کالج پروژه

www.collegeprozheh.ir

دانلود پروژه های دانشگاهی

بانک موضوعات پایان نامه

دانلود مقالات انگلیسی با ترجمه فارسی

آموزش نگارش پایان نامه ، مقاله ، پروپوزال

www.collegeprozheh.ir کالج پروژه

تحلیل مسائل مربوط به کیفیت داده ها در سیستم های اطلاعات پژوهشی از طریق پروفایل بندی داده ها

چکیده

موفقیت یا شکست یک RIS در یک موسسه علمی عمدتاً مربوط به کیفیت داده های موجود به عنوان پایه ای برای کاربردهای RIS است. زیباترین ابزارهای هوش کسب و کار (BI) (گزارش دهی و غیره), هنگامی که داده های نادرست، ناقص و یا متناقض را نمایش دهند, بی ارزش هستند. بنابراین بخشی جدایی ناپذیر از هر RIS ، ادغام داده ها از سیستم های عملیاتی است. قبل از شروع فرآیند ادغام (ETL) یک سیستم منبع، یک تحلیل غنی از داده های منبع ضروری است. با حمایت از یک بررسی کیفیت داده ها، علل مشکلات کیفی معمولاً قابل شناسایی می شوند. تحلیل های مربوطه با استفاده از پروفایل های داده ها انجام می شوند تا تصویر خوبی از وضعیت داده ها ارائه شود. در این مقاله، روش های پروفایل سازی داده ها به منظور به دست آوردن یک مرور کلی از کیفیت داده ها در سیستم های منبع قبل قبل از ادغام آنها در RIS ارائه شده است. با کمک پروفایل سازی داده ها ها، موسسات علمی نه تنها می توانند اطلاعات تحقیقاتی خود را ارزیابی کنند و نیز اطلاعاتی در مورد کیفیت آنها ارائه دهند، بلکه همچنین وابستگی ها و زواید از میان زمینه های داده را بررسی می کنند و آنها را در RIS خود اصلاح می کنند.

کلید واژه ها: سیستم های اطلاعات پژوهشی فعلی، CRIS، سیستم های اطلاعات پژوهشی، RIS، اطلاعات پژوهشی، منابع داده، کیفیت داده، بار تبدیل استخراج، ETL، تحلیل داده ها، پروفایل سازی داده، سیستم علمی، استانداردسازی

۱. مقدمه

در سال های اخیر، الزامات گزارشگری تحقیق در تمام موسسات، به شدت افزایش یافته است. هر دوی موسسه داخلی (هیئت نمایندگی، اداره و مدیریت تخصصی) و خارجی (وزارت های مربوطه، عموم علاقه مند) نیاز به اطلاعات جامع در مورد فعالیت های تحقیقاتی موسسات برای اهداف برنامه ریزی و کنترل دارند. فعالیت های تحقیقاتی شامل

www.collegeprozheh.ir کالج پروژه

اطلاعات تحقیقاتی مانند پروژه ها، بودجه های شخص ثالث، اختراعات، شرکا، قیمت ها و نشریات و غیره می باشد. این اطلاعات بازنگری در یک سیستم اطلاعات پژوهشی مربوطه ذخیره و مدیریت می شود. یک سیستم اطلاعات پژوهشی زمانی استفاده می شود که سیستم اطلاعاتی از منابع متفاوت, اطلاعات را از طریق یک فرایند ETL یکپارچه گردآوری می کند و امکان تحلیل آن را با استفاده از توابع خروجی و تحلیل مختلف فراهم می کند. با معرفی RIS ، موسسات علمی می توانند نمای کلی کنونی از فعالیت های تحقیقاتی خود به دست آورند، اطلاعات مربوط به فعالیت های علمی، پروژه ها و نشریات خود را جمع آوری، پردازش و مدیریت نمایند و همچنین آنها در وب خود ادغام نمایند. علاوه بر این، کاربران RIS نیاز به تلاش اضافی برای بررسی فعالیت های تحقیقاتی آنها ندارند.

حجم های رو به رشد داده ها و تعداد فزاینده سیستم های منبع می تواند به خطاهای داده ممکن، تکرارها، مقادیر گم شده، قالب بندی نادرست و تناقض در RIS منجر شود. علاوه بر این، برای تحلیل کیفیت داده ها در سیستم های منبع قبل از ادغام آنها در RIS به کار گرفته می شود. هرقدر نقص های کیفی زودتر کنترل شوند و بهبود یابند، بهتر است. برای این منظور، هدف از این مقاله, ارائه روشهای ممکن برای پروفایل بندی اطلاعات کاربردی برای اطلاعات تحقیقاتی است تا تسهیلاتی فراهم شود که تحلیل و ارزیابی دقیق داده های موجود فراهم شود.

2. اطلاعات تحقیقاتی و سیستم اطلاعات تحقیقاتی

علاوه بر تدریس، تحقیق یکی از وظایف اصلی دانشگاه ها است. اطلاعات در مورد انجام وظایف و خدمات در این زمینه باید با زمان کمتری دردسترس قرار گیرند و قابل اطمینان تر باشد. برای زمینه تحقیق، اطلاعات و اطلاعات به طور عمده برای نقشه برداری فعالیت های پژوهشی و نتایج آنها، و مدیریت فرآیندهای مرتبط با فعالیت تحقیقاتی جمع آوری می شوند. این ممکن است شامل اطلاعات در مورد پروژه های تحقیقاتی، مدت زمان آنها، پژوهشگران شرکت کننده و انتشارات مربوطه باشد. این اطلاعات همچنین اطلاعات تحقیق یا داده های تحقیقاتی نامیده می شود. رده های اطلاعات تحقیقاتی (RI) با ارقام و شاخص های کلیدی مختلف را می توان از جدول ۱ مشاهده کرد:

با این وجود، نه تنها خود داده ها، بلکه ساختار داده ها و ارتباط اطلاعات یک پژوهش، به منظور توانایی جمع آوری و ارزیابی اطلاعات در سطوح مختلف سازمان برای اهداف متفاوت در مرکز کار قرار می گیرند (سطوح می توانند, یک پروفسور, یک کرسی, یک موسسه, یک دپارتمان یا یک دانشکده باشند).

جدول ۱: اطلاعات تحقیق (RI) با ارقام و شاخص های کلیدی مختلف (آیروزال و همکاران, ۲۰۱۸)

ارقام و شاخص های کلیدی	RI
نام, عنوان, جنس, ملیت	فرد
اختیار قانونی, بازنگری شده, مقالات	نشریات
تعداد, مدت زمان, تامین مالی, همکاری	پروژه ها
	منابع مالی شخص
TPF تدارکات, TPF خرج شده, TPF رقابتی	ثالث
جوایز	هدایا
خانواده ثبت اختراع, اسپین آف, شماره ثبت اختراع	ثبت های اختراع
واحد دانشگاهی, رابطه استخدام, رده کارکنان, دپارتمان	اطلاعات سازمانی

فرایندها و سیستم هایی که اطلاعات تحقیقاتی را ذخیره و مدیریت می کنند، به سیستم های اطلاعات پژوهشی (RIS یا CRIS برای سیستم اطلاعات جاری تحقیقاتی) مرتبط هستند.

