

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

اعظم‌السادات پرنی

کارشناسی ارشد مهندسی فناوری اطلاعات؛
دانشکده مهندسی صنایع؛ دانشگاه صنعتی خواجه
نصیرالدین طوسی as.parei@gmail.com

حجت‌اله حمیدی

دکتری مهندسی کامپیوتر - فناوری اطلاعات؛
استادیار؛ دانشکده مهندسی صنایع؛
دانشگاه صنعتی خواجه نصیرالدین طوسی؛
پدیده‌آور رابط h_hamidi@kntu.ac.ir



مقاله برای اصلاح به مدت ۱ روز نزد پدیدآوران بوده است.

پذیرش: ۱۳۹۵/۰۶/۰۲

دریافت: ۱۳۹۵/۰۴/۱۵

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۸۲۳۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در SCOPUS، ISI، LISTA و
jjpm.irandoc.ac.ir
دوره ۳۲ | شماره ۴ | صص ۱۱۷۱-۱۲۰۲
تابستان ۱۳۹۶



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، 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 رویکردی جدید به کشف الگوهای متنی جهت سازمان‌دهی و تجزیه و تحلیل هوشمند متن می‌پردازیم. هدف اصلی، به کارگیری 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 نیمه‌ساختارمند هستند که برای استفاده از آن‌ها می‌توان علاوه بر روش‌های مورد استفاده در داده کاوی، از فناوری‌هایی مانند پردازش زبان طبیعی، تجزیه و تحلیل هوشمند و علم آمار بهره گرفت.

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

۱. مقدمه

داده کاوی^۱ به پردازش داده‌های عظیم و استخراج اطلاعات کاربردی از آن‌ها می‌پردازد (Saracoglu, Tutuncu, and Allahverdi 2007). داده‌ها ممکن است در چارچوب‌های داده‌های توصیفی مانند گزارش‌های کیفیت اسناد و یادداشت‌هایی برای کمک به مشتریان باشد (Kornfein and Goldfarb 2007). لذا، تلاش‌های تکنیکی و دستی برای هدایت کردن این منابع داده‌ای، به‌منظور کشف الگوها و داده‌های سودمند موجود در این منابع مورد نیاز است (Kornfein and Goldfarb 2007). از آنجا که بالای ۸۰ درصد از این اطلاعات به‌صورت متنی نگه‌داری می‌شوند، باور بر این است که متن کاوی ارزش بالقوه تجاری بالایی داشته باشد (Gupta, and Lehal 2009). انتقال این منابع مفید اطلاعاتی به چارچوب‌های قابل استفاده باعث پیشرفت فناوری‌های آینده است و به کیفیت خدمات‌رسانی کمک خواهد کرد. با گسترش روزافزون اینترنت و فناوری‌های اطلاعاتی، حجم اطلاعات در دسترس بسیار افزایش یافته است (Weng and Lin 2003). از آنجا که پردازش دستی این داده‌های متنی کاری طاقت‌فرساست، به روش‌های متن کاوی^۲ احتیاج هست (Hashimi, Hafez, and Mathkour 2015). بنابراین، افراد وقت بیشتری را صرف پالایش کردن اطلاعات می‌کنند (Weng and Lin 2003). متن کاوی با اطلاعات متنی یا بدون ساختار^۳، برای استخراج اطلاعات با معنا و دانش از مقدار زیادی اطلاعات سروکار دارد. امروزه، راه‌حل‌های متنوعی برای مدیریت و سازمان‌دهی مقدار زیادی از اسناد متنی و به‌دست آوردن اطلاعات مفید از این داده‌ها (متن کاوی) در حال پیشرفت و تحقیق است. جست‌وجوی سند مشابه به‌عنوان بخشی از متن کاوی در نظر گرفته‌شده و روش‌های هوش مصنوعی^۴ زیادی در این مراحل مورد استفاده قرار گرفته‌اند. از این رو، این موضوع به‌عنوان یکی از عملیات اساسی مدیریت اسناد مشهود است (Saracoglu, Tutuncu, and Allahverdi 2008). علاوه بر این، یافتن اسناد مرتبط و مشابه با اسناد موجود از حجم عظیمی از اسناد، نقش مهمی را در متن کاوی ایفا می‌کند. تمامی این مباحث ما را به‌سوی طراحی یک سیستم جست‌وجوی مؤثرتر سوق می‌دهد (همان). متن کاوی فرایند تجزیه و تحلیل اسناد به‌وسیله جست‌وجوی الگوها در متون به زبان طبیعی و به‌منظور استخراج اطلاعات برای اهداف خاص است (Weng and Lin 2003). متن کاوی شاخه‌ای از حوزه داده کاوی است که سعی در یافتن الگوهای

1. data mining

2. text mining

3. unstructured

4. artificial intelligence

جالب توجه از پایگاه داده‌های بزرگ دارد و طی آن کشف اطلاعات جدید یا استخراج خودکار اطلاعات از منابع مکتوب مختلف به وسیله رایانه امکان‌پذیر می‌شود. متن کاوی که تحت عناوین تجزیه و تحلیل هوشمند متن، کاوش داده‌های متنی و کشف دانش از متون نیز شناخته می‌شود، به‌طور کلی، به فرایند استخراج اطلاعات و دانش جالب توجه و غیربدیهی از متن بدون ساختار اشاره دارد (Gupta, and Lehal 2009).

این مقاله به‌صورت زیر سازمان‌دهی شده است: پس از مقدمه‌ای کوتاه در بخش اول، 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 شامل: ۱. پیش‌پردازش که خود از بخش‌های مدل فضای برداری و پیش‌پردازش زبان‌شناختی تشکیل شده است، ۲. تولید و استخراج ویژگی، و ۳. انتخاب ویژگی است، 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، نتیجه‌گیری ارائه خواهد شد.

۲. مروری بر پیشینه پژوهش

موضوع اصلی در کاوش متن، یافتن اسناد مرتبط از میان حجم عظیمی از اسناد است (Saracoglu, Tutuncu, and Allahverdi 2008). دو رویکرد پایه برای جست‌وجوی اسناد مشابه وجود دارد. رویکرد اول استخراج کلیدواژه‌ها از سند (رویکرد مبتنی بر کلیدواژه) (Weng and Lin 2003) و دیگری رویکرد مبتنی بر استفاده از تمامی کلمات سند است. فرایند کشف دانش از پایگاه داده‌ها ضمن این که سبب طبقه‌بندی معتبر، به‌روز و قابل‌استفاده از داده‌ها می‌شود، می‌تواند الگوهای قابل فهمی از داده‌ها ارائه می‌دهد (Fayyad 1996). در متن کاوی به‌طور کلی، استخراج ویژگی تأیید شده است (Gunal and et al. 2006). انتخاب ویژگی (Feng and et al. 2012) روش طبقه‌بندی (Wu and Li 2013) و پیش‌پردازش (Gunal and et al. 2006) تأثیر قابل توجهی بر موفقیت فرایندهای طبقه‌بندی متن داشته

