# An Investigation into the Process of Organizing and Retrieving Web Texts based on the Integration of Semantic Concept in order to Organize Knowledge

**Saeede Anbaee Farimani**
PhD Candidate; Department of Computer Engineering;
Mashhad Branch; Islamic Azad University; Mashhad, Iran;
Email: anbaee@mshdiau.ac.ir

**Hamid Tabatabaee***
PhD; Assistant Professor; Department of Computer Engineering;
Quchan Branch; Islamic Azad University; Quchan, Iran;
Email: h_tabatabaee@mshdiau.ac.ir

**Mojtaba Kaffashan Kakhki**
Assistant Professor; Knowledge and Information Science;
Academic Member of Ferdowsi University of Mashhad;
Email: kafashan@ferdowsi.um.ac.ir

Iranian Journal of Information Processing and Management

**Received: 07, Jan. 2019** **Accepted: 09, May 2019**

**Abstract:** Improvement in information retrieval performance relates to the method of knowledge extraction from large amounts of text information on web. Text classification is a way of knowledge extraction with supervised machine learning methods. This paper proposed Kullback-Leibler divergence KNN for classifying extracted features based on term weighting with Latent Dirichlet Allocation algorithm. LDA is Non-Negative matrix factorization method proposed for topic modeling and dimension reduction of high dimensional feature space. In traditional LDA, each component value is assigned using the information retrieval Term Frequency measure. While this weighting method seems very appropriate for information retrieval, it is not clear that it is the best choice for text classification problems. Actually, this weighting method does not leverage the information implicitly contained in the categorization task to represent documents. In this paper, we introduce a new weighting method based on Point wise Mutual Information for accessing the importance of a word for a specific latent concept, then each document classified based on probability distribution over the latent topics. Experimental result investigated when we used Pointwise Mutual

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 4 | pp. 1879-1904
Summer 2019

* Corresponding Author

Information measure for term weighing and K Nearest Neighbor with Kullback-Leibler distance for classification, accuracy has been 82.5%, with the same accuracy versus probabilistic deep learning methods.

**Keywords:** Text Mining, Text Classification, Topic Modeling, Latent Dirichlet llocation, Document Representation, Knowledge Organization, Pointwise Mutual Information

# جستاری بر فرایند سازماندهی و بازیابی متون وبی مبتنی بر تجمیع مفاهیم معنایی در راستای سازماندهی دانش

**سعیده انبایی فریمانی**
دانشجوی دکتری؛ گروه مهندسی کامپیوتر؛ واحد مشهد؛ دانشگاه آزاد اسلامی؛ مشهد، ایران؛
Anbaee@mshdiau.ac.ir

**حمید طباطبایی**
استادیار؛ گروه مهندسی کامپیوتر؛ واحد قوچان؛ دانشگاه آزاد اسلامی؛ قوچان، ایران؛
پدیدآور رابط h_tabatabaee@mshdiau.ac.ir

**مجتبی کفاشان کاخکی**
استادیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه فردوسی مشهد؛ ایران kafashan@ferdowsi.um.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

**دریافت:** ۱۳۹۷/۱۰/۱۷ **پذیرش:** ۱۳۹۸/۰۲/۱۹ مقاله برای اصلاح به مدت چهار روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳٤ | شمارهٔ ٤ | ص‌ص ۱۸۷۹-۱۹۰٤
تابستان ۱۳۹۸

**چکیده:** سازماندهی و بازیابی دانش منتشرشده در محیط وب به‌عنوان یکی از مهم‌ترین کاربردهای متن‌کاوی مطرح است. از جملهٔ چالش‌های سازماندهی مجموعهٔ عظیمی از متون در قالب یک پیکرهٔ متنی، ابعاد زیاد ویژگی‌ها و خلوت بودن ماتریس ویژگی‌هاست. نحوهٔ انتخاب ویژگی‌ها و کاهش آن‌ها در این مسئله تأثیر به‌سزایی در بالاتر رفتن دقت سازماندهی و بازیابی متون دارد. در بسیاری از پژوهش‌ها به بررسی جداگانهٔ این دو چالش پرداخته شده است. این پژوهش با رویکرد توجه همزمان به این دو چالش انجام گرفته است. پس از تعیین متون مرتبط با ۲۰ گروه خبری وبی و پس از فاز پیش‌پردازش متون با استفاده از الگوریتم الگوسازی عنوان «الدی‌ای»، کیسه‌ای (تجمیعی) از مفاهیم معنایی برای پیکرهٔ متنی مورد نظر ساخته شد. به‌منظور بررسی میزان تأثیر واژه‌های پیکرهٔ متون در هر مفهوم پنهان، به بررسی نحوهٔ وزن‌دهی واژگان یک پیکره، در مفاهیم استخراج‌شده توسط الگوریتم «الدی‌ای» پرداخته شد. از این رو، برای هر متن یک توزیع احتمال رخداد حول هر عنوان استخراج گردید که برای سازماندهی و بازیابی دانش موجود در آن مورد استفاده قرار گرفت. برای سازماندهی آن از الگوریتم نزدیک‌ترین K همسایه با معیار شباهت واگرای «کولبک لیبلر» که میزان فاصلهٔ دو توزیع احتمال را می‌سنجد؛ استفاده شد.

نتایج آزمون‌ها نشان داد که میزان صحت سازماندهی روش پیشنهادی در صورتی که از معیار وزن‌دهی واکشی اطلاعات متقابل نقطه‌ای و الگوریتم KL-KNN استفاده شده باشد، ۸۲/۵ درصد است. نتایج تحلیل‌ها نشان داد که این روش دارای دقتی مشابه با روش‌هایی است که از فنون یادگیری عمیق استفاده می‌نمایند. افزون بر این، روش به‌کاررفته در این پژوهش نشان‌دهندهٔ پیچیدگی کمتر در فرایند سازماندهی و بازیابی متون مورد مطالعهٔ پژوهش بود.

**کلیدواژه‌ها:** متن‌کاوی، طبقه‌بندی متن، الگوسازی عنوان، بازیابی، سازماندهی دانش، واکشی اطلاعات متقابل نقطه‌ای

**۱. مقدمه**

سازماندهی و بازیابی از مباحث مهم و مطرح در نظام‌های اطلاعاتی است؛ به‌گونه‌ای که بر اساس تحقیقات «جارولین و واکاری» (Järvelin & Vakkari 1993) از هر سه مقالهٔ نوشته‌شده در علم اطلاع‌رسانی یک مقاله به سازماندهی و بازیابی اختصاص یافته است. افزون بر این، با گسترش و رشد وب و شبکه‌های اجتماعی، حجم متون منتشرشده به‌عنوان داده‌های بدون ساختار در حال گسترش است (Thiyagarajan & Shanthi 2017). فهم قدرت متون منتشرشده با در نظر گرفتن نقش‌آفرینی متن‌کاوی و داده‌کاوی در تولید دانش سطحی، دانش چندوجهی، دانش پنهان، و دانش عمیق دوچندان می‌گردد. متن‌کاوی به‌مثابهٔ تحلیل هوشمند متن، کاوش متن و یا کشف و تولید دانش تلقی می‌شود (قاضی‌زاده و خزانه‌ها ۱۳۹۵). به‌عبارت دقیق‌تر، کشف و تولید دانش مستلزم توجه ویژه به فرایندهای سازماندهی و بازیابی دانش مندرج در متون است. سازماندهی و بازیابی متون، یکی از شاخه‌های مهم کاوش متن است که امروزه در بسیاری از زمینه‌ها از جمله دسته‌بندی اسناد خبری، طبقه‌بندی صفحات وب، تحلیل احساس در کاربران یک خدمت، فیلترکردن اسپم‌ها و ... کاربرد دارد (Hotho & Paaß 2005). همچنین، سازماندهی و بازیابی متون از زمرهٔ کاربردهای یادگیری ماشین و از جمله روش‌های یادگیری با ناظر است که در آن به ساخت یک الگو از یک پیکرهٔ متنی[1] بر اساس داده‌های آموزشی دارای برچسب پرداخته می‌شود و سپس، با استفاده از الگوی تولیدشده، به سازماندهی سایر داده‌های بدون برچسب می‌پردازد. فرایند آموزش شامل پیش‌پردازش داده‌ها، انتخاب ویژگی‌های مناسب، کاهش ابعاد ویژگی‌ها، و آموزش الگوست. گروهی از روش‌ها برای ساخت ماتریس ویژگی از فراوانی تکرار

1. text corpus

واژه‌های موجود در متون یک پیکره استفاده می‌کنند. بر این پایه، ممکن است بسیاری از واژه‌هایی که در متون فراوانی تکرار کمی داشته باشند، در مرحلهٔ کاهش ویژگی حذف شوند (Blei, Ng & Jordan 2003; Chen, Guo & Bai 2017; Chen & Li 2016; Zhao & Mao 2018). این نحوهٔ کاهش ویژگی سبب کاهش دقت سازماندهی و بازیابی دانش ارائه‌شده در متن مورد نظر می‌گردد. این فرایند در قالب مهار واژگانی تبیین می‌گردد. مهار واژگانی به مجموعه‌ای محدود از اصطلاح‌های مجاز نسبت داده می‌شود که باید در نمایه‌سازی و در جست‌وجوی مدارک در یک نظام اطلاعاتی معین به کار رود (پائو ۱۹۸۹). در روش پیشنهادی در پژوهش حاضر تلاش می‌شود که بعد از انجام مراحل پیش‌پردازش، برای پیکرهٔ متن، ماتریسی از ویژگی‌ها با استفاده از روش کاهش ویژگی تخصیص دریکله[1] به مفاهیم پنهان (یا الدی‌اِی[2]) به‌گونه‌ای طراحی شود که افزون بر مؤثر بودن فراوانی تکرار واژگان، میزان اهمیت آن واژه در پیکرهٔ متون نیز لحاظ شود و در نهایت، پس از استخراج ماتریس توزیع احتمال رخداد[3] هر عنوان در اسناد پیکره با استفاده از طبقه‌بند مبتنی بر نزدیکترین K[4] همسایه با معیار فاصله «کولبک‌لیبلر»[5] برای سازماندهی و بازیابی متون استفاده شود. «کولبک‌لیبلر» به‌عنوان معیاری برای اندازه‌گیری واگرایی یک توزیع احتمال از یک توزیع احتمال ثانویه در نظر گرفته می‌شود. در روش الگوسازی عنوان[6] «الدی‌اِی» حد آستانه‌ای از تکرار برای واژگان در نظر گرفته می‌شود و واژه‌هایی که کمتر از حد آستانه در متن تکرار شده باشند، از فرایند الگوسازی مفاهیم معنایی حذف می‌شوند؛ در

---

1. Dirichlet

۲. Latent Dirichlet Allocation یک الگوی تولیدی در آمار و یک روش الگوسازی عنوان است. این الگو برای الگوسازی تعدادی متغیرهای پنهان (عناوین) در مجموعه‌ای از متن‌ها که شامل کلمات هستند، به وجود آمده است. از این رو، میزان تأثیر هر واژه با توجه به فراوانی تکرار آن در یک مفهوم معنایی و همچنین، میزان تأثیر هر مفهوم استخراج‌شده در هر متن تعیین می‌گردد.

