توسعه‌ی شاخص‌های کیفیت داده به‌منظور ارزیابی سامانه‌های اطلاعاتی تحقیقاتی: یک مطالعه‌ی موردی

آیناز اشتریان اصفهانی[1]؛ محمد جواد ارشادی[2]*؛ امیر عزیزی[3]

تاریخ دریافت مقاله: ۹۸/۰۵/۰۷
تاریخ پذیرش مقاله: ۹۸/۱۱/۱۴

چکیده

در پی رشد فناوری‌ها و ابزارهای ارتباطی و اطلاعاتی، امروزه شاهد تولید و توسعه‌ی پایگاه‌های داده در اکثر سازمان‌ها هستیم. از طرفی برای کسب جایگاه مناسب در دنیای کسب‌وکار کنونی لازم است سازمان‌های مختلف با تغییرات بیرونی سازگار شوند و به نوسانات مختلف بازار حساسیت کافی داشته باشند. به‌علاوه، کلید حل اکثر مشکلات سازمانی در دل داده‌ها و روندهای تولیدشده توسط همان سازمان است و بررسی داده‌های سازمان‌های دیگر تنها می‌تواند راهنمای رسیدن به پاسخ باشد. ازاین‌رو ارزشمندترین و مهم‌ترین موجودی هر سازمان، داده‌های تولیدشده توسط همان سازمان است. بر این اساس امروزه کیفیت داده‌های سازمان و پایش مستمر آن‌ها به‌عنوان یک راهبرد کلیدی شناخته می‌شود. در این پژوهش پس از بررسی متون علمی مختلف، شاخص‌های کیفیت داده مانند دقت، صحت، جامعیت و به‌هنگام بودن، برای ارزیابی سامانه‌های اطلاعاتی تحقیقاتی توسعه داده شدند. سامانه‌ی ملی ثبت پایان‌نامه/ رساله‌ی دانش‌آموختگان کل کشور به‌عنوان مطالعه‌ی موردی انتخاب شد. نتایج نشان داد پس از بهبود فرایند ثبت در این سامانه‌ی ملی شاخص‌های کیفیت داده، وضعیت بهتری را نشان می‌دهد. استفاده از فهرست‌های آماده‌ی کرکره‌ای به‌منظور افزایش خطاناپذیرسازی در ثبت داده‌ها، احراز هویت دانشجو و اساتید راهنما و مشاور به کمک کد ملی ازجمله پیشنهادهای اجرایی هستند که در راستای بهبود کیفیت داده‌ها در سامانه‌ی ثبت ارائه شدند.

واژگان کلیدی:
کیفیت داده، شاخص‌های کلیدی عملکرد، سامانه‌های اطلاعاتی تحقیقاتی

۱ مقدمه

افزایش رقابت میان سازمان‌های مختلف و تغییرات شدید بازار در سال‌های اخیر سازمان‌ها را برآن داشته است که به تحلیل عملکرد درونی خود و شناخت و پایش مستمر رقبای خود بپردازند. این کار بدون تحلیل مداوم داده‌ها امکان‌پذیر نیست. بر این اساس راهبردهای مبتنی بر داده، ارزشمندی داده‌های سازمان‌ها را بیش از پیش نمایان ساخته است. از این‌رو، تمامی سازمان‌ها در تلاشند تا با بررسی داده‌های خود و نیز سایر داده‌های مرتبط با بازار، عملکرد بهتری از خود نشان دهند. از سوی دیگر، به‌دلیل ناسازگاری ابعاد مختلف داده و مدل‌سازی نامناسب داده‌های مختلف و همچنین ساختارهای غلط پایگاه‌های داده و سامانه‌های اطلاعاتی، بسیاری از داده‌های موجود سازمان قابل بررسی نیستند. بنابراین، ارزیابی پیوسته‌ی داده‌ها، پایگاه‌های اطلاعاتی و سامانه‌های موجود به یکی از برنامه‌های اصلی سازمان‌ها بدل شده است. در نگاه کلی سنجه‌های عملکردی را می‌توان به سه گروه اصلی شاخص‌های کلیدی، نتیجه‌های عملکرد و شاخص‌های کلیدی عملکرد[1] (KPI) طبقه‌بندی کرد (Arabi et al., 2017). امروزه کنترل کیفیت آماری به‌عنوان یک رویکرد مهم در پایش شاخص‌های عملکردی شناخته می‌شود. علی‌رغم اینکه این فن در گذشته منحصراً به پایش کیفیت محصول می‌پرداخت اما امروزه این ابزار برای شاخص‌های فرایندی و سیستمی نیز به‌کار می‌رود. بنابراین، امروزه در متون علمی بیشتر با نام کنترل فرایند آماری شناخته می‌شود. هدف از کنترل فرایند آماری، یافتن تعادل اقتصادی بین تلاش‌های انجام‌شده در بازنگری کنترل کیفیت و احتمال‌یافتن محصول معیوب است. بازرسی صددرصد از تولید، کار و هزینه‌ی فشرده‌ای نیاز دارد. نمودارهای کنترل و ابزارهای کنترل کیفیت آماری با استفاده از ایجاد تعادل و پایداری اقتصادی به ما در بهبود مستمر سازمان کمک می‌کنند. از سوی دیگر ارزیابی عملکرد یک سازمان به کمک شاخص‌های کلیدی عملکرد می‌تواند آغاز حرکت روبه‌جلو برای تعالی باشد. به کمک اطلاعات تولیدشده توسط نمودار کنترل و ابزارهای مرتبط با آن و همچنین نتایج اثربخش سیستم‌های نظارت در قالب KPI می‌توان در سازمان‌های بزرگ کلان داده‌های مختلف را ساختار داد و روندهای بازار

۱ دانشجوی کارشناس ارشد مدیریت سیستم و بهره‌وری- دانشگاه آزاد اسلامی- واحد علوم و تحقیقا ت
۲ نویسنده مسئول- استادیار، عضو هیئت علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
۳ استادیار، عضو هیئت علمی دانشگاه آزاد اسلامی - واحد علوم و تحقیقا ت

و رقبا را تحلیل کرد (Hoyles, 2007).

KPI نه تنها برای نظارت بر فرایند تولید و ارائه‌ی محصول کاربرد دارد بلکه به مشتری اجازه می‌دهد که عملکرد آن محصول را قضاوت کند. درنتیجه، KPI یک ابزار ضروری است و به سازمان کمک می‌کند تا علاوه‌بر اینکه از کارایی ضعیف جلوگیری شود، به سمت موفقیت حرکت شود (Gasta, 2004).

۱-۱ اهداف پژوهش

سامانه‌ی گنج به‌منظور گردآوری همه‌ی داده‌های پایان‌نامه/ رساله‌های دانشجویان داخل کشور به گونه‌های طراحی شده است که متشکل از فیلدها یا اقلام دادهای مختلفی ازجمله نام، نام خانوادگی، شماره دانشجویی، کد ملی، مقطع تحصیلی، رشته دانشگاهی و ... است که به هر رکورد) سابقه (فایل مربوط به پایان‌نامه/ رساله‌ی آن دانشجو الصاق شده است. مجموعه داده‌های مذکور یک فراداده خوانده می‌شود. بررسی شاخص‌های عملکردی مهم در سامانه‌ی گنج در هریک از ابعاد چهارگانه‌ی کیفیت ذاتی، درستی، عینیت و قابلیت باورپذیری داده‌ها و ارائه‌ی چارچوبی برای پایش مستمر آن‌ها مهم‌ترین اهدافی است که در پژوهش حاضر مورد توجه قرار خواهند گرفت.

۲ مبانی نظری

داده‌ها در سازمان‌ها منبعی حیاتی برای حمایت از فرایند کسب‌وکار و تصمیم‌گیری مدیریتی به‌حساب می‌آیند (Khosroanjom et al., 2011). رشد انبار داده‌ها و دسترسی مستقیم به اطلاعات از منابع مختلف توسط مدیران و کاربران اطلاعات، نیاز و همچنین آگاهی از کیفیت بالای داده‌ها را در سازمان‌ها افزایش داده است (Pipino et al., 2002). همان‌طور که تولید یک محصول را می‌توان به‌صورت سیستم پردازشی توصیف کرد که برروی مواد خام فعالیت می‌کند تا آن‌ها را به محصولات فیزیکی تبدیل کند؛ تولید اطلاعات را نیز می‌توان به‌صورت یک سیستم پردازشی که برروی داده‌های خام فعالیت می‌کند تا آن‌ها را به محصولات اطلاعاتی تبدیل کند، تعریف کرد (جدول ۱).

