوند موضوعی پژوهش‌های علمی با استفاده از تحلیل شبکه‌ی مقالات علمی

رضا شکرچیان چالشتری

E-mail: rz.shekarchian@ut.ac.ir

گزارش ماهانه شماره (۲) - ۹۷/۷/۶

چکیده

یک راه برای تحلیل مجموعه‌ی اسناد دسته‌بندی نشده مدل‌سازی موضوعی می‌باشد. مدل‌سازی موضوعی می‌تواند کلماتی را که از نظر معنایی مشابه هستند کنار هم قرار دهد. با تحلیل اسناد مجموعه چندین موضوع به دست می آید. هر موضوع توزیعی بر روی کلمات می باشد و هر سند توزیعی از موضوعات می باشد. در مدل سازی موضوعی این توزیعات به دست می آیند و با مشخص شدن موضوعات می توان کلمات مشابه را در کنار هم دسته بندی کرد.

کلمات کلیدی: مدل سازی موضوعی

---

۱. معرفی LDA وPLSA

روش‌های مدل‌سازی احتمالاتی موضوعات، مجموعه‌ای از الگوریتم‌هایی هستند که هدف اصلی آن‌ها کشف ساختار نهان موضوعات در حجم وسیعی از اسناد می‌باشد. یکی از مطرح‌ترین و پایه‌ای‌ترین روش‌ها LDA[۱] [1] می‌باشد. LDA خود به منظور بهبود روش PLSA[۲] [2] پیشنهاد داده شد. این روش نیز خود بهبودی بر روش مشهور دیگری با نام LSA[۳] [3] می‌باشد.

LSA یکی از روش‌های پایه در مدل‌سازی موضوعی می‌باشد. ایده‌ی اصلی LSA این است که ماتریس سند-کلمه را به دو ماتریس سند-موضوع و موضوع کلمه تبدیل کند. قدم اول ساخت ماتریس سند-کلمه می‌باشد. برای m سند و مجموعه کلمه‌ای به اندازه‌ی n کلمه می‌تواند ماتریسی mxn ساخت که در آن سطرها سندها و ستون آن کلمات می‌باشند .

یک روش ساده برای وزن‌دهی به درایه‌های ماتریس سند-کلمه تعداد دفعاتی می‌باشند که کلمه در سند ظاهر شده است. این روش اهمیت یک کلمه در سند را نادیده می‌گیرد . روش دیگر وزن‌دهی درایه‌های ماتریس با tf*idf کلمات می‌باشد. که به نوعی میزان اهمیت و یکتایی کلمات را در سند نشان می‌دهد .Tf تعداد دفعاتی است که کلمه در سند ظاهر شده و idf متناظر با عکس تعداد سندهایی در مجموعه است که کلمه در آن‌ها ظاهر شده است .این ماتریس تنک و با نویز بالا می‌باشد و بسیاری

---

[۱] Latent Dirichlet allocation

[۲] Probabilistic latent semantic analysis

[۳] Latent semantic analysis

از درایه‌های آن اطلاعات مناسبی ندارند. به همین منظور به دنبال موضوعاتی می‌گردیم که ارتباط معنادارتری از کلمات و اسناد ارائه کنند LSA .از روش کاهش بعد ماتریس استفاده می‌کند و برای کاهش بعد از روش SVD بهره می‌گیرد. روش SVD ماتریس M را به ۳ ماتریس تبدیل می‌کند M=U*S*V :
که S ماتریس قطری مقادیر ویژه‌ی ماتریس M می‌باشد. از t مقدار ویژه‌ی اول برای ساخت S استفاده می‌شود که تعداد موضوعات را مشخص می‌کند U .ماتریس سند-موضوع و V ماتریس موضوع-کلمه می‌باشد. با استفاده از این ماتریس‌ها و روش‌های اندازه‌گیری‌ای همجون شباهت کسینوسی می‌توان شباهت اسناد، شباهت کلمات، شباهت کلمات با اسناد را اندازه‌گیری کرد .
از جمله مشکلات LSA نیاز آن به تعداد بالایی سند برای به دست آوردن نتایج دقیق می‌باشد .

