به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

ارائه راهکارهای اجرایی برای بهبود کیفیت اطلاعات پایان‌نامه/رساله‌ها در سامانه ثبت برپایه تکنیک‌های داده‌کاوی

تاریخ شروع طرح پژوهشی	۱۳۹۹/۱۲/۱۲	**تاریخ پایان طرح پژوهشی**	۱۴۰۱/۰۳/۰۱

نام و نام خانوادگی مجری طرح پژوهشی	محمدجواد ارشادی
نام و نام خانوادگی همکاران طرح پژوهشی	مهندس محمدمهدی ارشادی، دکتر آزاده فخرزاده، مهندس سمانه قاضی‌زاده

دربارهٔ طرح پژوهشی

کیفیت مناسب اطلاعات مدارک پایان‌نامه/ رساله (پارسا)ها از جنبه‌های گوناگون موردتوجه است. از یک‌سو، بازیابی دقیق و صحیح را در سامانه گنج به همراه دارد. از سوی دیگر، هزینه‌های کنترل، نمایه‌سازی و کنترل کیفیت نمایه‌سازی را می‌تواند تا سطح مطلوبی کاهش دهد. برای رسیدن به سطح مطلوب از کیفیت در اطلاعات پارساها تحلیل نتایج کیفیت اطلاعات یک کار کلیدی است. به‌بیان‌دیگر، لازم است نتایج کیفیت اطلاعات به شکل درستی تحلیل شود تا بتوان ارزیابی درستی از سطح کیفیت و نوع ناهمخوانی‌ها ارائه‌کرد. این کار مقدمه تعیین راهکارهای کاربردی برای بهبود کیفیت است. بنابراین در این پژوهش، نتایج کیفیت اطلاعات در سامانه ثبت پارساهای درون کشور مورد تحلیل قرار گرفت و راهکارهای بهبود برپایه این تحلیل توسعه داده شد. ارزیابی کیفیت اطلاعات در سامانه ثبت در قالب ۳۳ قلم اطلاعاتی صورت می‌پذیرد که شامل مواردی همچون نام دانشجو، اساتید راهنما و مشاور، چکیده و کلیدواژه هستند. برای تعیین راهکارهای بهبود کیفیت اطلاعات، این پژوهش در دو گام اصلی انجام گرفت. در گام اول نتایج کیفیت اطلاعات به کمک ابزار داده‌کاوی در قالب روش خوشه‌بندی مورد تحلیل قرار گرفت. در گام دوم گروه کانونی شکل گرفت و برپایه نتایج گام اول راهکارهای بهبود کیفیت اطلاعات در سامانه ثبت برای ناهمخوانی‌های اصلی و بحرانی تعیین شد. در نهایت، استفاده از تکنیک‌های فرایندمحور مانند کاربرد چک‌لیستهای کنترل کیفیت و پردازش تصویر، در کنار تکنیک‌های داده‌محور مانند پاک‌سازی داده‌ها برای بهبود کیفیت پایان نامه/ رساله بومی‌سازی و توسعه داده شدند.

روش پژوهش

این پژوهش در دو گام صورت پذیرفت. در گام اول به تحلیل آماری نتایج کیفیت داده‌ها و فراداده‌ها (اطلاعات کتابشناختی) پارساها