یک RIS, یک پایگاه داده مرکزی است که می تواند برای جمع آوری، مدیریت و ارائه اطلاعات در مورد فعالیت های پژوهشی و نتایج تحقیقات استفاده شود.

شکل زیر (نگاه کنید به شکل ۱) یک نمای کلی از ورود اطلاعات پژوهشی از یک دانشگاه به سیستم اطلاعات پژوهشی و معماری RIS را نشان می دهد.

www.collegeprozheh.ir کالج پروژه

بلوک های ساختاری معماری RIS را می توان به عنوان یک فرآیند سه مرحله ای مشاهده کرد:

۱. لایه دسترسی به داده ها

۲. لایه سیستم برنامه

۳. لایه ارائه

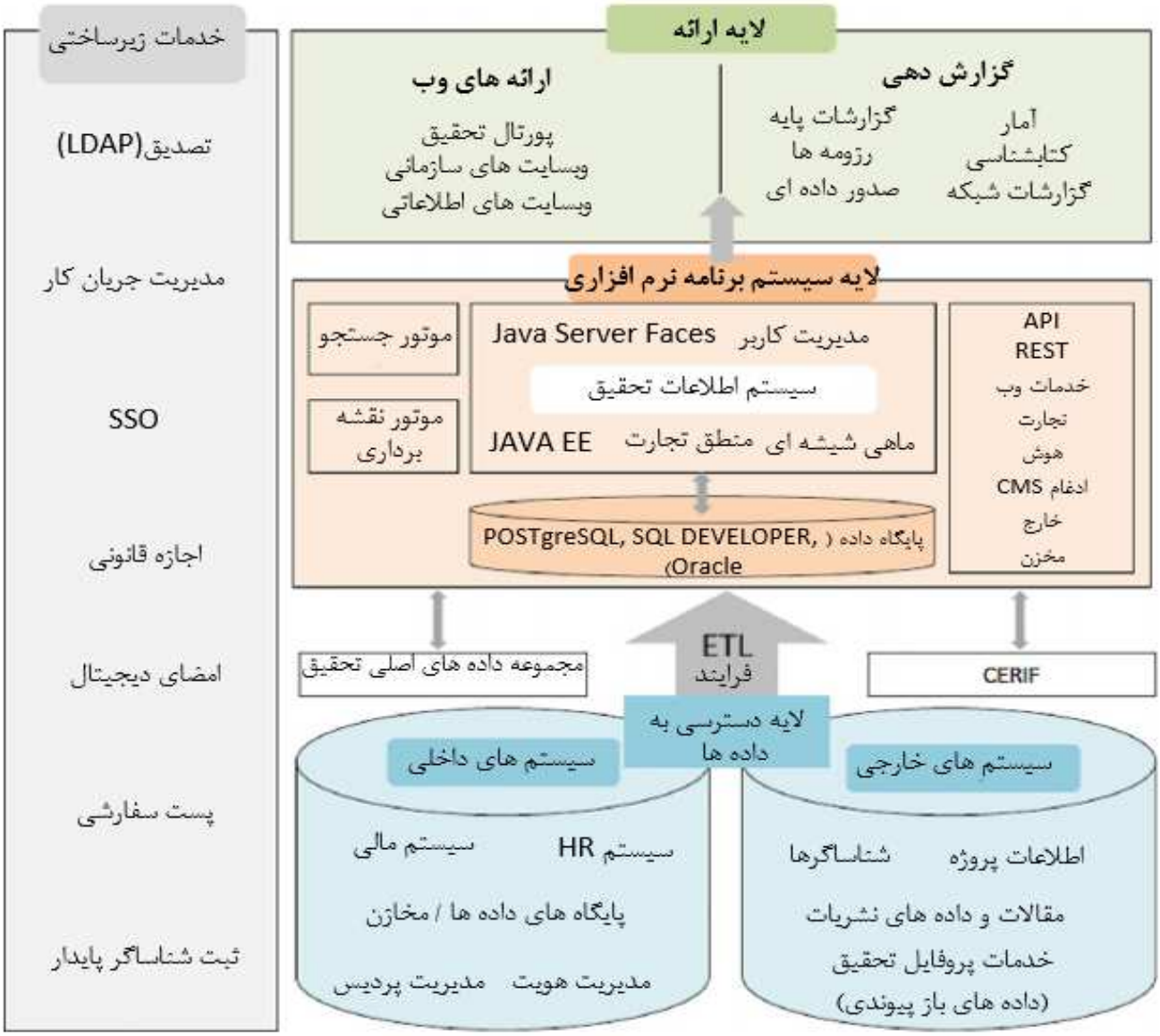
لایه دسترسی به داده ها شامل منابع داده های داخلی و خارجی (سیستم های عملیاتی) است. این سطح شامل، به عنوان مثال، پایگاه داده ها از مخزن کتابخانه های دولت یا انتشارات، شناسه هایی مانند ORCID یا داده های کتابشناختی از وب Science یا Scopus و غیره می باشد. یک فایل بندی از این منابع داده ها از طریق پروتکل ETL کلاسیک در RIS انجام می شود. لایه سیستم برنامه شامل سیستم اطلاعات پژوهشی و برنامه های کاربردی آن می باشد که ادغام، مدیریت و تحلیل داده های حفظ شده در سطح پایه را انجام می دهند. لایه ارائه, تهیه خاص-گروه هدف و ارائه نتایج تحلیل برای کاربر را نشان می دهد که در قالب گزارش ها با استفاده از ابزارهای هوش کسب و کار در دسترس قرار می گیرند. علاوه بر گزینه های گوناگون گزارشگری، پورتال ها و وب سایت های تسهیلات نیز می توانند در اینجا پر شوند.

عمود بر لایه های توصیف شده، خدمات زیربنایی، خدمات کلی برای همه سیستم های اطلاعاتی مانند احراز هویت (LDAP)، مجوز، ثبت نام تک و غیره وجود دارند.

پیشنهادها برای جمع آوری، ذخیره سازی و مبادله اطلاعات مجدد استانداردسازی شده تحقیق در RIS، مدل داده های مجموعه داده های هسته ای تحقیق (RCD) و مدل داده های فرمت اطلاعات تحقیقاتی مشترک اروپایی (CERIF) هستند. این دو مدل، مقوله و رابطه آنها را با یکدیگر شرح می دهند.

