گزارش نشست علمی

«تشکیل هوشمند پرونده‌های حدیثی و شبکه احادیث با کاربست هوش مصنوعی»

اشاره

نشست علمی «تشکیل هوشمند پرونده‌های حدیثی و شبکه احادیث با کاربست هوش مصنوعی»، روز چهارشنبه ۱۶ مهرماه ۱۴۰۴ با ارائه حجّت‌الاسلام والمسلمین علیرضا شهبازی برگزار شد. این نشست که در حاشیه «نمایشگاه دستاوردهای فنّاورانه علوم اسلامی و انسانی دیجیتال» برگزار گردید، یکی از چهار نشست علمی تخصّصی این رویداد بود؛ نمایشگاهی که با هدف ارائه آخرین پیشرفت‌ها و دستاوردهای فنّاورانه در حوزه علوم اسلامی و انسانی و فنّاوری‌های دیجیتال نوین و نیز ایجاد بستری برای گفت‌وگوی علمی و تخصّصی میان پژوهشگران و فعّالان این عرصه، برگزار شد.

در این جلسه، سخنران به معرّفی «پیکره حدیثی رضوان» پرداخت؛ پیکره‌ای جامع که بالغ بر یک میلیون و دویست هزار حدیث از فریقین را در برمی‌گیرد و با استفاده از الگوریتم‌های هوش مصنوعی، به‌صورت هوشمند پردازش، تحلیل و کارآمد شده است. در این پیکره، متون حدیثی همراه با تفکیک سند و متن، تطبیق منابع، تحلیل‌های رجالی و قابلیت‌های استنتاجی ارائه می‌شود و زمینه شکل‌گیری پرونده‌های هوشمند حدیثی و شبکه‌های ارتباطی میان احادیث و راویان را فراهم می‌سازد. «سامانه حدیثی ثقات» و رابطه آن با «رضوان»، از دیگر مطالب ارائه شده در این نشست علمی بود.

این دو پروژه و سایر طرح‌های مشابه، حاصل تلاش پژوهشگران مؤسّسه ناصر جامع مجتهد (نجم) است؛ مؤسّسه‌ای که از شهریور ۱۳۹۷ به‌صورت غیررسمی با هدف به‌کارگیری مرزهای فنّاوری «هوش مصنوعی توضیح‌پذیر» در فرآیند اجتهاد آغاز به کار کرده است و امروز نقشی پیشرو در توسعه ابزارهای هوشمند علوم اسلامی ایفا می‌کند.

آنچه در ادامه می‌آید، گزارشی تحلیلی و بازپرداخته از سخنان ارائه‌شده در این نشست است.

واژگان کلیدی: پیکره رضوان، سامانه ثقات، استنتاج ماشینی، گراف دانش روایی، برچسب‌گذاری حدیث.

طرح مسئله

پیکره رضوان، اقیانوسی از احادیث است. این سامانه، جامع‌ترین پایگاه داده هوشمند احادیث شیعه و سنّی است که با قدرت هوش مصنوعی غنی شده است. تصمیم بر راه‌اندازی و تهیه پیکره رضوان، سابقه‌ای دارد که به وجود مجموعه‌ای از چالش‌ها و مشکلات انباشته در حوزه مطالعات حدیثی و فعّالیت‌های مرتبط با حاشیه‌نگاری و تحلیل متون برمی‌گردد؛ مشکلاتی که طیّ چهار تا پنج سال گذشته به‌صورت جدی خود را نشان داده و عملاً مسیر فعّالیت‌های پژوهشی را با دشواری‌های قابل توجّهی مواجه کرده بود. تجربه میدانی نشان می‌داد که حجم بالایی از این دشواری‌ها، نه ناشی از کمبود داده یا ضعف علمی، بلکه حاصل کارهای موازی، غیر هم‌افزا و فاقد ساختار یکپارچه در میان نهادها و گروه‌های پژوهشی مختلف است.

چراکه در عمل، نهادها و تیم‌های متعدّد، هریک به‌صورت مستقلّ، به استخراج و پردازش احادیث می‌پرداختند؛ برای نمونه، در حوزه تاریخ حدیث یا تصحیح متون روایی، گروه‌هایی وجود داشتند که احادیث را از نرم‌افزارهایی مانند «جامع الأحادیث» استخراج

پیکره رضوان، اقیانوسی از احادیث است. این سامانه، جامع‌ترین پایگاه داده هوشمند احادیث شیعه و سنّی است که با قدرت هوش مصنوعی غنی شده است. تصمیم بر راه‌اندازی و تهیه پیکره رضوان، سابقه‌ای دارد که به وجود مجموعه‌ای از چالش‌ها و مشکلات انباشته در حوزه مطالعات حدیثی و فعّالیت‌های مرتبط با حاشیه‌نگاری و تحلیل متون برمی‌گردد؛ مشکلاتی که طیّ چهار تا پنج سال گذشته به‌صورت جدی خود را نشان داده و عملاً مسیر فعّالیت‌های پژوهشی را با دشواری‌های قابل توجّهی مواجه کرده بود

کرده، آنها را در قالب فایل‌های Word و یا در بهترین حالت در برخی نرم‌افزارهای تحلیل کیفی وارد می‌کردند و سپس، بر اساس نیاز پژوهشی خود، اقدام به تشکیل خانواده‌های حدیثی یا دسته‌بندی‌های موضوعی می‌نمودند. این فرایندها، گاه در قالب فایل‌های ساده متنی و گاه در پایگاه‌های داده‌ای محدود و اختصاصی انجام می‌شد که صرفاً برای همان تیم یا پروژه قابل استفاده بود.

