Journal of Information Technology Management
ISSN: 2008-5893
EISSN: 2423-5059
Vol. 9, No. 1; PP. 39 - 60
Spring 2017

University of Tehran
Faculty of Management

Textual Data Mining Applications in the Service Chain Knowledge Management of e-Government

Jalal Rezaeenour [1], Mohammadreza Sheikhbahaei [2]

Abstract: Systems related to knowledge management can improve quality and efficiency of knowledge used for decision making process. Approximately 80 percent of corporate information are in textual data formats. That is why text mining is useful and important in service chain knowledge management. For example, one of the most important applications of text mining is in managing on-line source of digital documents and the analysis of internal documents. This research is based on text-based documents and textual information and interviews processed by Grounded theory. In this research clustering techniques were applied at first step. In the second step, Apriori association rules techniques for discovering and extracting the most useful association rules were applied. In other words, integration of datamining techniques was emphasized to improve the accuracy and precision of classification. Using decision tree technique for classification may result in reducing classification precision. But, the proposed method showed a significant improvement in classification precision.

Key words: *e-Government, Knowledge management, Service Chain, Textual datamining.*

1. Associate Prof., Faculty of Engineering & Technology, University of Qom, Qom, Iran
2. M.Sc. in Information Technology Engineering, Faculty of Engineering & Technology, University of Qom, Qom, Iran

Submitted: 10 / March / 2015
Accepted: 25/ February / 2017
Corresponding Author: Jalal Rezaeenour
Email: j.rezaee@qom.ac.ir

مدیریت فناوری اطلاعات
Journal of Information Technology Management
دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۹، شمارهٔ ۱
بهار ۱۳۹۶
صص. ۶۰- ۳۹

کاربردهای داده‌کاوی متنی در حوزهٔ مدیریت دانش زنجیرهٔ خدمات دولت الکترونیکی

جلال رضائی نور[۱]، محمدرضا شیخ بهایی[۲]

چکیده: سیستم‌های مدیریت دانش، کیفیت و بازدهی دانش استفاده‌شده در فرایند تصمیم‌گیری را بهبود می‌بخشند. حدود ۸۰ درصد اطلاعات سازمان‌ها در فرمت‌های متنی هستند؛ به همین علت متن‌کاوی آن هم در حوزهٔ مدیریت دانش زنجیرهٔ خدمات، بسیار مفید و بااهمیت است. به‌طور مثال، یکی از کاربردهای مهم متن‌کاوی به‌منظور مدیریت منابع روی خط اسناد دیجیتال و تجزیه و تحلیل اسناد درون‌سازمانی به‌کار برده می‌شود. پژوهش حاضر به اسناد و مدارک مبتنی بر متن اختصاص دارد که براساس ارسال نظرها، فرم‌های اطلاعات متنی و پرسشنامه‌های مبتنی بر روش تئوری زمینه‌ای تدوین شده است. در نخستین گام تحقیق، تکنیک‌های خوشه‌بندی به اجرا درآمد و در گام دوم، تکنیک قوانین انجمنی Apriori به‌منظور کشف و استخراج مفیدترین قوانین انجمنی اعمال شد. به‌بیانی، بر یکپارچه‌سازی تکنیک‌های داده‌کاوی متنی برای بهبود دقت رده‌بندی تأکید شده است. در کلاسه‌بندی مستندات با استفاده از تکنیک درخت تصمیم به کلاس‌های مربوط به آن، دقت کلاسه‌بندی کاهش یافت، اما استفاده از روش ارائه‌شده در این تحقیق، بهبود شایان توجهی در دقت رده‌بندی ایجاد کرد.

واژه‌های کلیدی: داده‌کاوی متنی، دولت الکترونیکی، زنجیرهٔ خدمات، مدیریت دانش.

۱. دانشیار گروه مهندسی صنایع، دانشکدهٔ فنی و مهندسی، دانشگاه قم، قم، ایران
۲. کارشناس ارشد مهندسی فناوری اطلاعات، دانشکدهٔ فنی و مهندسی، دانشگاه قم، قم، ایران

تاریخ دریافت مقاله: ۱۳۹۳/۱۲/۱۹
تاریخ پذیرش نهایی مقاله: ۱۳۹۵/۱۲/۰۷
نویسندهٔ مسئول مقاله: جلال رضایی نور
E-mail: j.rezaee@qom.ac.ir

مقدمه

براساس گزارش سال ۲۰۱۰ سازمان ملل، دولت الکترونیک از اواخر دههٔ نود میلادی ظهور کرد؛ اما از سال ۱۹۹۶ در تحقیقات دانشگاهی مشاهده شد. به‌رغم اقدامات صورت‌گرفته در سال‌های اخیر، وضعیت توسعهٔ دولت الکترونیک در ایران در حد مطلوبی قرار ندارد (ثقفی، علی‌احمدی، قاضی‌نوری و حورعلی، ۱۳۹۴). پیاده‌سازی موفق دولت الکترونیک، کار ساده و آسانی نیست و با موانع دانش‌افزاری مواجه است (زارعی، ثقفی و زرین، ۱۳۹۲).

در دنیای امروز و در اقتصاد دیجیتالی و به‌خصوص در حوزه‌های خدمات دولت الکترونیکی، اطلاعات زیادی در فرمت متن وجود دارند که می‌توان به‌راحتی آنها را در کلاس‌های از پیش تعریف‌شده طبقه‌بندی و رده‌بندی کرد که البته حدود ۸۰ درصد از اطلاعات در دسترس به‌عنوان اسناد متنی در دسترس است. این اطلاعات اغلب در بیشتر داده‌های توصیفی مانند گزارش‌ها، اطلاعات به‌دست‌آمده از مشتریان، ساخت مستندات کیفیت، تحقیقات میدانی و تجزیه و تحلیل‌های تئوری زمینه‌ای[1]، یادداشت‌ها و غیره هستند. برای بهبود عملکرد و ارائهٔ خدمات باکیفیت‌تر در آینده و ارائهٔ راه حل، باید اطلاعات موجود را به فرمت‌های قابل استفاده تبدیل کرد. تصمیم‌گیرندگان و کارکنان دانشی سازمان و به‌خصوص مدیران دانشی، تصمیمات کسب‌وکار خویش را از طریق کشف الگوهای دانش به‌کار می‌گیرند که سبب کاهش هزینه‌های سربار از خدمات، بهبود کیفیت و مدیریت بهتر می‌شود. هم‌زمان با رشد فزایندهٔ تحولات اقتصادی‌ـ اجتماعی، تأثیر دانش و مدیریت تجربه‌های سازمانی به‌ویژه سازمان‌ها و ارگان‌های دولتی به‌شدت احساس می‌شود (رضائی‌نور، لسانی، زکی‌زاده و مجید، ۱۳۹۳). مدیریت دانش، توانایی سازمان‌ها برای یادگیری از محیط خود و مشارکت‌دادن دانش در فرایندهای کسب‌وکار و تصمیم‌گیری را افزایش می‌دهد (جعفری، رضائی نور و اخوان، ۲۰۰۹).

روش‌های متن‌کاوی مزایایی دارد که سبب مدیریت بهتر منابع دانش و فعالیت‌های مدیریت دانش می‌شود. متن‌کاوی در کشف دانش مفید برای کمک به پردازش اطلاعات و بهبود بهره‌وری کارکنان دانشی سازمان استفاده می‌شود. نتیجهٔ متن‌کاوی، افزایش ارزش افزودهٔ کسب‌وکار به‌منظور تسهیل فرایند تصمیم‌گیری و کاهش هزینه، نسبت به سایر تکنیک‌های پردازش متن است. در اصل برای به‌دست‌آوردن مزایای رقابتی‌تر و بهره‌برداری از اطلاعات چندگانه، روش‌های کشف دانش در نظر گرفته می‌شود.