جدول ۱: مقایسه‌ی تولید محصول و اطلاعات (Strong et al., 1997)

	تولید محصول	تولید داده
ورودی	مواد خام	داده خام
پردازش	خط تولید	سیستم اطلاعاتی
خروجی	محصولات	محصولات اطلاعاتی

داده‌های نامرغوب نیز می‌توانند بر قابلیت سیستم و همچنین تضمین عملکرد عملیاتی زیان وارد کنند (Geeki-yanage et al., 2018). از جهت دیگر داده‌های نامرغوب می‌توانند اثرات اجتماعی و اقتصادی قابل‌توجهی بگذارند. همچنین بر رضایت مشتریان و روحیه‌ی کارکنان و اطمینان میان سازمان‌ها تأثیرگذار است. کیفیت داده‌ها به‌عنوان یک مسئله‌ی عملکردی و ضروری در ارتباط با فرایندهای عملیاتی، تصمیم‌گیری‌ها (Chengalur-Smith et al., 1999) و همکاری‌های درون‌سازمانی (Mecella et al, 2002) شناخته شده است. نتایج برخی پژوهش‌ها نشان می‌دهد که تصمیم‌گیرندگانی که با DQ آشنایی داشته‌اند و از آن بیشتر برای حل در تصمیم‌گیری استفاده کرده‌اند نتایج اثربخش‌تری از تصمیم‌گیری‌های خود به‌دست آورده‌اند (Moges, 2016).

برخی مطالعات حاکی از آن است که کیفیت داده‌ها یک مفهوم چندبعدی است (Pipino et al., 2002). ازاین‌رو برای درک بهتر، اندازه‌گیری و بهبود DQ در متون علمی ابعادی مشخص شده است (Vaziri et al., 2017). ابعاد کیفیت داده که برای اولین‌بار توسط وانگ و استرانگ در سال ۱۹۹۷ عنوان شده است، شامل ۱۵ بعد است که در چهار دسته، طبقه‌بندی شده‌اند و این دسته‌بندی شامل: ذاتی، دسترس‌پذیرس، متنی و نمایشگری هستند. همانند هر مشخصه‌ی کیفی محصولی کیفیت داده نیز لازم است در سازمان مورد پایش قرار گیرد (Jones-Farmer et al., 2014) و برای این‌کار لازم است KPI مناسب برای این حوزه تدوین شود.

جدول ۲: دسته‌بندی ابعاد (Strong et al., 1997)

نام بعد	
ذاتی [۲]	دقت
	عینیت
	باورپذیری
	معتبر بودن
دسترس پذیری [۳]	در دسترس بودن
	امنیت
متنی [۴]	مربوط بودن
	ارزش افزوده
	به روز بودن
	جامعیت
	مقدار داده‌ها
نمایشگری [۵]	تفسیرپذیری
	سهولت درک
	نمایش مختصر
	نمایش سازگار

۳ پیشینه‌ی پژوهش

برای اندازه‌گیری ابعاد کیفیت داده‌ها، ابزارها و یا معیارهایی در متون علمی مشخص شده است، که این معیارها به‌صورت ذهنی[۶] و یا به‌صورت عینی[۷] قابل دسته‌بندی هستند. معیارهای ذهنی براساس نظرات و تجربیات کاربران داده که می‌توانند مدیران و یا متخصصان داده باشند، مشخص می‌شوند. وزیری و همکاران در سال ۲۰۱۷ عنوان کردند که سنجش ذهنی، معمولاً با استفاده از مصاحبه‌هایی که مربوط به داده‌ها و یا پرسش‌نامه‌هاست، صورت می‌پذیرد که منعکس‌کننده‌ی نیازها و تجارب ذی‌نفعان ازجمله گردآورندگان، نگهبانان و مصرف‌کنندگان داده است. از طرف دیگر سنجش عینی بر مبنای فرمول‌های ریاضی است که برای ارزیابی کیفیت یک مجموعه داده استفاده می‌شود.

در حوزه‌ی پایش آماری KPI به کمک SPC در صنایع و حوزه‌های مختلف مطالعات مختلفی انجام شده است. برای نمونه اندازه‌گیری عملکرد تأمین‌کنندگان فروشگاه‌های زنجیره‌ای (Morgan, 2007)، اندازه‌گیری تأثیر عوامل قومیتی و فرهنگی بر عملکرد کارکنان (Hoyles et al., 2007) و همچنین پایش KPI در خدمات فناوری اطلاعات شرکت‌های مخابراتی (Suhairi and Ford, 2013) همگی از SPC استفاده شده است.

همواره اندازه‌گیری کیفیت یک مقوله‌ی سخت و پیچیده بوده و به همین دلیل برای مدیران امری دشوار محسوب می‌شود. تحقیقات گذشته نشان می‌دهند به‌منظور ارزیابی خودکار کیفیت یک سیستم، تعیین شاخص‌ها امری اجتناب‌ناپذیر است. ارزیابی‌هایی که به‌صورت دستی توسط انسان‌ها انجام می‌شود، امکان دارد همراه با قضاوت‌های شخصی و در برخی موارد همراه با خطا باشند. ارزیابی مستمر، یک جزء حیاتی و جدانشدنی برای معیارهای کیفیت است. اختصاص منابع برای پوشش هزینه‌های دستی ارزیابی، تنها شاید یک‌بار امکان‌پذیر باشد، زیرا اگر مستمر و در طول زمان انجام شود، هزینه‌های زیادی دربر دارد. مهم‌ترین چالش‌هایی که در خودکارسازی ارزیابی شاخص‌های کیفیت وجود دارد ایجاد تعادل میان مقیاس‌پذیربودن[۸] و هدفمندبودن[۹] آن ارزیابی است. باوجود اینکه ارزیابی دستی هدفمند است، اما مقیاس‌پذیر نیست درصورتی‌که رویکردهای آماری ساده مقیاس‌پذیر هستند؛ اما معنادار نیستند. همچنین ارزیابی کیفی خودکار نیز می‌تواند معنادار باشد (Ochoa and Duval, 2006). ارزیابی عملکرد سیستم‌های مختلف، همواره یکی از دغدغه‌های اساسی برنامه‌ریزان و تحلیلگران سیستم‌ها بوده است. یکی از کاراترین راه‌هایی که مدیران موفق به‌صورت معمول برای نظارت بر عملکرد داخلی سازمان از آن استفاده می‌کنند، اندازه‌گیری شاخص‌های عملکردی در قسمت‌های مختلف سازمان است. قطعاً چنین رویکردی به بهبود مستمر شرکت منجر خواهد شد. در حوزه‌ی کیفیت داده آکوستا و همکاران، به شناسایی مشکلات مرتبط با کیفیت داده‌ها و جایگاه آن در ارزیابی عملکرد پرداختند. آن‌ها بررسی سامانه‌های اطلاعاتی و همه‌ی مبادی تولید داده به‌عنوان وسیله‌ای برای شناسایی مشکلات مرتبط با کیفیت داده‌ها استفاده کردند. رویکرد

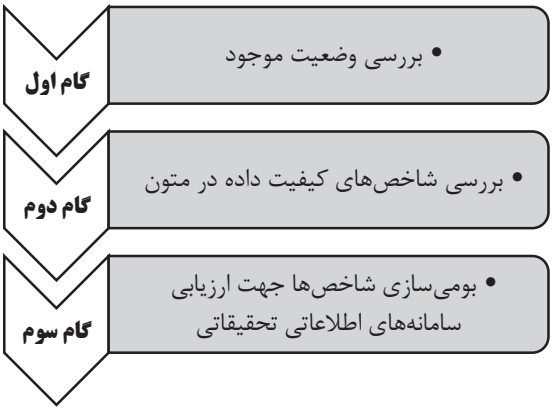
فصل‌نامه علمی مدیریت استاندارد و کیفیت
سال نهم- شماره ۱ - پیاپی ۳۱ - بهار ۱۳۹۸

خودکارسازی شناسایی مشکلات داده در پژوهش انجام شده به‌کار گرفته شد (Arabi et al., 2017). مطالعه‌ی موردی مدیریت کیفیت داده‌های حفاری با مقیاس آزمایشگاهی پژوهشی دیگر است که توسط گیکیانگ و همکارانش صورت گرفته است. هدف این پژوهش ارائه‌ی یک رویکرد مناسب برای درک، تمیز، بهبود و تفسیر داده‌ها پس از زمان واقعی برای حفظ یا افزایش ویژگی‌های داده، مانند دقت، هماهنگی، قابلیت اطمینان و اعتبار است. در این مطالعه، یک حفاری با مقیاس آزمایشگاهی با یک سیستم کنترل قادر به حفاری برای کسب اطلاعات و بهبود کیفیت استفاده می‌شود. عملکرد ایمن و کارآمد چنین سیستم کنترلی به‌شدت بر کیفیت داده‌های به‌دست‌آمده در حین حفاری و دسترسی کافی آن بستگی دارد. روش‌های مختلف بهبود کیفیت داده‌ها به سیستم هوشمند در تصمیم‌گیری بهتر و تشخیص سریع‌تر گسل کمک می‌کند. (Geekiyanage et al., 2018).

در این پژوهش ارائه‌ی شاخص‌هایی برای ارزیابی و بهبود کیفیت داده‌های یکی از مهم‌ترین سامانه‌های اطلاعاتی کشور موردتوجه پژوهشگران بوده است. در ادامه به بیان روش پژوهش خواهیم پرداخت.

۴ روش پژوهش

روش تحقیق مورد استفاده در این پژوهش توصیفی بوده است که برای تشخیص شاخص‌های مهم و درنتیجه بهبود کیفیت سامانه‌ی گنج تلاش می‌کند. شکل (۱) مراحل اجرای پژوهش را بیان می‌کند.



