روش‌های استخراج خودکار دانش از متون حدیثی

مروری بر پژوهش‌های صورت‌گرفته

مصطفی علیمرادی

malimoradi@noornet.net

اشاره

حدیث، از منابع متنی مهم در استنباط آموزه‌های شرعی، عقیدتی و اخلاقی در اسلام است. فزونی منابع حدیثی و ضرورت استفاده از آن در شناخت احکام اسلام، استفاده از فنون خودکار استخراج دانش از متون حدیثی را ضرورت می‌بخشد. خصوصیات زبانی ویژه متون حدیثی (زبان عربی قدیم و متون روایت‌گونه)، مستلزم گردآوری و بهره‌گیری از شیوه‌های خاص پردازش زبان طبیعی شده است که در متون و زبان‌های دیگر استفاده از چنین پردازش‌های خاص نبوده است. در این مقاله، نخست استخراج خودکار دانش بر پایه فنون متن‌کاوی بیان، و پژوهش‌های انجام‌شده برای استفاده از این شیوه‌ها در استخراج دانش از متون حدیثی، بررسی می‌شود.

کلیدواژگان: متن‌کاوی، استخراج خودکار دانش، متون حدیثی، الگوریتم‌های متن‌کاوی، کشف دانش.

مقدمه

عصر اطلاعات، استفاده و ذخیره‌سازی مقدار عظیم اطلاعات متنی و چندرسانه‌ای را آسان کرده است. فراوانی اسناد رقمی در وب و دیگر پایگاه‌های دادگان محلی، در حال فزونی است. درحالی‌که میزان اطلاعات در دسترس برای استفاده پیوسته افزایشی است، توانایی برای جذب و پردازش اطلاعات متوازن با آن پیشرفت ندارد.

حوزه‌های بازیابی اطلاعات و استخراج دانش، امروزه مورد توجه پژوهشگران هوش مصنوعی و متن‌کاوی است. با ظهور پیکره‌های متنی در سال‌های اخیر، نیاز به ماژول‌های یکپارچه برای استخراج دانش در نظام‌های بازیابی اطلاعات حس می‌شود. پردازش پیکره‌های متنی بزرگ، نیازهایی را که در محدوده استخراج دانش و حوزه‌های بازیابی اطلاعات می‌گنجند، راهبری می‌کند. استخراج دانش، شاید برجسته‌ترین فنی است که اخیراً در عملیات پیش‌پردازش متن‌کاوی استفاده شده است.

استخراج دانش، عبارت از ایجاد دانش از منابع ساختاریافته (مانند پایگاه دادگان روابطی، با زبان ایکس.ام.ل) و منابع ساختارنیافته مانند: متون، اسناد و تصاویر است. دانشِ به‌دست‌آمده از نظام استخراج دانش، لازم است در یک قالب ماشین‌خوان و قابل تفسیر با ماشین باشد و نیز باید دانش را در حالتی که استنتاج را آسان می‌کند، عرضه شود. استخراج دانش از متن، عبارتی کلیدی در فناوری مفهومی است و در وب مفهومی نیز از فنون کلیدی به شمار می‌رود. (wikipedia 2017)

استخراج دانش، فرایندی پیچیده است که تشخیص ساختارهای از پیش ناشناخته و احتمالاً اطلاعات اصلی سودمند را از دادگان عظیم ممکن می‌کند. این فن، زیرمجموعه هوش مصنوعی نیز به شمار می‌رود که در آن، پژوهش‌هایی درباره زبان و متن انجام می‌شود. این پژوهش‌ها، عموماً بر پایه پیکره‌های زبانی، مجموعه‌های دادگان متنی ماشین‌خوان که با استفاده از فراداده و برچسب‌گذاری یا حاشیه‌نویسی غنی شده‌اند، نشان‌دادن تحلیل‌های ریخت‌شناختی و مانند آن، صورت می‌گیرد.

استخراج دانش، فرایندی است که در آن اطلاعات پوشیده، بالقوه سودمند، یا از پیش ناشناخته، از دادگان استخراج می‌شود. در استخراج دانش، دادگانِ فراهم‌آمده از منابع

* پژوهشگر مرکز تحقیقات کامپیوتری علوم اسلامی (نور).