هدف از پژوهش حاضر، به‌کارگیری داده‌کاوی متنی در حوزهٔ مدیریت دانش زنجیرهٔ خدمات دولت الکترونیکی است و این تحقیق به‌دنبال بهبود زمینه‌های مختلف کسب‌وکار از طریق

---

1. Grounded Theory

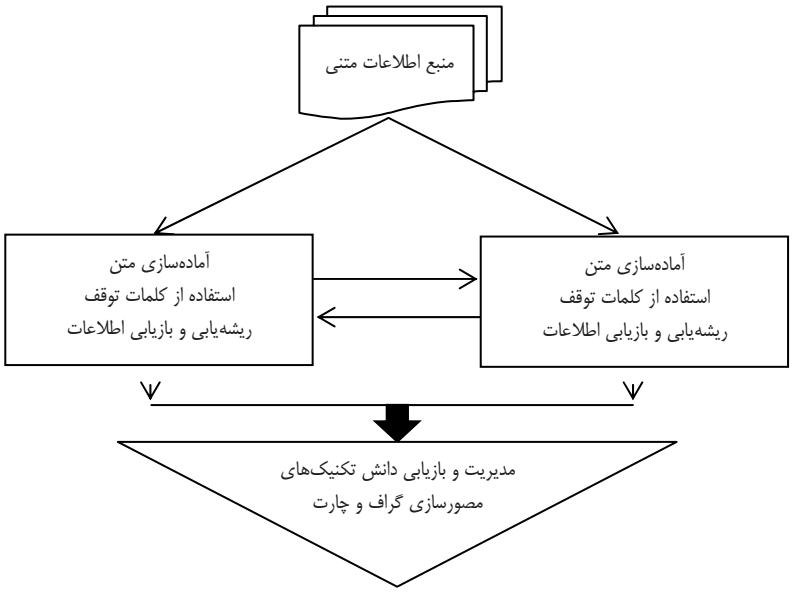
شناسایی دانش مفید از تجربه‌های قبلی و مستندات موجود و تجزیه و تحلیل‌های تئوری زمینه‌ای است. به‌طور مثال، اگر نیازهای مشتریان را بتوان شناسایی و طبقه‌بندی کرد، تصمیم‌گیری‌های بهتر در آینده سبب بهبود سطح رضایت مشتریان خواهد شد. طبقه‌بندی متن رویکرد مهمی برای دست‌یافتن به داده‌های متنی یا اطلاعات در فرایند کلی کشف دانش از پایگاه داده‌های متنی است. متن‌کاوی نویدبخش‌ترین بخش اقتصاد مبتنی بر دانش دیجیتال است که بیشتر برای طبقه‌بندی اسناد متنی به دسته‌های تعریف‌شده یا مجموعه‌ای از کلاس‌ها براساس محتوا استفاده می‌شود. فیلترکردن ایمیل‌ها، مدیریت اسناد و مدارک، شناسایی نیازهای مشتری، بررسی مستندات به‌دست‌آمده از تحقیقات میدانی و نتایج تئوری زمینه‌ای و غیره از کاربردهای دیگر این حوزهٔ دانش است (رضائی‌نور و نظری‌دوست، ۲۰۱۲). بنابراین، استفاده از این فناوری به‌منظور دسترسی به اطلاعات و مدیریت آنها برای بهره‌برداری بهتر در برنامه‌ها و تصمیم‌گیری‌های آینده مفید است. داده‌های تحقیق شامل مجموعه داده‌های کاربردی در حوزهٔ زنجیرهٔ خدمات دولت الکترونیک مربوط به سال ۱۳۹۳ است که از طریق تحقیقات میدانی و تجزیه و تحلیل تئوری زمینه‌ای جمع‌آوری شده‌اند و در این تحقیق اسناد متنی بررسی می‌شوند. ادامهٔ این نوشتار بدین ترتیب ارائه می‌شود: در بخش دوم پیشینه‌ای از روش‌های رده‌بندی متن و گزارش‌های کاری اخیر در حیطهٔ راه حل‌های مدیریت دانش زنجیرهٔ خدمات در حوزهٔ خدمات دولت الکترونیکی مرور می‌شود. بخش سوم، به بحث و تبادل نظر دربارهٔ معماری و متدولوژی‌های ارائه‌شده و نیز روش‌های مختلف گنجانده‌شده در این روش‌ها اختصاص دارد. در بخش چهارم، پیاده‌سازی روش‌های ارائه‌شده براساس داده‌های واقعی در محدودهٔ مدیریت دانش زنجیرهٔ خدمات در حوزهٔ خدمات دولت الکترونیکی در قالب داده‌های به‌دست‌آمده از تحقیقات میدانی تجزیه و تحلیل می‌شود و بخش پنجم نیز به جمع‌بندی و نتیجه‌گیری می‌پردازد.

پیشینهٔ پژوهش

روش‌های طبقه‌بندی متن اولین‌بار در سال ۱۹۵۰ برای طبقه‌بندی اسناد مطرح شد که به‌صورت خودکار صورت می‌پذیرفت. در سال ۱۹۶۰ مقاله‌ای در زمینهٔ طبقه‌بندی خودکار متن منتشر شد. شناسایی اطلاعات مفید از پایگاه داده‌های متنی از طریق تکنیک‌های مختلف داده‌کاوی و به‌طور گسترده در حوزهٔ نرم‌افزارهای مختلف استفاده شد؛ اما برنامه‌های کاربردی در زمینه‌های خدماتی که پایگاه داده‌های اطلاعاتی در حوزه‌های خدماتی دولت الکترونیکی را برای کشف اطلاعات و تبدیل آن به منابع دانش مفید گزارش دهد، بسیار اندک بود.

تکنولوژی داده‌کاوی انعطاف‌پذیری لازم را برای بهره‌برداری اطلاعات از فرمت‌های داده‌های مختلف یا پایگاه داده‌های رابطه‌ای، انبار داده، معاملاتی و...، فراهم می‌کند. پایگاه داده‌های متنی

می‌تواند اطلاعات را در قالب مقاله‌ها، گزارش‌ها، صفحه‌های وب، پیام‌های حاوی یادداشت و... که در انواع بدون ساختار، نیمه‌ساختاریافته و ساختاریافته هستند، نگهداری کند. متن‌کاوی را می‌توان استخراج متن از داده‌های متنی و کشف دانش از پایگاه داده متنی تعریف کرد. فرایند استخراج متن به‌شدت بر روش‌های داده‌کاوی برای کشف دانش مفید متکی است با این تفاوت که در متن‌کاوی، داده‌ها بدون ساختار هستند و چالش‌های بیشتری نسبت به داده‌های ساختاریافته دارند. استخراج متن استاندارد شامل مراحل آماده‌سازی، پردازش و تجزیه و تحلیل متن می‌شود (هان و کمبر، ۲۰۰۰). فرایند استخراج متن به‌عنوان روش‌های تعاملی و تکرارشونده مطابق شکل ۱ است.



شکل ۱. فرایند متن‌کاوی به‌عنوان فرایند تعاملی و تکرارشونده

شکل ۱ روش تکرارشونده برای کشف دانش باارزش را به نمایش گذاشته است که از فرمت‌های داده‌های متنی اقتباس شده و در مدیریت دانش در حوزهٔ خدمات دولت الکترونیکی کاربردهای بسیاری دارد. اطلاعات موجود در فرم‌ها از داده‌های متنی، به‌عنوان ورودی برای آماده‌سازی متن و روش‌های پردازش متن استفاده می‌شود. هر دو مرحلهٔ آماده‌سازی و مراحل پردازش متن باید به‌صورت تعاملی برای پیداکردن الگوهای مفید و قابل فهم در داده‌هایی که قرار است در مرحلهٔ پایانی پیاده‌سازی شوند، یعنی تجزیه و تحلیل متن، به‌صورت مصور نمایش

داده شود. در نهایت، نتایج به‌دست‌آمده در قالب نمودار یا جدول‌هایی نمایش داده می‌شوند (بـری و لینوف، ۲۰۰۴).

نخستین‌بار استفاده از تکنیک‌های طبقه‌بندی برای کاوش و طبقه‌بندی نقص کیفیت خدمات مجموعه داده‌های حوزهٔ دولت الکترونیکی و تجزیه و تحلیل متن و استخراج دانـش (TAKMI) به‌منظور شناسایی علل شکست عرضهٔ خدمات و شناسایی رفتار مشتری مربوط به خدمات خاص در سال ۲۰۰۱ صورت گرفت (ناسوکاوا و ناگانو، ۲۰۰۱). در چند سال اخیر، بیشتر پژوهش‌هـا بـر بررسی تأثیرات مدیریت دانش بـر زنجیـره‌هـای تـأمین غیـر از تجـارت الکترونیکـی دولـت بـه شهروندان (G2C) تمرکز داشتند و در حوزهٔ تجارت الکترونیکی نیز، بیشتر به تجارت الکترونیکی کسب‌وکار بر کسب‌وکار (B2B) و کسب‌وکار بر مشتری (B2C) متمایل بودند، از ایـن رو توجـه شایسته‌ای به این حوزه از دانش نشده است. بنا بـه دلایـل یادشـده، سـابقه و پیشـینهٔ پـژوهش اختصاصی در حوزهٔ مدیریت دانش زنجیرهٔ تأمین تجارت الکترونیکی (G2C) در دسترس نیسـت. این پژوهش، پیشینهٔ حوزه‌های نام‌برده را به‌طور مجزا نقد و بررسی کرده است.

