

ارائه رویکردی برای مدیریت و سازمان‌دهی اسناد متنی با استفاده از تجزیه و تحلیل هوشمند متن

اعظم السادات پرنی

کارشناسی ارشد مهندسی فناوری اطلاعات
گروه فناوری اطلاعات - دانشکده مهندسی صنایع - دانشگاه صنعتی
خواجه نصیرالدین طوسی

حجت‌اله حمیدی*

دکتری مهندسی کامپیوتر - فناوری اطلاعات، استادیار گروه فناوری
اطلاعات - دانشکده مهندسی صنایع - دانشگاه
صنعتی خواجه نصیرالدین طوسی

دریافت: ۹۵/۰۴/۱۵ پذیرش: ۹۵/۰۶/۰۲ این مقاله به مدت ۱۰ روز برای ویرایش نزد نویسنده بوده است

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۲۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در SCOPUS، LISA و ISC
http://jipm.irandoc.ac.ir
دوره XX | شماره X | صص XX-XX
۱۳XX X

چکیده: با توجه به حجم عظیم داده‌های جمع‌آوری شده در سازمان‌ها و سیستم‌های نگهداشت و مدیریت اطلاعات که سبب شکل‌گیری انبار داده‌های بسیار بزرگ شده است، نیاز به استخراج الگو از متون هر روز بیشتر احساس می‌شود. متن کاوی یکی از مهم‌ترین روش‌ها در استخراج الگوی مناسب است که به وسیله اقتباس یا استخراج دانش از مجموعه‌ای از داده‌ها، به اهداف سازمان‌ها بسیار کمک می‌کند. این روش‌ها همچنین می‌توانند افق جدیدی را برای تجارت و حفاظت از مالکیت معنوی آثار نویسندگان به وجود آورند. در این مقاله با رویکرد جدیدی به کشف الگوهای متنی جهت سازمان‌دهی و تجزیه و تحلیل هوشمند متن می‌پردازیم. هدف اصلی بکارگیری از الگوی مناسب در جهت حفظ آثار نویسندگان، محققان و اسناد متنی است. با توجه به حجم آثار نویسندگان و سیستم‌های مدیریت اسناد، حجم اطلاعات در دسترس نیز به شدت افزایش یافته است. برای کشف دانش موجود در این داده‌ها که منفعت زیادی را برای کاربران اطلاعات به دنبال دارد روش‌های خاصی مورد نیاز است که در حوزه داده کاوی به آن پرداخته شده است. بخش اعظم این داده‌های در دسترس به صورت متنی و بدون ساختار یا نیمه ساختارمند می‌باشند که برای استفاده از آن‌ها می‌توان علاوه بر روش‌های مورد استفاده در داده کاوی، از فناوری‌هایی مانند پردازش زبان طبیعی، تجزیه و تحلیل

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد

کنید:

دورن متن:

(اعظم السادات پرنی و حجت‌اله حمیدی، زودآیند)

در فهرست منابع:

پرنی، اعظم السادات، حمیدی، حجت‌اله زودآیند.
ارائه رویکردی برای مدیریت و سازمان‌دهی اسناد متنی با استفاده از تجزیه و تحلیل هوشمند متن. پژوهشنامه پردازش و مدیریت اطلاعات.

http://jipm.irandoc.ac.ir (دسترسی در

۱۳۹۵/۰۶/۰۲)

هوشمند و علم آمار بهره گرفت.

کلیدواژه‌ها: کاوش متن، مدیریت متن، مالکیت معنوی، استخراج اطلاعات، داده کاوی

*h_hamidi@kntu.ac.ir

۱. مقدمه

داده کاوی^۱ به پردازش داده‌های عظیم و استخراج اطلاعات کاربردی از آن‌ها می‌پردازد (سارا کوگلو و دیگران، ۲۰۰۷). داده‌ها ممکن است در چارچوب‌های داده‌های توصیفی مانند گزارش‌های کیفیت اسناد و یادداشت‌هایی برای کمک به مشتریان باشد (کورن فین، ۲۰۰۷). لذا تلاش‌های تکنیکی و دستی برای هدایت کردن این منابع داده‌ای، به منظور کشف کردن الگوها و داده‌های سودمند، در این منابع مورد نیاز است (گلد فرب، ۲۰۰۷). از آنجایی که بالای ۸۰ درصد از این اطلاعات به صورت متنی نگه‌داری می‌شوند، باور بر این است که متن کاوی ارزش بالقوه تجاری بالایی داشته باشد (گوپتا و لهال، ۲۰۰۹). انتقال این منابع مفید اطلاعاتی به چارچوب‌های قابل استفاده باعث پیشرفت فناوریهای آینده می‌شود و به کیفیت خدمات‌رسانی کمک خواهد کرد. با گسترش روزافزون اینترنت و فناوری‌های اطلاعاتی، حجم اطلاعات در دسترس بسیار افزایش یافته است (ونگ و لین، ۲۰۰۳). از آنجایی که پردازش دستی این داده‌های متنی کاری طاقت‌فرساست، به روش‌های متن کاوی^۲ احتیاج است. (هاشمی و دیگران، ۲۰۱۵) بنابراین افراد وقت بیشتری را صرف پالایش کردن اطلاعات می‌کنند (ونگ و لین، ۲۰۰۳).. متن کاوی با اطلاعات متنی یا بدون ساختار^۳، برای استخراج اطلاعات بامعنی و دانش از مقدار زیادی اطلاعات، سروکار دارد. امروزه راه‌حل‌های متنوعی برای مدیریت و سازمان‌دهی مقدار زیادی از اسناد متنی و به دست آوردن اطلاعات مفید از این داده‌ها (متن کاوی) در حال پیشرفت و تحقیق هست. جستجوی سند مشابه به‌عنوان بخشی از متن کاوی در نظر گرفته شده و روش‌های هوش مصنوعی^۴ زیادی در این مراحل مورد استفاده قرار گرفته‌اند. از این‌رو این موضوع به‌عنوان یکی از عملیات اساسی مدیریت اسناد مشهود است (سارا کوگلو و دیگران، ۲۰۰۸). علاوه بر این، یافتن اسناد

¹ data mining

² text mining

³ unstructured

⁴ artificial intelligence

مرتبط و مشابه با اسناد موجود از حجم عظیمی از اسناد، نقش مهمی را در متن کاوی ایفا می‌کند. تمامی این مباحث ما را به سوی طراحی یک سیستم جستجوی مؤثرتر سوق می‌دهد (سارا کوگلو و دیگران، ۲۰۰۸). متن کاوی فرآیند تجزیه و تحلیل اسناد به وسیله‌ی جستجوی الگوها در متون به زبان طبیعی و به منظور استخراج اطلاعات برای اهداف خاص است (ونگ و لین، ۲۰۰۳). متن کاوی شاخه‌ای از حوزه داده کاوی است که سعی در یافتن الگوهای جالب توجه از پایگاه داده‌های بزرگ دارد که کشف اطلاعات جدید یا استخراج خودکار اطلاعات از منابع مکتوب مختلف، به وسیله رایانه امکان پذیر می‌شود. متن کاوی که تحت عناوین تجزیه و تحلیل هوشمند متن، کاوش داده‌های متنی و کشف دانش از متون نیز شناخته می‌شود، به طور کلی به فرآیند استخراج اطلاعات و دانش جالب توجه و غیر بدیهی از متن بدون ساختار اشاره دارد (گوپتا و لهال، ۲۰۰۹).

این مقاله به صورت زیر سازمان‌دهی شده است: در بخش دوم به مروری درباره ادبیات موضوع در زمینه داده کاوی در اسناد متنی می‌پردازیم در بخش سوم ارتباط متن کاوی با حوزه‌های دیگر نظیر کشف دانش در پایگاه داده و وب کاوی را بیان می‌کنیم، در بخش چهارم به تعریف دقیق و مزایای متن کاوی اشاره می‌کنیم. در بخش پنجم به مراحل اصلی متن کاوی که به ترتیب شامل: ۱- پیش پردازش که خود از بخش‌های مدل فضای برداری و پیش پردازش زبان‌شناختی تشکیل شده است و ۲- تولید و استخراج ویژگی ۳- انتخاب ویژگی می‌باشد، پرداخته می‌شود. در بخش ششم به فازهای اصلی فرایند متن کاوی می‌پردازیم. در بخش هفتم مدل پیشنهادی توضیح داده می‌شود. در بخش هشتم یافته‌های مقاله بررسی می‌گردد و در بخش نهم نتیجه گیری ارائه می‌شود.

۲. مروری بر پیشینه پژوهش

موضوع اصلی در کاوش متن، یافتن اسناد مرتبط از میان حجم عظیمی از اسناد است (سارا کوگلو و دیگران، ۲۰۰۸). دو رویکرد پایه برای جستجوی اسناد مشابه وجود دارد. رویکرد ابتدایی استخراج کلیدواژه‌ها از سند (رویکرد مبتنی بر کلیدواژه) (ونگ و لین ۲۰۰۳؛ ونگ و لو، ۲۰۰۴) و دیگری رویکرد مبتنی بر استفاده از تمامی کلمات سند است. فرآیند کلی کشف دانش در پایگاه داده‌ها سبب طبقه‌بندی معتبر، به روز و قابل استفاده می‌شود و در نهایت الگوهای قابل فهم داده‌ای ارائه می‌دهد. (فیاد، ۱۹۹۶) در متن کاوی به طور کلی، استخراج ویژگی تأیید شده است (گونال و دیگران، ۲۰۰۶). انتخاب ویژگی (فنگ گو جینگ و دیگران،

۲۰۱۳) روش طبقه‌بندی (تنگ و نگ و ۲۰۱۱) و پیش‌پردازش (گونال و دیگران، ۲۰۱۴) تأثیر قابل توجهی بر موفقیت فرآیندهای طبقه‌بندی متن داشته است. عملکرد طبقه‌بندی متن از طریق یک اصطلاح‌نامه مبتنی بر پیکره و وردنت، با به‌کارگیری الگوریتم نزدیک‌ترین همسایگی و شبکه عصبی پس‌انتشار (BPNN) الگوریتم را بهبود دادند (جیانگ پنگ و دیگران ۲۰۱۲). الگوریتم k-NN برای طبقه‌بندی متن با ترکیب آن با یک الگوریتم خوشه‌بندی مورد استفاده قرار می‌گیرد. بیزین ساده، الگوریتمی دیگر است که برای طبقه‌بندی متن موفق بوده است. (چنگ هوان تیان، ۲۰۰۹) برای مثال مراحل فرایند متن‌کاوی در اخبار و اطلاعات در جدول ۱ نشان داده شده است. در جدول زیر، نوآوری مورد نیاز در جنبه‌های مهم انتخاب ویژگی، کاهش ویژگی و بازنمایی ویژگی بسیاری از آثار بررسی شده است.