۳. هر کدام از عناوین استخراج‌شده بر اساس وزن مؤثر لغات هر عنوان، با یک میزان احتمال مشخص در یک سند وبی ظاهر می‌گردد، به‌طوری که حاصل‌جمع مقادیر این احتمال‌ها یک شود.

۴. جست‌وجوی K نزدیک‌ترین همسایه، K همسایه نزدیک‌تر به نقطهٔ پرس‌وجو را برمی‌گرداند. از این روش، به‌منظور تخمین یا دسته‌بندی یک نقطه بر اساس اجماع همسایگان آن استفاده می‌شود.

۵. Kullback Leibler میزان فاصلهٔ واگرایی «کولبک‌لیبلر» به‌عنوان معیاری برای اندازه‌گیری واگرایی یک توزیع احتمال از یک توزیع احتمال ثانویه در نظر گرفته می‌شود. به‌عبارت دیگر، مقدار صفر برای «کولبک‌لیبلر» نشان می‌دهد که ما می‌توانیم انتظار رفتاری مشابه (نه دقیقاً یکسان) از دو توزیع داشته باشیم؛ در حالی که مقدار ۱ برای این معیار نشان می‌دهد که دو توزیع رفتارهای متضادی دارند.

6. topic modeling

حالی که ممکن است این واژه‌ها دارای نقش مهمی در طبقهٔ خود باشند. برای رفع این محدودیت در این پژوهش با اعمال ضریب واکشی اطلاعات متقابل نقطه‌ای[1]، این واژه‌ها در نحوهٔ استخراج مفاهیم معنایی مشارکت داده شد.

همسو با دیدگاه «ویکری و ویکری» (۱۹۸۹) اگر به مجموعهٔ جملات موجود در یک متن بیندیشیم، می‌توانیم دریابیم که چه ساختار ارتباطی پیچیده‌ای دارد. درون جمله، واژه‌ها در طبقات نحوی جای می‌گیرند. در ورای این، جملات نیز به شیوه‌ای با یکدیگر ارتباط می‌یابند که بتوانند بحث موجود در یک متن را پیش ببرند. در نتیجه، با مجموعه‌ای از مفاهیم آشکار و پنهان در درون متن مواجه هستیم که روند ذخیره، سازماندهی و بازیابی آن را به شدت تحت‌الشعاع خود قرار می‌دهد. در فرایند سازماندهی و بازیابی متن روش‌های مختلفی برای ارائهٔ متن و اعمال طبقه‌بندی آن‌ها بر پایهٔ نحوهٔ ارائهٔ آن وجود دارد. بنابراین، در این پژوهش ابتدا در بخش مفاهیم پایه به بررسی مفاهیم پایه‌ای ارائهٔ متن پرداخته شده است. روش‌های کاهش ابعاد ویژگی‌ها و نحوهٔ سازماندهی و بازیابی متن بیان شده و سپس، مقایسه‌ای بین این روش‌ها ذکر شده است. در ادامه، روش پیشنهادی پژوهش، ارزیابی روش، و در نهایت، نتیجه‌گیری و پژوهش‌های آتی ارائه شده است.

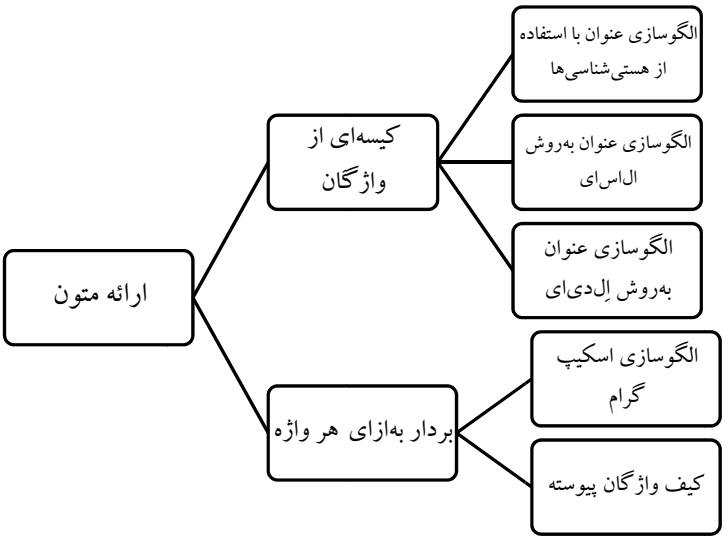
## ۲. مبانی نظری پژوهش

### ۲-۱. تبیین مفاهیم پایه در ارائهٔ متن

یک پیکرهٔ متنی به مجموعه‌ای عظیم از اسناد متنی نسبت داده می‌شود. بزرگی یک پیکرهٔ متنی را با استفاده از تعداد واژه‌های موجود در فرهنگ لغات آن پیکره می‌سنجند. زیربنای بسیاری از کاربردهای متن‌کاوی، فرایند بازیابی متن است. این فرایند را می‌توان به شکل‌های مختلفی انجام داد. به‌طور کلی، روش‌هایی مانند روش متداول کیسه‌ای (تجمیع) از واژگان[2] بردار به ازای هر واژه در متن، الگوسازی عنوان با استفاده از هستی‌شناسی‌ها و به کمک تجزیهٔ نامنفی ماتریس‌ها در بازیابی متن مورد توجه قرار می‌گیرد. در این بخش به بیان جزئیات مربوط به هر کدام از روش‌های متداول یادشده پرداخته می‌شود. نمودار ۱، شمایی از روش‌های استفاده‌شده در ارائهٔ متن جهت سازماندهی و بازیابی آن را نشان می‌دهد.

1. point wise mutual information

۲. یک روش بازنمایی متن است که در آن هر لغت بر اساس فرکانس تکرار وزن‌دهی می‌گردد.




**شکل ۱. شمایی از روش‌های ارائهٔ متن در فرایند سازماندهی دانش**

### ۲-۲. سازماندهی و بازیابی متون به روش بردار به‌ازای هر واژه

روش بردار به‌ازای هر واژه در سال ۲۰۱۳ توسط «میکولو» و همکاران یکی از محققان گوگل ارائه شد (Mikolov et al. 2013). این روش به‌ازای هر واژه، بُرداری با طول کوچک تولید می‌کند. این روش مبتنی بر این نظر است که معنای یک متن یا سند با واژه‌هایی که در آن قرار دارند، بیان می‌شود (گراسمن و فریدر[1]، ۲۰۰۴). این الگو در حقیقت الگویی پیش‌گو بوده و برای یادگیری بازنمایی‌های برداری از واژه‌ها استفاده می‌شود. هدف از روش‌های پیش‌گویانه این است که به کاوشگر اجازه می‌دهد یک ارزش ناشناخته را از یک متغیر مشخص پیش‌بینی نماید (قاضی‌زاده و خزانه‌ها ۱۳۹۵). در روش برداری فرض بر این است که بردارهای وزن–اصطلاحی[2] مربوط به آن دسته از اسنادی که به‌عنوان باربط شناخته شده‌اند، شباهت‌هایی با یکدیگر دارند (به‌عبارت دیگر اسناد باربط به یکدیگر شبیه‌اند). افزون بر این، فرض بر آن است که اسناد بی‌ربط بردارهایی هستند که با اسناد باربط شباهتی ندارند. بر این پایه، تشکیل بردار پیش‌گو را می‌توان به دو

1. Grossman and Frieder
2. term- weight

روش کیف واژه‌های پیوسته[1] و الگوی اسکیپ‌گرام[2] انجام داد (Mikolov et al. 2013). برای افزایش دقت این روش، مجموعهٔ دادهٔ اولیه که برای یادگیری و آموزش الگو مورد نیاز است، حدود چند میلیارد واژه را که درون چندین میلیون سند یا متن به‌کار رفته، توسط فنون یادگیری عمیق طبقه‌بندی و در نهایت، سازماندهی می‌شود (Chen, Guo & Bai 2017). یادگیری و آموزش الگو، گونه‌ای از یادگیری خودکار یک الگو یا قالبی از دادگان است که به‌منظور فهم ماهیت اطلاعات و ساختن تصمیم معقول برای پرس‌وجوها و دادگان آتی اجرا می‌شود (قاضی‌زاده و خزانه‌ها ۱۳۹۵).