RIS, یک دیدگاه جامع از فعالیت های پژوهشی و نتایج در موسسه تحقیقاتی را ارائه می دهد. آنها نه تنها فعالیت های پژوهشی مؤسسات بلکه همچنین محققان فعلی، مرکزی و روشن را نشان می دهند. با مرکزیت در این سیستم، محققان امکان دست یابی به کار را فراهم می کنند. داده ها یک بار با RIS وارد می شوند و می توانند چندین بار استفاده شوند، برای مثال در وب سایت ها، برای برنامه های کاربردی پروژه و یا گزارشات فرایندها. یک مدیریت داده

دوگانه و کار اضافی برای کاربران باید اجتناب شود. هدف دیگر این است که سیستم های اطلاعات پژوهشی را به عنوان یک ابزار مرکزی برای ارتباط مداوم و پیوسته و مستند سازی فعالیت های پژوهشی و نتایج مختلف تحقیقاتی ایجاد کنیم. بهبود بازیابی اطلاعات کمک می کند تا محققان, همکاران, شرکت ها در تخصیص قراردادهای تحقیقاتی را برای ارائه شفافیت و اطلاعات عمومی در مورد سازمان های خود را در اختیار عموم قرار دهند.



شکل ۱: سیستم اطلاعات تحقیقاتی (آذر و ابوسبا، ۲۰۱۷).

۳. پروفایل بندی داده ها

www.collegeprozheh.ir کالج پروژه

بر اساس جمع آوری و ادغام منابع داده های داخلی و خارجی (مانند پایگاه های داده متفاوت، فایل های متنی و یا فایل های XML و غیره) از امکانات در سیستم های اطلاعات پژوهشی، انواع مختلفی از خطاهای داده وجود دارند که باید بیشتر توسط RIS پردازش شوند. برخی از این خطاها عبارتند از:

- گم شدن مقادیر (کامل بودن مشخصه)
- اطلاعات غلط ناشی از ورودی، اندازه گیری یا پردازش خطاها (صحت مشخصه)
- تکرار در مجموعه داده ها (مشخصه بدون زواید)
- داده های ناسازگار ارائه شده (سازگاری مشخصه)
- مقادیر منطقی متناقض (سازگاری مشخصه)

کیفیت داده های منبع, تأثیر مستقیم بر کیفیت RIS دارد. برای جلوگیری از بروز مشکلات کیفیت داده های ساختاری و محتوا، تحلیل و پاکسازی داده ها در یکپارچگی (ادغام) داده ها صورت می گیرد. ادغام داده ها, انتقال داده های عملیاتی از سیستم های مختلف میراثی به RIS است. پر کردن RIS با استفاده از فرایند ETL انجام می شود. این اصطلاح مخفف استخراج، تبدیل و بار است و به عنوان فرایند تدارک داده ها درک می شود. هدف آن, تمیز کردن داده ها از ساختارهای متفاوت و استانداردسازی آنها به منظور ذخیره دائمی در RISاست. برای بارگیری داده ها در یک RIS، آنها ابتدا باید استخراج شوند. از آنجا که اطلاعات معمولاً از چندین سیستم منبع بارگذاری می شوند، داده های سیستم های مربوطه باید با یکدیگر هماهنگ شوند. سپس داده ها در یک RIS بارگذاری می شوند. این فرایند توسط شکل زیر نشان داده می شود (شکل ۲):

برای بازنگری موفق کیفیت داده ها و تصمیم گیری در این مورد که آیا مشکلات داده در سیستم های منبع را می توان با اصلاح در فرآیند ETL بهبود داد و یا حل کرد. اما پروفایل سازی داده ها را می توان برای درک بهتر ساختار منابع داده و شناسایی و تصحیح خودکار اشتباهات احتمالی مورد استفاده قرار داد. این خطاهای شناسایی شده در داده ها ثابت نیستند، بلکه فقط فراداده های متعلق را اصلاح می کند و سپس مشکلات کیفیت داده را حل می کند.

www.collegeprozheh.ir کالج پروژه

پروفایل بندی داده ها, یک اصطلاح جدید است و به عنوان مترادف برای تحلیل داده ها استفاده می شود. پروفایل بندی داده ها, یک فرایند خودکار برای تحلیل داده های موجود است (اولسن، ۲۰۰۳). روش های مختلف برای تحلیل سیستماتیک, اطلاعاتی در مورد ساختار، محتوا و کیفیت جمع آوری داده ها به دست می دهند تا تصویری دقیق از وضعیت فعلی بدست آید, مانند (Olsen، 2003):

- تعریف مقادیر داده های مجاز.
- ناهنجاری ها و موارد پرت درون ستون ها.
- ارتباطات بین ستون ها (کلید اولیه، وابستگی ها، و غیره).
- ستون هایی که در آن فرمت های تاریخ ممکن و آدرس های پست الکترونیکی و غیره را می توان یافت.

برای تحلیل اطلاعات تحقیق در RIS، پروفایل سازی داده ها, سه نوع تحلیل را پیشنهاد می دهد (نگاه کنید به شکل ۳). پشت هر کدام از این سه نوع, روش های مختلف برای پروفایل بندی داده ها وجود دارند. این کر عمدتا به شیوه تحلیل داده ها بستگی دارد: در یک ستون ("تحلیل ویژگی")، در وابستگی های ستون ("وابستگی عملکردی") یا در وابستگی ویژگی ها / ستون ها در جداول متفاوت ("تحلیل مرجع").

تحلیل ویژگی: اطلاعات عمومی و دقیق در مورد ساختار، محتویات یک جدول و تمام روابط بین جداول مختلف، ستون ها و مقادیری که در جدول ظاهر می شود، را می گیرد. در اینجا سوالات کلیدی برای پاسخ دادن به اطلاعات مربوط به هر یک از جنبه های زیر (Apel، Behme، Eberlein، Merighi &، 2015) وجود دارند:

۱. تحلیل ساختار داده: (آیا داده ها به فراداده های مرتبط پاسخ می دهند؟)

۲. تحلیل محتویات داده ها: (آیا مقادیر داده ها, کامل، صحیح و به روز، داده های استاندارد شده بر اساس قوانین قابل اجرا است؟)

۳. تحلیل وابستگی: (آیا داده ها در تمام ستون ها و جداول دارای نقشه برداری مورد نیاز برای رابطه کلیدی مشخص شده هستند؟ آیا روابط استنتاج شده در ستون ها، جداول و پایگاه های داده وجود دارند؟ آیا داده های زاید وجود دارند؟)

www.collegeprozheh.ir کالج پروژه

در پشت این پرسش های اصلی، اطلاعات زیادی در زمینه های زیر یافت می شود (Apel et al.، 2015):

۳,۱ تحلیل نام صفت (ویژگی)

تحلیل نام صفت به نام صفت اشاره دارد. در اینجا، اسامی صفت باید با نوع داده ها و محتوای داده ها مطابقت داشته باشد (مثلا یک "شناسه نویسنده" می تواند یک مقدار عددی داشته باشد).