جدول ۱. مقایسه داده‌های ورودی برای سیستم‌های مختلف

نویسنده	انواع متون	منبع متن	تعداد مورد بررسی	بدون ساختار	پیش‌برنامه‌ریزی شده
Tetlock et al. (2008)	اخبار مالی	Wall Street Journal, Dow Jones News Service از بانک اطلاعاتی اخبار در Factiva	۳۵۰,۰۰۰ متن	بله	خیر
Mahajan et al. (2008)	اخبار مالی	اشاره نشده	۷۰۰ مقاله خبری	بله	خیر
Butler and Keşelj (2009)	گزارش‌ها سالیانه	وبسایت شرکت	اشاره نشده	بله	بله
Schumaker and Chen (2009)	اخبار مالی	Yahoo Finance	۲۸۰۰	بله	گزارش‌های سالیانه شرکت
Li (2010)	بایگانی‌های شرکت	بحث و تجزیه و تحلیل بخش مدیریت 10-K and 10-Q بایگانی وبسایت SEC Edgar	۱۳ میلیون باینیه	بله	خیر
Huang, Liao, Yang, Chang, and Luo (2010) and Huang, Chuang, et al. (2010)	اخبار مالی	روزنامه الکترونیکی در تایوان	۱۲,۸۳۰ سرفصل	بله	خیر
Bollen and Huina (2011)	تویت‌ها	تویتر	۹,۸۵۳,۴۹۸	بله	خیر
Vu et al. (2012)	تویت‌ها	تویتر	۵,۰۰۱,۴۶۰	بله	خیر
Schumaker et al.	اخبار مالی	Yahoo Finance	۲۸۰۲	بله	خیر

					(2012)
خبر	بله	اشاره نشده	کارگزاری‌ها	اخبار کارگزاری	Lugmayr and Gossen(2012)
خبر	بله	۵۲,۷۴۶ پیام	ویلاک‌ها، فروم‌ها، اخبار و شبکه‌های اجتماعی	رسانه‌های اجتماعی متعارف	Yu, Duan, et al. (2013)
خبر	بله	۱۰۸۷۰ و ۳۴۷۸ به ترتیب	DGAP, EuroAdhoc	اطلاعه‌های شرکت و اخبار مالی	Hagenau et al. (2013)
خبر	بله	۳۶۱,۷۸۲	Bloomberg	اخبار عمومی	Jin et al. (2013)
بله	خبر	اشاره نشده	Bloomberg	اخبار اقتصاد کلان	Chatrath et al. (2014)

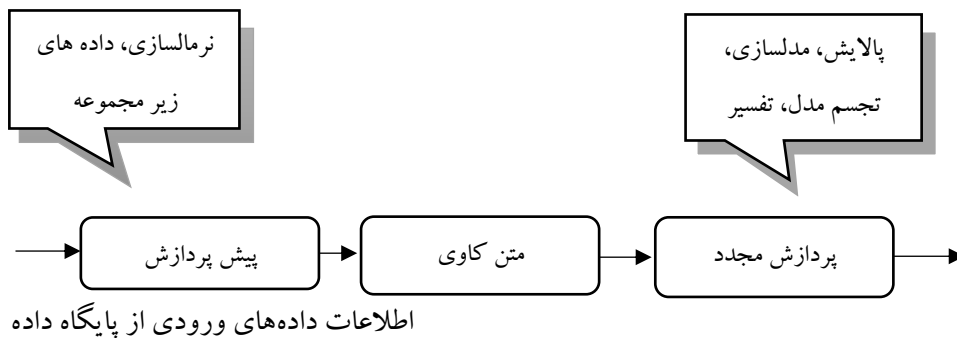
بنابراین کشف زمینه‌های متفاوتی از نظر انواع بازار و انواع متن ورودی برای سیستم‌ها، بسیار حائز اهمیت است.

۳. ارتباط متن کاوی و دیگر دانش‌ها

۳-۱ کشف دانش و ارتباط آن با متن کاوی

کشف دانش در پایگاه داده‌ها (KDD) در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است و هدف آن کشف ارتباط و نظم بین اطلاعات قابل مشاهده است. گاهی داده کاوی شامل همه ابعاد فرآیند کشف دانش است و در بعضی موارد نیز داده کاوی به عنوان بخشی از فرآیندهای KDD در نظر گرفته می‌شود و فاز مدل کردن را توصیف می‌کند. در نهایت KDD فرآیند یافتن اطلاعات و الگوهای مفید از داده را گویند و داده کاوی بهره‌گیری از الگوریتم‌هایی برای یافتن اطلاعات مفید و الگوها در فرآیند KDD است. تعاریف مختلفی برای کشف دانش یا کشف دانش در پایگاه داده وجود دارد. مثلاً بیان شده است که KDD فرآیند شناسایی الگوهای قابل فهم، مفید، جدید و معتبر در داده است. در واقع هدف پیدا کردن الگوها و روابط پنهان در این داده‌هاست. از جمله خصوصیتی که برای اندازه‌گیری کیفیت الگوهای پیدا شده در داده می‌توان استفاده کرد عبارت‌اند از: قابلیت فهم انسان، اعتبارسنجی با معیارهای آماری، تازگی و مفید بودن. کشف دانش در پایگاه داده را می‌توان به عنوان یک فرایند که به وسیله چندین گام پردازش کردن تعریف می‌شود، در نظر گرفت. این گام‌ها به منظور استخراج الگوهای مفید باید بر روی مجموعه داده‌ای اعمال شوند. طبق نظرات قبلی گام‌ها را می‌توان

به صورت زیر بیان کرد: (۱) درک کردن کسب و کار (۲) درک کردن داده (۳) آماده سازی داده (۴) مدل کردن (۵) ارزیابی (۶) گسترش. مرحله پیش پردازش غالباً یکی از مراحل زمان بر و درعین حال بسیار مهم در کسب نتیجه مطلوب است. مخصوصاً در متن کاوی که نیاز به متدهای پیش پردازش کردن خاصی برای تبدیل داده متنی به فرمتی که برای الگوریتم های داده کاوی مناسب است، داریم.



شکل ۱. ارتباط کشف دانش از پایگاه داده و متن کاوی

۲-۳ تفاوت متن کاوی و داده کاوی

با توجه به رشد روزافزون و توسعه فناوری، اطلاعات به صورت گسترده و با سرعت بالا منتشر می شود در نتیجه یکی از زمینه های تحقیقاتی مهم، ذخیره داده ها و به دست آوردن اطلاعات سودمند از آنهاست. داده کاوی برای این منظور به وجود آمد و می توان گفت که مقدار زیادی از دانش را پردازش کرده و آنها را به اطلاعات کاربردی تبدیل می کند و در صورتی که داده های مورد پردازش به صورت متن و بدون ساختار یا نیمه ساختارمند باشند، این فرآیند متن کاوی نامیده می شود (ونگ و لین، ۲۰۰۳؛ سارا کوگلو و دیگران، ۲۰۰۸). متن کاوی (بری، ۲۰۰۴) مشابه داده کاوی است، با این تفاوت که ابزارهای داده کاوی (شمکنت و المسری، ۲۰۰۰) برای مدیریت داده های ساختارمند از پایگاه داده ها طراحی شده اند. متن کاوی می تواند با مجموعه داده های بدون ساختار یا نیمه ساختارمند مانند نامه های الکترونیکی، اسناد تمام متنی و پرونده های وب کار کند. در نتیجه، متن کاوی راه حل بهتری برای شرکت هاست. با این حال تا به امروز، بیشتر تلاش تحقیق و توسعه روی

داده‌کاوی با استفاده از داده‌های ساختارمند متمرکز بوده‌اند (ونگ و لین، ۲۰۰۳؛ ساراگوگلو و دیگران، ۲۰۰۸؛ گوپتا و لهال، ۲۰۰۹).

جدول ۲. ارتباط داده‌کاوی و متن‌کاوی

	جستجوی هدف مشخص	کشف سازوکار مشخص
داده‌های ساخت یافته	بازیابی داده‌ها	داده‌کاوی
داده‌های بدون ساختار-متنی	بازیابی اطلاعات	متن‌کاوی

۳-۳ تفاوت متن‌کاوی و وب‌کاوی

بر این اساس متن‌کاوی با جستجوی در وب نیز متفاوت است. در وب‌کاوی، کاربر به‌طور معمول به دنبال چیزی است که در حال حاضر شناخته شده است و توسط شخصی دیگر نوشته شده است و مسئله صرفاً کنار گذاشتن تمامی موارد نامربوط به نیازها و یافتن اطلاعات مرتبط است. در متن‌کاوی، هدف کشف اطلاعات ناشناخته است، یعنی چیزی که هنوز کسی آن را نمی‌داند و نمی‌تواند نوشته شده باشد (گوپتا و لهال، ۲۰۰۹).

۴. متن‌کاوی و جریان کار در متن‌کاوی

متن‌کاوی یا کشف دانش از متن برای اولین بار برای استخراج الگوهای متنی بیان شد. می‌توان گفت که متن‌کاوی از تکنیک‌های بازیابی اطلاعات، استخراج اطلاعات همچنین پردازش کردن زبان طبیعی استفاده می‌کند و آن‌ها را به الگوریتم‌ها و متدهای KDD، داده‌کاوی، یادگیری ماشین و آماری مرتبط می‌کنند. با توجه به ناحیه‌های تحقیق گوناگون، بر هر یک از آن‌ها می‌توان تعاریف مختلفی از متن‌کاوی در نظر گرفت در ادامه برخی از این تعاریف بیان می‌شوند:

متن‌کاوی به معنای استخراج اطلاعات: در این تعریف متن‌کاوی متناظر با استخراج اطلاعات در نظر گرفته می‌شود (استخراج واقعیت‌ها از متن).