الگوریتم‌های متن‌کاوی	کارکرد	نوع	توصیف
درخت تصمیم	طبقه‌بندی	نظارت‌شده	اطلاعات پیش‌بینی‌شده در قالب قوانین قابل فهم برای انسان استخراج می‌کند. این قوانین، در شکل گزاره‌های «اگر- آنگاه- وگرنه» است و تصمیم‌هایی را بیان می‌کنند که منجر به پیش‌بینی‌ها می‌شود.
الگوهای تعمیم‌یافته خطی	طبقه‌بندی و بازگشتی	نظارت‌شده	واپازش (رگرسیون) لوجستیک برای طبقه‌بندی اهداف دوگانه، و رکرسیون خطی برای اهداف مستمر اعمال می‌کند. طبقه‌بندی این الگوریتم، از حدود مطمئن برای پیش‌بینی احتمالات پشتیبانی می‌کند. رکرسیون الگوی تعمیم‌یافته خطی نیز از مرزهای مورد اعتماد برای پیش‌بینی پشتیبانی می‌کند.
حداقل طول شرح	شناسایی و رتبه‌بندی اوصاف مهم	نظارت‌شده	اصل نظری این الگو، انتخاب اطلاعات است. این الگوی «حداقل طول شرح» فرض کرده است که نمایش ساده‌تر و فشرده‌تر دادگان، بهتر و برای توضیح دادگان محتمل‌تر است.
بیز ساده	طبقه‌بندی	نظارت‌شده	با استفاده از قضیه‌های بیز که احتمال یک پیش‌بینی از شواهد موجود به دست می‌آید، همان‌گونه‌که در دادگان مشاهده شده است، پیش‌بینی‌هایی انجام می‌دهد.
ماشین بردار پشتیبان	طبقه‌بندی و رگرسیون	نظارت‌شده	نسخه‌های متمایز از این الگوریتم، از توابع هسته متفاوت برای انجام انواع گوناگون از مجموعه داده استفاده می‌کنند. هسته‌های خطی و غیر خطی (گاوسی)، در این الگوریتم، پشتیبانی می‌شوند. طبقه‌بندی در این الگوریتم، طبقات هدف را با عریض‌ترین حاشیه‌های ممکن، جدا می‌کند. رگرسیون این الگو، در پی یافتن کارکردی مستمر را می‌یابد؛ مانند بیشترین تعداد نقاط داده که درون کمترین فضا ممکن است باشد.
پیشینی	وابستگی	نظارت‌نشده	این الگو، با کشف آیتم‌های هم‌رخدادِ درون یک مجموعه به تحلیل سبد خرید می‌پردازد. شیوه پیشینی، قوانینی با بیشترین پشتیبانی را از پشتیبانی کمتر و بیشترین اعتماد را از اعتماد کمتر، کشف می‌کند.
ک ـ معانی	خوشه‌بندی	نظارت‌نشده	الگوریتم خوشه‌بندی بر پایه فاصله که دادگان را در شماری از پیش‌تعیین‌شده از خوشه‌ها بخش‌بندی می‌کند. هر خوشه، مرکز گرانشی دارد.

جدول شماره ۱: الگوریتم‌های متن‌کاوی

گوناگون، در مخزن دادگان واحد گردآوری می‌شوند. این دادگانِ ذخیره‌شده در انبار دادگان، دادگان هدف خوانده می‌شوند. استخراج دانش، فرایندی از داده‌کاوی شمرده شده است. (Canarelli 1996, 1)

داده‌کاوی در نگاه کلی، فرایندی از استخراج دانش سودمند و الگودهی به دادگان عظیم است. این کار همچنین، فرایند کشف دانش، کاوش دانش از دادگان، استخراج دانش یا داده یا تحلیل الگو خوانده شده است. داده‌کاوی، تحلیل مجموعه‌هایی از دادگان مشاهداتی برای یافتن روابط پوشیده و خلاصه‌سازی دادگان به شیوه‌ای ناب است که هم فهم‌پذیرتر و هم برای مالکان داده سودمند است. افزون بر آن، فنون داده‌کاوی، استخراج دانش از دادگان در الگوهای آماری برای مشاهده چگونگی ارتباط‌های گوناگون به یکدیگر و برای فهم بهتر پدیده‌های موجود در آنهاست. شیوه‌های متن‌کاوی، عبارت‌اند از: شبکه‌های عصبی، درخت تصمیم، تحلیل خوشه، تحلیل سبد فروشگاه و تحلیل پس‌رفت.

الگوسازی ساختاریافته درخت، فنون داده‌کاوی است که برای بخش‌بندی مجموعه دادگان در گروه‌هایی همگن به‌هم‌مرتبط به شکل مکرر به کار می‌رود، تا پیش‌بینی‌ها درباره نمونه‌های آتی را دقیق‌تر کند.

از مزیت‌های الگوریتم درخت تصمیم، توانایی در کار با ارزش‌ها و مقادیری است

داده‌کاوی در نگاه کلی، فرایندی از استخراج دانش سودمند و الگودهی به دادگان عظیم است. این کار همچنین، فرایند کشف دانش، کاوش دانش از دادگان، استخراج دانش یا داده یا تحلیل الگو خوانده شده است. داده‌کاوی، تحلیل مجموعه‌هایی از دادگان مشاهداتی برای یافتن روابط پوشیده و خلاصه‌سازی دادگان به شیوه‌ای ناب است که هم فهم‌پذیرتر و هم برای مالکان داده سودمند است. افزون بر آن، فنون داده‌کاوی، استخراج دانش از دادگان در الگوهای آماری برای مشاهده چگونگی ارتباط‌های گوناگون به یکدیگر و برای فهم بهتر پدیده‌های موجود در آنهاست. شیوه‌های متن‌کاوی، عبارت‌اند از: شبکه‌های عصبی، درخت تصمیم، تحلیل خوشه، تحلیل سبد فروشگاه و تحلیل پس‌رفت

که در متن پیدا نیست؛ اما تلاشی بسیار که برای دستیابی به آن لازم است، به مثابه مانع و اشکال این الگوریتم شمرده می‌شود. در جدول شماره ۱، به برخی از الگوریتم‌های متن‌کاوی اشاره شده است.