شکل ۱: نمودار اجرای تحقیق

همان‌گونه که در شکل (۱) قابل‌مشاهده است مراحل این پژوهش شامل سه گام اصلی است. در گام اول وضعیت موجود سامانه‌ی گنج را بررسی می‌کنیم. در گام دوم شاخص‌های موجود جهت ارزیابی ابعاد کیفیت داده در مقالات و متون را بیان می‌کنیم و سپس در گام سوم با توجه به دو گام ابتدایی شاخص‌های گردآوری شده را برای سامانه‌ی مذکور بومی‌سازی می‌کنیم و نتایج را بررسی خواهیم کرد. سنجش پایگاه داده به‌صورت بخشی از گام سوم صورت خواهد پذیرفت که باتوجه به اهمیت آن در ادامه به اجزای آن اشاره خواهد شد.

۴-۱ چارچوب روش سنجش پایگاه داده

سنجش پایگاه داده به‌عنوان بخشی از گام سوم پژوهش (شکل ۱) شامل اجزای زیر است.

۱) جمع آوری داده بر پایه‌ی شاخص‌های کیفیت داده:

در این مرحله اجزای شاخص‌هایی که بر پایه‌ی چارچوب عملکردی سامانه‌ی گنج توسعه داده شده شناسایی شده است و داده‌های موردنظر از پایان‌نامه‌ها و رساله‌های دانشجویان گردآوری می‌شود. به‌منظور تحلیل بهتر داده هر شاخص به‌طور جداگانه ذخیره می‌شود.

۲) بررسی نرمال‌بودن داده‌ها:

پس از جمع‌آوری و گردآوری داده‌های مدنظر هر شاخص، ضروری است که نرمال‌بودن داده‌ها مورد بررسی قرار گیرد.

۳) انجام تحلیل‌های آماری:

پس از اطمینان‌یافتن از نرمال‌بودن داده‌ها، داده‌های هر شاخص را طبق روابط در نظر گرفته‌شده (گام‌های اول و دوم شکل (۱) محاسبه‌کرده و نمودار کنترل ترسیم می‌شود. انجام تحلیل‌های روند، آنالیز رگرسیون و سایر تحلیل‌های آماری از مهم‌ترین اقدامات این مرحله است.

۴) تعیین اقدامات اصلاحی:

برپایه‌ی نتایج تحلیل‌های آماری و چگونگی روند شاخص‌ها و آستانه‌های تعیین‌شده اقدامات اصلاحی مناسب به‌منظور بهبود وضعیت شاخص‌ها تعیین خواهد شد.

۵ تجزیه‌تحلیل یافته‌ها

بر پایه‌ی سه گام اصلی معرفی‌شده در بخش چهارم (شکل ۱) در این بخش به نتایج هر گام به تفکیک اشاره می‌شود.

۵-۱ بررسی وضعیت موجود سامانه‌ی گنج

پژوهشگاه علوم و فناوری اطلاعات ایران، وظایفی همچون فراهم‌آوری، سازمان‌دهی، ذخیره، حفظ، بازیابی، تحلیل و اشاعه‌ی اطلاعات و مدارک علمی و فناوری کشور و اطلاعات پشتیبانی آن‌ها در سطح ملی و بین‌المللی و مدیریت تأمین را برعهده دارد. سامانه‌های مختلفی مانند سامانه‌ی گنج به‌منظور اجرای مناسب این مأموریت طراحی شده‌اند. سامانه‌ی گنج شرایطی فراهم‌آورده تا پژوهشگران کل کشور دسترسی آسان و سریع به مدارک اطلاعات علم و فناوری داخل و خارج از کشور داشته باشند. علی‌رغم کارایی و اثربخشی این سامانه در جهت ارتقا و انتقال آسان اطلاعات علمی به کاربران، پایگاه با معایب و اشکالاتی نیز مواجه است و نیاز به بهینه‌سازی دارد تا اهداف راهبردی پژوهشگاه هر چه بهتر محقق شود. بخشی از نابسامانی‌ها در این سامانه ناشی از خطاهای انسانی است که به هنگام نمایه‌سازی و ورود اطلاعات در بخش سازمان‌دهی اطلاعات صورت گرفته است. بخشی دیگر از ایرادها منشأ رایانه‌ای و سیستمی دارد که به مرور زمان و با تغییر نرم‌افزارها و سخت‌افزارها و کاراکترها و در هنگام تبدیل‌های مختلف و ورود ماشینی اطلاعات در پایگاه به‌وجود آمده است. فرایندهای اصلی اجرایی که سامانه‌ی گنج بر پایه‌ی آن‌ها اشاره شده است در سه گروه اجرایی مختلف به‌صورت زیر انجام می‌شود.

۱) گروه ارتباطات و فراهم‌آوری،
۲) مدیریت سازمان‌دهی و تحلیل اطلاعات،
۳) مدیریت حفظ و اشاعه‌ی اطلاعات.

* گروه ارتباطات و فرآهم آوری

این واحد مرحله‌ی ورودی و آغازین ورود منابع اطلاعاتی به مرکز اطلاعات است. مأموریت‌های اصلی گروه ارتباطات و فراهم‌آوری را «فراهم‌آوری اطلاعات و مدارک علم و فناوری» و «ثبت اطلاعات علم و فناوری کشور» تشکیل می‌دهد.

* مدیریت سازمان‌دهی و تحلیل اطلاعات

مأموریت مدیریت سازمان‌دهی و تحلیل اطلاعات شامل: تحقیق و بررسی روش‌های سازمان‌دهی و طبقه‌بندی مدارک و اطلاعات، سازمان‌دهی مدارک به‌منظور پردازش، تحلیل و بازیابی اطلاعات، انتخاب یکی از روش‌های نمایه‌سازی و کنترل واژگان با اصطلاح نامه‌هاست.

* حفظ و اشاعه‌ی اطلاعات

ایجاد دسترسی کاربران به اصل مدارک؛ اشاعه‌ی اطلاعات و تحویل نسخه‌های تکثیرشده‌ی مدارک در قالب‌ها و از روش‌های ممکن متناسب با زیرساخت‌های اطلاعاتی و ارتباطی کشور و نیاز کاربران است که این واحد شامل سه بخش سالن جست‌وجوی حضوری، سفارش مدرک غیرحضوری و گزارش پیشینه است.

۵-۲ بررسی شاخص‌های کیفیت داده در متون علمی

پی‌پی‌نو و همکاران

پی‌پی‌نو و همکاران در سال ۲۰۰۲ در مقاله‌ای ارزیابی ذهنی و عینی کیفیت داده‌ها را توصیف می‌کنند و سه کاربرگ عملی برای معیارهای کیفی داده‌های هدف ارائه می‌دهند. در پژوهش آن‌ها ارزیابی ذهنی و عینی کیفیت داده ترکیب می‌شود و نشان می دهد که چگونه در عمل قابل‌استفاده است. آن‌ها معتقدند اقدامات مربوط به کیفیت داده‌ها برای حل مشکلات کیفی خاص صورت می‌پذیرد و عمومیت بخشیدن آن‌ها فاقد کارایی لازم خواهد بود. پی‌پی‌نو و همکارانش در مقاله‌ی خود معیارهای کیفیت داده به‌منظور توسعه‌ی قابلیت استفاده داده را طراحی کردند (Pipino et al., 2002). چارچوب کلی طراحی شاخص‌ها از دیدگاه آن‌ها به‌صورت زیر است.

• نرخ ساده: این روش نسبت نتایج مطلوب به کل را اندازه می‌گیرد. تجربه نشان می‌دهد که مدیران ترجیح می‌دهند که نسبت نتایج مثبت را نشان دهند، زیرا این روش برای مقایسه‌هایی مفید است که نشان‌دهنده‌ی روند بهبود مستمر است. ابعاد دیگری که می‌تواند با استفاده از این کاربرگ ارزیابی شوند عبارت‌اند از: نشانه‌های اختصار، ارتباط و سهولت دسترسی و همچنین ابعاد خطای نشان‌دهنده‌ی صحت اطلاعات. اگر یک واحد داده را در خطا در نظر بگیریم، متریک به‌عنوان تعداد واحدهای داده دارای خطا تقسیم بر تعداد کل واحد داده‌ها محاسبه شده و به‌صورت نسبتی از یک تعریف می‌شود.

• روش کمینه یا بیشینه: کمینه یا بیشینه، رویکردی است برای بررسی ابعادی که ارزیابی آن‌ها نیاز به اجماع دارد. شاخص‌های کیفیت داده چندگانه (متغیر)، می‌تواند شامل رویکرد کمینه یا بیشینه باشد. این روش یک مقدار کمینه

(یا بیشینه) را از میان مقادیر نرمال‌شده‌ی شاخص‌های کیفیت داده‌ی فردی محاسبه می‌کند.