مسئله اساسی آن بود که میان این فعّالیت‌های موازی، هیچ اتّصال معنادار و استانداردی وجود نداشت؛ حتّی در مواردی که چند گروه به طور هم‌زمان روی یک موضوع واحد – مانند تصحیح کتاب تهذیب – کار می‌کردند، به دلیل تفاوت منابع خام، شیوه‌های استخراج و قالب‌های ذخیره‌سازی، اتّصال و تطبیق داده‌ها به کاری بسیار پُرهزینه و پیچیده تبدیل می‌شد. در بهترین حالت، برای تشخیص اینکه دو حدیث استخراج‌شده در پروژه‌های مختلف، در واقع یک حدیث واحد هستند، نیاز به تطبیق‌های دستی و زمان‌بر وجود داشت.

این وضعیت، به دلیل فقدان ساختارهای داده‌ای ساخت‌یافته (Structured)، موجب افزایش چشمگیر هزینه‌های زمانی، انسانی و حتّی مالی می‌شد و پروژه‌ها را هم طولانی‌تر و هم پُرهزینه‌تر می‌کرد.

در مسیر سند، اسامی راویان به‌صورت خودکار استخراج می‌شدند و با بهره‌گیری از استنتاج‌ورزی ماشینی و ساختارهای استدلالی، تلاش می‌شد تشخیص داده شود که هر فرد، دقیقاً چه کسی است و بر اساس چه قراینی این تطبیق انجام می‌گیرد. این فرایند، به‌صورت ماشینی انجام می‌شد و نمونه‌های آن، در پروژه «ثقات» قابل مشاهده است

این مشکل، در مسیر اجرای پروژه‌های پیشین، به‌ویژه در فضای مرتبط با «ثقات»، به‌صورت کاملاً ملموس احساس شد. روشن شد که اگر قرار باشد پیکره‌های حدیثی موجود – که طیّ سال‌ها تلاش تولید شده‌اند – مجدداً و بدون زیرساخت مناسب مورد استفاده قرار گیرند، عملاً همان چرخه ناکارآمد تکرار خواهد شد. با وجود تلاش‌ها برای ایجاد ارتباط و همکاری با پیکره‌های ارزشمند موجود در مرکز تحقیقات کامپیوتری علوم اسلامی (نور)، به دلایل فنّی و ساختاری، امکان بهره‌برداری یکپارچه از آنها فراهم نشد.

در اینجا تأکید می‌شود که طرح این مسئله، به‌هیچ‌وجه به معنای نقد یا نادیده‌گرفتن خدمات گسترده و ارزشمند مرکز تحقیقات کامپیوتری علوم اسلامی نیست. خدمات این مؤسّسه به جهان اسلام و مکتب تشیّع، نقشی بنیادین و غیرقابل انکار داشته است. آنچه مطرح می‌شود، صرفاً ناظر به یک نیاز خاصّ در حوزه داده‌های خام، پردازش‌پذیر و قابل استفاده در پروژه‌های هوشمند است؛ نیازی که پاسخ‌گویی به آن، مستلزم رویکردی متفاوت بود.

مشکل اصلی، آن بود که در بسیاری از نرم‌افزارهای موجود، امکان دسترسی به داده خام به شکلی که بتوان آن را در پروژه‌های دیگر به کار گرفت، وجود نداشت؛ درحالی‌که پژوهش‌های نوین، نیازمند آن هستند که حدیث نه فقط به‌عنوان متنی برای جست‌وجو، بلکه به‌عنوان یک موجودیت مستقلّ داده‌ای تلقّی شود؛ موجودیتی که بتوان آن را برچسب‌گذاری کرد، در شبکه‌ای از مفاهیم و داده‌های رجالی قرار داد و میان نهادها و پروژه‌های مختلف به اشتراک گذاشت.

شکل‌گیری ایده پیکره حدیثی رضوان

بر همین اساس، طیّ پنج تا شش ماه اخیر، تصمیم بر آن شد که مسیر جدیدی در قالب طراحی پیکره حدیثی همراه با ارزش افزوده هوشمند دنبال شود. تأکید اصلی، بر این بود که به‌جای تولید یک پیکره خام، «پیکره‌ساز» طراحی شود؛ سامانه‌ای که تمام مراحل لازم برای پردازش، تفکیک و غنی‌سازی احادیث را به‌صورت نظام‌مند انجام دهد.

از اسفند سال ۱۴۰۳، فرایند طراحی پیکره آغاز شد. نخستین گام، انتخاب متنی بود که از نظر حقوقی مشکلی نداشته باشد و از طرفی، وامدار نهادهای غیرشیعی هم نباشیم و درعین‌حال، از کیفیت قابل‌قبول برخوردار باشد. گزینه‌های مختلفی بررسی شد؛ از جمله «مکتبة الشاملة» و برخی پروژه‌های بین‌المللی پیکره‌سازی علوم اسلامی که با وجود سرمایه‌گذاری‌های قابل‌توجّه، از نظر

کیفیت و میزان ارزش افزوده، پاسخ‌گوی نیازهای پروژه نبودند.

