بررسی تغییرات حوزه کیفیت داده با استفاده از تحلیل کلمات کلیدی

دو فصلنامه علمی- پژوهشی

مدیریت اطلاعات

دوره ۳، شماره ۲

پاییز و زمستان ۱۳۹۶

احمد خلیلی جعفرآباد

دکتری مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه تهران، تهران، ایران[۱]

چکیده: حوزه کیفیت داده ازجمله حوزه‌های رو به رشد و مهم در حوزه سیستم‌های اطلاعاتی است. شناخت دقیق این حوزه از یک‌سو و شناخت ویژگی‌های زیر حوزه‌های نوین این حوزه برای محققان از اهمیت بالایی برخوردار خواهد بود. شناخت ویژگی این حوزه و میزان بین‌رشته‌ای بودن آن‌ها به محققان برای تصمیم‌گیری در مورد روند تحقیقات و انتخاب حوزه فعالیت کمک بسزایی خواهد کرد. برای شناسایی حوزه‌ها و بررسی میزان بین‌رشته‌ای بودن آن‌ها، در این مطالعه با استفاده از گراف هم رخدادی کلمات و تحلیل ۹۰۰۰ مقاله، ویژگی‌های حوزه‌های نوین مطالعاتی موردبررسی قرارگرفته است. بر اساس این مطالعه مشخص‌شده است که این حوزه‌ها بیشتر بین‌رشته‌ای بوده است و تمرکز آن بر روی ارتباط بین چندین حوزه مطالعاتی است. به‌بیان‌دیگر این کلمات کلیدی بیشتر با کلمات کلیدی موجود در خوشه‌های دیگر ارتباط داشته‌اند تا با کلمات کلیدی که با آن‌ها در یک خوشه قرارگرفته است. بر اساس این مطالعه جدیدترین حوزه مرتبط با کلان داده است که مباحث یکپارچه‌سازی و داده گمشده در این حوزه از اولویت بالاتری نسبت به مابقی حوزه‌ها برخوردار است.

کلیدواژه‌ها: تحلیل شبکه، داده‌کاوی، کیفیت داده، علم داده.

---

۱. نویسنده مسئول: Ahmad.khalili@ut.ac.ir

مقدمه

یکی از موضوعات مهم در مدیریت داده‌ها، بحث کیفیت داده است. داده خوب و باکیفیت مبنای رسیدن به نتایج مطلوب از سیستم‌ها است. به‌بیان‌دیگر اثربخشی مدیریت یک کسب‌وکار در سازمان به‌واسطه کیفیت داده‌های آن سازمان به‌عنوان ماده خام تصمیم‌گیری تعیین خواهد شد (Chaffey and White 2010). سیستم اطلاعاتی همراه با داده بی‌کیفیت، هرچند طراحی و پیاده‌سازی مناسبی داشته باشد نمی‌تواند تصمیم‌گیران و کاربران سیستم‌های اطلاعاتی را به نتایج مطلوب مدنظر برساند؛ بنابراین تمرکز بر روی موضوع کیفیت داده در سیستم‌های اطلاعاتی امری اجتناب‌ناپذیر است.

برای راهبری پژوهش‌های این حوزه نیاز است تا دید مناسبی از این حوزه و تغییرات آن برای محققان ایجاد گردد. علی‌رغم این‌که عمر موضوع کیفیت داده به‌اندازه عمر خود داده است، اما امروزه به دلیل تغییرات وسیع در نوع سیستم‌های اطلاعاتی، کیفیت داده در حال تبدیل‌شدن به یکی از موضوعات استراتژیک سازمان است و در طول یک یا دو دهه گذشته، ماهیت موضوعات مرتبط باکیفیت داده تغییریافته است (Sadiq 2013) و به‌شدت بر پیچیدگی و اهمیت آن افزوده‌شده است. علت اصلی این موضوع تغییر در نگرش به کیفیت داده از نگرش به محتوا به‌سوی نگرش به زمینه داده است (Shankaranarayanan and Blake 2017). این امر به دلایل متفاوتی اتفاق افتاده است که از آن جمله می‌توان به گسترش استفاده از کلان داده در سازمان‌ها اشاره داشت که منجر به تغییرات بسیاری در مدیریت و تحلیل داده‌شده است.

از دید حوزه‌های تحقیقاتی، کیفیت داده به‌عنوان یک موضوع میان‌رشته‌ای بین مدیریت، آمار و علوم کامپیوتر مطرح است (Batini and Scannapieco 2006) و این حوزه کاملاً مرتبط با حوزه علوم کامپیوتر و سیستم‌های اطلاعاتی است و موضوع کلیدی متخصصان هر دو حوزه و به‌خصوص سیستم‌های اطلاعاتی است و به همین دلیل بررسی این حوزه و ابعاد مختلف آن از جنبه‌های مختلف برای متخصصان هر دو حوزه حائز اهمیت است (Sadiq 2013).

با توجه به این‌که احتمال دستیابی به پیشرفت‌های چشم‌گیر در حوزه‌های بین‌رشته‌ای بسیار بالاتر است (Porter, Roessner and Henriques 2008)، شناسایی حوزه‌های میان‌رشته‌ای از اهمیت بالایی برخوردار است. به همین دلیل از ابتدای دهه ۹۰ میلادی مطالعات بسیاری در مورد بین‌رشته‌ای بودن حوزه‌ها و شیوه کشف آن‌ها به انجام رسیده است. معمولاً برای سنجش میان‌رشته‌ای بودن یک حوزه علمی دو رویکرد کلی وجود دارد که یکی از کلاس‌های از پیش تعیین‌شده توسط WOS استفاده می‌کند و دیگری از رویکرد پایین به بالا که از پیش کلاس‌ها نامشخص است (Rafols and Meyer 2009) که در این مقاله به دلیل این‌که حوزه مطالعه یک زیر حوزه سیستم‌های اطلاعاتی است، عملاً باید از رویکرد دوم استفاده نمود.

اگرچه تاکنون مطالعات زیادی بر روی حوزه کیفیت داده به انجام رسیده است و هر یک به‌زعم خود تلاشی برای ارائه دیدی کلی از این حوزه انجام داده است؛ اما تاکنون مطالعه مشخصی در موردبررسی میان‌رشته‌ای بودن این حوزه و زیر حوزه‌های آن به انجام نرسیده است. فلذا در این پژوهش قصد داریم با استفاده از روش‌های مبتنی بر نقشه علم و داده‌کاوی، زیر حوزه‌های این علم را کشف نماییم و زیر حوزه‌های این حوزه بخش را به جهت میان‌رشته‌ای بودن مورد ارزیابی قرار دهیم.

