

## بررسی رویکردهای متن کاوی و عملکرد آن در کشف و استخراج موضوع

فاطمه زرمهر: دانشجوی دکتری علم اطلاعات و دانش شناسی، دانشگاه اصفهان، اصفهان، ایران

\*علی منصوری: استادیار گروه علم اطلاعات و دانش شناسی، دانشگاه اصفهان، اصفهان، ایران (نویسنده مسئول) [a.mansouri@edu.ui.ac.ir](mailto:a.mansouri@edu.ui.ac.ir)

حسین کارشناس: استادیار گروه هوش مصنوعی، دانشگاه اصفهان، اصفهان، ایران

### چکیده

دریافت: ۱۳۹۸/۱۰/۲۷

پذیرش: ۱۳۹۸/۱۲/۲۱

**زمینه و هدف:** در این پژوهش چهار روش متن کاوی بررسی می شود و بر درک و شناسایی خصوصیات و محدودیت های آن ها در کشف موضوع تمرکز می کند. این چهار روش عبارتند از (۱) تجزیه و تحلیل معنایی پنهان (LSA) (۲) تحلیل معنایی پنهان احتمالاتی (PLSA)، (۳) تخصیص دیریکله پنهان (LDA) و (۴) مدل سازی موضوعی همبسته (CTM).

**روش پژوهش:** پژوهش حاضر از نوع کتابخانه ای است که در آن، ادبیات حوزه متن کاوی و مدل سازی موضوعی مرور و تحلیل شده است. **یافته ها:** تجزیه و تحلیل معنایی پنهان می تواند برای تشخیص موضوعات خاص و منحصر به فرد در مدارکی که تنها به یک موضوع پرداخته اند استفاده شود. سه روش دیگر متن کاوی، بر موضوعات و گرایش کلی متن متمرکز هستند. تحلیل معنایی پنهان احتمالاتی برای مدارکی که به یک موضوع پرداخته اند قابل استفاده است اما برخلاف تجزیه و تحلیل معنایی پنهان، این روش در کشف موضوعات و مضامین کلی متن کاربرد دارد. در حالی که تخصیص دیریکله پنهان در مورد مدارکی که به چندین موضوع پرداخته اند کاربرد بیشتری دارد. روش مدل سازی موضوعی همبسته می تواند در تشخیص ارتباط بین دسته های موضوعی مختلف استفاده شود.

**نتیجه گیری:** رویکردهای متن کاوی به خاطر بهره گیری از تحلیل معنایی در کشف و استخراج موضوع متون مناسب است

**کلیدواژه:** متن کاوی، مدل سازی موضوعی، تحلیل معنایی، کشف موضوع

تعارض منافع: گزارش نشده است.

منبع حمایت کننده: حامی مالی نداشته است.

**شیوه استناد به این مقاله**

**APA:** Zarmehr, F., Mansouri, A., Karshenas, H., (2020). A review of text mining approaches and their function in discovering and extracting a topic. *Human Information Interaction*. 7(1); 15-26 (Persian)

**Vancouver:** Zarmehr, F., Mansouri, A., Karshenas, H. A review of text mining approaches and their function in discovering and extracting a topic. *Human Information Interaction*. 2020; 7(1): 15-26 (Persian)



انتشار مجله تعامل انسان و اطلاعات با حمایت مالی دانشگاه قزوین انجام می شود.

انتشار این مقاله به صورت دسترسی آزاد مطابق با **CC BY-NC-SA 3.0** صورت گرفته است.

## A review of text mining approaches and their function in discovering and extracting a topic

**Fatemeh Zarmehr:** PhD student in Information Science and Knowledge, University of Isfahan, Isfahan, Iran

**\*Ali Mansouri:** Assistant Professor, Department of Information Science, University of Isfahan, Isfahan, Iran (Corresponding Author) [a.mansouri@edu.ui.ac.ir](mailto:a.mansouri@edu.ui.ac.ir)

**Hossein Karshenas:** Assistant Professor, Department of Artificial Intelligence, University of Isfahan, Isfahan, Iran

Received: 17/01/2020

Accepted: 11/03/2020

### Abstract

**Background and aim:** Four text mining methods are examined and focused on understanding and identifying their properties and limitations in subject discovery.

**Methodology:** The study is an analytical review of the literature of text mining and topic modeling.

**Findings:** LSA could be used to classify specific and unique topics in documents that address only a single topic. The other three text mining methods focus on topics and general partiality of the text. PLSA is applicable to documents dealing with a topic, unlike the LSA, it is used to discover general themes and contexts. However, LDA is more applicable to documents that address several issues. The CTM, method can be used to identify relationship between different subject categories.

**Conclusion:** Text mining tactics are suitable for employing analysis in discovering and extracting the text subjects.

**Keywords:** Text mining, Topic Modeling, Semantic Analysis, Topic Discovery.

*Conflicts of Interest:* None

*Funding:* None.

### How to cite this article

**APA:** Zarmehr, F., Mansouri, A., Karshenas, H.,(2020). A review of text mining approaches and their function in discovering and extracting a topic. *Human Information Interaction*. 7(1); 15-26 (Persian)

**Vancouver:** Zarmehr, F., Mansouri, A., Karshenas, H. A review of text mining approaches and their function in discovering and extracting a topic. *Human Information Interaction*. 2020; 7(1): 15-26 (Persian)



لهال<sup>۱۳</sup>، ۲۰۰۹). این فن نه تنها به تجزیه و تحلیل متن‌های حجیم همچون کتاب‌ها و مقالات مجلات دانشگاهی پرداخته، بلکه متن پست‌های الکترونیکی، توثیقات و یا نظرات اعلام شده در شبکه‌های اجتماعی را نیز پوشش می‌دهد؛ بنابراین هر نوع فایل متنی غیر ساختارمند و ساختارمند به کمک متن‌کاوی قابل تجزیه و تحلیل است (دین<sup>۱۴</sup>، ۲۰۱۴). از جمله کاربردهای متن‌کاوی می‌توان به کشف موضوع، مدیریت ارتباط با مشتری و تبلیغات هدفمند و ... اشاره کرد.

مدل‌سازی موضوعی یکی از اشکال تجزیه و تحلیل متن به منظور بررسی رابطه بین کلمات داخل مدرک است، جایی که کلمات در کنار یکدیگر تشکیل‌دهنده درون‌مایه و موضوع اصلی متن است. با توجه به سه هدف اصلی تکنیک مدل‌سازی موضوعی مبنی بر کشف موضوعات پنهان (بیترمن و فیشر، ۲۰۱۸؛ هاگن، ۲۰۱۸؛ فیگرولا، گارسیا و پیتو، ۲۰۱۷ و محمدیان، ۱۳۹۳)، تفسیر اسناد بر اساس موضوعات (دین فانگ و همکارانش<sup>۱۵</sup>، ۲۰۱۸) و نهایتاً سازمان‌دهی و طبقه‌بندی متون (سلوی و همکاران<sup>۱۶</sup>، ۲۰۱۹؛ سهرابی و همکاران، ۲۰۱۷) می‌توان نوشتارهای مربوط به این حوزه را نیز به سه دسته عمده تقسیم نمود. باین‌حال استفاده از این تکنیک فراتر از سه هدف مذکور در حال پژوهش و ارزیابی است. سایر کاربردهای این تکنیک با استناد به پژوهش‌های انجام شده عبارتند از: ۱. بازیابی اطلاعات (کین<sup>۱۷</sup>، ۲۰۱۶؛ مین چول کیم<sup>۱۸</sup>، ۲۰۱۸)، ۲. تحلیل سیر تحولی مفهومی در طول دوره‌های زمان (کوراتا و همکارانش<sup>۱۹</sup>، ۲۰۱۸)، ۳. غنی‌سازی فهرست واژگان (رانی، دهار و یاس<sup>۲۰</sup>، ۲۰۱۷).