• میانگین وزنی: برای رویکردهای چندمتغیره، این روش جایگزینی برای روش قبلی است. برای اطمینان از نرمال‌بودن نتایج، در هر عامل وزن باید بین صفر و یک باشد. اگر سازمان بتواند میزان اهمیت هریک از متغیرها را به‌صورت باورپذیر مشخص کند، میانگین وزنی ممکن است یک کاربرگ مناسب برای استفاده باشد.

اچوا و دوال

اوچوا و دوال در سال ۲۰۰۶ مجموعه‌ای از معیارهای قابل اندازه‌گیری کیفیت فراداده را جمع‌آوری کرده‌اند. آن‌ها برای درک بهتر، معیارها را از چند بعد جامعیت، صحت، غنای اطلاعات و دقت ذاتی بررسی کرده‌اند که در ادامه به تشریح آن‌ها می‌پردازیم (Ochoa and Duval, 2006).

• کامل‌بودن[10]

معیار تمامیت با شمارش تعداد فیلدهای تکمیل‌شده در یک رکورد فراداده بررسی می‌شود. یک راه برای ساختن یک معیار قابل اندازه‌گیری بر این اساس، برآورد تعداد کل فیلدها و تمام فیلدهایی است که در یک مجموعه هستند که رابطه‌ی (۱) محاسبه می‌شود.

$$q_c(record) = \frac{\sum_{i=1}^{n}[record[field_i] \neq null]}{n} \qquad (۱)$$

• معیارهای کامل‌بودن

یکی از ابتدایی‌ترین روش‌ها برای ارزیابی سطح کامل‌بودن یک رکورد از فراداده شمارش تعداد فیلدهایی است که دارای محتوا هستند/ (خالی نیستند). در مورد فیلدهای چندمتغیره، اگر کمینه یکی از متغیرهای یک فیلد دارای مقدار باشد آن فیلد را کامل درنظر می‌گیریم. در ادامه رابطه‌ی (۲) فرمول ساده‌ای برای ارزیابی کامل‌بودن ارائه می‌دهد.

$$Q_{completness} = \frac{\sum_{i=1}^{n} p(i)}{N} \qquad (۲)$$

اگر فیلد i دارای یک مقدار غیر صفر باشد، P (i) برابر ۱ است و در غیر این صورت صفر است. N تعداد فیلدهاست.

• صحت[11]

صحت یک رکورد فراداده است که بیانگر درستی مقدار فیلد با توجه به منابع است. اوچوا و دوال پیشنهاد کردند که صحت می‌تواند به‌عنوان فاصله‌ی معنایی بین اطلاعات داده‌شده از طریق ثبت فراداده و اطلاعات داده‌شده از طریق منابع قابل‌درک باشد. این فاصله‌ی معنایی، تفاوت بین اطلاعاتی است که یک کاربر واردکرده و اطلاعات همان کاربر که می‌توان از منبع استخراج کرد. فاصله‌ی کوتاه‌تر به معنی دقت بالاتر رکورد فراداده است. با این شیوه معیار متری q_a را می‌توان به طریق رابطه‌ی (۳) محاسبه کرد.

$$q_a(record) = 1 - \frac{\sum_{i=1}^{n} d_i(record[field_i])}{n} \qquad (۳)$$

• معیارهای صحت

صحت شاخصی است که نشان می‌دهد به چه میزان مقادیر فراداده «درست[12]» است. در حقیقت شاخص صحت نشان می‌دهد که یک رکورد چقدر خوب توانسته است یک شیء را توصیف کند. شاخص صحت می‌تواند در دسته شاخص‌های عینی یا ذهنی طبقه‌بندی شود. در حوزه‌ی شاخص‌های عینی صحیح یا غلط‌بودن یک فیلد به راحتی ارزیابی می‌شود.

$$Q_{accuracy} = 1 - \frac{\sqrt{\sum_{i=1}^{n} d(field_i)^2}}{\sum_{i=1}^{n} d(field_i)} \quad \left(\text{با در نظر گرفتن } \sum_{i=1}^{n} d(field_i) > 0\right) \qquad (۴)$$

که در این رابطه عبارت d(fieldi) میزان صحت تعلق گرفته به فیلد iام است.

• غنای اطلاعات[13]

عبارات و توضیحات موجود به ازای هر رکورد از فراداده به توصیف آن فراداده می‌پردازد. به‌منظور ارجاع و توصیف منحصربه‌فرد یک فراداده لازم است عبارات و توضیحات موجود آن فراداده (اقلام اطلاعاتی) کافی باشد. از دیدگاه کاربر، یک رکورد اطلاعاتی درصورتی با کیفیت است که به اندازه‌ی کافی در مورد آنچه که توصیف می‌کند، اطمینان کافی را به کاربر منتقل کند. اوچوا و دوال پیشنهاد کردند که میزان غنای اطلاعات به‌وسیله‌ی اندازه‌گیری میزان اطلاعات منحصربه‌فرد موجود در فراداده قابل‌محاسبه خواهد بود آن‌ها این معیار را انطباق با انتظارات نامیدند که به طریق رابطه‌ی (۵) محاسبه می‌شود.

$$q_i(record) = \frac{\sum_{i=1}^{n} I(resource[field_i])}{n} \qquad (۵)$$

• دقت ذاتی[14]

دقت ذاتی درخصوص فیلدهای متنی به‌کار برده می‌شود. این شاخص مربوط به قابلیت خواندن داده است و به‌طورمستقیم

تحت‌تأثیر املای درست متن قرار می‌گیرد. میزان دقت ذاتی یک فیلد متنی به کمک شاخص q_{ip} از رابطه‌ی (۶) محاسبه می‌شود.

$$q_{ip}(record) = 1 - \frac{m}{n} \qquad (۶)$$

که در آن m تعداد اشتباهات املایی است و n تعداد کل کلمات است. شاخص دقت ذاتی در یک متن با ۵۰ کلمه و ۱۰ اشتباه املایی ۸۰ درصد می‌شود.

• معیارهای به‌هنگام بودن[۱۵]

به‌موقع بودن به‌طورخاص مربوط به درجه‌ای است که یک رکورد از فراداده در بین مخاطبان مورد استقبال قرار می‌گیرد. درحقیقت سطح به‌روزبودن داده توسط این شاخص ارزیابی می‌شود. درواقع با اندازه‌گیری و محاسبه می‌توان فهمید که یک فراداده چگونه با گذشت زمان باز هم می‌تواند مفید و به‌روز باشد.

$$\text{age} = \text{present\_year} - \text{publication\_year} \qquad (۷)$$

بتینی و همکاران

روش‌شناسی ارزیابی کیفیت اطلاعات از دیدگاه این پژوهشگران دارای معیارها و ابعاد مشخص و گوناگونی است. به‌علت رویکرد بالا به پایین در روش ارزیابی کیفیت اطلاعات برای تعریف ابعاد و معیارها دو دسته‌بندی متفاوت درنظر گرفته شده است که در زیر به آن‌ها می‌پردازیم.

• کیفیت محصول در مقابل کیفیت خدمات،

• مطابقت با مشخصات در مقابل برآورده‌سازی نیازهای مشتری یا عدم تأمین آن‌ها که این‌ها منجر به مجموعه‌ی به شدت پراکنده‌ای از ابعاد و معیارهای مرتبط می‌شود.

بتینی و همکاران در سال ۲۰۰۹ یک چارچوب جامع برای شناخت و ارزیابی روش‌های مختلف کیفیت داده ارائه کردند. این چارچوب شامل نُه دیدگاه مختلف است که اجازه می‌دهد تجزیه و تحلیل و مقایسه‌ی روش‌های متفاوت کیفیت داده به شکل کامل‌تری صورت گیرد. نُه دیدگاه فوق عبارت‌اند از:

۱) گام‌ها و مراحل ارزیابی کیفیت: یک روش می‌تواند تا سه مرحله یا فاز اصلی داشته باشد که عبارت‌اند از:

• بازسازی[۱۶]: جمع‌آوری اطلاعات اولیه و مقدماتی در رابطه با سازمان،

• اندازه‌گیری و ارزیابی: اندازه‌گیری و ارزیابی ابعاد کیفیت و مقایسه‌ی آن‌ها با مقادیر از پیش تعیین‌شده،

• بهبود: ارائه‌ی راهکارهای بهبود کیفیت داده.

۲) راهبردها و فنون: این دیدگاه راهبردهایی را که می‌توان از آن‌ها به‌منظور بهبود کیفیت داده استفاده کرد، ارائه می‌کند. به‌طورکلی برای بهبود کیفیت داده دو نوع راهبرد وجود دارد.

•راهبردهای داده‌محور،

•راهبردهای فرایندمحور

۳) ابعاد و معیارها: هر روش‌شناسی که برای بهبود کیفیت انتخاب می‌شود باید فهرستی از ابعاد و معیارها را در خود داشته باشد. برای هر بعد، متریک (ها) لازم است هدف‌گذاری جداگانه انجام شود.

۴) انواع هزینه‌ها: برخی از روش‌شناسی‌ها هزینه‌های بهبود کیفیت داده را ارزیابی می‌کنند. این هزینه‌ها ممکن است مستقیم (ناشی از اجرای روش) یا غیرمستقیم (ناشی از اجرای دوباره‌ی فرایندها یا از دست دادن فرصت‌های کسب‌و کار) باشند.