مرور ادبیات کیفیت داده

کیفیت داده به حوزه‌های مختلفی شامل آمار، مدیریت و علوم کامپیوتر ارتباط دارد. محققان علم آمار اولین گروهی بودند که برخی از مسائل کیفیت داده را موردبررسی قراردادند. آن‌ها با ارائه تئوری‌های ریاضی در اواخر دهه ۱۹۶۰ راهکارهایی برای یافتن داده‌های تکراری در مجموعه داده‌ها ارائه کردند. پس از آن‌ها محققان علوم مدیریت در دهه ۱۹۸۰ متمرکز حذف مشکلات کیفیت داده در فرایندهای تولید داده و سیستم‌های مرتبط با آن شدند. در دهه ۱۹۹۰ نیز محققان علوم کامپیوتر اقدام به تعریف، اندازه‌گیری و بهبود کیفیت داده‌های الکترونیکی در بانک‌های اطلاعاتی و انباره‌های داده نمودند(Batini and Scannapieca 2006).

اما اگر بخواهیم نگاه تاریخی دقیق‌تر به این حوزه داشته باشیم می‌توان شروع این حوزه به‌صورت رسمی را به دهه ۸۰ نسبت داد. حوزه مدیریت کیفیت داده برای نخستین بار در دهه ۸۰ میلادی و توسط برودی ارائه گردید. او نشان داد که اهمیت حوزه‌های سازمانی به‌اندازه حوزه‌های فنی مدیریت کیفیت داده است. او همچنین تأکید کرد که کیفیت داده بدون توجه به هر دو بعد ذکرشده یعنی حوزه‌های سازمانی و فنی اتفاق نخواهد افتاد (Brodie 1980).

پس‌ازآن جدی‌ترین اقدام این حوزه در آغاز را می‌توان وابسته به دانشگاه MIT دانست که از سال ۱۹۹۰ و با راه‌اندازی گروهی تحقیقاتی در دانشگاه MIT به حوزه علوم کامپیوتر مطرح شد. محققان دانشگاه MIT بر این باور بودند که در حوزه مسائل مرتبط باکیفیت داده مطالعات زیادی باید انجام شود و در این راستا نیز اقدامات بسیاری را به انجام رساندند (Batini and Scannapieca 2006). پیام گروه تحقیقاتی دانشگاه MIT این است که داده‌ای باکیفیت است که مناسب برای استفاده باشد و تناسب برای استفاده مفهومی بسیاری فراتر از دقت اطلاعات است. به‌عبارت‌دیگر داده‌ای باکیفیت است که برای استفاده مخاطب مناسب باشد و این تناسب بسیار پیچیده‌تر از دقت است. مثلاً داده مناسب باید در زمان مناسب و به‌صورت کامل ارائه شود. این دانشگاه در سالیان گذشته همواره تلاش کرده است تا تحت عنوان مدیریت اطلاعات برای کسب‌وکارها، مجموعه رهنمودهایی را برای توجه دادن کسب‌وکارها به استفاده‌کنندگان اطلاعات ارائه نماید. هدف از این کار این است که داده به‌عنوان یک محصول ارزشمند و مهم در کسب‌وکارها برای استفاده‌کنندگان داخلی یا خارجی سازمان شناخته شود (Batini and Scannapieca 2006). این کار با مقاله تأثیرگذار وانگ و استرانگ (۱۹۹۶) که ابعاد مختلف کیفیت داده را موردبررسی قرار داده بودند ادامه یافت. در این مقاله کیفیت داده‌ها از ابعاد مختلف چهارگانه کیفیت ذاتی داده‌ها، کیفیت زمینه‌ای داده‌ها، قابلیت بازنمایی داده‌ها و قابلیت دسترسی داده‌ها ارائه‌شده است (Wang and Strong 1996). بسیاری از مطالعات دیگر در حوزه مدیریت کیفیت داده و اطلاعات بر پایه مطالعه وانگ و استرانگ[1] (۱۹۹۶) بنا گذاشته‌شده است.

۱.Wang and Strong

همان‌طور که پیش‌تر نیز به آن اشاره شد مدیریت کیفیت داده و اطلاعات یک حوزه چندبخشی است و محققان حوزه کامپیوتر و سیستم‌های اطلاعاتی هر یک‌بر بخش خاصی از آن مانند نگاه بانک اطلاعاتی، داده‌کاوی، مدیریت و یا زمینه خاص کاری تمرکز کرده‌اند. به همین دلیل تاکنون، پژوهش‌های متعددی به‌منظور تعیین دامنه‌ی کیفیت داده و چارچوب‌بندی آن صورت گرفته است که اولاً با نگاه به یک حوزه خاص از کیفیت داده تدوین‌شده‌اند و ثانیاً بیشتر آن‌ها با استفاده از روش‌های مبتنی بر خبرگان، سعی در مرزبندی این حوزه داشته‌اند.

ازجمله قدیمی‌ترین پژوهش‌ها در رابطه با چارچوب‌بندی کیفیت داده، مطالعه وانگ در سال ۱۹۹۵ است. در این مقاله با نگاه به داده به‌مثابه یک محصول و استفاده از هفت المان ایزو ۹۰۰۰، چارچوبی برای کیفیت داده ارائه و حدود ۱۰۰ مقاله‌ی مرتبط تا قبل از سال ۱۹۹۴ با قضاوت انسانی، به‌وسیله‌ی این چهارچوب دسته‌بندی‌شده است. در جدول یک چارچوب وانگ قابل‌مشاهده است:

جدول ۱. چارچوب وانگ در حوزه کیفیت داده

المان	تعریف
مسئولیت‌های مدیریتی	• توسعه‌ی سیاست کیفیت داده • راه‌اندازی سیستم کیفیت داده
هزینه‌های اجرا و ضمانت	• هزینه‌های اجرایی شامل هزینه‌های پیشگیری، ارزیابی قیمت و شکست • هزینه‌های ضمانت که اثبات کیفیتِ مطابق با معیارهای مشتری و مدیریت است
تحقیق و توسعه	• تعریف ابعاد کیفیت داده و اندازه‌گیری مقادیر آن‌ها • تحلیل و طراحی جنبه‌های کیفیتی محصولات داده‌ای • طراحی سیستم‌های تولید داده که تمام ابعاد کیفیت داده را تجمیع می‌کند
تولید	• نیازمندی‌های کیفیت، در تهیه‌ی داده‌ی خام، اجزا و سرهم‌بندی کردن به‌منظور تولید محصولات داده‌ای • تائید کیفیت داده‌های خام، داده‌های در حال ساخت و محصولات داده‌ای نهایی • شناسایی داده‌هایی که تطابق با مشخصات خواسته‌شده را ندارند و طراحی عملیات اصلاحی
توزیع	• ذخیره‌سازی، شناسایی، بسته‌بندی، نصب، تحویل و خدمات پس از فروش برای محصولات داده‌ای • مستندسازی دقیق اطلاعات محصولات داده‌ای
مدیریت افراد	• آگاهی بخشی به کارکنان در مورد مسائل مربوط به کیفیت داده • انگیزش کارکنان برای تولید محصولات داده‌ای باکیفیت • ارزیابی عملکرد کارکنان دررسیدن به اهداف بالا
مسائل قانونی	• ایمنی و مسئولیت‌پذیری در مورد محصولات داده‌ای

در سال ۲۰۰۶، در پژوهش دیگری، سعی شده است فقدان بدنه‌ی تئوری مستحکم برای کیفیت داده، با مرزبندی این حوزه جبران شود. در این پژوهش، با استفاده از ۱۷۱ مقاله در حوزه‌ی کیفیت داده و اطلاعات، ۲۷۹ کلمه‌ی کلیدی استخراج شد تا به‌وسیله‌ی آن‌ها، یک مدل مفهومی[1] از حوزه‌ی کیفیت داده ارائه شود. مدل مفهومی، دارای سه نمای مختلف، اجرایی، رفتاری و سازمانی است. درنتیجه، این مقاله، با استفاده از کلمات کلیدی استخراج‌شده، سه نمای مفهومی جداگانه به نمایش گذاشته است (Lima, Maçada and Vargas 2006).