با توجه به اهمیت نیاز به تجزیه و تحلیل و درک داده‌های متنی، در این مقاله سعی شده است روش‌های عمده متن‌کاوی در کشف موضوع مورد مطالعه و بررسی قرار گیرد. این روش‌ها از رویکردهای ساده و اولیه مدل‌سازی موضوعی است که از روش‌های شناسایی تفاوت ظاهری تا رویکردهای جدیدتر مبتنی بر احتمالات، متنوع هست. روش‌های احتمالاتی اخیر به محدودیت‌های روش‌های قبلی غلبه کرده و به محققان این امکان را داده است تا الگوی تولید داده‌های متنی را طی فرآیند مدل‌سازی متن فراهم کنند. در این پژوهش ویژگی‌ها و محدودیت‌های هر روش در فرآیند تشخیص موضوع مورد بررسی قرار گرفته است. هر یک از چهار روش استخراج متن معرفی شده دارای ویژگی‌های منحصر به فرد

نظام‌های اطلاعاتی مدرن این امکان را فراهم کرده که به حجم زیادی از داده‌ها دسترسی ایجاد شود. بخش عمده‌ای از این داده‌ها، داده‌های ساختارمند هستند که می‌توانند با استفاده از نرم‌افزارهای پایگاه داده‌ای، به راحتی پردازش و در دسترس قرار گیرند. باین‌حال به‌طور فزاینده، مقدار زیادی از داده‌ها مثل داده‌های متنی بدون ساختار هستند و تجزیه و تحلیل و درک آن‌ها را دچار مشکل کرده است. تجزیه و تحلیل دستی داده‌های متنی بدون ساختار عملاً ممکن نیست، و همین امر نیاز به ابزارهای جدیدی برای جستجو، سازمان‌دهی و فهم حجم بسیار اطلاعات درون آن را اجتناب‌ناپذیر کرده است (هوانگ و همکاران<sup>۱</sup>، ۲۰۱۷). زیرا نظام‌های سنتی سازمان‌دهی و بازیابی اطلاعات که عمدتاً برای منابع چاپی طراحی شده‌اند، قادر به پاسخگویی در این زمینه و بستر نیستند (سورگل<sup>۲</sup>، ۲۰۰۴؛ بابو<sup>۳</sup>، ۲۰۱۲؛ هوانگ و همکاران، ۲۰۱۷؛ ابوسبا کاظمینی، ۱۳۹۰؛ نوروزی و خویدکی، ۱۳۹۳؛ ابراهیم‌زاده و حسینی بهشتی، ۱۳۹۵؛ خادمیان و کوکی، ۱۳۹۷). بنابراین برای فائق آمدن به این چالش، پرداختن به رویکردی جدید در سازمان‌دهی منابع اطلاعاتی لازم است.

رویکرد تخصیص خودکار موضوع به متون بر اساس روش‌های متن‌کاوی و الگوریتم‌های اختصاصی این روش، رویکردی جدید است که در پژوهش‌های مختلف سعی شده است جنبه‌هایی از نیاز کتابداران و متخصصان موضوعی را به‌منظور شناسایی موضوعات و تفسیر منابع و کارکرد آن در بازیابی بهینه منابع مورد بررسی قرار دهند (بیترمن و فیشر<sup>۴</sup>، ۲۰۱۸؛ سان آندرس و همکارانش<sup>۵</sup>، ۲۰۱۸؛ هاگن<sup>۶</sup>، ۲۰۱۸؛ فانگ<sup>۷</sup> و همکارانش، ۲۰۱۸؛ فیگرولا و همکارانش<sup>۸</sup>، ۲۰۱۷؛ کین<sup>۹</sup>، ۲۰۱۶؛ استیور<sup>۱۰</sup> و همکارانش، ۲۰۰۴؛ نیومن<sup>۱۱</sup> و همکارانش، ۲۰۰۷؛ محمدیان، ۱۳۹۳؛ رانی، دهار و ویاس<sup>۱۲</sup>، ۲۰۱۷). در همین راستا، روش‌های متن‌کاوی به‌منظور خودکار سازی فرآیند تجزیه و تحلیل این نوع داده‌ها توسعه پیدا کرد. متن‌کاوی یا تجزیه و تحلیل متن یکی از حوزه‌های خاص داده‌کاوی است. از آنجاکه بیشتر اطلاعات (بیش از ۸۰٪) به‌صورت متن ذخیره شده‌اند، و حاوی اطلاعات ارزشمند و نهفته‌ای هستند، اعتقاد بر این است که متن‌کاوی ارزش بالقوه بالایی دارد (گوپتا و

<sup>1</sup> Hwang & et al

<sup>2</sup> Soergel

<sup>3</sup> Babu

<sup>4</sup> Bitterman, & Fischer

<sup>5</sup> Sanandres

<sup>6</sup> Hagen

<sup>7</sup> Fang

<sup>8</sup> Figuerola

<sup>9</sup> Kinyanjui

<sup>10</sup> Steyver

<sup>11</sup> Newman

<sup>12</sup> Rani, , Dhar & Vyas

<sup>13</sup> Gupta & Lehal

<sup>14</sup> Dean

<sup>15</sup> Debin Fang, Haixia, Baojun, Xiaojun

<sup>16</sup> Selvi & et al

<sup>17</sup> Cain

<sup>18</sup> Meen Chul kim

<sup>19</sup> Kurata & et al

<sup>20</sup> Rani, Dhar, Vyas

فانگ و همکارانش، ۲۰۱۸). زیرا در روش های متن کاوی مختلف رویکردهای متفاوتی برای مقابله با مترادفات و چند معنی بودن کلمات به کار گرفته شده است که در توضیح هر روش به آن اشاره می شود. در بخش بعدی به مدل فضای برداری اشاره خواهد شد که مفهوم اساسی دیگر در رویکردهای متن کاوی هست.

### مدل فضا برداری<sup>۷</sup> (VSM)

یکی از مهم ترین کارهایی که در زمینه استخراج ویژگی ها از اسناد و همچنین نحوه مقایسه این ویژگی ها با یکدیگر انجام شد، نمایش اسناد در قالب شکل برداری است. یعنی هر یک از کلمات متن (بدون تکرار) به عنوان یک مؤلفه برداری در نظر گرفته می شود و در کنار سایر کلمات متن تشکیل یک فضای برداری را می دهند. به این ترتیب یک متن معمولی که ممکن است شامل پانصد کلمه باشد به صورت یک بردار حدوداً دویست بعدی مدل می شود که البته تصور هندسی آن مطمئناً ساده نیست اما در عوض جنبه های محاسباتی قابل درک و ساده ای را به وجود می آورد که باعث می شود به صورت منطقی و قابل قبول دو متن، به عنوان دو بردار با یکدیگر مقایسه شوند. در مدل فضای برداری، یکی از روش های مقایسه، استفاده از کسینوس زاویه دو برداری است که بازنمایش متن هستند (تصویر ۱). به همین دلیل به این معیار، شباهت کسینوسی گفته می شود (ونکات<sup>۸</sup> و همکاران، ۲۰۱۸). یکی از روش هایی که در رابطه با نحوه وزن دهی به هر یک از مؤلفه ها و یا به عبارت دیگر بزرگی و میزان هر یک از ابعاد به کار می رود، معیار TF-IDF است که خود ترکیبی از دو روش وزن دهی TF<sup>۹</sup> و IDF<sup>۱۰</sup> است. ویژگی TF بیانگر تکرار کلمه است. زیرا فرض بر این است که کلماتی که در متن به تعداد زیاد تکرار شوند احتمالاً ارزش معنایی زیادی دارند. ویژگی IDF نیز معکوس تکرار در متون است. یعنی فرض بر این گرفته شده که اگر کلمه در یک متن بکار گرفته شده و این کلمه در متون دیگر تقریباً کمیاب باشد این مسئله نشان دهنده ارزش بیشتر این کلمه در متنی است که به کار برده شده است. با داشتن نسبت TF و IDF مقدار TFIDF حاصل ضرب این دو مقدار خواهد بود. با محاسبه مقدار می توان مقداری که برای هر بعد از بردار متن لازم است را حساب کرده و در نهایت با مقایسه معیار کسینوسی نسبت به شباهت داشتن و یا شباهت نداشتن متون با یکدیگر تصمیم گرفت (نادژدا و الکسی<sup>۱۱</sup>، ۲۰۱۸).

