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 (۳، ۴) - ۹۷/۸/۱۳

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 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 SVDبهره می‌گیرد. روش SVD ماتریس M را به ۳ ماتریس تبدیل می‌کند M=U*S*V :

---

[۱] Latent Dirichlet allocation

[۲] Probabilistic latent semantic analysis

[۳] Latent semantic analysis

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 د ست آمدن مدل تولیدی منا سب، هدف پیدا کردن بهترین مجموعه از متغیرهای نهان می‌با شد که می‌توانند تو صیف 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 [7]–[4] ,[2], [1]. 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 مو ضوعات z می‌با شد. $P(w \mid z)$ توزیع کلمات را بر روی مو ضوع z ن شان می‌دهد و توزیع کلمه-مو ضوع[۲] معرفی می شود. 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 مو ضوعات صورت می‌گیرد و یک مو ضوع z انتخاب می شود. سپس یک کلمه از توزیع کلمه-موضوع $P(w \mid z)$ انتخاب می‌شود. از $P(z_i = j)$ برای نشان دادن احتمال انتخاب موضوع j برای کلمه‌ی i ام در نمونه‌گیری و از $P(w_i \mid z_i = j)$ برای احتمال کلمه‌ی w_i در مو ضوع j ام ا ستفاده می شود. 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 مو ضوعات می‌با شد. فرمول ۲-۱ به طور خلا صه 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 مو ضوع مرتبط ا ست و برای هر مو ضوع، کلمه با چه احتمالی به مو ضوع تعلق می‌گیرد. این مدل از فرض تشکیل شدن هر سند از چندین مو ضوع ا ستفاده می‌کند. نسبتی که هر سند از مو ضوعات 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 مدل‌سازی موضوعی

[۱] generative model

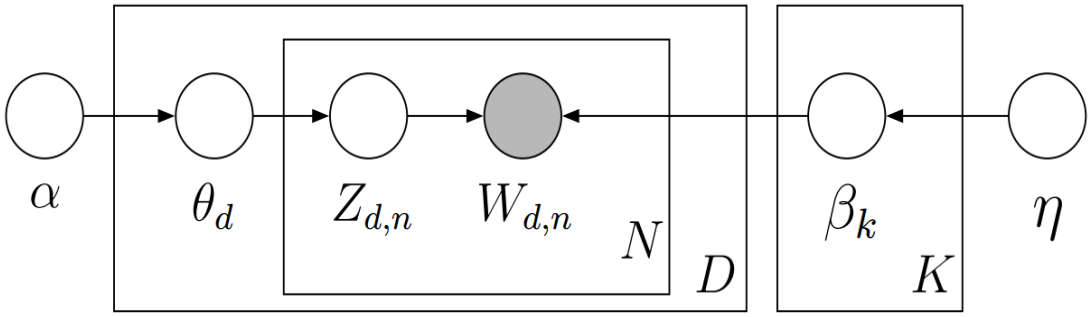
[۲] Topic-word distribution

پیدا کردن مو ضوعاتی از روی مجموعه‌ی ا سناد می‌با شد. ا سناد به عنوان م شاهدات در نظر گرفته می شوند و در ابتدا وجود دارند. سه عن صر مو ضوعات، توزیع مو ضوعات بر روی هر سند و مو ضوعی که به هر کلمه‌ی سند ن سبت داده می‌شود به عنوان دانش نهفته شناخته می‌شود که قرار است یادگیری شوند. همان‌طور که گفته شد مدلی که این دانش نهفته را به د ست می‌آورد، یک مدل مولد می‌با شد. مدل مولد تلاش دارد تا پارامترهای مدل را به گونه‌ای به د ست آورد که م شاهدات با بالاترین احتمال تولید شوند. هزینه‌ی ا صلی مدل سازی مو ضوعی در ق سمت ا ستنتاج ساختار نهان موضوعات با استفاده از اسناد مشاهده شده می‌باشد. هدف استنتاج پیدا کردن توزیع شرطی متغیرهای نهان، مشروط به دیده شـدن مشـاهدات می‌باشـد. به این احتمال شـرطی، توزیع موخر[۱] نیز گفته می‌شـود. می‌توان برای توصـیف روش LDA از نمادگذاری زیر ا ستفاده کرد. مو ضوعات $\beta_{1:k}$ می‌با شند که هر β_k توزیعی بر روی کلمات می‌با شد. 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 می‌با شند که $z_{d,n}$ تخ صیص مو ضوع برای nامین کلمه در سند dام می‌با شد. 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.

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

باید توجه داشت که در توزیع گفته شده یک‌سری وابستگی وجود دارد. به عنوان مثال تخصیص مو ضوع $z_{d,n}$ وابسته به ن سبت‌های مو ضوع θ_d برای هر مو ضوع می‌با شد. همین‌طور کلمه‌ی م شاهده 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 موضـوع دارد، به دسـت می‌آید). تمامی این وابسـتگی‌هاسـت که LDA را LDA کرده اســت. مدل‌های گرافیکی احتمالاتی[۶] این امکان را فراهم می‌کنند که برخی از توزیع‌های احتمالی را به زبان گرافیکی بیان کرد. شکل ۱ مدل گرافیکی برای LDA می‌باشد.

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



شکل ۱) مدل گرافیکی روش LDA. هر نود بیان کننده‌ی یک متغیر تصادفی می‌باشد و متناظر با نقشش در فرآیند 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 مدل سازی مو ضوعی ا سناد می‌با شد. این روش خود بهبودی برای روش PLSA [2] می‌باشد. این روش بسیار شبیه به LDA می‌باشد. در PLSA متغیرهای پنهان موضوعات، $z_k \in \{z_1,...,z_K\}$ متنا سب با تعداد اتفاق کلمات $w_j \in \{w_1,...,w_M\}$ در سند $d_i \in \{d_1,...,d_N\}$ می‌با شند. 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 موضوع z_k می‌باشد و $P(z_k \mid d_i)$ احتمال موضوع z_k برای سند d_i می‌باشد. 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 $\theta = \{P(z_k \mid d_i)\}$ می‌با شند که می‌توان آن‌ها را به ا ستفاده از الگوریتم EM[۳] [8] تخمین زد.

مدل سازی مو ضوعی سابقه‌ی چندین ساله‌ای دارد. در ابتدا به مو ضوعات به صورت بسته‌ی لغات نگاه می شد. 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، ا ستفاده از عبارات به جای لغات و ا ستفاده از دانش داخلی و خارجی در مدل سازی از جمله‌ی این گسترش‌ها می‌باشند. هریک از این گسترش‌ها در ادامه بررسی خواهند شد.