در مطالعه دیگری که به انجام رسیده است، مدنیک در سال ۲۰۰۹ کیفیت داده را به چهار زیر حوزه مطالعاتی تأثیر کیفیت داده، مسائل فنی مرتبط با بانک اطلاعاتی، کیفیت داده در زمینه علوم کامپیوتر و فناوری اطلاعات و اصلاح کیفیت داده‌ها تقسیم کرده است(Madnick & et al. 2009). البته او تصریح کرده است که به دنبال ارائه‌ی یک نمای جامع از کیفیت داده نبوده است و تنها هدف آن اشاره به مهم‌ترین و برجسته‌ترین موضوعات این حوزه بوده است.

در مقاله دیگری در سال ۲۰۰۸ نیز نگاه محصول محور به داده و اطلاعات اتخاذشده است. در این مقاله با ترکیب نگاه جوران در مورد تناسب استفاده و چارچوب وانگ چارچوب جدیدی در این حوزه ارائه‌شده است. نتایج این مقاله نشان می‌دهد بیشترین تعدادِ پژوهش‌ها در چارچوب وانگ، به ترتیب مربوط به المان‌های تحقیق و توسعه (ابعاد و اندازه‌گیری)، مسئولیت‌های مدیریتی و توزیع بوده است. هم‌چنین فاکتورهای چه چیزی و چطور در مقاله‌ی جوران، بیشترین تعداد مقاله را داشته است؛ بنابراین بلوک‌هایی که از برخورد این فاکتورها با المان‌ها ایجادشده‌اند، بیشترین میزان توجه را از سوی پژوهشگران این حوزه داشته‌اند (Neely and Cook 2008).

گروه دیگری از مقالات نیز بر جنبه‌های فنی تمرکز داشته‌اند. در سال ۲۰۰۰ نظام رده‌بندی مشکلات کیفیت داده حوزه تمیز کردن داده توسط رهم ارائه گردید. بر اساس این مقاله مشکلات در دو حوزه داده‌های از یک منبع داده و داده‌های مرتبط با چند منبع داده تقسیم‌بندی شده است. این مقاله همچنین مشکلات هر گروه را در سطح شما[2] و در سطح نمونه موردبررسی قرار داده است و به ابزارها و روش‌های رفع این مشکلات اشاره‌کرده است (Rahm and Do 2000).

پس‌ازآن در سال ۲۰۰۳ کیم و همکارانش یک تقسیم‌بندی جامع از داده کثیف ارائه کردند. هدف آن‌ها آماده کردن یک چارچوب برای فهم شیوه به وجود آمدن داده کثیف بود. آن‌ها همچنین تلاش کردند مسائلی را که در حین تمیز کردن داده باید موردتوجه قرار گیرد را مطالعه نمایند. در همین راستا آن‌ها نظام رده‌بندی خود را که نشان‌دهنده ۳۳ نوع داده کثیف بود ارائه کردند. نظام رده‌بندی کیم در ریشه خود دارای دو نوع داده است که شامل داده مفقود و داده غیر مفقود است و سپس در سطوح بعدی سعی در

1.Conceptual Map
۲.Schema

تدقیق این دسته‌بندی کرده است. ازجمله نکات مهم این کتسونومی واردکردن مبحث زمان در مورد داده‌ها بی‌کیفیت است که زمان تولید و یا زمانی که داده صحت دارد را شامل می‌گردد (Kim & et al. 2003). در سال ۲۰۰۵ اولیویرا در پژوهش خود که به لحاظ ساختاری شباهت زیادی به مطالعه رهم دارد، کلیه مشکلات کیفیت داده را به چهار دسته تقسیم کرده است. در این تقسیم‌بندی که به‌صورت دودویی صورت گرفته است، مشکلات در مورد یک منبع داده[۱]، مشکلات در سطح رابطه تکی[۲]، مشکلات ناشی از روابط چندگانه[۳] مشکلات ناشی از منابع داده چندگانه[۴] در دو نوع بالا نیز موردبررسی قرارگرفته است (Oliveira, Rodrigues and Henriques 2005). در مطالعه دیگری که توسط اولیویرا انجام‌شده است نیز با همین نگاه مشکلات کیفیت داده در چند سطح شناسایی‌شده است. در این مطالعه به ۲۰ مشکل در کیفیت داده و در چهار بخش ویژگی، رابطه تکی، رابطه چندگانه و منابع داده چندگانه شناسایی‌شده است(Oliveira, Rodrigues and Henriques 2005).

در یک نظام رده‌بندی دیگر در سال ۲۰۱۲ به داده‌های زمان‌مبنا به‌عنوان یک نوع داده‌ای خاص تمرکز شده است. این پژوهش با الهام گرفتن از مطالعات رهم[۵] و همچنین مطالعات دیگری که پیش از آن انجام‌شده است داده‌های کثیف بر محور زمان را موردبررسی قرار داده است. در این مطالعه مانند مطالعه رهم منابع داده واحد و یا چندگانه از یکدیگر جداشده است و در ذیل هر مورد مشکلات مختلف موردبررسی قرارگرفته است (Gschwandtner et al. 2012). در یک مقاله مهم دیگر موضوع تولید داده بدون کیفیت بر اساس منشأ موردبررسی قرارگرفته است. موضوع خاستگاه داده ازجمله موضوعات نوین در حوزه مدیریت داده و کیفیت داده است. در این پژوهش یک نظام رده‌بندی برای موضوع خاستگاه داده ارائه‌شده است که استفاده از خاستگاه، موضوع خاستگاه، نمایش خاستگاه، ذخیره کردن خاستگاه داده و انتشار خاستگاه داده ازجمله مواردی است که موردبحث قرارگرفته است (Simmhan et al. 2005). این پژوهش به افراد علاقه‌مند به حوزه مطالعه و تحلیل فراداده و همچنین مدیریت داده دید بسیار مناسبی برای شناخت خاستگاه داده و ثبت آن ایجاد خواهد کرد.

در یک مطالعه دیگر برخی محققان مشکلات کیفیت داده را به سه نوع مشکلات ناشی از محتوا، فرم و زمان تقسیم کرده‌اند (Kim, Kishore and Sanders, 2005). بعد محتوایی به مسائل مرتبط باکیفیت ذاتی محتوا اشاره دارد که به رساندن داده دقیق، مرتبط و کامل به مخاطب اشاره دارد. بعد فرم به ارائه اطلاعات و بهبود فهم کاربر از اطلاعات ارائه‌شده ارتباط دارد و بعد زمان نیز به مسائل مرتبط با رساندن داده به کاربر اشاره دارد (Kim, Kishore and Sanders, 2005).