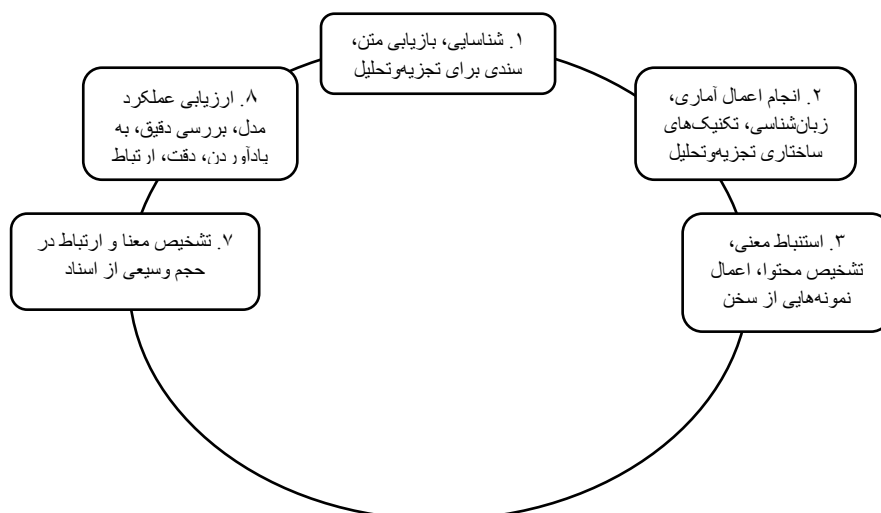
متن کاوی به معنای کشف داده متنی: متن کاوی را می‌توان به‌عنوان متدها و الگوریتم‌هایی از فیلدهای یادگیری ماشین و آماری برای متن‌ها باهدف پیدا کردن الگوهای مفید در نظر گرفت. برای این هدف پیش‌پردازش کردن متون ضروری است. در بسیاری از روش‌ها، متدهای استخراج اطلاعات، پردازش کردن زبان طبیعی یا برخی پیش‌پردازش‌های ساده برای استخراج داده از متون استفاده می‌شود. سپس می‌توان الگوریتم‌های داده‌کاوی را بر روی داده‌های استخراج‌شده اعمال کرد.

۱-۴ مزایای متن کاوی

از مزایای متن کاوی می‌توان به مواردی مانند کمک به استخراج اطلاعات مفید از انبوهی از داده‌ها در زمان کم و به‌طور کارآمد و کمک در پیش‌بینی جنبه‌های آینده بر اساس مشاهدات و آمار ارائه‌شده اشاره نمود. هاشمی و دیگران (۲۰۱۵) برای متن کاوی مزایای زیر را برشمرده‌اند:

- کمک در پیش‌بینی جنبه‌های آینده بر اساس مشاهدات و آمار ارائه‌شده
- کمک به ایجاد و ساخت الگو از داده‌های ارائه‌شده که به ما درباره افزایش یا کاهش گرایش‌ها مثلاً در تجارت یا اقتصاد می‌گوید.
- کمک به نهادهای امنیتی با نظارت و تجزیه و تحلیل داده‌های متنی جمع‌آوری‌شده از منابع اینترنتی مانند وبلاگ‌ها
- بهبود کاربرد علم پزشکی از پایگاه داده‌های پزشکی موجود با استفاده از روش‌های متن کاوی
- پیشرفته‌تر کردن تجزیه و تحلیل، ذخیره‌سازی و دسترسی به اطلاعات در وب‌سایت‌ها و موتورهای جستجوی مختلف به‌منظور کارآمدتر و دقیق‌تر کردن فرآیند جستجو
- کمک به مطالعه توزیع فراوانی کلمه با استفاده از تجزیه و تحلیل لغوی و تشخیص الگو و نویسنده اثر.

کلیه مراحل بیان‌شده جریان کار عمومی در متن کاوی است که در بلوک دیاگرام شکل ۲ نشان داده شده است.



۹. یافته‌های آماده
برای کاربر نهایی

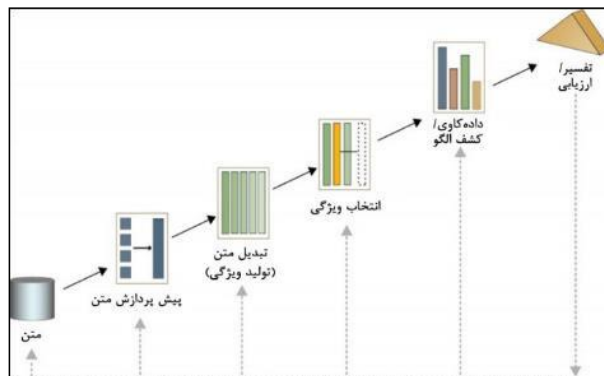
۶. طبقه‌بندی، دسته‌بندی
سندها و سازمان‌دهی در
سازمان اطلاعات

۲. جریان کار عمومی در متن

۴. استخراج محتوا
و الگوها

۵. مراحل متن کاوی

فرآیند متن کاوی شامل مراحل اصلی ۵. اعمال آماری، فراگیری ماشین، تکنیک تطبیق الگو و ویژگی و انتخاب ویژگی می‌باشد. مراحل کلی فرآیند متن کاوی در شکل ۳ (هاسمی و دیگران، ۲۰۱۵) نشان داده شده است. در ابتدا فرآیند پیش‌پردازش روی متن مورد نظر انجام می‌شود و پس از آن عملیات استخراج ویژگی و انتخاب ویژگی‌های روی آن انجام می‌شود. این داده‌ها سپس با استفاده از الگوریتم‌های مخصوص، الگوهای موجود شناسایی شده و جهت ارزیابی^۱ یا تفسیر^۲ به خروجی فرستاده می‌شوند. توضیحات بیشتر در ادامه آمده است.



شکل ۳. مراحل کلی فرآیند متن کاوی

۱-۵ پیش‌پردازش

برای کاوش کردن حجم قابل توجهی از اسناد ضروری است که اسناد پیش‌پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره شوند. زمانی که داده‌های ورودی در دسترس هستند باید آن‌ها را برای ورود به الگوریتم‌های یادگیری ماشین آماده کنیم. این مرحله برای داده‌های متنی به این معنا است که آن‌ها را از حالت غیر ساخت یافته به فرمت

¹ evaluation

² interpretation

ساختاریافته و قابل تشخیص برای ماشین تبدیل کنیم هدف پیش پردازش کاهش ابعاد فضای نمایش لغات موجود در سند بوده و معمولاً شامل این وظایف می‌باشد (گوپتا و لهال، ۲۰۰۹):

- حذف ایست واژه‌ها (کلماتی که در اکثر متون یافت می‌شوند اما دارای اهمیت معنایی خاصی نبوده و جزو کلیدواژه‌ها به حساب نمی‌آیند).
 - یکپارچه‌سازی تمامی حروف به یک نوع (از نظر بزرگی و کوچکی حروف)
 - ریشه‌یابی لغات: یکسان در نظر گرفتن کلمات مشابه از نظر نحوی مانند حالت جمع، تغییرات کلامی و ... به منظور به دست آوردن ریشه هر کلمه که بر معنی آن تأکید می‌کند.
- برای تعریف رسمی تر الگوریتم‌ها برخی تعاریف لازم را بیان می‌کنیم: مجموعه اسناد با D نمایش داده می‌شود. T نشان‌دهنده دیکشنری است $T = \{t_1, t_2 \dots t_m\}$. فرکانس ترم $t \in T$ در سند $d \in D$ به صورت $tf(d, t)$ نمایش داده می‌شود. بردار ترم t_d به صورت $t_d = (tf(d, t_1) \dots, tf(d, t_m))$ نمایش داده می‌شود. مرکز مجموعه X از بردارهای ترم به صورت مقدار میانگین بردارهای ترمش تعریف می‌شود:

$$(t_x := \frac{1}{|X|} \sum_{t \in X} t_d) \quad (1)$$

همچنین می‌توان tf را بر روی زیرمجموعه‌ای از ترم‌ها اعمال کرد. اگر T' زیرمجموعه T باشد دارای

$$tf(d, T') := \sum_{t \in T'} tf(d, t) \quad (2)$$

داشتن تعداد محدودی از ویژگی‌ها، مسئله بسیار بااهمیتی است. افزایش تعداد ویژگی‌ها که ممکن است به سادگی در طول مرحله انتخاب ویژگی اتفاق بیفتد می‌تواند باعث تبدیل شدن مسئله متن کاوی به یک مسئله غیرقابل حل برای تمام الگوریتم‌های یادگیری ماشین بشود (پستو، ۲۰۱۳). برای کاهش دادن سائز دیکشنری و ابعاد توصیف اسناد یک مجموعه می‌توان از روش‌های فیلترینگ و ریشه‌یابی استفاده کرد. روش‌های فیلترینگ، کلمات را از دیکشنری و بنابراین از اسناد حذف می‌کنند. از جمله این روش‌ها می‌توان به $stop\ word\ filtering$ و حذف $stop\ word$ ها اشاره نمود. برای کاهش دادن تعداد کلمات دیکشنری می‌توان از روش‌های انتخاب ترم نیز استفاده

نمود. در روش‌های انتخاب‌ترم تنها کلمات انتخاب‌شده برای توصیف کردن اسناد استفاده می‌شوند. یک روش برای انتخاب کلمات، استخراج کلمات بر اساس درجه خلوص¹ آن‌هاست. برای هر کلمه t در دیکشنری درجه خلوص آن به صورت زیر تعریف می‌شود:

$$W(t) = 1 + \frac{1}{\log_2 |D|} \sum_{d \in D} P(d, t) \log_2 P(d, t), P(d, t) = \frac{tf(d, t)}{\sum_{l=1}^n tf(d_l, t)} \quad (3)$$

درجه خلوص نشان‌دهنده این است که یک کلمه چقدر اسناد را به خوبی، توسط کلمه جستجو جدا می‌کند. اگر کلمه در تعداد زیادی از اسناد رخ داده باشد در نتیجه درجه خلوص کمی دارد. روش‌های ریشه‌یابی سعی می‌کنند فرم‌های اولیه کلمات را بسازند. مثلاً حذف کردن علامت جمع (s) از اسم‌ها یا ing از افعال و... بعد از فرآیند ریشه‌یابی هر کلمه توسط ریشه‌اش نمایش داده می‌شود. در واقع در اینجا کلماتی که ریشه مشترک دارند به ریشه خودشان تبدیل می‌شوند.