کارکردهای متن‌کاوی

متن‌کاوی، کارکردهای گوناگونی دارد و در حوزه‌های مختلف دانش مورد استفاده قرار می‌گیرد. در اینجا برخی از کارکردهای متن‌کاوی در حوزه دانش اطلاعات به گونه خلاصه بیان می‌شود.

- **شناخت خودکار اصطلاح‌ها:** بر پایه این کار، اصطلاح‌های خاص علوم به شکل خودکار از متون استخراج می‌شوند و می‌توان از این فن در نمایه‌سازی، ساخت اصطلاح‌نامه و هستی‌شناسی استفاده کرد. روش‌های گوناگونی برای این کار وجود دارد که عبارت‌اند از: آماری، هم‌رخدادی و یادگیری ماشینی.
- **خوشه‌بندی مدرک:** خوشه‌بندی، روشی برای گروه‌بندی موجودیت‌های مشابه است که در آن، مدارک مشابه در گروه‌هایی به نام خوشه جای می‌گیرند. خوشه‌بندی‌ها بیشتر بر پایه موضوع انجام می‌شوند و در یک نگاه، می‌توان آن را به خوشه‌بندی سلسله‌مراتبی و مسطح (خطی) دسته‌بندی کرد.
- **طبقه‌بندی خودکار مدرک:** در طبقه‌بندی خودکار مدرک، اسناد ذیل مقوله‌های از پیش‌تعیین‌شده جای می‌گیرند تا بازیابی‌شان آسان‌تر شود. طبقه‌بندی متن، عبارت از تخصیص خودکار یک یا چند سند به یکی از دسته‌های موضوعی است که از پیش تعریف شده است.
- **خلاصه‌سازی مدارک:** خلاصه‌سازی مدارک، به معنای استخراج جمله‌های پُرمفهوم از اسناد است. در این روش، فرض گرفته شده که خوانندگان، متن را در

مجموعه‌ای از اطلاعات خُردشده ملاحظه می‌کنند و این اطلاعات دارای هویتی مستقل‌اند. ازاین‌رو، در شیوه خلاصه‌سازی، متون به گزاره‌ها، اصطلاح‌ها و عبارت‌های معنادار شکسته می‌شوند و از آن میان، با استفاده از فنون رتبه‌دهی و وزن‌دهی، مهم‌ترین بخش متون استخراج و به منزله خلاصه عرضه می‌شود.

روش‌های استخراج خودکار دانش از متون حدیثی

در باور مسلمانان، سنّت پیامبر(ص) پس از قرآن، از منابع معتبر در استنتاج مفاهیم دینی و احکام شرعی است و ازاین‌رو، متون حدیثی برای مسلمانان بسیار بااهمیت و از منابع بنیادین شمرده می‌شود. با توجه به حجم بالای متون حدیثی از یک سو، و اهمیت استخراج همه مفاهیم مندرج در آن برای استنباط احکام صحیح از سوی دیگر، به‌کارگیری روش‌هایی برای آسان‌سازی و سرعت‌دادن به استخراج مفاهیم و دقت در یافتن آن، ضروری می‌نماید.

مرور پژوهش‌های انجام‌گرفته درباره استخراج دانش از متون حدیثی

مقالات گوناگونی در حوزه کشف خودکار دانش از متون حدیثی به زبان انگلیسی و فارسی موجود است و هریک بر پایه شیوه‌ای از فنون متن‌کاوی در پی تبیین این کار بوده‌اند. این پژوهش‌ها را می‌توان بر پایه الگوریتم‌های به‌کاررفته در آن، به شکل‌های ذیل دسته‌بندی کرد:

۱. الگوریتم طبقه‌بندی متن

ختام جبارا در مقاله‌ای با عنوان «کشف دانش در حدیث با استفاده از الگوریتم طبقه‌بندی متون» که در سال ۲۰۱۰م تدوین کرده، کوشیده است الگوریتمی برای کشف دانش از متون حدیثی ارائه دهد تا بر پایه آن بتواند احادیث را در طبقات از پیش تعریف‌شده دسته‌بندی کند. او در این مقاله، الگوریتم طبقه‌بندی متون را برگزیده است. این الگوریتم، متشکل از دو مرحله عمده است: یادگیری و طبقه‌بندی. مشاهده‌ها، بر پایه مجموعه‌های برگزیده از کتاب صحیح بخاری هدایت شده است که در آن، سیزده کتاب به منزله طبقاتی که این آزمایش را انجام دهند، برگزیده شده بود. در نتیجه، شیوه طبقه‌بندی که بسط ریشه خوانده می‌شد، در این پژوهش به کار برده شد تا کشف دانش از کتاب حدیث صحیح بخاری با تخصیص هر حدیث به یک طبقه از طبقاتِ از پیش تعریف‌شده صورت گیرد.