وو در سال ۲۰۰۱ موضوع مشکلات هماهنگی در زنجیرهٔ تأمین را مطرح کرد و بر چگـونگی طراحی سیستم‌های چندعامله برای بهبود اطلاعات و به‌اشتراک‌گذاری دانش تأکیـد داشـت (وو، ۲۰۰۱). سیواکومار و روی (۲۰۰۴) مفهوم افزونگی دانش را به‌عنوان فاکتوری بسیار مهم در ایجاد ارزش زنجیرهٔ تأمین معرفی کردند. هالـت، کچـن و اسـلاتر (۲۰۰۴) بـه بررسـی برتـری برخـی زنجیره‌های تأمین نسبت به زنجیره‌های تأمین دیگر پرداختند و گفتند که هرچه حافظـهٔ زنجیـره بیشتر باشد، تمایل به کسب دانش در آن بیشتر خواهد بود و این امر فعالیت‌های کسـب دانـش و توزیع اطلاعات را شکل می‌دهد. چونگ و مایرز (۲۰۰۸) مشکل اصلی را به‌اشتراک‌گذاری دانـش در شبکه‌های استراتژیک جهانی می‌دانند. این دو محقـق در رونـد تحقیـق بـه بررسـی عـواملی پرداختند که در پایداری اشتراک دانش در زنجیرهٔ تأمین تجارت جهانی کمک می‌کند و مـدیریت مناسب، بازار مناسب، منابع مناسب، اشتراک هویت، سرمایهٔ رابطه‌ای و انعطاف‌پذیری را به‌عنـوان مهم‌ترین عوامل معرفی کردند. مایرز و چونگ (۲۰۰۸) در پژوهش دیگری به بررسـی چگـونگی به‌اشتراک‌گذاری دانش و ایجاد ارزش برای خریداران و تأمین‌کنندگان در زنجیرهٔ تـأمین جهـانی پرداختند و مشکل اصلیِ به‌اشتراک‌گذاری دانش را در تفاوت‌های فرهنگی معرفی کردند. خلفـان، کاشیاپ، لی و ابوت (۲۰۱۰) تسخیر و انتشار دانش را تجزیه و تحلیل کردند. آنها تأثیر گروه‌های همکاری را در یکپارچه‌سازی زنجیرهٔ تأمین و عملکرد برجسته می‌دانند.

طبق بررسی‌های صورت‌گرفته، در حوزهٔ ارتباطی مدیریت دانش زنجیرهٔ تأمین خدمات دولـت الکترونیکی، شکاف‌های بنیادینی در تحقیقات گذشته وجود دارد. هرچند تحقیقات گسـترده‌ای در هریک از بخش‌های مدیریت دانش، مدیریت زنجیرهٔ تأمین، مـدیریت زنجیـرهٔ خـدمات و دولـت

الکترونیکی انجام شده است، دربارهٔ رابطه‌های هریک از مؤلفه‌های یادشده تا کنون مطالعه‌ای صورت نپذیرفته و هیچ چارچوب، ساختار یا مدلی در این زمینه ارائه نشده است. همین امر محققان را بر آن داشت تا شکاف عمیقی که در حوزهٔ ارتباطی میان موضوعات کلیدی مطرح‌شده وجود دارد را با ارائهٔ مدلی مفهومی پوشش دهند. این مدل می‌تواند روند توسعهٔ تحقق دولت الکترونیکی را به‌ویژه در حوزهٔ خدماتی دولت به شهروندان، سرعت بخشد.

روش‌شناسی پژوهش

این تحقیق از لحاظ هدف کاربردی است و از نظر شیوهٔ گردآوری و تحلیل اطلاعات، توصیفی و از نوع توسعه‌ای است؛ زیرا شامل مجموعه روش‌هایی می‌شود که هدف آن توصیف شرایط یا پدیده‌های بررسی شده است و قصد دارد با ترکیب روش‌های خوشه‌بندی و داده‌کاوی متنی، مدیریت دانش زنجیرهٔ خدمات دولت الکترونیکی را توسعه دهد.

مجموعه داده‌های پژوهش براساس مستندات به‌دست‌آمده از مصاحبه‌های باز با کارشناسان یکی از ارگان‌های دولتی پیشگام در حوزهٔ دولت الکترونیکی جمع‌آوری شده است. طبق اصول گراندد تئوری، ارزیابی و تصدیق کدگذاری‌های صورت‌پذیرفته فقط از طریق مصاحبه با افراد مصاحبه‌شونده به‌دست می‌آید و سازوکار خاصی برای تأیید صحت محتوای مصاحبه‌ها وجود ندارد. بنابراین در تحقیق حاضر، پژوهشگران استفاده از روش متن‌کاوی با متد MKTPKS (الرحمن و هاردینگ، ۲۰۱۲) را برای تصدیق کدگذاری مصاحبه‌ها، قبل از تأیید نهایی مصاحبه‌شوندگان در دستور کار خود قرار دادند. این امر سبب بهبود کدگذاری‌ها شد؛ زیرا چنانچه کدگذاری توسط متن‌کاوی تصدیق نشود، مراحل اولیهٔ کدگذاری دوباره توسط پژوهشگران اجرا می‌شود. آنچه پژوهش حاضر را از سایر پژوهش‌ها متمایز می‌کند، ترکیب دو روش گراندد تئوری و متن‌کاوی است.

در این پژوهش روشی پیشنهاد شده است که به تجزیه و تحلیل پایگاه دادهٔ متنی و طبقه‌بندی مطالب می‌پردازد و آنها را در کلاس‌های متفاوت طبقه‌بندی می‌کند. در این تحقیق سه سطح سیستم شامل ویژگی‌های مختلف برای متن‌کاوی به این ترتیب پیشنهاد شده است: پردازش دانش و ذخیره‌سازی در سطح اول، واحد پالایش در سطح دوم و کاربرد و طبقه‌بندی دانش در سطح سوم. جریان اطلاعات و دانش از بخش‌های مختلف سیستم، شامل تولید خلاصهٔ متن از MKTPKS و طبقه‌بندی اسناد موجود براساس MKTPKS می‌شود. شرح مفصلی از دنبالهٔ فعالیت‌ها در زیر آمده است.

پردازش اطلاعات و داده

در نخستین گام بررسی و تحلیل فرمت داده‌های متنی، باید اطلاعات متنی به‌صورت مستندات متنی در دسترس باشند. معمولاً این مستندات توسط افراد خبره (کارشناسان) در سازمان بررسی می‌شود و تصمیمات لازم به‌واسطهٔ این افراد اتخاذ می‌گردد که ممکن است مفید یا غیرمفید باشد. این نوع بررسی گران است؛ چراکه به زمان و تلاش کارشناسان حوزه‌های مختلف نیاز دارد. برای آغاز فرایند طبقه‌بندی خودکار، متن داده‌های ورودی باید در فرمت مناسبی برای استفاده از تکنیک‌های داده‌کاوی متنی مختلف آماده شوند که شامل حذف کلمات توقف و توابع ریشه‌یابی لغات ساده می‌شود. برای رسیدن به هدف (ایجاد داده‌های کاربردی) به‌منظور اعمال تکنیک‌های مختلف داده‌کاوی، باید مراحل زیر را طی کرد.

گام اول حذف اطلاعات غیرضروری موجود در فرم‌های توقف کلمه مانند افعال، حروف ربط، اتصالات قطع، ضمایر و غیره است. کلماتی که حذف می‌شوند در تفسیر معنای متن (دارای تأثیر کمتر) چندان مفید نیستند. ریشه‌یابی به‌عنوان روند آمیختن کلمات به ساقهٔ اصلی، پایه یا ریشهٔ آنها تعریف شده است. به‌طور مثال، ریشهٔ رسیدن و می‌رسد، رساندن کلمهٔ رسید است. این روش به گرفتن اطلاعات کل فضای حمل (دامنهٔ فضای اطلاعات متنی)، کاهش ابعاد داده و در نهایت طبقه‌بندی داده‌ها کمک می‌کند. گام بعدی، نمایش داده‌های متنی به فرم ماتریس است که در آن باید هر بردار ردیف شامل شرایط و هر بردار ستون شامل کد شناسایی مربوط به سند (شناسه، ID) باشد. برای کاهش تأثیر فقدان اطلاعات کلیدی در این مرحله از نمایش داده‌های متنی، رویکرد BOW به‌کار رفته است که از فضای کل اطلاعات برای تحلیل استفاده می‌کند. این روش مستقل از ساختار متن است و هر کلمه به‌عنوان نهادی مستقل حاوی برخی اطلاعات در نظر گرفته می‌شود.

سطح اول: واحد پردازش و ذخیره‌سازی دانش

این بخش به تجزیه و تحلیل داده‌های متنی کمک می‌کند تا با استفاده از الگوریتم‌های داده‌کاوی مختلف، داده‌هایی حاصل شود که نمایندهٔ سودمندی از کلمه‌ها و عبارت‌های تعریف‌شده در متن باشند. روش‌های نمایش داده‌های مختلف که در این سطح استفاده می‌شوند عبارت‌اند از: ۱. بسامد کلمه (TF)[1] و ۲. بسامد معکوس اسناد (IDF)[2]. انتخاب نماینده از داده‌ها باید از طریق آزمایش‌های گسترده و با درنظرگرفتن کل فضای اطلاعات باشد که کل موارد از

1. Term Frequency
2. Inverse Document Frequency

طریق ماتریس صورت می‌پذیرد. در حال حاضر، این پژوهش بر استفاده از تکنیک‌های خوشه‌بندی برای افرازبندی داده‌ها به زیرمجموعه‌های مفید از اطلاعات در هر خوشه تمرکز کرده است.