پرداخت. در این گام خطاها و ناهمخوانی‌های مشاهده‌شده در مدارک پارسا در سامانه ثبت مورد ارزیابی و تحلیل قرار گرفت. این خطاها ناهمخوانی‌هایی را تشکیل می‌دهند که توسط همکاران مدیریت ثبت و فراهم‌آوری اطلاعات به ازای هر مدرک پارسا و متناسب با خطای مشاهده‌شده به هر مدرک تخصیص داده می‌شوند. به‌بیان‌دیگر مجموعه ناهمخوانی‌ها و خطاهای الصاق شده به هر مدرک ویژگی‌های گوناگونی هستند که در حال حاضر مدارک پارسا برپایه این خطاها بررسی شده‌اند. از سوی دیگر هر مدرک پارسا که در قالب این ویژگی‌ها کنترل کیفیت شده است دارای ویژگی‌های ذاتی دیگری مانند نام دانشگاه، نام رشته و گرایش و حوزه تخصصی مدرک نیز هست. همچنین، مجموعه داده پیش‌گفته علاوه بر تحلیل آماری به کمک ابزارهای داده‌کاوی (قوانین انجمنی و خوشه‌بندی) نیز مورد تحلیل قرار خواهد گرفت. در این گام تحلیل جامعی بر روی نوع خطاها، اقلام اطلاعاتی مربوط به خطاها و حوزه تخصصی پارساها و اثرگذاری آن‌ها بر روی خطاها موردبررسی قرار گرفت. همچنین میزان همبستگی خطاها بر روی یکدیگر (در مورد خطاهای پرتکرار) نیز در این بخش انجام شد. یکی از گزارش‌های کاربردی در این حوزه، خطاهای دارای فراوانی زیاد در هر حوزه تخصصی است. در این گام مجموعه داده خطاها، تحلیل عددی شده و الگوهای موجود در این مجموعه داده مانند نوع خطاها در حوزه‌های تخصصی گوناگون و همبستگی میان خطاها بررسی شد. گام اول را می‌توان در دو بند زیر خلاصه نمود.

۱-۱) با استفاده از روش‌های آماری مشخصات و ویژگی‌های مجموعه داده خطاها استخراج شد.

۱-۲) با استفاده از روش‌های داده‌کاوی الگوهای موجود در داده بررسی شد.

- گام‌های مرتبط با قوانین انجمنی:
  - بررسی پایگاه داده و تحلیل داده‌های نویز؛
  - کمی‌سازی ویژگی‌های کیفی؛
  - استانداردسازی داده‌ها؛
  - استخراج کدهای رد مرتبط با هر پایان‌نامه و طراحی پایگاه داده جدید متناسب با استخراج قوانین انجمنی؛
  - بررسی فراوانی حالت‌های مختلف تکرار قوانین مختلف؛
  - شناسایی و انتخاب قوانین قوی یا معتبر؛
- گام‌های مرتبط با روش خوشه‌بندی:
  - بررسی پایگاه داده و تحلیل داده‌های نویز؛
  - کمی‌سازی ویژگی‌های کیفی؛
  - استانداردسازی داده‌ها؛
  - اعمال روش‌های خوشه‌بندی؛
  - انتخاب مشخصه مناسب برای تعیین تعداد خوشه مناسب؛
  - بررسی ویژگی‌های خوشه‌ها؛

یافته‌های طرح پژوهشی

برای استفاده از خوشه‌بندی k-means، تمام ویژگی های یک رکورد در نظر گرفته شد و در بازه ۰ تا ۱ استانداردسازی شد، بطوریکه بیشترین عدد از یک ویژگی به یک و کوچکترین عدد به صفر تصویر شد. برای تعیین تعداد مناسب خوشه‌ها از معیار Sep-Comp استفاده شده است. روند افزایش معیار به ازای افزایش تعداد خوشه از ۱ تا ۷ نمایی بوده است و پس از آن روند صعودی کند می شود. بنابراین به نظر می‌رسد ۷ خوشه مناسب‌ترین تعداد خوشه برای این پایگاه داده است. در هر خوشه یک حوزه خاص فراوانی بیشتری دارد و هر خوشه می تواند نماینده یک حوزه خاص باشد. حوزه هر فایل نقش مهمی در خوشه‌بندی داشته است و مثلا حوزه کشاورزی در خوشه ۱ با علوم انسانی بیشترین شباهت را دارد و با حوزه فنی مهندسی کمترین شباهت را دارد. برای بررسی همبستگی و درک ارتباط بین خطاها و هر خوشه، فراوانی خطاها در هر خوشه هم بررسی شد. الگوی تکرار خطاها در این خوشه‌ها بسیار شبیه به هم هستند. بررسی ها نشان داد که تفاوت اندکی در الگوی تکرار خطا در خوشه‌ها وجود داشت. در همه خوشه ها خطای ۶۰۰، ۶۰۷، ۶۸۰ جزء خطای پرتکرار و خطاهای ۶۰۹ ، ۶۷۲ ، ۵۳۳ سه خطای کم تکرار بودند. از نتایج خوشه بندی می توان نتیجه گرفت که همبستگی ای بین ویژگی های ذاتی از جمله حوزه هر فایل و خطاها وجود ندارد. بنابراین در بررسی خطاها و ایجاد قوانین بهبود کیفیت فایلها، برای تمام فایلها می توان یکسان عمل کرد.