۳,۲ تحلیل نوع داده ها

این تحلیل, نوع داده های ویژگی را ارزیابی می کند. آیا در varchar, یک میدان نوع داده وجود دارد، برای مثال فقط اعداد وجود دارند، تغییر نوع داده برای پردازش بیشتر ممکن است مفید باشد. این باعث تعریف یک قاعده می شود که تضمین می کند که تمام مقادیر, یک نوع داده مشابه داشته باشند. هدف از تحلیل نوع داده ها، یافتن معیارهایی مانند حداقل، حداکثر یا طول است، به این ترتیب، به عنوان مثال یافتن اختلافات در ذخیره سازی میسر می شود.

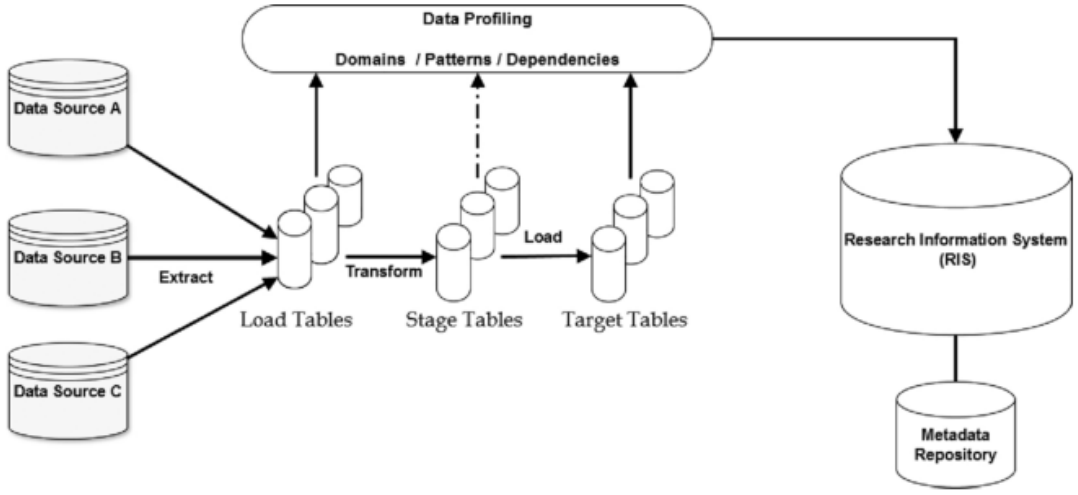
پروفایل بندی داده ها

حوزه ها / الگوها / وابستگی ها

منبع داده A

سیستم اطلاعات تحقیق (RIS) بار تبدیل استخراج منبع دادهد B

مخزن فراداده ها جداول هدف جداول مرحله جداول بار منبع داده C



شکل ۲. ادغام داده در RIS

www.collegeprozheh.ir کالج پروژه

پروفایل بندی داده ها

وابستگی عملکردی

تحلیل ارجاعی

صفت افزونه

پیوست ها

یتیم ها

کودکی

تحلیل صفت

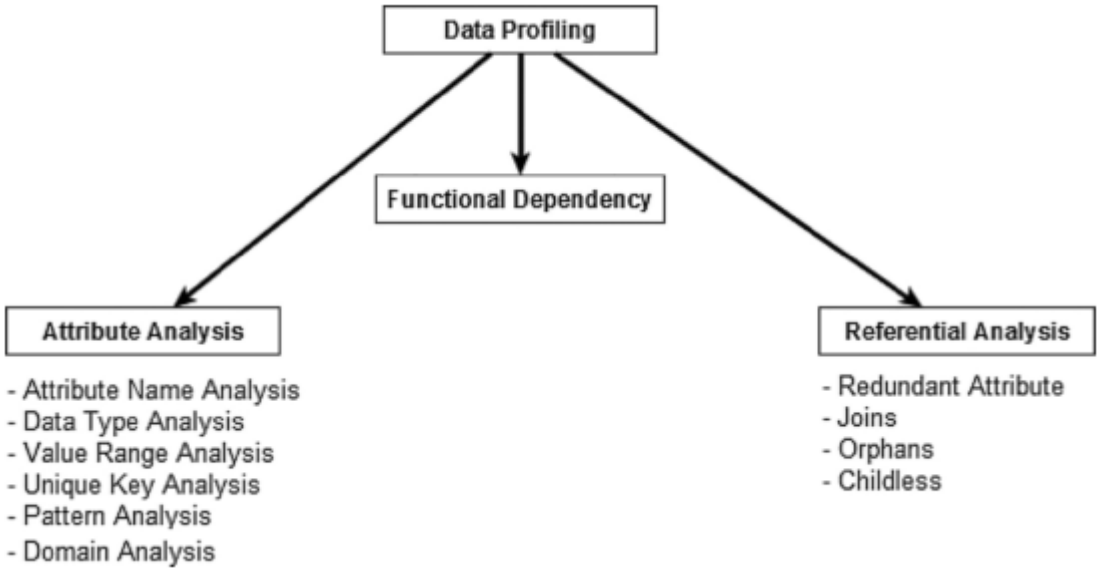
تحلیل نام صفت

تحلیل نوع داده ها

تحلیل گستره مقادیر

تحلیل کلیدی منحصر به فرد

تحلیل حوزه



شکل ۳. انواع تحلیل پروفایل بندی داده ها

DataSet: SELECT List of publication.Author ID, List of publication.Name, List of publication.ORCID, List of publication.Birth Date, List of publication.Address FROM list of publication.xls...

Author ID	Name	ORCID	Birth Date	Address
353035	Alien Scott	0000-0007-0212-2108	10/25/1965	145 F. Concord Street, Orlando, ...
353035	Dr. Alien Scott	0000-0000-0000-0000	1965-10-25 00:00:00	Concord Street, 32801 145F
353035	Alien William Scott	0000-0007-0212-2108	652510	25 Concord 32801 Street
353036	A. Scott	<null>	11/25/56	12 Ford Ave 32801
353036	Scott Alien	0000-0007-0212-2108	1965-11-25 00:00:00	<null>
<null>	Alien Scoth	702122108	1956-10-25 00:00:00	Street C., 32801 145F. US
410003	Olivia Svenson	0450-1254-3598-F156	1983-02-06 00:00:00	745-7801 P.B. Las Vegas 29502
410003	Svenson Olivia	045012543598F156	1983	7801 P.B. Las Vegas 29502

Previous page Next page

شکل ۴. نمونه ای از یک فهرست انتشار

www.collegeprozheh.ir کالج پروژه

۳,۳. تحلیل محدوده مقادیر

در این تحلیل، ارقام کلیدی آماری مختلف برای تحلیل داده ها (حداقل، حداکثر، متوسط، توزیع فرکانس، انحراف استاندارد و غیره) مورد استفاده قرار می گیرند.