خوشه‌بندی

خوشه‌بندی پردازشی است که با استفاده از معیارهای فیزیکی یا کمی، اطلاعات و داده‌ها را به گروه‌هایی با خواص مشابه دسته‌بندی می‌کند. این معیار کیفی می‌تواند براساس مرکز ثقل خوشه، عمل خوشه‌بندی را انجام دهد. تکنیک دیگر در یافتن شبیه‌ترین اعضا از طریق Terms (مجموعهٔ عبارت‌ها) و روش K-Means است تا اولین سطح دانش را از بین ارتباطات طبیعی کشف کند. یکی از معیارهای مهم خوشه‌بند، محاسبهٔ فاصلهٔ اقلیدسی است که از این رابطه در روش خوشه‌بندی K-Means نیز می‌توان استفاده کرد.

$$D(x,y) = \sqrt{\sum_{i=1}^{m}(X_i - Y_i)^2}$$ رابطهٔ ۱)

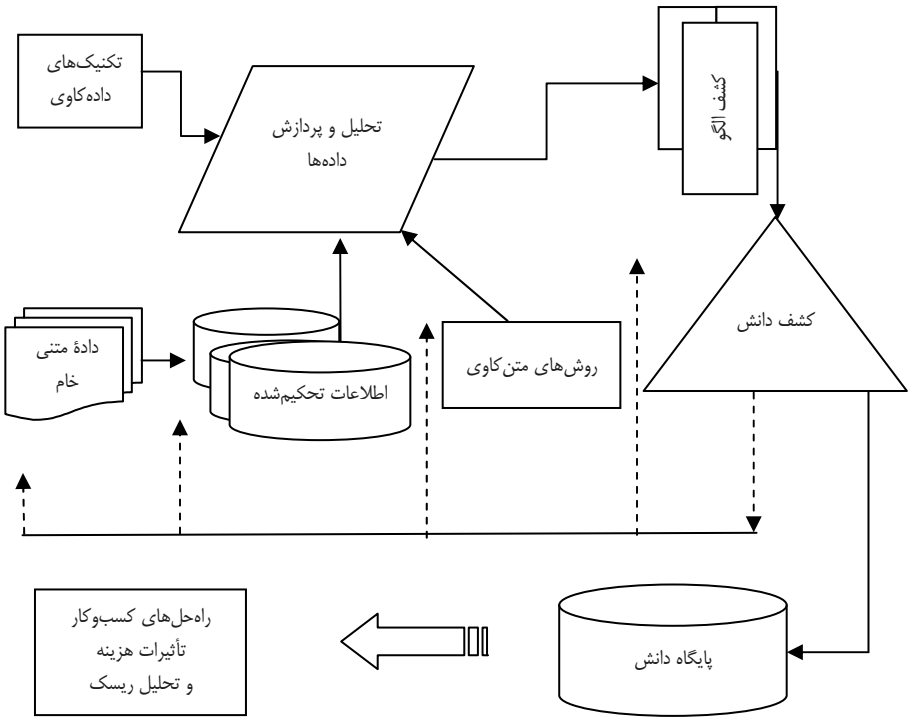
تشکیل پایگاه دادهٔ رابطه‌ای

خروجی کاربرد خوشه‌بندی K-Means باید به‌صورت فرمت‌های قابل استفاده در جدول‌های رابطه‌ای مختلف ذخیره‌سازی شود. این جدول‌ها شامل ستون‌هایی به همراه شناسهٔ خوشه‌بندی هستند، خوشه‌بندی و برچسب‌گذاری خوشه‌ها بیشتر در پردازش‌های هرس اطلاعات کلیدی یا کشف دانش استفاده می‌شوند، این عملکرد به ذخیره‌سازی و مدیریت اطلاعات برای بیشتر تحلیل‌ها کمک می‌کند.

سطح دوم: واحد پالایش دانش

ورودی این واحد در قالب جدول‌های رابطه‌ای است که در آن اسناد به‌عنوان معاملات یا تراکنش‌های انجام‌شده و شرایط به‌عنوان اقلام در نظر گرفته می‌شوند. این فرایند با پالایش اطلاعات و دانش کلیدی به همراه تولید NKTPKS و از طریق کاربرد قوانین انجمنی APRIORI برای کاوش انجام شده است (آگراوال، ایمی‌لینسکی و سوامی، ۱۹۹۳). در ساخت MKTPKS بخش اساسی و ضروری استفاده از تحلیل داده‌ها برای رده‌بندی مستندات متنی است. ساخت MKTPKS بر یافتن قوانین انجمنی ارجحیت دارد، به‌دلیل آنکه شناسایی بیش از حد قوانین موجب ازدیاد جمعیت در پایگاه دانش می‌شود. علاوه‌بر این، MKTPKS می‌تواند به کشف روابط باارزش‌تر در شرایط تعریف‌شده در متن کمک کند. این اعمال، یافتن ارتباطات در

میان مفاهیم مختلف تعریف‌شده در مستندات متنی را آسان می‌کند. نگاشت MKTPKSهای کشف‌شده به مجموعه‌های ویژه از مستندات، به شناسایی مجموعه مستندات حاوی اطلاعات خوب و بد کمک می‌کند.

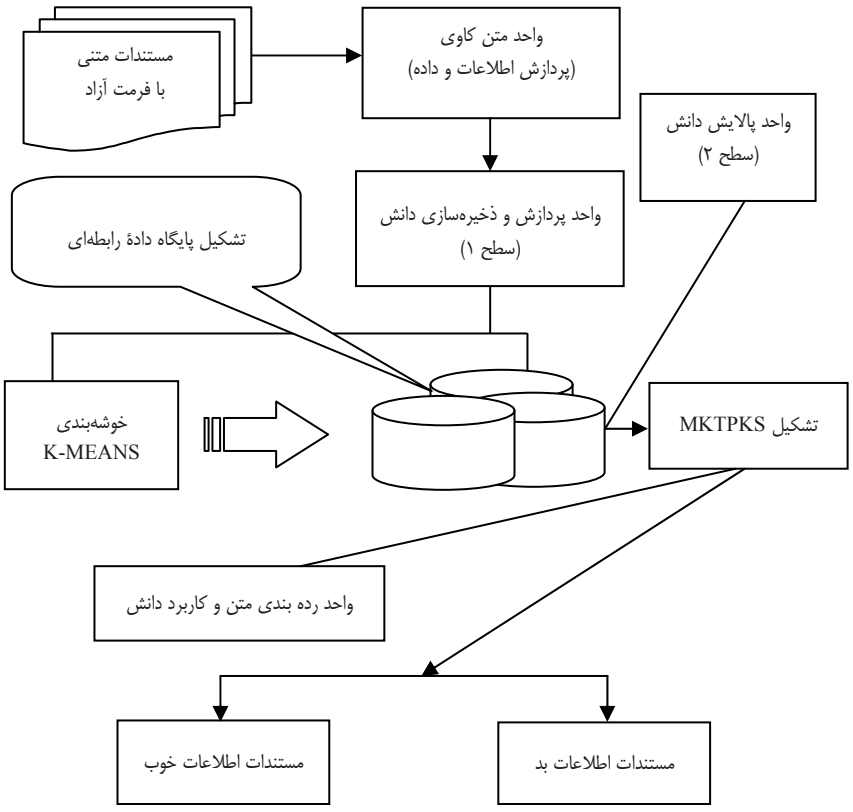


شکل ۲. داده‌کاوی متنی برای کشف دانش در پایین‌دست و راه حل‌های مدیریتی

سطح سوم: واحد رده‌بندی متن و کاربرد دانش

داده‌ها اصولاً به‌صورت پایگاه دادهٔ نیمه‌ساختاریافته (نه کاملاً ساختاریافته و نه بدون ساختار) در طبیعت ذخیره می‌شوند. برای رده‌بندی داده‌های متنی به کلاس‌های از پیش تعریف‌شده، لازم و ضروری است مجموعهٔ مد نظر به‌صورت دستی به کلاس‌های متفاوت برای آزمون و صحت رده‌بندی افراز شود. این افراز به کمک کارشناسان دامنه انجام می‌شود. خصیصه‌های طبقه‌بندی، مجموعه‌ای از ویژگی‌های کلاس یا متغیر هدف است. در پژوهش حاضر این کار به‌کمک کارشناسان حوزه که درک درستی از زمینه‌های داده‌های متنی و معنای شرایط تعریف‌شده در اسناد متنی دارند، صورت پذیرفته است. در واحد سطح سوم، رده‌بندی‌های متفاوتی استفاده

می‌شود که برای مطالعهٔ تأثیرات شرایط رده‌بندی داده‌های متنی به دو کلاس مختلف، می‌تواند ساخت رده‌بندی مستندات را با دقت بیشتری بهبود دهد. رده‌بندی‌های خاص مطرح‌شده در پژوهش حاضر عبارت‌اند از: ۱. درخت تصمیم‌گیری (C۴/۵)، ۲. نزدیک‌ترین همسایه (K-NN)، ۳. بیز ساده[1] و ۴. ماشین‌های پشتیبان بردار (SVMs). نتیجهٔ این آزمون، رده‌بندی‌های متفاوتی است که از تغییرپذیری مبتنی بر انتخاب متغیرهای اطلاعاتی روی محدودهٔ معیارهای فاصله‌ای متفاوت، از معیار فاصلهٔ اقلیدسی ساده تا متدهای مبتنی بر هستهٔ روش ارائه‌شده در مقالهٔ یادشده ایجاد شده است. هدف از رده‌بندی، اعتبارسنجی فرضیهٔ مبتنی بر روش ارائه‌شده روی MKTPKS برای بهبود صحت رده‌بندی بر الگوریتم‌های رده‌بندی است.