خود است و استفاده مؤثر از این ویژگی ها نقش مهمی در به دست آوردن عملکرد مذکور دارد.

### مفاهیم و فرضیات

هر چهار روش اصلی متن کاوی که در ادامه توضیح داده خواهد شد بر اساس یک سری فرضیات پایه و همچنین مدل فضای برداری است.

### مدرک و کیسه کلمات<sup>۱</sup>

واحد اصلی متن، کلمه است. کلمات حاوی کاراکترها هستند که معانی و مفاهیم از آن ها به وجود می آید. از ترکیب کلمات با قوانین دستور زبان، جمله ساخته می شود. جملات، واحدهای اصلی کارکردی در متن است که حاوی اطلاعاتی در مورد عملکرد بعضی موضوعات است. پاراگراف یا بندها، واحدهای اصلی تشکیل ساختار متن است که به موضوع خاصی می پردازد (استرانک<sup>۲</sup>، ۲۰۰۷). هر چه طول متن در ساختار بخش ها، فصل ها و متن اصلی افزایش یابد، اشکال ساختاری اضافی به وجود می آید و در نهایت مجموعه ای از مدارک یا اسناد که به آن ها پیکره<sup>۳</sup> نیز گفته می شود، شکل می گیرد.

در مطالعات متن کاوی برای تشخیص موضوع، مدرک به عنوان واحد اصلی تجزیه و تحلیل استفاده می شود، زیرا یک نویسنده معمولاً کلیت یک مدرک را در مورد یک موضوع واحد می نویسد. این مدرک می تواند یک مقاله یا یک کتاب باشد که بستگی به هدف نویسنده و نوع تجزیه و تحلیل وی از موضوع دارد. در بعضی موارد، یک مدرک ممکن است حاوی فقط یک فصل یا یک پاراگراف ساده و یا حتی یک جمله باشد. در مطالعات متن کاوی، گاهی ساختار نحوی یک جمله یا پاراگراف به عمد نادیده گرفته می شود تا بتوان به طور مؤثر فرایندهای بعدی را اجرا نمود. بنابراین یک جمله صرفاً به عنوان مجموعه ای از کلمات یا به اصطلاح کیسه ای از کلمات<sup>۴</sup> در نظر گرفته می شود. این ایده که یک جمله، صرفاً کیسه ای از کلمات بدون در نظر گرفتن ساختار اضافه است، فرض اساسی در متن کاوی برای تشخیص موضوع است (فرنگ و همکارانش، ۲۰۱۸).

از آنجا که گاهی ساختار نحوی متن در متن کاوی نادیده گرفته می شود، ترتیب کلمات بدون تأثیر در نتیجه تحلیل می تواند تغییر کند. مفهوم کیسه کلمات به معنای قابلیت جابه جایی و تعویض کلمات در نظر گرفته می شود (همان، ۲۰۱۸). دو مشکل اساسی که در تجزیه و تحلیل متن به وجود می آید، "مترادفات"<sup>۵</sup> و "کلمات چند معنی"<sup>۶</sup> است. این مشکل به طرق مختلف برطرف شده است

<sup>7</sup> Vector space model

<sup>8</sup> Venkat

<sup>9</sup> Term Frequency

<sup>10</sup> Inverse Document Frequency

<sup>11</sup> Nadezhda & Aleksey

<sup>1</sup> Bag of word

<sup>2</sup> Strunk

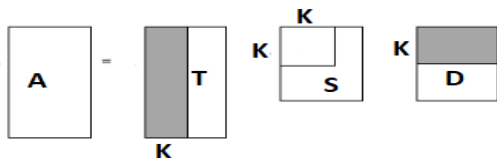
<sup>3</sup> corpus

<sup>4</sup> Bag-of- word

<sup>5</sup> synonymy

<sup>6</sup> polysemy

می‌کند. کاهش بعد ماتریس با استفاده از تجزیه مقادیر منفرد<sup>۶</sup> انجام می‌گیرد که ماتریس عبارت- مدرک را به سه ماتریس عبارت- اندازه (T)، ماتریس مقدار منفرد (مقدار- مقدار)(S)، و D ماتریس مدرک- اندازه تجزیه می‌کند. تعداد اندازه‌ها نیز که رتبه ماتریس عبارت- مدرک است با r نشان داده می‌شود. در تحلیل معنایی پنهان، ماتریس‌های S، T و D به میزان K کاهش می‌یابد. در تصویر ۲ که نحوه کاهش ماتریس را با استفاده از تجزیه مقادیر منفرد توضیح می‌دهد، نواحی هاشور خورده ماتریس‌های D و T باقی می‌مانند چون مرتبط با بزرگ‌ترین مقادیر منفرد هستند و نواحی غیر هاشور خورده نیز حذف می‌شوند. هدف از کاهش اندازه در فرایند تجزیه مقادیر منفرد و فرآیند LSA، کاهش نوفه(پارازیت) در فضای پنهان است که باعث می‌شود تا ساختار را به واژگانی غنی‌تر تقلیل دهد و معانی پنهان کنونی مجموعه را آشکار سازد (استیوی و گریس<sup>۷</sup>، ۲۰۰۷).



تصویر ۲. نحوه کاهش ماتریس با استفاده از تجزیه مقادیر منفرد(دراکوس<sup>۸</sup>، ۲۰۱۹)

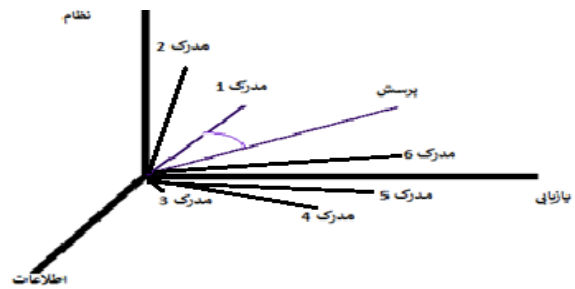
به دلیل خاصیت متعامد یا عمود بر هم موضوعات در این روش که ناشی از همپوشانی نداشتن دسته‌های موضوعی است، کلمات در یک دسته‌بندی موضوعی ارتباط کمی با دسته موضوعی دیگر دارد ولی کلمات درون یک دسته موضوعی با یکدیگر ارتباط زیادی دارند. موضوعات در LSA بیشتر با مترادفات سروکار دارد تا کلمات چندمعنایی. کلمات درون دسته‌های موضوعی به‌عنوان کلمات مترادف و هم‌معنی در نظر گرفته می‌شوند زیرا همبستگی زیادی با یکدیگر درون یک دسته موضوعی دارند. با این حال در موارد چندمعنا بودن کلمه، که با توجه به معنی و مفهوم کلمه باید در دسته‌بندی‌های موضوعی مختلف قرار گیرد، این امر امکان‌پذیر نیست. برای مثال کلمه شیر نه تنها باید در دسته موضوعی لبنیات قرار گیرد بلکه در دسته‌بندی حیوانات نیز قابل‌نمایش باشد.