در نهایت، نرم‌افزار «مکتب اهل‌بیت» به‌عنوان متن پایه انتخاب شد؛ نرم‌افزاری که امکان دسترسی به داده‌های باز و استخراج پایگاه داده را فراهم می‌کرد. این برنامه، خصوصیات را به‌صورت جامع در خودش داشت. البته نقطه‌ضعف‌هایی هم داشت که بعداً بیان خواهد شد. بر اساس این انتخاب، دو دوره پیکره‌سازی انجام شد که یکی از آنها به‌منظور آزمون امکان‌سنجی (Feasibility) طراحی شده بود. نسخه اوّلیه حاصل از این مرحله، مبنای رونمایی اوّلیه پروژه ثقات در دی‌ماه سال گذشته قرار گرفت؛ نسخه‌ای که بیش از آنکه محصول نهایی باشد، کارکرد آزمایشی و ارزیابانه داشت.

پس از شکل‌گیری ایده کلی، فرایند طراحی پیکره به‌صورت جدی وارد مرحله اجرا شد. در این مرحله، تمرکز اصلی بر کتاب‌ها و منابع متنی قرار گرفت. در نرم‌افزار «مکتب اهل‌بیت»، حجم قابل توجّهی از کتاب‌ها در دسترس بود و از میان آنها، حدود ۱۲۰۰ کتاب انتخاب شد؛ کتاب‌هایی که به نحوی در مظان وجود حدیث قرار داشتند یا احتمال می‌رفت که حاوی نقل‌های حدیثی باشند.

نکته قابل‌توجّه آن بود که این انتخاب، محدود به کتاب‌های صِرفاً حدیثی نبود. افزون بر آثار روایی، کتاب‌های فقهی، اصول فقهی، لغوی و حتّی برخی آثار غیرحدیثی نیز وارد این مجموعه شدند؛ زیرا تجربه نشان داده است که گاه یک نکته حدیثی مهم، در اثری مانند «عده» شیخ طوسی یا دیگر متون غیرحدیثی آمده و در منابع دیگر تکرار نشده است. به همین دلیل، سیاست کلّی بر آن قرار گرفت که هر کتابی – در هر موضوعی – که تا پیش از قرن ۵ تألیف شده و احتمال وجود حدیث در آن می‌رفت، وارد این مسیر شود.

در خصوص کتاب‌های لغت، به‌ویژه آثار قرون اوّلیه، چالش‌ها به‌مراتب بیشتر بود. کتب لغت قرن ۳ و ۴، به دلیل ساختار خاصّ زبانی و شیوه ارائه مطالب، فرایند جداسازی حدیث را دشوارتر می‌کرد. بااین‌حال، تمامی کتاب‌های لغوی تا پیش از قرن ۵، در صورت احراز شرایط، به مجموعه افزوده شدند. این رویکرد، باعث شد دامنه پیکره، از همان ابتدا گسترده و جامع طراحی شود.

در ادامه، کتاب‌ها بر اساس حوزه‌های موضوعی دسته‌بندی شدند؛ برای مثال، در حوزه مطالعات قرآنی، از میان کتاب‌های مکتب اهل‌بیت، در نهایت، ۴ کتاب انتخاب شد که حاصل آن، استخراج حدود ۱۱۰ هزار حدیث از این مسیر بود. در حوزه تاریخ، ۲۶۴ کتاب تاریخی وارد فرایند شد که از آنها، حدود ۱۸۸ هزار حدیث استخراج و تفکیک شد.

معماری پایپ‌لاین پردازش حدیث

در همه این موارد، کتاب‌ها پس از ورود به این مسیر، وارد یک «پایپ‌لاین» یا فرایند کارخانه‌ای می‌شدند؛ فرایندی که به‌صورت خودکار، کتاب را مرحله‌به‌مرحله پردازش می‌کرد. هر کتابی که وارد این مسیر می‌شد، ابتدا توسط یک ماژول اختصاصی، احادیث آن از متن کلّی جدا می‌گردید. سپس، ماژول دیگری وظیفه تفکیک سند از متن حدیث را بر عهده داشت. از این نقطه به بعد، مسیر پردازش، به دو شاخه اصلی تقسیم می‌شد:

یکی از کارکردهای مهم مشابهت مضمونی، امکان مشاهده فضای عمومی پیرامون یک مضمون خاصّ است؛ برای نمونه، در حدیث «الحسن والحسین سیدی شباب اهل الجنة»، کاربر می‌تواند علاوه بر متن اصلی حدیث، مجموعه‌ای از احادیث مرتبط را در لایه‌های عمیق‌تر مشاهده کند؛ احادیثی که در همان فضای مفهومی و ارزشی قرار دارند.

علاوه بر این، ایده‌های دیگری همچون: استخراج روابط جزء و کل، اعمّ و اخص و نیز تشخیص موارد تقطیع حدیثی مورد توجّه قرار گرفت

۱. مسیر پردازش سند حدیث؛ ۲. مسیر پردازش متن حدیث.

در مسیر سند، اسامی راویان به‌صورت خودکار استخراج می‌شدند و با بهره‌گیری از استنتاج‌ورزی ماشینی و ساختارهای استدلالی، تلاش می‌شد تشخیص داده شود که هر فرد، دقیقاً چه کسی است و بر اساس چه قراینی این تطبیق انجام می‌گیرد. این فرایند، به‌صورت ماشینی انجام می‌شد و نمونه‌های آن، در پروژه «ثقات» قابل مشاهده است.