شکل ۳. سیستم رده‌بندی متن و مدیریت دانش مبتنی MKTPKS

1. Naïve Bayes

تحلیل و طراحی

الگوریتم تحلیل درخت تصمیم بیشتر برای مشکلات رده‌بندی و فرایند ساخت شروع درخت تصمیم به‌وسیلهٔ انتخاب گرهٔ تصمیم و جداسازی آن به زیرگره و برگ استفاده می‌شود. الگوریتم درخت تصمیم C۴/۵، توسعه‌یافتهٔ الگوریتم ID۳ است. این الگوریتم مبتنی بر ساخت درخت تصمیم و جداسازی به‌وسیلهٔ گره‌های تصمیم از طریق انتخاب جداسازهای بهینه و ادامهٔ جست‌وجو در حد امکان است. برای استفاده از مفاهیم بهرهٔ اطلاعاتی[1] و کاهش آنتروپی به‌منظور کسب تقسیم مطلوب از رابطه‌های زیر استفاده می‌کنیم. فرض کنید متغیر X ارزش K احتمال ممکن $p_1, p_2, \ldots, p_k$ را داشته باشد، آن‌گاه آنتروپی X از رابطهٔ ۲ محاسبه می‌شود.

$$H(X) = -\sum P_j log\ 2(P_j)$$ رابطهٔ ۲)

برای محاسبهٔ میانگین اطلاعات می‌توان از مجموع وزن‌دار آنتروپی برای زیرمجموعه‌های فردی (مجزا) استفاده کرد.

$$Hs\ (X) = \sum i = 1k\ P_i\ H_s(T_i)$$ رابطهٔ ۳)

در این رابطه، P_i نشان‌دهندهٔ نسبت رکوردها در زیرمجموعه‌های i است. برای محاسبهٔ بهره اطلاعاتی از رابطهٔ ۴ استفاده می‌شود.

$$information\ gain\ IG(S) = H(T) - Hs(T)$$ رابطهٔ ۴)

الگوریتم نزدیک‌ترین همسایه

الگوریتم نزدیک‌ترین همسایه روشی است که با استفاده از معیار فاصله برای رده‌بندی داده‌ها به‌کار می‌رود. الگوریتم نزدیک‌ترین همسایه از طریق نمونه‌های آموزشی کار می‌کند. در این روش، مجموعه نه‌تنها شامل داده‌ها می‌شود، بلکه طبقه‌بندی مد نظر برای هریک از موارد صورت می‌پذیرد. در واقع می‌توان گفت داده‌های آموزش مدل می‌شوند. در اصل الگوریتم K-NN کمترین فاصله از نمونه‌های ورودی جدید را در نمونه‌های آموزشی پیدا می‌کند. براساس معیار یادشده، هر نمونهٔ ورودی جدید در کلاس مربوط به خود قرار می‌گیرد. معمول‌ترین تابع محاسبهٔ فاصله، فاصلهٔ اقلیدسی است که در رابطهٔ ۱ به آن اشاره شده است.

---

1. Information Gain

الگوریتم بیز ساده

الگوریتم بیز ساده روش آسان و خوب شناخته‌شده‌ای برای رده‌بندی است که به‌منظور حل مشکلات حوزهٔ عملی استفاده می‌شود. رده‌بندی بیز ساده به‌منظور یافتن احتمالات مشترک از کلمه‌ها و کلاس‌ها در مجموعهٔ رکوردها به‌کار می‌رود. این رویکرد مبتنی بر تئوری بیز ساده است. احتمال کلاس c در مستند d_j از رابطهٔ زیر به‌دست می‌آید. در این الگوریتم فرض بر این است که طبقات مستقل از یکدیگرند که با عنوان «استقلال مشروط کلاس» مطرح می‌شود.

رابطهٔ ۵)
$$P(c\ /d_j) = \frac{P(d_j\ /\ c)P(c)}{P(d_j)}$$

$$j = 1, 2, 3, \dots, m$$

الگوریتم ماشین‌های پشتیبان بردار

الگوریتم SVM اولین‌بار سال ۱۹۶۰ در روسیه توسعه داده شد. این الگوریتم رده‌بندی غیرخطی است که از روش‌های غیرخطی بهره می‌برد و داده‌ها را از فضای ورودی یا فضای پارامتری به فضای ویژگی‌هایی با ابعاد زیاد نگاشت می‌کند. هدف این الگوریتم، انتخاب ابرصفحهٔ جداساز بهینه برای حداکثرسازی حاشیهٔ بین دو کلاس است. برای حل مشکل طبقه‌بندی دودویی که در آن W_1 و W_2 نشان‌دهندهٔ دو کلاس در یک مجموعهٔ داده‌های آموزشی هستند، مجموعهٔ $X=\{x_1, x_2, \dots, xn\}$ به همراه برچسب کلاس ارائه شده است. ابرصفحه‌ای که داده‌های مجزا را به دو کلاس طبقه‌بندی می‌کند، به‌شرح زیر است.

رابطهٔ ۶)
$$f(x) = sgn(< w, x > + b)$$

در این رابطه، w بردار ضریب و b میزان تمایل به یک طرف (چولگی) ابرصفحه و sgn مخفف تابع دوقطبی است. مشکل بهینه‌سازی که به تولید ابرصفحه منجر می‌شود را به‌صورت زیر می‌نویسند:

رابطهٔ ۷)
$$\text{Minimize}_{w,\ x}\ 1/2||w||2$$

رابطهٔ ۸)
$$Yi(< w, xi > +b) >= 1\ ,\quad \text{for i} =\ 1,2, \dots, N$$

بزرگ‌ترشدن حاشیه سبب بهترشدن توانایی تعمیم انتظار می‌شود.

یافته‌های پژوهش

بررسی و بازبینی داده‌های پروژه و تعریف اطلاعات خوب و بد

برای آزمایش روش ارائه‌شده، از یک مجموعه دادهٔ نمونه که از حوزهٔ خدمات دولت الکترونیکی جمع‌آوری شده‌اند، استفاده شده است. این داده‌ها از طریق تحقیقات میدانی به‌دست آمده‌اند که هریک از آنها شامل اطلاعات مربوط به نظرها و پیشنهادهای کارشناسان و متخصصان از انواع خدمات ارائه‌شده در حوزهٔ دولت الکترونیکی و همچنین تجزیه و تحلیل اطلاعات یادشده براساس روش تئوری زمینه‌ای است. داده‌ها به دو کلاس مختلف از مستندات که هریک حاوی اطلاعات خوب و بد هستند، دسته‌بندی شدند. این فعالیت با مطالعهٔ دقیق از طریق بازبینی به‌کمک کارشناسان هر حوزه برای کسب اطمینان از این است که معانی هریک از آیتم‌های متن که تحت زمان و هزینه دسته‌بندی شدند، به‌درستی صورت پذیرفته است. در روش پژوهش حاضر پس از بازبینی، مجموعه به دو کلاس متفاوت گروه‌بندی شدند.

رده‌بندی مستندات به مستندات حاوی اطلاعات خوب و بد

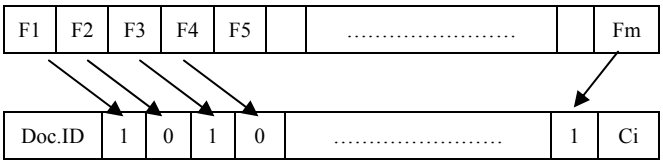
درخت تصمیم (C ۴/۵)، نزدیک‌ترین همسایه، بیز ساده و ماشین‌های پشتیبان بردار که در مراحل قبل دربارهٔ آنها توضیح داده شد، برای رده‌بندی داده‌های متنی در این تحقیق استفاده شدند. الگوریتم روی مجموعهٔ داده‌های منتقل‌شده از مجموعهٔ ویژگی‌های منتخب و خصیصه‌ها اعمال می‌شود و به نوعی کاربرد ترکیبی[1] از سطح ۱ (واحد ذخیره‌سازی و پردازش دانش) و سطح ۲ (واحد پالایش دانش) پیدا می‌کند.