۳,۴. تحلیل کلید منحصر به فرد

این تحلیل در مورد پیدا کردن صفرها یا تکرارها است. این دو برای هر ارزیابی و فرآیند خطرناک هستند، بنابراین لازم است که یک قانون ایجاد شود که تضمین کند که تمام مقادیر وارد شده، نه تکراری هستند و نه حاوی مقادیر صفر.

۳,۵ تحلیل الگو

در این تحلیل، الگوها یا ارائه های کلی توسط تحلیل صفات شناسایی می شوند. ابتدا مقادیر برای الگوهای احتمالی مورد جستجو قرار می گیرند و سپس این مقادیر با فیلتر کردن الگوها شناسایی می شوند و در رابطه قرار می گیرند. درصدهای محاسبه شده در این روش, اطلاعات مربوط به اعتبار نمونه ها را ارائه می دهند. سپس، قوانین را می توان برای حل هر مسئله شناسایی ایجاد کرد. الگوهای شناخته شده ممکن به عنوان مثال فرمت های تاریخ، ایجاد آدرس های ایمیل و شماره تلفن، و غیره می باشند.

۳,۶. تحلیل دامنه

تحلیل دامنه, اطلاعاتی در مورد محدوده های مقادیر / مقادیر ممکن که اغلب رخ می دهند، ارائه می دهد. به عنوان مثال، در اینجا ستون "وضعیت زناشویی" و "جنسیت" آمده است. پس از بررسی این ستون، مشخص می شود که مقادیر رخ داده در میان موارد زیر یافت می شوند: "مجرد"، "ازدواج" یا "طلاق گرفته" و "M" یا "W". این دامنه می تواند برای احراز قوانین و محدود کردن مقادیر مجاز استفاده شود. علاوه بر این، چنین قاعده تجمیع را تسهیل می کند و دقت اطلاعات را افزایش می دهد.

اشکال ۴ تا ۸ زیر, یک نمونه عملی از یک فهرست انتشار را نشان می دهند و ساختار داده ها و محتوای داده های آن را با استفاده از ابزار پاک کننده داده تحلیل می کنند.

www.collegeprozheh.ir کالج پروژه

۳.۷ وابستگی کارکردی

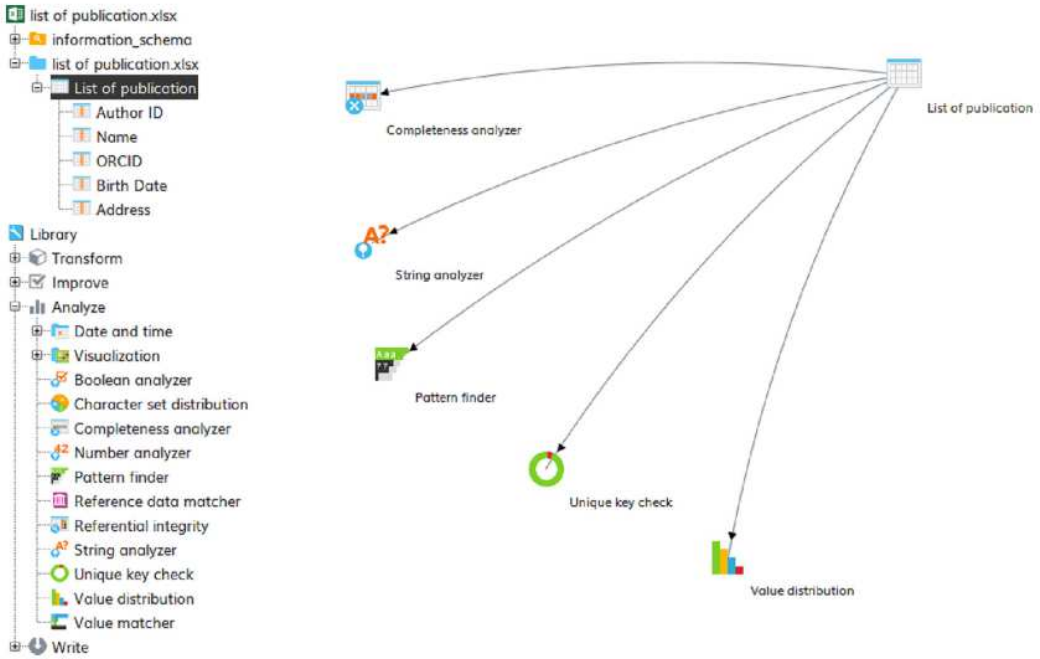
این تحلیل, وابستگی ها بین ستون های فردی را تعیین می کند. در اینجا مشخص می شود کدام صفات را می توان با مقادیر دیگر محاسبه کرد یا استخراج کرد. برای این منظور، قواعد اگر-آنگاه با نمره اطمینان بالا چک می شوند.

جدول ۲ نمونه ای از محتویات لیست انتشار را نشان می دهد که در آن ویژگی "سال انتشار" به ویژگی "عنوان" بستگی دارد.

۳.۸ تحلیل ارجاعی

در این تحلیل، اتصالات بین اشیاء چندگانه (در روابط متفاوت) ایجاد می شوند. و برای وابستگی های خاص تر فیلتر می شود. در اینجا عبارات مورد استفاده برای شناسایی اشیاء مورد بررسی قرار می گیرند. اصطلاحات دارای ویژگی های اضافی، پیوستن، یتیمان و بدون فرزند می باشند. با استفاده از بیان های این تحلیل، قوانین مرجع می تواند مشخص شود و یا محاسبه شود.

جدول ۳ نمونه ای از یک تحلیل مرجع را نشان می دهد. "لیست انتشارات", شیء کودک است که از "عنوان و سال انتشار" شی والدین به ارث می رسد.