---

۱ Plate
۲ Log likelihood
۳ Expectation maximization

۲. استفاده از موجودیت‌ها، شبکه‌های اطلاعاتی[۱] و عبارات در گسترش مدل‌سازی موضوعی

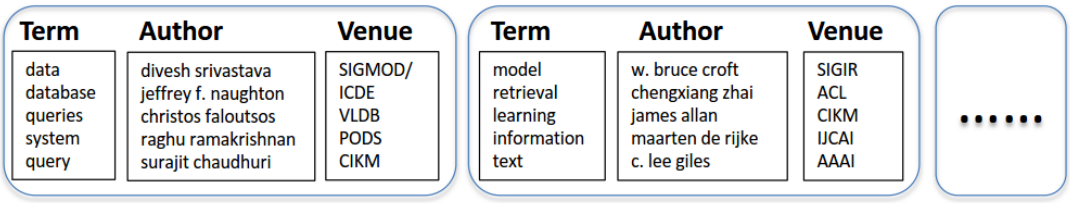
روش‌های بسیاری برای بهبود LDA ارائه شدند. برخی از اطلاعات بیشتری از اسناد استفاده می‌کنند و مدل را پیچیده‌تر می‌کنند. به طور مثال در [9] از اطلاعات نویسندگان در کنار متن اسناد برای مدل‌سازی استفاده شده است. در مقاله‌ی [10] از موجودیت‌های موجود در اسناد برای بهبود مدل‌سازی کمک گرفته شده است. به طور مثال یک متن خبری می‌تواند در برگیرنده‌ی نام‌های اشخاص، محل‌ها، شرکت‌ها، مکان‌ها و ... باشد. این اطلاعات دانش بسیاری از سند در بر دارند و می‌توانند در مدل‌سازی بسیار کمک کننده باشند.

در برخی از روش‌ها از ارتباط میان اسناد شبکه‌های همگون و ناهمگونی ساخته‌اند و از این ارتباط برای مدل خود استفاده کرده‌اند. از جمله‌ی این روش‌ها می‌توان به [12] ,[11], [9] اشاره کرد. به طور مثال در [12]از اطلاعات مجله‌ای که مقاله در آن به چاپ رسیده است، نویسندگان مقاله و خود اسناد یک شبکه‌ی ناهمگون ساخته است. هر مقاله به یک مجله و چند نویسنده لینک دارد. هر نویسنده و مقاله نیز به چند مقاله لینک دارد و مرتبط است. از موضوعات به دست آمده برای هر سند برای مدل کردن موضوعات مربوط به نویسندگان و مجله‌ها استفاده می‌شود. از موضوعات به دست آمده برای مقالات و مجله‌ها نیز در مدل سازی موضوعی اسناد استفاده می شود . به نوعی در این شبکه موضوع هر نود در تاثیر پذیرفته و تاثیرگذار بر موضوعات همسایه‌ها می‌باشد و تاثیر موضوعات در شبکه‌ی ساخته شده از نودی به نود دیگر منتقل می‌شود. البته این مقاله از روش PLSA به عنوان مدل پایه برای مدل‌سازی موضوعی اسناد استفاده کرده است.

مقاله‌ی [11] از جمله تحقیقاتی است که بر روی مدل‌سازی موضوعات به صورت سلسله مراتبی کار کرده است. در این مقاله هدف ساختن سلسله مراتب موضوعات در شبکه‌ی اطلاعات ناهمگون می‌باشد. شبکه‌ی اطلاعات ناهمگون حاصل از اشیا از نوع‌های مختلفی می‌باشد که با یکدیگر در ارتباط می‌باشند. این نوع شبکه‌ی اطلاعات را می‌توان از بسیاری از منابع متداول همچون گزارشات شرکت‌های تجاری، مقالات منتشر شده‌ی علمی و رسانه‌های اجتماعی به دست آورد. ساختن سلسله مراتب مفاهیم با کیفیت بالا به منظور نمایش موضوعات در درشت‌دانگی‌های مختلف می‌تواند در جستجو، بررسی و عبور از اطلاعات و شناسایی الگو کمک‌رسان باشد. در این مقاله الگوریتمی ارائه شده است که به صورت بازگشتی اقدام به ساخت سلسله مراتب موضوعات می‌کند. نوآوری و خلاقیتی که در مقاله به کار رفته این است که می‌تواند هم عبارات متنی و هم موجودیت‌های از نوع‌های مختلف را با یک روش ترکیبی از خوشه‌بندی و رتبه‌بندی بر روی شبکه داده‌های ناهمگون استفاده کند و سلسله مراتب موضوعات از نوع‌های مختلف را به دست آورد. در بسیاری از کارهایی که بر روی مدل‌های سلسله مراتبی موضوعات کار شده، استفاده از رابطه‌ی میان موجودیت‌ها با نوع‌های مختلف رایج نبوده است. در روشی که در این مقاله ارائه شده هم اطلاعات متن و هم ارتباطات میان موجودیت‌ها برای ساخت سلسله مراتب استفاده می‌شوند. نوآوری‌های به کار برده شده از این قرار می‌باشند. ساخت سلسله مراتب موضوعات به صورت بازگشتی به طوری که موضوعات به صورت لیست مرتبی از عبارات و موجودیت‌های از نوع‌های مختلف می‌باشند. از خوشه‌بندی و روش‌های رتبه‌بندی برای به دست آوردن زیر موضوع‌ها از موضوعات به صورت بازگشتی استفاده می‌شود.

