به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصه گزارش برای پژوهشگران

طراحی و پیاده سازی آزمایشی موتور سامانه هوشمند استخراج پژوهشگران مشابه

۱۳۹۶

جلال الدین نصیری

معرفی پژوهش

پژوهشگاه علوم و فناوری اطلاعات با جمع‌آوری و پردازش تمام پایان‌نامه‌ها و رساله‌های وزارت علوم، تحقیقات و فناوری از جایگاه ممتاز و روبه رشدی برای دانشجویان، اساتید و مدیران برخوردار شده است. ارائه سرویس‌های ارزش افزوده برای سیاست‌گذاران و تصمیم‌گیرندگان حوزه علم و فناوری و امکان تجاری‌سازی دانش در جهت تثبیت نقش مرجعیت پژوهشگاه می‌تواند مفید باشد. استخراج تخصص پژوهشگران، مطالعه روند فعالیت علمی پژوهشگران، پژوهشگر در یک نگاه نمونه ای از سرویس‌های ارزش افزوده با محوریت پژوهشگرمی‌باشد.

در پایگاه داده ثبت ایرانداک، یک پژوهشگر مشخص به اشکال مختلف در پایگاه داده ثبت شده است. دلایلی مانند چند املائی بودن برخی نام‌ها، عدم دقت دانشجو، غلط‌های املائی را می‌توان برای علت گوناگونی نام پژوهشگر نام برد. از طرف دیگر تا زمانی که زیرساخت یکتاسازی پژوهشگران ترمیم نگردد طراحی و پیاده‌سازی سامانه‌های ارزش افزوده در حوزه پژوهشگران با شکست روبرو خواهد شد.

با توجه به تعداد بالای پژوهشگران، در این طرح پژوهشی ابتدا با استفاده از روش‌های زبان‌شناسی رایانشی و مفاهیم کلان داده یک معماری برای استخراج پژوهشگران مشابه طراحی می‌گردد و در ادامه نسخه آزمایشی آن پیاده‌سازی خواهد گردید. این طرح پژوهشگران مشابه را فقط با تطابق هوشمند رشته‌ای (ساختاری) بررسی خواهد کرد و در حوزه معنایی ورود نخواهد کرد.

روش پژوهش

ایجاد سامانه‌های متعدد تحلیل داده بر مبنای داده‌های مخدوش و ناسازگار[1] نتیجه مورد نظر را در پی نخواهد داشت. بنابراین اولین قدم در طراحی و پیاده‌سازی سامانه‌های ارزش افزوده در حوزه پژوهشگران ایجاد زیرساخت یکتاسازی پژوهشگران می‌باشد. این مرحله که جزو استراتژی پاک سازی داده[2] نیز می‌باشد در بعضی از سناریوها بسیار سخت و پیچیده می‌باشد. به طور کلی موارد ذیل را برای ضرورت و اهداف طرح می‌توان نام برد:

- تحلیل اقلام اطلاعاتی قابل اتکا برای یکتاسازی پژوهشگران
- بررسی و توسعه آزمایشی الگوریتم تطبیق تقریبی رشته
- با توجه به حجم بالای پژوهشگران (حدود ۷۰۰ هزار پژوهشگر) چگونه با هرس هوشمندانه و تکنیک های پردازش موازی در زمان کارا، الگوریتم اجرا شود.

یافته‌ها

تعداد ۳۰ هزار پژوهشگر تکراری استخراج گردید و برای استفاده و تمیزسازی به گروه مربوطه تحویل داده شده است.

نتیجه گیری و پیشنهادها

همانطور که فصل‌های پیشین اشاره شد، عدم وجود اقلام اطلاعاتی مفید مانند شماره شناسنامه، نام پدر، محل تولد و غیر معتبر بودن شناسه ملی، سازمان پژوهشگر(با توجه به اینکه سازمان مستند ثبت شده است) فرایند اعتبارسنجی و یگانه‌سازی پژوهشگران مشابه را با چالش روبرو ساخته است.

از طرف دیگر الگوریتم به کار رفته با استفاده از تکنیک‌های موازی سازی موفق شده است از جنبه متنی پژوهشگران مشابه را به خوبی استخراج و معرفی کند. به نظر می‌رسد استفاده از نیروی انسانی در کنار موتور توسعه داده شده آزمایشی یک رهیافت مناسب برای بهبود کیفیت اطلاعات پژوهشگران می‌باشد.

از طرف دیگر دو پژوهشگر می‌توانند داری نام و نام خانوادگی دقیقا برابر بوده ولی ذاتا دو فرد مجزا باشند. این موضوع در دامنه طرح نبوده ولی در حین اجرای طرح پژوهشی نمونه‌هایی برخورد شده است.

با توجه به تجزیه و تحلیل‌های دادگان مرتبط با پژوهشگران، به نظر می رسد برای بهبود کیفیت اطلاعات مجموعه اقدامات ذیل باید انجام پذیرد. برای بهبود کیفیت داده معمولا در دو بخش فعالیت انجام می‌گردد، افزایش کیفیت

[1] inconsistent
[2] Data cleansing

ورود داده و افزایش کیفیت دادگان ثبت شده که در ادامه پیشنهادات برای هر دو بخش ارایه شده است.

پیشنهادات افزایش کیفیت دادگان پژوهشگر در زمان ورود اطلاعات:

- طراحی فرم‌هایی که اطلاعات بیشتر و مفیدتری از پژوهشگر در سامانه ثبت دریافت شود. به طور مثال جنسیت، محل تولد و شماره شناسنامه بسیار ضروری می‌باشد.
- اضافه نمودن کنترل‌هایی مفهومی که از ورود اقلام اطلاعاتی نادرست جلوگیری گردد. به طور مثال در قسمت نام فارسی عدد یا کد ملی از جنبه ساختاری بررسی گردد.
- طراحی ثبت اقلام اطلاعاتی برای سازمان وابسته پژوهشگر نه سازمان وابسته به پایان نامه، رساله.

پیشنهادات افزایش کیفیت دادگان پژوهشگر پس از ثبت اطلاعات:

- استفاده از موتور آزمایشی توسعه داده شده در کنار نیروی انسانی
- در صورت امکان افزودن اطلاعات دیگر پژوهشگران از پایگاه داده دیگر مانند اطلاعات اعضای هیات علمی وزارت علوم، تحقیقات وفناوری، نتایج استخراج شده بهبود چشمگیری خواهد داشت.