است. عملکرد طبقه‌بندی متن از طریق یک اصطلاح‌نامه مبتنی بر پیکره و وردنت^۱، با به‌کارگیری الگوریتم نزدیک‌ترین همسایگی و شبکه عصبی پسرانتشار (BPNN)^۲ الگوریتم را بهبود دادند (Jiang and et al. 2012). الگوریتم k-NN برای طبقه‌بندی متن با ترکیب آن با یک الگوریتم خوشه‌بندی مورد استفاده قرار می‌گیرد. بیزین^۳ ساده، الگوریتمی دیگر است که برای طبقه‌بندی متن موفق بوده است (Chen and et al. 2009). برای مثال، مراحل فرایند متن‌کاوی در اخبار و اطلاعات در جدول ۱ نشان داده شده است. در این جدول 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	پیش برنامهریزی 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 Kešelj (2009)	گزارش‌ها سالیانه	وبسایت شرکت	اشاره نشده	بله	بله
Schumaker and Chen (2009)	اخبار مالی	Yahoo Finance	۲,۸۰۰	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 بیانیه	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

1. WordNet
2. Back Proagation Neural Network
3. Bayes' theorem

نویسنده	انواع متون	منبع متن	تعداد مورد بررسی	بدون ساختار	پیش برنامه‌ریزی شده
Bollen and Huina (2011)	تویت‌ها	توییتر	۹,۸۵۳,۴۹۸	بله	خیر
Vu et al. (2012)	تویت‌ها	توییتر	۵,۰۰۱,۴۶۰	بله	خیر
Schumaker et al. (2012)	اخبار مالی	Yahoo Finance	۲,۸۰۲	بله	خیر
Lugmayr and Gossen(2012)	اخبار کارگزاری	کارگزاری‌ها	اشاره نشده	بله	خیر
Yu, Duan, et al. (2013)	رسانه‌های اجتماعی متعارف	وبلاگ‌ها، فروم‌ها، اخبار و شبکه‌های اجتماعی	۵۲,۷۴۶ پیام	بله	خیر
Hagenau et al. (2013)	اطلاعه‌های شرکت و اخبار مالی	DGAP, EuroAdhoc	۳,۴۷۸ و ۱۰,۸۷۰ به ترتیب	بله	خیر
Jin et al. (2013)	اخبار عمومی	Bloomberg	۳۶۱,۷۸۲	بله	خیر
Chatrath et al. (2014)	اخبار اقتصاد کلان	Bloomberg	اشاره نشده	خیر	بله

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

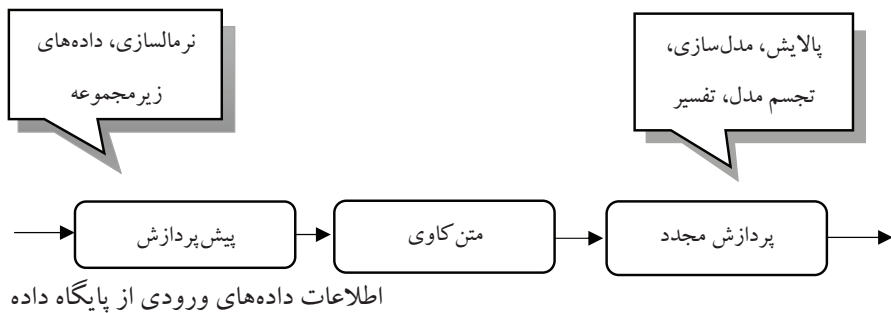
۳. ارتباط متن کاوی و دیگر دانش‌ها

۳-۱. کشف دانش و ارتباط آن با متن کاوی

کشف دانش در پایگاه داده‌ها (KDD)^۱ در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، سطح بالا و به دنبال جست‌وجوی دانش در اطلاعات شکل گرفته است و هدف آن کشف ارتباط و نظم بین اطلاعات قابل مشاهده است. گاهی داده کاوی شامل همه ابعاد فرایند کشف دانش است و در بعضی موارد نیز داده کاوی به عنوان بخشی از فرایندهای KDD در نظر گرفته می‌شود و فاز مدل کردن را توصیف می‌کند. در نهایت، KDD فرایند یافتن اطلاعات و الگوهای مفید از داده را گویند و داده کاوی بهره‌گیری از الگوریتم‌هایی برای

1. Knowledge Data Discovery

یافتن اطلاعات مفید و الگوها در فرایند KDD است. تعاریف مختلفی برای کشف دانش در پایگاه داده وجود دارد؛ مثلاً بیان شده است که KDD فرایند 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 خصوصیات که برای اندازه‌گیری کیفیت الگوهای پیداشده 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 به‌عنوان یک فرایند که به‌وسیله چندین گام پردازش کردن تعریف می‌شود، در نظر گرفت. این گام‌ها به‌منظور استخراج الگوهای مفید باید بر روی مجموعه داده‌ای اعمال شوند. طبق نظرات قبلی گام‌ها را می‌توان به‌صورت زیر بیان کرد: (۱) درک کردن کسب و کار (۲) درک کردن داده (۳) آماده‌سازی داده (۴) مدل کردن (۵) ارزیابی (۶) گسترش. مرحله پیش‌پردازش غالباً یکی از مراحل زمان‌بر و در عین حال، بسیار مهم در کسب نتیجه مطلوب است؛ به‌ویژه این که در متن کاوی، نیاز به متدهای پیش‌پردازش خاصی برای تبدیل داده متنی به فرامتی مناسب برای الگوریتم‌های داده کاوی داریم.



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

۲-۳. تفاوت متن کاوی و داده کاوی

با توجه به رشد و توسعه روزافزون فناوری، اطلاعات به‌صورت گسترده و با سرعت بالا منتشر می‌شود. در نتیجه، یکی از زمینه‌های تحقیقاتی مهم، ذخیره داده‌ها و به‌دست آوردن اطلاعات سودمند از آن‌هاست. داده کاوی برای این منظور به‌وجود آمد و می‌توان گفت که مقدار زیادی از دانش را پردازش کرده و آن‌ها

را به اطلاعات کاربردی تبدیل می‌کند و در صورتی که داده‌های مورد پردازش به صورت متن و بدون ساختار یا نیمه‌ساختارمند باشند، این فرایند متن کاوی نامیده می‌شود (Weng and Lin 2003; Saracoglu, Tutuncu, and Allahverdi 2008). متن کاوی مشابه داده کاوی است (Berry and Linoff 2004)، با این تفاوت که ابزارهای داده کاوی برای مدیریت داده‌های ساختارمند از پایگاه داده‌ها طراحی شده‌اند (Elmasri and Navathe 2000). متن کاوی می‌تواند با مجموعه داده‌های بدون ساختار یا نیمه‌ساختارمند مانند نامه‌های الکترونیکی، اسناد تمام‌متنی و پرونده‌های وب کار کند. در نتیجه، متن کاوی راه‌حل بهتری برای شرکت‌هاست. با این حال، تا امروز، بیشتر تلاش‌های تحقیق و توسعه روی داده کاوی با استفاده از داده‌های ساختارمند متمرکز بوده‌اند (Weng and Lin 2003; Saracoglu, Tutuncu, and Allahverdi 2008; Gupta, and Lehal 2009).