<sup>6</sup> Single Value Decomposition(SVD): یک روش

ریاضی بر پایه جبر خطی و با استفاده از خواص ماتریس هاست که به عنوان یک روش کاهش، یک ماتریس را به ضرب سه ماتریس و در حقیقت به تصاویر مشخصه سازنده اش تجزیه می‌کند، که دو ماتریس متعامد و سومی قطری است و می‌تواند در تصاویر مشخصه ابتدایی، پدیده‌های افقی را شناسایی نماید(مرتضوی و جواهریان، ۱۳۹۲)

<sup>7</sup> Stayver & Griffiths

<sup>8</sup> Derakos



تصویر ۱. مدل فضا برداری مدارک(ونکات و همکاران، ۲۰۱۸)  
 پس به‌طور خلاصه در نظر گرفتن متن به‌صورت کیسه‌ای از کلمات و همچنین مدل فضا برداری متن، دو فرض اساسی در روش‌های متن‌کاوی است که بر اساس آن عمل می‌شود.  
 اما باید توجه داشت که مدل فضا برداری چهار محدودیت اصلی ایجاد می‌کند. اولاً در بازایی اطلاعات، یک مدرک طولانی شباهت کمتری با پرسمان پیدا می‌کند زیرا مقدار نرمال شده مدرک، عددی بزرگ است. در نتیجه مدارک طولانی و پرجمع با پرسمان کمتر منطبق می‌شود. مشکل دوم ناشی از در نظر گرفتن ترتیب کلمات در متن است زیرا متن به‌صورت کیسه‌ای از کلمات در نظر گرفته می‌شود. درحالی‌که ساختار نحوی یک مدرک اطلاعات ارزشمندی در بردارد. سومین مشکل این است که کلمات پرسمان باید دقیقاً منطبق با کلمات درون متن باشد و این به‌خاطر در نظر نگرفتن مسئله مترادفات است. و چهارمین مسئله، مشکل چند معنی بودن کلمه است، زیرا مدل فضا برداری تنها به شکل کلمه توجه می‌کند (هاگن، ۲۰۱۸). برای غلبه بر مسئله مترادفات، تجزیه و تحلیل معنایی پنهان (LSA) توسعه داده شد که در ادامه به آن پرداخته خواهد شد.

### چهار رویکرد مطرح در متن‌کاوی

متن‌کاوی برای شناسایی موضوعات با استفاده از مدل‌های متمایزکننده مثل تجزیه و تحلیل معنایی پنهان<sup>۱</sup> تا مدل‌های مولد مثل تجزیه و تحلیل معنایی پنهان احتمالاتی<sup>۲</sup>، تخصیص دیریکله پنهان<sup>۳</sup> و مدل‌سازی موضوعی همبسته<sup>۴</sup> توسعه یافته است.

### تجزیه و تحلیل معنایی پنهان

مدل فضای برداری نمی‌توانست مشکل مترادفات و چند معنی بودن یک کلمه واحد را برطرف کند و بنابراین برای غلبه بر این مسئله تجزیه و تحلیل معنایی پنهان (LSA) طراحی شد (چائو و مائو<sup>۵</sup>، ۲۰۱۸). تجزیه و تحلیل معنایی پنهان، ماتریس مدرک- کلمه یا بردار فضایی اصلی را در یک فضای عامل کوچک طراحی

<sup>1</sup> Latent semantic analysis(LSA)

<sup>2</sup> Probability latent semantic analysis(PLSA)

<sup>3</sup> Latent Dirichlet allocation(LDA)

<sup>4</sup> Correlated topic modeling(CTM)

<sup>5</sup> Zhao & Mao

مانند این نکته منفی در این روش این است که خاصیت متعامد (عمود بر هم بردار موضوعات) در LSA مانع از حضور کلمه در دسته‌های موضوعی مختلف می‌شود.

در شاخه‌های مختلف علم اطلاعات مثل بازیابی اطلاعات، تشخیص موضوع، طراحی هستی‌شناسی اصطلاحات تخصصی نیمه‌خودکار، طبقه‌بندی موضوعات و امتیازدهی به مقالات به کار می‌رود و پایه‌ای برای عملکرد روش‌های پیشرفته است. به عنوان مثال چین (Chien, 2016) در مطالعه‌ای امکان استفاده از الگوریتم LSA در ارتقای دسترسی به متون دیجیتال را مورد بررسی قرار داد. وی به این نتیجه رسید که استفاده از این تکنیک در بازیافت موضوعات متون طبقه‌بندی نشده بسیار کاربردی بود. رانی، دهار و ویاس<sup>۱</sup> نیز (۲۰۱۷) در مطالعه خود به بررسی امکان استفاده از روش مدل‌سازی موضوعی و مقایسه دو الگوریتم مدل‌سازی موضوعی یعنی LSI، SVD و LDA در طراحی هستی‌شناسی اصطلاحات تخصصی نیمه‌خودکار و طبقه‌بندی موضوعات پرداختند. هدف آن‌ها سنجش ارتباط آماری بین مدارک و اصطلاحات برای ساخت هستی‌شناسی موضوعی و ایجاد گراف هستی‌شناسی با حداقل دخالت انسان و بازیابی معنایی برای موضوع‌ها و تشخیص کلمات در یک حوزه موضوعی بود. در واقع تحقیق آن‌ها با تمرکز بر مسائل یادگیری آنتولوژی مثل تبدیل خودکار متن به هستی‌شناسی بود. دومین ردیف جدول ۱ برنامه‌ها و کاربرد آن را نشان می‌دهد. LSA کاربرد مدل فضا برداری را با در نظر گرفتن مترادفات بهبود می‌بخشد اما سه محدودیت دارد: ۱) کاهش بعد مستقیم از یک ماتریس است و بر اساس نظریه احتمال قوی ساخته نشده است. ۲) تعداد کافی موضوع به صورت آماری قابل شناسایی نیست و تشخیص تعداد موضوعات بستگی به قضاوت انسان دارد. ۳) مسئله چند معنی بودن یک کلمه واحد در LSA به خاطر خاصیت متعامد بردارها قابل حل نیست (افسان و همکارانش، ۲۰۱۷؛ رانی، دهار و ویاس، ۲۰۱۷). برای غلبه بر کاهش بعد مستقیم ماتریس و مسئله کلمات چندمعنایی، روش‌های مولد دیگری توسعه یافت که در بخش‌های بعدی به آن پرداخته می‌شود.

**تجزیه و تحلیل معنایی پنهان احتمالاتی**

تجزیه و تحلیل معنایی پنهان احتمالاتی فرض می‌کند که مدارک طی سه مرحله تولید می‌شوند. اول، یک مدرک  $d$  با احتمال  $P(d)$  ایجاد می‌شود. سپس، موضوع  $Z$  با احتمال  $P(z|d)$  در نظر گرفته می‌شود. در آخر، هر کلمه  $W$  در یک موضوع با احتمال  $P(w|z)$  جای می‌گیرد (زمانی، دیانت و صادق زاده، ۱۳۹۱).

احتمال وقوع کلمه - سند  $P(d,w)$  را می‌توان با  $P(d)$   $\sum_{z \in Z} P(w|z)P(z|d)$  با استفاده از الگوریتم

ماکسیمم انتظار وقوع<sup>۲</sup> (EM) که یک راه حل کلی در برآورد پارامترهای ناشناخته است، احتمالات موضوع  $P(z)$ ، احتمال مدرک با توجه به موضوع  $P(d|z)$  و احتمال کلمه با توجه به موضوع  $P(w|z)$  را تخمین می‌زند. در الگوریتم PLSA مقادیر  $P(d|z)$ ،  $P(z)$  و  $P(w|z)$  به عنوان احتمال تفسیر می‌شود. به طور کلی PLSA موضوع کلی متن را مشخص می‌کند (حیدری، ۱۳۹۳).