شکل ۵. نمونه ای یک فهرست انتشار در ابزار پاک کننده داده ها

www.collegeprozheh.ir کالج پروژه

String analyzer

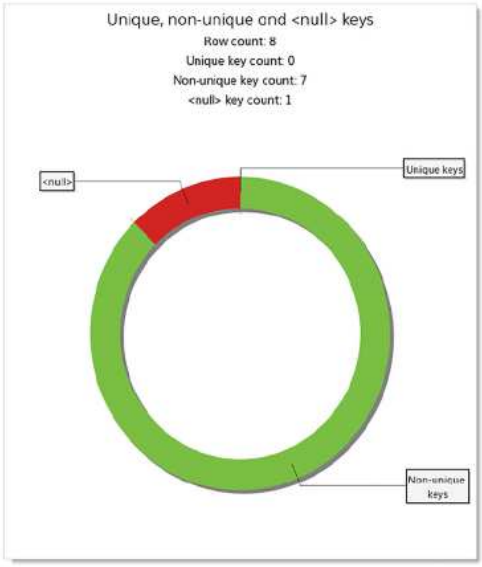
(5 columns)

	Author ID	Name	ORCID	Birth Date	Address
Row count	8	8	8	8	8
Null count	1	0	1	0	1
Blank count	0	0	0	0	0
Entirely uppercase count	0	0	2	0	0
Entirely lowercase count	0	0	0	0	0
Total char count	42	103	120	104	182
Max chars	6	19	19	19	37
Min chars	6	8	9	4	17
Avg chars	6	12,875	17,143	13	26
Max white spaces	0	2	0	1	5
Min white spaces	0	1	0	0	3
Avg white spaces	0	1,25	0	0,5	3,714
Uppercase chars	0	18	2	0	24
Uppercase chars (excl. first letters)	0	8	0	0	10
Lowercase chars	0	73	0	0	61
Digit chars	42	0	103	80	59
Diacritic chars	0	0	0	0	0
Non-letter chars	42	12	118	104	97
Word count	7	18	7	12	33
Max words	1	3	1	2	6
Min words	1	2	1	1	4

شکل ۶. تحلیلگر رشته

Unique key check

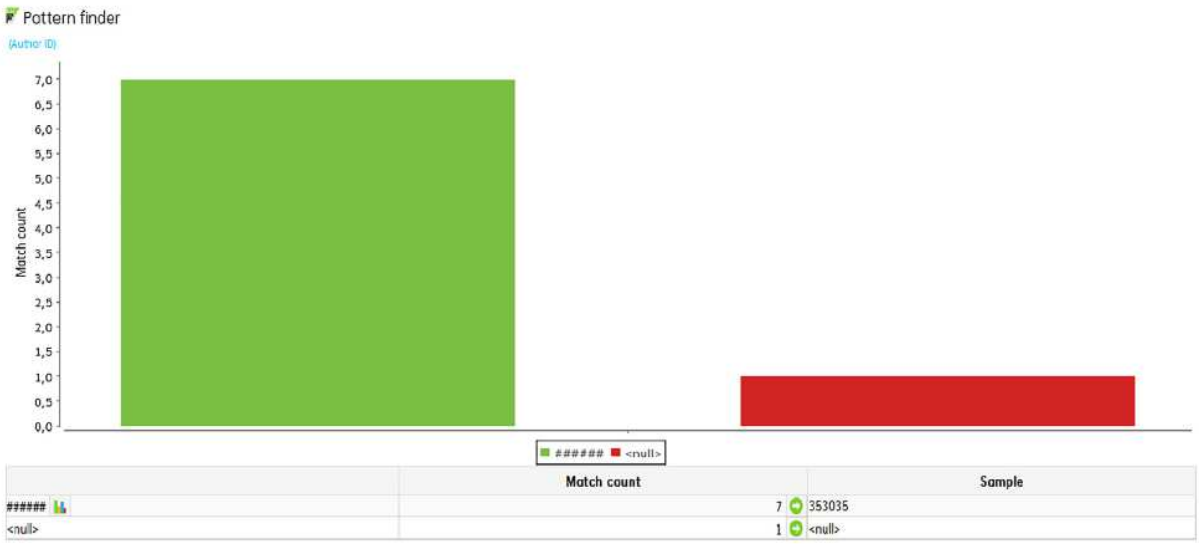
(Author ID)



Key	Count
353036	3
353036	2
410003	2

شکل ۷. چک کردن کلید منحصر به فرد

www.collegeprozheh.ir کالج پروژه



شکل ۸. یابنده الگو

تحلیل مرجع از این دو اشیا نشان می دهد که عنوان "پایگاه های داده" از جدول "لیست انتشار" (یتیم) و عنوان "مدیریت پروژه"، "تحلیل داده ها" و "ابر رایانه" از جدول "عنوان و سال از انتشارات "(بدون فرزند) هستند. همچنین یک لینک در ستون عنوان نمایش داده می شود.

بر اساس این نتایج، شما می توانید قوانین مرجع را تعیین کنید که توانایی بین دو جدول ("لیست انتشار" و "عنوان و سال انتشار") را تعیین می کند.

۴. بحث و بررسی

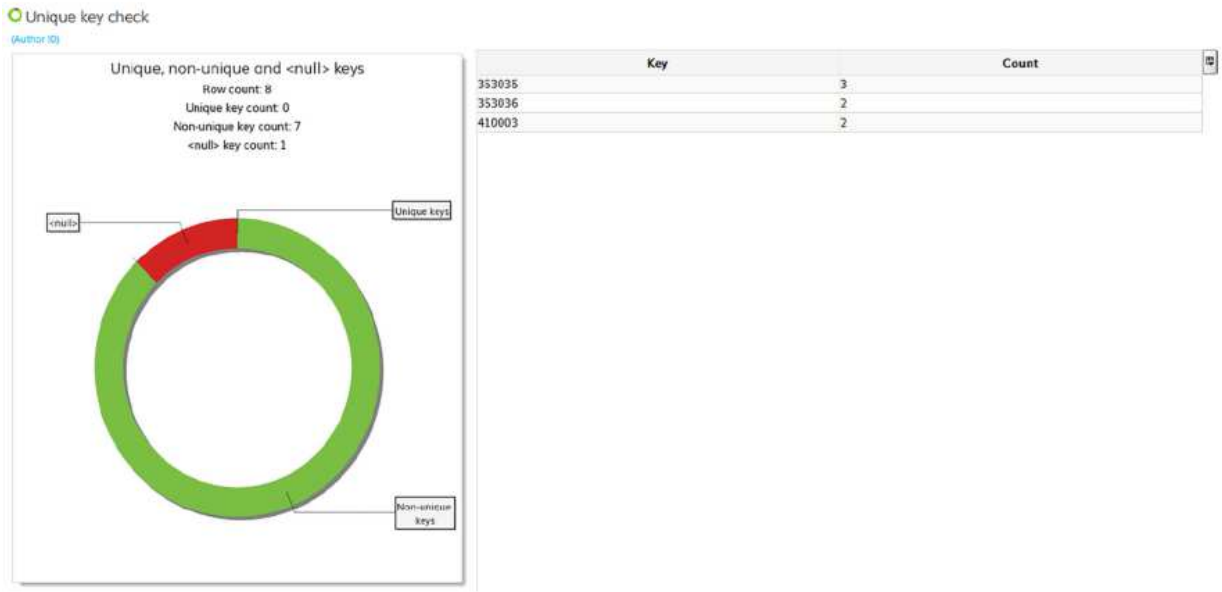
پروفایل سازی داده ها, داده های مشکل ساز را شناسایی می کند و فراداده ها را خودکارسازی می کند و در عین حال اصلاح خطاهای داده های معمول در داده ها را میسر می سازد. مؤسسات علمی می توانند داده های منبع خود را برای شناخت ساختارها، روابط و قوانین اطلاعات پروفایل بندی کنند. علاوه بر این، تحلیل ویژگی، تحلیل مرجع، تحلیل وابستگی عملکرد و یا داده های نمایه می تواند با استفاده از قوانین سفارشی انجام شوند.

www.collegeprozheh.ir کالج پروژه

به منظور نظارت بر کیفیت داده ها در RIS، جریان فرآیند متشکل از زیر می تواند به عنوان پایه ای برای فاکتور های خدمت استفاده شود و به عنوان یک مدل یا کمک به نشان دادن چگونگی تحلیل آنها در امکانات داده ها خطاهای RIS رفع و بهبود می یابد.