نتیجه‌گیری و پیشنهادها

پایگاههای اطلاعاتی پژوهشی یکی از مهمترین پایگاههای اطلاعاتی هستند که نقش بسزایی در امر پژوهش دارند. روزانه هزاران سند تحقیقاتی تولید می‌شود و در یک یا چند پایگاه اطلاعاتی شناخته شده ذخیره می‌شوند.کیفیت داده در این پایگاهای اطلاعاتی نقش مهمی در بازیابی صحیح و قابل اعتماد اطلاعات برعهده دارد. برخی از پایگاههای اطلاعاتی عمومی‌تر هستند، به این معنا که اطلاعات پژوهشی در هر حوزه و با هر موضوع و قالبی در آنها ذخیره می‌شود. در این پایگاهها که الگوی نوشتاری خاصی تعریف نمی‌شود احتمال خطا بیشتر است. بررسی این خطاها به صورت دستی بسیار زمانبر و پر هزینه است. در این تحقیق روشهای داده کاوی برای تحلیل و بررسی مجموعه داده خطاها بررسی شد. برای این منظور در یک مطالعه موردی، از پایگاه اطلاعاتی گنج، یک مجموعه داده شامل اطلاعات فراداده فایل(ویژگی های ذاتی) و مشکلات کیفی و خطاهای مرسوم تهیه شد. سپس با استفاده محاسبه فراوانی خطاها و قوانین انجمنی مجموعه بهینه از خطاهای بارز و تعیین کننده ایجاد شد. نتایج نشان داد با در نظر گرفتن ۲۵ ٪ از خطاها (۱۵ خطا در اینجا) می توان تا ۸۰ درصد از کل خطاهای فایلها را حذف کرد. به این ترتیب واحد کنترل کیفیت با در نظر گرفتن لیست بهینه خطاها با صرف انرژی و وقت کمتری می تواند مشکلات کیفی را مرتفع سازد. برای کشف رابطه و وابستگی بین الگوی تکرار خطاها و ویژگی‌های ذاتی فایل (حوزه تخصصی، دانشگاه، نام نویسنده و ...) از روش خوشه‌بندی استفاده شد. در روش خوشه‌بندی، ویژگی‌های بارز در هر خوشه که نقش اساسی در تمایز خوشه‌ها دارند، مشخص می‌شود. نتایج K-means بر داده آزمایشی نشان داد، ویژگی های ذاتی و به خصوص حوزه هر فایل منجر به تمایز فایلها و قرار گرفتن آنها در خوشه‌های متفاوت شده است. الگوی تکرار خطاها در هر خوشه هم بررسی شد و این الگو در خوشه‌ها شبیه هم بود و وابستگی‌ای بین خطاها و حوزه فایلها و سایر ویژگی ها دیده نشد. در نتیجه در مورد این داده آزمایشی قوانین رفع خطاها و بهبود کیفیت داده می تواند مستقل از ویژگی‌های ذاتی فایلها ارئه شود.