جدول ۳. پیش‌پردازش: انتخاب ویژگی، کاهش ابعاد و بازنمایش ویژگی

نویسنده	انتخاب ویژگی	کاهش ابعاد	بازنمایش ویژگی
Tetlock et al. (2008)	برای لغات منفی Bag-of-words	لغت‌نامه از پیش تعریف‌شده، فرهنگ لغت روانشناسی	فراوانی تقسیم‌بر تعداد کل کلمات
Mahajan et al. (2008)	Latent Dirichlet Allocation (LDA)	استخراج بیست‌وپنج مبحث	Binary
Butler and Kešelj (2009)	Character n-Grams, three readability scores عملکرد سال گذشته	حداقل اتفاق در هر سند	فراوانی n-gram در یک پروفایل
Schumaker and Chen (2009)	Bag of words, noun phrases, named entities	حداقل اتفاق در هر سند	Binary
Li (2010)	لحن و محتوایی Bag of words	لغت‌نامه از پیش تعریف‌شده	Binary, Dictionary value
Huang, Liao, Yang, Chang, and Luo (2010) and Huang, Chuang, et al. (2010)	واژگان همزمان، دستور جفت	جایگزینی مترادف‌ها	شاخص وزن‌دار
Groth and Muntermann (2011)	Bag-of-words	روش نمره دهی ویژگی با استفاده از Information Gain	TF-IDF

¹ entropy

	و معیار مجذور کای		
By OpinionFinder	By OpinionFinder	By OpinionFinder	Bollen and Huina (2011)
Neg_Pos and Bullish_Bearish	شناسایی نام‌های موجود	تعداد کل جملات مثبت یا منفی توینتر	Vu et al. (2012)
Binary	حداقل اتفاق در هر سند	OpinionFinder overall tone and polarity	Schumaker et al. (2012)
مقادیر احساسات	ریشه‌یابی	Bag-of-words	Lugmayr and Gossen(2012)
Binary Hagenau	اشاره نشده	Bag-of-words	Yu, Duan, et al. (2013)
TF-IDF	Frequency for news, Chi 2 – approach and bi-normal separation (BNS) for exogenous-feedback-based feature selection, dictionary	Bag-of-words, noun phrases, wordcombinations, n-grams	Hagenau et al. (2013)
توزیع هر مقاله	استخراج موضوع	Latent Dirichlet Allocation (LDA)	Jin et al. (2013)
داده‌های ساخت یافته	داده‌های ساخت یافته	داده‌های ساخت یافته	Chatrath et al. (2014)

بعد از انتخاب حداقل تعداد ویژگی‌ها، لازم است آن‌ها را با مقادیر عددی نمایش دهیم تا بتوان از آن‌ها در الگوریتم‌های یادگیری ماشین استفاده کرد. به همین منظور در ستون مربوط به بازنمایش ویژگی در جدول ۳ روش این کار را برای تمام مقالات بررسی شده مرور کرده‌ایم. مقادیر عددی اختصاص داده شده، مانند نمره یا امتیاز عمل می‌کنند. ۵ روش بسیار متداول آن عبارت‌اند از: اطلاعات به دست آمده^۱ (IG)، آماره مربع کای^۲ (CHI)، نوسانات سند^۳ (DF)، دقت متوازن^۴ (Acc2) نوسان معکوس فراوانی سند^۵ (TF-IDF). مقایسه‌ای میان این روش‌ها را می‌توان در پژوهش تسای و گنگور (۲۰۱۳) مشاهده کرد. پایه‌ای‌ترین آن‌ها روش باینری است که در آن با اختصاص دو مقدار ۱ و ۰ با حضور یا عدم حضور یک ویژگی که ممکن است یک لغت بررسی می‌شود (شوماخر و دیگران، ۲۰۱۲). روش رایج بعدی اصطلاحاً TF-IDF نام دارد (هگنا و دیگران، ۲۰۱۳) ارزش TF-IDF متناسب با تعداد دفعاتی که یک کلمه در اسناد ظاهر می‌شود افزایش می‌یابد، اما نسبت به نوسانات واژه در مجموعه سند خنثی است تا بتواند تعادل را برای

¹ Information Gain

² Chi-square Statistics

³ Document Frequency

⁴ Accuracy Balanced

⁵ Term Frequency-Inverse Document Frequency

کلمات برقرار کند. روش‌های رایج و مشابه دیگری نیز وجود دارد، به‌عنوان مثال TF-CDF که از روش CF توسعه یافته است ادعا می‌کند که می‌تواند کارایی بیشتری نسبت به TF-IDF در عمل داشته باشد.

۵-۱-۱ پیش‌پردازش زبان‌شناختی

گاهی اوقات ممکنه از پیش‌پردازش‌های زبان‌شناختی نیز برای افزایش دادن اطلاعات قابل دسترس درباره ترم‌ها استفاده شود. برای این منظور روش‌های زیر اعمال می‌شوند:

برچسب‌گذاری نحوی: برای هر ترم مشخص می‌کند که اسم، فعل، صفت و... است. اگرچه تعداد زیادی معتقد به این نیستند که این کار جزئی از متن‌کاوی است ولی در (ویتن، ۲۰۰۴)، برای مثال سیستمی به نام GATE در دانشگاه شفیلد، در یک کتابخانه‌ی دیجیتال به این قصد جاگذاری شده است. GATE شامل ابزاری برای برچسب زدن بر جملات است. برای مثال این سیستم می‌تواند در داخل یک متن، نام موقعیت‌های جغرافیایی، نام اشخاص و چیزهایی شبیه به این را بیابد. این سیستم بیشتر شامل استخراج اطلاعات است تا استخراج دانش. POS اغلب نقش بزرگی را در پردازش زبان‌های طبیعی بازی می‌کند. در حقیقت این اولین قدم در پردازش زبان طبیعی است و پردازش زبان طبیعی یکی از پایه‌های متن‌کاوی است.

چانک کردن متن^۱: هدف گروه‌بندی کردن کلمات مجاور در یک جمله است.

رفع ابهام معنی کلمه^۲ (WSD): در اینجا سعی می‌شود ابهامات در معنی یک کلمه یا عبارت رفع شود. مثل کلماتی که چندین معنی متفاوت دارند. بنابراین به‌جای ترم‌ها معناها می‌توانند در نمایش فضای بردار ذخیره شوند. این باعث بزرگ‌تر شدن دیکشنری می‌شود؛ اما همچنین موجب می‌شود که معنای یک ترم در نمایش در نظر گرفته شود.

۵-۲ تولید و استخراج ویژگی

اگرچه برنامه‌های کاربردی زیادی در زمینه بازیابی اطلاعات مانند پالایش و جستجوی اطلاعات مرتبط می‌توانند از تحقیقات در زمینه‌ی رده‌بندی متن سود ببرند، مشکل اصلی رده‌بندی متن، ابعاد

¹ Text chunking

² Word Sense Disambiguation

بالای فضای ویژگی با توجه به تعداد زیاد لغات است. این مشکل ممکن است موجب افزایش پیچیدگی محاسباتی روش‌های یادگیری ماشین برای رده‌بندی متن شده و با توجه به کلمات نامرتب یا زائد موجب ناکارآمدی نتایج حاصل از دقت کم بشود. به‌عنوان راه‌حلی برای این مشکل از روش‌های استخراج ویژگی و انتخاب ویژگی استفاده می‌شود.

۳-۵ انتخاب ویژگی

انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌های اصلی را با توجه به برخی از معیارها یا اهمیت ویژگی انتخاب می‌کند. الگوریتم‌های انتخاب ویژگی به دودسته تقسیم می‌شوند:

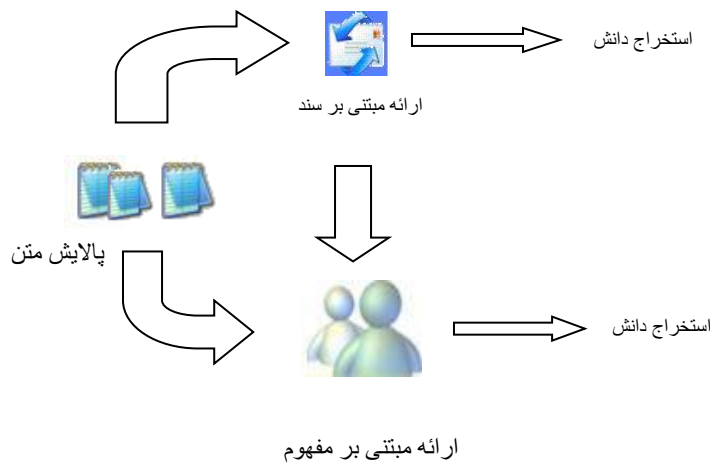
الگوریتم‌های رتبه‌بندی ویژگی: این الگوریتم‌ها ویژگی‌هایی را که به‌تنهایی برای طبقه‌بندی بسیار مرتبط هستند، بدون در نظر گرفتن تعامل ویژگی‌ها انتخاب می‌کنند. این موضوع، الگوریتم‌های FR را ساده کرده و از نظر محاسباتی کارآمدتر از الگوریتم‌های انتخاب زیرمجموعه ویژگی می‌سازد و بنابراین یک انتخاب ارجح برای رده‌بندی متن می‌باشد. تفکیک دوگانه‌ی نرمال و IG نمونه‌هایی از الگوریتم‌های FR هستند. با این حال الگوریتم‌های FR مجموعه‌ای از ویژگی‌ها یا کلمات بسیار مرتبط را در خروجی می‌دهند که احتمالاً می‌تواند زائد بوده و بدین ترتیب منجر به کاهش عملکرد یک طبقه‌بند شود.

الگوریتم‌های انتخاب زیرمجموعه ویژگی (FSS): این الگوریتم‌ها تعامل بین ویژگی‌ها را در نظر می‌گیرند و درعین حال از نظر محاسباتی گران می‌باشند (جاوید و دیگران، ۲۰۱۵). یک عنصر کلیدی فرآیند متن کاوی، ارتباط اطلاعات استخراج‌شده به‌همدیگر به‌منظور تشکیل حقایق یا فرضیات جدید برای استفاده‌های بعدی توسط ابزارهای متداول است.

۶. فازهای اصلی فرایند متن کاوی

اولین فاز پیش‌پردازش مستندات است. خروجی این فاز می‌تواند دو شکل مختلف داشته باشد: (۱) مبتنی بر سند (۲) مبتنی بر مفهوم. در فرمت نمایش مبتنی بر سند، آنچه مهم است، نحوه‌ی نمایش بهتر مستندات است. مثلاً تبدیل اسناد به یک فرمت میانی و نیمه ساخت‌یافته، یا بکار بردن یک ایندکس بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را کاراتر می‌کند. در نوع دوم نمایش اسناد بهبود بخشیده می‌شود، مفاهیم و معانی موجود در سند و نیز ارتباط میان آن‌ها و هر نوع اطلاعات مفهومی دیگری که قابل استخراج است، از متن استخراج می‌شود. در این نوع نمایش دیگر با مستندات به‌عنوان یک موجودیت مواجه نیستیم بلکه با مفاهیمی که از این

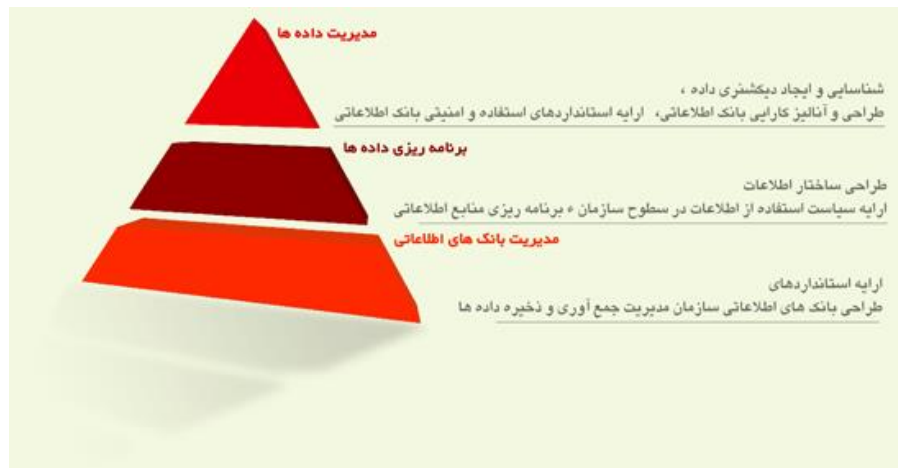
مستندات استخراج شده‌اند، روبه‌رو هستیم. قدم بعدی استخراج دانش از این فرم‌های میانی نمایش اسناد است. بر اساس نحوه‌ی نمایش یک سند، روش استخراج دانش از یک سند متفاوت است. نمایش مبتنی بر سند، برای گروه‌بندی، طبقه‌بندی، تجسم سازی و نظایر این‌ها استفاده می‌شود، درحالی‌که نمایش مبتنی بر مفهوم برای یافتن روابط میان مفاهیم، ساختن اتوماتیک تزاروس و آنتولوژی و نظایر آن بکار می‌رود که در شکل ۴ نمایش داده شده است.