جدول ۲. ارتباط داده کاوی و متن کاوی

کشف سازوکار مشخص	جست‌وجوی هدف مشخص	
داده کاوی	بازیابی داده‌ها	داده‌های ساخت‌یافته
متن کاوی	بازیابی اطلاعات	داده‌های بدون ساختار - متنی

۳-۳. تفاوت متن کاوی و وب کاوی

بر این اساس، متن کاوی با جست‌وجو در وب نیز متفاوت است. در وب کاوی، 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 متن کاوی، هدف کشف اطلاعات ناشناخته است، یعنی چیزی که هنوز کسی آن را نمی‌داند و نمی‌تواند نوشته شده باشد (Gupta, and Lehal 2009).

۴. متن کاوی و جریان کار در متن کاوی

متن کاوی یا کشف دانش از متن برای اولین بار برای استخراج الگوهای متنی بیان شد. می‌توان گفت که متن کاوی از تکنیک‌های بازیابی اطلاعات، استخراج اطلاعات، همچنین پردازش زبان طبیعی استفاده می‌کند و آن‌ها را به الگوریتم‌ها و متدهای KDD، داده کاوی، یادگیری ماشین و آماری مرتبط می‌کند. با توجه به ناحیه‌های تحقیق گوناگون، برای هر

یک از آن‌ها می‌توان تعاریف مختلفی از متن کاوی در نظر گرفت. در ادامه، برخی از این تعاریف بیان می‌شوند:

متن کاوی به معنای استخراج اطلاعات: در این تعریف متن کاوی 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 فیلهای یادگیری ماشین و آماری برای متن‌ها با هدف 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 داده کاوی را بر روی داده‌های استخراج‌شده اعمال کرد.

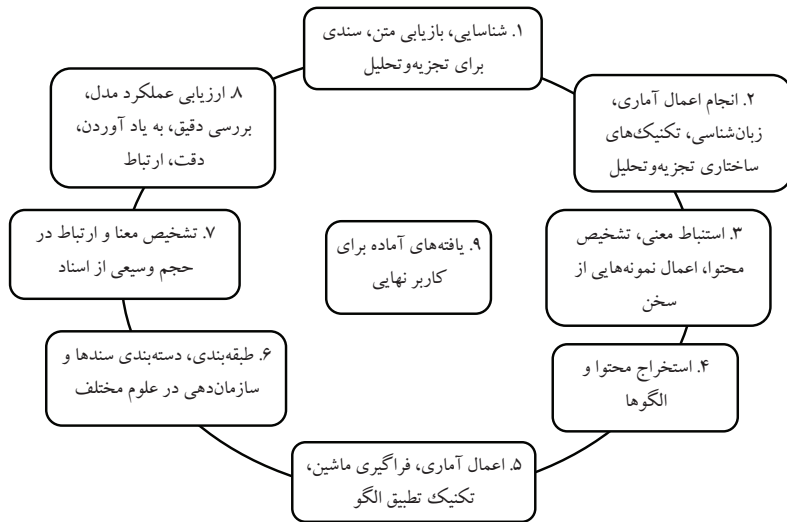
۴-۱. مزایای متن کاوی

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. (Hashimi, Hafez, and Mathkour (2015) 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 جست‌وجوی مختلف به‌منظور کارآمدتر و دقیق‌تر کردن فرایند جست‌وجو
- ◇ 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.



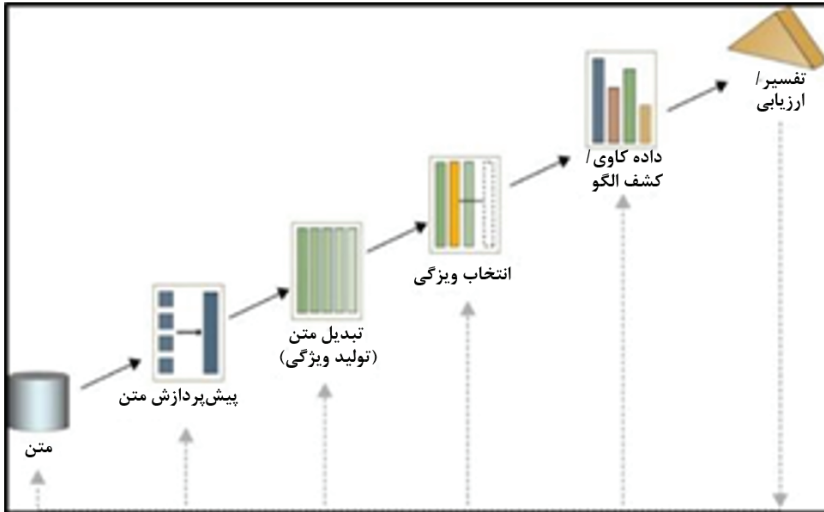
شکل ۲. کلیه مراحل جریان کار عمومی در متن کاوی

۵. مراحل متن کاوی

فرایند متن کاوی شامل مراحل اصلی پیش‌پردازش، استخراج ویژگی و انتخاب ویژگی است. مراحل کلی فرایند متن کاوی در شکل ۳ نشان داده شده است. ابتدا فرایند پیش‌پردازش روی متن مورد نظر انجام می‌شود و پس از آن عملیات استخراج ویژگی و انتخاب ویژگی‌ها روی آن انجام می‌گیرد. این داده‌ها سپس، با استفاده از الگوریتم‌های مخصوص و الگوهای موجود شناسایی شده و جهت ارزیابی^۱ یا تفسیر^۲ به خروجی فرستاده می‌شود. توضیحات بیشتر در ادامه آمده است.

1. evaluation

2. interpretation



شکل ۳. مراحل کلی فرایند متن کاوی (Hashimi, Hafez, and Mathkour (2015)

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 گردد. زمانی که داده‌های ورودی در دسترس است باید آن‌ها را برای ورود به الگوریتم‌های یادگیری ماشین آماده کنیم. این مرحله برای داده‌های متنی به این معناست که آن‌ها را از حالت غیرساخت‌یافته به فرمت ساختاریافته و قابل تشخیص برای ماشین تبدیل کنیم. هدف پیش‌پردازش کاهش ابعاد فضای نمایش لغات موجود در سند بوده و معمولاً شامل وظایف زیر است (Gupta, and Lehal 2009):

- ◇ حذف ایست‌واژه‌ها (کلماتی که در اکثر متون یافت می‌شوند، اما دارای اهمیت معنایی خاصی نبوده و جزو کلیدواژه‌ها به حساب نمی‌آیند).
- ◇ یکسان‌سازی 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 معنای آن تأکید می‌کند.