در این تحقیق، پیکره‌ها دربردارنده ۱۳۲۱ حدیث بودند. نتایج این الگوریتم، با نتایج دو شیوه مطرح‌شده الکابی و فن طبقه‌بندی مبتنی بر واژه مقایسه شد و در نتیجه، مشخص شد که الگوریتم طبقه‌بندی بسط ریشه، از دید فراخوانی برای همه طبقه‌ها بهتر از فن طبقه‌بندی مبتنی بر واژه و الگوریتم الکابی است؛ درحالی‌که طبقه‌بندی مبتنی بر وب و الکابی، از دید دقت برای دو طبقه از سیزده طبقه نتایج بهتری داشت.

فناوری استخراج موجودیت نام، برای شناسایی موجودیت‌های سودمند از مجموعه احادیث به کار می‌رود. کتب حدیثی به گونه معمول بر پایه موضوعات به ابواب گوناگون تقسیم‌شده و هریک از این ابواب نیز در موضوعات جزئی‌تر عرضه می‌شوند. هر حدیث نیز دارای شماره‌های خاص‌اند. یه اذعان حراج و همکاران در سال ۲۰۱۱ و ۲۰۱۴م برای تبدیل متون ساختارنیافته به یک متن نیمه ساختاریافته، فرایندی تعریف شده است تا موجودیت‌های مطلوب را از این متون استخراج کند؛ مانند شماره باب، عنوان باب، شماره بخش، عنوان بخش، شماره حدیث اسناد، متن، عبارت آغازین (طرف) و تاریخ. آنها الگویی به کار بردند که از مبدل حالت در شکل خودکار استفاده می‌کرد. این خودکارسازی، با مجموعه‌ای از حالت‌ها و انتقال میان این حالت‌ها نمایش داده می‌شود؛ درحالی‌که متن به یکدیگر پیوند داده شده است. این امر خودکار، تناوبی از بردارها (یعنی واژگان) را به تناوبی از الگوها (یعنی موجودات) تبدیل می‌کند. این الگو، دقت ۷۱٪، بازخوانی ۳۹٪ و امتیاز ۵۲٪ را به دست آورد. این نظام در تشخیص شمار

در باور مسلمانان، سنّت پیامبر(ص) پس از قرآن، از منابع معتبر در استنتاج مفاهیم دینی و احکام شرعی است و ازاین‌رو، متون حدیثی برای مسلمانان بسیار بااهمیت و از منابع بنیادین شمرده می‌شود. با توجه به حجم بالای متون حدیثی از یک سو، و اهمیت استخراج همه مفاهیم مندرج در آن برای استنباط احکام صحیح از سوی دیگر، به‌کارگیری روش‌هایی برای آسان‌سازی و سرعت‌دادن به استخراج مفاهیم و دقت در یافتن آن، ضروری می‌نماید

موجودیت‌ها مانند شمار: باب‌ها، بخش‌ها و حدیث خوب عمل کرده است؛ اما در تشخیص تاریخ موجودیت، به سبب اینکه در شکل البفایی نگارش یافته بود، ضعیف عمل نموده است.

در پژوهشی دیگر، Kawther A.Aldhaln و همکارانش که در قالب مقاله‌ای با عنوان «داده‌کاوی و استخراج دانش اسلامی؛ حدیث به منزله منابع دانش» که در سال ۲۰۱۰م انجام داده‌اند، با شیوه‌ای مبتنی بر طبقه‌بندی یادگیری نظارت‌نشده و طبقه‌بندی یادگیری نظارت‌شده، به طبقه‌بندی حدیث پرداختند.

کابی و همکاران در سال ۲۰۰۵م در پژوهشی اثبات کرده‌اند که طبقه‌بندی حدیث باید از طریق شیوه‌ای ساده با دقت قابل اطمینان انجام شود. ایشان احادیث عربی را بر پایه محاسبه بسامد اصطلاح‌ها، به هشت باب از ابواب کتاب صحیح بخاری تقسیم کرده‌اند. آنها در این پژوهش، از اسناد و واژگان خنثا چشم‌پوشی نموده و با استفاده از نظام ریشه‌ساز، واژگان را به ریشه خود ارجاع داده‌اند. در کتاب **صحیح بخاری**، احادیث مشابه ممکن است در بیش از یک فصل آمده باشد، ازاین‌رو، در نظامی که کابی و همکاران تهیه کرده‌اند، دو موضوع با بیشترین رتبه نشان داده شده است. به منظور انجام وزن‌دهی به واژگان، از شیوه تی.اف ـ ای.دی.اف استفاده شده و با ۱۲۰ حدیث یادگیری انجام شده است. آزمون نظام طراحی‌شده نیز با ۸۰ حدیث، ۸۳٫۲٪ دقت نشان داده است.