شکل ۴. فرایند متن کاوی

اولین فاز پیش‌پردازش مستندات است. خروجی این فاز می‌تواند دو شکل مختلف داشته باشد: (۱) مبتنی بر سند (۲) مبتنی بر مفهوم. در فرمت نمایش مبتنی بر سند، آنچه مهم است، نحوه‌ی نمایش بهتر برای مستندات است. مثلاً تبدیل اسناد به یک فرمت میانی و نیمه ساخت یافته، یا بکار بردن یک ایندکس بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را کاراتر می‌کند. هر موجودیت در این نمایش در نهایت باز هم یک سند خواهد بود. در نوع دوم نمایش اسناد بهبود بخشیده می‌شود، مفاهیم و معانی موجود در سند و نیز ارتباط میان آن‌ها و هر نوع اطلاعات مفهومی دیگری که قابل استخراج است، از متن استخراج می‌شود. در این نوع نمایش دیگر با مستندات به‌عنوان یک موجودیت مواجه

نیستیم بلکه با مفاهیمی که از این مستندات استخراج شده‌اند، روبه‌رو هستیم. قدم بعدی استخراج دانش از این فرم‌های میانی نمایش اسناد است. بر اساس نحوه نمایش یک سند، روش استخراج دانش از یک سند متفاوت است. نمایش مبتنی بر سند، برای گروه‌بندی، طبقه‌بندی، تجسم‌سازی و نظایر این‌ها استفاده می‌شود، درحالی‌که نمایش مبتنی بر مفهوم برای یافتن روابط میان مفاهیم، ساختن اتوماتیک تزاروس و آنتولوژی و نظایر آن بکار می‌رود.



شکل ۵. مدیریت مناسب داده‌ها (ابراهیم هژیر، ۱۳۹۳)

برای رسیدن به یک سیستم اطلاعاتی مناسب، داده‌ها می‌بایست به صورتی منطقی طبقه‌بندی و ذخیره شوند تا استفاده از آن‌ها ساده‌تر بوده، با کارایی بیشتری تحلیل شوند و سریع‌تر مورد استفاده قرار گیرند و در نتیجه مدیریت بهتری بر آن‌ها اعمال شود.

۶-۱ استخراج اطلاعات

نقطه شروع برای رایانه‌ها به منظور تجزیه و تحلیل متن بدون ساختار، استفاده از استخراج اطلاعات است. نرم‌افزار استخراج اطلاعات، عبارات کلیدی و روابط درون‌متن را شناسایی می‌کند که این کار را با جستجوی دنباله‌های از پیش تعریف شده در متن طی یک فرآیند که تطبیق الگو نام دارد، انجام می‌دهد. این نرم‌افزار به روابط بین تمام افراد، مکان‌ها و زمان‌های شناسایی شده برای ارائه به

کاربر با اطلاعات بامعنی پی می‌برد. این فناوری می‌تواند در رابطه با متون حجیم بسیار کاربردی باشد. داده‌کاوی سنتی فرض می‌کند که داده‌های در حال کاوش در قالب پایگاه داده‌های رابطه‌ای هستند. درحالی‌که این‌طور نبوده و با کمک این روش، اطلاعاتی که هیچ‌گونه ساختاری ندارند، تبدیل به اطلاعاتی می‌شوند که دارای ساختار بوده و می‌توانند به‌صورت بهتری مورد استفاده واحد کشف دانش قرار بگیرند (گوپتا و لهال، ۲۰۰۹).

۱-۱-۶ رده‌بندی برای استخراج اطلاعات

استخراج اشیاء براساس درک کردن پیغام فرموله بندی می‌شدند. کلماتی که شی با آن‌ها آغاز می‌شود برچسب B، کلماتی که داخل شی می‌آیند برچسب I و کلمات خارج از شی برچسب O می‌گیرند. برای مثال فرض کنید یکی از اطلاعات استخراج شده از اسناد، نام افراد جدیدالورود به شرکت باشد که برابر با (John J. Donner, Jr.) است در این صورت داریم:

"by (O) John (B) J. (I) Donner (I) Jr. (I) the (O)"

(۴)

در اینجا مشکل رده‌بندی ترتیبی برای برچسب‌های هر کلمه، با کلمات مجاور به‌عنوان بردار ویژگی ورودی وجود دارد. یک روش متداول برای نمایش بردار ویژگی، نمایش آن به‌صورت باینری است. هر کامپوننت ویژگی می‌تواند به‌عنوان یک آزمون در نظر گرفته شود که اثبات می‌کند آیا یک الگوی خاص در یک موقعیت خاص رخ می‌دهد یا نه. برای مثال یک کامپوننت ویژگی اگر کلمه قبلی، کلمه "John" است، مقدار ۱ می‌گیرد و در غیر این صورت مقدار ۰ می‌گیرد. علاوه بر تست کردن وجود کلمات خاص می‌توان این‌که آیا کلمات یا حروف بزرگ شروع می‌شوند، یک پسوند خاصی دارند یا نه را نیز تست کرد. هم‌اکنون می‌توان هر متد رده‌بندی کارایی را برای رده‌بندی کردن برچسب‌های کلمات با استفاده از بردار ویژگی ورودی به‌کاربرد.

۲-۶ ردیابی موضوع

این روش با استفاده از نگهداری پروفایل کاربر و بر اساس موضوعات موردعلاقه انتخاب شده توسط کاربر یا پیگیری اسنادی که کاربر مشاهده می‌کند، سایر اسناد موردعلاقه‌ی وی را پیش‌بینی می‌کند. یک نمونه از این روش توسط وب‌سایت یا هو مورد استفاده قرار گرفته است. این فناوری می‌تواند کاربردهای گوناگونی داشته باشد. به‌عنوان مثال می‌تواند یک شرکت را از وجود اطلاعاتی در رابطه با رقبای در اخبار آگاه نماید. همچنین می‌تواند مورد استفاده پزشکانی قرار گیرد

که در جستجوی درمان جدید برای بیماری‌ها هستند یا می‌خواهند در جریان آخرین پیشرفت‌ها قرار بگیرند. همچنین در حوزه آموزش می‌توان از فناوری‌های رایجی موضوع برای اطمینان از در اختیار داشتن جدیدترین منابع تحقیقاتی موردعلاقه استفاده نمود (گوپتا و لهال، ۲۰۰۹).

۳-۶ خلاصه‌سازی

خلاصه‌سازی متن یعنی دریافت یک متن و تولید یا استخراج یک متن دیگر از متن اصلی، به گونه‌ای که متن به دست آمده از متن اصلی کوتاه‌تر باشد، نکات اصلی و مهم آن را در برداشته باشد، خوانا بوده و بین جملات آن پیوستگی وجود داشته باشد. این فناوری به منظور تلاش برای فهمیدن اینکه آیا یک متن طولانی مطابق با نیاز کاربر است و ارزش بیشتر خواندن را دارد یا خیر، بسیار مفید می‌باشد. این سیستم قادر به پردازش و خلاصه‌سازی یک متن بسیار طولانی در زمانی معادل با زمان خواندن تنها یک پاراگراف متن توسط کاربر می‌باشد. نکته کلیدی خلاصه‌سازی، کاهش طول و جزئیات یک سند با حفظ مفاهیم اصلی و معنای کلی آن است. چالش مطرح در اینجا این است که اگرچه رایانه‌ها قادر به شناسایی افراد، مکان‌ها، زمان و ... هستند، اما آموزش نرم‌افزار به منظور تجزیه و تحلیل معانی و تفسیر آن‌ها بسیار پیچیده است. خلاصه‌سازی به دو صورت استخراجی (گزینشی) و چکیده انجام می‌شود. در نوع استخراجی از روش‌های آماری استفاده شده و متن خلاصه با انتخاب جملاتی از متن اصلی به دست می‌آید. در خلاصه‌سازی از نوع چکیده از روش‌های پردازش زبان طبیعی استفاده می‌شود و خلاصه‌ی متن پس از فهم مطالب موجود در متن اصلی، تولید می‌شود. در خلاصه‌سازی یک مقدار به صورت درصد به عنوان سطح خلاصه‌سازی متن توسط کاربر تعیین می‌شود. فرآیند خلاصه‌سازی خودکار متن طی سه مرحله شامل پیش‌پردازش، تبدیل ساختار متن به ساختار خلاصه و تولید متن خلاصه انجام می‌شود (گوپتا و لهال، ۲۰۰۹).

۴-۶ خوشه‌بندی

روشی است که برای گروه‌بندی اسناد مشابه مورد استفاده قرار می‌گیرد. این روش با رده‌بندی تفاوت دارد و در آن به جای استفاده از موضوعات از پیش تعریف شده، گروه‌بندی به صورت خودکار و پویا انجام می‌شود. مزیت آن این است که اسناد می‌توانند در چند زیر عنوان ظاهر شوند؛ بنابراین یک سند مفید هیچ‌وقت از دست نخواهد رفت. یک الگوریتم خوشه‌بندی پایه، یک بردار از موضوعات برای هر سند ایجاد می‌کند و وزن‌ها را اندازه‌گیری کرده و تصمیم می‌گیرد که سند برای کدام موضوع مناسب‌تر است. خوشه‌بندی برای اسناد بسیار کاربردی است.

در الگوریتم‌های خوشه‌بندی مانند میانگین-K¹، هنگام محاسبه شباهت بین اسناد متنی، نه تنها بردار ویژه مبتنی بر الگوریتم آماری فراوانی لغات در نظر گرفته می‌شود، بلکه ترکیب درجه ارتباط بین کلمات و همچنین رابطه بین کلیدواژه‌ها در نظر گرفته می‌شود. در نتیجه دقت خوشه‌بندی متن افزایش می‌یابد (گوپتا و لهال، ۲۰۰۹).