LDA یک مدل مولد[۱] می‌باشد. مدل مولد برای اسناد بر اساس یک سری قانون نمونه‌گیری احتمالاتی می‌باشد. این قانون‌ها مشخص می‌کنند کلمات اسناد چگونه ممکن است بر پایه‌ی متغیرهای نهان تولید شوند. پس از به دست آمدن مدل تولیدی مناسب، هدف پیدا کردن بهترین مجموعه از متغیرهای نهان می‌باشد که می‌توانند توصیف کننده‌ی مشاهدات باشند (مثلا کلمات موجود در اسناد)، با فرض این‌که مدل به دست آمده داده‌ها را تولید کرده است. در واقع مجموعه‌ای از مشاهدات و اسناد موجود می‌باشد. می‌خواهیم مدل مولد را به گونه‌ای بسازیم که گویی این مدل تولید کننده‌ی مشاهدات بوده است. مدل‌های احتمالاتی موضوعی متنوعی وجود دارد. مانند [1], [2], [4]–[7]. ایده‌ی اصلی تمام این مدل‌ها یکسان می‌باشد. ایده این است که اسناد توزیعی از موضوعات می‌باشند. این مدل‌ها بیشتر در فرضیات آماری دارای تفاوت می‌باشند. فرض می‌کنیم $P(z)$ برای یک سند نشان دهنده‌ی توزیع بر روی تمام موضوعات z می‌باشد. $P(w \mid z)$ توزیع کلمات را بر روی موضوع z نشان می‌دهد و توزیع کلمه-موضوع[۲] معرفی می‌شود. به ازای هر سند تولید هر کلمه‌ی در دو مرحله صورت می‌گیرد. برای تولید هر کلمه‌ی w_i در یک سند، ابتدا یک نمونه‌گیری بر روی توزیع موضوعات صورت می‌گیرد و یک موضوع z انتخاب می شود. سپس یک کلمه از توزیع کلمه-موضوع $P(w \mid z)$ انتخاب می شود. از $P(z_i = j)$ برای نشان دادن احتمال انتخاب موضوع j برای کلمه‌ی iام در نمونه‌گیری و از $P(w_i \mid z_i = j)$ برای احتمال کلمه‌ی w_i در موضوع jام استفاده می‌شود. برای یک سند، احتمال تولید کلمات آن مطابق با فرمول ۲-۱ می‌باشد.

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \qquad (۱-۱)$$

که T در این‌جا تعداد موضوعات می‌باشد. فرمول ۲-۱ به طور خلاصه بیان می‌کند که احتمال تولید یک کلمه برای یک سند، برابر با احتمال تولید کلمه توسط موضوعات می‌باشد. به همین خاطر برای تمام موضوعات بررسی می‌کند سند چقدر به هر موضوع مرتبط است و برای هر موضوع، کلمه با چه احتمالی به موضوع تعلق می‌گیرد. این مدل از فرض تشکیل شدن هر سند از چندین موضوع استفاده می‌کند. نسبتی که هر سند از موضوعات دارد، با دیگر اسناد متفاوت است. یک سند ممکن است بیشتر راجع به دو موضوع اقتصاد و سیاست صحبت کرده باشد در حالی که سند دیگر ممکن است بیشتر راجع به سیاست و ورزش باشد. این خاصیت LDA می‌باشد که اسناد موجود، مجموعه‌ای یکسان از موضوعات را در بر می‌گیرند، ولی هر سند میزان تعلق متفاوتی به هر موضوع دارد. هدف مدل سازی موضوعی پیدا کردن موضوعاتی از روی مجموعه‌ی اسناد می‌باشد. اسناد به عنوان مشاهدات در نظر گرفته می‌شوند و در ابتدا وجود دارند. سه عنصر موضوعات، توزیع موضوعات بر روی