**۲-۳. سازماندهی و بازیابی به روش کیسه‌ای (تجمیعی) از واژگان معنایی**

در روش تجمیعی، تعدادی از واژگان موجود در متون در یک پیکر متنی به‌عنوان یک فرهنگ لغت در نظر گرفته می‌شود. سپس، بر اساس تعداد تکرار هر کدام از این واژه‌ها در متن، یک بردار به ازای آن متن ساخته می‌شود. در این روش، فرایند متن‌کاوی شامل ساختن ماتریس‌هایی از میزان فراوانی واژه‌ها در متن است. بر این پایه، دسته‌بندی اسناد بر اساس این فرهنگ لغت انجام می‌شود. در شرایطی که تعداد واژه‌های موجود در فرهنگ لغت بسیار زیاد باشد، چالش‌هایی از جمله ابعاد بالای ماتریس‌های استخراج‌شده و یا مسئلهٔ ماتریس‌های خلوت به‌وجود می‌آید. بنابراین، باید به‌دنبال روش‌هایی برای کاهش ابعاد مسئله بود (Hotho, Paaß 2005). در این گروه از روش‌ها، در فرایند متن‌کاوی اغلب معنای واژه‌ها و عنوان‌های موجود در متن برای کاهش ویژگی‌ها در نظر گرفته می‌شود. فنون مختلفی جهت کاهش ویژگی در این حوزه وجود دارد که می‌توان به خوشه‌بندی واژه‌ها با استفاده از هستی‌شناسی‌ها، ویکی‌ها و همچنین الگوریتم‌های الگوسازی عنوان اشاره کرد. در نظر گرفتن سطح معنایی واژگان در متن و نیز معانی واژه در بافت جمله، زمینهٔ بررسی واژه‌های چندمعنایی[3] را فراهم می‌کند (مهراد و فلاحتی فومنی ۱۳۸۴).

---

1. Continues Bag of Words (CBOW)

۲. در الگوی Skip-gram به‌جای حدس یک کلمه، چندین کلمه حدس زده می‌شود. در این الگو یک کلمه به‌عنوان ورودی داده شده و دو کلمهٔ قبل و دو کلمهٔ بعد از آن حدس زده می‌شود.

3. چند معنایی به حالتی اطلاق می‌گردد که در آن واژه‌ای دارای معانی متعدد اما مرتبط به هم باشد.

**۲-٤. الگوسازی عنوان با استفاده از هستی‌شناسی‌ها**

در روش پیشنهادی Zhao & Mao (2018) برای ارائهٔ یک متن، پس از استخراج واژه‌های موجود در متن، گروهی از آن‌ها بر اساس فراوانی تکرار به‌عنوان واژه‌های پایه‌ای انتخاب شده و برای هر واژهٔ پایه، خوشه‌ای از واژه‌ها که ارتباط معنایی بین آن‌ها با استفاده از ویکی‌ها استخراج شده است، استخراج گردیده و از معیار شباهت کسینوسی به‌عنوان درجهٔ عضویت فازی یک واژه به هر کدام از آن خوشه‌ها جهت سازماندهی و بازیابی آن متن استفاده می‌شود. در واقع، برای هر واژه برداری از مقادیر فازی شباهت استخراج گردیده است. سپس، از طبقه‌بندی تحلیل افتراق خطی[1] برای دسته‌بندی استفاده شده است. بر این پایه، با خوشه‌بندی واژه‌ها به ساخت مفاهیم معنایی جهت کاهش ویژگی‌های متون در راستای مواجهه با چالش‌های پیش‌گفته پرداخته می‌شود. در روش پیشنهادی Elhadad, Badran & Salama (2017) با بهره‌گیری از ساختار سلسله‌مراتبی هستی‌شناسی واژه‌هایی که با سایر دسته‌های نحوی ارتباطی ندارند، از فضای ویژگی‌ها حذف می‌گردد. در این روش، ابتدا هر متن الگوسازی می‌شود. سپس، به هر ویژگی استخراج‌شده، بر اساس تعداد تکرار آن در هر متن وزنی اختصاص داده می‌شود. در ادامه، با استفاده از معیارهای شباهت Meng, Huang & Gu (2013) به بررسی شباهت معنایی و نحوی واژه‌های استخراج‌شده پرداخته شده و تنها واژه‌هایی با میزان شباهت بیشتر انتخاب می‌گردد. این واژگان بر پایهٔ تعداد تکرار در متن‌ها وزن‌دهی می‌شوند. همچنین، در روش پیشنهادی «مورینو-گارسیا» و همکاران به بررسی چند مجموعه دادهٔ مختلف پرداخته شد. نتایج بررسی نشان داد که صحت سازماندهی به نحوهٔ استخراج مفاهیم مرتبط با واژه‌ها وابسته است. برای نمونه، نتایج بررسی چگونگی سازماندهی متون خبری در پژوهش آن‌ها نشان داد که با توجه به کم‌تکرار بودن واژه‌های کلیدی در متون خبری و وجود معانی مختلفی از واژه‌ها و همچنین، مبهم بودن کاربرد برخی از واژه‌ها که دارای بیش از یک معنا هستند، روش‌های مبتنی بر هستی‌شناسی‌ها در مجموعه‌ای از متون گسترده و بزرگ ناکارآمد هستند (Mouriño-García et al. 2016). بر این اساس، کاربرد هستی‌شناسی‌ها در فرایند سازماندهی و بازیابی دانشِ مستتر در متون وبی گوناگون و متنوع با حجم بزرگ را باید با احتیاط بیشتری مورد توجه قرار داد.

1. linear discriminative analysis

### ۲-۵. الگوسازی عنوان به‌کمک تجزیهٔ نامنفی ماتریس‌ها

در این روش به‌منظور شناسایی مفاهیم پنهان موجود در متون به استخراج ویژگی‌هایی جدید برای سازماندهی و بازیابی متون پرداخته می‌شود. این روش با یافتن مقادیر و بردارهای ویژه به حذف خشه‌ها[1] از فضای مسئله و سپس، به الگوسازی مفاهیم پنهان می‌پردازد (Arun, et al. 2010; Blei, Ng & Jordan 2003). این در حالی است که در روش تجمیعی از واژگان معنایی با اعمال وزن‌دهی ضریب واکشی اطلاعات متقابل نقطه‌ای به تجزیهٔ ماتریس ویژگی‌ها توجه می‌شود. در واقع، ابعاد مسئله به رخداد مفاهیم معنایی محدود خواهد شد که در هر متن، با توجه به واژگان همان متن رخ می‌دهد. از جمله نوآوری‌های روش پیشنهادی اعمال وزن‌دهی واکشی اطلاعات متقابل نقطه‌ای نسبت به فراوانی تکرار در روش پایهٔ تجمیعی است. این نحوهٔ وزن‌دهی در واقع، اهمیت هر واژه را در هر متن نسبت به کل پیکره می‌سنجد و مفاهیم معنایی با توجه به چنین نحوهٔ وزن‌دهی برای متون استخراج می‌شود. از این رو، پس از تجزیهٔ ماتریس ویژگی‌ها به ماتریس مفاهیم و متون، یک توزیع احتمال رخداد از مفاهیم در هر متن ایجاد می‌شود.

## ۳. مروری بر پیشینه‌های مرتبط

در زمینهٔ کاوش متن فعالیت‌ها و نوآوری‌های زیادی در جریان است. اثری قدیمی از Salton (1989) در این زمینه وجود دارد و مرورهای بعدی و جدیدتر توسط Sievert (1996) به انجام رسیده است. Luhn (1958) از اولین کسانی بود که نقش الگوهای به‌کارگیری واژگان را در یافتن واژه‌هایی که موضوع متن را به نحو بهتری منعکس می‌کنند، برای مقاصد نمایه‌سازی و چکیده‌نویسی خودکار تشخیص داد. تلاش‌های جاری مربوط به کنسرسیوم وب جهانی در پیاده‌سازی وب معنایی است. وب، مجموعه‌ای از میلیون‌ها متن و سند است که هیچ سازماندهی خاصی بین آن‌ها وجود ندارد. با این حال، اسناد وبی پیوندهایی به اسناد و متون دیگر دارند که نشان می‌دهد به‌نوعی با یکدیگر ارتباط دارند. در واقع، کاوش در وب در بیشتر موارد کاوش واژگان یا نمادهای خاص یا ترکیبی از آن‌ها در رکوردهای موجود در وب است (میدو[2] و همکاران ۱۹۹۲). بر این پایه، بر مبنای بررسی

1. noise
2. Meadow

متـون پیشـین در رابطـه بـا کاوش واژگان، روش تجمیعـی از واژگان معنایـی اولیـن بـار در سـال ۲۰۰۳ توسـط (Blei, Ng & Jordan 2003) ارائـه شـد. ایـن روش یـک شـیوهٔ بـدون ناظـر الگوسـازی عنـوان در متـن اسـت کـه بـا هـدف یافتـن مفاهیـم معنایـی پنهـان متناسـب بـا متـن بـه‌کار مـی‌رود و در کاهـش ابعـاد چالـش‌زای آن در مسـائل مختلفـی نظیـر تشـخیص چهریزه‌هـا و طبقه‌بنـدی متـون کاربـرد دارد (Chien, Lee & Tan 2018). در صـورت کاربـرد ایـن روش در تحلیـل متـن فـرض می‌شـود کـه هـر متـن دارای ترکیبـی از چنـد مفهـوم پنهـان اسـت و هـر مفهـوم پنهـان نیـز دارای توزیـع احتمالـی روی واژه‌هـای آن پیکـره اسـت (Blei, Ng & Jordan 2003; Wilson & Chew 2010). بنابرایـن، واژه‌هـای مشاهده‌شـده در هـر مفهـوم پنهـان متغیرهـای چندجملـه‌ای هسـتند و مفاهیـم متناسـب نیـز متغیرهـای «ال‌دی‌اِی» هسـتند (Chien, Lee & Tan 2018). در ایـن روش بـا اسـتفاده از توزیـع دریکلـه و توزیـع چندجملـه‌ای بـه تعییـن میـزان تأثیـر هـر واژه از پیکـره در هـر مفهـوم پنهـان و همچنیـن، توزیـع احتمـال رخـداد مفاهیـم پنهـان در متـون پرداختـه می‌شـود. بـر ایـن پایـه، وزن چندجملـه‌ای هـر واژه در هـر مفهـوم پنهـان و همچنیـن وزن هـر مفهـوم در هـر متـن بایـد در نظـر گرفتـه شـود. ایـدهٔ اصلـی در وزن دادن بـه واژگان در کاوش متن، افزایـش تـوان تمایـز واژه از طریـق تعییـن اهمیـت نسـبی آن در پرسـش اسـت. این‌گونـه وزن‌هـا بـرای رتبه‌بنـدی متـون از نظـر شـباهت بـه یـک پرسـش یـا شـباهت بـه متـن یـک سـند دیگـر نیـز کاربـرد دارد (میـدو و همـکاران ۱۹۹۲). بـر ایـن اسـاس، اغلـب بـرای محاسـبهٔ راسـتی‌آزمایی حداکثـری وقـوع یـک مفهـوم در متـن از الگوریتم‌هـای شبیه‌سـازی اسـتفاده می‌شـود (Griffiths & Steyvers 2004). بـا ایـن حـال، این‌کـه معنـای واژه‌ای را ثبـت کنیـم، بـه ایـن مفهـوم نیسـت کـه ایـن واژه همه‌جـا همـان معنـا را خواهـد داشـت، زیـرا یـک واژه در هـر جـا می‌توانـد معنـای متفاوتـی داشـته باشـد و خواننـدگان نیـز ممکـن اسـت در همـه جـا تفاوت‌هـا را تشـخیص ندهنـد.