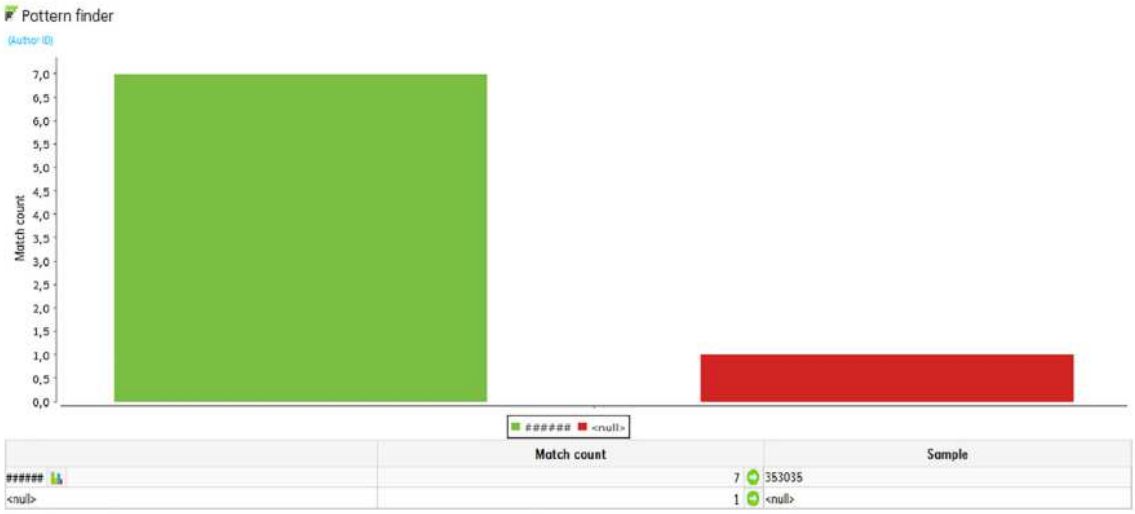
شکل زیر جریان فرا-فرایند ذکرشده برای تحلیل و بهبود کیفیت داده ها در RIS را معرفی می کند (شکل ۹).

در ابتدای جریان فرایند، منابع داده های خارجی و داخلی یک دستگاه توسط مدیریت یا پرسنل فنی جمع آوری می شوند و این منبع داده ها, پروفایل بندی می شود و تمام اطلاعات حاوی داده ها شناخته شده است. سپس برخی از قوانین داده ها را استخراج می کنند و سپس از این قوانین استنتاج شده شده برای اصلاحات استفاده می کنند. این اصلاحات با استفاده از روش پاک کردن برای پاک کردن داده ها در هدف انجام می شود. در نهایت داده های بارگذاری شده در RIS ارائه خواهند شد. با استفاده از پورتال ها، پورت مجدد و سایر برنامه های کاربردی مستقیم، اطلاعاتی که از سیستم می بینند، تجسم می شود. در اینجا، اطلاعات پردازش شده و تحلیل ها در اختیار کاربر به صورت واضح از طریق اجزای مختلف برنامه کاربردی قرار می گیرند.



شکل ۷. چک کردن کلید منحصر به فرد

www.collegeprozheh.ir کالج پروژه



شکل ۸. یابنده الگو.

جدول ۲ :نمونه برای وابستگی عملکردی یک فهرست انتشار

شناسه نویسنده	نام	ORCID	تاریخ تولد	عنوان	سال P
353035	Alien Scott	0000–0007-	1965	ادغام داده ها	2011
		1222-2301			
400015	Virginia	0000–0123-	1985	داده های بزرگ	2015
	Mic	1201-0111			
410003	Olivia	0450-1254-3598-	1983	پایگاه های داده	2017
	Svenson	F156			

جدول ۳ :نمونه ای برای تحلیل مرجع

شناسه نویسنده	نام	ORCID	تاریخ تولد	عنوان	سال P
					r
353035	Alien Scott	0000–0007-	ادغام داده ها	2011 (Child)	

www.collegeprozheh.ir کالج پروژه

		1965			
		1222-2301		Integration	
400015	Virginia	0000–0123-1985	داده های بزرگ	2015	
	Mic	1201-0111			
410003	Olivia	0450-1254-	پایگاه های داده	Databases	2017
	Svenson	3598-F156			
عنوان			سال P		
ادغام داده ها		2011		(Parents)	
مدیریت پروژه		2014			
داده های بزرگ			2015		
تحلیل داده ها		2015			
رایانش ابری		2016			