حراج در سال ۲۰۱۱م مطالعاتی درباره ارزیابی شیوه‌های ریشه‌سازی برای طبقه‌بندی متون حدیثی انجام داد. این مطالعه، ریشه‌سازی بر پایه لغت‌نامه، ریشه‌سازی مبتنی بر ریشه و ریشه‌سازی سبک به مثابه مرحله کاهنده خصوصیات بررسی کرده است. ریشه‌ساز بر پایه لغت‌نامه، بالاترین دقت را در شیوه طبقه‌بندی‌کننده شبکه عصبی مصنوعی، در قیاس دیگر شیوه‌ها به دست آورد. ریشه‌ساز سبک، بالاترین دقت را در شیوه طبقه‌بندی‌کننده ماشین پشتیبان در قیاس با دیگر شیوه‌ها به دست آورد.

Kawther A.Aldhaln و همکاران در سال ۲۰۱۲م احادیث را به چهار دسته صحیح، حسن، ضعیف و مردود تقسیم کردند. به منظور ساخت مجموعه دادگان یادگیری و آزمون، ۹۹۹ حدیث از سه کتاب صحیح بخاری، جامع الترمزی و سلسله الاحادیث الضعیفة والموضوعة واثرها السیئ فی الأمة انتخاب شد. از آنجا که مجموعه دادگان متشکل از کتب گوناگون بود، پیش‌پردازش دادگان برای کاستن از حشویات و ایجاد سبک اسناد حدیث همگن انجام شد. خصوصیات احادیث بر پایه پنج اصل در علم حدیث تعیین شد: راویان گوناگون برای خلوصشان، راویان گوناگون برای راستی‌شان، نبود وفقه در اسناد، نبود عبارت‌های ناهنجار در متن حدیث و نبود نقص در متن حدیث. در طبقه‌بندی ایشان، دو شیوه گوناگون به کار برده شد: درخت تصمیم و بیز ساده؛ هرچند طبقه‌بندی‌کننده بیز ساده با ۹۷٫۶٪ فراخوانی و ۹۷٫۵۹۷٪ دقت، نتایج بهتری در پی داشت.

نجیب در سال ۲۰۱۴م رهیافت طبقه‌بندی جدید مطرح کرد که میان احادیث صحیح و ضغیف تمایز ایجاد می‌کرد. آنها طبقه‌بندی‌کننده وابسته که قوانین طبقه‌بندی و وابستگی را به هم الحاق می‌کرد، به کار بردند. طبقه‌بندی مبتنی بر وابستگی، به هدف کشف روابط میان خصوصیات برای تعیین مجموعه از قوانین طبقه‌بندی ساخته می‌شود. کارکرد این شیوه در حوزه حدیث آزموده شده؛ اما میزان روشنی از دقت در این شیوه گزارش نشده است.

حراج و همکاران در سال ۲۰۰۹م دو تجربه مختلف را درباره مجموعه دادگان یکسان توصیف کردند:

فعالیت‌های متن‌کاوی در مرکز نور ابعاد گوناگونی داشته است و متخصصان این مرکز، در این حوزه، نظام‌های مختلف متبنی بر متن‌کاوی را طراحی کرده‌اند. این فعالیت‌ها در زمینه حدیث، شامل نمونه‌هایی همچون نظام کشف روایات مشابه است

بازیابی اطلاعات با استفاده از بخش‌بندی و بدون بخش‌بندی. برای بازیابی اطلاعات بدون بخش‌بندی، ریشه‌ها از متن پرس‌وجوی انجام‌شده استخراج شد تا وزنشان بر پایه تی.اف ـ ای.دی.اف به دست آید. پس از آن، نمایه‌سازی اصطلاح‌های مرتبط حدیثی و شمارش وزن‌های مرتبط با معدلات خاصشان تعیین گردید. سرانجام، نظام مجموعه از حدیث را عرضه کرد: مجموعه مرتبط و مجموعه نامرتبط. در مرحله مشاهده، میانگین فراخوانی ۵۴٪ و میانگین دقت ۴۱٪ بود.

بلال و محسن، در سال ۲۰۱۲م نظام خبره مبتنی بر قوانین توزیع‌شده را معرفی کردند که از آن برای طبقه‌بندی احادیث به حدیث صحیح و غیرصحیح استفاده کردند. این نظام خبره، «پرس‌وجوی محدث» نام‌گذاری شد که پنج ماژول اصلی آن عبارت‌اند از:

۱. موتور استنباط: مجموعه‌ای از گزاره‌های «اگر، آنگاه، وگرنه» که قوانین را می‌سازند؛

۲. پایگاه دانش: جدول تصمیم دوگانه برای نمایش دانش؛

۳. تجزیه‌کننده و استخراج‌کننده رخداد: که برای تجزیه و پرس‌وجوی کاربران و برای استخراج دانش مرتبط با در نظرگرفتن پرس‌وجو عرضه شده است؛

۴. امکانات توضیح و تبیین: برای به دست دادن جزئیات درباره چگونگی و چرایی استخراج نتایج به کاربران؛

۵. پایگاه دادگان. نظام محدث، به مثابه معماری سرویس‌گرا مبتنی بر نظام خبره ابری از طریق وب در دسترس است. نتایج نظام خبره، در مقاله عرضه‌شده از سوی ایشان، بررسی نشده است.