«چـن» و همـکاران در پژوهشـی بیـان کردنـد کـه بـا توجـه بـه کم‌تکـرار بـودن کلمـات کلیـدی در متن‌هـای خبـری و وجـود معانـی مختلـف از واژه‌هـا و همچنیـن، مبهـم بـودن کاربـرد برخـی از واژه‌هـا کـه دارای بیـش از یـک معنـا هسـتند، از الگوریتـم «ال‌دی‌اِی» بـرای شناسـایی مفاهیـم موجـود در متـن جهـت سـازماندهی متـون اسـتفاده کردنـد (Chen et al. 2016). در روش پیشـنهادی آنـان بـا در نظـر گرفتـن ارتبـاط بیـن عنـوان یـک متـن و واژه‌هـای موجـود در آن، مقـدار احتمـال وقـوع یـک موضـوع بـرای یـک متـن محاسـبه و از نتیجـهٔ آن بـرای بهتـر کـردن نتایـج سـازماندهی و بازیابـی اسـتفاده شـده اسـت. در واقـع، ایـدهٔ اصلـی آن‌هـا ترکیـب اطلاعـات مربـوط بـه موضـوع متـن و فراوانـی تکـرار واژگان موجـود در آن متـن بـود. ایـن کاهـش بـا

انتخاب واژه‌هایی از متن انجام می‌گردد که مرتبط با عنوان آن متن هستند.

همسو با دیدگاه «میدو» و همکاران (۱۹۹۲)، یکی دیگر از روش‌های به‌کاررفته در حوزهٔ متن‌کاوی و سازماندهی دانش وبی، روش خوشه‌بندی متونی است که دارای ارزش‌ها و خصوصیت‌های مشابهی هستند یا توزیع فراوانی واژگان آن‌ها به یکدیگر مشابهت دارد. بررسی پیشینه‌های مرتبط با به‌کارگیری این روش نشان می‌دهد که مطالعات اولیه در این حوزه توسط «بونر» مطرح شده است. در این روش او ماتریسی از خصوصیات یک رکورد تدوین کرده و سپس، ماتریس دیگری تهیه نمود که نشان‌دهندهٔ مشابهت رکوردهای بازیابی‌شده در ماتریس اول بود. در ماتریس سوم نیز تعداد خوشه‌ها را می‌توان مشخص کرد. سرانجام از بین فهرست خوشه‌های برگزیده، زیرمجموعه‌ای که بهترین تناسب را با کاربرد مورد نظر دارد، می‌توان انتخاب نمود (Bonner 1964). پس از مطالعات انجام‌شده توسط «بونر»، سایر پژوهشگران الگوریتم‌های متعددی را برای ایجاد خوشه‌ها طراحی و پیشنهاد دادند (Miyamoto 1990).

همان‌طور که مشاهده می‌شود، نمونه‌های محدود انجام‌شده در پژوهش‌های پیشین پس از فاز پیش‌پردازش، حد بالا و پایینی از فراوانی تکرار واژه برای انتخاب ویژگی‌های اولیه را در نظر می‌گیرند و واژگانی را که دارای فراوانی تکرار کمتر از حد آستانهٔ پائین و بیشتر از حد آستانهٔ بالا باشند، از مجموعهٔ فرهنگ لغات حذف می‌کنند. همچنین، از فراوانی تکرار واژه‌ها به‌عنوان وزن مؤثر آن‌ها استفاده شده و دلیل حذف واژه‌ها به‌علت لحاظ نشدن این مسئله در الگوریتم «ال‌دی‌اِی» است. در برخی از پژوهش‌های پیشین (مانند پژوهش Chen, Guo & Bai 2017; Wilson & Chew 2010; Yang et al. 2017) پژوهشگران نشان دادند در صورتی که از وزن‌دهی متناسب استفاده شود، نیازی به حذف چنین واژه‌هایی از مجموعهٔ فرهنگ لغات هر متن نیست و می‌توان با وزن‌دهی متناسب از حذف واژه‌هایی که ممکن است از وزن مطلوبی برخوردار باشند، جلوگیری کرد. همچنین، برخی از پژوهش‌های مرتبط دیگر نیز نشان از به‌کارگیری الگوریتم‌های خوشه‌بندی به‌منظور سازماندهی و بازیابی دانش متون دارد. بر این اساس، بررسی پژوهش‌های پیشین در این زمینه نشان می‌دهد که مطالعات انجام‌شده در این حوزه موضوعی چندان گسترده نیست و با توسعه و آزمون سایر روش‌های پیشنهادی در متن‌های مختلف وبی به احتمال بتوان روند سازماندهی دانش آشکار و پنهان موجود در متون وبی را تسهیل نمود. از این رو، پژوهش حاضر در تلاش است که با حذف یا تعدیل چالش‌های پیش‌گفته در راستای سازماندهی

و بازیابی دانش مندرج در متون وبی، روش‌های جدیدی معرفی و پس از پیش‌پردازش، استخراج ویژگی‌ها، یادگیری و آموزش الگو را انجام داده و در نهایت، به آزمون و ارزیابی روش پیشنهادی بپردازد. در ادامه، هر کدام از مراحل ذکرشده به تفکیک بیان می‌گردد.

## ٤. روش پژوهش

در این بخش به بیان مراحل مختلف روش پژوهش پرداخته شده است. روش پیشنهادی مورد مطالعه در پژوهش شامل مراحل پیش‌پردازش، استخراج ویژگی، تشکیل ماتریس واژگانی، فاز آموزش الگو و طبقه‌بندی است. نمودار ۲، شمایی از روش پیشنهادی پژوهش را نشان می‌دهد.


**شکل ۲. شمایی از مراحل روش پیشنهادی پژوهش**

از آنجا که روش پیشنهادی پژوهش حاضر برای پاسخ‌گویی به نیاز محققان در زمینهٔ تحلیل عظیم داده‌ها سازمان یافته و تا جایی که می‌دانیم مجموعهٔ دادهٔ خبری استانداردی با حجم عظیم به زبان فارسی موجود نیست، به‌منظور طراحی، پیکرهٔ عظیمی از داده‌های متنی ۲۰ گروه خبری خارجی که توالی انتشار منظمی داشتند، انتخاب شد و کل داده‌های متنی منتشرشدهٔ آن‌ها در فضای وب، پیکرهٔ متنی پژوهش را تشکیل داد. جدول ۱، ویژگی‌های مجموعهٔ داده‌های گروه‌های ۲۰گانهٔ خبری را نشان می‌دهد. این مجموعهٔ داده دارای حدود ۲۰۰۰۰ سند خبری از ۲۰ گروه خبری مختلف است و از مجموعهٔ داده‌های متداول در زمینهٔ متن‌کاوی به شمار می‌رود. اکثر پژوهش‌هایی که به‌نوعی روشی جدید در زمینهٔ سازماندهی دانش ارائه داده‌اند، از این مجموعهٔ داده استفاده نموده‌اند. در ادامه، روند اجرای پژوهش در قالب روش پیشنهادی (شکل ۲) به‌صورت تجربی مورد بررسی قرار می‌گیرد.

## ۵. تبیین مراحل اجرای روش پیشنهادی پژوهش

### ۵-۱. پیش‌پردازش

اولین گام در فرایند متن‌کاوی، فاز پیش‌پردازش است. از جملهٔ مراحل متداول در

فاز پیش‌پردازش شناسایی توکن[1]ها، حذف کلمات توقف و حروف اضافه است (Chen, Guo & Bai 2017 2017; Chien, Lee & Tan 2018; Zhao & Mao 2018). در این مرحله جهت آماده‌سازی متن، در هر متن ابتدا واژه‌ها از یکدیگر جداسازی و توکن‌ها شناسایی می‌شوند؛ حروف اضافه و کلمات توقف حذف شده و در نهایت، واژه‌های باقی‌مانده، فرهنگ لغات کل پیکرهٔ متون مورد مطالعه را تشکیل می‌دهند. این فرایند در پژوهش داده‌های متنی به‌دست‌آمده از گروه‌های خبری مورد مطالعه نیز به انجام رسید.