از جمله کاربردهای این رویکرد در حوزه علم اطلاعات می‌توان به تشخیص موضوع در متن، خوشه بندی مفاهیم حوزه‌های علمی و غنی سازی فهرست واژگان اشاره کرد. هافمن همکارانش (۲۰۰۱) برای اولین بار از PLSA برای شناسایی موضوعات در مجله عصب‌شناسی و بازیابی اطلاعات استفاده کردند. هاگن نیز (Hagen, 2018) در مطالعه‌ای به تجزیه و تحلیل محتوا و موضوعات اصلی دادخواست‌های الکترونیکی به کمک مدل‌سازی موضوعی پرداخت. هدف وی ارتقا و ارزیابی مدل‌های PLSA بود. نتایج پژوهش‌های او نشان داد که موضوعات تولیدشده در مدل‌سازی موضوعی نسبت به موضوعات برگرفته از تجزیه و تحلیل دستی این مزیت را دارد که زمینه‌های مختلف مطرح در متن را بیان نموده که ممکن است از دید تحلیلگران موضوعی پنهان مانده باشد. لئو و همکارانش (۲۰۱۶) در بررسی مفاهیم حوزه بیوانفورماتیک، با بهره‌گیری از تکنیک مدل‌سازی موضوعی به خوشه‌بندی مفاهیم این حوزه پرداختند. سایر مقالات مشابه باهدف خوشه‌بندی مفاهیم در حوزه‌های علمی به کمک تجزیه و تحلیل پنهان احتمالاتی می‌توان به آثار کستلانی و همکارانش؛ ۲۰۱۰، ماسرولی و همکارانش؛ ۲۰۱۲، بیسگین و همکارانش؛ ۲۰۱۳، لی و همکارانش؛ ۲۰۱۴، فانگ و همکارانش؛ ۲۰۱۵ اشاره کرد. سایر کاربردهای PLSA در جدول ۱ نمایش داده شده است. در رابطه با مسئله مترادفات، کلمات درون دسته‌های موضوعی در PLSA بیشتر از روش LSA به هم مرتبط هستند و در مورد کلمات چندمعنایی، کلمات درون دسته‌های موضوعی به روش PLSA می‌تواند به‌طور هم‌زمان در موضوعات دیگر نمایش داده شود. یکی از محدودیت‌های PLSA این است که تولید مدرک  $P(d)$  در مدل را در نظر نمی‌گیرد. برای نمایش روند تولید در سطح مدرک، تخصیص دیریکله پنهان توسعه پیدا کرد که در بخش بعدی به آن پرداخته می‌شود.

<sup>2</sup> Expectation-Maximization: انتظار - سازی  
وقوع روشی است استاندارد برای تخمین ماکزیمم احتمال وقوع در مدل‌های متغیر پنهان (هافمن، ۲۰۰۱)

<sup>1</sup> Rani, Dhar & Vyas



## تخصیص دیریکله پنهان

با توجه به محدودیت PLSA، روش تخصیص دیریکله<sup>۱</sup> پنهان، از روند تولید مدارک با توزیع دیریکله استفاده می‌کند. طبق الگوریتم LDA هر مدارک طی سه مرحله تولید می‌شود. اول، تعداد کلمات استفاده شده در یک مدارک با نمونه‌گیری از توزیع پواسون (توزیع احتمالی گسسته) مشخص می‌شود. دوم، توزیع در موضوعات برای یک مدارک از توزیع دیریکله استخراج می‌شود. سوم، بر اساس توزیع اختصاصی مدارک، موضوعات ایجاد می‌شوند، و سپس کلمات برای هر موضوع ایجاد می‌شود (بلای و همکاران، ۲۰۰۳). مشابه PLSA، تخصیص دیریکله پنهان نیز موضوعاتی را ارائه می‌دهد که در آن کلمات دارای ارزش احتمالاتی است. تخصیص دیریکله پنهان برای مدل‌سازی مدارک طولانی حاوی چندین موضوع مناسب است. در رابطه با مسئله مترادفات در روش LDA، کلمات درون یک دسته موضوعی معنای یکسان دارند، اما کلمات متشکل از صفت‌ها و اسم هستند مانند "صفحه خوب". این ساختار موصوف و صفت محققان را قادر می‌سازد تا به راحتی به دسته‌های موضوعی برچسب اختصاص دهند. تخصیص دیریکله در بسیاری از حوزه‌ها مثل تشخیص موضوع، تشخیص احساس، و ابهام‌زدایی از مفهوم کلمه به کار می‌رود. به عنوان مثال بیترمن و فیشر (۲۰۱۸) در مطالعه‌ای به کشف نقاط قوت موضوعی در نشریات حوزه روان‌شناختی پرداختند و برای رسیدن به این هدف از ابزار مدل‌سازی موضوعی استفاده کردند. آن‌ها از الگوریتم موضوعی LDA برای کشف و استخراج موضوعات پنهان در اسناد استفاده کردند و به این نتیجه رسیدند که با استفاده از این الگوریتم، می‌توان موضوعات خاصی را کشف کرد که با سیستم‌های طبقه‌بندی موجود امکان شناسایی آن آسان نیست. در جدول ۱ به سایر کاربردهای این الگوریتم اشاره شده است. در مورد کلمات چندمعنایی نیز، این دسته از کلمات می‌توانند به طور هم‌زمان بر اساس معنای مورد نظر در دسته‌های موضوعی مختلف قرار گیرند (افشان<sup>۲</sup> و همکاران، ۲۰۱۷). با این حال تخصیص دیریکله پنهان ارتباط بین دسته‌های موضوعی مختلف را نمی‌تواند در نظر بگیرد. درحالی‌که با شناسایی ارتباط بین دسته‌های موضوعی می‌توان ساختار عمیق بین مدارک را بهتر درک کرد.

مدل‌سازی موضوعی همبسته (CTM) برای رفع محدودیت‌های LDA توسعه یافت.

مدل‌سازی موضوعی همبسته<sup>۳</sup>

مدل‌سازی موضوعی همبسته، با استفاده از توزیع نرمال منطقی ارتباط بین دسته‌های موضوعی را شناسایی می‌کند. در تخصیص دیریکله پنهان (LDA)، فرض بر این است که کلمه درون دسته‌های موضوعی، توزیع چندجمله‌ای قرار دارد و موضوعات مختلف می‌تواند در یک مدارک ظاهر شود و سهم موضوعات طبق توزیع دیریکله متفاوت است. مدل‌سازی موضوعی همبسته از همان روند مولد LDA استفاده می‌کند اما برای کشف ارتباطات موضوعی به جای استفاده از توزیع دیریکله از توزیع نرمال منطقی<sup>۴</sup> استفاده می‌کند (آینده<sup>۵</sup>، ۲۰۱۱). در تخصیص دیریکله، دسته‌های موضوعی عملاً مستقل هستند، بنابراین مباحث نمی‌توانند با یکدیگر رابطه‌ای داشته باشند. این استقلال و ارتباط نداشتن مانع از ظهور کلمه در سایر موضوعات می‌شود. برای نمایش ارتباط موضوع‌ها، مدل‌سازی موضوعی همبسته از توزیع نرمال منطقی با در نظر گرفتن ساختار کوواریانس در بین اجزاء دسته‌های موضوعی استفاده می‌کند (همان، ۲۰۱۱). مدل‌سازی موضوعی همبسته در سایر حوزه‌ها مثل تشخیص موضوع و بازیابی تصویر استفاده می‌شود. در مورد مترادف‌ها نیز، کلمات درون هر دسته موضوعی دارای معنای یکسان هستند و آن کلمه‌ها با کلمات موضوعات مشابه نیز مشابهت دارد. در مورد کلمات چندمعنایی نیز، کلمات با معنای مختلف می‌تواند به طور هم‌زمان در سایر دسته‌های موضوعی نمایش داده شود. مدل‌سازی موضوعی همبسته، راهی برای هموار کردن مشکل چندمعنایی بودن کلمات فراهم نمود. برای مثال، چون موضوع "برنامه‌نویسی جاوا" ارتباط نزدیک‌تری با موضوع "خدمات وب" دارد، این کلمه با سایر موضوعات مشابه ابهام‌زدایی می‌شود. خصوصیات و محدودیت‌های هر چهار روش متن‌کاوی در جدول ۱ خلاصه شده است.