جبارا در سال ۲۰۱۰م نظام متن‌کاوی برای بازیابی طبقه حدیث در پاسخ به پرس‌وجوها عرضه کرد. در این نظام، ۱۳۲۱ حدیث از کتاب صحیح بخاری که در سیزده گروه (ایمان، دانش، عبادت، دعوت به نماز، کسوف، صدقات، حالات خوب، روزه، طب، تغذیه، حج، شکایت و فضایل پیامبر) تقسیم شده بود، برای یادگیری و آزمودن انتخاب شد. این آزمایش‌ها، شامل سه مرحله بود: مرحله پیش‌پردازش که متشکل از حذف اسناد، رمزگذاری، حذف نشانه‌گذاری و علائم تفکیک‌کننده، حذف واژگان خنثا و ریشه‌سازی بود. در مرحله دوم یادگیری که در آن خصوصیات ماتریکس با استفاده از روش تی.اف ـ آی.دی.اف ساخته شده است، اعمال شد. مرحله سوم که در آن نتایج یادگیری که از مرحله پیش به دست آمده بود، برای طبقه‌بندی به کار برده شد. همچنین، خصوصیات پرس‌وجو، محاسبات و توسعه پرس‌وجوی انجام شده در مرحله سوم

را وزن‌دهی کرد. نتیجه آزمایش‌ها، ۴۵٪ دقت و ۴۹٪ فراخوانی بود.

در پژوهشی دیگر که از سوی خمسین و همکاران در سال ۲۰۱۴م انجام شده، بر اهمیت نظام تصدیق خودکار برای قرآن و حدیث به منظور مبارزه با شکل‌های جعلی قرآن و احادیث دروغین در فضای مجازی، تأکید گردیده است.

۲. الگوریتم درخت تصمیم

حراج و همکاران در سال ۲۰۰۹م با استفاده از شیوه درخت تصمیم، طبقه‌بندی‌کننده‌ای طراحی کرده‌اند که با به‌کارگیری۴۵۳ حدیث گروه‌بندی‌شده در ۱۴ مقوله از کتاب دائرة المعارف نبوی احادیث را طبقه‌بندی می‌کرد. مرحله پیش‌پردازش در پژوهش آنها، از تبدیل اسناد به متون اصلی، حذف واژگان خنثا و ریشه‌سازی تشکیل می‌شد. پس از پیش‌پردازش، برداری متشکل از همه اصطلاح‌های موجود در متون حدیثی ساخته شد. پس از آن، ابعاد بردار بر پایه معیارهای خاص و وزنی که برای هر بعد با استفاده از بسامد در نظر گرفته شده بود، به ۱۹۳۸ بعد کاهش یافت. در مرحله آزمون، ۳۸ درصد فراخوانی و ۴۷ درصد دقت و ۴۰ درصد نمره منفی به دست آورد. این نتایج و دسته‌بندی‌های نادرست، به دلیل ماهیت و خصوصیات اسناد حدیثی دانسته شده است.

۳. الگوریتم فازی

قاضی‌زاده و همکارانش در سال ۲۰۰۵م با استفاده از مجموعه قوانین و دیدگاه‌های متخصصان، نظام خبره فازی را طراحی کرده‌اند. در این پژوهش، به منظور استنتاج نتایج، دو ماشین استنتاج طراحی شد. نخستین ماشین، رتبه هریک از راویان را تولید می‌کند و نتیجه استنتاجش را به ماشین استنتاج دوم منتقل می‌کند. محصول ماشین استنتاج دوم، ارزیابی اعتبار حدیث است. نظام طراحی‌شده قاضی‌زاده و همکاران، با استفاده از مجموعه احادیث کتاب کافی آزموده شد تا احادیث را در چهار دسته ناشناخته، ضعیف، خوب و

از جمله کاربردهای «سامانه تشخیص ماشینی روایات مشابه» عبارت‌اند از: تشخیص زیرمجموعه بودن احادیث، پیدا کردن متن و ترجمه، شناسایی احادیث غیرتکراری، شناسایی تعابیر مختلف اسناد، شناسایی کتب مفقوده، موضوعات مشابه، میزان احادیث مشابه بین معصومان(ع)، شناسایی معصوم به عنوان راوی حدیث.

البته فعالیت‌های مرکز در این حوزه، به همین سامانه خلاصه نمی‌شود و زمینه‌های لازم برای استفاده از فنون متن‌کاوی در متون حدیثی ایجاد شده است. فعالیت‌های دیگری نیز در مرکز در حال انجام است که هنوز به شکل محصول مستقل آماده ارائه نشده و یا ضمن محصولات دیگر، در حال عرضه است

قابل اطمینان، جای دهد. نتیجه آزمون، ۹۴٪ دقت را نشان داد.