۵) انواع داده‌ها: سه نوع داده اصلی وجود دارد که یک روش‌شناس ممکن است با همه یا برخی از این انواع مواجه شود. این سه نوع داده عبارت‌اند از: ساختاریافته، نیمه ساختاریافته و بدون ساختار.

۶) انواع سیستم‌های اطلاعاتی: داده‌های مورد ارزیابی قرارگرفته در یک روش ممکن است در قالب شش نوع سیستم اطلاعاتی ایجاد شده باشد که عبارت‌اند از: یکپارچه، انبار داده، توزیع‌شده، تعاونی، سیستم‌های اطلاعات وب و همکار برخی از روش‌شناسی‌ها.

۷) نوع سازمان: این دیدگاه، نوع سازمانی را که داده در آن مورد ارزیابی قرار می‌گیرد موردتوجه قرار می‌دهد. درحالی‌که در بسیاری از روش‌شناسی‌ها نوع سازمان ذکر نشده است.

۸) فرایندها: روش‌شناسی‌های مختلف فرایندهایی را که داده‌ها را ایجاد و به‌روزرسانی می‌کنند موردتوجه قرار می‌دهند.

۹) خدمات: خدماتی که توسط فرایندهای مربوط به کیفیت داده ارائه می‌شوند در این دسته موردتوجه قرار می‌گیرند. به‌عنوان‌مثال، در دانشگاه بیشتر خدمات مرتبط با فرایندهای حوزه‌ی کیفیت داده به امور ثبت‌نام یا ارائه‌ی گزارش‌های

آموزشی به دانشجویان می‌پردازند.

هنریش و همکاران

هنریش و همکاران در سال ۲۰۰۹ در قالب یک روش اجرایی معیارهایی را برای ارزیابی کیفیت داده از بعد «رایج‌بودن» معرفی و توسعه دادند. آن‌ها الزاماتی را معرفی کردند که به‌منظور ایجاد اطمینان از مناسب‌بودن شاخص‌های کیفیت داده‌ی تعریف‌شده می‌توانند مورداستفاده قرار بگیرند. در ادامه این الزامات معرفی می‌شوند.

• [قابلیت نرمال‌سازی[۱۷]] این الزام برای اطمینان از قابل مقایسه بودن مقادیر یک شاخص در حوزه‌ها یا مقاطع زمانی مختلف ضروری است. به‌عنوان‌مثال برای مقایسه‌ی سطوح مختلف DQ در طول زمان (پی‌پی‌نو و همکاران، ۲۰۰۲) معیارهای DQ اغلب یک مقدار بین صفر (کاملاً بد) و یک (کاملاً خوب) تعریف می‌شوند.

• [مقیاس فاصله‌ای[۱۸]] به‌منظور پایش شاخص‌های کیفیت داده در طول زمان و ارزیابی اقتصادی از اقدامات انجام‌گرفته لازم است شاخص‌های تعیین‌شده از مقیاس فاصله‌ای برخوردار باشند. به این معنا که، تفاوت بین دو سطح کیفیت داده در یک شاخص باید معنی‌دار باشد. بنابراین، به‌عنوان‌مثال، تفاوت ۰/۲ از مقادیر ۰/۷ و ۰/۹ و مقادیر ۰/۴ و ۰/۶ پس از اندازه‌گیری یک شاخص باید معنای مشابهی را داشته باشد.

• [تفسیرپذیری[۱۹]] الزام «تفسیر آسان توسط کاربران کسب‌وکار» توسط برخی نویسندگان بیان شده است. به بیان دیگر، ارزش معیارهای DQ باید به‌درستی توسط کاربران درک شود.

• [تجمیع‌پذیری[۲۰]] درصورت استفاده از یک مدل داده‌ی رابطه‌ای، معیارها باید انعطاف‌پذیری کافی داشته باشند. به‌عبارتی، معیارها باید به‌گونه‌ای تعریف شوند که بتوان آن‌ها را از یک سطح پایین پایگاه داده به سطح بالاتر تعمیم داد و درعین‌حال مقایسه کرد. به‌عنوان‌مثال اگر یک شاخص کیفیت داده در سطح جدول تعریف می‌شود بتوان عدد به‌دست‌آمده را در سطح پایگاه داده (که مجموعه‌ای از جداول به‌هم پیوسته است) تعمیم داد و نتایج آن را مقایسه کرد. همچنین این الزام نشان می‌دهد که باید بتوان نتایج به‌دست‌آمده از جداول یک پایگاه داده را مقایسه و درصورت لزوم تجمیع کرد.

• [سازگاری[۲۱]] معیارهای تعیین‌شده در حوزه‌ی کیفیت داده باید درزمینه‌ای که مورد استفاده قرار می‌گیرند قابلیت سازگاری داشته باشند.

• [امکان‌سنجی[۲۲](سهولت محاسبه)] برای اطمینان از عملی‌بودن، معیارها باید براساس شاخص‌های ورودی تعیین شوند. هنگام تعریف معیارها، باید روش درست تعیین شاخص‌های ورودی تعریف شود. اگر تعیین دقیق شاخص‌های ورودی امکان‌پذیر نباشد یا هزینه‌ی آن زیاد باشد، روش‌های متداول جایگزین (مانند روش‌های آماری) پیشنهاد می‌شود. به‌عبارتی باید بتوان شاخص کیفیت داده را به‌گونه‌ای تعیین کرد که امکان خودکارسازی اندازه‌گیری آن به راحتی فراهم شود. همان‌طورکه در مقدمه عنوان شد این شش الزام به‌منظور ارزیابی شاخص‌های مختلف کیفیت داده می‌توانند مورد استفاده قرار گیرند. در ادامه به بررسی ادبیات مربوط به شاخص رایج‌بودن در مقالات مختلف خواهیم پرداخت. جدول (۳) برخی تعاریف برگرفته‌شده از مراجع مختلف درخصوص شاخص رواج ارائه می‌شود

جدول ۳: تعاریف انتخاب‌شده برای بعد رواج

تعریف	مرجع
به روز بودن : "منقضی نبودن مقادیر ثبت شده. منقضی بودن یک داده از این جهت که خطایی در آن رخ داده است که با مقدار فعلی (صحیح) متفاوت است."	بالو و همکاران (۱۹۹۸)
به موقع بودن: "میزانی که عمر داده برای کار فعلی مناسب است."	استرانگ و همکاران (۱۹۹۷)
رواج: به میزان به روز بودن داده اطلاق میشود. به روز بودن داده به این معناست که داده در گذر زمان و با وجود هرنوع استهلاکی که ممکن است به آن دچار شود کماکان صحیح باقی بماند.	ردمان (۱۹۹۸)
به موقع بودن: "میزان اطلاعاتی که برای کار در دست به اندازه کافی روزآمد است"	پی پی نو (۲۰۰۲)

در ادامه چهار رابطه‌ی اصلی که توسط هینریچس (۲۰۰۹) بالو و همکاران (۱۹۹۸)، (اون و شانکاران (۲۰۰۷) و هینریچس و همکاران (۲۰۰۹) درخصوص کمی‌کردن شاخص رواج توسعه داده شده‌اند، ارائه شده است. هیریچنس در سال ۲۰۰۹ شاخص رواج را در قالب رابطه‌ی (۸) محاسبه کرد.

$$\text{Currency} = \frac{1}{(mean\ attribute\ update\ frequency) - (age\ of\ attribute\ value) + 1} \quad (۸)$$

در این رابطه شاخص mean attribute update frequency میزان فرکانس به‌روزرسانی ویژگی را نشان می‌دهد. (مثلاً ۱۰ بار در سال) از سوی دیگر شاخص age of attribute value نشانگر عمر یک ویژگی در پایگاه داده است. به بیان دیگر این شاخص نشان می‌دهد که فاصله‌ی زمانی میان ذخیره‌شدن یک ویژگی در پایگاه داده و زمان اندازه‌گیری شاخص کیفیت داده چقدر بوده است.

بالو و همکاران (۱۹۹۸) تعریف شاخص رواج را به‌صورت زیر ارائه کردند.

$$\text{Currency} = \max\left[\left(1-\frac{age\ of\ attribute\ value}{shelf\ life}\right)\right]^{s} \qquad (۹)$$

در این رابطه age of attribute value مشابه رابطه‌ی قبلی است و شاخص shelf life نشان‌دهنده‌ی میزان نوسانات صورت‌گرفته در یک ویژگی در گذر زمان است. شاخص s نیز توسط خبرگان و باتوجه به نوع کاربرد داده تعیین خواهد شد.

اون و شانکاران (۱۹۹۷) رابطه‌ی (۱۰) را به‌منظور ارزیابی شاخص رواج توسعه دادند.

$$1-(age\ of\ attribute\ value/T_{max}\ )^{s} \qquad (۱۰)$$

در این رابطه‌ی شاخص T_{max}، بیشینه دوره‌ی عمر یک ویژگی از داده است و شاخص s یک عدد حقیقی مثبت است که براساس نظر خبرگان و متناسب با کاربرد این رابطه تعیین می‌شود.