۶-۵ ارتباط‌دهنده مفاهیم

ابزارهای ارتباط‌دهنده مفاهیم، اسناد مرتبط را با شناسایی مفاهیم معمولاً مشترک به یکدیگر متصل می‌کنند و به کاربران کمک می‌کنند تا اطلاعاتی را بیابند که شاید با جستجوی سنتی قادر به یافتن آن نبودند. این روش مشاهده و مرور اطلاعات را به جای جستجوی آن ترویج می‌دهد. ارتباط مفاهیم، یک مفهوم ارزشمند در متن کاوی است به خصوص در زمینه پزشکی که در آن تحقیقات زیادی انجام شده است و برای محققان غیرممکن است که تمامی مطالب موجود را خوانده و آن‌ها را به سایر تحقیقات ارتباط دهند. در حالت ایدئال، نرم‌افزار ارتباط‌دهنده مفاهیم می‌تواند پیوند بین بیماری‌ها و درمان‌ها را شناسایی کند، حتی زمانی که انسان به دلیل حجم زیاد اطلاعات در دسترس قادر به انجام این کار نیست (گوپتا و لهال، ۲۰۰۹).

۶-۶ تصویرسازی اطلاعات

تصویرسازی اطلاعات یا متن کاوی بصری^۲ منابع متنی عظیم را در سلسله‌مراتب تصویری یا نقشه قرار می‌دهد و علاوه بر قابلیت جستجوی ساده، قابلیت مرور را نیز فراهم می‌کند. تصویرسازی اطلاعات هنگامی مفید است که کاربر به محدود کردن طیف گسترده‌ای از اسناد و کشف موضوعات مرتبط نیاز دارد. به‌عنوان مثال دولت می‌تواند با استفاده از تصویرسازی اطلاعات شبکه‌های تروریستی را شناسایی کرده یا در مورد جنایات به اطلاعاتی دست یابد که پیش‌ازاین ارتباطی بین آن‌ها تصور نمی‌شد (گوپتا و لهال، ۲۰۰۹).

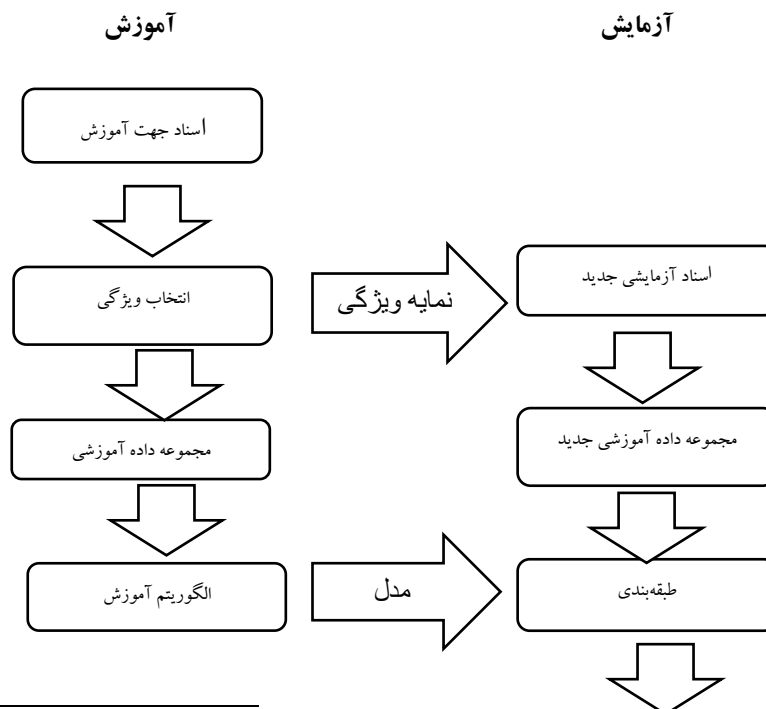
۷. مدل پیشنهادی

رده‌بندی متن فناوری کلیدی برای سازمان‌دهی داده‌های متنی بوده (چانگا و دیگران، ۲۰۰۸؛ الزمیل و کن، ۲۰۱۱) و به‌طور گسترده‌ای در هنگام سازمان‌دهی اسناد دیجیتالی مورد استفاده قرار می‌گیرد. با توجه به افزایش تعداد اسناد دیجیتالی، رده‌بندی خودکار متن امروزه بیش‌ازپیش

¹ K-means

² visual text mining

مورد توجه قرار گرفته است. رده‌بندی شامل شناخت موضوع یا موضوعات اصلی یک سند و قرار دادن آن در یک مجموعه شامل تعداد ثابتی از موضوعات از پیش تعریف شده است (چانگا و دیگران، ۲۰۰۸؛ گوپتا و لهال، ۲۰۰۹؛ الزمیل و کن، ۲۰۱۱). در رده‌بندی، یک سند به صورت کیسه‌ای از کلمات^۱ در نظر گرفته می‌شود. پس از شمارش کلمات ظاهر شده در اسناد، محاسبات خاصی صورت می‌گیرد و موضوعات اصلی که یک سند را پوشش می‌دهند، شناسایی می‌شوند. بلوک دیاگرام شکل ۶ جریان رده‌بندی متن را به صورت کلی نشان داده است. در این دیاگرام دو بخش آزمایش و آموزش وجود دارد که در بخش آموزش ویژگی‌هایی از اسناد انتخاب می‌گردد و با توجه به داده‌های آموزشی یک الگوریتم شکل می‌گیرد که معمولاً از الگوریتم‌های یادگیری ماشین تحت سرپرست استفاده می‌شود. هدف آن‌ها یادگیری طبقه بندها از نمونه‌های شناخته شده (به اصطلاح اسناد دارای برچسب) و انجام رده‌بندی خودکار نمونه‌های ناشناخته (به اصطلاح اسناد بدون برچسب) است. این موضوع در شکل ۶ نشان داده شده است (گوپتا و لهال، ۲۰۰۹)؛ بنابراین طراحی یک الگوریتم انتخاب ویژگی خوب برای رده‌بندی متن، جایی که تعداد زیادی از ویژگی‌ها که نشان‌دهنده کلمات بوده و لغات دارای چالش‌های جدی برای اثربخشی و کارایی طبقه بندها هستند، از اهمیت خاصی برخوردار است. (جاوید، ۲۰۱۵)



¹ Bag of Words

تخصیص دسته

شکل ۶. بلوک دیاگرام جریان رده‌بندی متن

رده‌بندی اغلب متکی بر یک واژه‌نامه است که در آن موضوعات از قبل تعریف شده و روابط میان کلمات مشخص هستند. ابزارهای رده‌بندی به‌طور معمول از روشی برای رتبه‌بندی اسناد به‌منظور تعیین اسنادی که دارای بیشترین محتوا درباره‌ی موضوعی خاص هستند، استفاده می‌کنند. رده‌بندی نیز همانند خلاصه‌سازی می‌تواند به همراه روش ردیابی موضوع به‌منظور تشخیص ارتباط یک سند با فردی که در جستجوی اطلاعات درباره‌ی موضوعی است، مورد استفاده قرار گیرد. هدف از رده‌بندی متون نسبت دادن کلاس‌های از پیش تعریف شده به اسناد متنی است. مثلاً یک خبر جدید که وارد می‌شود بگوییم متعلق به کلاس ورزشی یا سیاسی یا هنری. برای رده‌بندی اسناد روش‌های گوناگونی به کار می‌روند. در رده‌بندی یک مجموعه آموزش از اسناد وجود دارد که برای این مجموعه کلاس‌ها مشخص است. با استفاده از این مجموعه مدل رده‌بندی مشخص می‌شود، سپس با استفاده از آن کلاس سند جدید که وارد می‌شود، مشخص می‌گردد. دو رویکرد پایه برای جستجوی اسناد مشابه وجود دارد. رویکرد اول استخراج کلیدواژه‌ها از سند (رویکرد مبتنی بر کلیدواژه) (ونگ و لین ۲۰۰۳؛ ونگ و لو، ۲۰۰۴) و دیگری رویکرد مبتنی بر استفاده از تمامی کلمات در سند است. در این روش بردار ویژگی^۱ سند با کمک تمام کلمات موجود در سند تعیین شده و شباهت مورد جستجو قرار می‌گیرد. این ویژگی بین تمام لغات و همچنین تمام اسناد با استفاده از مدل فضای بردار^۲ نمایش داده می‌شود. اگرچه روش مبتنی بر کلیدواژه به فرآیند تصمیم‌گیری سرعت می‌بخشد، اما تشکیل بردار ویژگی سند با استفاده از تمام کلمات، نتایج جستجو را قطعی‌تر و دقیق‌تر می‌سازد (سارا کوگلو و دیگران، ۲۰۰۷).

با توجه به مطالب بررسی شده جستجوی سند مشابه به‌عنوان بخشی از متن‌کاوی در نظر گرفته می‌شود و شامل مراحل پیش‌پردازش، رده‌بندی و خوشه‌بندی است. به‌عنوان مثال یکی از نویسندگان محبوب کتاب‌های کودکان به نام لایمن فرانک باوم مشهور به ال. فرانک باوم که خالق اثر جادوگر شهر از می‌باشد، در سن ۶۳ سالگی درگذشت. ایشان در مدت‌زمان زندگی ۱۹ کتاب درباره جادوگر شهر از خود به یادگار گذاشتند. اتفاق جالبی که یک سال پس از مرگ

¹ feature vector

² Vector Space Model

ایشان رخ داد چاپ اثر دیگری از ایشان بود که به دلایلی قابل قبول نبود. بر همین اساس پژوهشگران داده کاوی و متن کاوی کتب ایشان را به پنج هزار تکه رده بندی کرده و فراوانی کلمات را در هر دسته شمارش کردند. سپس پنجاه کلمه ای که بیشتر استفاده شده بود را به عنوان پرکاربردترین کلمات استخراج کردند.

در نهایت مقایسه ای بین آثار قبلی نویسنده و اثر منتشر شده جدید صورت گرفته است، نوشته های فرانک باوم در سمت راست وجود دارد که فراوانی استفاده از کلمات مشخص شد. با توجه به بررسی های صورت گرفته روی متون، کاهش بعد برای مصورسازی و تحلیل مؤلفه های اصلی، می توان نتیجه گرفت که اسلوب نگارش نویسنده کاملاً متفاوت می باشد. لذا با توجه به دقت بالای محاسبات، اثر جدید با آثار قبلی نویسنده همخوانی نداشته و مالکیت معنوی آثار ایشان مشخص می شود. البته روش های هوش مصنوعی زیادی در این مراحل مورد استفاده قرار گرفته اند. همچنین از این موضوع به عنوان یکی از عملیات اساسی مدیریت اسناد می توان بهره برد. علاوه بر این، یافتن اسناد مرتبط با اسناد موجود از حجم عظیمی از اسناد، موضوع مهمی است. این به معنی عدم ارائه اسناد نامرتب یا کاهش تعداد اسناد نامرتب تا حد ممکن در نتایج جستجو است. تمامی این مباحث ما را به سوی طراحی یک سیستم جستجوی مؤثرتر سوق می دهد (سارا کوگلو و دیگران، ۲۰۰۸).