٤. الگوریتم شبکه عصبی مصنوعی

شبکه‌های عصبی مصنوعی هم از شیوه‌های به‌کار گرفته‌شده برای طبقه‌بندی احادیث است. حراج و قواسمه در سال ۲۰۰۹م از این شیوه برای طبقه‌بندی حدیث استفاده کردند. در رهیافت ایشان، افزون بر پیش‌پردازش متن، در گام نخست از شیوه تجزیه ارزش منحصربه‌فرد استفاده شده است که فرایندی مؤثر در پاکسازی داده به شمار می‌رود. در این پژوهش، ۷۳۹ واژه منحصربه‌فرد وجود دارد که هر ویژگی، به یک واژه ارجاع شده است. آنها از دائرة العمارف نبوی که شامل ۴۵۳ سند بوده و به ۱۴ دسته (ایمان، قرآن دانش، جرایم، جهاد، رفتار خوب، نسل‌های گذشته، زندگی‌نامه، قضاوت، عبادت، رفتار، غذا، لباس و حالت‌های شخصی) تقسیم شده است، استفاده کردند. فراخوانی و دقت به‌دست‌آمده برای پیش‌بینی دسته‌بندی حدیث، حدود ۸۸٪ بود.

۵. چندالگوریتمی

الکابی در سال ۲۰۱۰م، چهار الگوریتم را برای دسته‌بندی احادیث آزمود که عبارت بودند از: بیز ساده، الگوریتم راچیو، ک ـ نزدیک‌ترین همسایه و ماشین بردار پشتیبان. برای ارزیابی بسامد نسبی هر واژه در اسناد، از شیوه تی. اف–آی.دی.اف استفاده شد. برای یادگیری ماشینی، ۱۳۵۰ حدیث استفاده شد و ۱۵۰ حدیث برای آزمودن دقت شیوه‌های طبقه‌بندی به‌کار گرفته شد. میانگین فراخوانی همه این روش‌ها، ۱۰۰٪ بود؛ اما دقت الگوریتم راچیو ۶۷,۱۱ ٪ بیز ساده ۶۶,۵۵ ٪، ک ـ نزدیک‌ترین همسایه ۶۶,۵۵ ٪ و ماشین بردار پشتیبان ۶۳,۳۶٪ بود. ازاین‌رو، الگوریتم راچیو، احادیث را با بیشترین فراخوانی و بالاترین دقت طبقه‌بندی کرده است.

٦. الگوریتم فضای بردار

حراج و حمدی شریف در سال ۲۰۰۷م فهرستی از احادیث مرتبط که بر پایه مشابهت ذخیره شده بودند، عرضه کردند. این فهرست، بر پایه الگوی فضای بردار عمل می‌کرد. گام نخست ریشه‌سازی ریخت‌شناختی حدیث، بر پایه یک لغت‌نامه ریشه بود. پس از پیش‌پردازش، وزن‌دهی و نمایه‌سازی با استفاده از روش تی.اف ـ آی.دی.ال انجام شد. سپس، مشابهت میان پرس‌وجوی انجام گردید و حدیث با کمک فن اندازه‌گیری کسینوس انجام شد. حدیث بازیابی‌شده در دو نوبت انجام شد: همه اصطلاح‌ها از پنج سند مرتبط نخست با استفاده از وزن ارتباط به پرس‌وجوی نخستین منظم شدند؛ ده اصطلاح برگزیده نخست از اسناد بازیابی‌شده با پرس‌وجوی اصلی مرتبط شدند تا یک پرس‌وجوی قوی‌تر ساخته شود؛ درحالی‌که اصطلاح جدید، وزن کمتری از اصطلاح‌های پرس‌وجوشده اوّلیه دارد. این پژوهش با ۶۰ حدیث، ۶۶ درصد دقت و ۸۰ درصد فراخوانی به دست آورده است.

۷. پژوهش‌های دیگر

پژوهشگرانی نیز به بررسی احادیث موجود در فضای وب پرداختند. کریم و حزمی، در سال ۲۰۰۵م تحلیل کمی داده با استفاده از مصاحبه با دانشجویان کارشناسی ارشد مالزیایی به منظور ارزیابی اطلاعات درباره حدیث را در اینترنت انجام دادند. نتیجه تحلیل آنها، این بود که تقریباً همه شرکت‌کنندگان، اینترنت را به مثابه منبع حدیثی مناسب ملاحظه کردند؛ هرچند خطر استفاده از احادیث نادرست نیز وجود دارد.