در رابطه با متن، مراحل متعدّدی طیّ می‌شد؛ مواردی مانند: اِعراب‌گذاری متن حدیث، ترجمه حدیث به زبان‌های مختلف، شناسایی لغات کلیدی و آماده‌سازی داده برای استفاده در محیط‌های چندزبانه. ترجمه‌ها شامل: زبان‌های فارسی، عربی، انگلیسی و سایر زبان‌های مورد نیاز بود. در کنار این موارد، داده‌هایی که در مظانّ استفاده‌های حدیث‌پژوهانه بودند نیز به این فرایند افزوده می‌شدند.

تشکیل شبکه حدیثی و ارتباط میان منابع

پس از تکمیل این مراحل، حدیث وارد فاز «شبکه‌سازی» می‌شد. در این مرحله، بررسی می‌شد که یک حدیث مشخّص در کدام منابع دیگر تکرار شده است. این، همان مفهومی است که در نرم‌افزارهای نور تحت عنوان «گروه‌بندی احادیث» شناخته می‌شود؛ امّا در اینجا با دقّت و گستره بیشتری دنبال شد.

شبکه‌سازی، شامل: شناسایی شباهت‌های لفظی، معنایی و حتّی مضمونی میان احادیث بود. همچنین، با این هدف که در آینده ارتباط کامل‌تری برقرار شود، این پیکره حدیثی به پایگاه نور متّصل شد؛ به گونه‌ای که هر حدیث دارای شناسه مشخّص خود در نور باشد و ارتباط آن با دیگر پایگاه‌ها مانند «پایگاه مجلّات تخصّصی نور» و سایر سرویس‌های مرتبط برقرار شود.

در حوزه کتب حدیثی، حدود ۹۸۴ هزار حدیث از مجموعه آثار مکتب اهل‌بیت استخراج شد. برخلاف برخی دسته‌ها، در این بخش محدودیت «قبل از قرن ۵» لحاظ نشد، تا آثاری مانند «وسائل الشیعه» نیز امکان ورود به این ساختار را داشته باشند. همچنین، کتاب‌های فقهی، اصول فقهی، ادبیات عرب و عقاید که در نرم‌افزار مکتب اهل‌بیت دسته‌بندی شده بودند، همگی از این مسیر خودکار عبور داده شدند.

در مقطعی، تعداد احادیث به حدود ۱,۳۰۰,۰۰۰ رسید؛ امّا پس از اعمال مسیرهای ارزیابی و پالایش، بخشی از داده‌ها – مانند مواردی که صرفاً نقل قول مصنّف بودند – حذف شد و حجم نهایی، به حدود ۱,۲۵۰,۰۰۰ حدیث رسید. بااین‌حال، ساختار به‌گونه‌ای طراحی شده که هر کتاب جدیدی در صورت تأمین متن، بتواند به‌راحتی وارد این پایپ‌لاین شود و همان فرایند را طیّ کند.

به‌عنوان نمونه، حدیث «اوّل مَا خَلَقَ اللهُ نُوری» از کتاب «عوالی اللئالی» که در «مکتب اهل‌بیت» موجود است، وارد این مسیر شد. متن حدیث، به‌عنوان یک واحد مستقلّ شناسایی گردید، سند آن استخراج شد و سپس، وارد مراحل بعدی از جمله: خلاصه‌سازی ماشینی، اعراب‌گذاری و ترجمه گردید. این فرایند، نشان می‌دهد که چگونه یک حدیث، از متن خام کتاب، به یک موجودیت داده‌ای غنی و قابل شبکه‌سازی تبدیل می‌شود.

معیارهای شناسایی حدیث و چالش‌های فنّی

شناسایی حدیث در متن، بر اساس مجموعه‌ای از پارامترها انجام

می‌شد. در برخی کتاب‌های ساختاریافته، از ابزارهای تشخیص ساختار استفاده شد؛ هرچند این روش، به‌تنهایی قابل اتکا نبود. ازاین‌رو، جست‌وجوهای مبتنی بر الگو به کار گرفته شد؛ مشابه آنچه در جست‌وجوهای پیشرفته نرم‌افزارهای نور مشاهده می‌شود.

در کنار آن، از تحلیل کلمات پُرتکرار در اسناد – مانند «حدثنا»، «أخبرنا»، «روی عنه» و نظایر آن – برای تشخیص پایان سند و آغاز متن حدیث استفاده شد. بااین‌حال، تنوّع ساختار کتاب‌ها به‌گونه‌ای بود که نهایتاً وجود یک «قاضی» یا عامل تصمیم‌گیر هوشمند، ضروری به نظر می‌رسید؛ عاملی که بتواند در شرایط پیچیده، تصمیم نهایی را اتخاذ کند.

برای حلّ این مسئله، از یک عامل هوشمند استفاده شد که میان خروجی‌های مختلف تصمیم‌گیری می‌کرد؛ اینکه کدام بخش حدیث است، سند از کجا تا کجاست و کدام الگو معتبرتر است. پیچیدگی کار، زمانی بیشتر شد که برخی کتاب‌ها، مانند برخی منابع اهل‌سنّت، ساختارهای غیرمعمول داشتند؛ برای مثال، قرارگرفتن سند پس از متن حدیث، یا نبودِ سند در بخش‌هایی از کتاب.