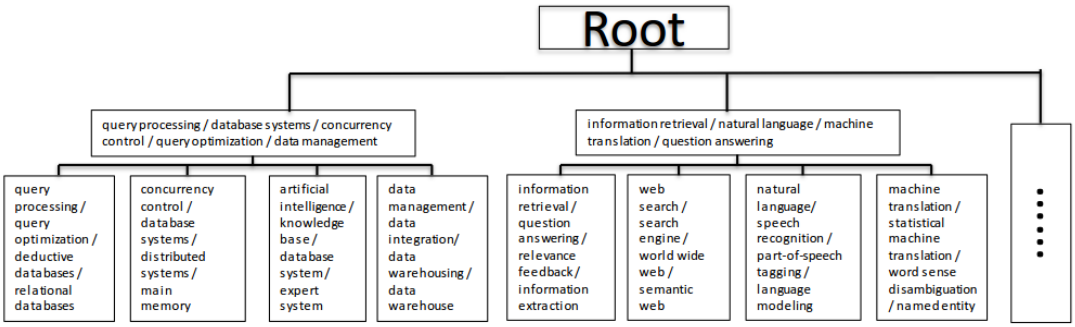
[۱] Information networks

این روش از مزایای روش [13] استفاده می‌کند که روشی برای تحلیل شبکه‌ها ناهمگون می‌باشد و برای به دست آوردن سلسله مراتب موضوعات بسیار مناسب می‌باشد. روشی که در این مقاله ارائه شده همچنین این قابلیت را دارد تا به صورت اتوماتیک اهمیت لینک‌های موجودیت‌ها با نوع‌های مختلف را مشخص کند. این مقاله به نوعی ادامه‌ی کار قبلی همین افراد است که روش [14] می‌باشد. این روش یک روش مبتنی بر آمار می‌باشد و سلسله مراتب موضوعات را تنها با تکیه بر محتوای متن انجام می‌دهد. در این روش هر موضوع به صورت سلسله مراتب و لیست مرتبی از عبارات با طول‌های مختلف نشان داده می‌شوند. این روش یک روش مبتنی بر عبارت می‌باشد و از این نگاه عبارتی به جای نگاه یک‌گرمی[۱] برای خوشه‌بندی و رتبه‌بندی موضوعات استفاده می‌کند. برای تخمین تکرار موضوعی عبارات از الگوریتم‌های کاوش الگوهای مکرر[۲] استفاده می‌کند و از این تکرار برای رتبه‌بندی موضوعی عبارات استفاده می‌کند. برای رتبه‌بندی عبارات موضوع، تابعی طراحی شده که چهار عامل را در کیفیت عبارات موضوع موثر می‌داند. پوشش[۳]، خلوص[۴]، کامل بودن[۵] و عبارت بودن[۶]. پوشش بیان می‌کند که یک عبارت برای یک موضوع باید اسناد بسیاری را با آن موضوع پوشش دهد. به طور مثال «information retrieval» پوشش بهتری نسبت به «retrieval cross-language information» در موضوع بازیابی اطلاعات دارد. یک عبارات در یک موضوع خلوص دارد اگر تنها در اسناد با آن موضوع مکرر باشد و نه در اسنادی با موضوع دیگر. کامل بودن بیان می‌کند که یک موضوع کامل نباید زیر مجموعه‌ای از عبارت کامل بزرگ‌تر دیگری باشد. عبارت بودن هم نشان دهنده‌ی این است که کلمات عبارات باید به میزان کافی بیشتر از متوسط هم‌رخداد کلمات، با یکدیگر هم‌اتفاق باشند. در نهایت این روش از خوشه‌بندی بازگشتی برای ساخت سلسه مراتب استفاده می‌کند. سه شکل ۲ و ۳ و ۴ نتایج سه روش NetClus، CATHY و CATHYIN را نمایش می‌دهند. شکل‌ها از مقاله‌ی [11] می‌باشند. این سه روش مدل‌سازی موضوعی بر روی مقالات علوم کامپیوتر با استفاده از سه خصیصه‌ی کلمات، نویسندگان و مجله انجام شده‌اند.

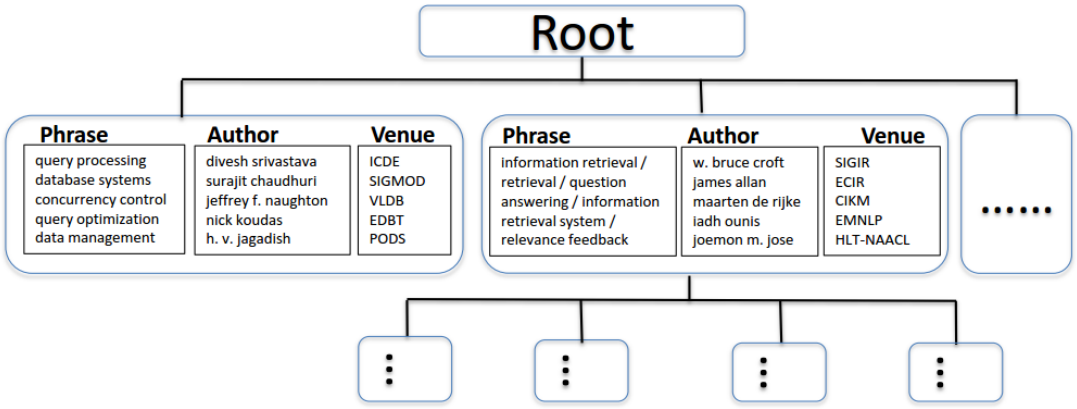


شکل ۲) روش مقاله‌ی [13]. خوشه‌های موجودیت‌های ناهمگون. هر مستطیل محیطی نشان دهنده‌ی یک خوشه می‌باشد. هر خوشه لیست مرتبی از کلمات و دو لیست مرتب از موجودیت‌ها می‌باشد.

[۱] unigram
[۲] Frequent pattern mining
[۳] coverage
[۴] purity
[۵] completeness
[۶] phraseness



شکل ۳)روش مقاله‌ی [14]. سلسله مراتب موضوعی از متن‌ها. هر نود در شکل، لیست مرتبی از عبارات می‌باشد.



شکل ۴) روش مقاله [11]. سلسله مراتبی از موجودیت‌ها ناهمگون. هر نود در شکل لیست مرتبی از عبارات و دو لیست مرتب از موجودیت‌ها دارد.

دیده شد که چگونه از عبارات و موجودیت‌ها در مدل‌سازی و نمایش‌سازی موضوعات استفاده شد. از دیگر کارهایی که از عبارات برای مدل‌سازی موضوعی بهره گرفته، [15] می‌باشد. از قسمتی از ایده‌ی این مقاله که مربوط به استخراج عبارت‌ها[۱] می‌باشد، برای پیش‌برد این تحقیق بهره گرفته شده است. استفاده از عبارت‌ها برای بیان موضوعات بسیار به نظر بشر نزدیک‌تر از استفاده از کلمات می‌باشد و انسان بسیاری از مفاهیم را با دیدن چند کلمه در کنار هم متوجه می شود. بسیاری از مقالات از پس‌پردازش[۲] بر روی نتیجه‌ی مدل سازهای موضوعی مبتنی بر یک گرم استفاده می‌کنند تا آن‌ها را به صورت عبارت‌گونه نمایش دهند. برخی دیگر از کارها از الگوریتم‌های پیچیده‌ی مبتنی بر چند گرم[۳] بهره می‌گیرند. این روش‌ها یا کیفیت پایینی دارند و یا به خاطر پیچیدگی بسیار هزینه بر می‌باشند و مقیاس‌پذیر نمی‌باشند. روش این مقاله هم از نظر کیفیت و هم از نظر عملکرد بسیار کارا می‌باشد.