شنطاوی و همکاران در سال ۲۰۱۲م مشاهداتی را تبیین کردند که دربردارنده دو گام عمده بود: بازیابی احادیث از صفحات وب و تشخیص درستی احادیث بازیابی‌شده. آنها از پایگاه دادگانی که شیخ الألبانی ساخته بود و متشکل از ۱۷ هزار متن حدیثی و درجه صحیح بودن آنها بود، استفاده کردند. آنها پایگاه دادگان را رمزگذاری نموده و واژگان خنثا و علائم واکه را حذف کردند. همچنین، نمایه‌های موضعی که شامل بیش از ۵۶۰۰۰ اصطلاح بود، ساخته شد. به منظور استخراج متون حدیثی از صفحات وب، یک برنامه پاک‌کننده رمزهای اج.تی.ام.ال به زبان جاوا طراحی شد تا همه رمزگذاری‌های مبتنی بر اچ.تی.ام.ال را حذف کند. سپس، چهار واژه

مجاور از صفحه وب با نمایه‌های موضوعی حدیثی مقایسه شد تا متن حدیثی تشخیص داده شود. هنگامی که همه متون حدیثی استخراج شد، هریک در پایگاه دادگان یافته شد تا درجه صحتش تعیین شود. در این پژوهش، از پنج صفحه وب که دربردارنده متون حدیث بودند، به شکل تصادفی انتخاب شدند که ۷۶٫۱٪ دقت و ۴۲٫۱٪ فراخوانی به دست آمد.

استخراج دانش از متون حدیثی در مرکز نور

فعالیت‌های متن‌کاوی در مرکز نور ابعاد گوناگونی داشته است و متخصصان این مرکز، در این حوزه نظام‌های مختلف متبنی بر متن‌کاوی را طراحی کرده‌اند. این فعالیت‌ها در حوزه حدیث، شامل نمونه‌هایی همچون نظام کشف روایات مشابه است.

شناسایی میزان شباهت یک متن با حجم انبوهی از متون دیگر در متون حدیث، با استفاده از این نظام انجام می‌شود. این سامانه، به شکل مبتنی بر وب ارائه شده است و در نشانی اینترنتی: http://textmining.noorsoft.org/FA/SimilarHadith در دسترس است. از جمله کاربردهای «سامانه تشخیص ماشینی روایات مشابه» عبارت‌اند از: تشخیص زیرمجموعه بودن احادیث، پیدا کردن متن و ترجمه، شناسایی احادیث غیرتکراری، شناسایی تعابیر مختلف اسناد، شناسایی کتب مفقوده، موضوعات مشابه، میزان احادیث مشابه بین معصومان(ع)، شناسایی معصوم به عنوان راوی حدیث.

البته فعالیت‌های مرکز در این حوزه، به همین سامانه خلاصه نمی‌شود و زمینه‌های لازم برای استفاده از فنون متن‌کاوی در متون حدیثی ایجاد شده است. فعالیت‌های دیگری نیز در مرکز در حال انجام است که هنوز به شکل محصول مستقل آماده ارائه نشده و یا ضمن محصولات دیگر، در حال عرضه است.

منابع:

1. Canarelli, Patrick .1996 .*Knowledge extraction from data using genetic algorithms* .Seville :European Commission , Joint Research Centre.

2. wikipedia .2017 .*Knowledge extraction* 14 .September دستیابی در . November 15, 2017 .https://en.wikipedia.org/wiki/Knowledge_extraction.

3. Al Kharashi IA, Al Sughaiyer IA (2002) Rule merging in a rule-based Arabic stemmer. In: Proceedings of the 19th international conference on computational linguistics, vol 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1–7. doi: 10.3115/1072228.1072265.

4. Aldhaln K, Zeki A, Zeki A (2010) Datamining and Islamic knowledge extraction: alhadith as a knowledge resources. In: Proceeding 3rd international conference on ICT4M, Jakarta, Indonesia, pp 21–25. Retrieved from: http://irep.iium.edu.my/17123/1/WA_17123_AKRAM_Datamining_and_Islamic_

KnowledgeExtraction.pdf.

5. Aldhaln K, Zeki A, Zeki A (2012) Knowledge extraction in hadith using data mining technique. Int J Inf Technol Comput Sci 2:13–21. Retrieved from: http://www.ijitcs.com/2ndicekmt/Kawther+AAldhaln.php.

6. Aldhaln K, Zeki A, Zeki A, Alreshidi H (2012a) Improving knowledge extraction of Hadith classifier using decision tree algorithm. In: 2012 international conference on information retrieval knowledge management, pp 148–152. doi: 10.1109/InfRKM.2012.6205024.

7. Aldhaln K, Zeki A, Zeki A, Alreshidi H (2012b) Novel mechanism to improve Hadith classifier performance. In: 2012 international conference on advanced computer science applications and technologies (ACSAT),pp 512–517. doi: 10.1109/ACSAT.2012.93.

8. Al-Kabi MN, Al-Sinjilawi SI (2007) A comparative study of the efficiency of different measures to classify Arabic text. Univ Sharjah J Pure Appl Sci 4(2):13–26.

9. Alrazou HM (2008) Data mining application on the Islamic knowledge resource. Alukah. Retrieved from http://www.alukah.net/culture/0/3123.

10. Al-tarawneh R, Hamatta HSA, Muiadi H (2014) Novel approach for Arabic spell-checker: based on radix search tree. Int J Comput Appl 95(7): 1–5.

11. Bilal K, Mohsin S (2012) Muhadith: a cloud based distributed expert system for classification of ahadith. In: Proceedings of the 2012 10th international conference on frontiers of information technology, IEEE Computer Society, Washington, DC, USA, pp 73–78. doi: 10.1109/FIT.2012.22.■