<sup>۱</sup> توزیع دیریکله در نظریه احتمال و آمار یک توزیع پیوسته است. این توزیع به طور کلی حالت گسترش یافته توزیع بتا برای توابع چندمتغیره است. معمولاً از توزیع دیریکله به عنوان توزیع پیشین در مدل‌سازی بیزی استفاده می‌شود

<sup>۲</sup> Efsun

<sup>۳</sup> Correlated topic modeling

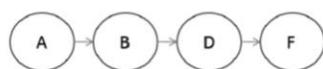
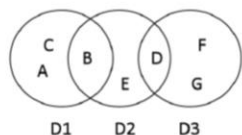
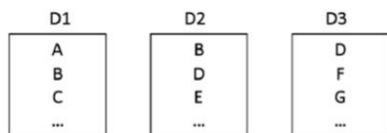
<sup>۴</sup> Logistic normal distribution

<sup>۵</sup> Hinde

جدول ۱. ویژگی‌ها و محدودیت‌های روش‌های چهارگانه متن کاوی

ویژگی‌ها/ محدودیت‌ها	مدل‌ها
<p><b>ویژگی‌ها:</b> کاهش بعد TF-IDF با استفاده از تجزیه ارزش منفرد تشخیص کلمات مترادف</p> <p><b>محدودیت‌ها:</b> مشکل تشخیص تعداد موضوعات مشکل تفسیر مقادیر احتمال مشکل برچسب‌گذاری به دسته‌های موضوعی در بعضی موارد با استفاده از کلمات درون دسته‌ها</p>	LSA
<p><b>ویژگی‌ها:</b> اجزای متن، متغیرهای تصادفی چندجمله‌ای هستند که می‌توانند به عنوان "موضوعات" بازنمایی شوند. هر کلمه از یک موضوع واحد تولید می‌شود. کلمات مختلف در یک مدرک می‌توانند از موضوعات مختلف ایجاد شوند تجزیه و تحلیل معنایی پنهان احتمالاتی مسئله چندمعنایی بودن کلمات را تا حدودی حل می‌کند.</p> <p><b>محدودیت‌ها:</b> هیچ مدل احتمالی در سطح اسناد وجود ندارد.</p>	PLSA
<p><b>ویژگی‌ها:</b> ارائه مدل با توزیع چندجمله‌ای برای کلمات در دسته‌های موضوعی و توزیع دیریکله در سطح دسته‌های موضوعی اجرای فرایند بر روی مدارک پرحجم و طولانی نمایش ویژگی‌های نحوی کلمه در دسته‌های موضوعی</p> <p><b>محدودیت‌ها:</b> توانایی نداشتن در مدل‌سازی ارتباط بین دسته‌های موضوعی</p>	LDA
<p><b>ویژگی‌ها:</b> در نظر گرفتن ارتباط بین دسته‌های موضوعی با استفاده از توزیع نرمال منطقی امکان ظهور کلمات در دسته‌های موضوعی مختلف امکان نمایش گراف موضوعات</p> <p><b>محدودیت‌ها:</b> نیاز به محاسبه‌های پیچیده حضور بیش از اندازه کلمات عادی در دسته‌های موضوعی</p>	CTM

خاطر موضوع D وجود دارد. مرتبه دوم همبستگی یا هم‌رخدادی بین مدارک ۱ و ۲ به خاطر ارتباط تریایا (A B D F) وجود دارد. درحالی که LSA موضوعات منحصر به فرد هر مدرک (A, C - E - F, G) را بر اساس ترتیب و نظم (B, D) در نظر می‌گیرد، مدل‌های احتمالاتی، ابتدا بخش‌های مشترک (B-D) را در نظر می‌گیرند.



تصویر ۳. ساختار هم‌رخدادی کلمات درون مدارک (لی، سانگ و

کیچم، ۲۰۱۰، ۴)

درک نظری تفاوت‌های عملکردی چهار روش متن کاوی روش‌های چهارگانه متن کاوی اساساً مبتنی بر هم‌رخدادی و رابطه تریایی<sup>۱</sup> است (کنتوستاتیس و پوتنجر، ۲۰۰۶). در ریاضیات، بین سه عضو a و b و c از مجموعه A یک رابطه تریایا برقرار است هرگاه بتوان از وجود رابطه دوتایی بین a و b از یک سو، و b و c از سوی دیگر، نتیجه گرفت که a و c نیز دارای همان رابطه هستند. تریایی یا تعدی پذیری مانند بازتاب و تقارن یکی از ویژگی‌های برخی از رابطه‌ها است (بلای و نفرتی، ۲۰۰۷).

به‌طور کلی، LSA از ساختاری منحصر به فرد بهره می‌گیرد اما سایر روش‌ها دارای همپوشانی در ساختار یا روش هستند. برای مثال، تصویر ۳ سه مدرک را نشان می‌دهد که مدرک ۱ (D1) حاوی سه موضوع (A, B, C)، مدرک ۲ شامل موضوعات (B, D, E) و مدرک ۳ حاوی موضوعات (D, F, G) است. از منظر اسناد، بین مدارک ۱ و ۲ هم‌رخدادی در موضوع B و بین مدارک ۲ و ۳ هم‌رخدادی به

<sup>1</sup> Transitive relation

<sup>2</sup> Kontostathis, & Pottenger

<sup>3</sup> Blei & Lafferty



موضوعات استخراج شده به روش LSA دارای دو ویژگی منحصر به فرد بود: حضور کلمات متمایز و منحصر به فرد در یک موضوع و وجود تمایز و انحصار بین موضوعات از طرف دیگر، در مدل LSA، فقط کاهش بعد، بدون در نظر گرفتن معانی قابل مشاهده بود و موضوعات نسبت به یکدیگر متمایز یا به اصطلاح متعامد<sup>۱</sup> هستند. PLSA سه ویژگی مجزا را نشان داده است: (۱) نام محصول و لوازم جانبی آن از احتمال بالایی برخوردار بود. (۲) «دوربین» در بیش از چندین موضوع به طور هم زمان نشان داده شده، به طوری که مسئله چند معنایی بودن کلمه احساس می شود هر چند معنی مشخص و واضح نیست. حداقل می توان تصور کرد که کلمه «دوربین» در موضوع اول در رابطه با «تصویر و عکس» است و در موضوع دوم در رابطه با «باتری». (۳) موضوعات استخراج شده به روش PLSA درباره موضوع یا گرایش کلی است. برای مثال، می توان تصور کرد که به طور کلی، موضوع ۱ در رابطه با موضوع عکس و موضوع دوم در رابطه با باتری است. از منظر مدل ایجاد شده در این روش نیز باید گفت که، کلمات با احتمال بالا در ابتدا آمده است. از منظر ساختار هم رخدادی نیز، PLSA کلمات مشترک را استخراج می کند تا کلمات متمایز و منحصر به فرد را.

روش LDA نیز دارای سه ویژگی منحصر به فرد بود: (۱) مشابه PLSA، موضوعات استخراج شده در LDA ترکیبی از نام محصول و لوازم جانبی آن است. این به خاطر احتمال وقوع زیاد این کلمات درون متن است. (۲) هم رخدادی صفات و اسم وجود دارد که امکان برچسب گذاری دسته های موضوعی را ساده کرده است. (۳) LDA مدل خوبی برای اسناد حجیم و طولانی ارائه می کند. از منظر ساختار هم رخدادی نیز، می توان حدس زد که چون کیفیت تصویر، قیمت و باتری از نگاه کاربر مهم بوده است، الگوی هم رخدادی این کلمات در مدارک به طور واضح دیده می شود.