**۵-۲. استخراج ویژگی**

از جملهٔ روش‌های متداول برای کاهش ابعاد ویژگی‌ها روش «الدی‌اِی» است (Griffiths & Steyvers 2004). این الگوریتم، تعداد متون موجود در پیکره را به‌صورت یک توزیع احتمال دریکله رخداد k مفهوم پنهان در آن الگوسازی می‌کند و واژه‌های پیکره نیز دارای توزیع چندجمله‌ای روی مفاهیم پنهان هستند. همان‌طور که پیش‌تر بیان شد، روش «الدی‌اِی» به واژگان موجود در یک متن وبی با توجه به فراوانی تکرار آن‌ها اهمیت داده و آن‌ها را با توجه به فراوانی تکرار در مفاهیم توزیع می‌کند. به همین دلیل، واژه‌هایی با فراوانی تکرار بسیار بالا و همچنین واژه‌های کم‌تکرار از این مجموعه حذف می‌شوند. در صورتی که از وزن‌دهی مناسبی استفاده شود، می‌توان از حذف واژه‌های دارای ارزش جلوگیری کرد و در واقع، واژه‌ها را دارای یک ضریب تأثیر در نظر گرفت که در توزیع شدن آن‌ها بین مفاهیم مؤثر باشد. میزان اطلاع یک واژه در پیکره را با استفاده از آنتروپی آن می‌سنجند. به‌عبارت دیگر، هرچه آنتروپی بیشتر باشد، واژه دارای بار اطلاعاتی کمتری در پیکره است. حال، رخداد واژه‌ای از پیکره در یک سند را در نظر می‌گیریم. آنچه به واژه در سند ارزش می‌بخشد، میزان اطلاع آن واژه از موضوع آن سند است. بنا بر اصل نظریهٔ اطلاعات و نظریهٔ احتمال شرطی، میزان اطلاعی که یک واژه از یک سند در بر دارد، با استفاده از معیار واکشی اطلاعات متقابل رخداد آن واژه در سند نسبت به کل پیکرهٔ متن سنجیده می‌شود. یک واژه در یک پیکرهٔ متنی را می‌توان یک رویداد در نظر گرفت که در کنار واژه‌های دیگر در یک سند متنی بخشی از اطلاع راجع به آن سند را تشکیل می‌دهد. بنابراین، می‌توان از معیار واکشی اطلاعات متقابل نقطه‌ای جهت

1. token

محاسبهٔ ارزش واژه در سند استفاده کرد (Wilson & Chew 2010). معیار واکشی اطلاعات متقابل با محاسبهٔ مقدار متوسط اطلاعات متقابل نقطه‌ای به بررسی میزان وابستگی متقابل بین دو متغیر تصادفی می‌پردازد؛ در حالی که واکشی اطلاعات متقابل نقطه‌ای به محاسبهٔ میزان اطلاع واژه به‌ازای رخدادش در سند می‌پردازد. این است که از این معیار به‌منظور وزن‌دهی به واژه‌ها در سند استفاده شد.

الگوریتم «ال‌دی‌ای» با استخراج ویژگی‌هایی جدید تحت عنوان مفاهیم معنایی، نقش به‌سزایی در کاهش ابعاد فضای مسئله دارد. بر طبق مطالب پیش‌گفته، در صورتی که هر مفهوم را یک سبد در نظر بگیریم که دارای توپ‌هایی با رنگ‌های مختلف (هر رنگ به ازای هر لغت) و اندازه‌های مختلف است، توزیع شدن توپ‌ها در هر کدام از سبدها با توجه به اندازهٔ آن‌ها (که توسط معیار واکشی اطلاعات متقابل تعیین شد) و ضریب چندجمله‌ای انجام می‌پذیرد. بنابراین، روابط یادشده به شکل زیر بازنویسی و به اجرا درآمد:

$$p(z_{ij} = k \mid \emptyset^k) = \frac{w(x_i) N_{i(.)k}^{-ij} + \beta}{\sum_w m(w)\, N_{(.)(.)k}^{-ij} + W_\beta}$$

$$p(k \mid d_j) = \frac{\sum_w m(w)\, N_{(.)jk}^{-ij} + \alpha}{\sum_w m(w)\, N_{(.)(.)k}^{-ij} + T_\alpha}$$

$w(x_i)$ وزن لغت $x_i$ است که در ضریب چندجمله‌ای تأثیر گذاشته و بر اساس رابطهٔ زیر محاسبه می‌گردد:

$$PMI\left(x_i, d\right) = -log2\ \frac{p(x_i|d)}{p(xi)} =$$

$$\frac{(word\ x_i\ in\ document\ d)}{\#(word\ x_i)}$$

**۵-۳. سازماندهی متون**

خروجی فرایند استخراج ویژگی‌ها تعیین مفاهیم پنهان است. از آنجا که در ماتریس متون، مفاهیم هر متن دارای برچسب است و به شکل یک توزیع احتمال روی تعدادی

مفهوم بازیابی شده است، متونی که دارای توزیع‌های مشابهی روی مفاهیم باشند، به احتمال دارای برچسب مشابه نیز هستند. «هینز» برچسب‌گذاری متون مشابه را از زمرهٔ ویژگی‌های میان‌کنش‌پذیری نظام‌های بازیابی می‌داند (۲۰۰۴). بنابراین، برای طبقه‌بندی می‌توان از معیار فاصلهٔ «کولبک‌لیبر» که فاصلهٔ دو توزیع احتمال در بین متون مشابه را می‌سنجد، استفاده کرد (Wang, Kulkarni & Verdu 2009). باید این واقعیت را پذیرفت که ساختارهای لغوی با وجود ارتباط داشتن با یکدیگر، جدا از هم به نظر می‌رسند؛ لیکن تعیین ارتباط آن در قالب نظام‌های بازیابی دارای پیچیدگی‌های روان‌شناختی و زبانی است. در این پژوهش نیز برای مواجهه با این چالش از الگوریتم زیر استفاده شد:

$$KLD\left(p\ ،\ q\right) = p * Log_2\left({}^{p}/_{q}\right)$$

**٥-٤. آزمون شبیه‌سازی**

آزمایش شبیه‌سازی به‌منظور بررسی قابلیت سازماندهی و بازیابی مجموعه‌ای از متون در یک پیکرهٔ بزرگ طراحی می‌شود. در این پژوهش از مجموعهٔ داده‌های ۲۰ گروه خبری وبی و اخبار منتشرشدهٔ آن‌ها تحت وب استفاده شد. این مجموعهٔ داده دارای پیکره‌ای به بزرگی ۱۲۸۰۰۰ واژه و ۲۰۰۰۰ متن بود. داده‌های به‌کاررفته در پژوهش حاضر را می‌توان به‌عنوان یکی از بزرگ‌ترین پیکره‌های متنی خبری استاندارد تحت وب برشمرد. جدول ۱، ویژگی‌های این مجموعهٔ داده را نشان می‌دهد.

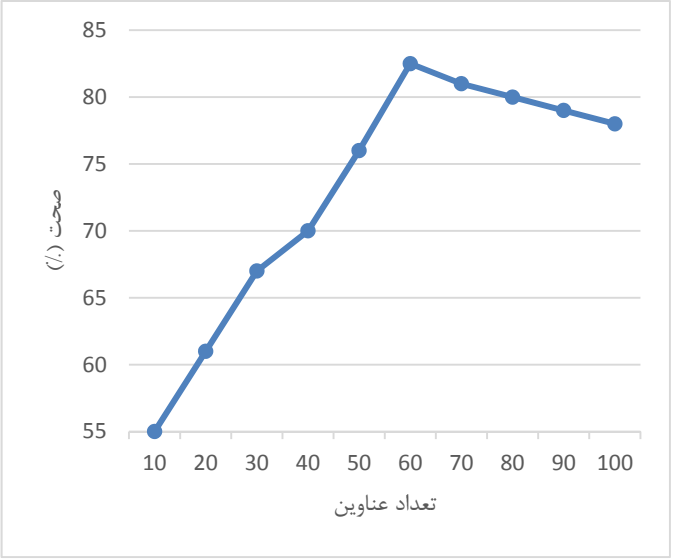
**جدول ۱. ویژگی‌های مجموعهٔ دادهٔ ۲۰ گروه خبری مورد مطالعه**

| گروه | تعداد اسناد |
|---|---|
| alt.atheism | 1000 |
| comp.graphics | 999 |
| comp.os.ms-windows.misc | 1000 |
| comp.sys.ibm.pc.hardware | 1000 |
| comp.sys.mac.hardware | 1000 |
| talk.politics.mideas | 1000 |
| comp.windows.x | 1000 |
| misc.forsale | 1000 |

| تعداد اسناد | گروه |
|---|---|
| 1000 | rec.autos |
| 1000 | rec.motorcycles |
| 1000 | rec.sport.baseball |
| 1000 | talk.politics.guns |
| 1000 | talk.religion.misc |
| 1000 | rec.sport.hockey |
| 1000 | sci.crypt |
| 1000 | sci.electronics |
| 1000 | sci.med |
| 1000 | sci.space |
| 997 | soc.religion.christian |
| 1000 | talk.politics.misc |

### ۵-۵. تنظیم پارامترهای «ال‌دی‌اِی»

بـر مبنـای مطالـب پیش‌گفتـه، قبـل از اعمـال روش «الدی‌ای» بایـد در خصـوص تعـداد مفاهیـم پنهـان تصمیم‌گیـری شـود. می‌تـوان تعـداد مفاهیـم موجـود را بـا اسـتفاده از آزمـون و خطـا و مقایسـهٔ دقـت روش در هـر تکـرار محاسـبه کـرد. شـکل ۳، دقـت روش پیشـنهادی را بـرای مقادیـر مختلـف واژه کـه تعـداد مفاهیـم پنهـان اسـت، آزمـوده اسـت.

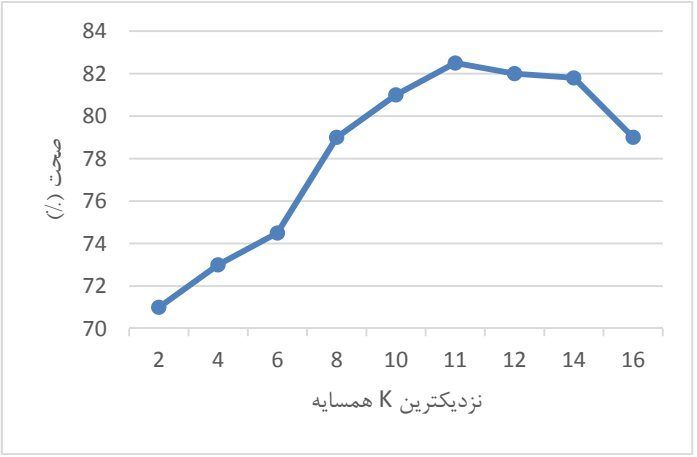



**شکل ۳. صحت طبقه‌بندی به ازای تعداد عناوین**

نتایج آزمایش این روش نشان داد که در k = 60، بهترین دقت خود را که ۸۲/۵ درصد است، کسب نموده است. دومین پارامتر در استفاده از الگوریتم استخراج مفاهیم پنهان، انتخاب تعداد اصطلاح‌های موجود در هر مفهوم است. روش‌هایی که از این الگوریتم و این مجموعهٔ داده استفاده کرده‌اند، اغلب از ۱۰ اصطلاح پرارزش یک عنوان استفاده و تعداد مفاهیم را بر اساس تعداد ویژگی‌های اولیه انتخاب کرده‌اند.