در مرحله شبکه‌سازی، صرف شباهت لفظی کافی دانسته نشد. گاه دو حدیث دارای عبارات مشابه بودند؛ امّا در دو سیاق معنایی متفاوت قرار داشتند؛ مانند احادیثی که یکی درباره ذکر هنگام ورود به بازار، و دیگری درباره ذکر در نماز سخن می‌گفت. ازاین‌رو، تابعی طراحی شد که شباهت لفظی، شباهت معنایی و حتّی طول حدیث را به‌صورت هم‌زمان در نظر می‌گرفت تا بتوان با دقّت ادعا کرد که یک حدیث، همان حدیثِ موجود در منبع دیگر است. عملاً کارکرد آن، این است که در موتورهای جست‌وجو، اگر یکی از آن را پیدا کرد، می‌توان به او گفت بقیه آن را نشان نده؛ چون این، همان حدیث است.

بحث دیگر، «احادیث مشابه لفظی» است. بحث مشابهت‌های لفظی، الگوریتم‌های خاصّ خود را دارد و موضوع جدیدی هم نیست؛ با روش‌هایی مانند جابه‌جایی کاراکترها و سنجش فاصله‌های متنی می‌توان میزان شباهت را ارزیابی کرد؛ امّا در بحث «حدیث در سایر منابع»، قدری سخت‌گیری اِعمال شده است؛ به این معنا که اگر حدیثی در این بخش شناسایی نشود، همچنان می‌توان آن را از طریق مشابهت‌های لفظی بازیابی کرد. افزون بر این، مرحله مشابهت‌های معنایی نیز در نظر گرفته شده است، تا با استخراج و کنار هم قراردادن متن‌های مرتبط، بتوان احادیث هم‌مضمون را شناسایی و تحلیل نمود.

در همین چارچوب، حدیث «اوّل مَا خَلَقَ اللهُ نُوری» با حدیث «اوّل مَا خَلَقَ اللهُ القَلَم ثُمّ خَلَقَ النُّور» به‌عنوان مشابه معنایی شناسایی شد. در این فرایند، حتّی خطاهای تایپی موجود در منبع اصلی، عمداً اصلاح نشد تا اصالت منبع حفظ شود. اگرچه ماژول تصحیح خطا نیز آزمایش شد، امّا تصمیم بر آن قرار گرفت که در این مرحله، متن اصلی بدون دست‌کاری باقی بماند و اصلاحات احتمالی، به مراحل بعدی موکول شود.

شکل‌گیری ایده «مشاهده مضمونی»

در ادامه مسیر طراحی، ایده دیگری نیز مطرح شد که بر پایه

ترکیب چند ماژول مختلف استوار بود. حاصل این ایده، تعریف نوع تازه‌ای از ارتباط میان احادیث با عنوان «مشاهده مضمونی» بود؛ مفهومی که فراتر از مشابهت صرفِ معنایی عمل می‌کند. در این رویکرد، خروجی ماژول‌های مختلفی همچون: تشخیص مشابهت معنایی، خلاصه‌سازی و استخراج نکات کلیدی، در کنار یکدیگر قرار می‌گیرند تا علاوه بر تشخیص شباهت معنایی، فضای کلّی و پیرامونی یک بحث حدیثی نیز قابل مشاهده باشد.

این قابلیت، به‌ویژه در احادیثی که در حوزه فضایل – مانند احادیث مرتبط با فضایل امیرالمؤمنین (ع) – قرار می‌گیرند، اهمّیّت بسیاری پیدا می‌کند. در چنین مواردی، در لایه مشابهت‌های مضمونی، گاه احادیثی از منابع غیرشیعی یا حتّی در مدح خلفا نیز ظاهر می‌شود؛ زیرا فضای کلّی متن، فضای مدح شخصیت‌های پس از پیامبر اکرم (ص) است. این امر، از یک سو فرصتی ارزشمند فراهم می‌کند تا با تشکیل پرونده حدیثی جامع، بتوان زمینه‌های جعل یا الگوبرداری‌های تاریخی را بهتر شناسایی کرد و ازسوی‌دیگر، ممکن است از حیث محتوایی ناخواسته تلقّی شود. به همین دلیل، در استفاده‌های پژوهشی خاصّ – مانند پروژه «ثقات» – با اعمال محدودیت، دایره احادیث به موارد اخص تقلیل داده شد.

«رضوان»، صرفاً یک پیکره حدیثی مستقلّ نیست؛ بلکه قرار است به‌عنوان یکی از پیکره‌های اصلی در یک انباره بزرگ‌تر با عنوان «میقات» در فضای پژوهش‌های اسلامی دیجیتال قرار گیرد. حرکت به‌سوی این ساختار هم‌افزا، ضرورتی اجتناب‌ناپذیر است. همان گونه که در سخنان برخی صاحب‌نظران نیز مطرح شده، مسیر پژوهش‌های اسلامی دیجیتال، مسیری نیست که بتوان آن را به‌صورت جزیره‌ای و منفرد طیّ کرد

در طراحی نهایی، تفکیک دقیقی میان سطوح مختلف شباهت صورت گرفت. مشابهت لفظی، به طور طبیعی دارای شباهت معنایی شدید نیز هست؛ امّا در نمایش نرم‌افزاری، لازم بود این دو سطح، از یکدیگر تفکیک شوند تا کاربر بتواند با دقّت بیشتری داده‌ها را مشاهده کند. مشابهت مضمونی نیز که ترکیبی از چند ماژول تحلیلی بود، در اغلب موارد، در دل خود مشابهت معنایی را نیز در بر داشت.

بر اثر این طراحی، پایگاه داده روابط و پیوندهای میان احادیث به حجمی در حدود ۶۵۰ میلیون رکورد رسید. البته بدیهی است که همه این روابط، به‌صورت مستقیم به کاربر نمایش داده نمی‌شود و تنها سطوح معنادار و کاربردی در خروجی‌ها مورد استفاده قرار می‌گیرد.