[۱] phrase
[۲] Post processing
[۳] N-gram

۳. استفاده از دانش در مدل‌سازی موضوعی

۳-۱. استفاده از دانش داخلی متن

مقاله‌ی [16] از جمله اولین کارهایی است که با تکیه بر متن و دانش برخاسته از متن، کیفیت مدل‌سازی موضوعی را بهبود داده است. هم‌چنین این مقاله روشی خودکار برای ارزیابی موضوعات پیشنهاد کرده است. فرمولی که در این مقاله برای ارزیابی ارائه شده بسیار شبیه به PMI[۱] می‌باشد. میزان هم‌رخدادی[۲] دو به دوی M کلمه‌ی با بیشترین احتمال در هر موضوع موثر در خروجی معیار ارزیابی می‌باشد و انسجام موضوع[۳] را نشان می‌دهد. این مقاله روشی برای جلوگیری از ساخته شدن موضوعات ضعیف و با انسجام پایین دارد. با معیار ارائه شده، شناسایی موضوعات با کیفیت پایین، بدون نظر خبره و به صورت خودکار امکان‌پذیر شده است. در این مقاله، از هم‌رخدادی کلمات در مجموعه‌ی اسناد مستقیما در مدل‌سازی موضوع‌ها استفاده می‌کند. شاید در ابتدا این‌گونه به نظر آید که با اجازه ندادن به ظاهر شدن در یک موضوع به کلماتی که با هم‌دیگر هرگز هم‌رخداد نبوده‌اند، می‌توان به موضوعات با کیفیت دست یافت. ولی این طور نیست. طبق قانون توانی[۴]، بیشتر کلمات به ندرت ظاهر می‌شوند و با بیشتر کلمات هم‌رخداد نیستند. حتی اگر به صورت معنایی با یکدیگر مرتبط باشند. در مدل‌ساز موضوعی ارائه شده تعداد رخداد کلمه‌ی w در موضوع t نه تنها احتمال دیده شدن دوباره‌ی کلمه را بالا می‌برد، بلکه باعث افزایش احتمال دیده شدن کلمات مرتبط نیز می شود. این روش جدید موضوعات اسناد را از طریق LDA به دست می‌آورد. با این تفاوت که مولفه‌ی مر سومی که در نمونه‌گیری موضوع-کلمه[۵] استفاده می شود و معروف به Polya urn می‌باشد را با Polya urn تعمیم یافته[۶] [17] جایگزین می‌کند. یک دنباله از نمونه‌های i.i.d[۷] توزیع گسسته را می‌توان به دست آمده از تکرار انتخاب تصادفی یک توپ از ظرفی دانست. تعداد توپ‌ها از هر رنگ متناسب با احتمال آن توپ می‌باشد. پس از انتخاب هر توپ، جایگذاری انجام می‌شود. در Polya urn ساده هر توپ با توپی از همان رنگ جای‌گذاری می‌شود. هنگامی که یک توپ با رنگ مشخص از ظرفی برداشته می شود، توپ به همراه توپ جدیدی از همان رنگ به ظرف بازگردانده می شود. این‌گونه نمونه‌گیری باعث می‌شود احتمال انتخاب توپ رنگ w با هر بار انتخاب توپی با آن رنگ افزایش یابد. LDA از Polya urn ساده استفاده می‌کند. در روش Polya urn تعمیم یافته با انتخاب توپ رنگ w، دو توپ از همان رنگ و A_{vw} توپ اضافی از هر رنگ $v \in \{1,...,W\}$ به ظرف بازگردانده می‌شوند. A_{vw} میزان تاثیر و ارتباط رنگ‌ها به یک‌دیگر را نشان می‌دهد. در مساله‌ی مدل‌سازی موضوعی کلمات در نقش رنگ‌ها و موضوعات در نقش ظرف‌ها می‌باشند. توزیع کلمات یک موضوع متناسب با نسبت توپ‌های از هر رنگ در ظرف می‌باشد. در [16] از هم‌رخدادی

---

۱ Point-wised Mutual Information
۲ Co-occurrence
۳ Topic coherence
۴ Power-law
۵ Topic-word
۶ Generalized Polya urn
۷ Independent and identically distributed

کلمات در ساخت ماتریس A_{vw} استفاده می‌شود.

۲-۳ استفاده از دانش دامنه‌های دیگر در مدل‌سازی موضوعی

یکی از روش‌هایی که از دانش دیگر دامنه‌ها برای مدل‌سازی مجموعه استفاده می‌کند روش AMC[1] [18] می‌باشد. یکی از ضعف‌های مدل‌سازی موضوعات کیفیت نامطلوب نتایج در صورت کم بودن تعداد اسناد مجموعه می‌باشد. الگو گرفتن از نحوه‌ی یادگیری انسان و نحوه تعامل آن با اطلاعات جدید می‌تواند در این زمینه کمک کننده باشد. یادگیری طولانی مدت[2] یکی از روش‌ها در یادگیری ماشین می‌باشد که AMC به نوعی از این الگو برای بهبود نتایج استفاده کرده است. انسان هنگام برخورد با یک رویداد جدید سعی می‌کند تا از آن‌چه در گذشته یاد گرفته است استفاده کند و نتیجه‌ی نحوه‌ی تعامل با رویداد جدید خود یک دانشی می‌شود برای رویدادهای بعدی. LDA و روش‌های گسترش داده‌ی آن از جمله روش‌های مرسوم و معروف برای مدل سازی موضوعی می‌باشند و نیاز به حجم زیادی از اسناد برای مدل‌سازی می‌باشند. مجموعه‌های سندی‌ای که شامل مقدار قابل توجهی سند باشند معمولا اندک می‌باشند. به عنوان مثال نظرهای انجام گرفته بر روی کالاها را اگر در نظر بگیریم، اکثر کالاها بیشتر از ۱۰۰ نظر ندارند. اگر نظرهای مربوط به هر کالا را یک مجموعه در نظر بگیریم، با این تعداد از اسناد نمی‌توان موضوعات نهفته در نظرات را مدل‌سازی کرد. برای مقابله با این شرایط ۳ روش می‌توان پیشنهاد داد.

- **ارائه‌ی یک روش مدل سازی بهتر**: این روش زمانی می‌تواند تاثیرگذار باشد که تعداد زیادی از اسناد موجود باشد. چون‌که یادگیری مدل سازی موضوعی به صورت بدون ناظر[3] صورت می‌گیرد، اگر حجم اسناد کم باشد، اطلاعات کافی آماری قابل اطمینان وجود نخواهد داشت تا بتوان موضوعات منسجم به دست آورد. به نوعی نیاز به نظارت و یا اطلاعات خارج از اسناد داده شده نیاز می‌باشد.
- **درخواست از کاربر برای ارائه‌ی دانش اولیه از دامنه:** یک نمونه از اطلاعات خارجی دانش اولیه کاربر از دامنه می‌باشد. به عنوان مثال کاربر می‌تواند دانش خود را به صورت ارتباط بایسته[4] و ارتباط نبایسته[5] وارد کند. ارتباط بایسته بیان می‌کند که دو کلمه باید در یک موضوع قرار بگیرند و ارتباط نبایسته بیان می‌کند که دو کلمه باید در دو موضوع متفاوت باشند. البته این نوع درخواست از کاربر می‌تواند مشکل‌زا باشد. چرا که مدل‌سازی غیر اتوماتیک می‌شود و همچنین کاربر نمی‌داند که چه دانشی برای چه قسمتی می‌تواند مفید واقع شود.