کاربرد واحد متن‌کاوی (فرایند اطلاعات و داده)

نتایج کاربرد این رویکرد ترکیبی، ساخت مجموعه MKTPKS است. این عمل با استفاده از رویکرد خوشه‌بندی صورت می‌پذیرد. در نتیجه، فضای وجود داشتن یا نداشتن برای عبارت‌های کلیدی در اسناد ساخته می‌شود. هر نمایندهٔ بردار از اسناد با استفاده از مجموعهٔ MKTPKS 3Term Sets انجام می‌پذیرد. شکل ۴ رابطهٔ فهرست عبارت‌های کلیدی تشکیل‌شده و وجود آنها در اسناد و نمایندهٔ مربوط به کلاس‌ها را نشان می‌دهد (اگر نامزد مجموعه باشد، مقدار ۱ و در غیر این‌صورت، مقدار صفر به خود اختصاص داده است) که در آن Ci معرف برچسب کلاس با توجه به داده‌های آموزشی و Fm نمایندهٔ مجموعهٔ MKTPKS 3TermSets است.

---

1. Hybrid



شکل ٤. نمایش نامزدهای مجموعهٔ termset

بنابراین، پس از انتقال کل فضای مستندات بـه فـرم MKTPKS 3TermSets، اطلاعـات کلیدی به‌دقت به فرم داده‌های متنی تعریف می‌شوند. ماتریس جدیـد در فعالیـت‌هـای رده‌بنـدی برای تفکیک مستندات به دو کلاس متفاوت کارایی دارد. در این بخش برای تحلیل داده‌ها، باید الگوریتم‌های داده‌کاوی متفاوتی را بررسی کرد. داده‌ها باید به فرمت مناسب تبدیل شوند و بـرای پردازش بیشتر می‌توان آنها را در فایل متنی تثبیت کرد. تجزیهٔ متن با استفاده از یـک کـد جـاوا برای شمارش شرایط انجام گرفت؛ به‌طوری که بسامد متناظر با آنها در قالب یک ماتریس بسامد به‌دست آمد. کلمات توقف نیز از داده‌های متن حذف شدند و نوعی روش ریشـه‌یـابی سـاده نیـز اعمال گردید. روش داده‌کاوی متنی به کاهش ابعاد داده با حفظ فضای اطلاعـات مفیـد و بـدون ازدست‌دادن اطلاعات کلیدی کمک می‌کند. فایل‌های یادشده با پسوند (csv.) یـا متنـی ذخیـره می‌شوند (فایل‌های جداشده با کاما) که می‌توان برای پردازش دانش و پایگاه دادهٔ رابطه‌ای از آن استفاده کرد.

کاربرد واحد ذخیره‌سازی و پردازش دانش (سطح ۱)

واحد سطح اول

فایل csv یا متنی ذخیره‌شده در نرم‌افزار Weka یـا RapidMiner بارگـذاری مـی‌شـود کـه از طریق آن با به‌کارگیری تکنیک‌های خوشه‌بندی متفاوت می‌توان به درک اولیه و کشف و تسخیر عبارت‌های کلمهٔ کلیدی دانش دست یافت. الگـوریتم خوشـه‌بنـدی k-means بـرای جداسـازی فضای اطلاعاتی ورودی به اعداد و زیرفضاها اعمال می‌شود. آزمایش‌های زیادی برای پیداکردن تعداد مناسبی خوشه به‌منظور کاهش اثر از دست‌دادن اطلاعات اجرا مـی‌شـود (ویـتن و فرانـک، ۲۰۰۰). در نهایت، شش خوشه به‌عنوان بهترین راه‌حل برای کار این پـژوهش انتخـاب شـد کـه مشابه نتیجهٔ کاوش ندیم‌الرحمن (۲۰۱۲) است. به‌کارگیری تکنیک خوشه‌بندی تسخیر اطلاعـات کلیدی با اولین سطح از دانش، در یافتن عبارت‌های کلمهٔ کلیدی کمک می‌کند. فقط سه خوشـه در طول روند این تحقیق انتخاب شدند. اطلاعات کلیدی تسخیرشده در خوشه‌هـای مختلـف بـه مجموعه‌های مختلف از اطلاعات موجود در هر سند اشاره دارد، بنـابراین تفسـیر ایـن اطلاعـات کلیدی و اینکه به‌طور دقیق اسناد حاوی اطلاعات خوب یا بد هستند، دشوار است.

کاربرد واحد پالایش دانش (سطح ۲)

در این گام از قوانین انجمنی الگوریتم APRIORI برای کاوش MKTPKS استفاده می‌شود. ورودی در قالب جدول‌های رابطه‌ای است که در آن اسناد نشان‌دهندهٔ معاملات و تراکنش‌ها به‌منزلهٔ اقلام در نظر گرفته می‌شوند. خروجی به شکل MKTPKS 3-Termsets خواهد بود. در این سطح نمایندگانی که برای شناسایی به‌کار می‌روند، انتخاب می‌شوند. همکاری این عبارت‌ها برای تولید MKTPKS 3-TermSets به‌عنوان نهاد واحد، برای نمایش موضوعات کلیدی بحث‌شده در اسناد پایگاه دادهٔ متنی وارد شده است. با توجه به مثال قبل، در این سطح برای پیداکردن موضوعات کلیدی بحث‌شده در پایگاه دادهٔ متنی دچار مشکل می‌شویم و رده‌بندی اسناد به اطلاعات خوب و بد به‌درستی و با صحت و دقت صورت نمی‌پذیرد. به‌منظور غلبه بر این مشکل، روند استخراج اطلاعات مفید در اسناد مدون یادشده به تصفیهٔ بیشتر نیاز دارد؛ این پالایش از اطلاعات کلیدی یا کشف دانش در سطح ۱، از طریق کاوش قوانین انجمنی APRIORI صورت می‌پذیرد. شایان ذکر است قبل از کاربرد این کاوش، اطلاعات کلیدی تسخیرشده از عبارت‌های کلیدی به ذخیره‌سازی نیاز خواهند داشت. این فعالیت یک پایگاه دادهٔ رابطه‌ای با استفاده از جدول‌های حاوی برچسب‌های خوشه‌ای، اصطلاحات کلیدی شناسایی‌شده و کد شناسایی اسناد (شناسه) ایجاد می‌کند. این جدول‌های رابطه‌ای به فرم MKTPKS استفاده می‌شوند که هم سبب کاهش تعداد ابعاد در فضای ویژگی می‌شوند و هم به اعتبارسنجی فرضیه برای دستیابی به دقت بیشتر در رده‌بندی کمک می‌کند.

نتایج و کاربردهای واحد رده‌بندی و بهره‌برداری از دانش (سطح ۳)

این بخش نشان می‌دهد روش‌های استفاده‌شده برای رده‌بندی داده‌های متنی، به دو کلاس تفکیک می‌شود. نتایج به‌دست‌آمده از کاربرد سطح ۲ (واحد پالایش دانش) به فرم مدل ماتریس جدید مبتنی بر MKTPKS 3-termSets در دسترس خواهد بود که در بخش‌های قبل بررسی شد. ماتریس جدید در نرم‌افزار Weka یا RapidMiner بارگذاری شده و چهار رده‌بند مختلف در رده‌بندی کلاس‌های مربوطه به‌کار برده می‌شود. مجموعهٔ متغیر هدف بدین‌منظور به‌کار برده شده است که متغیر کلاس، تعداد اسناد حاوی اطلاعات خوب یا بد را مشخص کند. هدف از آموزش سیستم و تعیین میزان، رده‌بندی درست و نادرست است. نتایج به‌دست‌آمده از رده‌بندی‌های مختلف در MKTPKS 3-termsets براساس مدل ماتریس است که به مقایسهٔ دقت رده‌بندی در برابر مدل عبارت‌های ساده می‌پردازد. گوشه‌ای از طبقه‌بندی داده‌های متنی از پایگاه دادهٔ متنی با استفاده از درخت تصمیم‌گیری (الگوریتم C۴/۵ یا J۴۸) براساس بازنمایی و