### ٥-٦. تنظیم پارامتر الگوریتم

در الگوریتم نزدیک‌ترین k همسایه، تنظیم پارامتر k که تعداد شبیه‌ترین همسایه است، اغلب با آزمایش تنظیم می‌شود. شکل ۴، نشان‌دهندهٔ دقت روش در مقادیر مختلف k است. نتایج نشان داد که در k = 12 روش، بیشترین دقت را کسب نموده است.




**شکل ٤. دقت طبقه‌بندی به‌ازای مقادیر مختلف از K همسایه در الگوریتم KL-KNN**

### ۵-۷. ارزیابی

در بخش قبل، ویژگی‌های مجموعهٔ داده و آزمایش‌های مربوط به تنظیم پارامترهای روش به‌کاررفته در پژوهش بیان شد. در این بخش، ابتدا به بررسی دقت طبقه‌بندی روش پیشنهادی بر هر کدام از دسته‌های مجموعهٔ دادهٔ متنی تحت وب از ۲۰ گروه خبری پرداخته می‌شود. سپس، به مقایسهٔ دقت KL-KNN[1] با سایر دسته‌بندها و روش‌هایی که از «الدی‌اِی» برای استخراج مفاهیم معنایی استفاده کرده‌اند، پرداخته و در نهایت، مقایسهٔ دقت روش با سایر روش‌های ذکرشده در بخش مفاهیم پایه بیان می‌گردد. به باور «پائو» ارزیابی به‌معنای داوری دربارهٔ ارزشمندی و شایستگی موضوع مورد ارزیابی است و باید امور و عملکرد نظام سازماندهی و بازیابی مورد توجه قرار گیرد (۱۹۸۹). ارزیابی روش‌ها بر اساس ویژگی دقت[2] که با استفاده از رابطهٔ زیر محاسبه می‌شود، انجام گرفته است. در واقع، دقت، حاصل تقسیم تعداد اسنادی که به‌درستی دسته‌بندی شده‌اند بر تعداد کل اسناد است. اسنادی که به‌درستی دسته‌بندی شده‌اند، شامل دو دسته است: مثبت درست[3] (روش

---

۱. KNN یکی از ساده‌ترین الگوریتم‌های داده‌کاوی و طبقه‌بندی است. این الگوریتم عملیات طبقه‌بندی را به‌صورت ساده انجام داده و نتایج قابل اطمینانی به‌عنوان پیش‌بینی برمی‌گرداند.

2. accuracy

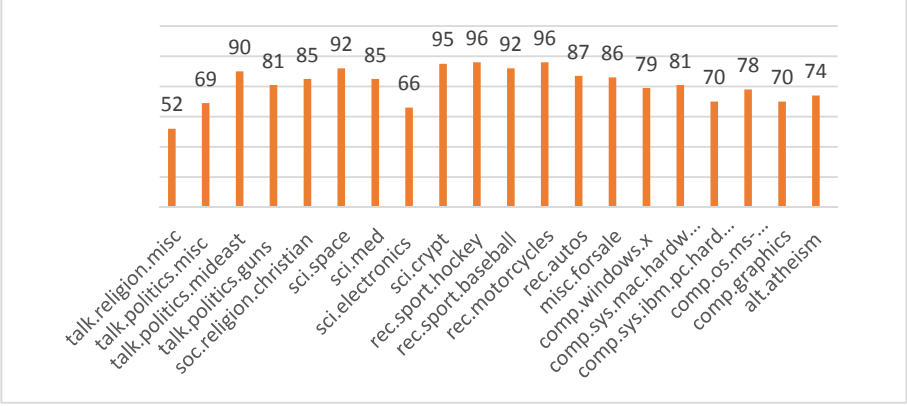
3. true positive

به‌درستی آن را در دسته‌ای که به آن تعلق داشته، جا داده است) و منفی درست[1] (روش به‌درستی آن را در دسته‌ای که به آن تعلق نداشته، جا نداده است).

$$Accuracy = \frac{tp + tn}{n}$$

### ٦. ارزیابی دقت روش پیشنهادی

همان‌طور که پیش‌تر اشاره شد، مجموعهٔ دادهٔ پژوهش متشکل از ۲۰ گروه خبری دارای ۲۰ دستهٔ مختلف است که در این بخش به بررسی دقت الگوریتم هر دسته پرداخته شده است. پس از فاز پیش‌پردازش به استخراج توزیع احتمال رخداد هر کدام از مفاهیم معنایی روی هر متن پرداخته شد و با فرض داشتن یک فضای توزیع احتمال، از معیار «کولبک‌لیبر» به‌عنوان میزان شباهت در الگوریتم طبقه‌بند KNN استفاده گردید. شکل ۵، دقت روش پیشنهادی در هر کدام از دسته‌های مجموعهٔ دادهٔ گروه‌های خبری را نشان می‌دهد.



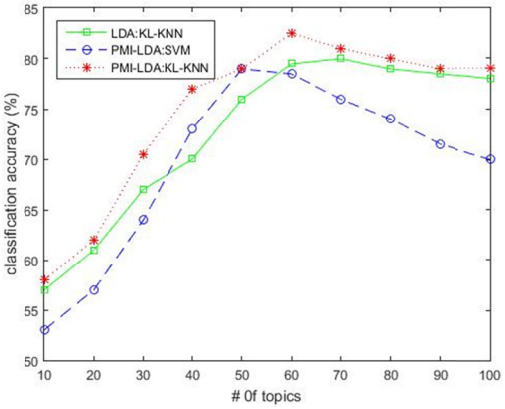

**شکل ۵. دقت روش پیشنهادی بر هر کدام از دسته‌ها در مجموعهٔ دادهٔ گروه‌های خبری**

#### ٦-١. مقایسهٔ دقت KL_KNN با سایر دسته‌بندها

در این بخش به مقایسهٔ دقت روش پیشنهادی با روش‌هایی که از «الدی‌ای» به‌عنوان روش کاهش ویژگی استفاده نموده‌اند، می‌پردازیم. این روش‌ها اغلب از طبقه‌بند ماشین

1. true negative

بردار پشتیبانی[1] (Chen et al. 2016; Pereira et al. 2018; Yang et al. 2017) یا طبقه‌بند بیز ساده[2] (Kim, Kim & Jinseog 2016; Mouriño-García et al. 2016) به‌عنوان طبقه‌بند استفاده نموده‌اند. در روش پیشنهادی پژوهش حاضر از الگوریتم KL-KNN به‌عنوان طبقه‌بند استفاده شده است. در شکل ۶، دقت الگوریتم پایهٔ «الدی‌اِی» و الگوریتم پیشنهادی «الدی‌اِی» واکشی اطلاعات متقابل نقطه‌ای[3] که در هر دو ویژگی‌های جدید استخراج‌شده توسط طبقه‌بند KL-KNN طبقه‌بندی‌شده مقایسه شده‌اند. نتایج نشان می‌دهد زمانی که از معیار وزن‌دهی واکشی اطلاعات متقابل نقطه‌ای استفاده شده، دقت بهتری نسبت به روش پایهٔ «الدی‌اِی» کسب شده است. همچنین، در شکل ۶، دقت الگوریتم «الدی‌اِی» و واکشی اطلاعات متقابل نقطه‌ای با طبقه‌بندهای KL-KNN و ماشین بردار پشتیبانی مقایسه شده است. نتایج نشان می‌دهد که نظریهٔ دسته‌بندی داده‌ها با استفاده از KL_KNN در یک فضای توزیع احتمالی نتایج بهتری نسبت به روش خطی ماشین بردار پشتیبانی کسب نموده است.




**شکل ٦. مقایسهٔ دقت الگوریتم با طبقه‌بند ماشین بردار پشتیبانی و KL-KNN**

در پژوهش‌های Chen et al. (2016) و Chen & Li (2016) از دسته‌بند ماشین بردار پشتیبانی جهت طبقه‌بندی داده‌ها استفاده شد. نتایج روش پیشنهادی پژوهش حاضر نشان می‌دهد

---

۱. ماشین بردار پشتیبانی یا Support vector machine (SVM) یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. مبنای کار این طبقه‌بند دسته‌بندی خطی داده‌هاست و جزو الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود.

۲. به‌طور ساده، طبقه‌بند بیز ساده (naive bayes) روشی برای دسته‌بندی پدیده‌ها، بر پایهٔ احتمال وقوع یا عدم وقوع یک پدیده است.

3. Point wise Mutual Information-Latent Drichlet Allocation -PMI-LDA

کـه طبقه‌بنـدی بـا اسـتفاده از معیـار «کولبک‌لیبـر» نتایـج بهتـری را بـر مجموعـهٔ دادهٔ گروه‌هـای خبـری مـورد مطالعـه کسـب نمـوده اسـت. ایـن امـر بیانگـر ایـن مطلـب اسـت کـه دسـته‌بندی داده‌هـای پژوهـش بـا اسـتفاده از الگـوی توزیـع احتمـال آن‌هـا حـول مفاهیـم معنایـی دقـت بهتـری نسـبت بـه دسـته‌بندی خطـی بـا ماشـین بـردار پشـتیبانی کسـب نمـوده اسـت.