برای اندازه گیری کارایی مدل رده بندی، یک مجموعه تست که مستقل از مجموعه آموزش است در نظر گرفته می شود؛ و برچسب هایی که برای این اسناد توسط مدل تخمین زده می شود با برچسب واقعی اسناد مقایسه می شود. نسبت اسنادی که به درستی رده بندی شده اند به تعداد کل اسناد accuracy نامیده می شود سه معیار برای مقایسه رده بندی کننده ها استفاده می شود: (۱) precision: کسری از اسناد بازیابی شده ای که مربوط هستند (۲) recall: نشان دهنده کسری از اسناد مربوط بازیابی شده است. این دو معیار به صورت زیر تعریف می شوند:

$$precision = \frac{\#\{relevant \cap retrieved\}}{\#retrived} \quad (5)$$

$$recall = \frac{\#\{relevant \cap retrieved\}}{\#relevant} \quad (6)$$

بین precision و recall مصالحه^۱ وجود دارد. به همین دلیل معیار دیگری به نام F-score که مصالحه‌ای بین هر دو برقرار می‌کند، برای اندازه‌گیری کارایی کل رده‌بندی کننده به کار می‌رود.

$$f = \frac{2}{1/recall + 1/precision}$$

(۷)

برخی از روش‌های به کار گرفته شده برای رده‌بندی متن مبتنی بر نظریه‌های آماری و یادگیری ماشین عبارت‌اند از k نزدیک‌ترین همسایه^۲، نیو بیز^۳، ماشین بردار پشتیبان^۴ و شبکه عصبی^۵.

در خصوص مقالات فارسی هر یک از چالش‌های رسم‌الخط زبان فارسی رایانه‌ای، رفتار متفاوتی از خود نشان می‌دهد و می‌توان آن‌ها را در رده‌بندی متن لحاظ کرد، نتیجه این رفتارها در چندین آثار فارسی به‌قرار زیر است:

- ۱- بررسی نحوه استفاده از حروف اضافه در مقالات و آثار قبلی نویسندگان و تطبیق دادن با نمونه جدید.
- ۲- تنوع نحوه استفاده از پیشوندها و پسوندها مانند "می"، "ها"، موارد فوق به‌طور چسبیده یا جدا از کلمه بکار برده می‌شود.
- ۳- بکار بردن "حمزه" به‌صورت‌های مختلف. جستجو برای کلمات مشابه، باحالت‌های مختلف "حمزه" تبدیل می‌شود. به‌عبارت‌دیگر کاوش کلمه "مسئله" به کاوش برای کلمات "مسئله" و "مسأله" منجر می‌شود. می‌توان با جایگزینی "ی" بجای "ء" نیز دامنه کاوش را وسیع‌تر نمود، مثل "رئیس" و "رییس".
- ۴- استفاده یا عدم استفاده از "ء" در ترکیب‌های اضافی یا وصفی
- ۵- استفاده از "ا" و "آ" در آثار
- ۶- استفاده از اصطلاح‌نامه در زبان انگلیسی (Thesaurus) برای حل مشکل تنوع املائی کلمات. این معضل در زبان‌های دیگر شامل تنوع استفاده از "ی" در کلمات عربی مختوم به "ا"، تنوع املائی بعضی کلمات که همه درست هستند.

^۱ trade off

^۲ K-Nearest Neighbor (KNN)

^۳ naive bayes

^۴ Support Vector Machine

^۵ neural network

۷- استفاده از کلمات خارج از زبان فارسی به صورت ترجمه فارسی یا معادل فارسی آن‌ها در مقالات و نوشته‌ها

۸- تبدیل کلمات اروپایی به رسم الخط فارسی با همان تلفظ اصلی (Cross language Retrieval):

مانند برنامه‌های "Open Source" که از کلمه "سورس باز" بجای آن استفاده می‌شود.
۹- استفاده یا عدم استفاده از جمع مکسر برای بعضی کلمات. برای مثال کلمه "موسی"، به کلمه "موسا" و کلمه "کامپیوتر" به کلمه "رایانه" متناظر شده است.

نمونه‌ای از پایگاه داده که با مشاوره انجمن‌ها، بزرگان و فرهنگستان ادب فارسی تشکیل شده است، در جدول ۴ نشان داده شده است:

جدول ۴. نمونه‌ای از محتویات پایگاه داده مترادف‌ها

واژه استفاده شده	واژه اصلی
موسا	موسی
امپراطور	امپراتور
آنتولوژی، انتولوژی، انتالوژی، هستی‌شناسی	Ontology
کامپیوتر، رایانه	Computer
منبع، سورس	Source

۸. یافته‌ها

فنون متن‌کاوی اشکال متفاوتی دارند که متناسب با نوع داده‌های متنی، برخی از فنون برای مشابه‌یابی موضوعیت بیشتری می‌یابد. از میان متن‌های مختلف تنها متنی که روابط بسیار دقیق و معناداری میان آن برقرار است تا حدی که برای محققان متن‌کاوی جلب توجه کرده، اعجاز عددی الفاظ قرآن است که مشابه‌یابی با فنون مختلف متن‌کاوی به‌خوبی جواب داده است. یافتن متون مشابه از نظر مفهومی به صورت خودکار یکی از کاربردهای اصلی و مهم متن‌کاوی در مدیریت و سازمان‌دهی اسناد متنی به شمار می‌رود. هر یک از کارهای تحقیقاتی انجام‌شده در این زمینه به‌نوعی باهدف گرفتن چالش‌های مطرح سعی در بهبود بخشی از عملیات این فرآیند دارند که

منجر به بهبود عملکرد و کارایی کل اثر می‌شود. از چالش‌های موجود در این زمینه می‌توان به موارد زیر اشاره کرد:

- ابعاد بالای فضای ویژگی با توجه به تعداد زیاد لغات در متن
- چندمعنایی لغات موجود در متن و مسئله لغات مترادف
- ارتباط مفهومی میان کلمات موجود در متن
- پراکندگی مفاهیم به خصوص در متون کوتاه
- دقت الگوریتم مشابه یابی متون
- سرعت الگوریتم مشابه یابی متون

هر روشی که پاسخ مناسب‌تری برای رسیدگی به چالش‌های مطرح در این حوزه ارائه دهد، بر بهبود عملکرد یافتن متون مشابه تأثیری مثبت خواهد داشت.

با توجه به موارد بیان‌شده و روش‌های متن‌کاوی رویکردی که مطرح می‌شود حفاظت از مالکیت معنوی و فکری در آثار نویسندگان می‌باشد. در این نظریه با استفاده از علم متن‌کاوی و ابزارهای پیوند مفهومی **Concept Linkage** می‌توان اسناد مرتبط را از نظر معنایی بررسی کرد و اشتراک بین آن‌ها را به صورت یک قانون کلی، با دقت بالا و به صورت مدل عینی پیاده‌سازی کرد. این امر یکی از مهم‌ترین مواردی که می‌تواند برخلاف شیوه‌های سنتی و قدیمی برای حفاظت از اثر نویسندگان بسیار مفید واقع گردد. بدین منظور می‌توان با بررسی آثار قبلی و قدیمی نویسنده یک مدل مفهومی از آن‌ها استخراج کرد. به عنوان مثال با استفاده از کلماتی که در شیوه نگارش نویسندگان موجود است و حروف اضافه‌ای که همواره استفاده می‌کنند یک مدل خاص برای صاحبان اثر، ایجاد می‌شود. به این صورت با استفاده از متن‌کاوی و کشف الگوهای معنادار از متن می‌توان در مورد اثری که نویسنده‌ای مدعی آن است بررسی انجام داد و صاحب اثر اصلی را تشخیص داد که این امر انقلاب بزرگی در حفظ آثار متنی به وجود می‌آورد.

همچنین پیوند مفاهیم ارزشمند در متن‌کاوی، به‌ویژه در حوزه سلامت و زیست پزشکی بسیار کاربردی می‌باشد که در مورد آن پژوهش بسیاری انجام شده است. مطمئناً خواندن حجم زیادی از محتوا و مرتبط ساختن آن‌ها به دیگر پژوهش‌ها، برای پژوهشگران غیرممکن است، بنابراین برنامه‌های پیوند مفاهیم ممکن است پیوند میان بیماری‌ها و درمان‌ها را تشخیص دهد که معمولاً

انسان‌ها از عهده آن بر نمی‌آیند. برای نمونه، نرم‌افزارهای متن‌کاوی ممکن است به آسانی پیوند میان عنوان «آ» و «ب» و «پ» و «ت» را که روابط مشهور هستند، تشخیص دهد؛ ابزارهای متن‌کاوی همچنین قادر به تشخیص پیوند بالقوه میان «آ» و «ت» نیز هستند. با توجه به محدودیت‌های آنی و فیزیکی پژوهشگر انسانی برای بررسی حجم قابل توجه متن‌ها و کشف ارتباط بین آن‌ها که باعث عدم کشف ارتباط و دسته‌بندی متون می‌شود، نیاز به استفاده از روش‌های متن‌کاوی بسیار ضروری احساس می‌شود.

لذا برای ایجاد شبکه‌های مفهومی استفاده از داده‌کاوی و فنون متن‌کاوی ضروری می‌باشد. بسیاری از این ابزارها در ایجاد پیوند میان متون، مفاهیم، اصطلاح‌ها و ایجاد شبکه‌ای منسجم از دادگان و در نتیجه کشف اطلاعات و دانش از آن‌ها، نقشی مؤثر خواهند داشت؛ به‌ویژه هنگامی که حجم منابع بسیار زیاد است و انسان به‌سختی قادر به طبقه‌بندی و درک ارتباط بین آن‌ها می‌باشد. ابزار متن‌کاوی، منابع متن‌کاوی و یا برنامه‌های تخصصی متن‌کاوی از ملزومات یک پژوهش جامع در مورد تمام بخش‌های کار می‌باشد؛ اما در آینده نیاز است تا هر کدام از بخش‌های انتخاب ویژگی، بازنمایش ویژگی و کاهش ویژگی به شکل تخصصی مورد بررسی قرار بگیرد؛ اما برای بررسی تخصصی‌تر می‌توان به‌صورت خاص منابع متن‌کاوی مثل رسانه‌های اجتماعی، تیتراهای خبری، متن‌های خبری و متون علمی و غیره را مورد بررسی قرارداد. بنابراین با استفاده از ابزارهایی مانند رایانه و علوم مثل زبان‌شناسی، فلسفه ذهن، فلسفه زبان، هوش مصنوعی، همچنین تکنیک‌های سازمان‌دهی متون و منابع، با استفاده از نظام‌های فهرست‌نویسی، نمایه‌سازی و طبقه‌بندی و... علاوه بر اینکه راه‌حلی برای دسترسی سریع و آسان به اطلاعات و کشف دانش فراهم می‌شود، می‌تواند روش مؤثری برای حفاظت از مالکیت فکری و معنوی آثار متنی نیز به شمار رود.