برای تعریف رسمی تر الگوریتم‌ها برخی تعاریف لازم را بیان می‌کنیم: مجموعه

اسناد با D نمایش داده می‌شود. T نشان‌دهندهٔ 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 به صورت $(tf(d, t_1) \dots tf(d, t_m))$ نمایش داده می‌شود. مرکز مجموعه X از بردارهای ترم به صورت مقدار میانگین بردارهای ترم مربوط 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 مرحلهٔ انتخاب ویژگی اتفاق بیفتد، می‌تواند باعث تبدیل شدن مسئلهٔ متن‌کاوی به یک مسئلهٔ غیرقابل حل برای تمام الگوریتم‌های یادگیری ماشین بشود (Pestov 2013). 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 جملهٔ 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 انتخاب ترم تنها کلمات انتخاب‌شده برای توصیف کردن اسناد استفاده می‌شوند. یک روش برای انتخاب کلمات، استخراج کلمات بر اساس درجهٔ خلوص^۱ آن‌هاست. برای هر کلمه t در دیکشنری درجهٔ خلوص آن به صورت زیر تعریف می‌شود:

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

درجهٔ خلوص نشان‌دهندهٔ این است که یک کلمه چقدر اسناد را به خوبی توسط کلمهٔ جست‌وجو جدا می‌کند. اگر کلمه در تعداد زیادی از اسناد رخ داده باشد، درجهٔ خلوص کمی دارد. روش‌های ریشه‌یابی سعی می‌کنند فرم‌های اولیهٔ کلمات را بسازند. مثلاً حذف کردن علامت جمع (s) از اسم‌ها یا ing از افعال و ... بعد از فرایند ریشه‌یابی،

1. entropy

هر کلمه توسط ریشه آن کلمه نمایش داده می‌شود. در واقع، در این جا 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)	Bag-of-words برای لغات منفی	لغتنامه از پیش تعریف شده، فرهنگ لغت روان‌شناسی	فراوانی تقسیم بر تعداد کل کلمات
Mahajan et al. (2008)	Latent Dirichlet Allocation (LDA)	استخراج بیست و پنج 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)	Bag of words لحن و محتوایی	لغتنامه از پیش تعریف شده	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 نمره‌دهی ویژگی با استفاده از Information Gain و معیار مجذور کای	TF-IDF
Bollen and Huina (2011)	By OpinionFinder	By OpinionFinder	By OpinionFinder
Vu et al. (2012)	تعداد کل جملات مثبت یا منفی تویتر	شناسایی نام‌های موجود	Neg_Pos and Bullish_Bearish
Schumaker et al. (2012)	OpinionFinder overall tone and polarity	حداقل اتفاق در هر سند	Binary
Lugmayr and Gossen(2012)	Bag-of-words	ریشه یابی	مقادیر احساسات
Yu, Duan, et al. (2013)	Bag-of-words	اشاره نشده	Binary Hagenau

نویسنده	انتخاب ویژگی	کاهش ابعاد	بازنمایش ویژگی
Hagenau et al. (2013)	Bag-of-words, noun phrases, word combinations, n-grams	Frequency for news, Chi 2 –approach and bi-normal separation (BNS) for exogenous-feedback-based feature selection, dictionary	TF-IDF
Jin et al. (2013)	Latent Dirichlet Allocation (LDA)	استخراج موضوع	توزیع هر مقاله
Chatrath et al. (2014)	داده‌های ساخت یافته	داده‌های ساخت یافته	داده‌های ساخت یافته

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. پنج 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) و نوسان معکوس فراوانی سند^۵ (TF-IDF). مقایسه‌ای میان این روش‌ها را می‌توان در پژوهش Gungor و Tasci (2013) مشاهده کرد. پایه‌ای‌ترین آن‌ها روش باینری است که در آن با اختصاص دو مقدار ۱ و ۰ با حضور یا عدم حضور یک ویژگی بررسی می‌شود (Schumaker and et al 2012). روش رایج بعدی اصطلاحاً TF-IDF نام دارد (Hegna and Murtomaa 2003) 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 کلمات برقرار کند. روش‌های رایج و مشابه دیگری نیز وجود دارد. به عنوان مثال TF-CDF، که از روش CF

1. information gain
2. Chi-square statistics
3. document frequency
4. accuracy balanced
5. term frequency-inverse document frequency

توسعه یافته است، ادعا می‌کند که در عمل می‌تواند کارایی بیشتری نسبت به TF-IDF داشته باشد.

۵-۱-۱. پیش‌پردازش زبان‌شناختی

گاهی اوقات ممکن است از پیش‌پردازش‌های زبان‌شناختی نیز برای افزایش اطلاعات قابل دسترس درباره‌ی ترم‌ها استفاده شود. برای این منظور روش‌های زیر اعمال می‌شود:

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 (Witten 2004)، برای مثال سیستمی به نام GATE در دانشگاه شفیلد، در یک کتابخانه دیجیتال به این قصد جاگذاری شده است. GATE شامل ابزاری برای برچسب‌زدن 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. ^۱ 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: هدف، گروه‌بندی کردن کلمات مجاور در یک جمله است.

رفع ابهام معنای کلمه: در این جا سعی می‌شود که ابهامات در معنای یک کلمه یا عبارت رفع شود، مثل زمانی که کلماتی دارای چندین معنای متفاوت هستند. 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 بازبازی اطلاعات مانند پالایش و

1.

2. Text chunking

3. Word Sense Disambiguation (WSD)

جست‌وجوی اطلاعات مرتبط می‌توانند از تحقیقات در زمینه رده‌بندی متن سود ببرند، مشکل اصلی رده‌بندی متن، ابعاد 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 فرایندی است که زیرمجموعه‌ای از ویژگی‌های اصلی را با توجه به برخی از معیارها یا اهمیت ویژگی انتخاب می‌کند. الگوریتم‌های انتخاب ویژگی به دو دسته تقسیم می‌شوند:

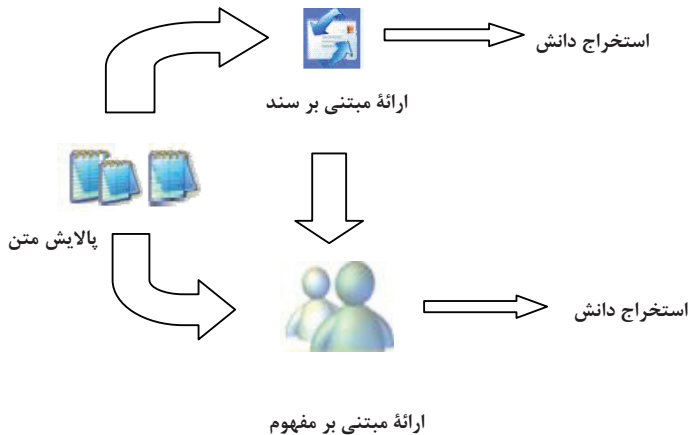
الگوریتم‌های رتبه‌بندی ویژگی: این الگوریتم‌ها ویژگی‌هایی را که به‌تنهایی برای طبقه‌بندی بسیار مرتبط هستند، بدون در نظر گرفتن تعامل ویژگی‌ها انتخاب می‌کنند. این موضوع، الگوریتم‌های FR را ساده کرده و از نظر محاسباتی کارآمدتر از الگوریتم‌های انتخاب زیرمجموعه ویژگی می‌سازد و بنابراین، یک انتخاب ارجح برای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