شاخص چهارم که توسط هیریچنس در سال ۲۰۰۷ ارائه شد، براساس نظریه‌ی احتمالات شکل گرفته است و میزان احتمال به‌روزبودن داده را به‌عنوان شاخص رواج توصیف کرده است.

$$exp(-decline(A)\cdot age(w.A)) \qquad (۱۱)$$

وزیری و همکاران

محققانی همچون وزیری و همکاران (۲۰۱۷) مطالعاتی را درزمینه‌ی اندازه‌گیری کیفیت داده‌ها در حالتی که معیارها دارای وزن‌های متفاوتی هستند، انجام داده‌اند. در این پژوهش، کیفیت داده‌ها (DQ) با عنوان «مناسب برای استفاده[۲۳]» تعریف شده است. به‌منظور اندازه‌گیری و بهبود DQ، روش‌های مختلفی تعریف شده است. به‌طورمعمول، این روش‌ها مجموعه‌ای از دستورالعمل‌ها و فنونی هستند که فرایند منطقی را برای اندازه‌گیری و بهبود کیفیت داده‌ها تعریف می‌کند. در روش ارائه‌شده در این پژوهش به‌منظور ارائه‌ی راهکارهای سیستمی و سازمان‌یافته‌تر (در فرایند، سامانه، پایگاه داده و ...)، ابعاد مختلفی برای کیفیت داده تعریف شده است تا بتوان اقدامات اصلاحی را برپایه‌ی نوع مشکل ارائه کرد. برخی از مهم‌ترین ابعاد DQ، دقت، کامل‌بودن، به‌موقع‌بودن و مرتبط‌بودن است. در هر سازمان متناسب با نوع داده و کاربردهای آن شاخص‌های متفاوتی برای اندازه‌گیری کیفیت داده موردتوجه قرار می‌گیرد. در اکثر سازمان‌ها، برخی از داده‌ها بیشتر از سایرین ارجحیت دارند. به‌عبارت‌دیگر، برخی داده‌ها وزن بیشتری دارند. ازاین‌رو در اندازه‌گیری شاخص‌های DQ این دسته از داده‌ها باید نقش پررنگ‌تری بازی کنند. اکثر معیارهای توسعه‌یافته تاکنون به وزن داده‌ها توجه نکرده‌اند. وزیری و همکاران، معیارهای جدیدی براساس وزن داده‌ها ایجاد می‌کنند تا اثربخشی شاخص‌ها را افزایش دهند. مطالعات آن‌ها نشان داد که اثربخشی اندازه‌گیری‌های DQ توسط معیارهای وزنی تاحدود زیادی بیشتر است. در هر سازمان وزن داده‌ها می‌تواند متناسب با نقش آن داده در دستیابی سازمان به اهداف کسب‌وکار خود به شکل متفاوتی تعریف شود. بنابراین، چنین داده‌هایی باید در اندازه‌گیری DQ اهمیت بیشتری داشته باشد. به‌عنوان‌مثال در بانک اطلاعاتی مشتریان یک شرکت فیلدهایی مانند آدرس (ستون آدرس) ممکن است به لحاظ اهمیت در سطح بالاتری نسبت به سایر فیلدها (مانند نام، سن و ...) باشد و می‌تواند با اهمیت جداگانه‌ای در کیفیت داده موردتوجه قرار گیرد. از سوی دیگر مشتریان یک منطقه‌ی خاص جغرافیایی ممکن است به‌منظور دستیابی سازمان به اهداف راهبردی خود جایگاه بالاتری نسبت به سایر مناطق داشته باشد. پس لازم است داده‌های آن منطقه به‌طور ویژه‌تری موردتوجه قرار گیرد. در ادامه به معرفی مهم‌ترین شاخص‌های وزن‌دهی شده توسط وزیری و همکاران خواهیم پرداخت.

- کامل‌بودن (وزن‌دهی به ستون‌ها)[۲۴]

این شاخص بدین‌منظور توسعه داده شده است که بتواند کامل‌بودن یک پایگاه داده را از دیدگاه کیفیت داده با در نظر گرفتن اهمیت ستون‌های (فیلدها) مختلف مورد ارزیابی قرار دهد. این شاخص می‌تواند برای کل جدول با رابطه‌ی

(۱۲) محاسبه شود.

$$\sum_{i=1}^{n}(cw_i \times cc_i) \qquad (۱۲)$$

در این رابطه، CW_i درجه‌ی اهمیت و CC_i میزان کامل بودن ستون iام است. n تعداد کل ستون‌ها در پایگاه داده است.

• کامل‌بودن وزنی چندتایی‌ها[۲۵] (Tulpes)

یک چندتایی (Tuple) در پایگاه داده همان برداری است که هر درایه آن می‌تواند یک یا چند درایه مختلف داشته باشد. به‌منظور ارزیابی کیفیت داده در یک چندتایی رابطه‌ی وزن‌دهی شده‌ی زیر توسط وزیری و همکاران ارائه شد.

$$\sum_{j=1}^{m}(tw_j \times tc_j) \qquad (۱۳)$$

در رابطه‌ی بالا فرض شده است که در یک پایگاه داده m چندتایی (Tuple) مختلف وجود دارد و از نظر سازمان هر چندتایی اهمیت متفاوتی (tw_j) دارد. همچنین میزان کامل‌بودن چندتایی jام با متغیر tc_j مورد ارزیابی قرار می‌گیرد. در ادامه و در شاخص بعدی می‌توان با ترکیب دو شاخص فوق شاخص کامل‌تری به‌منظور ارزیابی کیفیت داده‌ها ارائه کرد.

• کامل‌بودن وزنی (وزن‌دهی به هر سلول[۲۶])

این شاخص روشی برای محاسبه‌ی کامل‌بودن با وزن‌دهی به تک‌تک سلول‌های جدول ارائه می‌کند. هر سلول دارای دو وزن مرتبط با آن است که یکی وزن آن چندتایی (tw_j) است که سلول به آن تعلق دارد و دیگری وزن ستونی است که سلول در آن قرار دارد. با ترکیب این دو وزن رابطه‌ی جدید به شکل زیر ارائه خواهد شد.

$$\sum_{i=1}^{n}\sum_{i=1}^{m} tw_j \times cw_i \times cell_{ji} \qquad (۱۴)$$

باید توجه داشت که هنگام ارزیابی سلول به‌صورت تک‌تک، مقدار $cell_{ji}$ متناسب با کامل یا تهی‌بودن آن یک یا صفر است.

• مرتبط‌بودن ستون‌ها[۲۷]

یکی دیگر از مهم‌ترین مسائلی که باید درنظر گرفته شود، «مرتبط‌بودن» است. Relevancy به معنی میزان اطلاعات قابل‌استفاده و مفید برای کار است. یکی از معیارهایی که باید برای رعایت آن موردتوجه قرار گیرد تعداد «دسترسی» به داده است. به‌عبارت‌دیگر، دامنه‌ها یا ستون‌هایی که بیشترین دسترسی را دارند اهمیت بیشتری دارند؛ بنابراین، این در این رویکرد وزن بیشتری به‌دست می‌آورد. تعداد دسترسی‌ها باید نسبت به دسترسی‌های کل به مجموعه داده‌های مربوط درنظر گرفته شود. در ادامه مهم‌ترین رابطه‌های شاخص مرتبط‌بودن ارائه شده است.

$$\text{مرتبط بودن در سطح پایگاه داده} = \frac{\text{تعداد دسترسی‌ها به یک پایگاه داده}}{\text{کل دسترسی‌ها به کل پایگاه‌های داده}} \qquad (۱۵)$$

$$\text{مرتبط بودن در سطح جدول} = \frac{\text{تعداد دسترسی‌ها به یک جدول}}{\text{کل دسترسی‌ها به کل جدول‌ها}} \qquad (۱۶)$$

$$\text{مرتبط بودن در سطح چندتایی} = \frac{\text{تعداد دسترسی‌ها به یک چندتایی}}{\text{کل دسترسی‌ها به کل چندتایی‌ها}} \qquad (۱۷)$$

$$\text{مرتبط بودن در سطح ستون} = \frac{\text{تعداد دسترسی‌ها به یک ستون}}{\text{کل دسترسی‌ها به کل ستون‌ها}} \qquad (۱۸)$$

در رابطه‌های ارائه‌شده‌ی بالا میزان دسترسی به پایگاه داده، چندتایی، جدول و ستون در روابط (۱۵) تا (۱۸) مورد ارزیابی قرار می‌گیرد. به‌عبارتی به‌عنوان‌مثال درصورتی‌که در مجموعه پایگاه‌های داده تعداد دسترسی‌ها به یک پایگاه بیشتر از سایر پایگاه‌ها باشد شاخص مرتبط‌بودن برای آن پایگاه داده باید عدد بیشتری را نشان دهد. وزیری و همکاران نشان دادند که نتایج به‌دست‌آمده از روابط (۱۷) و (۱۸) می‌تواند جایگزین مناسبی برای شاخص وزن در روابط (۱۵) و (۱۶) باشد.