نهایتاً عملکرد CTM، ارائه رابطه بین دسته های موضوعی است. این روش نیز، حاوی اسم ها و صفات در هر دسته موضوعی است که امکان برچسب گذاری دسته های موضوعی را راحت تر می کند. بعلاوه، CTM امکان تشخیص رابطه بین موضوعات را نیز فراهم کرده است. برای مثال، می توان ۵ گروه موضوعی از ۱۰ موضوع ایجاد کرد مانند (۰، ۵)، (۱، ۶)، (۲، ۷)، (۳، ۸) و (۴، ۹). بنابراین، CTM نه تنها در استخراج موضوع استفاده می شود بلکه در بیان رابطه بین دسته های مختلف موضوعی استخراج شده نیز توانا است.

کنتوستاتیس و پوتنگر (۲۰۰۶) دریافتند که هم رخدادی مرتبه دوم کلمات تأثیر به سزایی در نتایج LSA دارد. بنابراین، هنگام استفاده از چهار روش، محققان مدارک را تا سطح پاراگراف جداسازی می کنند تا الگوی هم رخدادی را غنی سازند. بنابراین، عملکرد هر چهار روش متن کاوی به طور اساسی تحت تأثیر هم رخدادی بالا و ارتباط ترایا است. ساختار هم رخدادی در تصویر ۱ می تواند با احتمالات شرطی  $P(D|B)$ ،  $P(B|A)$  و  $P(F|D)$ ، ارتباط A-B-D-F را نشان دهد. تنوع مدل های احتمالاتی ناشی از مدل سازی ساختار هم رخدادی است. LDA، PLSA و CTM این دانش قبلی را با هر مدل ترکیب می کند.

### بررسی عملکرد چهار روش در تشخیص موضوع

تشخیص موضوع یکی از کاربردهای کلی تمام رویکردهای متن کاوی از LSA تا CTM است. با توجه به این که تمام چهار روش متن کاوی به طور مستقیم در مطالعات مربوط به تشخیص موضوع استفاده می شود، می توان تفاوت ها و شباهت های عملکردی این چهار روش را در یک آزمایش کلی بر روی یک دسته از مدارک نشان داد. لی، سانگ و کیم (۲۰۱۰) در مطالعه ای، به بررسی نظرات کاربران در رابطه با دوربین عکاسی از وبسایت آمازون پرداختند، که بعد از فرآیند پیش پردازش و حذف کلمات زائد، ماتریس کلمه-مدرک ایجاد گردید. سپس چهار روش متن کاوی روی این ماتریس اعمال شد. جدول زیر موضوعات استخراج شده در این چهار روش را نشان می دهد.

جدول ۲. زمینه دوربین عکاسی با ۴ روش متن کاوی

روش	موضوع	کلمات
LSA	#2	اجرا، مدیریت، توقف، شرکت، آرزو، مقایسه کردن، کارت ها
	#3	نامزد، حرفه ای، پیشنهاد، نوه، شکست، پیچیده، دوست داشتن، کتاب ها
PLSA	#1	دوربین، بسیار خوب، تصاویر، دیجیتال، ویژگی ها، کانن، دوربین ها، آسان، نقطه، خوب، کیفیت، تصاویر، قیمت
	#8	دوربین، باتری ها، خوب، تصاویر، فقط، باتری، عالی، مشابه، زمان بر، چراغ
LDA	#1	دوربین، باتری ها، خوب، باتری، کانن، تصاویر، فقط، می خواستم، ساده، مشابه، زندگی، نما
	#2	دوربین، فلش، خوب، مد، زوم، کانن، تصویر، تصویرها، بهتر، دستی، زمان، دیجیتال، عالی
CTM	#0	دوربین، تصویر، عالی، قدرت نما، نما، محصول، دیجیتال، خوب، مشابه، زوم، آمازون
	#5	دوربین، کانن، تصویر، دوربین ها، پارازیت، دیجیتال، تصاویر، کیفیت، عالی، نقطه، استفاده ها، خوب

<sup>1</sup> orthogonal

## نتیجه گیری

دسته‌های موضوعی مختلف استفاده شود. بنابراین رویکردهای متن کاوی به سبب بهره‌گیری از تحلیل معنایی در کشف و استخراج موضوع متون مناسب است. بر این پایه با توجه به قابلیت‌های تکنیک‌های مذکور پیشنهاد می‌گردد که به انجام پژوهش در زمینه کشف موضوعات و روابط پنهان در حوزه‌های علوم، بازیابی اطلاعات، دسته بندی مدارک بر اساس موضوعات، کشف الگوهای برجسته و رویدادهای در حال ظهور، خوشه بندی مفاهیم حوزه‌های علمی، تحلیل سیر تحول مفهومی در طول دوره‌های زمانی، غنی سازی فهرست واژگان، تعیین روابط سلسله مراتبی مفاهیم یک حوزه یا زمینه خاص علمی، و همچنین در الگوسازی و بازنمایی دانش در محیط‌های وب پایه مثل امور تجاری و بانکی، متن کاوی برای مدیریت ارگان‌های آموزشی و پژوهشی، مدیریت دانش، بررسی پایگاه‌های اطلاعاتی استفاده شود. این امر منجر به غنای ادبیات نظری و پژوهش‌های عملیاتی بیشتر در این حوزه منجر خواهد شد.

## تعارض منافع

گزارش نشده است.

## منبع حمایت کننده

گزارش نشده است.

متن کاوی یا تجزیه و تحلیل متن یکی از حوزه‌های خاص داده کاوی است. از آنجاکه اکثر اطلاعات (بیش از ۸۰٪) به صورت متن ذخیره شده‌اند، و حاوی اطلاعات ارزشمند و نهفته‌ای هستند، اعتقاد بر این است که متن کاوی ارزش بالقوه بالایی دارد. متن کاوی برای شناسایی موضوع از مدل‌های متمایزکننده مثل تجزیه و تحلیل معنایی پنهان تا مدل‌های مولد مثل تجزیه و تحلیل معنایی پنهان احتمالاتی، تخصیص دیریکله پنهان و مدل‌سازی موضوعی همبسته توسعه یافته است. در این مقاله ویژگی‌ها و محدودیت‌های چهار روش متن کاوی شامل LDA، PLSA، LSA و CTM مورد بحث و بررسی قرار گرفت. زمینه‌های نظری در هر چهار روش بیان گردید و عملکرد چهار روش در تشخیص موضوع مورد مقایسه واقع شد. یافته‌ها نشان می‌دهد که در فرایند تشخیص موضوع، LSA می‌تواند برای تشخیص موضوعات خاص و منحصر به فرد در مدارکی که تنها به یک موضوع پرداخته‌اند استفاده شود. سه روش دیگر مورد بررسی، بر موضوعات و گرایش کلی متن متمرکز هستند. PLSA برای مدارکی که به یک موضوع پرداخته‌اند قابل استفاده است اما برخلاف LSA، این روش در کشف موضوعات و مضامین کلی متن کاربرد دارد. در حالی که LDA در مورد مدارکی که به چندین موضوع پرداخته‌اند کاربرد بیشتری دارد. روش CTM می‌تواند در تشخیص ارتباط بین

578. Available at:  
<https://www.researchgate.net/publication/274394886>  
 Hierarchical Theme and Topic Modeling  
 Dean, J(2014). Bigdata, datamining & machine learning: Value creation for business leader and practitioners, Retrieved from:  
[https://www.wiley.com/en-ir/Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners-p-9781118618042](https://www.wiley.com/en-ir/Big+Data,+Data+Mining,+and+Machine+Learning:+Value+Creation+for+Business+Leaders+and+Practitioners-p-9781118618042)

Drakos, G(2019). NLP Tutorials: topic modeling with SVD and truncated SVD. GDcoder. Retrieved from:  
<https://medium.com/@george.drakos62/nlp-tutorial-topic-modeling-with-singular-value-decomposition-svd-and-truncated-svd-fbpc-a-and-5fa612277c22>.

Efsun ,S., Yadav, K., Chio, H. A (2017). Topic modeling based classification of clinical report. Association for computational linguistics, 67-73. Retrieved from: <http://aclweb.org/anthology/P13-3010>.