۴. ارزیابی کیفیت مدل‌ساز موضوع

[1] Automatic Must-link Cannot-link
[2] Lifelong learning
[3] Unsupervised
[4] Must-link
[5] Cannot-link

روش‌های مدل سازی مجموعه‌ای از لغات که مو ضوع نام دارد را از مجموعه‌ی سندی ا ستخراج می‌کنند. این کار به وسیله‌ی تکرارهای کلمات در سند انجام می‌شود. روش‌های ارزیابی انسجام[1] مو ضوعات کمک می‌کنند تا مو ضوعات خوب از بد تمیز داده شوند و ن شان می‌دهد که مو ضوعات به دست آمده تا چه حدی قابل فهم و معنادار می‌باشند. معیارهای ارزیابی انسجام بیشتر در حوزه‌ی پردازش زبان‌های طبیعی ا ستفاده می شوند. در پردازش زبان‌های طبیعی مو ضوعات بی شتر به عنوان پیش‌پردازش اسناد مجموعه استفاده می‌شوند. موضوعات مجموعه و موضوعات هر سند، در خلا صه سازی ا سناد [21]، روش‌های ت شخیص معنای کلمات [22] و ترجمه‌ی ما شین [23] ا ستفاده می‌شوند. در این تحقیق مسائل و دغدغه‌های مربوط به پردازش زبان مطرح نیست. در این‌جا ملاک خوبی و بدی یک مو ضوع میزان نزدیکی آن به فهم ب شری می‌با شد. چنگ و همکاران در [24] رو شی پیشنهاد داده‌اند که مبتنی بر نفوذ[2] کلمه می‌باشد. کلمات مزاحم[3] به صورت تصادفی به موضوع‌ها وارد می‌شوند و از کاربران خواسته می‌شود که لغت نفوذی را پیدا کنند. روش نفوذ کلمات از این فرض ا ستفاده می‌کند که کلمات مزاحم در یک مو ضوع منسجم بهتر قابل ت شخیص می‌با شند. هرچه قدرت ت شخیص بی شتر با شد، مو ضوع منسجم‌تر ا ست. این روش به صورت غیر خودکار انجام می شود و این می‌تواند ا صلی‌ترین نقطه‌ی ضعف این روش با شد. در رو شی دیگر نیومن و همکاران [25] برای ارزیابی کیفیت موضوعات از کاربران خواستند که به موضوعات بر اساس کیفیت آن‌ها امتیاز دهند. در این روش هر مو ضوع با ده کلمه‌ی اول آن م شخص می شود. تلاش‌های ب سیاری صورت گرفت تا رو شی خودکار ارائه دهند که ارزیابی آن نزدیک به ارزیابی کاربران باشد. از ویکی‌پدیای انگلیسی و هم‌رخدادی کلمات موضوع در آن برای این ارزیابی خودکار استفاده شد. میمنو و همکاران [16] روش بسیار مشابهی ارائه کردند. تفاوت اصلی آن‌ها این است که به جای استفاده از اسناد ویکی‌پدیا برای نمونه‌گیری تعداد هم‌اتفاقی کلمات از اسناد خود مجموعه‌ی مدل شده برای نمونه‌گیری استفاده می‌کند. مشاهده شد که روش آن‌ها نزدیکی قابل توجهی به روشی که از امتیاز کاربران برای کیفیت‌سنجی موضوعات استفاده می‌کند، دارد. این مقاله علاوه بر ارائه‌ی یک معیار ارزیابی خودکار، روش جدیدی برای مدل‌سازی موضوعی ارائه کرده است که از هم‌رخدادی کلمات برای بهبود نتایج مدل‌سازی استفاده می‌کند. این روش خودکار نزدیکی قابل توجه به نظر کاربران در مورد کیفیت

---

[1] Coherence

[2] Intrusion

[3] intruder

مو ضوعات دارد. همین‌طور کیفیت به د ست آمده از این روش نزدیکی قابل توجهی به رو شی دارد که برای ارزیابی کیفیت موضوعات از نفوذ کلمات استفاده می‌کند.

۵. کاربردهای مدل‌سازی موضوعی

مدل سازهای مو ضوعی در بـ سیاری از زمینه‌ها وارد شده‌اند. از این میان می‌توان به مدل سازی زبان و انطباق مدل زبانی[۱] [27] ,[26]، بازیابی اطلاعات [28] و [34]–[32]، ابهام‌زدایی معنای کلمات [32]، تحلیل شبکه‌های اجتماعی [34] ,[33] و کاوش نظرات [36] ,[35] اشاره کرد.

مقاله‌ی [37] به پیدا کردن انجمن‌ها[۲] و موضـوعات با هم، در یک روش ترکیبی پرداخته اسـت. بـ سیاری از اطلاعاتی که از کاربرها به د ست می‌آید دارای متن می‌با شند. کاربرها با یکدیگر در ارتباط می‌باشـند و در گراف روابط بین آن‌ها یال وجود دارد. بسـیاری از روش‌ها از لینک میان نودهای گراف برای مدل سازی مو ضوعی ا ستفاده کرده‌اند. ولی در این مقاله شنا سایی انجمن‌ها به کمک شنا سایی موضوعات و شناسایی موضوعات به کمک شناسایی انجمن‌ها آمده‌اند. کاربرهایی که در یک انجمن قرار می‌گیرند موضوعات شبیه به هم زیادی دارند.

از دیگر کاربرهای مدل سازی مو ضوعی می‌توان به سیـ ستم‌های تو صیه‌گر ا شاره کرد. مقاله‌ی [38] یکی از این کارها می‌باشـد که یک توصـیه‌گر برای مقالات علمی با کمک مدل‌سـازی موضـوعی طراحی کرده ا ست. این مقاله روش‌های مر سوم تو صیه‌گرها را با مدل سازی مو ضوعی ترکیب کرده و ساختار نهان کاربرها و موردها را به صورتی توصیف‌پذیر فراهم آورده است.

مقاله‌ی [39] تلفیقی از مدل‌سـازی موضـوعی و خلاصـه‌سـازی متن می‌باشـد. موضـوعات را به صورت سل سله مراتبی از مجموعه‌ی ا سناد ا ستخراج می‌کند. از الگوریتم‌های تحلیل شبکه ا ستفاده می‌کند. از اسناد شبکه‌ای از تاثیر کلمات می‌سازد. کلمات خلاصه‌سازی که بیان کننده‌ی اصل موضوع هسـتند و بیشـترین تاثیر را در شـبکه دارند را می‌یابد. برای سـاخت گراف تاثیر کلمات، کلمات نقش نودها را دارند. ارتباط بین دو کلمه بر اسـاس هم‌رخدادی آن‌ها در کل مجموعه می‌باشـد. گراف بیان کننده‌ی این اسـت که اگر کلمه‌ی x دیده شـد، چه میزان امکان دارد که به معنای کلمه‌ی y بدون دیدن آن برسیم. علاوه بر پیدا کردن کلمات موضوعی که بیشترین تاثیر را بر دیگر کلمات شبکه دارند،

[۱] Language model adaption
[۲] community

ارتباط میان مو ضوعات را نیز می‌یابد. کلمات خو شه‌بندی می شوند و سل سله مراتبی برای مو ضوعات ساخته می‌شود.