در بخش (۳) (پیشینه‌ی پژوهش) شاخص‌های ابعاد کیفیت داده و مقالاتی را که به توسعه‌ی این شاخص‌ها پرداختند، بررسی کردیم. همان‌طورکه دیدیم معیارها و شاخص‌ها بخشی جدایی‌ناپذیر از روش‌های کیفیت داده هستند. به‌منظور بهره‌مندشدن از روش‌شناسی‌های کارا و اثربخش، معیارهای مناسبی باید توسعه یابند. در بخش بعد، شاخص‌ها و معیارهایی برای ارزیابی کیفیت داده در سامانه‌های اطلاعاتی تحقیقاتی ارائه خواهد شد.

۳-۵ بومی‌سازی شاخص‌ها جهت ارزیابی سامانه‌های اطلاعاتی تحقیقاتی

با توجه به مطالعات انجام‌گرفته در گذشته، بررسی رابطه‌ها، سوابق، مقالات و با درنظر داشتن تمام ابعاد شاخص‌های کیفیت، حال به شاخص‌های کیفی سامانه‌ی گنج می‌پردازیم تا تمام ابعاد را بررسی و شاخص‌های مهم را استخراج کنیم. در ابتدا هرکدام از شاخص‌ها را به تفکیک بررسی‌کرده و جدول شناسنامه‌ی شاخص را جداگانه رسم می‌کنیم.

- جامعیت،

- صحت،
- دقت،
- به‌هنگام بودن.

۵-۳-۱ جامعیت

معیار جامعیت یکی از مهم‌ترین معیارهای DQ است، که با شمارش تعداد فیلدهای تکمیل‌شده در یک رکورد فراداده، سنجش و ارزیابی می‌شود. با توجه به این معیار یکی از شاخص‌های اصلی در سامانه‌ی گنج معیار جامعیت یا تمامیت است که میزان کامل‌بودن (پرشده) رکوردها توسط دانشجویان را اندازه‌گیری می‌کند.

جدول ۴: شناسنامه‌ی شاخص کامل‌بودن

شناسنامه شاخص‌های کیفیت داده در سامانه گنج
نام شاخص : میزان کامل بودن فیلدها در سامانه گنج (جامعیت)
فرمول شاخص $$Completeness(j) = \frac{\sum_{i=1}^{n} Per\_Com(i)}{n}$$ (۱۹)
معرفی متغیرها Completeness(j) = میزان کامل بودن نمونه‌ها (پایان‌نامه‌ها) در نمونه یا دوره jام n= تعداد پایان‌نامه‌ها در هربار نمونه‌برداری Per_Com (i): شاخص کامل بودن پایان‌نامه‌ها در نمونه iام. این شاخص از نسبت تعداد فیلدهای پر شده در پایان‌نامه iام به کل فیلدهای موجود در سامانه گنج بدست می‌آید.

۵-۳-۲ صحت

صحت شاخصی برای ارزیابی یک رکورد فراداده است که بیانگر درستی مقدار فیلد با توجه به منابع است. در سامانه‌ی گنج این فاصله‌ی معنایی، نسبت فیلدهای صحیحی که یک دانشجو واردکرده به کل فیلدهای پرشده توسط همان شخص که درواقع این شاخص بر دو اساس درصد[۲۸] و تعداد[۲۹] محاسبه می‌شود.

جدول ۵: شناسنامه‌ی شاخص صحت

شناسنامه شاخص‌های کیفیت داده در سامانه گنج
نام شاخص: **نسبت تعداد فیلدهای صحیح به کل فیلدهای تکمیل شده**
فرمول شاخص $$Per - Accuracy(j) = \frac{\sum_{i=1}^{n} Per\_Acc(i)}{n}$$ (۲۰) $$Con - Accuracy(j) = \frac{\sum_{i=1}^{n} Con\_Acc(i)}{n}$$ (۲۱)
معرفی متغیرها **Per_Acc(i)** = درصد فیلدهای صحیح پایان نامه iام به کل فیلدهای تکمیل شده توسط دانشجو **Con_Acc(i)** = تعداد فیلدهای صحیح پایان نامه iام به کل فیلدهای تکمیل شده توسط دانشجو **n**= تعداد پایان‌نامه‌ها در هربار نمونه‌برداری

۵-۳-۳ دقت

دقت درخصوص فیلدهای متنی به‌کار برده می‌شود که مشابه با معیار دسترس‌پذیری، این مورد نیز مربوط به قابلیت خواندن است. این موضوع به‌طورمستقیم تحت‌تأثیر املای درست متن قرار می‌گیرد. این شاخص در سامانه‌ی گنج جهت درنظر گرفتن تعداد اشتباهات دانشجویان حین واردکردن اطلاعات خود و بررسی دقت این افراد طرح شده است که از رابطه‌ی یک منهای نسبت اشتباهات املایی در کلمات به کل کلمات متن به‌دست می‌آید.

جدول ۶: شناسنامه‌ی شاخص دقت

شناسنامه شاخص‌های کیفیت داده در سامانه گنج
نام شاخص : **نسبت تعداد اشتباهات املایی فیلد به کل کلمات متن فیلد**
فرمول شاخص $$q_{ip}(record) = 1 - \frac{m}{n}$$ (۲۲)
معرفی متغیرها **m** = **اشتباهات املایی** **n** = **کل کلمات متن**

۵-۳-۴ به‌هنگام بودن

به‌هنگام بودن به‌طورخاص مربوط به درجه‌ای از به‌روزبودن است که یک رکورد فراداده در میان جامعه‌ی خاص درحال‌حاضر جریان دارد. درواقع با اندازه‌گیری و محاسبه می‌توان فهمید که یک فراداده چگونه با گذشت زمان باز هم می‌تواند مفید و به‌روز باشد.

نقش این شاخص در سامانه‌ی گنج اندازه‌گیری فاصله‌ی زمانی بین دفاع دانشجویان و ثبت پایان‌نامه‌ی آن‌هاست.

جدول ۷: شناسنامه‌ی شاخص به‌هنگام بودن

شناسنامه شاخص‌های کیفیت داده در سامانه گنج
نام شاخص : **فاصله زمانی بین دفاع و ثبت پایان نامه (به هنگام بودن)**
فرمول شاخص $$Timeliness(j) = \frac{\sum_{i=1}^{n} Dis\_Tim(i)}{n}$$ (۲۳)
معرفی متغیرها **Dis-Tim(i)= فاصله زمانی بین دفاع دانشجو و ثبت پایان نامه توسط دانشجو در پایان‌نامه iام** **n= تعداد پایان‌نامه‌ها در هربار نمونه‌برداری**

۵-۴ ارزیابی سامانه‌های اطلاعاتی به کمک شاخص‌های کیفیت داده

سامانه‌های گردآوری و ثبت، سازمان‌دهی و اشاعه‌ی اطلاعات پایان‌نامه‌ها/ رساله‌های دانش‌آموختگان داخل کشور یکی از مهم‌ترین سامانه‌های اطلاعاتی تحقیقاتی هستند که رسالت اصلی آن حفظ و اشاعه‌ی پایان‌نامه‌ها و رساله‌های داخل کشور بوده و روزانه بازدیدکنندگان فراوانی دارد. از آنجا که پایان‌نامه‌ها و رساله‌های اشاعه داده‌شده در سامانه‌ی گنج مبنای پژوهش خیل کثیری از پژوهشگران در مقطع کارشناسی ارشد و دکتری قرار خواهد گرفت؛ بنابراین، ارتقای کیفیت این سامانه تأثیری انکارناپذیر در افزایش کیفیت پژوهش‌های صورت‌گرفته از سوی پژوهشگران خواهد داشت. به‌منظور ارتقای کیفیت سامانه‌ی ثبت که مقدمه و بخش بنیادین مجموعه سامانه‌های ایرانداک در حفظ و اشاعه محسوب می‌شود از سال ۱۳۹۴ پروژه‌ی بهبود این سامانه آغاز شد و از سال ۱۳۹۶ مورد استفاده دانشجویان قرار گرفت.

بر پایه‌ی اندازه‌گیری و توسعه‌ی شاخص‌های کیفیت داده بهبودهای زیر در سامانه‌ی ثبت جدید نسبت به ورژن قبلی آن صورت پذیرفت.

- اضافه‌شدن تأیید دانشگاه بعد از تأیید ایرانداک در فرایند ثبت پارساها به‌منظور افزایش قابلیت اطمینان مدارک اشاعه داده‌شده و بهبود کیفیت داده‌های نهایی؛
- استفاده از فهرست‌های آماده‌ی کرکره‌ای به‌منظور افزایش خطاناپذیرسازی در ثبت داده‌ها توسط دانشجویان. به‌عنوان‌مثال برخلاف سامانه‌ی قدیم در این سامانه نام دانشگاه، نوع دانشگاه و ... از یک فهرست از پیش طراحی شده انتخاب می‌شود که این کار از یکسو امکان اشتباه در ایجاد داده را حذف می‌کند و از سوی دیگر امکان گزارش‌گیری در مدیریت علم و فناوری در ایرانداک را افزایش می‌دهد؛
- اطلاع‌رسانی به‌کمک پیامک و رایانامه به افزایش سرعت در اجرای فرایند با حفظ کیفیت خروجی؛
- خطاناپذیرکردن فرایند احراز هویت دانشجو و اساتید راهنما و مشاور به کمک کد در راستای افزایش کیفیت داده‌ها؛
- در برخی از فیلدها مانند زبان، ارائه‌ی امکان انتخاب توسط دانشجو به‌صورت تیک گزینه‌ها امکان خطا در فرایند را به شدت کاهش داد؛
- ارائه‌ی دستورالعمل‌های مناسب در هنگام ثبت‌نام در سیستم ثبت جدید توسط دانشجویان .