**٦-۲. مقایسهٔ دقت روش پیشنهادی با سایر روش‌ها**

در بخـش مفاهیـم پایـه، دسـته‌بندی انـواع روش‌هـای مرتبـط بـا طبقه‌بنـدی متـن ارائـه شـد. در ایـن بخـش بـه مقایسـهٔ دقـت روش پیشـنهادی بـا روش‌هـای بیان‌شـده در بخـش مفاهیـم پایـه پرداختـه شـده اسـت. جـدول ۲، مقایسـه‌ای بیـن دقـت روش‌هـای مختلـف و همچنیـن، روش پیشـنهادی پژوهـش بـر مجموعـهٔ دادهٔ وبـی گروه‌هـای خبـری پژوهـش را نشـان می‌دهـد. بـر اسـاس نتایـج آزمایش‌هـا، مشـاهده می‌شـود کـه دقـت روش پیشـنهادی نسـبت بـه برخـی روش‌هـا، کـه در بخـش مفاهیـم پایـه بیـان شـد، بهتـر شـده اسـت. در واقـع، دقـت روش نسـبت بـه روش‌ایـی کـه از هستی‌شناسـی‌ها و ویکی‌هـا بـرای کاهـش ویژگـی اسـتفاده نموده‌انـد (Chakraverty et al. 2015; Mouriño-García et al. 2016; Zhao & Mao 2018) و یـا روش‌هایـی کـه بـا اعمـال فنـون وزن‌دهـی بـدون کاهـش ویژگـی فراینـد سـازماندهی و بازیابـی را انجـام داده‌انـد (Chen et al. 2016)، بهبـود داشـته اسـت. همچنیـن، روش پیشـنهادی پژوهـش دارای دقتـی مشـابه بـا روش‌هایـی اسـت کـه از الگـوی مخلـوط دریکلـه بـرای الگوسـازی مفاهیـم پنهـان اسـتفاده کرده‌انـد (Chien, Lee & Tan 2018). ایـن روش دارای پیچیدگـی محاسـباتی بسـیار بالایـی اسـت و نیازمنـد دقـت در یافتـن تعـداد مخلوط‌هـای دریکلـه اسـت؛ در حالـی کـه روش پیشـنهادی پژوهـش بـا اعمـال وزن‌دهـی متفـاوت بـر الگوریتـم پایـهٔ «الدی‌اِی» دارای دقـت مشـابه یـا بالاتـری بـا ایـن روش اسـت. افـزون بـر ایـن، روش پیشـنهادی پژوهـش نسـبت بـه روش‌هایـی ماننـد پژوهـش (Yang et al. 2017) کـه از فنـون یادگیـری عمیـق و الگـوی بـردار بـه‌ازای هـر واژه اسـتفاده کرده‌انـد، دقـت مشـابهی را کسـب نمـود. ایـن مسـئله اهمیـت روش وزن‌دهـی و به‌کارگیـری روش کاهـش ابعـاد نسـبت بـه سـایر روش‌هـای پیچیـده را نشـان می‌دهـد. در واقـع، ویژگـی مهـم روش پیشـنهادی پژوهـش اعمـال وزن مؤثـر توزیـع واژه‌هـا در مفاهیـم بـر اسـاس میـزان اطـلاع آن‌هـا از موضـوع سـند، نسـبت بـه الگوریتـم پایـهٔ «الدی‌اِی» اسـت و همچنیـن بـا اسـتفاده از دسـته‌بند KL-KNN بـه سـازماندهی و بازیابـی مفاهیـم بـر اسـاس نحـوهٔ توزیـع احتمـال آن‌هـا حـول هـر طبقـه پرداختـه شـده اسـت. روش پیشـنهادی Gupta, Kumar & Pant (2018) بـا در نظـر گرفتـن چالـش عظیـم

داده‌ها، پیاده‌سازی روش خود را در بستر محیط توزیع‌شدهٔ «هادوپ»[1] انجام داده است. ویژگی پردازش موازی و روش توزیع داده‌ها در این محیط سبب افزون‌شدن دقت این روش نسبت به روش پیشنهادی گردیده است.

**جدول ۲. مقایسهٔ دقت روش پیشنهادی پژوهش با سایر روش‌های سازماندهی و بازیابی متون وبی**

| ردیف | نام | روش سازماندهی | سال | روش طبقه‌بندی | صحت |
|---|---|---|---|---|---|
| ۱ | تجمیعی از واژه‌ها – فازی | Fuzzy BOW +wiki | ۲۰۱۵ | ماشین بردار پشتیبان | ۷۵ درصد |
| ۲ | زنجیره‌ای از لغات زمینه‌ای | بهبودی بر تجمیعی از واژه‌ها | ۲۰۱۵ | TF-IDF-RCF + ماشین بردار پشتیبان | ۹۰ درصد (بخشی از پیکره) |
| ۳ | تجمیعی از مفاهیم زمینه‌ای برای طبقه‌بندی بیزی | LSA | ۲۰۱۶ | طبقه‌بند بیزی | ۷۵ درصد |
| ۴ | تجمیعی از مفاهیم معنایی دریکله | مفاهیم معنایی | ۲۰۱۶ | LDA+SVM | ۹۴ درصد (بخشی از پیکره) |
| ۵ | تجمیعی از مفاهیم معنایی | Skip-gram +برچسب اسناد | ۲۰۱۷ | شبکهٔ عصبی عمیق | ۸۲/۳ درصد |
| ۶ | توزیع مخلوط دریکله برای الگوسازی مفاهیم معنایی | مفاهیم معنایی | ۲۰۱۷ | مخلوط دریکله نظارت‌شده | ۸۲ درصد |
| ۷ | تجمیعی از واژه‌ها وزن‌دهی بر اساس چگالی | بهبود یافته BOW | ۲۰۱۷ | ماشین بردار پشتیبان | ۷۹ درصد |
| ۸ | طبقه‌بندی متن در محیط توزیع‌شده | BOW | ۲۰۱۸ | طبقه‌بند بیزی در محیط هادوپ | ۸۸/۸ درصد |
| ۹ | روش پیشنهادی | تجمیعی از مفاهیم معنایی وزن‌دهی‌شده بر اساس اطلاعات متقابل نقطه‌ای | ۲۰۱۷-۲۰۱۸ | طبقه‌بند نزدیک‌ترین k همسایه با معیار واگرایی کولبک لیبلر | ۸۲/۵ درصد |

## ۷. بحث و نتیجه‌گیری

در این پژوهش ابتدا به مرور گروهی از روش‌های مختلف سازماندهی و بازیابی متون پرداخته شد و ویژگی‌های هر گروه بیان گردید. از هر کدام از این روش‌ها می‌توان جهت

1. Hadoop

سازماندهی و بازیابی متون مبتنی بر وب استفاده کرد. از جمله مشکلات موجود در تمامی این روش‌ها، بالا بودن تعداد ویژگی‌ها در مسئلهٔ سازماندهی متن است. از این رو، در اینجا به مطالعهٔ الگوریتم «الدی‌اِی» به‌عنوان یک روش احتمالی در تجزیهٔ غیرمنفی ماتریس‌ها جهت کاهش فضای ویژگی‌ها تمرکز گردید. در این الگوریتم تعدادی از ویژگی‌های متون مورد مطالعه که دارای فراوانی تکرار کم و یا بیشتر از حد آستانه هستند، حذف می‌گردند که ممکن است در این بین کلماتی که دارای نقش مهمی هستند نیز حذف شوند. همچنین، از معیار فراوانی تکرار واژگان در پیکرهٔ متنی به‌عنوان وزن آن واژه استفاده می‌گردد؛ بدون این‌که به درجهٔ اهمیت آن واژه در متن توجه شود. بنابراین، از آنجا که انتخاب مجموعهٔ ویژگی‌های هر متن نقش به‌سزایی در دقت الگوریتم‌های سازماندهی و بازیابی محتوای آن دارد، در این پژوهش یک روش وزن‌دهی متفاوت با روش پایه در الگوریتم «الدی‌اِی» برای سازماندهی و بازیابی متون مورد نظر در پژوهش (گروه‌های خبری) استفاده شد. در روش پیشنهادی، پس از فاز پیش‌پردازش، که در آن تنها حروف اضافه و واژه‌های توقف حذف می‌شوند، سایر واژگان متن حفظ می‌گردد. وزن‌دهی با هدف تأثیر دادن میزان اهمیت هر واژه در متن که با معیار واکشی اطلاعات متقابل نقطه‌ای محاسبه گردید و همچنین، فراوانی تکرار آن در پیکره انجام شد. نتایج آزمایش‌ها نشان داد که زمانی که از معیار واکشی اطلاعات متقابل نقطه‌ای در الگوریتم «الدی‌اِی» استفاده شود، دقت سازماندهی و بازیابی محتوا به‌گونه‌ای محسوس بهبود می‌یابد. برای دسته‌بندی ماتریس متون و مفاهیم استخراج‌شده توسط الگوریتم «الدی‌اِی» و واکشی اطلاعات معیار، از طبقه‌بند KNN بر اساس معیارهای فاصله‌ای استفاده شد. این دسته‌بند با استفاده از معیار KLD فاصلهٔ دو توزیع احتمال رخداد مفاهیم معنایی در هر متن را با یکدیگر سنجید. بر پایهٔ یافته‌های پژوهش، نتایج نشان داد که این نحوهٔ سازماندهی و بازیابی محتوای متن معیار خوبی برای سازماندهی و بازیابی سایر متون تحت وب بر اساس موضوع آن‌ها با توجه به احتمال رخداد مفاهیم مختلف در آن‌ها، نسبت به سایر روش‌های سازماندهی و بازیابی محتوا مانند دسته‌بند ماشین بردار پشتیبانی (SVM) است.