مدل سازی مو ضوعی می‌تواند به عنوان یک پایگاه دانش عمل کند و در کاربردهایی که مفاهیم جای کلمات را گرفته‌اند می‌تواند نقشی ایفا کند. این مفاهیم می‌توانند از پایگاه‌های دانشی هم‌چون وردنت و یا ویکی‌پدیا به دست آیند. مقاله‌ی [40] یکی از این مقالات می‌باشد. هم‌چنین می‌توان از مو ضوعات به د ست آمده در ساخت پایگاه‌های دانش کمک گرفت. از پایگاه‌های دانش نیز می‌توان در ساخت مو ضوعات کمک گرفت و به جای نگاه کلمه‌ای به متن، نگاهی مفهومی با عنا صر متن دا شت و از این نگاه در مدل سازی مو ضوعی ا ستفاده کرد. برخورد مفهومی با متن بی شتر در متون کوتاه حائز اهمیت می‌شوند. چرا که این متن‌ها کلمات کافی برای اعمال هویت خود ندارند. باید از همان چند کلمه و مفاهیم آن‌ها در تصمیم‌گیری‌ها(مثلا طبقه‌بندی) راجع به متن استفاده کرد. مقاله‌ی [41] از پایگاه دانشی احتمالاتی برای ساخت مفاهیم اسناد متن کوتاه استفاده می‌کند. پایگاه دانشی که استفاده می‌کند، Probase [42] و [43] نام دارد. این پایگاه دانش مفاهیم بسیار و غنی‌ای دارد که نزدیک به تصور بشری می‌باشند. پایگاه‌های دانشی هم‌چون وردنت [44]، ویکی‌پدیا [45] و Freebase [46] توسط خبره‌ها و نیروهای انسانی و یا تلاش‌های گروهی ساخته شده‌اند. تلاش‌های بسیاری صورت گرفته است که پایگاه‌های دانش به صورت خودکار ساخته شوند. KnowItAll [47]، TextRunner [48]، WikiTaxonomy [49] و YAGO [50] از جمله‌ی این تلاش‌ها می‌باشند. دو مشکل اصلی که بسیاری از پایگاه‌های دانش دارند، اندازه و جامعیت آن‌ها می‌باشد. به عبارتی پوشش و ریزدانگی مفاهیم را به خوبی ندارند. مثلا Freebase بعد از ۲۵ سال تلاش، ۲۰۰۰ دسته از مفاهیم دارد. Probase شامل میلیون‌ها مفهوم می‌باشد که دائما از میلیاردها صفحه‌ی وب به دست می‌آید. روابطی که میان مفاهیم در Probase تعریف شده بسیار می‌باشد که شاید مهم‌ترین آن‌ها شباهت میان مفاهیم باشد.

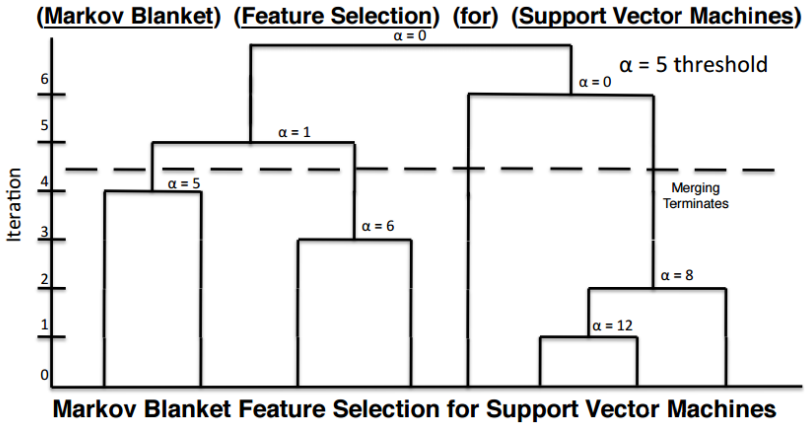
۶. نتیجه‌گیری

در این ف صل نگاهی اجمالی به برخی از روش‌های موجود در مدل سازی مو ضوعی انداخته شد. تلاش‌ها و نگاه‌های مختلفی برای بهبود مدل‌سازی موضوعی وجود دارد. در این فصل از هر نگاهی نمونه‌ای معرفی شد که جدیدترین و بهترین نتایج اخیر را ارائه می‌کند. این نگاه از ساخت گراف از

اسناد و استفاده از آن برای مدل‌سازی تا استفاده از عبارت‌ها، موجودیت، دانش‌های داخلی و خارجی و استفاده‌های مدل‌سازی موضوعی را شامل می‌شود.

یکی از راه‌های بهبود مدل سازی استفاده از چندگرم‌ها و عبارت‌ها در مدل سازی موضوعی می‌باشد. یکی از کارهایی که از عبارات برای مدل سازی مو ضوعی بهره گرفته، [15] می‌با شد. از قسمتی از ایده‌ی این مقاله که مربوط به ا ستخراج عبارت‌ها می‌باشـــد، برای پیش‌برد این تحقیق بهره گرفته شـــده اســـت. روش این مقاله هم از نظر کیفیت و هم از نظر عملکرد بسیار کارا می‌با شد و می‌تواند در مقیاس‌پذیری سیستم کمک کننده با شد. عبارت‌ها در این مقاله چند ویژگی دارند. لیستی از عبارت‌ها نشان‌دهنده‌ی یک مو ضوع منسجم هستند. عبارت‌هایی که به د ست می‌آیند نزدیک به فهم بشری هستند. این روش موضوعات را با دقت و سرعت و پیچیدگی نزدیک به LDA به دست می‌آورد، با این تفاوت که به جای ا ستفاده از یک گرم برای مو ضوعات، از عبارات برای مدل سازی و نمایش مو ضوعات ا ستفاده می‌کند. این مقاله یک روش جدید کاوش عبارت برای تقســـیم‌بندی متن به چندین عبارت یک یا چند کلمه‌ای و روشـــی جدید برای مدل سازی مو ضوعی بر روی بخش‌های ایجاد شده، ارائه کرده ا ست. به د ست آوردن عبارات در این مقاله به دو قسمت اصـلی تقسـیم می‌شـود. پیدا کردن عبارات مکرر کاندید و تعداد تکرار آن‌ها و در مرحله‌ی بعد ترکیب کلمات هر سـند برای ساخت عبارات با کیفیت. قسمت کاوش عبارات مکرر را می‌توان در به د ست آوردن تعداد تکرار کلمات پشت سر همی که از یک حداقلی بیشتر در کنار هم تکرار شده‌اند، خلاصه کرد. دو اصل می‌تواند در کارایی کاوش موثر باشد. اول