---

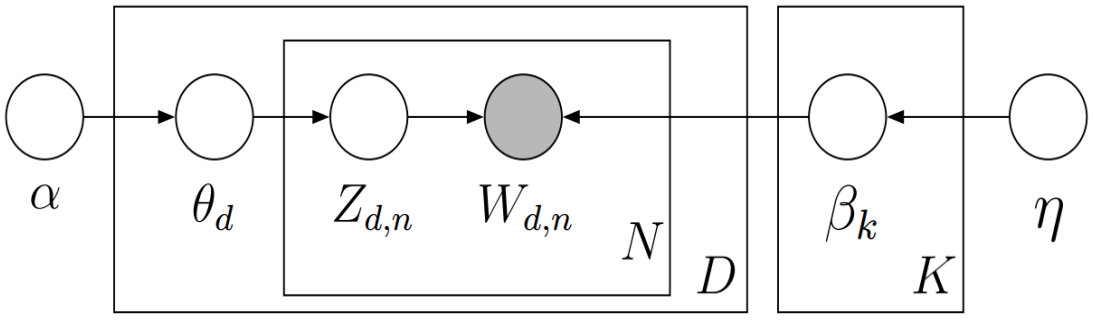
[۱] generative model

[۲] Topic-word distribution

هر سند و موضوعی که به هر کلمه‌ی سند نسبت داده می‌شود به عنوان دانش نهفته شناخته می‌شود که قرار است یادگیری شوند. همان‌طور که گفته شد مدلی که این دانش نهفته را به دست می‌آورد، یک مدل مولد می‌باشد. مدل مولد تلاش دارد تا پارامترهای مدل را به گونه‌ای به دست آورد که مشاهدات با بالاترین احتمال تولید شوند. هزینه‌ی اصلی مدل‌سازی موضوعی در قسمت استنتاج ساختار نهان موضوعات با استفاده از اسناد مشاهده شده می‌باشد. هدف استنتاج پیدا کردن توزیع شرطی متغیرهای نهان، مشروط به دیده شدن مشاهدات می‌باشد. به این احتمال شرطی، توزیع موخر[۱] نیز گفته می‌شود. می‌توان برای توصیف روش LDA از نمادگذاری زیر استفاده کرد. موضوعات $\beta_{1:k}$ می‌باشند که هر β_k توزیعی بر روی کلمات می‌باشد. k تعداد موضوعات می‌باشد. نسبتی که سند d از هر موضوع به خود اختصاص می‌دهد، با θ_d نمایش داده می‌شود و نسبت‌های موضوع[۲] خوانده می‌شود. $\theta_{d,k}$ نشان دهنده‌ی نسبتی است که موضوع k برای سند dام دارد. تخصیص‌های موضوع[۳] برای سند dام z_d می‌باشند که $z_{d,n}$ تخصیص موضوع برای nامین کلمه در سند dام می‌باشد. در نهایت کلمات دیده شده برای سند d، می‌باشند که $w_{d,n}$، nامین کلمه برای سند d می‌باشد. با در نظر گرفتن نمادگذاری‌های گفته شده، فرآیند تولیدی برای LDA از توزیع توأم متغیرهای آشکار[۴] و متغیرهای نهان[۵] به صورت فرمول ۲-۲ می‌باشد.

$$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D})=\prod_{i=1}^{K}p(\beta_i)\prod_{d=1}^{D}p(\theta_d)\left(\prod_{n=1}^{N}p(z_{d,n}\mid\theta_d)p(w_{d,n}\mid\beta_{1:K},z_{d,n})\right) \qquad (۱-۲)$$

باید توجه داشت که در توزیع گفته شده یک‌سری وابستگی وجود دارد. به عنوان مثال تخصیص موضوع $z_{d,n}$ وابسته به نسبت‌های موضوع θ_d برای هر موضوع می‌باشد. همین‌طور کلمه‌ی مشاهده شده‌ی $w_{d,n}$ وابسته به تخصیص موضوع $z_{d,n}$ و تمام موضوعات $\beta_{1:k}$ می‌باشد (عملا کلمه با نگاه به موضوعی که $z_{d,n}$ به آن اشاره می‌کند و احتمالی که کلمه‌ی $w_{d,n}$ در آن موضوع دارد، به دست می‌آید). تمامی این وابستگی‌ها‌ست که LDA را LDA کرده است. مدل‌های گرافیکی احتمالاتی[۶] این امکان را فراهم می‌کنند که برخی از توزیع‌های احتمالی را به زبان گرافیکی بیان کرد. شکل مدل گرافیکی برای LDA می‌باشد.