همچنین در یک مطالعه مشابه یک ساختار دوبعدی هستی شناسانه ارائه‌شده است که یکی بر ابعاد اساسی شامل کامل بودن، غیرمبهم بودن، صحیح بودن تمرکز دارد و بعد دوم بر انتزاع از سطح اول که

1.Attribute/Tuple
2.Single Relation
3.Multiple Relation
4.Multiple Data Sources
۵.Rahm

شامل اسکیما و نمونه است تمرکز دارد (Du and Zhou 2012). با توجه به تمرکز این پژوهش بر حوزه مالی، بیشتر تمرکز پژوهش بر روی تصمیم‌گیری در مسائل مالی بوده است و نظام رده‌بندی ارائه‌شده از گستردگی کافی برخوردار نیست.

در مطالعه دیگری نیز کیفیت داده به شش گروه از مطالعات تقسیم‌بندی گردید و تحولات آن‌ها در طول سالیان موردبررسی قرار گرفت. بر اساس بررسی انجام‌شده در این تحقیق مشخص شد که حوزه فنّاوری‌های محاسباتی مربوط به کیفیت داده از حوزه‌های با بالاترین سرعت رشد است (Zhang et al. 2013).

در یک مطالعه جامع دیگر که از مقالات مهم این حوزه به شمار می‌رود، این حوزه به سه سؤال اصلی پاسخ‌داده‌شده است که در آن در مورد سنجش کیفیت داده، شیوه مدیریت کیفیت داده و همچنین اثر کیفیت داده در سازمان بحث شده است (Helfert and Ge 2006). در حوزه سنجش کیفیت داده، سه زیر حوزه مشکلات کیفیت داده، ابعاد کیفیت داده و متدولوژی‌های سنجش کیفیت داده قرار داده‌شده است. مدیریت کیفیت داده ترکیبی از سه موضوع مدیریت کیفیت، مدیریت اطلاعات و مدیریت دانش است. در حوزه زمینه کیفیت داده نیز این حوزه بر زمینه‌های مختلفی در سازمان اثرگذار خواهد بود که در این مطالعه به دو حوزه سیستم اطلاعاتی و تصمیم‌گیری پرداخته‌شده است. درنهایت این مطالعه وضعیت روز کیفیت داده در هر یک از موضوعات را بررسی و تبیین کرده است (Helfert and Ge 2006).

در سالیان اخیر گروهی از محققان تلاش‌هایی را برای ارائه کردن نظام رده‌بندی بر اساس روش‌های کمی انجام داده‌اند. یکی از مهم‌ترین مطالعات انجام‌شده با این روش‌ها توسط راجر بلیک انجام‌شده است. در این مطالعه با به‌کارگیری روش LSA[1] به‌منظور کشف حوزه‌های مختلف در ارتباط باکیفیت داده استفاده‌شده است (Blake 2010). در این مقاله شش حوزه اصلی به نام‌های ارزیابی کیفیت داده، مدیریت کیفیت داده، تأثیر کیفیت داده بر سطوح مختلف سازمانی، بانک اطلاعاتی و کیفیت داده، تأثیر کیفیت داده بر تصمیم‌گیری و کاربردهای کیفیت داده اشاره‌شده است و هر یک به زیر حوزه‌هایی شکسته شده است که درمجموع ۱۵ حوزه کلی را شکل داده است. این محقق با انجام مطالعه مشابهی در سال ۲۰۱۲ که ازلحاظ روش کاملاً مشابه مطالعه سال ۲۰۱۰ بود به شش حوزه کلی دست‌یافت که در جدول زیر این حوزه‌ها قابل‌مشاهده است (Blake and Shankaranarayanan 2012). در جدول دو خروجی اصلی این مطالعه قابل‌مشاهده است:

جدول ۲. مدل ارائه‌شده توسط راجر بلیک و شانکارانایانان (۲۰۱۲) از مسائل اصلی حوزه کیفیت داده

موضوع اصلی	تم اصلی
تناسب استفاده	ارزیابی کیفیت داده
شاخص‌ها	
تولید داده	مدیریت کیفیت داده
فرایندهای کیفیت داده و بهبود آن‌ها	

1.Latent Semantic Analysis

موضوع اصلی	تم اصلی
کیفیت داده در سازمان	کیفیت داده در سازمان
طراحی بانک داده و داده‌کاوی	کیفیت داده در بانک اطلاعاتی
پرس‌وجو در داده و تمیز کردن	
یکپارچه‌سازی داده	
سیستم پشتیبان تصمیم	کیفیت داده و تصمیم‌گیری
اقتصاد کیفیت داده	
کیفیت داده در وب	کاربرد کیفیت داده
کیفیت داده برای تحلیل	

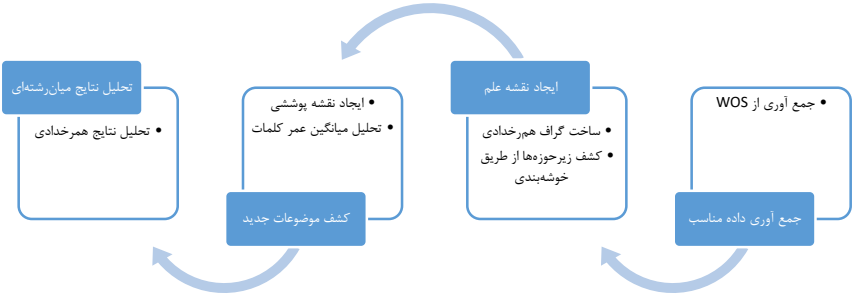
همچنین در دو مقاله پی‌درپی نیز به بررسی تغییرات این حوزه پرداخته است که در مقاله اخیر خود در سال ۲۰۱۷ با اشاره به تغییرات اساسی حوزه کیفیت داده به بررسی تغییرات این حوزه پرداخته است (Shankaranarayanan and Blake 2017). در این مقاله حوزه‌هایی که در حال حاضر در کیفیت داده از اهمیت برخوردار هستند موردبررسی قرارگرفته است. ازجمله یافته‌های این مقاله این است که موضوعات کیفیت داده در طول زمان ثابت بوده است و تنها در زمینه‌های مختلف اهمیت پیداکرده است.

در مطالعه جامع دیگری با بررسی کلیدواژه‌های مقالات به چاپ رسیده، یک نظام رده‌بندی برای کیفیت داده ارائه‌شده است و تغییرات این حوزه در طول ۲۰ سال مورد ارزیابی قرارگرفته است (Sadiq, Yeganeh and Indulska 2011) در این پژوهش ۱۸ حوزه اصلی برای کیفیت داده مطرح‌شده است که حاصل بررسی کلمات کلیدی مقالات موردبررسی در این پژوهش است. این پژوهش ازجمله کامل‌ترین پژوهش‌های انجام‌شده در حوزه کیفیت داده است که با استفاده از تحلیل کلمات کلیدی حدود ۵۰۰۰ هزار مقاله اقدام به شناسایی حوزه‌های اصلی این حوزه نموده است.