این‌که اگر عبارت P مکرر نبا شد، هر عبارت بزرگ‌تری که شامل P با شد نیز مکرر نی ست. دوم هم این‌که اگر سندی هیچ عبارت مکرری به طول n نداشته باشد، آنگاه سند عبارت مکرری با طول بیشتر از n نخواهد داشت. نوآوری اصلی در کاوش عبارات، نحوه‌ی ساخت عبارات می‌باشد. عبارات از پایین به بالا ساخته می‌شوند. در هر چرخه به صورت حریصانه، منا سب‌ترین جفت عبارت کاندید با هم ترکیب می شوند. این ترکیب از عبارات یک یا چند کلمه‌ای به د ست می‌آید. از آن جا که تنها عباراتی که از تقسیم‌بندی سند به دست می‌آیند، معتبر می‌باشند، عباراتی که یک حداقل آستانه‌ای را دارند پذیرفته می‌شوند. شکل ۳-۸ نشان‌دهنده‌ی نحوه‌ی ساخت عبارات از پایین به بالا می‌باشد و برگرفته از مقاله‌ی [15] می‌باشد. تمام عبارات ترکیب شده مکرر می‌باشند.



ساخت پایین به بالا عبارت‌ها بر روی عنوان‌های مقالات علوم کامپوتر

عبارت‌های پرتکرار معنادار به د ست آمدند. این کلمات در کنار یکدیگر بسیار ظاهر شده‌اند. یک کلمه‌ی عبارت تکمیل کننده‌ی معنای کلمه‌ی همسایه می‌باشد. می‌توان مدعی شد قرار گرفتن کلمات یک عبارت در یک موضوع می‌تواند باعث بالا رفتن کیفیت موضوعات شود. این ایده مطرح می‌شود که از کلمات یک عبارت به عنوان ارتباط-بایسته[1] در مدل‌سازی موضوعی به روش AMC استفاده شود. در این تحقیق از یک عبارت N کلمه‌ای، N-1 ارتباط-بایسته ساخته شد. قبلا ارتباط-بایسته‌ها از کلمات موضوعات به دست می‌آمدند. کلماتی که در موضوع‌های بسیاری با هم ظاهر می‌شوند این دانش را منتقل می‌کنند که این کلمات بهتر است در یک موضوع قرار گیرند. برای درک بهتر چگونگی عملکرد الگوریتم به بخش ۲-۴-۲ مراجعه شود.

دو اجرای متفاوت انجام شد. در هر دو اجرا مجموعه‌ی سندی به تعداد مشخصی خو شه شکسته می شود. از خو شه‌بندی مو ضوعی برای خو شه‌بندی ا ستفاده می شود. برای مدل سازی مو ضوعی مجموعه از مدل سازی مو ضوعی خو شه‌ها ا ستفاده می شود و مو ضوعات مجموعه در قالب مو ضوعات خو شه‌ها نمایش داده می شوند. در بخش ۳-۶-۱ نشان داده شد که برای استفاده از دانش خوشه‌ها چگونه می‌توان استفاده کرد تا یک خوشه مدل‌سازی موضوعی شود. باید خوشه‌ها در ابتدا با یک روش مدل‌سازی موضوعی، مدل‌سازی شوند. موضوعات خوشه‌ها به دست می‌آید. از موضوعات دانش در قالب ارتباط-بایسته‌ها و ارتباط-نبایسته‌ها استخراج می‌شوند. با کمک دانش به دست آمده از دیگر

[1] Must-link

خوشـه‌ها، خوشـه‌ی مورد نظر مدل‌سـازی موضـوعی می‌شـود. یک ایده در این‌جا اسـتفاده از کلمات عبارت‌ها به عنوان ارتباط-بایسته در کنار ارتباط-بایسته‌های به دست آمده از موضوعات می‌باشد. در یک اجرا این ایده عملیاتی شد. فرض کنید قرار است خوشه‌ی iام مدل‌سازی موضوعی شود. در فاز اول از موضوعات اولیه‌ی ساخته شده از تمام خوشه‌ها به جز خوشـه‌ی iام اسـتفاده می‌شـود و کلماتی که در کنار یکدیگر ارتباط-بایسـته‌ها را شـکل می‌دهند در یک فایل ثبت می شوند. در فاز بعد عبارت‌های پر تکرار و معنادار خوشه‌ی iام با استفاده از روش مقاله‌ی [15] به دست می‌آیند. هر عبارت امتیازی دارد که متناسب با تکرار عبارت در مجموعه می‌باشد. عبارت‌ها بر اساس امتیاز شان مرتب شده‌اند. N عبارت با امتیاز بیشتر برای استخراج ارتباط-بایسته استخراج می شوند. عدد N متناسب و تقریبا برابر با تعداد ارتباط-بایسته‌های استخراج شده از موضوعات خوشه‌ها می‌باشد. در نهایت از دو نوع ارتباط-بایسته استفاده شد و موضوعات جدید خوشه با استفاده از دانش حاصل از عبارت‌های خوشه و دانش حاصل از موضوعات دیگر خوشه‌ها به دست آمد.

فرض کنید قرار است موضوعات خوشه‌ی iام را با استفاده از دانش حاصل از ارتباط-بایسته‌ها به دست آورید. برای استخراج ارتباط-بایسته‌ها، موضوعات تمام دیگر خوشه‌ها باید به حافظه منتقل شوند تا دانش از آن‌ها استخراج شود. چرا که دانش، ارتباط دوتایی کلماتی است که در موضوعات خوشه‌ها بسیار با یکدیگر ظاهر شده‌اند. خوشه‌ی i به تمام دیگر خوشه‌ها وابسته می شود. این وابستگی مشکلاتی به همراه دارد. اول اینکه حافظه‌ی مصرفی را بالا می‌برد. چرا که اطلاعات موضوعات خوشه‌ها باید به حافظه منتقل شوند. هرچه تعداد موضوعات بیشتر باشد حجم اطلاعات موضوعات بیشتر می‌شود. نمونه‌ای از این اطلاعات، توزیع موضوعات برای هر سند می‌باشد. توزیع هر موضوع در یک سند عددی اعشاری می‌باشد. جمع احتمال توزیع‌ها ۱ می‌شود. چون تعداد موضوعات زیاد است، به غیر از اندک تعداد موضوعی که احتمال بالایی دارند، بقیه‌ی موضوع‌ها عددی کوچک می شوند که نیاز به اعشار بالا برای غیر صفر شدن آن‌ها ایجاد می‌شود. این وابستگی به تمام خوشه‌ها امکان اجرای هر خوشه بر روی یک ماشین را با مشکل مواجه می‌کند. این ایده مطرح می‌شود که جایگزینی برای استخراج ارتباط-بایسته‌ها یافت شود.