Fang EX, Li M-D, Jordan MI, Liu H (2018) Mining massive amounts of genomic data: a semi parametric topic modeling approach

Fang, D., Yang, H., Gao, B. and Li, X. (2018), "Discovering research topics from library electronic references using latent Dirichlet allocation", Library Hi Tech, 36(3), 400-410. <https://doi.org/10.1108/LHT-06-2017-0132>

## References

- Abosaba Kazemaini, A(2011). Comparison of Comprehensiveness and Prevention of Recovered Information Based on Front and Back Storage Storage Systems in Persian Library Software. Master thesis. Department of Library & Information Science, Faculty of Educational Sciences and Psychology, Isfahan University.
- Babu, P. B., Sarangi, A.K., & Madalli, D. P. (2012). "Knowledge Organization Systems for semantic digital Libraries". International Conference Trends in Knowledge and Information Dynamics. Bangalore, Pakistan. Retrieved from: [http://eprints.rclis.org/19759/1/KOS semantic Digital Libraries.pdf](http://eprints.rclis.org/19759/1/KOS_semantic_Digital_Libraries.pdf)
- Bitterman, Andre; Fischer, Andreas (2018). How to identify hot topics in psychology using topic modeling. Zeitschrift fur psychologie. 226(1), 3-13.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 17-35.
- Blei, D; Ng, A; Jordan, M (2003), "Latent dirichlet allocation," Journal, 3, 993-1022.
- Blei, David & Lafferty, John (2007). A correlated topic model of science. The annual of applied statistics, 1(1), 17-35.
- Chien, Jt(2016). Hierarchical theme and topic modeling. IEEE trans neural netw learn syst. 27(3): 565-

- Figuerola, C.G., García Marco, F.J. & Pinto, M. *Scientometrics* (2017) 112: 1507. Retrieved from: <https://doi.org/10.1007/s11192-017-2432-9>.
- Gupta, V. and G. Lehal (2009) "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies In Web Intelligence*, 1.
- Hagen, Loni (2018). Content analysis of e-petition with topic modeling: how to train and evaluate LDA models? *Information processing & management*, 54(6), 1292-1307.
- Heydari, F (2014). Web users clustering and initial fetching of web pages using hidden probabilistic semantic analysis. Master thesis. Isfahan University of Technology.
- Hinde J. (2011) Logistic Normal Distribution. In: Lovric M. (eds) *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg
- Hofmann T. (2001) "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, 42(1-2), 177-196.
- Hwang, S.Y., Wei, C.P., Lee, C.H., & Chen, Y.S. (2017). Coauthor ship network based literature recommendation with topic model. *Online Information Review*, 41(3), 318-336.
- Khademian, M., Kokabi, M (2018). Liberian Thing's Social Labels Versus Subject Headings in the Library of Congress: Review of Texts. *Journal of Library and Information Science*, 8 (1,) 313- 335. Retrieved 3/3/98, from : <https://infosci.um.ac.ir/index.php/riis/article/view/57823>
- Kinyanjui, Daniel (2016) Subject cataloguing and the principles on which the choice of subject headings should be based, GRIN Verlag: Munich.
- Koller, D., and Friedman, N. (2009), "Probabilistic Graphical Models: Principles and Techniques", The MIT Press.
- Kurata, K & et al (2018). Analyzing library and information science full-text articles using a topic modeling approach. 81 Annual meeting of the association for information science & technology in Vancouver of Canada (10-14, November, 2018). Retrieved from: <https://www.researchgate.net/publication/330812928>
- 8 Analyzing library and information science full-text articles using a topic modeling approach
- lee, S., Song, J & Kim, Y (2010). An Empirical comparison of four text mining methods. *Journal of computer information system*. 51(1):1-10. Retrieved from: <https://www.researchgate.net/publication/286840108>
- 8 An empirical comparison of four text mining methods
- Meen Ch & Yongjun, Zh (July 18th 2018). *Scientometrics of Scientometrics: Mapping Historical Footprint and Emerging Technologies in Scientometrics, Scientometrics*, Mari Jibu and Yoshiyuki Osabe, IntechOpen, DOI: 10.5772/intechopen.77951. Available from: <https://www.intechopen.com/books/scientometrics/scientometrics-of-scientometrics-mapping-historical-footprint-and-emerging-technologies-in-scientometrics>
- Mohammadian, B (2014) Identification of scientific theft in Persian documents based on thematic modeling. Master thesis. Department of Computer, Faculty of Engineering, Kharazmi University.
- Mortazavi, A., Javaherian, A (2013). Application of single value decomposition to random noise attenuation in synthetic and real seismic data. *Oil Research*. (80), 123-134. Retrieved from: <https://pr.ripi.ir/article> 459
- 85173420168d8944de96e91c8a871aa2.pdf
- Nadezhda, Y & Aleksey, F (2018). Improving the quality of information retrieval using syntactic analysis of search query. Retrieved from: <https://www.semanticscholar.org/paper/Improving-the-Quality-of-Information-Retrieval-of-Yarushkina-Filipov/d0955103ee4e4cd78a0d24f880a1cda7f3b35d5e>
- Newman, D., Hagedorn, K., Chemudugunta, C., & Smyth, P. (2007). Subject metadata enrichment using statistical topic models. *JCDL*. Retrieved from: <https://www.researchgate.net/publication/220924369>
- 9 Subject Metadata Enrichment using Statistical Topic Models
- Norouzi, Y., Khavidaki, S (2014). *Social Semantic Digital Library: A Perspective for Digital Libraries in Iran*. *Rahyaft*, 57, 63-74. Retrieved 5/8/98 from <http://rahyaft.nrisp.ac.ir/article/13557.html>
- Rani, M., Dhar, A, K., Vyas, O.P (2017). Semi-Automatic terminology ontology learning based on topic modeling. *Engineering Application of Artificial Intelligence*, 63, 108-125. Retrieved from: <https://www.researchgate.net/publication/317195300>
- 0 Semi-Automatic Terminology Ontology Learning Based on Topic Modeling
- Rani, M., Dhar, A., Kumar; Vyas, O.P (2017). Semi-Automatic terminology ontology learning based on topic modeling. *Engineering Application of Artificial Intelligence*, 63, 108-125.
- Sanandres, E; Madariaga, C; Abello, R (2018). Topic modelling of twitter conversations. Retrieved from: <https://www.researchgate.net/publication/326450126>
- 6 Topic Modeling of Twitter Conversations/citations
- Selvi, M & et al (2019). Classification of medical dataset along with topic modeling using LDA. *Lecture notes in electrical engineering* 511. Springer.
- Soergel, D (2004). *Indexing language and thesauri: construction and maintenance*. Los Angeles, CA: Melville

- Sohrabi, B; Raeesi vanani, I; Baranizade Shineh, M (2017). Topic Modeling and classification of cyberspace papers using text mining. *Cyberspace studies*, 2(1), 103- 125.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7), 424-440.
- Steyvers, M; Smyth, P; Rosen-Zvi, M; Griffiths, T, (2004) "Probabilistic author-topic models for information discovery," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington.
- Strunk Jr, W.(2007), "The elements of style", Fiquarian Publishing, LLC.
- Venkat N. Gudivada, Amogh R. Gudivada(2018). *Hand book of ststistic*. USA, Elsevier. Retrieved from : <https://www.sciencedirect.com/topics/computer-science/vector-space-models>
- Zamani, M., Dianat, R., Sadeghzadeh, M(2013). Classification of Persian Texts Using Probabilistic Hidden Semantic Analysis Method, 1st National Conference on Application of Intelligent Systems (Soft Computing) in Science and Technology, Quchan, Islamic Azad University of Quchan.
- Zhao, R., & K. Mao. 2018. Fuzzy Bag-of-Words Model for Document Representation. *IEEE Transactions on Fuzzy Systems* ۲۶ (2): 794-804. doi:10.1109/TFUZZ.2017.2690222.