همان‌طور که از بررسی ادبیات کیفیت داده مشخص است، تاکنون مطالعه‌ای در مورد میان‌رشته‌ای بودن این حوزه به انجام نرسیده است و تمام مطالعات انجام‌شده تمرکز خود را بر کشف محدوده این علم و تعریف زیر حوزه‌های آن گذاشته‌اند. به همین دلیل تاکنون مطالعه دقیقی در مورد میان‌رشته‌ای بودن این حوزه به انجام نرسیده است. فلذا نیاز است تا بامطالعه دقیق این حوزه میان‌رشته‌ای بودن یا نبودن و روند آن در این حوزه موردبررسی قرار گیرد.

روش انجام پژوهش

به‌منظور انجام این پژوهش از یک روش چندمرحله‌ای مبتنی بر ساخت گراف هم رخدادی کلمات کلیدی استفاده‌شده است. در این روش با استفاده از تحلیل خصوصیات ساختاری گراف هم رخدادی کلمات کلیدی یافته‌هایی در مورد ویژگی‌های حوزه‌های نوین این حوزه استخراج و موردبحث قرارگرفته است. برای مشخص‌تر شدن روش انجام کار شکل یک روش کلی انجام این پژوهش را به نمایش گذاشته است.



شکل ۱. روند انجام پژوهش

۱. جمع‌آوری داده مناسب

برای انجام این پژوهش کلیه مقالات به چاپ رسیده در بین سال‌های ۱۹۷۰ تا سال ۲۰۱۶ در نشریات تحت پوشش Web Of Science انتخاب‌شده است. این مقالات شامل کلیه مقالات حوزه کیفیت داده محدود به علوم کامپیوتر و سیستم‌های اطلاعاتی است. برای شناسایی مقالات مختلف از کلمات کلیدی این حوزه که از نظام‌های طبقه‌بندی استخراج‌شده است استفاده شد (Madnick and et al. 2009; Khalilijafarabad, Helfer and Ge 2016). البته لازم به ذکر است برای استخراج کلمات کلیدی آن دسته از کلماتی که عمومیت بسیاری داشته و مقالات نامرتبط زیادی را وارد داده‌ها می‌کردند حذف‌شده است. شکل دو نشان دهنده پرس‌وجوی استفاده‌شده برای شناسایی مقالات مرتبط در WOS است.

You searched for: TOPIC: ("data quality") OR TOPIC: ("information quality") OR TOPIC: ("content quality") OR TOPIC: ("data consistency") OR TOPIC: ("linkage")

Refined by: WEB OF SCIENCE CATEGORIES: (MANAGEMENT OR COMPUTER SCIENCE INFORMATION SYSTEMS OR BUSINESS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR INFORMATION SCIENCE LIBRARY SCIENCE) AND WEB OF SCIENCE CATEGORIES: (COMPUTER SCIENCE INFORMATION SYSTEMS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR BUSINESS OR INFORMATION SCIENCE LIBRARY SCIENCE OR COMPUTER SCIENCE SOFTWARE ENGINEERING OR OPERATIONS RESEARCH MANAGEMENT SCIENCE OR STATISTICS PROBABILITY)

Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI.

شکل ۲. پرس و جوی استفاده‌شده برای استخراج مقالات مرتبط باکیفیت داده

نتیجه این جستجو دستیابی به بیش از ۹۰۰۰ مقاله است. این مقالات از طریق سایت WOS و در ساختار اطلاعاتی txt ذخیره‌سازی شده است.

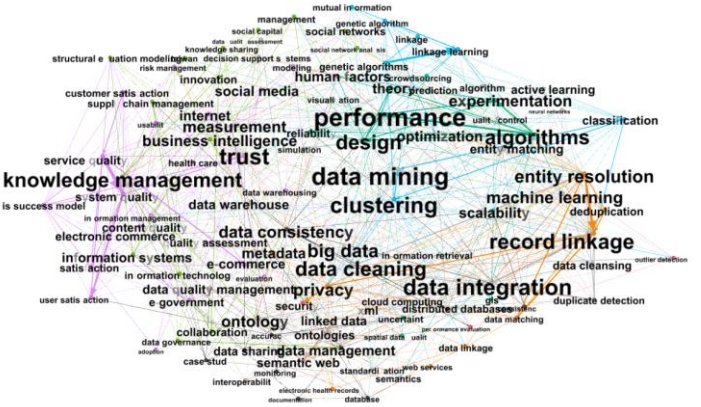
۲. ساخت نقشه علم

یکی از گام‌های پیچیده برای بررسی تغییرات علمی ارائه مدلی پایه است که بتواند تغییرات علمی را نشان دهد. برای این منظور معمولاً دانشمندان از نقشه علم استفاده می‌کنند. این نقشه می‌تواند ساختار و ارتباط فی‌مابین اجزا را که در این پژوهش مقالات هستند را به نمایش دربیاورد (Morris and Van der Veer Martens 2008). برای رسیدن به نقشه علم لازم است دو گام اساسی برداشته شود. گام نخست ساخت یک نقشه پایه و گام بعد کشف زیر حوزه‌های موجود در نقشه است.

۲,۱. ساخت گراف هم رخدادی

برای ساخت نقشه علم از روش‌های بسیار متنوعی استفاده می‌شود که هرکدام یکی از انواع داده‌های کتاب‌شناختی مقالات مانند نویسنده، متن، کلمات کلیدی یا مراجع را مدنظر قرار می‌دهند (Qi 2016). یکی از رایج‌ترین روش‌های مورداستفاده برای ساخت این نقشه‌ها استفاده از هم رخدادی کلمات یا هم رخدادی کلمات کلیدی است. نقشه علم مبتنی بر هم رخدادی کلمات به این صورت است که اگر کلمه کلیدی A و کلمه کلیدی B در یک مقاله باهم توسط نویسنده پیشنهادشده باشد، آنگاه ارتباطی بین این دو ایجاد خواهد شد که نشان از شباهت کاربرد این دو کلمه کلیدی در مقالات دارد. همچنین در صورت تکرار فروانی تکرار نیز به‌عنوان وزن ارتباط این دو در نظر گرفته خواهد شد.

در این مرحله با استخراج کلمات کلیدی موجود در مقالات و محاسبه تعداد تکرار این کلمات با یکدیگر یک گراف هم رخدادی کلمات کلیدی استخراج گردید. این گراف از ۱۶۰۰۰ کلمه کلیدی ساخته‌شده است که برای آسودگی در تحلیل کلمات کلیدی که کمتر از پنج بار درمجموع مقالات مورداستفاده قرارگرفته بودند حذف گردید و درنهایت یک گراف با ۵۰۰ کلمه کلیدی پرکاربرد ساخته شد که در شکل سه قابل‌ملاحظه است. همان‌طور که در شکل نیز قابل‌مشاهده است، موضوعات اصلی در حوزه کیفیت داده در این گراف قابل‌مشاهده است. حوزه‌هایی مانند یکپارچه‌سازی داده، تمیز سازی داده، ارتباط دادن رکوردهای اطلاعاتی ازجمله موضوعات مهم در مبحث کیفیت داده است.