شکل (۱-۱) مدل گرافیکی روش LDA. هر نود بیان کننده‌ی یک متغیر تصادفی می‌باشد و متناظر با نقشش در فرآیند

[۱] Posterior distribution
[۲] Topic proportions
[۳] Topic assignments
[۴] Observed variables
[۵] Hidden variables
[۶] Probabilistic graphical models

مولد برچسب خورده است. نودهای نهان (نسبت‌های موضوع، تخصیص‌های موضوع و موضوع‌ها) به صورت ساده و نودهای آشکار (کلمات دیده شده در اسناد) به صورت سایه‌دار نمایش داده شده‌اند. مستطیل‌ها در شکل به عنوان صفحه[۱] شناخته می‌شوند و نشان دهنده‌ی تعداد تکرار می‌باشند. صفحه‌ی N نشان دهنده‌ی تعداد کلمات در هر سند و صفحه‌ی D نشان دهنده‌ی تعداد اسناد مجموعه می‌باشد.

LDA یکی از معتبرترین روش‌های موجود مدل‌سازی موضوعی اسناد می‌باشد. این روش خود بهبودی برای روش PLSA [2] می‌باشد. این روش بسیار شبیه به LDA می‌باشد. در PLSA متغیرهای پنهان موضوعات، $z_k \in \{z_1,...,z_K\}$ متناسب با تعداد اتفاق کلمات $w_j \in \{w_1,...,w_M\}$ در سند $d_i \in \{d_1,...,d_N\}$ می‌باشند. برای جفت سند و کلمه‌ی (d,w) داریم:

$$P(d_i,w_j) = P(d_i)\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \qquad (۱-۳)$$

$P(w_j \mid z_k)$ احتمال کلمه‌ی w_j در مو ضوع z_k می‌با شد و $P(z_k \mid d_i)$ احتمال مو ضوع z_k برای سند d_i می‌با شد. این پارامترها را می‌توان با بیشینه کردن لگاریتم راست‌نمایی[۲] مجوعه‌ی C به دست آورد.

$$L(C) = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j) log \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \qquad (۱-۴)$$

پارامترهای مدل $\phi = \{P(w_j \mid z_k)\}$ و $\theta = \{P(z_k \mid d_i)\}$ می‌باشند که می‌توان آن‌ها را به استفاده از الگوریتم EM[۳] [8] تخمین زد.

مدل‌سازی موضوعی سابقه‌ی چندین ساله‌ای دارد. در ابتدا به موضوعات به صورت بسته‌ی لغات نگاه می‌شد. گسترش‌هایی بر این نگاه صورت گرفت. به دست آوردن موضوعات به صورت سلسله مراتبی به جای حالت تک لایه، به کارگیری موجودیت‌ها در کنار متن، استفاده از عبارات به جای لغات و استفاده از دانش داخلی و خارجی در مدل‌سازی از جمله‌ی این گسترش‌ها می‌باشند. هریک از این گسترش‌ها در ادامه بررسی خواهند شد.

مراجع

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[2] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[۱] Plate

[۲] Log likelihood

[۳] Expectation maximization

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The Author-topic Model for Authors and Documents,” in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004, pp. 487–494.

[5] T. L. Griffiths and M. Steyvers, “A Probabilistic Approach to Semantic Representation.” 2002.

[6] T. L. Griffiths and M. Steyvers, “Prediction and Semantic Association,” in *Advances in Neural Information Processing Systems*, 2003, p. 15.

[7] D. Cohn and T. Hofmann, “The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity.” 2001.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.