یکی از کارکردهای مهم مشابهت مضمونی، امکان مشاهده فضای عمومی پیرامون یک مضمون خاصّ است؛ برای نمونه، در حدیث «الحسن والحسین سیدی شباب اهل الجنة»، کاربر می‌تواند علاوه بر متن اصلی حدیث، مجموعه‌ای از احادیث مرتبط را در لایه‌های عمیق‌تر مشاهده کند؛ احادیثی که در همان فضای مفهومی و ارزشی قرار دارند.

علاوه بر این، ایده‌های دیگری همچون: استخراج روابط جزء و کل، اعمّ و اخص و نیز تشخیص موارد تقطیع حدیثی مورد توجّه قرار گرفت. ممکن است، یک حدیث بخشی از حدیثی دیگر باشد و یا حدیثی در دل حدیثی گسترده‌تر قرار گیرد. این نوع روابط اگرچه در نسخه فعلی «رضوان» پیاده‌سازی نشده‌اند، امّا در برنامه توسعه نسخه‌های بعدی قرار دارند.

طراحی نظام ارزیابی و اصلاح داده‌ها

در مجموع، این فرایند به طراحی یک «پایپلاین» یا کارخانه هوشمند پردازش حدیث و سند انجامید؛ مسیری که احادیث پس از ورود به آن، مراحل مختلف استخراج، تحلیل و شبکه‌سازی را طیّ می‌کنند. پس از پیاده‌سازی این ساختار، ضرورت وجود یک نظام ارزیابی و اصلاح داده‌ها به‌صورت جدی مطرح شد؛ چراکه بدون سنجش دقّت خروجی‌های ماشینی، نمی‌توان درباره کارایی این سامانه داوری علمی داشت.

به همین منظور، یک سامانه ارزیابی طراحی شد تا احادیث تولیدشده در اختیار خبرگان حوزه حدیث قرار گیرد. در مرحله نخست، ۱۲۰۰ حدیث به‌صورت تصادفی از منابع شیعی و سنّی انتخاب شد. این احادیث، توسط پنج نفر از اعضای متخصّص

تیم، به طور مستقلّ بررسی شدند و هر حدیث، دست‌کم دو بار مورد ارزیابی قرار گرفت. بدین ترتیب، بخش‌های مختلف سامانه، از جمله: تشخیص حدیث، جداسازی سند و متن، اِعراب‌گذاری، ترجمه و تحلیل محتوایی، به‌صورت دقیق ارزیابی شدند.

نتایج این ارزیابی‌ها، در مقاله‌ای علمی به طور کامل گزارش شده و آمارهای مربوط به دقّت هریک از ماژول‌ها، در آن قابل مشاهده است.

یکی از ماژول‌های مهم در این مسیر، موضوع‌گذاری احادیث بود. هدف نهایی در این حوزه، دستیابی به یک موسوعه موضوعی خودکار در فضای احادیث است که بر پایه یک آنتولوژی هستی‌شناسانه سامان یابد؛ به این معنا که اگر حدیثی با موضوع «احترام به مادر» برچسب‌گذاری شود، به‌صورت خودکار در موضوعات بالادستی مانند «احترام به والدین» و «آداب معاشرت خانوادگی» نیز قابل مشاهده باشد. چنین ساختاری، امکان تحلیل‌های عمیق‌تر و شبکه‌ای در مطالعات حدیثی را فراهم می‌کند.

در فرایند ارزیابی، برای مقایسه منصفانه، حدود ۲۰۰ حدیث با ترجمه‌ها و اِعراب‌گذاری‌های خبرگانی (مانند ترجمه‌های علمای برجسته) نیز وارد مجموعه ارزیابی شد. نتیجه‌ای که به دست آمد، جالب توجّه بود: در موارد متعدّدی، ترجمه‌های ماشینی تولیدشده توسط سامانه، از منظر ارزیابان، امتیاز بالاتری نسبت به برخی ترجمه‌های کلاسیک دریافت کردند. اگرچه تفاوت زبان فارسی قدیم و معاصر در این قضاوت بی‌تأثیر نبود، امّا این نتیجه نشان‌دهنده ظرفیت بالای سامانه در حوزه ترجمه ماشینی احادیث

نسخه جدید سامانه «ثقات» هم‌اکنون در بخش احادیث خود از رضوان استفاده می‌کند. سیاست کلان، آن است که به‌جای انتقال فایل‌ها، دسترسی از طریق API برقرار شود، تا همه سامانه‌ها از آخرین نسخه داده بهره‌مند شوند و علاوه‌برآن، ارجاعات پژوهشی بر اساس شناسه‌های یکتا انجام گیرد و در نهایت، اتّصال میان پروژه‌های تاریخی، رجالی، خانوادگی و... حفظ شود و بتوان تحلیل‌های ترکیبی تولید کرد

است. در بخش اِعراب‌گذاری نیز، دقّت خروجی ماشینی تقریباً هم‌سطح نمونه‌های انسانی ارزیابی شد.