همسو با دیدگاه «ویکری و ویکری» (۱۹۸۹) با توجه به اهمیت وزن‌دهی مناسب واژگان که مورد تأکید پژوهش نیز است، سازمان معناشناختی واژگان می‌تواند به شکلی مفید با یکدیگر در راستای وزن‌دهی مطلوب واژه‌ها و نشانه‌های متنی در تعامل باشند. تمام عناصر طراحی‌شده در هر الگوی ذخیره، سازماندهی و بازیابی در ابتدا با زبان بیان می‌شود. از این

رو، زبان‌شناسی نیز می‌تواند در تمامی زمینه‌های چالش برانگیز مورد نظر پژوهش بصیرت لازم را ایجاد نماید. چالش بزرگ پیشِ رو این است که محدودیت‌های ساختارهای معنایی نظام‌های بازیابی همواره نمی‌تواند مطابق با نیازهای پرسشگران گوناگونی تنظیم گردد. با این حال، توسعهٔ پژوهش‌هایی از این دست می‌تواند تسهیل‌کنندهٔ حل چالش‌هایی مانند تعدد ویژگی‌های متنی یا محدودیت‌های ماتریس ویژگی‌ها در راستای سازماندهی و بازیابی مطلوب‌تر دانش آشکار و پنهان در متون عظیم وب‌پایه باشد. این پژوهش به‌عنوان نمونه‌ای کوچک به‌منظور به تصویرکشیدن کارکردهای متن‌کاوی در راستای سازماندهی و بازیابی دانشِ منتشرشده در محیط وب طراحی گردید. بر این پایه، با توجه به قابلیت‌های انجام و تکرار پژوهش در محیط‌های مشابه پیشنهاد می‌گردد که از روش‌های به‌کاررفته در این پژوهش در الگوسازی و بازنمایی دانش در محیط‌های وب‌پایه از قبیل امور بانکی، متن‌کاوی برای مدیریت مؤسسات آموزشی، مدیریت بهینهٔ وب‌سایت‌ها، مدیریت دانش، بررسی کتابخانه‌های دیجیتال، فرایندهای آموزش عالی، امور پزشکی و تجاری و ... نیز استفاده شود. این امر نه‌تنها منجر به مقایسه و بازآزمون روش پیشنهادی پژوهش حاضر می‌گردد، بلکه زمینهٔ غنای ادبیات نظری و پژوهش‌های عملیاتی بیشتر در این حوزه را با رویکردی بین رشته‌ای در پی خواهد داشت.

## فهرست منابع


پائو، میراندا لی. ۱۹۸۹. *مفاهیم بازیابی اطلاعات*. ترجمهٔ اسدا... آزاد و رحمتا... فتاحی. ۱۳۸۰. مشهد: دانشگاه فردوسی مشهد.

قاضی‌زاده، حمید، و مهدیه خزانه‌ها. ۱۳۹۵. *داده‌کاوی در علم اطلاعات و دانش‌شناسی مفاهیم و کاربرد*. تهران: چاپار، اساطیر پارسی.

گراسمن، دیوید، و افیر فریدر. ۲۰۰۴. *بازیابی اطلاعات: الگوریتم‌ها و روش‌های اکتشافی*. ترجمهٔ جعفر مهراد و سارا کلینی. ۱۳۸۴. مشهد: انتشارات کتابخانه رایانه‌ای؛ شیراز: کتابخانهٔ منطقه‌ای علوم و تکنولوژی.

مهراد، جعفر، و محمدرضا فلاحتی فومنی. ۱۳۸۴. *معناشناسی و بازیابی اطلاعات: هفت گفتار*. شیراز: کتابخانهٔ منطقه‌ای علوم و تکنولوژی.

میدو، چالز تی.، برت آر. بویس، دونالد اچ. کرفت، و کارول باری. ۱۹۹۲. *نظام‌های بازیابی اطلاعات متنی*. ترجمهٔ نجلا حریری. ۱۳۹۰. تهران: چاپار.

ویکری، برایان، و الینا ویکری. ۱۹۸۹. *علم اطلاع‌رسانی در نظر و عمل*. ترجمهٔ عبدالحسین فرج‌پهلو. ۱۳۸۰. مشهد: دانشگاه فردوسی مشهد.

هینز، دیوید. ۲۰۰۴. *ابرداده برای مدیریت و بازیابی اطلاعات*. ترجمهٔ علیرضا سعادت علیجانی و فاطمه ذاکری فرد. ۱۳۸۵. تهران: چاپار.

## References

Arun, R., V. Suresh, C. E. Veni Madhavan, & M. N. Narasimha Murthy. 2010. *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*. Berlin: Heidelberg.

Blei, D. M., A. Y. Ng, & M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3.993-1022 :

Bonner, R. E. 1964. On some clustering techniques. *IBM Journal of Research and Development* 8:(1) 22-32. doi:10.1147/rd.81.0022

Chakraverty, S., B. Juneja, U. Pandey, & A. Arora. 2015. *Dual lexical chaining for context based text classification.* Paper presented at the 2015 International Conference on Advances in Computer Engineering and Applications. Ghaziabad, India.

Chen, K., Z. Zhang, J. Long, & H. Zhang. 2016. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications* 66: 245-260. doi:https://doi.org/10.1016/j.eswa.2016.09.009

Chen, Q., X. Guo, & H. Bai. 2017. Semantic-based topic detection using Markov decision processes. *Neurocomputing* 242: 40-50. doi:https://doi.org/10.1016/j.neucom.2017.02.020

Chen, Y., & S. Li. 2016. *Using latent Dirichlet allocation to improve text classification performance of support vector machine.* Paper presented at the 2016 IEEE Congress on Evolutionary Computation (CEC). Vancouver, BC, Canada.

Chien, J.-T., C.-H. Lee, & Z.-H. Tan. 2018. Latent Dirichlet mixture model. *Neurocomputing* 278: 12-22. doi:https://doi.org/10.1016/j.neucom.2017.08.029

Elhadad, M. K., K. Badran, & G. I. Salama. 2017. *A novel approach for ontology-based dimensionality reduction for web text document classification.* Paper presented at the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). Wuhan, China.

Griffiths, T. L., & M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America. 101(1)*: 5228-5235. doi:10.1073/pnas.0307752101

Gupta, S., V. Kumar, & B. Pant. 2018. *Classification of Textual Data in Distributed Environment. 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T),* Allahabad, India*120-124* :.

Hotho A., N. A., G. Paaß. 2005. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* 20 (1): 19-62.

Järvelin, K., & P. Vakkari. 1993. The evolution of library and information science 1965–1985: A content analysis of journal articles. *Information Processing & Management* 29 (1): 129-144 .doi:https://doi.org/10.1016/0306-4573(93)90028-C

Kim, H., J. Kim, & K. Jinseog. 2016. *Semantic text classification with tensor space model-based naïve Bayes.* Paper presented at the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Budapest, Hungary.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2 (2): 159-165. doi:10.1147/rd.22.0159

Meng, L., R. Huang, & J. Gu. 2013. A review of semantic similarity measures in WordNet. *Journal of Intelligent & Fuzzy Systems* 36 (4): 3045-3059.

*Mikolov, T., G. S. Corrado, K. Chen, & J. Dean. 2013.* Efficient Estimation of Word Representations in Vector Space. A computing research. *arXiv preprint arXiv*:1301.3781.

Miyamoto, S. 1990. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Boston: Kluwer Academic Publishers.

Mouriño-García, M., R. Pérez-Rodríguez, L. Anido-Rifón, & M. Gómez-Carballa. 2016. *Bag-of-Concepts Document Representation for Bayesian Text Classification.* Paper presented at the 2016 IEEE International Conference on Computer and Information Technology (CIT). Nadi, Fiji.

Pereira, R. B., A. Plastino, B. Zadrozny, & L. H. C. Merschmann. 2018. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review* 49 (1): 57-78. doi:10.1007/s10462-016-9516-4.

Salton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley: Longman Publishing Co., Inc.

Sievert, M. C. 1996. Full-text information retrieval: Introduction. *Journal of the American Society for Information Science* 47 (4): 261-262. doi:10.1002/(sici)1097-4571(199604)47:4<261::aid-asi1>3.0.co;2-v

Thiyagarajan, D., & N. Shanthi. 2017. A modified multi objective heuristic for effective feature selection in text classification. *Cluster Computing journal* 11: 1-11, doi/10.1007:s10586-017-1150-7.

Wang, Q., S. R. Kulkarni, & S. Verdu. 2009. Divergence Estimation for Multidimensional Densities Via $k$-Nearest-Neighbor Distances. *IEEE Transactions on Information Theory* 55 (5): 2392-2405. doi:10.1109/TIT.2009.2016060.

Wilson, A. T., & P. A. Chew. 2010. *Term weighting schemes for Latent Dirichlet Allocation*. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California.

Yang, L., X. Chen, Z. Liu, & M. Sun. 2017. Improving Word Representations with Document Labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (4): 863-870. doi:10.1109/TASLP.2017.2658019.

Zhao, R., & K. Mao. 2018. Fuzzy Bag-of-Words Model for Document Representation. *IEEE Transactions on Fuzzy Systems* 26 (2): 794-804. doi:10.1109/TFUZZ.2017.2690222.

**سعیده انبایی فریمانی**

متولد ۱۳۶۶ است. ایشان هم‌اکنون دانشجوی دورهٔ دکتری تخصصی دانشگاه آزاد اسلامی واحد مشهد در رشتهٔ مهندسی کامپیوتر است و در حال تدوین رسالهٔ دکتری در زمینهٔ پردازش زبان طبیعی و طبقه‌بندی سری‌های زمانی با ارائهٔ مدل رفتار سرمایه‌گذاران در بازارهای مالی مبتنی بر شناسایی آنلاین رویدادهای خبری است.

**حمید طباطبایی**

متولد ۱۳۵۳ است. دارای مدرک دکتری در رشتهٔ مهندسی برق گرایش کنترل هوشمند از دانشگاه فردوسی مشهد است. ایشان هم‌اکنون استادیار دانشگاه آزاد اسلامی واحد قوچان است.

محاسبات نرم، سیستم‌های هوشمند و مدیریت دانش و سیستم‌های مبتنی بر دانش از جمله علایق پژوهشی وی است.

**مجتبی کفاشان کاخکی**

متولد ۱۳۵۷ دارای مدرک دکتری علم اطلاعات و دانش‌شناسی از دانشگاه فردوسی مشهد است. ایشان هم‌اکنون استادیار گروه علم اطلاعات و دانش‌شناسی و مدیر گروه است.

استراتژی‌های مدیریت دانش، فرایندهای مدیریت دانش، سازماندهی دانش و فلسفهٔ دانش از جمله علایق پژوهشی وی است.