بر پایهٔ عبارت‌های ساده تشکیل شده است. نمودار درختی تشکیل‌شده در نرم‌افزار Rapid Miner نشان می‌دهد هر گره به زیرگره‌ها یا برگ‌هایی تقسیم شده است که در آن، مستندات اطلاعات به گروه خوب و بد رده‌بندی می‌شوند. هر گره‌ای که Information Gain آن حداکثر (Maximum) باشد، به زیرنودهایی تفکیک می‌شود. هر گرهٔ برگ نشان‌دهندهٔ رده‌بندی نهایی اطلاعات به اسناد حاوی اطلاعات خوب یا بد دربارهٔ یک پروژه در پایگاه دادهٔ متنی است. رده‌بندی داده‌ها براساس خروجی سیستم MKTPKS 3-termsets صورت می‌پذیرد. فضای اطلاعات به دو کلاس از مستندات اطلاعاتی خوب و بد دسته‌بندی شده است که با انتخاب گره‌ها و زیرگره‌های اطلاعاتی رده‌بندی می‌شوند. برگ شاخه نشان‌دهندهٔ تعدادی از مستندات رده‌بندی‌شده به‌عنوان خوب و بد است. بنابراین، فرایند تشکیل درخت تصمیم‌گیری تا زمانی ادامه می‌یابد که فضای سند از اطلاعات به‌طور کامل به دو دستهٔ مختلف رده‌بندی شود. سناریوی تحقیق حاضر با در نظرگرفتن عبارات زیر تعریف شده است: «مطلوب مشتری تعداد بسیار کم دستورالعمل و تغییرات است» که می‌تواند به کارکنان برای اجرای هموار (نرم) پروژه کمک کند و آن را در زمان مقرر به پایان رساند. زمان اتمام پروژه یا زمان ارائهٔ خدمات می‌تواند نوعی شاخص عملکرد کلیدی خوب باشد که اگر پروژه یا خدمت در آن (زمان مقرر) به اتمام برسد، موجب رضایتمندی مشتری می‌شود که این خود سبب حفظ مشتری توسط شرکت خواهد شد. از این رو، اگر تصمیم‌گیرندگان می‌توانند به‌آسانی رده‌بندی داده‌های متنی را براساس مستندات حاوی اطلاعات خوب یا بد انجام دهند، به این دلیل است که برای ساخت پروژه‌های جدید یا ارائهٔ خدمات نوین در آینده، تصمیمات برتر و بهتری گرفته شود. این عمل در نهایت به بهبود کسب‌وکار از طریق شناسایی راه‌های حفظ مشتریان خود با توجه به تجربه‌های به‌دست‌آمده در گزارش‌های قبل، کمک می‌کند. یکی از اهداف این پژوهش رده‌بندی با دقت داده‌های متنی است (کاهش میزان غیر رده‌بندی). برای رسیدن به این هدف و مدیریت بهتر منابع دانش، ماتریس‌های مختلف با ساختار داده‌ای متنی در نظر گرفته شده‌اند. دقت رده‌بندی با استفاده از اطلاعات رده‌بندی مستندات حاوی اطلاعات خوب و بد محاسبه می‌شود.

ارزیابی سیستم پیشنهادی

ارزیابی نهایی از روش ارائه‌شده، براساس متوسط F-Measure است که به‌عنوان میانگین هارمونیک بازخوانی[1] و دقت[2] تعریف‌شده، ساخته شده است. دلیل انتخاب F-Measure این است که به هر دو مفهوم دقت و بازخوانی توجه شده است (میائو، دوان، ژانگ و جیائو، ۲۰۰۹). ارزیابی

1. Recall
2. Precision

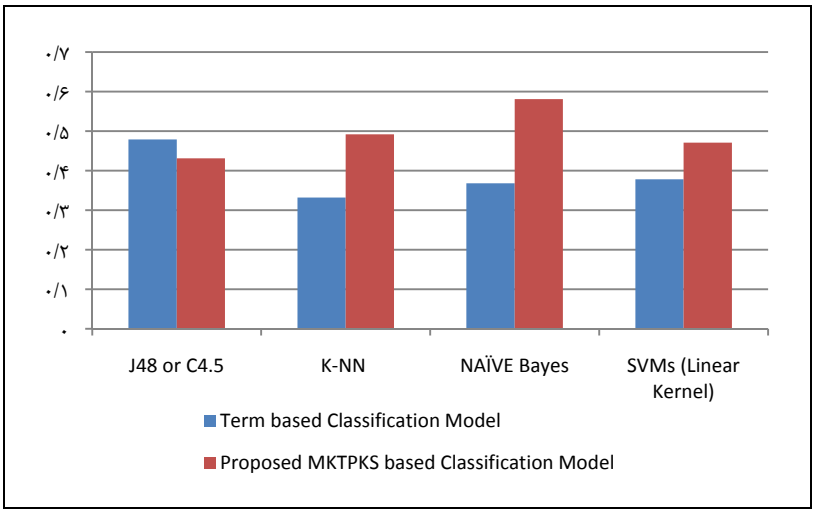
سیستم با ۱۰ برابر کردن روش اعتبارسنجی‌شده در Weka و RapidMiner بررسی شد. تنظیم هر الگوریتم برای رسیدن به یک سطح معین، متفاوت است و این عمل باید به‌صورتی انجام گیرد که دقت به بهترین شکل ممکن رعایت شود. با رده‌بندی بیز ساده، بهترین دقت رده‌بندی با حفظ تنظیمات بدون تغییر به‌دست می‌آید. در رده‌بندی‌های دیگر، باید تنظیمات پارامترهای بهینه انتخاب شود. برای الگوریتم درخت تصمیم‌گیری (الگوریتم C۴/۵ یا J۴۸) نسبت هسته‌های مختلف استفاده می‌شود تا بهترین نتایج با استفاده از نسبت هسته از ۱۰ به‌دست آید. به‌طور مشابه برای K-NN تنظیمات بهینه با K = ۱۰ در نظر گرفته می‌شود و یک هستهٔ خطی بهترین نتایج را براساس مدل رده‌بندی مبتنی بر SVMs در اختیار ما قرار می‌دهد. جدول ۱ مقایسهٔ عملکرد رده‌بندی‌های مختلف را نمایش می‌دهد.

جدول ۱. مقایسهٔ عملکرد طبقه‌بندی‌های مختلف

مدل رده‌بندی	مدل رده‌بندی مبتنی بر واژه (F-measure)	مدل رده‌بندی مبتنی بر MKTPKS پیشنهادی (F-measure)
Decision trees (J۴۸ or C۴/۵)	۰/۴۷۹	۰/۴۳۱
K-NN (k=۱۰)	۰/۳۳۲	۰/۴۹۲
NAÏVE Bayes	۰/۳۶۸	۰/۵۸۱
SVMs (Linear Kernel)	۰/۳۷۸	۰/۴۷۱

جدول ۱ دقت مدل‌های رده‌بندی مبتنی بر عبارت‌های ساده و MKTPKS 3-Termsets را نمایش می‌دهد. دقت رده‌بندی براساس مدل‌های K-NN، Naïve Bayes و SVM (هسته‌ای خطی)، بهتر از مدل مبتنی بر عبارت‌های ساده است. شکل ۵ نشان می‌دهد دقت رده‌بندی درخت تصمیم‌گیری (C۴/۵) با استفاده از مدل ارائه‌شدهٔ مبتنی بر MKTPKS 3-Termsets، از مدل رده‌بندی مبتنی بر عبارت‌های ساده کمتر است. با این‌حال، دقت رده‌بندی‌های دیگر (نزدیک‌ترین همسایه، بیز ساده و ماشین پشتیبان بردار) با استفاده از متدولوژی ارائه‌شده نسبت به مدل رده‌بندی مبتنی بر عبارت‌های ساده، بهبود یافته است. از این رو، می‌توان نتیجه گرفت که اگر روش پیشنهادی را در رده‌بندی داده‌ها استفاده کنیم، دقت بیشتری برای رده‌بندی داده‌ها به دو کلاس مختلف به دست می‌آید که شامل مستندات حاوی اطلاعات خوب و بد هستند. بهترین دانش کاربردی که در این مقاله ارائه شده است، روش‌های رده‌بندی ترکیبی است که با استفاده از MKTPKS 3-Termsets برای تکنیک‌های داده‌کاوی متنی استفاده می‌شوند. پژوهش حاضر

براساس رده‌بندی داده‌های متنی به دو کلاس مختلف برای تعریف اسناد اطلاعات بـه دو گـروه خوب و بد اجرا شده است. ادغام و یکپارچه‌سازی تکنیک‌هـای رده‌بنـدی داده‌هـای متنـی سـبب بهبود رده‌بندی توسط الگوریتم‌ها و تکنیـک‌هـای رده‌بنـدی مـی‌شـود. در اغلـب مـوارد رویکـرد ارائه‌شده بهبود شایان توجهی را برای دقت رده‌بندی با اسـتفاده از ارزیـابی F-Measure از خـود نشان داده است. با این‌حال، رده‌بندی مستندات بـه کـلاس‌هـای مربوطـه بـا اسـتفاده از درخـت تصمیم (الگوریتم C۴/۵ یا J۴۸) دقت رده‌بندی را کاهش می‌دهد. دلیل کاهش دقـت، وابسـتگی شدید الگوریتم C۴/۵ به بسامد شرایط است. همچنین ماهیت داده‌های طبیعی ممکـن اسـت بـر دقت و صحت طبقه‌بندی تأثیر بگذارد.



شکل ۵. مقایسهٔ دقت طبقه‌بندی با استفاده از معیار F

از روش ارائه‌شده در این مقاله می‌توان به نکات زیر پی برد:

۱. روش نمایش مبتنی بر شرایط واحد، یکی از راه‌هـای مفیـد و کـاربردی بـرای کشـف دانش است، اما این روش‌ها بر دقت رده‌بندی داده‌های متنی تأثیر می‌گذارند؛
۲. کاربردهای ترکیبی تکنیک‌های داده‌کاوی متنی نتایج بهتری را در سـناریوی تحقیـق حاضر ارائه داد و همچنین، هرس اطلاعات و پالایش دانش ممکن است با اسـتفاده از قوانین انجمنی APRIORI از تکنیک‌های داده‌کاوی انجام پذیرد؛
۳. ساخت MKTPKS 3-TermSets و استفاده از آن دقت رده‌بندی را بهبود می‌دهد؛

۴. در برخی زمینه‌های کسب‌وکار اگر میزان خطا کاهش یابد، تصمیم‌های بهتری اتخاذ می‌شود.