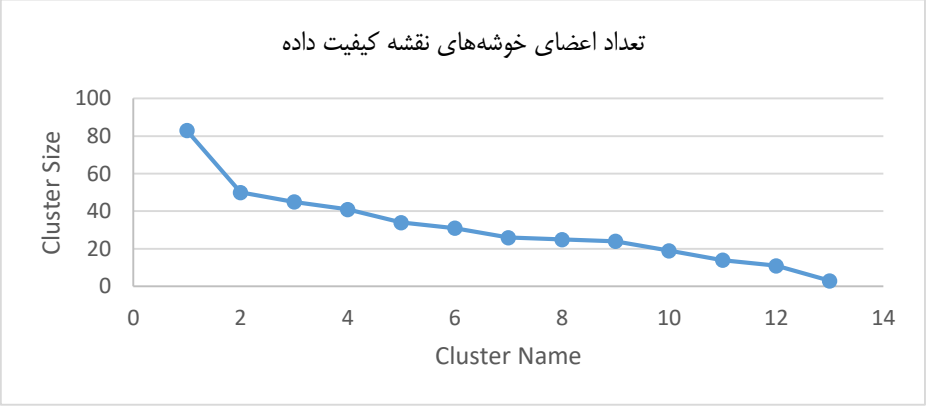


شکل ۳. گراف نقشه علم کیفیت داده

۲,۲. کشف زیر حوزه‌ها از طریق خوشه‌بندی

در گام دوم کشف مجامع علمی یا زیر حوزه‌های علمی قرار دارد که برای اکتشاف آن روش‌های بسیار متنوعی وجود دارد. در این مقاله از مدل ماژولاریتی لواین[۱] استفاده‌شده است (Blondel et al. 2008). که ازجمله مدل‌های دقیق و سریع برای کشف اجتماع در شبکه‌های پیچیده است.

این روش که بر مبنای ماژولاریتی در شبکه است ازجمله روش‌های مناسب و سریع برای کشف مجامع است. در این بخش برای هر دوره زمانی یک‌بار مدل لواین اجراشده است. الگوریتم لوواین با رویکرد دستیابی به‌سرعت بالا جهت تشخیص اجتماع‌ها می‌کوشد تا با استفاده از سنجه ماژولاریتی، گره‌ها را در پایین‌ترین سطح (که هر یک از آن‌ها در یک اجتماع قرار دارند) در اجتماعات کوچک اولیه باهم ادغام نماید. آنگاه در لایه بالاتر اجتماع‌های کوچک را به‌عنوان گره در نظر گرفته و رویه بالا را روی آن‌ها انجام می‌دهد. این الگوریتم در مرحله بعد اجتماع‌های ممکن میان اجتماع‌های جدید را می‌یابد و این کار تا زمانی که امکان تشکیل اجتماع دیگری نباشد ادامه پیدا می‌کند. بدین ترتیب یک ساختار سلسله مراتبی ایجاد می‌گردد و به دلیل کاهش اندازه مسئله در هر لایه، سرعت الگوریتم جهت تشکیل اجتماع‌های نهایی به‌شدت بالا می‌رود. با استفاده از مدل لواین ۱۴ خوشه موضوعی برای حوزه کیفیت داده استخراج‌شده است که در شکل چهار اندازه هر خوشه قابل‌مشاهده است. همان‌طور که در شکل نیز مشخص‌شده است، بیشترین تعداد عضو برای یک خوشه ۸۳ و کمترین تعداد عضو سه است که نشان‌دهنده بزرگ‌ترین و کوچک‌ترین خوشه یا زیر حوزه شناسایی‌شده است.



شکل ۴. تعداد اعضای هر خوشه از نقشه کیفیت داده

۱.Louvain Modularity

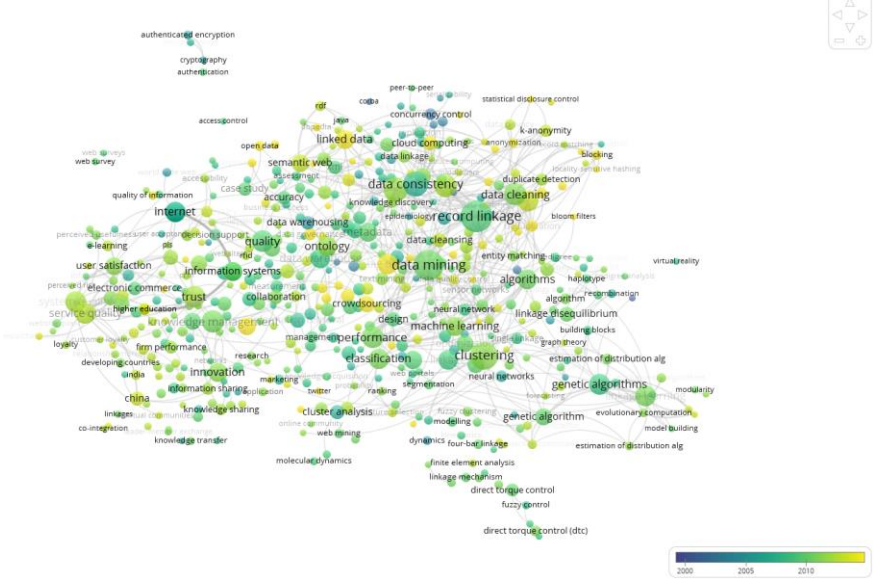
۳. کشف موضوعات جدید

برای تحلیل موضوعات جدید از دو بخش مختلف استفاده‌شده است که بخش اول استفاده از نقشه پوششی و بخش دوم تحلیل میانگین عمر کلمات کلیدی است.

۳,۱. ایجاد نقشه پوششی

یکی از روش‌ها برای شناسایی تغییرات حوزه‌های علمی و بررسی تحولات آن‌ها استفاده از نقشه‌های پوششی است. نقشه‌های پوششی نقشه‌هایی هستند که حاصل ترکیب دو نقشه با یکدیگر هستند. به‌عنوان‌مثال هنگامی‌که می‌خواهیم نقش زمان را بر روی یک نقشه علم نمایش دهیم از این نقشه‌های استفاده می‌کنیم (Rafols, Porter and Leydesdorff 2010).

با استفاده از این نقشه‌ها می‌توان میزان استفاده از لغات را برحسب زمان بررسی نماییم و همچنین میانگین زمان استفاده از هر لغت را نیز بررسی کنیم. این بررسی می‌تواند جدیدترین کلیدواژه‌های هر حوزه علمی را برای محققان نمایان کند. خروجی این مدل در شکل پنج قابل‌مشاهده است. لازم به ذکر است این نقشه با استفاده از نرم‌افزار VOSVIEWR ایجادشده است.



شکل ۵. نقشه پوششی حوزه کیفیت داده برحسب زمان

۳,۲. تحلیل میانگین عمر کلمات

در این مرحله کلیدواژه‌هایی که در سالیان اخیر در حوزه کیفیت داده موردتوجه قرارگرفته‌اند انتخاب گردید و در هرکدام دو مؤلفه موردبررسی قرار گرفت. در جدول زیر ۱۰ لغت برتر که جدیدترین موارد استفاده را داشته‌اند به همراه سن استفاده ارائه‌شده است. این جدول نشان‌دهنده این موضوع است که جدیدترین

موضوعات حوزه کیفیت داده چه موضوعاتی هستند. از کلمات این جدول به‌عنوان ورودی تحلیل فاز بعدی استفاده‌شده است.