الگوریتم‌های انتخاب زیرمجموعه ویژگی (FSS): این الگوریتم‌ها تعامل بین ویژگی‌ها را در نظر می‌گیرند و در عین حال، از نظر محاسباتی پیچیده هستند (Javed, Maruf, and Babri, 2015). یک عنصر کلیدی فرایند متن کاوی، ارتباط اطلاعات استخراج شده 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، و ۲) مبتنی بر مفهوم. در فرمت نمایش مبتنی بر سند آنچه مهم است، نحوه نمایش بهتر مستندات است؛ مثلاً تبدیل اسناد به یک فرمت میانی و نیمه‌ساخت یافته، یا به کاربردن یک ایندکس بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را کارتر می‌کند. در نوع دوم، نمایش اسناد بهبود بخشیده می‌شود و 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، در حالی که نمایش مبتنی بر مفهوم برای یافتن روابط میان مفاهیم، ساختن اتوماتیک تزاروس و آنولوژی و نظایر آن به کار می‌رود که در شکل ۴ نمایش داده شده است.



شکل ۴. فرایند متن کاوی

شکل ۴. فرایند متن کاوی



شکل ۵. مدیریت مناسب داده‌ها (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.

۱-۶. 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 درون متن را شناسایی می‌کند و این کار را با جست‌وجوی دنباله‌های از پیش تعریف شده در متن طی فرایندی، 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 بامعنا 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، در حالی که این طور نبوده و با کمک این روش، اطلاعاتی که هیچ‌گونه ساختاری ندارند، تبدیل به اطلاعاتی می‌شوند که دارای ساختار بوده و می‌توانند به صورت بهتری مورد استفاده واحد کشف دانش قرار بگیرند (Gupta and Lehal 2009).

۱-۱-۶. رده‌بندی برای استخراج اطلاعات

استخراج اشیاء بر اساس درک پیغام فرمول‌بندی می‌شوند. کلماتی که شیء با آن‌ها

آغاز می‌شود، برچسب B، کلماتی که داخل شیء می‌آیند، برچسب A، و کلمات خارج از شیء 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 این جا 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 جزء 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 جزء ویژگی اگر کلمه قبلی، کلمه "John" است، مقدار ۱ و در غیر این صورت مقدار ۰ می‌گیرد. علاوه بر تست کردن وجود کلمات خاص می‌توان این را نیز تست کرد که آیا کلمات با حروف بزرگ شروع می‌شوند و پسوند خاصی دارند یا نه. هم‌اکنون می‌توان هر متد رده‌بندی کارایی را برای رده‌بندی کردن برچسب‌های کلمات با استفاده از بردار ویژگی ورودی به کار برد.

۲-۶. ردیابی موضوع

این روش با استفاده از نگهداری پروفایل کاربر و بر اساس موضوعات مورد علاقه انتخاب شده توسط او یا پیگیری اسنادی که او مشاهده می‌کند، سایر اسناد مورد علاقه وی را پیش‌بینی می‌کند. یک نمونه از این روش توسط وب‌سایت «یاهو» مورد استفاده قرار گرفته است. این فناوری می‌تواند کاربردهای گوناگونی داشته باشد. به‌عنوان مثال، می‌تواند یک شرکت را از وجود اطلاعاتی در رابطه با رقبا در اخبار آگاه نماید. همچنین، می‌تواند مورد استفاده پزشکانی قرار گیرد که در جست‌وجوی درمان جدید برای بیماری‌ها هستند یا می‌خواهند در جریان آخرین پیشرفت‌ها قرار گیرند. همچنین، در حوزه آموزش می‌توان از فناوری ردیابی موضوع برای اطمینان از در اختیار داشتن جدیدترین منابع تحقیقاتی مورد علاقه استفاده نمود (Gupta and Lehal 2009).

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 دربرداشته باشد، و خوانا بوده و بین جملات آن پیوستگی وجود داشته باشد. این فناوری به منظور تلاش برای فهمیدن این که آیا یک متن طولانی مطابق با نیاز کاربر بوده و ارزش خواندن تمام آن را دارد یا خیر، بسیار مفید است. این سیستم قادر به پردازش و خلاصه‌سازی یک متن بسیار طولانی در زمانی معادل با زمان خواندن تنها یک پاراگراف متن توسط کاربر است. نکته کلیدی خلاصه‌سازی، کاهش طول و جزئیات یک سند با حفظ مفاهیم اصلی و معنای کلی آن است. چالش مطرح در این جا 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 خلاصه 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. فرایند خلاصه‌سازی خودکار متن طی سه مرحله شامل پیش‌پردازش، تبدیل ساختار متن به ساختار خلاصه، و تولید متن خلاصه انجام می‌شود (Gupta and Lehal 2009).

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 زیرعنوان 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-

K^۱، 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 (همان).

۶-۵. ارتباط دهنده مفاهیم

ابزارهای ارتباط دهنده 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 جست‌وجوی سنتی قادر به یافتن آن نبودند. این روش، مشاهده و مرور اطلاعات را به جای جست‌وجوی 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 ایده‌آل، نرم‌افزار ارتباط دهنده 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 (همان).

۶-۶. تصویرسازی اطلاعات

تصویرسازی اطلاعات یا متن کاوی بصری^۲، منابع متنی عظیم را در سلسله مراتب تصویری یا نقشه قرار می‌دهد و علاوه بر قابلیت جست‌وجوی 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. به عنوان مثال، دولت می‌تواند با استفاده از تصویرسازی اطلاعات شبکه‌های تروریستی را شناسایی کرده یا در مورد جنایات به اطلاعاتی دست یابد که پیش از این ارتباطی بین آن‌ها تصور نمی‌شد (همان).