جدول ۵. مقایسه در زمینه مشابه یابی

نام مقاله	موضوعات و مزیت‌ها
سارا کولگو و دیگران، ۲۰۰۸	پیشنهاد روشی دو مرحله‌ای برای رده‌بندی اسنادی که متعلق به چند طبقه هستند با استفاده از خوشه‌بندی فازی و رده‌بندی شباهت فازی.
اوگوز، ۲۰۱۱	ارائه یک روش دو مرحله‌ای برای استخراج ویژگی و انتخاب ویژگی به منظور بهبود عملکرد رده‌بندی اسناد مشابه با استفاده از رتبه‌بندی به روش IG و استخراج و انتخاب ویژگی با استفاده از الگوریتم ژنتیک (GA) و تجزیه و تحلیل مؤلفه‌های اصلی (PCA) به صورت جداگانه.
وی و دیگران، ۲۰۱۵	یکپارچه‌سازی وردنت با زنجیره‌ی لغوی با استفاده از ساختار سلسله مراتبی هستان‌شناسی و ارتباطات برای ارائه یک ارزیابی دقیق‌تر بین لغات برای ابهام‌زدایی مفهوم کلمه و در نتیجه کاهش ابعاد فضای ویژگی.
وو و اوک، ۲۰۱۵	استفاده از تجزیه و تحلیل مدل موضوعی بر اساس تخصیص دیریکله نهان به منظور بهبود عملکرد رده‌بندی متون

کوتاه.	
ارائه یک الگوریتم انتخاب ویژگی دو مرحله‌ای برای انتخاب ویژگی در رده‌بندی متون با در نظر گرفتن تعاملات کلمات را به منظور حذف کلمات زائد.	جاوید و دیگران، ۲۰۱۵

۹. نتیجه‌گیری

با گسترش روزافزون متون چاپی و متون منتشرشده در فضای اینترنت، رشد اطلاعات افزایش یافته و سرریز اطلاعات رخ می‌دهد. از آنجایی که پردازش دستی این داده‌های متنی کاری طاقت‌فرسا است، به روش‌های متن‌کاوی احتیاج است. متن‌کاوی که تحت عناوین تجزیه و تحلیل هوشمند متن، کاوش داده‌های متنی و کشف دانش از متون نیز شناخته می‌شود، به‌طور کلی به فرآیند استخراج اطلاعات و دانش جالب توجه و غیر بدیهی از متن بدون ساختار اشاره دارد. متن‌کاوی یک حوزه تحقیقاتی جوان و میان‌رشته‌ای بوده و در عین حال که ریشه عمیقی در NLP دارد، برای فرآیند کشف دانش با روش‌هایی از بازیابی اطلاعات، داده‌کاوی، آمار، یادگیری ماشین، استدلال، استخراج اطلاعات، مدیریت دانش، زبان‌شناسی محاسباتی و ... در ارتباط است. این حوزه تمام فعالیت‌هایی که به‌نوعی به دنبال کسب دانش از متن هستند را شامل می‌گردد. آنالیز داده‌های متنی توسط تکنیک‌های یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روش‌های مرتبط دیگر همگی در زمره مقوله یادگیری متن قرار می‌گیرند. طبقه‌بندی متن، یعنی انتساب اسناد متنی بر اساس محتوی به یک یا چند طبقه از قبل تعیین شده که یکی از مهم‌ترین مسائل در متن‌کاوی است. مرتب کردن بلادرنگ نامه‌های الکترونیکی یا فایل‌ها در سلسله مراتبی از پوشه‌ها، تشخیص موضوع متن، جستجوی مفهومی از متون منتشرشده، از جمله کاربردهای مبحث طبقه‌بندی (دسته بندی-کلاسه بندی) متن است. در بسیاری از موارد، افراد حرفه‌ای آموزش دیده، برای طبقه‌بندی متون جدید به کار گرفته می‌شوند. این فرآیند بسیار زمان‌بر و پرهزینه است و لذا کاربرد خود را محدود می‌سازد، به همین منظور علاقه روزافزونی به توسعه فناوری‌هایی در دسته‌بندی خودکار متن ابراز می‌شود. از طرفی با توجه به حجم گسترده آثار تحقیقاتی که به‌صورت متن ذخیره می‌شوند؛ انتظار می‌رود که متن‌کاوی یکی از مهم‌ترین تکنولوژی‌ها در آینده باشد، بنابراین ایجاد یک چهارچوب برای استفاده از متن‌کاوی در حفاظت از مالکیت معنوی و فکری نویسندگان و بررسی مشکلات و چالش‌های آن، زمینه تحقیقات بعدی را در این حوزه تشکیل می‌دهد. با توجه به حجم وسیع اطلاعات متنی آنلاین در دسترس ایجاد سیستم‌های متن‌کاوی

به شدت مورد نیاز است. چنین سیستمی می تواند در میان گزینه های ممکن قرار بگیرد. این جنبه از متن کاوی با توجه به زمینه در حال رشد آن می تواند به طور مستقل بسیار جذاب باشد و از سایه متن کاوی عمومی خارج شود. این شاخه از متن کاوی برای پیش بینی بازار فارغ از بحث آنالیز احساسات زمینه ای بسیار امیدوار کننده است.

۱۰. مراجع

سارا مصباح، دکتر مسعود رهگذر، متن کاوی، دانشکده فنی، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران ۱۳۸۸

ابراهیم هژیر، داده کاوی و مفاهیم کاربرد، دانشگاه آزاد ورزقان ۱۳۹۳

- Al Zamil M.G.H. Can A.B. 2011. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems, 24(1), Elsevier, February, pp. 58–65
- Changa Y. Chena S. Liaub C. 2008. Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. Expert Systems with Applications, 34(3), Elsevier, April, pp. 1948–1953
- Fodeh S. Punch B. Tan P. N. 2011. On ontology-driven document clustering using core semantic features. Knowledge and Information Systems, 28(2), Springer-Verlag, August, pp. 395-421.
- Gupta V. Lehal G.S. 2009. A Survey of Text Mining Techniques and Applications. Emerging Technologies in Web Intelligence, 1(1), Academy Publisher, August, pp. 60-76.
- Hashimi H. Hafez A. Mathkour H. 2015. Selection criteria for text mining approaches. Computers in Human Behavior, 51(B), Elsevier, October, pp. 729–733.
- Javed K. Maruf S. A.Babri H. 2015. A two-stage Markov blanket based feature selection algorithm for text classification. Neurocomputing, 157, Elsevier, June, pp. 91–104.
- Salton, G. 1989. Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.
- Saracoglu R. Tutuncu K. Allahverdi N. 2007. A fuzzy clustering approach for finding similar documents using a novel similarity measure. Expert Systems with Applications, 33(3), Elsevier, October, pp. 600–605.
- Saracoglu R. Tutuncu K. Allahverdi N. 2008. A new approach on search for similar documents with multiple categories using fuzzy clustering. Expert Systems with Applications, 34(4), Elsevier, May, pp. 2545–2554.
- Suh J. H. Park C. H. Jeon S. H. 2010. Applying text and data mining techniques to forecasting the trend of petitions filed to e-People. Expert Systems with Applications, 37(10), Elsevier, October, pp. 7255–7268.
- Uguz H. 2011. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), Elsevier, October, pp. 1024–1032.

Weng S. Lin Y. 2003. A study on searching for similar documents based on multiple concepts and distribution of concepts. *Expert Systems with Applications*, 25(3), Elsevier, October, pp. 355–368.

Vo D. H. Ock C. Y. 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, 42(3), Elsevier, February, pp. 1684–1698.

Wei T. Lu Y. Chang H. Zhou Q. Bao X. 2015. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4), Elsevier, March, pp. 2264–2275.

N. Kanya*, S. Geetha "INFORMATION EXTRACTION -A TEXT MINING APPROACH" 2007 produced IEEE

Raymond J. Mooney and Un Yong Nahm "Text mining with Information Extraction", 2005

M. Rajman. "Text Mining, Knowledge extraction from unstructured textual data". Proc. of EUROSTAT Conference, Frankfurt (Deutschland), May, 1997

Data mining Concepts and Techniques: Jiawei Han and Micheline Kamber.

H. Karanikas and B. Theodoulidis, 'Knowledge discovery in text and text mining software', Technical report, UMIST - CRIM, Manchester, 2002.

N. Kanya, S. Geetha "Information Extraction -A Text mining approach" ICTES 2007, Dec. 20-22, 2007. pp.1111-1118.

MA in Information Technology Engineering - electronic commerce,
master degree, K. N. Toosi University of Technology:

Azamsadat Parei- as.parei@gmail.com

PhD in Information Technology Engineering Group, Assistant
Professor Department of Industrial Engineering, K. N. Toosi
University of Technology:

Dr. hodjat hamidi- h_hamidi@kntu.ac.ir

**An approach to managing and organizing text documents using of
intelligent text analysis**

Abstract

Regarding the fact that stored data occupies a large space in organizations and retention systems and information management that has been resulted in gigantic data warehouses, extract the appropriate model need is felt increasingly. Text mining is one of the most significant methods for extracting a useful and appropriate model that helps organizations achieving their goals through extraction and adaption f knowledge out of data sets. Those methods allow for a new horizon for trading and protecting intellectual property of authors' works. In this paper, a new approach is needed to decipher the text patterns to organize and intelligent text analysis. The main purpose of applying the proper method of preserving the works of writers, scholars and text documents. Regarding the number of those works and documentary management systems the size of available data has been increased considerably. In order to uncover the implicit knowledge out of this data with considerable usefulness for users a specific method is required that has been practiced in the data mining field. Much of this data is unstructured or semi-structured text are available which allow one to use it can be used in addition to data mining methods, technologies such as natural language processing, intelligent analysis and Science Statistics used.

Keywords: text search, text management, intellectual property, information extraction, data mining



اعظم السادات پرنی دارای مدرک تحصیلی کارشناسی ارشد در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک از دانشگاه صنعتی خواجه نصیرالدین طوسی است. تجارت الکترونیک و کسب و کار از جمله علایق پژوهشی وی است.



دکتر حجت اله حمیدی متولد سال ۱۳۵۵ دارای مدرک تحصیلی دکتری در رشته مهندسی کامپیوتر - فناوری اطلاعات است. ایشان هم‌اکنون استادیار گروه فناوری اطلاعات دانشگاه صنعتی خواجه نصیرالدین طوسی است. تجارت الکترونیک، کسب و کار هوشمند، و محاسبات نرم از جمله علایق پژوهشی وی است.