پیکره رضوان و ضرورت هم‌افزایی در پژوهش‌های اسلامی دیجیتال

«رضوان»، صرفاً یک پیکره حدیثی مستقلّ نیست؛ بلکه قرار است به‌عنوان یکی از پیکره‌های اصلی در یک انباره بزرگ‌تر با عنوان «میقات» در فضای پژوهش‌های اسلامی دیجیتال قرار گیرد. حرکت به‌سوی این ساختار هم‌افزا، ضرورتی اجتناب‌ناپذیر است. همان گ‌ونه که در سخنان برخی صاحب‌نظران نیز مطرح شده، مسیر پژوهش‌های اسلامی دیجیتال، مسیری نیست که بتوان آن را به‌صورت جزیره‌ای و منفرد طیّ کرد.

در شرایطی که نهادهای پژوهشی بین‌المللی و مستشرقان، از بودجه‌ها و زیرساخت‌های فنّاورانه بسیار گسترده‌تری برخوردارند، تنها راه بقا و پیشرفت، هم‌افزایی، تجمیع ظرفیت‌ها و حرکت در قالب ساختارهای مشترک است. در غیر این صورت، خطرِ عقب‌ماندگی یا حذف از عرصه رقابت علمی و فنّاورانه، کاملاً جدی خواهد بود.

بدیهی است که برخی نهادها ممکن است در مقاطعی مایل باشند داده‌های خود را به‌صورت خصوصی نگه دارند؛ همانند همکاری فعلی مؤسّسه معارف با پروژه «ثقات» که داده‌های آن، هنوز عمومی نشده است. این محدودیت‌ها، قابل درک است؛ امّا اصل کلان، آن است که با شکل‌گیری چنین ساختاری، از موازی‌کاری‌های پُرهزینه، به‌ویژه در حوزه‌هایی مانند OCR، جلوگیری شود؛ زیرا منابع مالی و انسانی، به‌اندازه‌ای نیست که هر نهاد مسیرهای مشابه را به طور مستقل طیّ کند.

رضوان، هنوز به طور رسمی رونمایی نشده و ما به‌خوبی آگاهیم که دقّت فعلی پیکره، صددرصد نیست؛ برخی مؤلّفه‌ها در ارزیابی‌ها، بالای ۹۰درصد هستند؛ امّا رسیدن به دقّت‌های ۹۸ یا ۹۹درصد، نیازمند زمان، پالایش‌های متوالی و مشارکت جمعی است.

بر اساس آخرین ارزیابی‌ها که حدود دو ماه پیش انجام شده، دقّت کلّی سامانه حدود ۹۰درصد برآورد می‌شود. در این فاصله، برخی اشکالات سیستماتیک در اِعراب‌گذاری و ترجمه برطرف شده و جهش قابل‌توجّهی در کیفیت این دو حوزه حاصل شده است.

در حوزه ارجاعات قرآنی نیز، اگرچه هنوز ارزیابی رسمی صورت نگرفته، امّا بررسی‌های تصادفی تیم توسعه نشان می‌دهد که برخی از نتایج، دقّت و ظرافت شگفت‌انگیزی دارند.

نسخه جدید سامانه «ثقات» هم‌اکنون در بخش احادیث خود از رضوان استفاده می‌کند. سیاست کلان، آن است که به‌جای انتقال فایل‌ها، دسترسی از طریق API برقرار شود، تا همه سامانه‌ها از آخرین نسخه داده بهره‌مند شوند و علاوه‌برآن، ارجاعات پژوهشی

بر اساس شناسه‌های یکتا انجام گیرد و در نهایت، اتّصال میان پروژه‌های تاریخی، رجالی، خانوادگی و... حفظ شود و بتوان تحلیل‌های ترکیبی تولید کرد. در این ساختار، دیگر خبری از کپی‌برداری به Word و قطع‌شدن پیوند داده‌ها نخواهد بود.

برای رضوان مجوزی در نظر گرفته شده که حتّی استفاده تجاری را نیز مجاز می‌داند؛ بدین معنا که هر نهادی می‌تواند از این داده‌ها محصول بسازد و از آن کسب درآمد کند؛ مشروط بر آنکه: اولاً، منبع «رضوان» را ذکر کند؛ ثانیاً، خود را متعهد بداند که توسعه‌هایی که بر روی داده انجام می‌دهد، به جامعه بازگرداند. این الزام، از جنس تعهد حقوقی سخت‌گیرانه نیست؛ بلکه یک انتظار اخلاقی برای شکل‌گیری چرخه «دانش آزاد و هم‌افزا» است، تا در رقابت جهانی علوم اسلامی دیجیتال، عقب نمانیم.

دسترسی آزمایشی به رضوان

در حال حاضر، یک فرم درخواست برای دسترسی به داده‌ها در نظر گرفته شده است و پس از رونمایی رسمی رضوان که احتمالاً تا دو هفته آینده انجام می‌شود، نمونه داده‌ها و مسیر رسمی دسترسی نیز فعّال خواهد شد. هدف از این کنترل اوّلیه، رصد شیوه استفاده و آماده‌سازی بستر مناسب برای انتشار عمومی‌تر داده‌ها در مراحل بعدی است.

ثقات، پایگاه گراف دانش علوم حدیث است و بر پایه کتب رجالی و تاریخی، راویان و اشخاص مهم در طول پنج سده ابتدایی هجری قمری را در قالب گراف دانش مدل کرده است. با توجّه به این ساختار مدل‌سازی، امکان استنتاج‌های منطقی و حدسی (محاسباتی)، تغییر مبنای رجالی، بازیابی و جست‌وجوهای پیچیده، نمایش‌های نقشه زمان – مکان، نموداری، گرافی و مقایسه‌ای فراهم شده است

سامانه حدیثی «ثقات»

پروژه «ثقات»، از یک بانک اطّلاعاتی ساده عبور کرده و به یک سامانه استنتاجی کاربرمحور برای دانش رجال و حدیث تبدیل شده است.