جدول ۳. کلمات کلیدی و میانگین عمر استفاده از آن‌ها

کلمه کلیدی	میانگین سال استفاده از کلمه
Big data	2014.7959
Data governance	2014.4118
Social media	2014.2286
Quality assessment	2014.0476
Crowdsourcing	2014
Deduplication	2014
Linked data	2014
Cloud computing	2013.5417
Data matching	2013.5
Standardization	2013.3

۴. تحلیل نتایج میان‌رشته‌ای

برای بررسی میزان بین‌رشته‌ای بودن یک حوزه علمی، معمولاً به میزان ارتباطات داخلی و میزان ارتباطات بین‌حوزه‌ای آن توجه می‌شود. برای این کار معمولاً دو شاخص انسجام و تنوع موردبررسی قرار می‌گیرد که از دقت بالایی برخوردار است اما معمولاً در سطح کلان و در رویکرد از بالا به پایین مناسب است (Rafols and Meyer 2009). در این پژوهش با استفاده از الگوی انسجام سعی شده است میزان بین‌رشته‌ای بودن حوزه‌های نوین کیفیت داده مورد ارزیابی قرار گیرد. مؤلفه اول چیدمان لغاتی که حول هر کلمه کلیدی که از مرحله جزء ۱۰ لغت اول بود را ارزیابی می‌کند. این بخش شامل شناسایی لغاتی است که با لغت موردنظر در یک خوشه مشترک قرارگرفته است. دومین مؤلفه مورد نشان می‌دهد که کلمه شناسایی‌شده چقدر با کلمات کلیدی سایر خوشه‌ها مرتبط است.

همان‌طور که مشخص است هر چه اشتراک این دو مؤلفه پایین‌تر باشد، به معنای میان‌رشته‌ای‌تر بودن آن کلمات کلیدی است.

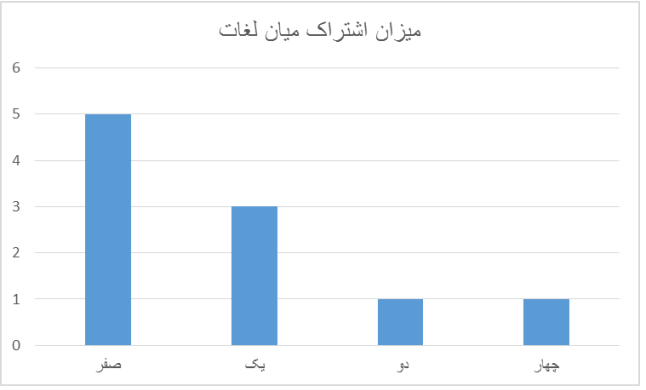
به‌منظور بررسی نتایج ابتدا خوشه‌های مرتبط با هر کلیدواژه موردبررسی قرار گرفت و سپس کلماتی که بیشترین هم رخدادی را با این کلمه داشتند موردبررسی قرارگرفته است که جدول چهار نشان‌دهنده این نتایج است:

جدول ۴. کلمات کلیدی جدید و کلمات با هم رخدادی بالا و درون خوشه‌ای

کلمات کلیدی	کلمات با خوشه مشابه	کلمات با بیشترین هم رخدادی
big data	Self- data cleansing, data mining, organizing map, Unsupervised learning, Semi supervised learning, visualization, xml, decision tree	Meta data, Data Cleansing, Missing Data, Data Integration, Cloud Computing
data governance	business intelligence, Information system, collaboration, innovation,	Data management

کلمات کلیدی	کلمات با خوشه مشابه	کلمات با بیشترین هم رخدادی
	knowledge management, knowledge sharing, social capital, social network analysis, structural equation modeling	
social media	Adoption, content quality,customer satisfaction, data warehousing, e-commerce, e-government, evaluation, information management, internet, is success model, measurement, modeling, risk, service quality, system quality, trust, usability, user satisfaction	Component, emergencies, facebook
quality assessment	data integration, data quality management, interoperability, linked data, ontologies, ontology, semantic web, semantics, standardization	Meta data
crowdsourcing	data integration, data quality management, interoperability, linked data, ontologies, ontology, quality assessment, semantic web, semantics, standardization	Data cleansing, citizen science
Deduplication	Algorithms, crowdsourcing, Design, experimentation, GIS, performance, Reliability, spatial data quality, Theory, uncertainty	Record linkage, data matching, data provenance, active learning
linked data	data linkage, data matching, deduplication, electronic health records, entity resolution, metadata, privacy, record linkage, web services	Entity resolution, data integration, Dbpedia, metadata, instance matching, entity linkage, semantic web, Sparql, Rdf
cloud computing	Algorithm, mapreduce, machine learning	Hadoop, big data, privacy, data consistency, security, machine learning
data matching	data linkage, deduplication, electronic health, records, entity resolution, metadata, privacy, record linkage, web services	Scalability, deduplication, record linkage, entity resolution, privacy, blocking, data provenance
standardization	data integration, data quality management, interoperability, linked data, ontologies, ontology, quality assessment, semantic web, semantics	Data integration, interoperability

نکته بسیار مهمی که در این بخش قابل‌مشاهده است، اشتراک پایین میان کلمات کلیدی داخل خوشه و کلمات کلیدی با هم رخدادی بالا با کلمه موردنظر است. همان‌طور که در شکل شش. نیز مشخص‌شده است، بیشترین فراوانی اشتراک مرتبط با صفر و پس‌ازآن یک است. به‌عبارت‌دیگر در هشت کلمه از ۱۰ کلمه کلیدی انتخاب‌شده تنها یک کلمه مشترک بین دو ستون جدول مشاهده‌شده است.



شکل ۶. میزان اشتراک میان لغات درون خوشه و هم رخدادی بالا

همان‌طور که پیش‌تر نیز بحث شد، این اشتراک پایین نشان‌دهنده میان‌رشته‌ای‌تر شدن موضوعات جدید کیفیت داده است. به‌بیان‌دیگر این موضوعات بین محققان زیر حوزه‌های مختلف کیفیت داده در حال مطالعه است و برای موفقیت در آن‌ها باید به بیش از یک تخصص در کیفیت داده مجهز بود.

بحث و نتیجه‌گیری

هدف از انجام این پژوهش شناسایی حوزه‌های تحقیقاتی کیفیت داده و همچنین مطالعه میزان میان‌رشته‌ای بودن این حوزه تحقیقاتی است. بر اساس مطالعه انجام‌شده حوزه کلان داده به‌عنوان جوان‌ترین حوزه در مبحث کیفیت داده مطرح‌شده است. ازجمله کلماتی که با این کلمه کلیدی هم رخدادی بالایی دارند یکپارچه‌سازی داده‌ها از منابع مختلف و همچنین بحث داده‌های گم‌شده است. موضوع دوم مرتبط با حکمرانی داده است که مبحث مدیریت داده ازجمله مباحث مرتبط با آن هست. سومین حوزه مرتبط با مباحث رسانه‌های اجتماعی و مباحث مرتبط باکیفیت داده در رسانه‌های اجتماعی است.