در اجرای دوم تلاش شد وابستگی یک خوشه از خوشه‌های دیگر برای استخراج دانش از بین برود. وابستگی در استخراج دانش بود. برای اسـتخراج دانش برای یک خوشـه باید موضـوعات تمام دیگر خوشـه‌ها در حافظه بارگزاری شـوند. البته می‌توان با ایجاد محدودیت در انتخاب خوشه‌های مشابه تعداد خوشه‌هایی که برای استخراج دانش استفاده می‌شوند را کاهش داد. ولی باز هم حجم موضوعات هر خوشه می‌تواند یک سربار باشد. تصمیم گرفته شد در اجرای دوم به جای استفاده از دانش حاصل از عبارت‌ها در کنار دانش حاصل از موضوعات دیگر خوشه‌ها برای بهبود مدل‌سازی موضوعی، تنها از دانش عبارت‌ها استفاده شود. با استفاده از این سیاست دانش هر خوشه وابسته به خود خوشه می شود. از عبارت‌های پرتکراری که در سندهای خوشه وجود دارند به عنوان دانش استفاده می‌شود. هر خوشه می‌تواند مستقل از دیگر خوشه‌ها مدل سازی موضوعی شود. این سیاست باعث می شود در مصرف منابعی همچون حافظه و پردازش‌گر صرفه‌جویی شود. هم‌چنین می‌توان خوشه‌ها را به صورت همزمان مدل سازی موضوعی کرد. در قسمت نتایج دقت دو اجرا مقایسه شده‌اند. اجرایی که دانش از عبارت‌ها در کنار دانش موضوعات استفاده می شود و زمانی که دانش تنها از دانش عبارت‌های خوشه می‌باشد.

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.

[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.

[9] Y. Sun, J. Han, J. Gao, and Y. Yu, “iTopicModel: Information Network-Integrated Topic Modeling,” in *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, 2009, pp. 493–502.

[10] H. Kim, Y. Sun, J. Hockenmaier, and J. Han, “ETM: Entity Topic Models for Mining Documents Associated with Entities.,” *ICDM*, 2012.

[11] C. Wang, M. Danilevsky, J. Liu, N. Desai, H. Ji, and J. Han, “Constructing Topical Hierarchies in Heterogeneous Information Networks,” in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013, pp. 767–776.

[12] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, “Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks,” pp. 1271–1279, 2011.

[13] Y. Sun, Y. Yu, and J. Han, “Ranking-based Clustering of Heterogeneous Information Networks with Star Network Schema,” in *Proceedings of the 15th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 2009, pp. 797–806.

[14] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han, "A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 437–445.

[15] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable Topical Phrase Mining from Text Corpora," *CoRR*, vol. abs/1406.6, 2014.

[16] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.

[17] H. Mahmoud, *Polya Urn Models*, 1st ed. Chapman & Hall/CRC, 2008.

[18] Z. Chen and B. Liu, "Mining Topics in Documents: Standing on the Shoulders of Big Data," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1116–1125.

[19] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Discovering Coherent Topics Using General Knowledge," in *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, 2013, pp. 209–218.

[20] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging Multi-domain Prior Knowledge in Topic Models," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013, pp. 2071–2077.

[21] A. Haghighi and L. Vanderwende, "Exploring Content Models for Multi-document Summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 362–370.

[22] J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin, "Word Sense Induction for Novel Sense Detection," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 591–601.

[23] B. Zhao and E. P. Xing, "HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation," in *Advances in Neural Information Processing Systems 20*, J. c. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2007, pp. 1689–1696.

[24] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. 2009, pp. 288–296.

[25] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic Evaluation of Topic Coherence,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100–108.

[26] J. T. Chien and M. S. Wu, “Adaptive Bayesian Latent Semantic Analysis,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 1, pp. 198–207, Jan. 2008.

[27] M. S. Wu, H. S. Lee, and H. M. Wang, “Exploiting semantic associative information in topic modeling,” in *Spoken Language Technology Workshop (SLT), 2010 IEEE*, 2010, pp. 384–388.

[28] H. M. Wallach, “Topic Modeling: Beyond Bag-of-words,” in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 977–984.

[29] X. Wang, A. McCallum, and X. Wei, “Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval,” in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, 2007, pp. 697–702.

[30] X. Wei and W. B. Croft, “LDA-based Document Models for Ad-hoc Retrieval,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 178–185.

[31] A. Kotov, V. Rakesh, E. Agichtein, and C. K. Reddy, “Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings,” A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, Eds. Cham: Springer International Publishing, 2015, pp. 635–647.

[32] L. Li, B. Roth, and C. Sporleder, “Topic Models for Word Sense Disambiguation and Token-based Idiom Detection,” in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1138–1147.

[33] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, “Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1271–1279.

[34] Q. Mei, D. Cai, D. Zhang, and C. Zhai, “Topic Modeling with Network Regularization,” in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 101–110.

[35] M. H. Alam and S. Lee, “Semantic Aspect Discovery for Online Reviews,” in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 2012, pp. 816–821.

[36] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang, “Multi-aspect Sentiment Analysis for Chinese Online Social Reviews Based on Topic Modeling and HowNet Lexicon,” *Know.-Based Syst.*, vol. 37, pp. 186–195, Jan. 2013.

[37] Z. Yin, L. Cao, Q. Gu, and J. Han, “Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling,” *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 63:1–63:21, Sep. 2012.

[38] C. Wang and D. M. Blei, “Collaborative Topic Modeling for Recommending Scientific Articles,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 448–456.

[39] C. Wang, X. Yu, Y. Li, C. Zhai, and J. Han, “Content Coverage Maximization on Word Networks for Hierarchical Topic Summarization,” in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, 2013, pp. 249–258.

[40] F. Wang, Z. Wang, Z. Li, and J.-R. Wen, “Concept-based Short Text Classification and Ranking,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 1069–1078.

[41] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, “Short Text Conceptualization Using a Probabilistic Knowledgebase,” in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, 2011, pp. 2330–2336.

[42] W. Wu, H. Li, H. Wang, and K. Zhu, “Towards a Probabilistic Taxonomy of Many Concepts,” Microsoft Technical Report, Mar. 2011.

[43] W. Wu, H. Li, H. Wang, and K. Q. Zhu, “Probase: A Probabilistic Taxonomy for Text Understanding,” in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 481–492.

[44] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[45] D. B. Lenat and R. V Guha, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[46] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, “Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge,” in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 2008, pp. 1247–1250.

[47] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, “Web-scale Information Extraction in Knowitall: (Preliminary Results),” in *Proceedings of the 13th International Conference on World Wide Web*, 2004, pp. 100–110.

[48] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, “Open Information Extraction from the Web,” *Commun. ACM*, vol. 51, no. 12, pp. 68–74, Dec. 2008.

[49] S. P. Ponzetto and M. Strube, “Deriving a Large Scale Taxonomy from Wikipedia,” in *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, 2007, pp. 1440–1445.

[50] F. M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A Core of Semantic Knowledge,” in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 697–706.