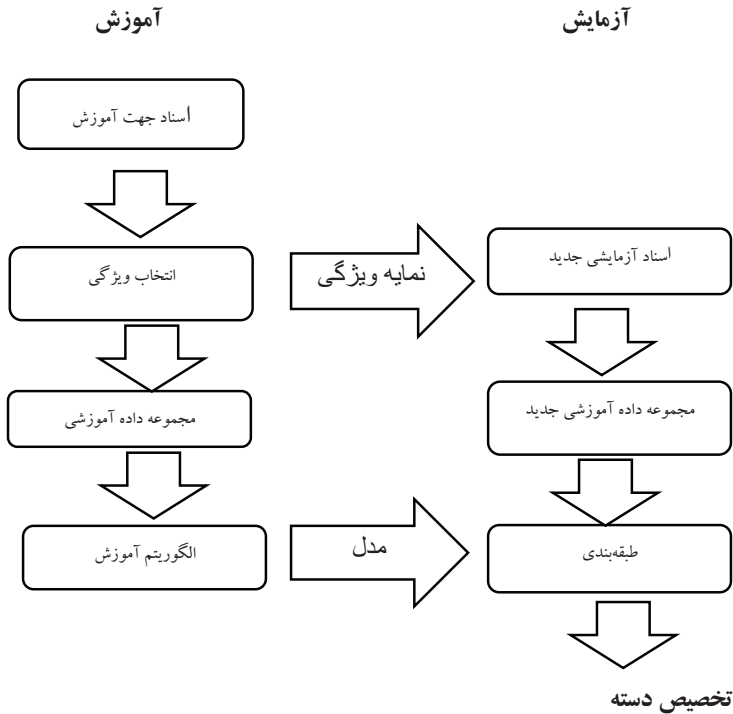
1. K-means

2. visual text mining

۷. مدل پیشنهادی

رده‌بندی متن، فناوری کلیدی برای سازمان‌دهی داده‌های متنی بوده (Changa, 2011; Al Zamil and Can 2011; Chena, and Liaub 2008) و به‌طوری گسترده در هنگام سازمان‌دهی اسناد دیجیتالی مورد استفاده قرار می‌گیرد. با توجه به افزایش تعداد اسناد دیجیتالی، رده‌بندی خودکار متن امروزه بیش از پیش 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 (Changa, Chena, and Liaub 2008; Gupta and Lehal 2009; Al Zamil and Can 2011). در رده‌بندی، یک سند به‌صورت کیسه‌ای از کلمات^۱ در نظر گرفته می‌شود. پس از شمارش کلمات ظاهرشده 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 به‌صورت 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 طبقه‌بندها از نمونه‌های شناخته‌شده (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. 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 طبقه‌بندها هستند، از اهمیت خاصی برخوردار است (Javed, Maruf, and Babri 2015).

1. bag of words



شکل ۶. بلوک دیاگرام جریان رده‌بندی متن (Gupta and Lehal 2009)

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 درباره موضوعی خاص هستند، استفاده می‌کنند. رده‌بندی نیز همانند خلاصه‌سازی می‌تواند به همراه روش ردیابی موضوع به‌منظور تشخیص ارتباط یک سند با فردی که در جست‌وجوی اطلاعات درباره موضوعی است، مورد استفاده قرار گیرد. هدف از رده‌بندی متون نسبت‌دادن کلاس‌های از پیش تعریف‌شده به اسناد متنی است. مثلاً خبر جدیدی که وارد می‌شود، بگویم متعلق است به کلاس ورزشی یا سیاسی یا هنری. برای رده‌بندی اسناد روش‌های گوناگونی به کار می‌رود. 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 کلاس‌ها سند جدیدی که وارد می‌شود، مشخص می‌گردد. دو رویکرد پایه برای جست‌وجوی اسناد مشابه وجود دارد. رویکرد اول استخراج کلیدواژه‌ها از سند

(رویکرد مبتنی بر کلیدواژه) (Weng and Lin 2003) 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 جست‌وجو 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 فرایند تصمیم‌گیری سرعت می‌بخشد، اما تشکیل بردار ویژگی سند با استفاده از تمام کلمات، نتایج جست‌وجو را قطعی‌تر و دقیق‌تر می‌سازد (Saracoglu, Tutuncu, and Allahverdi 2007).

با توجه به مطالب بررسی شده جست‌وجوی سند مشابه به‌عنوان بخشی از متن کاوی 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»^۳ مشهور به «ال. فرانک باوم» که خالق اثر «جادوگر شهر» می‌باشد، در سن ۶۳ سالگی درگذشت. وی ۱۹ کتاب درباره «جادوگر شهر» از خود به یادگار گذاشت. اتفاق جالبی که یک سال پس از مرگ وی رخ داد، چاپ اثر جدیدی با نام ایشان بود. بر همین اساس پژوهشگران داده کاوی و متن کاوی کتاب‌های او را به ۵۰۰۰ تکه رده‌بندی کرده و فراوانی کلمات را در هر دسته شمارش کردند. سپس، پنجاه کلمه‌ای را که بیشتر استفاده شده بود، به‌عنوان پرکاربردترین کلمات استخراج کردند.

در نهایت، مقایسه‌ای بین آثار قبلی این نویسنده و اثر منتشرشده جدید صورت گرفت و فراوانی استفاده از کلمات در نوشته‌های فرانک باوم مشخص شد. با توجه به بررسی‌های صورت گرفته روی متون و کاهش بعدی برای مصورسازی و تحلیل مؤلفه‌های اصلی می‌توان نتیجه گرفت که اسلوب نگارش نویسنده کاملاً متفاوت است. لذا، با توجه به دقت بالای محاسبات، اثر جدید با آثار قبلی نویسنده همخوانی نداشته و مالکیت معنوی آثار ایشان مشخص می‌شود. البته روش‌های هوش مصنوعی زیادی در این مراحل مورد استفاده قرار گرفته‌اند. همچنین، از این موضوع به‌عنوان یکی از عملیات اساسی مدیریت اسناد می‌توان بهره برد. علاوه بر این، یافتن اسناد مرتبط با اسناد موجود از میان حجم عظیمی از اسناد، موضوع مهمی است. این به معنای عدم ارائه اسناد نامرتبط یا کاهش

1. feature vector

2. Vector Space Model

3. Lyman Frank Baum

تعداد اسناد نامرتبط تا حد ممکن در نتایج جست‌وجو است. تمامی این مباحث ما را به‌سوی طراحی یک سیستم جست‌وجوی مؤثرتر سوق می‌دهد (Saracoglu, Tutuncu, and Allahverdi 2007).

برای اندازه‌گیری کارایی مدل رده‌بندی، یک مجموعه تست، که مستقل از مجموعه آموزش است در نظر گرفته می‌شود و برچسب‌هایی که برای این اسناد توسط مدل تخمین زده می‌شود با برچسب واقعی اسناد مقایسه می‌شود. نسبت اسنادی که به‌درستی رده‌بندی شده‌اند به تعداد کل اسناد accuracy نامیده می‌شود. سه معیار برای مقایسه رده‌بندی‌کننده‌ها استفاده می‌شود: (۱) precision: کسری از اسناد بازیابی شده‌ای که مربوط هستند، (۲) recall: نشان‌دهنده کسری از اسناد مربوط بازیابی شده است. این دو معیار به‌صورت زیر تعریف می‌شوند:

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

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

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 رده‌بندی‌کننده به کار می‌رود.

$$f = \frac{2}{1/recall + 1/precision} \quad (7)$$

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

-
1. trade off
 2. K-Nearest Neighbor (KNN)
 3. naive bayes
 4. Support Vector Machine
 5. neural network

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 به قرار 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 به کار برده می‌شود.
 ۳. به کار بردن «حمزه» به صورت‌های مختلف. به کار بردن «حمزه» به صورت‌های مختلف. جست‌وجو برای کلمات مشابه، با حالت‌های مختلف «حمزه». به عبارت دیگر کاوش کلمه «مسئله» به کاوش برای کلمات «مسئله» و «مسأله» منجر می‌شود. می‌توان با جایگزینی «ی» به جای «ء» نیز دامنه کاوش را وسیع‌تر نمود، مثل «رییس» 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^۱ در زبان انگلیسی برای حل مشکل تنوع املائی کلمات. برای حل مشکل تنوع املائی کلمات. این معضل در زبان‌های دیگر شامل تنوع استفاده از «ی» در کلمات عربی مختوم به «ا»، می‌باشد.
 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 رسم‌الخط فارسی با همان تلفظ اصلی (crosslanguage retrieval): مانند برنامه‌های «Open Sourc» که از کلمه «سورس باز» به جای آن استفاده می‌شود.
 ۹. استفاده یا عدم استفاده از جمع مکسر برای بعضی کلمات. برای مثال، کلمه «موسی»، به کلمه «موسا»
- نمونه‌ای از پایگاه داده که با مشاوره انجمن‌ها، بزرگان و فرهنگستان ادب زبان فارسی تشکیل شده، در جدول ۴ نشان داده شده است:

1. thesaurus

جدول ۴. نمونه‌ای از محتویات پایگاه داده مترادف‌ها

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

۸. یافته‌ها

فنون متن کاوی اشکال متفاوتی دارند و برخی از فنون متناسب با نوع داده‌های متنی 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 متن کاوی جلب توجه کرده، اعجاز عددی الفاظ «قرآن» است که مشابه‌یابی با فنون مختلف متن کاوی به خوبی جواب داده است. یافتن متون مشابه از نظر مفهومی به صورت خود کار یکی از کاربردهای اصلی و مهم متن کاوی در مدیریت و سازمان‌دهی اسناد متنی به شمار می‌رود. هر یک از کارهای تحقیقاتی انجام شده در این زمینه به نوعی با هدف گرفتن چالش‌های مطرح سعی در بهبود بخشی از عملیات این فرایند دارد 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؛
- ◇ چند معنایی لغات موجود در متن و مسئله لغات مترادف؛
- ◇ ارتباط مفهومی میان کلمات موجود در متن؛
- ◇ پراکنندگی مفاهیم به خصوص در متون کوتاه؛
- ◇ دقت الگوریتم مشابه‌یابی متون؛
- ◇ سرعت الگوریتم مشابه‌یابی 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 بیان شده و روش‌های متن کاوی، رویکردی که مطرح می‌شود برای حفاظت از مالکیت معنوی و فکری در آثار نویسندگان است. در این نظریه با استفاده از

علم متن کاوی و ابزارهای پیوند مفهومی^۱ 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 است که می‌تواند بر خلاف 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 متن کاوی 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 متن کاوی، به ویژه در حوزه سلامت و زیست‌پزشکی بسیار کاربردی است و 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 متن کاوی ممکن است به آسانی پیوند میان عنوان «آ» و «ب» و «پ» و «ت» را که روابط مشهور هستند، تشخیص دهند. ابزارهای متن کاوی همچنین قادر به تشخیص پیوند بالقوه میان «آ» و «ت» نیز هستند. با توجه به محدودیت‌های آنی و فیزیکی پژوهشگران 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 متن کاوی بسیار ضروری احساس می‌شود.

این است که برای ایجاد شبکه‌های مفهومی، استفاده از داده کاوی و فنون متن کاوی ضروری می‌باشد. بسیاری از این ابزارها در ایجاد پیوند میان متون، مفاهیم، اصطلاح‌ها و ایجاد شبکه‌ای منسجم از داده‌ها و در نتیجه، کشف اطلاعات و دانش از آن‌ها نقشی مؤثر خواهند داشت؛ به ویژه هنگامی که حجم منابع بسیار زیاد است و انسان به سختی قادر به طبقه‌بندی و درک ارتباط بین آن‌هاست. ابزار متن کاوی، منابع متن کاوی و یا

1. concept linkage

برنامه‌های تخصصی متن کاوی از ملزومات یک پژوهش جامع در مورد تمام بخش‌های کار می‌باشد. اما، در آینده نیاز است که هر کدام از بخش‌های انتخاب و ویژگی، بازنمایش و ویژگی و کاهش ویژگی به شکل تخصصی مورد بررسی قرار گیرد. اما برای بررسی تخصصی‌تر می‌توان به صورت خاص منابع متن کاوی مثل رسانه‌های اجتماعی، تیتراهای خبری، متن‌های خبری و متون علمی و غیره را مورد بررسی قرار داد. 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 این که راه‌حلی برای دسترسی سریع و آسان به اطلاعات و کشف دانش فراهم می‌شود، می‌تواند روشی مؤثر برای حفاظت از مالکیت فکری و معنوی آثار متنی نیز به‌شمار رود.

جدول ۵. مقایسه در زمینه مشابه‌یابی

نام مقاله	موضوعات و مزیت‌ها
Saracoglu, Tutuncu, and Allahverdi 2007	پیشنهاد روشی دومرحله‌ای برای رده‌بندی اسنادی که متعلق به چند طبقه هستند با استفاده از خوشه‌بندی فازی و رده‌بندی شباهت فازی.
Uguz 2011	ارائه یک روش دومرحله‌ای 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.
Wei, others 2015	یکپارچه‌سازی وردنت با زنجیره لغوی با استفاده از ساختار سلسله‌مراتبی هستان‌شناسی و ارتباطات برای ارائه یک ارزیابی دقیق‌تر بین لغات برای ابهام‌زدایی مفهوم کلمه و در نتیجه، کاهش ابعاد فضای ویژگی.
Vo, and Ock 2015	استفاده از تجزیه و تحلیل مدل موضوعی به منظور بهبود عملکرد رده‌بندی متون کوتاه.
Javed, Maruf, and Babri 2015	ارائه یک الگوریتم انتخاب ویژگی دومرحله‌ای برای انتخاب ویژگی در رده‌بندی متون با در نظر گرفتن تعاملات کلمات به منظور حذف کلمات زائد.

۹. نتیجه‌گیری

با گسترش روزافزون متون چاپی و متون منتشرشده در فضای اینترنت، رشد اطلاعات افزایش یافته و سرریز اطلاعات رخ می‌دهد. از آنجا که پردازش دستی این داده‌های متنی کاری طاقت‌فرساست، به روش‌های متن کاوی احتیاج است. متن کاوی که تحت عناوین