نکته مهمی که در مورد تمام این حوزه‌ها قابل‌مشاهده است، ویژگی بین‌رشته‌ای بودن تمامی این حوزه‌ها است که به این معنا است حوزه‌های نوین بیشتر بر بستر حوزه‌های قبلی مطرح‌شده‌اند و ارتباطی بین چندین حوزه مختلف را ایجاد کرده‌اند. به‌عنوان نمونه حوزه کلان داده با مباحث مختلفی از کیفیت داده در ارتباط است که ازجمله آن‌ها می‌توان به یکپارچه‌سازی، مدیریت فراداده، داده‌های گم‌شده اشاره کرد. این بدین معنا است که این حوزه‌ها بر روی مرز قرارگرفته‌اند و ارتباطات بیرونی زیادی دارند که باعث همبندی گراف خواهد شد.

فهرست منابع

Batini, C. and M. Scannapieca. 2006. "*Introduction to Data Quality.*" Data Quality: Concepts, Methodologies and Techniques: 1-18.

Batini, C. and M. Scannapieco. 2006. "*Data Quality Concepts, Methodologies and Techniques*" Springer-Verlag.

Blake, R. 2010. "*Identifying the core topics and themes of data and information quality research*". AMCIS.

Blake, R. and G. Shankaranarayanan. 2012. "*Discovering Data and Information Quality Research Insights Gained through Latent Semantic Analysis.*" International Journal of Business Intelligence Research (IJBIR) **3**(1): 1-16.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre. 2008. "*Fast unfolding of communities in large networks*." Journal of statistical mechanics: theory and experiment 2008(10): P10008.

Brodie, M. L. 1980. "*Data quality in information systems.*" Information & Management **3**(6): 245-258.

Chaffey, D. and G. White. 2010. "*Business information management: improving performance using information systems.*" Pearson Education.

Du, J. and L. Zhou. 2012. "*Improving financial data quality using ontologies.*" Decision Support Systems 54(1): 76-86.

Gschwandtner, T., J. Gärtner, W. Aigner and S. Miksch. 2012. "*A taxonomy of dirty time-oriented data*." Multidisciplinary Research and Practice for Information Systems, Springer**:** 58-72.

Helfert, M. and M. Ge. 2006. "*A review of information quality research.*" 11th International Conference on Information Quality.

Khalilijafarabad, A., M. Helfert and M. Ge. 2016. "*Developing a Data Quality Research Taxonomy–an Organizational Perspective*." International Conference on Information and Data Quality, Spain.

Kim, W., B.-J. Choi, E.-K. Hong, S.-K. Kim and D. Lee. 2003. "*A taxonomy of dirty data*." Data mining and knowledge discovery **7**(1): 81-99.

Kim, Y. J., R. Kishore and G. L. Sanders. 2005. "*From DQ to EQ: understanding data quality in the context of e-business systems.*" Communications of the ACM **48**(10): 75-81.

Lima, L. F. R., A. C. G. Maçada and L. M. Vargas. 2006. "*Research into Information Quality: A Study of the State of the Art in IQ and Its Consolidation*". International Conference on Information Quality.

Madnick, S. E., R. Y. Wang, Y. W. Lee and H. Zhu. 2009. "*Overview and framework for data and information quality research.*" Journal of Data and Information Quality (JDIQ) **1**(1): 2.

Morris, S. A. and B. Van der Veer Martens. 2008. "*Mapping research specialties.*" Annual review of information science and technology 42(1): 213-295.

Neely, M. P. and J. Cook. 2008. "*A Framework for Classification of the Data and Information Quality Literature and Preliminart Results (1996-2007).*" AMCIS 2008 Proceedings: 131.

Oliveira, P., F. Rodrigues, P. Henriques and H. Galhardas. 2005. "*A taxonomy of data quality problems.*" 2nd Int. Workshop on Data and Information Quality, Citeseer.

Oliveira, P., F. Rodrigues and P. R. Henriques. 2005. "*A Formal Definition of Data Quality Problems*." IQ.

Porter, A. L., D. J. Roessner and A. E. Heberger. 2008. "*How interdisciplinary is a given body of research?*" Research evaluation **17**(4): 273-282.

Qi, W. 2016. "*Studies in the Dynamics of Science: Exploring emergence, classification, and interdisciplinarity*". KTH Royal Institute of Technology.

Rafols, I. and M. Meyer. 2009. "*Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience.*" Scientometrics **82**(2): 263-287.

Rafols, I., A. L. Porter and L. Leydesdorff . 2010. "*Science overlay maps: A new tool for research policy and library management.*" Journal of the American Society for information Science and Technology 61(9): 1871-1887.

Rahm, E. and H. H. Do. 2000. "*Data cleaning: Problems and current approaches.*" IEEE Data Eng. Bull. **23**(4): 3-13.

Sadiq, S. 2013. "*Prologue: Research and Practice in Data Quality Management*." Handbook of Data Quality, Springer**:** 1-1.١

Sadiq, S. W., N. K. Yeganeh and M. Indulska. 2011. "*Cross-disciplinary collaborations in data quality research.*" European Conference on Information Systems.

Shankaranarayanan, G. and R. Blake. 2017. "*From Content to Context: The Evolution and Growth of Data Quality Research.*" Journal of Data and Information Quality (JDIQ) **8**(2): 9.

Simmhan, Y. L., B. Plale and D. Gannon. 2005. "*A survey of data provenance in e-science.*" ACM Sigmod Record **34**(3): 31-36.

Wang, R. Y. and D. M. Strong. 1996. "*Beyond accuracy: What data quality means to data consumers.*" Journal of management information systems **12**(4): 5-33.

Zhang, T., Y. Wu, H. Zhang, Y. Liu and W. Huang. 2013. "*Identifying Data Quality/Information Quality Research: Framework and Evolution. Diversity, Technology, and Innovation for Operational Competitiveness*". Proceedings of the 2013 International Conference on Technology Innovation and Industrial Management, ToKnowPress.

Analyzing the evolution of Data Quality research area using Co-word analysis

Ahmad Khalilijafarabad[1]
PhD. IT Management, Faculty of Management, University of Tehran, Tehran, Iran

Abstract: The domain of data quality is one of the growing and important areas in the field of information systems. The exact recognition of this field on the one hand and the recognition of the features of the new sub-fields of this field and its interdisciplinary of it will be of great importance to researchers. This knowledge is important for them to decide on the research process and the choice of the field of activity. For this purpose, in this study, using the graph of keyword co-occurrence is conducted on more than 9000 papers. Based on this study, it has been found that these domains have been more interdisciplinary and focus on the relationship between several fields of study. In other words, these keywords are more closely associated with the keywords in the other clusters than with the keywords they are in the same cluster. According to this study, the latest are is big data that focused on integration issues and missing data in this area.
Keywords: Data quality, Data mining, Data science, Social network analysis

1.Corresponding Author: Ahmad.khalili@ut.ac.ir