ثقات، پایگاه گراف دانش علوم حدیث است و بر پایه کتب رجالی و تاریخی، راویان و اشخاص مهم در طول پنج سده ابتدایی هجری قمری را در قالب گراف دانش مدل کرده است. با توجّه به این ساختار مدل‌سازی، امکان استنتاج‌های منطقی و حدسی (محاسباتی)، تغییر مبنای رجالی، بازیابی و جست‌وجوهای پیچیده، نمایش‌های نقشه زمان – مکان، نموداری، گرافی و مقایسه‌ای فراهم شده است. ثقات، در تعامل با کاربرهای خود می‌باشد و هرکسی می‌تواند اطّلاعات رجالی خویش را در آن وارد کرده و بر اساس مبنای رجالی خود، نتیجه مورد نظر را دریافت کند.

در حال حاضر، حدود ۴۵ هزار راوی شیعه با معیارهای درایه‌ای وارد سامانه شده‌اند و تلاش شده ساختار اطّلاعاتی آنان، مشابه نرم افزار درایة النور مرکز تحقیقات کامپیوتری علوم اسلامی باشد؛ به گونه‌ای که اگر روزی امکان اتّصال این سامانه به پروژه‌های مرکز نور فراهم شد، مشکل عدم معیارمندی به حدّاقل برسد.

راویان اهل‌سنّت نیز افزوده شده‌اند؛ امّا به دلیل فقدان معیارهای یکپارچه در برخی منابع اهل‌سنّت، بخشی از داده‌ها با روش‌های تولید و استنتاج ماشینی بازسازی شده است.

تمامی توصیفات راویان بر پایه مقدّماتِ قابل حذف و بازبینی ساخته شده‌اند؛ برای مثال، اگر زراره «ادیب» معرّفی شده، این حکم، حاصل ترکیب گزارش نجاشی، اعتبار منبع و الگوریتم استنتاج است. کاربر می‌تواند هر مقدّمه‌ای را حذف کند. در این صورت، نتیجه نیز به‌صورت خودکار به‌روزرسانی می‌شود؛ به بیان دیگر، هیچ صفتی برای راویان، «قطعیِ بدون مقدّمه» نیست؛ همه چیز، بر پایه شاهد تاریخی است.

دانش راویان، در سامانه به دو لایه تقسیم شده است:

ـ حسی: گزارش‌های مستقیم تاریخی، مانند: «نجاشی چنین گفت»؛

ـ حدسی: نتایج استنتاجی ماشین، مانند: کنیه، شغل، محلّ اقامت محتمل و....

این تفکیک، اجازه می‌دهد کاربر دقیقاً بداند کدام داده، «گزارش» است و کدام‌یک «نتیجه تحلیل» است.

با توجّه به کمبود داده‌های صریح تاریخی درباره محلّ زندگی یا زمان اقامت بسیاری از راویان، سامانه از طریق نسبت‌های روایی، استاد – شاگردی و هم‌عصری، داده‌های محلّ اقامت احتمالی و بازه زمانی زندگی را استخراج می‌کند.

نتیجه این فرایند، تولید نقشه‌های مکانی – زمانی از راویان است که امکان تحلیل شبکه‌ای تاریخی را فراهم می‌کند.

در سامانه «ثقات»، هر راوی یک گره در یک گراف دانشی بزرگ است. این گراف، امکان موارد زیر را می‌دهد:

ـ مقایسه دو راوی با یکدیگر؛

ـ تحلیل شبکه ارتباطات روایی؛

ـ ترسیم نمودارهای دایره‌ای از همبستگی‌ها؛ مثلاً: نسبت شغل‌ها در میان فرق مختلف شیعه (امامی، واقفی و...).

نکته‌ای که وجود دارد، این است که ما به دلیل جداسازی ماشینیِ متن حدیث از سند، علاقه‌مند بودیم این تفکیک، به‌صورت شفاف قابل مشاهده باشد. همین مسئله، باعث شد از مقطعی به بعد، به سوی توسعه یک «کتاب‌خوان» مستقلّ حرکت کنیم. در این کتاب‌خوان، همان بخشی که عرض شد، به‌وسیله یک ماژول هوشمند، حدیث را از دل متن به‌صورت خودکار استخراج و جدا می‌کند.

کتاب‌خوان هوشمند حدیثی، دارای ویژگی‌های زیر است:

* سند و متن حدیث را به‌صورت ماشینی از هم جدا می‌کند؛
* اِعراب‌گذاری را فعّال و غیرفعّال می‌کند؛
* شناسنامه هر حدیث را نشان می‌دهد که از کدام منبع استخراج شده است.

نمونه‌هایی مانند صحیح بخاری نشان می‌دهد که در برخی نسخه‌ها، حتّی تفکیک ابتدایی سطرها هم رعایت نشده است؛ امّا ماژول جداساز حدیث، با دقّت قابل‌قبولی، این نقص را جبران کرده است.

افق آینده

در نسخه‌های بعدی، نقش تصحیف، سقط و تحریف در اسناد وارد سیستم خواهد شد. آنتولوژی‌های خبرگانی، جایگزین تزاروس‌های کلاسیک می‌شوند و گراف دانشی روات، کتب اصول و مصنّفات پیش از قرن پنجم، به‌صورت یکپارچه شکل خواهد گرفت.■