نتیجه‌گیری و پیشنهادها

روش به‌کاررفته در پژوهش حاضر، روشی نوین در مستندسازی دانش‌های سازمانی است. این روش هنگام مستندسازی دانش اخذشده از طریق مصاحبه، به‌عنوان سازوکار تصدیق محتوای ثبت‌شده به‌کار برده می‌شود. دسته‌بندی دانش‌های مستندشده در پایگاه‌های دانشی، به‌صورتی است که در موقعیت بحرانی، زمان دسترسی به اطلاعات مفید و کاربردی را به‌شدت کاهش می‌دهد و این امر سبب بهینه‌شدن پایگاه‌های دانشی می‌شود. به‌بیانی، ابتدا داده‌ها و دانش به‌دست‌آمده از طریق مصاحبه به‌صورت متنی وارد الگوریتم پیشنهادی می‌شود؛ سپس از طریق داده‌های متنی، دانش واردشده به دو کلاس جداگانه از مستندات حاوی اطلاعات خوب و بد دسته‌بندی می‌شود.

در تحقیق حاضر فرمت داده‌ها و اسناد متنی به دو کلاس مختلف اختصاص داده شده است. روش ارائه‌شده در این پژوهش در قالب مطالعهٔ موردی در زمینهٔ مدیریت دانش زنجیرهٔ خدمات در حوزهٔ خدمات دولت الکترونیکی، براساس داده‌های به‌دست‌آمده از تحقیقات میدانی یکی از سازمان‌های دولت الکترونیکی به اجرا درآمد. رویکرد جدید ارائه‌شده در این تحقیق، استفاده از مجموعه عبارت‌های متوالی دانش با عنوان MKTPKS است که از آن می‌توان به‌منظور طبقه‌بندی اطلاعات ارسالی به اطلاعات خوب و بد استفاده کرد. در این تحقیق از تکنیک‌های C۴/۵، K-NN، Naiva Bayes و SVM برای آزمایش سودمندی روش پیشنهادی استفاده شده است.

تأکید مقالهٔ حاضر بر یکپارچه‌سازی تکنیک‌های داده‌کاوی متنی برای بهبود دقت رده‌بندی است. در رده‌بندی مستندات با استفاده از درخت تصمیم به کلاس‌های مربوطه، دقت رده‌بندی کاهش می‌یابد، اما استفاده از روش ارائه‌شده در این تحقیق، بهبود شایان توجهی را براساس تشکیل مجموعهٔ MKTPKS در دقت رده‌بندی با استفاده از معیار F-Measure اعمال می‌کند. کاهش دقت در C۴/۵، ممکن است به‌دلیل انتخاب اطلاعات معیارهای مبتنی بر آنتروپی براساس دروغ باشد، در حالی که سایر الگوریتم‌های رده‌بندی از فاصلهٔ ساده، احتمالاتی و معیارهای فاصلهٔ مبتنی بر هسته، برای پیداکردن شباهت بین مستندات استفاده می‌کنند. پیشنهاد می‌شود در پژوهش‌های آینده برای اعتبارسنجی روش‌های ارائه‌شده براساس معیارهایی غیر از F-Measure و همچنین کار روی ساختارهای جمله‌بندی و معنایی از داده‌های متنی که در

تحقیق حاضر به آن پرداخته نشده است، بر استراتژی‌های بهبود روش‌های مطرح‌شده کـه بـرای مدل‌های نمایندگی ماتریس است، تمرکز شود.

یکی از محدودیت‌های این تحقیق، استفاده از شمار محدود الگوریتم‌های متن‌کاوی است که پیشنهاد می‌شود سایر پژوهشگران روش استفاده‌شده در پژوهش حاضر را با الگوریتم‌هـای دیگـر متن‌کاوی آزمایش کنند. پژوهش حاضر را می‌توان برای توسعهٔ دانش در حوزه‌های سیستم بهبود کیفیت خدمات مبتنی بر دانش، بهبود راندمان و کارایی کارکنان در ارائهٔ گزارش‌های عملکرد که به‌صورت متنی ثبت شده است و بررسی داده‌های متنی کسب‌وکار به‌کار برد. همچنین می‌توان از الگوریتم متن‌کاوی پیشنهادی، در مراحل مختلف کدگذاری روش گراندد تئـوری بـرای تصـدیق محتوای هر مرحله استفاده کرد.

فهرست منابع

ثقفـی، ف.؛ علـی احمـدی، ع.؛ قاضـی نـوری، س.س. و حـورعلی، م. (۱۳۹۴)، تـدوین و شناسـایی سـناریوهای امکان‌پذیرآیندهٔ خدمات دولت الکترونیک ایران در افق ۱۴۰۴. فصلنامهٔ *مدیریت فناوری اطلاعـات*، ۷(۱)، ۶۸-۴۹.

رضائی نور، ج.؛ لسانی، ر.؛ زکی‌زاده، ع. و صفا مجید، غ. (۱۳۹۳)، بررسی شبکه‌های همکاری نویسـندگی در حـوزهٔ فناوری اطلاعات با استفاده از تکنیک‌های شبکه‌های اجتمـاعی، *فصـلنامهٔ مـدیریت فنـاوری اطلاعـات*، ۶(۲)، ۲۵۰-۲۲۹.

زارعی، ب.؛ ثقفی، ف. و زرین، ل. (۱۳۹۲)، سنجش میزان تـأثیر رویکـرد قابلیـت بـر توسـعهٔ دولـت الکترونیکـی، *فصلنامهٔ مدیریت فناوری اطلاعات*، ۵(۲)، ۹۴-۷۵.

Agrawal, R., Imielinski, T. & Swami, A. (1993). Mining association rule between sets of items in large databases. *In Proceedings of international conference on management of data (SIGMOD 93)*. (pp. 207–216).

Berry, M. J. A. & Linoff, G. (2004). *Data mining techniques for marketing, sales and customer relationship management*. Hoboken, NJ: Wiley Computer Publishing.

Cheung, M. S. & Myers, M. B. (2008). Managing knowledge sharing networks in global supply chains. *International Journal of Management and Decision Making*, 9(6), 581-599.

Han, J. & Kamber, M. (2000). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Hult, G.T.M., Ketchen, D.J. & Slater, S.F. (2004). Information processing, knowledge development, and strategic supply chain performance. *Academy of management journal*, 47(2), 241-253.

Jafari, M., Rezaeenour, J. & Akhavan, P. (2009). Identifying progressive route of organizational knowledge creation theory. *World Applied Sciences Journal,* 7 (10), 1287-1294.

Khalfan, M. M., Kashyap, M., Li, X. & Abbott, C. (2010). Knowledge management in construction supply chain integration. *International Journal of Networking and Virtual Organisations*, 7(2-3), 207-221.

Miao, D., Duan, Q., Zhang, H. & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. *Expert Systems with Applications*, 36(5), 9168–9174.

Myers, M. B. & Cheung, M. S. (2008). Sharing global supply chain knowledge. *MIT Sloan Management Review*, 49(4), 67-73.

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining systems. *IBM Systems Journal*, 40(4), 967 - 984.

Rezaeenour, J., Lesani, R., Zakizadeh, A. & Majid, G.S. (2014). Evaluating Authorship Collaboration Networks in the Field of Information Technology Using Social Netwowk Techniques. *Information Technology Management,* 6(2), 229-250. (*in Persian*)

Rezaienour, J. & Nazaridoust, M. (2012). Data Mining Application in Analysis of Knowledge Management Gaps. In *Proceeding of the 2nd World Conference on Soft Computing* (pp. 551-557).

Saghafi, F., Aliahmadi, A., Ghazinoory, S.S. & Hourali, M. (2015). Developing and Identifying Possibility & Plausibility of E-Government Services Scenarios in Iran by 1404, *Journal of "Information Technology Management,* 7(1), 49-68. (*in Persian*)

Sivakumar, K. & Roy, S. (2004). Knowledge redundancy in supply chains: a framework. *Supply Chain Management: An International Journal*, 9(3), 241-249.

Ur-Rahman, N. & Harding, J.A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, 39(5), 4729-4739.

Witten, I. H. & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations*. San Francisco*: Morgan Kaufman.*

Wu, D. J. (2001). Software agents for knowledge management: coordination in multi-agent supply chains and auctions. *Expert Systems with Applications*, 20(1), 51-64.

Zarei, B., Saghafi, F. & Zarrin, L. (2013). Measuring the Amount of Effects of Capability Approach on Developing E-government, *Information Technology Management,* 5(2), 75-94. (*in Persian*)