۵-۵ سنجش پایگاه داده

همان‌طورکه پیشتر توضیح دادیم، جهت پایش شاخص‌ها برای هر شاخص، شناسنامه‌ای طراحی‌کرده و به کمک آن شاخص‌ها را اندازه‌گیری کردیم. میانگین شاخص‌های کامل‌بودن، صحت، دقت و به‌هنگام بودن پس از اندازه‌گیری بر پایه‌ی شاخص‌های مذکور، به‌ترتیب ۹۸/۹، ۸۸/۳، ۰/۹۵۷ و ۱۰۶/۹، به‌دست آمد. به بیان دیگر، ۹۸/۹ درصد از فیلدهای سامانه ثبت جدید به‌طورکامل پرشده‌اند. ۸۸/۳ درصد از فیلدهای پرشده از منظر شاخص صحت دارای مطلوبیت کافی هستند. دقت ۹۵/۷ درصد از فیلدهای تکمیل‌شده مناسب است و درنهایت به‌طور میانگین از زمان دفاع تا ثبت در سامانه ۱۰۶/۹ روز می‌گذرد.

همچنین، در تحلیل کیفی شاخص‌های کیفیت داده، در سامانه‌ی ثبت قدیم، متوسط تعداد نقص‌های مشاهده‌شده درمجموع فیلدهای عنوان و چکیده برابر ۸/۸ است. از سوی دیگر میانگین تعداد نقص در سامانه‌ی ثبت جدید درمجموع فیلدهای عنوان و چکیده برابر ۷/۰۵ است. بنابراین، به‌صورت کمی می‌توان عنوان کرد که به‌طور متوسط ۱/۷۵ در شاخص متوسط تعداد نقص در سامانه‌ی ثبت بهبود ایجاد شده است. که با تقسیم به عدد ۸/۸ معادل ۱۹/۸ درصد بهبود در این سامانه مشاهده می‌شود.

۶ بحث و نتیجه‌گیری

در بخش‌های پیشین ضرورت ارزیابی سازمان بر پایه‌ی شاخص‌های کیفیت داده و تأثیر آن بر توسعه‌ی سازمان و تصمیم‌گیری بهتر و اثربخش‌تر مدیران را بیان کردیم. پس از شناسایی نقش بسیار مهم ارزیابی پایگاه داده هر سازمان به کمک شاخص‌های کیفیت داده، معیارهای ارائه‌شده در مقالات پیشین را مرورکرده و روابط توسعه‌یافته‌ی شاخص‌ها توسط هر محقق را بررسی‌کرده و بیان کردیم. با شناخت ابعاد کیفیت داده و شاخص‌های ارزیابی این ابعاد، با توجه به ساختار سلسله‌مراتبی سامانه‌ی گنج شاخص‌هایی مهم و کارامد جهت افزایش عملکرد سامانه مذکور درنظر گرفته و

برای هر شاخص شناسنامه‌ای تعریف کردیم. روابطی جهت پایش شاخص‌های درنظر گرفته‌شده نیز تعریف‌کرده که در روند بهبود مستمر سامانه‌ی بهره نقش بسزایی دارد. نتایج این پژوهش نشان داد می‌توان برای پایگاه‌های داده و سامانه‌های اطلاعاتی مختلف، برپایه‌ی کارکرد آن‌ها و نیاز کاربران شاخص‌های کیفیت داده را توسعه داد. در موردکاوی صورت‌گرفته نتایج نشان داد شاخص‌های کیفیت داده می‌توانند به ارزیابی بازطراحی‌های صورت‌گرفته در سامانه‌ی گنج بپردازند. همچنین می‌توان میزان بهبود صورت‌گرفته در این باز طراحی را در قالب شاخص‌های کیفیت داده به‌خوبی اندازه گرفت. به‌عنوان‌مثال، در مورد میانگین میزان نقص در اقلام اطلاعاتی نتایج نشان داد کیفیت داده‌ی پایگاه مورد مطالعه بهبود قابل‌توجهی (نزدیک به ۲۰ درصد) داشته است.

۷ پیشنهاد پژوهش

در مطالعات و پژوهش‌های آتی در حوزه‌ی کیفیت داده‌های علم و فناوری پیشنهاد می‌شود اقدامات زیر موردتوجه قرار گیرد.

- شاخص‌هایی همچون قابلیت دسترسی، غنای اطلاعات، اصالت و ... می‌توانند جهت ارزیابی سازمان مورداستفاده قرار بگیرند.
- می‌توان شاخص‌هایی هم برای ارزیابی‌های ذهنی تعریف کرد تا افراد، تا حدودی در یک چارچوب خاص تمرکز کنند و تصمیم‌گیری‌های لازم را انجام دهند.
- همچنین می‌توان با وزن‌دهی شاخص‌های کیفیت داده به کمک فنون تصمیم‌گیری اهمیت نسبی آن‌ها را سنجید.

۸ منابع

Arabi, P, Zafar Heidarpour, M, Khoshgftar, A. (2017). Accelerating the achievement of critical success factors and key performance indicators, with an emphasis on the fundamental transformation document in the education research area of Mashhad 6th District. In the International Conference on Modern Approaches in Humanities.

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, 44(4), 462-484.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR),* 41(3), 16.

Chengalur-Smith, I. N., Ballou, D. P., & Pazer, H. L. (1999). The impact of data quality information on decision making: an exploratory analysis. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 853-864

Even, A., & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 38(2), 75-93.

Gasta, W. (2004, April). Statistical process control using key process indicators for vacuum devices. In Vacuum Electronics Conference, 2004. IVEC 2004. *Fifth IEEE International*.

Geekiyanage, Suranga CH, Dan Sui, and Bernt S. Aadnoy. (2018). Drilling Data Quality Management: Case Study With a Laboratory Scale Drilling Rig. ASME 2018 37th International Conference on Ocean, Offshore and Arctic Engineering. *American Society of Mechanical Engineers*.

Heinrich, B., Klier, M., & Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data and Information Quality (JDIQ)*, 1(1):5.

Hoyles, Celia, et al. (2007). Attributing meanings to representations of data: The case of statistical process control. Mathematical Thinking and Learning, 9 (4): 331-360

Jones-Farmer, L. A., Ezell, J. D., & Hazen, B. T. (2014). *Applying control chart methods to enhance data quality. Technometrics*, 56(1): 29-41.

Khosroanjom, D., Ahmadzade, M., Niknafs, A., & Mavi, R. K. (2011). Using fuzzy AHP for evaluating the dimensions of data quality. *International Journal of Business Information Systems*, 8(3): 269-285

Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T., & Batini, C. (2002). Managing data quality in cooperative information systems. *In OTM Confederated International Conferences on the Move to Meaningful Internet Systems"*, 486-502.

Moges, Helen-Tadesse. (2016). Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes—An

فصل‌نامه علمی مدیریت استاندارد و کیفیت
سال نهم- شماره ۱ - پیاپی ۳۱ - بهار ۱۳۹۸

1. Key Performance Indicator
2. Intrinsic
3. Accessibility
4. Contextual
5. Representational
6. Subjective
7. Objective
8. Scalable
9. Meaningful
10. Completeness
11. Accuracy
12. Correct
13. Richness of Information
14. Intrinsic Precision
15. Timeliness Metrics
16. Reconstruction
17. Normalization
18. Interval scale
19. Interpretability
20. Aggregation
21. Adaptivity
22. Feasibility
23. Suitable for use
24. Weighted column completeness
25. Weighted tuple completeness
26. Weighted cell completeness
27. Relevancy
28. percentage
29. count

exploratory study. Decision Support Systems, 83.

Morgan, Chris, and Adam Dewhurst. (2007). Using SPC to measure a national supermarket chain's suppliers' performance. *International Journal of Operations & Production Management* 27(8): 874-900.

Ochoa, X., & Duval, E. (2006). Quality metrics for learning object metadata. In EdMedia+ Innovate Learning (1004-1011). Association for the Advancement of Computing in Education (AACE).

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-83

Suhairi, Kasman, and Ford Lumban Gaol. (2013). The measurement of optimization performance of managed service division with ITIL framework using statistical process control." Journal of Networks, 8(3).

Strong, D., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. Communications of the ACM, 40(5), 103-111.

Vaziri, R., Mohsenzadeh, M., & Habibi, J. (2017). Measuring data quality with weighted metrics. *Total Quality Management & Business Excellence*, 1-13.