تجزیه و تحلیل هوشمند متن، کاوش داده‌های متنی و کشف دانش از متون نیز شناخته می‌شود، به‌طور کلی به فرایند استخراج اطلاعات و دانش جالب توجه و غیربديهی از متن بدون ساختار اشاره دارد. متن کاوی یک حوزه تحقیقاتی جوان و میان‌رشته‌ای بوده و در عین حال که ریشه‌ای عمیق در پردازش زبان طبیعی^۱ دارد، با فرایند کشف دانش با روش‌هایی از بازیابی اطلاعات، داده کاوی، آمار، یادگیری ماشین، استدلال، استخراج اطلاعات، مدیریت دانش، زبان‌شناسی محاسباتی و ... در ارتباط است. این حوزه تمام فعالیت‌هایی را که به‌نوعی به دنبال کسب دانش از متن هستند، شامل می‌گردد. تحلیل داده‌های متنی توسط تکنیک‌های یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روش‌های مرتبط دیگر همگی در زمره مقوله یادگیری متن قرار می‌گیرند. اسناد متنی بر اساس محتوا به یک یا چند طبقه از قبل تعیین شده طبقه‌بندی می‌شوند، که طبقه‌بندی متون یکی از مهم‌ترین مسائل در متن کاوی می‌باشد. مرتب کردن بلادرنگ نام‌های الکترونیکی یا فایل‌ها در سلسله‌مراتبی از پوشه‌ها، تشخیص موضوع متن، جست‌وجوی مفهومی از متون منتشرشده، از جمله کاربردهای مبحث طبقه‌بندی (دسته‌بندی-کلاسه‌بندی) متن است. در بسیاری از موارد، افراد حرفه‌ای آموزش دیده، برای طبقه‌بندی متون جدید به کار گرفته می‌شوند. این فرایند 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 متن کاوی یکی از مهم‌ترین تکنولوژی‌ها در آینده باشد. بنابراین، ایجاد یک چارچوب برای استفاده از متن کاوی در حفاظت از مالکیت معنوی و فکری نویسندگان و بررسی مشکلات و چالش‌های آن، زمینه تحقیقات بعدی در این حوزه را تشکیل می‌دهد. با توجه به حجم وسیع اطلاعات متنی آنلاین در دسترس، ایجاد سیستم‌های متن کاوی به شدت مورد نیاز است. چنین سیستمی می‌تواند در میان گزینه‌های ممکن قرار بگیرد. این جنبه از متن کاوی با توجه به زمینه در حال رشد آن می‌تواند به‌طور مستقل بسیار جذاب باشد و از سایه متن کاوی عمومی خارج شود. این شاخه از متن کاوی برای پیش‌بینی بازار، فارغ از بحث آنالیز احساسات زمینه‌ای، بسیار امیدوارکننده است.

1. Natural Language Processing (NLP)

فهرست منابع

- مصباح، سارا، و مسعود رهگذر. ۱۳۸۸. متن کاوی. تهران: دانشکده فنی، دانشکده مهندسی برق و کامپیوتر دانشگاه تهران.
- هزیر، ابراهیم. ۱۳۹۳. داده کاوی و مفاهیم کاربرد. ورزقان: دانشگاه آزاد ورزقان.
- Al Zamil, M. G. H., and A. B. Can. 2011. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. *Knowledge-Based Systems* 24 (1): 58–65.
- Berry, M. J. A. and G. S. Linoff. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 2nd Edition, John Wiley & Sons, Inc, USA.
- Changa, Y., S. Chena, and C. Liaub. 2008. Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Systems with Applications* 34 (3): 1948–1953.
- Chen, J., H. Huang, S. Tian and Y. Qu. 2009. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3): 5432-5435.
- Elmasri, R. and B. Navathe. 2000. *Fundamentals of Database Systems*. 3rd Edition, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.
- Fayyad, U. M. 1996. Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert: Intelligent Systems and Their Applications*, 11(5): 20-25.
- Fodeh, S., B. Punch, and P. N. Tan. 2011. On ontology-driven document clustering using core semantic features. *Knowledge and Information Systems* 28 (2): 395-421.
- Gunal, S., S., Ergin, M.B, Gulmezoglu and O.N, Gerek. 2006. On feature extraction for spam email detection. *Lecture Notes in Computer Science*, 4105: 635–642.
- Gupta, V., and G. S. Lehal. 2009. A Survey of Text Mining Techniques and Applications. *Emerging Technologies in Web Intelligence* 1 (1): 60-76.
- Hashimi, H., A. Hafez, and H. Mathkour. 2015. Selection criteria for text mining approaches. *Computers in Human Behavior* 51 (B): 729–733.
- Hegna, K. and E. Murtomaa. 2003. Data Mining MARC to find FRBR?. *International Cataloguing and Bibliographic Control*, 32: 52–55.
- Javed, K., S. A. Maruf, and H. Babri. 2015. A two-stage Markov blanket based feature selection algorithm for text classification. *Neurocomputing* 157: 91–104.
- Jiang, S., G. Pang, M. Wu and L. Kuang. 2012. An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, 39: 1503–1509.
- Kanya, N., and S. Geetha. 2007. Information Extraction – a Text mining approach. *Proc. of the IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007)*, pp.1111-1118.
- Karanikas, H., and B. Theodoulidis. 2002. Knowledge discovery in text and text mining software. Technical report, University of Manchester Institute of Science and Technology (UMIST – CRIM), Manchester.
- Kornfein, M., H. Goldfarb. 2007. A Comparison of Classification Techniques for Technical Text Passages. *Proc. Of The World Congress On Engineering*, 2, pp.1072-1075.
- Pestov, V. 2013. Lower bounds on performance of metric tree indexing schemes for exact similarity search in high dimensions. *Algorithmica*, 66 (2): 310-328.
- Rajman, M. 1997. Text Mining, Knowledge extraction from unstructured textual data”. *Proc. of EUROSTAT Conference, Francfort (Deutschland)*, may, 1997

- Data mining Concepts and Techniques: Jiawei Han and Micheline Kamber.
- Raymond, J. Mooney, and Un Yong Nahm. 2005. Text mining with Information Extraction. In *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, W. Daelemans and T. du Plessis and C. Snyman and L. Teck (Eds.), pp. 141-160.
- Salton, G. 1989. *Automatic text processing: The transformation, analysis, and retrieval of Reading*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.
- Saracoglu, R., K. Tutuncu, and N. Allahverdi. 2007. A fuzzy clustering approach for finding similar documents using a novel similarity measure. *Expert Systems with Applications* 33 (3): 600-605.
- _____. 2008. A new approach on search for similar documents with multiple categories using fuzzy clustering. *Expert Systems with Applications* 34 (4): 2545-2554.
- Schumaker, R.P., Y. Zhang, C.-N. Huang and H. Chen. 2012. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53 (3): 458-464.
- Suh, J. H., C. H. Park, and S. H. Jeon. 2010. Applying text and data mining techniques to forecasting the trend of petitions filed to e-People. *Expert Systems with Applications* 37 (10): 7255-7268.
- Tasci, S. and T. Gungor. 2013. Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*, 40(12): 4871-4886.
- 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): 1024-1032.
- Vo, D. H., and C. Y. Ock. 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications* 42 (3): 1684-1698.
- Wei, T., Y. Lu, H. Chang, Q. Zhou and X. Bao. 2015. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications* 42 (4): 2264-2275.
- Weng, S., and Y. Lin. 2003. A study on searching for similar documents based on multiple concepts and distribution of concepts. *Expert Systems with Applications* 25 (3): 355-368.
- Witten, I. H. 2004. Adaptive text mining: Inferring structure from sequences. *J Discrete Algorithms*, 2 (2): 137-159.

اعظم‌السادات پرئی

دارای مدرک تحصیلی کارشناسی ارشد در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک از دانشگاه صنعتی خواجه نصیرالدین طوسی است.

تجارت الکترونیک و کسب‌وکار از جمله علایق پژوهشی وی است.



حجت‌اله 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، کسب‌وکار هوشمند، و محاسبات نرم از جمله علایق پژوهشی وی است.