منابع داده خارجی و داخلی

وارد کردن داده ها استارت

داده های پروفایل

تصحیح داده ها خیر داده ها اوکی؟ قواعد داده ها

بله

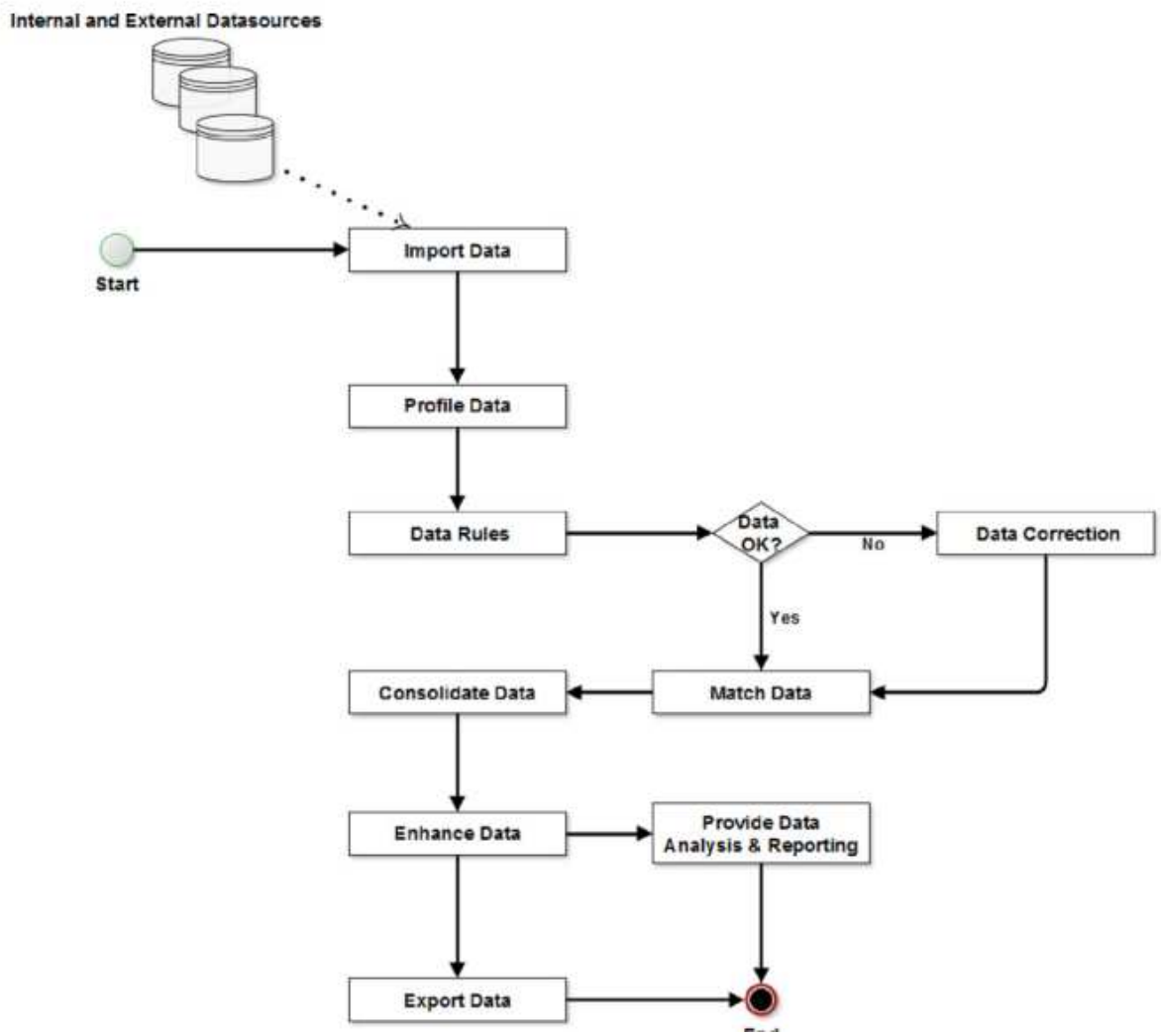
تطبیق داده ها تلفیق داده ها

فراهم کردن داده

تحلیل و گزارش دهی ارتقای داده ها

پایان صدور داده ها

www.collegeprozheh.ir کالج پروژه



شکل ۹. دیاگرام جریان فرایند در تحلیل و بهبود کیفیت داده ها در RIS

۵. نتیجه گیری و چشم انداز

پس از انتقال داده ها به RIS، می توان تحلیل های مختلفی انجام داد. این کار شامل ارزیابی کامل بودن، شناسایی الگوهای، تکراری سخت، صفر کردن، نشان دادن تفاوتها و اعتبارسنجی ویژگیها می باشد. برای شناسایی مجموعه داده های زاید که به طور کامل معادل نیستند، روش های پروفایل سازی داده توسعه یافته است. این روش ها در این مقاله برای شناسایی، تحلیل و اصلاح خطاهای داده شده در سیستم های اطلاعات تحقیقاتی ارائه شده است. پروفایل داده ها به عنوان یک مولفه مهم در بهبود کیفیت داده ها قبل از اینکه داده ها بتوانند در RIS ادغام شوند. مؤسسات

www.collegeprozheh.ir کالج پروژه

باید از محتویات اطلاعات قبل از اینکه بتوانند اطلاعات خود را برای اولین بار بدست آورند و قوانین DQ را از این آشنایی استخراج کرده و سپس در جریان اندازه گیری کیفیت داده ارزیابی شوند، استفاده می شود. ابزارها با استفاده از پروفایل های داده کمک می کنند. ابزارهای تبلیغاتی محتوای واقعی، ساختار و کیفیت داده ها را ارزیابی می کنند. برای انجام این کار، روابط بین داده های موجود در پرونده ها، و نیز روابط بین داده ها بین پرونده ها بررسی می شود. با استفاده از ابزارهای مناسب، ناهنجاری ها و خطاهای داده می توانند حتی از بزرگترین میزان داده ها شناسایی شوند، که می تواند با قوانین کیفیت اطلاعات اضافی همراه با مدیر عامل مسئول یا مدیر داده، اصلاح شود. ابزار Data Profiling عمدتا تجاری و در دسترس برای هر دو محیط برنامه های کوچک و مجموعه های کاربردی جامع برای کیفیت داده ها و یکپارچه سازی داده ها است. در سال های اخیر، یک بازار برای طبقه بندی اطلاعات نیز به عنوان یک سرویس در حال توسعه است. استفاده از ابزارهای پروپزال داده برای ارزیابی ارزشمند است، زیرا آنها به طور قابل ملاحظهای نیازهای منابع را کاهش میدهند. به خصوص با استفاده از تکرار، مقدار E ff بسیار کمتر از استفاده از ابزار است. علاوه بر این، نتایج پروفایل های داده شده به دست آمده می تواند به سرعت و به آسانی مورد استفاده قرار گیرد.

References

Apel, D., Behme, W., Eberlein, R., & Merighi, C. (2015). *Successfully control data quality. practical solutions for business intelligence projects* (3rd Ed.). Dpunkt Verlag Revised and Expanded Edition.

Azeroual, O., & Abuosha, M. (2017). Improving the data quality in the research information systems. *International Journal of Computer Science and Information Security, 15*(11), 82-86.

Azeroual, O., Saake, G., & Abuosha, M. (2018). Data quality measures and data cleansing for research information systems. *Journal of Digital Information Management, 16*(1), 12-21.

Olsen, J. (2003). *Data quality – the accuracy dimension.* San Francisco: Morgan Kaufmann Publishers.

Otmane Azeroual is a researcher at the German Institute for Higher Education Research and Science Studies (DZHW) in Berlin. After studying Business Information Systems at the University of Applied Sciences (HTW) Berlin, he began his Ph.D. in Computer Science at the Institute for Technical and Business Information Systems (ITI), Database Research Group of the Otto-von-Guericke-University Magdeburg and at the Department of Computer Science and Engineering of the University of Applied Sciences (HTW) Berlin.